PROCEEDINGS of the ICCS-X

Tenth Islamic Countries Conference on Statistical Sciences

Volume II

STATISTICS FOR DEVELOPMENT AND GOOD GOVERNANCE

الإحصاء من أجل التنمية و الحكم الرشيد

Editors
Zeinab Amin and Ali S. Hadi

The American University in Cairo

Proceedings of the ICCS-X

Tenth Islamic Countries Conference on Statistical Sciences

Statistics for Development and Good Governance

Volume II

Editors

Zeinab Amin and Ali S. Hadi

Department of Mathematics and Actuarial Science

The American University in Cairo

The Islamic Countries Society of Statistical Sciences
Plot No. 44-A, Civic Centre,
Liaqat Chowk,
Sabzazar Scheme,
Multan Road
Lahore, Pakistan

Telephones: +92-42-37490670

E-mail: secretary@isoss.net

ISBN: 978-977-416-365-8

Printed in Cairo, Egypt

Mir Maswood Ali (1929 - 2009)

DEDICATION

The Tenth Islamic Countries Conference on Statistical Sciences, including this Proceedings are dedicated to the memory of Mir Maswood Ali (photo above), Professor Emeritus of Statistics at the University of Western Ontario. The following is an Obituary, written by his brother Mir Masoom Ali, George and Frances Ball Distinguished Professor Emeritus of Statistics, Ball State University:

Mir Maswood Ali, 80, Professor of Statistics Emeritus, University of Western Ontario and a brilliant statistician of Bangladeshi origin, died August 18, 2009 in London, Ontario, Canada due to pulmonary complications. It is my great honor and privilege to write this obituary for my older brother who was very dear to me and who had tremendous influence on my career.

Ali received his B.Sc. degree in Mathematics in 1948 and his M.Sc. degree in Statistics in 1950 both from the University of Dhaka. He belonged to the first batch of graduate students in statistics and had obtained first class and secured the highest mark for which he was awarded a gold medal. He served as Lecturer in the Department of Statistics at Dhaka University from 1950 to 1952. He then worked from 1952 to 1957 as an Actuarial Assistant at Norwich Union Life and Canada Life. In 1958, he obtained a second Master's degree in Actuarial Science at the University of Michigan and worked there as a Teaching Fellow until 1959. He then went to the University of Toronto where he obtained his Ph.D. degree in Statistics in 1961 under the supervision of D. A. S Fraser after merely two years of studies. He then joined the Mathematics Department at the University of Western Ontario (UWO) in London, Ontario, Canada as assistant professor in 1961. He was the first faculty member in statistics in the department and was quickly promoted to the rank of associate professor in 1963 and to full professor in 1966 and he remained there until his retirement in 1994 when he was named Professor Emeritus. Ali had developed the graduate and undergraduate programs in statistics in his department and he was instrumental in the creation of a separate Department of Statistics and Actuarial Sciences at UWO.

He supervised 15 Ph.D. students, a number of whom are now well-known statisticians and 40 Master's theses. He published in leading statistical journals such as the Annals of Mathematical Statistics, the Journal of the Royal Statistical Society, the Journal of Multivariate Analysis, the Pacific Journal of Mathematics, and Biometrika, to name a few. His research interests encompass many areas of Statistics and Mathematics, including order statistics, distribution theory, characterizations, spherically symmetric and elliptically contoured distributions, multivariate statistics, and *n*-dimensional geometry and his two highly rated papers are in geometry which appeared in the Pacific Journal of Mathematics.

He was a man of strong principle. He was also a very decent and humble man who never sought recognition for anything that he did or achieved. He was a dedicated family man and he devoted lot of his time to his own family. He left behind his loving wife of 47 years Surayia, and eight grown children, Rayhan, Yasmin, Selina, Sharmeene, Sadek, Nasreen, Ayesha, and Adnan, and seven grandchildren. His youngest daughter Ayesha followed his father's footsteps and now teaches statistics at the University of Guelph in Canada.

Mir Maswood Ali was my immediate older brother and it was due to his influence that I got into statistics as a student in 1953. He was a great mentor, a great teacher and a friend and he was all that I wanted to be in life. I will miss him dearly.

In loving memory of my brother,
Mir Masoom Ali
George and Frances Ball Distinguished Professor Emeritus
Ball State University, Muncie, Indiana, USA
Excerpted from IMS Bulletin, Vol. 38, Issue 9, 2009

CONTENTS

VOLUME I

VOLUME II

PREFACE

The Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X) was held during the period December 20-23, 2009 at the brand new campus of The American University in Cairo (AUC), in New Cairo, Egypt. The ICCS-X was organized by the Islamic Countries Society of Statistical Sciences (ISOSS) and cosponsored by AUC and the Egyptian Cabinet Information and Decision Support Center (IDSC).

The collaboration between a governmental organization, represented in IDSC, and a private non-for-profit university, represented in AUC, in sponsoring such an international conference has proven to be a very effective and mutually beneficial joint effort. The conference, which brought together researchers and practitioners in statistical sciences from 32 countries all over the world, was open to all people interested in the development of statistics and its applications regardless of affiliation, origin, nationality, gender or religion.

The theme of the ICCS-X was *Statistics for Development and Good Governance*. As can be seen in this Proceedings, three Discussion Panels "*Public Opinion Polling and Good Governance*," by Prof. Magued Osman, IDSC, "*Measuring the Unmeasurable*," by Dr. Anis Yusoff, National University of Malaysia, and "*Indicators and Politics*," by Prof. Ali S. Hadi, AUC, have been devoted entirely to this theme. Other papers dealt with various broad topics in statistics theory and its applications. As a result, ICCS-X has attracted a distinguished team of speakers giving more than 190 presentations.

These proceedings do not contain all articles that have been presented at the conference. Only articles that have undergone and passed peer reviews are included. The reviews have taken into consideration both the quality of the paper and the quality of the presentation at the conference. These Proceedings contain 29 abstracts and 85 complete papers. The papers were arranged alphabetically according to the first author. Due to the large size, these proceedings are split into two volumes; Volume I and Volume II.

Organizing a conference requires a lot of effort by many people, collaboration, coordination, and paying attention to very small details. We would like to thank all organizers and participants of the conference. We are particularly grateful to Prof. Jef Teugels, Catholic University of Leuven, Belgium and the current President of the International Statistical Institute for giving the opening keynote talk despite his very busy schedule. We are also thankful to Prof. Kaye E. Basford, University of Queensland, Australia, Prof. Jim Berger, Duke University, USA and former President of the Institute of Mathematical Statistics, and Prof. Edward J. Wegman, George Mason University, USA for giving the other three keynote talks. In addition to the four keynote talks and the three panel discussions, the program included nine invited sessions and 20 contributed sessions. According to feedback from participants, the conference was a great success.

We are also grateful for the following referees who devoted the time and effort to review these articles: Dr. Mina Abdel Malek, Dr. Maged George, Dr. Mohamed Gharib, Dr. Ramadan Hamed, Dr. Mohamed Ismail, Dr. Hafiz Khan, Dr. Mohamed Mahmoud, Dr. Nadia Makary, Dr. Amani Moussa, Dr. Abdel Nasser Saad, Dr. Kamal Selim, Dr. Tarek Selim, Dr. Zeinab Selim, and Dr. Mark Werner.

The Conference was organized by three Committees: the Scientific and Program Committee (Chair: Ali S. Hadi and Co-Chair: Zeinab Amin), the International Organizing Committee (Chair: Shahjahan Khan), and the Local Organizing Committee (Chair: Magued Osman and Co-

Chair: Zeinab Amin). The members of these committees are given on page xv. Each of these committees has worked tirelessly for the organization of this conference. We are indebted to each and every one of them. We have also benefited from the contributions of the ISOSS Headquarters and in particular the President of ISOSS Prof. Shahjahan Khan. Prof. Mohamed Ibrahim has generously shared with us his valuable experience in the organization of the ICCS-IX Conference that was held in Malaysia in 2007.

Prof. Wafik Younan helped in putting together the Local Organizing Committee (LOC), which consists of members from several Egyptian universities and government agencies including Ain Shams University, Al-Azhar University, The American University in Cairo, Cairo University, CAPMAS, Helwan University, and IDSC. Over the 15 months prior to the conference, the LOC has held monthly meetings at Cairo University's Institute of Statistical Studies and Research, where Dr. Amani Moussa and Dr. Mahmoud Riyad were the primary hosts.

Dr. Wafik Younan also served as the Treasurer. We are very grateful for the following organizations for their financial and other support: AUC, the Egyptian Ministry of Tourism, IDSC, the Islamic Development Bank. Mr. Amr Agamawi of IDSC was instrumental in fund raising and administrative activities and has so ably taken care of various logistics and attention to details. Dr. Mostafa Abou El-Neil, Mr. Waleed Gadow, and the multimedia team of IDSC, were responsible of the design and printing of various publications including posters, brochure, and the Book of Abstracts. Eng. Medhat El Bakry, Eng. Ibrahim Hamdy, Eng. Ahmed Khalifa, the information system and communication team of IDSC, and Eng. Mai Farouk, of AUC, were responsible of maintaining and updating the website of the conference. Lamyaa Mohamed Sayed prepared the list of contributors. Finally, the staff at the Office of AUC's Vice Provost (Samah Abdel-Geleel, Basma Al-Maabady, Sawsan Mardini, Dahlia Saad, and Nancy Wadie) helped with the correspondence and various other organizational details. We apologize if we left out some of the people who have provided us with help. This omission is, of course, not intentional.

Zeinab Amin and Ali S. Hadi, Joint Editors
Cairo, Egypt
July 2010

ICCS-X Organizing Committees

Scientific Program Committee (SPC)
Ahmed, Munir - Pakistan
Ahmed, Nasar U - USA
Ahmed, S Ejaz - Canada
Ahsanullah, Mohammad - USA
Aly, Emad Eldin - Kuwiat
AlZayed, Abdulhamid – KSA
Amin, Zeinab - Egypt (Co-Chair)
Arnold, Barry C. - USA
Castillo, Maria-Carmen - Spain
Djauhari, Maman - Indonesia
ElShaarawi, Abdulhameed - Canada
Ghosh, Malay - USA
Hadi, Ali S - Egypt (Chair)
Horova, Ivan - Czech Republic
Ibrahim, Noor Akma - Malaysia
Imon, Rahmatullah - USA
Khan, Shahjahan - Australia
Mengersen, Kerrie - Australia
Mian, Mohammad Hanif - Pakistan
Nyquist, Hans - Sweden
Osman, Magued - Egypt
Parsian, Ahmad - Iran
Provost, Serge B.- Canada
Puntanen, Simo - Finland
Raqab, Mohammad - Jordan
Rezakhak, Syed - Iran
Sinha, Bikas K - India
Yanagawa, Takashi - Japan

International Organizing Committee (IOC)
Abuammoh, Abdulrahman - KSA
Ageel, Mohammed I. - KSA
Ahmad, Munir - Pakistan
Ahmed, Kazi Saleh - Bangladesh
Ahmed, Nasar U - USA
Ahsanullah, Abd Mohammad - USA
Ahsanullah, Mohammad - USA
AlBiyyat, Hilal - Jordan
Al-Awati, Shafiqa - Kuwait
Al-Saleh, Mohammad - Qatar
Ali, Abdunnabi - Libya
Amin, Zeinab - Egypt
Basci, Sidika - Turkey
Bellout, Djamel - UAE
Chattopadhaya, Asis K - India
Djauhari, Maman - Indonesia
El-Bassouni, M Yahya - UAE
El-Shaarawi, Abdulhameed - Canada
Erçetin, Sefika Sule – Turkey
Grover, Gurpirit - India
Gupta, Pushpa L. - USA
Gupta, R.C. - USA
Hadi, Ali S - Egypt/USA
Hasan, Baktiar - Belgium
Hussain, Tayyab - Libya
Kabir, M - Bangladesh
Khan, Abdul Hamid - India
Khan, Bashir U, Canada
Khan, Hafiz Abdullah - UK
Khan, Shahjahan - Australia (Chair)
King, Max - Australia
Mian, Abul Basher - Bangladesh
Mian, Mohammad Hanif - Pakistan
Mohamed, Ibrahim - Malaysia
Ojikutu, Rasheed Kola - Nigeria
Osman, Magued - Egypt
Provost, Serge B - Canada
Rahman, Mohammad Shafiqur - Oman
Rejali, Ali - Iran
Rezakhah, Syed - Iran
Saefuddin, Asep - Indonesia
Samb, LO Gane - Senegal
Sinha, Bimal K. - USA
Soleiman, Hana - UAE
Srivastava, Muno - Canada
Srivastava, R.C. - USA
Sufian, Abu - Bahrain
Wegman, Edward - USA
Younan, Wafik – Egypt

Local Organizing Committee (LOC)
Abou El-Ela, Esmat - Al Azhar University

Abou El-Kassem, Essam - Helwan University
Amin, Zeinab (Co-Chair) - AUC
El-Agamawi, Amr - IDSC
El-Batrawy, Rawya - CAPMAS
El-Sheneity, Sahar - Cairo University
El-Tawila, Sahar - IDSC
Farouk, Mai - AUC
George, Maged - AUC
Gharib, Mohamed - Ain Shams University
Hadi, Ali S - AUC
Hamed, Ramadan - Cairo University
Ismail, Mohamed - Cairo University
Magdy, Dina - Cairo University
Mahmoud, Mohamed - Ain Shams University
Mousa, Amani - ISSR
Nour Eldien, Mohamed - ISSR
Osman, Magued (Chair) - IDSC
Riad, Mahmoud - ISSR
Saad, Abdel Naser – Cairo University
Werner, Mark - AUC
Younan, Wafik (Treasurer) - AUC
Zaky, Hassan - Cairo University

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 572–582.

THE FITTING OF BINNED AND CONDITIONAL INCOMPLETE MIXTURE DATA

Yousef M. Emhemmed and Wisame H. Elbouishi
Statistics Dept, Faculty of Science, El-Fateh University, Libya.
E-mail: emhemmedy@yahoo.co.uk

ABSTRACT

Incomplete data analysis covers a wide variety of problems that are often seen in practice, one such example is the fitting of binned data where each observation is assumed to have been raised from one of k different groups. As each of the data units and their sources (or, the source of at least some units) is being unobservable, this can be treated as a missing data problem. Finite mixture distributions are typically used to model this sort of multi-source data, see (Little and Rubin, 2002). The missing of data leaves the favorite likelihood estimation with no closed form solution. The Expectation-Maximization (EM) algorithm proved to be one of the most convenient and flexible tool to provide the ML estimates. This paper deals with two different types of incomplete mixture data, the binned (grouped) and conditional (grouped with some extra sub-frequencies) normal mixture data, from both theoretical and application point of view. The additional information in the form of sub frequencies, may lead intuitively to a further improvement of the performance of the EM procedure. The procedure has been applied to a simulated data set.

Keywords: EM algorithm, Binned Data, Conditional Data, Incomplete-data and Mixture Distribution.

1. INTRODUCTION

Binned and conditional data arise frequently in a wide variety of application settings since many measuring instruments produce quantized data. For both binning and conditional data, one can think of the original "raw" measurements as being masked by the binning processes and hence it is natural to think of this problem as one involving incomplete-data.

The EM procedure is an obvious candidate for finite-mixture model fitting. The theory of using EM for fitting maximum likelihood finite-mixture models to univariate binned data was developed in McLachlan and Jones (1988). The problem in somewhat simpler form was addressed earlier by Dempster et al (1977) when the EM algorithm was originally introduced. The EM algorithm is a very general iterative algorithm for parameter estimation by maximum likelihood and it formalizes an intuitive idea for obtaining parameter estimates when some of the data are incomplete. The EM algorithm is a hill-climbing approach, thus it can only be guaranteed to reach a local maxima. To reach the global maxima, when there are multiple maximas, clearly depends on the starting values. When there are multiple local maximas, it is often hard to identify a reasonable starting value. Many strategies have been set down for selecting good initial values. Our strategy to compare the performances of binned and

conditional EM procedures is to try random set of initial values which is deliberately selected to be far from the actual parameter values. For more recent applications, see Juan Du (2002), Myung (2003), and J. Andrew (2005). This paper is aiming to address and demonstrate the fitting of finite mixture distributions to binned and conditional data sets by the method of maximum likelihood using the iterative EM technique.

2. FINITE MIXTURE DISTRIBUTION

The finite mixture models are being increasingly used to model the distributions of wide variety of random phenomena, where each observation is believed to belong to one of several different types, each of which has its own distribution. A natural way is to assume that the data are drawn from a finite mixture distribution. The observed outcomes, $x_1, x_2, \ldots, x_n$, of a random variable X are assumed to have come from a mixture of a finite number, say k of groups with mixing weights $\pi_1, \pi_2, ..., \pi_k$. The finite mixture density of X takes the form,

$$g(x;\phi) = \sum_{i=1}^{k} \pi_i f_i(x;\theta_i), \text{ with } 0 < \pi_i < 1 \text{ and } \sum_{i=1}^{k} \pi_i = 1.$$

The complete collection ϕ, of the unknown parameters is need to be fully estimated, where $\phi=\{(\pi_i,\theta_i);i=1,2,...,k.\}$.

3. INCOMPLETE DATA PROBLEM

The problem of missing data is one of the most encountered phenomena in practice where observing the complete data in a real study is the exception rather the rule. Thus, in many situations, the available data sets are "incomplete" that is the observed part of the data contains only partial information about the phenomena under study.

We begin by assuming that $Y=(Y_{mis}, Y_{obs})$ is the complete data specification follow some parametric probability function, where Y_{mis} denotes the missing part of data, when sampling from a mixture this could be an original "raw" data or their component-membership or both and Y_{obs} denotes the observed part of the data. The fitting of binned mixture data is obviously an example of incomplete data; see Titterington and Makov (1985).

3.1 Binned Mixture Data

Data of a mixed population is said to be complete if for every item, both the measurement x and its component membership are observed. One example of an incomplete data is the binned data, where data are collected or transformed into frequencies located in disjoined areas X_j; j=1, 2, …, r. called bins. In binned data structure, the raw data and their component membership are not observed. The only available data is the set of frequencies n_j, where each frequency represents all the outcomes x_i; i=1, 2, …, k (from the entire set X) which belongs to the bin X_j. For given $n = \sum_{i=1}^{k} n_j$ and the marginal distribution of the random variable X, it is assumed that the set of frequencies n_j has a multinomial distribution with probability $P_j(\phi)$ of an observation falling in the j^{th} bin. The corresponding likelihood function then takes the form,

$$\mathcal{L}(\phi) = n! \prod_{i=1}^{k} \frac{\mathcal{P}_j(\phi)^{n_j}}{n_j}; \text{ with } \mathcal{P}_j(\phi) = \int_{X_j} g(x;\phi)\, dx.$$

3.2 Conditional Data

Conditional data are binned data enhanced with some additional sub-frequencies. These extra-information are believed to bring further improvement to the overall estimation. In this context, the unobservable outcomes $x_{ij1}, x_{ij2}, ..., x_{ijn_{ij}}$, which supported by X_{ij} to represent the observations made by the i^{th} component and falling in the j^{th} bin, are to be measured as sub-frequencies n_{ij} conditioned on the source which they have come from. The sub-frequencies n_{ij} are to be characterized by the random variables W_{ij} which has the following probability function,

$$P(\mathcal{W}_{ij} = n_{ij};\ i = 1, 2, ..., k) = n_j! \prod_{i=1}^{k} \frac{\mathcal{P}_{ij}(\theta_i)^{n_{ij}}}{n_{ij}},$$

where $P_{ij}(\phi)$ denotes the probability that an individual known to be from the i^{th} component falls into the j^{th} bin.

4. EM ALGORITHM

For the binned and conditional mixture, the form of likelihood functions is usually complicated, and hence a closed form solution to the normal equations cannot be found. Numerical techniques can then be applied. The EM algorithm is a broadly applicable approach to the iterative computation of MLE's, useful in a variety of incomplete-data problems. The aim of the EM algorithm is to find parameter values which maximize the "manufactured" complete data log likelihood in the E-step. As the complete data log likelihood is based partly on unobservable data, it has been replaced by its conditional expectation given the observed data. Starting from suitable initial parameter values, the E- and M-steps are repeated until convergence. The main idea of the EM algorithm is to maximize the incomplete log-likelihood indirectly by maximizing the expected complete log-likelihood function. The general relationship between mixtures and the EM algorithm has been covered in a number of sources, for more details see McLachlan and Basfor (1988), Everitt and Hand (1981), Titterington, et al. (1985), and McLachlan and Krishnan (1997).

One way to assess the quality of the estimates in the binned data context is to obtain their standard errors, for this an estimate of the information matrix is required. For detailed discussion concerning the calculation of this matrix, see McLachlan and Krishnan (1997).

5. ML ESTIMATES FOR MIXTURE BINNED DATA

This section describes the use of the EM-algorithm to obtain the ML estimates of the binned mixture data. Mixture models could be fitted via EM algorithm for observations that takes the form of binned data. The relationship between EM algorithm and binned data has been developed in the mixture case by Mclachlan and Jones (1988). On the basis of what has been

illustrated earlier, if all the frequencies n_j were observable, the corresponding log-likelihood function takes the form,

$$\log\left(\mathcal{L}(\phi)\right)=\sum_{j=1}^{r} n!\,\log\left(\sum_{i=1}^{k}\pi_i\,\mathcal{P}_{ij}(\theta_i)\right).$$

The above log-likelihood has no explicit solution. This problem for the binned mixture data can be solved within the EM framework by introducing new random variables, X_{js}, j=1, 2, …, r , s=1, 2, …, n_j, representing all the n_j unobservable individual observations in the j^{th} bin, and a hidden variable $Z_{js}=(Z_{1js}, Z_{2js}, ..., Z_{kjs})$, with $\sum_{i=1}^{k} Z_{ijs}=1,$ where Z_{ijs} equals one for X_{js} belongs to the i^{th} component and equals zero otherwise. In the light of the above, The complete log-likelihood function is simplified further to,

$$\log\left(\mathcal{L}_c(\phi)\right)=\sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{s=1}^{n_j} Z_{ijs}\left(\log(\pi_i)+\log\left(f_i(x_{js};\theta_i)\right)\right).$$

Given the complete data log-likelihood function, the E-step of the EM algorithm at the $(h+1)^{th}$ stage requires taking its expectation conditional on the observed data and current values of $\phi^{(h)}$. This conditional expectation (the Q-function), takes the following form,

$$Q\left(\phi;\phi^{(h)}\right)=\sum_{j=1}^{r}\sum_{i=1}^{k} n_j\,\mathcal{P}_{i|j}(\phi^{(h)})\left(\log(\pi_i)+\frac{1}{\mathcal{P}_{ij}(\theta_i^{(h)})}\int_{\mathcal{X}_j} f_i(x,\theta_i^{(h)})\log\left(f_i(x;\theta_i)\right)dx\right),$$

where $\mathcal{P}_{i|j}(\phi^{(h)})=\dfrac{\pi_i^{(h)}\,\mathcal{P}_{ij}(\theta_i^{(h)})}{\mathcal{P}_j(\phi^{(h)})}$ and $\mathcal{P}_{ij}(\theta_i^{(h)})=\int_{\mathcal{X}_j} f_i(x,\theta_i^{(h)})dx.$

The M-step aims to estimate the unknown parameters by maximizing the Q-function. This sometimes may not be easy. For the mixture normal densities, using the Lagrange multiplier, differentiate the Q-function with respect to each of the unknown parameters, equating the resulting expressions to zero and performing some simple algebra. The required estimates are given as,

$$\pi_i^{(h+1)}=\frac{\sum_{j=1}^{r} n_j\,\mathcal{P}_{i|j}(\phi^{(h)})}{n},$$

$$\mu_i^{(h+1)}=\frac{\sum_{j=1}^{r}\frac{n_j}{\mathcal{P}_j(\phi^{(h)})}\int_{\mathcal{X}_j} x\,\mathcal{N}\left(\mu_i^{(h)},(\sigma_i^2)^{(h)}\right)dx}{\sum_{j=1}^{r}\frac{n_j}{\mathcal{P}_j(\phi^{(h)})}\mathcal{P}_{ij}(\theta_i^{(h)})},$$

$$(\sigma_i^2)^{(h+1)} = \frac{\sum_{j=1}^{r} \frac{n_j}{\mathcal{P}_j(\phi^{(h)})} \int_{\mathcal{X}_j} (x-\mu_i^{(h+1)})^2 \mathcal{N}(\mu_i^{(h)}, (\sigma_i^2)^{(h)}) dx}{\sum_{j=1}^{r} \frac{n_j}{\mathcal{P}_j(\phi^{(h)})} \mathcal{P}_{ij}(\theta_i^{(h)})}.$$

6. ML ESTIMATES FOR CONDITIONAL DATA

It seems interesting to see the influence that would have to be on the estimation process when an extra information to the binned data, in the form of sub frequencies, is available. In order to formulate the observed likelihood function under the conditional data structure, suppose that in addition to the bin frequencies n_j, only the sub frequencies n_{lm}, which represent the number of observations contributed by the l^{th} component ($0 < l \le k$) of the mixture model that fall in the m^{th} bin ($0 < m \le r$), are to be known. Accordingly, the log-likelihood function of the observed data takes the following form,

$$\log(\mathcal{L}(\phi)) = \sum_{j=1}^{r} n_j \log\left(\sum_{i=1}^{k} \pi_i \mathcal{P}_{ij}(\theta_i)\right) + \sum_{j \in m} \sum_{i \in l_j} \log(\mathcal{P}_{i|j}(\phi)) \cdot$$

The above form of the log-likelihood function, clearly ,does not yield an explicit solution. This estimation problem can be solved within the EM framework by introducing the set of the missing component frequencies n_{ij}; $i \in l_j$, $j \in m$ and the variables X_{ij}.

By treating W_{ij} as a random variables corresponding to the component sub-frequencies n_{ij}; i=1, 2, …, k, j=1, 2, …., r the complete data log-likelihood can be written as,

$$\log(\mathcal{L}_c(\phi)) = \sum_{i=1}^{k} \sum_{j=1}^{r} w_{ij} \left(\log(\pi_i) + \log(f_i(x_{ij}; \theta_i))\right).$$

On the $(h+1)^{th}$ iteration of the EM algorithm, the E-step requires the calculation of Q-function, the conditional expectation of the complete log-likelihood, given the observed data of frequencies and the current value $\phi^{(h)}$ of ϕ, it follows that,

$$Q(\phi; \phi^{(h)}) = \sum_{i=1}^{k} \sum_{j=1}^{r} E_{\phi^{(h)}}\left(W_{ij} \middle| \{n_{ij}; i \in l_j, j \in m\}, n_j\right) \left(\log(\pi_i) + E_{\phi^{(h)}}(\log(f_i(X; \theta_i)) \middle| X \in \mathcal{X}_{ij})\right).$$

It is well known that, at each bin the conditional expectations of the random variable W_{ij}, given the set $\{n_{ij}; i \in l_j, j \in m\}$ of the observed component frequencies and the bin frequency n_j has the following form,

$$w_{ij}(\phi) = \begin{cases} (n_j - \sum_{i \in l_j} n_{ij}) \dfrac{\pi_i \mathcal{P}_{ij}(\theta_i)}{\sum_{i \notin l_j} \pi_i \mathcal{P}_{ij}(\theta_i)} & ; \text{ if } i \notin l_j \\ n_{ij} & ; \text{ if } i \in l_j, \end{cases}$$

where $v_{ij}(\phi^{(h)}) = E_{\phi^{(h)}}(W_{ij} | \{n_{ij}; i \in l_j, j \in m\}, n_j)$.

For mixture normal densities, by differentiating *Q*-function with respect to each of the unknown parameters, equating the resulting expressions to zero and by performing some simple algebra, we obtain,

$$\pi_i^{(h+1)} = \sum_{j=1}^{r} v_{ij}(\phi^{(h)}) \Big/ n ,$$

$$\mu_i^{(h+1)} = \frac{\sum_{j=1}^{r} \frac{v_{ij}(\phi^{(h)})}{P_{ij}(\theta_i^{(h)})} \int_{X_j} x \, N(\mu_i^{(h)}, (\sigma_i^2)^{(h)}) dx}{\sum_{j=1}^{r} v_{ij}(\phi^{(h)})},$$

$$(\sigma_i^2)^{(h+1)} = \frac{\sum_{j=1}^{r} \frac{v_{ij}(\phi^{(h)})}{P_{ij}(\theta_i^{(h)})} \int_{X_j} (x - \mu_i^{(h+1)})^2 N(\mu_i^{(h)}, (\sigma_i^2)^{(h)}) dx}{\sum_{j=1}^{r} v_{ij}(\phi^{(h)})}.$$

7. APPLICATIONS

One desirable way to judge the performance of the above mentioned procedure is to apply it to different simulated data sets with known features. This will provide a direct comparison between the actual and estimated parameter values. Furthermore, it seems interesting to develop the application from binned to conditional data case, since conditional data bring additional information to the binned data problem which then intuitively believed to improve the estimation. The procedure has been applied to many simulated data sets and to avoid tedious repetitions, this section deals with a single set of simulated data obtained from a three-component normal mixture density, the first two components are totally overlapped while the third one is well separated from them. A data set of size 700 has been simulated from a 3-component normal mixture density with, $\mu_1=\mu_2=0$, $\mu_3=6$, $\sigma_1^2=1$, $\sigma_2^2=4$, $\sigma_3^2=4$ and $\pi_1=0.2$, $\pi_2=0.5$. The overlap in this data set is very obvious as $\mu_1 = \mu_2$. The aim is to estimates of the parameters and their standard errors using the EM procedure. With the presence of this overlapping, the fitting is not as easy as that of the well separated components. Furthermore, the binned EM may face a harder task until convergence compared to the conditional EM. The Fortran program given by Mclachlan and Jones (1988) has been developed to carry out the estimation for the conditional binned data case. Table (7.1) summarizes the 700 data points with 18 equal-width intervals whereas Table (7.2) illustrates the results of the fitting.

Table (7.1): The empirical binned data of 3-component normal mixture.

Bin	1	2	3	4	5	6	7	8	9
Upper Boundary	-5.5	-4.5	-3.5	-2.5	-1.5	-0.5	0.5	1.5	2.5
Frequency	2	3	11	26	46	92	120	105	62

10	11	12	13	14	15	16	17	18	Total
3.5	4.5	5.5	6.5	7.5	8.5	9.5	10.5	11.5	------
40	36	34	39	34	31	11	5	3	700

Table (7.2): The fitting's output of the binned mixture data.

Parameter	Symbol	True value	Initial value	Estimate	St. Error
Means	μ_1	0	0	0.14015	0.23152
	μ_2	0	5	5.39071	16.3719
	μ_3	6	8	7.30659	19.3710
Variances	σ_1^2	1	2	3.45548	0.46026
	σ_2^2	4	2	2.33871	15.2593
	σ_3^2	4	2	2.49526	12.4985
Mixing Proportions	π_1	0.2	0.333	0.73740	0.05890
	π_2	0.5	0.333	0.13676	2.54197
	π_3	0.3	0.334	0.12583	2.50056
Number of Iterations				106	
Max of Log-likelihood				-1761.2380	
Residual				0. 166748	

To compare the performances of binned and conditional EM procedures, the initial values are deliberately selected to be far from the actual parameter values. As the new information become available, the attentions are turned to the case of conditional EM. Table(7.3) represents the conditional data structure and its fitting's summary is then illustrated in Table(7.4), the starting values are the same as those proposed for the binned data case. Indeed, the estimates are improved (as expected) and hence confirm the superiority of the conditional EM.

Table (7.3): The simulated conditional data of the 3-component normal mixture.

Bin	Upper Boundary	Frequency	Components		
			1	2	3
1	-5.5	2	#	#	#
2	-4.5	3	0	#	#
3	-3.5	11	#	#	#
4	-2.5	26	#	#	#
5	-1.5	46	#	38	#
6	-0.5	92	#	#	#
7	0.5	120	49	#	#
8	1.5	105	#	#	1
9	2.5	62	9	43	10
10	3.5	40	#	#	#
11	4.5	36	#	#	#
12	5.5	34	#	#	#
13	6.5	39	#	0	#
14	7.5	34	#	#	#
15	8.5	31	#	#	#
16	9.5	11	#	#	#
17	10.5	5	#	#	#
18	11.5	3	#	#	#
Total	----------	700		----------	

*The symbol # represents the missing sub frequencies.

Fig(7.1) displays the estimated mixture model (full line) for the binned and conditional EM, respectively. Clearly, both estimated densities are successfully capturing the shape of the corresponding histogram of the observed data. The outlook of the fits for both cases is similar except at the peak of the figures and this reflects the benefits of the additional information. The dashed curves at both figures represent the estimated component densities where the partial and complete overlapping are present. Obviously, there are significant differences between the outputs of the two EM procedures, where in Fig(7.1)(a) it seems that the overlapping is mostly between the second and the third components, where as Fig(7.1)(b) shows a total overlapping between the first two components and a well separated third component, which is more consistent with the actual simulated features.

The performances of the EM procedure, in terms of convergence speed for both binned and conditional EM are presented on the basis of the number of iterations (cycles) needed to reach the required convergence. The history of the observed log-likelihood (solid line) based on the binned and conditional data can be seen in Fig(7.2).

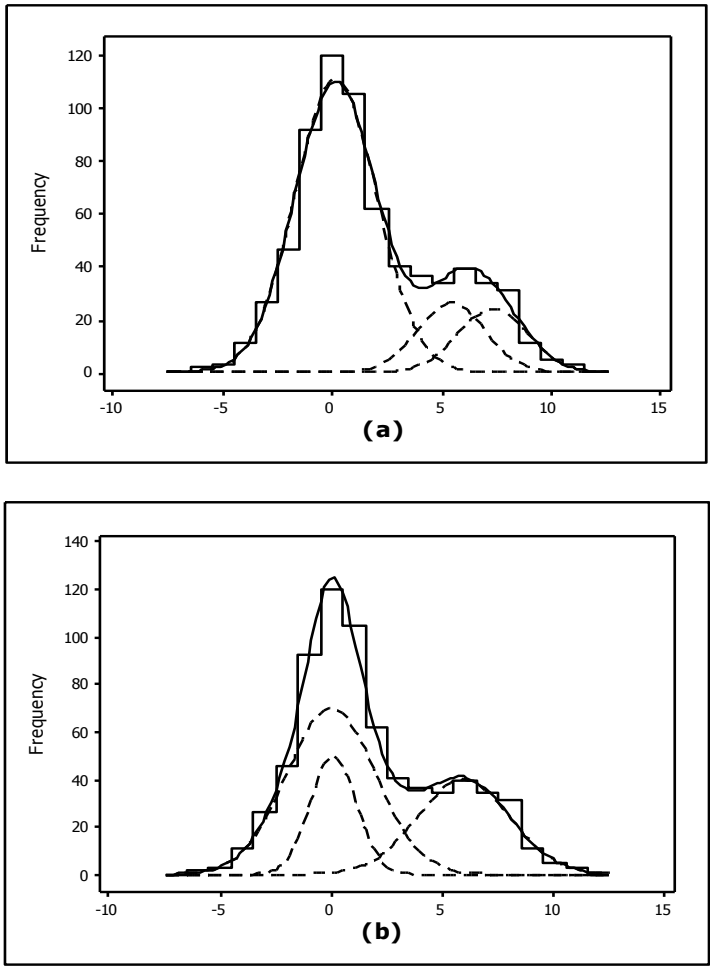
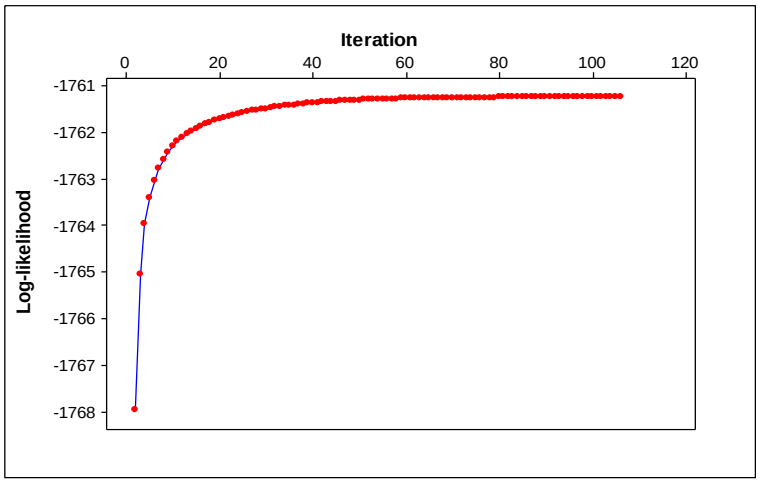
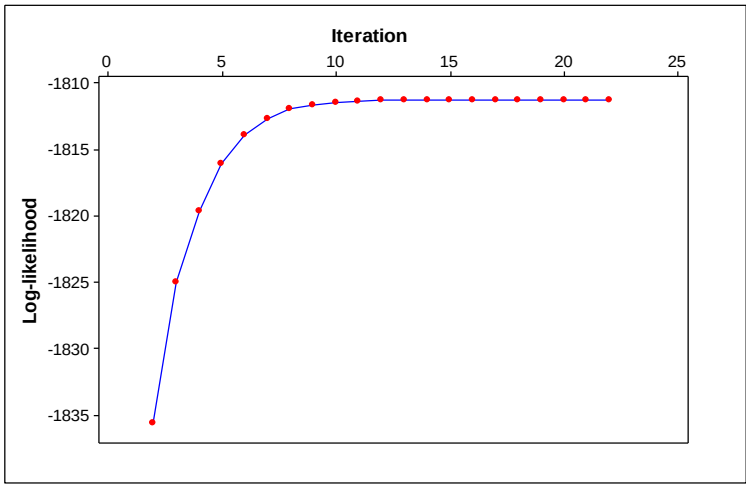


Fig (7.1): The histogram and the estimated mixture of three normal distributions via (a) the Binned EM. (b) the Conditional EM.



(a) for the Binned EM



(b) for the Conditional EM.

Fig (7.2): The history of log-likelihood versus the number of iterations:.

As anticipated, the conditional EM tends to converge more rapidly compared to the binned EM. This is another advantage to the conditional EM procedure over that of the binned EM.

8. CONCLUSIONS

This paper provides a simple and straightforward way to deal with both binned and conditional normal mixture data by employing the combination of maximum likelihood and the iterative EM estimation procedures. The main features that could be concluded here are,

Table (7.4): The fitting's output of the conditional mixture data.

Parameter	Symbol	True value	Initial value	Estimate	St. Error
Means	μ_1	0	0	0.0792602	0.33188
	μ_2	0	5	0.0189974	0.41602
	μ_3	6	8	5.9219180	0.49253
Variances	σ_1^2	1	2	1.0568550	0.95802
	σ_2^2	4	2	3.9703740	0.97192
	σ_3^2	4	2	4.2763790	1.20252
Mixing Proportions	π_1	0.2	0.333	0.1934000	0.19387
	π_2	0.5	0.333	0.5103550	0.21792
	π_3	0.3	0.334	0.2962450	0.05271
Number of Iterations				22	
Max of Log-likelihood				-1811.2760	
Residual				0. 050407	

1. Simulation results show that the EM algorithm does not perform well in the case of binned mixture data and performs very well in the case of conditional mixture data with a substantial computational savings and improvement of the overall estimation quality.
2. A direct comparison between binned and conditional EM algorithms showed that the conditional EM is less sensitive to the starting values than the binned EM, since the involvement of an additional information may provide the EM algorithm with extra strength to recover from any bad start.
3. The EM algorithm is usually simple to implement and its convergence is most likely guaranteed especially for well selected initial values.
4. The performance of the EM algorithm in the case of well-separated mixture components is much better than that in the case of heavily overlapped mixture components. This is partly due to the fact that the EM faces difficulty to discriminate between component memberships of the data items in the overlapped regions.
5. To judge the performance of the EM algorithm, the following measures are helpful, the maximum value of the estimated log likelihood, the standard error of the estimated parameters and the number of the required iterations. Finally a visual inspection of the closeness of the estimated curve of the suggested mixture model to the binned-data histogram enhanced with a goodness of fit test based on a chi-square distribution.

REFERENCES

Andrew, J. R. and William A. L.(2005). *A General Class of Multinomial Mixture Models For Anuran*. Calling Survey Data Ecology, 86(9),pp. 2505-2512.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). *Maximum Likelihood from Incomplete Data via the EM Algorithm*. J. R. Statist. Soc., B39, 1-38.

Everitt and Hand. D. J. (1981). *Fitting Mixture Distributions, Monographs on Applied Probability and Statistics.* Chapman and Hall & Methuen, Inc.

Juan Du.(2002). *Combined Algorithms For Fitting Finite Mixture Distributions*. B. Sc. Project, McMaster University. Canada.

Little, R.J.A., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

McLachlan, G.J., and Jones, P.N. (1990). *Algorithm AS254:Maximum Likelihood Estimation From Grouped And Truncated Data With Finite Normal Mixture Model*. Appl. Stat. 39, No. 2, 273-312.

McLachlan, G.J., and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: Wiley.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models.* Marcel Dekker, Inc.

McLachlan, G.J., and Jones, P.N. (1988). *Fitting Mixture Mode to Grouped and Truncated Data via EM Algorithm.* Biometrics 44, 571-578.

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical analysis of finite mixture distributions.* New York: Wiley

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 583–583.

LOOKING AT OUTLIERS

Nick Fieller
Department of Probability & Statistics
University of Sheffield
Sheffield, S3 7RH, U.K.
E-mail: n.fieller@sheffield.ac.uk

ABSTRACT

Outliers in low dimensional moderate sized data sets are easy to see since they 'stick out' in some direction or other. A simple scatter plot of data on a single or a pair or perhaps a triple of components will reveal them and it is then easy to assess informally whether or not they 'matter' in the sense of whether they are likely to affect later analyses adversely. However, once one moves beyond 'low dimensional' to moderate or high numbers of dimensions outliers are difficult to see and even more difficult to assess even informally whether or not they matter. In this context 'low dimensions' means certainly single figures and possibly outliers in five or six dimensional data sets can prove troublesome in this respect. Consideration of a union-intersection test approach for multivariate outliers provides a route for determining an effective 'outlier displaying component' (essentially the linear discriminant function between the outlier and the other data points). This basic idea can be exploited for display and informal assessment purposes and can be extended to handle multiple outliers and take advantage of robust estimation techniques.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 584–606.

TESTING FOR AGGREGATION BIAS

May Gadallah
Department of Statistics, Faculty of Economics and Political Science, Cairo University, Cairo, Egypt
E-mail: mayabaza@hotmail.com

ABSTRACT

Large aggregated data do not always provide unbiased estimators for the individual level parameters, such as variance, covariance, slope, and correlation coefficient. A small individual level data set will not provide efficient estimators for the individual level parameters. There are conditions under which the aggregated data will provide unbiased estimators for the slope, correlation coefficient, or both. Under these conditions, using the aggregated data can be more efficient by itself than using only small sample of individual level data. The efficiency can even increase by using a combination of both data sets. The purpose of this study is to use different approaches for testing the existence of these conditions by using large set of aggregated data and small set of individual level data, in order to provide a consistent estimator for slopes. The approaches include the proportionality of covariance matrices test and the hierarchical likelihood ratio test, assuming a hierarchical linear model under multivariate normal distribution. By using these tests a model selection approach is very promising as the efficiency gain can reach more than 160% compared by using the individual level data only or even using the consistent estimates of the aggregated data.

Keywords: Aggregation bias, hierarchical linear model, hierarchical likelihood ratio test, proportionality of covariance matrices.

1. INTRODUCTION

In many studies, data are available at both the individual level and at the group level, with individuals belonging to groups, such as factories, clinics, schools or any other group level. Studies limited to characteristics of groups of individuals are usually termed ecological studies. A major concern about ecologic studies is that individual level data are not available on the joint distribution of the variables of interest within the group. Using the aggregated level data for individual inference may cause the problem of ecologic fallacy, which Morgenstern (1998) defines as "the mistaken assumption that a statistical association observed between two ecologic (group-level) variables is equal to the association between corresponding variables at the individual level". Although research on this topic started 50 years ago with Robinson (1950), only recently has this area of research attracted the attention of scientists in diverse fields such as epidemiology, econometrics, politics, and sociology. These fields differ in how they treat the problem of ecologic inference, and within each field there is also a divergence in the adopted approaches.
The ecologic fallacy is sometimes called "aggregation bias" if the aggregated measures are used as the outcome in the analysis, as in this paper.

The goal of the paper is testing for the aggregation bias based on using large set of aggregated data to estimate individual level parameters by using only small set of individual level data, focusing on the slopes of multiple regression models.

The structure of the paper is as follows: the second section gives a review of the aggregation data problem, third section covers the framework adopted in the paper, fourth section discusses the aggregation bias under the framework, fifth section w discusses the conditions under which aggregated data will provide unbiased and consistent estimators, sixth section introduces the testing approaches, and finally last section presents the simulations results and conclusions.

2. PROBLEMS OF ECOLOGIACAL AND AGGREGATED DATA

For the rest of this paper we assume that the aim of the study is to get estimators of the individual level parameters. The individual level parameter estimators based on aggregated data may be subject to bias for reasons described below.

2.1 Specification Bias

Pure specification bias arises from assuming an incorrect ecological model. This commonly happens when we assume that the ecological relationship takes the same form as the individual relation. In some cases this assumption holds, such as in the case of the linear additive model (fixed parameters and no interaction). If i indexes the group (or group, area, etc), and j indexes the individual unit in group i, then the relationship at the individual level is given as follows:

$$E(Y_{ij} \mid X_{ij}) = \gamma + \beta X_{ij}$$

where k is the number of groups, n_i is sample size in group i, i=1, 2,…, k, j=1,…,n_i, and $\sum_{i=1}^{k} n_i$ =N, $\bar{X}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} X_{ij}$, $\bar{Y}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} Y_{ij}$

Data aggregation will lead to the following relationship $E(\bar{Y}_i \mid \bar{X}_i) = \gamma + \beta \overline{X}_i$.

This is the same relationship; thus the ecological study provides an unbiased estimate for the individual level parameters γ and β. In the case of non-linear models this simple relation does not occur. For example, in the log linear model (Sheppard, 2001), the individual relationship is

$$E(Y_{ij} \mid X_{ij}) = \exp(\gamma + \beta X_{ij}),$$

this does not imply

$$E(\bar{Y}_i \mid \bar{X}_i) = \exp(\gamma + \beta \overline{X}_i).$$

The following model should be used instead:

$$E(\bar{Y}_i \mid X_{i1}, X_{i2}, .., X_{in_i}) = n_i^{-1}\sum_{j=1}^{n_i} \exp(\gamma + \beta X_{ij})$$

2.2 Effect Modification Or Statistical Interaction

Product term covariates are usually discussed under the terms "effect modification". Consider the linear model

$$E(Y_{ij} \mid X_{ij}, Z_{ij}, W_{ij}) = \gamma + \beta X_{ij} + \delta Z_{ij} + \lambda W_{ij},$$

where W_{ij} is the interaction term $X_{ij}*Z_{ij}$. At the ecological level the above equation will result in the following:

$$E(\overline{Y} \mid \overline{X}_i, \overline{Z_i}, \overline{W_i}) = \gamma + \beta \overline{X_i} + \delta \overline{Z_i} + \lambda \overline{W_i},$$

where $\overline{W}_i$ is the average of X_{ij} x Z_{ij} in group i. Since we don't usually have the data at the individual level, the $\overline{W}_i$ term is usually not available. Instead the ecologic study might substitute $\overline{X}_i * \overline{Z}_i$ for $\overline{W}_i$, which yields biased estimators for the coefficients unless X and Z are uncorrelated within-groups (Greenland, 1992).

2.3 Contextual Effect

Contextual effect occurs when an individual outcome is affected, not only by its own exposure, but also by the average exposure in the same area (Wakefield et al., 2001):

$$E(Y_{ij} \mid X_{ij}) = \gamma + \beta X_{ij} + \delta \bar{X}_i .$$

When we use only aggregated data we cannot distinguish between the effect of the individual exposure and the effect of the contextual average exposure. This can be easily shown under the linear model:

$$E(\bar{Y}_i \mid \bar{X}_i) = \gamma + \beta \bar{X}_i + \delta \bar{X}_i = \gamma + (\beta + \delta) \bar{X}_i$$

In the ecological regression, the estimate of the coefficient will be for the sum of the individual effect β and the contextual effect δ; we cannot estimate β and δ separately. If the model is non-linear, with link function

$$g(E(Y_{ij} \mid X_{ij})) = \gamma + \beta X_{ij} + \delta \bar{X}_i,$$

not even the combined effect $(\beta + \delta)$ can be estimated without bias using aggregated data without further assumptions. Richardson et al. (1987) discussed having δ=0 (i.e. no contextual effect) as an assumption for estimating the individual effect β.

2.4 Ignoring Varying Parameters across Groups

We assume in § 2.1-2.3 that the parameters of interest, γ and β, are the same in all groups. If the parameters vary across groups then the ecological analysis will cause ecological fallacy as well.

2.4.1 Different Intercepts

The intercept in the linear model represents the baseline expected outcome, when all the covariates are assumed to have the value zero. If the baseline expected outcome changes across groups, an analysis using aggregated data might result in a biased estimate for the slope coefficient. Assume the following model:

$$E(Y_{ij} \mid X_{ij}) = \gamma_i + \beta X_{ij},$$

thus

$$E(\bar{Y}_i \mid \bar{X}_i) = \gamma_i + \beta\ \bar{X}_i \quad .$$

The ecological analysis yields the following ordinary least square (OLS) estimate for the slope coefficient:

$$\hat{\beta}_a = \frac{\sum_i^k (\bar{X}_i - \bar{\bar{X}})(\bar{Y}_i - \bar{\bar{Y}})}{\sum_i^k (\bar{X}_i - \bar{\bar{X}})^2}, \text{ where } \bar{\bar{X}} = \sum_i^k \bar{X}_i / k, \text{ and the same holds for } \bar{\bar{Y}}.$$

The subject "a" denotes that the estimator is based on aggregated data. The aggregate estimator is a biased estimator for the individual effect, as shown below:

$$E(\hat{\beta}_a) = \frac{\sum_i^k (\bar{X}_i - \bar{\bar{X}})(\gamma_i + \beta\bar{X}_i - \bar{\gamma} - \beta\bar{\bar{X}})}{\sum_i^k (\bar{X}_i - \bar{\bar{X}})^2}$$

$$= \frac{\sum_i^k (\bar{X}_i - \bar{\bar{X}})^2\beta + \sum_i^k (\bar{X}_i - \bar{\bar{X}})(\gamma_i - \sum_i^k \frac{\gamma_i}{k})}{\sum_i^k (\bar{X}_i - \bar{\bar{X}})^2} = \beta + \frac{\sum_i^k (\bar{X}_i - \bar{\bar{X}})(\gamma_i - \bar{\gamma})}{\sum_i^k (\bar{X}_i - \bar{\bar{X}})^2} = \beta + bias$$

where $\bar{\gamma} = \sum_i^k \frac{\gamma_i}{k}$. The bias given here is from the regression of the intercept γ_i on the $\bar{X}_i$. The bias in this case can inflate or deflate the slope and even change its sign depending on the pattern change of the intercepts across groups. If a random intercept is assumed, we can still get biased estimate for the slope if only aggregated data are used.

2.4.2 Different Slopes

If the slopes are different across groups, whether in case of systematic change or random change, the aggregated data cannot estimate the varying slopes. In this case it is not clear what combination of the slopes should be estimated.

3. STATISTICAL FRAMEWORK

Aggregation bias can be studied under various statistical frameworks depending on the data type and the statistical model used. King (1997), King et al. (1999), Gelman et al. (2001), and Wakefield (2004b) studied aggregation bias in dichotomous variables in 2 x 2 tables. Prentice et al. (1995), Sheppard et al. (1995), Sheppard et al. (1996), Plummer et al. (1996), Guthrie et al. (2001), Sheppard (2001), and Wakefield (2004a) studied aggregation bias under the log-linear model. Greenland et al. (1994), Wakefield (2003) studied aggregation bias under a non-linear model.

In this study we focus on studying inference using aggregated data in the hierarchical linear model with continuous variables using multivariate normal distribution.

3.1 The Hierarchical Linear Model

The hierarchical linear model was adopted in several approaches that discussed the ecological fallacy (or the aggregation bias) (Steel et al. 1996, Steel et al. 1996, Steel et al. 1997, Raghunathan et al. 2003, Gadallah 2006). The following illustrates the individual level as well as the aggregated level parameters under the hierarchical linear model.

3.1.1 Individual Level

Assume that there are k groups, and that in every group i we have n_i independent multivariate normal random variables (Y_{ij} , X_{1ij}, X_{2ij}) with mean (μ_{1xi}, μ_{2xi} ,μ_{yi}) covariance matrix Σ, where j=1,…,n_i, i=1,2,…,k, and $\sum_{i}^{k} n_i$ =N.

Under the hierarchical linear model, the conditional distribution of Y_{ij} *and* X_{1ij}, X_{2ij} given the (μ_{1xi}, μ_{2xi} ,μ_{yi}) is:

$$\begin{pmatrix} X_{1ij} \\ X_{2ij} \\ Y_{ij} \end{pmatrix} \Big| \begin{pmatrix} \mu_{1xi} \\ \mu_{2xi} \\ \mu_{yi} \end{pmatrix} \sim N_3 \left(\begin{pmatrix} \mu_{1xi} \\ \mu_{2xi} \\ \mu_{yi} \end{pmatrix}, \quad \Sigma \right) \tag{1}$$

where

$$\begin{pmatrix} \mu_{1xi} \\ \mu_{2xi} \\ \mu_{yi} \end{pmatrix} \sim N_3 \left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_y \end{pmatrix}, \quad \Omega \right) \qquad \Sigma = \begin{pmatrix} \sigma_{1x1x} & \sigma_{1x2x} & \sigma_{1xy} \\ \sigma_{1x2x} & \sigma_{2x2x} & \sigma_{1x2x} \\ \sigma_{1xy} & \sigma_{2xy} & \sigma_{yy} \end{pmatrix}$$

$$\Omega = \begin{pmatrix} \omega_{1x1x} & \omega_{1x2x} & \omega_{1xy} \\ \omega_{1x2x} & \omega_{2x2x} & \omega_{1x2x} \\ \omega_{1xy} & \omega_{2xy} & \omega_{yy} \end{pmatrix}$$

The unconditional distribution of Y_{ij} and X_{1ij} ,, X_{2ij} is easily found as

$$\begin{pmatrix} X_{1ij} \\ X_{2ij} \\ Y_{ij} \end{pmatrix} \sim N_3 \left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_y \end{pmatrix}, \quad (\Sigma + \Omega) \right) \tag{2}$$

The conditional distribution of Y_{ij} *given* X_{1ij} *,* X_{2ij}, μ_{1xi}, μ_{2xi} and μ_{yi} will be

$$Y_{ij} \mid X_{1ij}, X_{2ij}, \mu_{1xi}, \mu_{2xi}, \mu_{yi} \sim N(\mu_{yi} - \Sigma_{xx}^{-1}\Sigma_{xy} \begin{pmatrix} \mu_{1xi} \\ \mu_{2xi} \end{pmatrix} + \Sigma_{xx}^{-1}\Sigma_{xy} \begin{pmatrix} X_{1ij} \\ X_{2ij} \end{pmatrix}, \sigma_{y|x}) \tag{3}$$

where,

$$\sigma_{y|\bar{x}} = \sigma_{yy} - \Sigma_{xy}^T \Sigma_{xx}^{-1} \Sigma_{xy} \qquad \Sigma_{xy} = \begin{pmatrix} \sigma_{1xy} & \sigma_{2xy} \end{pmatrix}^T \qquad \Sigma_{xx} = \begin{pmatrix} \sigma_{1x1x} & \sigma_{1x2x} \\ \sigma_{1x2x} & \sigma_{2x2x} \end{pmatrix}$$

The conditional distribution of Y_{ij} given X_{1ij}, X_{2ij} is:

$$Y_{ij} \mid X_{1ij}, X_{2ij} \sim N \left(\mu_y - (\Sigma_{xx} + \Omega_{xx})^{-1}(\Sigma_{xy} + \Omega_{xy}) \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + (\sigma_{xx} + \Omega_{xx})^{-1}(\sigma_{xy} + \Omega_{xy}) \begin{pmatrix} X_{1ij} \\ X_{2ij} \end{pmatrix}, \quad \sigma^{"}_{y|x} \right) \tag{4}$$

where,

$$\sigma^{"}_{y|x} = (\sigma_{yy} + \omega_{yy}) - (\Sigma_{xy} + \Omega_{xy})^T (\Sigma_{xx} + \Omega_{xx})^{-1} (\Sigma_{xy} + \Omega_{xy})$$

However, for the unconditional distribution of Y_{ij} and X_{1ij}, X_{2ij} the slope vector we are trying to estimate is $(\Sigma_{xx} + \Omega_{xx})^{-1}(\Sigma_{xy} + \Omega_{xy})$. We call this quantity β_p, where the subscript p indicates the case of pooled data. Here β_p is generally not the same parameter as the one we get from the within-group data unless we have some restrictions.

3.1.2 Aggregated Level

Since we are assuming Y_{ij} and X_{1ij}, X_{2ij} are continuous, the aggregated data that we get will be the means. For the means, the conditional distribution given μ_{1xi}, μ_{2xi} ,μ_{yi} will be as follows, allowing n_i to be different for each group i:

$$\begin{pmatrix} \bar{X}_{1i} & \mu_{1xi} \\ \bar{X}_{2i} & \mid \mu_{2xi} \\ \bar{Y}_i & \mu_{yi} \end{pmatrix} \sim N_3 \left(\begin{pmatrix} \mu_{1xi} \\ \mu_{2xi} \\ \mu_{yi} \end{pmatrix}, \frac{1}{n_i} \quad \Sigma \right) \tag{5}$$

The unconditional distribution will be

.

$$\begin{pmatrix} \bar{X}_{1i} \\ \bar{X}_{2i} \\ \bar{Y}_i \end{pmatrix} \sim N_3 \left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_y \end{pmatrix}, \quad (\frac{\Sigma}{n_i} + \Omega) \right) \tag{6}$$

When we use the aggregated data in the regression, we are actually using the unconditional distribution to regress $\bar{Y}$ on $\bar{X}$. In this case the conditional distribution of $\bar{Y}_i \mid \bar{X}_{1i}, \bar{X}_{2i}$ is

$$\bar{Y}_i \mid \bar{X}_{1i}, \bar{X}_{2i} \sim N\left(\mu_y - (\Sigma_{xx}/n_i + \Omega_{xx})^{-1}(\Sigma_{xy}/n_i + \Omega_{xy})\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix} + (\Sigma_{xx}/n_i + \Omega_{xx})^{-1}(\Sigma_{xy}/n_i + \Omega_{xy})\begin{pmatrix}\bar{X}_{1i} \\ \bar{X}_{2i}\end{pmatrix}_i, \quad \sigma_{(\bar{Y}_i|\bar{X}_i)} \right) \tag{7}$$

where, $\sigma_{(\bar{Y}_i|\bar{X}_i)} = (\sigma_{yy}/n_i + \omega_{yy}) - (\Sigma_{xy}/n_i + \Omega_{xy})(\Sigma_{xx}/n_i + \Omega_{xx})^{-1}(\Sigma_{xy}/n_i + \Omega_{xy})$. Equation (7) shows that the conditional mean and the conditional variance differ from one group to another according to population size.

4. THE AGGREGATION BIAS

In this section, estimators using the individual data will be compared with the corresponding estimators using the aggregated data.

4.1 The Variance and the Covariance

Under the hierarchical linear model, the sample variance for any variable X using the aggregated data $\sum_i^k n_i(\bar{X}_i - \bar{\bar{X}})^2/(k-1)$ will provide an estimator for $\sigma_{xx} + c_{N,k}\omega_{xx}$ a function of the diagonal element in the variance-covariance matrix in equation (6), where

$$c_{N,k} = (N^2 - \textstyle\sum_i n_i^{\,2})/(N(k-1)).$$

This estimator is different from the estimator of within-group variance σ_{xx}. The sample covariance for a variable X and Y $\sum_i^k n_i(\bar{X}_i - \bar{\bar{X}})(\bar{Y}_i - \bar{\bar{Y}})/(k-1)$ will be an unbiased estimator for $\sigma_{xy} + c_{N,k}\omega_{xy}$, and not for σ_{xy}, where

$$\bar{X}_i = \sum_j^n \frac{X_{ij}}{n_i},\ \bar{\bar{X}} = \sum_i^k n_i \bar{X}_i / N,\ \bar{Y}_i = \sum_j^n \frac{Y_{ij}}{n_i},\ \bar{\bar{Y}} = \sum_i^k n_i \bar{Y}_i / N.$$

4.2 The Slope and the Correlation Coefficient

4.2.1 The Slope

Regressing Y on X's, conditionally on the means, produces the following equation:

$$E(Y_{ij} \mid X_{1ij}, X_{2ij}, \mu_{1xi}, \mu_{2xi}, \mu_{yi}) = \mu_{yi} - \Sigma_{xx}^{-1}\Sigma_{xy}\begin{pmatrix}\mu_{1xi} \\ \mu_{2xi}\end{pmatrix} + \Sigma_{xx}^{-1}\Sigma_{xy}\begin{pmatrix}X_{1ij} \\ X_{2ij}\end{pmatrix} \tag{8}$$

where the slope vector is $\beta_w = \Sigma_{xx}^{-1}\Sigma_{xy}$. Using the data coming from the unconditional distribution of Y_{ij} and X_{ij} leads to:

$$E(Y_{ij} \mid X_{1ij}, X_{2ij}) = \mu_y - (\Sigma_{xx} + \Omega_{xx})^{-1}(\Sigma_{xy} + \Omega_{xy})\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}_1 + (\Sigma_{xx} + \Omega_{xx})^{-1}(\Sigma_{xy} + \Omega_{xy})\begin{pmatrix}X_{1ij} \\ X_{2ij}\end{pmatrix} \tag{9}$$

where the slope is $\beta_p = (\Sigma_{xx} + \Omega_{xx})^{-1}(\Sigma_{xy} + \Omega_{xy})$.

Finally, using the aggregated data the slope from regressing $\bar{Y}$ on $\bar{X}$'s will follow from

$$E(\bar{Y}_i \mid \bar{X}_{1i}, \bar{X}_{2i}) = \mu_y - (\Sigma_{xy}/n_i + \Omega_{xy})(\Sigma_{xx}/n_i + \Omega_{xx})^{-1}\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}_1 + (\Sigma_{xy}/n_i + \Omega_{xy})(\Sigma_{xx}/n_i + \Omega_{xx})^{-1}\begin{pmatrix}\bar{X}_{1i}\\ \bar{X}_{2i}\end{pmatrix} \quad (10)$$

$$E(\bar{Y} \mid \bar{X}) = B_0 + B_1\bar{X} \quad (11)$$

where $B_1 = \left[\left[(\Sigma_{xx}/n_i + \Omega_{xx})\right]_{ss}^{-1}\left[(\Sigma_{xy}/n_i + \Omega_{xy})\right]_{ss}\right]$ consists of k x k diagonal matrices, and

$$\bar{Y} = \begin{pmatrix}\bar{Y}_1\\ \bar{Y}_2\\ .\\ .\\ \bar{Y}_k\end{pmatrix}, \quad \bar{X} = \begin{pmatrix}\bar{X}_{11} & \bar{X}_{11}\\ \bar{X}_{12} & \bar{X}_{22}\\ . & .\\ . & .\\ \bar{X}_{1k} & \bar{X}_{2k}\end{pmatrix}$$

while B_0 has $\mu_y - (\Sigma_{xx}/n_i + \Omega_{xx})^{-1}(\Sigma_{xy}/n_i + \Omega_{xy})\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}$, on its elements. Here the maximum likelihood estimates of B_0, B_1, will be different according to the assumption of equal population sizes $(n_i \equiv n, \ \forall \text{ i})$ or under unequal population sizes. Under this framework if we regress $\bar{Y}$ on only one of the $\bar{X}$'s, the least squares estimate for the slope is $\hat{\beta}_a = \dfrac{\sum_i^k n_i(\bar{X}_i - \bar{\bar{X}})(\bar{Y}_i - \bar{\bar{Y}})}{\sum_i^k n_i(\bar{X}_i - \bar{\bar{X}})^2}$.

Using Taylor's series expansions of the denominator and numerator around their expectations we find $E(\hat{\beta}_a) \approx \dfrac{\sigma_{xy} + c_{N,k}\Omega_{xy}}{\sigma_{xx} + c_{N,k}\Omega_{xx}}$ which is generally not equal to $\dfrac{\sigma_{xy}}{\sigma_{xx}}$, unless certain conditions are satisfied. Same will hold for the slope in the multiple regression, where we use:

$$\hat{\beta}_a = \frac{\sum n_i(\bar{X}_{1i} - \bar{\bar{X}}_1)(\bar{Y}_i - \bar{\bar{Y}})\sum n_i(\bar{X}_{2i} - \bar{\bar{X}}_2)^2 - \sum n_i(\bar{X}_{1i} - \bar{\bar{X}}_1)(\bar{Y}_i - \bar{\bar{Y}})\sum n_i(\bar{X}_{1i} - \bar{\bar{X}}_1)(\bar{X}_{2i} - \bar{\bar{X}}_2)}{\sum n_i(\bar{X}_{1i} - \bar{\bar{X}}_1)^2\sum n_i(\bar{X}_{2i} - \bar{\bar{X}}_2)^2 - [\sum n_i(\bar{X}_{1i} - \bar{\bar{X}}_1)(\bar{X}_{2i} - \bar{\bar{X}}_2)]^2} \quad (12)$$

The aggregated slope vector will be a consistent estimator for

$$(\Sigma_{xx} + c_{N,k}\Omega_{xx})^{-1}(\Sigma_{xy} + c_{N,k}\Omega_{xy}) \quad (13)$$

4.2.2 The correlation coefficient

Raghunathan et al. (2003) differentiated between three parameters that can be estimated using different data sets. The three parameters are:

* The individual simple correlation coefficient, which is based on within-group data, $\rho_{yx} = \sigma_{xy} \Big/ \sqrt{\sigma_{xx}\sigma_{yy}}$
* The total or contaminated correlation coefficient, which is based on the pooled data,

$$\Psi_{yx} = (\sigma_{xy} + \omega_{xy}) \Big/ \sqrt{(\sigma_{xx} + \omega_{xx})(\sigma_{yy} + \omega_{yy})}$$

* The aggregated correlation coefficient, which is based on the group means,

$$\varphi_{yx} = (\sigma_{xy} + c_{n,k}\omega_{xy}) \Big/ \sqrt{(\sigma_{xx} + c_{n,k}\omega_{xx})(\sigma_{yy} + c_{n,k}\omega_{yy})}$$

while the population ecological correlation coefficient (group means correlation) is defined as $\tau_{yx} = \omega_{xy} \Big/ \sqrt{\omega_{xx}\omega_{yy}}$.

Using only the aggregated data will not provide an unbiased estimator for ρ, since

$$E\left(\frac{\sum_i^k n_i(\bar{X}_i - \bar{\bar{X}})(\bar{Y}_i - \bar{\bar{Y}})}{\sqrt{\sum_i^k n_i(\bar{X}_i - \bar{\bar{X}})^2 \sum_i^k n_i(\bar{Y}_i - \bar{\bar{Y}})^2}} \right) \neq \rho_{xy}.$$

5. CONDITIONS OF CONSISTENT ESTIMATORS

Using aggregated data will produce biased estimators for the variance and the covariance of the within-group unless the Ω=0; i.e. there is no between-group variation. If $\Sigma \approx \Omega$ and n_i's $\to \infty$ (i.e. the within-group population that the aggregated data is based on is large), the bias will decrease and will tend to zero. The within-group slope and the correlation coefficient estimated by aggregated data will also suffer from bias, but not in the following situations:

a) When Ω =0 (i.e. there is no between-group variation), using the pooled or aggregated data will give unbiased estimator for the β_i.[1] This condition provides unbiased estimators for the variances, covariance, slope, and correlation coefficient.

b) Equal correlation condition

Assuming equal correlations condition we can get a consistent estimator for the correlation coefficient: Using the between covariance matrix

$$\Omega = \begin{pmatrix} \omega_{1x1x} & \omega_{1x2x} & \omega_{1xy} \\ \omega_{1x2x} & \omega_{2x2x} & \omega_{1x2x} \\ \omega_{1xy} & \omega_{2xy} & \omega_{yy} \end{pmatrix} = \begin{pmatrix} f_1 & 0 & 0 \\ 0 & f_2 & 0 \\ 0 & 0 & f_3 \end{pmatrix} \begin{pmatrix} \sigma_{1x1x} & \sigma_{1x2x} & \sigma_{1xy} \\ \sigma_{1x2x} & \sigma_{2x2x} & \sigma_{1x2x} \\ \sigma_{1xy} & \sigma_{2xy} & \sigma_{yy} \end{pmatrix} \begin{pmatrix} f_1 & 0 & 0 \\ 0 & f_2 & 0 \\ 0 & 0 & f_3 \end{pmatrix} \quad (14)$$

for simplicity assume that there are two variables only, in this case

[1] By using ordinary least squares for equal population sizes or weighted least squares for unequal population sizes.

$$\begin{pmatrix} f_1^2\sigma_{xx} & f_1 f_2\sigma_{xy} \\ f_1 f_2\sigma_{xy} & f_2^2\sigma_{yy} \end{pmatrix} = \begin{pmatrix} f_1^2\sigma_{xx} & f_1 f_2\rho\sqrt{\sigma_{xx}\sigma_{yy}} \\ f_1 f_2\rho\sqrt{\sigma_{xx}\sigma_{yy}} & f_2^2\sigma_{yy} \end{pmatrix} = \begin{pmatrix} \omega_{xx} & \rho\sqrt{\omega_{yy}\omega_{xx}}) \\ \rho\sqrt{\omega_{yy}\omega_{xx}} & \omega_{yy} \end{pmatrix} \quad (15)$$

where $f_1, f_2, f_3 > 0$. Since the correlation coefficient is the same in both covariance matrices, using the aggregated data will provide a consistent estimator for the correlation coefficient.

c) When the components of the unconditional variance and covariance are in a constant ratio. More specifically, writing

$$\text{Var(X)=A+B=E[Var(X}|\mu_x)] + \text{Var[E(X}|\mu_x)]$$

$\text{Cov(X,Y)=A`+B`=E[Cov(X,Y}|\mu_x, \mu_y)] + \text{Cov[E(X,Y}|\mu_x, \mu_y)]$**,** This condition is $A`/A=B`/B$.

Under the assumption that the association of X's and Y at the individual level does not vary from one group to another, Richardson et al. (1987). Under this condition $\Omega = f\,\Sigma$, where $f_1 = f_2 = f_3 = f$ which is a special case of the previous condition, where all elements in the two matrices are proportional. The aggregated data will provide consistent estimators for the slope and the correlation coefficient. We will call this condition the proportionality condition.

If the condition in a) or c) is met then. $\beta_i = \beta_a = \beta_p$

This follows from

$$\text{Var(X)=A+B=E[Var(X}|\mu_x)] + \text{Var[E(X}|\mu_x)] = \Sigma_{xx} + \Omega_{xx} \quad (16)$$

$$\text{Cov(X,Y)=A`+B`=E[Cov(X,Y}|\mu_x, \mu_y)] + \text{Cov[E(X,Y}|\mu_x, \mu_y)] = \Sigma_{xy} + \Omega_{xy} \quad (17)$$

Assuming the ratio condition the following is proved:

$$\begin{aligned} \beta_w &= \Sigma_{xx}^{-1}\Sigma_{xy} \\ \beta_p &= (\Sigma_{xx} + \Omega_{xx})^{-1}(\Sigma_{xy} + \Omega_{xy}) = (\Sigma_{xx} + f\,\Sigma_{xx})^{-1}(\Sigma_{xy} + f\,\Sigma_{xy}) = \Sigma_{xx}^{-1}\Sigma_{xy} \\ \beta_a &= (\Sigma_{xx} + c_{N,k}\Omega_{xx})^{-1}(\Sigma_{xy} + c_{N,k}\Omega_{xy}) = (\Sigma_{xx} + c_{N,k}\,f\,\Sigma_{xx})^{-1}(\Sigma_{xy} + c_{N,k}\,f\,\Sigma_{xy}) = \Sigma_{xx}^{-1}\Sigma_{xy} \end{aligned} \quad (18)$$

Under the previous conditions, consistent and efficient estimators can be derived from the aggregated data only if we know that the condition is satisfied; otherwise, we may get completely biased estimators.

The following section discusses the available tests for meeting these conditions under the multivariate normal model in case of a fixed within covariance matrix Σ. These tests will allow us to test whether the aggregated data will provide unbiased or consistent estimators for the different parameters or not, focusing on the slopes. The tests require the availability of small set of auxiliary individual level data in addition to aggregated data based on large population sizes.

6. TESTING APPROACHES

Federer (1951) studied testing the proportionality of the covariance matrices by solving the maximum likelihood approach for two groups and dimension of the matrix not exceeding 3. The approach was based on the application of the likelihood– ratio test under the normality assumption. The test computes the MLE of the covariance matrices and proportional constant under the null and the alternative. It then applies the likelihood ratio test.

Manly and Rayner (1987) and Rayner and Manly (1990) used a hierarchical likelihood ratio test for comparing several covariance matrices. They partitioned the likelihood ratio test into three components, thus allowing for nested models for covariance matrices, testing equal correlations, proportional matrices, and equal matrices. The paper assumed data consistent with multivariate normality.

Flury (1988) introduced a hierarchy of similarities among several covariance matrices, defining five levels of similarities among k covariance matrices: level 1 for equality, level 2 for proportionality, level 3 for common principle components (CPC), then the level 4 for partial CPC, and the level 5 for arbitrary covariance matrices.

For 2 x 2 covariance matrices, Flury (1983) considered various types of relationships between covariance matrices. One of these was equality of the regression slopes in two or more populations without assuming the equality of residual variance, which will be the case if the ratio condition is satisfied. In 1987 Erikson used the likelihood ratio test to test for proportionality. The paper provided an algorithm and proved its convergence and uniqueness of the maximum likelihood estimate.

6.1 Testing for Matrices Equality

Let $x_1, x_2, ..., x_{n_1} \sim N_d(\mu_1, \Sigma_1)$ and $w_1, w_2, ..., w_{n_2} \sim N_d(\mu_2, \Sigma_2)$. We can use a likelihood ratio test for testing the equality of the two, or more[2], dispersion matrices without considering any inferences among the two populations' means. $H_0 : \Sigma_1 = \Sigma_2 (= \Sigma)$ The likelihood ratio statistic is

$$l = \frac{\left|\hat{\Sigma}\right|^{n/2}}{\left|\hat{\Sigma}_1\right|^{n_1/2}\left|\hat{\Sigma}_2\right|^{n_2/2}} = c_{12}\frac{\left|Q_1\right|^{n_1/2}\left|Q_2\right|^{n_2/2}}{\left|Q_1+Q_2\right|^{n/2}} \tag{19}$$

where, $n = n_1 + n_2$, and

$$Q_1 = n_1\hat{\Sigma}_1 = \sum_i (x_i - \bar{x})(x_i - \bar{x})'$$

$$Q_2 = n_2\hat{\Sigma}_2 = \sum_i (w_i - \bar{w})(w_i - \bar{w})'$$

$$c_{12} = \frac{n^{nd/2}}{n_1^{n_1 d/2} n_2^{n_2 d/2}}$$

The distribution of –2log(l) is asymptotically chi-square with degrees of freedom ½ d (d+1) under the null. A modification of the statistic, which is a better chi-square approximation, is done using the degrees of freedom associated with Q_i, using n_i-1 instead of n_i, which will lead to an unbiased test.

[2] See Seber 1984 chapter 9.

Another chi-square approximation will be using a multivariate analogue of Bartlett's test of homogeneity (Seber 1984, p.450) $-2(1-c_1)\log M$, where M is the likelihood ratio after degrees of freedom modification and

$$c_1=\frac{2d^2+3d-1}{6(d+1)}\{\sum_i g_i^{-1}-g^{-1}\}, \text{ where } g_i=n_i-1, g=\sum_i g_i .$$

If the equality of the two matrices is met, this means that $\Omega = 0$. The sample covariance matrix based on the aggregated data will provide unbiased estimators for the same individual parameters and with different efficiencies.[3] If $\Omega = 0$, or if the proportionality condition is satisfied, combining the slope estimators of the aggregated data and the auxiliary data using the degrees of freedom will provide more efficient estimates[4] than using the only the individual level data.

6.2 Testing for Proportionality

Flury (1988) and Erikson (1987) introduced a likelihood ratio test for testing the proportionality of two or more matrices.

Let S_i, i= 0,…, k are independent p x p matrices with $S_i \sim Wishart_p(n_i^{-1}\Sigma_i, n_i)$, where Σ_i are positive definite covariance matrices. For testing the null hypotheses $H_0:\Sigma_i=\rho_i\Sigma_1$ i= 2,…,k, where the ρ_i's are unknown positive constants, the spectral decomposition of every covariance matrix can be written as follows:

$$\Sigma_i=\beta\Lambda_i\beta'$$
$$\Lambda_i=diag(\lambda_{i1},...,\lambda_{ip}), \text{ where i=2,…,k j=1,…,p}$$
$$\lambda_{ij}=\rho_i\lambda_{1j}$$

since the proportional model can be looked at as an offspring of the CPC (Common Principal Component) model. Flury (1988) derived the log-likelihood ratio statistic for testing the null versus the alternative of arbitrary covariance Σ_i to be:

$$\begin{aligned} X^2_{prop} &= -2\log\frac{L(\hat{\Sigma}_1,...\hat{\Sigma}_k)}{L(S_1,...,S_k)} \\ &= \sum_i^k n_i\log\frac{\det(\hat{\Sigma}_i)}{\det(S_i)} \\ &= n\{\sum_{j=1}^p\log(\hat{\lambda}_j+\sum_{i=1}^k r_i[p\log\hat{\rho}_i-\log(\det(S_i))]\} \\ r_i &= n_i/n \end{aligned} \qquad (20)$$

The test statistic, under the null hypothesis is asymptotically chi-square with degrees of freedom

[3] See Appendix 1.
[4] See Appendix 2.

$(k-1)(p^2+p-2)/2$.

Manly and Rayner (1987) and Eriksen (1987) proposed a different parameterisation for the same test and gave an algorithm for finding the maximum likelihood estimates. They also proved the convergence of the algorithm and uniqueness of the maximum likelihood estimator. Boente and Orellana (2003) proposed robust estimator of the proportionality constant and derived the asymptotic distribution. Two estimators will be studied in the paper

$$\hat{\rho} = \left[\frac{|W^*/(N^*-k^*)|}{|B/(k-1)*c_{N,k}|} \right]^{1/p}, \tag{21}$$

another estimator is

$$\hat{\rho} = \frac{trace(W^{*-1}B)}{p} * \frac{(k-1)*c_{n,k}}{(N^*-k^*)} \tag{22}$$

where p is the covariance matrix dimension.

6.3 Hierarchical Likelihood Ratio Test

In 1987 Manly and Rayner proposed a hierarchical likelihood ratio test to show that the test for equality matrices can be more informative by hierarchically partitioning it into three components. They based their test on having random samples from k multivariate normal populations each with d variables $X_1, X_2, ..., X_d$ with vector μ_j and covariance matrix Σ_j for population *j*, *j*= 1, …, k. The following table shows the nested models for covariance matrices:

Table 1: The Hierarchical Likelihood Ratio Test

Model	The Null	The Test Statistic	Degrees of Freedom
0	Σ_0, all j	$T_1 = n\log(\lvert\hat{\Sigma}_0\rvert / \lvert\hat{\Sigma}^{'}\rvert) - 2d\sum_{j=2}^{s} n_j \log(\hat{f}_j)$	k-1
1	$f_j^2\Sigma'$, all j, f_1=1	$T_2 = n\log(\lvert\hat{\Sigma}^{'}\rvert / \lvert\hat{\Sigma}^{''}\rvert) - 2d\sum_{j=2}^{k} n_j \{\log(\hat{f}_{1j},..., \hat{f}_{dj}) - d\log(\hat{f}_j)\}$	(d-1)(k-1)
2	$F_j\Sigma^{''}F_j^{'}, F_j = diag(f_{ij})$ for all j	$T_3 = \sum_{j=1}^{k} n_j \log(\lvert\hat{\Sigma}^{''}\rvert / \lvert S_j\rvert) + 2\sum_{j=2}^{k} n_j \log(\hat{f}_{1j}..., \hat{f}_{dj})$	½ (k-1)d(d-1)
3	All different	$T^* = T_1 + T_2 + T_3 = \sum_{j=1}^{k} n_j \log(\lvert\hat{\Sigma}_0\rvert / \lvert S_j\rvert)$	½ (k-1)d(d+1)

0 No between-group variation 1 Proportional matrices
2 Equal correlation 3 Different matrices

Here the *S's* are the sample covariance matrices after correcting their bias by dividing sums of squares and cross- products by degrees of freedom.

Under the assumption of having fixed a within-group covariance matrix, meeting any of the covariance matrices relationships mentioned in the previous section will allow us to use the aggregated data estimators in combination with the individual data and receive asymptotically high relative efficiency for the different estimators.[5]

In this paper both approaches of Erikson (1987) and the hierarchical likelihood ratio test (HLRT) for some suggested covariance matrices will be applied.

7. RESULTS AND CONCLUSIONS

In this section simulations are done to apply Erikson's and HLRT approaches to test for aggregation bias by using within covariance matrix of a small individual level data and the between covariance matrix of the aggregated level data.[6]

7.1 The Data

Assume:

1) We have k groups, and within every group i, we have n_i independent multivariate normal random variables (X_{1ij}, X_{2ij} ,Y_{ij}), with mean (μ_{1xi}, μ_{2xi} ,μ_{yi}) and covariance matrix Σ, where, i=1, 2, …, k, and j=1, …, n_i, $\sum_i^k n_i$ =N.

2) The group means μ_{1xi}, μ_{2xi} μ_{yi} are independently distributed with mean μ= (μ_{1x}, μ_{2x} ,μ_y) and covariance matrix Ω, and the aggregated data $\bar{X}_{1i}, \bar{X}_{2i}, \bar{Y}_i$, which are based on the population data, are available for all groups i=1,…, k and are based on large population sizes.

Let $(X^*_{1ij}, X^*_{2ij}, Y^*_{ij})$ where the j=n_i+1, n_i+2,…, n_i+ n_i^* denote the independent extra data from group i=1,2,…,k where either n^*=0 or $n^* \geq 2$.

Let W* and B* denote the 3x3 matrices of the within-group and the between-groups sums of squares based on the auxiliary data.

$$W^*= \begin{bmatrix} \sum_{i=1}^k \sum_{j=n_i+1}^{n_i+n_i^*} (X^*_{1ij} - \bar{X}^*_{i1})^2 & \sum_{i=1}^k \sum_{j=n_j+1}^{n_i+n_i^*} (X^*_{1ij} - \bar{X}^*_{1})(X^*_{2ij} - \bar{X}^*_{i2}) & \sum_{i=1}^k \sum_{j=n_j+1}^{n_i+n_i^*} (X^*_{1ij} - \bar{X}^*_{i1})(Y^*_{ij} - \bar{Y}^*_{i}) \\ \sum_{i=1}^k \sum_{j=n_j+1}^{n_i+n_i^*} (X^*_{1ij} - \bar{X}^*_{1})(X^*_{2ij} - \bar{X}^*_{i2}) & \sum_{i=1}^k \sum_{j=n_i+1}^{n_i+n_i^*} (X^*_{2ij} - \bar{X}^*_{i2})^2 & \sum_{i=1}^k \sum_{j=n_j+1}^{n_i+n_i^*} (X^*_{2ij} - \bar{X}^*_{i2})(Y^*_{ij} - \bar{Y}^*_{i}) \\ \sum_{i=1}^k \sum_{j=n_j+1}^{n_i+n_i^*} (X^*_{1ij} - \bar{X}^*_{1})(Y^*_{ij} - \bar{Y}^*_{i}) & \sum_{i=1}^k \sum_{j=n_i+1}^{n_i+n_i^*} (X^*_{2ij} - \bar{X}^*_{i2})(Y^*_{ij} - \bar{Y}^*_{i}) & \sum_{i=1}^k \sum_{j=n_i+1}^{n_i+n_i^*} (Y^*_{ij} - \bar{Y}^*_{i})^2 \end{bmatrix} \quad (23)$$

W^* can be written in the following sub-matrices $\begin{pmatrix} W^*_{xx} & W^*_{xy} \\ W^*_{yx} & W^*_{yy} \end{pmatrix}$

[5] See Gadallah, M. 2006.

[6] Even if the individual level data is a subset of the aggregated data, under the normality assumption the within and between covariance matrices are independent.

$$B^{*}=\begin{bmatrix} \sum_{i=1}^{k} n_i^{*}(\bar{X}_{1i}^{*}-\bar{\bar{X}}_1^{*})^2 & \sum_{i=1}^{k} n_i^{*}(\bar{X}_{1i}^{*}-\bar{\bar{X}}_1^{*})(\bar{X}_{2i}^{*}-\bar{\bar{X}}_2^{*}) & \sum_{i=1}^{k} n_i^{*}(\bar{X}_{1i}^{*}-\bar{\bar{X}}_1^{*})(\bar{Y}_i^{*}-\bar{\bar{Y}}^{*}) \\ \sum_{i=1}^{k} n_i^{*}(\bar{X}_{1i}^{*}-\bar{\bar{X}}_1^{*})(\bar{X}_{2i}^{*}-\bar{\bar{X}}_2^{*}) & \sum_{i=1}^{k} n_i^{*}(\bar{X}_{2i}^{*}-\bar{\bar{X}}_2^{*})^2 & \sum_{i=1}^{k} n_i^{*}(\bar{X}_{2i}^{*}-\bar{\bar{X}}_2^{*})(\bar{Y}_i^{*}-\bar{\bar{Y}}^{*}) \\ \sum_{i=1}^{k} n_i^{*}(\bar{X}_{1i}^{*}-\bar{\bar{X}}_1^{*})(\bar{Y}_i^{*}-\bar{\bar{Y}}^{*}) & \sum_{i=1}^{k} n_i^{*}(\bar{X}_{2i}^{*}-\bar{\bar{X}}_2^{*})(\bar{Y}_i^{*}-\bar{\bar{Y}}^{*}) & \sum_{i=1}^{k} n_i^{*}(\bar{Y}_i^{*}-\bar{\bar{Y}}^{*})^2 \end{bmatrix} \quad (24)$$

Where $\bar{X}_{ij}^{*}=\sum_{i=n_i+1}^{n_i+n_i^{*}} X_{ij}^{*}/n_i^{*}$, $\bar{\bar{X}}^{*}=\sum_i n_i^{*}\bar{X}_i^{*}/N^{*}$, $\bar{Y}_{ij}^{*}=\sum_{i=n_i+1}^{n_i+n_i^{*}} Y_{ij}^{*}/n_i^{*}$, $\bar{\bar{Y}}^{*}=\sum_i n_i^{*}\bar{Y}_i^{*}/N^{*}$, variable, and $N^{*}=\sum_i n_i^{*}$.

A matrix B can be also defined for the between-groups variation coming from the aggregated data, since n_i and group means are known, $B=\begin{pmatrix} B_{xx} & B_{xy} \\ B_{yx} & B_{yy} \end{pmatrix}$.

By taking the expectation under the hierarchical model, we find

$$E(B/(k-1))=\Sigma+c_{N,k}\Omega \quad (25)$$

$$E(B^{*}/(k^{*}-1))=\Sigma+c_{N^{*},k^{*}}\Omega \quad (26)$$

$$E(W^{*}/(N^{*}-k^{*}))=\Sigma \quad (27)$$

where k* should be 2 or greater representing the number of groups that we will get the extra individual data from, and $c_{N^{*},k^{*}}=(N^{*2}-\sum_i {n^{*}}_i^{2})/(N^{*}(k^{*}-1))$. We assume $c_{N^{*},k^{*}}=n^{*}$, extra sample sizes are equal across groups, and we also assume that population sizes are large enough that $(n_i \square n, \forall \text{ i})$.

Consistent least squares estimators for slopes will be as follows:
The slope at the individual level $\hat{\beta}_{yx}=W_{xx}^{*-1}W_{xy}^{*}$ The pooled slope $\hat{\beta}_p=T_{xx}^{*-1}T_{xy}^{*}$, where T is the sum of the within and the between covariance matrices. The slope at the aggregated level $\hat{\beta}_a=B_{xx}^{-1}B_{xy}$

Two approaches are applied; the Erikson's test for proportionality and the HLRT, testing the two matrices $W^{*}/(N^{*}-k^{*})$ and $B/(k-1)$ If the proportionality hypothesis fails to be rejected, we can use the aggregated data combined with the individual data to estimate the slope as follows:

$$\hat{\beta}_{weighted}=[(N^{*}-k^{*}-4)\hat{\beta}_a+(k-4)\hat{\beta}_w]/(N^{*}-k^{*}-k-8) \quad (28)$$

(See Appendix) If the proportionality condition is rejected, there is a proposed method of moment to combine the two estimators given by Gadallah (2006), but in this paper the individual level data estimator will be used instead.

In order to examine these approaches we shall assume different scenarios for the relation between the Ω and the Σ, 3000 simulations were done for each scenario in order to study the following:

a) Comparing the efficiency of the slope estimators using only the individual level data, the aggregated level data, and the combined estimator if the proportionality test fails to be rejected.

b) The bias and the efficiency of the proportionality constant estimator.
c) Comparing the power of the two tests (Erikson's approach and the hierarchical linear model)
d) The effect of the size of the individual level data and number of groups used as auxiliary data set.
e) The sensitivity of the change of the diagonal elements versus the off diagonal elements.

The following scenarios are applied:

First Scenario: The proportionality condition is satisfied

$$\Sigma = \begin{pmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.4 \\ 0.3 & 0.4 & 1 \end{pmatrix} \quad \Omega = 0.2\Sigma$$

Second Scenario: The two proportions of variances and covariances differ slightly

$$\Omega = \begin{pmatrix} f_1\sigma_{1x1x} & f_2\sigma_{1x2x} & f_2\sigma_{1xy} \\ f_2\sigma_{1x2x} & f_1\sigma_{2x2x} & f_2\sigma_{1x2x} \\ f_2\sigma_{1xy} & f_2\sigma_{2xy} & f_1\sigma_{yy} \end{pmatrix}$$ where f_1=0.8 and f_2=0.10 and same Σ

Third Scenario: The variances differ un-proportionally

$$\Omega = \begin{pmatrix} 2f\sigma_{1x1x} & f\sigma_{1x2x} & f\sigma_{1xy} \\ f\sigma_{1x2x} & 3f\sigma_{2x2x} & f\sigma_{1x2x} \\ f\sigma_{1xy} & f\sigma_{2xy} & 4f\sigma_{yy} \end{pmatrix}$$ where f=0.2 and same Σ

7.2 Discussion

In order to study the effect of the individual data set size, the three scenarios were repeated for number of groups 5 with size 10 from each group, 10 groups with 10 individuals from each group, and finally 5 groups with 20 individuals from each group.

Tables (2), (5), and (8) show the results of slope estimates and its mean square error using aggregated data only, individual level only, and the combined estimate if the proportionality fails to be rejected, for Erikson and HLRT approaches. The relative efficiency of the estimates was computed by comparing mean square error of the individual level data estimates with that of the combined estimates.

Tables (3), (6), and (9) are for the results of estimating the proportionality constant if the proportionality fails to be rejected. Tables (4), (7), and (10) give the rate of failing to reject the null in the three scenarios.

First Scenario: As seen in Table (2), when the proportionality condition is met, the aggregated data can be used to get consistent slope estimates with efficiency gain up to 160% than using only the individual level data. Both the Erikson's and the HLRT approaches get consistent estimates for the proportionality constant with higher efficiency if the HLRT is applied. The

HLRT is more sensitive in detecting the proportionality than the Eriskson's approach, but as seen in Table (4) increasing the size of the individual data set doesn't have large influence on increasing the acceptance rate

Second Scenario: If the proportionality factor of the diagonal (variances) differs slightly from the off diagonal proportionality factor (covariances), the efficiency gained by using the combined slope estimate decreases and its bias increase (Table (5)). Increasing the individual data set, either by increasing the number of groups or the number of individuals sampled from every group, decreases the efficiency gained by using the combined estimate, and also the bias of the combined estimates decreases. This decrease is due to the increase of the rejection rate and using the individual data only in estimating the slope (Table (7)).

The results of slope estimates using Erikson's approach and the HLR slightly differ, but the HLRT is more efficient in estimating the proportionality constant, which reflects an average of the diagonal and off diagonal proportionality factors.

Third Scenario: The results of the third scenario (Tables 8-10) show that although the HLRT approach is more efficient in estimating the slope, it is less powerful in rejecting the null. The conclusion is reflected in the larger bias of the slope estimates produced in the HLRT than the bias produced by the Erikson's approach, since the individual level estimates are dominating the latter one. The power of both tests increases as the individual data size increases, whether by increasing the number of groups or the number of individuals sampled from each group. Finally, although the estimates of the proportionality constant do not differ in the two approaches, the variance is less in the HLRT (Table (10)). The proportionality constant estimates in the third scenario appear to be dominated by the average proportionality factors of the diagonal elements.

7.3 Conclusions

Using aggregated data to estimate individual level parameters may cause bias. Detecting the existence of aggregation bias is very useful in order to get consistent estimators of individual level parameters with higher efficiency than using only small set of individual level data. Only a small set of the individual level data can be enough in detecting the bias, with no preferences in increasing the number of groups or the number of individuals per group.

The HLRT is more efficient in estimating the proportionality constant. In Erikson's approach using the trace or the determinant (equations 21 and 22) in estimating the proportionality constant almost give the same results. The relationship of the diagonal values in the two covariance matrices, has a greater effect on the test than the off diagonal values. More investigation is needed in order to determine to what extent do diagonal values affect the significance and power of the test.

The HLRT is a more promising approach in detecting the aggregation bias, since it can be used in estimating the variances, covariances, and the correlation coefficients as well as the slope through applying all the tests. Testing the equality, proportionality of within and between covariance matrices is a new application and a very promising approach, not only in detecting aggregation bias, but also in multilevel models which needs further investigations.

First Scenario

$$\Sigma = \begin{pmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.4 \\ 0.3 & 0.4 & 1 \end{pmatrix} \quad \Omega = 0.2\Sigma$$

Table 2: Slope Estimation Results using Erikson's approach and HLRT Applying the First Scenario

	Slope Estimate using aggregated data		Slope Estimate using individual data (true value=0.095)		Weighted Slope Estimate using both data sets		Relative Efficiency of Combined estimate	
	Erikson	HLRT	Erikson	HLRT	Erikson	HLRT	Erikson	HLRT
k*=5, n*=10	0.0954 (0.0127)	0.092 (0.0135)	0.0952 (0.0247)	0.092 (0.0262)	0.0952 (0.0112)	0.0947 (0.010)	2.21	2.62
k*=10, n*=10	0.0948 (0.0135)	0.0961 (0.0135)	0.0938 (0.0125)	0.099 (0.0123)	0.095 (0.0076)	0.0947 (0.0071)	1.64	1.73
k*=5, n*=20	0.0963 (0.0134)	0.0943 (0.0130)	0.0997 (0.0121)	0.0955 (0.0116)	0.0979 (0.0074)	0.0955 (0.0067)	1.62	1.73

- The number between the brackets is the mean square error
- The mean square error of the aggregated estimate is computed from the true aggregated slope

Table 3: Proportionality Constant Estimation Results using Erikson's approach and HLRT Applying the First Scenario

	Proport. constant Erikson's approach (true value=0.2)		Proport. constant
	Trace	Determinant	HLRT
k*=5, n*=10	0.203 (0.0037)	0.202 (0.0038)	0.202 (0.0010)
k*=10, n*=10	0.202 (0.0040)	0.201 (0.0040)	0.202 (0.0007)
k*=5, n*=20	0.202 (0.0044)	0.201 (0.0044)	0.201 (0.0006)

The mean of the proportionality constant was computed when the null fails to be rejected. The number between brackets presents the variance

Table 4: Acceptance Rate using Erikson's approach and HLRT Applying the First Scenario

	Acceptance Rate			
	Erikson	HLRT		
		Equality	Proportionality	Equal Correlation
k*=5, n*=10	0.935	0	0.977	0.975
k*=10, n*=10	0.917	0	0.978	0.976
k*=5, n*=20	0.914	0	0.981	0.972

Second Scenario

$$\Omega = \begin{pmatrix} f_1\sigma_{1x1x} & f_2\sigma_{1x2x} & f_2\sigma_{1xy} \\ f_2\sigma_{1x2x} & f_1\sigma_{2x2x} & f_2\sigma_{1x2x} \\ f_2\sigma_{1xy} & f_2\sigma_{2xy} & f_1\sigma_{yy} \end{pmatrix}$$ where f_1=0.8 and f_2=0.10 and same Σ

Table 5: Slope Estimation Results using Erikson's approach and HLRT Applying the Second Scenario

	Slope Estimate using aggregated data (true value=0.0232)		Slope Estimate using individual data		Weighted Slope Estimate using both data sets		Relative Efficiency Of Combined estimate	
	Erikson	HLRT	Erikson	HLRT	Erikson	HLRT	Erikson	HLRT
k*=5, n*=10	0.023 (0.0125)	0.0254 (0.0120)	0.0944 (0.0261)	0.0959 (0.0259)	0.0744 (0.0204)	0.0755 (0.0198)	1.28	1.31
k*=10, n*=10	0.0238 (0.0126)	0.02318 (0.0123)	0.0949 (0.0125)	0.09627 (0.0131)	0.0880 (0.0115)	0.08653 (0.0118)	1.09	1.11
k*=5, n*=20	0.0201 (0.0132)	0.02318 (0.0123)	0.0955 (0.0119)	0.0960 (0.0116)	0.088 (0.0110)	0.0862 (0.0105)	1.08	1.10

- The number between the brackets is the mean square error
- The mean square error of the aggregated estimate is computed from the true aggregated slope

Table 6: Proportionality Constant Estimation Results using Erikson's approach and HLRT Applying the Second Scenario

	Proport. constant Erikson's approach		Proport. constant
	Trace	Determinant	HLRT
k*=5, n*=10	0.883 (0.742)	0.876 (0.731)	0.865 (0.0192)
k*=10, n*=10	0.883 (1.663)	0.872 (1.622)	0.879 (0.0125)
k*=5, n*=20	0.883 (1.802)	0.872 (1.756)	0.882 (0.0123)

- The mean of the proportionality constant was computed when the null fails to be rejected
- The number between brackets presents the variance

Table 7: Acceptance Rate using Erikson's approach and HLRT Applying the Second Scenario

	Acceptance Rate			
	Erikson	HLRT		
		Equality	Proportionality	Equal Correlation
k*=5, n*=10	0.519	0.897	0.977	0.580
k*=10, n*=10	0.321	0.893	0.972	0.394
k*=5, n*=20	0.304	0.898	0.979	0.392

Third Scenario

$$\Omega = \begin{pmatrix} 2f\sigma_{1x1x} & f\sigma_{1x2x} & f\sigma_{1xy} \\ f\sigma_{1x2x} & 3f\sigma_{2x2x} & f\sigma_{1x2x} \\ f\sigma_{1xy} & f\sigma_{2xy} & 4f\sigma_{yy} \end{pmatrix} \text{ where } f=0.2 \text{ and same } \Sigma$$

Table 8: Slope Estimation Results using Erikson's approach and HLRT Applying the Third Scenario

	Slope Estimate using aggregated data (true value=0.0574)		Slope Estimate using individual data (true value=0.095)		Weighted Slope Estimate using both data sets		Relative Efficiency Of Combined estimate	
	Erikson	HLRT	Erikson	HLRT	Erikson	HLRT	Erikson	HLRT
k*=5, n*=10	0.0569 (0.008)	0.0553 (0.0081)	0.095 (0.0259)	0.100 (0.0257)	0.0878 (0.0241)	0.086 (0.0183)	1.07	1.40
k*=10, n*=10	0.0567 (0.0082)	0,0596 (0.0085)	0.097 (0.0127)	0.0972 (0.0119)	0.946 (0.0119)	0.0919 (0.010)	1.07	1.19
k*=5, n*=20	0.059 (0.0083)	0.059 (0.0081)	0.098 (0.0117)	0.0962 (0.0115)	0.0958 (0.0110)	0.0919 (0.010)	1.06	1.15

- The number between the brackets is the mean square error
- The mean square error of the aggregated estimate is computed from the true aggregated slope

Table 9: Proportionality Constant Estimation Results using Erikson's approach and HLRT Applying the Third Scenario

	Proport. constant Erikson's approach		Proport. constant
	Trace	Determinant	HLRT
k*=5, n*=10	0.628 (0.628)	0.623 (0.592)	0.612 (0.013)
k*=10, n*=10	0.630 (1.812)	0.622 (1.7669)	0.628 (0.012)
k*=5, n*=20	0.633 (1.851)	0.624 (1.803)	0.628 (0.013)

- The mean of the proportionality constant was computed when the null fails to be rejected
- The number between brackets presents the variance

Table 10: Acceptance Rate using Erikson's approach and HLRT Applying the Third Scenario

	Acceptance Rate			
	Erikson	HLRT		
		Equality	Proportionality	Equal Correlation
k*=5, n*=10	0.399	0.166	0.696	0.772
k*=10, n*=10	0.180	0.069	0.502	0.651
k*=5, n*=20	0.178	0.066	0,469	0.623

APPENDIX

1. Combined estimator for the Variance/Covariance

Since W* , B*, and B follow the Wishart distribution:

$Var(W^*/(N^*-k^*)) = 2[(\sigma_{lr}^2 + \sigma_{ll}\sigma_{rr})/(N^*-k^*)]$

$Var(B_{xx}/(k-1)) = 2[(\sigma_{lr}^2 + \sigma_{ll}\sigma_{rr})/(k-1)]$ for l and r represent the three variables X_1, X_2 and Y.

An optimal weighted average estimate for each element of Σ will be corresponding element in $\frac{W^* + B}{N^* + k - k^* - 1}$ [7], which will be more efficient than using only the auxiliary data.

2. Combined Estimator for the Slope

Assuming that populations have equal sizes, then

$\hat{\beta}_a = B_{xx}^{-1} B_{xy}$

$V(\hat{\beta}_a \mid X_1, X_2) = \sigma_{(\bar{Y}_i|\bar{X}_i)} B_{xx}^{-1}$ and

$B_{xx}^{-1} \sim Wishart^{-1}((\Sigma_{xx}/n + \Omega_{xx})^{-1}, k-1)$

$E(B_{xx}^{-1}) = \frac{(\Sigma_{xx}/n + \Omega_{xx})^{-1}}{k-1-p-1}$ where p=matrix dimension=2 (Gupta et al. (2000))

$V(\hat{\beta}_a) = E_x(V(\hat{\beta}_a \mid X_1, X_2)) + V_x(E(\hat{\beta}_a \mid X_1, X_2))$

$= E_x(\sigma_{(\bar{Y}_i|\bar{X}_i)} B_{xx}^{-1}) + V_x[(\Sigma_{xx}/n + \Omega_{xx})^{-1}(\Sigma_{xy}/n + \Omega_{xy})]$

$= \sigma_{(\bar{Y}_i|\bar{X}_i)} \frac{(\Sigma_{xx}/n + \Omega_{xx})^{-1}}{k-4}$

Since under normality assumption the within and between covariance matrices are independent, then under the proportionality assumption the combined estimator will be more efficient than using only one estimator.[8] $\hat{\beta}_{weighted} = [(N^* - k^* - 4)\hat{\beta}_a + (k-4)\hat{\beta}_i]/(N^* - k^* - k - 8)$

AKNOWLEDGEMENT

I would like to express my gratitude and appreciation to Professor Duan Naihua and Professor William Cumberland for supervising and mentoring me throughout my dissertation, which made this paper possible.

[7] If the individual level data is not a subset of the aggregated data the total sample covariance matrix can be used instead of the within covariance matrix as an estimate for Σ with N*-1 degrees of freedom.

[8] See Gadallah, M.(2006) pp 31.

REFERENCES

Boente, G., Orellana, L., (2004), "Robust plug-in estimators in scatter models," *Journal of Statistical Planning and Inference*, 122, 95-110.

Eriksen, P.S. (1987), "Proportionality of covariance matrices," *The Annals of Statistics*, 15, 732-748.

Federer, W.T.(1951), "Testing proportionality of covariance matrices," *Annals of Mathematical Statistics*, 22, 102-106.

Flury, B. (1983), "Maximum likelihood estimation of some patterned 2 x 2 covariance matrices," Technical report 83-9, Purdue University, Department of Statistics, Lafayette, IN.

Flury, B. (1986), "Proportionality of k Covariance Matrices," *Statistics and Probability Letters*, 4, 29-33.

Flury, B. (1988), *Common Principal Components and Related Multivariate Models*, Wiley New York.

Gadallah, M. (2006), "*Combining Aggregated and Individual Level Data to Estimate Individual Level Parameters: Variance, Covariance and Slope Coefficient*". PhD thesis, School of Public Health, University of California, Los Angeles.

Gelman, A., Park, D.K., Ansolabehere, S., Price, P.N, & Minnite, L. C. (2001), "Models, assumptions and model checking in ecological regressions," *Journal of the Royal Statistical Society*, Series A 164, 101-18.

Greenland, S. (1992), "Divergent biases in ecologic and individual-level studies," *Statistics in Medicine*, 11, 1209-23.

Greenland, S. & Robins, J. (1994), "Invited commentary: Ecologic studies – biases, misconceptions and counterexamples," *American Journal of Epidemiology*, 139, 747-64.

Gupta, A.K., Nagar, D.K. (2000), *Matrix Variate Distributions*, Chapman & Hall / CRC, Boca Raton.

Guthrie, K. A. & Sheppard, L. (2001), "Overcoming biases and misconceptions in ecological studies," *Journal of the Royal Statistical Society*, Series A 164, 141-54.

King, G. (1997), *A Solution to the Ecological Inference Problem*, Princeton University Press, Princeton, New Jersey.

King, G. Rosen, O., and Tanner, M. A. (1999), "Binomial-beta hierarchical models for ecological inference," *Sociological Methods and Research* 28, 61-90.

Manly, B.F.J., and Rayner, J.C. W. (1987)' "The comparison of sample covariance matrices using likelihood ratio tests," *Biometrika*, 74, 841-847.

Morgenstern, H. (1998), "Ecologic study", in P. Armitage & T. Colton, eds, *Encyclopedia of Biostatistics Vol. 2*. Wiley and Sons Ltd, New York, 1255-76.

Plummer, M. & Clayton D. (1996), "Estimation of population exposure," *Journal of the Royal Statistical Society*, Series B 58, 113-26.

Prentice, R.L.& Sheppard, L. (1995), "Aggregate data studies of disease risk factors," *Biometrika*, 82, 113-25.

Raghunathan, T. E. Diehr, P. K. & Cheadle, A. D. (2003), "Combining aggregate and individual level data to estimate an individual level correlation coefficient," *Journal of Educational and Behavioral Statistics,* 28, 1-19.

Rayner, J.C.W., Manly, B.F.J. (1990), "Hierarchical likelihood ratio tests for equality of covariance matrices," *Journal of Statistics and Computation Simulation*, 35, 91-99.

Richardson, S., Stucker, I. & Hemon, D. (1987), "Comparison of relative risks obtained in ecological and individual studies: Some methodological considerations," *International Journal of Epidemiology*, 16, 111-20.

Robinson, W.S. (1950), "Ecological correlations and the behavior of individuals," *American Sociological Review,* 3, 351-357.

Seber, G.A.F. (1984), *Multivariate Observations,* Wiley, New York.

Sheppard, L. (2001), "Insight on bias and information in group-level studies," *Biostatistics* 4:265-278.

Sheppard, L. and Prentice, R. L. (1995), "On the reliability and precision of within-and between-population estimates of relative risk parameters," *Biometrics* 51, 853-63.

Sheppard, L., Prentice, R. L., and Rossing, M.A. (1996), "Design considerations for estimation of exposure effects on disease risk, using aggregate data studies," *Statistics in Medicine,* 15, 1849-58.

Steel, D.G., Tranmer, M. and Holt, D. (1997), "Logistic regression analysis with aggregate data: Tacking the Ecological Fallacy," *Proceedings of the Survey Research Section, American Statistical Association,* 324-29.

Steel, D. G. and Holt, D. (1996), "Analysing and adjusting aggregation effects: The ecological fallacy revisited," *The International Statistical Review,* Vol 64, no1, pp 39 -60.

Steel, D.G., Holt, D. and Tranmer, M. (1996), "Making unit-level inference from aggregate data," *Survey Methodology*. Vol 22, no1, 3-15.

Wakefield, J. & Salway, R. (2001), "A statistical framework for ecological and aggregate studies," *Journal of the Royal Statistical Society,* Series A 164, 119-137.

Wakefield, J. (2003), "Sensitivity analyses for ecological regression," *Biometrics*, 59, 9-17.

Wakefield, J. (2004a), "A critique of statistical aspects of ecological studies in spatial epidemiology," *Environmental and Ecological Statistics*, 11, 31-54.

Wakefield, J. (2004b), "Ecological inference for 2 x 2 tables," *Journal of the Royal Statistical Society,* Series A, 167, 385-445.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 607–633.

SENSITIVITY OF DESCRIPTIVE GOODNESS-OF-FIT INDICES TO SPECIFICATION ERROR IN STRUCTURAL EQUATION MODELING

Hesham F. Gadelrab
Mansoura University, Faculty of Education, Psychology Department
Mansoura, Egypt 35516
The British University in Egypt (BUE), Business Administration Department
Sherouk City, Cairo, Postal No. 11837, P.O. Box 43
E-mail: heshfm@mans.edu.eg, hesham.gadelrab@bue.edu.eg

ABSTARCT

The purpose of the present study was to empirically investigate the sensitivity of commonly used Structural Equation Modeling (SEM) fit indices derived from Maximum Likelihood (ML) estimation method to the degree and type of model misspecification under different sample size conditions. The performance of these fit indices was examined over four levels of model misspecification (ranging from no error to high misspecification), two types of model misspecification (recursive and nonrecursive misspecification) and four levels of sample size (ranging from 100 to 1000). A three-factor balanced design was used in the present study with repeated measures over degree and type of misspecification. Data were generated using EQS 6.1 under different model misspecifications and sample size conditions. Findings from this study showed that the goodness-of-fit test did not equally detect the same size of error in different types of models. Fit indices were less sensitive to recursive than nonrecursive misspecification under the same sample size and misspecification conditions. Therefore, conclusions differ greatly on whether there is a misspecification in the model with respect to the size and type of specification error in the model as well as the size of sample. Fit indices were greatly varied in their reliability of estimation and sensitivity to sample size. Recommendations for using specific fit indices are discussed.

Keywords: Structural Equation Modeling, Misspecification, Goodness-of-fit, Sensitivity, Fit Indices, Monte Carlo Study

1. INTRODUCTION

Structural equation modeling (SEM) has been increasingly recognized as a useful quantitative method in specifying, estimating, and testing hypothesized theoretical models which describe relationships among variables that are substantively meaningful in the real world (Fan & Wang, 1998). SEM is a comprehensive statistical approach to testing hypotheses about relations among observed and latent variables (Hoyle, 1995). It represents a broad class of models that allows simultaneous estimation of relations between observed and latent variables and among latent variables themselves (Bollen, 1989b). A structural equation model is comprised of a measurement model, which specifies how latent variables or theoretical constructs are measured

in terms of observed variables, and a structural model, which specifies the direct and indirect relationships among latent variables (Schumacker & Lomax, 1996).

SEM applications in substantive research include model specification, model identification, model estimation, and model evaluation. Model specification involves the explicit statement of the hypothesized relationships among the variables; both observed and latent in the model. The model is specified on the basis of a specific theoretical framework. A model is identified if model parameters that need to be estimated can be computed. This refers to the possibility of finding unique values of the parameters of the specified model. Model estimation involves the calculation of model parameters that need to be estimated, so that the estimated parameters lead to the sample covariance matrix. A number of different methods are commonly used to fit structural equation models to data. Some of more popular methods are maximum likelihood (ML), generalized least square (GLS), and weighted least squares (WLS). Model evaluation assesses overall model adequacy by showing to what extent a specified model fits the empirical sample data. If the model does not fit well, it could be improved through model respecification. Although each step in applying SEMs has been the subject of considerable discussion, the most heated controversies involve testing the model fit and respecification (Bollen & Long, 1993).

There are two general classes of assumptions that underlie the statistical methods used to estimate SEM models: distributional and structural (Satorra, 1989). Distributional assumptions affect the precision of the estimator and hence the significance of test statistics. Maximum likelihood (ML) and generalized least squares (GLS) are the most frequently used estimation methods for SEM. ML and GLS make the distributional assumption that the measured variables have a multivariate normal distribution in the population. The restrictive character of these assumptions has motivated the development of new estimation methods that provide appropriate estimates of parameters even if the multivariate normal assumption is violated. Examples of these methods are asymptotically distribution-free (ADF) estimation, which adjusts its results for the degree of kurtosis in the data (e.g., Browne, 1984), and methods that are based on elliptical distribution theory, which requires only symmetrical distributions (e.g., Bollen, 1989b; Bentler & Dijkstra, 1985).

Structural assumptions set up a model of interrelationships among observed and/or latent variables, and imply a specific structure for the covariance matrix of vector of observed variables (Satorra, 1990). All methods of estimation make the structural assumption that the structure tested in the sample accurately reflects the structure that exists in the population (Curran, West & Finch, 1996). The major purpose of the typical goodness-of-fit indices is to reflect the degree of misspecification in the model. Violation of both distributional and structural assumptions is common and often unavoidable in practice. Although the impact of violating the distributional assumptions on evaluating model fit in SEM has been a focal point of many studies, much less is known about violations of the structural assumptions.

Two basic kinds of SEM models could be distinguished, recursive and nonrecursive. Of the two, recursive models are the most straightforward and have two basic features: their disturbances are uncorrelated, and all causal effects are unidirectional. Nonrecursive models

have feedback loops or may have correlated disturbances. The distinction between recursive and nonrecursive models has several implications, some conceptual and others practical. The assumptions of recursive models that all causal effects are unidirectional and that the disturbances are independent when there are direct effects among the endogenous variables greatly simplify the statistical demands for their analysis. On the other hand, nonrecursive models require more specialized statistical methods of estimation, and may also require more specialized assumptions. Also, the likelihood of a problem in the analysis of a nonrecursive model, such as problems of identification, is much greater than for a recursive model (Kline, 1998). Perhaps due to such difficulties associated with specifying nonrecursive models not present with recursive models, few nonrecursive models exist in the social science literature (Berry, 1985; Kline, 1998).

2. ASSESSMENT OF MODEL FIT IN SEM

The fundamental hypothesis for structural equation procedures is that the covariance matrix of the observed variables is a function of a set of parameters. If the model is correct and if we know the parameters, the population covariance matrix would be exactly reproduced (Bollen, 1989a):

$$\Sigma = \Sigma(\theta) \tag{1}$$

where Σ is the population covariance matrix of observed variables, θ is a vector that contains the model parameters, and $\Sigma(\theta)$ is the covariance matrix written as a function of θ (Bollen, 1989b).

In practice, a specified model is fitted to a sample covariance matrix, *S*. For any selected vector of parameter estimates, $\hat{\theta}$, a reproduced or implied population covariance matrix, $\hat{\Sigma}$ could be obtained:

$$\hat{\Sigma} = \Sigma(\hat{\theta}) \tag{2}$$

The objective in parameter estimation is to find $\hat{\theta}$ so that the resulting covariance structure implied by the specified model, $\hat{\Sigma}$ is as similar as possible to *S*. The difference between $\hat{\Sigma}$ and *S* is measured by an estimated discrepancy function, $\hat{F}$ which takes a value of zero only when *S* = $\hat{\Sigma}$ and otherwise is positive, increasing as the difference between *S* and $\hat{\Sigma}$ increases (MacCallum, Browne & Sugawara, 1996). Estimation of the parameters of the model involves minimizing this discrepancy function, thus the magnitude of $\hat{F}$ reflects the degree of lack of fit of the specified model to the sample data. The most common discrepancy function is the maximum likelihood (ML). Other common discrepancy functions include generalized least squares (GLS) and asymptotic distribution free (ADF). The test statistic $T = (n - 1)F_{min}$ has an asymptotic χ^2 distribution if the specified model is correct in the population with d degrees of freedom, where d equals the number of distinct parameters to be estimated subtracted from the

total number of observation. Therefore, the test statistic *T* is often called "the χ^2 test." (Hu & Bentler, 1995).

Because SEM is used to test the fit between a theoretical model and empirical data, there must be mechanisms to determine how adequately the model accounts for the data. A wide array of fit indices have been proposed for evaluating the fit of SEM models. Fit indices quantify the degree of correspondence between a hypothesized model and the data (Hu & Bentler, 1995).

Initially, only the p value associated with the likelihood ratio χ^2 statistic was used to evaluate fit, under the null hypothesis that the population covariance is identical with those predicted from the model estimates (Gerbing & Anderson, 1993). Since a null hypothesis that a model fits exactly in some population is known a priori to be false, it seems pointless even to try to test whether it is true (Browne & Cudeck, 1993; MacCallum, Browne & Sugawara, 1996). If the sample size is sufficiently large in a practical investigation, it can be expected that even models that closely approximate the population covariance matrix will be rejected.

Because of the problems related to χ^2 test for model fit assessment in SEM such as the abovementioned dependency of sample size and lack of information regarding the degree of fit, Bentler and Bonett (1980) introduced and popularized several alternative measures of fit, so-called fit indices. A fit index is an overall summary statistic that evaluates how well a particular SEM model explains the sample data. Rather than testing if the model fits the data exactly, fit indices test if the model as specified is an approximation to reality. Many investigators (e.g., Cudeck & Browne, 1983) argued that it is preferable to depart from the unrealistic assumption that the model will fit the data exactly. Instead, Cudeck and Browne, 1983 proposed that any given target model "be regarded as one of many formulations for describing behavioral theory, some of which are reasonable" (p.50).

The use of fit indices as alternative measures of model fit became very popular with time, however, none of these fit indices have been endorsed as the "best index" by the majority of researchers. To decide on the appropriateness of the model, the substantive researcher must sort through and understand the meaning of the values for these various indices (Gerbing & Anderson, 1993). This led to a real challenge for applied researchers in selecting appropriate fit indices among the large number of fit indices available in many popular SEM programs (Hu, & Bentler, 1998).

Moreover, as the number of fit indices increased, it became apparent that it is essential to develop classification systems that can aid researchers in choosing appropriate fit indices for their studies. The different indices have been classified in a number of ways. One of the most widely cited and followed classification schemes for fit indices is the absolute versus incremental distinction of fit indices (Bollen, 1989b; Gerbing & Anderson, 1993; Marsh, Balla, and McDonald, 1988; Tanaka, 1993). This framework presents "stand-alone" or "absolute" indices followed by two different subtypes of "incremental" indices called Type-1 and Type-2 indices. Hu and Bentler (1995) further added a third group of incremental fit indices; called types-3 fit indices. One also could add to that framework yet another dimension for classifying fit, which would be labeled "adjusted" indices (Maruyama, 1998).

Absolute fit indices compare the observed covariance matrix to that estimated using the conventional discrepancy function (Maruyama, 1998). An absolute fit index directly assesses how well an a priori model reproduces the sample data. Although no reference model is employed to assess the amount of increment in model fit, an implicit or explicit comparison may be made to a saturated model that exactly reproduces the observed covariance matrix (Hu & Bentler, 1995). Some of the more commonly used absolute fit indices include goodness-of-fit index (GFI), adjusted goodness-of-fit index (AGFI) and standardized root mean squared residuals (SRMR) (Jöreskog & Sörbom, 1989).

Incremental fit index measures the proportionate improvement in fit by comparing a target model with a more restricted, nested baseline model (Hu & Bentler, 1995). It addresses the question: how well does a particular model do in explaining a set of observed data compared with other possible model/models? (Maruyama, 1998). The typical baseline model is an independence model, in which all the observed variables are assumed to have variances but are uncorrelated with each other (Kline, 1998).

A Type-1 index directly compares a given model to the fit of a more restricted baseline model. As defined by (Hu & Bentler, 1995), Type-1 index uses information only from the optimized statistics T, used in fitting baseline (T_B) and target (T_T) models. A general form of such indices can be written as Type-1 incremental index $= |T_B - T_T| / T_B$. Some widely used examples of Type-1 fit indices are Bentler-Bonett normed fit index (NFI) (Bentler & Bonett, 1980), and relative fit index (RFI) (Bollen, 1986).

Type-2 indices compare models but also include information from the expected value of the T statistics for the true model specification under a central chi-square distribution (Maruyama, 1998). It assumes that the test statistics T for the target model follows an asymptotic chi-square distribution with a mean equal to the degree of freedom for a target model. A general form of such indices can be written as a Type-2 incremental fit index = $|T_B - T_T| / (T_B - df_T)$, where df_T is the degrees of freedom for the target model (Hu & Bentler, 1995), which is the expected value of the statistic *T* for the true target model that is correctly specified so that there is no misspecification. There are several different Type-2 indices that are widely used. Some widely used Type-2 indices are Tucker-Lewis index (TLI) (Tucker & Lewis, 1973), Bentler-Bonett non-normed fit index (NNFI) (Bentler & Bonett, 1980), incremental fit index (IFI) (Bollen, 1989a).

A Type-3 index uses Type-1 information but additionally uses information from expected values of T_B , T_T or both, under the relevant noncentral chi-square distribution. It assumes that with true models or at least not extremely misspecified target models, the test statistic *T* can be approximated in large samples by the noncentral chi-square distribution. This type of fit index has not been as widely used in the SEM literature, as Type-1 and Type-2 indices (Maruyama, 1998). Examples of Type-3 indices include the relative noncentrality index (RNI) (McDonald, 1989; McDonald & Mrash, 1990), and comparative fit index (CFI) (Bentler, 1990; Hu & Bentler, 1995). Type-3 indices are assumed to be sample size independent.

Adjusted indices explicitly address the question: how does the model combine fit and parsimony? (Maruyama, 1998). Many models could fit the data if enough parameters were

estimated, so these indices penalize for lack of parsimony. For models that use a lot of degrees of freedom in model specification, the adjusted or parsimonious fit indices look worse than do the relative fit indices (Maruyama, 1998). Examples of this type of fit indices are parsimony goodness of fit index (PGFI) (James, Mulaik & Brett, 1982), and parsimony normed fit index (PNFI) (Mulaik, James, Van Alstine, Bonnett, Lind & Stillwell, 1989).

In addition to these types of fit indices, there are fit indices for comparing non-nested models. Researchers who have alternative models that cannot be nested are faced with a different challenge, since it is difficult to compare models that make different assumptions about patterns and relationships (Maruyama, 1998). Examples of this type of fit indices include Akaike information criteria (AIC) (Akaike, 1987), modified Akaike information criteria (CAIC), root mean square error of approximation (RMSEA) (Steiger & Lind, 1980; Steiger, 1990), and expected cross validation index (ECVI) (Browne & Cudeck, 1993).

Because fit indices were developed with different rationales and with different motivations (Gerbing & Anderson, 1993), they may differ on one or several dimensions. Tanaka (1993) proposed a six-dimension typology for SEM fit indices and attempted to categorize some popular fit indices along these six dimensions. This typology represents the multifaceted nature of fit indices which not only makes the comparison among fit indices difficult but also makes it very difficult to select the "best" index from all those available based on the theoretical rationales on which they are developed (Fan & Wang, 1998).

The dimensions were provided by Tanaka (1993) were "population- vs. sample-based", "simplicity vs. complexity", "normed vs. non-normed", "absolute vs. relative", "estimation method free vs. estimation method specific", and "sample size independent vs. sample size dependent" (p. 16).

Some attempts are made to define the criteria of the best fit index. For many researchers, the ideal fit index would (a) be relatively independent of sample size; (b) provide an accurate and consistent measure of difference in goodness of fit for competing models of the same data and for the same model applied to different data; (c) vary along an externally meaningful, well defined, absolute continuum such that its value can be easily interpreted; and (d) be replicable, that is, provide an indication of which model can be most successfully cross-validated when tested with new data (Marsh, Balla & McDonald, 1988). No one index has been shown to satisfy all of these conditions. Furthermore, not all researchers would even agree with all of these criteria. For example, Cudeck and Henly (1991) argue that sample size should affect fit.

Studying the behavior of fit indices under model misspecification is of particular interest given the high likelihood that the model estimated in the sample does not precisely conform to the model that exists in the population. Ideally, the extent to which a model is correctly specified or misspecified should be the primary determinant for model fit assessment. Therefore, the degree of model misspecification should be the major contributor to the variation of the value of the fit index. The performance of fit indices under misspecification is a fundamental criterion for identifying a good fit index. In reality, there exist a few confounding factors that affect the

performance of SEM fit indices such as data normality, the estimation methods used in SEM analysis, and the sample size (Fan & Wang, 1998).

Few studies have investigated the sensitivity of SEM fit indices to model misspecification. Previous recommendations on the adequacy of fit indices have been primarily based on the evaluation of the effect of sample size or the effect of estimation method, without taking into account the sensitivity of an index to model misspecification (Hu & Bentler, 1998). The technique used exclusively in the early studies of the evaluation of SEM indices has focused on correctly specified models only (Anderson & Gerbing, 1984; Bearden, Sharma & Teel, 1982; Boomsma, 1982, 1985). These analyses demonstrate how well the indices indicate correct specification, but they do not show the ability of fit indices to discriminate good fitting from badly fitting models. Yet, at least as important as correctly indicating true fit is the ability of the index to distinguish between the lack of fit due to sampling variability and that due to misspecification (Gerbing & Anderson, 1993). Furthermore, under the true model, many sample fit indices have a ceiling of about 1.00, and therefore, studying true models with these statistics may have artificially created ceiling effects that mask performance differences among different fit indices (Fan & Wang, 1998). Few studies have included misspecified models as well as properly specified models (Fan, Wang & Thompson, 1999; Fan & Wang, 1998; Hu & Bentler, 1998; Hu & Bentler, 1999; Marsh, Balla & Hau, 1996; La Du & Tanaka, 1989).

Most previous studies that investigated model misspecification have addressed the effect of model misspecification on the computation of parameter estimates and standard errors (e.g. Curran, West & Finch, 1996; Kaplan, 1988,1989) or post hoc model modification (MacCallum, 1986). However, little is known about the behavior of SEM fit indices under violation of different types of misspecifications. Moreover, previous Monte Carlo investigations of the behavior of various fit indices considered the confirmatory factor analysis model as the test model. The problem with this tendency is that the confirmatory factor analysis model is not a typical of the models currently being estimated (Hayduk, 1996). Although common in practice, general structural equation models are rare in Monte Carlo simulations (Paxton, Curran, Bollen, Kirby & Chen, 2001).

In this study, the effect on fit indices is examined of two types of model misspecification; namely nonrecursive and recursive misspecification. Useful information, such as to which type of model misspecification, fit indices were more sensitive, could be obtained. As far as I can tell, essentially no previous study comparing the sensitivity of fit indices to these types of model misspecification.

The purpose of the present study is to investigate empirically the effects of model misspecification, misspecification type and sample size on several alternative but commonly used SEM fit indices derived from ML estimation method. Moreover the effect of the interaction among degree of model misspecification, sample size, and misspecification type is also evaluated. Accordingly, fit indices that performed best are identified and recommended for use in practice. The performance of fit indices was examined over four levels of model misspecification ranging from no error to high misspecification, four levels of sample size

ranging from 100 to 1000, and two types of misspecified models. The first type of model misspecification involves specification of the target model as recursive, whereas the correct model specification is nonrecursive. Specification of the target model as nonrecursive as the correct model specification, identified the second type of model misspecification considered in the study.

The study is conducted within the context of general structural equation modeling representation rather than the confirmatory factor analysis model, and investigates the sensitivity of SEM fit indices to misspecification in the structural sub-model. The models are studied under a variety of systematic degrees of misspecification. Three levels of misspecification, in addition to the correct specification are used in this study. Findings from the present study are intended to provide the researcher further insight into the behavior of fit indices to detect model misspecification under different types of misspecification, and the role sample size plays with degree of misspecification, type of misspecification, or both in affecting the value of fit indices. Moreover, evidence regarding the efficacy of cutoff values of fit indices is provided (Hu & Bentler, 1998; 1999).

To increase the external validity, the study investigates models for analysis that are based on published substantive research. As suggested by Gerbing & Anderson, 1993 "defining the models studied by Monte Carlo methods according to the characteristics of published models would result in greater generality of results for substantive researchers, as well as contribute to the comparability of knowledge accumulated across multiple studies" (p.62).

The present study is aimed to address the following questions: which SEM fit index (-es) is more sensitive to structural sub-model misspecification? Is the value of fit indices affected by the type of the model misspecification (recursive/nonrecursive)? Which SEM fit index (-es) is more precise with low sampling fluctuations? How does the interaction between degree of model misspecification and type of model misspecification affect the value of fit indices? Is the effect of degree of model misspecification and misspecification type on the value of fit indices is conditional upon the sample size?

The study investigates the behavior of ten commonly used fit indices; namely Goodness-of-Fit Index (GFI; Jöreskog & Sörbom, 1981), Standardized Root Mean-square Residuals (SRMR; Jöreskog & Sörbom, 1981), McDonald's Centrality Index (MCI; McDonald, 1989), Root Mean Square Error of Approximation (RMSEA; Steiger, 1990), Normed Fit Index (NFI; Bentler & Bonett, 1980), Relative Fit Index (RFI; Bollen, 1986), Incremental Fit Index (IFI; Bollen, 1989a), Non-Normed Fit Index (NNFI; Bentler & Bonnet, 1980), Relative Noncentrality Index (RNI; McDonald and Marsh,1990) and Comparative Fit Index (CFI; Bentler, 1990). Table 1 shows the defining equations and abbreviations for all fit indices considered in the study.

3. METHOD

3.1 Population Model Used in the Study

A key step in designing a Monte Carlo experiment is to create a model that is representative from an applied point of view, manageable, and answering relevant, specific research questions (Paxton et al., 2001). A modified version of a well known SEM model of peer influence proposed by Duncan, Haller and Portes (1971) is chosen as our population model from which the simulated data is generated. To increase the external validity of Monte Carlo research results, Gerbing and Anderson (1993) and Paxton et al. (2001) recommend the use of simulating models that resemble those in published research. The original model of Duncan et al. (1971) is a general structural equation model that contains two latent variables and ten observed variables. The model is nonrecursive due to the reciprocal paths between the two latent variables.

To accommodate the different levels of misspecification and answering the specific research questions, the model is modified to include a third, dependent construct with two more observed variables (indicators) by which the added latent variable is measured. Three paths from the exogenous observed variables to the added latent variable are included in the model. The modified model then contains 12 measured variables and three latent factors with two indicators per factor. The causality among variables make the model a general structural equation model rather than confirmatory factor analysis model. Causality among variables is common in published research but rare in SEM Monte Carlo simulations (Paxton et al., 2001). Both the original and the modified models have been researched in many studies (e.g., Jöreskog, 1979; Jöreskog & Sörbom, 1989; Farley & Reddy, 1987). Based on the above theoretical model, the fully specified population model used in the present study is displayed as Figure 1. The observed variables are indicated by Vs, latent variables are indicated by Fs, measurement errors of the observed variables are indicated by Es, and disturbances of the latent variables are indicated by Ds.

Parameter values are carefully selected using Wright's rules of paths tracing and decomposing relationships between variables into causal and noncausal components (Maruyama, 1998),and following recommendations of Paxton et al. (2001). In addition, parameter values were selected to ensure proper parameter estimation, to manage the misspecification in such a way that the omission of one or more parameters would result in the desired degree of misspecification. Precautions were taken to make sure that all specifications of the model were identified using the rules of identification (Kline, 1998). The true population variance/covariance matrix for the target population model is obtained using the procedure described by Jöreskog & Sörbom (1989, P. 211-213) with LISERAL 8.53.

3.2 Misspecified Models

Although misspecification can occur when one or more parameters are estimated whose population values are zeros (misspecification of inclusion), as well as when one or more parameters are fixed to zero whose population values are nonzeros (misspecification of exclusion), or in both situations; the present study only examines the sensitivity of fit indices to

misspecification of exclusion only. Therefore model misspecification is achieved in this study by fixing certain structural paths to zero in the model that should have been freed.

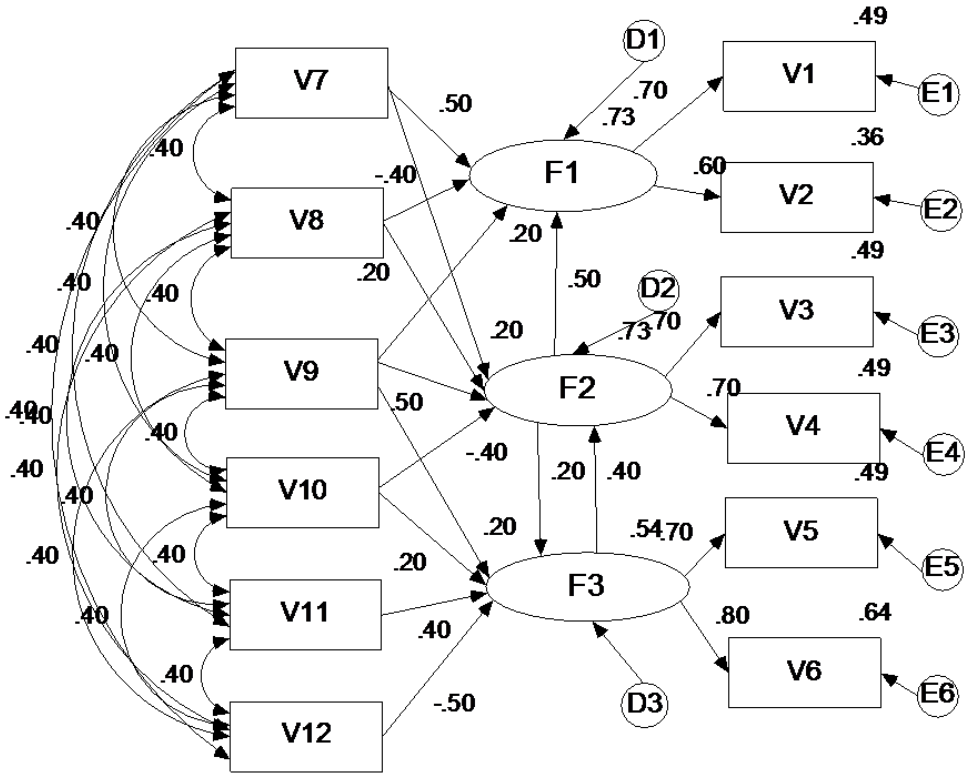


Figure (1) The true population model specification and parameter values

The decision to limit the study to only this kind of misspecification error was based on the difficulty of isolating the effects of different types of misspecification on the fit index value. Moreover, this type of error is chosen because it is thought to be a frequent type of specification error in practice. In application of SEM, researchers often attempt to force simple structure on their models because of the obvious interpretation simplicity (Hutchinson, 1998). Studies by Farley and Reddy (1987), Kaplan (1988), and La Du and Tanaka (1989) have shown that errors of omission are much more serious than errors of inclusion. Hu & Bentler (1999) noted that misspecification models due to inclusion of parameters whose population values are zeros have zero population noncentrality and therefore do not have significantly different estimates for model fit indices. In the context of multiple regression, Pedhazur (1982) showed that omitting relevant variables from the regression model (misspecification of exclusion), biases the estimation of the regression coefficients of the variables in the model, however including irrelevant variables in the model (misspecification of inclusion) does not affect the estimation of the regression coefficients.

In the present study, degree of misspecification refers to the number and size of paths that have been omitted from the correctly specified models. Systematic variation of the degree of model misspecification is achieved by varying the degree to which the fitted model resembles the true model. Four specifications ranging from a correct model to a highly misspecified model are utilized. The misspecification of the model consists of a nonrecursive or recursive misspecification. This last type of misspecification is thought to be a frequent specification error in the social sciences, given the difficulties associated with specifying nonrecursive models (Kline, 1998). Our study design includes these two types of model misspecification. For each of these two types, three misspecification levels are imposed on the correct model: slight misspecification, moderate misspecification, and high misspecification. The three levels of misspecification differ in terms of number and size of omitted paths, and in terms of power to detect misspecification.The true model specification was labeled as M0, whereas, the misspecified model specifications were named M1, M2, and M3, which represent the slightly misspecified model, the moderately misspecified model, and the highly misspecified model, respectively.

For both types of model misspecification, M0 is the properly specified model, that is, the estimated model matches the population model. For nonrecursive misspecification, M1N omits the paths from V7 and V9 to F1 and V9 to F3; M2N additionally omits V8 and V10 to F2; and finally, M3N additionally removes V12 to F3. For recursive misspecification, M1R omits the paths from F2 to F3 and from V7 and V9 to F1; M2R additionally omits V8 to F1 and V10 to F3; and finally, M3R additionally removes V9 to F2. Paths were selected for omission such that the misspecified models could be legitimate specifications of the model and are common in practice. For example, recursively misspecified models were specified to represent a chain of causality among the latent variables. Paxton et al. (2001) noticed "chains of causality among latent variables are common in published research but rare in structural equation model Monte Carlo simulations" (P.292).

To keep the size of specification error equivalent for the two types of misspecification over the degree of misspecification (M1, M2, M3), and to avoid confounding misspecification type with degree of misspecification, omitted paths are matched in terms of both the size of unstandardized and standardized population path coefficients for the two types of misspecification. That is, under the same level of misspecification, the omitted paths in the two types of misspecified models were selected so as to have the same population path coefficient size. In addition, power to detect misspecifications was approximately equivalent for the two types of misspecification over degrees of misspecifications across all sample sizes studied. Power is calculated in our study for each misspecified model under the various sample size conditions using the procedure outlined by Saris & Satorra (1993).

3.3 Study Design

A three-factor balanced experimental design was used in our study. Four levels of sample sizes (100, 200, 500, and 1000), four levels of model misspecification (correctly specified, slightly

misspecified, moderately misspecified, and highly misspecified), and two types of model misspecification (recursive and nonrecursive misspecification) were incorporated in a 4 x 4 x 2 factorial design with repeated measures over degree and type of misspecification. Such a design allows a systematic assessment of the impact of the three factors and the interaction among them on fit indices. Data were analyzed using the GLM repeated measures procedure in SPSS 17 (SPSS, 2008).

Sample sizes chosen for our study ranged from small (100) to large (1000) relative to the size of the model. Gerbing and Anderson (1993) suggested that at least 100 or more replications per cell in the design are needed to provide an accurate estimate of the population values. In the present study 500 replications (samples) in each cell condition were drawn from the known population model. This number of replications would thus be considered large enough to accurately calculate statistics. The aim of producing such a large number of replications is to improve the precision of the study (Skrondal, 2000).

3.4 Data Generation

Simulated raw data were generated in EQS 6.1 (Bentler, 2008) as random draws from the population covariance matrix. EQS has a built-in simulation procedure for generating normally distributed samples based on a population matrix. Five hundred raw data sets for each sample size were created. For each sample generated under specific sample size, one of the model misspecifications (M0, M1N, M2N, M3N, M1R, M2R, or M3R) is estimated for that sample. That is, the data for a given sample size is generated based on the true population covariance matrix, followed by the testing of the goodness-of-fit between a target model and the generated data. During the course of fitting the models to the data, some data sets may present convergence difficulties, that is not converge, or converge to improper solutions. Improper solutions refer to estimates that take on values that would be impossible for the corresponding parameters, such as outside the constraints. These primarily take the form of a correlation greater than one or constrained at one or a variance that is negative or constrained to zero (Chen, Bollen, Paxton, Curran, & Kirby, 2001).

In the present study, a sample was regarded as invalid if after 100 iterations; the estimation did not converge, or converged to improper solution. To achieve convergence as quickly as possible for each replication, population parameters were used as starting values. Starting values were needed because EQS estimates the model and produces fit statistics as part of its data generation. All invalid samples were discarded in the course of the final Monte Carlo analysis. Because it is good to use the same number of replications for all treatments, our goal was to generate 500 good replications at each sample size for all model specifications. The obvious advantage of using the same number of replications for all treatments is to have a balanced design. Under such a balanced design, the effects of violating the assumptions of ANOVA such as distributional assumptions, are minimized. To obtain 500 good replications, 750 raw data sets were initially generated for sample sizes of 200 through 2000, and 900 raw data sets were generated for sample size 100 where invalid samples are more likely. The first 500 replications that converged and provided proper solutions at each sample size over the misspecification levels and types were kept. This strategy resulted in 500 good samples for all five sample size levels.

4. RESULTS

4.1 Preliminary Analyses

A preliminary examination of the behavior of each fit index was based on the mean distances between the observed fit index values for each of the model specifications (Hu & Bentler, 1998) and the corresponding population fit index values, as well as the correlations among the fit indices. Population values of goodness-of-fit indices obtained from each specification of the model were calculated by fitting each model to the population covariance matrix with n = 100,000.

The overall mean distances were calculated for each fit index. The purpose for calculating these distances was to determine how much each index might depart from its value under the correct and incorrect specifications of the model. The Overall mean distance (OMD) is calculated using the following formula (Hu & Bentler, 1998):

$$OMD = \sqrt{\Sigma(O - P)^2 / N} \qquad (3)$$

where O is the observed fit index value, P is the population fit index value, and N is the total number of observed fit indices.

Analysis of these distances showed that for the correctly specified and slightly misspecified models, the values of CFI, IFI, RNI, MCI, RMSEA and SRMR were closer to their corresponding population values than the other fit indices. For higher misspecification levels, values of SRMR, RMSEA, CFI, RNI and IFI were closer to their corresponding population values than the other fit indices.

Information regarding means, variability and correlations with sample size for each of the 10 fit indices is presented in Table 2. Inspection of these results shows that values of GFI, RMR, SRMR, NFI and RFI are considerably related to sample size under all misspecification levels for both types of misspecification. However, the effect of sample size on these five fit indices becomes somewhat less serious as the degree of misspecification increases. MCI, NNFI and RNI are relatively independent of sample size at all misspecification levels for both types of misspecification. Correlations between RMSEA, IFI and CFI and sample size suggest that the effect of sample size on these fit indices depends on the degree of misspecification. For the true-population specification of the model, RMSEA and CFI shows moderate correlation with sample size. For the slightly misspecified models, correlation between RMSEA and CFI and sample size are lower but still relatively high, especially for the recursive misspecification type. The correlation with sample size for RMSEA and CFI decreases with increasing degree of misspecification. Whereas sample size affected RMSEA and CFI at the true model specification, it affected IFI only at the high recursive misspecification.

Examining the mean and variability of the sampling distribution for different fit indices revealed that values of all fit indices were responsive to the change in the degree of model misspecification. For the true-population specification of the model, RNI, IFI, NNFI, and MCI had mean values of at least .999 which indicates nearly perfect fit. CFI and RMSEA have mean values of .997 and .011 respectively. Whereas the standard deviation of CFI, RNI and IFI were as low as .01, NNFI, RMSEA and MCI had relatively higher standard deviations (approximately .02) at the true-population model specification.

A statistic is an unbiased estimate of a parameter if the expected value of the sampling distribution of the statistics is equal to the parameter of which it is an estimate (Winer, Brown & Michels, 1991). Generally, given the population values of the fit indices, the mean of the sampling distributions for IFI, NNFI, RNI and CFI were the least biased among the studied fit

indices under the examined types and model misspecifications. The mean of the sampling distributions for NFI, RFI, and GFI were the most biased.

For both types of misspecification, values of NNFI, RFI, MCI, and RMSEA degraded substantially in terms of fit with increasing degree of misspecification. Mean value of NNFI dropped considerably from a value as high as 1.00 at the true specification of the model to as low as 0.766 and 0.802 for the high misspecification for the nonrecursive and recursive mis-specification, respectively. The mean value of MCI dropped noticeably from 1.00 for the true specification to .780 and .811 for the high misspecification for the nonrecursive and recursive misspecification, respectively. Although values of RFI degraded considerably as the degree of model misspecification increased, they showed relatively high standard deviations for all model misspecifications. Moreover, the sampling distribution mean value for RFI at the true-population model specification was biased downward with mean values of .941, .954 respectively. The mean of the sampling distribution for RMSEA increased from .01 under the true-population model specification to .116 and .106 for the high misspecification for the nonrecursive and recursive misspecification, respectively, with relatively low standard deviations.

In terms of type of misspecification, all fit indices showed better fit for the recursive than the nonrecursive misspecification type. However, the difference between the mean values of the fit indices for recursive and recursive misspecification varied according to the level of misspecification. For the slightly misspecified models, the difference was small for most fit indices. As misspecification increased, the difference in the mean of the sampling distributions of fit indices at both recursive and nonrecursive misspecification increased.

4.2 Sources Of Variation In The Fit Indices

A series of 4 (sample size) x 4 (misspecification levels) x 2 (misspecification types) ANOVA experiments were performed on each fit index to examine the potential effects of degree of model misspecification, misspecification type, and sample size on fit indices. Due to the large number of observations, standard probability values were not particularly useful cutoffs for distinguishing between meaningful and inconsequential effects. Therefore, a measure of effect size, η^2, was reported and interpreted in this study (Tabachnick & Fidell, 2001). Anderson and Gerbing (1984) have suggested that an omega-squared of less than 0.03 is negligible even when F test is significant. For the purpose of this analysis and because omega-squared is always smaller than eta-squared (Tabachnick & Fidell, 2001), only an effect with η^2 of 0.05 or larger was used to identify important effects. Table 3 shows partial η^2 for the different sources of variation for each fit index. As these results show, the three-way interactions for all fit indices had partial η^2 less than .05. The small amount of variation accounted for by the three-way interactions implies that the results can be safely interpreted by considering the three two-way interactions.

Interaction effects of level and type of misspecification: An inspection of Table 3 indicates that the effect size for level of model misspecification by type of misspecification was significant for all ten fit indices. High values of η^2 for this interaction suggest that the effect of type of misspecification on the fit index value depends on levels of model misspecification. Therefore, interpretation of type of misspecification without considering degree of misspecification is problematic. Since both the main effects of degree of misspecification and type of

misspecification yielded large effect sizes (η^2), investigating these interactions might be better approached using interaction contrasts (Tabachnick & Fidell, 2001).

Interaction contrasts for each fit index were constructed by comparing each level of misspecification except the true specification, to the previous levels of misspecification for both recursive and nonrecursive type of misspecification. The purpose of these interaction contrasts is to determine at which level of misspecification the difference in the mean value of the fit index between the two types of misspecification account for a significant amounts of interaction variance. In addition, these contrasts are orthogonal, which facilitates the interpretation of the outcomes. Analysis of these results revealed that a substantive amount of interaction between degree of model misspecification and misspecification type variance were attributable to the moderate and high degrees of model misspecification for all fit indices except RMSEA. In other words, the difference between the mean value of the fit index for recursive and nonrecursive misspecifications was stronger at the high and extensive degrees of misspecification. This may suggest that with the exception of RMSEA, fit indices were relatively less sensitive to recursive/ nonrecursive type of misspecification at low degree of misspecification and became more sensitive to type of misspecification as the specification error in the model increase. RMSEA was the only fit index that demonstrated sensitivity to type of misspecification at a low degree of misspecification. In general, we found that all fit indices yielded less error for the recursive misspecification above the equivalent nonrecursive misspecification. The difference in fit between the two types of misspecification increased as degree of misspecification increased.

Interaction effects of level of misspecification and sample size: Two fit indices showed significant interaction between level of model misspecification and sample size: SRMR and RMSEA. This might indicate that the main effects of these two factors did not adequately predict the variability among the cell means for SRMR and RMSEA. To investigate at which levels of misspecification most of the interaction variance could be accounted for, follow-up contrasts were used. Our results indicated that for SRMR, the sample size has a minimal effect when no or low degree of misspecification was present. However as misspecification in the model increased, sample size started to have a significant effect on the value of SRMR. On the other hand for RMSEA, sample size affected the value of RMSEA only at no or low specification error. Although, RMSEA has no significant main effect of sample size (see Table 3) − which is not the case for SRMR, sample size affects RMSEA at certain levels of misspecification. In other words, the effect of sample size on values of RMSEA appeared to be conditional upon level of model misspecification. RMSEA is associated with sample size when there is no or low specification error. Nevertheless, the effect of sample size diminished as the degree of model misspecification increased.

Interaction effects of type of misspecification and sample size: The interaction effect of type of misspecification and sample size for SRMR could be considered statistically reliable (η^2 = .065, Table 3). This significant interaction indicated that the main effects of both the type of misspecification and sample size could did not adequately predict the variability among the SRMR cell means. Inspection of the mean values of SRMR showed that as sample size grows bigger, the difference between the mean value of SRMR at recursive and nonrecursive misspecifications increased. The interaction effect of type of misspecification and sample size for SRMR was further investigated using the interaction contrasts comparing the recursive versus the nonrecursive misspecification types at the mean of the lowest two sample sizes (n=100 and

200) versus the highest two sample sizes (n=500 and 1000). Result of the interaction contrast analyses indicated that the difference between the mean values of SRMR for recursive and nonrecursive misspecification types under relatively small sample sizes (≤ 200) is significantly lower than the corresponding difference at larger sample sizes (> 200).

Main effects: The main effect of level of model misspecification accounted for a substantive amount of fit indices variation. Over the sizeable range of degrees of misspecification, all fit indices showed reasonable sensitivity to model misspecification. Mean values of fit indices decreased as level of specification error increased. The main effect of type of model misspecification also accounted for a large amount of variance for all fit indices. However, all fit indices yielded less error for recursive than nonrecursive misspecifications under the same conditions of level of misspecification and sample size. However, mean values of RMSEA, MCI, and CFI showed the lowest discrimination between recursive and nonrecursive misspecification types under the same conditions of level of model misspecification and sample size. The main effect of sample size was associated with substantive η^2 for GFI, SRMR, NFI and RFI.

4.3 Reliability Of Estimation And Sampling Fluctuations

An important characteristic of incremental fit indices is the precision of estimation and the relative lack of sampling fluctuations. The approach normally used in Monte Carlo studies to represent this feature is to compare the within-cell variation. However, Marsh et al. (1996) showed that this approach for estimating and comparing relative precision in different fit indices is not appropriate because the different fit indices may vary along different metrics; hence, their within-cell variances are not comparable. Therefore, they suggested a more appropriate approach to evaluate sampling fluctuations by comparing standardized variance components instead of those associated with raw scores. One way to standardize variance components is to compute the variance explained relative to total variance. However, they noted that the standardized residual variance may not be appropriate because some of the variance explained is due to undesirable sample size effects in some indices. Therefore, they recommended using the variance associated with degree of model misspecification as the operationalization of the true score variance and hence this variance may be a better basis for comparison (Marsh et al., 1996). We used this approach in the present study for comparing the reliability of estimation for fit indices.

Table 4 shows the raw and standardized variance components due to degree of misspecification, sample size and residuals. In addition, the ratio of the degree of misspecification variance (the operationalization of the true score variance) relative to the sum of variance due to sample size (undesirable systematic bias) and the residual variance was computed and presented in Table 4. The higher this ratio, the better the fit index can be viewed with regard to reliability of estimation and relative absence sampling fluctuations.

Inspection of the total sum of squares for the different fit indices in Table 4 reveales that fit indices vary substantially in total variability. MCI, RFI, and NNFI have larger variability than the other fit indices, suggesting that the different fit indices varied along a substantially different metrics. Therefore, comparing the raw-scores variance components for evaluating reliability of estimation is not appropriate. Standardized variance components were computed by dividing the raw-score variance components by total score variability. A comparison of standardized residual variance components (SS_{RE} / SS_T in Table 4) reveals that it varies from .074 for SRMR to .138 for NNFI. However, as shown in Table 4, fit indices that demonstrate smaller residual variance show

higher standardized sample size variance, suggesting that part of the smaller residual variance of these fit indices might be due to the systematic sample size effect. Furthermore, Table 4 shows that the proportion of variance accounted for by degree of model misspecification relative to total fit index variability (SS_{DM} / SS_T) varies substantially from .584 for SRMR to .865 for IFI. Based on this approach for comparing reliability of estimation of fit indices, IFI, CFI, MCI were the best among all fit indices (SS_{DM} / SS_T was at least .861), followed by RNI, RMSEA and NNFI (SS_{DM} / SS_T was at least .846). The performance of SRMR, NFI, GFI, and RFI were the poorest among the ten fit indices studied.

These results were further confirmed using the ratio of the degree of misspecification variance relative to the sum of variance due to sample size and the residual variance ($SS_{DM} / [SS_{SS} + SS_{RE}]$). The differences among fit indices using this ratio were substantially larger than observed with the proportion of variance accounted for by degree of model misspecification relative to total index variability. This index ranged from 1.83 for RFI to 7.29 for IFI. As shown in Table 4, the six fit indices with the highest ratio among all fit indices. IFI (7.29) and CFI (7.26) were almost equivalent and slightly better than MCI (6.86) RNI (6.85), RMSEA (6.74) and NNFI (6.15), whereas the performances of SRMR (2.79), NFI (2.35), and particularly GFI (1.92), and RFI (1.83) were the poorest with ratios considerably lower than the best six fit indices.

5. DISCUSSION

This study sought to evaluate and compare the performance of various SEM goodness-of-fit indices under conditions of degree of model misspecification, type of model misspecification and various sample sizes. Based on a known population model used to generate simulated data, models are hypothesized to be true, trivially, moderately, and highly misspecified. Two types of model misspecification are considered: recursively and nonrecursively misspecified models. Results of this study showed that the sensitivity of fit indices to model misspecification was higher for nonrecursive than for recursive misspecification type. That is, when the true-population model is nonrecursive and the model is misspecified as recursive, fit indices were less sensitive to misspecification than if the model is misspecified but still nonrecursive under the same size of sample and misspecification conditions. Among all fit indices studied, SRMR showed the highest sensitivity to type of misspecification, whereas RMSEA was the least sensitive fit index to type of misspecification. Nevertheless for all fit indices, the effect of type of misspecification increased with increasing degree of misspecification.

With regard to sensitivity to model misspecification, all fit indices studied here were found to be sensitive to model misspecification. Previous studies that investigated the effect of model misspecification on the value of fit indices found that model misspecification contributed the most to the variation of all fit indices (e.g., Fan & Wang, 1998; Fan et al., 1999). However these studies found that the amount of variation accounted for by model specification, unlike the present study, varied substantially among fit indices. In addition to the type of models used in these past studies and how the model was misspecified, this might be due to the fact that the range of misspecified models on the continuum of degree of misspecification is limited compared to the present study. For example, Fan and Wang (1998) and Fan et al. (1999) defined two misspecified models in which they evaluate the sensitivity to model misspecification. Nevertheless, results of the effect of model misspecification in the present study are consistent

with results found by Marsh et al. (1996) who used correctly specified as well as several substantially misspecified models. Consistent with the present study, they found that degrees of model misspecification accounted for a large proportion of variance in all fit indices they studied; namely NFI, RFI, NNFI, IFI, RNI, and CFI.

Whereas fit indices showed high sensitivity to model misspecification, they were varied in their reliability of estimation and sampling fluctuations using the model misspecification as the operationalization of the true score variance. Except for MCI and RMSEA, all absolute fit indices and all Type-1 fit indices studied performed the poorest among all fit indices. All Type-2 and Type-3 fit indices studied here; namely IFI, NNFI, RNI and CFI, in addition to RMSEA and MCI yielded the highest reliability among all fit indices studied. Consistent with the present study's results, Marsh et al. (1996) found that IFI, RNI, CFI and NNFI were the best fit indices with regard to selectivity to misspecification and relative lack of sampling fluctuations. Fan & Wang (1998) and Fan et al. (1999) found that MCI and RMSEA were among the highest sensitive fit indices to model misspecification.

Although not all researchers agree that the influence of sample size is not necessarily desirable, most researchers have proposed that a systematic relation between sample size and the values of fit is undesirable. Results of the present study showed that the mean of the sampling distributions of all absolute fit indices except RMSEA and MCI, and all Type-1 fit indices, were systematically related to sample size under both true-population and misspecified models, whereas the means of the sampling distributions of IFI, NNFI, RNI. CFI and MCI were relatively independent of sample size. GFI, NFI, and RFI were downward biased especially at small sample sizes and true-population model specification. When *n* is small, it is unexpected to have a value close to 1.00 for these fit indices, even if a perfect model specification is tested. SRMR was positively biased at small *n* and correct specification. The results reported here regarding sample size effect are consistent with previous studies (Anderson & Gerbing, 1984; Bentler, 1990; Bollen, 1989a; Fan & Wang, 1998; Fan et al., 1999; La Du & Tanaka, 1989; Marsh et al., 1988; Marsh et al., 1996). Under the true-population model specification, Anderson and Gerbing (1984) found that the effect of sample size was substantial for GFI and NFI; and small for NNFI. Although they concluded that NNFI had much larger sampling fluctuations than did other fit indices, Marsh et al. (1996) noted that this conclusion is based on inappropriate method to assess sampling fluctuations, and subsequently suggested and used more appropriate methodology and found that sampling fluctuations in NNFI were similar or better to those in the other incremental fit indices. The methodology suggested by Marsh et al. (1996) is applied here and similar results were found in the present study.

Marsh et al. (1988) found that sample size has a substantive effect on all absolute and Type-1 fit indices, a conclusion replicated in the present study. They recommended the use of NNFI along with other 4 Type-2 fit indices because of their independence of sample size and sensitivity to model misspecification. Bollen (1989a) also noted the undesirable relation to sample size of NFI and suggested IFI to correct this problem. He indicated that IFI and NNFI were relatively unaffected of sample size, which is consistent with the present study results.

Furthermore, Marsh et al. (1996) found that RNI, NNFI, and IFI were relatively unrelated to sample size, whereas RFI and NFI were strongly related to sample size. IFI showed small bias at the smallest sample size (n=100) and the extensive model misspecification. McDonald & Marsh (1990) showed mathematically and that IFI is positively biased at small sample size and misspecified models and that the size of bias decreased as the degree of misspecification approached zero, a conclusion supported in the present study only with the extensive misspecification and sample size of 100.

Recently, using general structural equation models, Fan and Wang (1998) ,Fan et al. (1999) and Hu & Bentler (1998) found that sample size was related to the sampling distribution of GFI, RFI and NFI, whereas sample size had a small effect on CFI, NNFI, IFI and RMSEA. These results were consistent with the present study except for RMSEA. In the present study, although the main effect of sample size on RMSEA was negligible − consistent with previous studies, sample size still plays a role in the value of RMSEA. Sample size was found to affect RMSEA at no or slight misspecification particularly for recursive misspecification type and small sample size (e.g., $n \leq 200$). At relatively small n and no or slight specification error, the mean of sampling distribution of RMSEA was positively biased. Regardless of degree of model misspecification, RMSEA was less or not biased at high sample size; in addition, RMSEA seems independent of sample size at moderate or high misspecification levels at all sample sizes. Because this conclusion differs from previous research, the generalizability of this conclusion needs more evaluation. Moreover, results of the present study found that sample size not only affected SRMR, but also sample size interacts with both degree and type of misspecification for these two fit indices. As sample size increases the effect of misspecification type increase. In addition, the mean of the sampling distribution of SRMR was more positively biased at small sample sizes and true-population model specification than at large sample sizes and high degrees of misspecification.

The results of the study are limited by the conditions used in the study. Misspecification is achieved in this study by fixing certain structural paths to zero that should have been freed. Because misspecification can take a variety of forms, under different kinds of misspecifications, the results may not be consistent with the results obtained from this study. In addition, the present study will not be able to detect if the fit indices behave differently under different estimation methods or when the multivariate normality assumption is violated. It is well known that degree of model misspecification is not easily quantified, so it is difficult to make a priori prediction about severity of misspecification (Gerbing & Anderson, 1993). In the present study degree of model misspecification is empirically determined by varying the degree the fitted model resembles the true model. Although the number of removed paths, the true-population parameter size and the type of misspecified model are taken into account when manipulating the degree of misspecification, in the present study the terms slightly, moderately, highly, and extensively misspecified are used only to indicate different degrees of misspecification. The selection of misspecified models could still be a limitation to the present study. The study also

focused on single-sample situations, results of the present study may not applied in situations of multi-sample structural equation models.

6. RECOMMENDATIONS FOR USE OF FIT INDICES IN PRACTICE

GFI: GFI was sensitive to model misspecification, but also highly sensitive to sample size, biased downward at correct specification as well as at all misspecified models especially at small *n* and true specification. GFI was the poorest among fit indices with regard to reliability of estimation and sampling fluctuations. Therefore, GFI is not recommended for use in practical application of SEM.

SRMR: The present study showed that SRMR were positively biased by *n*, a conclusion consistent with the results of previous studies. In addition, the effect of sample size was more serious at small sample sizes and true-population model specification. Although the sensitivity of SRMR to model misspecification was comparable to other fit indices, its reliability of estimation was not as good as Type-2 and -3 fit indices but not as bad as most absolute and Type-1 fit indices. Nevertheless, SRMR can be recommended for use as a measure of local fit.

NFI and RFI: Results of the present study show that NFI is biased by *n*, a conclusion consistent with previous studies. Although Bollen (1986) developed RFI to overcome this problem with NFI, previous research as well as the present results show that RFI is also substantially biased by sample size. Whereas NFI and RFI are still widely used, they are typically not among the recommended fit indices.

IFI: In the present study, IFI was sensitive to model misspecification, but did not systematically relate to sample size at any model specification, did not show biasness at any sample size or model specification, had a small sampling fluctuations and was among the best indices with regard to precision of estimation. Current results show also that IFI was slightly related to *n* at high model specification particularly for the recursive misspecification type and was faintly biased by *n* at the smallest sample size and the high model misspecification. However, Marsh (1995) and Marsh et al. (1996) found that the adjustment for degrees of freedom in IFI is inappropriate in that it penalizes model parsimony and rewards model complexity and this undesirable property of IFI is more noticeable for small *n*, a conclusion not tested in the present study. In the present comparison, although IFI showed small bias at smallest sample size and high and extensive misspecification conditions, this bias was very small and could be considered no important in practice. Thus, IFI was successful in meeting the criteria considered in the study, and is recommended for use particularly under moderate to large sample sizes ($n > 200$).

NNFI: NNFI showed high sensitivity to model misspecification, high reliability of estimation, independence of sample size, and the mean of its sampling distribution was not biased at any sample size or model specification/misspecification. Sampling fluctuations in NNFI were comparable to other incremental fit indices, a conclusion consistent with Marsh et al. (1996). Although, Marsh et al. (1996) specified situations in which the behavior of NNFI is likely to be

unstable such as when the baseline model is true, it is less likely to encounter such these situations in practice. In addition, Bentler (1990), McDonald and Marsh (1990), Marsh et al. (1996) and Mulaik et al. (1989) demonstrated mathematically and empirically that NNFI appropriately penalized model complexity and rewarded model parsimony; a property deemed to be desirable and useful by many researchers (Bollen, 1989b, 1990; Gerbing & Anderson, 1993). For these reasons, NNFI is recommended for use.

RNI and CFI: RNI and CFI are the representatives of Type-3 incremental fit indices in this study. They showed high sensitivity to model misspecification and high reliability of estimation. Although both fit indices were not systematically related to sample size, RNI showed more relative independence on sample size than CFI. Under true-population model specification and slightly misspecified models, CFI showed some relation with sample size particularly under recursive misspecification (see Table 4). CFI is negatively biased (underestimate the population value) by *n* at small sample sizes, and the size of bias tends to decrease as the degree of misspecification increase. This small marginal correlation between CFI and *n* was not observed in RNI. It is recommended to report values of RNI and CFI for overall model data fit.

MCI: In the present study, MCI was the most successful among all absolute fit indices studied here. MCI was sensitive to model misspecification, has small sampling fluctuations and high reliability of estimation, and not systematically related to sample size, with a mean of approximately 1.00 for the true model specification at all sample sizes. Because MCI is a transformation of the rescaled noncentrality parameter, it is not expected to be systematically related to sample size and to be insensitive to model misspecification. Results of the present study regarding MCI were consistent with previous evaluation of the fit index (e.g., Hu & Bentler, 1998; Fan et al., 1999) .MCI is highly recommended to be utilized for SEM practice.

RMSEA: In the present comparison, RMSEA appropriately reflected systematic variation in model misspecification, and showed small sample fluctuations. Although the main effect of sample size on RMSEA was very small, which is consistent with the conclusions of previous studies (e.g., Fan & Wang, 1998; Fan et al., 1999; Hu & Bentler, 1998), RMSEA showed small relation to sample size under true-population model specification and slightly recursive misspecified models (see Table 4). RMSEA was positively biased (overestimate the population value) by *n* at small sample size, and the size of bias tends to diminish as the degree of misspecification increase. In the present study, the interaction effect of sample size and model misspecification was significant, suggesting that sample size had different effect on RMSEA under different degrees model misspecification. Sample size had an effect on RMSEA under the true specification and had lesser effect under the slight misspecification. As the degree of misspecification increases, no effect of sample size has been observed. Therefore RMSEA is not recommended for use at small sample sizes (e.g., $n \leq 200$) because it tends to over-reject truly specified models.

TABLE 1
Defining Equations and Abbreviations for Fit Indices Investigated in the Study

Fit Index	Abbreviation	Defining Equation
ABSOLUTE FIT INDICES		
Goodness-of-Fit Index	GFI	$GFI = 1 - [tr(\hat{E}^{-1}S - I)^2 / tr(\hat{E}^{-1}S)^2]$
Standardized Root-Mean-square Residual	SRMR	$SRMR = [2\Sigma\Sigma((s_{ij} - e_{ij})/(s_{ii}s_{jj}))/P(P+1)]^{1/2}$
McDonald's Centrality Index	MCI	$MCI = \exp(-1/2[(T_T - df_T)/(N-1)])$
Root-Mean-Square Error of Approximation	RMSEA	$RMSEA = \{\max.[(T_T - df_T)/(N-1), 0]\}^{1/2}$
INCREMENTAL FIT INDICES		
TYPE-1		
Normed Fit Index	NFI	$NFI = (T_B - T_T)/T_B$
Relative Fit Index	RFI	$RFI = [(T_B/df_B) - (T_T/df_T)]/(T_B/df_B)$
TYPE-2		
Non-Normed Fit Index	NNFI	$NNFI = [(T_B/df_B) - (T_T/df_T)]/[(T_B/df_B) - 1]$
Incremental Fit Index	IFI	$IFI = (T_B - T_T)/(T_B - df_T)$
TYPE-3		
Relative Noncentrality Index	RNI	$RNI = [(T_B - df_B) - (T_T - df_T)]/(T_B - df_B)$
Comparative Fit Index	CFI	$CFI = 1 - \max[(T_T - df_T), 0]/\max(T_T - df_T), (T_B - df_B), 0]$

Note, $\hat{E}$ = implied covariance matrix; S = sample covariance matrix; I = identity matrix; tr = trace of a matrix; P = number of observed variables; df = degrees of freedom; s_{ij} = observed covariances; e_{ij} = reproduced covariances; s_{ii} and s_{jj} = observed standard deviations; T_T = T statistics for the target model; T_B = T statistic for the baseline model; df_T = degrees of freedom for the target model; df_B = degrees of freedom for the target model; N = sample size.

TABLE 2. Correlations With Sample Size, Mean and Standard Deviation, for All Fit Indices at All Model Specifications

Index		True Specification			Slight Misspecification			Moderate			High Misspecification		
	Type	r	M	SD	r	M	SD	r	M	SD	r	M	SD
GFI	Nonre	0.709	0.982	0.018	0.646	0.954	0.019	0.609	0.933	0.020	0.582	0.907	0.020
	Rec.				0.653	0.956	0.019	0.647	0.946	0.019	0.610	0.919	0.019
SRMR	Nonre	-0.790	0.024	0.013	-0.567	0.048	0.010	-0.538	0.058	0.009	-0.447	0.077	0.008
	Rec.				-0.576	0.048	0.010	-0.614	0.050	0.010	-0.522	0.065	0.009
MCI	Nonre	0.033	0.999	0.021	-0.026	0.922	0.030	-0.019	0.865	0.035	0.000	0.780	0.038
	Rec.				-0.023	0.926	0.030	-0.052	0.894	0.031	-0.077	0.811	0.035
RMSEA	Nonre	-0.314	0.011	0.017	0.093	0.068	0.016	0.053	0.089	0.013	0.006	0.116	0.012
	Rec.				0.094	0.066	0.016	0.095	0.078	0.014	0.087	0.106	0.012
NFI	Nonre	0.709	0.972	0.027	0.645	0.930	0.029	0.601	0.896	0.031	0.545	0.844	0.032
	Rec.				0.648	0.932	0.029	0.626	0.913	0.030	0.552	0.863	0.030
RFI	Nonre	0.709	0.941	0.057	0.645	0.864	0.055	0.601	0.809	0.056	0.545	0.722	0.057
	Rec.				0.648	0.868	0.056	0.626	0.840	0.054	0.552	0.755	0.054
IFI	Nonre	0.031	1.000	0.010	-0.075	0.958	0.016	-0.081	0.925	0.020	-0.082	0.872	0.023
	Rec.				-0.070	0.961	0.016	-0.107	0.942	0.017	-0.153	0.892	0.022
NNFI	Nonre	0.030	0.999	0.023	-0.019	0.917	0.033	-0.005	0.859	0.038	0.023	0.766	0.044
	Rec.				-0.016	0.921	0.032	-0.040	0.891	0.033	-0.057	0.802	0.040
RNI	Nonre	0.030	1.000	0.011	-0.019	0.957	0.017	-0.005	0.923	0.021	0.023	0.869	0.025
	Rec.				-0.016	0.959	0.016	-0.040	0.941	0.018	-0.057	0.889	0.023
CFI	Nonre	0.309	0.997	0.008	-0.014	0.957	0.017	-0.005	0.923	0.021	0.023	0.869	0.025
	Rec.				-0.010	0.959	0.016	-0.039	0.941	0.018	-0.057	0.889	0.023

TABLE 3. Partial Eta-squared for the Different Sources of Variation for Each Fit Index

Source	GFI	SRMR	MCI	RMSEA	NFI	RFI	IFI	NNFI	RNI	CFI
Mis. Level x Mis. Type x Samp. Size*	.003	.032	.003	.007	.004	.004	.006	.003	.008	.008
Mis. Level x Mis. Type	.479	.654	.364	.333	.377	.376	.367	.358	.359	.394
Mis. Level x Samp. Size	.015	.321	.004	.155	.015	.091	.028	.008	.003	.029
Mis. Type x Samp. Size	.001	.065	.005	.006	.001	.001	.003	.005	.006	.005
Mis. Level	.968	.963	.963	.955	.967	.961	.962	.956	.960	.956
Mis. Type	.603	.694	.478	.454	.497	.498	.487	.477	.477	.478
Samp. Size	.782	.700	.001	.003	.719	.728	.011	.001	.001	.003

*Mis. Level = Degree of Model Misspecification, Mis. Type= Type of Model Misspecification, Samp. Size= Sample Size

TABLE 4. Raw and Standardized Variances (Sum-of-Squared Deviations) Attributable of Degree of Misspecification, Sample Size and Residual for All Fit Indices.

	Variance due to Degree of Misspc. (SS_{DM})	Variance due to Sample Size (SS_{SS})	Residual Variance (SS_{RE})	Total Variance (SS_T)	SS_{RE} / SS_T	SS_{DM} / SS_T	SS_{DM} / ($SS_{SS} + SS_{RE}$)
GFI	16.814	6.786	1.981	25.976	0.076	0.647	1.918
SRMR	7.784	1.972	0.814	11.038	0.074	0.705	2.794
MCI	144.810	0.010	21.098	168.245	0.125	0.861	6.860
RMSEA	36.432	0.046	5.360	42.454	0.126	0.858	6.739
NFI	48.869	14.792	5.993	70.486	0.085	0.693	2.351
RFI	142.305	56.493	21.187	223.219	0.095	0.638	1.832
IFI	47.665	0.040	6.497	55.125	0.118	0.865	7.292
NNFI	160.232	0.014	26.041	189.324	0.138	0.846	6.150
RNI	50.341	0.004	7.341	58.613	0.125	0.859	6.854
CFI	48.032	0.035	6.577	55.638	0.118	0.863	7.264

REFERENCES

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317–332.

Anderson, J., & Gerbing, D. W. (1984). The effects of sampling error on convergence, improper solutions and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, 49, 155-173.

Bearden, W. D., Sharma, S., & Teel, J. E. (1982). Sample size effects on chi-square and other statistics used in evaluating causal models. *Journal of Marketing Research*, 19, 425-430.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238-246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588-606.

Bentler, P.M., & Dijkstra, T. (1985). Efficient estimation via linearization in structural models. In P.R. Krishnaiah (Ed.), *Multivariate analysis VI* (pp.9-42). Amsterdam: North-Holland.

Bentler, P.M. (2008). *EQS Structural Equations Program*. Encino, CA: Multivariate Software.

Berry, W.D. (1985). *Nonrecursive Causal Models*. Beverly Hills, CA: Sage.

Bollen, K. A. (1986). Sample size and Bentler and Bonett's nonnormed fit index. *Psychometrika*, 51, 375-377.

Bollen, K. A. (1989a). A new incremental fit index for general structural equation models. *Sociological Research and Methods*, 17, 303-316.

Bollen, K.A. (1989b). *Structural equations with latent variables*. New York: Wiley.

Bollen, K.A., & Long, J. S. (1993). Introduction. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 1-9). Newbury Park, CA: Sage.

Boomsma, A. (1982). The robustness of LISERAL against small sample sizes in factor analysis models. In K.G. Joreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction (part 1)*. Amsterdam: North-Holland.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISERAL maximum likelihood estimation. *Psychometrika*, 50, 229-242.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37, 62-83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research*, 29, 468-508.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, 18, 147-167.

Cudeck, R., & Henly, S.J. (1991). Model selection in covariance structure analysis and the "problem" of sample size: A clarification. *Psychological Bulletin*, 109, 512-519.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1, 16-29.

Duncan, O. D., Haller, A.O., & Portes, A. (1971). Peer influence on aspirations, a reinterpretation. In H. M. Blalock (ed.), *Causal models in the social sciences* (pp. 105-127). Aldine-Atherton, Inc.

Fan, X., Wang, L. (1998). Effects of potential confounding factor on fit indices and parameter estimates for true and misspecified SEM models. *Educational and Psychological Measurement*, 58, 701-735.

Fan, X., Wang, L., & Thompson, B. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling*, 6, 56-83.

Farley, J. U., & Reddy, S. K. (1987). A factorial evaluation of effects of model specification and error on parameter estimation in a structural equation model. *Multivariate Behavioral Research*, 22, 71-90.

Gerbing, D. W., & Anderson, J. C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 40-65). Newbury Park, CA: Sage.

Hayduk, L.A. (1996). *LISREL Issues, Debates, and Strategies*. Baltimore: Johns Hopkins University Press.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Issues, concepts, and applications* (pp. 158-176). Newbury Park, CA: Sage.

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Issues, concepts, and applications* (pp. 76-99). Newbury Park, CA: Sage.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424-453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional versus new alternatives. *Structural Equation Modeling*, 6, 1-55.

Hutchinson, S. R. (1998). The stability of post hoc model modifications in confirmatory factor analysis models. *Journal of Experimental Education*, 66, 361-380.

Jöreskog, K. G. (1979). Structural equation models in the social sciences: Specification, estimation and testing. In K. G. Jöreskog & D. Sörbom (Eds.), *Advances in factor analysis and structural equation models* (pp. 105-127). Cambridge, MA: Abt Associates.

Jöreskog, K. G. & Sörbom, D. (1981). *LISERAL V: Analysis of linear structural relationships by the method of maximum likelihood*. Chicago: National Educational Resources.

Jöreskog, K. G. & Sörbom, D. (1989). *LISREL 7: a guide to the program and application* (2nd ed.). Chicago, IL: SPSS.

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research*, 23, 69-86.

Kaplan, D. (1989). Model modification in covariance structure analysis: Application of the expected parameter change statistic. *Multivariate Behavioral Research*, 24, 285-305.

Kline, R.B. (1998). *Principals and practice of structural equation modeling*. New York: The Guilford Press.

La Du, T. J., & Tanaka, S. J. (1989). Influence of sample size, estimation method, and model specification on goodness-of-fit assessments in structural equation models. *Journal of Applied Psychology*, 74, 625-636.

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, 100, 107-120.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure analysis. *Psychological Methods*, 1, 130-149.

Marsh, H. W. (1995). The Δ2 and χ^2 I2 fit indices for structural equation models: A brief note of clarification. *Structural Equation Modeling*, 2, 246-254.

Marsh, H. W., Balla, J. R., & Hau, K.-T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical properties. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 315-353). Mahwah, NJ: Erlbaum.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: Effects of sample size. *Psychological Bulletin*, 103, 391-411.

Maruyama, G. M., (1998). *Basics of structural equation modeling*. Newbury Park, CA: Sage.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification*, 6, 97-103.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin*, 107, 247-255.

Mulaik, S. A., James, L. R., Van Alstine, J., Bonnett, N., Lind, S., & Stillwell, C. D. (1989). Evaluation of goodness of fit indices for structural equation models. *Psychological Bulletin*, 105, 430-445.

Paxton, P., Curran, P.J., Bollen, K.A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, 8, 287-312.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd Ed.). New York: Holt, Rinehart, & Winston.

Sairs, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 181-204). Newbury Park, CA: Sage.

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika*, 54, 131-151.

Schumaker, R.E., & Lomax, R.G. (1996). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Erlbaum.

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, 35, 137-167.

SPSS for Windows (2008). *SPSS for Windows, Rel. 17.0*. Chicago : SPSS Inc.

Steiger, J. H. (1990). Structural model evaluation and modification: An internal estimation approach. *Multivariate Behavioral Research*, 25, 173-180.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Tabachnick, B.G., & Fidell, L.S. (2001). *Computer-assisted research design and analysis*. Needham Heights, MA: Allyn & Bacon.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structure equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1-10.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 634–645.

CALCULATING VALUE AT RISKS OF DELTA-GAMMA METHODS VIA GAMMA-POLYNOMIAL DENSITY APPROXIMATION TECHNIQUE

Hyung-Tae Ha

Department of Applied Statistics, Kyungwon University,
Sungnam-ci, Kyunggi-do,
South Korea, 461-701.
E-mail: htha@kyungwon.ac.kr

ABSTRACT

We provide a simple methodology for approximating Value-At-Risk (VAR) of delta-gamma method, which is well known method to obtain random returns of financial portfolios. VAR of a financial portfolio is simply quantiles of a loss distribution in statistical sense. Under the normality assumption of risk measures, delta-gamma method can be expressed as a general linear combination of non-central chi-square random variables. After expressing an indefinite quadratic form as the difference of two positive definite quadratic forms, one can obtain an approximation to its density function by making use of the transformation of variables technique. The main part of the proposed algorithm is approximating positive definite and indefinite quadratic forms in normal random variables by making use of gamma-polynomial density approximation technique. It is shown that the density function of a positive definite quadratic form can be approximated from its moments in terms of a product of gamma baseline density and a polynomial. A detailed step-by-step algorithm which is easy to implement is provided. The proposed approximants produce very accurate VaRs throughout the range of the distributions being considered. Some numerical examples illustrate the results.

Keywords: Value-At-Risk; Delta-Gamma Method; Approximation Algorithm; Moments; Quadratic Forms.

1. INTRODUCTION

Financial investors are taking great efforts to estimate the probability of large portfolio losses, which is caused by changes in the portfolio's risk factors during the holding period such as interest rates, currency exchange rates, stock prices, equities, commodities. An important concept for quantifying and managing portfolio risk is Value-At-Risk (VAR), which is defined as a quantile of the loss in portfolio value during a single period of time considered. If the

value of the portfolio at time t is $V(t)$, the holding period is t_0, and the value of the portfolio at time $t + t_0$ is $V(t + t_0)$, then the loss in portfolio value during the holding period is

$$\mathcal{L} \;=\; V(t) - V(t + t_0)\,. \tag{1}$$

Since the loss in portfolio consist of a large number of instruments, we first map by replacing the instruments by positions on a limited number of risk factors. The loss random variable can be expressed with a complex function of $g(\mathbf{X})$, which is the return for the portfolio or interest over the holding period of time, where $\mathbf{X}$ is assumed to be a multivariate normal random vector with zero mean vector and covariance matrix Σ. That is, $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0},\, \Sigma)$ and Σ is a symmetric positive $p \times p$ matrix. Then, the complex function of $g(\cdot)$ can be approximated by making use of the second order Taylor series approximation as

$$\mathcal{L} \;\approx\; d + \mathbf{a}'\,\mathbf{X} + \frac{1}{2} X\,\Gamma\,\mathbf{X} \tag{2}$$

where d is a scalar, $\mathbf{a}$ is a p column vector and Γ is a $p \times p$ matrix derived from the current portfolio positions. This is commonly called the delta-gamma approximation, see for instance Britten-Jones and Schaefer (1999), Jorion (1997), Glasserman *et al.* (2000), Jaschke (2002).

Several methods have been proposed to compute quantiles of the loss distribution defined in Equation (2), among others, Monte-Carlo simulation discussed by Glasserman *et al.* (2000), Saddlepoint approximations studied by Feuerverger and Wong (2000), Cornish-Fisher expansions investigated by Zangari (1996) and Jaschke (2002), Johnson transformations considered by Longerstaey (1996), and Fourier-inversion studied by Rouvinez (1997). Mina and Ulmer (1999) compare Johnson transformations, Fourier inversion, Cornish-Fisher approximations, and Monte Carlo simulation.

Those methods have their own merits and shortcomings. Monte Carlo simulation would be only way to proceed for extremely complicated models. But the enormous computational cost often is required to obtain accurate VAR estimates of the loss distribution in the region of interest because the portfolio may consist of a very large number of financial instruments and the large number of runs are required. The Edgeworth series approximation and Cornish-Fisher expansion might be good options for approximating a density function when the normal approximation does not provide enough accuracy. But if the target distributions to approximate do not have the similar behavior of normal distribution, the Edgeworth series approximation and Cornish-Fisher expansion often fail to provide good approximations. It should be noted that Delta-Gamma method doesn't have the same tail convergence rate with normal distribution. The saddlepoint approximation methods are usually quite accurate in the tail areas of the target density, but not for the middle range of the distribution of interest. As pointed out in Reid (1988), saddlepoint approximation techniques are not widely used in many scientific applications because it may not be easy to understand the concepts of the techniques and apply them in many types of situations although they are very accurate approximation tool in tail probability.

A novel and accessible approach is proposed in this paper to calculating VAR of Delta-Gamma method of the value of the financial portfolio. One may can re-express Delta-Gamma method, which is stated as a generalized quadratic forms in normal random variables, as an indefinite quadratic forms from simple algebraic operations. After expressing an indefinite quadratic forms as the difference of two positive definite quadratic forms, the density function of a positive definite quadratic forms can be approximated in terms of the product of a gamma density function and a polynomial. Such representations are based on the moments of a positive quadratic form, which can be determined from its cumulants by means of a recursive formula. Then, one can obtain an approximation to the loss density function by means of the transformation of variable technique. Finally, VAR, which is a quantile of loss distribution, can be obtained by numerical integration and solving qunatile equations that equate the loss distribution functions to a given probability. The proposed approximants produce very accurate percentiles over the entire range of the distribution. A convenient specially-designed approximation algorithm is also provided.

The rest of this paper is organized as follows. Section 2 introduces a brief introduction to Delta-Gamma method and its moments. A moment-based approximation algorithm, which is specially designed for calculating quantiles of indefinite quadratic forms in normal random variables, are proposed in Section 3. Numerical examples are presented in Section 4. Some relevant computational considerations are also discussed in Section 5.

2. DELTA-GAMMA METHOD AND ITS CUMULANTS

Delta-Gamma approximation by making use of the second order Taylor series approximation to the complex function of $g(\mathbf{X})$ may be written as

$$\mathcal{L} \approx d + \mathbf{a}'\mathbf{X} + \frac{1}{2}\mathbf{X}\,\Gamma\,\mathbf{X} \tag{3}$$

where $\mathbf{X}$ is the vector of returns over one time period for our risk factors, d is a scalar, $\mathbf{a}$ is a p column vector and Γ is a $p\times p$ matrix derived from the current portfolio positions. As derived in Mathai and Provost (1992), the moment generating function of the loss distribution is given by

$$\begin{aligned}
M_{\mathcal{L}}(t) &= E[e^{t\mathcal{L}}] \\
&= (2\pi)^{-p/2}|\Sigma|^{-\frac{1}{2}} \int_{\mathbf{x}} \exp(\frac{t}{2}\mathbf{x}'\Gamma\mathbf{x} - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}))\mathrm{d}\mathbf{x} \\
&= |I - t\Gamma\Sigma|^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}(\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu} - 2td) + \frac{1}{2}(\boldsymbol{\mu}+t\Sigma\mathbf{a})'(I-t\Gamma\Sigma)^{-1}\Sigma^{-1}(\boldsymbol{\mu}+t\Sigma\mathbf{a})\Big) \\
&= |I - t\Sigma^{\frac{1}{2}}\,\Gamma\Sigma^{\frac{1}{2}}|^{-\frac{1}{2}} \exp\Big(t(d + \frac{1}{2}\boldsymbol{\mu}'\Gamma\boldsymbol{\mu} + \mathbf{a}'\boldsymbol{\mu}) \\
&\quad + \frac{t^2}{2}(\Sigma^{\frac{1}{2}}\mathbf{a} + \Sigma^{\frac{1}{2}}\Gamma\boldsymbol{\mu})'(I - t\Sigma^{\frac{1}{2}}\Gamma\Sigma^{\frac{1}{2}})^{-1}(\Sigma^{\frac{1}{2}}\mathbf{a} + \Sigma^{\frac{1}{2}}\Gamma\boldsymbol{\mu})\Big)
\end{aligned} \tag{4}$$

where t can be an arbitrarily small number neighborhood of zero. The associated cumulant generating function is then given by

$$\begin{aligned}
K_{\mathcal{L}}(t) &= \log M_{\mathcal{L}}(t) \\
&= -\frac{1}{2}\log|I - tC'\Gamma C| + t(d + \mathbf{a}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\mu}'\Gamma\boldsymbol{\mu}) \\
&\quad + \frac{t^2}{2}(C'\mathbf{a} + C'\Gamma\boldsymbol{\mu})'(I - tC'\Gamma C)^{-1}(C'\mathbf{a} + C'\Gamma\boldsymbol{\mu}) \\
&= t(d + \mathbf{a}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\mu}'\Gamma\boldsymbol{\mu}) + \sum_{j=1}^{\infty}\frac{t^j}{2j}\mathrm{tr}(C'\Gamma C)^j + \sum_{j=0}^{\infty}t^{j+2}\Big(\frac{1}{2}\mathbf{a}'C(C'\Gamma C)^jC'\mathbf{a} \\
&\quad + \frac{1}{2}\boldsymbol{\mu}'\Gamma C(C'\Gamma C)^jC'\Gamma\boldsymbol{\mu} + \mathbf{a}'C(C'\Gamma C)^jC'\Gamma\boldsymbol{\mu}\Big) \\
&= t(d + \mathbf{a}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\mu}'\Gamma\boldsymbol{\mu}) + \frac{1}{2}\sum_{j=1}^{\infty}\frac{t^j}{j}\mathrm{tr}(\Gamma\Sigma)^j \\
&\quad + \sum_{j=0}^{\infty}t^{j+2}\Big(\frac{1}{2}\mathbf{a}'(\Sigma\Gamma)^j\Sigma\mathbf{a} + \frac{1}{2}\boldsymbol{\mu}'(\Gamma\Sigma)^{j+1}\Gamma\boldsymbol{\mu} + \mathbf{a}'(\Sigma\Gamma)^{j+1}\boldsymbol{\mu}\Big), \qquad (5)
\end{aligned}$$

where C denotes the symmetric positive square root of Σ and $\mathrm{tr}(\cdot)$ denotes the trace of $(\cdot)$, while its s^{th} cumulant K_s is

$$K_s = \frac{s!}{2}\Big(\frac{\mathrm{tr}(\Gamma\Sigma)^s}{s} + \mathbf{a}'(\Sigma\Gamma)^{s-2}\Sigma\mathbf{a} + \boldsymbol{\mu}'(\Gamma\Sigma)^{s-1}\Gamma\boldsymbol{\mu} + \mathbf{a}'(\Sigma\Gamma)^{s-1}\Gamma\boldsymbol{\mu}\Big). \qquad (6)$$

It should be noted that $\mathrm{tr}(\Gamma\Sigma)^s = \sum_{j=1}^{p}\lambda_j^s$ where the λ_j's, $j = 1, \ldots, p$, are the eigenvalues of $\Gamma\Sigma$.

3. APPROXIMATION ALGORITHM FOR VALUE-AT-RISK

Let $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$ where Σ is a positive definite covariance matrix. Consider a loss random variable

$$\mathcal{L} = d + \mathbf{a}'\,\mathbf{X} + \frac{1}{2}\mathbf{X}\,\Gamma\,\mathbf{X} \qquad (7)$$

where θ is a scalar, $\mathbf{a}$ is a p column vector and Γ is a $p \times p$ matrix derived from the current portfolio positions. The following algorithm can be utilized to determine VAR of the loss distribution, which is a quantile of the loss random variable.

Phase 1. Quadratic Forms

We re-express a generalized quadratic form for $\mathcal{L}$ to an indefinite quadratic form.

$$\mathcal{L} = \lceil + \mathbf{a}' \mathbf{X} + \frac{\infty}{\in} \mathbf{X} - \mathbf{X} \; = \lceil_{\infty} + \mathbf{R} \mathcal{A} \mathbf{R} \,. \tag{8}$$

where $A = \frac{1}{2}\Gamma$ and $\mathbf{R} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$. Let $\mathcal{Q} = \mathcal{L} - d_1$. Then $\mathcal{Q}$ is a indefinite quadratic form in normal random variables.

Phase 2. Eigenvalues

We obtain a representation of an indefinite quadratic form in normal random variables in terms of standard normal variables. On letting $\mathbf{Z} \sim \mathcal{N}_p(\mathbf{0}, I)$ and A is a $p \times p$ real symmetric matrix, where I is a $p \times p$ identity matrix, one has $\mathbf{R} = C\mathbf{Z} + \boldsymbol{\mu}$ where C denotes the symmetric positive square root of Σ, and then the quadratic form

$$\begin{aligned} \mathcal{Q} &= \mathbf{R}' A \mathbf{R} \\ &= (\mathbf{Z} + C^{-1}\boldsymbol{\mu})' \, C \, A \, C (\mathbf{Z} + C^{-1}\boldsymbol{\mu}) \\ &= (\mathbf{Z} + C^{-1}\boldsymbol{\mu})' P P' \, C \, A \, C \, P \, P' \, (\mathbf{Z} + C^{-1}\boldsymbol{\mu}) \end{aligned} \tag{9}$$

where $\mathbf{R}'$ denotes the transpose of $\mathbf{R}$, P is an orthogonal matrix that diagonalizes $C\,A\,C$, that is, $P' \, C \, A \, C \, P = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, $\lambda_1, \ldots, \lambda_p$ being the eigenvalues of $A\,\Sigma$ in *decreasing* order. Let $\mathbf{v}_i$ denote the *normalized* eigenvector of $C\,A\,C$ *corresponding to* λ_i (such that $C\,A\,C\,\mathbf{v}_i = \lambda_i \mathbf{v}_i$ and $\mathbf{v}_i{}'\mathbf{v}_i = 1$), $i = 1, \ldots, p$, and $P = (\mathbf{v}_1, \ldots, \mathbf{v}_p)$.

Remarks. We note that if A is not symmetric, it suffices to replace this matrix by $(A + A')/2$ in a quadratic form.

Phase 3. Spectral Decomposition Theorem

Letting $\mathbf{U} = P'\mathbf{Z}$, one has $\mathbf{U} \sim \mathcal{N}_p(\mathbf{0}, I)$ since P is a orthogonal matrix, and then, in light of the *spectral decomposition theorem*,

$$\begin{aligned} \mathcal{Q} &= (\mathbf{U} + \mathbf{b})' \mathrm{diag}(\lambda_1, \ldots, \lambda_p)(\mathbf{U} + \mathbf{b}) \\ &= \sum_{j=1}^{p} \lambda_j (U_j + b_j)^2 \end{aligned} \tag{10}$$

where $\mathbf{b} = P' C^{-1} \boldsymbol{\mu}$ with $\mathbf{b} = (b_1, \ldots, b_p)'$, $\mathbf{U} = (U_1, \ldots, U_p)'$, and $U_j + b_j \sim \mathcal{N}\ (b_j, 1)$,

$j = 1, \ldots, p$. Thus,

$$\begin{aligned} \mathcal{Q} &= \sum_{j=1}^{r} \lambda_j (U_j + b_j)^2 - \sum_{j=r+\theta+1}^{p} |\lambda_j| (U_j + b_j)^2 \\ &\equiv Q_1 - Q_2 \,, \end{aligned} \tag{11}$$

where r is the number of positive eigenvalues of $A\Sigma$ and $p - r - \theta$ is the number of negative eigenvalues of $A\Sigma$, θ being the number of null eigenvalues. That is, a noncentral indefinite quadratic form, $\mathcal{Q}$ can be expressed as a difference of positive definite quadratic forms.

Phase 4. Moments

On letting $Q_1 \equiv \mathbf{R}_1' A_1 \mathbf{R}_1$ and $Q_2 \equiv \mathbf{R}_2' A_2 \mathbf{R}_2$, appearing in Equation (11) where $A_1 = \text{diag}(\lambda_1, \ldots, \lambda_r)$, $A_2 = \text{diag}(|\lambda_{r+\theta+1}|, \ldots, |\lambda_p|)$, $\mathbf{R}_1 \sim \mathcal{N}_r(\boldsymbol{\mu_1}, I)$ with $\boldsymbol{\mu_1} = (b_1, \ldots, b_r)'$, and $\mathbf{R}_2 \sim \mathcal{N}_{p-r-\theta}(\boldsymbol{\mu_2}, I)$ with $\boldsymbol{\mu_2} = (b_{r+\theta+1}, \ldots, b_p)'$, the b_j's being as defined in Equation (10), as derived in Mathai and Provost (1992), the s^{th} cumulants of Q_1 denoted by $k_1(s)$ is

$$k_1(s) = 2^{s-1}(s-1)! \sum_{j=1}^{r} \lambda_j^s (s b_j^2 + 1) \,, \quad s \geq 1 \,, \tag{12}$$

and the s^{th} cumulants of Q_2 denoted by $k_1(s)$ is

$$k_2(s) = 2^{s-1}(s-1)! \sum_{j=r+\theta+1}^{p} \lambda_j^s (s b_j^2 + 1) \,, \quad s \geq 1 \,. \tag{13}$$

The moments of a random variable can be obtained from its cumulants by means of the recursive relationship obtained by Smith (1995). Accordingly, the h^{th} moment when given its h^{th} cumulant is given by

$$\mu(h) = \sum_{i=0}^{h-1} \frac{(h-1)!}{(h-1-i)!\, i!} \, k(h-i)\, \mu(i) \,. \tag{14}$$

Phase 5. Gamma-polynomial Approximation

Gamma-polynomial density approximants, was proposed in Ha and Provost (2007) can be applied to obtain the approximation to the distribution of each positive definite quadratic form Q_1 and Q_2 on the basis of their respective moments. Let Y be a random variable whose support is the real half line and let its raw moments $E(Y^h)$ be denoted by $\mu_Y(h)$, $h = 0, 1, \ldots$. We are interested in approximating the density function of the random variable Y, denoted by $f_Y(x)$. A *gamma-polynomial* density approximant of degree ℓ, denoted by

$f_Y(x;\,\ell)$, is

$$f_Y(x;\,\ell) = \psi(x) \sum_{i=0}^{\ell} \xi_i x^i \,. \tag{15}$$

This density approximant is expressed as the product of a gamma density, $\psi(x)$, and a polynomial adjustment, $\sum_{i=0}^{\ell} \xi_i x^i$. That is, the gamma baseline density function is

$$\psi(x) = \frac{1}{\Gamma(\alpha)\ \beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}\ \mathcal{I}_{(0,\infty)}(x)\,, \tag{16}$$

where $\mathcal{I}_A(x)$ denotes the indicator function, which is equal to 1 when $x \in A$ and 0 otherwise, and $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} \mathrm{dt}$. The parameters α and β of the gamma baseline density function are estimated from the first two moments of Y as follows:

$$\alpha = \frac{\mu_Y(1)^2}{\mu_Y(2) - \mu_Y(1)^2} \quad \text{and} \quad \beta = \frac{\mu_Y(2)}{\mu_Y(1)} - \mu_Y(1)\ , \tag{17}$$

see for instance Johnson *et al* (1995, Section 17). The j^{th} moments of the gamma baseline density function is denoted by $m(j)$, that is,

$$\int_0^{\infty} x^j \psi(x) dx \equiv m(j)\,. \tag{18}$$

The h^{th} moment of this gamma baseline distribution can be expressed as

$$m(h) = \frac{\beta^h \Gamma(\alpha + h)}{\Gamma(\alpha)} = \beta^h \prod_{i=1}^{h} (\alpha + h - i)\ , \quad h = 0, 1, \ldots\,. \tag{19}$$

From the moment matching technique between the moments of the target distribution and the estimated gamma baseline distribution, we can obtain the coefficients ξ_i of the polynomial adjustment. That is, the coefficients ξ_i satisfy the following system of linear equations:

$$(m(h), \ldots, m(h+\ell)).(\xi_0, \ldots, \xi_\ell)' = \mu(h),\ h = 0, 1, \ldots, \ell. \tag{20}$$

Phase 6. Transformation of Variables Technique

Since an indefinite quadratic form can be expressed as the difference of two positive definite quadratic forms, its density function can be obtained from those of the positive definite quadratic forms via the transformation of variables technique. For the problem at hand, letting $f^a_{\mathcal{Q}}(q)$, $f^a_{Q_1}(q_1)$ and $f^a_{Q_2}(q_2)$ respectively denote the approximate densities of $\mathcal{Q}$, Q_1 and Q_2, the supports of Q_1 and Q_2 being respectively $(0, \infty)$ and $(0, \infty)$, the approximate

density function of the indefinite quadratic form $\mathcal{Q}$ is given by

$$f^a_{\mathcal{Q}}(q) = \begin{cases} \int_0^\infty f^a_{Q_1}(q+x) f^a_{Q_2}(x) dx & q \geq 0 \\ \int_{-q}^\infty f^a_{Q_1}(q+x) f^a_{Q_2}(x) dx & q < 0. \end{cases} \tag{21}$$

Phase 7. Value-at-Risk

The corresponding cumulative distribution function can then be evaluated by numerical integration. Percentage points were obtained by equating the distributions functions to a given probability and then by solving the resulting equations iteratively. One can finally obtain VAR by adding d_1 to the resulting percentage point.

4. NUMERICAL STUDY

In this section, we consider a simple example in order that the approximated distribution can be compared with the exact distributions. We consider the case of a positive definite central quadratic form in independently distributed in standard normal variables, which can be expressed as

$$\mathcal{Q} = \mathbf{R}' A \mathbf{R} = \sum_{j=1}^{r} \lambda_j Y_j \,, \tag{22}$$

where $A > 0$, $\mathbf{R} \sim \mathcal{N}_p(\mathbf{0}, I)$, λ_j, $j = 1, \ldots, r$, are the positive eigenvalues of A, the Y_j's, $j = 1, \ldots, r$ are independently distributed central chi-square random variables, each having one degree of freedom, and the λ_j's are the eigenvalues of the matrix A. In this example, $\lambda_1 = \lambda_2 = 1$, $\lambda_3 = \lambda_4 = 3$, and $\lambda_5 = \lambda_6 = 7$. Since the eigenvalues occur in pairs, the exact density function can be determined from the positive part of Equation (23), which was derived by Imhof (1961), wherein $\lambda'_k = \lambda_{k/2}$, $s = t = r/2$, $\rho = 0$ and an empty product is interpreted as 1. In this case, the density function of $\mathcal{Q}$ can be directly approximated by means of Equation (15). In this case, $\alpha = 2.05085$ and $\beta = 10.7273$. The tenth-degree gamma-polynomial density approximant of $\mathcal{Q}$ for the given λ_k's is shown in Figure 1, superimposed on the exact density. The difference between the exact and approximated density functions is plotted in Figure 2. Certain VAR determined from the exact distribution and fifteenth-degree gamma-polynomial approximants are included in Table 1. The difference between the exact and approximated distribution functions is also plotted in Figure 3.

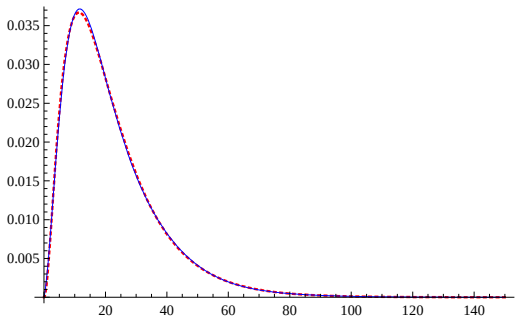


Figure 1: PDF for Exact density (dashed) & gamma-polynomial approximant

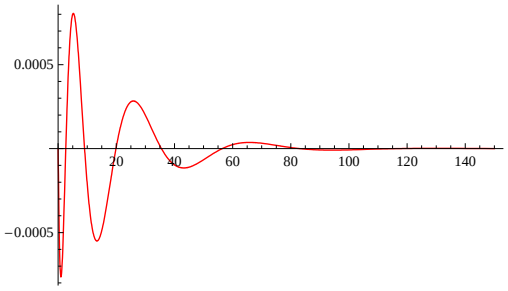


Figure 2: Difference between Exact density & gamma-polynomial density approximant

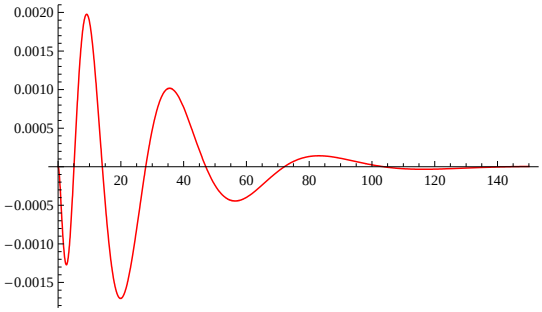


Figure 3: Difference between Exact distribution & gamma-polynomial distribution approximant

Table 1: Certain Value-At-Risks

CDF	Gamma-poly (d=10)	Exact
0.01	2.3753	2.50106
0.05	4.84055	4.84823
0.10	6.73377	6.69306
0.50	18.2224	18.2735
0.90	42.1646	42.0881
0.95	51.7819	51.8777
0.99	74.5424	74.4587

Table 2: Certain extreme Value-At-Risks

CDF	Gamma-poly	Exact
0.0001	0.307168	0.479181
0.001	0.872502	1.06936
0.999	106.509	106.701
0.9999	138.257	138.937

$$g(q) = \begin{cases} \sum_{j=1}^{s} \dfrac{{\lambda'_j}^{t-2}\, e^{-2q/(2\lambda'_j)}}{2\left(\prod_{k=1,k\neq j}^{s}(\lambda'_j-\lambda'_k)\right)\left(\prod_{k=s+1}^{t}(|\lambda'_j|+|\lambda'_k|)\right)}, & q \geq 0 \\ \sum_{j=s+1}^{t} \dfrac{|\lambda'_j|^{t-2}\, e^{2q/(2|\lambda'_j|)}}{2\left(\prod_{k=s+1,k\neq j}^{t}(|\lambda'_j|-|\lambda'_k|)\right)\left(\prod_{k=1}^{s}(\lambda'_j+\lambda'_k)\right)}, & q < 0. \end{cases} \tag{23}$$

We found that the 95$^{\text{th}}$ percentiles obtained from approximants of degrees 4, 6, 8, 10 and 12 are respectively 51.1927, 52.2512, 52.05, 51.7819 and 51.7897 whereas the exact 95$^{\text{th}}$ percentiles is 51.8777. It is seen that the approximations converge to the exact 95$^{\text{th}}$ percentile. The degree of the approximant can be selected according to the desired level of precision. Certain extreme VAR obtained from the exact density function and fifteenth-degree gamma-polynomial approximants are presented in Table 2. More precision can be obtained by including additional terms in the approximations. However, when several successive approximate density functions are seen to be nearly identical, not much additional precision will be gained by making use of more moments.

5. COMPUTATIONAL CONSIDERATIONS AND CONCLUDING REMARKS

The proposed density approximation methodology is not only conceptually simple since it is essentially based on a moment-matching technique, but it also is easy to program and consistently yields remarkably accurate percentage points. Although most applications require relatively few moments, the proposed approximants can also accommodate a large number of moments, if need be. The proposed approximation algorithm, which is designed specially for VAR of Delta-Gamma method, has a few remarkable features. First, the step-by-step approximation algorithm is a competitive technique even when the portfolio distribution is very skewed whereas the Cornish-Fisher approximation provides accuracy only when the portfolio distribution is relatively close to normal. This technique achieves a sufficient accuracy potentially fast once moments of two positive definite moments are calculated. Second, this technique achieves accurately all the range of VARs since the proposed algorithm produce very accurate percentiles over the entire range of the distribution, whereas Saddlepoint approximation accurately determines only tail probabilities. Finally, it is seen in the example that the approximations converge to the exact percentiles.

ACKNOWLEDGEMENTS

This research was supported by the Kyungwon University Research Fund.

REFERENCES

Britton-Jones, M. and Schaefer, S. (1999). Nonlinear Value-at-Risk, *European Finance Review*, **2**, 161–187.

Feuerverger, A. and Wong, A. C. (2000). Computation of Value at Risk for nonlinear portfolios, *Journal of Risk*, **3(1)**, 37–55.

Glasserman, P., Heidelberger, P. and Shahabuddin, P. (2000). Variance reduction techniques for estimating Value-at-Risk *Management Science*, **46**, 1349–1364.

Ha, H-T. and Provost, S. B. (2007). A viable alternative to resorting to statistical tables. *Communication in Statistics: Simulation and Computation*, **36**, 1135–1151.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, **48**, 419–426.

Jaschke, S. (2002). The Cornish-Fisher expansion in the context of Delta-Gamma-Normal approximations, *Journal of Risk*, **4(4)**, 33-52.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995). *Continuous Univariate Distributions, Vol 2*. John Wiley & Sons, New York.

Joridon, P. (1997). *Value at Risk*. McGraw-Hill, New York.

Longerstaey, J. (1996). RiskMetrics technical document, Technical Report fourth edition, J.P.Morgan, http://www.riskmetrics.com.

Mathai, A. M. and Provost, S. B. (1992). *Quadratic Forms in Random Variables, Theory and Applications*. Marcel Dekker Inc., New York.

Mina, J. and Ulmer, A. (1999). Delta-gamma four ways, http://www.riskmetrics.com.

Reid, N. (1988). Saddlepoint methods and statistical inference. *Statistical Science*, **3**, 213–227.

Rouvinez, C. (1997). Going greek with VaR, *Risk*, **10(2)**, 57–65.

Smith, P. J. (1995). A recursive formulation of the old problem of obtaining moments from cumulants and vice versa. *The American Statistician*, **49**, 217–219.

Zangari, P. (1996). A VaR methodology for portfolios that include options, RiskMetrics Monitor, 1996 (first quarter), 4–12.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 646–651.

ON A GENERALIZED DIFFERENTIAL EQUATION FOR GENERATING SCUI DISTRIBUTIONS

Saleha Naghmi Habibullah
Kinnaird College for Women, Lahore, Pakistan
E-mail: salehahabibullah@hotmail.com

Ahmed Zogo Memon and Munir Ahmad
National College of Business Administration & Economics
Lahore, Pakistan

ABSTRACT

Habibullah et al. (2009) have developed a differential equation for generating a class of SCUI distributions (i.e. those which are invariant under the reciprocal transformation), and have demonstrated its usefulness by deriving from it a density function which is flexible in its application and provides a useful model for life-data. In this paper, we propose a much more generalized form of the differential equation for generating the class of SCUI distributions so that the above-mentioned differential equation becomes a special case of this one. The utility of this generalized form lies in the derivation of a large number of density functions that are not derivable from the differential equation proposed by Habibullah et al. (2009) but provide useful models for real data. Some of the well-established density functions are presented as examples.

1. INTRODUCTION

Utilization of differential equations in special problems of probability theory seems to have originated in the last decade of the nineteenth century. Pearson (1895) notes that in the limiting case, the hypergeometric distribution can be expressed in the form

$$\frac{df}{dx} = \frac{(x-a)\,f}{b_0 + b_1 x + b_2 x^2} \tag{1.1}$$

He utilizes this fact to obtain the Pearson system of continuous distribution functions. Dunning and Hansen (1977) present the generalized Pearson distributions as the solution of the differential equation

$$\frac{df}{dt} = \frac{c_0 + c_1 t + c_2 t^2 + \ldots + c_m t^m}{c_0' + c_1' t + c_2' t^2 + \ldots + c_n' t^n} f(t) \tag{1.2}$$

where both m and n are greater or equal to 1. Cobb (1980) presents a differential equation of the form

$$\frac{df}{dx} = \frac{g(x)}{h(x)} f(x)dx,\ h(x) > 0, x \in Int(A) \tag{1.3}$$

where $Int(A)$ is an interval of the real line with the choice of A different for different densities, and $g(x)$ and $h(x)$ are polynomials such that the degree of $h(x)$ is one higher than the degree of $g(x)$. He illustrates three types of probability distributions (i.e. Normal type, Gamma type and Beta type) that can be generated from this differential equation under certain admissible conditions on $g(x)$ and $h(x)$.

By applying the transformation $Y = 1/X$ in the Pearson differential equation, Ahmad (1985) obtains a differential equation which, under certain conditions, generates a class of inverted distributions. He focuses on a special case of Cobb's differential equation, the one in which the degree of $g(y)$ is two and that of $h(y)$ is three. Specifically, he discusses the differential equation of the form

$$\frac{d}{dx}\left[\ln g(x)\right] = -\frac{\left(a_2 x^2 + a_1 x + a_0\right)}{x\left(B_0 x^2 + B_1 x + B_2\right)} \tag{1.4}$$

where the coefficients a_0, a_1, a_2 are given by $a_0 = 2B_2 - 1$, $a_1 = 2B_2 - a$, $a_2 = 2B_0$, and uses it to generate the Inverted Pearson System of probability distributions. The inverted class of distributions generated by the above differential equation includes the Inverted Normal, Inverted Type 1 (Inverted Beta), Inverted Type III, Inverted Type V, Inverted Type II, Inverted Type VI, Inverted Type VII and the Inverted Type IV distributions. Also, Ahmad (1985) develops relationships between the four parameters of the above differential equation (i.e. B_0, B_1, B_2 and a), and the first five moments of the probability distribution.

Chaudhry and Ahmad (1993) consider the following special case of differential equation (1.2)

$$\frac{df}{dt} = \frac{c_0 + c_4 t^4}{c_3' t^3} f(t),\ c_3 \neq 0 \tag{1.5}$$

They obtain a probability function as the solution to the differential equation (1.5), also obtain a relationship between the derived probability function and the Inverse Gaussian distribution, and show that the derived model is more suitable than the lognormal distribution for a particular data set.

Habibullah et al. (2009) propose the following differential equation which yields an unlimited number of differential equations under a set of conditions.

$$\frac{d}{dy}\left[\ln g(y)\right] = \frac{b_n y^n + b_{n-1} y^{n-1} + ... + b_0}{a_n y^n + a_{n-1} y^{n-1} + ... + a_0} \tag{1.6}$$

They demonstrate its usefulness by deriving from it a density function which is flexible in its application and provides a useful model for life-data.

In this paper, we propose a much more generalized form of the differential equation for generating the class of SCUI distributions so that the above-mentioned differential equation becomes a special case of this one. It is shown that a large number of SCUI distributions --- including some of the well-established density functions ---- that are not derivable the differential equation of Habibullah et al. (2009) fall under the category of the newly derived differential equation.

2. SCUI DISTRIBUTIONS

Kleiber and Kotz (2003) use the term 'closed under inversion' in the sense that the original distribution and the inverse distribution have the same domain of support, and belong to the same parametric class. Habibullah and Ahmad (2006) define Strict Closure Under Inversion as the case where the distribution of the reciprocal of a continuous random variable is identical to that of the original random variable. They use the abbreviation SCUI for distributions that are strictly closed under inversion

3. A GENERALIZED DIFFERENTIAL EQUATION FOR GENERATING SCUI DISTRIBUTIONS

We propose the following theorem:

Theorem 3.1: Let $g(y)$ be the *pdf* of $Y = \ln X$ where the random variable X has the *pdf* $f(x)$ defined on $(0,\infty)$. If

$$\frac{d}{dy}\left[\ln g(y)\right] = \frac{\sum_{i=0}^{n} b_i \left[w(y)\right]^i}{\sum_{i=0}^{n} a_i \left[w(y)\right]^i} \tag{3.1}$$

then $f(x)$ is SCUI provided that the following conditions hold:

Case I: $w(y)$ is an odd function of y i.e. $w(y) = -w(-y)$

(a) $a_i \neq 0$ and $b_j \neq 0$ for some $i, j, 0 \leq i, j \leq n$, and

(b) $\sum_{i=0}^{2j} (-1)^i a_{2j-i} b_i = 0, \quad j = 0,1,2,....,m,$

$$\sum_{i=0}^{2j} (-1)^i a_{n-i} b_{n-2j+i} = 0, \quad j = 0,1,2,....,m \tag{3.2}$$

where m is $\frac{n}{2}$ or $\frac{n-1}{2}$ according as n is an even or odd non-negative integer,

Case II: $w(y) = \left[w(-y)\right]^{-1}$

(a) $a_i \neq 0$ and $b_j \neq 0$ for some i,j, $0 \leq i, j \leq n$,

(b) $\sum_{i=0}^{j}\left(a_i b_{i+n-j}+a_{i+n-j}b_i\right)=0,\ \ j=0,1,2,....,n-1,$

$$\sum_{i=0}^{n} a_i b_i = 0 \tag{3.3}$$

Proof:

Case I: Note that

$$-\frac{d}{dy}\left[\ln g(-y)\right]=\frac{b_n\left[w(y)\right]^n-b_{n-1}\left[w(y)\right]^{n-1}+.....+b_0}{a_n\left[w(y)\right]^n-a_{n-1}\left[w(y)\right]^{n-1}+.....+a_0},\ n\ even$$

$$-\frac{d}{dy}\left[\ln g(-y)\right]=\frac{-b_n\left[w(y)\right]^n+b_{n-1}\left[w(y)\right]^{n-1}+.....+b_0}{-a_n\left[w(y)\right]^n+a_{n-1}\left[w(y)\right]^{n-1}+.....+a_0},\ n\ odd$$

X being strictly closed under inversion implies $g(y)=g(-y)$ which leads to

$$\frac{d}{dy}\left[\ln g(y)\right]-\frac{d}{dy}\left[\ln g(-y)\right]=0 \tag{3.4}$$

This implies that

$$\left(b_n\left[w(y)\right]^n+b_{n-1}\left[w(y)\right]^{n-1}+....+b_0\right)\left(a_n\left[w(y)\right]^n-a_{n-1}\left[w(y)\right]^{n-1}+....+a_0\right)$$
$$+\left(a_n\left[w(y)\right]^n+a_{n-1}\left[w(y)\right]^{n-1}+....+a_0\right)$$
$$\left(b_n\left[w(y)\right]^n-b_{n-1}\left[w(y)\right]^{n-1}+....+b_0\right)=0,\ n\ even$$

and

$$\left(b_n\left[w(y)\right]^n+b_{n-1}\left[w(y)\right]^{n-1}+....+b_0\right)\left(-a_n\left[w(y)\right]^n+a_{n-1}\left[w(y)\right]^{n-1}-....-a_0\right)$$
$$+\left(a_n\left[w(y)\right]^n+a_{n-1}\left[w(y)\right]^{n-1}+....+a_0\right)$$
$$\left(-b_n\left[w(y)\right]^n+b_{n-1}\left[w(y)\right]^{n-1}-....-b_0\right)=0,\ n\ odd$$

It thus follows that

$$\sum_{i=0}^{2j}(-1)^i\, a_{2j-i}b_i=0,\quad j=0,1,2,....,m,$$
$$\sum_{i=0}^{2j}(-1)^i\, a_{n-i}b_{n-2j+i}=0,\quad j=0,1,2,....,m$$

where m is $\frac{n}{2}$ or $\frac{n-1}{2}$ according as n is an even or odd non-negative integer,

Case II: The proof is similar to that of Case I.

4. EXAMPLES

Letting $w(y)=e^{y}$ in differential equation (3.1), we obtain the special case

$$\frac{d}{dy}\left[\ln g(y)\right]=\frac{b_n e^{ny}+b_{n-1}e^{(n-1)y}+.....+b_0}{a_n e^{ny}+a_{n-1}e^{(n-1)y}+.....+a_0} \tag{4.1}$$

which yields SCUI distributions under the set of conditions (3.3). We utilize differential equation (4.1) to present examples of some of the well-established density functions that are SCUI but are not derivable from the differential equation proposed by Habibullah et al. (2009):

4.1 F Distribution

Differential equation (4.1) with, $n=1$ and $b_1=-n,\ b_0=n,\ a_1=2,\ a_0=2$ yields

$$g(y)=ke^{y\left(\frac{n}{2}-1\right)}e^{y}\left(1+e^{y}\right)^{-n},\ -\infty<y<\infty$$

with normalizing factor $k=\Gamma(n)\Big/\left[\Gamma\left(\frac{n}{2}\right)\right]^2$

The transformation $Y=\ln X$ yields the F distribution with equal degrees of freedom.

4.2 Half – Cauchy Distribution

Putting $n=2$ and $b_2=-1,\ b_1=0,\ b_0=1,\ a_2=1,\ a_1=0,\ a_0=1$ in differential equation (4.1), we have

$$g(y)=\frac{2e^{y}}{\pi\left[1+\left(e^{y}\right)^2\right]},\ -\infty<y<\infty$$

Applying the transformation $X=e^{Y}$, we obtain

$$f(x)=\frac{2}{\pi\left(1+x^2\right)},\ 0<x<\infty$$

which is the standard half – Cauchy distribution.

4.3 Birnbaum Saunders Distribution

Putting $n=3$ in (4.1) and letting $b_3=-1, b_2=-\left(1-\alpha^2\right), b_1=\left(1-\alpha^2\right), b_0=1,\ a_3=0,\ a_2=2\alpha^2, a_1=2\alpha^2,$ $a_0=0$ we have

$$g(y)=\frac{e^{\alpha^{-2}}}{2\alpha\sqrt{2\pi}}\left(e^{y}\right)^{-\frac{1}{2}}\left(e^{y}+1\right)e^{-\frac{1}{2\alpha^2}\left(e^{y}+\frac{1}{e^{y}}\right)},\ -\infty<y<\infty$$

so that for $X = e^Y$

$$f(x) = \left[\exp\left(\alpha^{-2}\right)/2\alpha\sqrt{2\pi}\right] x^{-\frac{3}{2}}(1+x)\exp\left[-\left(x+\frac{1}{x}\right)/2\alpha^2\right],\ x>0$$

which is the Birnbaum Saunders distribution with $\sigma = 1$.

5. CONCLUDING REMARKS

Differential equation (3.1) generates a much wider class of SCUI distributions than differential equation (1.6) proposed by Habibullah et al. (2009). It seems to have the potential for generating a number of new SCUI density functions useful in modeling real data.

REFERENCES

Ahmad, M. (1985). Theory of Inversion. *Unpublished Manuscript.* National College of Business Administration and Economics, Lahore, Pakistan.

Chaudhry, M.A. and Ahmad, M. (1993). On A Probability Function Useful In Size Modeling. *Can. J. Forest Res.*, 23, 1679-1683.

Cobb, L. (1980). The Multimodal Exponential Families of Statistical Catastrophe Theory. *Statistical Distributions in Scientific Work, 4: Models, Structures, and Characterizations, Charles Taillie (ed),* Springer London Ltd., 69-90.

Dunning, K.A. and Hansen, J.R. (1977). Generalized Pearson distributions and nonlinear programming. *J. Statist. Comput. Simulation*, 6, 115-128.

Habibullah, S.N. and Ahmad, M. (2006). On a New Class of Univariate Continuous Distributions that are Closed Under Inversion. *Pak. J. Statist. and Oper. Res.*, II(2), 151-158.

Habibullah, S.N., Memon, A.Z. and Ahmad, M. (2009). On a Class of Distributions Strictly Closed Under Inversion. Submitted to *Pak. J. Statist*.

Kleiber, C. and Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley-IEEE., 189.

Pearson, K. (1895). Contribution to the Mathematical Theory of Evolution II. Skew variation in Homogeneous Materials. *Phil. Trans. RSS, London, Series A*, 186, 343-414.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 652–663.

NON-LINEAR GOAL PROGRAMMING APPROACH TO CLUSTERWISE LOGISTIC REGRESSION MODEL

Ramadan Hamed[1], Ali El Hefnawy[2] and Mohamed Ramadan[3]
[1, 2]Faculty of Economics and Political Science, Cairo University, Cairo, Egypt.
[3] Population Council WANA Regional Office, 59 Misr Agricultural Road, Maadi, Cairo, Egypt.
E-mail: [1]ramadanh@aucegypt.edu, [2]ahefnawy@aucegypt.edu, [3]mramadan@popcouncil.org

ABSTRACT

The Clusterwise Regression Model is applied to carry out cluster analysis within a regression framework. This paper employs logistic regression model in the clusterwise framework by using mathematical programming. The proposed "Clusterwise Logistic Model" integrates the utilization of logistic regression as a regression and discriminant tool. Therefore, it considers two types of errors to be used as clustering criteria; the sum of residuals, as in the regression analysis, and the sum of classification errors, as in the discriminant analysis. In the proposed model, the clustering membership parameters are introduced as probabilities. Thus, every subject has a number of probabilities equivalent to the total number of clusters. The theoretical design of the proposed mathematical programming model is based on non-linear goal programming, with a linear objective function and non-linear constraints. A set of simulation scenarios has been developed to assess the designed model. The simulation study shows that the correct classification has been enhanced by using the proposed mathematical programming model, as compared with the non-clustered logistic model, whether it is estimated by the maximum likelihood or the mathematical programming. In addition, the proposed mathematical programming model isn't influenced with the common logistic regression drawbacks. Even when the mean squared error of Clusterwise regression model surpasses the mean squared error of ML and the MP logistic models, the suggested model retains its advantages, because its statistics are very close to the non-clustered models statistics.

Keywords: Clusterwise model, Logistic regression, Non-linear programming approach, Cluster analysis, Goal programming.

1. INTRODUCTION

In real data analysis there are situations when multi-statistical techniques must be used. In such cases, the mathematical programming can be introduced as an alternative approach to integrate more than one statistical technique with multi-objective functions in one model with a single objective function. One of these models is the clusterwise regression model that is used to perform cluster analysis within a regression framework. This sort of models seeks to estimate the cluster membership parameters and the regression model parameters simultaneously. The clusterwise regression model has two main advantages. While the traditional regression model assumes the regression coefficient to be identical for all subjects in the sample, the clusterwise regression model allows it to vary with subjects of different clusters (Lau et al., 1999). On the other hand, the cluster analysis and regression analysis in the classical two-step procedure are

unrelated. Also it is sometimes hard to select criteria to cluster data since the approach for the cluster analysis is not unique. Therefore the second advantage of the clusterwise regression model is that it considers the interrelations between cluster analysis and regression analysis (Luo 2005).

The term “clusterwise” was first coined by Spath (1979, 1981, 1982, 1985). Spath (1979) proposed the "exchange algorithm" which is used to minimize the sum of the square errors, for partitions of length (k) and corresponding sets of parameters. The exchange algorithm is further generalized by Spath (1986) and Meier (1987), as stated in Lau et al., 1999, to minimize the sum of the absolute errors. In 1980 Aitkin and Wilson, as stated in Lau et al., 1999, introduced another approach "the mixture model", which is a parametric procedure with strong distributional assumption on the noise term. The mixture model does not directly classify subjects into clusters, instead, it computes the cluster membership probabilities for each subject.

In 1999 Lau et al., generalized all previous attempts in what they called "Generalized Clusterwise Regression Model", which incorporates the parameter heterogeneity in traditional regression using a mathematical programming model. As the cluster membership parameters in the clusterwise regression model are unknown, Lau et al., (1999) showed that the estimation of the clusterwise regression is a tough combinatorial optimization problem. They extended this effort to integrate the cluster analysis with discriminant analysis, in what they titled "Clusterwise Discriminant Model" that was developed to incorporate parameter heterogeneity into traditional discriminant analysis

The purpose of the present paper is to introduce the "Generalized Clusterwise Logistic Model". In this model, the clusterwise model is developed in the framework of logistic regression model. The importance of the proposed model arises from the reliable results of utilizing the logistic regression as a regression and a discriminant tool in real life.

One of the reasons why logistic regression model is important is that it can be used to predict a binary dependent variable based on continuous and/or categorical explanatory variables. In addition, it can determine the percent of variance in the dependent variable explained by the explanatory variables; to rank the relative importance of explanatory variables, to assess interaction effects and to understand the impact of covariate control variables (Christensen 1997). Unlike Ordinary Least Square (OLS) regression, logistic regression does not assume linearity of relationship between the explanatory variables and the dependent, does not require normally distributed variables, does not assume homoscedasticity, and in general has less stringent requirements (Draper & Smith 1998) and (Greene 1997). Also, the success of the logistic regression can be assessed by observing the classification table, which shows correct and incorrect classifications of the dichotomous, ordinal, or polytomous dependent (Hosmer & Lemeshow 2000). Another reason for its importance is that instead of classifying an observation into one group or the other, as in the discriminant analysis, the logistic regression predicts the probability that an indicator variable is equal to one (success case). To be precise, logistic regression model predicts the log odds1 that an observation will have an indicator equal to one (Agresti 2002).

The growing use of the logistic regression in many applications, as in epidemiological studies, social science, medical experiments …etc., encourages the use of a technique like clusterwise model. The studied population or respondents in most cases are usually heterogeneous. Accordingly, the construction of the Generalized Clusterwise Logistic Model

[1] The odd of an event is defined as the ratio of the probability that an event occurs to the probability that it fails to occur.

based on logistic regression model is expected to gain much attention in various fields, especially social sciences.

Therefore, the current paper aims to construct the theoretical framework of the Clusterwise Logistic Model, as well as to compare the efficiency of the constructed models with the logistic regression models that use the maximum likelihood and the mathematical programming approaches. This comparison depends on simulation studies to reveal the advantages and the limitations of the proposed models.

2. LOGISTIC REGRESSION USING MATHEMATICAL PROGRAMMING

Hamed et al., (2009) introduced a non-linear goal programming model that estimates the logistic regression model. Understanding this model is a prerequisite to understand the proposed model. The base concept of their model was focused on the role of logistic regression as a part of regression analysis and discrimination analysis models. The aim in dealing with logistic regression is always to minimize the residuals and maximize the probability of correct classification. Initiated from these objectives, two types of constraints were developed to achieve the two objectives simultaneously. The first constraint was constructed to limit the deviations between the expected and predicted values for the response variable (Y), to equal zero. However, to avoid the non-optimality or infeasibility, two complementary decision variables were included, in the form of goal programming. These two non-negative decision variables (s^+ and s^-) were designed to capture the corresponding negative and positive deviations. Thus, these S's were expected to play the role of the sum of the residuals in linear regression. Therefore, this constraint in the context of regression analysis is typically equivalent to residuals with zero mean. Thus, the first objective of the model was to minimize the sum of residuals.

$$\sum_{i=1}^{n}\left[Y_i - \frac{\exp(\boldsymbol{\beta}'\mathbf{x_i})}{1+\exp(\boldsymbol{\beta}'\mathbf{x_i})}\right] - s^+ + s^- = 0 \qquad (1)$$

where:

i: the subject subscript for the sample of size n.

$\boldsymbol{\beta}$: the (j+1) vector of explanatory variables coefficients and the intercept.

$\mathbf{x_i}$: the i-th vector of the explanatory variables and the intercept, where $\mathbf{x_1}$ is the unit vector. On the other hand, the logistic regression model is used as a discrimination model. Therefore, the logistic model must achieve a significant level of the subjects' correct classification, i.e. maximize the probability of correctly classified subjects.

The philosophy of the classification according to the logistic regression is based on a cut-off value for the predicted response variable, after estimating the regression's coefficients. Most of statistical packages consider the cut-off value 0.5 as an unbiased probability of the binary variable's categories, which reflect the probability of each category of the binary response variable. In contrast of building a thin separate line, Hamed et al., (2009) suggested to represent the separate line as an interval around the cut-off value, with $\mp\varepsilon$, where ε is an arbitrary small positive value. Moreover, a non-negative decision variable (d) was introduced as an error of deviation from the separation interval. This led the authors to construct the next two constraints:

$$\frac{\exp(\boldsymbol{\beta}'\mathbf{x_i})}{1+\exp(\boldsymbol{\beta}'\mathbf{x_i})} + d_i \geq 0.5 + \varepsilon \qquad \forall Y_i = 1 \qquad _{i=1,2,\cdots,n} \tag{2}$$

$$\frac{\exp(\boldsymbol{\beta}'\mathbf{x_i})}{1+\exp(\boldsymbol{\beta}'\mathbf{x_i})} - d_i \leq 0.5 - \varepsilon \qquad \forall Y_i = 0 \qquad _{i=1,2,\cdots,n} \tag{3}$$

The mechanism of these constraints is to classify the i-th subject to equation (2) if $(Y_i = 1)$, or to equation (3) otherwise. In these constraints (2 and 3) the term (d_i) represents the required positive real value that is needed to extract the subject (i) above or below the separation interval. Therefore, Hamed et al., (2009) defined their second objective by minimizing the sum of all deviations that will be needed to clean the separation interval from any subject. So, they defined a non-linear goal programming model with linear objective function and non-linear constraints to estimate the logistic regression model:

$$\text{minimize } F_{\boldsymbol{\beta},d_i,s^+,s^-} = \left(\sum_{i=1}^{n} d_i\right) + s^+ + s^- \tag{4}$$

Subject to

$$\frac{\exp(\boldsymbol{\beta}'\mathbf{x_i})}{1+\exp(\boldsymbol{\beta}'\mathbf{x_i})} + d_i \geq 0.5 + \varepsilon \qquad \forall Y_i = 1 \qquad _{i=1,2,\cdots,n} \tag{5}$$

$$\frac{\exp(\boldsymbol{\beta}'\mathbf{x_i})}{1+\exp(\boldsymbol{\beta}'\mathbf{x_i})} - d_i \leq 0.5 - \varepsilon \qquad \forall Y_i = 0 \qquad _{i=1,2,\cdots,n} \tag{6}$$

$$\sum_{i=1}^{n}\left[Y_i - \frac{\exp(\boldsymbol{\beta}'\mathbf{x_i})}{1+\exp(\boldsymbol{\beta}'\mathbf{x_i})}\right] - s^+ + s^- = 0 \tag{7}$$

$$d_i \geq 0 \qquad _{i=1,2,\cdots,n} \tag{8}$$

$$s^+, s^- \geq 0 \tag{9}$$

This model has many theoretical advantages, as it can accept any function of linear and/or non-linear constraints of the logistic regression's parameters, such as budget constraints. Moreover, it doesn't require any assumptions on the residuals. In addition, Hamed et al., (2009) used simulation scenarios to prove some characteristics of their model, especially the conclusion that it wasn't influenced with the common logistic regression drawbacks. Also, they showed that merging the two types of errors significantly increased the total number of correctly classified subjects and introduced more robust estimates of the regression's coefficient.

3. GENERALIZED CLUSTERWISE LOGISTIC MODEL

In this section the proposed Clusterwise Logistic Model is introduced. It is designed to include a predetermined and unrestricted number of clusters (e.g. K clusters) and number of explanatory variables (e.g. J explanatory variables). The idea of clusterwise is based on a classification criterion that is employed simultaneously with a sort of regression model to classify the subjects into clusters through what is convenient to call as cluster membership parameters. Lau et al., (1999), introduced the first attempt in that manner in their Generalized Clusterwise Regression Model. In the beginning, they introduced the cluster membership parameters as positive binary

decision variables. Therefore, their first defined model was a non-linear integer programming model. In the same paper, they proved that it is possible to transfer this model to a non-linear programming model by redefining the cluster membership parameters. Their transformation was based on introducing new constraints to restrict the sum of the cluster membership parameters to equal one with respect to every subject in the sample, and to introduce every membership cluster as a positive quantity that is bounded in the interval [0,1]. Moreover, they related the cluster membership parameters to the associated likelihood function. Thus, their model's mechanism allows the subject to be classified to a specific cluster if it is enhancing the overall fitted model in that cluster more than the other clusters. It is noted from this mechanism that the cluster membership parameters are playing the role of weights, i.e. they weight the relative importance of a specific cluster according to the fitting criterion.

Initiated from this mechanism, there are two main tasks to pursue; find a criterion that can be used to classify the subjects into homogeneous clusters and design the cluster membership parameters. To find this criterion let us reformulate the two types of errors; the regression residuals (equation 1) and the classification errors (equations 2 and 3). The (K) clusters are mutually exclusive, thus the subject (i) will be classified to only one cluster. Therefore, we can break $\sum_{i=1}^{n}[Y_i - \hat{Y}_i]$ in equation (1) to (K) parts that present the estimation in each cluster.

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\left({}_kC_i \left[Y_i - \frac{\exp({}_k\boldsymbol{\beta}'\mathbf{x_i})}{1+\exp({}_k\boldsymbol{\beta}'\mathbf{x_i})}\right]\right) - s^+ + s^- = 0 \qquad (10)$$

where:

k: the cluster subscript.

K: the number of clusters.

${}_kC_i$: the membership of the i-th observation to the k-th cluster.

${}_k\boldsymbol{\beta}$: the (j+1) vector of explanatory variables coefficients and intercept in the k-th cluster.

As in the above equation ${}_kC_i$ are the cluster membership parameters, and they are multiplied by the sum of residuals to present the probability of a subject's belonging to a specific cluster. Actually these C's are not expected to produce ones and zeros, but they produce probabilities. Thus, for every subject there will be (K) probabilities, everyone presents the belonging of that subject to the associated cluster, and the final decision is related to the highest probability.

Regarding the second type of errors, the classification rule should target each cluster individually. Therefore the subject that is classified to the cluster (k) uses this cluster classification rule, i.e. each cluster has its separation interval.

$$\frac{\exp({}_k\boldsymbol{\beta}'\mathbf{x_i})}{1+\exp({}_k\boldsymbol{\beta}'\mathbf{x_i})} + {}_kd_i \geq 0.5 + \varepsilon \quad \forall Y_i = 1 \qquad i=1,2,\cdots,n \quad k=1,2,\cdots,K \qquad (11)$$

$$\frac{\exp({}_k\boldsymbol{\beta}'\mathbf{x_i})}{1+\exp({}_k\boldsymbol{\beta}'\mathbf{x_i})} - {}_kd_i \leq 0.5 - \varepsilon \quad \forall Y_i = 0 \qquad i=1,2,\cdots,n \quad k=1,2,\cdots,K \qquad (12)$$

where:

${}_kd_i$: the deviation of the i-th observation from the separation interval in the k-th cluster.

From the above discussion it is clear that the desired objectives become more complicated. It is of interest to minimize the sum of residuals between the observed and the predicted response variable categories, and the sum of deviations from the separation interval in every cluster. So, the clustering criteria will be these two types of errors. Thus, the subject will have a highest probability of belonging to a specific cluster, if it achieves a minimum deviation from the fitted line in that cluster as compared with other clusters, and it is out of the separation interval or needs a slight pushing to be out of it in that cluster. Therefore Clusterwise Logistic Model can be written in the form of non-linear goal programming, with a linear objective function and non-linear constraints as follows:

$$\text{minimize } F_{{}_k\boldsymbol{\beta},\,{}_kC_i\,,\,{}_kd_i,s^+,s^-} = \left(\sum_{k=1}^{K}\sum_{i=1}^{n} {}_kC_i \times {}_kd_i\right) + s^+ + s^- \tag{13}$$

Subject to

$$\frac{\exp({}_k\boldsymbol{\beta}'\mathbf{x_i})}{1+\exp({}_k\boldsymbol{\beta}'\mathbf{x_i})} + {}_kd_i \geq 0.5 + \varepsilon \quad \forall Y_i = 1 \qquad i=1,2,\cdots,n \quad k=1,2,\cdots,K \tag{14}$$

$$\frac{\exp({}_k\boldsymbol{\beta}'\mathbf{x_i})}{1+\exp({}_k\boldsymbol{\beta}'\mathbf{x_i})} - {}_kd_i \leq 0.5 - \varepsilon \quad \forall Y_i = 0 \qquad i=1,2,\cdots,n \quad k=1,2,\cdots,K \tag{15}$$

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\left({}_kC_i\left[Y_i - \frac{\exp({}_k\boldsymbol{\beta}'\mathbf{x_i})}{1+\exp({}_k\boldsymbol{\beta}'\mathbf{x_i})}\right]\right) - s^+ + s^- = 0 \tag{16}$$

$$\sum_{k=1}^{K} {}_kC_i = 1 \qquad i=1,2,\cdots,n \tag{17}$$

$$\sum_{i=1}^{n} {}_kC_i \geq 1 \qquad k=1,2,\cdots,K \tag{18}$$

$${}_kC_i\,,\,{}_kd_i \geq 0 \qquad i=1,2,\cdots,n \quad k=1,2,\cdots,K \tag{19}$$

$$s^+, s^- \geq 0 \tag{20}$$

In the previous model it can be noticed that the membership parameters $({}_kC_i)$ are defined to represent the probability that subject (i) belongs to cluster (k). Therefore, they are forced to be bounded between zero and one, and to be summed to unity w.r.t. the i-th subject. The estimation of these membership parameters is obtained using a direct way and an indirect way. The direct way is the objective function (13). The membership probability decreases by increasing the deviation needed to separate the corresponding subject from the separation interval in the corresponding cluster, and vice versa.

The indirect way results from linking the objective function (13) with the constraint (16). The membership probability of the i-th subject increases in the cluster that minimizes the difference between the observed and predicted values of this subject. Therefore, these direct and indirect relations wrap up the two types of errors through linking the increase in the membership probability with the decrease in these two types of errors at the level of each subject. Moreover, the constraint (18) leads to breaking the observations to more than one cluster by enforcing the allocation to only one cluster. In contrast of classifying all subjects into one cluster, this constraint requires every cluster to attain at least one subject.

4. CONCEPTS AND DESIGN OF THE SIMULATION STUDY

This section is focused on the simulation study that has been made to assess the efficiency of the proposed model. For this purpose, the proposed model is compared with the non-clustered logistic regression model that is estimated by the maximum likelihood (ML) and the mathematical programming (MP) approaches. The design of the simulation study is founded on what is stated in the literature about simulation of logistic regression models (Christensen 1997). The simulation study is based on generating the covariates and the parameters to calculate the response variable's probabilities. Because these probabilities are in a deterministic pattern, they are compared with a randomly generated uniform (0,1) to classify the observations (subjects) and produce the binary random response variable (Xie 2005). In our simulation, as regards the first part of this scenario, the set of covariates is randomly generated from pre-specified distributions in each run, rather than fixing the values of the covariates at pre-specified values from only one generating process. This means that the number of runs equals the number of generated sets of covariates. This simulation scenario can be considered more reflective of observational studies rather than the traditional experimental designs that take control over covariates. In all runs the assumed population parameters of the logistic model are drawn from a symmetric uniform distribution (-3,3) (Xie 2005)[2]. Thus the actual probabilities $\pi(x)$ are estimated and compared with an independent uniform (0,1) to generate a binary random response variable, which is loaded with a noise part.

This simulation strategy is based on three factors; sample size, covariates' distribution, and covariates' multicolinearity[3] and extreme values. These three factors are selected to reflect the advantages and disadvantages of the simulated model. Four sample sizes have been selected to reflect four corresponding different levels. The first level 20 represents the small sample size, and the sample size 50 represents the reasonably moderate sample size. In addition, the 200 and 500 levels represent the reasonably large and the large sample sizes respectively. Also, both the collinear and independent covariates are used. The collinear continuous cases are overloaded with 20 percent of extreme values[4] that have been equally distributed above and below the observations[5].

The explanatory set in each run includes three covariates. Five different marginal distributions and a multivariate normal distribution of the covariates are used. These distributions represent both skewed and symmetric cases in continuous variables, as well as balance and high imbalance in binary variables. The pre-specified multivariate normal distribution's parameters are $\mu = 1.5$ and $\sigma = 1.5$ for the first variable, and $\mu = 2$ and $\sigma = 4$ for the second variable. In the collinearity case the correlations between x_1 and x_2 are significant at the 90's percent. The five marginal distributions are: the symmetric continuous uniform distribution (-3,3), the discrete

[2] We want to deeply thank; Prof. Ronghua Luo (*Peking University*), Prof. Shelley B. Bull (*University of Toronto*), Prof. Celia Greenwood (*University Avenue Toronto*) and Prof. Hun Myoung Park (*Indiana University*) for their valuable clarifications and beneficial comments through emails about simulation process, especially in the case of ordinal logistic regression.

[3] In the simulation study the introduced collinearity is not the perfect multicolinearity, because this is not the case in the real situation. Therefore, the simulation scenario is based on the theory that guarantees an associated level of multicolinearity with the existence of highly correlated bivariate normal variables (Bain & Engelhardt 1992).

[4] These extreme cases have been generated from a uniform distribution with suitable parameters.

[5] The correlation has been tested after the insertion of extreme values to be sure that the correlated variables still preserve the same level of correlation, which is 90 percent.

uniform distribution (0,6), and three Bernoulli distributions: two imbalanced distributions Bin(1,0.8) and Bin(1,0.15), and approximately balanced distribution Bin(1,0.5). Therefore, there are 20 combinations between covariates' distribution, multicolinearity and sample size. To determine the number of runs, it is assumed that every combination from these 20 combinations is a separate process. As in the literature (El-Haik & Al-Omar 2006) the number of runs can be determined by using the law:

$$m = \frac{\left(Z_{\alpha/2}\right)^2 s^2}{\delta^2} \qquad (21)$$

where:

$Z_{\alpha/2}$: is the standard normal score.

δ: is the desired margin of error, which is the half-length of the confidence interval with a $100(1-\alpha)\%$ confidence level.

s^2: is the variance obtained from the runs.

By using 95% confidence level, the performance mean obtained from the simulation model is estimated within ∓ 0.5 of the true unknown mean and from 20 runs as pilot cases, the standard deviation for the correct classification percent is 3.6 percent. Therefore the number of runs in each combination and every sample size is calculated to be 200 runs, with total 4000 runs in each suggested model.

The simulation's results are based on two indices, as follows:

A. The correct classification percent: is the percentage of correctly classified subjects to the sample size.
B. The mean squared error: is the mean squared deviance between the actual probability $\pi(\mathbf{x})$ and the estimated probability $\hat{\pi}(\mathbf{x})$.

These simulated runs have been done through building routines using four packages: SPSS, MATLAB, GAMS and Excel. The estimation of logistic model using maximum likelihood approach depends on two packages MATLAB (R2007b) and SPSS (16). MATLAB (R2007b) has been used for estimating the logistic regression model. This program uses the maximum likelihood (ML) method through a robust non-linear fitting that iteratively reweights response values and re-computes a least squares fit. The least squares component of the algorithm differs from linear least squares, but the reweighting loop is identical to that for robust linear methods. We built our own routines that allow us to simulate the mathematical programming models, which are based on MATLAB (R2007b) and GAMS (22.7). Moreover, MS Excel has been used to link and summarize the simulations results.

5. SIMULATION STUDY RESULTS

The logistic regression using the maximum likelihood (ML) and the mathematical programming (MP) approaches[6], and the Clusterwise Logistic Regression[7] Model have been experienced using the same runs. As a classification method, the correct classification percent is one of the important logistic regression's characteristics. The results from table 1 show that the correct

[6] The used runs in the case of logistic regression using maximum likelihood and mathematical programming have been published in Hamed et al., (2009).

[7] The simulated data are hypothesized to be classified to only two clusters.

classification percent of Clusterwise Logistic Model is not less than 96 percent and its average hits 98.7 percent. This result is notably higher than its counterpart in both ML and MP logistic models. Meanwhile the correct classification percent is significantly higher in MP logistic model than in ML logistic model for sample sizes 20 and 50. In reasonably large sample size 200 the correct classification percent shows some significant differences toward MP logistic model and in the large sample size 500 it reverses toward ML logistic model.

On the other hand, 75.6 percent of the 4000 runs do not include any tie, and 23.3 percent of runs do not have more than 10 percent of ties. However, the ties always have the same probability under any cluster. This characteristic is very important, whereas if the model produces any tie, these ties are not expected to have different estimated probability $\hat{\pi}(\mathbf{x})$ in the different clusters.

To demonstrate the advantage of the proposed model, the correct classification percent results are supported by the estimated mean squared error results, which are based on comparing the differences between the actual and the estimated probabilities. The actual probabilities are based on the actual parameters obtained from the simulation process. For the small, moderate and reasonably large sample sizes, the MSE in Clusterwise Logistic Model is higher than the corresponding ones in ML and MP logistic models. However, the MSE in Clusterwise Logistic Model is sharply decreased in the large sample size (500) to be lower than ML and MP logistic models.

6. COMMENTS AND CONCLUSION

This section focuses on the main findings and conclusions of the current paper. The paper introduces "Clusterwise Logistic Model" by using mathematical programming. The theoretical framing of this model is non-linear goal programming with a linear objective function and linear constraints. The model uses both of regression sum of residuals and sum of classification errors as clustering criteria. These criteria have been linked to the clustering membership parameters, which are introduced in a probabilistic form. Thus, at the level of subjects, less deviation from the fitted line and less margin of errors that is needed to clear the separation interval for a specific cluster, imply more probability to belong to this cluster.

The proposed model has been compared with the non-clustered logistic regression that is estimated by maximum likelihood and mathematical programming approaches through a set of pre-designed simulation scenarios.

Three main conclusions can be obtained from these scenarios. The first is that the mathematical programming approach contributes to the enhancement of correct classification, compared with the maximum likelihood approach. In addition, Clusterwise Logistic Model yields a significant enhancement. This means that the clustering approach succeeds in producing homogeneous clusters, which improves the overall correct classification. Secondly, the proposed mathematical programming model is not influenced with the common logistic regression drawbacks.

Hamed et al., (2009) used a published biostatistics case study to prove that their proposed MP logistic model is not influenced with the separation or monotone likelihood problem. As Clusterwise Logistic Model is based on MP logistic regression model, it is expected to maintain the same condition, and to produce finite parameters. Lastly, the suggested Clusterwise Logistic Model shows higher mean squared error than MP and ML logistic models'. However, this does not deprive the Clusterwise Logistic Model from its advantages, because its statistics are very

close to both MP and ML logistic models'. In addition, the theoretical advantage of the proposed Clusterwise Logistic Model is that it is flexible enough to accept any linear/nonlinear constraint(s), which is theoretically needed in many applied studies.

Table (1): Correct classification percent of ML and MP logistic regression models, and MP clusterwise logistic regression model for different covariates' combinations and different sample sizes, at 4000 runs.(1)

Sample Size	Distribution	Multicolinearity	Logistic (ML)	Logistic (MP)	Clusterwise Logistic (MP)
n=20	*Mixed*	*Collinear*	89.8	87.5	96.3
		Independent	94.3	93.0	98.4
	Continuous	*Collinear*	94.8	93.6	98.0
		Independent	95.3	94.3	97.8
	Discrete	*Independent*	92.5	91.2	97.1
n=50	*Mixed*	*Collinear*	89.6	88.7	98.9
		Independent	91.4	90.7	99.0
	Continuous	*Collinear*	92.3	91.7	98.7
		Independent	93.8	93.3	99.3
	Discrete	*Independent*	93.3	93.0	98.7
n=200	*Mixed*	*Collinear*	90.3	90.2	99.1
		Independent	92.4	92.2	99.5
	Continuous	*Collinear*	92.9	92.6	99.5
		Independent	93.5	93.5	99.4
	Discrete	*Independent*	92.4	92.3	98.6
n=500	*Mixed*	*Collinear*	90.4	90.6	99.4
		Independent	90.4	90.8	99.5
	Continuous	*Collinear*	91.2	91.5	99.1
		Independent	91.2	91.7	99.2
	Discrete	*Independent*	92.4	92.6	97.6

(1) Each cell consists of 200 runs.

Table (2): Mean squared error of ML and MP logistic regression models, and MP clusterwise logistic regression model for different covariates' combinations and different sample sizes, at 4000 runs.(1)

Sample Size	Distribution	Multicolinearity	Logistic (ML)	Logistic (MP)	Clusterwise Logistic (MP)
n=20	*Mixed*	*Collinear*	0.026	0.027	0.050
		Independent	0.022	0.030	0.039
	Continuous	*Collinear*	0.026	0.029	0.039
		Independent	0.029	0.033	0.041
	Discrete	*Independent*	0.014	0.019	0.033
n=50	*Mixed*	*Collinear*	0.015	0.016	0.042
		Independent	0.016	0.017	0.038
	Continuous	*Collinear*	0.019	0.018	0.034
		Independent	0.015	0.019	0.030
	Discrete	*Independent*	0.008	0.010	0.023
n=200	*Mixed*	*Collinear*	0.293	0.314	0.052
		Independent	0.006	0.005	0.043
	Continuous	*Collinear*	0.008	0.004	0.039
		Independent	0.007	0.005	0.035
	Discrete	*Independent*	0.004	0.003	0.026
n=500	*Mixed*	*Collinear*	0.008	0.002	0.000
		Independent	0.010	0.002	0.000
	Continuous	*Collinear*	0.013	0.002	0.000
		Independent	0.013	0.002	0.000
	Discrete	*Independent*	0.004	0.001	0.000

(1) Each cell consists of 200 runs.

REFERENCES

Agresti, A. (2002), *Categorical Data Analysis*, 2nd edition, John Wiley & Sons, Inc., Hoboken, New Jersey.

Bain, L. J., and Engelhardt, M. (1992), *Introduction to Probability and Mathematical Statistics*, 2nd edition, PWS-KENT publishing company, Boston, USA.

Christensen, R. (1997), *Log-Linear Models and Logistic Regression*, 2nd edition, Springer-Verlag, Inc., New York.

Draper, N. R. and Smith, H. (1998), *Applied Regression Analysis*, 3rd edition, John Wiley & Sons, Inc., New York.

El-Haik, B., and Al-Omar, R. (2006), *Simulation-based lean six-sigma and design for six-sigma*, John Wiley and Sons, Inc., Hoboken, New Jersey, USA.

Greene, W. H. (1997), *Econometric Analysis*, 3rd edition, Prentice-Hall International, Inc., USA.

Hamed, R., El Hefnawy, A., and Ramadan, M. (March 2009), "Logistic Regression Using Non-Linear Goal Programming", *the 21th annual conference on Statistics and Modeling in the Human and Social Sciences, Faculty of Economic and Political Science, Cairo university*.

Hosmer, D. W. and Lemeshow, S. (2000), *Applied Logistic Regression*, 2nd edition, John Wiley & Sons, Inc., New York.

Lau, K., Leung, P., and Tse, K. (1999), "A Mathematical Programming Approach to Clusterwise Regression Model and its Extensions", *European Journal of Operational Research*, Vol. 116, No. 3; 640-652.

Luo, Z. (2005), *Flexible Pavement Condition Model Using Clusterwise Regression and Mechanistic-Empirical Procedure for Fatigue Cracking Modeling*, unpublished Ph.D. thesis, College of Engineering, The University of Toledo, Toledo, OHIO.

Spath, H. (1979), "Algorithm 39: Clusterwise Linear Regression", *Computing*, Vol. 22; 367-373.

Spath, H. (1981), "Correction to Algorithm 39: Clusterwise Linear Regression", *Computing*, Vol. 26; 275.

Spath, H. (1982), "Algorithm 48: A Fast Algorithm for Clusterwise Linear Regression", *Computing*, 29; 175-181.

Spath, H. (1985), *Cluster Dissection and Analysis*, New York: Wiley.

Xie, X. (2005), *A Goodness-of-fit Test for Logistic Regression Models with Continuous Predictors*, unpublished Ph.D. thesis, the Graduate College of the University of Iowa, Iowa, USA.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 664–675.

A FAMILY OF ESTIMATORS FOR SINGLE AND TWO-PHASE SAMPLING USING TWO AUXILIARY ATTRIBUTES

Muhammad Hanif[1], Inam-ul-Haq[2] and Munir Ahmad[3]
[1] Lahore University of Management Sciences, Lahore, Pakistan
E-mail: hanif@lums.edu.pk
[2, 3]National College of Business Administration & Economics, Lahore, Pakistan
E-mail: [2]inam-ul-haq786@hotmail.com, [3]drmunir@ncbae.edu.pk

ABSTRACT

A general family of estimators has been proposed and general expression of mean square error of these estimators has been derived by Jhajj et al. (2006). In this paper we have suggested improved version of Jhajj et al. (2006) by using two auxiliary attributes, we have also suggested some new estimators. Mathematical comparisons of these estimators have been made. Empirical study has also been conducted to show that new estimators are more efficient. More over it is investigated numerically that full information cases are more efficient than partial and no information cases.

1. INTRODUCTION

Use of auxiliary information to increase the efficiency of estimators for population mean is an integral part of recently developed estimators. The use of auxiliary information in estimation process is as old as history of survey sampling. The first use of auxiliary information in survey sampling can be traced from the work of Neyman (1938). Generally the auxiliary variables are quantitative in nature but the use of qualitative auxiliary variables has been proposed in ratio, product and regression estimators by Naik and Gupta. (1996). A family of estimators using single auxiliary attribute has been introduced by Jhajj et al. (2006).

In this paper we have developed a set of estimators which are improved form of Jhajj et al. (2006) and Shabbir and Gupta (2007). For this let $(y_i, \tau_{i1}, \tau_{i2})$ be the ith sample point from a population of size N, where $\tau_j\,(j=1,2)$ is the value of jth auxiliary attribute. We suppose that the complete dichotomy is recorded for each attribute so that $\tau_{ij}=1$ if ith unit of population possesses jth attribute and τ_j = 0 otherwise. Let $A_j=\sum_{i=1}^{N}\tau_{ij}$ and $a_j=\sum_{i=1}^{n}\tau_{ij}$ be the total number of units in the population and sample respectively, possessing attribute τ_j. Let $P_j=N^{-1}A_j$ and $p_j=n^{-1}a_j$ be the corresponding proportion of units possessing attributes τ_j. Let us define $\bar{e}_y=\bar{y}-\bar{Y}$ and $\bar{e}_{\tau_j}=p_j-P_j$ with following properties:

$E(\bar{e}_y^2)=\theta S_y^2$ where $\theta=n^{-1}-N^{-1}$, $E(\bar{e}_y)=0=E(\bar{e}_{\tau_j})$, $E(\bar{e}_{\tau_j}^2)=\theta S_{\tau_j}^2$, $E(\bar{e}_y\bar{e}_{\tau_j})=\theta S_y S_{\tau_j}\rho_{pb_j}$,

$E(\bar{e}_{\tau_1}\bar{e}_{\tau_2})=\theta S_{\tau_1}S_{\tau_2}Q_{12}$ and $S_{y\tau_j}=\frac{1}{N-1}\sum_{j=1}^{N}(y_i-\bar{Y})(\tau_{ij}-P_j)$.

Suppose further that $\rho_{Pbj} = S_{y\tau_j} \big/ \left(S_y S_{\tau_j}\right)$ be the point bi-serial correlation coefficient and Q_{12} is the coefficient of association, where $-1 \le Q_{12} \le +1$. Let n_1 and n_2 be the size of first-phase and second-phase sample respectively, so that $n_2 < n_1$ and $p_{j(1)}$, $p_{j(2)}$ are proportion of units possessing attribute τ_j in first-phase and second-phase sample respectively. The mean of main variable of interest at second phase is denoted by $\bar{y}_2$. Also

$\bar{e}_{y_2} = \bar{y}_2 - \bar{Y}$, $\bar{e}_{\tau_j(1)} = p_{j(1)} - P_j$, $\bar{e}_{\tau_j(2)} = p_{j(2)} - P_j$ $(j = 1,2)$, $E\left(\bar{e}_{y_2}\right) = \theta_2 S_y^2$, $E\left(\bar{e}_{\tau_{j(1)}} - \bar{e}_{\tau_{j(2)}}\right)^2 = \theta_3 S_{\tau_j}^2$,
$E\left\{\bar{e}_{y_2}\left(\bar{e}_{\tau_{j(2)}} - \bar{e}_{\tau_{j(1)}}\right)\right\} = \theta_2 S_y S_{\tau_j} \rho_{pbj}$, $E\left[\left(\bar{e}_{\tau_{1(2)}} - \bar{e}_{\tau_{1(1)}}\right)\left(\bar{e}_{\tau_{2(2)}} - \bar{e}_{\tau_{2(1)}}\right)\right] = \theta_3 S_{\tau_1} S_{\tau_2} Q_{12}$, $\theta_3 = \theta_2 - \theta_1$, $\theta_1 = n_1^{-1} - N^{-1}$ and
$\theta_2 = n_2^{-1} - N^{-1}$.

2. SOME PREVIOUS ESTIMATORS BASED ON AUXILIARY ATTRIBUTES

In this section we have reproduced some previous estimators available in literature.

2.1 Single-Phase Sampling (Full Information Case)

i) If information on a single auxiliary attribute τ_1 is known then a family of estimator suggested by Jhajj et al. (2006) is given as

$$T_{1(1)} = g_\omega\left(\bar{y}, v_1\right), \tag{2.1}$$

where $v_1 = p_1 / P_1$ and $g_\omega(\bar{y}, v_1)$ is a parametric function of $\bar{y}$ and v_1 such that $g_\omega\left(\bar{Y}, 1\right) = \bar{Y}$, and satisfy certain regularity conditions. The mean square error of (2.1) is:

$$MSE\left(T_{1(1)}\right) \approx \theta\left(1 - \rho_{pb1}^2\right) S_y^2, \tag{2.2}$$

where ρ_{pb1}^2 is squared point bi-serial correlation coefficient.

ii) An estimator suggested by Shabbir and Gupta (2007) for full information case is

$$t_{2(1)} = \bar{y}\left[d_1 + d_2\left(P_1 - p_1\right)\right]\frac{P_1}{p_1}. \quad (p_1 > 0) \tag{2.3}$$

The values of d_1 and d_2 that minimizes $MSE\left(t_{2(1)}\right)$ are

$$d_1 = \frac{1}{1 + \theta\left(1 - \rho_{Pb_1}^2\right)\bar{Y}^{-2} S_y^2} \text{ and } d_2 = \frac{\left(\rho_{Pb_1} C_y - C_{P_1}\right)}{\left[1 + \theta\left(1 - \rho_{Pb_1}^2\right) C_y^2\right] S_{P_1}}.$$

The mean square error of $t_{2(1)}$ is

$$MSE\left(t_{2(1)}\right) \approx \frac{\theta\left(1 - \rho_{pb1}^2\right) S_y^2}{1 + \theta\left(1 - \rho_{pb1}^2\right)\bar{Y}^{-2} S_y^2}. \tag{2.4}$$

2.2 Two-Phase Sampling (No Information Case)

i) A family of estimator for two phase sampling by Jhajj et al. (2006) is

$$T_{3(2)} = g_{\omega}\left(\bar{y}_2, v_{1d}\right), \tag{2.5}$$

where $v_{1d} = p_{1(2)} / p_{1(1)}$, such that $g_{\omega}\left(\bar{Y},1\right) = \bar{Y}$. The mean square error of (2.5) is

$$MSE\left(T_{3(2)}\right) \approx \left(\theta_2 - \theta_3 \rho_{pb1}^2\right) S_y^2. \tag{2.6}$$

ii) An estimator was also suggested by Shabbir and Gupta (2007) for no information is

$$t_{4(2)} = \bar{y}_2 \left[W_1 + W_2\left(p_{1(1)} - p_{1(2)}\right)\right] \frac{p_{1(1)}}{p_{1(2)}}. \quad \left(p_{1(2)} > 0\right) \tag{2.7}$$

The expressions for W_1 and W_2 that minimizes $MSE\left(t_{4(2)}\right)$ are

$$W_1 = \frac{1}{1+\left(\theta_2 - \theta_3 \rho_{pb1}^2\right)\bar{Y}^{-2} S_y^2} \text{ and } W_2 = \frac{\left(\rho_{Pb_1} C_y - C_{P_1}\right)}{\left[1+\left(\theta_2 - \theta_3 \rho_{Pb_1}^2\right) C_y^2\right] S_{P_1}} .$$

Then the mean square estimator of $t_{4(2)}$ is

$$MSE\left(t_{4(2)}\right) \approx \frac{\left(\theta_2 - \theta_3 \rho_{pb1}^2\right) S_y^2}{1+\left(\theta_2 - \theta_3 \rho_{pb1}^2\right)\bar{Y}^{-2} S_y^2}. \tag{2.8}$$

3. A NEW ESTIMATORS FOR SINGLE AND TWO PHASE SAMPLING USING ONE ATTRIBUTE

An approximate estimator suggested by Shabbir and Gupta (2007) was not defined at $p_1 = 0$, therefore we are proposing a new exact estimator, which may be considered as an alternate suggested in Shabbir and Gupta (2007). This new approach has an advantage over estimator suggested in Shabbir and Gupta (2007), as it is defined for any value of sample proportion " p_1" and mean square error of proposed estimator is also exact because this new estimator do not contain any ratio.

The estimator for full information case using single attribute is

$$t_{5(1)} = d_0 \left[\bar{y} - d_1\left(p_1 - P_1\right)\right], \tag{3.1}$$

where d_0 and d_1 are unknown constants to be determined? The mean square error of $t_{5(1)}$ will be

$$MSE\left(t_{5(1)}\right) = \left(d_0 - 1\right)^2 \bar{Y}^2 + \theta d_0^2 \left[S_y^2 + d_1^2 S_{\tau_1}^2 - 2 d_1 S_y S_{\tau_1} \rho_{Pb_1}\right]. \tag{3.2}$$

Optimum value of d_0 and d_1 which minimize $MSE\left(t_{5(1)}\right)$ are

$$d_0 = \frac{1}{1+\theta\left(1-\rho_{Pb_1}^2\right)\bar{Y}^{-2}S_y^2} \text{ and } d_1 = \frac{S_y \rho_{pb_1}}{S_{\tau_1}} .$$

Using the value of d_0 and d_1 in (3.2) and on simplification we get

$$MSE\left(t_{5(1)}\right) \approx \frac{\theta\left(1-\rho_{pb1}^2\right)S_y^2}{1+\theta\left(1-\rho_{pb1}^2\right)\bar{Y}^{-2}S_y^2}, \tag{3.3}$$

which is exact unlike suggested Shabbir and Gupta (2007). Another suggested estimator for no information case is

$$t_{6(2)} = W_0\left[\bar{y}_2 - W_1\left(p_{1(2)} - p_{1(1)}\right)\right], \tag{3.4}$$

where W_0 and W_1 are constants to be determined. The mean square error of $t_{6(2)}$ will be

$$MSE\left(t_{6(2)}\right) = \left[\left(W_0-1\right)^2\bar{Y}^2 + W_0^2\left\{\theta_2 S_y^2 + \theta_3\left(W_1^2 S_{\tau_1}^2 - 2W_1 S_y S_{\tau_1}\rho_{Pb_1}\right)\right\}\right]. \tag{3.5}$$

The optimum value of W_0 and W_1, which minimize $MSE\left(t_{6(2)}\right)$ are

$$W_0 = \frac{1}{1+\left(\theta_2 - \theta_3\rho_{pb1}^2\right)\bar{Y}^{-2}S_y^2} \text{ and } W_1 = \frac{S_y \rho_{pb_1}}{S_{\tau_1}} .$$

Using the value of W_0 and W_1 in (3.5) and on simplification we get.

$$MSE\left(t_{6(2)}\right) = \frac{\left(\theta_2 - \theta_3\rho_{Pb_1}^2\right)S_y^2}{1+\left(\theta_2 - \theta_3\rho_{Pb_1}^2\right)\bar{Y}^{-2}S_y^2}, \tag{3.6}$$

which is exact unlike the one suggested Shabbir and Gupta (2007). This may be considered as an alternative to the one suggested Shabbir and Gupta (2007).

4. A NEW ESTIMATOR FOR FULL PARTIAL AND NO INFORMATION CASES USING TWO ATTRIBUTE

In this section we will propose a new estimator for single-phase sampling for full information case also for two-phase sampling (partial and no information cases) using two auxiliary attributes.

4.1 A New Estimator for Single Phase Sampling Using Two Auxiliary Attributes (Full Information Case)

In this section we are proposing a family of estimators for full information case by adding another attribute in estimator given by Jhajj et al. (2006).

$$T_{7(1)} = g_{\omega}(\bar{y}, v_1, v_2), \tag{4.1}$$

where $v_1 = \frac{p_1}{P_1}$, $v_2 = \frac{p_2}{P_2}$, $v_1 > 0, v_2 > 0$, p_1, p_2 are sample proportions possessing attributes τ_1 and τ_2 respectively. $g_{\omega}(\bar{y}, v_1, v_2)$ is the parametric function such that $g_{\omega}(\bar{Y}, 1, 1) = \bar{Y}$, and satisfying the point $(\bar{y}, v_1, v_2)$ to be in a bounded set in R_3 containing a point $(\bar{Y}, 1, 1)$. The attributes τ_1 and τ_2 are significantly correlated with main variable. We consider the following estimator of the family defined in equation (4.1) i.e.

$$t_{7(1)} = \bar{y} + \alpha_1(v_1 - 1) + \alpha_1(v_2 - 1) \ , \tag{4.2}$$

where α_1 and α_2 are constant and are to be determined. The mean square error is

$$MSE(t_{7(1)}) = \theta\left[S_y^2 + \alpha_1^2 \frac{S_{\tau_1}^2}{P_1^2} + \alpha_2^2 \frac{S_{\tau_2}^2}{P_2^2} + 2\alpha_1 S_y \frac{S_{\tau_1}}{P_1} \rho_{Pb_1} + 2\alpha_2 S_y \frac{S_{\tau_2}}{P_2} \rho_{Pb_2} + 2\alpha_1 \alpha_2 \frac{S_{\tau_1}}{P_1} \frac{S_{\tau_2}}{P_2} Q_{12} \right]. \tag{4.3}$$

Optimum values of α_1 and α_2 are,

$$\alpha_1 = \frac{-P_1 S_y \left(\rho_{Pb_1} - Q_{12} \rho_{Pb_2} \right)}{S_{\tau_1} \left(1 - Q_{12}^2 \right)} \ , \ \alpha_2 = \frac{-P_2 S_y \left(\rho_{Pb_2} - Q_{12} \rho_{Pb_1} \right)}{S_{\tau_2} \left(1 - Q_{12}^2 \right)} \ .$$

Using α_1 and α_2 in (4.3) we get

$$MSE(t_{7(1)}) = \theta\left(1 - \rho_{y.\tau_1\tau_2}^2\right) S_y^2, \tag{4.4}$$

where $\rho_{y.\tau_1\tau_2}$ is multiple bi-serial correlation coefficient. We are also proposing a regression type estimator for full information case using two auxiliary attributes i.e.

$$t_{8(1)} = \gamma_0 \left[\bar{y} - \gamma_1 (p_1 - P_1) - \gamma_2 (p_2 - P_2) \right], \tag{4.5}$$

where γ_0, γ_1 and γ_2 are unknown constants to be determined. The mean square error of $t_{8(1)}$ will be,

$$MSE\left(t_{8(1)}\right)=\left(\gamma_0-1\right)^2\bar{Y}^2+\theta\gamma_0^2\begin{bmatrix}S_y^2+\gamma_1^2S_{\tau_1}^2+\gamma_2^2S_{\tau_2}^2-2\gamma_1S_yS_{\tau_1}\rho_{Pb_1}\\-2\gamma_2S_yS_{\tau_2}\rho_{Pb_2}+2\gamma_1\gamma_2S_{\tau_1}S_{\tau_2}Q_{12}\end{bmatrix}. \tag{4.6}$$

Optimum value of γ_0, γ_1 and γ_2 are

$$\gamma_0=\frac{1}{1+\theta\left(1-\rho_{y.\tau_1\tau_2}^2\right)\bar{Y}^{-2}S_y^2},\ \gamma_1=\frac{P_1S_y\left(\rho_{Pb_1}-Q_{12}\rho_{Pb_2}\right)}{S_{\tau_1}\left(1-Q_{12}^2\right)},$$

and

$$\gamma_2=\frac{P_2S_y\left(\rho_{Pb_2}-Q_{12}\rho_{Pb_1}\right)}{S_{\tau_2}\left(1-Q_{12}^2\right)}.$$

Using the value of γ_0, γ_1 and γ_2 in (4.6) and on simplification we get,

$$MSE(t_{8(1)})=\frac{\theta\left(1-\rho_{y.\tau_1\tau_2}^2\right)S_y^2}{1+\theta\left(1-\rho_{y.\tau_1\tau_2}^2\right)\bar{Y}^{-2}S_y^2}. \tag{4.7}$$

4.2 A New Estimator for Single Phase Sampling Using Two Auxiliary Attributes (Partial and No Information Cases)

In this section two cases will be discussed, one for partial information and other for no information.

4.2.1 Partial Information Case

We propose a family of estimators as

$$T_{9(2)}=g_\omega\left(\bar{y},v_1,v_{2d}\right), \tag{4.8}$$

where $v_1=\frac{p_{1(2)}}{P_1}, v_{2d}=\frac{p_{2(2)}}{p_{2(1)}},\ v_1>0, v_{2d}>0$, P_1 is known but P_2 is not known, where $g_\omega\left(\bar{y},v_1,v_{2d}\right)$ is parametric function such that $g_\omega\left(\bar{Y},1,1\right)=\bar{Y}$, and satisfying condition mentioned for (4.1). We consider the following estimator of the family defined in equation (4.8)

$$t_{9(2)}=\bar{y}_2+\alpha_1'\left(v_1-1\right)+\alpha_2'\left(v_{2d}-1\right), \tag{4.9}$$

where α_1' and α_2' are constants to be determined. The mean square error of $t_{9(2)}$ will be.

$$MSE\left(t_{9(2)}\right)=\theta_2\left[S_y^2+\alpha_1'^2\frac{S_{\tau_1}^2}{P_1^2}+2\alpha_1'S_y\frac{S_{\tau_1}}{P_1}\rho_{Pb_1}\right]$$
$$+\theta_3\left[\alpha_2'^2\frac{S_{\tau_2}^2}{P_2^2}+2\alpha_2'\bar{Y}S_y\frac{S_{\tau_2}}{P_2}\rho_{Pb_2}+2\alpha_1'\alpha_2'\frac{S_{\tau_1}}{P_1}\frac{S_{\tau_2}}{P_2}Q_{12}\right]. \tag{4.10}$$

The optimum values of α_1' and α_2' are

$$\alpha_1' = \frac{-P_1 S_y \left(\theta_2 \rho_{Pb_1} - Q_{12}\theta_3 \rho_{Pb_2}\right)}{S_{\tau_1}\left(\theta_2 - \theta_3 Q_{12}^2\right)} \ , \ \alpha_2' = \frac{-\theta_2 P_2 S_y \left(\rho_{Pb_2} - Q_{12}\rho_{Pb_1}\right)}{S_{\tau_2}\left(\theta_2 - \theta_3 Q_{12}^2\right)} \ .$$

Using the value of α_1' and α_2' in (4.10) and on simplification we get

$$MSE\left(t_{9(2)}\right) = \theta_2 \left[1 - \frac{\theta_2 \rho_{Pb_1}^2 + \theta_3 \rho_{Pb_2}^2 - 2\theta_3 Q_{12}\rho_{Pb_1}\rho_{Pb_2}}{\left(\theta_2 - \theta_3 Q_{12}^2\right)}\right] S_y^2 \ . \tag{4.11}$$

Regression type estimator using two auxiliary attributes also has been suggested i.e.

$$t_{10(2)} = \delta_0 \left[\bar{y}_2 - \delta_1\left(p_{1(2)} - P_1\right) - \delta_2\left(p_{2(2)} - p_{2(1)}\right)\right], \tag{4.12}$$

where δ_0, δ_1, δ_2 are unknown constants to be determined . The mean square error of $t_{10(2)}$ will be.

$$MSE\left(t_{10(2)}\right) = \left[\begin{array}{l} \left(\delta_0 - 1\right)^2 \bar{Y}^2 + \delta_0^2 \left\{\begin{array}{l} \theta_2\left(\bar{Y}^2 S_y^2 + \delta_1^2 S_{\tau_1}^2 - 2\delta_1 S_y S_{\tau_1}\rho_{Pb_1}\right) + \\ \theta_3\left(+\delta_2^2 S_{\tau_2}^2 - 2\delta_2 S_y S_{\tau_2}\rho_{Pb_2} + 2\delta_1\delta_2 S_{\tau_1} S_{\tau_2} Q_{12}\right) \end{array}\right\} \end{array}\right] \tag{4.13}$$

The optimum value of δ_0, δ_1, δ_2 are

$$\delta_0 = \frac{1}{1 + \theta_2\left[1 - \frac{\theta_2 \rho_{Pb_1}^2 + \theta_3 \rho_{Pb_2}^2 - 2\theta_3 Q_{12}\rho_{Pb_1}\rho_{Pb_2}}{\left(\theta_2 - \theta_3 Q_{12}^2\right)}\right]\bar{Y}^{-2} S_y^2},$$

$$\delta_1 = \frac{P_1 S_y \left(\theta_2 \rho_{Pb_1} - Q_{12}\theta_3 \rho_{Pb_2}\right)}{S_{\tau_1}\left(\theta_2 - \theta_3 Q_{12}^2\right)}, \text{ and } \delta_2 = \frac{\theta_2 P_2 S_y \left(\rho_{Pb_2} - Q_{12}\rho_{Pb_1}\right)}{S_{\tau_2}\left(\theta_2 - \theta_3 Q_{12}^2\right)}.$$

Using the value of δ_0, δ_1, δ_2 in (4.13) and on simplification we get.

$$MSE(t_{10(2)}) = \frac{\theta_2\left[1 - \frac{\theta_2 \rho_{Pb_1}^2 + \theta_3 \rho_{Pb_2}^2 - 2\theta_3 Q_{12}\rho_{Pb_1}\rho_{Pb_2}}{\left(\theta_2 - \theta_3 Q_{12}^2\right)}\right] S_y^2}{1 + \theta_2\left[1 - \frac{\theta_2 \rho_{Pb_1}^2 + \theta_3 \rho_{Pb_2}^2 - 2\theta_3 Q_{12}\rho_{Pb_1}\rho_{Pb_2}}{\left(\theta_2 - \theta_3 Q_{12}^2\right)}\right]\bar{Y}^{-2} S_y^2}. \tag{4.14}$$

4.2.2 No Information Case

Like full information we propose a family of estimator for no information case, under same condition mentioned for (4.1), i.e.

$$T_{11(2)} = g_\omega\left(\bar{y}_2, v_{1d}, v_{2d}\right) \ , \tag{4.15}$$

where $v_{1d} = \frac{p_{1(2)}}{P_{1(1)}}$, $v_{2d} = \frac{p_{2(2)}}{P_{2(1)}}$, $v_{1d} > 0, v_{2d} > 0$ and $g_\omega(\bar{y}_2, v_{1d}, v_{2d})$ is parametric function such that $g_\omega(\bar{Y}, 1, 1) = \bar{Y}$. We consider the following estimator of the family defined in (4.15)

$$t_{11(2)} = \bar{y}_2 + \alpha_1(v_{1d} - 1) + \alpha_2(v_{2d} - 1), \tag{4.16}$$

The mean square error of $t_{11(2)}$ will be

$$MSE(t_{11(2)}) = \left[\theta_2 S_y^2 + \theta_3 \left\{\alpha_1^2 \frac{S_{\tau_1}^2}{P_1^2} + \alpha_2^2 \frac{S_{\tau_2}^2}{P_2^2} + 2\alpha_1 S_y \frac{S_{\tau_1}}{P_1}\rho_{Pb_1} + 2\alpha_2 S_y \frac{S_{\tau_2}}{P_2}\rho_{Pb_2} + 2\alpha_1\alpha_2 \frac{S_{\tau_1}}{P_1}\frac{S_{\tau_2}}{P_2}Q_{12}\right\}\right]. \tag{4.17}$$

the optimum value α_1 and α_2 in (4.17) are same as derived for full information case, using the value of α_1 and α_2 in (4.17) and on simplification we get.

$$MSE(t_{11(2)}) = \left\{\theta_2\left(1 - \rho_{y.\tau_1\tau_2}^2\right) + \theta_1\rho_{y.\tau_1\tau_2}^2\right\} S_y^2. \tag{4.18}$$

We also propose regression type estimator for no information case, i.e.

$$t_{12(2)} = \gamma_0^*\left[\bar{y}_2 - \gamma_1\left(p_{1(2)} - p_{1(1)}\right) - \gamma_2\left(p_{2(2)} - p_{2(1)}\right)\right], \tag{4.19}$$

where γ_0^*, γ_1 and γ_2 are constants to be determined. The mean square error of $t_{12(2)}$ will be,

$$MSE(t_{12(2)}) = \left[\left(\gamma_0^* - 1\right)^2 \bar{Y}^2 + \left(\gamma_0^*\right)^2 \left\{\theta_2 \bar{Y}^2 C_y^2 + \theta_3 \begin{pmatrix} \gamma_1^2 S_{\tau_1}^2 + \gamma_2^2 S_{\tau_2}^2 - 2\gamma_1 S_y S_{\tau_1}\rho_{Pb_1} \\ -2\gamma_2 S_y S_{\tau_2}\rho_{Pb_2} + 2\gamma_1\gamma_2 S_{\tau_1} S_{\tau_2} Q_{12} \end{pmatrix}\right\}\right]. \tag{4.20}$$

Optimum values of γ_0^*, is $\gamma_0^* = \frac{1}{1 + \left\{\theta_2\left(1 - \rho_{y.\tau_1\tau_2}^2\right) + \theta_1\rho_{y.\tau_1\tau_2}^2\right\}\bar{Y}^{-2} S_y^2}$ while optimum values γ_1 and γ_2 in (5.13) are same as given in full information case. Using the value of γ_0^*, γ_1 and γ_2 in (5.13) and on simplification we get.

$$MSE(t_{12(2)}) = \frac{\left\{\theta_2\left(1 - \rho_{y.\tau_1\tau_2}^2\right) + \theta_1\rho_{y.\tau_1\tau_2}^2\right\} S_y^2}{1 + \left\{\theta_2\left(1 - \rho_{y.\tau_1\tau_2}^2\right) + \theta_1\rho_{y.\tau_1\tau_2}^2\right\}\bar{Y}^{-2} S_y^2}. \tag{4.21}$$

There could be number of estimators of families proposed in (4.1), (4.8) and (4.15). Some estimators of family proposed in (4.1) are

i) $\bar{y} + \alpha_1(v_1 - 1) + \alpha_1(v_2 - 1)$

ii) $\bar{y} V_1^{\alpha_1} V_2^{\alpha_2}$,

iii) $\bar{y}\, e^{\alpha_1(V_1 - 1) + \alpha_2(V_2 - 1)}$

iv) $\bar{y}\left(V_1 e^{(V_1-1)}\right)^{\alpha_1}\left(V_2 e^{(V_2-1)}\right)^{\alpha_2}$

v) $\bar{y}V_1^{\alpha_1} e^{\alpha_2(V_2-1)}$

vi) $\frac{\bar{y}}{2}\left[V_1^{\alpha_1}V_2^{\alpha_2} + e^{\alpha_1(V_1-1)+\alpha_2(V_2-1)}\right]$,

vii) $\bar{y} + \alpha_1\left(V_1^{\alpha_3} - 1\right) + \alpha_2\left(V_2^{\alpha_4} - 1\right)$,

viii) $\bar{y} + \alpha_1\left(V_1^{\alpha_3} - 1\right) + \alpha_2\left(V_2 - 1\right)$,

ix) $\frac{\bar{y}}{k_1 + k_2}\left[k_1 V_1^{\frac{\alpha_1}{2}} + k_2\, e^{\alpha_2(V_2-1)}\right]$,

x) $\bar{y}\left[k\, e^{\alpha_1(V_1-1)} + (1-k)e^{\alpha_2(V_2-1)}\right]$,

The mean square error of all these estimators have been derived and found same for all ten estimators. This is also there for partial and no information case. It can be easily verified that $MSE\left(t_{7(1)}\right) \le MSE\left(t_{1(1)}\right)$ also $MSE\left(t_{8(1)}\right) \le MSE\left(t_{2(1)}\right)$. It shows that $t_{7(1)}$ and $t_{8(1)}$ are more efficient than $t_{1(1)}$ and $t_{2(1)}$ respectively. Similarly it can be shown that $MSE\left(t_{11(2)}\right) \le MSE\left(t_{3(2)}\right)$ also $MSE\left(t_{12(2)}\right) \le MSE\left(t_{4(2)}\right)$, which shows that $t_{11(2)}$ and $t_{12(2)}$ are more efficient than $t_{3(2)}$ and $t_{4(2)}$ respectively.

5. EMPIRICAL STUDY COMMENTS AND CONCLUSION

Twelve populations are taken from Government of Pakistan (1998). It is shown empirically in table-2 that proposed estimator $t_{8(1)}$ out perform other competing estimators as it has maximum efficiency in almost all the populations. Also $t_{12(2)}$ performs best in almost all the populations. We conclude that $t_{8(1)}$ and $t_{12(2)}$ are more efficient than the other estimators in single phase and two-phase sampling. It is further observed full information case is always more efficient than no information case.

The optimum value of α_1 and α_2 involve some population parameters, which are assumed to be known for the efficient use of proposed family $T_{7(1)}$. In case these parameters are unknown, these can be estimated from the sample. If we follow approach of Srivastava and Jhajj (1983), the estimator of proposed family, $T_{7(1)}$ will have the same minimum mean square, if we replace the unknown value of parameters involved in optimum value of α_1 and α_2 with their estimators. Similar is the case for other proposed estimators and families.

Proposed estimator $t_{10(1)}$ is recommended to estimate the population mean for full information case as $t_{10(1)}$ outperform all the existing estimators for full information. Similarly $t_{12(2)}$ is recommended to estimate the population mean for no information case as $t_{12(2)}$ outperform all the existing estimators for no information.

It is also recommended that full information should always be preferred if possible, otherwise partial information are the best choice, no information case are recommended when we have no other choice. It can easily observed from table 3 that the estimators based on full information case are always more efficient than estimators based on partial and no information. It can also be

observed that estimators based on partial information case are always more efficient than estimators based no information.

APPENDIX

Table-1: Description of Populations and Variables

Pop #	Description	Main Variable	Attribute-I (τ_1) is present if	Attribute-II(τ_2) is present if
1	District-wise area and production of Vegetables for year 1995-96	Production of Vegetables (In tones)	Districts of N.W.F.P (including Fata Areas)	Area of Districts less than 500 hectors.
2	District-wise area and production of Vegetables for year 1996-97	Production of Vegetable (In tones)	Districts of Punjab	Area of Districts less than 401 hectors.
3	District-wise area and production of Vegetables for year 1997-98	Production of Vegetables (In tones)	Districts of Punjab	Area of Districts greater than 1000 hectors.
4	District-wise area and production of all Fruits for year 1995-96	Production of all Fruits (In tones)	Area of Districts greater than 1000 hectares	Districts of Punjab
5	District-wise area and production of all Fruits for year 1996-97	Production of all Fruits (In tones)	Districts of Sind	Area of Districts less than 1000 hectors.
6	District-wise area and production of all Fruits for year 1997-98	Production of all Fruits (In tones)	Districts of N.W.F.P (including Fata Areas)	Area of Districts less than 500 hectors.
7	District-wise area and production of Wheat for year 1995-96	Production of Wheat (In tones)	Area of Districts greater than 30 hectors.	Districts of Punjab
8	District-wise area and production of Wheat for year 1996-97	Production of Wheat (In tones)	Districts of Punjab	Area of Districts greater than 35 hectors.
9	District-wise area and production of Wheat for year 1997-98	Production of Wheat (In tones)	Districts of N.W.F.P (including Fata Areas)	Area of Districts greater than 25 hectors.
10	District-wise area and production of Onion for year 1995-96	Production of Onions (In tones)	Area of Districts greater than 40 hectors.	Districts of N.W.F.P (including Fata Areas)
11	District-wise area and production of Onion for year 1996-97	Production of Onions (In tones)	Area of Districts greater than 50 hectors.	Districts of N.W.F.P (including ata Areas)
12	District-wise area and production of Onion for year 1997-98	Production of Onions (In tones)	Districts of Punjab	Area of Districts greater than 60 hectors.

Table 2 Relative Efficiency of Various Estimators

Pop #	Single-Phase Sampling (Full information case)					Two-Phase Sampling (No information case)			
	$\bar{y}$	$T_{1(1)}$	$t_{2(1)}$	$t_{7(1)}$	$t_{8(1)}$	$T_{3(2)}$	$t_{4(2)}$	$t_{11(2)}$	$t_{12(2)}$
1	100	108.77	120.10	118.20	129.53	104.62	115.94	109.20	120.52
2	100	142.25	153.36	149.52	160.63	119.40	130.50	122.12	133.23
3	100	142.03	152.92	143.39	154.27	119.31	130.20	119.84	130.72
4	100	122.30	134.59	124.56	136.84	111.08	123.37	112.09	124.375
5	100	102.49	114.96	126.02	138.34	101.11	113.58	112.73	125.20
6	100	111.41	123.65	114.58	126.82	105.93	118.17	107.48	119.72
7	100	146.61	155.30	239.88	248.56	115.15	123.83	140.33	149.01
8	100	225.50	233.66	268.09	276.28	147.90	156.08	157.466	165.65
9	100	125.40	132.80	186.98	194.42	113.36	120.80	137.125	144357
10	100	105.90	137.40	107.13	138.63	103.40	134.90	104.10	135.60
11	100	107.07	136.09	107.76	136.96	104.04	133.14	104.43	133.63
12	100	101.80	129.20	109.90	137.30	101.05	128.45	105.60	133.00

Table 3 Comparison of full, partial and no information for Proposed Generalized Estimators ($t_{7(1)}, t_{9(2)}, t_{11(2)}$) of Jhajj et al. (2006) and Generalized New estimators ($t_{8(1)}, t_{10(2)}, t_{12(2)}$) (Relative efficiency)

Pop #	Generalized Estimators ($t_{7(1)}, t_{9(2)}, t_{11(2)}$) of Jhajj et al. (2006)			Generalized New estimators ($t_{8(1)}, t_{10(2)}, t_{12(2)}$)		
	Relative efficiency of full & partial information to no information		Relative efficiency of full information to partial information	Relative efficiency of full & partial information to no information		Relative efficiency of full information to partial information
	$t_{7(1)}$ (Full information)	$t_{9(2)}$ (Partial information)	$t_{7(1)}$ (Full information)	$t_{8(1)}$ (Full information)	$t_{10(2)}$ (Partial information)	$t_{12(2)}$ (Full information)
1	108.2466	103.5194	104.56646	107.4748	103.192	104.1504
2	122.4347	116.9979	104.64689	120.5643	115.5812	104.3114
3	119.6564	118.7813	100.73673	118.0197	117.2217	100.6808
4	111.129	109.8239	101.18835	110.0295	109.008	100.9371
5	111.7863	100.9679	110.71469	110.6128	100.9286	109.5951
6	106.6035	104.9007	101.62329	105.9285	104.2866	101.5745
7	170.9397	108.7959	157.11966	166.8066	108.2819	154.0484
8	170.2505	144.9148	117.48314	166.7876	142.6934	116.8853
9	136.3569	111.3599	122.44703	134.4831	110.7719	121.4053
10	102.9238	102.5836	100.3317	102.2447	101.9841	100.2555
11	103.192	102.7445	100.43559	102.4953	102.1453	100.3426
12	104.0688	100.7416	103.30275	103.2306	100.5894	102.6257

ACKNOWLEDGEMENTS

We are very thankful to Professor M. Samiuddin, whose suggestions helped improve the presentation of the paper.

REFERENCES

Government of Pakistan (1998). *Crop Area Production by Districts (1995-96 to 1997-98)*: Ministry of Food, Agriculture and Livestock. Food, Agriculture and Livestock Division, Economic Wing, Islamabad.

Jhajj, H.S., Sharma, M.K. and Grover, L.K. (2006). A family of estimators of population mean using information on auxiliary attribute. *Pak. J. Statist.*, 22(1), 43-50.

Neyman, J. (1938). Contributions to the theory of sampling human populations. *J. Amer. Statist. Assoc.*, 33, 101-116.

Naik, V.D. and Gupta, P.C. (1996). A note on estimation of mean with known population of an auxiliary character. *J. Ind. Soc. of Agri. Statist.*, 48(2), 151-158.

Shabbir, J. and Gupta, S. (2007). On estimating the finite population mean with known population proportion of an auxiliary variable. *Pak. J. Statist.*, 23(1), 1-9.

Srivastava, S.K. and Jhajj, H.S. (1983). A class of estimators of the population means using multi-auxiliary information. *Calcutta Statistics Association Bulletin*, 32, 47-56.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 676–684.

GENERALIZATION OF ESTIMATORS FOR FULL PARTIAL AND NO INFORMATION USING MULTI-AUXILIARY ATTRIBUTES

Muhammad Hanif[1], Inam-ul-Haq[2] and Muhammad Qaiser Shahbaz[3]
[1]Lahore University of Management Sciences, Lahore, Pakistan.
[2]National College of Business Administration & Economics, Lahore, Pakistan.
[3]Department of Mathematics, COMSATS Institute of Information Technology, Lahore, Pakistan
E-mail: [1]hanif@lums.edu.pk, [2]inam-ul-haq786@hotmail.com, [3]qshahbaz@gmail.com

ABSTRACT

A general family of estimators has been proposed and general expression of mean square error of these estimators has been derived by Jhajj et al. (2006). In this paper we have proposed a generalized family of estimators based on the information of "*k*" auxiliary attributes. Three different cases have been discussed that include the full, partial and no information cases. The family has been proposed for single-phase sampling in case of full information and for two-phase sampling in case of partial and no information cases. The expression for mean square error has been derived in all three cases. It is found that the proposed family has smaller mean square error than given by Jhajj et al. (2006).

1. INTRODUCTION

A family of estimators using single auxiliary attribute has been introduced by Jhajj et al. (2006). In this paper, we have proposed a new class of estimator by using information on "k" auxiliary attributes. The new class of estimators is a general extension of the class of estimators proposed by Jhajj et al. (2006). For this let $(y_i, \tau_{i1}, \tau_{i2}, \ldots \tau_{ik})$ be the ith sample point from a population of size N, where τ_j (j=1,2,…,k) is the value of jth auxiliary attribute. We suppose that the complete dichotomy is recorded for each attribute so that $\tau_{ij} = 1$ if ith unit of population possesses jth attribute, τ_j, and 0 otherwise. Let $A_j = \sum_{i=1}^{N} \tau_{ij}$ and $a_j = \sum_{i=1}^{n} \tau_{ij}$ be the total number of units in the population and sample respectively, possessing attribute τ_j. Let $P_j = N^{-1}A_j$ and $p_j = n^{-1}a_j$ be the corresponding proportion of units possessing attributes τ_j. Let us define $\bar{e}_y = \bar{y} - \bar{Y}$ and $\bar{e}_{\tau_j} = p_j - P_j$ with following properties:

$E(\bar{e}_y^2) = \theta S_y^2$, $E(\bar{e}_{\tau_j}^2) = \theta S_{\tau_j}^2$, $E(\bar{e}_y \bar{e}_{\tau_j}) = \theta S_y S_{\tau_j} \rho_{pb_j}$, $E(\bar{e}_{\tau_j} \bar{e}_{\tau_\psi}) = \theta S_{\tau_j} S_{\tau_\psi} Q_{j\psi}$ & $j \neq \psi$, $E(\bar{e}_y) = 0 = E(\bar{e}_{\tau_j})$

,where $\theta = n^{-1} - N^{-1}$

and $S_{y\tau_j} = \frac{1}{N-1} \sum_{j=1}^{N} (y_i - \bar{Y})(\tau_{ij} - P_j)$.

Suppose further that $\rho_{Pbj} = S_{y\tau_j} \big/ \left(S_y S_{\tau_j}\right)$ be the point bi-serial correlation coefficient and $Q_{j\psi}$ $\left(-1 \le Q_{j\psi} \le +1\right)$ is coefficient of association. Let n_1 and n_2 be the size of first-phase and second-phase sample respectively, so that $n_2 < n_1$ and $p_{j(1)}$, $p_{j(2)}$ are proportion of units possessing attribute τ_j in first-phase and second-phase sample respectively. The mean of main variable of interest at second phase is denoted by $\bar{y}_2$. Also $\bar{e}_{y_2} = \bar{y}_2 - \bar{Y}$, $\bar{e}_{\tau_j(1)} = p_{j(1)} - P_j$, $\bar{e}_{\tau_j(2)} = p_{j(2)} - P_j$ (j=1,2…k), $\theta_3 = \theta_2 - \theta_1$. We also define following expectations:

$$E\left(\bar{e}_{y_2}\right) = \theta_2 S_y^2,\ E\left(\bar{e}_{\tau_{j(1)}} - \bar{e}_{\tau_{j(2)}}\right)^2 = \theta_3 S_{\tau_j}^2,\ E\left\{\bar{e}_{y_2}\left(\bar{e}_{\tau_{j(2)}} - \bar{e}_{\tau_{j(1)}}\right)\right\} = \theta_3 S_y S_{\tau_j} \rho_{pbj},$$

$$E\left[\left(\bar{e}_{\tau_{j(2)}} - \bar{e}_{\tau_{j(1)}}\right)\left(\bar{e}_{\tau_{\psi(2)}} - \bar{e}_{\tau_{\psi(1)}}\right)\right] = \theta_3 S_{\tau_j} S_{\tau_\psi} Q_{j\psi},\quad \underline{\phi}'_{y\tau} \Phi_\tau^{-1} \underline{\phi}_{y\tau} = \left(\rho^2_{y.\tau_1\tau_2\ldots\tau_k}\right) S_y^2$$

$$\underline{\phi}'_{y\tau_1} \Phi_{\tau_1}^{-1} \underline{\phi}_{y\tau_1} = \left(\rho^2_{y.\tau_1\tau_2\ldots\tau_m}\right) S_y^2\ ,\ \underline{\phi}'_{y\tau_2} \Phi_{\tau_2}^{-1} \underline{\phi}_{y\tau_2} = \left(\rho^2_{y.\tau_{m+1}\tau_{m+2}\ldots\tau_k}\right) S_y^2,$$

where $\rho^2_{y.\tau_1\tau_2\ldots\tau_k}$ is squared multiple bi-serial correlation coefficient. These notations will be used in developing the mean square error of the new family of estimators.

2. SOME PREVIOUS ESTIMATORS BASED ON AUXILIARY ATTRIBUTES

2.1 Single-Phase Sampling (Full Information Case)

i) If information on a single auxiliary attribute τ_1 is known then a family of estimator suggested by Jhajj et al. (2006) is given as

$$T_{1(1)} = g_\omega\left(\bar{y}, v_1\right), \tag{2.1}$$

where $v_1 = p_1 / P_1$ and $g_\omega(\bar{y}, v_1)$ is a parametric function of $\bar{y}$ and v_1 such that $g_\omega(\bar{Y}, 1) = \bar{Y}$, and satisfy certain regularity conditions. The mean square error of (2.1) is:

$$MSE\left(T_{1(1)}\right) \approx \theta\left(1 - \rho^2_{pb1}\right) S_y^2, \tag{2.2}$$

where ρ^2_{pb1} is squared point bi-serial correlation coefficient.

2.2 Two-Phase Sampling (No Information Case)

A family of estimator for two phase sampling by Jhajj et al. (2006) is given as

$$T_{3(2)} = g_\omega\left(\bar{y}_2, v_{1d}\right), \tag{2.3}$$

where $v_{1d} = p_{1(2)} \big/ p_{1(1)}$, such that $g_\omega\left(\bar{Y}, 1\right) = \bar{Y}$. The mean square error of (2.3) is:

$$MSE\left(T_{3(2)}\right) \approx \left(\theta_2 - \theta_3 \rho^2_{pb1}\right) S_y^2. \tag{2.4}$$

In the following section we develop the general family of estimators by using information on "k" auxiliary attributes.

3. NEW FAMILY OF ESTIMATORS

3.1 Generalized Estimator using "k" Auxiliary Attributes for Full Information Case

Suppose that population proportion P_j is known for all the auxiliary attributes. Using this full information we propose a general family of estimators as:

$$T_{3(1)} = g_{\omega}(\bar{y}, v_1, v_2, ..., v_k), \tag{3.1}$$

where $v_j = p_j / P_j, v_j > 0$ and p_j is the sample proportions of jth attributes. Also $g_{\omega}(\bar{y}, v_1, v_2, ..., v_k)$ is the parametric function such that $g_{\omega}(\bar{y}, v_1, v_2, ..., v_k) = \bar{Y}$, and the point $(\bar{y}, v_1, v_2, ..., v_k)$ are to be in a bounded set in R_k containing a point $(\bar{Y}, 1, 1, ..., 1)$. The attributes τ_j are significantly correlated with main variable. Some families of estimators under above conditions may be formulated as,

i) $$t_{3(1)} = \bar{y} + \sum_{j=1}^{k} \alpha_j \left(v_j - 1 \right) = \bar{y} + \underline{\alpha}' \left(\underline{v} - 1 \right) = \bar{y} + \underline{\alpha}' \underline{\varphi}, \tag{3.2}$$

ii) $$t_{4(1)} = \bar{y} \left[(v_1)^{\alpha_1} (v_2)^{\alpha_2} ... (v_k)^{\alpha_k} \right], \tag{3.3}$$

where $\underline{\alpha}_{(k x 1)} = \left[\alpha_j \right]$, $\underline{v}_{(k x 1)} = \left[v_j \right]$, $\underline{\varphi} = \underline{v} - \underline{1}$ and $\underline{1}$ is a vector of one's. Using $\bar{y} = \bar{Y} + \bar{e}_y$ in (3.2), squaring and applying expectation we have:

$$MSE\left(t_{3(1)}\right) = \theta \left[S_y^2 + \underline{\alpha}' \Phi_{\tau} \underline{\alpha} + 2 \underline{\alpha}' \underline{\phi}_{y\tau} \right], \tag{3.4}$$

where $\theta = n^{-1} - N^{-1}$, $E\left(\underline{\varphi}\underline{\varphi}'\right) = \theta \Phi_{\tau}$ is the covariance matrix of $\underline{\varphi}$ and $E\left(\underline{\varphi}\bar{e}_y\right) = \theta \underline{\phi}_{y\tau}$ is the vector of covariance between Y and $\underline{\varphi}$. Partially differentiating (3.4) with respect to $\underline{\alpha}$ and equating the derivative to zero we have

$$\underline{\alpha} = -\Phi_{\tau}^{-1} \underline{\phi}_{y\tau}. \tag{3.5}$$

Using (3.5) in (3.4) we have,

$$MSE\left(t_{3(1)}\right) = \theta \left(1 - \rho^2_{y.\tau_1 \tau_2 ... \tau_k}\right) S_y^2, \tag{3.6}$$

where $\rho^2_{y.\tau_1 \tau_2 ... \tau_k}$ is defined earlier. Comparing (3.6) with (2.2) we can readily see that

$$MSE\left(t_{3(1)}\right) < MSE\left(t_{1(1)}\right).$$

Using $\bar{y} = \bar{Y} + \bar{e}_y$ in (3.3), squaring and applying expectation we have:

$$MSE\left(t_{4(1)}\right) \approx \theta\left[S_y^2 + \underline{\alpha}'\bar{Y}^2\Phi_\tau\underline{\alpha} + 2\bar{Y}\underline{\alpha}'\underline{\phi}_{y\tau}\right], \tag{3.7}$$

Partially differentiating (3.7) with respect to $\underline{\alpha}$ and equating the derivative to zero we have,

$$\underline{\alpha} = -\frac{1}{\bar{Y}}\Phi_\tau^{-1}\underline{\phi}_{y\tau}. \tag{3.8}$$

Using (3.8) in (3.7) we have,

$$MSE\left(t_{4(1)}\right) \approx \theta\left(1-\rho^2_{y.\tau_1\tau_2\ldots\tau_k}\right)S_y^2, \tag{3.9}$$

Some special cases of family formulated in (3.3) can be constructed easily as shown below.

(i) Generalized Ratio Estimator for Single-Phase Sampling for Full Information Case Using "K" Auxiliary Attributes

If we put $\alpha_1 = \alpha_2 = \alpha_3 \ldots = \alpha_k = -1$ in (3.3) we get generalized ratio estimator i.e.

$$t_{5(1)} = \bar{y}\left(\frac{P_1}{p_1}\right)\left(\frac{P_2}{p_2}\right).....\left(\frac{P_k}{p_k}\right), \tag{3.10}$$

and by putting $\underline{\alpha} = \left[-\underline{1}\right]_{k\times 1}$ in (3.7) we get mean square error of $t_{5(1)}$ i.e.

$$MSE(t_{5(1)}) \approx \theta\bar{Y}^2\left[C_y^2 + \sum_{j=1}^{k}C_{\tau_j}^2 - 2\sum_{j=1}^{k}C_yC_{\tau_j}\rho_{Pb_j} + 2\sum_{j\neq\psi=1}^{k}C_{\tau_j}C_{\tau_\psi}Q_{j\psi}\right] \tag{3.11}$$

(ii) Generalized Product Estimator for Single-Phase Sampling for Full Information Case Using "K" Auxiliary Attributes

If we put $\alpha_1 = \alpha_2 = \alpha_3 \ldots = \alpha_k = 1$ in (3.3) we get generalized product estimator i.e.

$$t_{6(1)} = \bar{y}\left(\frac{p_1}{P_1}\right)\left(\frac{p_2}{P_2}\right).....\left(\frac{p_k}{P_k}\right), \tag{3.12}$$

and by putting $\underline{\alpha} = \left[\underline{1}\right]_{k\times 1}$ in (3.7) we get mean square error of $t_{6(1)}$ i.e.

$$MSE(t_{6(1)}) \approx \theta\bar{Y}^2\left[C_y^2 + \sum_{j=1}^{k}C_{\tau_j}^2 + 2\sum_{j=1}^{k}C_yC_{\tau_j}\rho_{Pb_j} + 2\sum_{j\neq\psi=1}^{k}C_{\tau_j}C_{\tau_\psi}Q_{j\psi}\right]. \tag{3.13}$$

3.2 Generalized Estimator Using "k" Auxiliary Attributes for No Information Case

We propose a general family of estimators for two-phase sampling when information of auxiliary attributes is not known for the population. We propose the following general family of estimators:

$$T_{7(2)} = g_{\omega}\left(\bar{y}_2, v_{1d}, v_{2d}, ..., v_{kd}\right), \tag{3.14}$$

where $v_{jd} = p_{j(2)}/p_{j(1)}; v_{jd} > 0$. Under the conditions; stated for (3.1) the following families of estimators may be formulated as,

i) $$t_{7(2)} = \bar{y}_2 + \sum_{j=1}^{k} \alpha_j \left(v_{jd} - 1\right) = \bar{y}_2 + \underline{\alpha}' \left(\underline{v}_d - 1\right) = \bar{y}_2 + \underline{\alpha}' \underline{\varphi}_d, \tag{3.15}$$

ii) $$t_{8(2)} = \bar{y}_2 \left[(v_{1d})^{\alpha_1} (v_{2d})^{\alpha_2} ... (v_{kd})^{\alpha_k} \right] \tag{3.16}$$

where $\underline{\varphi}_d = \underline{v}_d - 1$. Now, using $\bar{y}_2 = \bar{Y} + \bar{e}_{y_2}$ in (3.15) we have

$$t_{7(2)} - \bar{Y} = \bar{e}_{y_2} + \underline{\alpha}' \underline{\varphi}_d .$$

Squaring and applying expectation we have:

$$MSE\left(t_{7(2)}\right) = \theta_2 S_y^2 + \left(\theta_2 - \theta_1\right) \underline{\alpha}' \Phi_{\tau} \underline{\alpha} - 2\left(\theta_2 - \theta_1\right) \underline{\alpha}' \underline{\phi}_{y\tau}; \tag{3.17}$$

The optimum value of $\underline{\alpha}$ is same as derived for full information case in (3.5). Using (3.5) in (3.17) we have,

$$MSE\left(t_{7(2)}\right) = \left\{ \theta_2 \left(1 - \rho^2_{y.\tau_1\tau_2...\tau_k}\right) + \theta_1 \rho^2_{y.\tau_1\tau_2...\tau_k} \right\} S_y^2, \tag{3.18}$$

where $\rho^2_{y.\tau_1\tau_2...\tau_k}$ is defined earlier. Using $\bar{y}_2 = \bar{Y} + \bar{e}_{y_2}$ in (3.16) and simplifying we have,

$$t_{8(2)} - \bar{Y} \approx \left(\bar{e}_{y_2} + \bar{Y} \underline{\alpha}' \underline{\varphi}_d\right).$$

Squaring and applying expectation we have:

$$MSE\left(t_{8(2)}\right) \approx \left[\theta_2 S_y^2 + \left(\theta_2 - \theta_1\right) \left\{ \bar{Y}^2 \underline{\alpha}' \Phi_{\tau} \underline{\alpha} + 2\bar{Y} \underline{\alpha}' \underline{\phi}_{y\tau} \right\} \right] \tag{3.19}$$

The optimum value of $\underline{\alpha}$ is same as derived for full information case in (3.8). Using (3.8) in (3.19) we have,

$$MSE\left(t_{8(2)}\right) \approx \left\{ \theta_2 \left(1 - \rho^2_{y.\tau_1\tau_2...\tau_k}\right) + \theta_1 \rho^2_{y.\tau_1\tau_2...\tau_k} \right\} S_y^2, \tag{3.20}$$

Some special cases of family formulated in (3.16) can be constructed easily as shown below.

(i) Generalized Ratio Estimator for Two-Phase Sampling for No Information Case Using "K" Auxiliary Attributes

If we put $\alpha_1 = \alpha_2 = \alpha_3 ... = \alpha_k = -1$ in (3.16) we get generalized ratio estimator i.e.

$$t_{9(2)} = \bar{y}_2 \left(\frac{p_{1(1)}}{p_{1(2)}} \right) \left(\frac{p_{2(1)}}{p_{2(2)}} \right) ..... \left(\frac{p_{k(1)}}{p_{k(2)}} \right), \tag{3.21}$$

and by putting $\underline{\alpha} = [-\underline{1}]_{k\times 1}$ in (3.19) we get mean square error of $t_{9(2)}$ i.e.

$$MSE(t_{9(2)}) \approx \bar{Y}^2 \left[\theta_2 C_y^2 + \theta_3 \left\{ \sum_{j=1}^{k} C_{\tau_j}^2 - 2\sum_{j=1}^{k} C_y C_{\tau_j} \rho_{Pb_j} + 2 \sum_{j\neq\psi=1}^{k} C_{\tau_j} C_{\tau_\psi} Q_{j\psi} \right\} \right] \tag{3.22}$$

(ii) Generalized Product Estimator for Two-Phase Sampling for No Information Case Using "K" Auxiliary Attributes

If we put $\alpha_1 = \alpha_2 = \alpha_3 ... = \alpha_k = 1$ in (3.16) we get generalized product estimator i.e.

$$t_{10(2)} = \bar{y}_2 \left(\frac{p_{1(2)}}{p_{1(1)}} \right) \left(\frac{p_{2(2)}}{p_{2(1)}} \right) ..... \left(\frac{p_{k(2)}}{p_{k(1)}} \right), \tag{3.23}$$

and by putting $\underline{\alpha} = [\underline{1}]_{k\times 1}$ in (3.19) we get mean square error of $t_{10(2)}$ i.e.

$$MSE(t_{10(2)}) \approx \bar{Y}^2 \left[\theta_2 C_y^2 + \theta_3 \left\{ \sum_{j=1}^{k} C_{\tau_j}^2 + 2\sum_{j=1}^{k} C_y C_{\tau_j} \rho_{Pb_j} + 2 \sum_{j\neq\psi=1}^{k} C_{\tau_j} C_{\tau_\psi} Q_{j\psi} \right\} \right] \tag{3.24}$$

3.3 Generalized Estimator for "k" Auxiliary Attributes (With "m" known and "m<k") for Partial Information Case

Suppose that population proportion P_j are known for j= (1, 2…m) auxiliary attributes and the population proportion P_j is unknown for j= (m+1 m+2…k) attributes. Using such partial information we propose following general family of estimators:

$$T_{11(2)} = g_\omega \left(\bar{y}_2, v_1, v_2 .... v_m, v_{(m+1)}, v_{(m+2)}, ..., v_k \right), \tag{3.25}$$

where $v_j = p_{j(1)}/P_j, (j = 1, 2...m)$; $v_j = p_{j(2)}/p_{j(1)}, (j = m+1, m+2...k), v_j \,\&\, v_{jd} > 0$. Under the conditions; stated for (3.1) the following families of estimators may be formulated as,

i) $$t_{11(2)} = \bar{y}_2 + \sum_{j=1}^{m} \alpha_j \left(v_j - 1 \right) + \sum_{j=m+1}^{k} \alpha_j \left(v_j - 1 \right) =$$

$$\bar{y}_2 + \underline{\alpha}_1' \left(\underline{v}_1 - 1 \right) + \underline{\alpha}_2' \left(\underline{v}_2 - 1 \right) = \bar{y}_2 + \underline{\alpha}_1' \underline{\varphi}_1 + \underline{\alpha}_2' \underline{\varphi}_2, \tag{3.26}$$

ii) $$t_{12(2)} = \bar{y}_2 \left[(v_1)^{\alpha_1} (v_2)^{\alpha_2} \ldots (v_m)^{\alpha_m} \left(v_{(m+1)} \right)^{\alpha_{m+1}} \left(v_{(m+2)} \right)^{\alpha_{m+2}} \ldots (v_k)^{\alpha_k} \right], \quad (3.27)$$

where $\underline{\alpha}_1 = \left[\alpha_j \right]_{m\times 1}$ and $\underline{\alpha}_2 = \left[\alpha_j \right]_{(k-m)\times 1}$ also $\underline{\varphi}_1 = \left[v_j - 1 \right]_{m\times 1}$ & $v_j = \frac{p_{j(1)}}{P_j}$ and $\underline{\varphi}_2 = \left[v_j - 1 \right]_{(k-m)\times 1}$ & $v_j = \frac{p_{j(2)}}{p_{j(1)}}$. Using $\bar{y}_2 = \bar{Y} + \bar{e}_{y_2}$ in (3.26) we have $t_{11(2)} - \bar{Y} = \bar{e}_{y_2} + \underline{\alpha}_1' \underline{\varphi}_1 + \underline{\alpha}_2' \underline{\varphi}_2$

Squaring and applying expectation we have:

$$MSE\left(t_{11(2)} \right) = \left[\theta_2 S_y^2 + \theta_1 \left\{ \underline{\alpha}_1' \Phi_{\tau_1} \underline{\alpha}_1 + 2\underline{\alpha}_1' \underline{\phi}_{y\tau_1} \right\} + (\theta_2 - \theta_1) \left\{ \underline{\alpha}_2' \Phi_{\tau_2} \underline{\alpha}_2 + 2\underline{\alpha}_2' \underline{\phi}_{y\tau_2} \right\} \right]; \quad (3.28)$$

Partially differentiating (3.28) w.r.t. $\underline{\alpha}_1$ and $\underline{\alpha}_2$; equating the derivatives to zero and solving we have the following optimum values of $\underline{\alpha}_1$ and $\underline{\alpha}_2$,

$$\underline{\alpha}_1 = -\Phi_{\tau_1}^{-1} \underline{\phi}_{y\tau_1} \quad (3.29)$$

$$\underline{\alpha}_2 = -\Phi_{\tau_2}^{-1} \underline{\phi}_{y\tau_2} \quad (3.30)$$

Using (3.29) and (3.30) in (3.28), the minimum mean square error of $t_{11(2)}$ will be

$$MSE\left(t_{11(2)} \right) = \left\{ \theta_2 \left(1 - \rho_{y.\tau_{m+1}\tau_{m+2}\ldots\tau_k}^2 \right) + \theta_1 \left(\rho_{y.\tau_{m+1}\tau_{m+2}\ldots\tau_k}^2 - \rho_{y.\tau_1\tau_2\ldots\tau_m}^2 \right) \right\} S_y^2, \quad (3.31)$$

where $\rho_{y.\tau_{m+1}\tau_{m+2}\ldots\tau_k}^2$ and $\rho_{y.\tau_1\tau_2\ldots\tau_m}^2$ is the squared multiple bi-serial correlation coefficient.

Using $\bar{y}_2 = \bar{Y} + \bar{e}_{y_2}$ in (3.27) and simplifying we have, $t_{12(2)} - \bar{Y} = \bar{e}_{y_2} + \bar{Y} \underline{\alpha}_1' \underline{\varphi}_1 + \bar{Y} \underline{\alpha}_2' \underline{\varphi}_2$

Squaring and applying expectation we have:

$$MSE\left(t_{12(2)} \right) = \left[\theta_2 S_y^2 + \theta_1 \left\{ \bar{Y}^2 \underline{\alpha}_1' \Phi_{\tau_1} \underline{\alpha}_1 + 2\bar{Y} \underline{\alpha}_1' \underline{\phi}_{y\tau_1} \right\} + (\theta_2 - \theta_1) \left\{ \bar{Y}^2 \underline{\alpha}_2' \Phi_{\tau_2} \underline{\alpha}_2 + 2\bar{Y} \underline{\alpha}_2' \underline{\phi}_{y\tau_2} \right\} \right] \quad (3.32)$$

Partially differentiating (3.32) w.r.t. $\underline{\alpha}_1$ and $\underline{\alpha}_2$; equating the derivatives to zero and solving we have the following optimum values of $\underline{\alpha}_1$ and $\underline{\alpha}_2$:

$$\underline{\alpha}_1 = -\frac{1}{\bar{Y}} \Phi_{\tau_1}^{-1} \underline{\phi}_{y\tau_1} \quad (3.33)$$

$$\underline{\alpha}_2 = -\frac{1}{\overline{Y}}\Phi_{\tau_2}^{-1}\underline{\phi}_{y\tau_2} \tag{3.34}$$

Using (3.33) and (3.34) in (3.32), the minimum mean square error of $t_{12(2)}$ will be

$$MSE\left(t_{12(2)}\right) = \left\{\theta_2\left(1-\rho^2_{y.\tau_{m+1}\tau_{m+2}\ldots\tau_k}\right)+\theta_1\left(\rho^2_{y.\tau_{m+1}\tau_{m+2}\ldots\tau_k}-\rho^2_{y.\tau_1\tau_2\ldots\tau_m}\right)\right\}S_y^2, \tag{3.35}$$

Some special cases of family formulated in (3.27) can be constructed easily as shown below.

(i) Generalized Ratio Estimator for Two-Phase Sampling for partial Information Case Using "K" Auxiliary Attributes (With "m" known and "m<k")

If we put $\alpha_1 = \alpha_2 \ldots = \alpha_{m+1} = \alpha_{m+2} \ldots = \alpha_k = -1$ in (3.27) we get generalized ratio estimator i.e.

$$t_{13(2)} = \overline{y}_2\left(\frac{P_1}{p_{1(1)}}\right)\left(\frac{P_2}{p_{2(1)}}\right).....\left(\frac{P_m}{p_{m(1)}}\right)\left(\frac{p_{(m+1)(1)}}{p_{(m+1)(2)}}\right).....\left(\frac{p_{k(1)}}{p_{k(2)}}\right), \tag{3.36}$$

and by putting $\underline{\alpha}_1 = [-\underline{1}]_{m\times 1}$ and $\underline{\alpha}_2 = [-\underline{1}]_{(k-m)\times 1}$ in (3.32) we get mean square error of $t_{13(2)}$ i.e.

$$MSE\left(t_{13(2)}\right) \approx \overline{Y}^2\left[\begin{array}{l}\theta_2\left\{C_y^2+\sum\limits_{j=m+1}^{k}C_{\tau_j}^2-2\sum\limits_{j=m+1}^{k}C_yC_{\tau_j}\rho_{Pb_j}+2\sum\limits_{j\neq\psi=m+1}^{k}C_{\tau_j}C_{\tau_\psi}Q_{j\psi}\right\}\\ +\theta_1\left\{\begin{array}{l}\left(\sum\limits_{j=1}^{m}C_{\tau_j}^2-\sum\limits_{j=m+1}^{k}C_{\tau_j}^2\right)-2\left(\sum\limits_{j=1}^{m}C_yC_{\tau_j}\rho_{Pb_j}-\sum\limits_{j=m+1}^{k}C_yC_{\tau_j}\rho_{Pb_j}\right)\\ +2\left(\sum\limits_{j\neq\psi=1}^{m}C_{\tau_j}C_{\tau_\psi}Q_{j\psi}-\sum\limits_{j\neq\psi=m+1}^{k}C_{\tau_j}C_{\tau_\psi}Q_{j\psi}\right)\end{array}\right\}\end{array}\right]. \tag{3.37}$$

(ii) Generalized Product Estimator for Two-Phase Sampling for partial Information Case Using "K" Auxiliary Attributes (With "m" known and "m<k")

If we put $\alpha_1 = \alpha_2 \ldots = \alpha_{m+1} = \alpha_{m+2} \ldots = \alpha_k = 1$ in (3.27), we get generalized product estimator i.e.

$$t_{14(2)} = \overline{y}_2\left(\frac{p_{1(1)}}{P_1}\right)\left(\frac{p_{2(1)}}{P_2}\right).....\left(\frac{p_{m(1)}}{P_m}\right)\left(\frac{p_{(m+1)(2)}}{p_{(m+1)(1)}}\right).....\left(\frac{p_{k(2)}}{p_{k(1)}}\right), \tag{3.38}$$

and by putting $\underline{\alpha}_1 = [\underline{1}]_{m\times 1}$ and $\underline{\alpha}_2 = [\underline{1}]_{(k-m)\times 1}$ in (3.32) we get mean square error of $t_{14(2)}$ i.e.

$$MSE\left(t_{14(2)}\right) \approx \overline{Y}^2\left[\begin{array}{l}\theta_2\left\{C_y^2+\sum\limits_{j=m+1}^{k}C_{\tau_j}^2+2\sum\limits_{j=m+1}^{k}C_yC_{\tau_j}\rho_{Pb_j}+2\sum\limits_{j\neq\psi=m+1}^{k}C_{\tau_j}C_{\tau_\psi}Q_{j\psi}\right\}\\ +\theta_1\left\{\begin{array}{l}\left(\sum\limits_{j=1}^{m}C_{\tau_j}^2-\sum\limits_{j=m+1}^{k}C_{\tau_j}^2\right)+2\left(\sum\limits_{j=1}^{m}C_yC_{\tau_j}\rho_{Pb_j}-\sum\limits_{j=m+1}^{k}C_yC_{\tau_j}\rho_{Pb_j}\right)\\ +2\left(\sum\limits_{j\neq\psi=1}^{m}C_{\tau_j}C_{\tau_\psi}Q_{j\psi}-\sum\limits_{j\neq\psi=m+1}^{k}C_{\tau_j}C_{\tau_\psi}Q_{j\psi}\right)\end{array}\right\}\end{array}\right]. \tag{3.39}$$

4. COMMENTS AND CONCLUSION

The information on *k* auxiliary attributes has been utilized to develop the generalized family of estimators for single and two phase sampling. There could be number of families of estimators for general families proposed in (3.1), (3.14) and (3.25), the special members have been given in (3.2), (3.3), (3.15), (3.16), (3.26) and (3.27). The expression for mean square error of the resulting estimators has been given in (3.5), (3.11) and (3.15). It can be easily seen that the mean square errors given in (3.6), (3.9), (3.18), (3.20), (3.31) and (3.35) are smaller as compared with the expression given by Jhajj et al. (2006). The optimum value of α_j involve some population parameters, which are assumed to be known for the efficient use of proposed families $T_{3(1)}, T_{7(2)}$ and $T_{11(2)}$. In case these parameters are unknown, these can be estimated from the sample. If we follow approach of Srivastava and Jhajj (1983), the estimator of proposed families $T_{3(1)}$, $T_{7(2)}$ and $T_{11(2)}$ will have the same minimum mean square, if we replace the unknown value of parameters involved in optimum value of α_j with their consistent estimators

REFERENCES

Jhajj, H.S., Sharma M.K. and Grover, L.K. (2006). A family of estimators of population mean using information on auxiliary attribute. *Pak. J. Statist.*, 22(1), 43-50.

Srivastava, S.K. and Jhajj, H.S. (1983). A class of estimators of the population means using multi-auxiliary information. *Calcutta Statistics Association Bulletin*, 32, 47-56.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 685–692.

IDENTIFYING ABERRANT VARIABLE FROM AN OUT-OF-CONTROL SIGNAL

Siti Rahayu Mohd. Hashim
Department of Probability and Statistics, Hicks Building, Hounsfield Road,
S3 7RY, University of Sheffield, UK.
E-mail: stp08sm@sheffield.ac.uk

ABSTRACT

Identification of one or more aberrant variables poses a persistent problem in interpreting the out of control signal in a multivariate control chart. The limitations of the available multivariate control charts plus the various level of interdependency between variables and mean shifts make the task even more difficult. This paper studies further the impact of these two factors on the diagnostic methods of multivariate processes. A simulation approach was taken to investigate the effects of various mean shifts and correlations structure. A few existing methods are reviewed and their performances are compared in terms of the percentage of detection and correct identification. These methods are compared with a novel proposal for identification and interpretation of aberrancy.

Keywords: Aberrant variable, multivariate control chart, diagnostic methods, out-of-control signal

1. INTRODUCTION

One of the most important tools in quality control is quality control chart. Since, many processes involve two or more quality characteristics or variables, multivariate control charts have become a popular tool in Statistical Process Control for monitoring purposes as well as for identifying the variables which are responsible for the out of control signals. In this study they are referred to as aberrant variables.

Nevertheless, as stated by Runger and Alt (1996), multivariate statistical process control and the use of multivariate control charts in particular, has one significant practical disadvantage that is the difficulty to determine which of the monitored variables is responsible for the out-of-control signal. Hayter and Tsui (1994) stated that the procedure for multivariate control problem must satisfy three conditions which are the ability to control the overall family wise error rate at the nominal level α, providing a simple mechanism in determining the variables responsible for the out-of-control signal and the ability to identify the magnitude of the mean change for the out-of-control variable. Jackson (1991) also stated that any multivariate process control procedure should fulfil four conditions. Firstly, the procedure must be able to tell whether the process is in control or not. Secondly, it must have a specification of the overall probability for the event 'Procedure diagnoses an out-of-control state errorneously'. Thirdly, it must take into account the relationship among the variables involved and finally, the procedure must be able to identify the cause of the problem. The second condition in Hayter and Tsui (1994) and the third and fourth condition proposed by Jackson (1991), will be the focus of our discussion in this paper.

Jackson (1980,1981 & 1985) and MacGregor and Kourti (1995) have discussed in detail the approach of using principal component analysis in multivariate quality control. Since the principal components (or latent variables) are uncorrelated, they are often interpreted as measurements of distinct characteristics of the process (Runger and Alt, 1996). Maravelakis et al. (2002) also used principal component analysis to calculate a ratio for each variable in every observation in order to determine the contribution of every single variable to every signal produced by the multivariate control chart. The drawback of this approach is that sometimes, though rarely, the principal components do not provide meaningful information on the variables. It will definitely affect the interpretation of the principal components themselves and any calculation based on the principal components would be meaningless. Furthermore, as stated by Runger, Alt and Montgomery (1996), if a latent variable is difficult to interpret, then it is difficult to translate an unusual value for a latent variable into corrective action.

Identifying the aberrant variables can also be done by performing individual t-tests on each variable (Alt, 1985) but a t-test for zero means could find neither variable unusual. Doganaksoy, Faltin and Tucker (1991) also used a similar approach where the ranking of the univariate t statistic will be obtained to determine which variable is most likely to have changed. Another popular approach is by decomposing the T^2 statistics discussed by Murphy (1987), Chua and Montgomery (1992), Mason, Tracy and Young (1995) and Tracy, Young and Mason (1992). Computations or decomposition of the T^2 statistics from some or all of the subsets of variables is used to assess the contributions in the signal. Hawkins (1991) also recommended the approach to detect a shift of the process mean in the direction of one of the measured variables.

All the methods discussed in the previous section are heavily dependent on the out of control signal produced by the Hotelling's (1947) multivariate control chart. Suppose that one has $p\times1$ random vectors X_1, X_2, …, X_p, and each of these vectors representing the p quality characteristics or variables either as individual observations or mean vectors. It is always assumed that X_i, i=1,2,…,p are independent and follow multivariate normal distribution and to be monitored and observed over time. For simplicity, one always assumes that each of the random vectors has the known covariance matrix, Σ and the in-control process mean vector $\mu=(0,0,...,0)'$. Hotelling's multivariate control chart is used to detect shifts over time from this in-control vector. An out of control signal is produced from a statistically significant shift in the mean vector as soon as;

$$\chi_i^2 = X_i^{'} \Sigma^{-1} X_i > \chi_{p,\alpha}^2 \qquad (1)$$

where $\chi_{p,\alpha}^2$ is the specified upper control limit (UCL) and α represents the level of significance of the hypothesis tests (Ryan, 2000).

2. INTERPRETATION METHODS OF OUT OF CONTROL SIGNAL

Diagnostic Method 1

Doganaksoy, Faltin and Tucker (1991) proposed a diagnostic method by ranking the univariate t-statistics. This approach will calculate the univariate t-statistics for each variable;

$$t = \frac{\left(\bar{x}_{i,new} - \bar{x}_{i,ref}\right)}{\left[s_{ii}\left(\frac{1}{n_{new}} + \frac{1}{n_{ref}}\right)\right]^{1/2}} \tag{2}$$

with s_{ii} is the variance of a variable and n_{new} and n_{ref} is the sample size of the tested variable and the size of a sample from where the variance was obtained. K_{ind} will be calculated by the following formula

$$K_{ind} = \left|2T\left(t; n_1 - 1\right) - 1\right| \tag{3}$$

where $T(t;n_1\text{-}1)$ obtained from the cumulative distribution function of the t distribution with n_1-1 degrees of freedom. The aberrant variable is identified by looking at the K_{ind} value. The variable with the highest K_{ind} value is the most likely the one that has changed in the process.

Diagnostic Method 2

Maravelakis et al. (2002), has proposed two methods in identifying the aberrant variable(s) that caused the out-of control signal in multivariate control chart. Both methods need to perform the principal components analysis in prior of the computation of the ratio(s) for every variable for each observation in the data. The first method is used when the values of correlations in variance covariance matrix are all positive. The formula used to calculate the ratio for a given variable is:

$$r_{ki} = \frac{\left(u_{k1} + u_{k2} + ... + u_{kd}\right) x_{ki}}{Y_{1i} + Y_{2i} + ... + Y_{di}} \tag{4}$$

where *u*'s are the values in the first principal component, x_{ki} is the *i*-th observation's value of variable X_k, Y_{ji} is the score of the *i*-th vector of observations in the *j*-th principal component or

$$Y_{ji} = u_{ik}x_{1i} + u_{2k}x_{2i} + ... + u_{dk}x_{di} \tag{5}$$

with j=1,2,…,*d* and *d* is the number of significant principal components from the analysis. Maravelakis et al. (2002) relied on the Average Root method (Jackson, 1990) in determining the number of principal component where only the first principal component is considered in the ratio calculation based on the fact that the first principal component contains the most information of the data. So, the ratio for variable-*k* in observation –*i* will be

$$r_{ki} = \frac{\left(\mu_{k1} x_{ki}\right)}{Y_{1i}} \tag{6}$$

where μ_{ki} is the first principal component value for the first variable, x_{ki} is the *i*-th observation for variable-*k* and Y_{1i} is the score of the *i*-th vector of observations in the *j*-th principal component.

Maravelakis et al. (2002) has proposed a different way of calculating the ratios for mixed values in covariance matrix. The denominator of (4) is calculated by the in-control mean vector

and not by the observation vector, X_i. Maravelakis et al. (2002) has proposed to use the percentile of (α) and *(1-*α*)* from the normal probability distribution as the control limits of the ratios with α is the significant level of the test. Whilst, the bivariate integral normal distribution proposed by Hinkley(1969) is used to determine the control limits for the ratios with mixed sign values of covariance matrix.

In this study, the proposed approach for the mixed values in covariance matrix is not followed and the ratios will not be plotted within their control limits in order to identify the aberrant variables. Since the referenced mean vector is 0's, the ratio will obviously be indefinite. For this reason, this study used equation (6) to calculate the ratios regardless the type of the variance and covariance matrix. As the ratio clearly representing the contribution of variable-*k* in observation-*i*, this study tried to adopt the approach of Doganaksoy, Faltin and Tucker (1996) by ranking the contribution of each variable for every observation. The ratios are treated as the weight of a variable in an observation. The higher the weight of the variable, the most likely it was the aberrant one. Instead of studying whether the ratios are within the control limits, we are now ranking the ratio of each variable based on its contribution in every observation.

Diagnostic method 3

This is an extension of the method proposed by Doganaksoy, Faltin and Tucker (1991) which is to be used together with the K_{ind} discussed in the first diagnostic method. Another value, K_{Bonf} , is calculated for every variable,

$$K_{Bonf\text{-}k} = \frac{(p + K_{sim} - 1)}{p} \tag{7}$$

where *p* is the number of variables and K_{sim} as stated in Doganaksoy, Faltin and Tucker (1991) "represents a trade off between the power of the intervals to identify attributes which have truly changed, versus the likelihood of misidentifying an attribute as having changed when it in fact did not". The value of K_{sim} is fixed prior to the analysis and lies between 0 and 1. Doganaksoy,Faltin and Tucker(1991) have discussed in detail how to select the best K_{sim}. The aberrant variables can be identified by comparing the values of the K_{ind} and K_{Bonf}. The variable with it's $K_{ind} > K_{Bonf}$ is classified as being the one most likely to have changed.

Correct Identification

The correct identification of the diagnostic method has been done before by Das and Prakash (2007). In this study, the correct identification refers to the number of time the deviated variable(s) detected as the most likely to have changed from all of the observations simulated from the deviated mean vectors.

3. RESULTS

One thousand observations have been simulated from a multivariate normal distribution for a particular mean vector and variance and covariance matrix. The covariance matrices from Doganaksoy, Faltin and Tucker (1996) were used in the simulation as well as some of the mean vectors in their illustrations. This study also extends the selection of the mean vectors to take into account the impact of the different signs and different allocation of the deviated value(s) from the origin, 0, in the mean vector.

$$cv1 = \begin{bmatrix} 1.00 & 0.80 & 0.55 & 0.65 \\ 0.80 & 1.00 & 0.65 & 0.50 \\ 0.55 & 0.65 & 1.00 & 0.60 \\ 0.60 & 0.50 & 0.60 & 1.00 \end{bmatrix}; \qquad cv4 = \begin{bmatrix} 1.00 & 0.20 & -0.50 & 0.30 \\ 0.20 & 1.00 & 0.20 & -0.50 \\ -0.50 & 0.20 & 1.00 & 0.20 \\ 0.30 & -0.50 & 0.20 & 1.00 \end{bmatrix}$$

Table 1: The percentage of out-of-control signals and correct identification with 1 variable deviates from origin with all positive signs of covariance matrix.

Mean vector	% of Out-of-Control Signals	% of Correct Identification		
		Method 1	Method 2	Method3
[2,0,0,0]	83.3	77.8	74.0	51.7
[0,2,0,0]	86.1	79.5	73.6	50.7
[0,0,2,0]	60.7	78.8	70.5	48.5
[0,0,0,2]	58.0	80.1	66.8	53.5

Table 2: The percentage of out-of-control signals and correct identification with 1 variable deviates from origin with mixed signs of covariance matrix.

Mean vector	% of Out-of-Control Signals	% of Correct Identification		
		Method 1	Method 2	Method 3
[2,0,0,0]	98.8	75.2	67.0	48.8
[0,2,0,0]	98.5	74.7	68.7	46.9
[0,0,2,0]	98.0	77.8	49.9	52.5
[0,0,0,2]	99.3	76.6	81.5	51.6

The multivariate Hotelling's control chart has detected higher percentages of out-of-control signals from the datasets with mixed signs covariance matrix. Table 1 shows that method 1 has higher percentages of correct identification compared to method 2. Table 2 also shows that method 1 generally has higher percentages of correct identification. Method 3 mostly gives the lowest percentages for all cases.

Table 3: The percentage of out-of-control signals and correct identification with 2 variables deviate in the same direction from origin with all positive signs of covariance matrix.

Mean vector	% of Out-of-Control Signals	% of Correct Identification		
		Method 1	Method 2	Method 3
[2,2,0,0]	62.9	93.0	89.1	51.2
[2,0,2,0]	98.8	88.2	89.2	51.2
[2,0,0,2]	84.0	87.5	87.4	52.6
[0,2,2,0]	82.8	89.1	89.1	49.6
[0,2,0,2]	99.3	88.6	87.4	52.1
[0,0,2,2]	67.5	89.0	86.4	51.0

Table 4: The percentage of out-of-control signals and correct identification with 2 variables deviate from origin with mixed signs of covariance matrix.

Mean vector	% of Out-of-Control Signals	% of Correct Identification		
		Method 1	Method 2	Method 3
[2,2,0,0]	51.7	96.4	91.7	47.9
[2,0,2,0]	100.0	95.4	100.0	47.9
[2,0,0,2]	49.8	92.8	95.7	52.0
[0,2,2,0]	53.5	93.8	86.8	49.7
[0,2,0,2]	100.0	94.9	95.6	49.2
[0,0,2,2]	54.2	94.7	91.5	52.0

Table 5: The percentage of out-of-control signals and correct identification with 2 variables deviate from origin in a different direction with all positive signs of covariance matrix.

Mean vector	% of Out-of-Control Signals	% of Correct Identification		
		Method 1	Method 2	Method 3
[2,-2,0,0]	100.0	98.9	99.0	53.0
[-2,2,0,0]	100.0	99.2	99.0	49.2
[2,0,-2,0]	95.3	98.5	99.0	49.9
[-2,0,2,0]	95.9	98.9	98.9	48.1
[2,0,0,-2]	99.5	98.7	97.1	51.7
[-2,0,0,2]	99.6	98.3	79.7	50.6
[0,2,-2,0]	99.8	98.7	98.6	49.4
[0,-2,2,0]	99.6	98.6	98.3	48.7
[0,2,0,-2]	95.8	99.1	97.9	51.1
[0,-2,0,2]	94.7	98.7	97.2	51.2
[0,0,2,-2]	97.9	98.7	95.5	50.0
[0,0,-2,2]	98.3	98.3	95.8	50.8

Table 6: The percentage of out-of-control signals and correct identification with 2 variables deviate from origin in a different direction with mixed signs of covariance matrix.

Mean vector	% of Out-of-Control Signals	% of Correct Identification		
		Method 1	Method 2	Method 3
[2,-2,0,0]	100.0	92.8	94.4	49.9
[-2,2,0,0]	100.0	92.5	94.3	47.6
[2,0,-2,0]	43.8	90.5	82.4	51.5
[-2,0,2,0]	40.2	91.5	81.9	54.0
[2,0,0,-2]	100.0	93.5	94.1	49.0
[-2,0,0,2]	100.0	93.3	82.1	50.0
[0,2,-2,0]	100.0	92.2	79.7	50.6
[0,-2,2,0]	100.0	94.0	80.9	51.7
[0,2,0,-2]	40.1	89.4	95.5	48.0
[0,-2,0,2]	43.3	91.2	96.0	51.2
[0,0,2,-2]	100.0	93.0	93.9	50.8
[0,0,-2,2]	100.0	92.7	94.8	52.9

The percentages of out-of-control signals are clearly higher in Table 3 except for cases 2 and 5 which shows 100% signals in Table 4, even so, Table 3 also shows a very high percentage of the out-of-control signals which are nearly 100% correct identification for those particular cases. In Table 3, method 1 and 2 give similar results whereas in Table 4, method 2 has lower percentages of correct identification when the deviated variables are variable 2 and variable 3.

In Table 5, most of the cases give a very high percentage of out-of-control signals. Diagnostic methods 1 and 2 show a similar ability in identifying the aberrant variables and method 3 also performed better even though still quite low compared to the first two diagnostic methods. In Table 6, the multivariate control chart detected all the out-of-control signals except for cases 3, 4, 9 and 10. Method 1 performed better in most of the cases and method 3 still having the lowest percentage of correct identification among the three methods.

4. CONCLUSION

In this study, it is found that multivariate control charts have more power to detect the out-of-control signal when the variance covariance matrix has mixed signs values and only one variable deviated from origin. For the cases with two variables deviated from the origin in the same direction, higher detection of out of control signals obtained when the deviated variables has a moderate or low correlation in both types of variance covariance matrices. Whereas, for the cases with two variables deviated from origin in different direction, multivariate control charts have very high percentages of signal's detection for both cases except for two cases with mixed signs variance covariance matrix. Results showed that the signal's detection are less than 50% when the deviated variables are correlated moderately in negative direction.

The results between diagnostic methods showed that method 1 generally has higher percentages of correct identification compared to the other diagnostic methods when only one variable deviated from origin regardless of the type of the variance covariance matrices except for one case in mixed signs variance matrix. Method 2 has showed a higher percentage of correct identification when the deviated variable has a moderate correlation with other variables. For the cases with two variables deviated from origin in the same direction, method 1 and 2 generally has a similar performance except for one cases with mixed signs values in variance covariance matrix. Method 1 has higher percentage of correct identification when the deviated variables have a low correlation value. For the cases with two variables deviated from origin in different directions, method 1 and 2 have shared a very good performance in the case with all positive signs in variance covariance matrix. Whereas in the case with mixed signs values, method 1 has a better performance when the correlation between the variables is moderate in negative direction. These results support the finding by Das and Prakash(2008) which stated that the method introduced by Doganaksoy, Faltin and Tucker(1996) has a better performance in identifying the correct aberrant variable when the correlation between variables are low or moderate.

REFERENCES

Alt, F.B. (1985). Multivariate quality control. *Encyclopedia of Statistical Sciences,* Kotz and Johnson (Eds), Vol. 6, John Wiley, N.Y.

Das, N. and Prakash, V. (2008). Interpreting the Out-of-Control Signal in multivariate Control Chart- A Comparative Study. *International Journal of Advanced Manufacturing Technology*, **37**:966-979.

Doganaksoy, N, Faltin, F.W, and Tucker, W.T (1991). Identification of Out-of-Control Quality Characteristics in a Multivariate Manufacturing Environment. *Communication in Statistics (Theory and Methods)*, **20**(9):2775-2790.

Hawkins, D.M (1991). Multivariate Quality Control Based on Regression-Adjusted Variables. *Technometrics*, **33**(1):61-75.

Hayter, A. and Tsui, K. (1994). Identification and quantification in multivariate quality control problems. *Journal of Quality and Technology*, **26**(3):197-207.

Hinkley, D.V. (1969). Inference about the Intersection in Two-Phase Regression. *Biometrika*, **56**(3):495-504.

Hotelling, H. (1947). Multivariate Quality Control. *Techniques of Statistical Analysis*. Eisenhart C., Hastay M.W. and Wallis W.A, McGraw Hill, New York, 111-184.

Jackson, J.E (1980). Principal Components and Factor Analysis: Part I- Principal Components. *Journal of Quality Technology*, **12**(4):201-213.

Jackson, J.E (1981). Principal Components and Factor Analysis: Part II-Additional Topics Related to Principal Components. *Journal of Quality Technology*, **13**(1):46-58.

Jackson, J.E (1985). Multivariate Quality Control. *Communication in Statistics – Theory and Methods*, **14**(11):2657-2688.

Jackson J.E (1991). A User Guide to Principal Components. John Wiley, New York.

MacGregor, J.F. and Kourti, T. (1995). Statistical Process Control of Multivariate Processes. *Control Engineering Practice*, **3**(3):403-414.

Maravelakis, P.E, Bersimis, S., Panaretos, J. and Psarakis, S. (2002). Identifying the Out-of-Control Variable in a Multivariate Control Chart. *Communication in Statistics (Theory and Methods)*, **31**(12):2391-2408.

Mason, R.L, Tracy, N.D, and Young, J.C (1995). Decomposition of T^2 for Multivariate Control Chart Interpretation. *Journal of Quality Technology*, **27**(2):99-108.

Mason, R.L, Tracy, N.D, and Young, J.C (1995). A Practical Approach for Interpreting Multivariate T^2 Control Chart Signals. *Journal of Quality Technology*, **97**(4):396-406.

Ryan, T.P. (2000) Statistical Methods for Quality Improvement, John Wiley, 2nd Edition, New York.

Tracy, N.D., Young, J.C., and Mason, R.L. (1992). Multivariate Control Charts for Individual Observations. *Journal of Quality Technology*, **24**:88-95.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 693–700.

A NEW GENERALIZATION OF POLYA-EGGENBERGER DISTRIBUTION AND ITS APPLICATIONS

Anwar Hassan
PG Department of Statistics
University of Kashmir, Srinagar-India
E-mail: anwar.hassan5@gmail.com, anwar.hassan2007@gmail.com

Sheikh Bilal Ahmad
Department of Statistics
Amar Singh College, Srinagar, Kashmir-India
E-mail: sbilal_sbilal@yahoo.com

ABSTRACT

In this paper, a new generalization of the Polya-Eggenberger distribution has been introduced by compounding the Binomial distribution with the generalized Beta distribution of Π-kind defined by Nadarajah and Kotz (2003). Some special cases, moment generating function and factorial moments of the distribution have been derived in terms of generalized hypergeometric function. Stirling numbers of second kind has been used to obtain the moments about zero. Finally, a computer programme in R-Software has been used to ease the computations for estimating the parameters of the distribution for data fitting and it has been shown that the distribution gives a remarkably best fit as compared to other generalizations present in the literature.

Keywords: Binomial distribution, generalized Beta distribution of Π-kind, generalized Polya-Eggenberger distribution, moment generating function, Chi-squre fitting.

1. INTRODUCTION

Urn models have been used by many authors to describe several classical contagious distributions. The first work seems to have been done by Eggenberger and Polya (1923). They considered one Urn model and obtained Polya-Eggenberger and inverse Polya-Eggenberger distributions. Janardan and Schaeffer (1977) have called these distributions as Markov-Polya distributions. The sampling scheme used in deriving Polya's distribution is known as Polya–Eggenberger sampling schemes where we draw the balls with replacement, note the colour of the ball drawn and add c additional balls of the same colour before the next draw is performed.

Many generalization of the Polya-Eggenberger distribution are present in the literature, almost all are derived through urn models. Janardan (1973) obtained quasi-Polya distribution by introducing an urn model dependent on predetermined strategies. Again in 1975, Janardan used two urn models with predetermined strategy but with Eggenberger and Polya sampling scheme and obtained qusi-Polya distribution and inverse quasi-Polya distribution. Sen and Mishra (1996) unified both the sampling schemes (direct and indirect) by introducing a new parameter to obtain a generalized Polya-Eggenberger model with Polya-Eggenberger and inverse Polya-Eggenberger distributions as its particular cases. Sen and Ritu (1996) introduced three generalized Markov-

Polya urn models with predetermined strategies by the unified sampling scheme. Most recently, Hassan and Bilal (2006) introduced a generalized Negative Polya-Eggenberger distribution through a mixture model which has Polya-Eggenberger distribution as its particular case.

In this paper, a new generalization of the Polya- Eggenberger distribution has been proposed involving hypergeometric function which is more versatile as compared to other generalizations present in the literature. This fact has been illustrated with the help of three data sets presented in section 4 of this paper. In section 2, we derived the proposed model and section 3 deals with some interesting structural properties of the proposed model. Finally, in section 4, a computer programme in R-Software has been used to ease the computations for estimating the parameters of the proposed model for data fitting and it has been shown that the proposed model gives a remarkably best fit as compared to other generalizations present in the literature

2. THE PROPOSED MODEL

When a sample of fixed size n is taken from an infinite population where each element in the population has an equal and independent probability p of possessing a specified attribute or the sample is taken from a finite population where each element in the population has an equal and independent probability p of having a specified attribute and elements are sampled independently and sequentially with replacement then these situations can be represented by a random variable X (possessing the attribute) that follows Binomial distribution with parameters (n, p) and pmf given by

$$P(x;p) = \binom{n}{x} p^x (1-p)^{n-x} \ , \quad 0 < p < 1, \quad x = 0,1,2,.....n \tag{2.1}$$

In real life situations, the assumption that the probability p of each element (possessing the attribute) is constant from trial to trial does not seem to be realistic. This assumption holds good in case of chance mechanism only. In fact, every living being use their past experience (success or failure) and wisdom for determining their future strategies to achieve their goals and so the probability p of a success does not remain constant and thus may be considered as a random variable taking values between 0 and 1 , $0 < p < 1$. The natural distribution of p is Beta distribution.

Beta distributions are very versatile and a variety of uncertainties can be usefully modeled by them. Many generalizations of Beta distribution involving algebraic and exponential functions have been proposed in the literature; see chapter 25 in Johnson *et. al.* (1995) and Gupta and Nadarajah (2004) for detailed accounts. In (2003), Nadarajah, S. and Kotz, S. introduced a new generalization of Beta distribution of II-kind in terms of hypergeometric function with parameters (a,b,γ) and p.d.f. given by

$$P(p) = \frac{b\beta(a,b)}{\beta(a,b+\gamma)} \, p^{a+b-1} \, {}_2F_1[1-\gamma, a; a+b; p] \tag{2.2}$$

for $0 < p < 1$, $a > 0, b > 0$ and $\gamma > 0$ and $\beta(a,b)$ is a Beta function where as ${}_2F_1[1-\gamma, a; a+b; p]$ is a hypergeometric function which is defined by

$${}_2F_1[a,b;c;x] = \sum_{j=0}^{\infty} \frac{a^{[j]} b^{[j]}}{c^{[j]}} \frac{x^j}{j!}$$

where $a^{[j]}$ stands for ascending factorials of ' a ' given by $a^{[j]} = a(a+1)....(a+j-1)$. Assuming the distribution of a random variable p as (2.2), the distribution of a random variable X is obtained by compounding the Binomial model (2.1) through the values of p with the generalized Beta distribution of II-kind (2.2), we obtain

$$P(X=x) = \binom{n}{x} \frac{b\beta(a,b)}{\beta(a,b+\gamma)} \int_0^1 p^{x+a+b-1}(1-p)^{n-x}\, {}_2F_1[1-\gamma, a; a+b; p]dp$$

This on simplification gives

$$P(X=x) = \binom{n}{x} \frac{b\beta(a,b)}{\beta(a,b+\gamma)} \frac{(n-x)!(x+a+b-1)!}{(a+b+n)!} \sum_{j=0}^{\infty} \frac{(1-\gamma)^{[j]} a^{[j]} (x+a+b)^{[j]}}{(a+b)^{[j]}(a+b+n+1)^{[j]}} \frac{1^j}{j!}.$$

The last sum is a generalized hypergeometric function, the equation above can be written as

$$P(X=x) = \frac{n!}{x!} \frac{b\beta(a,b)}{\beta(a,b+\gamma)} \frac{(a+b)^{[x]}}{(a+b)^{[n+1]}}\, {}_3F_2[1-\gamma, a, a+b+x; a+b, a+b+n+1; 1] \tag{2.3}$$

This can be put into an alternative form as

$$P(X=x) = \frac{n!}{x!} \frac{(a+b)^{[\gamma]}(a+b)^{[x]}}{(b+1)^{[\gamma-1]}(a+b)^{[n+1]}}\, {}_3F_2[1-\gamma, a, a+b+x; a+b, a+b+n+1; 1] \tag{2.4}$$

for $x = 0,1,....,n$, $(n,a,b,\gamma) > 0$ and ${}_3F_2[1-\gamma, a, a+b+x; a+b, a+b+n+1; 1]$ is a generalized hypergeometric function which is absolutely convergent if $\mathrm{Re}(b+n+\gamma-x) > 0$, see chapter 5, page 74, Special functions by Earl D. Rainville (1971) for details. The expressions (2.3) and (2.4) represent a new generalization of the Polya-Eggenberger distribution with parameters (n,a,b,γ).

Remarks: For $a+b+\gamma = 1$ or $\gamma = 1$ or $a = 0$ or $b = 0$, the proposed distribution (2.4) reduces to Polya-Eggenberger distribution with parameters in different forms.

3. STRUCTURAL PROPERTIES

In this section, some of the interesting properties of the proposed model have been explored in order to understand its nature to some extent. These are described below.

3.1 Mean and Variance

3.1.1 Mean

By the conditional mean, we have

$$Mean = E(X) = E\big[E(X/p)\big] \tag{3.1}$$

where $E(X/p)$ is a conditional expectation of X given p and for given p the random variable X has a binomial distribution (2.1) with mean and variance given by

$$\left.\begin{aligned} E(X/p) &= np \\ V(X/p) &= np(1-p) \end{aligned}\right\} \tag{3.2}$$

Using (3.2) in (3.1), we get $E(X) = nE[p]$ .Since p is varying as the generalized beta distribution of II-kind (2.2) with mth moment about origin given by

$$E(p^m) = \frac{b\beta(a,b)}{(a+b+m)\beta(a,b+\gamma)}\ {}_3F_2[1-\gamma, a, a+b+m; a+b, a+b+m+1, 1] \tag{3.3}$$

Taking $m=1$ in the equation above, we get the mean of the proposed model as

$$E(X) = \frac{nb\beta(a,b)}{(a+b+1)\beta(a,b+\gamma)}\ {}_3F_2[1-\gamma, a, a+b+1; a+b, a+b+2, 1]\ .$$

3.1.2 Variance

Similarly, variance of the proposed model can be obtained by the conditional variance

$$V(X) = E\left[V(X/p)\right] + V\left[E(X/p)\right] \tag{3.4}$$

The equation (3.4) together with (3.2) gives

$$V(X) = E[np(1-p)] + V[np] = nE[p] + n(n-1)E[p^2] - n^2\{E(p)\}^2 .$$

Using (3.3) for $m=1,2$ in the equation above, we get

$$\begin{aligned} V(X) = {} & \frac{nb\beta(a,b)}{(a+b+1)\beta(a,b+\gamma)}\ {}_3F_2[1-\gamma, a, a+b+1; a+b, a+b+1+1, 1] \\ & + \frac{n(n-1)b\beta(a,b)}{(a+b+2)\beta(a,b+\gamma)}\ {}_3F_2[1-\gamma, a, a+b+2; a+b, a+b+3, 1] \\ & - \left\{\frac{nb\beta(a,b)}{(a+b+1)\beta(a,b+\gamma)}\ {}_3F_2[1-\gamma, a, a+b+1; a+b, a+b+1+1, 1]\right\}^2 . \end{aligned}$$

3.2 Moment Generating Function

The derivation of the moment generating function (mgf) of the proposed model is not straightforward as it involves generalized hypergeometric function

$$ {}_3F_2[1-\gamma, a, a+b+x; a+b, a+b+n+1, 1] $$

which is an infinite series not easy to be summed. Since the proposed model is obtained by compounding binomial model (2.1) with mgf $M_{X/P}(t) = [1+p(e^t-1)]^n$ through the values of p with the generalized beta distribution of II-kind (2.2), therefore, a theorem by Feller (1943) yields the mgf of the proposed model as

$$ M_X(t) = \frac{b\beta(a,b)}{\beta(a,\gamma+b)} \int_0^1 (1+p(e^t-1))^n \; p^{a+b-1} {}_2F_1[1-\gamma, a; a+b, p]dp $$

$$ = \frac{b\beta(a,b)}{\beta(a,\gamma+b)} \sum_{k=0}^{n} \binom{n}{k} (e^t-1)^k \int_0^1 p^{a+b+k-1} {}_2F_1[1-\gamma, a; a+b, p]dp $$

By an application of beta integrals, the equation above yields the moment generating function of the proposed model as

$$ M_X(t) = \frac{b\beta(a,b)}{\beta(a,\gamma+b)} \sum_{k=0}^{n} \binom{n}{k} \frac{(e^t-1)^k}{(a+b+k)} \; {}_3F_2[1-\gamma, a, a+b+k; a+b, a+b+k+1, 1] $$

3.3 Moments for the Proposed Model

In this section, we obtained the rth moment about origin of the proposed model in terms of Stirling numbers of second kind and generalized hypergeometric function. By conditional mean, we have

$$ \mu_r' = E(X^r) = E[E(X^r/p)] \tag{3.5} $$

where $E(X^r/p)$ is conditional rth moment about origin of X given p and for given p the random variable X has binomial distribution (2.1) with rth moment about origin given by

$$ E(X^r/p) = \sum_{j=0}^{r} \frac{s(r,j)\,n!\,p^r}{(n-j)!} \tag{3.6} $$

where $s(r,j) = \dfrac{\Delta^j 0^r}{j!} = \dfrac{\Delta^j x^r}{j!}$ at $x=0$ is the Stirling numbers of second kind.

Substituting (3.6) in (3.5), we get $\mu_r' = \sum_{j=0}^{r} \dfrac{s(r,j)\,n!}{(n-j)!} \; E(p^r)$

Using (3.3) in the equation above, we get

$$\mu'_r = \sum_{j=0}^{r} \frac{s(r,j)\,n!}{(n-j)!} \frac{b\beta(a,b)}{(a+b+r)\beta(a,b+\gamma)} \; {}_3F_2[1-\gamma, a, a+b+r; a+b, a+b+r+1, 1] \tag{3.7}$$

Taking $r = 1,2,3,4$ in (3.7), we get the first four moments about origin of the proposed model as

$$\mu'_1 = \frac{nb\beta(a,b)}{(a+b+1)\beta(a,b+\gamma)} \; {}_3F_2[1-\gamma, a, a+b+1; a+b, a+b+2, 1]$$

$$\mu'_2 = \frac{nb\beta(a,b)}{(a+b+1)\beta(a,b+\gamma)} \; {}_3F_2[1-\gamma, a, a+b+1; a+b, a+b+2, 1]$$
$$+ \frac{n(n-1)b\beta(a,b)}{(a+b+2)\beta(a,b+\gamma)} \; {}_3F_2[1-\gamma, a, a+b+2; a+b, a+b+3, 1]$$

$$\mu'_3 = \frac{nb\beta(a,b)}{(a+b+1)\beta(a,b+\gamma)} \; {}_3F_2[1-\gamma, a, a+b+1; a+b, a+b+1+1, 1]$$
$$+ \frac{n(n-1)(n-2)b\beta(a,b)}{(a+b+3)\beta(a,b+\gamma)} \; {}_3F_2[1-\gamma, a, a+b+3; a+b, a+b+4, 1]$$
$$+ \frac{3n(n-1)b\beta(a,b)}{(a+b+2)\beta(a,b+\gamma)} \; {}_3F_2[1-\gamma, a, a+b+2; a+b, a+b+3, 1]$$

$$\mu'_4 = \frac{nb\beta(a,b)}{(a+b+1)\beta(a,b+\gamma)} \; {}_3F_2[1-\gamma, a, a+b+1; a+b, a+b+1+1, 1]$$
$$+ \frac{n(n-1)(n-2)(n-3)b\beta(a,b)}{(a+b+4)\beta(a,b+\gamma)} \; {}_3F_2[1-\gamma, a, a+b+4; a+b, a+b+5, 1]$$
$$+ \frac{6n(n-1)(n-2)b\beta(a,b)}{(a+b+3)\beta(a,b+\gamma)} \; {}_3F_2[1-\gamma, a, a+b+3; a+b, a+b+4, 1]$$
$$+ \frac{7n(n-1)b\beta(a,b)}{(a+b+2)\beta(a,b+\gamma)} \; {}_3F_2[1-\gamma, a, a+b+2; a+b, a+b+3, 1]$$

3.4 Factorial Moments

Following the similar arguments of the previous section, the rth factorial moment of the proposed model can be obtained as

$$\mu'_{(r)} = E(X^{(r)}) = E[E(X^{(r)}/p)] \tag{3.8}$$

where $E(X^{(r)}/p) = n^{(r)}\, p^r$ and $n^{(r)} = n(n-1).....(n-r+1)$. Substituting this value in (3.8), we get $\mu'_{(r)} = n^{(r)}\, E(p^r)$ A use of (3.3) in the equation above gives the rth factorial moment of the proposed model as

$$\mu'_{(r)} = \frac{n^{(r)} b\beta(a,b)}{(a+b+r)\beta(a,b+\gamma)} \; {}_3F_2[1-\gamma, a, a+b+r; a+b, a+b+r+1, 1]$$

TABLE 4.1 Results of 10 shots fired from a rifle at each of 100 targets; Sveshnikcv, A. A. Ed. (1968), PP 312-313.

No. of Accidents	Obs. Freq.	Expected frequencies			
		PED	GPED (SM 1996)	GNPED (HB 2006)	Proposed Model
0	0	0.15	0.14	0.14	0.14
1	2	1.29	1.22	1.22	1.22
2	4	5.04	4.87	4.86	4.86
3	10	12.16	11.91	11.91	11.92
4	22	20.03	19.87	19.89	19.91
5	26	23.49	23.58	23.61	23.63
6	18	19.85	20.12	20.13	20.13
7	12	11.89	12.13	12.11	12.09
8	4	4.81	4.87	4.86	4.85
9	2	1.16	1.13	1.14	1.14
10	0	0.13	0.16	0.13	0.11
Total	100	100	100		100
ML Estimate		$n = 10$ $a = 43.045327$ $b = 41.890767$	$n = 10$ $\mu = -0.93094$ $a = 50.30695$ $b = 42.20174$	$n = 9.244629$ $\beta = 0.037662$ $\alpha = 47.262270$ $\gamma = 41.913235$	$n = 10$ $a = 43.768887$ $b = 0.000211$ $\gamma = 41.319327$
χ^2 (d.f)		1.056273 (4)	1.020418 (3)	1.008995 (2)	1.001827 (3)

TABLE 4.2 (Source: Snedecor, G.W. and Cochran, W.G. (1967): Statistical methods, sixth edition, The Iowa University Press, Iowa, U.S.A., PP.237)

No. of Accidents	Obs. Freq.	Expected frequencies			
		PED	GPED (SM 1996)	GNPED (HB 2006)	Proposed Model
0	36	35.05	35.59	35.60	35.88
1	48	48.66	48.99	48.98	48.56
2	38	38.19	38.07	38.07	37.92
3	23	21.97	21.69	21.70	21.86
4	10	10.21	9.97	9.98	10.13
5	3	4.01	3.87	3.87	3.92
6	1	1.36	1.30	1.30	1.28
7	1	0.55	0.52	0.50	0.45
Total	160	160	160	160	160
ML Estimate		$n = 7$ $a = 5.513321$ $b = 53.295234$ 0.230829(3)	$n = 7$ $\mu = -0.79706$ $a = 5.723447$ $b = 36.39953$	$n = 12.024178$ $\beta = 0.1830634$ $\alpha = 5.6509704$ $\gamma = 37.1968856$	$n = 7$ $a = 4.471435$ $b = 0.001004$ $\gamma = 29.854893$
χ^2 d.f.			0.1877409 (2)	0.1813224 (1)	0.1429262 (2)

SM: Sen and Mishra HB: Hassan and Bilal

4. GOODNESS OF FIT

In this section, we present three data sets available in the literature to examine the fitting of the proposed model and comparing it with the fitting of Polya-Eggenberger distribution (PED), the generalized Polya-Eggenberger distribution (GPED) defined by Sen and Mishra (1996) and the generalized negative Polya-Eggenberger distribution (GNPED) defined by Hassan and Bilal (2006). Due to complicated likelihood function, the maximum likelihood estimate of the parameters of the proposed model is not straightforward and need some iterative procedure such as Fisher's scoring method, Newton-Rapson method etc. for their solution. R-software provides one among such solutions. Therefore, a computer programme in R-software is used to estimate the parameters of the distribution. The ML estimates of the parameters so obtained are shown at their respective places in the table. It may be mentioned here that the parameter n in case of PED**,** GPED defined by Sen and Mishra (1996) and Proposed model are known where as in case of GNPED, defined by Hassan and Bilal (2006), the value of n is estimated. It is clear from all the tables above that the proposed model gives a satisfactory fit and provides a better alternative than the compared distributions.

REFERENCES

Earl, D. Rainville (1971). Special Functions (Re-print). Chelsa Publishing Company, Bronx, New York.

Eggenberger, F. and Polya, G. (1923). Uber die Statistik verketteter vorgange, Z.Angew. Math. Mech.1, 279-289.Newbold, E. (1927). Practical applications of the statistics of repeated events, particularly to industrial accidents. J. Roy. Statist. Soc. 90, 487-547.

Gupta, A.K. and Nadarajah, S. (2004).Handbook of Beta Distribution and its Applications. New York: Marcel Dekker.

Hassan, A. and Bilal, S. (2006). A generalized negative Polya-Eggenberger distribution and its applications. International Journal of Modern Mathematics, volume 1, number 1, Oct. 2006, 97-113.

Janardan, K. G. (1973). A new four urn model with predetermined strategy. Technical Report 37-I, Sangamon State University, Springfield II.

Janardan, K. G. (1975). Markov-Polya urn models with predetermined strategies-I, Gujarat Statist. Rev. 2(1), 17-32.

Janardan, K. G. and Schaeffer, D. J. (1977). A generalization of the Markov-Polya distribution. Its extensions and applications. Biom. Zeit 19, 87-106.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995). Continuous Univariate Distributions, volume 2 (second edition).New York: John Wiley and Sons.

Saralees Nadarajah and Samuel Kotz (2003). A generalized Beta distribution Π. Inter- Stat. Dec. 2003, 1-7.

Sen, K. and Mishra, A. (1996). A generalized Polya-Eggenberger model generating various discrete probability distributions. Sankhya.

Sen, K. and Ritu, J. (1996). Generalized Markov-Polya urn models with predetermined strategies. Journal of Statistical planning and Inference 54, 119-133.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 701–707.

DEALING WITH ROUNDED ZEROS IN COMPOSITIONAL DATA UNDER DIRICHLET MODELS

Rafiq H. Hijazi
Department of Statistics
United Arab Emirates University
P. O. Box 17555, Al-Ain, UAE
E-mail: rhijazi@uaeu.ac.ae

ABSTRACT

One of the obstacles facing the application of the Dirichlet modeling of compositional data is the occurrence of zero. In the Dirichlet model, the presence of zeros makes the probability density function vanish. Zeros in compositional data are classified into "rounded" zeros and "essential" or true zeros. The rounded zero corresponds to a small proportion or below detection limit value while the essential zero is an indication of the complete absence of the component in the composition. Several parametric and non-parametric imputation techniques have been proposed to replace rounded zeros and model the essential zeros under logratio model. In this paper, a new method based on Beta regression is proposed for replacing rounded zeros in compositional data. The performance of the proposed method is analyzed using Monte Carlo simulation and an illustrative example using real data is given.

1. INTRODUCTION

Compositional data are non-negative proportions with unit-sum. This type of data arises whenever objects are classified into disjoint categories and the resulting relative frequencies are recorded, or partition a whole measurement into percentage contributions from its various parts. The sample space of compositional data is the simplex S^D defined as

$$S^D = \left\{ (x_1, \ldots, x_D) : x_j > 0 \text{ for } j = 1, 2, \ldots, D \text{ and } \sum_{j=1}^{D} x_j = 1 \right\}$$

Compositional data occur in nearly all disciplines, but recognition and modeling of their basic structure have gotten particular attention in geology, chemistry, political science, business and economics. For example, economists might be interested in how the composition of household income spent on food, housing, clothes, entertainment and services. Due to the unit-sum constraint and its consequences, traditional regression models are not suitable for modeling compositional data. Aitchison (1986) suggested an analysis based on the logratios of the compositional data. Campbell and Mosimann (1987) developed an alternative approach by extending the Dirichlet distribution to a class of Dirichlet Covariate Models. Hijazi and Jernigan (2009) developed maximum likelihood inference in Dirichlet regression models. Hijazi (2006, 2008) investigated the diagnostics checking and the residuals analysis in Dirichlet regression.

In compositional data analysis, the presence of zero components represents one of the main obstacles facing the application of both logratio analysis and Dirichlet regression. In a logratio

analysis, we cannot take the logarithm of zero when applying the additive logistic transformation. In the Dirichlet model, the presence of zeros makes the probability density function vanish.

In this paper, we propose a new technique, based on Beta regression, for replacing the zeros under Dirichlet model. Section 2 gives an overview of zeros and zero replacement strategies in compositional data besides the new proposed replacement method. A Monte Carlo simulation study to compare the proposed method with the multiplicative replacement strategy is presented in Section 3. An application to illustrate the use of the proposed technique is presented in Section 4. Finally, concluding remarks are given in Section 5.

2. ZERO REPLACEMENT IN COMPOSITIONAL DATA

2.1 Types of Zeros in Compositional Data

Aitchison (1986) classified the zeros in compositional data into "rounded" or trace elements zeros and "essential" or true zeros. The trace zero is an artifact of the measurement process, where observation is recorded as zero when it is below the detection limit (BDL). For example, in the porphyry deposits, assume that we record the amount on the different elements. If the scale used does not identify the presence of the element if it is less than 0.2%, then this component is recorded as zero. Thus the observed zero is a proxy for a very small number below 0.2%. On the other hand, often the observation is recorded as zero as an indication of the complete absence of the component in the composition. In the household budget, a household spending nothing on tobacco will have a zero for the share of the tobacco in the budget.

2.2 Common Zero Replacement Strategies

The treatment of the zero observations in compositional data should be done according to the cause of the zero (Aitchison 1986). Several attempts have been made to deal with the essential zeros using ranks (Bacon-Shone 1992) and conditional modeling (Aitchison and Kay 2003). In case of rounded zeros, Aitchison (1986) suggested the reduction of the number of components in the composition by amalgamation. That is, eliminating the components with zero observations by combining them with some other components. Such approach is not appropriate when the goal is modeling the original compositions or the model includes only three components. However, a more logical approach is to replace the rounded zeros by a small nonzero value that does not seriously distort the covariance structure of the data (Martìn-Fernández *et al.* 2003a). The first replacement method, the additive replacement, proposed by Aitchison (1986) is simply replacing the zeros by a small value δ and the normalizing the imputed compositions. Fry *et al.* (2000) showed that the additive replacement is not subcompositionally coherent and consequently, distorts the covariance structure of the data set.

Martin-Fernandez *et al.* (2003a) proposed an alternative method using a multiplicative replacement which preserves the ratios of nonzero components. Let $x = (x_1,\ldots,x_D) \in S^D$ be a composition with rounded zeros. The multiplicative method replaces the composition x containing c zeros with a zero-free composition $r \in S^D$ according to the following replacement rule

$$r_j = \begin{cases} \delta & if \quad x_j = 0 \\ (1-c\delta)x_j & if \quad x_j > 0 \end{cases} \tag{1}$$

In addition, Martin-Fernandez *et al.* (2003a) emphasized that the best results are obtained when δ is close to 65% of the detection limit. However, since the multiplicative replacement imputes exactly the same value in all the zeros of the compositions, this replacement introduces artificial correlation between components which have zero values in the same composition.

Besides these nonparametric approaches, several parametric approaches based on applying a modified EM algorithm on the additive logratio transformation (Martin-Fernandez *et al.* (2003b), Palarea-Albaladejo *et al.* (2007) and Palarea-Albaladejo and Martìn-Fernández (2008)). However, none of these methods is applicable when the compositional data arise from Dirichlet model.

2.3 Beta Regression Based Strategy

As mentioned earlier, the existing parametric replacement strategies assume that the compositional data arise from the additive logistic normal distribution, the underlying distribution in logratio analysis (Aitchsion 1986). When the underlying distribution is Dirichlet, a natural imputation method should be based on Beta distribution as a marginal of Dirichlet distribution. Ferrari and Cribari-Neto (2004) have proposed a regression model when the response variable is beta distributed. The proposed model is based on the parameterization of the mean and dispersion parameters in the beta distribution as follows. If Y is beta distributed with parameters p and q, then the dispersion parameter $\phi=p+q$ and the mean parameter is then $\mu = \pi/\phi$. The density of Y using this parameterization is given by

$$f(y;\mu,\phi)= \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0<y<1 \tag{2}$$

Similar to OLS regression, the beta regression model is obtained by assuming that the mean μ can be written as a linear combination of some independent variables $x_1,\ldots,x_k$ using a link function $g(.)$ as

$$g(\mu)= \sum_{i=1}^{k} x_i\beta_i \tag{3}$$

where $\beta_1,\ldots,\beta_k$ is a vector of unknown parameters. Let $\mathbf{X}=(\mathbf{x_1},\ldots,\mathbf{x_D})$ be the compositional data with rounded zeros in component x_j. Our proposed approach works as follows:

1. Split **X** based on the existence of rounded zeros in x_j into a zero-free subdata $X_{(1)}$ and $X_{(2)}$, a subdata with all zeros in component x_j.
2. Apply beta regression on portion of $X_{(1)}$ where the values of the j^{th} component are close to the detection limit. The j^{th} component is the response variable and the rest of components as covariates.
3. Using the estimated regression parameters in (2), predict the imputed values of the rounded zeros in $X_{(2)}$.
4. Use the multiplicative replacement strategy using the imputed values to replace the rounded zeros.
5. Repeat this process sequentially on components with rounded zeros.

It is clear that this method takes into account the information included in the covariance structure and produces different imputed values for each composition. The method also assumes that the component with rounded zeros is correlated with the other components in the compositional data especially in $X_{(1)}$. It is noteworthy that external covariates related to the component with rounded zeros might be used in the regression model sole or jointly with the compositional components. In addition, this method would not replace the zeros by negative values but it might replace them by values over the detection limit.

3. SIMULATION-BASED RESULTS

Our interest is mainly focused on to what extent the zero replacement strategies affect to the estimation of the relationship between the components. Consider a 4-component random composition drawn from Dirichlet distribution with parameters 1, 4, 15 and 20 i.e., $\mathcal{D}(1,4,15,20)$. For our simulation purposes, 100 datasets **X** of size 200 are drawn from the above Dirichlet distribution. Next, small values under the detection limit in the first component of **X** are replaced by zero. A range of 10 detection limits is considered from 0.0025 to 0.025 with increments of 0.0025. Thus, 1000 datasets containing different number of rounded zeros are generated. To measure the distortion between the original data **X** and the imputed data $\mathbf{X}^*$, the mean squared distances (MSD) is used. The MSD is given by

$$MSD = \frac{\sum_{i=1}^{200} d_a^2(x_i, x_i^*)}{200} \tag{4}$$

where the Aitchison's distance (d_a) is defined as

$$d_a(x_i, x_i^*) = \sqrt{\sum_{i<j} \left(\ln \frac{x_i}{x_j} - \ln \frac{x_i^*}{x_j^*} \right)^2} \tag{5}$$

Figure (1) shows the mean MSD for the two replacement methods for the different detection limits. For small detection limits, the multiplicative replacement seems to perform better than our proposed method. However, as the detection limit increases and consequently the percentage of zeros, the beta regression based method outperforms the multiplicative replacement method. Same conclusion is drawn if the Euclidean distance is used instead of Aitchison distance in (4). To compare the effect of replacement method on the variability in the compositional data, the mean estimates of Dirichlet parameters are used. The variability in Dirichlet model is inversely proportional to the sum of its parameters. Figure (2) shows the mean estimate of the first parameters in the original data and the imputed data. Similar behavior is shown for the rest of the parameters. Compared to the proposed method, under the multiplicative replacement, the parameter is clearly overestimated and hence the total variability is underestimated. Such underestimation increases as the percentage of zeros increases. This is due to the replacement of all zeros with same value which is not the case in beta regression based method.

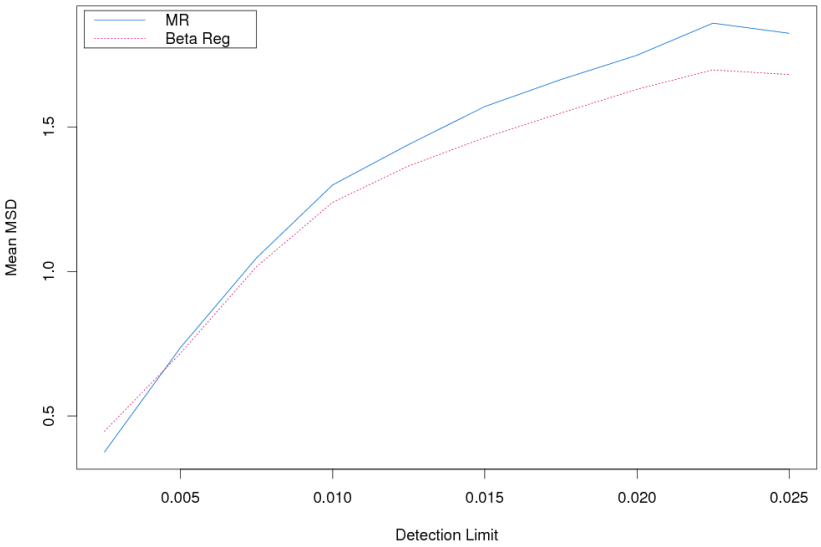


Figure 1: Replacement methods distortion

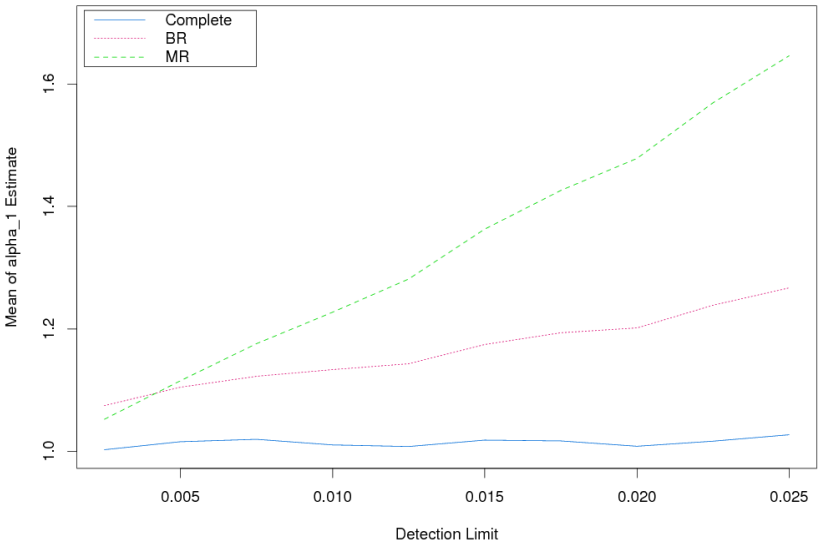


Figure 2: Parameter estimation under Replacement methods

4. APPLICATION

Consider the data collected from a deep-sea core measuring 478 cm in length from the Mediterranean Sea floor (Davis 2002). The core was split and grain-size analysis was made of 51 intervals. This paper focuses on the new proposed replacement method, we will consider that the instrument used does not detect the presence of any element if the percentage is less that 2.5%,

i.e. the detection limit is 2.5%. This will result in 9 compositions with sand component recorded as rounded zero as shown in Figure (3a). The Aitchison distance between the original data and the imputed datasets using the beta regression based method and multiplicative methods are 0.0042 and 0.0094, respectively. This indicates that the new replacement method yielded an imputed data which is closer to the original than the one produced by the multiplicative method. The compositions with rounded zeros and the corresponding imputed compositions are shown in Figure (3b).

The maximum likelihood estimates of the original data and the imputed data are given in Table (1). It is clear that the multiplicative replacement method underestimated the model parameters compared to the original data but the beta regression based method overestimated such parameters. The sum of the estimates under the proposed method is slightly larger than the corresponding sum in the original data resulting in a slightly smaller estimate of the variability. However, the multiplicative replacement method yielded slightly larger estimate of the variability.

Table 1: Maximum likelihood estimates of original and imputed sediments data

	Clay	Silt	Sand	Sum
Original Data	10.599	28.702	2.672	41.973
Beta regression replacement	11.041	29.934	2.831	43.806
Multiplicative replacement	10.243	27.709	2.538	40.490

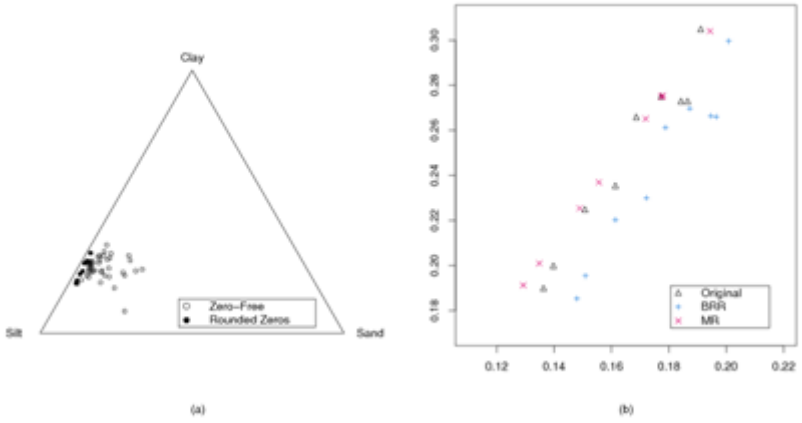


Figure 3: (a) The ternary diagram of the sediments data (b) The original compositions before rounding, imputed compositions with BRR (beta regression based replacement) and imputed compositions with MR (multiplicative replacement).

5. COMMENTS AND CONCLUSION

In this work we have proposed a new replacement method based on beta regression under Dirichlet model. The proposed method was compared with the multiplicative replacement method through simulation study and real data example implemented in S-Plus. The new method outperforms the multiplicative replacement method especially in datasets with large percentage

of zeros. This method gives positive imputed value but does not take into account the detection limit of the part. The method should be modified to overcome this deficiency. The proposed method is expected to be less efficient in the absence of correlation between the component with zeros and other components or external variables.

REFERENCES

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, New York.

Aitchison J., Kay J. W. (2003). Possible solutions of some essential zero problems in compositional data analysis. In: Thió-Henestrosa S., and Martìn-Fernández, J.A. (Eds.), Proceedings of CODAWORK'05, The 2nd Compositional Data Analysis Workshop, October 19-21, University of Girona, Girona (Spain).

Bacon-Shone, J. (1992).Ranking Methods for Compositional Data. *Applied Statistics*, 41, 533-537.

Campbell, G., and Mosimann, J. E. (1987). Multivariate methods for proportional shape. *ASA Proceedings of the Section on Statistical Graphics*, 10-17.

Davis, J. C. (2002). *Statistics and data analysis in geology*. John Wiley & Sons, New York.

Ferrari, S. L. P., Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. *Journal of Applied Statistics*, 31(7), 799–815.

Fry, J. M., Fry, T. R. L., and McLaren, K. R. (2000). Compositional data analysis and zeros in micro data: *Applied Economics*, 32(8), 953-959.

Hijazi, R. (2006). Residuals and Diagnostics in Dirichlet Regression. *ASA Proceedings of the Joint Statistical Meetings 2006, American Statistical Association*, 1190-1196.

Hijazi, R. (2008). Residuals Analysis of the Dirichlet Regression. *Egyptian Statistical Journal*, 52 (2), 109-120.

Hijazi, R., Jernigan, W. (2009). Modeling Compositional Data Using Dirichlet Regression. *Journal of Applied Probability and Statistics*, 4 (1), 77-91.

Martìn-Fernández, J.A., Barceló-Vidal, C., Pawlowsky-Glahn, V., (2003a). Dealing with zeros and missing values in compositional data sets. *Mathematical Geology*, 35, 253-278.

Martìn-Fernández, J.A., Palarea-Albaladejo J, Gómez-Garcìa, J. (2003b). Markov chain Monte Carlo method applied to rounding zeros of compositional data: first approach. In: Thió-Henestrosa S., and Martìn-Fernández, J.A. (Eds.), Proceedings of CODAWORK'05, The 2nd Compositional Data Analysis Workshop, October 19-21, University of Girona, Girona (Spain).

Palarea-Albaladejo, J., Martìn-Fernández, (2008). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computers & Geosciences*, 34, 902-917.

Palarea-Albaladejo, J., Martìn-Fernández, J.A., Gómez-Garcìa, J., (2007). A parametric approach for dealing with compositional rounded zeros. *Mathematical Geology*, 39, 625-645.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 708–732.

REGRESSION BASED ESTIMATES FOR THE BOX-COX POWER TRANSFORMATION

Osama Abdelaziz Hussien[1] and Remah El-Sawee[2]
Department of Statistics, Faculty of Commerce, Alexandria University, Egypt
E-mail: [1]osama52@gmail.com, [1]ossama.abdelaziz@alexcommerce.edu.eg,
[2] remah-elsawee@hotmail.com

ABSTRACT

The aim of this article is to propose a new method to estimate the Box-Cox transformation parameter simultaneously with the parameters of a location-scale family. The new estimator modifies Lloyd's generalized least-squares method by adding the transformation parameter to the simple linear regression model to became $y_{(i)}^{\lambda}=\mu+\sigma E(z_{(i)}) + \varepsilon_{(i)}$ I = 1, 2, …, n, where $z_{(i)}$,I = 1, 2 ,…,n, are the order statistics from a standard normal distribution. The regression-based estimator for λ is the one that minimizes the residual sum of squares of the above model. Given λ the corresponding estimators of μ and σ^2 are BLUE. We show that the artificial regression estimator, Halawa (1996), and the Shapiro-Wilk estimator, Rahman (1999), are special cases of the regression-based estimator. A simulation study was conducted to compare the performance of the regression-based estimators with the normal likelihood estimators of Box-Cox.

Keywords: Box-Cox transformation, goodness-of-fit tests, Shapero-Wilk test, artificial regression, generalized least squares.

1. INTRODUCTION

There has been a considerable literature on the subject of power transformation since they were introduced by Box and Cox in (1964). In a single sample model the main object of power transformation is to achieve normality. The Box-Cox family of power transformations enlarges the parametric model to include transformation parameters and then estimating these parameters simultaneously with the parameters of the original model. Suppose that F is a family of absolutely continuous distributions with cdf's of the form $F(y)=G((x-\mu)/\sigma)$ ($\sigma>0$). In other words, F is a location-scale family. Let $h(Y,\lambda)$ denotes the Box-Cox power transformation, i.e.

$$h(Y,\lambda) = \begin{cases} \frac{Y^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ ln(Y) & \lambda = 0 \end{cases} \quad (1)$$

Suppose that there exists a λ^{*} which produces the model

$$h(Y,\lambda^{*}) = \mu + \sigma\,\varepsilon \tag{2}$$

where ε has density f. The Box-Cox normal likelihood estimator of the parameter vector $\theta=(\mu,\sigma,\lambda)^t$ is based on the assumption that under the model (2) the density f is the standard normal distribution. The log likelihood function used to estimate θ is given by

$$\ell_{BC}(\lambda,\ \mu,\ \sigma; Y_1, ..., Y_n) = \frac{-n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2)$$

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i^{(\lambda)} - \mu)^2 + (\lambda - 1)\sum_{i=1}^{n}\ln(Y_i) \tag{3}$$

For some initial value for λ, one can use (1-3) to find maximum likelihood estimates of μ and σ^2. Substituting those estimates in (3) one can find the maximum likelihood estimate of λ by maximizing (3). Hence, one gets the Box-Cox normal maximum likelihood estimators

$$\hat{\theta}_{BC} = (\hat{\lambda}_{BC}, \hat{\mu}_{BC}(\hat{\lambda}), \hat{\sigma}_{BC}(\hat{\lambda}))^t$$

where

$$\hat{\mu}_{BC} = \overline{Y^{(\hat{\lambda})}} \tag{4}$$

$$\hat{\sigma}_{BC} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i^{(\hat{\lambda})} - \overline{Y^{(\hat{\lambda})}})^2} \tag{5}$$

Kouider and Chen (1995) proved that the ℓ_{BC} is concave downward and hence it has a local maximum. Cho et al. (2001) established strongly consistent and asymptotically normal of $\hat{\theta}_{BC}$.

Bickel and Doksom (1981) showed that "the performance of all Box-Cox type procedures is unstable and highly dependent on the parameters of the model in structured models with small to moderate error variances". Carrol (1982) showed that $\hat{\theta}_{BC}$ is very sensitive to outliers and it can be highly inefficient if the distribution of ε has heavier tails than the normal distribution. Robust estimators for the Box-Cox transformation parameter have been proposed by Carrol (1980, 1982) and by Bickel and Doksum (1981). A bounded influence estimator of (λ, μ, σ) was given by Carrol and Ruppert (1985).

In this article we propose a new estimator for Box-Cox transformation parameter that utilizes a modified Lloyd's generalized least-squares model:

$$h(Y_{(r)},\lambda^{*}) = \mu + \sigma E(Z_{(r)}) + \varepsilon_r \qquad (6)$$

where $Y_{(1)},Y_{(2)},\ldots,Y_{(n)}$ are order statistics for a random sample from a location-scale family F, and $Z_{(r)}=(Y_{(r)}-\mu)/\sigma$, r=1,2,…,n. Accordingly, the resulting estimators of μ and σ^2 are BLUE. A robust form of this regression-based estimator will be presented in a subsequent article.

Also, we show that artificial regression estimator, Halawa (1996), and the Shapiro-Wilk estimator, Rahman (1999), are special cases of the regression-based estimator. We compare the performance of the regression-based estimators with the normal likelihood estimator of Box-Cox according to two criteria:

1. The mean squared error of the estimate.
2. Empirical power of a goodness-of-fit test for the transformed data.

The rest of the paper is organized as follows. Section 2 describes the proposed general regression based estimators, the artificial regression estimator and the Shapiro-Wilk estimator. Section 3 discuses the results of the Monte Carlo simulations according to the mean square error criterion. Section 4 present the comparisons according to the goodness-of-fit tests. Conclusions and remarks for further studies are presented in section 5.

2. REGRESSION-BASED ESTIMATORS FOR THE BOX-COX TRANSFORMATION PARAMETER

Given the model (6), let $E(Z_{(r)})=\alpha_r$, $cov(Z_{(r)},Z_{(s)})=\beta_{rs}$ r,s=1,2,…,n.

be the expected values and covariances of order statistics from a random sample of the standard normal distribution. Then

$$E(Y_{(r)}) = \mu + \sigma\,\alpha_r , \qquad cov(Y_{(r)},Y_{(s)}) = \sigma^2\,\beta_{rs}$$

where the α_r , β_{rs} can be evaluated "once and for all". These equations can be written as

$$E(Y_o)=\mu\, \mathit{1} + \alpha\sigma = A\theta \qquad cov(Y_o)=\sigma^2\, V,$$

where Y_o and α are the column vectors of the $Y_{(r)}$ and α_r respectively; *1* is a is a column of n 1's and $A=(1,\alpha)$, $\theta^t=(\mu\ ,\ \sigma)$, $\sigma^2\, V$ is the covariance matrix of the $Y_{(r)}$.

If the covariance matrix is positive definite, we can apply the generalized Gauss-Markov least-squares theorem to get the BLUE of θ as

$$\hat{\theta} = \left(A^t \Omega A\right)^{-1} A^t \Omega Y_o \qquad (7)$$

$$\text{cov}(\hat{\theta}) = \sigma^2 \left(A^t \Omega A\right)^{-1} \tag{8}$$

where $\Omega = V^{-1}$. The result that generalized Gauss-Markov estimators above are BLUE is known as the Aitken theorem. For a proof see Kariya and Kuata (2004), page 34. The residual sum of squares is

$$\text{RSS} = (Y_o - A\hat{\theta})^t \Omega (Y_o - A\hat{\theta}) = Y_o^t (I - M_x)^t \Omega (I - M_x) Y_o \tag{9}$$

where $M_x = A(A^t \Omega A)^{-1} A^t \Omega$. We suggest a regression based estimator of the Box-Cox power transformation as follows:

1. Start by initial value of λ (say λ^*).
2. Compute $\hat{\theta}(\lambda^*) = \left(A^t \Omega A\right)^{-1} A^t \Omega Y_o^{\lambda^*}$
3. Use an optimization procedure to find λ that minimizes RSS

$$RSS_{|\lambda^*} = Y_o^{\lambda^* t} (I - M_x)^t \Omega (I - M_x) Y_o^{\lambda^*}$$

4. Repeat (2) and (3) until convergence.

A robust form of this estimator could be defined by the location and scale linear estimators from double censored samples presented by Sarhan and Greenberg (1956; 1958). The moments of order statistics from the normal distribution (A and Ω) are tabulated for small samples only, Harter (1961). Further study is needed to explore the small sample and large sample behavior of the proposed regression based estimators.

We establish here only $\text{COV}(\hat{\theta}_{|\lambda^*})$. If the transformation parameter is known ($\lambda = \lambda^*$), then $h(Y_r, \lambda^*)$ will have a standard normal distribution. David and Nagaraja (2003) showed that

$$\text{cov}(\hat{\mu}\big|_{\lambda^*}, \hat{\sigma}^2\big|_{\lambda^*}) = 0 \qquad \text{var}(\hat{\mu}\big|_{\lambda^*}) = \frac{\sigma^2}{n} \qquad \text{var}(\hat{\sigma}\big|_{\lambda^*}) = \frac{\sigma^2}{\alpha' \Omega \alpha}.$$

Note that $\text{var}(\hat{\mu}\big|_{\lambda^*}) = \text{var}(\hat{\mu}_{BC})$ $\qquad \text{var}(\hat{\sigma}_{BC}) = \dfrac{\sigma^2}{2n}$

Using the available tables for the mean and covariances of order statistics, we compute $\alpha' \Omega \alpha$. The table below shows that for $n \leq 30$ $\text{var}(\hat{\sigma}^2\big|_{\lambda^*}) < \text{var}(\hat{\sigma}_{BC})$

n	$1//(\alpha^t\Omega\alpha)$	1/2n
2	0.180045	0.25
4	0.105696	0.125
6	0.074603	0.083333
8	0.057587	0.0625
10	0.046868	0.05
16	0.030044	0.03125
20	0.024222	0.025
24	0.020152	0.020833
28	0.016257	0.017857
30	0.01097	0.016667

2.1 The Artificial Regression Estimator

Halawa (1996) uses the model given by (6) with $E(X_{(i)})$ replaced by its Blom's approximation $z_{(i)}$, Blom (1958) and Lin and Vonesh (1989),

$$z_{(i)} \approx \Phi^{-1}\left[\frac{i-0.375}{n+0.25}\right]$$

The artificial regression model, given by

$$Y_{(i)}^{(\lambda)} = \mu + \sigma z_{(i)} + \varepsilon_{(i)} \quad , i = 1,2,\ldots,n \tag{10}$$

adds a covariate z to the original model (1) that could reduce the large variance of the Box-Cox transformation estimates.

The log likelihood for the model (10) is given by

$$\ell_R(\lambda, \mu, \sigma; Y_1,..,Y_n) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(Y_{(i)}^{(\lambda)} - \mu - \sigma z_{(i)})^2 + \sum_{i=1}^{n}\ln(Y^{\lambda-1}) \tag{11}$$

Start by initial value of λ (say λ^*), one gets the maximum likelihood estimates of $\mu|_{\lambda^*}$ and $\sigma|_{\lambda^*}$. Substitute these values in (5) one gets $\hat{\lambda}$ by a maximization numerical procedure. Substitute in $\hat{\mu}$ and $\hat{\sigma}$ and iterate until converge. The artificial regression estimators will be denoted by

$$\hat{\theta}_R = (\hat{\lambda}_R, \hat{\mu}_R(\hat{\lambda}), \hat{\sigma}_R(\hat{\lambda}))^t.$$

Halawa (1996) proved that the artificial regression estimators are strongly consistent and asymptotically normal.

Note that $\text{var}(\hat{\mu}_{BC}) = \frac{\sigma^2}{n}$ while $\text{var}(\hat{\mu}_R) = \frac{1}{n}$,

$$\text{var}(\hat{\sigma}_{BC}) = \frac{\sigma^2}{2n} \quad \text{while } \text{var}(\hat{\sigma}_R) = \frac{1}{n}$$

i.e. the artificial regression estimator gives better results than the normal likelihood estimators for $\sigma^2 \geq 2$.

2.2 The Shapero-Wilk Estimator

Given the model (6) the Shapiro-Wilk statistic is defined by

$$W = \frac{\hat{\sigma}^2}{S^2} \cdot B^2 \tag{12}$$

where $S^2 = \sum_{i=1}^{n} (y - \bar{y})^2 / (n-1)$,

$$\hat{\sigma}\Big|_{\lambda^*} = \frac{\alpha^t \Omega X_*}{\alpha' \Omega \alpha} \tag{13}$$

$$B^2 = \frac{\alpha' \Omega \alpha}{\alpha' \Omega \Omega \alpha} \tag{14}$$

Note that $\hat{\sigma}^2$ and S^2 are unbiased estimators for the slope of regression of $y_{(i)}$ on $E(Z_{(i)})$. The constant B^2 ensures that the test statistic W always takes values between 0 and 1, Shapiro and Wilk (1965).

A computing formula for W is given by $W = (a^t Y_0)^2 / S^2$, where $a = \Omega\alpha / \sqrt{\alpha^t \Omega \Omega \alpha}$. Shapiro and Wilk used approximations for the components a_i of a, and gave a table for sample sizes from n=3 to 50.

D'Agostino, and Stephen, (1986) page 206 stated that "*Alternatively, since in the generalized least squares analysis RSS is minimized by the parameter estimates* $\hat{\mu}$ *and* $\hat{\sigma}$ *, the test might be based on* $Z_2(X,\alpha)=RSS/S^2$. *Some examination of such tests has been made by Spinelli (1980), for the exponential and the extreme-value distributions, but otherwise they have not been much developed*".

The W-statistic of the transformed data for given λ is

$$W = \frac{(\sum_{i=1}^{n} a_i Y_{(i)}^{(\lambda)})^2}{\sum_{i=1}^{n} (Y_{(i)}^{(\lambda)} - \overline{Y^{(\lambda)}})^2}$$

If there exist a λ such that the power transformed observations are "approximately" normally distributed $N(\mu,\sigma^2)$, the maximum W-statistic estimate is that value which maximizes the Shapero-Wilk W statistic, i.e. maximizes the observed significance level, of the transformed data. One method to find $\hat{\lambda}$ is solve nonlinear equation by letting the first derivative of W(λ) with respect to λ equal zero .

$$\frac{dW}{d\lambda} = \frac{\sum_{i=1}^{n} (Y_{(i)}^{(\lambda)} - \overline{Y^{(\lambda)}})^2 \frac{d}{d\lambda}(\sum_{i=1}^{n} a_i Y_{(i)}^{(\lambda)})^2 - (\sum_{i=1}^{n} a_i Y_{(i)}^{(\lambda)})^2 \frac{d}{d\lambda}(\sum_{i=1}^{n} (Y_{(i)}^{(\lambda)} - \overline{Y^{(\lambda)}})^2)}{[\sum_{i=1}^{n} (Y_{(i)}^{(\lambda)} - \overline{Y^{(\lambda)}})^2]^2}$$

Putting $dW/d\lambda=0$ we get

$$\sum_{i=1}^{n} (Y_{(i)}^{(\hat{\lambda})})^2 \sum_{i=1}^{n} a_i [Y_{(i)}^{\hat{\lambda}} \ln(Y_{(i)}) - Y_{(i)}^{(\hat{\lambda})}] - n(\overline{Y^{(\hat{\lambda})}})^2 \sum_{i=1}^{n} a_i \left[Y_{(i)}^{\hat{\lambda}} \ln(Y_{(i)}) - Y_{(i)}^{(\hat{\lambda})} \right] -$$
$$\sum_{i=1}^{n} a_i Y_{(i)}^{(\hat{\lambda})} [n(\overline{Y^{(\hat{\lambda})}})^2 - \overline{Y^{(\hat{\lambda})}} \sum_{i=1}^{n} Y_{(i)}^{\hat{\lambda}} \ln(Y_{(i)}) - \sum_{i=1}^{n} (Y_{(i)}^{(\hat{\lambda})})^2 + \sum_{i=1}^{n} Y_{(i)}^{(\hat{\lambda})} Y_{(i)}^{\hat{\lambda}} \ln(Y_{(i)})] = 0 \quad (15)$$

Solving the above equation (numerically) we get $\hat{\lambda}_W$. Using $\hat{\lambda}_W$ we get

$$\hat{\mu}_W = \sum_{i=1}^{n} Y_i^{(\hat{\lambda})} / n = \overline{Y^{(\hat{\lambda})}} \qquad \hat{\sigma}_w = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i^{(\hat{\lambda})} - \hat{\mu}_W)^2}$$

Sun (1978) conducted a small Monte Carlo simulation to study the performance of the transformed estimates for two sample sizes (25, 50). He found that the maximum W-statistic has smaller bias but larger variances than the normal likelihood estimator and the bias increases and the sample variance of both estimates decreases as the absolute value of λ increases.

Rahman (1999) conducted a Monte Carlo simulation to compare the performance of $\hat{\lambda}_W$, $\hat{\lambda}_R$ and $\hat{\lambda}_{BC}$ for sample sizes (20, 40, and 100). In almost all situations he considered, variability for $\hat{\lambda}_R$ are smaller compared to that of $\hat{\lambda}_W$ and $\hat{\lambda}_{BC}$.

3. SIMULATION STUDY

We conduct a simulation study to compare the performance of the three methods of estimation at sample sizes 20, 40 and 100. To cover the situations not considered by Rahman (1999), we use: λ = 0, 0.25, 0.5, 2, noise ratio (μ/σ) = 1, 2, 3, 4, 5, and standard deviation for the transformed data = 1, 2, 3, 4, 5. For each method of estimation a 1000 random sample from the standard normal distribution were generated with initial seed 9831815 using the mathematical programming language GAUSS(9.0). We use the inverse transformation

$$Y = \begin{cases} 1+\lambda(\mu+\sigma\varepsilon) & \lambda \neq 0 \\ \exp(\mu+\sigma\varepsilon) & \lambda = 0 \end{cases}$$

The choices of the parameters are made such that the Y vector is always positive. Tables (1) to (5) give the biased, variances and average mean square error of the estimates for the cases considered.

Table (1) compare the performance of the estimators for different values of λ, with fixed values of μ and σ. This table shows that:

- $\hat{\lambda}_W$ has the highest biased and its variances are inflated.
- The ratios var($\hat{\lambda}_W$)/var($\hat{\lambda}_R$) and var($\hat{\lambda}_{BC}$)/var($\hat{\lambda}_R$) is very large. This means that the estimator $\hat{\lambda}_R$ is stable in a small interval near λ.
- The mean square error of $\hat{\mu}_R$ and $\hat{\sigma}_R$ are the smallest for all values of λ.

Table (2) illustrate the effect of the sample size on the estimates, with fixed σ and noise ratio. The table shows that:

- The mean square error of $\hat{\lambda}_R$ is much smaller than the mean square error of the other estimators for all sample sizes.
- The variances of all estimates decreases with the increase in the sample size, but the variances of the artificial regression method are much smaller for all sample sizes.

Table (3) illustrates the performance of λ estimators with different values of σ and noise ratios. The table shows that:

- $\hat{\lambda}_R$ has the smallest mean square error when (σ=2, μ/σ>2) , (σ=1,μ/σ=5) and when σ>2 for all noise ratios.
- The mean square error of $\hat{\lambda}_W$ is the smallest when (σ=2, μ/σ=1).

- The mean square error of $\hat{\lambda}_{BC}$ is the smallest when ((σ=1, μ/σ<1) and when (σ=2, μ/σ=2).
- For (σ=4, μ/σ=1, 2, 5) and for (σ=5, μ/σ=1) the biased of $\hat{\lambda}_R$ is the smallest.
- For (σ=5, μ/σ= 5) the biased of $\hat{\lambda}_{BC}$ is the smallest.
- For all other combinations of σ and μ/σ the biased of $\hat{\lambda}_R$ is the smallest.
- The variance of $\hat{\lambda}_R$ is the smallest except when (σ<3, μ/σ=1), where the variance of $\hat{\lambda}_{BC}$ is the smallest.

Table (4) illustrates the effect of the estimated λ on $\hat{\mu}$ .The table shows that:

- The bias of $\hat{\mu}$'s increases as the noise ratio increases.
- The bias of both $\hat{\mu}_{BC}$ and $\hat{\mu}_W$ increase as σ increases
- For σ≤3 the bias of $\hat{\mu}_R$ decreases as σ increases.
- For σ> 3 the bias of $\hat{\mu}_R$ increases as σ increases.
- The variance of $\hat{\mu}$'s increases as the noise ratio increases.
- $\hat{\mu}_R$ has the smallest mean square error for σ>2 and for (σ=2 , μ/σ >2) .
- $\hat{\mu}_{BC}$ has the smallest mean square error for. for σ=1 and for (σ=2 , μ/σ ≤2).

Table (5) illustrates the effect of the estimated λ on $\hat{\sigma}$ .The table shows that:

- $\hat{\sigma}_R$ has the smallest mean square error for σ>1, except when (σ=2, μ/σ =2) where $\hat{\sigma}_{BC}$ has the smallest mean square error.
- $\hat{\sigma}_{BC}$ has the smallest mean square error for (σ=1 , μ/σ ≤4)
- The mean square error of $\hat{\sigma}_{BC}$ and $\hat{\sigma}_W$ inflated as the noise ratio increases while $\hat{\sigma}_R$ remains stable.

4. GOODNESS-OF-FIT COMPARISONS

The second criterion to assess the performance of the power transformation parameter estimators is the goodness-of-fit of the transformed data. More specifically, which estimation method (for λ) will have the smallest significance level of a test for normality if the parent distribution is not normal. Three goodness-of-fit tests will be considered:

Anderson Darling test for normality: Anderson and Darling (1954) test is an EDF test. It is applicable for symmetric or asymmetric distributions for small and large samples. It is sensitive to outliers.

Shapiro-Wilk test for normality: Royston (1982) showed that the Shapiro-Wilk test is the most powerful test for symmetric light-tailed alternatives. Also, it has high power for skewed alternatives.

Filliben probability plot test for normality: Filliben (1975) introduces a test for normality as the correlation coefficient between the order statistics Y_o and its medians (instead of its means as the Shapiro-Wilk test statistic). The power of the test is high for symmetric heavy-tailed distributions. The critical values are tabulated for small and large sample sizes.

The simulation study was done in two phases:

Phase I:

Generate 1000 random sample (n=20, 40 and 100) from a transformed data on the form

$$Y = (1 + \lambda (\mu + \sigma \varepsilon))^{1/\lambda} \quad , \quad \varepsilon_i \sim N(0, 1),$$

where λ=1 and 5 , σ=3, and μ/σ=1, 2, 3, 4,and 5.

- For each sample compute $\hat{\lambda}_W$, $\hat{\lambda}_R$ *and* $\hat{\lambda}_{BC}$
- For each test for normality compute the empirical significance level defined as the percentage of cases the null hypothesis is rejected at significance level α=0.05

$$\hat{\alpha}_n = \frac{\#of T_m > Z_\alpha}{1000}$$

T_m is the test statistic for sample m, m=1,2,…,1000.

Table (6) gives 1- $\hat{\alpha}_n$ for the three test statistics with λ=1.

Table (7) gives 1- $\hat{\alpha}_n$ for the three test statistics with λ=0.5.

These tables show that:

- For the artificial regression method 1- $\hat{\alpha}_n$ is smaller than the corresponding values for the other methods of estimation.
- The empirical significance level is almost the same for the other methods.
- The performance of the artificial regression improves as the noise ratio increases.

Phase II:

- Generate 1000 random sample (n=20, 40 and 100) from the distributions in the following table.

Distribution	Skewness	Kurtosis
Lognormal (0 , 1)	6.185	113.936
SConN (0.1 , 9)	0	8.333
Gamma (4 , β)	1	7.5
Gamma (9 , β)	0.6667	5.0
Gamma (16 , β)	0.5	4.125
Weibull (2.1 , β)	0.5672	3.1323
Weibull (3.3 , β)	0.07791	3.7110
Weibull (4.5 , β)	-0.1784	2.8081

SConN means contaminated normal distribution of the form Y=0.9N(10,1)+0.1N(10,9), β=5 and 10.

- For each sample compute $\hat{\lambda}_W$, $\hat{\lambda}_R$ *and* $\hat{\lambda}_{BC}$ and transform the data.
- For each test for normality compute the empirical significance level defined as the percentage of cases the null hypothesis is rejected at significance level α=0.05.

Table (8) gives (1 - $\hat{\alpha}_n$) for the three test statistics for positively skewed distributions. Table (9) gives (1 - $\hat{\alpha}_n$) for the three test statistics for the negatively skewed Wiebull distribution. Table (10) gives (1 - $\hat{\alpha}_n$) for the three test statistics for the contaminated normal distribution.

These tables show that:

- The artificial regression method does not perform will, may be because of the use of the approximated means of order statistics.
- The effect of skewness and kurtosis on (1- $\hat{\alpha}_n$) has no clear pattern.
- The performance of all methods of transformation decreases as the sample size increases.
- The Box-Cox transformation was designed to remove skewness only, so it has low (1 – $\hat{\alpha}_n$) for heavy tailed distributions. In such cases a robust transformation could perform better.

5. COMMENTS AND CONCLUSION

We propose a general regression-based estimator for λ that minimizes the residual sum of squares of Lloyd's generalized least squares. Two estimators for the Box-Cox transformation parameter of this type have been presented; the artificial regression estimator, Halawa (1996),

and the Shapiro-Wilk estimator, Sun (1978), Rahman (1999) and Gaudard and Karson (2000). Both estimators depend on an approximate values for $E(Z_{(r)})$ and cov $(Z_{(r)}, Z_{(s)})$.A simulation study was conducted to compare the performance of the regression-based estimators with the normal likelihood estimator of Box-Cox $\hat{\lambda}_{BC}$ The simulation study shows that:

For fixed μ and σ:

- The artificial regression estimator for λ($\hat{\lambda}_R$) is stable in a small interval near λ.
- The Shapiro-Wilk estimator for λ ($\hat{\lambda}_W$) has the highest biased and its variance is inflated.
- The mean square error of $\hat{\mu}_R$ and $\hat{\sigma}_R$ are the smallest for all values of λ.

The effect of the sample size:

- For fixed σ and fixed noise ratio the mean square error of $\hat{\lambda}_R$ is much smaller than the mean square error of the other estimators for all sample sizes.
- The variances of all estimates decreases with the increase in the sample size, but the variances of the artificial regression method are much smaller for all sample sizes.

The performance of λ estimators with different values of σ and noise ratios:

- The artificial regression estimator $\hat{\lambda}_R$ has the smallest mean square error when (σ=2, μ/σ>2) , (σ=1,μ/σ=5) and when σ>2 for all noise ratios.
- The mean square error of $\hat{\lambda}_W$ is the smallest when (σ=2, μ/σ=1).
- The mean square error of $\hat{\lambda}_{BC}$ is the smallest when ((σ=1, μ/σ<1) and when (σ=2, μ/σ=2).
- The variance of $\hat{\lambda}_R$ is the smallest except when (σ<3, μ/σ=1), where the variance of $\hat{\lambda}_{BC}$ is the smallest.

The estimators of μ and σ are not robust for all methods:

- The bias of $\hat{\mu}$'s increases as the noise ratio increases.
- The variance of $\hat{\mu}$'s increases as the noise ratio increases.
- The mean square error of $\hat{\sigma}_{BC}$ and $\hat{\sigma}_W$ inflated as the noise ratio increases while $\hat{\sigma}_R$ remains stable.
- For σ≤3 the bias of $\hat{\mu}_R$ decreases as σ increases.
- For σ> 3 the bias of $\hat{\mu}_R$ increases as σ increases.

The empirical power for the goodness of fit tests shows that:

- The artificial regression method does not perform will, may be because of the use of the approximated means of order statistics.
- The effect of skewness and kurtosis on empirical power has no clear pattern.
- The performance of all methods of transformation decreases as the sample size increases.
- The Box-Cox transformation was designed to remove skewness only, so it has low empirical power for heavy tailed distributions. In such cases a robust transformation could perform better.

A further study is needed to compare the proposed estimator for the Box-Cox transformation by minimizing the residual sum of squares of the generalized least squares. A robust form of this estimator could be defined by the location and scale linear estimators from double censored samples. Further study is needed also do derive the influence functions for the above estimators.

APPENDIX

Table (1) λ estimates : n=40, σ=3 and noise ratio=5

λ		$\hat{\lambda}_{BC}$	$\hat{\lambda}_{R}$	$\hat{\lambda}_{W}$	$\hat{\mu}_{BC}$	$\hat{\mu}_{R}$	$\hat{\mu}_{W}$	$\hat{\sigma}_{BC}$	$\hat{\sigma}_{R}$	$\hat{\sigma}_{W}$
	bias	0.0801	-0.0383	0.0599	-25.879	-1.2477	-71.983	9.9239	0.1001	32.863
-1	var	0.5021	0.0018	0.5918	21995	2.0915	3.9001e+6	3775.4	0.0484	3.9001e+6
	mse	0.5085	0.0033	0.5953	22662	3.6481	3.9049e+6	3873.5	0.0584	3.9011e+6
	bias	0.0010	0.0088	0.0013	1.5894	1.126	1.9782e+59	0.7082	0.1869	1.2354e+60
0	var	0.0021	8.8117e-5	0.0055	47.575	1.7032	3.9134e+122	7.8388	0.0928	3.9134e+122
	mse	0.0021	0.0002	0.0055	50.096	2.9709	3.9134e+122	8.3396	0.1277	3.9286e+122
	bias	-0.0432	0.0302	-0.0286	12.555	1.4344	26.15	4.9459	0.1750	11.632
0.5	var	0.1474	0.0009	0.1674	3212.9	2.3344	96379	520.56	0.0569	96379
	mse	0.1492	0.0018	0.1682	3370.2	4.3915	97053	544.97	0.0875	96514.303
	bias	-0.0699	0.0450	-0.0416	27.401	1.4677	63.77	10.488	0.1586	27.662
1	var	0.5167	0.0020	0.5905	19016	2.3969	8.3749e+5	3205.9	0.0522	8.3749e+5
	mse	0.5215	0.0040	0.5921	19765	4.5508	8.4147e+5	3315.6	0.0773	8.3825e+5

Table (2) λ estimates for different sample sizes λ=0.5 σ=5 noise ratio=5

n		$\hat{\lambda}_{BC}$	$\hat{\lambda}_{R}$	$\hat{\lambda}_{W}$	$\hat{\mu}_{BC}$	$\hat{\mu}_{R}$	$\hat{\mu}_{W}$	$\hat{\sigma}_{BC}$	$\hat{\sigma}_{R}$	$\hat{\sigma}_{W}$
	bias	-0.0701	-0.1107	-0.0263	139.91	-7.9409	1.3054e+6	63.166	-2.5267	9.6032e+5
20	var	0.2605	0.0016	0.4043	2.3923e+6	6.4718	1.4606e+16	5.7696e+5	0.0504	1.4606e+16
	mse	0.2654	0.0138	0.4049	2.4116e+6	69.529	1.4606e+16	5.8089e+5	6.4346	1.4607e+16
	bias	-0.0409	-0.0558	0.0291	36.115	-4.4363	92.119	13.919	-1.5061	41.771
40	var	0.1339	0.0008	0.1551	32502	4.7984	2.0012e+6	5360.3	0.0839	2.0012e+6
	mse	0.1355	0.0039	0.1559	33803	24.479	2.0095e+6	5553.5	2.3523	2.1129e+6
	bias	-0.0199	0.0119	-0.0266	8.1943	1.1609	8.1878	3.0236	0.1581	3.1127
100	var	0.0475	0.0003	0.0490	1069.1	2.7878	1295	121.23	0.1231	1295
	mse	0.0479	0.0004	0.0497	1136.2	4.1353	1316.9	130.36	0.1481	1304.6889

Table (3) λ estimates λ=0.5 n=40 different σ and noise ratios

σ	μ/σ	$\hat{\lambda}_{BC}$			$\hat{\lambda}_{R}$			$\hat{\lambda}_{W}$		
		bias	var	mse	bias	var	mse	bias	var	mse
1	1	-0.0540	0.0321	0.0350	0.6492	0.0418	0.4632	-0.0256	0.0397	0.0404
	2	-0.0450	0.0673	0.0693	0.6384	0.0237	0.4279	-0.0257	0.0778	0.0784
	3	-0.0392	0.1131	0.1146	0.5724	0.0112	0.3388	-0.0273	0.1302	0.1309
	4	-0.0432	0.1639	0.1658	0.5117	0.0072	0.2690	-0.0264	0.1904	0.1910
	5	-0.0395	0.2229	0.2244	0.46508	0.0045	0.2208	-0.0246	0.2547	0.2553
2	1	-0.1151	0.0089	0.0223	0.1379	0.0119	0.0309	-0.0762	0.0149	0.0207
	2	-0.0537	0.0321	0.0350	0.1949	0.0059	0.0438	-0.0300	0.0405	0.0414
	3	-0.0417	0.0681	0.0699	0.1804	0.0033	0.0358	-0.0296	0.0805	0.0813
	4	-0.0437	0.1140	0.1159	0.1611	0.0021	0.0280	-0.0288	0.1302	0.1311
	5	-0.0461	0.1702	0.1723	0.1434	0.0014	0.0220	-0.0295	0.1908	0.1916
3	1	-0.1605	0.0059	0.0316	-0.0131	0.0059	0.0061	-0.1184	0.0112	0.0252
	2	-0.0618	0.0224	0.0262	0.0438	0.0029	0.0049	-0.0348	0.0300	0.0312
	3	-0.0455	0.0557	0.0578	0.0391	0.0017	0.0032	-0.0274	0.0661	0.0668
	4	-0.0430	0.0973	0.0991	0.0333	0.0012	0.0023	-0.0273	0.1114	0.1122
	5	-0.0432	0.1474	0.1492	0.0301	0.0009	0.0018	-0.0286	0.1674	0.1682
4	1	-0.1853	0.0052	0.0395	-0.0893	0.0040	0.0120	-0.1459	0.0099	0.0312
	2	-0.0711	0.0181	0.0231	-0.0370	0.0019	0.0033	-0.0411	0.0256	0.0273
	3	-0.0467	0.0490	0.0512	-0.0370	0.0011	0.0025	-0.0256	0.0580	0.0587
	4	-0.0427	0.0915	0.0933	-0.0335	0.009	0.0021	-0.0289	0.1016	0.1024
	5	-0.0348	0.1383	0.1395	-0.0258	0.0008	0.0015	-0.0287	0.1596	0.1604
5	1	-0.2022	0.0048	0.0457	-0.1350	0.0029	0.0211	-0.1639	0.0090	0.0359
	2	-0.0792	0.0155	0.0218	-0.0908	0.0013	0.0096	-0.0447	0.0235	0.0255
	3	-0.0451	0.0461	0.0481	-0.0853	0.0009	0.0082	-0.0325	0.0548	0.0559
	4	-0.0396	0.0850	0.0875	-0.0729	0.0009	0.0062	-0.0228	0.0990	0.0995
	5	-0.0409	0.1339	0.1355	-0.0554	0.0008	0.0039	-0.0921	0.1551	0.1559

Table (4) μ estimates, λ=0.5, n=40.

σ	μ/σ	$\hat{\mu}_{BC}$			$\hat{\mu}_{R}$			$\hat{\mu}_{W}$		
		bias	var	mse	bias	var	mse	bias	var	mse
1	1	-0.02994	0.0415	0.0424	0.7100	0.1534	0.6576	-0.0040	0.0494	0.0494
	2	-0.0101	0.2363	0.2364	1.7839	0.4017	3.5841	0.0378	0.3255	0.3269
	3	0.1360	1.5742	1.5925	3.0876	0.7756	10.308	0.2420	2.391	2.4491
	4	0.5204	7.7789	8.0489	4.4624	1.2683	21.181	0.8231	13.309	13.985
	5	0.8204	11.682	12.354	5.8685	1.6476	38.086	3.4	6514.9	6525.9
2	1	-0.2301	0.1415	0.1945	0.4595	0.3093	0.5205	-0.1232	0.2031	0.2182
	2	-0.1376	1.0942	1.113	1.4212	0.6314	2.6512	0.0569	1.8255	1.8286
	3	0.3452	10.42	10.538	2.424	1.0119	6.8878	0.6962	19.861	20.344
	4	1.7572	68.962	72.043	3.3525	1.4272	12.666	2.8626	260.65	268.82
	5	5.8406	672.21	706.26	4.1746	1.8189	19.246	8.636	2618.5	2692.8
3	1	-0.5476	0.2773	0.5771	0.0801	0.4255	0.4319	-0.3629	0.4568	0.5884
	2	-0.3810	2.6258	2.7707	0.5625	0.7665	1.0828	0.0265	5.0025	5.0026
	3	0.5422	28.924	29.215	0.8791	1.0988	1.8714	1.4217	69.189	71.204
	4	3.6613	266.14	279.51	1.1472	1.6192	2.9351	6.2167	1156.9	1195.4
	5	12.555	3212.4	33702	1.4344	2.3344	4.3915	26.15	96379	97053
4	1	-0.9445	0.4403	1.3325	-0.3842	0.5322	0.6798	-0.6827	0.7841	1.2502
	2	-0.7722	4.957	5.5529	-0.5818	0.8482	1.1866	-0.8546	9.7639	9.7702
	3	0.8342	63.213	63.902	-1.0523	1.2322	2.3394	2.3354	167.66	173.09
	4	6.6131	871.88	915.53	-1.4528	2.0768	4.1871	10.56	4071.9	4183
	5	25.118	13014	13644	-1.564	3.313	5.7041	40.175	1.3654e+5	1.3814e+5
5	1	-1.3785	0.6346	2.5348	0.9027	0.5956	1.4105	-1.0482	1.192	2.2905
	2	-1.2693	7.8739	9.4842	-1.9607	0.8609	4.7052	-0.1391	21.547	21.565
	3	1.3253	131.59	133.34	-3.2128	1.4091	11.731	3.1841	345.86	355.96
	4	9.8141	1.1745.8	1842	-4.1575	2.7474	20.032	17.577	10503	10811
	5	36.115	32502	33803	-4.4119	4.7644	24.229	92.119	2.0012e+6	2.0095e+6

Table (5) estimates for σ: λ=0.5, n=40.

σ	μ/σ	$\hat{\sigma}_{BC}$			$\hat{\sigma}_{R}$			$\hat{\sigma}_{W}$		
		bias	var	mse	bias	var	mse	bias	var	mse
1	1	-0.0607	0.0235	0.0272	0.6062	0.0464	0.1439	-0.0406	0.0494	0.0510
	2	-0.0321	0.1061	0.1071	1.2732	0.0918	1.7128	0.0007	0.3255	0.3255
	3	0.0679	0.4054	0.4099	1.6657	0.1210	0.8957	0.1280	2.3908	2.4071
	4	0.2552	1.3786	1.4437	1.8756	0.1228	3.6408	0.4054	13.309	13.4733
	5	0.7139	17.488	17.996	1.986	0.1016	4.0459	1.6733	6514.9	6517.6999
2	1	-0.3375	0.0469	0.1609	0.1645	0.0516	0.0787	-0.2623	0.2031	0.2719
	2	-0.1523	0.4122	0.4353	0.8173	0.0969	0.7649	-0.0207	1.8255	1.8259
	3	0.1724	2.3507	2.3802	1.0472	0.1004	1.1969	0.3753	19.861	20.0018
	4	0.8204	11.682	12.354	1.1087	0.0773	1.3066	1.3745	260.65	262.5392
	5	2.3717	104.55	110.17	1.109	0.0633	1.2933	3.6614	2618.5	2631.9058
3	1	-0.8673	0.0741	0.8262	-0.4447	0.0552	0.253	-0.7157	0.4568	0.9690
	2	-0.3620	0.8741	1.0051	0.1212	0.0945	0.1092	-0.0779	5.0024	5.0084
	3	0.2808	6.2922	6.3704	0.1822	0.0793	0.1125	0.7710	69.189	69.7834
	4	1.6682	45.201	47.979	0.1750	0.0608	0.0914	2.9493	1156.9	1165.5984
	5	4.9459	520.56	544.97	0.1750	0.0569	0.0875	11.632	96379	96514.303
4	1	-1.5175	0.1115	2.4143	-1.1799	0.0605	1.4526	-1.3099	0.7841	2.4999
	2	-0.6646	1.5108	1.9523	-0.7014	0.0889	0.5809	-0.2025	9.7639	9.8049
	3	0.4385	13.164	13.355	-0.7623	0.0651	0.6462	1.2447	167.66	169.2092
	4	2.9453	149.13	157.79	-0.7770	0.0600	0.6638	4.9239	4071.9	4096.1448
	5	9.7744	2090.6	2186	-0.7024	0.0649	0.5583	16.763	13.654e+5	13.65681e+5
5	1	-2.239	0.1686	5.1816	-1.9866	0.0647	4.0113	-1.9757	1.192	5.0954
	2	-1.0262	2.3015	3.3544	-1.6015	0.0865	2.6512	-0.3147	21.547	21.6460
	3	0.6827	26.735	27.199	-1.735	0.0588	3.0689	1.7315	345.86	348.8580
	4	4.2631	291.53	30.9.67	-1.7024	0.0669	2.9651	8.0847	10503	10568.362
	5	13.919	5360.3	5553.5	-1.5062	0.0846	2.3534	41.771	2.0012e+6	2.0029e+6

Table (6) 1- $\hat{\alpha}_n$ for the three test statistics with λ=1.

μ/σ	Test / Method of estimation	n = 20			n = 40			n = 100		
		A*	W	R	A*	W	R	A*	W	R
1	MLE_{BC}	97.77	98.78	99.60	97.40	97.72	99.61	93.90	85.63	98.88
	MLE_R	99.62	14.52	14.58	99.78	1.83	1.98	99.99	0.0	0.0
	W	98.79	99.29	99.67	98.38	97.97	99.62	96.06	86.67	99.20
	Original-data	94.54	95.06	94.98	94.59	95.40	94.25	94.71	90.03	95.07
2	MLE_{BC}	98.66	99.38	99.61	98.38	98.86	99.46	97.82	92.36	99.47
	MLE_R	96.56	79.77	80.27	97.04	64.41	64.74	95.95	29.59	33.35
	W	99.06	99.60	99.66	98.71	99.02	99.49	98.41	91.95	99.49
	Original-data	94.91	94.90	94.85	94.87	95.64	94.64	94.83	90.03	95.02
3	MLE_{BC}	98.75	99.42	99.45	98.62	99.24	99.15	98.36	89.49	99.10
	MLE_R	94.23	93.80	93.70	95.27	94.60	94.37	94.47	87.16	90.82
	W	98.78	99.48	99.47	98.86	99.27	99.21	98.67	89.79	99.15
	Original-data	94.32	94.56	94.77	94.64	95.30	94.47	95.29	90.03	95.36
4	MLE_{BC}	99.04	99.56	99.51	98.73	99.32	99.26	98.33	89.11	98.96
	MLE_R	94.34	94.74	94.55	94.92	95.91	95.31	94.76	90.61	95.20
	W	98.90	99.51	99.44	98.74	99.26	99.10	98.52	89.02	99.08
	Original-data	94.41	94.62	95.02	94.68	95.62	94.78	95.05	90.25	95.09
5	MLE_{BC}	98.70	99.56	99.58	98.49	99.27	99.24	98.61	89.05	99.07
	MLE_R	94.44	94.47	94.76	95.03	95.64	94.63	95.12	90.36	95.04
	W	98.99	99.54	99.54	98.90	97.91	99.61	98.48	88.72	98.89
	Original-data	94.69	95.15	94.97	94.63	95.49	94.67	94.71	89.81	95.12

Table (7) 1- $\hat{\alpha}_n$ for the three test statistics with λ=0.5.

μ/σ	Test Method of estimation	n = 20			n = 40			n = 100		
		A*	W	R	A*	W	R	A*	W	R
1	MLE_{BC}	97.44	98.67	99.42	95.87	96.45	99.07	89.16	79.43	97.29
	MLE_R	96.95	98.33	99.15	94.41	96.05	98.72	79.92	29.80	80.81
	W	98.83	99.50	99.81	98.09	97.91	99.61	94.90	82.80	98.69
	Original-data	31.40	24.33	32.30	4.09	1.09	2.87	0.0	0.0	0.0
2	MLE_{BC}	98.63	99.36	99.47	98.40	99.15	99.25	98.24	91.85	99.39
	MLE_R	95.80	96.83	97.15	96.53	97.49	97.77	92.01	88.58	93.56
	W	99.01	99.53	99.67	98.87	99.39	99.31	98.60	91.77	99.26
	Original-data	68.44	63.89	67.81	40.81	28.40	35.15	3.52	0.26	1.25
3	MLE_{BC}	98.67	99.35	99.56	98.58	99.18	99.30	98.44	89.49	99.22
	MLE_R	94.22	94.63	94.82	95.23	96.49	95.68	94.26	90.54	94.84
	W	99.05	99.64	99.55	98.71	99.30	99.21	98.35	88.37	98.78
	Original-data	81.56	78.89	80.53	66.07	59.34	61.28	27.76	15.09	19.98
4	MLE_{BC}	98.87	99.45	99.64	98.56	99.16	99.09	98.36	89.01	99.23
	MLE_R	94.24	94.41	94.66	94.98	95.94	95.41	94.43	90.13	94.97
	W	99.11	99.65	99.66	98.90	99.41	99.82	98.49	88.79	98.94
	Original-data	86.10	84.83	85.87	77.89	73.74	73.94	50.99	40.88	43.06
5	MLE_{BC}	98.72	99.43	99.46	98.38	99.09	99.05	98.49	88.98	99.10
	MLE_R	94.49	94.82	94.87	94.78	95.66	95.01	94.57	90.24	94.97
	W	98.93	99.48	99.42	99.02	99.39	99.30	98.62	88.80	98.97
	Original-data	89.64	88.61	89.07	83.17	80.80	80.26	64.72	58.08	58.06

Table (8) (1- $\hat{\alpha}_n$) for the three test statistics for positively skewed distributions

Distribution	γ_1	γ_2	Method of estimation	n = 20			n = 40			n = 100		
				A*	W	R	A*	W	R	A*	W	R
Weibull (3.3,5)	0.07791	2.7110	MLE_{BC}	98.68	99.38	99.77	98.27	98.63	99.59	97.82	91.33	99.64
			MLE_R	94.06	94.61	95.50	88.45	86.90	89.03	49.64	24.13	39.55
			W	99.14	99.61	99.78	98.65	98.85	99.61	97.85	91.57	99.65
			Original-data	94.81	95.78	96.87	94.65	95.30	96.92	94.16	88.71	97.46
Weibull (3.3,10)	0.07791	2.7110	MLE_{BC}	98.51	99.33	99.68	98.31	98.75	99.80	97.67	91.31	99.52
			MLE_R	95.19	96.38	97.21	95.69	96.19	97.85	91.75	85.08	94.76
			W	98.90	99.55	99.74	98.71	98.97	99.83	97.96	91.85	99.62
			Original-data	94.78	95.89	96.56	95.19	95.79	97.16	94.20	89.23	97.41
Gamma (16,5)	0.5	4.125	MLE_{BC}	98.73	99.50	99.38	98.70	99.21	99.31	98.47	89.35	99.08
			MLE_R	94.02	94.69	94.75	93.91	94.46	93.95	91.33	87.52	89.99
			W	98.93	99.58	99.49	98.98	99.38	99.35	98.63	89.22	98.95
			Original-data	89.32	88.54	88.89	85.20	83.32	82.01	68.50	62.89	61.97
Gamma (16,10)	0.5	4.125	MLE_{BC}	98.88	99.55	99.62	98.62	99.35	99.19	98.49	88.91	98.98
			MLE_R	94.75	95.31	95.41	94.32	94.77	94.05	92.61	88.39	92.16
			W	99.11	99.66	99.56	98.76	99.38	99.24	98.77	88.44	99.16
			Original-data	90.22	89.44	89.78	84.79	82.84	81.22	67.87	62.59	61.76
Weibull (2.1,5)	0.5672	3.1323	MLE_{BC}	98.67	99.32	99.56	98.19	98.68	99.66	97.43	91.34	99.67
			MLE_R	92.18	92.10	93.26	76.11	67.98	73.94	10.52	0.38	1.63
			W	99.04	99.48	99.62	98.66	98.92	99.67	98.11	91.34	99.66
			Original-data	88.44	87.20	83.84	79.97	74.49	78.43	49.09	27.71	42.61
Weibull (2.1,10)	0.5672	3.1323	MLE_{BC}	98.31	99.32	99.69	98.11	98.69	99.48	97.74	91.32	99.68
			MLE_R	95.12	96.28	97.25	93.36	93.49	95.03	70.81	49.12	68.43
			W	98.74	99.51	99.73	98.48	98.81	99.52	98.11	91.48	99.64
			Original-data	87.66	86.73	88.67	78.73	73.34	76.78	49.00	28.46	42.64

Table (8) Continued

Distribution	γ_1	γ_2	Method of estimation	n = 20			n = 40			n = 100		
				A*	W	R	A*	W	R	A*	W	R
Gamma (9,5)	0.6667	5	MLE_{BC}	98.67	99.39	99.40	98.44	99.22	99.22	98.47	89.23	99.25
			MLE_R	94.32	94.67	94.66	93.51	93.84	92.43	86.08	82.84	83.27
			W	98.84	99.50	99.47	98.60	99.20	99.30	98.77	89.09	99.08
			Original-data	85.93	84.35	85.04	76.18	72.64	71.88	46.79	38.29	38.74
Gamma (9,10)	0.6667	5	MLE_{BC}	98.66	99.41	99.50	98.79	99.37	99.25	98.52	89.54	99.13
			MLE_R	94.54	95.01	95.12	94.60	95.01	94.03	90.29	86.20	89.48
			W	99.04	99.53	99.66	98.70	99.37	99.24	98.58	89.38	99.02
			Original-data	85.84	84.58	85.44	75.79	72.48	71.47	46.29	38.65	39.39
Gamma (4,5)	1	7.5	MLE_{BC}	98.65	99.42	99.56	98.55	99.16	99.34	98.60	89.76	99.26
			MLE_R	93.70	94.52	94.96	90.37	89.40	88.63	65.52	57.11	58.10
			W	98.88	99.51	99.56	98.92	99.32	99.47	98.80	89.85	99.33
			Original-data	73.44	70.12	72.54	52.29	43.96	47.72	10.62	4.49	6.69
Gamma (4,10)	1	7.5	MLE_{BC}	98.69	99.40	99.50	98.65	99.19	99.32	98.67	89.37	99.36
			MLE_R	94.74	95.54	96.05	93.53	93.83	93.52	82.59	78.94	79.90
			W	98.92	99.59	99.57	98.78	99.14	99.18	98.69	88.82	99.28
			Original-data	73.87	70.21	72.48	51.44	42.99	46.02	10.73	4.69	6.56
Lognormal	6.185	113.936	MLE_{BC}	98.84	99.52	99.57	98.84	99.31	99.23	98.40	88.79	98.91
			MLE_R	87.41	86.34	88.89	80.64	77.42	77.20	74.98	70.07	70.36
			W	99.06	99.60	99.58	98.94	99.38	99.24	98.54	88.77	98.87
			Original-data	8.91	6.35	8.54	0.21	0.05	0.14	0.00	0.00	0.00

Table (9) (1- $\hat{\alpha}_n$) for the three test statistics for the negatively skewed Wiebull distribution.

Criterion / Method of estimation		A*	W	R	Distribution
MLEBC	n=20	98.54	99.27	99.75	Weibull (4.5,5)
MLER		94.98	96.02	96.93	
W		98.89	99.49	99.76	
Original-data		93.92	94.75	95.45	
MLEBC	n=40	98.26	98.71	99.60	
MLER		93.67	93.73	95.40	
W		98.62	98.80	99.60	
Original-data		93.18	93.94	95.15	
MLEBC	n=100	97.57	91.66	99.64	
MLER		80.21	65.59	80.49	
W		98.16	91.43	99.67	
Original-data		90.00	85.54	94.14	
MLEBC	n=20	98.54	99.27	99.67	Weibull (4.5,10)
MLER		94.35	95.20	96.20	
W		98.98	99.38	99.68	
Original-data		94.32	95.05	96.08	
MLEBC	n=40	98.35	98.69	99.60	
MLER		94.91	95.59	96.93	
W		98.78	98.88	99.63	
Original-data		93.52	93.99	95.10	
MLEBC	n=100	97.56	90.72	99.55	
MLER		94.92	90.07	98.10	
W		98.09	91.49	99.62	
Original-data		90.08	85.39	93.97	

Table (10) (1- $\hat{\alpha}_n$) for the three test statistics for the contaminated normal distribution.

Criterion / Method of estimation		A*	W	R
MLE_{BC}	n=20	94.92	96.65	93.79
MLE_R		75.28	71.33	68.01
W		95.23	96.71	94.12
Original-data		74.72	70.84	67.36
MLE_{BC}	n=40	84.33	89.27	77.55
MLE_R		62.17	58.41	46.65
W		84.23	89.29	77.51
Original-data		60.91	57.45	45.92
MLE_{BC}	n=100	52.61	62.13	37.25
MLE_R		35.18	39.44	17.06
W		52.71	61.62	36.89
Original-data		33.19	37.77	15.96

REFERENCES

Anderson, T.W. and Darling, D.A. (1954). A Test of Goodness of Fit. Journal of the American Statistical Association, 49, 765-769.

Bickel, P.J. and Doksum, K.A. (1981). An Analysis of Transformations Revisited. Journal of the American Statistical Association, 76,296-311**.**

Blom, G. (1958). Statistical Estimates and Transformed Beta- Variables. Almqvist and Wiksell, Uppsala, Sweden; Wiley, New York.

Box, G.E.P. and Cox, D.R. (1964). Analysis of Transformation. Journal of the Royal Statistical Society, Ser. B, 26, 211-252**.**

Carroll, R. J. (1980). A robust method for testing transformations to achieve approximate normality. Journal of the Royal Statistical Society, Series B, *42*, 71-78.

Carroll, R. J. (1982). Power transformations when the choice of power is restricted to a finite set. Journal of the American Statistical Association, *77*, 908-915.

Carroll, R. J. and Ruppert, D. (1985). Transformations: a robust analysis. Technometrics, *27*, 1-12.

Cho, K ., Yeo, I. ,Johnson, R. and Loh, W. (2001). Asymptotic theory for Box-Cox transformations in linear models. Statist. And Probab. Letters, 51, 337-343.

D'Agostino, R.B. and Stephen, M.A. (1986). Goodness-of-Fit Techniques. Marcel Dekker, New York**.**

David, H. A. and Nagaraja, H. (2003). Order Statistics. 3nd. ed, John Wiley, New York**.**

Filliben, J.J. (1975). The Probability Plot Correlation Coefficient Test for Normality. Technometrics, 17, 111 - 117**.**

Gaudar, M. and Karson, M (2000). On Estimating the Box-Cox Transformation To Normality. Communications in Statistics - Simulation and Computation, 29 (2), 559 - 582**.**

Halawa, A. M. (1996). Estimating the Box-Cox Transformation via Artificial Regression Model. Communications in Statistics-Simulation and Computation, 25 (2), 331 - 350**.**

Harter, H.L. (1961). Expected Values of Normal Order Statistics. Biometrika, 48, 151 -165**.**

Kariya, T. and Kuata, H (2004). Generalized Least Squares. John Wiley, New York.

Kouider, E. and Chen, H. (1995). Concavity of Box-Cox Log-likelihood Function. Statistics & Pribability Letters , 25 , 171-175.

Lin, L.I. and Vonesh, E.F. (1989). An Empirical Nonlinear Data-Fitting Approach for Transforming Data to Normality. American Statistician. 43, 237-243.

Lloyd, E. H. (1952). Least-squares estimation of location and scale parameters using order statistics. Biometrika 39, 88-95.

Rahman, M. (1999). Estimating the Box-Cox Transformation via Shapiro-Wilk W Statistic. Communications in Statistics-Simulation and Computation, 28 (1), 223 -241**.**

Royston, J.P. (1982). An Extension of Shapiro and Wilk's Test for Normality to Large Samples. Applied Statistics, 31, 115 - 124**.**

Sarhan, A. and Greenberg, B. (1956). Estimation of the location and scale parameters by order statistics from singly and doublely censored samples, Part I, The normal distribution up to samples of size 10. Ann. Math. Statist.,27, 427-451.

Sarhan, A. and Greenberg, B. (1958). Estimation of the location and scale parameters by order statistics from singly and doublely censored samples, Part II. Ann. Math. Statist.,29, 79-105.

Shapiro, S.S. and Wilk, M.B. (1965). An Analysis of Variance Test for Normality (Complete Samples). Biometrika, 52, 591 - 611**.**

Spinelli, J. (1980). Contributions to goodness-of-fit. M.Sc. Thesis, Department of Mathematics, Simon Fraser University.

Sun, G. (1978). Study on the Box and Cox Power Transformation to Normality. Ph. D. thesis, Oklahoma State University.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 733–746.

A NEW ALGORITHM FOR COMPUTING THE MOMENTS AND PRODUCT MOMENTS OF ORDER STATISTICS IN CONTINUOUS DISTRIBUTIONS

Osama Abdelaziz Hussien
Department of Statistics, Faculty of Commerce,
Alexandria University, Alexandria, Egypt
E-mail: osama52@gmail.com, ossama.abdelaziz@alex-commerce.edu.eg

ABSTRACT

In this article, we present a new algorithm to compute the lower moments, product moments, variances and covariances of order statistics in continuous distributions, and for any sample size. The algorithm is written using the GAUSS Mathematical and Statistical System matrix programming language. The accuracy of the calculations was tested for the normal, the uniform and the exponential distributions. As an application of the algorithm, exact computation of the population L-moments and the TL-moments will be given for all distribution studied.

Keywords: Order statistics, L-moments, Simpson's algorithm, Gauss-Legender quadrature algorithm, series approximations, recurrence relations

1. INTRODUCTION

Let X_1 , X_2 ,…, X_n be a random sample from an absolutely continuous distribution with cumulative distribution function (cdf) F(x) and probability density function (pdf) f(x), and let $X_{(1)}$, $X_{(2)}$, …,$X_{(n)}$ denote the order statistics obtained from this sample. The k^{th} moments of the order statistic $X_{(r)}$ is given by

$$E\left(X_{(r)}^{k}\right) = \frac{n!}{(r-1)!(n-r)!}\int_{-\infty}^{\infty} y^{k}\,[F_X(y)]^{r-1}[1-F_X(y)]^{n-r} f_X(y)dy \tag{1}$$

Sen (1959) has shown that, if $E|X|^{\delta}$ exists for some $\delta > 0$, then $E(X_{(r)}^{k})$ exists for all r satisfying $r_o < r < n - r_o + 1$, where $r_o\delta = k$. The product moments between any two order statistics $X_{(r)}$ and $X_{(s)}$, r < s; r, s=1,2,…,n, will be defined by $E\left(X_{(r)}X_{(s)}\right) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!}$.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{y} xy\,[F_X(x)]^{r-1}[F_X(y)-F_X(x)]^{s-r-1}[1-F_X(y)]^{n-s} f_X(x)f_X(y)dxdy \tag{2}$$

The covariance of $X_{(r)}$ and $X_{(s)}$ r, s=1,2,…,n, will be given by

$$Cov\left(X_{(r)},X_{(s)}\right) = E\left(X_{(r)}X_{(s)}\right) - E\left(X_{(r)}\right)E\left(X_{(s)}\right) \tag{3}$$

If the population distribution is symmetric about 0, then

$$E\left(X_{(r)}\right) = -E\left(X_{(n-r+1)}\right) \tag{4}$$
$$E\left(X_{(r)}X_{(s)}\right) = E\left(X_{(n-r+1)}X_{(n-s+1)}\right) \tag{5}$$
$$Cov\left(X_{(r)},X_{(s)}\right) = Cov\left(X_{(n-r+1)},X_{(n-s+1)}\right) \tag{6}$$

If the population distribution is symmetric about 0 and *n* is odd, then

$$E(X_{(n+1)}/2 = 0 \tag{7}$$

The moments of order statistics in random samples of small size n can be obtained explicitly only for a few simple populations, such as the uniform and the exponential. Tables of means, variances, and covariances are available in the literature for some standard distributions (Harter and Balakrishnan, 1996). The tables use different computation methods and some series approximations and they are available only for small sample sizes. We think that statisticians, somehow still believe on what Hirakawa claims in 1973 "moments of order statistics can be evaluated by numerical integration. *But straightforward integration has shortcomings in view of quantity of computation and the accuracy*".

Recurrence relations between the moments of order statistics have been studied extensively with the principal aim of reducing the number of independent calculations required for the evaluation of the moments. Such relations may also be used as partial checks on direct calculations of the moments.

Davis and Stephens (1978) and Royston (1982), present algorithms to compute the moments and product moments of order statistics for small samples from the normal distribution. The Maple procedures presented by Childs and Balakrishnan (2002) utilizes the series approximations presented by David and Johnson (1954) to approximate the moments and product moments of order statistics for some continuous distributions.

This paper presents a new algorithm for the computation of the moments, product moments, variances and covariances of order statistics in continuous distributions and for any sample sizes. The algorithm is written using the GAUSS Mathematical and Statistical System matrix programming language. The powerful numerical integration procedures in GAUSS plus the procedures for computing the cdf and the pdf for many distributions make this task attainable.

The rest of this paper is organizes as follows. In Section 2 we review the basic computation methods used to compute the order statistics in the literature. Section 3 describes the proposed algorithm and some formulae that can be used for checking its accuracy. We compare the results of the proposed algorithm with the available table-values and approximations for the moments of order statistics from the normal distribution in Section 4. In Section 5 computation of the population L-moments and the TL-moments using the proposed algorithm will be given for all distribution studied. Finally, conclusions and remarks for further studies are presented in Section 6.

2. COMPUTATION METHODS

There has been a large amount of work relating to moments and product moments of order statistics. See David and Nagaraja (2003) and Arnold et al. (1992). We review in this section some of the basic approaches used in the literature.

2.1 Explicit Forms

Explicit forms exist only for a few simple distributions, such as the uniform and the exponential. The explicit form for the logistic case (see Gupta and Shah ,1965) depends on an asymptotic calculation of the digamma and trigamma functions. Calculation of the explicit expressions for the first moment for the extreme value distribution encounter numerical problems due to rounding errors in the evaluation even for moderate sample size (Fard and Holmquist,2007). Nadarajah (2008) derived explicit forms for the moments of order statistics from the normal,

lognormal, gamma and beta distributions. His expressions take the form of a finite sum of a special function that can be computed through numerical routines avaliable for Mathematica.

2.2 Numerical Tables

Early tables for the moments of order statistics are computed using numerical integration. The accuracy of the "good old tables" is high, at least for the expected values, as we will see in the next section. However, the existing tables are only done to a few distributions and for small samples only. For the standard normal the order statistics' means have been tabulated extensively in (Harter, 1961). Variances and covariances for n < 20, are tabulated in (Teichroew,1956) ,and (Sarhan and Greenberg ,1956) and for n ≤ 50 , in (Tietjen et al., 1977).

2.3 Recurrence Relations

Many authors have studied recurrence relations between the moments of order statistics. Their aim often was to reduce the number of independent calculations required for the evaluation of the moments. Numerous recurrence relations have been developed for specific distributions. Such relations may also be used as partial checks to direct calculations of the moments. The following general relations, given by David and Nagaraja (2003) will be used to check the accuracy of the new proposed algorithm.

$$\sum_{r=1}^{n} E(X_{(r)}) = n\mu \tag{8}$$

$$\sum_{r=1}^{n} E(X_{(r)}^{2}) = nE(X^{2}) \tag{9}$$

$$\sum_{r=1}^{n}\sum_{s=1}^{n} \mathrm{E}(X_{(r)}X_{(s)}) = nE(X^{2}) + n(n-1)\mu^{2} \tag{10}$$

$$\sum_{r=1}^{n}\sum_{s=r+1}^{n} \mathrm{E}(X_{(r)}X_{(s)}) = \frac{1}{2}n(n-1)\mu^{2} \tag{11}$$

$$\sum_{r=1}^{n}\sum_{s=1}^{n} \mathrm{cov}(X_{(r)}X_{(s)}) = n\sigma^{2} \tag{12}$$

2.4 Approximations in Terms of the Quantile Function and its Derivatives

Quantile function based approximations are presented by David and Johnson (1954). They present results to order $(n + 2)^{-3}$ for all the first four cumulants and cross cumulants. By the probability integral transformation $U_{(r)} = F(X_{(r)})$, were $U_{(r)}$ is the rth order statistics in a sample of n from a uniform (0,1) distribution. Writing $X_{(r)} = Q(U_{(r)})$, where $Q = F^{-1}$, and $U_{(r)} = F(X_{(r)})$, is the uniform (0,1) rth order statistics then expanding $Q(U_{(r)})$ in a Taylor series about $E(U_{(r)}) = r/(n+1) \equiv p_r$ one gets

$$X_{(r)} = Q(p_r) + (U_{(r)}-p_r)Q'(p_r) + 0.5(U_{(r)}-p_r)^2 Q''(p_r) + (1/6)(U_{(r)}-p_r)^3 Q'''(p_r) + \ldots \tag{13}$$

Replacing $Q(p_r)$ by Q_r and setting $q_r=1-p_r$ we obtain to order $(n+2)^{-2}$

$$E(X_{(r)}) = Q_r + \frac{p_r q_r}{2(n+2)} Q_r'' + \frac{p_r q_r}{(n+2)^2}\left[\frac{1}{3}(q_r - p_r)Q_r''' + \frac{1}{8} p_r q_r Q_r''''\right] \quad (14)$$

$$\text{var}(X_{(r)}) = \frac{p_r q_r}{(n+2)} Q_r'^2 + \frac{p_r q_r}{(n+2)^2}\left[2(q_r - p_r)Q_r' Q_r'' + p_r q_r (Q_r' Q_r''' + \frac{1}{2} Q_r''^2)\right] \quad (15)$$

$$\text{cov}(X_{(r)}, X_{(s)}) = \frac{p_r q_s}{(n+2)} Q_r' Q_s' + \frac{p_r q_s}{(n+2)^2}\left[(q_r - p_r)Q_s' Q_r'' + (q_s - p_s)Q_r' Q_s'' + \right.$$

$$\left. \frac{1}{2} p_r q_r Q_s' Q_r''' + \frac{1}{2} p_s q_s Q_r' Q_s''' + \frac{1}{2} p_r q_s Q_r'' Q_s''\right] \quad (16)$$

2.5 Algorithms

Davis and Stephens (1978) present a FORTRAN procedure to approximate the covariance matrix of order statistics from a normal distribution. Their procedure requires the user to supply values of $\text{Var}(X_{(n)})$. Royston (1982), present a FORTRAN procedure to calculate the first moment of order statistics from a normal distribution. The calculation error is 0.001 for $n < 50$.

Next we write a procedure in GAUSS to compute the approximations given by (14)-(16) for the normal distribution. This procedure utilizes numerical integration procedures, can be applied to many continuous distributions and perform better than the procedures depend on the quantile approximation.

3. THE PROPOSED NUMERICAL INTEGRATION ALGORITHM

The proposed algorithm computes the $k^{\underline{th}}$ moment, the product moments, variances and covariances of order statistics from a continuous distribution with a very high accuracy. This level of accuracy will be at least as good as the available tabulated values, if exist even in large sample size. For most values of n, r and s one cannot obtain the required level of accuracy in computation of (1) or (2) unless F(x) and f(x) are evaluated with a high level of accuracy. The evaluation of F(x) for many distributions depends on some mathematical approximation. The major mathematical programming languages provide mathematical approximation procedures to compute F(x) for many distributions. Ideally, such procedure will provide accuracy essentially to "machine precision". This means that the combined effect of error introduced by the use of the mathematical approximation and rounding error in computation is negligible. The proposed algorithm uses those procedures many times to evaluate the numerical integrations involved in calculating….. Thus, the overall level of accuracy is very difficult to achieve, (see Kennedy and Gentle, 1980). Most of the advanced mathematical programming languages have high accuracy procedures to compute F(x) for many distributions and very high accuracy procedures for numerical integrations. This makes it possible to produce a high overall accuracy algorithm to compute (1)-(3).

Reasoning for Using the programming language GAUSS in the Algorithm

The GAUSS Mathematical and Statistical System is a fast matrix programming language designed for computationally intensive tasks. In a comparison of major mathematical program: Mathematica, Matlab, Maple, and Gauss, Steinhaus (2008) ranked Gauss second only to mathematica in the comparison of mathematical functions included numerical approximations, distribution functions and basic mathematical functions. Gauss includes high level accuracy procedures to compute F(x) for many distributions. Powerful and flexible procedures for double numerical integration are also available with very high accuracy. The matrix programming makes many programming tasks easy and fast.

3.1 The Steps of the Algorithm

3.1.1 Specify the Sample Size

For large samples, n > 169, $\binom{n}{r}$ cannot be computed directly. An asymptotic expansion of ln(n!) called the Stirling's series will be used for n > 169. It is given by

$$\boldsymbol{ln\, n!} = \boldsymbol{nln(n) - n + \frac{1}{2} ln(2\pi n) + \frac{1}{(2^2.3^1)n} - \frac{1}{(2^2.3^2.5^1)n^3} + \frac{1}{(2^2.3^2.5^1.7^1)n^5} - \frac{1}{(2^4.3^1.5^1.7^1)n^7} + \cdots}$$

The error in truncating the series is always of the same sign and at most the same magnitude as the first omitted term, (see Abramowitz and Stegun, 1965).

3.1.2 Specify the Probability Distribution

The available distributions so far are: standard normal, student's t, beta, gamma, extreme value, uniform, exponential and the logistic. If the distribution is symmetric (about 0) and n is odd the calculation of the moments will be simplified because of (3) and (4). The inverse Gaussian, Pareto, and the generalized lambda distribution will be added.

3.1.3 Specify the Higher Moments (Optional)

The default of the algorithm is to compute the expected value and the covariance matrix of order statistics. The user should specify a value of k > 2 to evaluate a higher moments.

3.1.4 Specify the Accuracy Level for Numerical Integrations (Optional)

Any numerical integration procedure approximates $\int_a^b g(x)dx$ where the limits of integration a and b are finite. This is not true for (1) and (2) for most of the distribution. So we set a = $F^{-1}(p_1)$ and b=$F^{-1}(p_2)$. Typically, p_1 =0.000001 =1-p_2. The user can change those limits to get different level of accuracy. To compute $E(X_{(r)})$ and $Var(X_{(r)})$, r= 1,2,…,n we uses the adaptive Simpson's algorithm for numerical integration (*intsimp* procedure of Gauss 9.0) with tolerance limits 0.00000001. In the adaptive process, we divide the interval [a,b] into two subintervals and then decide whether each of them is to be divided into more subintervals. The procedure is continued until some specified accuracy is obtained throughout the entire interval [a,b]. The Gauss-

Legender quadrature algorithm (*intquad1* procedure of Gauss 9.0) allows accurate and fast integration for many functions, but it is not adaptive to yield any level of accuracy as the *intsimp* procedure, (see Gauss 9.0 Language Reference, (2007)).

For the product moments the double integration limits will be $\int_a^b \int_a^y g(x,y)dxdy$ with the same definition of a and b as before. The Gauss-Legendre quadrature algorithm for double numerical integration (*intgrat2* procedure of Gauss 9.0) will be used. This procedure allows integrating over a region which is bounded by functions, rather than just scalars as the *intquad2* procedure.

3.1.5 Recurrence Relations Check (Optional)

The algorithm may check the accuracy of the general recurrence relations given by equations (6)–(10) above. Other recurrence relations checks for some specific distributions will be added to the algorithm. This will be useful when no tabulated values are available, e.g. the beta distribution and the student's t distribution.

3.1.6 Output

a-The vector of expected values $E(X_{(r)})$, r= 1,2,…,n.
b-The covariance matrix of order statistics $Cov(X_{(r)}, X_{(s)})$; r, s = 1,2,…,n.
c- Higher order moments (optional).
d- A check for the accuracy by recurrence relations (optional).

3.2 Time and Accuracy

The execution time is directly proportional to the sample size. The following table represents the execution time for the normal distribution.

N=50	N=100	N=500	N=1000
7 :27 seconds	**31:28 seconds**	**14:17:1 minutes**	**21:47:3 minutes**

The approximate series algorithm is much faster than the new algorithm specially for large samples, but it is much less accurate. The new algorithm is accurate to at least 15 decimal places. The accuracy is always higher for the expected value than for the covariance matrix. For very large sample sizes (n>1000 in most of the cases) one may have to sacrifies high accuracy of the numerical integration by reducing tolerance limits or limits of integration for the convergance of the integration. A detailed study of the accuracy will be given for the standard normal distribution only.

3.3 Exponential and Uniform Checks

It is well known that the moments of order statistics for the exponential and the uniform distributions have explicit forms. So, we can verify the accuracy of the algorithm by comparing the results of the algorithm to the exact values for those distributions. For the standard exponential distribution with pdf $f(x) = \exp(-x)$, $x > 0$, the moments of order statistics are given by

$$E(X_{(r)}) = \sum_{i=1}^{r} \frac{1}{(n-i+1)} \quad (17)$$

$$var(X_{(r)}) = \sum_{i=1}^{r} \frac{1}{(n-i+1)^2} \quad (18)$$

$$cov(X_{(r)}, X_{(s)}) = var(X_{(r)}) \qquad s > r \quad (19)$$

For the uniform distribution U(0,1) the moments of order statistics are given by

$$E(X_{(r)}) = \frac{r}{n+1} \equiv p_r \quad (20)$$

$$var(X_{(r)}) = \frac{p_r q_r}{n+2} \quad (21)$$

where $p_r = \frac{r}{n+1}$, and $q_r = 1 - p_r$. Define $\mu_r^* = E(X_{(r)})$ calculated by the algorithm, r=1,2,…,n.

$\mu_r = E(X_{(r)})$ calculated by the explicit form, (17) or (20), r=1,2,…,n.

$\sigma_r^{2*} = var(X_{(r)})$ calculated by the algorithm, r=1,2,…,n.

$\sigma_r^2 = var(X_{(r)})$ calculated by the explicit form,(18) or (21), r=1,2,…,n.

$$\text{MAXDmue} = \max_r \left| \mu_r^* - \mu_r \right| \qquad \text{MAXDsgma} = \max_r \left| \sigma_r^{2*} - \sigma_r^2 \right|$$

Table (1) shows the difference between the new algorithm calculation with the exact moments for the exponential distribution for n =2(2)8. For n=100, MAXDmue is 4.96565E-05 at r=98. While, the MAXDsgma is 4.83715E-05 at r=98. Table (2) calculates MAXDmue and MAXDsgma for the uniform distribution at selected sample sizes. One may conclude that the algorithm computation for the uniform and the exponential moments are accurate , except for a very small rounding error.

4. ACCURACY OF COMPUTATION FOR THE NORMAL DISTRIBUTION

We limit the accuracy of computation of the moments to the case of the standard normal distribution. The accuracy for the logistic, gamma, beta, extreme value and the student's t distributions will be presented in a forthcoming article, The existing tables for the above distributions either do not exist or incomplete (for the expected values only or for small sample sizes only).

4.1 Comparison with the Tabulated Values and the Series Approximation

For the tabulated expected values we used the tables of Harter (1961). The Harter's tables have 5 decimal places accuracy for all r and for selected sample sizes ≤400. The tabulated variances and covariances of order statistics of sample sizes up to 20, to 10 decimal places, were given by Sarhan and Greenberg (1956). Covariances for n ≤ 50 to 10 decimal places are given by Tietjen et al. (1977). Parrish (1992) presents tables of the variances and covariances to 10 decimal places for n ≤50. Tiechroew (1956) calculated the f(x) and F(x) for the standard normal distribution for x=-12(0.02)12. We use the same values -12 and 12 as limits for the numerical integration for the expected values and the variances. We, also use the same limits, -12 and 12, for the numerical double integration to compute the covariances. The available tables for the covariances seem to use smaller limits. See Table (5) below.

The series approximation of David and Johnson (1954) can be computed for any sample size. We wrote a procedure in GAUSS to compute the series approximations, given by (13)-(16), for the standard normal distribution. This approximation procedure is very fast and can be applied to any sample size.

Table (1): Differences between the moments calculated by the algorithm and the exact moments for the exponential distribution. n=2(2)8.

n	*i*	mue	difference	variance	difference
2	1	0.500000031	3.07429E-08	0.249999949	-5.12004E-08
2	2	1.499999828	-1.71516E-07	1.249998699	-1.30084E-06
4	1	0.250000021	2.07088E-08	0.062499979	2.45628E-07
4	2	0.583333303	-3.0227E-08	0.17361115	1.50049E-07
4	3	1.083333336	5.77447E-09	0.423611104	1.04223E-07
4	4	2.083333132	-1.97594E-07	1.423608768	-2.2323E-06
6	1	0.16666666	-7.11021E-09	0.027777765	-1.30238E-08
6	2	0.36666659	-7.68074E-08	0.067777834	5.66294E-08
6	3	0.616666667	5.00215E-10	0.130277777	-8.25173E-10
6	4	0.949999988	-1.18248E-08	0.241389032	1.42918E-07
6	5	1.450000065	6.52578E-08	0.491388709	-1.79483E-07
6	6	2.449999742	-2.57948E-07	1.491384686	-4.20264E-06
8	1	0.125000001	6.93517E-10	0.015624995	-5.26161E-09
8	2	0.267857107	-3.56116E-08	0.036033183	2.01492E-08
8	3	0.434523812	2.79872E-09	0.063810938	-3.24522E-09
8	4	0.634523808	-1.26048E-09	0.103810953	1.21252E-08
8	5	0.88452382	1.08188E-08	0.166310922	-1.85654E-08
8	6	1.217857143	3.20937E-10	0.277422045	-7.01215E-09
8	7	1.71785714	-2.52425E-09	0.52742206	7.79245E-09
8	8	2.717856796	-3.46394E-07	1.527416632	-5.4204E-06

Table (2): The maximum absolute differences between the new algorithm calculations with the exact moments for the uniform distribution.

Sample size	MAXDmue	MAXDsgma
10	2.86139090199811e-009	4.32870374007105e-009
50	1.93867591202768e-007	1.72401465628426e-007
100	2.16414286491329e-007	4.02522250497446e-007
500	2.10522026498706e-005	2.17208716083794e-005

Define $\mu_r^T = E(X_{(r)})$ calculated from Harter's tables r=1,2,…,n.

$\mu_r^A = E(X_{(r)})$ calculated by David Johnson approximation, r=1,2,…,n.

MAXDm1=$\max_r|\mu_r^* - \mu_r^T|$

MAXDm2 = $\max_r|\mu_r^* - \mu_r^A|$

Table (3) computes MAXDm1 and MAXDm2 for selected sample sizes. This shows that the algorithm computations for the expected values are very close to the tabulated values. This means that it is more accurate than the approximation procedure of David and Johnson (1954).

Table (3): The maximum absolute differences between the new algorithm calculations and the tabulated and approximated means for the normal distribution.

Sample size	MAXDm1	MAXDm2
10	3.33203E-10 (r=2)	0.0007881025758 (r=2)
20	-8.9508E-10 (r=10)	0.002518972095 (r=1)
50	5.07898254498151e-006 (r=6)	0.00334367073 (r=1)
100	5.1140069301514e-006 (r=7)	0.003416387866 (r=1)

Define $\sigma_r^{2T} = var(X_{(r)})$ calculated from the tables r=1,2,…,n.

$\sigma_r^{2A} = var(X_{(r)})$ calculated by David Johnson approximation, r=1,2,…,n.

MAXDs1=$\max_r|\sigma_r^{2*} - \sigma_r^{2T}|$ MAXDs2 = $\max_r|\sigma_r^{2*} - \sigma_r^{2A}|$

Table (4) computes MAXDs1 and MAXDs2 for selected sample sizes. The algorithm computations of the variances are very close to the tabulated values. Hence, it is more accurate than the approximation procedure of David and Johnson (1954).

Table (4). The maximum absolute differences between the new algorithm calculations and the tabulated and approximated variances for the normal distribution.

Sample size	MAXDs1	MAXDs2
10	1.9412E-11 (r=2)	0.01214077653 (r=1)
20	9.477E-10 (r=2)	0.006664613847 (r=1)
50	Not computed	0.004339938323 (r=1)
100	Tabulated values not available.	0.003873108374 (r=1)

Let COV($X_{(}$) denote the covariance matrix of order statistics. The elements of COV(X) are defined by (3). Define

V^* =COV($X_{(}$) calculated by the algorithm,

V^A= COV($X_{(}$) calculated by David Johnson approximation,

V^T=COV($X_{(}$) calculated by the tables,

MAXV1=$\max_{(r,s):r\neq s}|V^* - V^A|$ MAXV2 = $\max_{(r,s):r\neq s}|V^* - V^T|$

Table (5) computes MAXV1 and MAXV2 for selected sample sizes. It shows that the tabulated covariances use small limits for double numerical integration (-6, 6). The algorithm uses more accurate limits (-12, 12). Moreover, the algorithm uses 15 decimal places. Note that max|V^T-V^A|= 0.854803190653206 for n=20. We can conclude that the algorithm gives more accurate computations for the means, variances and covariances of order statistics for the normal distribution.

Table (5) MAXV1 and MAXV2 for selected sample sizes

Sample size	limits of integration	MAXV2	MAXV1
10	(-12,12)	0.0885097147486911 (7,1)	0.08794242992 (7,1)
	(-6,6)	0.000290172417513175 (10,1)	
20	(-12,12)	0.971226179420447 (6,2)	0.3759743329 (8,1)
	(-6,6)	0.854814523292793 (6,16)	
50	Not computed.		1.519450143 (5,1)
100	Tabulated values not available.		3.694929265 (9,1)

4.2 The Recurrence Relations

The basic moments identities (8)-(12) will be used as partial checks on the algorithm calculations of the moments at least for moderate sample sizes. For very large sample size possible accumulation of rounding error could arise in the computation of the relations. For the standard normal distribution (8)-(12) will be reduced to

$$\sum_{r=1}^{n} E(X_{(r)}) = 0$$

$$\sum_{r=1}^{n}\sum_{s=1}^{n} \operatorname{cov}(X_{(r)}, X_{(s)}) = n$$

Table (6) computes the above relations for selected sample sizes. This shows that the identities are satisfied for all sample sizes. Tabulated values and series approximation are available for the gamma, logistic and extreme value distribution. So, the same checks for accuracy of the algorithm computations can be applied. Accuracy verifications for other distributions will be conducted in a forthcoming research paper.

Table (6) Recurrence relations check

n	$\sum_{r=1}^{n} E(X_{(r)})$	$\sum_{r=1}^{n}\sum_{s=1}^{n} \operatorname{cov}(X_{(r)}X_{(s)}) - n$
10	4.44089209850063e-016	2.1387336346379e-012
20	-1.77635683940025e-015	-7.06386060755904e-011
50	-6.21724893790088e-015	-3.25504601050852e-009
100	4.88498130835069e-015	1.84653003998392e-009
500	9.10382880192628e-016	27.8387111808675

5. APPLICATIONS

If $X_1, X_2, \ldots, X_n$ are iid with density f(x), then the order statistics are complete and sufficient, see Lehmann and Casella (1998) pages 36 and 72. Thus many minimum variance unbiased

estimators (UMVUEs) and powerful test procedures for the unknown parameters were constructed based on order statistics. For the location-scale family of absolutely continuous distributions $F(x) = G((x-\mu)/\sigma)$ ($\sigma>0$) Lloyd (1952) proved that the weighted least squares estimators for μ and σ based on expected values and covariance matrix of the order statistics are Best Linear Unbiased Estimators (BLUEs). Lloyd's procedure requires full knowledge of the expectations and the covariance matrix of the order statistics. David and Nagaraja (2003) page 198 stated that "*The covariances especially may be difficult to determine*". This shows the important of the new algorithm since Lloyed's estimators are applied extensively for many distributions from complete and censored samples.

For the location-scale family a class of regression goodness of fit tests is based on the model

$$X_{(i)} = \mu + \sigma E(X_{(i)}) + \in_i , \ i = 1,2,\ldots, n \tag{22}$$

D'Agostino and Stephens (1986), page197, stated that "The values of $E(X_{(i)})$ are the most natural values to plot along the horizontal axis (in a p-p plot), *but for most distributions they are difficult to calculate".* In a forthcoming paper we will study the effect of our calculated moments of order statistics on the regression based goodness of fit tests, including the correlation test and the Shapiro-Wilk test.

Another important application of the algorithm is the computation of the population L-moments and Tl-moments. Hosking (1990) define L-moments λ_m by

$$\lambda_m = \frac{1}{m}\sum_{j=0}^{m-1}(-1)^j \binom{m-1}{j} \mu_{m-j;m} \tag{23}$$

where $\mu_{r;n} \equiv E(X_{(r)})$ from a random sample of size n. Hosking (1992) tabulates formulae giving λ_1, λ_2, λ_3, λ_4 for some common distributions. An extra procedure in the algorithm will compute the population L-moments, m=1,2,… from any continuous distribution. The first 10 L-moments for the normal, logistic, beta(8,5), extreme-value, gamma(4) and the exponential distributions are given by Table (7).

Table (7) Population L-moments

m	normal	logistic	beta(8,5)	uniform	extreme	gamma(4)	exponential
2	0.56419	0.999001	0.073806	0.166567	0.692648	1.09375	0.496981
3	0	0	-0.00325	-0.001e-9	-0.11728	0.180097	0.163652
4	0.069171	0.165668	0.007521	-0.001e-9	0.103733	0.143557	0.080319
5	0	0	-0.00091	-0.001e-7	-0.03823	0.055989	0.046994
6	0.02463316	0.670735	0.00244210	-0.001e-7	0.03978	0.05286	0.030327
7	0	0	-0.00041	-0.001e-7	-0.01865	0.027012	0.020812
8	0.012324	0.034717	0.001146	-0.001e-6	0.020681	0.02705	0.01466
9	0	0	-0.00023	-0.001e-6	-0.01091	0.01586	0.010908
10	0.0073138	0.021225	0.00064733	-0.001e-6	0.012522	0.01633	0.008128

The sample L-moments are sensitive to outliers. Elamir and Seheult (2003) introduce a robust generalization to the L-moments called trimmed L-moments (TL-moments). The TL-moments allows for different proportion of trimming from each tail. This is more suitable for skewed distributions. The TL-moments is defined by

$$\lambda_m^{t_1 t_2} = \frac{1}{m}\sum_{k=0}^{m-1}(-1)^k \binom{m-1}{k} \mu_{m+t_1-k;m+t_1+t_2}, \qquad m = 1,2, \ldots . \tag{24}$$

This form allows for different proportion of trimming from each tail. Table (8) computes the first 10 population TL-moments with $t_1=t_2=1$. Table (9) computes the first 10 population TL-moments with $t_1=t_2=2$. In both tables we consider the normal, logistic, beta(8,5), extreme-value, gamma(4), and the exponential distributions.

Table (8) Population TL-moments ($t_1=t_2=1$)

m	normal	logistic	beta(8,5)	uniform	extrem	gamma(4)	exponential
2	0.297011	0.499999	0.039771	0.10000	0.353349	0.570116	0.249997
3	0	0	-0.0011134	6.5114e-8	-0.037646	0.059099	0.055551
4	0.01855726	0.041666	0.00211630	1.4666e-7	0.026647	0.03779	0.020831
5	0	0	-0.0001899	2.0536e-7	-0.007475	0.011064	0.009995
6	0.00441879	0.011110	0.00046528	3.5221e-7	0.006855	0.009265	0.005629
7	0	0	-6.2081e-5	4.7351e-7	-0.002652	0.003823	0.003395
8	0.00165765	0.004464	0.00016497	7.2062e-7	0.002699	0.003547	0.002229
9	0	0	-2.72595e-5	9.3701e-7	-0.001233	0.00175	0.001506
10	0.00078620	0.002221	7.4867e-5	1.3631e-6	0.001326	0.001709	0.0011

Table (9) Population TL-moments ($t_1=t_2=2$)

m	normal	logistic	beta(8,5)	uniform	extreme	gamma(4)	exponential
2	0.20154683	0.333333	0.02719850	0.071429	0.237165	0.385631	0.166666
3	0	0	-0.0005615	-5.71E-08	-0.018499	0.029352	0.027778
4	0.00763964	0.016666	0.00088707	-6.94E-08	0.010761	0.015448	0.008298
5	0	0	-6.42907e-5	-8.37E-08	-0.002450	0.003665	0.003334
6	0.00130438	0.003174	0.00014074	-1.05E-07	0.001978	0.00271	0.001587
7	0	0	-1.5829221	-1.72E-07	-0.000625	0.000951	0.000861
8	0.00038117	0.000992	3.90490e-5	-2.15E-07	0.000605	0.000807	0.000498
9	0	0	-5.57974e-5	-3.25E-07	-0.000243	0.000348	0.000304
10	0.00014800	0.000404	1.45540e-5	-4.26E-07	0.000243	0.000318	0.000278

6. COMMENTS AND CONCLUSION

The new algorithm facilitates the computation of the moments of order statistics for any sample size and for many continuous distributions. The accuracy for the logistic, gamma, beta, extreme value and the student's t distributions will be presented in a forthcoming article, The existing tables for the above distributions either do not exist or incomplete (for the expected values only or for small sample sizes only). Also, the inverse Gaussian, Pareto, and the generalized lambda

distribution will be added and another algorithm will be written for discrete distributions. The algorithm still needs improvement especially for the computation of the covariances.

The proposed procedure for calculating the moments of order statistics can improve the statistical analysis in many applications: Lloyd's weighted least squares estimators for the location scale family, linear estimators for censored samples from any continuous distribution, regression based goodness of fit tests and regression based power transformation.

REFERENCES

Abramowitz, M. and Stegun, I. (Eds) (1965). *Handbook of Mathematical Functions*, with Formulas, Graphs, and Mathematical Tables, Dover, New York.

Arnold, B. C, Balakrishnan, N., and Nagaraja, H. N. (1992). A First Course in Order Statistics. Wiley, New York.

Childs, A. and Balakrishnan, N. (2002). Series approximations for moments of order statistics using MAPLE. Comput. Statist. Data Anal. 38, 341-7.

D'Agostino, R. B. and Stephens, M. A. (eds.) (1986). Goodness-of-Fit Techniques. Marcel Dekker, New York.

David, F. N. and Johnson, N. L. (1956). Some tests of significance with ordered variables (with Discussion). J. Roy. Statist. Soc. B 18, 1-31.

David, H. A. and Nagaraja, H. N. (2003). Order Statistics, 3ed ed. Wiley, New York.

Davis, C. S. and Stephens, M. A. (1978). Approximating the covariance matrix of normal order statistics. Algorithm AS 128, Appl. Statist. 27, 206-12.

Elamir, E. A. and Seheult, A. H. (2003). Trimmed L-moments. Comput. Statist. & Data Analysis,43,299-314.

Fard, M. and Holmquist, B. (2007). First moment approximation for order statistics from the extreme value distribution. Statistical Methodology 4,196-203.

Gauss 9.0 Language Reference (2007). @ www.aptech.com/manuals/

Gupta, S. S. and Shah, B. K. (1965). Exact moments and percentage points of the order statistics and the distribution of the range from the logistic distribution. Ann. Math. Statist. 36,907-20.

Harter, H. L. (1961). Expected values of normal order statistics. Biometrika 48, 151-65. Correction 48,476.

Harter, H. L. and Balakrishnan, N. (1996). CRC Handbook of Tables for the Use of Order Statistics in Estimation. CRC Press, Boca Raton, FL.

Hirakawa, K. (1973). Moments of order statistics. Ann. Statist. 1, 392-4.

Hosking, J. R. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. J. Roy. Statist. Soc. B,52,105-124.

Hosking, J. R. (1992). Moments or L-moments? An example comparing two measures of distributional shape. American Statistician,46,186-189.

Kennedy, W. and Gentle, J. (1980). Statistical Computing. Marcel Dekker, Inc. New York.

Lehmann, E. L. and Casella, G. (1998). Theory of Point Estimation. 2ed. Edition. Springer. New York.

Lloyd, E. H. (1952). Least-squares estimation of location and scale parameters using order statistics. Biometrika 39, 88-95.

Nadarajah, S. (2008). Explicit expressions for moments of order statistics. Statistics & probability letters 78, 196-205.

Parrish, R. S. (1992). Computing variances and covariances of normal order statistics. Commun. Statist.-Simul. Comput. 21,71-101.

Royston, J. P. (1982). Algorithm AS 177. Expected normal order statistics (exact and approximate). Appl. Statist. 31, 161-5.

Sarhan, A. E. and Greenberg, B. G. (1956). Estimation of location and scale parameters by order statistics from singly and doubly censored samples. Part I: The normal distribution up to samples of size 10. Ann. Math. Statist. 27, 427-51. Correction 40, 325.

Sen, P. (1959). On the moments of the sample quantiles. Calcutta Statist. Ass. Bull. 9, 1-19.

Steinhaus, S. (2008). Comparison of mathematical programs for data analysis.

http://www.scientificweb.de/ncrunch/

Teichroew, D. (1956). Tables of expected values of order statistics and products of order statistics for samples of size twenty and less from the normal distribution. Ann. Math. Statist. 27, 410-26.

Tietjen, G. L., Kahaner, D. K., and Beckman, R. J. (1977). Variances and covariances of the normal order statistics for sample sizes 2 to 50. Selected Tables in Mathematical Statistics 5, 1-73.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 747–758.

LONGITUDINAL DATA ANALYSIS USING NONPARAMETRIC REGRESSION MODEL

Noor Akma Ibrahim
Institute for Mathematical Research
Universiti Putra Malaysia
43400 UPM, Serdang, Selangor
Malaysia
E-mail: nakma@putra.upm.edu.my

Suliadi
Dept. of Statistics, Bandung Islamic University
Jl. Tamansari No. 1 Bandung
Indonesia
E-mail: suliadi@gmail.com

ABSTRACT

This paper proposes nonparametric regression model to analyze longitudinal data. We combine natural cubic spline with generalized estimating equations (GEE) to handle unknown function of the covariate and accounting for the correlation within subjects. Specific condition in which we assume independence, AR-1 and exchangeable correlation structures from each subject with varying sample size are used in the simulation study to assess the efficiency of the estimators. A real data application of the proposed model is illustrated with comparison to parametric model and GEE-smoothing spline under independence assumption.

Keywords: longitudinal data; nonparametric regression; correlation, generalized estimating equations; smoothing spline

1. INTRODUCTION

It is very common in economics, epidemiology or clinical trials to make a study on subjects who are followed over time or several occasions to collect response variables, which is commonly known as longitudinal study. The characteristic of these data is that they are no longer independent, in which there is correlation within subject measurements. Another characteristic is that the variances are usually not homogeneous. Thus methods in the class of generalized linear model (GLM) are no longer valid for these data, since GLM assumes that observations are independent. Some developments have been proposed to analyze such data that can be classified into three types of model, marginal model, subject specific effect, and transition model (Davis, 2002). In the class of marginal model, Liang and Zeger (1986) and Zeger and Liang (1986) extended *quasi-likelihood estimation* of Wedderburn (1974) by introducing "working correlation" to accommodate within subject correlation, which is called generalized estimating equation (GEE). GEE yields consistent estimates of the regression coefficients and their variances even

though there is misspecification of the working correlation structure, provided the mean function is correctly specified.

GEE is part of the class of parametric estimation, in which the model can be stated in a linear function and the function is known. Very often the effect of the covariate cannot be specified in a specific function. Nonparametric regression can accommodate this problem by relaxing the relationship between covariate and response. In nonparametric regression, we assume that the effect of the covariate follows an unknown function without specific term that is just a smooth function. To date there are several methods in nonparametric regression, for example: local polynomial kernel regression, penalized splines regression, and smoothing splines. Green and Silverman (1994) gave a simple algorithm for nonparametric regression using cubic spline with penalized least square estimation. They also gave nonparametric and semiparametric methods for independent observations for a class of generalized linear models.

Some developments of nonparametric and semiparametric regression for longitudinal or clustered data have been achieved. Lin and Carroll (2000) considered nonparametric regression for longitudinal data using GEE-Local Polynomial Kernel (LPK). They showed that for kernel regression, in order to obtain an efficient estimator, one must ignore within subject correlation. This means within subject observations should be assumed independent; hence the working correlation matrix must be an identity matrix. This result was definitely different from GEE of Liang & Zeger's, in which the GEE estimator was consistent even there is a misspecification of the true correlation taken as a working correlation. Lin and Carroll (2001) also studied the behavior of local polynomial kernel which was applied to semiparametric-GEE for longitudinal data. The result was the same as in nonparametric GEE-LPK in Lin and Carroll (2000). Welsh et al. (2002) studied the locality of the kernel method for nonparametric regression and compared it to P-splined regression and smoothing splines. The result was that the kernel is local even when the correlation is taken into account. The result was different for smoothing splines, in which if there is no within subject correlation then smoothing splines is local, and if within subject correlation increases, than smoothing splines become more nonlocal. This implies that for smoothing splines, within subject correlation must be taken into account in the working correlation.

This paper considers nonparametric regression to analyze longitudinal data. We propose GEE-Smoothing spline in the analysis and study the properties of the estimator numerically. We use natural cubic spline and combine this with GEE of Liang & Zeger's in the estimation. Simulation study is carried out to investigate these properties.

The outline of this paper is follows. We give a short review of GEE in section 2.1. Section 2.2 provides a brief review of smoothing splines. The algorithm of the proposed method is discussed in section 3.1. Section 3.2 considers smoothing parameter selection. Properties of GEE-smoothing spline estimator using simulation is given in Section 4. Section 5 demonstrates the application of the proposed method to a real data set followed by the conclusion and discussion in Section 6.

2. GENERALIZED ESTIMATING EQUATION AND SMOOTHING SPLINES

2.1 Generalized Estimating Equation

Suppose there are n subjects, and the i-th subject is observed n_i times for the responses and covariates. Let $y_i = (y_{i1}, y_{i2}, \ldots, y_{i,ni})^T$ be the $n_i \times 1$ vector of response variable and $X_i = (x_{i1}, \ldots,$

$x_{ini})^T$ be n_i *x p* matrix of covariate for the *i-th* subject, and $x_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})^T$. It is assumed that the marginal density of y_{ij} follows the exponential family with probability density function

$$f(y_{ij}) = exp\left(\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{a(\phi)} + c(y_{ij}, \phi)\right).$$

The first two moments of y_{ij} are $E(y_{ij}) = b'(\theta_{ij}) = \mu_{ij}$ and $Var(y_{ij}) = b''(\theta_{ij})a(\phi)$, where θ_{ij} is a canonical parameter. It is assumed that between subjects, observations are independent. The relationship between μ and covariates is through the link function $g(\mu_{ij}) = \eta_{ij}$ with $\eta_{ij} = x^T_{ij}\beta$, where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is a *p x 1* vector of regression coefficient.

Generalized estimating equation to solve β was given by Liang and Zeger (1986) as follows:

$$\sum_{i=1}^{n} D_i^T V_i^{-1} S_i = 0, \tag{1}$$

where $D_i = \frac{\partial(b'(\theta_i))}{\partial\beta} = \frac{\partial\mu_i}{\partial\beta} = \frac{\partial\mu_i}{\partial\theta_i}\frac{\partial\theta_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta} = A_i \Delta_i X_i$; $\Delta_i = \frac{\partial\theta_i}{\partial\eta_i}$; $V_i = A_i^{1/2} R(\alpha) A_i^{1/2}$

A_i is an n_i *x* n_i diagonal matrix with diagonal elements $\partial\mu_{ij}/\partial\theta_{ij}$. $R(\alpha)$ is also called a "working correlation", a n_i *x* n_i symmetric matrix which fulfills the requirement of being a correlation matrix, and $S_i = y_i - \mu_i$, with $y_i = (y_{i1}, y_{i2}, \ldots, y_{ini})^T$ and $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{ini})^T$. The estimating equation (1) is similar to the quasi-likelihood estimating equation, except the form of V_i. Thus it can be seen as an estimating equation of β by letting Φ as the "quasi-likelihood" score function of the $y_1, \ldots, y_n$. Solution of β can be obtained by minimizing Φ over β. Thus,

$$\frac{\partial\Phi}{\partial\beta} = \sum_{i=1}^{n} D_i^T V_i^{-1} S_i = 0$$

Liang and Zeger (1986) gave the iterative procedure using modified Fisher scoring for β and the moment estimation method of α and ϕ. Given the current estimates $\hat{\alpha}$ and $\hat{\phi}$ then the iterative procedure for β is

$$\hat{\beta}_{s+1} = \hat{\beta}_s + \left[\sum_{i=1}^{n} D_i^T(\hat{\beta}_s)\tilde{V}_i^{-1} D_i(\hat{\beta}_s)\right]^{-1}\left[\sum_{i=1}^{n} D_i^T(\hat{\beta}_s)\tilde{V}_i^{-1} S_i(\hat{\beta}_s)\right], \tag{2}$$

where $\tilde{V}_i(\beta) = \tilde{V}_i\{\beta, \alpha(\beta, \hat{\phi}(\beta))\}$. The close form of moment estimator for α and ϕ for some correlation structures can be seen in Liang & Zeger (1986).

2.2 Smoothing Spline

Green and Silverman (1994) gave a simple approach in estimating smooth function *f* using natural cubic splines. Suppose given real numbers $t_1, \ldots, t_n$ on the interval $[a, b]$ satisfying $a < t_1 \ldots < t_n < b$. A function *f* on $[a, b]$ is *cubic spline* if two conditions are satisfied. First, *f* is cubic polynomial on each interval $(a, t_1), (t_1, t_2), \ldots, (t_n, b)$; second, the polynomial pieces fit together at the points t_i in such a way that *f* itself and its first and second derivative are continuous at each

t_i, thus the function is continuous on the whole of [*a*, *b*]. The function is said to be *natural cubic spline (NCS)*, if its second and third derivative are zero at *a* and *b*. Suppose $f_i = f(t_i)$ and $\gamma_i = f''(t_i)$ for $i = 1, 2, \ldots, n$. By definition of NCS, the second derivative of *f* at t_1 and t_n is zero, so $\gamma_1 = \gamma_n = 0$. Let $\boldsymbol{f} = (f_1, f_2, \ldots, f_n)^T$ and $\gamma = (\gamma_2, \ldots, \gamma_{n-1})^T$. Vector γ is numbered in non standard way, starting at $i = 2$. The vector $\boldsymbol{f}$ and vector γ completely specify the curve *f*. These two vectors are related and specified by two matrices *Q* and *R* defined below.

Let $h_i = t_{i+1} - t_i$, for $i = 1, 2, \ldots, n-1$. Let *Q* be the $n \times (n-2)$ matrix with elements q_{ij}, $i = 1, .., n$, and $j = 2, \ldots, n-1$, with $q_{j-1,j} = h_{j-1}^{-1}, q_{jj} = -h_{j-1}^{-1} - h_j^{-1}$, and $q_{j+1,j} = h_j^{-1}$ The *R* matrix is defined as follows. The symmetric matrix *R* is $(n-2) \times (n-2)$ with elements r_{ij}, for *i* and *j* running from 2 to $(n-1)$, given by

$$r_{ii} = (h_{i-1} + h_i)/3, \text{ for } i = 2, 3, .., n-1$$
$$r_{i,i+1} = r_{i+1,i} = h_i/6, \text{ for } i = 2, 3, .., n-1$$

Matrix *R* and *Q* are numbered in non standard way. The matrix *R* is strictly diagonal dominant, in which $|r_{ii}| > \sum_{i \neq j} |r_{ij}|$. Thus *R* is strictly positive-definite, hence R^{-1} exists. Defined a matrix *K* by

$$K = QR^{-1}Q^T \tag{3}$$

One important result is the theorem given by Greean & Silverman (1994)as stated below:

Theorem. *The vector $\boldsymbol{f}$ and γ specify a natural cubic spline f if and only if the condition $Q^T\boldsymbol{f} = R\gamma$ is satisfied. If condition above is satisfied then the roughness penalty will satisfy*

$$\int_a^b [f''(t)]^2 dt = \gamma^T R\gamma = f^T Kf \tag{4}$$

The proof of this theorem is in Green and Silverman (1994).

Green and Silverman (1994) proposed smoothing spline for several conditions, e.g nonparametric and semiparametric regressions for independent continuous data, nonparametric and semiparametric generalized linear models for independent data, and quasi-likelihood for independent data. They also considered method for correlated continuous data. For quasi-likelihood approach, the important result is the solution of the function $\boldsymbol{f}$ for nonparametric regression and parameter β for semiparametric regression, obtained by maximizing "penalized quasi-likelihood":

$$\Pi = \Phi - \frac{1}{2}\lambda \int [f''(t)]^2 dt \tag{5}$$

Thus the solution of $\boldsymbol{f}$ is obtained by maximizing (5).

3. GENERALIZED ESTIMATING EQUATION-SMOOTHING SPLINE

3.1 Estimation of GEE-Smoothing Spline

Suppose there are n subjects and the measurement of the i-th subject taken n_i times. Let $y_i = (y_{1i}, y_{2i}, \ldots, y_{ni,i})^T$ be a vector of responses of the i-th subject, corresponding to the vector of covariate $t_i = (t_{i1}, t_{i2}, \ldots, t_{i,ni})^T$ and y_{ij} comes from exponential family distribution with canonical parameter θ_{ij}. Thus $E(y_{ij}) = b'(\theta_{ij}) = \mu_{ij}$ and $Var(y_{ij}) = b''(\theta_{ij})a(\phi)$.

Consider the population average model, where the systematic component of the exponential family is nonparametric, rather than parametric, that is $g(\mu_{ij}) = \eta_{ij} = f(t_{ij}), i = 1,2,\ldots,n; j = 1,2,\ldots,n_i$. We replace the systematic component with unknown smooth function, i.e. natural cubic splines, rather than linear (known) function. In this paper we use the canonical link function $\theta_{ij} = \eta_{ij}$. Suppose X_i is the $n_i \times q$ incidence matrix of all t_{ij}'s that can be constructed as follows. Let all t_{ij}'s have q different values that can be stated as $t_{(1)} < t_{(2)} < \ldots < t_{(q)}$ and the relation to $\underline{x}_{ijk}$ is $x_{ijk} = 1$, if $t_{ij} = t_{(k)}$ and $x_{ijk} = 0$, if $t_{ij} \neq t_{(k)}$.

Let $x_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijq})^T$. The vector of the functions f at different point is denoted by

$$f = [f(t_{(1)}), f(t_{(2)}), \ldots, f(t_{(q)})]^T .$$

Then the function f at point t_{ij} can be expressed as $f(t_{ij}) = x_{ij}^T \boldsymbol{f}$. Set

$$X_i = (x_{i1}, x_{i2}, \ldots, x_{in_i})^T ; y_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^T$$
$$\eta_i = (\eta_{i1}, \eta_{i2}, \ldots, \eta_{in_i})^T ; \mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{in_i})^T$$

With this set-up, the link function now has the form $g(\mu_{ij}) = f(t_{ij}) = x_{ij}^T \boldsymbol{f}$

Since function f can be any arbitrary smooth function, then to maximize "quasi-likelihood" score function Φ, one might take y_{ij} as the estimates of $f(t_{ij})$ and the Φ will be maximum. However, the function obtained is just an interpolation of the y_{ij} and the function is too rough or wiggly. A smooth function can be obtained by adding roughness penalty to the objective function. This roughness penalty is called *penalized "quasi-likelihood" function* defined by

$$\Pi = \Phi - \frac{1}{2}\lambda \int_a^b [f''(t)]^2 dt .$$

From (1), (3), (4) and (5), the generalized estimating equation-smoothing splines is defined as

$$\frac{\partial \Pi}{\partial f} = \sum_{i=1}^{n} D_i^T V_i^{-1} S_i - \frac{\partial}{\partial f}\left[\frac{1}{2}\lambda \int [f''(t)]^2 dt\right] = \sum_{i=1}^{n} D_i^T V_i^{-1} S_i - \lambda K \boldsymbol{f} = 0 . \quad (6)$$

Given the current estimates of $\hat{\boldsymbol{\alpha}}$ and assuming canonical link function is being used, the iterative procedure using modified Fisher scoring for $\boldsymbol{f}$ is

$$\hat{\boldsymbol{f}}_{s+1} = \hat{\boldsymbol{f}}_s + \left[\sum_{i=1}^{n} D_i^T(\hat{\boldsymbol{f}}_s)\tilde{V}_i^{-1} D_i(\hat{\boldsymbol{f}}_s) + \lambda K\right]^{-1}\left[\sum_{i=1}^{n} D_i^T(\hat{\boldsymbol{f}}_s)\tilde{V}_i^{-1} S_i(\hat{\boldsymbol{f}}_s) - \lambda K \hat{\boldsymbol{f}}_s\right] \quad (7)$$

where $\tilde{V}$ is defined as in (2). Solution of $\boldsymbol{f}$ can also be obtained using iteratively re-weighted least square method. Define $\ddot{y}_i = X_i \boldsymbol{f} + A_i^{-1}(y_i - \mu_i)$

Using $D_i = \frac{\partial(b'(\theta_i))}{\partial\beta} = \frac{\partial\mu_i}{\partial\beta} = \frac{\partial\mu_i}{\partial\theta_i}\frac{\partial\theta_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta} = A_i\Delta_i X_i$; $\Delta_i = \frac{\partial\theta_i}{\partial\eta_i}$; $V_i = A_i^{1/2} R(\alpha) A_i^{1/2}$ and using canonical link function, thus (7) can be rewritten as

$$\hat{\boldsymbol{f}}_{s+1} = \left[\sum_{i=1}^{n} X_i^T A_i \tilde{V}_i^{-1} A_i X_i + \lambda K\right]^{-1}\left[\sum_{i=1}^{n} X_i^T A_i \tilde{V}_i^{-1} A_i \ddot{y}\right] \tag{8}$$

where A, $\boldsymbol{f}$, and μ are evaluated from the *s-th* iteration. Solution of $\boldsymbol{f}$ is obtained from (8) by iterating until convergence.

Derivation of the variance of estimate may follow Liang & Zeger (1986) also known as the sandwich estimator. For GEE this estimator is consistent even $R(\alpha)$ is not the true correlation matrix of y_i. Since $Var(\hat{f}) = Var(\hat{f} - Af)$ for any constant matrix A, then the covariance matrix of the estimate obtained from (7) is $Var(\hat{f}) = \Sigma_0^{-1}\Sigma_1\Sigma_0^{-1}$, where $\Sigma_0^{-1} = \left[\sum_{i=1}^{n} X_i^T A_i \tilde{V}_i^{-1} A_i X_i + \lambda K\right]^{-1}$ and $\Sigma_1 = \sum_{i=1}^{n} X_i^T A_i \tilde{V}_i^{-1} S_i S_i^T \tilde{V}_i^{-1} A_i X_i$. The sand-wich estimator is robust to the misspecification of correlation structure. Another possibility of variance of the estimate is the naïve (model based) estimator defined by $Var(\hat{f}) = \Sigma_0^{-1}\phi$. Liang & Zeger (1986) also gave moment method to estimate the association parameter, α, and the scale parameter, ϕ.

3.2 Smoothing Parameter Selection

Smoothing parameter (λ) is an important part in GEE-Smoothing Spline. This parameter measures the "trade off" or exchange between goodness of fit and the roughness or the smoothness of the curve. Hence, the performance of the estimator depends on λ. In selecting smoothing parameter, we use a method proposed by Wu & Zhang (2006, p326) which is called *leave-one-subject-out cross validated deviance (SCVD)*. Smoothing parameter λ is chosen that minimizes the SCVD score, where $SCVD(\lambda) = \sum_{i=1}^{K}\sum_{j=1}^{n_i} d(y_{ij}, \hat{\mu}_{ij}^{(-i)})$ where d is "deviance" and $\mu_{ij}^{(-i)} = g^{-1}(X\hat{f}^{(-i)})_{ij}$ is the estimate value for the *i-th* subject and the *j-th* time observation using $\boldsymbol{f}^{(-i)}$. The $\boldsymbol{f}^{(-i)}$ is $\boldsymbol{f}$ obtained without the *i-th* observation. Since GEE is based on quasi-likelihood thus the deviance is also based on the quasi-likelihood (see: Hardin & Hilbe, 2003, Ch. 4; McCullagh & Nelder, 1989 Ch. 9).

Direct computation of $\boldsymbol{f}^{(-i)}$ is time consuming. Wu & Zhang (2006) suggested using approximate of $\boldsymbol{f}^{(-i)}$ computed as follows. Suppose from the final iteration of (7) or (8) we have D_i, $\tilde{V}_i^{-1}$, and S_i. Then the $\boldsymbol{f}^{(-i)}$ is approximated by

$$\hat{\boldsymbol{f}}^{(-i)} = \hat{\boldsymbol{f}} + \left[\sum_{r\neq i} D_r^T \tilde{V}_r^{-1} D_r + \lambda K\right]^{-1}\left[\sum_{r\neq i}^{n} D_r^T \tilde{V}_r^{-1} S_r - \lambda K\hat{\boldsymbol{f}}\right] \tag{9}$$

We still need to compute $\hat{\boldsymbol{f}}^{(-i)}$ for $i = 1, 2, \ldots, n$, but we do not need to iterate (9) from the beginning.

4. SIMULATION STUDY

The objective of this simulation is to study the properties of GEE-smoothing spline, such as the bias, consistency, and efficiency by considering different sample size and correct and incorrect correlation structure in the estimation. In this simulation we only consider binary data using logit link function.

4.1 Model and Structure of Data

We generated correlated binary data using R language version 2.7.1 (see Leisch et al, 1998). Three correlation structures were considered: (i) autoregressive with corr(y_{ij},$y_{i(j+1)}$) = 0.7, for j = *1, 2, …,* n_i; (ii) exchangeable with corr(y_{ij},$y_{ij'}$) = 0.35, for j', j = *1, 2, …,* n_i and $j' \neq j$; and (iii) independent with corr(y_{ij},$y_{ij'}$) = 0, for j', j = *1, 2, …,* n_i and $j' \neq j$. Each subject was considered to be measured ten times, t = *7.5, 25.5, …, 169.5.* The function was $f(t) = Sin(\pi t/90)$. The response variable, y_{ij}, related to covariate t through a link function is as follows,

$$E(y_{ij}) = \mu_{ij} \text{ and } logit(\mu_{ij}) = log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = f(t_{ij})$$

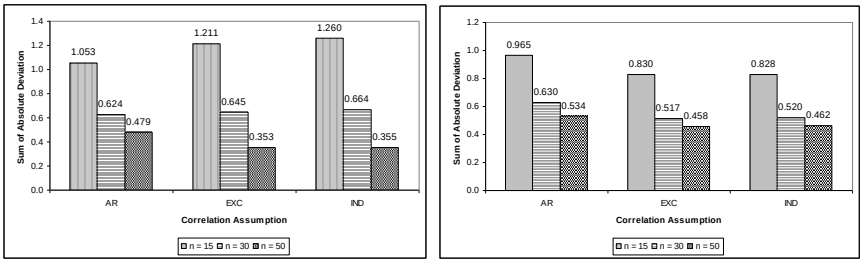
We considerd three sample size n = *15*, n = *30*, and n = *50*. For each correlation structure, we estimated the function f by asssuming that the correlation structures were (1) autoregressive, (2) exchangeable, and (3) independent. Thus for each one, there were nine combinations of sample size and correlation structure. Each combination was run 250 times. The association parameter and the scale parameter were estimated using method of moment given by Liang & Zeger (1986).

4. 2 Simulation Results

Figure 1 shows pointwise sum of absolute deviation (SAD). SAD is defined as follows. Suppose $\hat{f}_t^*$ is the average of 250 replications at point t, thus $\hat{f}_t^* = \sum_{r=1}^{250} \hat{f}^{(r)}(t)/250$. SAD is defined as $SAD = \sum_{j=1}^{10} |\hat{f}_{t_j}^* - f_{t_j}|$. Thus SAD shows the bias of the estimates. Figure 1(a-c) shows SAD for true correlation structure of autoregressive, exchangeable, and independent, respectively.

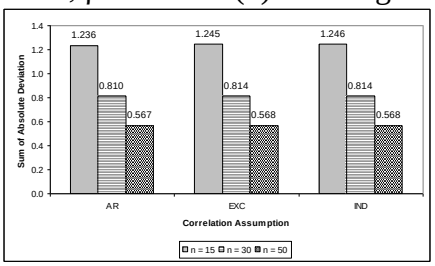
From Figure 1 we can observe the behavior of the bias of the estimators. Referring to the correlation structure, there is no obvious pattern of the bias with respect to the correct or incorrect correlation structure. The degree of biasness is related to the sample size. Using correct or incorrect correlation structure, the bias will decrease when sample size increases. This behavior is the same for data that have high correlation (autoregressive, ρ = 0.7), moderate correlation (Exchangeable, ρ = 0.35), and independent.

We use standard deviation of 250 of each point estimates to study the consistency and efficiency. The estimator is said to be consistent if the standard deviation tends to zero when sample size is infinity, i.e. standard deviation decreases when sample size increases. This standard deviation can also be used to study the efficiency, that is small standard deviation indicates the efficiency of the estimator. Figure 2 shows the standard deviation of 250 pointwise function estimates.



(a) Autoregressive, $\rho = 0.7$

(b) Exchangeable, $\rho = 0.35$



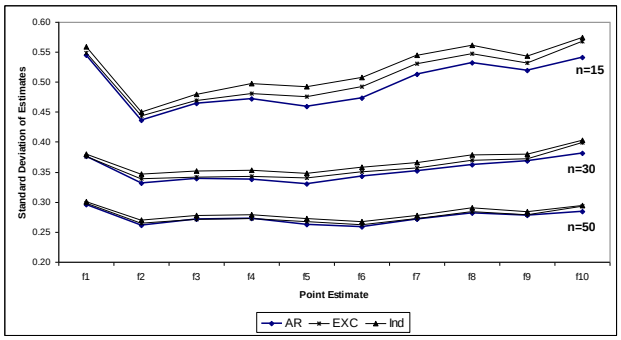
(c) Independence

Figure 1. Sum of Absolute Deviation of Three of True Correlation Structures

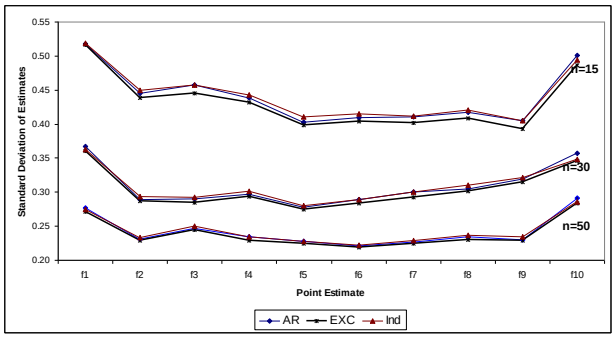
From Figure 2 we can observe the consistency of the estimator. The trend of the standard deviation for all true correlation structures is the same. It decreases when sample size increases. The same trend is also observed for all correlation structures, using correct or incorect correlation structure. This means that the estimators are consistent and the consistency still holds even if we use incorrect correlation structure. The rate of the decrease of standard deviation from *n =15* to *n = 30*, and from *n = 30* to *n = 50*, is the same for all true correlation structures. This indicates that the convergency rate is (almost) the same for all conditions of true correlation structures.

From the standard deviation we can also study the efficiency of the estimator. From the result of the efficiency study we may conclude whether we need to take into account the correlation into the model or just ignore the dependency. The method that produces smaller variance or standard deviaton of estimator indicates it is more efficient than the others.

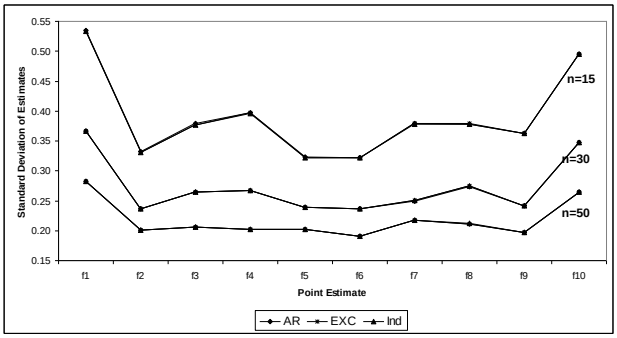
From Figure 2 we see that if the data are correlated (true correlation is autoregressive or exchangeable), for a specific sample size, the largest standard deviation is obtained if one assumes that the data are independent. Whilst using true correlation structure, the standard deviation is the smallest. This means that taking into account the dependency into the model is better than assuming the data are independent, even if we use incorrect correlation structure. The most efficient estimate is obtained if we use true correlation structure. The difference between standard deviations of correlation structure (AR, EXC, and IND) tends to get closer when we increase the sample size, hence we conjecture that the efficiency of correct or incorrect correlation structure is almost similar if sample size is large. If the true correlation structure is independent, the standard deviation of AR, EXC, and IND are almost similar, for all sample size. Thus in this case, the efficiency of using incorrect correlation structures is almost similar to the efficiency of using correct correlation structure.



(a) True Correlation Structure is AR



(b) True Correlation Structure is Exchangeable



(c) True Correlation Structure is Independence

Figure 2. Standard Deviation of 250 Replications of Pointwise Function Estimates

5. APPLICATION TO REAL DATA

As an application of the proposed method, we used data of AIDS Clinical Trials Group (ACTG) 388 study sponsored by NIAID/NIH. Data used was CD4+ cell count as the response variable (see. Fischl, et al., 2003 for detail). These data have been used by Park & Wu (2006) for nonparametric mixed-effects models. We did not utilize all the data, only those subjects that had received lamifudine plus zidovudine and indinavir (indinavir group) for observations at week 0, 8, 16, ..., 80. Only subjects with complete observations were considered. The covariate was the time (week).

As a comparison we considered three scenarios: (i) parametric approach; (ii) nonparametric (smoothing spline) approach with assumption within subject observations are independent; and (iii) nonparametric approach with assumption within subject observations are dependent, using GEE-Smoothing spline assuming the within subject correlation following autoregression structure. The model was $E(Y_{ij})=\mu_{ij}$. We considered linear and quadratic model for parametric approach:

1. $\mu^{L}_{ij} = \beta_0 + \beta_1 \text{WEEK}_{ij}$
2. $\mu^{Q}_{ij} = \beta_0 + \beta_1 \text{WEEK}_{ij} + \beta^2 \text{WEEK}^2_{ij}$,

and for nonparametric approach we used the model: $\mu_{ij}=f(WEEK)$. We used PROC GENMOD for parametric approach, and SAS IML for nonparametric approach.

The results of parametric approach are model (1): $\hat{y} = 163.696 + 2.938WEEK$ with $r_{j,j+t} = 0.887^t$ and model (2): $\hat{y} = 149.456 + 5.817WEEK - 0.036WEEK^2$ with $r_{j,j+t} = 0.886^t$. All estimates of regression coefficients have p-value < 0.0001. Thus we recommend quadratic model for the parametric approach.

Figure 3 shows the comparison between the estimate of variance of $\hat{f}(t_j)$ using independence and AR assumption of correlation structure, for naive and robust variance estimates. This figure shows that there are large differences between naive estimates of independence and AR correlation assumption, where variance from independence assumption is much smaller than that obtained from AR assumption. This shows that considering independence for these data is not appropriate and will result in the under estimation of variance. Whilst the robust estimate of variances both assuming independence and AR are almost similar, for all points measurement. It shows the power of the robust or sandwich variance estimate, if we use incorrect correlation structure.

The estimates of *f(t)* based on nonparametric approach by assuming independence and AR correlation structure are almost similar for both functions estimates (Figure 4) for all time measurements (week). This is not a general case for correlated data, since the within subject correlation is large, where the correlation estimate for AR assumption is $r_{j,j+t} = 0.889^t$. The advantage of using dependence assumption to these data is that we know that within subject observations have high correlation.

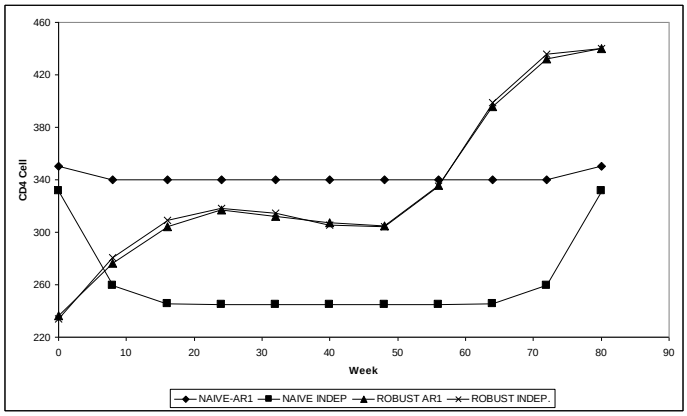


Figure 3. Variance Estimates of Ten Points Function Estimates for Nonparametric Approach of CD4 Cell Data.

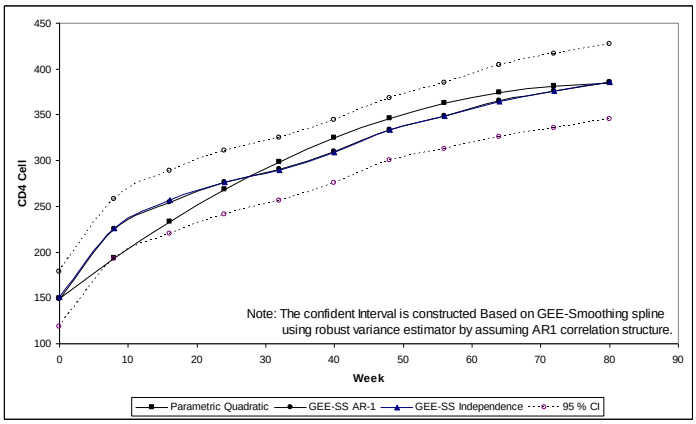


Figure 4. Pointwise Function Estimates for Parametric and Nonparametric Approach of CD4 Cell Data and the Confident Interval Based on GEE-Smoothing Spline.

Comparing the nonprametric approach and parametric approach for these data, we suggest that quadratic model is appropriate, since its estimates are in the range of nonparametric confidence interval (see Figure 4).

6. CONCLUSION

From the simulation results, we can see that estimates obtained from GEE-Smoothing spline has good properties, with respect to the biasness, consistency and efficiency. The best estimate is obtained if the correct correlation structure is used. The sandwich variance estimator gives a robust variance estimate with respect to the misspecification of the correlation structure.

REFERENCES

Davis, Charles S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag. New York. USA.

Fischl, M. A., Ribaudo, H. J., Collier, A. C., et al. (2003). A Randomized Trial of 2 Different 4-Drug Antiretroviral Regimens versus a 3-Drug Regimen, in Advanced Human Immunodeficiency Virus Disease. *The Journal of Infectious Diseases*, 188, 625-634.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach*. Chapman & Hall/CRC. New York, USA.

Hardin, James W. and Hilbe, J. M. (2003). *Generalized Estimating Equations.* Chapman & Hall/CRC. Washington, DC. USA.

Leisch, F., Weingessel, A. and Hornik, K. (1998). On the Generation of Correlated Artificial Binary Data. *Working Paper*. No 13. SFB. Adaptive Information System and Modeling in Economics and Management Science. Vienna University of Economics and Business Administration. Austria.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal Data Analysis using Generalized Linear Models. *Biometrika*, 73, 13-22.

Lin, X. and Carroll, R. J. (2000). Nonparametric Function Estimation for Clustered Data When the Predictor is Measured Without/With Error. *Journal of the American Statistical Association*, 95, 520-534.

Lin, X. and Carroll, R. J. (2001). Semiparametric Regression for Clustered Data Using Generalized Estimating Equations. *Journal of the American Statistical Association*, 96, 1045-1056.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models. 2nd Edition.* Chapman and Hall. London. UK.

Park, J. G. and Wu, H. (2006). Backfitting and Local Likelihood Methods for Nonparametric Mixed-Effects Models with Longitudinal Data. *Journal of Statistical Planning and Inferencia,* 136, 3760-3782.

Weddernburn, R. W. M. (1974). Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika*, 61, 439-447.

Welsh, A.H., Lin, X., and Carroll, R. J. (2002). Marginal Longitudinal Nonparametric Regression: Locality and Effeciency of Spline and Kernel Methods. *Journal of the American Statistical Association*, 97, 482-493.

Wu, H. and Zhang, J. T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis*. John Wiley & Sons. USA.

Zeger, S. L. and Liang, K.Y. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*, 42, 121-130.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 759–770.

CHI-SQUARE TEST FOR GOODNESS OF FIT FOR LOGISTIC DISTRIBUTION USING RANKED SET SAMPLING AND SIMPLE RANDOM SAMPLING

K. Ibrahim [1], M. T. Alodat [2], A. A. Jemain [3], S. A. Al-Subh [4]
[1,3,4] School of Mathematical Sciences, Universiti Kebangsaan Malaysia, Selangor, Malaysia
[2] Department of Statistics, Yarmouk University, Irbid, Jordan
E-mail: [1] Kamarulz@ukm.my, [2] alodatmts@yahoo.com , [3] kpsm@ukm.my,
[4] salsubh@yahoo.com

ABSTRACT

In this paper, we improve the power of the chi-square test for goodness-of-fit based on Ranked Set sampling (RSS), a sampling technique introduced by McIntyre (1952). Moreover, we conduct a simulation study to compare the power of the chi-square test based on RSS with its counterpart under SRS.

Keywords: Goodness of fit test; Chi-square test; Power; Logistic distribution; Ranked set sample.

1. INTRODUCTION

Ranked Set Sampling (RSS), introduced by McIntyre (1952), is an ingenious sampling technique for selecting a sample which is more informative than a Simple Random Sampling (SRS) to estimate the population mean. RSS technique is very useful when visual ranking of population units is less expensive than their actual quantifications. Therefore, selecting a sample based on RSS can reduce the cost and increase the efficiency of estimation. McIntyre had made use of RSS technique to estimate the mean pasture and forage yields.

The basic idea behind selecting a sample under RSS can be described as follows: Select m random samples each of size m. Using a visual inspection, rank the units within each sample with respect to the variable of interest. Then select, for actual measurement, the i^{th} smallest unit from the i^{th} sample, $i = 1,...,m$. In this way, we obtain a total of m measured units, one from each sample. The procedure could be repeated r times until a sample of $n = mr$ measurements are obtained. These mr measurements form an RSS. Takahasi and Wakimoto (1968) gave the theoretical setups for RSS. They showed that the mean of an RSS is the minimum variance unbiased estimator for the population mean. Dell and Clutter (1972) showed that the RSS mean remains unbiased and more efficient than the SRS mean for estimating the population even if ranking is not perfect. Stocks and Sager (1988) studied the characterization of an RSS. They suggested an unbiased estimator for the population distribution function based on the empirical distribution function of a RSS. Based on this empirical distribution function, they proposed a Kolmogorov-Smirnov goodness-of-fit test. They derived the null distribution of their proposed test. RSS has been used extensively in ecological and environmental fields (Johnson et al., 1993; Patil and Taillie, 1993; Kaur et al., 1996).

It is known that two factors which affect the efficiency of RSS are set size (m) and the ranking errors. The larger the set size, the larger is the efficiency of the RSS. Thus if the set size is larger, it is more difficult to conduct the visual ranking and as a result, the ranking errors increase. (Al-Saleh and Al-Omari, 2002). Several authors have modified RSS to reduce the error in ranking and to make visual ranking tractable by the experimenter (Muttlak, 1997; Samawi et al., 1996; Al-Odat and Al-Saleh, 2001; Muttlak, 2003). Samawi et al. (1996) investigated the use of Extreme Ranked Set Sampling (ERSS) which consists of quantifying the smallest and the largest order statistics. Muttlak (1997) introduced the Median Ranked Set Sampling (MRSS) which consists of quantifying only the median of each set in the McIntyre's RSS. Bhoj (1997) proposed a modification to RSS and called it New Ranked Set Sampling (NRSS). He used this method to estimate the location and the scale parameters of the rectangular and logistic distributions. A comprehensive survey about developments in RSS can be found in Chen (2000). Al-Odat and Al-Saleh (2001) introduced the concept of varied set size RSS, which they called Moving Extremes Ranked Set Sampling (MERSS). They investigated this modification non-parametrically and found that the procedure can be more efficient and applicable than the SRS.

When a researcher is interested in doing parametric statistical inferences about the population of interest based on RSS, it is important to know the shape of the parent distribution, i.e., the distribution from which a ranked set sample comes. This requires new statistical developments on how to do a goodness-of-fit test when the data in our hand are collected using RSS technique. In the literature, however, not much attention has been given on the goodness of fit tests on data collected based on RSS technique and its modifications. Thus, in this paper, we improve the power of the chi-square test statistic for goodness-of-fit under RSS.

This paper is organized as follows. In Section 2, we propose a chi-square test statistic for goodness-of-fit under RSS. In Section 3, we apply the proposed method for the logistic distribution. In Section 4, an algorithm is designed to calculate the power function under the distribution given in H. In Section 5, a simulation study is conducted to compare the power of the chi-square test statistic under RSS with its SRS counterpart. In Section 6, we apply the Kullback-Leibler information to compare the SRS and the RSS. In Section 7, we state our conclusions.

2. CHI-SQUARE TEST FOR GOODNESS-OF-FIT

In our study, we assume that the set size, in McIntyre's RSS, is odd. This assumption will lead to simple calculations when comparing our method with median ranked set sampling. If the set size is even, then the theory developed here could be extended without hesitation. Let $X_1, X_2,...,X_r$ be a random sample from the distribution function $F(x)$. Assume that our objective is to test the statistical hypotheses $H_o : F(x) = F_o(x) \quad \forall x$, vs. $H_1 : F(x) \neq F_o(x)$ for some x, where $F_o(x)$ is a known distribution function. One of the well known tests is the χ^2 test statistic for goodness-of-fit which can be described as follows. Let $I_1, I_2,...,I_{k+1}$ be a partition of the support of $F_o(x)$ and N_j = number of X_i's that fall in I_j, $j = 1, 2,...,k+1$ For large n, the hypothesis

$$H_o : F(x) = F_o(x) \quad \forall x, \text{ vs. } H_1 : F(x) \neq F_o(x)$$

for some x, is rejected if

$$\chi^2 = \sum_{j=1}^{k+1} \frac{(N_j - nP_j)^2}{nP_j} > \chi^2_{1-\alpha,\, k}$$

where $P_j = P_{F_o}(X_i \in I_j)$, $j = 1, 2, ..., k+1$ and $\chi^2_{1-\alpha,\, k}$, is the $(1-\alpha)100$ quantile of the chi-square distribution with k degrees of freedom. It can be noted that testing the hypothesis

$$H_o : F(x) = F_o(x) \quad \forall x, \text{ vs. } H_1 : F(x) \neq F_o(x)$$

for some x, is equivalent to testing the hypothesis

$$H_o^* : G_i(y) = G_{io}(y), \quad \forall y \quad \text{vs.} \quad H_1^* : G_i(y) \neq G_{io}(y)$$

for some i, where $G_i(y),\ G_{io}(y)$ are the cdf's of the i^{th} order statistics of random samples of size $2m-1$ chosen from $F(x),\ F_o(x)$, respectively. According to Arnold et al. (1992) $G_i(y)$ and $G_{io}(y)$ have the following representations:

$$G_i(y) = \sum_{j=i}^{2m-1} \binom{2m-1}{j} [F(y)]^j [1-F(y)]^{(2m-1)-j}$$

and

$$G_{io}(y) = \sum_{j=i}^{2m-1} \binom{2m-1}{j} [F_o(y)]^j [1-F_o(y)]^{(2m-1)-j},$$

respectively. For example, in case of $m = 2$, $i = 1, 2,$ and 3, the cdf's $G_i(y)'s$ and $G_{io}(y)'s$ are given by

$$\begin{aligned}
G_1(y) &= 1-[1-F(y)]^3, \\
G_{1o}(y) &= 1-[1-F_o(y)]^3, \\
G_2(y) &= 3F^2(y)(1-F(y)] + F^3(y), \\
&= 3F^2(y) - 2F^3(y), \\
G_{2o}(y) &= 3F_o^2(y) - 2F_o^3(y),
\end{aligned}$$

and

$$\begin{aligned}
G_3(y) &= F^3(y), \\
G_{3o}(y) &= F_o^3(y).
\end{aligned}$$

It is easy to show that the equation $G_i(y) = G_{io}(y)$ has the unique solution $F(x) = F_o(x)$.

If we apply ranked set sampling to collect the data based on the i^{th} order statistic, then we may use the resulting data to build a chi-square test for the hypothesis H_o^* vs. H_1^*. To do this, let $Y_1, ..., Y_r$ be a random sample of size r selected via the i^{th} order statistic and let

$I_j, j = 1,..., k+1$ be a partition of $(-\infty, \infty)$. Let M_j = #of Y_i's that fall in the interval I_j, $j = 1,..., k+1$. We use the chi-square test statistic

$$\chi^{*2} = \sum_{j=1}^{k+1} \frac{(M_j - r\ P_j^*)^2}{r\ P_j^*}, \tag{1}$$

where $P_j^* = \int_{I_j} dG_{io}(y)$ to test the hypotheses H_o^* vs. H_1^*. The hypothesis H_o^* is rejected at level α if $\chi^{*2} > \chi^2_{1-\alpha, k}$.

The logistic distribution is widely used in many fields of science because it is more suitable than normal distribution to model heavy-tailed data sets arising in biology, econometrics, agricultural etc. The logistic distribution has the following cdf

$$F_o(x) = (1+e^{-(x-\theta)/\sigma})^{-1}, \quad -\infty < x < \infty, \ -\infty < \theta < \infty, \ \sigma > 0.$$

A historical background on this distribution and its application can be found in Balakrishnan (1992). Because of its wide applications, this distribution will be mainly considered in this paper. In the next section, we show, using simulation, that the test statistic χ^{*2} is more powerful than the test statistic χ^2 when they are compared using samples of the same size. The power of the χ^{*2} test statistic can be calculated according to the equation

$$\text{Power of } \chi^{*2}(H) = P_H(\chi^{*2} > \chi^2_{1-\alpha, k}), \tag{2}$$

where H is a cdf under the alternative hypothesis H_1^*. Since it is difficult to obtain this power theoretically, we employ the Monte Carlo simulation to approximate this power.

3. TESTING FOR LOGISTIC DISTRIBUTION

Let $F_o(x) = (1+e^{-(x-\theta)/\sigma})^{-1}$, where θ and σ are assumed to be known. Let $Y_1,...,Y_r$ be as in the previous section. To test the hypothesis $H_o : F(x) = (1+e^{-(x-\theta)/\sigma})^{-1} \ \forall x$, it is equivalent to test

$$H_o^* : G_i(y) = G_{io}(y) \qquad \forall y,$$

and

$$G_{io}(y) = \sum_{j=i}^{2m-1} \binom{2m-1}{j} [(1+e^{-(y-\theta)/\sigma})^{-1}]^j [e^{-(y-\theta)/\sigma}(1+e^{-(y-\theta)/\sigma})^{-1}]^{(2m-1)-j}.$$

To do this, consider the following partition of $(-\infty, \infty)$:

$$I_1 = (-\infty, a], \ I_j = ((j-1)a, ja], \ j = 2,..., k, \ I_{k+1} = (ka, \infty) \tag{3}$$

Let M_j = number of Y_i's that fall in the interval I_j, $j = 1,..., k+1$. Thus, we have

$$P_j^*(\theta,\ \sigma) = \int_{(j-1)a}^{ja} dG_{io}(y) = G_{io}(ja) - G_{io}((j-1)a),$$

and $P_1^*(\theta,\ \sigma) = G_{io}(a)$ and $P_{k+1}^*(\theta,\ \sigma) = 1 - G_{io}(ka)$. So, we reject H_o^* at level of significance α if

$$\chi^{*2} = \sum_{j=1}^{k+1} \frac{(M_j - r\ P_j^*(\theta,\sigma))^2}{r\ P_j^*(\theta,\sigma)} > \chi^2_{1-\alpha,\ k} \qquad (4)$$

If θ and σ are unknown, then we may estimate them using the method of minimum chi-square distance, i.e, we estimate them by $\hat{\theta}$ and $\hat{\sigma}$, which minimize the left hand side of (4). In this case, we lose two degrees of freedom. For more details about this see Lehmann (1999).

4. POWER COMPARISON

In this section, we compare the power of the test statistic χ^{*2} with the power of the test statistic χ^2 based on samples of the same size. To calculate the power of χ^{*2} under H, a distribution under H_1, we need to use simulation. So, we design the following algorithm.

1. Select a sample of size r from H, a distribution under the alternative hypothesis.
2. Classify the sample obtained in step 1 into the $k+1$ subintervals $I_1, I_2, ..., I_{k+1}$, given in (3) to obtain the frequencies $M_1,\ M_2\ ,...,\ M_{k+1}$.
3. Obtain the values of $P_1^*,\ P_2^*, ...,\ P_{k+1}^*$ as follows:

   $$P_1^*(\theta,\ \sigma) = G_{io}(a),\ P_j^*(\theta,\ \sigma) = G_{io}(ja) - G_{io}((j-1)a),\ i = 2, ...,\ k,$$

   and $P_{k+1}^*(\theta,\ \sigma) = 1 - G_{io}(ka)$.
4. Calculate χ^{*2} from equation (1).
5. Repeat the steps (1) - (4), 10,000 times to get $\chi_1^{*2}, ...,\ \chi_{10,000}^{*2}$.
6. Approximate the power of the χ^{*2} test at H as follows

   $$\text{Power of } \chi^{*2}(H) \approx \frac{1}{10,000} \sum_{t=1}^{10,000} I(\chi_t^{*2} > \chi^2_{1-\alpha,k}),$$

where $I(.)$ stands for the indicator function.

5. SIMULATION RESULTS

We approximated the power of each test statistic based on a Monte Carlo simulation of 10,000 iterations according to the algorithm of Section 4. We compared the powers of the two test statistics for different samples sizes, $r = 30, 50, 100$, different set sizes, $2m-1$, where $m = 1, 2, 3, 4$, (m=1 refers to SRS case), different number of intervals: $k = 5, 10, 15$ and different alternative distributions: Normal (0, 1), Laplace (0, 1), Cauchy (0, 1), Uniform (-4, 4), Lognormal (0, 1), exponential (0, 1) and the Student-T with 5 degrees of freedom. We also made the comparison only for the cases when the data are quantified via either minimum, median or maximum. The Simulation results are presented in the Tables (1)-(3).

From the above tables, we make the following remarks:

1. The power is increasing in the sample size r and also in the set size m.
2. The chi-square tests based on the extreme order statistics are more powerful than their counterparts in SRS when compared under samples of the same size.
3. No clear pattern concerning the power and the number of intervals.

6. KULLBACK-LEIBLER INFORMATION

From the simulation results in the previous section, we see that the chi-square test based on the extreme order statistics is more powerful than the chi-square test based on the median. In this section, we try to give an interpretation to this by employing the Kullback-Leibler information number to test $H_o : F(x) = F_o(x)$ for all x against $H_1 : F(x) \neq F_o(x)$ *for some* x. The information theory defines the Kullback-Leibler as follows. Let $f_0(x)$ and $f_1(x)$ be two density functions induced by two hypotheses say H_0 and H_1, respectively. The Kullback-Leibler information number of the two densities $f_0(x)$ and $f_1(x)$, denoted by $I(f_0, f_1)$, is given by

$$I(f_0, f_1) = \int_{-\infty}^{\infty} f_0(x) \log \frac{f_0(x)}{f_1(x)} dx .$$

The quantity $I(f_0, f_1)$ can be interpreted as the mean information per observation under $f_0(x)$ that discriminates in favor of H_0 against H_1. Let $I^*(g_{io}, g_{i1})$ be the Kullback-Leibler information number of the two densities $g_{io}(x)$ and $g_{i1}(x)$ induced by the corresponding equivalent hypotheses H_0^* and H_1^*, respectively.

Table1. $1000 \times$ Power values for SRS and RSS (using first order statistic), $\alpha = 0.05$

H	k=5			k=10			k=15		
	$m=1$					r			
	30	50	100	30	50	100	30	50	100
Normal	80	632	995	115	457	999	62	283	952
Laplace	29	69	299	94	238	656	57	141	529
Cauchy	712	865	985	567	768	972	648	825	981
Uniform	861	976	1000	907	992	1000	920	989	1000
Lognormal	1000	1000	1000	1000	1000	1000	1000	1000	1000
Exponential	1000	1000	1000	1000	1000	1000	1000	1000	1000
Student(5)	42	183	654	60	187	673	36	111	528
	$m=2$					r			
	30	50	100	30	50	100	30	50	100
Normal	495	924	1000	252	795	1000	132	582	999
Laplace	45	131	423	72	207	623	52	131	500
Cauchy	725	915	996	555	804	986	628	868	994
Uniform	897	990	1000	943	997	1000	938	996	1000
Lognormal	1000	1000	1000	1000	1000	1000	1000	1000	1000
Exponential	1000	1000	1000	1000	1000	1000	1000	1000	1000
Student(5)	108	329	822	76	248	794	43	145	654
	$m=3$					r			
	30	50	100	30	50	100	30	50	100
Normal	929	999	1000	746	995	1000	553	973	1000
Laplace	114	272	701	117	282	725	94	216	645
Cauchy	867	982	1000	670	915	1000	781	956	1000
Uniform	890	994	1000	969	1000	1000	941	999	1000
Lognormal	1000	1000	1000	1000	1000	1000	1000	1000	1000
exponential	1000	1000	1000	1000	1000	1000	1000	1000	1000
Student(5)	351	663	982	207	541	965	134	395	916
	$m=4$					r			
	30	50	100	30	50	100	30	50	100
Normal	993	1000	1000	968	1000	1000	912	1000	1000
Laplace	218	453	854	207	421	855	170	338	789
Cauchy	940	996	1000	775	970	1000	870	988	1000
Uniform	928	998	1000	990	1000	1000	976	1000	1000
Lognormal	1000	1000	1000	1000	1000	1000	1000	1000	1000
Exponential	1000	1000	1000	1000	1000	1000	1000	1000	1000
Student(5)	547	859	997	389	753	994	288	640	986

Table 2. $1000\times$ Power values for SRS and RSS (using largest order statistic), $\alpha = 0.05$

H	k=5			k=10			k=15		
	$m=1$					r			
	30	50	100	30	50	100	30	50	100
Normal	83	621	997	111	454	990	61	280	952
Laplace	28	72	294	79	230	665	59	145	536
Cauchy	716	868	986	647	820	983	573	827	986
Uniform	916	978	1000	909	992	1000	913	988	1000
Lognormal	1000	1000	1000	1000	1000	1000	1000	1000	1000
Exponential	1000	1000	1000	1000	1000	1000	1000	1000	1000
Student(5)	40	176	653	61	185	678	41	117	524
	$m=2$					r			
	30	50	100	30	50	100	30	50	100
Normal	497	926	1000	258	792	1000	138	591	999
Laplace	45	126	425	77	202	637	50	140	514
Cauchy	721	917	998	627	863	997	548	870	988
Uniform	920	991	1000	944	997	1000	938	996	1000
Lognormal	1000	1000	1000	1000	1000	1000	1000	1000	1000
Exponential	1000	1000	1000	1000	1000	1000	1000	1000	1000
Student(5)	104	334	826	71	259	807	45	152	673
	$m=3$					r			
	30	50	100	30	50	100	30	50	100
Normal	929	999	1000	735	994	1000	547	971	1000
Laplace	123	268	695	112	279	727	87	220	645
Cauchy	869	980	1000	775	956	1000	670	955	1000
Uniform	891	1000	1000	968	999	1000	956	999	1000
Lognormal	1000	1000	1000	1000	1000	1000	1000	1000	1000
Exponential	1000	1000	1000	1000	1000	1000	1000	1000	1000
Student(5)	333	656	980	203	519	958	130	377	904
	$m=4$					r			
	30	50	100	30	50	100	30	50	100
Normal	992	1000	1000	973	1000	1000	908	1000	1000
Laplace	214	454	856	215	431	851	162	338	792
Cauchy	941	996	10000	865	988	1000	771	969	1000
Uniform	986	1000	1000	986	1000	1000	979	1000	1000
Lognormal	1000	1000	1000	1000	1000	1000	1000	1000	1000
Exponential	1000	1000	1000	1000	1000	1000	1000	1000	1000
Student(5)	558	879	1000	414	768	997	306	662	987

Table 3. $1000 \times$ Power values for SRS and RSS (using median), $\alpha = 0.05$

H	k=5			k=10			k=15		
	$m=1$					r			
	30	50	100	30	50	100	30	50	100
Normal	82	623	997	116	466	990	60	290	948
Laplace	26	72	300	97	238	657	60	142	525
Cauchy	701	863	986	654	822	982	574	774	972
Uniform	862	977	1000	907	991	1000	909	991	1000
Lognormal	1000	1000	1000	1000	1000	1000	1000	1000	1000
Exponential	1000	1000	1000	1000	1000	1000	1000	1000	1000
Student(5)	41	181	646	57	184	682	37	124	532
	$m=2$					r			
	30	50	100	30	50	100	30	50	100
Normal	4	14	108	38	216	901	14	94	767
Laplace	18	31	70	78	241	761	55	133	588
Cauchy	663	838	925	701	867	955	739	865	944
Uniform	907	984	1000	956	997	1000	966	996	1000
Lognormal	1000	1000	1000	1000	1000	1000	1000	1000	1000
Exponential	1000	1000	1000	1000	1000	1000	1000	1000	1000
Student(5)	8	15	61	30	105	574	17	51	404
	$m=3$					r			
	30	50	100	30	50	100	30	50	100
Normal	2	3	4	27	148	773	6	56	645
Laplace	2	3	4	58	225	789	20	83	559
Cauchy	278	387	615	430	591	803	497	576	786
Uniform	916	988	1000	970	1000	1000	968	999	1000
Lognormal	1000	1000	1000	1000	1000	1000	1000	1000	1000
Exponential	1000	1000	1000	1000	1000	1000	1000	1000	1000
Student(5)	1	3	3	14	69	3	6	24	315
	$m=4$					r			
	30	50	100	30	50	100	30	50	100
Normal	2	1	1	11	128	705	2	29	544
Laplace	1	3	5	42	217	798	12	49	523
Cauchy	150	276	325	231	303	510	261	346	557
Uniform	901	987	1000	980	1000	1000	981	997	1000
Lognormal	1000	1000	1000	1000	1000	1000	1000	1000	1000
Exponential	1000	1000	1000	1000	1000	1000	1000	1000	1000
Student(5)	2	1	3	7	57	2	2	12	229

Then

$$I^*(g_{io}, g_{i1}) = \int_{-\infty}^{\infty} g_{io}(y) \log \frac{g_{io}(y)}{g_{i1}(y)} dy,$$

where $g_{io}(y)$ and $g_{i1}(y)$ are the density functions of the i^{th} order statistics from samples of size $2m-1$ from the densities $f_0(x)$ and $f_1(x)$, respectively. So

$$I^*(g_{io}, g_{i1}) = \int_{-\infty}^{\infty} g_{io}(y) \log \frac{g_{io}(y)}{g_{i1}(y)} dy,$$

where

$$g_{io}(y) = c_i F_o(y)^{i-1} (1 - F_o(y))^{2m-1-i} f_o(y),$$

$$g_{i1}(y) = c_i F_1(y)^{i-1} (1 - F_1(y))^{2m-1-i} f_1(y)$$

and

$$c_i = \frac{(2m-1)!}{(i-1)!(2m-1-i)!}.$$

Using numerical integration, the values of $I(f_0, f_1)$ and $I^*(g_{io}, g_{i1})$ are calculated and presented in Table 4 for different distributions and different values of m and j. It can be noted from Table 4, that the mean information per observation under $g_{io}(y)$ ($i = 1..., m\text{-}1, m+1, ..., 2m\text{-}1$) that discriminates in favor of H_0^* against H_1^* is larger than the mean information per observation under $f_0(x)$ that discriminates in favor of H_0 against H_1. Comparing with the simulation results for the case when $i = m$ (median case), the results show that the chi-square test based on a SRS is better than the chi-square test based on a RSS. In fact this agrees with the simulation results in the previous section.

7. CONCLUSION

In this paper, we have proposed a chi-square test for goodness-of-fit when the data is collected via an RSS technique. We gave our attention to those RSS schemes which quantify only one order statistic namely minimum, median or maximum. Since it is easier for the experimenter to detect the extreme order statistics by visual inspection, then this makes the method applicable in real situation. Moreover, the theory developed could be extended easily to other distributions.

Table 4. Kullback-Leibler information $I(f_0, f_1)$ (upper cell) and $I^*(g_{io}, g_{i1})$ (lower cell) for different distributions and different m and j.

F_0 and F_1	m/j	1	2	3	4	5
Logistic and Normal	2	0.542	0.542	0.542		
		0.773	0.399	0.773		
	3	0.542	0.542	0.542	0.542	0.542
		1.164	0.543	0.360	0.543	1.164
Logistic and Cauchy	2	0.140	0.140	0.140		
		0.195	0.060	0.195		
	3	0.140	.140	0.140	0.140	0.140
		0.302	0.079	0.039	0.079	0.302
Logistic and Laplace	2	0.079	0.079	0.079		
		0.096	0.129	0.096		
	3	0.079	0.079	0.079	0.079	0.079
		0.102	0.151	0.165	0.151	0.102
Logistic and Student (5)	2	0.109	0.109	0.109		
		0.146	0.149	0.146		
	3	0.109	0.109	0.109	0.109	0.109
		0.189	0.189	0.170	0.189	0.189
Lognormal and Exponential (1)	2	0.119	0.119	0.119		
		0.303	0.215	0.125		
	3	0.119	0.119	0.119	0.119	0.119
		0.472	0.328	0.278	0.311	0.159
LogNormal and Chi-Square(5)	2	1.459	1.459	1.459		
		2.228	3.297	2.979		
	3	1.459	1.459	1.459	1.459	1.459
		2.528	4.277	5.209	5.171	3.867

ACKNOWLEDGEMENTS

The authors would like to thank the Universiti Kebangsaan Malaysia for supporting this work with code project UKM-GUP-TK-08-16-061.

REFERENCES

Al-Odat M. T. and Al-Saleh M. F. (2001). A variation of ranked set sampling. *Journal of Applied Statistical Science*. **10**, 137–146.

Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1992). *A First Course in Order Statistics.* John Wiley and Sons, New York.

Al-Saleh M. F and Alomari A. I. (2002). Multistage ranked set sampling. *Journal of Statistical planning and Inference*. **102**, 273–286.

Balakrishnan, N. (1992). *Handbook of the logistic distribution*. New York, Marcel Dekker Inc.

Bhoj, D. S. (1997). Estimation of parameters of the extreme value distribution using ranked set sampling. *Commun. Statist-Theory Meth*. **26(3),** 653-667.

Chen, Z. (2000). On ranked-set sample quantiles and their applications. *J. Statist. Plann.Inference*. **83**, 125-135.

Dell, D. R. and Clutter, J. L. (1972). Ranked set sampling theory with order statistics background. *Biometrics*. **28**, 545-555.

Johnson, G. D., Patil, G. P. and Sinha, A. K. (1993). Ranked set sampling for vegetation research. *Abstracta Botanica*. **17**, 87-102.

Kaur, A., Patil, G. P., Shirk, S. J. and Taillie, C. (1996). Environmental sampling with a concomitant variable: a comparison between ranked set sampling and stratified simple random sampling. *Journal of Applied Statistics*. **23**, 231-256.

Lehmann E. L. (1999). *Elements of Large- Sample Theory*. Springer-Verlag, New York, Inc.

McIntyre, G. A. (1952). A method of unbiased selective sampling using ranked sets. *Australian Journal of Agricultural Research*. **3**, 385-390.

Muttlak, H. A. (1997). Median ranked set sampling. *Journal of applied statistics science*. **6**, 245-255.

Muttlak, H. A. (2003). Modified ranked set sampling methods. *Pak. J. of statist*. **3**, 315-323.

Patil, G. P. and Taillie, C. (1993). Statistical evaluation of the attainment of interim cleanup standards hazardous waste sites. *Journal of Environmental and Ecological Statistics.* **1**, 117-140.

Samawi, H. M., Mohmmad, S. and Abu-Dayyeh, W. (1996). Estimation the population mean using extreme ranked set sampling. *Biometric*al *Journal*. **38**, 577-586.

Samawi, H. M. and Al-Sagheer, O. A. (2001). On the estimation of the distribution function using extreme and median ranked set sampling. *Biometrical Journal*. **43(3)**, 357-373.

Stockes S. L. and Sager T. W. (1988). Characterization of a Ranked -Set Sample with Application to Estimating Distribution Functions. *Journal of the American Statistical Association*. **83(402),** pp. 374-381.

Takahasi, K. and Wakitmoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*. **20**, 1-31.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 771–785.

BAYESIAN INFERENCE FOR SEASONAL ARMA MODELS: A GIBBS SAMPLING APPROACH

Mohamed A. Ismail
Statistics Department, Cairo University,
Giza, Egypt, and
Information and Decision Support Center,
The Egyptian Cabinet, Cairo, Egypt
E-mail: m.ismail@idsc.net.eg

Ayman A. Amin
Statistics & Insurance Department, Menoufia University,
Menoufia, Egypt
E-mail: aymanamin2008@gmail.com

ABSTRACT

This paper develops a Bayesian inference for a multiplicative seasonal ARMA model by implementing a fast, easy and accurate Gibbs sampling algorithm. The proposed algorithm does not involve any Metropolis-Hastings generation but is generated from normal and inverse gamma distributions. The proposed algorithm is illustrated using simulated examples and a real data set.

Keywords: Multiplicative seasonal ARMA models; prior distribution; posterior distribution; Gibbs sampling; Federal Reserve Board Production Index.

1. INTRODUCTION

Seasonal ARMA modeling of time series has been successfully applied in a great number of fields including economic forecasting. Bayesian analysis of ARMA type models is difficult even for non seasonal models since the likelihood function is analytically intractable, which causes problems in prior specification and posterior analysis. Different solutions including Markov Chain Monte Carlo (MCMC) methods have been suggested in the literature for the Bayesian time series analysis. Several authors have considered Bayesian analysis of ARMA models e.g. Newbold (1973), Monahan (1983), Broemeling and Shaarawy (1984), Shaarawy and Ismail (1987) and Marriott and Smith (1992) among others.

Bayesian time series analysis has been advanced by the emergence of MCMC methods especially; the Gibbs sampling method. Assuming a prior distribution on the initial observations and initial errors, Chib and Greenberg (1994) and Marriott et al. (1996) developed Bayesian analysis for ARMA models using MCMC technique. Barnett et al. (1996, 1997) used MCMC methodology to estimate the multiplicative seasonal autoregressive and ARMA model. Their algorithm was based on sampling functions of the partial autocorrelations. A virtue of their approach is that one for one draws of each partial autocorrelation can be obtained but at the cost of a more complicated algorithm.

Recently, Ismail (2003a, 2003b) used Gibbs sampling algorithm to analyze multiplicative seasonal autoregressive and seasonal moving average models. His algorithm was based on approximating the likelihood function via estimating the unobserved errors. Then, the approximate likelihood is used to derive the conditional distributions required for implementing Gibbs sampler. Rather than restricting the parameters space to satisfy stationarity and

invertibility conditions as in Barrnett et al. (1997) and Marriott et al. (1996) among others, the process could be made stationary and invertible by choosing the hyperparameters which ensure that the prior for the model coefficients lie in the stationarity and invertibility region. The latter approach was used by Broemeling (1985), McCulloch and Tsay (1994) and Ismail (2003a, 2003b) among others and is going to be used in this paper.

The objective of this paper is to extend Ismail's (2003a, 2003b) algorithm to multiplicative seasonal ARMA models. The proposed algorithm does not involve any Metropolis-Hastings iteration which is an advantage over other algorithms in the literature. In addition, our analysis is unconditional on the initial values, that is we assume that the series starts at time t = 1 with unknown initial observations and errors. Moreover, various features of the SARMA models, which may be complicated to check in the classical framework, may be routinely tested in the sampling based Bayesian framework. As an example, there is often interest in testing the significance of interaction parameters which are the product of the nonseasonal and seasonal coefficients in the model. The proposed algorithm can easily construct confidence intervals for interaction parameters and therefore test their significance.

The paper is organized as follows. Section 2 briefly describes the multiplicative SARMA model. Section 3 is devoted to summarizing posterior analysis and the full conditional posterior distributions of the parameters. The implementation details of the proposed algorithm including convergence monitoring are given in section 4. The proposed methodology is illustrated in section 5 using simulated examples and Federal Reserve Board Production Index. Finally, the conclusions are given in Section 6**.**

2. THE MULTIPLICATIVE SEASONAL ARMA MODEL

A time series $\{\mathbf{y_t}\}$ is said to be generated by a multiplicative seasonal ARMA model of orders p, q, P and Q, denoted by SARMA(p,q)(P,Q)s, if it satisfies

$$\boldsymbol{\phi_p}(\mathbf{B})\,\boldsymbol{\Phi_{Ps}}(\mathbf{B^s})\mathbf{y_t} = \boldsymbol{\theta_q}(\mathbf{B})\,\boldsymbol{\Theta_{Qs}}(\mathbf{B^s})\boldsymbol{\varepsilon_t} \qquad (\mathbf{1})$$

where $\{\boldsymbol{\varepsilon_t}; \mathbf{t} \in \mathbf{I}\}$ is a sequence of independent normal variates with zero mean and variance $\boldsymbol{\sigma^2} > \mathbf{0}$, and I is the set of integers. The backshift operator B is defined such that $\mathbf{B^r}\,\mathbf{y_t} = \mathbf{y_{t-r}}$, s is the number of seasons in the year. The nonseasonal autoregressive polynomial is $\boldsymbol{\phi_p}(\mathbf{B}) = \left(\mathbf{1} - \boldsymbol{\phi_1}\mathbf{B} - \boldsymbol{\phi_2}\mathbf{B^2} - \ ... - \boldsymbol{\phi_p}\mathbf{B^p}\right)$ with order p, the nonseasonal moving average polynomial is $\boldsymbol{\theta_q}(\mathbf{B}) = \left(\mathbf{1} + \boldsymbol{\theta_1}\mathbf{B} + \boldsymbol{\theta_2}\mathbf{B^2} + \ ... + \boldsymbol{\theta_q}\mathbf{B^q}\right)$ with order q, the seasonal autoregressive polynomial is $\boldsymbol{\Phi_{Ps}}(\mathbf{B^s}) = (\mathbf{1} - \boldsymbol{\Phi_{1s}}\mathbf{B^s} - \boldsymbol{\Phi_{2s}}\mathbf{B^{2s}} - \ ... - \boldsymbol{\Phi_{Ps}}\mathbf{B^{Ps}})$ with order P, and $\boldsymbol{\Theta_{Qs}}(\mathbf{B^s}) = \left(\mathbf{1} + \boldsymbol{\Theta_{1s}}\mathbf{B^s} + \boldsymbol{\Theta_{2s}}\mathbf{B^{2s}} + \ ... + \boldsymbol{\Theta_{Qs}}\mathbf{B^{Qs}}\right)$ is the seasonal moving average polynomial with order Q.

The nonseasonal and seasonal autoregressive coefficients are $\boldsymbol{\phi} = (\boldsymbol{\phi_1}\boldsymbol{\phi_2}\ ...\ \boldsymbol{\phi_P})^{\mathbf{T}}$ and $\boldsymbol{\Phi} = (\boldsymbol{\Phi_{1s}}\boldsymbol{\Phi_{2s}}\ ...\ \boldsymbol{\Phi_{Ps}})^{\mathbf{T}}$, and the nonseasonal and seasonal moving averages coefficients are $\boldsymbol{\theta} = (\boldsymbol{\theta_1}\boldsymbol{\theta_2}\ ...\ \boldsymbol{\theta_q})^{\mathbf{T}}$ and $\boldsymbol{\Theta} = (\boldsymbol{\Theta_{1s}}\boldsymbol{\Theta_{2s}}\ ...\ \boldsymbol{\Theta_{Qs}})^{\mathbf{T}}$. Each of the nonseasonal and seasonal orders p, q, P and Q is always less than or equal to the number of seasons in the year s. The time series $\{\mathbf{y_t}; \mathbf{t} \in \mathbf{I}\}$ is assumed to start at time t = 1 with unknown initial values $\mathbf{y_0} = (\mathbf{y_0}\ \mathbf{y_{-1}}\ ...\ \mathbf{y_{1-p-Ps}})$ and unknown initial errors $\boldsymbol{\varepsilon_0} = (\boldsymbol{\varepsilon_0}\ \boldsymbol{\varepsilon_{-1}}\ ...\ \boldsymbol{\varepsilon_{1-q-Qs}})$.

The model (1) can be written as:

$$\mathbf{y_t} = \sum_{\mathbf{i=1}}^{\mathbf{p}} \boldsymbol{\phi_i}\,\mathbf{y_{t-i}} + \sum_{\mathbf{j=1}}^{\mathbf{P}} \boldsymbol{\Phi_j}\,\mathbf{y_{t-js}} + \sum_{\mathbf{i=1}}^{\mathbf{p}} \quad \sum_{\mathbf{j=1}}^{\mathbf{P}} \boldsymbol{\phi_i}\boldsymbol{\Phi_j}\,\mathbf{y_{t-i-js}} + \boldsymbol{\varepsilon_t} +$$

$$\sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \sum_{j=1}^{Q} \Theta_j \varepsilon_{t-js} + \sum_{i=1}^{q} \sum_{j=1}^{Q} \theta_i \Theta_j \varepsilon_{t-i-js} = \mathbf{X_t}\boldsymbol{\beta_1} + \boldsymbol{\Lambda_t}\boldsymbol{\beta_2} \quad (2)$$

where,

$$\mathbf{X_t} = (y_{t-1}, \dots, y_{t-p}; \mathbf{O_1}; y_{t-s}, y_{t-s-1}, \dots, y_{t-s-p}; \mathbf{O_1}; \dots; y_{t-Ps}, y_{t-Ps-1}, \dots, y_{t-Ps-p}),$$
$$\boldsymbol{\beta_1} = (\phi_1, \dots, \phi_p; \mathbf{O_1}; \Phi_1, -\phi_1\Phi_1, \dots, -\phi_p\Phi_1; \mathbf{O_1}; \dots; \Phi_P, -\phi_1\Phi_P, \dots, -\phi_p\Phi_P)^T,$$
$$\boldsymbol{\Lambda_t} = (\varepsilon_{t-1}, \dots, \varepsilon_{t-q}; \mathbf{O_2}; \varepsilon_{t-s}, \varepsilon_{t-s-1}, \dots, \varepsilon_{t-s-q}; \mathbf{O_2}; \dots; \varepsilon_{t-Qs}, \varepsilon_{t-Qs-1}, \dots, \varepsilon_{t-Qs-q}),$$
$$\boldsymbol{\beta_2} = (\theta_1, \dots, \theta_q; \mathbf{O_2}; \Theta_1, \theta_1\Theta_1, \dots, \theta_q\Theta_1; \mathbf{O_2}; \dots; \Theta_Q, \theta_1\Theta_Q, \dots, \theta_q\Theta_Q)^T, \quad (3)$$

and $\mathbf{O_1}$ and $\mathbf{O_2}$ are (s – p – 1) and (s – q – 1) row vectors of zeros respectively. Model (2) shows that the multiplicative SARMA model can be written as ARMA model of order p+Ps and q+Qs with some zero coefficients and some coefficients that are products of nonseasonal and seasonal coefficients. Thus, the model is nonlinear in $\boldsymbol{\phi}, \boldsymbol{\Phi}, \boldsymbol{\theta}$ **and** $\boldsymbol{\Theta}$ which complicates the Bayesian analysis. However, the following sections explain how Gibbs sampling technique can facilitate the analysis. The SARMA model (2) is stationary if the roots of the polynomials $\boldsymbol{\phi_p}(\mathbf{B})$ and $\boldsymbol{\Phi_{Ps}}(\mathbf{B^s})$ lie outside the unit circle, and when the roots of the polynomials $\boldsymbol{\theta_q}(\mathbf{B})$ and $\boldsymbol{\Theta_{Qs}}(\mathbf{B^s})$ lie outside the unit circle the process is invertibile. For more details about the properties of seasonal ARMA models see Box and Jenkins (1976).

3. POSTERIOR ANALYSIS

3.1 Liklihood Function

Suppose that $\underline{\mathbf{y}} = (\mathbf{y_1}\ \mathbf{y_2}\ \dots\ \mathbf{y_n})$ is a realization from the multiplicative SARMA model (2), assuming that the random errors $\boldsymbol{\mathcal{E}_t} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma^2})$ and employing a straightforward random variable transformation from $\boldsymbol{\mathcal{E}_t}$ to $\mathbf{y_t}$, the likelihood function $\mathbf{L}\left(\boldsymbol{\phi}, \boldsymbol{\Phi}, \boldsymbol{\theta}, \boldsymbol{\Theta}, \boldsymbol{\sigma^2}, \mathbf{y_0}, \boldsymbol{\varepsilon_0} \mid \underline{\mathbf{y}}\right) = \boldsymbol{\mathcal{L}}$ is given by:

$$\boldsymbol{\mathcal{L}} \propto (\boldsymbol{\sigma^2})^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{n} \varepsilon_t^{\,2}\right) \quad (4)$$

where,

$$\varepsilon_t = y_t - \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{P} \Phi_j y_{t-js} + \sum_{i=1}^{p}\sum_{j=1}^{P} \phi_i\Phi_j y_{t-i-js} - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} - \sum_{j=1}^{Q} \Theta_j \varepsilon_{t-js} - \sum_{i=1}^{q}\sum_{j=1}^{Q} \theta_i\Theta_j \varepsilon_{t-i-js} \quad (5)$$

The likelihood function (4) is a complicated function in the parameters $, \boldsymbol{\Phi}, \boldsymbol{\theta}, \boldsymbol{\Theta}, \mathbf{y_0}$ **and** $\boldsymbol{\varepsilon_0}$. Suppose the errors are estimated recursively as:

$$e_t = y_t - \sum_{i=1}^{p} \hat{\phi}_i y_{t-i} - \sum_{j=1}^{P} \hat{\Phi}_{js} y_{t-js} + \sum_{i=1}^{p}\sum_{j=1}^{P} \hat{\phi}_i\hat{\Phi}_{js} y_{t-i-js} - \sum_{i=1}^{q} \hat{\theta}_i e_{t-i} - \sum_{j=1}^{Q} \hat{\Theta}_{js} e_{t-js} - \sum_{i=1}^{q}\sum_{j=1}^{Q} \hat{\theta}_i\hat{\Theta}_{js} e_{t-i-js}, \quad (6)$$

where $\widehat{\boldsymbol{\phi}}_{\mathbf{i}} \in \mathbf{R^p}, \widehat{\boldsymbol{\Phi}}_{\mathbf{jS}} \in \mathbf{R^P}, \widehat{\boldsymbol{\theta}}_{\mathbf{i}} \in \mathbf{R^q}, \textbf{and } \widehat{\boldsymbol{\Theta}}_{\mathbf{jS}} \in \mathbf{R^Q}$ are sensible estimates. Several estimation methods, such as the Innovations Substitution (IS) method proposed by Koreisha and Pulkila (1990), give consistent estimates for $\boldsymbol{\phi}, \boldsymbol{\Phi}, \boldsymbol{\theta}$ **and** $\boldsymbol{\Theta}$. The idea of the IS method is to fit a long autoregressive model to the series and obtain the residuals. Then appropriate lagged residuals are substituted into SARMA model (2). Finally, the parameters are estimated using ordinary least squares method. Substituting the residuals in the likelihood function (4) results in an approximate likelihood function

$$\mathcal{L}^* \propto (\boldsymbol{\sigma^2})^{-\frac{\mathbf{n}}{2}} \exp\left(-\frac{\mathbf{1}}{\mathbf{2\sigma^2}} \sum_{\mathbf{t=1}}^{\mathbf{n}} \boldsymbol{\varepsilon}_{\mathbf{t}}^{*\mathbf{2}}\right)$$

$$= (\boldsymbol{\sigma^2})^{-\frac{\mathbf{n}}{2}} \exp\left(-\frac{\mathbf{1}}{\mathbf{2\sigma^2}} (\mathbf{y} - \mathbf{X\beta_1} - \widehat{\boldsymbol{\Lambda}}\boldsymbol{\beta_2})^{\mathbf{T}} (\mathbf{y} - \mathbf{X\beta_1} - \widehat{\boldsymbol{\Lambda}}\boldsymbol{\beta_2})\right) \quad (7)$$

where,

$$\boldsymbol{\varepsilon}_{\mathbf{t}}^{*} = \mathbf{y_t} - \sum_{\mathbf{i=1}}^{\mathbf{p}} \boldsymbol{\phi}_{\mathbf{i}} \mathbf{y_{t-i}} - \sum_{\mathbf{j=1}}^{\mathbf{P}} \boldsymbol{\Phi}_{\mathbf{jS}} \mathbf{y_{t-js}} + \sum_{\mathbf{i=1}}^{\mathbf{p}} \sum_{\mathbf{j=1}}^{\mathbf{P}} \boldsymbol{\phi}_{\mathbf{i}} \boldsymbol{\Phi}_{\mathbf{js}} \mathbf{y_{t—i-js}} - \sum_{\mathbf{i=1}}^{\mathbf{q}} \boldsymbol{\theta}_{\mathbf{i}} \mathbf{e_{t-i}} -$$

$$\sum_{\mathbf{j=1}}^{\mathbf{Q}} \boldsymbol{\Theta}_{\mathbf{jS}} \mathbf{e_{t-js}} - \sum_{\mathbf{i=1}}^{\mathbf{q}} \sum_{\mathbf{j=1}}^{\mathbf{Q}} \boldsymbol{\theta}_{\mathbf{i}} \boldsymbol{\Theta}_{\mathbf{js}} \mathbf{e_{t—i-js}}. \quad (8)$$

$\boldsymbol{\beta_1}, \boldsymbol{\beta_2}$ are defined in (3), and $\widehat{\boldsymbol{\Lambda}}$ is a $\mathbf{n} \times (\mathbf{q} + \mathbf{Qs})$ matrix with t[th] row

$$\widehat{\boldsymbol{\Lambda}}_{\mathbf{t}} = (\mathbf{e_{t-1}}, \dots, \mathbf{e_{t-q}}; \mathbf{O_2}; \mathbf{e_{t-s}}, \mathbf{e_{t-s-1}}, \dots, \mathbf{e_{t-s-q}}; \mathbf{O_2}; \dots; \mathbf{e_{t-Qs}}, \mathbf{e_{t-Qs-1}}, \dots, \mathbf{e_{t-Qs-q}}),$$

where, $\mathbf{O_2}$ is a (s – q – 1) row vector of zeros.

3.2 Prior Specification

For multiplicative SARMA models, suppose that, given the error variance parameter $\boldsymbol{\sigma^2}$, the parameters $\boldsymbol{\phi}, \boldsymbol{\Phi}, \boldsymbol{\theta}, \boldsymbol{\Theta}, \mathbf{y_0}$ **and** $\boldsymbol{\varepsilon_0}$ are independent apriori, i.e.

$$\boldsymbol{\xi}(\boldsymbol{\phi}, \boldsymbol{\Phi}, \boldsymbol{\theta}, \boldsymbol{\Theta}, \boldsymbol{\sigma^2}, \mathbf{y_0}, \boldsymbol{\varepsilon_0}) = \boldsymbol{\xi}(\boldsymbol{\phi}|\boldsymbol{\sigma^2}) \times \boldsymbol{\xi}(\boldsymbol{\Phi}|\boldsymbol{\sigma^2}) \times \boldsymbol{\xi}(\boldsymbol{\theta}|\boldsymbol{\sigma^2}) \times \boldsymbol{\xi}(\boldsymbol{\Theta}|\boldsymbol{\sigma^2}) \times$$
$$\boldsymbol{\xi}(\mathbf{y_0}|\boldsymbol{\sigma^2}) \times \boldsymbol{\xi}(\boldsymbol{\varepsilon_0}|\boldsymbol{\sigma^2}) \times \boldsymbol{\xi}(\boldsymbol{\sigma^2})$$
$$= \mathbf{N_p}(\boldsymbol{\mu_\phi}, \boldsymbol{\sigma^2}\textstyle\sum_{\boldsymbol{\phi}}) \times \mathbf{N_P}(\boldsymbol{\mu_\Phi}, \boldsymbol{\sigma^2}\textstyle\sum_{\boldsymbol{\Phi}}) \times \mathbf{N_q}(\boldsymbol{\mu_\theta}, \boldsymbol{\sigma^2}\textstyle\sum_{\boldsymbol{\theta}}) \times \mathbf{N_Q}(\boldsymbol{\mu_\Theta}, \boldsymbol{\sigma^2}\textstyle\sum_{\boldsymbol{\Theta}})$$
$$\times \mathbf{N_{p+Ps}}(\boldsymbol{\mu_{y_0}}, \boldsymbol{\sigma^2}\textstyle\sum_{\mathbf{y_0}}) \times \mathbf{N_{q+Qs}}(\boldsymbol{\mu_{\varepsilon_0}}, \boldsymbol{\sigma^2}\textstyle\sum_{\boldsymbol{\varepsilon_0}})$$
$$\times \mathbf{IG}(\frac{\boldsymbol{\nu}}{\mathbf{2}}, \frac{\boldsymbol{\lambda}}{\mathbf{2}}) \quad (9)$$

where, $\mathbf{N_r}(\mu, \Delta)$ is the r-variate normal distribution with mean μ and variance Δ and IG(α, β) is the inverse gamma distribution with parameters α and β. Such prior distribution is a normal inverse gamma distribution and can then be written as:

$$\boldsymbol{\xi}(\boldsymbol{\phi}, \boldsymbol{\Phi}, \boldsymbol{\theta}, \boldsymbol{\Theta}, \boldsymbol{\sigma^2}, \mathbf{y_0}, \boldsymbol{\varepsilon_0}) \propto (\boldsymbol{\sigma^2})^{-\left(\frac{\upsilon+2p+P+Ps+2q+Q+Qs}{2}+1\right)} \exp\{-\frac{\mathbf{1}}{\mathbf{2\sigma^2}}[\boldsymbol{\lambda} + (\boldsymbol{\phi} - \boldsymbol{\mu_\phi})^{\mathbf{T}} \textstyle\sum_{\boldsymbol{\phi}}^{-1} \times$$
$$(\boldsymbol{\phi} - \boldsymbol{\mu_\phi}) + (\boldsymbol{\Phi} - \boldsymbol{\mu_\Phi})^{\mathbf{T}} \textstyle\sum_{\boldsymbol{\Phi}}^{-1} (\boldsymbol{\Phi} - \boldsymbol{\mu_\Phi}) + (\boldsymbol{\theta} - \boldsymbol{\mu_\theta})^{\mathbf{T}} \textstyle\sum_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu_\theta}) +$$
$$(\boldsymbol{\Theta} - \boldsymbol{\mu_\Theta})^{\mathbf{T}} \textstyle\sum_{\boldsymbol{\Theta}}^{-1} (\boldsymbol{\Theta} - \boldsymbol{\mu_\Theta}) + (\mathbf{y_0} - \boldsymbol{\mu_{y_0}})^{\mathbf{T}} \textstyle\sum_{\mathbf{y_0}}^{-1} (\mathbf{y_0} - \boldsymbol{\mu_{y_0}}) + (\boldsymbol{\varepsilon_0} - \boldsymbol{\mu_{\varepsilon_0}})^{\mathbf{T}} \times$$
$$\textstyle\sum_{\boldsymbol{\varepsilon_0}}^{-1} (\boldsymbol{\varepsilon_0} - \boldsymbol{\mu_{\varepsilon_0}})]\} \quad (10)$$

The prior distribution (10) is chosen for several reasons, it is flexible enough to be used in numerous applications, it also facilitates the mathematical calculations and it is, at least conditionally, a conjugate prior. Multiplying the joint prior distribution (10) by the approximate likelihood function (7) results in the joint posterior $\boldsymbol{\xi}\left(\boldsymbol{\phi}, \boldsymbol{\Phi}, \boldsymbol{\theta}, \boldsymbol{\Theta}, \boldsymbol{\sigma^2}, \mathbf{y_0}, \boldsymbol{\varepsilon_0} | \underline{\mathbf{y}}\right)$ which may be written as

$$\boldsymbol{\xi}\left(\boldsymbol{\phi}, \boldsymbol{\Phi}, \boldsymbol{\theta}, \boldsymbol{\Theta}, \boldsymbol{\sigma^2}, \mathbf{y_0}, \boldsymbol{\varepsilon_0} | \underline{\mathbf{y}}\right) \propto (\boldsymbol{\sigma^2})^{-\left(\frac{\mathbf{n+v+2p+P+Ps+2q+Q+Qs}}{\mathbf{2}}+\mathbf{1}\right)} \exp\{-\frac{\mathbf{1}}{\mathbf{2\sigma^2}}[\boldsymbol{\lambda} + (\boldsymbol{\phi}-\boldsymbol{\mu_\phi})^{\mathbf{T}}\boldsymbol{\Sigma_\phi}^{-1} \times (\boldsymbol{\phi}-\boldsymbol{\mu_\phi}) + (\boldsymbol{\Phi}-\boldsymbol{\mu_\Phi})^{\mathbf{T}}\boldsymbol{\Sigma_\Phi}^{-1}(\boldsymbol{\Phi}-\boldsymbol{\mu_\Phi}) + (\boldsymbol{\theta}-\boldsymbol{\mu_\theta})^{\mathbf{T}}\boldsymbol{\Sigma_\theta}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu_\theta}) + (\boldsymbol{\Theta}-\boldsymbol{\mu_\Theta})^{\mathbf{T}}\boldsymbol{\Sigma_\Theta}^{-1}(\boldsymbol{\Theta}-\boldsymbol{\mu_\Theta}) + (\mathbf{y_0}-\boldsymbol{\mu_{y_0}})^{\mathbf{T}}\boldsymbol{\Sigma_{y_0}}^{-1}(\mathbf{y_0}-\boldsymbol{\mu_{y_0}}) + (\boldsymbol{\varepsilon_0}-\boldsymbol{\mu_{\varepsilon_0}})^{\mathbf{T}} \times \boldsymbol{\Sigma_{\varepsilon_0}}^{-1}(\boldsymbol{\varepsilon_0}-\boldsymbol{\mu_{\varepsilon_0}}) + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta_1}-\widehat{\boldsymbol{\Lambda}}\boldsymbol{\beta_2})^{\mathbf{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta_1}-\widehat{\boldsymbol{\Lambda}}\boldsymbol{\beta_2})]\} \quad (11)$$

3.3 Full Conditional Distributions

The conditional posterior distributions for each of the unknown parameters is obtained from the joint posterior distribution (11) by grouping together terms in the joint posterior that depend on this parameter, and finding the appropriate normalizing constant to form a proper density. In this study all conditional posterior are normal and inverse gamma distributions for which sampling techniques exist.

3.3.1 The Conditional Posterior of $\boldsymbol{\phi}$:

The conditional posterior of $\boldsymbol{\phi}$ is

$$\boldsymbol{\phi^{j+1}} \sim \boldsymbol{\xi}(\boldsymbol{\phi^{j+1}}|\mathbf{y}, \boldsymbol{\Phi^j}, \boldsymbol{\theta^j}, \boldsymbol{\Theta^j}, (\boldsymbol{\sigma^2})^{\mathbf{j}}, \mathbf{y_0}^{\mathbf{j}}, \boldsymbol{\varepsilon_0}^{\mathbf{j}}) = \mathbf{N}(\boldsymbol{\mu}^*_{\boldsymbol{\phi}}, \boldsymbol{\upsilon}^*_{\boldsymbol{\phi}}),$$

where

$$\boldsymbol{\mu}^*_{\boldsymbol{\phi}} = \left[\left(\mathbf{H^T H} + \boldsymbol{\Sigma_\phi}^{-1}\right)(\boldsymbol{\Sigma_\phi}^{-1}\boldsymbol{\mu_\phi} + \mathbf{H^T}(\mathbf{y} - \mathbf{L}\boldsymbol{\Phi} - \widehat{\boldsymbol{\Lambda}}\boldsymbol{\beta_2}))\right], \quad \boldsymbol{\upsilon}^*_{\boldsymbol{\phi}} = \boldsymbol{\sigma^2}\left(\mathbf{H^T H} + \boldsymbol{\Sigma_\phi}^{-1}\right)^{-1},$$

$\mathbf{H}$ is an $\mathbf{n} \times \mathbf{p}$ matrix with $\mathbf{H_{ti}} = (\mathbf{y_{t-i}} - \sum_{\mathbf{j=1}}^{\mathbf{p}} \boldsymbol{\Phi_{js}}\, \mathbf{y_{t-i-js}})$ and $\mathbf{L}$ is an $\mathbf{n} \times \mathbf{P}$ matrix with $\mathbf{L_{tj}} = (\mathbf{y_{t-js}})$.

3.3.2 The Conditional posterior of $\boldsymbol{\Phi}$:

The conditional posterior of $\boldsymbol{\Phi}$ is

$$\boldsymbol{\Phi^{j+1}} \sim \boldsymbol{\xi}(\boldsymbol{\Phi^{j+1}}|\mathbf{y}, \boldsymbol{\phi^{j+1}}, \boldsymbol{\theta^j}, \boldsymbol{\Theta^j}, (\boldsymbol{\sigma^2})^{\mathbf{j}}, \mathbf{y_0}^{\mathbf{j}}, \boldsymbol{\varepsilon_0}^{\mathbf{j}}) = \mathbf{N}(\boldsymbol{\mu}^*_{\boldsymbol{\Phi}}, \boldsymbol{\upsilon}^*_{\boldsymbol{\Phi}}),$$

where,

$$\boldsymbol{\mu}^*_{\boldsymbol{\Phi}} = \left[\left(\mathbf{G^T G} + \boldsymbol{\Sigma_\Phi}^{-1}\right)^{-1}(\boldsymbol{\Sigma_\Phi}^{-1}\boldsymbol{\mu_\Phi} + \mathbf{G^T}(\mathbf{y} - \mathbf{R}\boldsymbol{\phi} - \widehat{\boldsymbol{\Lambda}}\boldsymbol{\beta_2}))\right], \quad \boldsymbol{\upsilon}^*_{\boldsymbol{\Phi}} = \boldsymbol{\sigma^2}\left(\mathbf{G^T G} + \boldsymbol{\Sigma_\Phi}^{-1}\right)^{-1},$$

$\mathbf{G}$ is an $\mathbf{n} \times \mathbf{P}$ matrix with $\mathbf{G_{tj}} = (\mathbf{y_{t-js}} - \sum_{\mathbf{i=1}}^{\mathbf{p}} \boldsymbol{\phi_i}\, \mathbf{y_{t-i-js}})$ and $\mathbf{R}$ is an $\mathbf{n} \times \mathbf{p}$ matrix with $\mathbf{R_{ti}} = (\mathbf{y_{t-i}})$.

3.3.3 The Conditional Posterior of $\boldsymbol{\theta}$:

The conditional posterior of $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta^{j+1} \sim \xi(\theta^{j+1}|y, \phi^{j+1}, \Phi^{j+1},\ \Theta^{j}, (\sigma^2)^{j},\ y_0{}^{j}, \varepsilon_0{}^{j}) = N\ (\mu_\theta^*, \upsilon_\theta^*\ ),}$$

where,

$$\boldsymbol{\mu_\theta^* = \left[\left(A^T A + \Sigma_\theta{}^{-1}\right)^{-1}(\Sigma_\theta{}^{-1}\mu_\theta + A^T(y - K\Theta - X\beta_1))\right],\quad \upsilon_\theta^* = \sigma^2\left(A^T A + \Sigma_\Theta{}^{-1}\right)^{-1},}$$

$\mathbf{A}$ is an $\mathbf{n \times q}$ matrix with $\mathbf{A_{ti} = (e_{t-i} - \sum_{j=1}^{Q} \Theta_{js}\, e_{t-i-js})}$, and $\mathbf{K}$ is an $\mathbf{n \times Q}$ matrix with $\mathbf{K_{tj} = (e_{t-js})}$.

3.3.4 The Conditional Posterior of Θ:

The conditional posterior of $\boldsymbol{\Theta}$ is

$$\boldsymbol{\Theta^{j+1} \sim \xi(\Theta^{j+1}|y, \phi^{j+1}, \Phi^{j+1},\ \theta^{j+1}, (\sigma^2)^{j},\ y_0{}^{j}, \varepsilon_0{}^{j})\ = N\ (\mu_\Theta^*, \upsilon_\Theta^*\ ),}$$

where,

$$\boldsymbol{\mu_\Theta^* = \left[\left(w^T w + \Sigma_\Theta{}^{-1}\right)^{-1}(\Sigma_\Theta{}^{-1}\mu_\Theta + w^T(y - Z\theta - X\beta_1))\right],\quad \upsilon_\Theta^* = \sigma^2\left(w^T w + \Sigma_\Theta{}^{-1}\right)^{-1},}$$

$\mathbf{w}$ is an $\mathbf{n \times Q}$ matrix with $\mathbf{w_{tj} = (e_{t-js} + \sum_{i=1}^{q} \theta_i\, e_{t-i-js})}$ and $\mathbf{Z}$ is an $\mathbf{n \times q}$ matrix with $\mathbf{Z_{ti} = (e_{t-i})}$.

3.3.5 The Conditional Posterior of σ^2:

The conditional posterior of $\boldsymbol{\sigma^2}$ is

$$\boldsymbol{(\sigma^2)^{j+1} \sim \xi((\sigma^2)^{j+1}|y, \phi^{j+1}, \Phi^{j+1},\ \theta^{j+1}, \Theta^{j+1},\ y_0{}^{j}, \varepsilon_0{}^{j})\ = IG\left(\frac{\upsilon^*}{2},\ \frac{\lambda + n(S^2)^{j+1}}{2}\right),}$$

where,

$\boldsymbol{\upsilon^* = n + \upsilon + 2p + P + Ps + 2q + Q + Qs}$ and $\boldsymbol{nS^2 = [(\phi - \mu_\phi)^T \sum_\phi{}^{-1}(\phi - \mu_\phi) + (\Phi - \mu_\Phi)^T \sum_\Phi{}^{-1}(\Phi - \mu_\Phi) + (\theta - \mu_\theta)^T \sum_\theta{}^{-1}(\theta - \mu_\theta) + (\Theta - \mu_\Theta)^T \sum_\Theta{}^{-1}(\Theta - \mu_\Theta) + (y_0 - \mu_{y_0})^T \sum_{y_0}{}^{-1}(y_0 - \mu_{y_0}) + (\varepsilon_0 - \mu_{\varepsilon_0})^T \sum_{\varepsilon_0}{}^{-1}(\varepsilon_0 - \mu_{\varepsilon_0}) + (y - X\beta_1 - \hat{\Lambda}\beta_2)^T (y - X\beta_1 - \hat{\Lambda}\beta_2)]}$.

Thus, the parameter $\boldsymbol{(\sigma^2)^{j+1}}$ can be sampled from Chi-square distribution using the transformation $\boldsymbol{\frac{\lambda + n(S^2)^{j+1}}{(\sigma^2)^{j+1}} \sim x^2_{\upsilon^*}}$.

3.3.6 The Conditional Posterior of y_0:

Using model (2), the equations for the elements of $\mathbf{y_0}$ and errors $\boldsymbol{\varepsilon_0}$ can be written as

$$\boldsymbol{F\, y_{p+Ps} = D\, y_0 + M\, \varepsilon_0 + N\, \varepsilon_{p+Ps}}$$

where,

$$\mathbf{F}=\begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ -\boldsymbol{\beta}_{\mathbf{1}} & \mathbf{1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ -\boldsymbol{\beta}_{\mathbf{2}} & -\boldsymbol{\beta}_{\mathbf{1}} & \mathbf{1} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots \\ -\boldsymbol{\beta}_{\mathbf{p+Ps-1}} & -\boldsymbol{\beta}_{\mathbf{p+Ps-2}} & \cdots & \cdots & -\boldsymbol{\beta}_{\mathbf{2}} & -\boldsymbol{\beta}_{\mathbf{1}} & \mathbf{1} \end{pmatrix}_{(\mathbf{p+Ps})\times(\mathbf{p+Ps})},$$

$$\mathbf{D}=\begin{pmatrix} \boldsymbol{\beta}_{\mathbf{1}} & \boldsymbol{\beta}_{\mathbf{2}} & \boldsymbol{\beta}_{\mathbf{3}} & \cdots & \cdots & \boldsymbol{\beta}_{\mathbf{p+Ps-1}} & \boldsymbol{\beta}_{\mathbf{p+Ps}} \\ \boldsymbol{\beta}_{\mathbf{2}} & \boldsymbol{\beta}_{\mathbf{3}} & \boldsymbol{\beta}_{\mathbf{4}} & \cdots & \cdots & \boldsymbol{\beta}_{\mathbf{p+Ps}} & \mathbf{0} \\ \boldsymbol{\beta}_{\mathbf{3}} & \boldsymbol{\beta}_{\mathbf{4}} & \boldsymbol{\beta}_{\mathbf{5}} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots \\ \boldsymbol{\beta}_{\mathbf{p+Ps-1}} & \boldsymbol{\beta}_{\mathbf{p+Ps}} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \boldsymbol{\beta}_{\mathbf{p+Ps}} & \mathbf{0} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \end{pmatrix}_{(\mathbf{p+Ps})\times(\mathbf{p+Ps})},$$

$$\mathbf{M}=\begin{pmatrix} \boldsymbol{\alpha}_{\mathbf{1}} & \boldsymbol{\alpha}_{\mathbf{2}} & \boldsymbol{\alpha}_{\mathbf{3}} & \cdots & \cdots & \boldsymbol{\alpha}_{\mathbf{q+Q_s-1}} & \boldsymbol{\alpha}_{\mathbf{q+Q_s}} \\ \boldsymbol{\alpha}_{\mathbf{2}} & \boldsymbol{\alpha}_{\mathbf{3}} & \boldsymbol{\alpha}_{\mathbf{4}} & \cdots & \cdots & \boldsymbol{\alpha}_{\mathbf{q+Q_s}} & \mathbf{0} \\ \boldsymbol{\alpha}_{\mathbf{3}} & \boldsymbol{\alpha}_{\mathbf{4}} & \boldsymbol{\alpha}_{\mathbf{45}} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots \\ \boldsymbol{\alpha}_{\mathbf{p+Ps-1}} & \boldsymbol{\alpha}_{\mathbf{p+Ps}} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \boldsymbol{\alpha}_{\mathbf{p+Ps}} & \mathbf{0} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \end{pmatrix}_{(\mathbf{p+Ps})\times(\mathbf{q+Q_s})},$$

$$\mathbf{N}=\begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \boldsymbol{\alpha}_{\mathbf{1}} & \mathbf{1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \boldsymbol{\alpha}_{\mathbf{2}} & \boldsymbol{\alpha}_{\mathbf{1}} & \mathbf{1} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots \\ \boldsymbol{\alpha}_{\mathbf{p+Ps-1}} & \boldsymbol{\alpha}_{\mathbf{p+Ps-2}} & \cdots & \cdots & \boldsymbol{\alpha}_{\mathbf{2}} & \boldsymbol{\alpha}_{\mathbf{1}} & \mathbf{1} \end{pmatrix}_{(\mathbf{p+Ps})\times(\mathbf{p+Ps})},$$

$\boldsymbol{\alpha}_{\mathbf{r}}=\mathbf{0}\ \forall\ \mathbf{r}=\mathbf{q}+\mathbf{Qs}+\mathbf{1},\dots,\mathbf{p}+\mathbf{Ps}\ \mathbf{if}\ \mathbf{p}+\mathbf{Ps}\ > q+Qs,\quad \mathbf{y}_{\mathbf{p+Ps}}=(\mathbf{y}_{\mathbf{1}},\dots,\mathbf{y}_{\mathbf{p+Ps}})^{\mathbf{T}}$ and $\boldsymbol{\varepsilon}_{\mathbf{p+Ps}}=(\boldsymbol{\varepsilon}_{\mathbf{1}},\boldsymbol{\varepsilon}_{\mathbf{2}},\dots,\boldsymbol{\varepsilon}_{\mathbf{p+Ps}})^{\mathbf{T}}$, which has the normal distribution with zero mean and variance $(\boldsymbol{\sigma}^{\mathbf{2}}\mathbf{I}_{\mathbf{p+Ps}})$ where $\mathbf{I}_{\mathbf{p+Ps}}$ is the unit matrix of order $(\mathbf{p}+\mathbf{Ps})$. Using linear regression results and standard Bayesian techniques, the conditional posterior of $\mathbf{y}_{\mathbf{0}}$ is

$$\mathbf{y_0}^{\mathbf{j+1}}\sim\boldsymbol{\xi}(\mathbf{y_0}^{\mathbf{j+1}}|\mathbf{y},\boldsymbol{\phi}^{\mathbf{j+1}},\boldsymbol{\Phi}^{\mathbf{j+1}},\ \boldsymbol{\theta}^{\mathbf{j+1}},\boldsymbol{\Theta}^{\mathbf{j+1}},(\boldsymbol{\sigma}^{\mathbf{2}})^{\mathbf{j+1}},\ \boldsymbol{\varepsilon_0}^{\mathbf{j}})\ =\mathbf{N}\ (\boldsymbol{\mu}^{*}_{\mathbf{y_0}},\boldsymbol{\upsilon}^{*}_{\mathbf{y_0}}),$$

where,

$$\boldsymbol{\mu}^{*}_{\mathbf{y_0}}=\left[\left(\boldsymbol{\Omega}\,\mathbf{D}+\boldsymbol{\Sigma}_{\mathbf{y_0}}{}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_{\mathbf{y_0}}{}^{-1}\boldsymbol{\mu}_{\mathbf{y_0}}+\boldsymbol{\Omega}\left(\mathbf{F}\,\mathbf{y}_{\mathbf{p+Ps}}-\ \mathbf{M}\,\boldsymbol{\varepsilon}_{\mathbf{0}}\right)\right)\right],\ \ \boldsymbol{\upsilon}^{*}_{\mathbf{y_0}}=\ \boldsymbol{\sigma}^{\mathbf{2}}\left(\boldsymbol{\Omega}\,\mathbf{D}+\boldsymbol{\Sigma}_{\mathbf{y_0}}{}^{-1}\right)^{-1},$$

and $\boldsymbol{\Omega}=\mathbf{D}^{\mathbf{T}}(\mathbf{N}\mathbf{N}^{\mathbf{T}})^{-1}$.

3.3.7 The Conditional Posterior of $\boldsymbol{\varepsilon}_{\mathbf{0}}$:

The conditional posterior of $\boldsymbol{\varepsilon}_{\mathbf{0}}$ is

$$\boldsymbol{\varepsilon_0}^{\mathbf{j+1}}\sim\boldsymbol{\xi}(\boldsymbol{\varepsilon_0}^{\mathbf{j+1}}|\mathbf{y},\boldsymbol{\phi}^{\mathbf{j+1}},\boldsymbol{\Phi}^{\mathbf{j+1}},\ \boldsymbol{\theta}^{\mathbf{j+1}},\boldsymbol{\Theta}^{\mathbf{j+1}},(\boldsymbol{\sigma}^{\mathbf{2}})^{\mathbf{j+1}},\ \mathbf{y_0}^{\mathbf{j+1}})\ =\mathbf{N}\ (\boldsymbol{\mu}^{*}_{\boldsymbol{\varepsilon}_{\mathbf{0}}},\boldsymbol{\upsilon}^{*}_{\boldsymbol{\varepsilon}_{\mathbf{0}}}),$$

where,

$\boldsymbol{\mu}_{\boldsymbol{\varepsilon_0}}^{*} = \left[\left(\boldsymbol{\Psi}\,\mathbf{M} + {\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon_0}}}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon_0}}{}^{-1}\boldsymbol{\mu}_{\boldsymbol{\varepsilon_0}} + \boldsymbol{\Psi}\left(\mathbf{F}\,\mathbf{y}_{\mathbf{p+Ps}} - \mathbf{D}\,\mathbf{y}_{\mathbf{0}}\right)\right)\right], \boldsymbol{\upsilon}_{\boldsymbol{\varepsilon_0}}^{*} = \boldsymbol{\sigma}^{\mathbf{2}}\left(\boldsymbol{\Psi}\,\mathbf{M} + {\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon_0}}}^{-1}\right)^{-1}$, and $\boldsymbol{\Psi} = \mathbf{M}^{\mathbf{T}}(\mathbf{N}\mathbf{N}^{\mathbf{T}})^{-1}$.

4. THE PROPOSED GIBBS SAMPLER

The proposed Gibbs sampling algorithm for multiplicative SARMA model given in equation (2) can be conducted as follows:

Step (1): Specify starting values $\phi^0, \Phi^0, \theta^0, \Theta^0, (\sigma^2)^0, {y_0}^0$ and ${\varepsilon_0}^0$ and set $j = 0$. A set of initial estimates of the model parameters can be obtained using the IS technique of Koreisha and Pulkila (1990).

Step (2): Calculate the residuals recursively using (6).

Step (3): Obtain the full conditional posterior distributions of the parameters.

Step (4): Simulate

- $\phi^{j+1} \sim \xi\left(\phi^{j+1}|y, \Phi^j, \theta^j, \Theta^j, (\sigma^2)^j, {y_0}^j, {\varepsilon_0}^j\right)$,
- $\Phi^{j+1} \sim \xi\left(\Phi^{j+1}|y, \phi^{j+1}, \theta^j, \Theta^j, (\sigma^2)^j, {y_0}^j, {\varepsilon_0}^j\right)$,
- $\theta^{j+1} \sim \xi\left(\theta^{j+1}|y, \phi^{j+1}, \Phi^{j+1}, \Theta^j, (\sigma^2)^j, {y_0}^j, {\varepsilon_0}^j\right)$,
- $\Theta^{j+1} \sim \xi\left(\Theta^{j+1}|y, \phi^{j+1}, \Phi^{j+1}, \theta^{j+1}, (\sigma^2)^j, {y_0}^j, {\varepsilon_0}^j\right)$,
- $(\sigma^2)^{j+1} \sim \xi\left((\sigma^2)^{j+1}|y, \phi^{j+1}, \Phi^{j+1}, \theta^{j+1}, \Theta^{j+1}, {y_0}^j, {\varepsilon_0}^j\right)$,
- ${y_0}^{j+1} \sim \xi\left({y_0}^{j+1}|y, \phi^{j+1}, \Phi^{j+1}, \theta^{j+1}, \Theta^{j+1}, (\sigma^2)^{j+1}, {\varepsilon_0}^j\right)$, and
- ${\varepsilon_0}^{j+1} \sim \xi\left({\varepsilon_0}^{j+1}|y, \phi^{j+1}, \Phi^{j+1}, \theta^{j+1}, \Theta^{j+1}, (\sigma^2)^{j+1}, {y_0}^{j+1}\right)$.

Step (5): Set $j = j + 1$ and go to Step (4).

This algorithm gives the next value of the Markov chain $\{\phi^{j+2}, \Phi^{j+2}, \theta^{j+2}, \Theta^{j+2}, (\sigma^2)^{j+2}, {y_0}^{j+2}, {\varepsilon_0}^{j+2}\}$ by simulating each of the full conditional posteriors where the conditioning elements are revised during a cycle.

This iterative process is repeated for a large number of iterations and convergence is monitored. After the chain converges, after n_0 iterations, the simulated values

$$\{\phi^{j+1}, \Phi^{j+1}, \theta^{j+1}, \Theta^{j+1}, (\sigma^2)^{j+1}, {y_0}^{j+1}, {\varepsilon_0}^{j+1}, \forall\, j > n_0\},$$

are used as a sample from the joint posterior. Posterior estimates of the parameters are computed directly via sample averages of the simulation outputs.

A large and growing literature deals with techniques for monitoring convergence of Gibbs sampling sequences. In what follows we summarize the diagnostics that will be used in the case of multiplicative SARMA model:

1. Autocorrelation estimates which indicate how much independence exists in the sequence of each parameter draws. A high degree of autocorrelation indicates that more draws are needed to get accurate posterior estimates.
2. Raftery and Lewis (1992, 1995) proposed a set of diagnostics which includes
   - a thinning ratio (Thin) which is a function of the autocorre1ation in the draws.

- the number of draws (Burn) to use for initial draws or "burn-in" before starting to sample the draws for purpose of posterior inference.
- the total number of draws (Total) needed to achieve desired level of accuracy.
- the number of draws (Nmin) that would be needed if the draws represented an iid chain.
- (I-stat) which is the ratio of the (Total) to (Nmin). Raftery and Lewis suggested that convergence problem may be indicated when values of (I- stat) exceed 5.

3. Geweke (1992) proposed two groups of diagnostics.
   (a) The first group includes the numerical standard errors (NSE) and relative numerical efficiency (RNE). The NSE estimates reflect the variation that can be expected if the simulation were to be repeated. The RNE estimates indicate the required number of draws to produce the same numerical accuracy when iid sample is drawn directly from the posterior distribution. The estimates of NSE and RNE are based on spectral analysis of time series where two sets of these estimates are obtained. The first set is based on the assumption that the draws come from iid process. The second set is based on different tapering or truncating of the periodgram window. When there are large differences between the two sets, the second set of estimates would be chosen because it would take in consideration autocorrelations in the draws.
   (b) The second group of diagnostics includes a test of whether the sampler has attained an equilibrium state. This is done by carrying out Z-test for the equality of the two means of the first and last parts of draws and the Chi squared marginal probability is obtained. Usually, the first and last parts are the first 20% and the last 50% of the draws.

LeSage (1999) implemented calculations of the above convergence measures using the Matlab package. These diagnostics will be used in section 5 to monitor the convergence of the proposed algorithm.

5. ILLUSTRATIVE EXAMPLES

5.1 Simulated Examples

In this subsection we present two examples with simulated data to evaluate the efficiency of the proposed methodology. The two examples deal with generating 250 observations from SARMA(1,1)(1,1)4 and SARMA(1,1)(1,1)12 models respectively. The two simulated examples are as follows:

(1): $y_t = 0.2\, y_{t-1} + 0.8\, y_{t-4} - 0.16\, y_{t-5} + 0.3\, \varepsilon_{t-1} + 0.9\, \varepsilon_{t-4} + 0.27\, \varepsilon_{t-5} + \varepsilon_t$
(2): $y_t = 0.5\, y_{t-1} + 0.8\, y_{t-12} - 0.4\, y_{t-13} + 0.4\, \varepsilon_{t-1} + 0.7\, \varepsilon_{t-12} + 0.28\, \varepsilon_{t-13} + \varepsilon_t$

The analysis was implemented using Matlab and running on Pentium PC 2.53 GHZ took several seconds (90 seconds on average) to complete. The error variance σ^2 was chosen to be 0.5 in the first example and 1 in the second example. A non informative prior was assumed for $\phi, \Phi, \theta, \Theta, y_0, \varepsilon_0$ and σ^2 via setting $\sum_\phi{}^{-1} = \sum_\Phi{}^{-1} = \sum_\theta{}^{-1} = \sum_\Theta{}^{-1} = 0$ (all these matrices are scalars), $\lambda = S^2$ and $\upsilon = 3$. A normal prior with zero mean and variance $\sigma^2 I_{p+Ps}$ was used for the initial observations vector y_0, and with zero mean and variance $\sigma^2 I_{q+Qs}$ was used for the

initial errors vector ε_0. The starting values for the parameters ϕ, Φ, θ and Θ were obtained using the IS method. The starting values for y_0 and ε_0 were assumed to be zeros.

Now, the implementation of the proposed Gibbs sampler is straightforward. For each data set, the Gibbs sampler was run 11,000 iterations where the first 1,000 draws are ignored and every tenth value in the sequence of the last 10,000 draws is recorded to have an approximately independent sample. All posterior estimates are computed directly as sample averages of the simulated outputs.

Table 1 presents the true values and Bayesian estimates of the parameters for example 1. Moreover, a 95% confidence interval using the 0.025 and 0.975 percentiles of the simulated draws is constructed for every parameter. From table 1, it is clear that Bayesian estimates are close to the true values and the 95% confidence interval includes the true value for every parameter. Sequential plots of the parameters generated sequences together with marginal densities are displayed in figure 1. The marginal densities are computed using non parametric technique with Gaussian kernel. Figure 1 shows that the proposed algorithm is stable and fluctuates in the neighborhood of true values. In addition, the marginal densities show that the true value of each parameter (which is indicated by the vertical line) falls in the constructed 95% confidence interval.

Table (1): Bayesian Results for example (1)

Parameter	True values	Mean	Std. Dev.	Lower 95% limit	Median	Upper 95% limit
ϕ	0.4	0.377	0.062	0.253	0.377	0.504
Φ	0.6	0.619	0.039	0.539	0.619	0.692
θ	0.4	0.357	0.079	0.198	0.354	0.514
Θ	0.6	0.585	0.070	0.444	0.586	0.722
σ^2	0.5	0.511	0.044	0.429	0.510	0.597

The convergence of the proposed algorithm is monitored using the diagnostic measures explained in section 4. The autocorrelations and Raftery and Lewis diagnostics are displayed in table 2 whereas table 3 presents the diagnostic measures of Geweke (1992). Table 2 shows that the draws for each of the parameter have small autocorrelations at lags 1, 5, 10 and 50 which indicates no convergence problem. This conclusion was confirmed by the diagnostic measures of Raftery and Lewis where the thining estimate (Thin) is 1, the reported (Nmin) is 937 which is close to the 1000 draws we used and I-stat value is 0.953 which is less than 5. Scanning the entries of table 3, confirms the convergence of the proposed algorithm where Chi squared probabilities show that the equal means hypothesis can not be rejected and no dramatic differences between the NSE estimates are found. In addition, the RNE estimates are close to 1 which indicates the iid nature of the output sample.

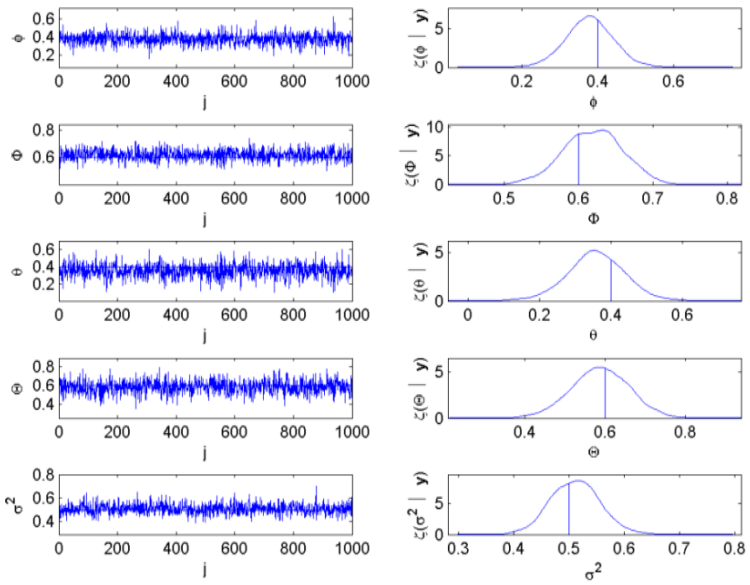


Figure (1): Sequential plots and marginal posterior distributions of example (1)

Table (2): Autocorrelations and Raftery - Lewis Diagnostics for example (1)

Par.	Autocorrelations				Raftery - Lewis Diagnostics				
	Lag 1	Lag 5	Lag 10	Lag 50	Thin	Burn	Total(N)	(Nmin)	I-stat
ϕ	-0.041	0.025	0.006	-0.094	1	2	893	937	0.953
Φ	0.001	0.084	0.084	-0.001	1	2	893	937	0.953
θ	-0.005	0.032	0.032	-0.037	1	2	893	937	0.953
Θ	0.006	-0.011	-0.011	-0.025	1	2	893	937	0.953
$\boldsymbol{\sigma^2}$	0.007	-0.035	-0.035	-0.010	1	2	893	937	0.953

Table (3): Geweke Diagnostics for example (1)

Par.	NSE iid	RNE iid	NSE 4%	RNE 4%	NSE 8%	RNE 8%	NSE 15%	RNE 15%	χ^2
ϕ	0.00196	1	0.00188	1.078	0.0016	1.498	0.00121	2.60	0.369
Φ	0.00123	1	0.00116	1.118	0.0012	0.996	0.00110	1.25	0.899
θ	0.00250	1	0.00242	1.062	0.0022	1.255	0.00175	2.04	0.515
Θ	0.00222	1	0.00255	0.758	0.0028	0.647	0.00260	0.73	0.019
$\boldsymbol{\sigma^2}$	0.00139	1	0.00151	0.844	0.0014	1.020	0.00146	0.90	0.996

A similar procedure to that used for example 1 is repeated for example 2 and the true values and Bayesian results are shown in table 4. Similar conclusions to those of example 1 are obtained. The convergence diagnostics for example 2 are displayed in table 5 and 6. Our Gibbs sampler is applied to several simulated data from other SARMA(1,1)(1,1)s models which do not appear here. The results for these data sets are similar to results of examples 1 and 2 and therefore are not included.

Table (4): Bayesian Results for example (2)

Parameter	True Values	Mean	Std. Dev.	Lower 95% limit	Median	Upper 95% limit
ϕ	0.5	0.517	0.050	0.423	0.517	0.612
Φ	0.8	0.817	0.024	0.774	0.816	0.867
θ	0.4	0.407	0.075	0.266	0.407	0.555
Θ	0.7	0.713	0.064	0.591	0.713	0.839
σ^2	1.0	0.950	0.085	0.797	0.943	1.138

5.2 Federal Reserve Board Production Index

The Federal Reserve Board Production Index consists of 372 monthly values from jauary 1948to December 1978. Using Box-Jenkins methodology, Shumway and Stoffer (2006) identified the following SARIMA $(1,1,1)(2,1,1)_{12}$ model for Federal Reserve Board Production Index y_t:

$$(1 - \phi B)\,(1 - \Phi_1 B^{12} - \Phi_2 B^{24}) w_t = (1 + \theta B)\,(1 + \Theta B^{12})\varepsilon_t \qquad (12)$$

where, $w_t = (1 - B)(1 - B^{12})y_t$

In this section, the proposed Bayesian analysis is applied to the differenced Federal Reserve Board Production Index series w_t. The hyper-parameters and starting values are chosen as in the simulated examples. Table 5 summarizes the Bayesian results for the differenced Federal Reserve Board Production Index series together with the corresponding results of Shumway and Stoffer (2006). Sequential plots and marginal densities of the differenced Federal Reserve Board Production Index series are displayed in figure 2.

From Table 5, our Bayesian estimates are comparable to the estimates of Shumway and Stoffer (2006). Moreover, the standard deviation of our Bayesian estimates is equal to the standard deviation of the estimates of Shumway and Stoffer (206). This may be considered as an advantage to the proposed algorithm where uncertainty about the parameters and initial values are incorporated.

Table (5): Bayesian Results for the differenced Federal Reserve Board Production Index series using SARMA(1,1)(2,1)12

Parameter	Mean	Std. Dev.	Lower 95% limit	Median	Upper 95% limit	S-S estimates
ϕ	0.574	0.105	0.368	0.577	0.777	0.580
Φ_1	-0.290	0.079	-0.444	-0.290	-0.145	-0.220
Φ_2	-0.315	0.050	-0.411	-0.316	-0.214	-0.280
θ	-0.244	0.117	-0.476	-0.245	-0.026	-0.270
Θ	-0.413	0.092	-0.587	-0.408	-0.233	-0.500
σ^2	1.271	0.093	1.097	1.267	1.461	1.350

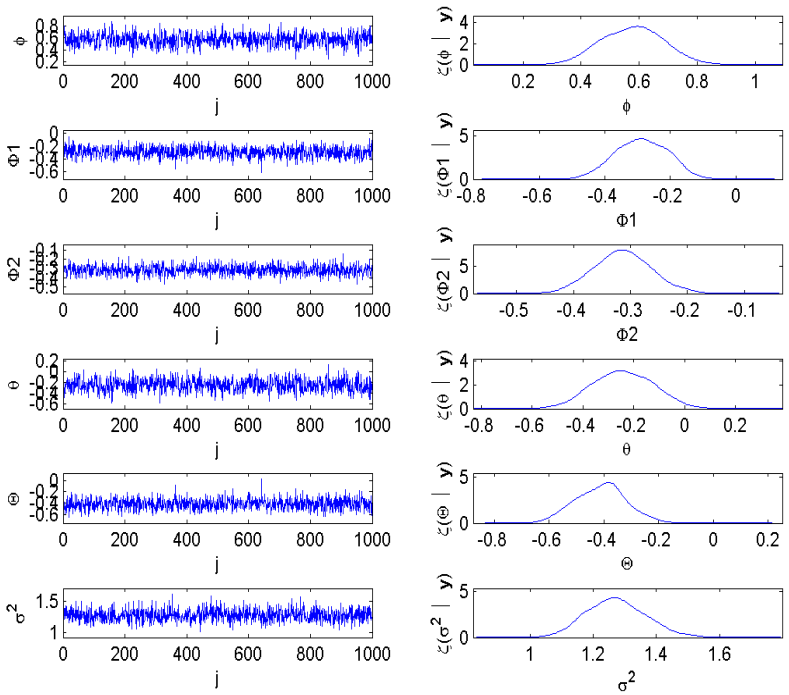


Figure (2): Sequential plots and marginal posterior distributions of the differenced Federal Reserve Board Production Index series

It is worthwhile to test the significance of the interaction parameter $\eta = \phi\Phi_1$ in the above SARIMA $(1,1,1)(2,1,1)_{12}$ model. Although the testing procedure of the significance of η is complicated or even impossible in the classical approach framework, it is straightforward in the suggested Bayesian framework. Using the proposed Gibbs sampling algorithm, the marginal posterior distribution of η is obtained and displayed in figure 3. Moreover a 95% credible interval for η is $-0.28 \leq \eta \leq -0.08$, which supports the significance hypothesis of the interaction parameter.

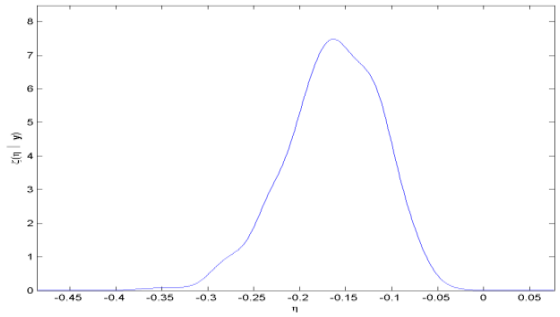


Figure (3): marginal posterior distribution of the interaction parameter η

6. COMMENTS AND CONCLUSION

In this paper we developed a simple and fast Gibbs sampling algorithm for estimating the parameters of the multiplicative SARMA model. The empirical results of the simulated examples and real data set showed the accuracy of the proposed methodology. An extensive check of convergence using several diagnostics showed that the convergence of the proposed algorithm was achieved.

Although the employed prior distribution in section 3 is informative, a noninformative prior is used for the parameters $\phi, \Phi, \theta, \Theta$ and σ^2 in the illustrative examples for the sake of the simplicity. However, if one needs to use informative prior, the hyperparameters of the prior distribution must be elicited. One way to elicit the hyperparameters is the training sample approach where the data is divided into two parts; the first part constitutes the training sample and is used to provide proper priors. Then, posterior distributions are obtained by combining these priors with the likelihood based on the second part of the data (non-training sample). The training sample approach is used by Lempers (1971) and Speigelhalter and smith (1982) among others.

Future work may investigate model identification using stochastic search variable selection, outliers detection, and extension to multivariate seasonal models.

REFERENCES

Barnett, G., Kohn, R., and Sheather, S. (1996). Bayesian estimation of an autoregressive model using Markov Chain Monte Carlo. *Journal of Econometrics*, 74, 237-254.

Barnett, G., Kohn, R., and Sheather, S. (1997). Robust Bayesian estimation of an autoregressive moving-average models. *Journal of Time Series Analysis*, 18, 11-28.

Box, G.E.P and Jenkins, G. M. (1976). *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco.

Broemeling, L. D. (1985). *Bayesian Analysis of Linear Models*. Marcel Dekker Inc., New York.

Broemeling, L. D. and Shaarawy, S. (1984). Bayesian inferences and forecasts with moving average processes. *Communications in Statistics: Theory and Methods*, 13, 1871-1888.

Chib, S. and Greenberg, E. (1994). Bayes inference in regression models with ARMA(p,q) errors. *Journal of Econometrics*, 64, 183-206.

Geweke, J. (1992). Evaluating the Accuracy of Sampling-Based Approaches to the Calculations of Posterior Moments. *In Bayesian Statistics 4*, J. M., Bernardo, J. O.

Ismail, M. A. (2003a). Bayesian Analysis of Seasonal Autoregressive Models. *Journal of Applied Statistical science*, 12(2)}, 123-136.

Ismail, M. A. (2003b). Bayesian Analysis of Seasonal Moving Average Model: A Gibbs Sampling Approach. *Japanese Journal of Applied Statistics*, 32(2), 61-75.

Korisha and Pukkila (1990). Linear methods for estimating ARMA and regression model with serial correlation. *Communications in Statistics: Simulation*, 19, 71-102.

Lempers, F.B. (1971). *Posterior probabilities of alternative linear models*. Rotterdam Unirersity press.

LeSage, J. P. (1999). Applied Econometrics using Matlab. Dept. of Economics, University of Toledo, available at http://www.econ.utoledo.edu.

Marriott, J., Ravishanker, N., Gelfand, A. and Pai, J. (1996). Bayesian analysis of ARMA processes: Complete sampling based inference under full likelihoods. *In Bayesian Statistics and Econometrics*: Essays in honor of Arnold Zellner, D., Berry, K., Chaloner, and J., Geweke, (eds). New York, Wiley.

Marriott, J. and Smith, A. F. M. (1992). Reparameterization aspects of numerical Bayesian methodology for autoregressive moving-average models. *Journal of Time Series Analysis*, 13, 327-343.

McCulloch, R. E. and Tsay, R. S. (1994). Bayesian analysis of autoregressive time series via the Gibbs sampler. *Journal of Time Series Analysis*, 15(2), 235-250

Monahan, J. F. (1983). Fully Bayesian analysis of ARMA time series models. *J. Econometrics*, 19, 147-164.

Newbold, P. (1973). Bayesian estimation of Box Jenkins transfer function noise models. *JRSSB*, 35, 323-336.

Raftrey, A. E. and Lewis, S. (1992). One long run with diagnostics: Implementation strategies for Markov Chain Monte Carlo. *Statistical Science*, 7, 493-497.

Raftrey, A. E. and Lewis, S. (1995). The number of iterations, convergence diagnostics and generic Metropolis algorithms. *In Practical Markov Chain Monte Carlo*, W. R., Gilks, D. J., Spiegelhalter and S., Richardson, (eds). London, Chapman and Hall.

Shaarawy, S. and Ismail, M. A. (1987). Bayesian Inference for Seasonal ARMA Models. *The Egyptian Statistical Journal*, 31, 323- 336.

Shumway, R.H. and Stoffer, D.S. (2006). *Time series analysis and its applications with R examples*. New York: Springer.

Spiegelhalter. D.I. and Smith. A.F.M. (1982). Bayes factors for linear and log-linear models with vague prior information. *JRSSB*, 44, 377-387.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 786–795.

A NEW THRESHOLD VALUE IN CURVE ESTIMATION BY WAVELET SHRINKAGE

B. Ismail and Anjum Khan
Department of Statistics
Mangalore University, mangalagangothri
Mangalore-574199 India
E-mail: ismailbn@yahoo.com

ABSTRACT

Wavelets are orthonormal basis functions with special properties that show potential in many areas of Mathematics and Statistics. A major advantage of wavelet methods in curve estimation is their adaptivity to erratic functions in the signal. This high degree of adaptivity is achieved through thresholding, which typically amounts to term by term assessment of estimates of coefficients in the empirical wavelet expansion of the unknown function. This paper concentrate on selecting a threshold for wavelet function estimation. A new threshold value is proposed for wavelet shrinkage estimators operating on data sets of length a power of 2. It is compared with established threshold parameter in wavelet shrinkage by using simulation. For certain range of data the proposed threshold value lies between minimax and universal threshold values. The simulation result shows that the proposed threshold value gives a better estimation of the true curve.

Keywords. Minimax estimation, non-parametric regression, non-linear estimation, orthonormal bases, thresholding, universal threshold, wavelet shrinkage.

1. INTRODUCTION

Suppose we are given data

$$y_i = f(x_i) + \epsilon_i \tag{1}$$

where ϵ_i is some noise process with variance σ^2, $x_i = \frac{i}{n}$. The idea of non parametric regression is to estimate the unknown function f from the observations y_i, $i = 1, 2, 3, \ldots, n$ without assuming any particular parametric form. In vector notation, the model (1) can be written as

$$Y = F + E \tag{2}$$

where $y = (y_1, y_2, \ldots, y_n)$, $F = (f_1, f_2, \ldots, f_n)$ and $E = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ the observed data, signal and noise respectively. The estimated function that is the discrete estimator $\hat{F} = (\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_n)$ will be judged by its expected mean square error $\frac{1}{n}E\|\hat{f} - f\|^2_{2,n}$.

In this paper we work on estimating the function f by shrinking the wavelet coefficients. The threshold value to shrink the empirical wavelet coefficient is crucial. We have suggested a threshold value which gives a better approximation of the true curve for moderate data.

The paper is organized as follows. In Section 2 we provide some necessary mathematical background and terminology in relation to wavelets. The concept of Discrete wavelet transform (DWT) and the non-linear estimation procedure for the estimation of function is presented in Section 3. In Section 4 we give a brief review about the choice of threshold. We propose a new threshold value and compare its performance with minimax and universal threshold values through simulations. We conclude the paper with some comments in the last section.

2. WAVELET OVERVIEW

Wavelets are functions specially made as to form an orthonormal basis for various function spaces. One such example is $L^2(R)$, set of all square integrable function on R. It can be shown (Daubechies (1992), Meyer (1992)) that it is possible to construct a function $\psi(x)$ so that if $f \in L^2(R)$, then

$$f(x) = \sum_{k \in Z} c_{0,k}\phi_{0,k}(x) + \sum_{j<J}\sum_{k \in Z} d_{j,k}\psi_{j,k}(x) \tag{3}$$

where $c_{0,k} = \int_R f(x)\phi_{0,k}(x)dx$ and $d_{j,k} = \int_R f(x)\psi_{j,k}(x)dx$, and j controls the maximum resolution. The function $\psi_{j,k} = 2^{j/2}\psi(2^j x - k)$ which is derived from the function $\psi(x)$ by dilation and the translation is called the mother wavelet. The function $\phi_{0,k}(x)$ are all derived from a function $\phi(x)$ known as father wavelet or scaling function by using dilation and translation formula $\phi_{0,k} = \phi(x - k)$. Wavelets have a built in "Spatially adaptive" that allows efficient estimation of functions with inhomogeneity, discontinuities in derivatives, sharp spikes and discontinuity in the function itself.

3. ESTIMATION OF FUNCTION

3.1 Discrete Wavelet Transform

Usually in statistical problem we have finite set of discrete data. If we have $n = 2^J$ value of $y(x)$ equally spaced between 0 and 1. we use wavelets at levels $j = 0, 1, 2, \ldots, J-1$, where $k = 1, \ldots, 2^j - 1$. Level 0 contains the mother and father wavelets while increasing value of j corresponding to wavelet which describes finer details. If $y = (y(x_1), y(x_2), \ldots, y(x_n))^T$, then

$$y(x_i) = c_{0,0}\phi(x_i) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(x_i) \tag{4}$$

The vector $w = (c_{0,0}, d_{0,0}, \ldots, d_{J-1,2^j-1})^T$ is referred as the DWT of y. The DWT may be represented by $n \times n$ unitary matrix W, $(WW^T = W^TW = I_n)$. In practice the DWT

can be performed using the algorithm of Mallat (1989) with $O(n)$ operation rather than the slow $O(n^2)$ matrix multiplication. In equation (4 the coefficient vector w can be obtained by using DWT to the given data $y = (y_1, y_2, \ldots, y_n)$ which assumes the model (1) and in vector notation as the model (2). Let DWT of Y be $w = WY = d + \epsilon$, where $d = WF$, and $\epsilon = WE$ respectively. Our goal is to recover F from the noisy data Y. Equivalently we estimate the true wavelet coefficient from the empirical wavelet coefficient.

3.2 **Non-linear Estimation**

Donoho and Johnstone (1994, 1995) and Donoho et al. (1995) proposed a non-linear estimator of f based on a reconstruction from a more judicious selection of empirical wavelet coefficients. This approach is now widely used in statistics particularly in signal processing and image analysis, former is known as de-noising and later one is image compression. From the statistical approach, the model (1) is a regression model over time and this method can be viewed as a non-parametric estimation of the function f using orthonormal basis.

Donoho et al. (1995) advise that shrinking wavelet coefficient will remove the noise of the signal. To achieve shrinkage they propose thresholding the empirical wavelet coefficients. Given the empirical wavelet coefficients and a threshold $\lambda > 0$ the hard threshold value is given by

$$\delta^H(w, \lambda) = wI(|w| > \lambda) \tag{5}$$

which is a "keep or kill" rule, and the soft threshold value is given by

$$\delta^s(w, \lambda) = Sign(w)(|w| - \lambda)I(|w| > \lambda) \tag{6}$$

which is shrink or kill rule.

There are some other thresholding rules, some of them are firm thresholding studied by Gao and Bruce (1997), SCAD threshold by Antoniadis and Fan (2001). The hard and soft thresholding rules are the most commonly used among the various wavelet thresholding estimators. The thresholded wavelet coefficient obtained by applying any of the thresholded rule $\delta(w, \lambda)$ given in (5) or (6) are used to obtain a selective reconstruction of the response function f. The resulting estimate can be written as

$$\hat{f}_\lambda(t) = \sum_{k=0}^{2^{j_0}-1} \frac{\hat{c_{0,k}}}{\sqrt{n}} \phi_{j_0,k}(t) + \sum_{j=j_0}^{j-1} \sum_{k=0}^{2^j-1} \frac{\delta(\lambda, \hat{d_{j,k}})}{\sqrt{n}} \psi_{j,k}(t). \tag{7}$$

The value $\sqrt{n}$ appear because of the different normality condition of continuous and discrete wavelet transform.

In the above case the vector $\hat{f}_\lambda$ of the corresponding estimator can be derived by simply performing IDWT of $\left\{\hat{c_{o,k}}, \delta_\lambda(\hat{d_{j,k}})\right\}$ and the resulting 3 step selective reconstruction estimation procedure that can be summarized by the following diagram.

$$y \xrightarrow{DWT} \left\{\hat{c_{j_o,k}}, \hat{d_{j,k}}\right\} \xrightarrow{Threshold} \left\{\hat{c_{o,k}}, \delta_\lambda(\hat{d_{j,k}})\right\} \xrightarrow{IDWT} \hat{f}_\lambda. \tag{8}$$

4. THE CHOICE OF THRESHOLD

Clearly, an appropriate choice of the threshold value λ is fundamental to the effectiveness of the procedure described in the previous section. Too large threshold might cut off important parts of the true function underlying the data whereas too small a threshold retain noise in a selective reconstruction. Donoho and Johnstone (1994) proposed minimax threshold that depends on the data size n, defined as $\lambda^M = \hat{\sigma}\lambda_n^*$ where λ_n^* is defined as the value of λ which achieves

$$\Lambda_n^* = \inf_{\lambda} \sup_{d} \{R_\lambda(d)/(n^{-1} + R_{oracle}(d))\} \tag{9}$$

where $R_\lambda(d) = E[\delta_\lambda(\hat{d}) - d]^2$ and $R_{oracle}(d)$ is the ideal risk achieved with the help of an oracle.

As an alternative to minimax threshold Donoho and Johnstone (1994) proposed the universal threshold $\lambda = \hat{\sigma}\sqrt{2\log n}$, which is also asymptotically optimal and simpler to implement. Comparing to the minimax the universal threshold value is substantially large. Universal threshold safeguards against allowing spurious noise into the reconstruction. This is due to the fact that if $z_1, z_2, \ldots, z_n$, represent iid $N(0,1)$ sequence, then as n converges to ∞

$$Pr\{\max_i \ |z_i| > \sqrt{2\log n}\} \longrightarrow 0. \tag{10}$$

Essentially (10) says that the probability of all the noise being shrink to zero is very high for large samples. Since the universal threshold is based on this asymptotic result, it does not always perform well in small sample situations. To improve the finite sample properties of the universal threshold, Donoho and Johnstone (1994) suggested that one should always retain coefficients on the first j_0 coarse level even if they do not pass the threshold. Hall and Patil (1996a) and Efromovich (1999) proposed to start thresholding from the resolution level $j_0 = \log_2(n)/(2r+1)$, where r is the regularity of the mother wavelet.

Various alternative data adaptive methods for selecting threshold for wavelet function estimation have been developed. For example Donoho and Johnstone (1995) proposed a sure shrink thresholding rule based on minimizing unbiased risk estimate, Nason (1996) have considered cross validation approach to the choice of λ. The multiple hypothesis procedure is developed by Abramovich and Benjamin (1995,1996). Further modification of the basic thresholding scheme include level dependent and block thresholding. In the first case different thresholds are used on different levels, where as in the second the coefficient are thresholded in blocks rather than individually. Both modification imply better asymptotic properties of the resulting wavelet estimators, for example see Donoho and Johnstone (1998), Hall et al. (1998).

The universal threshold which is also known as visushrink has worse mean square error performance for small and moderate samples. The minimax method does a better job at picking up abrupt jumps at the expense of smoothness in contrary, the universal policy gives smooth estimate that don't pick up jumps or other features.

Table 1: Threshold values

J	$n = 2^J$	Minimax	$\sqrt{J}$	$\sqrt{2 \log n}$
07	00000128	1.669	2.646	3.115
08	00000256	1.860	2.828	3.330
09	00000512	2.048	3.000	3.532
10	00001024	2.232	3.162	3.723
11	00002048	2.414	3.316	3.905
12	00004096	2.594	3.464	4.079
13	00008192	2.773	3.605	4.245
14	00016384	2.952	3.742	4.405
15	00032768	3.131	3.873	4.560
16	00065536	3.313	4.000	4.710
17	00131072	3.503	4.123	4.855
18	00262144	3.686	4.243	4.995
19	00524288	3.869	4.359	5.132
20	01048576	4.052	4.472	5.265
21	02097152	4.234	4.582	5.395
22	04194304	4.417	4.690	5.523
23	08388608	4.600	4.795	5.647
24	16777216	4.783	4.895	5.768
25	33554432	4.966	5.000	5.887

By considering the above facts we proposed a threshold value $\hat{\sigma}\sqrt{J}$ or $\hat{\sigma}\sqrt{\log n/\log 2}$, where J is the power of 2 in the sample of size n and $n = 2^J$. Here $\hat{\sigma}$ can be calculated using robust estimate of the noise level σ based only on the empirical wavelet coefficients at the finest resolution level proposed by Donoho and Johnstone (1995). The robust estimate of the noise level is given by

$$\hat{\sigma} = median(|\hat{d_{J-1,k}}| : for, k = 0, 1, 2, ...2^{J-1} - 1)/0.6745 \tag{11}$$

In contrary to the minimax and the universal threshold this proposed threshold works well for the moderate samples, also up to a certain range of data our threshold value lies between minimax and the universal threshold proposed by Donoho and Johnstone (1994). From the Table 1 we can see that our threshold value is close to the minimax threshold for large values of n, the size of the data.

Fig. 1 shows the four spatially inhomogeneous functions, Blocks, Bumps, Doppler and Heavisine for n = 4096. The formulae are given in Donoho and Johnstone (1994). Fig. 2 shows the noisy version of the four inhomogeneous functions of interest, rescale to have signal to noise ratio (SNR) = 7. Fig. 3 shows the selective wavelet reconstruction using the minimax threshold value, that gives a reconstruction with a minimum risk. Fig. 4 shows the

reconstruction using universal threshold, that gives a smooth reconstruction. Fig. 5 shows the reconstruction using the proposed threshold, that gives smooth reconstruction compare to minimax and little bit noisier but improved risk compare to universal policy for moderate data. In all the reconstructions the threshold is applied to the coefficients from level $j = 0$ to $J - 1$ using the soft threshold rule.

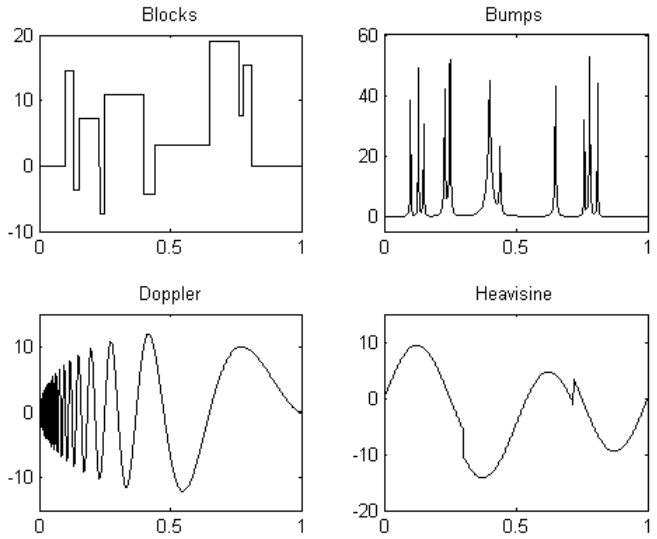


Figure 1: Original functions, *Blocks*, *Bumps*, *Doppler* and *Heavisine* with $n = 4096$.

The threshold value $\hat{\sigma}\sqrt{J}$ which depends on the power of 2 in data of size $n = 2^J$ compare to the universal threshold has improved mean square error, this is because the proposed threshold value is small compare to universal threshold value. Also as the size of the data increases the risk will be reduced at the expense of smoothness. Table 2 illustrates the comparative performance of the minimax, universal and the proposed threshold values in terms of mean square error $\|\hat{f} - f\|_{2,n}^2/n$ from 10 replications of the four spatially inhomogeneous functions Blocks, Bumps, Doppler and Heavisine.

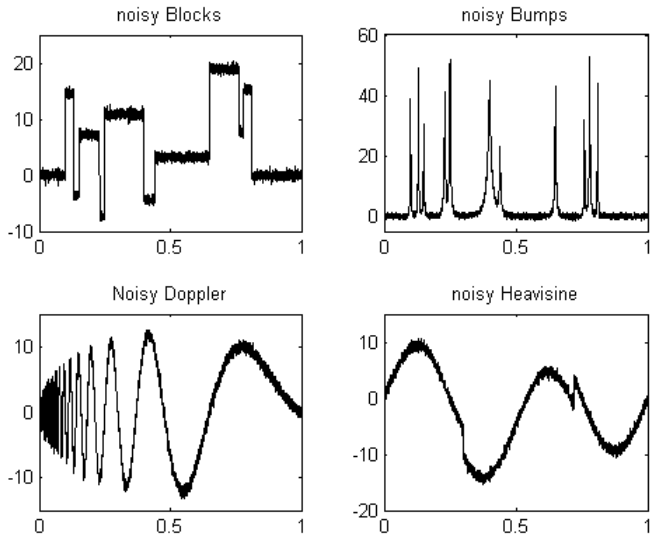


Figure 2: Four functions, *Blocks*, *Bumps* , *Doppler* and *Heavisine* with Gaussian white noise, $\sigma = 1$, SNR $= 7$.

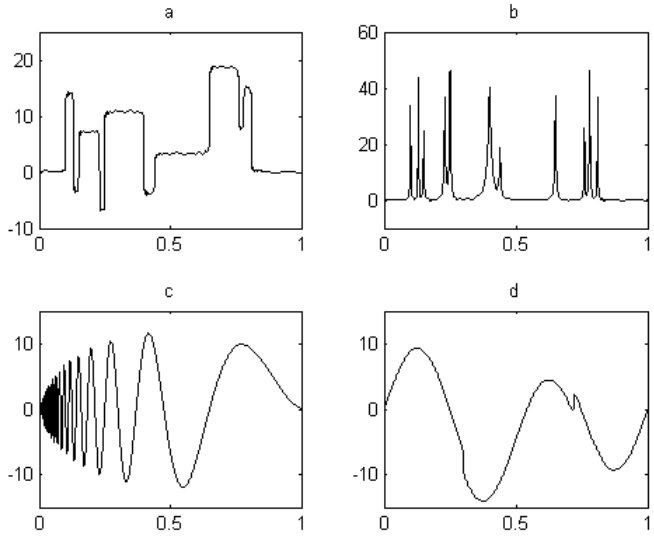


Figure 3: Reconstruction of (a) *Blocks*, (b) *Bumps*, (c) *Doppler*, and (d) *Heavisine* with soft thresholding using minimax policy.

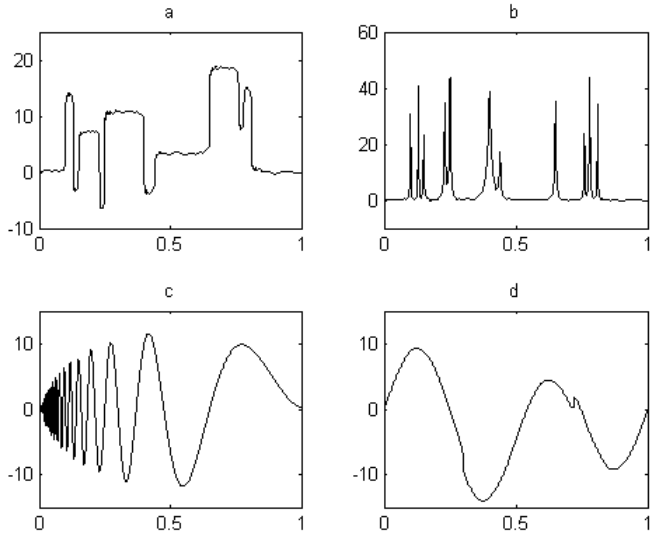


Figure 4: Reconstruction of (a) *Blocks*, (b) *Bumps*, (c) *Doppler*, and (d) *Heavisine* with soft thresholding using universal policy.

Table 2: MSE Performance of the minimax, universal and the proposed threshold values.

Blocks

J	$n = 2^J$	Minimax	$\sqrt{2 \log n}$	$\sqrt{J}$
08	00256	0.854	2.394	1.853
09	00512	0.688	1.768	1.345
10	01024	0.498	1.223	0.937
11	02048	0.357	0.815	0.620
12	04096	0.244	0.521	0.392
13	08192	0.161	0.334	0.250
14	16384	0.104	0.204	0.156

Bumps

J	$n = 2^J$	Minimax	$\sqrt{2 \log n}$	$\sqrt{J}$
08	00256	1.110	2.809	2.093
09	00512	0.785	2.083	1.580
10	01024	0.572	1.414	1.060
11	02048	0.383	0.863	0.662
12	04096	0.236	0.527	0.398
13	08192	0.144	0.307	0.231
14	16384	0.088	0.179	0.136

Doppler

J	$n = 2^J$	Minimax	$\sqrt{2 \log n}$	$\sqrt{J}$
08	00256	0.444	1.290	0.952
09	00512	0.302	0.787	0.609
10	01024	0.210	0.503	0.399
11	02048	0.136	0.307	0.232
12	04096	0.079	0.184	0.136
13	08192	0.043	0.105	0.074
14	16384	0.027	0.057	0.043

Heavisine

J	$n = 2^J$	Minimax	$\sqrt{2 \log n}$	$\sqrt{J}$
08	00256	0.234	0.542	0.421
09	00512	0.157	0.351	0.277
10	01024	0.114	0.244	0.188
11	02048	0.078	0.155	0.122
12	04096	0.049	0.095	0.079
13	08192	0.031	0.063	0.047
14	16384	0.021	0.041	0.030

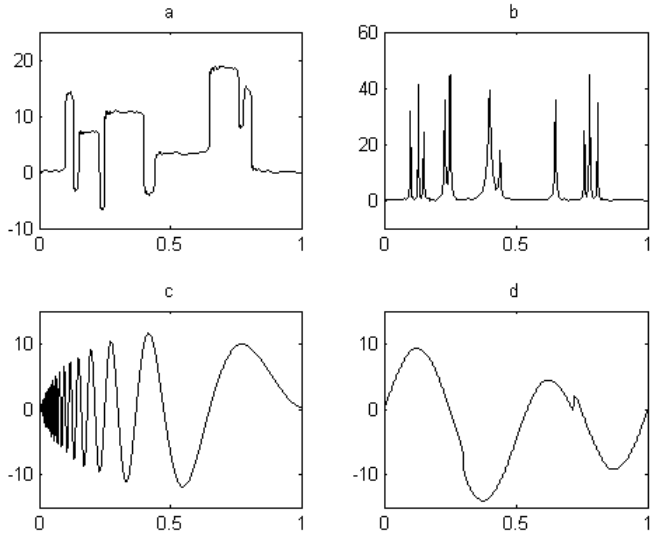


Figure 5: Reconstruction of (a) *Blocks*, (b) *Bumps*, (c) *Doppler*, and (d) *Heavisine* with soft thresholding using the threshold value $sqrt(J)$.

5. COMMENTS AND CONCLUSIONS

In this paper a new threshold value is proposed for the estimation of function using wavelets. We used this threshold value to estimate the true curve by de noising the noisy signals. We have measured the mean square error performance of the proposed threshold value with the existing threshold values and found to be better than universal threshold and for large data the proposed threshold value has MSE close to the minimax threshold value.

The existing minimax and the universal threshold values are close to each other for large values of n, but for moderate data the universal threshold gives a smooth estimate with large MSE and minimax estimator gives small MSE at the cost of smoothness. The proposed threshold value gives a smooth estimate of the true curve with improved MSE. Usually in practice the moderate data is much in use, and in this sense our threshold value is an important contribution to the choice of the threshold values in curve estimation using wavelets.

REFERENCES

Abramovich, F. and Benjamini, Y. (1995). Thresholding the wavelet coefficients as multiple hypothesis testing procedure. *Lect. Notes Statistics*, **103**, 5–14.

Abramovich, F. and Benjamini, Y. (1996). Adaptive Thresholding of wavelet coefficinets. *Comput. Statist. Data Anal*, **22**, 351–361.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximation (with discussion)*J.Am.Stat.Ass.*.96, 960–962.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia:SIAM.

Donoho, D. L. and Johnstone, I. M., (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrica*, **81**, 425–455.

Donoho, D. L. and Johnstone, I. M., (1995). Adapting to unknown smoothness via wavelet shrinkage. *J.Am.Stat.Ass.*, **90**, 1200–1224.

Donoho, D. L. and Johnstone, I. M., (1998). Minimax estimation via wavelet shrinkage. *Ann.statist*, **26**, 879–921.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage asymtopia(with discussion)?, *J.R.Statsi.soc.*B, **57**, 301–337.

Efromovich, S. (1999). Quasi linear wavelet estimation *J.Am.Stat.Ass.*, **94**, 189–204.

Gao, H. Y. and Bruce, A. G. (1997). Waveshrink with firm shrinkage.*Statist.Sin*, **7**, 855–874.

Hall, P. and Patil, P. (1996a). Effect of threshold rules on performance of wavelet based curve estimation.*Statist.Sin*, **6**, 331–345.

Hall, P., Kerkyacharian, G. and Picard, D. (1998). Block threshold rule for curve estimation using kernel and wavelet methods *Ann.Statist*, **26**, 922–942.

Mallat, S.G. (1989). A theory for multiresolution signal decomposition:The wavelet representation *IEEE. Transpattn Anal. Mach .Intell*, **1**, 674–693.

Meyer, Y. (1992).*Wavelets and operators*. Cambridge. Cambridge university press.

Nason, G. P. (1996) Wavelet shrinkage using cross validation *J.R.Statsi.soc.* B, **58**, 463–479.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 796–804.

MODELLING THE IMPACT OF US STOCK MARKET ON ASEAN COUNTRIES STOCK MARKETS

Mohd Tahir Ismail
School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Penang
E-mail: mtahir@cs.usm.my

ABSTRACT

The ASEAN countries which located on the Southeast Asian consist of 10 countries. Most of these countries depend on US as their main trading partner. As a result, if something does happen to US economy it surely will affect the economy of all these countries. Usually stock market fluctuation is used as the main indicator whether the economy of one country is in expansion or recession. Therefore, in this paper, we investigate the impact of US stock market fluctuation on four ASEAN countries stock markets namely Singapore, Indonesia, Thailand and Malaysia. Rather than using linear VAR model we used a two regimes multivariate Markov switching vector autoregressive (MS-VAR) model with regime shifts in both the mean and the variance to show how US stock market affect the four stock markets. Results reveal that when US stock market declines the four stock markets would also follow the same trend of declines and vice versa. In addition, the MS-VAR model fitted the data better than the linear vector autoregressive model (VAR).

1. INTRODUCTION

The Association of Southeast Asian Nation or ASEAN was established in 1967 consists of 10 countries namely Indonesia, Malaysia, Philippines, Singapore, Thailand, Brunei, Laos, Cambodia, Vietnam and Myanmar. One of the purposes of the ASEAN association was to accelerate economy growth, social progress and cultural development in the region. All the countries not only have a trade agreement between each other but they also have the same main trading partner outside the region. Many investors are attracted to invest in ASEAN region because of low wages labour, a lot of raw materials and many attractive incentives from the government of each country to the investors. US were the biggest trading partner follow by Japan then China and India was not far behind. Except for Singapore all the other countries are developing countries.

It is well known that US is main trading partner of many developing countries. Therefore, whatever happens to the US economy will also affect the economy of these countries. Usually the declining and increasing of the stock market is used as an indicator whether a country is in recession or expansion. This inter-relationship phenomenon in international market is not only a result of the liberalization of capital markets in developed and developing countries and the increasing variety and complexity of financial instrument but also a result of the increasing relatively of the developing and developed economies as developing countries become more integrated in international flow of trade and payment. As a result, this has triggered the interest of economists and policy makers to find the linkages between the stock market of developed countries mainly the US and the stock market of developing countries.

Numerous related studies on the relationship between stock market of US and developing countries have been done by researchers. For instance, Ghosh et al. (1999) examined whether the stock markets of nine Asian-Pacific countries are driven by US or Japan stock market during the financial turmoil in 1997 using the theory of cointegration. They had identified nine stock markets which can be divided into three groups; those that move with the US stock market, those that move with Japan stock market and those that are not affected by the two stock markets. Then Arshanapalli and Kulkarni (2001) studied the interdependence between Indian stock market and the US stock market and the results showed that the Indian stock market was not interrelated with the US stock market.

Later, Yang et al (2003), investigated the long run relationship and short-run dynamic causal linkages among the US, Japanese, and ten Asian emerging stock markets. They discovered that both long-run cointegration relationships and short-run causal linkages among these markets were strengthened during the financial crisis in 1997 and that these markets have generally been more integrated after the 1997 crisis than before the crisis. Wang et al. (2003) studied relationship among the five largest emerging African stock markets and US market and uncovered that both long-run relationships and short-run causal linkages show that regional integration between most of African stock markets was weakened after the 1997–1998 crisis. Finally, Serrano and Rivero (2003), revealed the mixed results on the existence of long run relationship due to structural breaks between the US and Latin Americans stock markets.

It appears that most of the research mention above did not focus on the ASEAN region specifically. Furthermore all the papers used similar methodology to analyze the interaction among the stock market. They begin their studies by finding whether the variables are cointegrated or not using cointegration test and followed by modelling the variables using Vector Autoregressive (VAR) or Vector Error Correction (VEC) to show the existent of short run or long run relationships among the variables. However in this paper we focus on finding the relationship between 4 ASEAN countries stock markets and the US stock markets. We also apply a different approach to study the interaction between the US and the four stock markets. Rather then finding linear interaction, we concentrate on investigate whether nonlinear interaction because of common regime switching behaviour exists among the stock markets by assuming that all the series are regime dependent. We use a two regimes multivariate Markov Switching Vector Autoregressive (MS-VAR) model with regime shifts that happened in both the mean and the variance to extract common regime switching behaviour from all the series.

This paper is organized as follows. The specification and estimation of the Markov Switching Vector Autoregressive model are given in Section II. Section III presents the empirical results and discussion on the results. Section IV contains the summary and the conclusion.

2. MARKOV SWITCHING VECTOR AUTOREGRESSIVE (MS-VAR) MODEL

Hamilton in 1989 developed the Markov Switching Autoregressive model (MS-AR) to identify changes between fast and slow growth regimes in the US economy. The model assume that a time series, y_t is normally distributed with μ_i in each of k possible regime where $i = 1,2,...,k$. A MS-AR model of two states with an AR process of order p, $MS-AR(p)$ is given as follows:

$$\begin{aligned} y_t &= \mu(s_t) + \left[\sum_{i=1}^{p} \alpha_i \left(y_{t-i} - \mu(s_{t-i})\right)\right] + u_t \\ u_t | s_t &\sim NID\left(0, \sigma^2\right) \quad and \quad s_t = 1,2 \end{aligned} \tag{1}$$

where α_i are the autoregressive parameters with $i = 1,2,...,p$.

The MS-AR framework of Equation (1) can be readily extended to MS-VAR model with two regimes that allows the mean and the variance to shifts simultaneously across the regime. The model is given below:

$$\begin{aligned} Y_t - \psi(s_t) &= A_1(s_t)\left(Y_{t-1} - \psi(s_{t-1})\right) + \ldots \\ &+ A_p(s_t)\left(Y_{t-p} - \psi\left(s_{t-p}\right)\right) + \varepsilon_t \end{aligned} \tag{2}$$

where $Y_t = (Y_{1t}, \ldots, Y_{nt})$ is the n dimensional time series vector, ψ is the vector of means, $A_1, \ldots, A_p$ are the matrices containing the autoregressive parameters, and ε_t is the white noise vector process such that $\varepsilon_t | s_t \sim NID\left(0, \Sigma(s_t)\right)$ Other specifications of MS-VAR model are being discussed by Krolzig (1997).

From Equation (1) and (2), s_t is a random variable that triggers the behaviour of Y_t to change from one regime to another. Therefore the simplest time series model that can describe a discrete value random variable such as the unobserved regime variable s_t is the Markov chain. Generally, s_t follow a first order Markov process where it implies that the current regime s_t depends on the regime one period ago, s_{t-1} and denoted as:

$$\begin{aligned} &P\left[s_t = j \middle| s_{t-1} = i, s_{t-2} = k, \ldots\right] \\ &= P\left[s_t = j \middle| s_{t-1} = i\right] = p_{ij} \end{aligned} \tag{3}$$

where p_{ij} is the transition probability from one regime to another. From m regimes, these transition probabilities can be collected in a $(m \times m)$ transition matrix denoted as P.

$$\mathrm{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{bmatrix} \tag{4}$$

with $\sum_{j=1}^{m} p_{ij} = 1, \quad i = 1, 2, \ldots, m \quad and \quad 0 \le p_{ij} \le 1$.

The transition probabilities also provide the expected duration that is the expected length the system is going to be stay in a certain regime. Let D define the duration of regime j. Then, the expected duration of the regime j is given by

$$E(D) = \frac{1}{1 - p_{jj}} \quad j = 1, 2, \ldots \tag{5}$$

The conventional procedure for estimating the model parameters is to maximize the log-likelihood function and then use these parameters to obtain the filtered and smoothed inferences for the unobserved regime variable s_t. However, this method becomes disadvantageous as the number of parameters to be estimated increases. Generally, in such cases, the Expectation Maximization (EM) algorithm is used. This technique starts with the initial estimates of the unobserved regime variable, s_t and iteratively produces a new joint distribution that increases the probability of observed data. These two steps are referred to Hamilton (1994) and Kim and Nelson (1999).

3. MODELLING DYNAMIC RELATIONSHIP

This section presents the results of the econometric specifications used for modelling the relationship between US and four ASEAN countries stock markets. It begins with a description of the data and testing for stationary using two unit root tests. Then if the data is stationary at the same order, Johansen test is used to examine the existent of cointegration. Later, the MS-VAR model is used to show the dynamic relationships.

3.1 Data

The data under investigation are 10 years old monthly average data from August 1999 until Julai 2009 which includes US Dow Jones Index (DWJON) and four ASEAN stock markets namely Kuala Lumpur Composite Index (KLCI), Jakarta Composite Index (JKI), Singapore Straits Time Index (STI) and Thailand Composite Index (TSI). Figure 1 and Figure 2 show the behaviour of the original and return series (which is the first difference of natural logarithms multiplied by 100 to express them in percentage terms) of the DWJON Index, the KLCI Index, the JKI index, the STI index and the TSI Index over the study period. Close inspection of the two figures reveals that the trend of up and down in the original series and the large positive and negative returns happen quite similar for the five series.

3.2 Stationarity and Cointegration Tests

Many of the econometric models require the knowledge of stationarity and order of integration for the variables. The unit root test is usually used to determine whether the order of integration of a variable is at level or first differences. Two of the common unit root tests are used in this paper namely the ADF test and the PP test. Besides that the two tests have been implemented with and without time trend. The ADF test was developed by Dickey and Fuller (1979) and the PP tests was suggested by Philips and Perron (1988).

From Table 1, most of the statistics for series at level are not significant. This suggests that the null hypothesis of unit root test cannot be rejected and the indices are not stationary at level. After first differencing has been employed for the series, the null hypothesis of unit root test can be rejected at 1% level of significance for series with or without trend, Thus, the series are stationary at first difference and integrated of order 1, I(1). Thus, the cointegration test can be carried out after all the series are integrated at the same order.

The Johansen and Juselius (1990) cointegration test or JJ test is carried out to examine the existence of the long-run relationship among the indices. This test identifies the number of the cointegration vector by using the maximum likelihood method. Two test statistics are used to test the presence of *r* cointegrating vectors, namely trace statistic and maximum eigen statistic. The existence of cointegration among the variables indicates the rejection of the non-causality among the variables. The result of the cointegration test is shown in Table 2 and *r* represents the number of the cointegration relationships of the hypothesis test.

According to Table 2, both trace statistic and maximal eigen statistic suggests that there is no cointegrating vector at 5% level of significance. Thus, each indices does not sustain a stable equilibrium relationship with each other's therefore, this suggests that there is no long-run cointegration among the indices. Next we modeled the relationship among the return series using MS-VAR model.

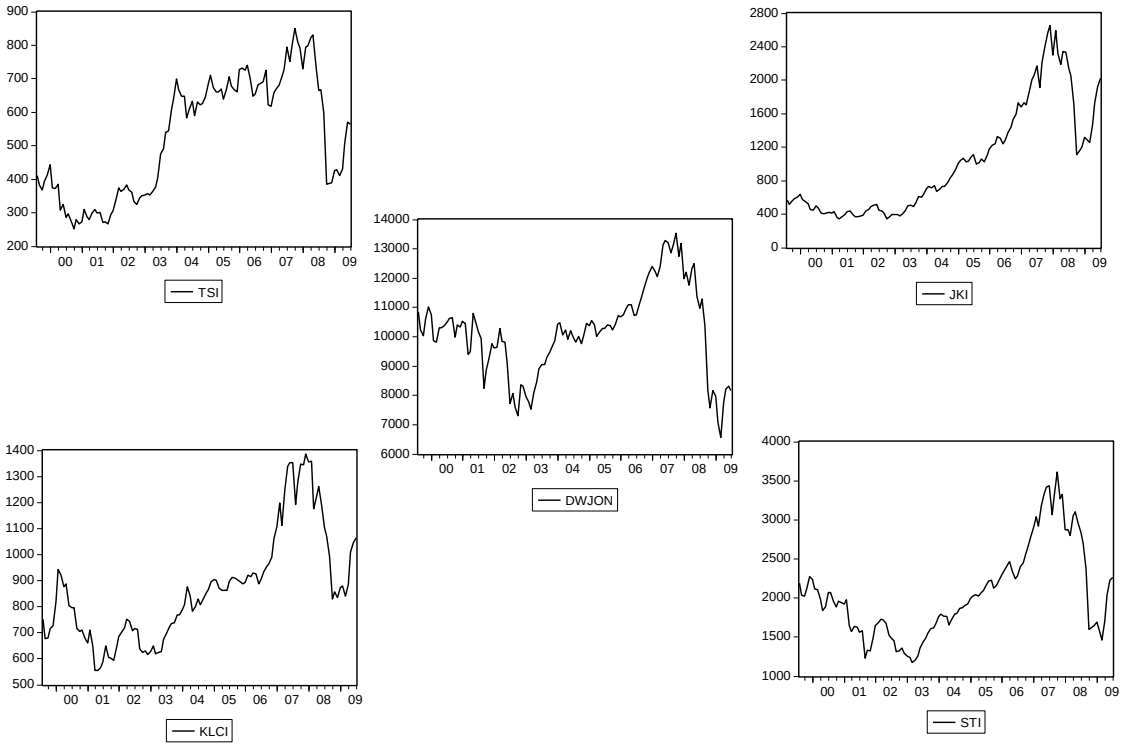


Figure 1: Original Series

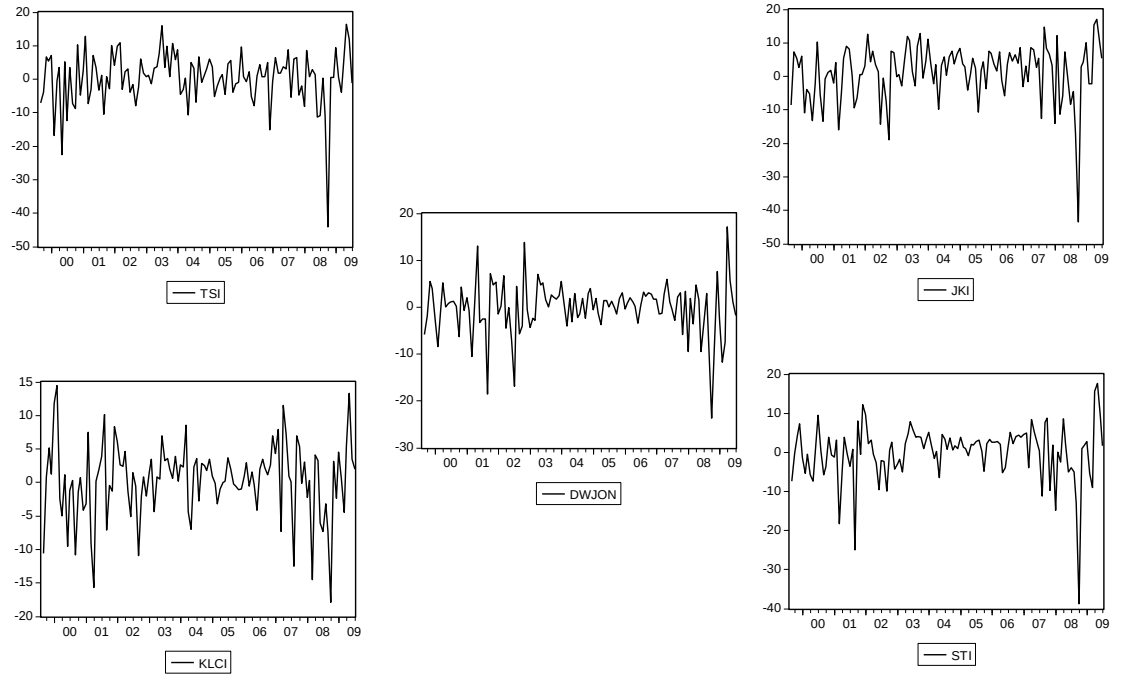


Figure 2: Return Series

Table 1 The Two Unit Root Tests

ADF test for sector indices

Variables	Level		1st Differentiation	
	No Trend	Trend	No trend	Trend
DWJON	-1.702	-1.642	-10.117**	-10.102**
STI	-1.356	-1.838	-9.503**	-9.015**
TSI	-1.278	-1.524	-9.639**	-9.610**
JKI	-0.429	-2.064	-9.323**	-9.139**
KLCI	-1.209	-2.031	-9.840**	-9.796**

PP test for sector indices

Variables	Level		1st Differentiation	
	No Trend	Trend	No trend	Trend
DWJON	-1.729	-1.671	-10.096**	-10.064**
STI	-1.542	-2.063	-9.059**	-9.021**
TSI	-1.487	-1.926	-9.668**	-9.638**
JKI	-0.669	-2.371	-9.323**	-9.139**
KLCI	-1.383	-2.299	-9.855**	-9.811**

Note:** indicates significance at 5%

Table 2 JJ Cointegration Test for Indices

Null hypothesis	Trace		Max-eigen	
	Statistic	5% critical value	Statistic	5% critical value
$r = 0$	56.47	68.52	26.68	33.46
$r \leq 1$	29.78	47.21	17.68	27.07
$r \leq 2$	12.09	29.68	6.55	20.97
$r \leq 3$	5.54	15.41	3.79	14.07
$r \leq 4$	1.74	3.76	1.74	3.76

3.3 Estimating MS-VAR Model

Following the principle of parsimony, we found that two regimes Markov Switching Vector Autoregressive model of order one with switching in the mean and the variance or MS-VAR(1) manage to capture the interaction among the five series very well. Before further discussing the estimation model, we need to determine whether regime shifts happened in the five return series. For this purpose, we use the likelihood ratio (LR) test suggested by Garcia and Perron (1996). As denoted in Table 3, the likelihood ratio test for testing the null hypothesis of linear model against an alternative of regime switching model, it is found that the null hypothesis can be rejected because the Davies (1987) *p*-value (value in the [] bracket) show significance results. Therefore, a nonlinear MS-VAR(1) model is better than linear VAR(1) model in describing the data. Moreover, the minimum value of AIC (Akaike), HQC (Hannan-Quinn) and SBC (Schwartz Bayesian) criteria indicate that the performance of the MS-VAR(1) models are better than the nested linear VAR(1) model.

Table 3 Model Comparison

	MS-VAR(1)	Linear VAR(1)
Log-likelihood	-1699.1097	-1765.3312
AIC	29.9341	30.6836
HQC	30.5728	31.1126
SBC	31.5072	31.7402
Log-likelihood Ratio (LR) Test	132.4430 [.000]	

Table 4 reports the parameters estimated of the two regimes MS-VAR (1). It can be seen from Table 4 that the estimated means of the MS-VAR(1) model for each of the two regimes has a clear economic interpretation. The first regime $(s_t = 1)$ indicates that all the stock market indices are in the Bear market or contraction phase with negative sign of the monthly expected return, $\mu(s_t = 1)$ and higher volatility, $\sigma^2(s_t = 1)$ Conversely, the second regime captures the Bull market or expansion phase of the stock market indices with positive sign of the monthly expected return, $\mu(s_t = 2)$ and lower volatility $\sigma^2(s_t = 2)$. However, the probabilities of staying in regime 1 and regime 2 are almost the same 0.8678 and 0.8660 respectively. It means on average the duration of staying in either regime is 7 to 8 months.

Furthermore, the main advantage of using MS-VAR model is that it provides us with smoothed regime probability plots of regime 1 and regime 2 which are the probability of staying in either regime 1 or regime 2 at time *t*. As seen in Figure 2, the smoothed probabilities of regime 1 are near one just after the smoothed probabilities of regime 2 are near zero. While Table 5 stated all the dating of staying in each regime. This means the smoothed regime probability plot tell us at which point in time all the series follow the same behavior which is either all the indices are increasing (regime 2) or decreasing (regime 1).

Table 4 Estimation of MS-VAR (1) model for Dynamic Relationship

	$DWJON_t$	STI_t	TSI_t	JKI_t	$KLCI_t$
	Regime-dependent means				
$\mu(s_t = 1)$	-1.4513	-1.8210	-1.6633	-1.5344	-0.4866
$\mu(s_t = 2)$	0.9551	1.8229	2.1968	3.6816	1.1531
	Coefficients				
$DWJON_{t-1}$	0.00672	0.1267	0.0341	0.1871	0.0650
STI_{t-1}	0.2161	0.3698	0.4723	0.4017	0.2226
TSI_{t-1}	-0.1129	-0.1158	-0.1738	0.0308	-0.0575
JKI_{t-1}	0.0269	0.0319	0.0401	-0.0397	0.0945
$KLCI_{t-1}$	-0.1150	-0.1676	-0.0502	-0.1857	-0.1261
	Regime-dependent variances				
$\sigma^2(s_t = 1)$	6.8992	8.9063	9.1021	10.0091	7.0208
$\sigma^2(s_t = 2)$	2.7194	2.5893	4.7544	4.0698	2.1378
p_{ij}	$s_{t-1} = 1$		$s_{t-1} = 2$		$E(D)$
$s_t = 1$	0.8678		0.1322		7.56
$s_t = 2$	0.1340		0.8660		7.46

As note on Table 5, the contraction period in early 2000 and 2001 happen because of the US economic downturn as the IT industry crash follow by the September 11 2001 attack on US. Nevertheless the longest contraction period happen from August 2007 until Mei 2009 with inline with the recession period in US. This is the longest and deepest recession period since 1930an recession. The recession period was triggered by the US housing market collapse and the ensuing global credit crisis. Until recently US recover from this recession and the MS-VAR model manages to capture it. The finding from Figure 3 and Table 5 ensure us that the suggestion of regime 1 as the state where all the stock markets are in the recession phase or the bear market and regime 2 as the state where all the stock markets are in the expansion phase or the bull market by using the estimated parameters is justified.

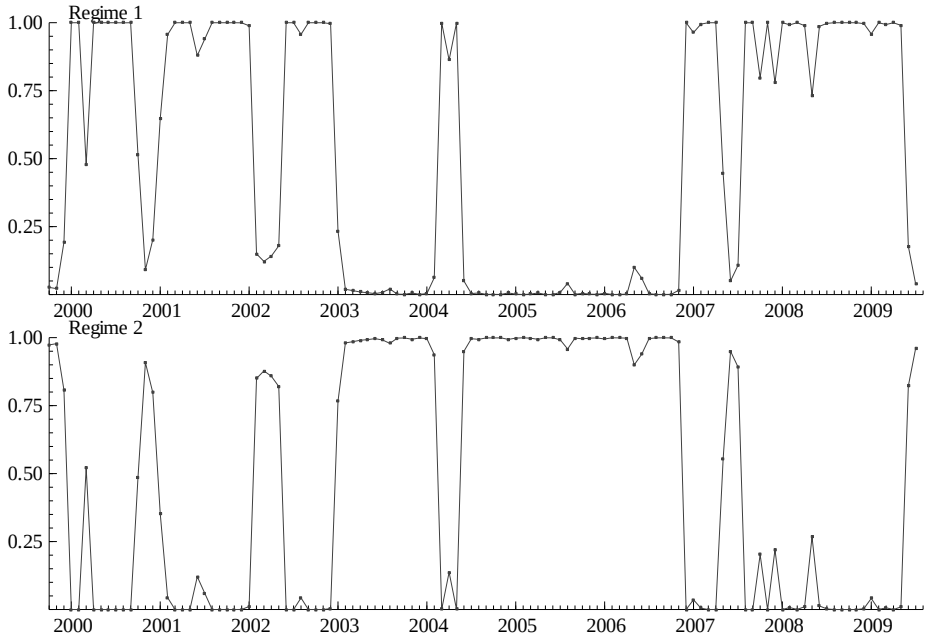


Figure 3 Smoothed Probability Plots of the MS-VAR(1) model

Table 5 Duration of Regime 1 and Regime 2

Regime 1 $(s_t = 1)$ (Contraction Period)	Regime 2 $(s_t = 2)$ (Expansion Period)
2000:1 - 2000:2 [1.0000]	1999:10 - 1999:12 [0.9197]
2000:4 - 2000:10 [0.9306]	2000:3 - 2000:3 [0.5211]
2001:1 - 2002:1 [0.9546]	2000:11 - 2000:12 [0.8532]
2002:6 - 2002:12 [0.9933]	2002:2 - 2002:5 [0.8517]
2004:3 - 2004:5 [0.9536]	2003:1 - 2004:2 [0.9719]
2006:12 - 2007:4 [0.9917]	2004:6 - 2006:11 [0.9891]
2007:8 - 2009:5 [0.9643]	2007:5 - 2007:7 [0.7993]
	2009:6 - 2009:7 [0.8924]

4. COMMENTS AND CONCLUSION

In this paper we have discussed modelling the interactions of US stock market (DWJON) and 4 ASEAN stock markets namely the KLCI (Malaysia), STI (Singapore), TSI (Thailand), and JKI (Indonesia). Results showed that the 4 ASEAN stock markets really depend on the increasing and decreasing of the US stock market. In addition the MS-VAR(1) model outperform linear VAR(1) in modelling the interaction.

REFERENCES

Arshanapalli, B and Kulkarni, M. S, (2001). Interrelationship between Indian and US Stock Markets. *Journal of Management Research*, 1, 141-148.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 74, 33-43.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74,:427-431.

Garcia, R. and Perron, P (1996). An analysis of the real interest rate under regime shifts. *Review of Economics and Statistics*, 78, 111-125.

Ghosh, A., Saidi, R and Johnson, K. H. (1999). Who moves the Asia-Pacific stock markets- US or Japan? Empirical evidence based on the theory of cointegration, *The Financial Review*, 34, 159-170.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*,57, 357-384.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton: Princeton University Press.

Johansen, S and Juselius, K (1990). Maximum likelihood estimation and inference on cointegration with applications to the demand for money. *Oxford Bulletin of Economics and Statistics*, 52, 169–210.

Kim, C. J. and Nelson, C. R. (1999). *State-space Models with Regime Switching: Classical and Gibbs-sampling Approaches with Application.* Cambridge: The MIT Press.

Krolzig, H. -M. (1997). *Markov-switching Vector Autoregression*. Berlin: Springer.

Phillips, P. C. B., and Perron, P. (1988). Testing for unit root in time series regression, *Biometrika*, 75, 335-346.

Serrano, J. L. F. and Rivero, S. S. (2003). Modelling the linkages between US and Latin American stock markets. *Applied Economics*, 35, 1423–1434.

Wang, Z., Yang, J and Bessler, D. A. (2003). Financial crisis and African stock market integration. *Applied Economics Letters*, 10, 527–533.

Yang, J., Kolari, J. W. and Min, I. (2003). Stock market integration and financial crises: the case of Asia. *Applied Financial Economics*, 13, 477–486.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 805–816.

THE RELATIONSHIP BETWEEN EDUCATION AND OCCUPATION USING FULLY AND PARTIALLY LATENT MODELS

Faisal G. Khamis[1], Muna F. Hanoon and Abdelhafid Belarbi
Faculty of Economics and Administrative Sciences, Al-Zaytoonah University of Jordan, Amman, Jordan
E-mail: [1] faisal_alshamari@yahoo.com

ABSTRACT

Several studies have been carried out to examine the association between education and occupation. These studies were useful for the purpose of intervention and policy making. In this study we examined the relationship between education factor which includes three indicators: the percentages of population who achieved (primary, secondary and tertiary) education and occupation factor which also includes three indicators: the percentages of (CLASS1, CLASS2 and CLASS3) of occupation by using fully and partially latent models. The data were collected from the information of 81 districts based on the census conducted in peninsular Malaysia in 1995. The goodness of fit indices for the assumed models were examined. No significant relationship was found between the educational achievement and the occupation factor. This study is composed of a number of path-diagrams to create a picture for the socioeconomic status in Malaysia.

Keywords: occupation, education, relationship, fully latent models, MIMIC models.

1. INTRODUCTION

The education of people in the community is potentially important because it may influence society in ways that affect everyone. For example, most hospitals and health centers in Malaysia were public, with financing from national sources, and were subjected to national quality regulations, but when many local people are well educated, it is perhaps easier to recruit qualified health for all the members of the family. Besides, education was a major determinant of income (Kravdal 2008). A low relative education was at least linked with low relative income. Blane, Brunner and Wilkinson (1996) stated that men and women with low educational attainment were the least likely or slowest to respond to the messages of health education. The results of Ross and Wu (1995) demonstrated a positive association between education and health and help explain why the association exists. (1) Compared to the poorly educated, well educated respondents were less likely to be unemployed, were more likely to work full time, to have fulfilling, subjectively rewarding jobs, high incomes and low economic hardship. (2) The well educated reported a greater sense of control over their lives and their health, and they had higher levels of social support. Duper (2008) used regression models to examine how education relates to low income and unemployment.

Anderson (1980) stated that the integration of work-experience education within the curriculum helps to prepare the student for a practical, productive life, which means that

interpersonal skills can be developed through student-teacher-employer relationships. Also, work-experience education program guide the student into an awareness of his/her responsibilities as a citizen. Accumulating evidence suggests that a highly qualified workforce contributes substantially to a nation's economic competitiveness, particularly when a large share of the workforce had acquired skills and knowledge through higher education; and these findings applied to states as well as nations; where U.S states that improved opportunities for education and training beyond high school advanced their residents' employment prospects and the competitiveness of their overall workforce (Wanger 2006). The World Development Report (2006) suggested that although curricula and teaching methods had remained largely unchanged in developing countries over the years, employers were increasingly demanding strong thinking, communication, and entrepreneurial skills demands largely unmet by educational systems in the developing and transition economics.

The literature on human capital accumulation indicated that high quality education at the primary level generates the highest returns, both at the primary level and all levels thereafter in both developing and transition countries. Fasih (2008) stated that, if the relationship of education and earnings is convex or linear, then expanding enrollment only at lower levels of education will not raise earnings substantially, and consequently not prove to be an effective means of helping people out of poverty. In developing and transitional countries such as Malaysia where there were large disparities in the quality of education between the rich and the poor, and where individuals were systematically sorted into high-quality schools by wealth, the poor were attained fewer skills for the same "quantity" of education. The policy option in such a case would be to counter the sorting process through the provision of choice of better schooling through, for example, school vouchers or better-quality publicly funded private schools for the poor (Angrist, Bettinger & Kremer 2006; Barrera-Osorio 2007). When Bertrand (1994) examined the education in terms of its usefulness as a preparation for employment, he stated that the theoretical analysis of education contribution to the productivity of labour and the methods used to forecast the quantitative, needs of the economy gave rise to considerable controversy and seemed to provide no more than very general indications. Also, Bertrand found that economy needs to provide enough jobs to meet demand, which is becoming an increasingly unlikely prospect in many countries and would call for some rethinking of education's role in this field. So the purpose of this research is to provide some implications for the policy makers regarding the increasing of opportunities for all members in the community to increase and enhance their levels of education. This increasing will probably increase the opportunities of getting a job and also enhance the levels of jobs with better environment and salary. It is not easy to decide which way higher education ought to go. It is clear that the modern economy demands a higher proportion of highly qualified personnel, but it is difficult to say to what extent. Levin and Rumberger (1989) stated that over-education stemmed from a more rapid increase in the number of university graduates was greater than offers of employment. The European trend towards extended study, for example, is certainly caused more by social demand than by the needs of the economy and will probably lead to frustration among young people, who will not always be able to find the high-level employment that they expect.

In this study, structural equation modeling (SEM) was used, where SEM was defined as hybrid model since it was a mixed system of equations between structural equation and measurement equations. There were many advantages of SEM technique making it applicable in many situations. First SEM technique has several flexible assumptions, such as allowing for correlations between independent variables, thus providing solution for multicollinearity problem

in regression analysis. Second, SEM allowed for the use of factor analysis to reduce the measurement error by having multiple indicators (manifest variables) per latent (factor) variable. Third, SEM had a structural graph of attraction because it provides graphical modeling interface. Fourth, SEM provided mechanism for testing overall model rather than testing each individual coefficient in the model, so that complex relationships can be easily identified and understood. Fifth, SEM had the ability to test models with multiple dependents. Sixth, SEM had the ability to model the mediating latent variables.

The SEM approach was convenient because it allows multiple measures of the same characteristic to be included in the model, where this approach may reduce potential bias from measurement error in the observed variables (Chandola 2005). As well as, SEM had characteristics which allow the results to be more informative compared to the more traditional applied multiple regression and path analysis techniques. Also, SEM allows a range of relations between variables to be recognized in the analysis compared to multiple regression analysis, and those relations can be recursive and non-recursive (Smith & Langfield-Smith 2004). Thus, SEM provides the researcher with an opportunity to adopt a more holistic approach to model building.

2. MATERIALS AND METHODS

2.1 Data

The data were collected from the department of statistics (Malaysia, 1995) based on the census of 81 districts conducted in peninsular Malaysia. We must construct on the basis of the prior concept or statistical analyses, which particular *indicators* load on each latent variable. More precisely, we constructed the following latent variables with their respective indicators:

Occupation factor: occupation latent factor includes three classes of occupation, starting from top to bottom in the income and social level were used as follows: CLASS1 included professional, administrative and managerial workers; CLASS2 included clerical workers; and CLASS3 included sales, and service workers. All classes were measured in percentages. These indicators described the type of occupation status for people living in the district.

Education factor: education latent factor included three indicators: percentages of population who achieved (primary, secondary and tertiary) education. A strong public economy resulting from a high average education may allow more generosity with respect to social support, and high individual incomes may trigger the establishing of some smaller private health services. Another possibility is that a higher level of education may increase the chance that the individual has a well paid job in the advanced service sector, which may offer some health advantages. Education attainment may reflect a person's capacity to absorb new information and to act on it (Nordstrom, Cnattingius & Haglund 1993). The focus was on education, which is readily available, often used, and theoretically meaningful indicator.

2.2 Analysis

Fully latent models: Fully latent models or SEM is an extension of standard regression models through which multivariate outcomes and latent variables can be modeled. SEM is more appropriate for this application than alternative causal modeling technique because they permit

specification of "measurement models". SEM needs two types of models: the measurement model (outer model), which connects the manifest variables to the latent variables and the structural model (inner model), which connects latent variables between them. Slight to moderate departures from normality can be handled by the maximum likelihood (ML) method (Raykov et al. 1991). In the observed variables, we found slight departure from normality. ML estimates were quite robust to violation of normality assumption in the factor model (Bentler 1980; Joreskog & Sorbom 1982). The causal variable was called exogenous variable, ξ, and the effect variable was called the endogenous variable, η. Unexplained variation was referred to as disturbance. The aim was to test the synthesized model of relations between the latent variables, where the structural equation model can be written as: $\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}$. Vectors $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are not observed, instead vectors $\mathbf{y}$ and $\mathbf{x}$ are observed, such that:

Measurement model for $\mathbf{y}$: $\mathbf{y} = \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon}$, and measurement model for $\mathbf{x}$: $\mathbf{x} = \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta}$.

MIMIC or partially latent models: the term MIMIC stands for Multiple Indicators and Multiple Causes (Jöreskog & Sörbom 2001). MIMIC model involves two types of models: the measurement model (outer model), which relates the indicators to the latent variables and the structural model (inner model), which explain the relationship between latents. The structural equation model is: $\boldsymbol{\eta} = \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta}$, and measurement model for $\mathbf{y}$: $\mathbf{y} = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is a $p \times 1$ vector of response variables, $\mathbf{x}$ is a $q \times 1$ vector of predictors, $\boldsymbol{\eta}$ is an $m \times 1$ random vector of latent dependent, or endogenous variables, $\boldsymbol{\varepsilon}$ is a $p \times 1$ vector of measurement errors in $\mathbf{y}$, $\boldsymbol{\Lambda}$ is a $p \times m$ matrix of coefficients of the regression of $\mathbf{y}$ on $\boldsymbol{\eta}$. The coefficients of $\boldsymbol{\Lambda}$ are the weights or factor loadings that relate the observed measures to the latents. The $\boldsymbol{\Gamma}$ is an $m \times q$ matrix of coefficients of the x-variables in the structural relationship. The elements of $\boldsymbol{\Gamma}$ represent direct causal effects of x-variables on η-variables. The $\boldsymbol{\zeta}$ is an $m \times 1$ vector of random disturbances in the structural relationship between $\boldsymbol{\eta}$ and $\mathbf{x}$, where in this study: $p = 3, q = 3$ and $m = 1$. The random components in LISREL model were assumed to satisfy the following minimal assumptions: $\boldsymbol{\varepsilon}$ is uncorrelated with $\boldsymbol{\eta}$, $\boldsymbol{\zeta}$ is uncorrelated with $\mathbf{x}$, and $\boldsymbol{\zeta}$ and $\boldsymbol{\varepsilon}$ are mutually uncorrelated. The model is identified if there are two or more latents and each latent has at least two indicators (Bollen 1989; Kline 1998). The models under study were identified since each of education and occupation latent variables included three indicators.

Parameter estimation: Parameter estimation was performed by ML estimation. The unknown parameters of the model are estimated so as to make the variances and covariances that are reproduced from the model in some sense close to the observed data. Obviously, a good model would allow very close approximation to the data. The proposed models are designed specifically to answer such questions as: Is the link between occupation and education myth or reality? From the previous studies, this link was reality in some countries but what about Malaysia?

Fit indexes: Perhaps the most basic fit index was the likelihood ratio, which was sometimes called Chi-square (χ^2) in the SEM literature. The value of the χ^2-statistic reflects the sample size and the value of the ML fitting function. The fitting function is the statistical criterion that ML attempts to minimize and is analogous to the least squares criterion of regression. For a particular model to be adequate, values of indexes that indicate absolute or relative proportions

of the observed covariances explained by the model such as the Goodness-of-Fit Index (GFI), the Adjusted Goodness-of-Fit Index (AGFI), and Normed Fit Index (NFI) should be greater than 0.90 (Bollen 1989; Hair *et al*. 1998). Comparative fit index (CFI) indicates the proportion in the improvement of the overall fit of the researcher's model relative to a null model like NFI but may be less affected by sample size. CFI should be greater than 0.90 (Kline 1998) or Hu and Bentler (1999) endorsed stricter standards, pushing CFI to about 0.95. Another widely used index is the standardized Root Mean Squared Residual (SRMR), which is a standardized summary of the average covariance residuals. Covariance residuals are the differences between the observed and model-implied covariances. A favorable value of the SRMR is less than 0.10 (Hu & Bentler 1999). Another measure based on statistical information theory is the Akaike Information Criterion (AIC). It is a comparative measure between models with different numbers of latents. AIC values closer to zero indicate better fit and greater parsimony (Bollen 1989; Hair *et al*. 1998).

The parsimonious goodness-of-fit index (PGFI) modifies the GFI differently from the AGFI; where the AGFI's adjustment of the GFI is based on the degrees of freedom in the estimated and null models. The PGFI is based on the parsimony of the estimated model (Hair *et al*. 1998), where this index varies between 0 and 1, with higher values indicating greater model parsimony. The Non-Normed Fit Index (NNFI) includes a correction for model complexity, much like the AGFI; a recommended value is 0.90 or greater (Hair *et al*. 1998). The Root Mean Square Error of Approximation (RMSEA) value below or equal to 0.08 is deemed acceptable (Hair *et al*. 1998) or Hu and Bentler (1999) pushes RMSEA values to smaller than 0.06 and they considered it greater than 0.10 as poor fit. RMSEA is a measure to assess how well a given model approximates the true model (Bollen 1989).

Path diagrams: A popular way to conceptualize a model was using a path diagram, which was a schematic drawing of the system (model) to be estimated. There were a few simple rules that assist in creating these diagrams: ovals represented latent variables. Indicators were represented by rectangles. Directional relations were indicated using a single-headed arrow. The expression "a picture is worth a thousand words" is a very apt one for SEM. Researchers who used SEM techniques often used path-diagrams to illustrate their hypotheses and summarize the results of the analysis. Figures 1 and 2 were shown a conceptualized path diagrams for the proposed models 1 and 2 respectively, explaining the parameters required to be estimated.

The sample design included two latent factors. The education factor, ξ, which constructed from three indicators, x_1, x_2, and x_3, that represented three levels of education, primary, secondary and tertiary respectively. The occupation factor, η, which included also three indicators, y_1, y_2 and y_3 that represented CLASS1, CLASS2 and CLASS3 of occupation respectively. For model 1 the analysis included the following SEM model:

$$\eta = \gamma\xi + \varsigma,\ \mathbf{y} = \eta\boldsymbol{\Lambda}_y + \boldsymbol{\varepsilon},\ \text{and}\ \mathbf{x} = \xi\boldsymbol{\Lambda}_x + \boldsymbol{\delta},$$

and for model 2: $\eta = \boldsymbol{\Gamma}\mathbf{x} + \varsigma$, and $\mathbf{y} = \eta\boldsymbol{\Lambda}_y + \boldsymbol{\varepsilon}$. where, $\boldsymbol{\Lambda}_y$ and $\boldsymbol{\Lambda}_x$ represented a vector of factor loadings of order 3×1; $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$ represented a vector of measurement errors of order 3×1 for vectors $\mathbf{y}$ and $\mathbf{x}$ respectively; $\boldsymbol{\Gamma}$ represented a vector of parameters required to be estimated of order 1×3.

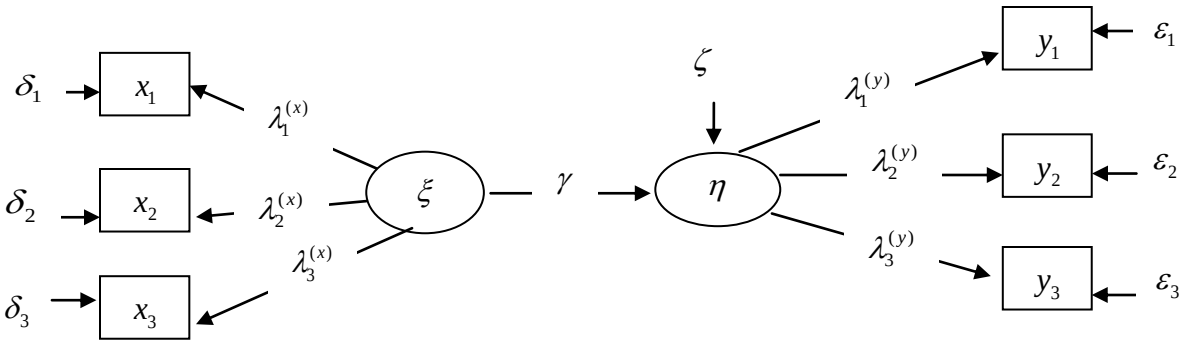


Figure 1: Conceptualized path-diagram for model 1 represents all variables

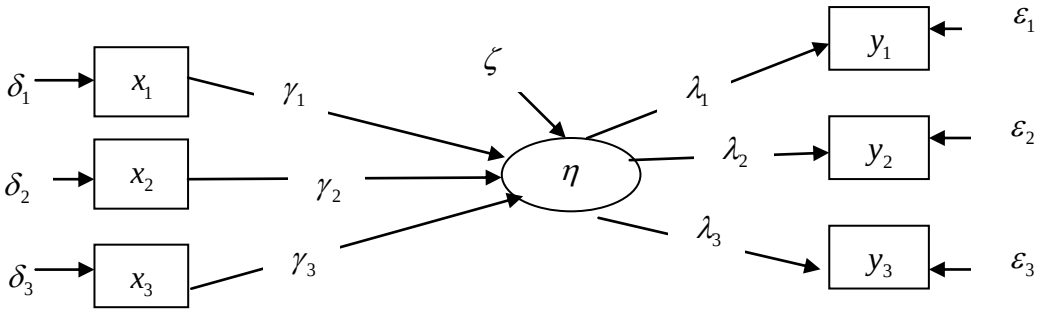


Figure 2: Conceptualized path-diagram for model 2 represents all variables

3. RESULTS

Every application of SEM should provide at least the following information: a clear and complete specification of models and variables, including a clear listing of the indicators of each latent; a clear statement of the type of data analyzed, with presentation of the sample correlation or covariance matrix; specification of the software and method of estimation; and complete results (Raykov et al. 1991). Table 1 showed Pearson correlation matrix, mean, and standard deviation for each indicator. As shown in Table 2, we provided several indexes of goodness of fit, allowing for a detailed evaluation of the adequacy of the fitted models. The simplest gauge of how well the model fits the data would be to inspect the residual matrix (Field 2000). The acceptable range of residual values was one in 20 standardized residuals exceeding ± 2.58 strictly by chance (Hair et al. 1998). Both models had not resulted in standardized residuals exceed the threshold value, and most of them were found close to zero, indicating high correspondence between elements of the implied covariances matrix of vector, $\mathbf{z} = (\mathbf{y}, \mathbf{x})$, denoted as $\boldsymbol{\Sigma}$ and the sample covariance matrix, $\mathbf{S}$. For assessing the fitted model, a model was considered adequate if the p-value was greater than 0.05, as 0.05 significance level was recommended as the minimum acceptance level for the proposed model (Hair et al. 1998). From Table 2, it was found that p-value for the fitted models was greater than 0.05, indicating that the

proposed models were acceptable or adequate in interpreting the relationship between education and occupation.

Bollen's incremental fit-index values were examined as these are least biased due to non-normality of variables and they were found most of them close to 0.95. Figures 3 and 4 explained the estimated parameters of fitted models 1 and 2 respectively. Model 1 and model 2 provided an excellent fit to the observed data as shown in Table 2, where for model 1 with ($\chi^2(8)$ = 6.99, *p*-value = 0.54) and for model 2 ($\chi^2(6) = 6.23$, *p*-value $= 0.40$). The estimated effect of education factor (labeled in Figure 3 as educ_ach) on occupation factor (labeled in Figure 3 as occupati) was found not significant with ($\hat{\gamma} = -0.12, t = -1.14$) based on fitted model 1. The estimated effects of education indicators on occupation factor were all found not significant with ($\hat{\gamma}_1 = -0.02, t = -0.51$; $\hat{\gamma}_2 = 0.00, t = -0.08$; and $\hat{\gamma}_3 = 0.01, t = 0.09$) respectively. Model 1 and model 2 were considered non-nested models. Non-nested models differ in number of latent factors or indicators. We can use AIC measure to compare between non-nested models. Given two non-nested models, the one with the lowest AIC was preferred (Kline 1998). However, model 1 was slightly better than model 2 because its' AIC was found somewhat less than AIC of model 2 as shown in Table 2.

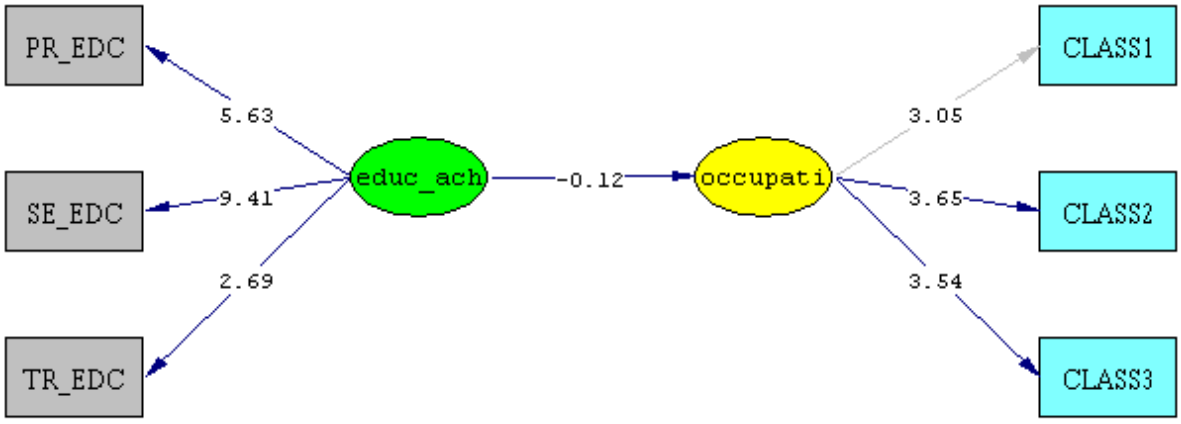


Figure 3: Path diagram shows the results of fitted model 1

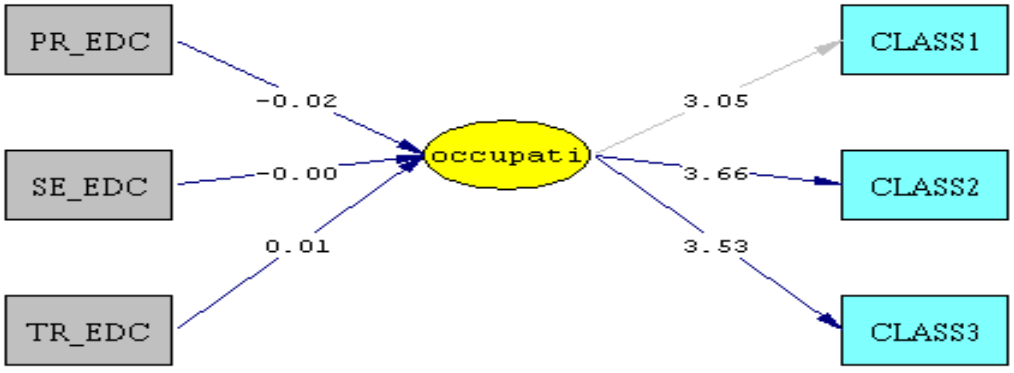


Figure 4: Path diagram shows the results of fitted model 2

Table 1: Pearson correlation matrix, Mean, and Standard Deviation (SD) for each variable

Variables	y_1	y_2	y_3	x_1	x_2	x_3	**Mean**	**SD**
CLASS1, y_1	1.00						10.07	3.30
CLASS2, y_2	0.88**	1.00					6.82	3.84
CLASS3, y_3	0.66**	0.68**	1.00				18.36	4.98
PR_EDC, x_1	-0.15	-0.16	-0.08	1.00			68.54	6.50
SE_EDC, x_2	-0.14	-0.14	-0.14	0.91**	1.00		45.80	8.94
TR_EDC, x_3	-0.12	-0.10	-0.11	0.71**	0.86**	1.00	6.17	3.28

**Correlation is significant at the 0.01 level (2-tailed)

Table 2: Comparison between the proposed models using fit indexes

Fit-indexes	**Model 1**	**Model 2**
Absolute-Fit measures		
χ^2-statistic(p-value) (d.f.)	6.99(0.54) (8)	6.23(0.40) (6)
GFI	0.97	0.98
SRMR	0.03	0.01
RMSEA	0.000	0.001
Incremental-Fit measures		
CFI	1.00	1.00
AGFI	0.93	0.91
NFI	0.98	0.98
NNFI	1.00	1.00
Parsimonious-Fit measures		
PGFI	0.37	0.28
AIC	32.81	36.04

χ^2-statistic = Likelihood-Ratio Chi-Square Statistic, GFI = Goodness-of-Fit Index, SRMR = Standardized Root Mean Square Residual, RMSEA = Root Mean Square Error of Approximation, CFI = Comparative fit index, AGFI = Adjusted Goodness-of-Fit Index, NFI = Normed Fit Index, NNFI = Non-Normed Fit Index (An old name for the NNFI is the Tucker-Lewis Index TLI), PGFI = Parsimonious Goodness-of-Fit Index, AIC = Akaike Information Criterion.

4. DISCUSSION

The role of this study was to review what was known about the role of education in improving the occupation opportunities with high level in both salary and social position. This subject was studied using several techniques and in this study structural equation modeling was used because we had several indicators for such latent factor. Bollen et al. (2001) argued that the latent factor approach had two advantages. First, this approach permits the integration of a range of measures or indicators of socioeconomic status (SES), thus avoiding the problems with choosing a single indicator. Secondly, this method allows greater control for measurement error. Ross and Wu (1995) concluded that high educational attainment proves health directly, and it improves health indirectly through work and economic conditions. In Pakistan for example, most studies analyzed the determinants of enrollment in school had found the association between household

income and girl's enrollment in school to be positive and statistically significant (Hazarika 2001; World Bank 2002). But the question in this paper is: what is the effect of education achievement on prosperity of the community represented by the occupation factor? Improvements to the quality and efficiency of basic education are urgently needed, in both developing and transition countries such as Malaysia. Therefore, policies are required to focus on (i) improving the efficiency of educational spending, so that the development of core skills does not require more years, and (ii) adapting the curriculum of basic as well as post basic education to develop the skills increasingly in demand in the global labor market: critical thinking, problem solving, and behavioral (that is, noncognitive) skills, as well as skills in information technology.

If improving the quality and quantity of skills was part of any educational package, this doesn't mean the package should success unless the issue of job creation was addressed. The supply of adequate jobs for the labor market is important for any policy maker. However, it is not simply whether an adequate number of jobs exist, but whether these jobs are of adequate quality. For example, subsidies in tertiary education need to be accompanied by the creation of an environment conducive to investment and technological progress. In the absence of such an environment, countries will find their population emigration for better opportunities and governments will need to continue subsidizing education to compensate for weak effective demand. Different countries at different levels of economic development had diverse requirements for education (Fasih 2008). For example, a study by De-Ferranti et al. (2003) suggested that whereas East Asian countries might benefit from more secondary school graduates to fill their skill needs gap, Latin American countries, because of their wealth of natural resources, would benefit from more experts in manufacturing processes and more tertiary education graduates.

It is essential to invest in quality early childhood education because the suggestion was: if the investment was made in developing the cognitive skills of children, the better the long-term impacts were for learning, skills development, and labor market outcomes. In a perfectly competitive labor market, skills such as motivation and ability may have higher value, thus people with higher ability may reap higher returns. From an education policy maker's point of view, this finding supports the importance of noncognitive skill development in schools and the education system as a whole. Also, the country context needs to be considered before recommending policy changes because decreasing returns from getting education could be the result of wage distortions caused by labor market rigidities. Expansion of higher education with no relation to job openings, and the resulting graduate unemployment, is the main cause of the brain drain which affects many of the developing countries, constituting a serious waste of resources.

Rwomire (1992) stresses the fact that the development of education has simply given rise to the replacement of a poorly-educated work force by one with a higher level of education. The number of jobs not increased as quickly as the number of graduates, and therefore the higher level of instruction had been of no benefit to the economy. Von Borstel (1992) examined the conditions for the success of a form of education that included productive work, where productive work was subordinate to school curricula and responds to the aims of education. However, most probably there was a lack in productive work in the school curricula in most of districts' schools in Malaysia in 1995. Also, we encouraged to offer job opportunities for young people, which enabled them to avoid leaving school early and this means that those people will face difficulties to get better jobs either in income level, social level or both because they left their school early. As Chung (1993) pointed out, in many developing countries, the majority of

the population cannot get regular jobs in the modern sector and a large percentage were condemned to remain in a state of long-term under-employment. General and vocational education thus seemed increasingly out of touch with reality.

5. CONCLUSION

With respect to model fit, researchers do not seem adequately sensitive to the fundamental reality that there is no true model, and all models are wrong to some degree, even in the population, and that the best one can hope for is to identify a parsimonious, substantively meaningful model that fits observed data adequately well (MacCallum & Austin 2000). Given this perspective, it is clear that a finding of good fit does not imply that a model is correct or true, but only plausible. We found models 1 and 2 acceptable or adequate fit in interpreting the hypothesized relationships. The education factor and its indicators in Malaysia in 1995 do not affect occupation factor based on both models. This was consistent with the study by Fasih (2008) who was stated that just increasing the quantity of education at the lower educational levels didn't raise earnings substantially, and thus not proved to be effective in helping people climb out of poverty. Education is a necessary but not sufficient condition for an individual to enjoy good occupation, where good occupation opportunities for the skilled require an economy as a whole to be operating well, with macroeconomic stability, an attractive investment climate, and efficient labor markets. The structures we had reported here as well as the strength of causal path-ways may vary depending on the specific nature and circumstances of the population under study. Further research is required in other developing countries.

REFERENCES

Anderson, E. J. (1980). Continuing education: A practical approach to career education, k-12. *The Journal of Adventist Education*, vol.42, No.2: 17-16.

Angrist, J., Bettinger E., & Kremer, M.. (2006). "Long-term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *American Economic Review*, vol.96 (3): 847–862.

Barrera-Osorio, F. (2007). "The Impact of Private Provision of Public Education: Empirical Evidence from Bogotá's Concession Schools." Policy Research Working Paper No. 4121, World Bank, Washington, DC.

Bentler, P. M., (1980). Multivariate Analysis with Latent Variables: Causal Modeling. Annual *Review Psychology*, vol.31: 419-456.

Bertrand, O. (1994). Education and work: International commission on education for twenty-first century: 1-23. UNESCO, Paris. http://www.unesco.org/education/pdf/16_53.pdf.

Blane, D., Brunner, E. & Wilkinson, R., (1996). *Health and Social Organization*. Routledge.

Bollen, K. A., (1989), Structural Equations with Latent Variables. John Wiley & Sons, USA.

Bollen, K. A., Glanville, J. L., & Stecklov, G. (2001). Socioeconomic status and class in studies of fertility and health in developing countries. Annual Review of Sociology, vol.27(1): 153-185.

Chandola, T., Clarke, P., Blane, D. & Morris, J.N. (2005). Pathways between education and health: a causal modeling approach: 1-44. www.ucl.ac.uk/epidemiology/ chandola/working%20for%20website.pdf

Chung, F. (1993). Education, Work and Employment (paper prepared for the International Commission on Education for the Twenty-first Century).

De-Ferranti, D.,William F. M., Guillermo E. P., Indermit G. J., Luis, G., Carolina S. P., & Norbert, S. (2003). *Closing the Gap in Education and Technology*. Latin American and Caribbean Studies. Washington, DC: World Bank.

Duper, M. E. (2008). Educational differences in health risks and illness over the life course: A test of cumulative disadvantage theory. *Social science Research*, vol.37(4): 1253-1266.

Fasih, T. (2008). Linking education policy to labor market outcomes. The World Bank. www.worldbank.org.

Field, A., (2000), "Structural Equation Modelling (SEM)," pp.1-9. www.sussex.ac.uk/users/andyf/teaching/pg/sem.pdf.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. G. (1998). Multivariate data analysis, Fifth Edition. Prentice Hall International, New Jersey.

Hazarika, G. (2001). The sensitivity of primary school enrollment to the costs of post-primary schooling in rural Pakistan: A gender perspective. *Education Economics*, vol.9(3): 237-244.

Hu, L., & Bentler, P. M. (1999). "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives," Structural Equation Modeling, 6(1), pp.1-55.

Joreskog, K. & Sorbom, D. (1982). Recent Developments in Structural Equation Modeling. *Journal of Marketing Research*, vol.19(4): 404-416.

Jöreskog, K. and Sörbom, D., (2001). LISREL8: User's Reference Guide. Second Edition. Lincolnwood: SSI.

Kline, R. B., (1998), Principles and Practice of Structural Equation Modeling, NY: Guilford Press.

Kravdal O. (2008). A broader perspective on education and mortality: Are we influenced by other people's education? Social science and medicine, vol.66: 620-636.

Levin, H. M. & Rumberger, R. W. (1989). Education, work and employment: present issues and future challenges in developed countries. In Caillods, The Prospects for Educational Planning, Paris, UNESCO / I I E P.

MacCallum, R. C. & Austin, J. T. (2000). Applications of Structural Equations Modeling in Psychological Research. Annual Review of Psychological Research. *Annual Review of Psychology* 51: 201-226.

Malaysia. Department of statistics. (1995). *Population report for administrative districts.* Kuala Lumpur: Department of statistics.

Nordstrom, M. L., Cnattingius, S. & Haglund, B., (1993). Social Differences in Swedish Infant Mortality by Cause of Death, 1983 to 1986. *American Journal of Public Health*, vol.83, No.1: 26-30.

Raykov, T., Tomer, A. & Nesselroade, J. R., (1991). Reporting Structural Equation Modeling Results in Psychology and Aging: Some Proposed Guidelines. *Psychology and Aging*, vol.6(4): 499-503.

Ross, C. E. & Wu, C., (1995). The links between education and health. *American Sociological Review*, vol.60: 719-745.

Rwomire, A. (1992). Education and development: African perspectives. *Prospects*, vol.22, no.2: 62-82.

Smith, D. & Langfield-Smith, K. (2004). Structural Equation Modeling in Management Accounting Research: Critical Analysis and Opportunities. *Journal of Accounting Literature*, 23: 49-86**.**

Von Borstel, A. (1991). A theoretical framework for productive education. *Prospects*, vol.21, no.3: 79-89.

Wanger, A. (2006). Measuring up internationally: Developing skills and knowledge for the global knowledge economy. National center for public policy and higher education, 1-31. www.highereducation.org/reports/muint/index.shtml.

World Bank. (2002). Pakistan poverty assessment; poverty in Pakistan: Vulnerabilities, social gaps and rural dynamics. Report No. 24296-PAK. Washington DC.

World Bank. (2007). *World Development Report 2006: Development and the Next Generation.* Washington, DC: World Bank.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 817–822.

A STUDY ON THE PERFORMANCES OF MEWMA AND MCUSUM CHARTS FOR SKEWED DISTRIBUTIONS

Michael B. C. Khoo[1], Abdu. M. A. Atta[2] and H. N. Phua
School of Mathematical Sciences, Universiti Sains Malaysia, Penang, Malaysia
E-mail: [1]mkbc@usm.my, [2]abduatta@yahoo.com

ABSTRACT

A multivariate chart, instead of separate univariate charts is used for a joint monitoring of several correlated variables. Two time weighted multivariate charts that are commonly used for a quick detection of small shifts in the mean vector are the multivariate exponentially weighted moving average (MEWMA) and multivariate cumulative sum (MCUSUM) charts. The MEWMA and MCUSUM charts use information from past data, which make them sensitive to small shifts. These charts require the assumption that the underlying process follows a multivariate normal distribution. This paper studies the robustness of the MEWMA and MCUSUM charts toward nonnormality by considering the multivariate Weibull and multivariate gamma distributions based on different sample sizes and correlation coefficients.

1. INTRODUCTION

In most process monitoring situations, the quality of a process is determined by two or more quality characteristics (Woodall and Montgomery, 1999). Process monitoring problems involving several related variables of interest are called multivariate statistical process control. The most useful tool used in the monitoring of a multivariate process is a multivariate control chart. The first step in constructing a multivariate chart involves the analysis of a preliminary set of data that is assumed to be in statistical control. This analysis is known as a Phase-I analysis and it is conducted to estimate process parameters that will be used for the monitoring of a future process, a.k.a., a Phase-II process.

Numerous multivariate charts and their extensions are presently available. These charts can be grouped into 3 broad categories, namely, the Hotelling's T^2, multivariate EWMA (MEWMA) and multivariate CUSUM (MCUSUM) charts. The Hotelling's T^2 chart was proposed by Hotelling (1947) for the detection of a large sustained shift. The MCUSUM chart was first suggested by Woodall and Ncube (1985) while the MEWMA chart was introduced by Lowry et al. (1992). However, the MCUSUM charts suggested by Crosier (1988) will be discussed in this paper as they are more widely used.

This paper is organized as follows: Section 2 reviews the MEWMA chart while Section 3 reviews the MCUSUM chart. In Section 4, a simulation study is conducted to compare the performances of MEWMA and MCUSUM charts for skewed distributions. Finally, conclusions are drawn in Section 5.

2. MEWMA CONTROL CHART

The MEWMA chart proposed by Lowry et al. (1992) is based on the following statistic:

$$\boldsymbol{Z}_t = \lambda \boldsymbol{X}_t + (1-\lambda)\boldsymbol{Z}_{t-1}, \quad \text{for } t = 1, 2, \ldots, \tag{1}$$

where $\boldsymbol{Z}_0 = \boldsymbol{\mu}_0$ and $0 < \lambda \leq 1$. $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$, are assumed to be independent multivariate normal random vectors, each with *p* quality characteristics. The control charting statistic of a MEWMA chart is (Lowry et al., 1992)

$$T_t^2 = \boldsymbol{Z}_t' \boldsymbol{\Sigma}_{\boldsymbol{Z}_t}^{-1} \boldsymbol{Z}_t . \tag{2}$$

The chart signals a shift in the mean vector when $T_t^2 > h_1$, where h_1 is the limit chosen to achieve a desired in-control ARL (ARL_0) and

$$\boldsymbol{\Sigma}_{\boldsymbol{Z}_t} = \frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2t}\right]\boldsymbol{\Sigma}_{\boldsymbol{X}} \tag{3}$$

is the variance-covariance matrix for $\boldsymbol{Z}_t$. Lowry et al. (1992) showed that the run length performance of the MEWMA chart depends on the off-target mean vector $\boldsymbol{\mu}_1$ and the covariance matrix of $\boldsymbol{X}_t$, i.e., $\boldsymbol{\Sigma}_{\boldsymbol{X}_t}$ only through the value of the non-centrality parameter,

$$\delta = \left\{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right\}^{1/2}, \tag{4}$$

where $\boldsymbol{\mu}_0$ denotes the in-control mean vector.

Lee and Khoo (2006a) provide a method based on the Markov chain approach for the selection of the optimal parameters, λ and h_1, which produce the minimum out-of-control ARL (ARL_1) for a desired size of a shift of interest based on a fixed ARL_0.

3. MCUSUM CONTROL CHART

Crosier (1988) suggested two multivariate CUSUM charts. The one with the better ARL performance is based on the following statistics:

$$C_t = \left\{(\boldsymbol{S}_{t-1} + \boldsymbol{X}_t - \boldsymbol{a})' \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1} (\boldsymbol{S}_{t-1} + \boldsymbol{X}_t - \boldsymbol{a})\right\}^{1/2}, \text{ for } t = 1, 2, \ldots, \tag{5}$$

where

$$\boldsymbol{S}_t = \begin{cases} \boldsymbol{0}, & \text{if } C_t \leq k \\ (\boldsymbol{S}_{t-1} + \boldsymbol{X}_t - \boldsymbol{a})\left(1 - \dfrac{k}{C_t}\right), & \text{if } C_t > k \end{cases}. \tag{6}$$

Note that $\boldsymbol{S}_0 = \boldsymbol{0}$, $k > 0$ is the reference value and $\boldsymbol{a}$ is the aim point or target value for the mean vector. The control charting statistic for the MCUSUM chart is (Crosier, 1988)

$$Y_t = \left(\boldsymbol{S}_t' \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1} \boldsymbol{S}_t\right)^{1/2} \tag{7}$$

A shift in the mean vector is signalled when $Y_t > h_2$, where h_2 represents the limit of the chart. The MCUSUM procedure assumes that the multivariate observations $\boldsymbol{X}_t$, for $t = 1, 2, \ldots$, follow an independently and identically distributed (i.i.d.) multivariate normal distribution. Lee and Khoo (2006b) give an approach based on the Markov chain method in determining the optimal parameters, *k* and h_2 that give the minimum out-of-control ARL (ARL_1) for a size of shift of interest based on a fixed ARL_0.

4. A SIMULATION STUDY

The assumption of the underlying process having i.i.d. multivariate normal random variates is required for both the MEWMA and MCUSUM charts. Since many multivariate processes, such as chemical processes come from populations that are skewed, it is difficult to satisfy the

multivariate normality assumption. In this section, the performances of the MEWMA and MCUSUM charts will be studied when the multivariate normality assumption is violated.

The performances of the MEWMA and MCUSUM charts are compared based on the false alarm rates when the process is in-control for multivariate skewed distributions, such as the Lee's multivariate Weibull (Lee, 1979) and Cheriyan and Ramabhadran's multivariate gamma distributions (Cheriyan, 1941 and Ramabhadran, 1951). For the sake of comparison, the multivariate normal distribution is also considered. For convenience, the bivariate case, i.e., the number of quality characteristics, $p = 2$ is considered. Note that the bivariate Weibull distribution can represent various skewnesses and correlations but the bivariate gamma can only represent some positive correlations (Kotz et al., 2000).

SAS programs are used to compute the false alarm rates for the three multivariate distributions considered. Each false alarm rate is computed based on 5000 simulation trials. The nominal false alarm rate is assumed to be $\alpha = 0.0027$ when the underlying distribution is bivariate normal. The MEWMA and MCUSUM charts are designed for a quick detection of a shift in the mean vector of size $\delta = 1$. The optimal smoothing constant, $\lambda = 0.13$ and limit $h_1 = 10.55$ are found for the MEWMA chart using the approach described in Lee and Khoo (2006a). Similarly, using the procedure given in Lee and Khoo (2006b), the optimal parameters are found to be $k = 0.5$ and $h_2 = 6.227$ for the MCUSUM chart.

The correlation coefficients, $\rho = 0.3$, 0.5 and 0.8 are considered for the bivariate distributions. For ease of computation, the sacle parameters of (1,1) for (X_1, X_2) are selected for the Weibull and gamma distributions. The shape parameters for (X_1, X_2) are chosen so that the desired skewnesses (γ_1, γ_2) = {(1,1), (1,2), (1,3), (2,2), (2, 3), (3,3)} for these parameters are attained. The sample sizes, $n = 3$, 5 and 7 are considered.

The false alarm rates for the MEWMA and MCUSUM charts are given in Tables 1 and 2, respectively. Note that the false alarm rates, marked as "*" in Tables 1 and 2 for the Cheriyan and Ramabhadran's bivariate gamma distribution cannot be computed because the corresponding shape parameters of one of the gamma distributed components, used in the transformation to compute variate X_2 have negative values. From Tables 1 and 2, it is found that for the multivariate Weibull and gamma distributions, the false alarm rates of the MEWMA and MCUSUM charts increase as the level of skewness and correlation coefficient increase. This is because the covariance matrix of the multivariate observation, $\boldsymbol{X}$ is inflated as the skewness and correlation coefficient increase, hence making it easier for the MEWMA and MCUSUM charts to issue out-of-control signals. Also note that the false alarm rate decreases as the sample size increases. This is consistent with the multivariate central limit theorem, where the sample mean vector of a multivariate skewed distribution approaches multivariate normality as the sample size increases. A comparison of the false alarm rates of the two charts show that generally the MEWMA chart has lower false alarm rates than the MCUSUM chart for various levels of skewnesses. Thus, the MEWMA chart is more robust than the MCUSUM chart.

5. CONCLUSIONS

In this paper, we have studied the performance of the MEWMA and MCUSUM charts for multivariate normal and multivariate skewed distributions. We found that the false alarms of both charts are affected by the skewness of the underlying distribution. Also, the sample size and correlation of the quality characteristics have an impact on the false alarm rates of the charts. The

simulation results show that the MEWMA chart has a lower false alarm rate than the MCUSUM chart when the underlying distribution is skewed. Since it is known that both the MEWMA and MCUSUM charts have equal performances in the detection of small shifts when the underlying process is multivariate normally distributed, the use of the MEWMA chart in process monitoring is recommended because the MEWMA chart is more robust towards skewed populations.

APPENDIX

Table 1. False alarm rates for the MEWMA chart when $\lambda = 0.13$ and $h_1 = 10.55$

Correlation coefficient	Multivariate distribution	Skewness coef.t (γ_1,γ_2)	Sample size, *n*		
			3	5	7
$\rho = 0.3$	Normal	(0,0)	0.0025830	0.0026450	0.0026600
	Weibull	(1,1)	0.0029040	0.0027250	0.0026130
		(1,2)	0.0035350	0.0031600	0.0030100
		(1,3)	0.0043070	0.0037940	0.0035010
		(2,2)	0.0041670	0.0035550	0.0033540
		(2,3)	0.0049280	0.0041320	0.0038270
		(3,3)	0.0056780	0.0047290	0.0043020
	Gamma	(1,1)	0.0029980	0.0028110	0.0027850
		(1,2)	0.0034710	0.0031680	0.0030860
		(1,3)	0.0039560	0.0035100	0.0033090
		(2,2)	0.0040630	0.0036180	0.0032550
		(2,3)	0.0047300	0.0040400	0.0036070
		(3,3)	0.0053240	0.0043790	0.0039400
$\rho = 0.5$	Normal	(0,0)	0.0025830	0.0026450	0.0026600
	Weibull	(1,1)	0.0028690	0.0026110	0.0024610
		(1,2)	0.0037170	0.0032160	0.0030640
		(1,3)	0.0047470	0.0040950	0.0038720
		(2,2)	0.0044490	0.0037830	0.0035030
		(2,3)	0.0053900	0.0045610	0.0042110
		(3,3)	0.0062970	0.0052450	0.0048160
	Gamma	(1,1)	0.0030340	0.0029600	0.0027750
		(1,2)	0.0036180	0.0031160	0.0030910
		(1,3)	*	*	*
		(2,2)	0.0041920	0.0035860	0.0033770
		(2,3)	0.0047770	0.0039130	0.0036190
		(3,3)	0.0055110	0.0046630	0.0042070
$\rho = 0.8$	Normal	(0,0)	0.0025830	0.0026450	0.0026600
	Weibull	(1,1)	0.0033460	0.0027720	0.0025900
		(1,2)	0.0049340	0.0041010	0.0036990
		(1,3)	0.0070200	0.0060950	0.0057130
		(2,2)	0.0057440	0.0046370	0.0041530
		(2,3)	0.0071680	0.0059480	0.0054110
		(3,3)	0.0081140	0.0068870	0.0062570
	Gamma	(1,1)	0.0032220	0.0030090	0.0029240
		(1,2)	*	*	*
		(1,3)	*	*	*
		(2,2)	0.0048070	0.0040280	0.0036410
		(2,3)	*	*	*
		(3,3)	0.0064080	0.0053320	0.0047210

Table 2. False alarm rates for the MCUSUM chart when $k = 0.5$ and $h_2 = 6.227$

Correlation coefficient	Multivariate distribution	Skewness Coef. (γ_1, γ_2)	Sample size, n		
			3	5	7
ρ = 0.3	Normal	(0,0)	0.0026930	0.0027220	0.0027220
	Weibull	(1,1)	0.0029440	0.0027420	0.0026800
		(1,2)	0.0035020	0.0031400	0.0030080
		(1,3)	0.0042500	0.0037290	0.0034920
		(2,2)	0.0036650	0.0033290	0.0031400
		(2,3)	0.0048150	0.0040180	0.0037430
		(3,3)	0.0055340	0.0045740	0.0041640
	Gamma	(1,1)	0.0026930	0.0027220	0.0027220
		(1,2)	0.0029440	0.0027420	0.0026800
		(1,3)	0.0035020	0.0031400	0.0030080
		(2,2)	0.0042500	0.0037290	0.0034920
		(2,3)	0.0036650	0.0033290	0.0031400
		(3,3)	0.0048150	0.0040180	0.0037430
ρ = 0.5	Normal	(0,0)	0.0026930	0.0027220	0.0027220
	Weibull	(1,1)	0.0028180	0.0025440	0.0024610
		(1,2)	0.0036490	0.0031830	0.0030670
		(1,3)	0.0046620	0.0040150	0.0038180
		(2,2)	0.0043290	0.0036840	0.0034060
		(2,3)	0.0052570	0.0044200	0.0040660
		(3,3)	0.0061670	0.0050810	0.0046200
	Gamma	(1,1)	0.0031260	0.0028970	0.0028610
		(1,2)	0.0034620	0.0032160	0.0029970
		(1,3)	*	*	*
		(2,2)	0.0039720	0.0035540	0.0033110
		(2,3)	0.0047190	0.0039000	0.0035950
		(3,3)	0.0055250	0.0045230	0.0040310
ρ = 0.8	Normal	(0,0)	0.0026930	0.0027220	0.0027220
	Weibull	(1,1)	0.0030700	0.0026280	0.0024780
		(1,2)	0.0047220	0.0039330	0.0036050
		(1,3)	0.0068540	0.0059950	0.0056410
		(2,2)	0.0054640	0.0044190	0.0039710
		(2,3)	0.0068990	0.0056910	0.0051910
		(3,3)	0.0078760	0.0065890	0.0059620
	Gamma	(1,1)	0.0032610	0.0029770	0.0028680
		(1,2)	*	*	*
		(1,3)	*	*	*
		(2,2)	0.0045670	0.0038680	0.0035700
		(2,3)	*	*	*
		(3,3)	0.0062740	0.0052470	0.0044940

ACKNOWLEDGEMENTS

This research is supported by the Universiti Sains Malaysia, Research University (RU) grant, no. 1001/PMATHS/811024.

REFERENCES

Cheriyan, K.C. (1941). A bivariate correlated gamma-type distribution function. *Journal of the Indian Mathematical Society*, 5, 133-144.

Crosier, R.B. (1988). Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics*, 30, 291–303.

Hotelling, H. (1947). "Multivariate quality control," *Techniques of Statistical Analysis*, Eisenhart, Hastay and Wallis (eds.), McGraw-Hill, New York.

Kotz, S., Balakrishnan, N. and Johnson, N.L. (2000). *Continuous multivariate distributions*, Vol. 1., 2nd ed., John Wiley, New York.

Lee, L. (1979). Multivariate distributions having Weibull properties. *Journal of Multivariate Analysis*, 9, 267–277.

Lee, M.H. and Khoo, M.B.C. (2006a). Optimal statistical design of a multivariate EWMA chart based on ARL and MRL. *Communications in Statistics – Simulation and Computation*, 35, 831–847.

Lee, M.H. and Khoo, M.B.C. (2006b). Optimal statistical design of a multivariate CUSUM chart. *International Journal of Reliability, Quality and Safety Engineering*, 13, 479–497.

Lowry, C.A., Woodall, W.H., Champ, C.W. and Rigdon, S.E. (1992). A multivariate exponentially weighted moving average control chart. *Technometrics*, 34, 46–53.

Ramabhadran, V.R. (1951). A multivariate gamma-type distributions. *Sankhya*, 11, 45–46.

Woodall, W.H. and Montgomery, D.C. (1999). Research issues and ideas in statistical process control. *Journal of Quality Technology*, 31, 376-386.

Woodall, W.H. and Ncube, M.M. (1985). Multivariate CUSUM quality control procedures. *Technometrics*, 27, 285–292.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 823–830.

BAYESIAN MULTIPLE CHANGE-POINT ESTIMATION USING SAMC

Jaehee Kim [1] and Sooyoung Cheon [2]
Department of Statistics, Duksung Women's University, Seoul 132-714, South Korea KU
Industry-Academy Cooperation Group Team of Economics and Statistics, Korea University,
Jochiwon 339-700, South Korea.
E-mail: [1] jaehee@duksung.ac.kr, [2] s7cheon@gmail.com

ABSTRACT

Bayesian inference for multiple change-point problems is studied. We use a truncated Poisson distribution for the number of change-points and conjugate prior for the exponential family distributions. SAMC is adopted in order to overcome the analytic difficulties in computing the posterior distributions. We demonstrate how the proposed method can be made for real data.

1. INTRODUCTION

In many applications of statistics, including areas as diverse as quality control and tracking an object following a ballistics trajectory, we are interested in detecting changes in the parameters of the distribution of a sequence of independent observations. Finding the number of change-points and their positions is one of the challenging statistical problems in which the dimension of the object of inference is not fixed.

Chernoff and Zacks (1964) considered a Bayes test for mean change for the normal observations. Kander and Zacks (1966) generalized the result of Chernoff and Zacks (1964) to the one-parameter exponential family. Hinkley (1970) investigated the maximum likelihood estimates of one change-point problem. As a Bayesian approach for the change-point problem, Smith (1975) considered one change-point problem in distributional changes using Gibbs sampler. Yao (1984) derived Bayes estimates in the presence of additive Gaussian noise and a signal which is a step function. Carlin *et al.* (1992) formulated the hierarchical Bayesian Markov chain model and used Gibbs sampler. Belisle *et al.* (1998) made inference about Bayesian hierarchical change-point model with the ensemble of sample paths for neuron spike train data. In the multiple change-point setting, Venter and Steel (1996) identified multiple abrupt change-points in a sequence of observations via hypothesis testing.

Hawkins (2001) developed an approach with maximum likelihood estimates of the change-points and within-segment parameters in the exponential family. For the Bayesian multiple change-point problem, Barry and Hartigan (1993) used a product partition model. Chib (1998) formulated the multiple change-point model in terms of a latent discrete state variable according to Markov process with transition probabilities. Stephens (1994) discussed the use of a sampling-based technique, the Gibbs sampler, including the binomial data model. Fearnhead (2006) suggested the recursion algorithm to search the change-points successively. Bayesian methods are attractive for change-point models since they allow for flexible relationships between parameters in various subspaces and are computationally advantageous.

In many of the Bayesian approaches, Markov sampling techniques have been used for the calculation of posterior probabilities. Due to the many possible partitions, the model space

becomes complex with multiple modes, and the traditional Monte Carlo methods are prone to get trapped in local energy minima. Tierney (1994) developed a hybrid sampler in order to traverse freely across the combined parameter spaces. Green (1995) proposed reversible Markov chain samplers that jump between parameter subspaces of differing dimensionalities which are applicable for multiple change-point problems. Liang *et al.* (2007) proposed the stochastic approximation Monte Carlo (SAMC) algorithm effective for importance sampling and model selection. In this paper, we briefly review Bayesian multiple change-point inference and describe implementation of the computational technique, SAMC, that can be used to facilitate Bayesian technique in the complex problem. We give illustrations of the Bayesian solution to multiple change-point problems via several examples. In section 2 a general multiple change-point model is defined and the Bayesian inference is provided for exponential family distributions. Section 3 describes briefly the SAMC algorithm applied to the multiple change-point problem. Section 4 presents a numerical result with real data for multiple change-point estimation. Finally, section 5 concludes the paper with a discussion.

2. THE BAYESIAN MULTIPLE CHANGE-POINT MODEL

Change-point identification is important in data analysis. Interest lies in making inference about the time or position in the sequence that the change occurred. This problem can be generalized to incorporate notions of multiple changes in the system, and arises when different subsequences of a data series follow different statistical distributions of the same functional form but have different parameters. Let $\mathbf{Z} = (z_1, z_2, \cdots, z_n)$ denote the independent observation sequence ordered in time. There exists a partition on the set $\{1,2,\cdots,n\}$ into blocks so that the sequence follows the same distribution within blocks. That is, the change-points divide the partitions. Let $\boldsymbol{x} = (x_1, x_2, \cdots, x_{n-1})$ be a binary vector with $x_{c_1} = x_{c_2} = \cdots = x_{c_k} = 1$ and being 0 elsewhere, $0 = c_0 < c_1 < \cdots < c_k < c_{k+1} = n$.

There are k change-points in the model and k is unknown. The multiple change-point model can be written as follows:

$$z_i \sim f_r(\cdot \mid \boldsymbol{\varphi}_r), \qquad c_{r-1} < i < c_r \tag{1}$$

for $r = 1,2,\cdots,k+1$ and f_r depends on the parameters $\boldsymbol{\varphi}_r \in \boldsymbol{\Phi}$. The parameters change at $c_1+1,\cdots,c_k+1$. Each $c_1,\cdots,c_k$ is called the change-point. Consider that f_r is a density parameterized by $\boldsymbol{\varphi}_r$. Let $\mathbf{x}^{(k)}$ denote a configuration of $\mathbf{x}$ with k change-points. Let $\eta^{(k)} = (\mathbf{x}^{(k)}, \boldsymbol{\varphi}_1, \cdots, \boldsymbol{\varphi}_{k+1})$ and A_k be the space of models with k change-points, $\mathbf{x}^{(k)} \in A_k$ and $\chi = \cup_{k=0}^{n} A_k$. The likelihood function of $\mathbf{Z}$ is $L(\mathbf{Z} \mid \eta^{(k)}) = \prod_{j=c_0+1}^{c_1} f_1(z_j \mid \boldsymbol{\varphi}_1) \cdots \prod_{j=c_k+1}^{c_{k+1}} f_{k+1}(z_j \mid \boldsymbol{\varphi}_{k+1})$. We set the prior distribution for $\mathbf{x}^{(k)}$ in $\eta^{(k)}$ as the truncated Poisson distribution,

$$\pi(\mathbf{x}^{(k)}) = \frac{\lambda^k}{\sum_{j=0}^{n-1} \frac{\lambda^j}{j!}} \frac{(n-1-k)!}{(n-1)!}, \qquad k = 0,1,\cdots,n-1.$$

Kim and Cheon (2009) and Cheon and Kim (2009) provide the derivation of the full posterior for the normal, exponential, binomial and Poisson distributions for multiple change-points

identification in Table 2.1. We can sample from this non-normalized posterior $P(\mathbf{x}^{(k)} \mid \mathbf{Z})$ by the SAMC technique with the partitioned sample space according to the negative posterior log-likelihood function and estimate the change-points which have the greatest posterior probabilities.

The BIC is commonly used in Bayesian model selection, discussed in Kass and Raftery (1995). The model with the highest posterior probability is the one that minimizes:

BIC = -2(log maximized likelihood) + $(\log n)$ (number of parameters).

We used BIC since BIC penalizes more severely for the parameters and the posterior comparison is considered in change-point estimation. BIC tends to favor simpler models and gives a rough approximation to the logarithm of the Bayes factor, which is easy to use and does not require evaluation of the prior distributions (Raftery, 1995).

3. APPLICATION OF SAMC TO MULTİPLE CHANGE-POİNT ESTIMATION

The basic idea of SAMC (Liang *et al.*, 2007) can be explained briefly as follows. Let

$$f(\mathbf{x}) = c\,\psi(\mathbf{x}), \qquad \mathbf{x} \in \chi \tag{2}$$

denote the target probability density/mass function, where χ is the sample space and c is an unknown constant. Let $E_1, E_2, \cdots, E_m$ denote a partition of χ, and let $w_i = \int_{E_i} \psi(\mathbf{x})dx$ for i = 1,…,m. SAMC seeks to sample from the trial distribution

$$f_w(\mathbf{x}) = \sum_{i=1}^{m} \frac{g_i \psi(\mathbf{x})}{w_i} I(\mathbf{x} \in E_i) \tag{3}$$

where g_i's are pre-specified constants such that $g_i > 0$ for all i and $\sum_{i=1}^{m} g_i = 1$ and $\mathbf{g} = (g_1, g_2, \ldots, g_m)$ is called the desired sampling distribution of the subregions. Let δ_{ti} denote the working estimate of $\log(w_i / g_i)$ obtained at iteration t, let $\boldsymbol{\delta}_t = (\delta_{t1}, \delta_{t2}, \ldots, \delta_{tm})$, and let $\{\gamma_t\}$ denote a positive, non-increasing sequence satisfying the conditions

$$\text{(i)} \sum_{t=1}^{\infty} \gamma_t = \infty, \ \text{(ii)} \sum_{t=1}^{\infty} \gamma_t^{\zeta} < \infty \tag{4}$$

for some $\zeta \in (1,2)$. Since $f_w(\mathbf{x})$ is invariant with respect to a scale change of $\mathbf{w} = (w_1, w_2, \ldots, w_m)$, the domain of $\boldsymbol{\delta}_t$ can be kept in the compact set Ω in simulations by adjusting $\boldsymbol{\delta}_t$ with a constant vector. In this paper, we set $\Omega = [-10^{100}, 10^{100}]^m$, although this is practically equivalent to setting $\Omega = \Re^m$.

One iteration of the SAMC algorithm consists of the following steps:

(a) (Sampling) Simulate a sample x_t by a single MH update with the target distribution

$$f_{\boldsymbol{\delta}_t}(\mathbf{x}) \propto \sum_{i=1}^{m} \frac{\psi(\mathbf{x})}{e^{\boldsymbol{\delta}_{ti}}} I(\mathbf{x} \in E_i)$$

Table 2.1 Log posterior distributions for the exponential family distributions

Dist.	Prior	Log Posterior
Normal $N(\mu_i,\sigma_i^2)$ μ_i,σ_i^2: unknown	$\mu_i \sim$ uniform $\sigma_i^2 \sim IG(\gamma,\delta)$ inv-gamma	$a_k + \frac{k+1}{2}\log 2\pi - \sum_{i=1}^{k+1}\left\{\frac{1}{2}\log(c_i - c_{i-1}) - \log\Gamma\left(\frac{c_i - c_{i-1} - 1}{2} + \gamma\right) + \left(\frac{c_i - c_{i-1} - 1}{2} + \gamma\right)\log\left[\delta + \frac{1}{2}\sum_{j=c_{i-1}+1}^{c_i} z_j^2 - \frac{\left(\sum_{j=c_{i-1}+1}^{c_i} z_j\right)^2}{2(c_i - c_{i-1})}\right]\right\}$ where $a_k = (k+1)(\gamma\log\delta - \log\Gamma(\gamma)) + \log(n-1-k)! + k\log\lambda$
Exponential $Exp(\sigma_i)$	$\sigma_i \sim G(\gamma,\delta)$ Gamma	$a_k + \sum_{i=1}^{k+1}\left\{\log\Gamma(c_i - c_{i-1} + \gamma) - (c_i - c_{i-1} + \gamma)\log\left(\delta + \sum_{j=c_{i-1}+1}^{c_i} z_j\right)\right\}$ where $a_k = (k+1)(\gamma\log\delta - \log\Gamma(\gamma)) + \log(n-1-k)! + k\log\lambda$
Binomial $B(b,p_i)$	$p_i \sim Beta(\alpha,\beta)$	$a_k + \sum_{i=1}^{k+1}\left\{\sum_{j=c_{i-1}+1}^{c_i}\left[\log b! - \log(b-z_j)! - \log z_j!\right] + \log\Gamma(\alpha + \sum_{j=c_{i-1}+1}^{c_i} z_j) + \log\Gamma(n(c_i - c_{i-1}) + \beta - \sum_{j=c_{i-1}+1}^{c_i} z_j) - \log\Gamma(\alpha + b(c_i - c_{i-1}) + \beta)\right\}$ where $a_k = (k+1)\{\log\Gamma(\alpha+\beta) - \log\Gamma(\alpha) - \log\Gamma(\beta)\} + \log(n-1-k)! + k\log\lambda$
Poisson $Poi(\phi_i)$	$\phi_i \sim G(\gamma,\delta)$	$a_k + \sum_{i=1}^{k+1}\left\{\log\Gamma(\gamma + \sum_{j=c_{i-1}+1}^{c_i} z_j) - (\sum_{j=c_{i-1}+1}^{c_i} z_j + \gamma)\log(c_i - c_{i-1} + \delta) - \sum_{j=c_{i-1}+1}^{c_i}\log z_j!\right\}$ where $a_k = (k+1)(\gamma\log\delta - \log\Gamma(\gamma)) + \log(n-1-k)! + k\log\lambda$.
Multi-variate Normal $N_d(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$	$(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) \sim IW(\upsilon_0,\Lambda_0^{-1})$ inverse-Wishart	$c_k - \frac{d(k+1)}{2}\log(2\pi) + \sum_{i=1}^{k+1}\log\left[\frac{2^{\frac{\nu_i d}{2}}\pi^{\frac{d(d-1)}{4}}}{(c_i - c_{i-1})^{1/2}}\,|\boldsymbol{\Lambda}_0 + (c_i - c_{i-1})\mathbf{S}_i|^{-\frac{\nu_i}{2}}\prod_{u=1}^{d}\Gamma\left(\frac{\nu_i + 1 - u}{2}\right)\right]$ where $c_k = -(k+1)\left[\frac{\nu_0 d}{2}\log 2 + \frac{d(d-1)}{4}\log\pi + \sum_{u=1}^{d}\log\Gamma\left(\frac{\nu_0 + 1 - u}{2}\right) - \frac{\nu_0}{2}\log|\boldsymbol{\Lambda}_0|\right]$. $\nu_i = c_i - c_{i-1} + \nu_0 - 1$, $\mathbf{S}_i = \frac{1}{c_i - c_{i-1}}\sum_{j=c_{i-1}+1}^{c_i}(\mathbf{z}_i - \bar{\mathbf{z}}^i)(\mathbf{z}_i - \bar{\mathbf{z}}^i)'$ $\mathbf{z}_j = (z_{j1},\ldots,z_{jd})'$ for $j = c_{i-1}+1,\ldots,c_i$, $\mathbf{1}_{(c_i - c_{i-1})\times 1} = (1,\ldots,1)'$ $\mathbf{z}^i = \begin{pmatrix} z_{c_{i-1}+1,1} & \cdots & z_{c_{i-1}+1,d} \\ \vdots & \ddots & \vdots \\ z_{c_i,1} & \cdots & z_{c_i,d} \end{pmatrix}$, $\bar{\mathbf{z}}^i = \left(\frac{1}{c_i - c_{i-1}}\sum_{j=c_{i-1}+1}^{c_i} z_{j1},\cdots,\frac{1}{c_i - c_{i-1}}\sum_{j=c_{i-1}+1}^{c_i} z_{jd}\right)'$

(b) (Weight updating) Set $\boldsymbol{\delta}^* = \boldsymbol{\delta}_t + \boldsymbol{\gamma}_{t+1}(\tilde{\mathbf{e}}_t - \mathbf{g})$, where $\tilde{\mathbf{e}}_t = (\tilde{e}_{t,1},\cdots,\tilde{e}_{t,m})$ and $\tilde{e}_{t,i} = 1$ if $\mathbf{x}_t \in E_i$ and 0 otherwise. If $\boldsymbol{\delta}^* \in \Omega$, set $\boldsymbol{\delta}_{t+1} = \boldsymbol{\delta}^*$; otherwise, set $\boldsymbol{\delta}_{t+1} = \boldsymbol{\delta}^* + \mathbf{c}^*$ where $\mathbf{c}^* = (c^*,\cdots,c^*)$ can be an arbitrary vector which satisfies the condition $\boldsymbol{\delta}^* + \mathbf{c}^* \in \Omega$.

In the change-point detection, the sample space should be partitioned according to the model index: e.g., $E_1 = \{\mathbf{x} : k = 1\}, E_2 = \{\mathbf{x} : k = 2\},\cdots$. In this paper, without loss of generality, we consider only models with $k_{\min} \le k \le k_{\max}$, where k is the number of change-points, and $k_{\min}$ and $k_{\max}$ can be determined after a pilot study of the above algorithm, respectively. Outside this range, $P(\chi_i \mid \mathbf{Z}) \approx 0$.

We apply the SAMC algorithm to the Bayesian model selection problem. For change-point detection, the maximum a *posteriori* (MAP) estimate of $\mathbf{x}^{(k)}$ is often a reasonable solution to the problem. The sampling step of SAMC is as follows. Let $x_k^{(t,l)}$ denote the l^{th} sample generated at an iteration t, where k indicates the number of change-points in the sample. The next sample can be generated from the following procedure.

(a) Set $j = k-1, k,$ or $k+1$ according to probabilities $q_{k,j}$, where $q_{k,k} = 1/3$ for $k_{\min} \le k \le k_{\max}$, $q_{k_{\min},k_{\min}+1} = q_{k_{\max},k_{\max}-1} = 2/3$, and $q_{k,k+1} = q_{k,k-1} = 1/3$ if $k_{\min} < k < k_{\max}$.

(b) Update $x_k^{(t,l)}$ by a "death", "simultaneous" or "birth" move if $j = k-1, k$ or $k+1$, respectively.

The "death", "simultaneous", and "birth" moves are designed as described in Green (1995) and Liang (2007 (b)).

4. BAYESIAN CHANGE-POINT ANALYSIS WITH WELL-LOG DATA FOR NORMAL CHANGE-POINT MODEL

We consider the problem of detecting change-points in well-log data, which come from O Ruanaidh and Fitzgerald (1996). The data, obtained by lowering a probe into a bore-hole, consist of 4050 measurements of the nuclear-magnetic response of underground rocks. Measurements were taken at discrete time-points by the probe as it was lowered through the hole. The data are used to interpret the geophysical structure of the rock surrounding the well. The variations in mean reflect the stratification of the earth's crust. The change-points in the signal occur each time a new rock type is encountered. Detecting the change points is important in oil-drilling; see the introduction of Fearnhead and Clifford (2003) for more details. These data have been previously analyzed by O Ruanaidh and Fitzgerald (1996), who used MCMC to fit a change-point model with a fixed number of change points; and by Fearnhead and Clifford (2003) who considered online analysis of the data using particle filters. Since well-log data were assumed to be followed a univariate Gaussian model in Adams and MacKay (2007), the normal change model was used for this well-log data and Bayesian analysis was performed.

We assume that there are no more than 4049 change-points in the observation sequence. We partitioned the sample space according to the model index with k_{min}=10 and k_{max}=20. We set t_0=50000, λ=15, r=2.0 and δ=0.00001, for a conjugate prior on σ_i^{-2}. For a proposal distribution, the uniform distribution was used. The SAMC algorithm was run for 10^7 iterations. A C-code used in all examples of this paper for implementing the SAMC algorithm is available upon request from the authors. Table 4.1 lists the five models with the largest log-posterior values identified by SAMC, and shows the maximum posterior change-point estimates (26, 1034, 1070, 1210, 1220, 1420, 1433, 1525, 1684, 1866, 2046, 2408, 2469, 2532, 2591, 2771, 2780, 3942, 3963). Figure 4.1 shows the performance of the maximum *a posteriori* (MAP) estimates of the change-points, indicating that nineteen change-points separate into homogenous groups well by corresponding with the abrupt changes in the mean of the data, as would be expected.

Table 4.1: The 5 models with the largest log-posterior values with well-log data.

#of change-pts	Log-posterior	BIC	Position
19	-5659.11	11484.35	(26,1034,1070,1210,1220,1420,1433,1525,1684,1866, 2046,2408,2469,2532,2591,2771,2780,3942,3963)
20	-5663.97	11502.38	(26,1034,1070,1210,1220,1420,1433,1525,1684,1866, (26,1041,1070,1210,1220,1420,1433,1525,1684,1866,
20	-5669.06	11512.55	(26,1041,1070,1210,1220,1420,1433,1525,1684,1866, 2046,2408,2469,2532,2591,2771,2780,3739,3942,3963)
19	-5670.28	11506.70	(26,1040,1070,1210,1220,1415,1433,1525,1684,1866, 2046,2408,2469,2532,2591,2771,2780,3942,3963)

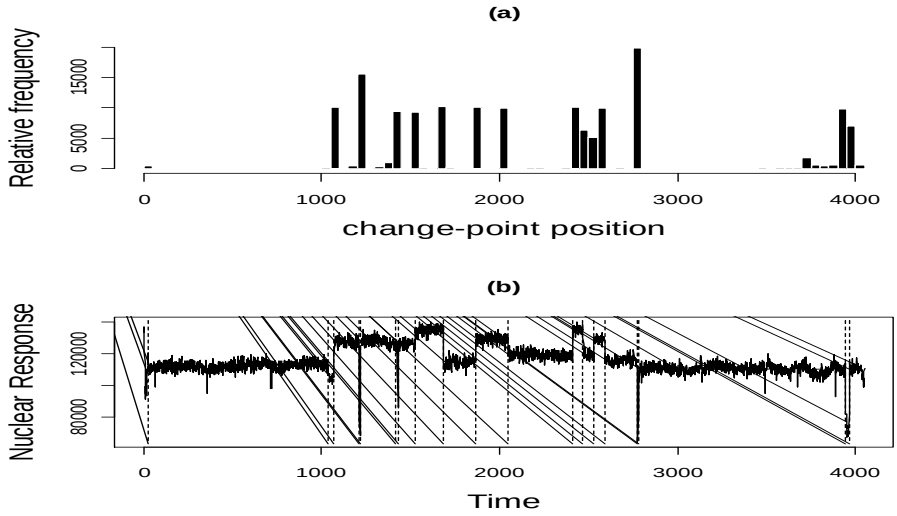


Fıgure 4.1: Well-log data: (a) The histogram for log posterior probabilities of change-point positions; (b) A maximum posteriori estimate of the change-point positions.

5. COMMENTS AND CONCLUSION

We have discussed Bayesian multiple change-point models for the exponential family distributions developed by Kim and Cheon (2009) and Cheon and Kim (2009) in this paper. We applied the SAMC algorithm, as a computational tool for posterior calculation, to the change-point identification since change-point estimation problem involves variable subspace dimensions. Although it would seem to be computationally intensive due to the unknown number of change-points, numerical results shows that SAMC can overcome this problem successfully. We illustrate the application of the posterior distributions to several data sets such as the well-log data. We find the results work as well as those given in earlier literatures on change-point estimation. Hence our method is simple to understand and is easily applied on change-point estimation for the sequence of independent random variable or random vectors from exponential distributions.

REFERENCES

Adams, R. P. and Mackay, D. J. C. (2007). Bayesian Online Changepoint Detection. *Univ. of Cambridge Technical Report*.

Barry D. and Hartigan, J. A. (1993). A Bayesian analysis for change-point problems. *Journal of the American Statistical Association,*88, 309-319.

Belisle, P., Joseph, L., MacGibbon, B., Wolfson, D. and du Berger, R. (1998). Change-point analysis of neuron spike train data. *Biometrics*, 54, 113-123.

Carlin, B. P., Gelfand, A. E. and Smith, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics*, 41, 389-405.

Cheon, S. and Kim, J. (2009). Multiple Change-point Detection of Multivariate Mean Vectors with Bayesian Approach. *Computational Statistics & Data Analaysis*, In press.

Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subject to changes in time. *Annals of Mathematical Statistics*, 35, 999-1018.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86, 221-241.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint Problems. *Statistics and Computing*, 16, 203-213.

Fearnhead, P. and Clifford, P. (2003). Online inference for hidden Markov models, *Journal of the Royal Statistical Society, Series B*, 65,887--899.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711-732.

Hawkins, D. M. (2001). Finding multiple change-point models to data, *Comput. Statistics & Data Analaysis*, 37, 323-341.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, 57, 1-17.

Kander, Z. and Zacks, S. (1966). Test procedure for possible changes in parameter of statistical distributions occurring at unknown time points, *Annals of Mathematical Statistics*, 37, 1196-1210.

Kass, R. E. and Raftery, A. (1995). Bayes factors, *Journal of the American Statistical Association*, 90, 773-795.

Kim, J. and Cheon, S. (2009). Bayesian Multiple Change-point Estimation with Annealing Stochastic Approximation Monte Carlo, *Computational Statistics*, Accepted.

Liang, F. (2007). Annealing Stochastic Approximation Monte Carlo for neural network training, *Mach. Learn*, 68, 201-233.

Liang, F. (2009). Improving SAMC Using Smoothing Methods: Theory and Applications to Bayesian Model Selection Problems, *The Annals of Statistics*, 37, 2626-2654.

Liang, F., Liu, C. and Carroll, R. (2007). Stochastic approximation in Monte Carlo computation, *Journal of the American Statistical Association*, 102, 305-320.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Socio. methodology,* 25, 111-163.

Robbins, H. and Monro, S. (1951). A stochastic approximation method, *Annals of Mathematical Statistics*, 22, 400-407.

Ruanaidh, J. J. K. and Fitzgerald, W. J. (1996). *Bayesian Online Changepoint Detection*, Springer, New York.

Smith, A. F. M. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables, *Biometrika*, 62, 407-416.

Stephens, D. A. (1994). Bayesian retrospective multiple-change-point identification. *Applied Statistics*, 43, 159-178.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion), *Annals of Stat*, 22, 1701-1762.

Venter, J. H. and Steel, S. J. (1996). Finding multiple abrupt change-points, *Comput. Stat. & Data Analaysis*, 22, 481-504.

Yao, Y. C. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches, *The Annals of Statistics*, 12, 1434-1447.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 831–844.

AN APPROPRIATE WEIGHT MODEL FOR FORECASTING FUZZY TIME SERIES AR(1) PROCESS

[1]Muhammad Hisyam Lee, [2]Riswan Efendi and [3]Zuhaimy Ismail
[1, 2, 3]Department of Mathematics, Universiti Teknologi Malaysia
81310 Skudai, Johor, Malaysia
E-mail: [1]mhl@utm.my, [2]wanchaniago@gmail.com and [3]zhi@fs.utm.my

ABSTRACT

This paper presents the appropriate weight for forecasting of AR(1) process based on fuzzy time series. Left and right (LAR) method has been proposed to obtain the appropriate weight of fuzzy logical relationship (FLRs) for forecasting of AR(1) process. Determining of weights was assigned using a collection of variations of chronological numbers in a fuzzy logical group (FLG). In addition, the weight and midpoint interval were done into the forecasting method, namely non-reversal and reversal methods. Both methods were validated through simulation using data which were generated from the certain time series models. By using data are generated from AR(1 and simulation technique both methods have been compared respectively. The experimental results show that the average of mean square error (MSE) from non-reversal method is smaller than reversal method. However, both of methods can be applied for forecasting of AR(1) process. In the end of this paper, the proposed method can be trained and tested by using real data.

Keywords: Fuzzy time series; stationary; weight; left and right method; non-reversal method, reversal method.

1. INTRODUCTION

Fuzzy time series has been widely explored for forecasting of diverse fields. Its application varies from forecasting of university student enrollment [Song and Chissom (1993a), Song and Chissom (1994), Chen (1996), Chen and Hsu (2004), Sah and Konstantin (2005), Kuo *et al*. (2009), and Chu *et al*. (2009)] to forecasting stock index [Huarng et al. (2007), Yu (2005), Cheng *et al*. (2006), Lee *et al.* (2009), Jilani and Burney (2008), Yu and Huarng (2008), and Chen (2000)], temperature [Lee *et al.* (2006)] and financial forecasting [Singh (2007a)]. The most interesting in fuzzy time series forecasting is the assumption regarding data that are not needed and this vie is totally different from the statistical methods. In principle, this fuzzy time series model because it has been established by using fuzzification, fuzzy logical relationship (FLRs), fuzzy logical group (FLG) and defuzzification.

In the year 1993 Song and Chissom (1993a) initiated a method to forecast the enrollments of the University of Alabama based on fuzzy time series. They converted the historical time series data into some linguistic values; the formation of linguistic variables was made after allocating the universe of discourse U (where U is a finite set) and partitioning into several equal length intervals. These linguistic variables were then converted into fuzzy data. From these fuzzy data, the fuzzy logical relationships of enrollments were made and fuzzy logical relationship groups

were established. Song and Chissom (1993a) and Song and Chissom (1994) proposed a universal forecasting method using fuzzy sets, which was termed as fuzzy time series. They presented the method to forecast the enrollments of the University of Alabama when the historical data were linguistic values, using fuzzy time series method. It discussed the concepts of time invariant fuzzy time series. They also extended the scope of their previous work, Song and Chissom (1993a), wherein they reported the application of the first-order time variant model. The difference between time-invariant and time-variant model and time invariant model was discussed. They discussed the first order time-invariant fuzzy time series model and a first order time-variant model. These models were compared with each other and with a time invariant Markov model using linguistic labels with probability distribution. These models were compared with each other and with a time invariant Markov model using linguistic labels with probability distribution; Sullivan and Woodall (1994). Further the results of these methods were compared with a first order autoregressive AR(1) model and second-order autoregressive AR(2) models.

Yu (2005) suggested the used of weighted fuzzy time series models for Taiwan stock index (TAIEX) forecasting. It is assigned by the recurrent fuzzy logical relationships (FLRs) in fuzzy logical group (FLG). In establishing fuzzy relationship and forecasting are important step to consider the weighted. Cheng *et al.* (2006) proposed the trend-weighted fuzzy time series model for TAIEX forecasting. The study for fuzzy time series is actively researched. Some of the most recent work includes the study by Lee and Park (1997) who proposed an efficient algorithm to compute the fuzzy weighted average for the purpose of aggregating imprecise sensory information represented by fuzzy numbers, which turned out to be superior to the previous work by reducing number of comparisons and arithmetic operations. The theoretical background for the fuzzy weighted average algorithm was constructed and verified. Kato and Sakawa (1998) proposed the formulation of large-scale multi-objective block - angular linear programming problems involving fuzzy numbers. Sugeno and Tanaka (1984) proposed successive identification method of a fuzzy model. The structure and initial parameters of a fuzzy model were determined to successively identify a fuzzy model. The model was called the 'initial model'. The initial model was identified by the off-line fuzzy modeling method using some pairs of input - output data. Bintley (1987) constructed an Expert System with the REVEAL modeling system. The expert system was applied to the problem of time series analysis. The concept of 'fuzzy modeling' was used to avoid over-fitting the model to the noise. A time series could be considered as a series of data observed or measured at regular intervals. Time series might be standing alone ('univariate') measurements of a single parameter or might be considered together in an attempt to establish the relationships between different phenomena ('multivariate analyses). Even though extensive studies were conducted, some of the studies do not focused on establishing of the appropriate rules for each component time series data such as stationary data, seasonal variation, trend series, and combination between seasonal and trend series data in fuzzy time series forecasting.

In our study, the focus will be to determine the appropriate weight and establish the forecasting method for AR(1) process. In addition, to obtain of weight for fuzzy logical relationship can be done based on left and right (LAR) method. Weights are determined using a collection of variation of chronological number in the fuzzy logical group (FLG). For forecasting method, there are two methods that can be proposed. Both methods are done by using weight and the midpoint interval, namely a non-reversal and reversal methods. Furthermore, both methods will be tested and trained by using data that are generated from AR(1) model using

simulation technique. The fitness function of mean square error (MSE) will be used on measuring the forecasting performance for both methods.

In this paper, an effort has been made to find the appropriate weight for forecasting of the stationary data based on fuzzy time series especially AR(1) process. The discussion begins with some definitions of the basic theory of fuzzy time series, and stationarity of time series data followed by the discussion on the appropriate weighted for stationary data on fuzzy time series. Few examples are used in the discussion on the proposed procedure on forecasting and simulation method. The verification, comparison and the model testing for real data are discussed and this paper ends with the conclusion of the study.

2. METHODOLOGY

A fuzzy set can be recognized by a membership function defined as in Huarng (2001):

Definition 1:
Let U be the universe of discourse. A fuzzy subset A on the universe of discourse U can be defined as follows:

$$A = \left\{ \left(u_i, \mu_A(u_i) \right) \middle| u_i \in U \right\} \tag{1}$$

where μ_A is the membership function of A, $\mu_A : U \rightarrow [0, 1]$, and $\mu_A(u_i)$ is the degree of membership of the element u_i in the fuzzy set A.

Definition 2:
Let U be the universe of discourse, $U = \{u_1, u_2, \ldots, u_n\}$, and U be a finite set. A fuzzy set A can be expressed as follows:

$$A = \sum_{i=1}^{n} \frac{\mu_A(u_i)}{u_i} = \frac{\mu_A(u_1)}{u_1} + \frac{\mu_A(u_2)}{u_2} + \cdots + \frac{\mu_A(u_n)}{u_n} \tag{2}$$

where the symbol "+" means the operation of union instead of the operation of summation, and the symbol "—" means the separator rather than the commonly used algebraic symbol of division.

Definition 3:
Let U be the universe of discourse, where U is an infinite set. A fuzzy set A of U can be expressed as follows:

$$A = \int_U \frac{\mu_A(u_i)}{u_i}, \quad \forall_{u_i} \in U \tag{3}$$

In addition, there are several definitions have been defined for fuzzy time series, see, for example, Song and Chissom (1993), Chen (1996), and Singh (2007).
Definition 4:
Let $Y(t)$ be the universe of discourse defined by the fuzzy set $\mu_i(t)$. If $F(t)$ consists of $\mu_i(t)$ (i = 1, 2, …), $F(t)$ is defined as a fuzzy time series on $Y(t)$ (t = …, 0, 1, 2, …), where $Y(t)$ is a subset of

real number. Following *Definition* 3, fuzzy relationships between two consecutive observations can be defined.

Definition 5:
Suppose $F(t)$ is caused by $F(t\text{-}1)$ denoted by $F(t\text{-}1) \rightarrow F(t)$, then this relationship can be represented by

$$F(t) = F(t-1) \circ R(t, t-1) \tag{4}$$

where $R(t, t\text{-}1)$ is a fuzzy relationship between $F(t)$ and $F(t\text{-}1)$ and is called the first-order model of $F(t)$.

Definition 6:
Let $F(t-1) = A_i$ and $F(t) = A_j$. The relationship between two consecutive data (called a fuzzy logical relationship, FLR), i.e., $F(t)$ and $F(t\text{-}1)$, can be denoted by $A_i \rightarrow A_j$, $i, j = 1, 2, \ldots, p$ (where p is interval or subinterval number) is called the left-hand side (LHS), and A_j is the right-hand side (RHS) of the FLR. The proposed a fuzzy time series model with procedure as follows:

- to define the universe of discourse and intervals
- to fuzzify
- to establish fuzzy relationships
- to forecast

Definition 7:
Let $A_i \rightarrow A_j, A_i \rightarrow A_k, ..., A_i \rightarrow A_p$ are FLRs with the same LHS can be grouped into an ordered FLG (called a fuzzy logical group) by putting all their RHS together as on the RHS of the FLG. It can be written as follows:

$$A_i \rightarrow A_j, A_k, ..., A_p \quad i, j, k, ..., p = 1, 2, \ldots, p \tag{5}$$

Stationarity is the first features in time series that are related to the mean value and variance of observation data. The series is said to be stationary if the mean and variance is constant over time, and the covariance between observations y_t and $y_{t\text{-}d}$ only dependent on the distance between the two observations that does not change over time. The usual practice in detecting stationarity of the data is by using time plot. Suppose there are n observations with values $y_1, y_2, \ldots, y_n$ of a time series, then these values when plotted against time will determine whether the time series is stationary. If the n values seem to fluctuate with constant variation around a constant mean μ, then is it reasonable to believe that the time series is stationary. In contrary, if the n values do not fluctuate around a constant mean or do not fluctuate with constant variation, then it is non-stationary; Bowerman and O'Connel (1987). The stationary condition may also be investigated by using autocorrelation function (ACF) and partial-autocorrelation function (PACF).

In time series modeling, the stationarity can be found in Autoregressive (AR) process, Moving Average (MA) process, and mix process between (AR) and (MA) known as the Autoregressive Moving Average (ARMA). Consider the following for first-order autoregressive (AR(1)) process; Palit and Popovic (2005).

$$y_t = \phi_1 y_{t-1} + a_t \tag{6}$$

with the stationarity condition requires that the variance are constant over time $Var(y_t) = Var(y_{t+1})$.

or the equality $E\{[\phi_1 y_{t-1} + a_t]^2\} = E\{[\phi_1 y_{t-1} + a_{t-1}]^2\}$ holds. Therefore, because of mutual independence of a_t and y_{t-1}, the equality $Var(y_t) = \phi_1^2 Var(y_{t-1}) + Var(a_t)$ follows, and finally the equality $\gamma_0 = \phi_1^2 \gamma_0 + \sigma^2$, where γ_0 does not depend on time *t*.

In this study, a generated AR(1) will be used in the simulation exercise. This is a powerful tool for analysis of many mathematical models and real-world systems when analytical solutions are not possible. Generally, there are some statistical aspects of simulation such as formulation of the problem, input data analysis, the model and computer program, validation, experimental design, and sample size; Kleijnen (1974). Using the statistical techniques, a sound simulation model can be built and adequately tested before implementation.

3. THE APPROPRIATE WEIGHT FOR STATIONARY TIME SERIES DATA

In the previous studies, Yu (2005) has discussed the weighted fuzzy time series models for Taiwan stock index (TAIEX) forecasting. Cheng *et.al.* (2006) suggested the trend-weighted fuzzy time series model for TAIEX forecasting. The weight used in their work was determined using recurrence on fuzzy logical relationship (FLR). Weighted was used to improve forecast accuracy but no work has explored the appropriate weighted for each component of the time series data. In this section, the discussion of weighted for stationary time series data will be detailed.

Fuzzy time series forecasting has been widely explored in the last decades but those studies did not explain the best rule and procedure of forecasting for the major characteristic features of time series data such as stationarity, linearity, trend, and seasonality. The existing procedure has been established based on the real data only and are not used for predicting each component in the time series data. In this study, the discussion will be on determining the appropriate weight for forecasting with stationary time series data. In addition, the term weight means are commonly used in statistics. In fuzzy time series, the numerical data will be transformed into the linguistic values. Further, the relationship will be mapped between each past linguistic value and each present linguistic value. This relationship is known as fuzzy logical relationship (FLG); Song and Chissom (1993a) and Yu (2005). If the actual data is stationary, then many relationships may be established among the same linguistic values or the recurrence can be found more than twice. Consecutively, the appropriate weight must be assigned to this type of data.

Our reviews on previous studies in related areas show that the descriptions of the recurrent fuzzy relationships are not clearly given. The repeated FLRs were simply ignored when fuzzy relationships were established. The following examples as given in Yu (2005) can be used to explain the repeated FLRs. Let there be the FLR_s given in chronological order as in Table 1.

Based on Table 1, there are four out of five FLR having the same LHS, A_1. The occurrences of the same FLR in column 2 are regarded as if there were only one occurrence. In other words, the recent identical FLR are not considered in the work by Song and Chissom (1993a). It is questionable if these recurrences are ignored. The occurrence of a particular FLR represents the number of its appearances in the past. For instance, in column 2, $A_1 \rightarrow A_1$ appears three times and $A_2 \rightarrow A_1$ appears only once. The recurrence can be used to indicate how the FLR may appear in the future. Hence, to cover all of the FLR, an approach to represent the fuzzy relationship is suggested below: $A_1 \rightarrow A_1, A_2, A_1, A_1, A_1$.

The various recurrences of FLR have been considered for determining weight on fuzzy time series by Yu (2005) and Cheng *et al*. (2006).

Table 1. Recurrence of fuzzy logical relationship (FLRs)

Linguistics Value	FLR_s	FLG (Fuzzy Logical Group)
A_1	-	
A_1	$A_1 \rightarrow A_1$	
A_2	$A_1 \rightarrow A_2$	$A_1 \rightarrow A_1, A_2, A_1, A_1, A_1$
A_1	$A_2 \rightarrow A_1$	$A_2 \rightarrow A_1$
A_1	$A_1 \rightarrow A_1$	
A_1	$A_1 \rightarrow A_1$	

In this paper, we do not consider the recurrence for computational weight. The recurrences occur due to the relationship between the same linguistic occur repeatedly on stationary time series data. From the simulation study, the number of the same FLR occurring more then 3 or 4 times for each group in FLG. The occurrence can be shown in Table 2.

Table 2 shows the various recurrences in FLG. For example, A_3 has 3 relationships with A_3, A_4 has 4 relationships A_4, A_6 has 8 relationships, A_7 has 7 relationships with A_7. On the other hand, the relationship between $A_i \rightarrow A_j$ also can be found frequently. While, for A_1, A_2 and A_9 can not be found their relationship with others. This condition is the main reason to ignore the recurrence for establishing the appropriate weight on stationary time series data. In 2009, Lee *et al.* (2009) developed the modified weight based on collection of variation of the chronological number in fuzzy logical group (FLG). In addition, to tackle this phenomenon, the new rule are presented based on left and right (LAR) relationship as follows

1. Imperfect LAR
   The weight can be determined if there $A_i \rightarrow A_{j-1}, A_j$ or $A_i \rightarrow A_j, A_{j+1}$ which $i = j, i, j \geq 1$ and $i, j \in Z$. In this condition, we can see that A_i has two relationships namely; the first value (A_{j-1}) that appears to left of the original value (A_i), the second value has the same value as original in a FLG. Therefore, this relationship is call as an imperfect left and right (LAR) relationship.
2. Perfect LAR
   The weight can be assigned if there are relationships in a FLG as follows
   $A_i \rightarrow A_{j-1}, A_j, A_{j+1}$ ($i = j, i, j \geq 2$ and $i, j \in Z$)
   $A_i \rightarrow A_{j-1}, A_{j+1}, A_j$
   $A_i \rightarrow A_j, A_{j-1}, A_{j+1}$
   $A_i \rightarrow A_j, A_{j+1}, A_{j-1}$
   $A_i \rightarrow A_{j+1}, A_{j-1}, A_j$
   $A_i \rightarrow A_{j+1}, A_j, A_{j-1}$
   Six conditions above are called as perfect LAR because we can find the left and right relationship of original value for each FLG.

Example
From Table 2, we have $A_1 \rightarrow A_2$, no weight can be considered because the relationship occurred only one time. We also have $A_9 \rightarrow A_8, A_5, A_8, A_6$, but also no weight can be obtained because it is not compatible with rule 1 and 2.

Example for rule 2
From Table 2, we have $A_3 \rightarrow A_5, A_3, A_6, A_6, A_3, A_2, A_3, A_5, A_2, A_4$. This FLG complies with the rule 2. In this case, the first value that appears to left of the original value (A_4), then the first value that appears to right of the original value (A_2) and has the same value as original (A_3). Thus, three conditions above are called as a perfect left and right relationship. In addition, this FLG can be written simply as

$$A_3 \rightarrow A_3, A_2, A_4 \tag{7}$$

or other FLG

$$A_8 \rightarrow A_9, A_8, A_7 \tag{8}$$

Table 2. Recurrence on fuzzy logical group (FLG)

Fuzzy Logical Group (FLG)
$A_1 \rightarrow A_2$
$A_2 \rightarrow A_3, A_4, A_6, A_6$
$A_3 \rightarrow A_5, A_3, A_6, A_6, A_3, A_2, A_3, A_5, A_2, A_4$
$A_4 \rightarrow A_5, A_3, A_5, A_2, A_4, A_3, A_7, A_4, A_4, A_6, A_4, A_6$
$A_5 \rightarrow A_3, A_4, A_6, A_7, A_6, A_5, A_4, A_8, A_8, A_6, A_6, A_6, A_3$
$A_6 \rightarrow A_7, A_7, A_3, A_6, A_7, A_6, A_7, A_8, A_6, A_7, A_7, A_8, A_6, A_6, A_5, A_5, A_8, A_5, A_5, A_3, A_8, A_6, A_6, A_6, A_9$
$A_7 \rightarrow A_4, A_7, A_9, A_7, A_7, A_7, A_6, A_8, A_4, A_6, A_7, A_7, A_7, A_5, A_8, A_4, A_5$
$A_8 \rightarrow A_9, A_8, A_6, A_9, A_7, A_7, A_4, A_9, A_6, A_8, A_5, A_8, A_6$
$A_9 \rightarrow A_8, A_5, A_8, A_6$

The computational weights are assigned by using the rule given the Section 3.2. Suppose $A_i \rightarrow A_{j-1}, A_j, A_{j+1}$, $i = j$ is a FLG and the weights are specified as follows:

$$(j\text{-}1) = c_1,\ j = c_2,\ (j + 1) = c_3.$$

Then

$$\begin{aligned}
\mathbf{W(t)} &= \begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix} \\
&= \left[\frac{(j-1)}{(j-1)+j=(j+1)} \quad \frac{j}{(j-1)+j+(j+1)} \quad \frac{(j+1)}{(j-1)+j+(j+1)} \right] \\
&= \left[\frac{c_1}{(c_1+c_2+c_3)} \quad \frac{c_2}{(c_1+c_2+c_3)} \quad \frac{c_3}{(c_1+c_2+c_3)} \right] \\
&= \left[\frac{c_1}{\sum_{h=1}^{3} c_h} \quad \frac{c_2}{\sum_{h=1}^{3} c_h} \quad \frac{c_3}{\sum_{h=1}^{3} c_h} \right]
\end{aligned} \tag{9}$$

Example 1
From the FLG (7) and (8) then the weight can be determined by using equation (9)
Given $A_3 \rightarrow A_3, A_2, A_4$ and $c_1 = 3$, $c_2 = 2$, $c_3 = 4$ then

$$w_1 = 3/(3 + 2 + 4),\ w_2 = 2/(3 + 2 + 4),\ w_3 = 4/\ (3 + 2 + 4)$$

thus $\mathbf{W}(A_3) = \mathbf{W}(t) = [\ w_1 \quad w_2 \quad w_3\ ] = [\ 3/9 \quad 2/9 \quad 4/9\ ] = [\ 0.33 \ \ 0.22 \ \ 0.45\ ]$

Given $A_8 \rightarrow A_9, A_8, A_7$ and $c_1 = 9$, $c_2 = 8$, $c_3 = 7$ then

$$w_1 = 9/(9 + 8 + 7),\ w_2 = 8/(9 + 8 + 7),\ w_3 = 7/\ (9 + 8 + 7)$$

thus $\mathbf{W}(A_8) = \mathbf{W}(t) = [\ w_1 \quad w_2 \quad w_3\ ] = [\ 9/24 \quad 8/24 \quad 7/24\ ] = [\ 0.37 \ \ 0.33 \ \ 0.30]$. Therefore, weights are satisfying condition in both of weight matrix.

In this study two models are proposed for forecasting, namely non-reversal model and reversal model. Both models can be detailed as follows: Let $A_i \rightarrow A_{j-1}, A_j, A_{j+1}$ is a FLG and the corresponding weights for A_{j-1}, A_j, A_{j+1} are w_1, w_2, w_3. The defuzzified of the midpoints of A_{j-1}, A_j, A_{j+1} are $m_{j\text{-}1}$, m_j, m_{j+1}. It can be denoted in the product of the defuzzified matrix and the transpose of the weight matrix:

In the forecasting model, two different methods may be used. The forecast model $F(t)$ is given as

$$F(t) = \mathbf{M(t)W(t)^T} = [m_{j-1}\ m_j\ \ m_{j+1}] \times [w_1\ w_2\ w_3]^T \tag{10}$$

where the number elements in matrix $\mathbf{M}(t)$ and $\mathbf{W}(t)$ are equal. In addition, the equation (10) can be denoted as a non-reversal method and it is used for in-sample forecast. Equation (10) can be modified as follows

$$F(t) = [m_{j-1}\ m_j\ \ m_{j+1}] \times [w_3\ w_2\ w_1]^T \tag{11}$$

where $\mathbf{M}(t)$ is a 1 x n matrix and $\mathbf{W}(t)^T$ is a n x 1 matrix, respectively. Equation (11) is known as a reversal method and it is used for in-sample forecast. Both of the equations are validated by simulation. For out-sample forecast, that equation can be denoted as

$$F(t) = \mathbf{M}(t\text{-}1) \text{ x } \mathbf{W}(t\text{-}1)^T \tag{12}$$

Example 2
By using a FLG from Example 1 then the forecast $F(t)$ can be determined as follows

$$F(t) = \mathbf{M}(t) \text{ x } \mathbf{W}(t)^T = [\ m_3 \ \ m_2 \ \ m_4\ ] \text{ x } [\ w_3 \ \ w_2 \ \ w_4\ ]^T.$$

4. THE ALGORITHM AND SIMULATION PROCEDURE

The computational step wise for forecasting are described as follows:

Step 1. Partition the universe of discourse U into several intervals of equal length. In general, U is defined as $U = [\ D_{\min} - D_1,\ D_{\max} + D_2]$, where $D_{\min}$ and $D_{\max}$ are the minimal and maximal

values of the historical data, D_1 and D_2 are proper positive numbers. Then U is partitioned into n equal intervals, $u_1, u_2, \ldots, u_n$, with length l defined as[28]

$$l = 1/\mathrm{n}[(D_{\min} - D_1) - (D_{\max} + D_2)]. \quad (13)$$

Step 2. Establish fuzzy sets for observations. Each linguistics observation A_i can be defined by the intervals: $u_1, u_2, \ldots, u_n$. Each A_i can be represented as following equation (13) and the value, k_j, is determined by the following situations[16]

IF $j = i$-1**, then** $k_j = 0.5$;
IF $j = i$, **then** $k_j = 1$;
IF $j = i + 1$, **then** $k_j = 0.5$; **elsewhere** $k_j = 0$; **and** $A_i = \Sigma\, k_j / u_j$.

Step 3. Establish the fuzzy relationships. Two consecutive fuzzy sets $A_i(t\text{-}1)$ and $A_j(t)$ can be denoted by a single FLR as $A_i \rightarrow A_j$.

Step 4. Establish the fuzzy logical groups for the corresponding trends. The FLRs with same LHSs (left hand sides) can be grouped to form a FLG. For example, $A_i \rightarrow A_j$, $A_i \rightarrow A_k$, $A_i \rightarrow A_m$ can be grouped as $A_i \rightarrow A_j$, A_k , A_m .

Step 5. Assign the weights. The weights can be calculated by using rule (1 & 2) given in Section 3.2 and 3.3.

Step 6. Calculate the forecast value by using equation (10), (11) for in-sample and equation (12) for out-sample. In this study, two rules are employed for forecasting as follows:

***Rule* 1** : If there are no weight for A_i , then the forecast is equal to midpoint of A_i .

***Rule* 2** : If there are weight for A_i , then the forecast can be computed using the equation (10), (11) and (12) for out-sample.

The procedure also can be illustrated in the form of flowchart as in Figure 1.
The procedure for simulation model to determine the appropriate weight for AR(1) process is presented in the following 4 steps.

Step 1. Choose one AR(1) model. For demonstration purposes, we simulate using the following AR(1) model $y_t = 5 + 0.5y_{t-1} + a_t$. where $\{a_t\}$ is the white noise.

Step 2. Generate a_t by using number random generation where a_t's are i.i.d N(0, 1). In this study, we generate 100 data and simulate 100 times for each experiment.

Step 3. Use the procedure in Section 4.1 then y_t is predicted by using both methods as described in Section 3.4.

Step 4. Compute the mean square error (MSE) of in-out sample forecast for each method, and then compare with MSE of y_t. In this simulation, MSE of y_t is get equal to 1.

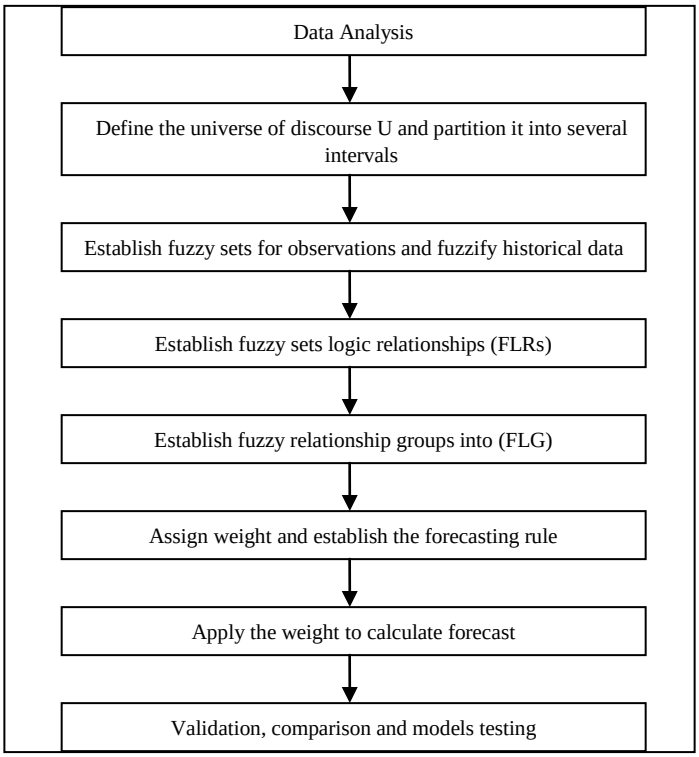


Fig. 1. Forecasting procedure

3.1 The Validation of Proposed Method

Simulations results in the form in-sample and out-sample MSE for non-reversal and reversal method is given in Table 3.

Table 3. MSE of the proposed methods

No of experiment	MSE of Non-reversal Method		MSE of Reversal method	
	In-sample	Out-sample	In-sample	Out-sample
1	0.10	1.38	0.09	1.38
2	0.06	1.31	0.06	1.39
3	0.18	2.82	0.18	2.78
4	0.12	1.31	0.09	1.28
5	0.06	0.82	0.06	0.86
6	0.10	1.90	0.10	1.87
7	0.09	2.05	0.07	2.04
8	0.08	1.86	0.07	2.20
9	0.09	1.42	0.07	1.39
10	0.04	1.76	0.04	1.79
⋮	⋮	⋮	⋮	⋮
99	0.10	0.70	0.10	0.74
100	0.10	1.25	0.08	1.39
Sum	8.07	133.46	7.71	136.05
Average	0.0807	1.3346	0.0771	1.3605

From Table 3, it shows that the MSE in-sample from reversal method is smaller than non-reversal method. On the other hand, average of MSE out-sample from non-reversal method is smaller than reversal method. In addition, the difference in MSE for the proposed method is not significant. Thus, they can be used for AR(1) process. In the other word, the weighted which have been assigned based on left and right (LAR) method can be called as an appropriate method for weight on the forecasting of AR(1) process.

In this section, the performance of the proposed method then both methods are tested to real data and compared with AR(1) model. The data was taken from Box *et.al* (1994) as data training. There are 70 observations from yield chemical process every hour. By using the computational step wise as given in Section 4. The forecast results can be computed and presented in Table 4.

Table 5 shows that MSE of non-reversal and reversal methods are smaller than AR(1) model. MSE of both methods are too significant difference with AR(1). It could be denoted that the difference of MSE is 5 times smaller than AR(1) model. Besides that, this improvement is also influenced by partition number. We propose 15 intervals for observations. This interval is determined by MSE minimum of out-sample forecast. In addition, Table 5 indicates that MSE out-sample forecast of reversal method is smaller than non-reversal method and also AR(1) but the difference is not too significant between non-reversal and AR(1). However, both of proposed methods are better than AR(1).

The comparison of MSE for each method can be seen in Table 5.

Table 4. Performance of the proposed method and AR(1) model

Time	Actual(y_t)	Forecasted		
		Non-Reversal Method	Reversal Method	AR(1)
1	47			51.38
2	64	64.36	64.36	55.41
3	23	25.72	25.72	47.30
4	71	70.80	70.80	66.86
5	38	38.60	38.60	43.96
6	64	64.36	64.36	59.71
7	55	55.03	54.37	47.30
8	41	38.60	38.60	51.60
9	59	55.03	54.37	58.27
10	48	45.04	45.04	49.69
11	71	70.80	70.80	54.94
12	35	32.16	32.16	43.96
13	57	55.03	54.37	61.14
14	40	38.60	38.60	50.64
15	58	55.03	54.37	58.75
16	44	45.04	45.04	50.17
17	80	77.26	77.26	56.84
18	55	55.03	54.37	39.67
19	37	38.60	38.60	51.60
20	74	70.80	70.80	60.18
⋮	⋮	52.32	51.86	42.53
49	43	45.04	45.04	47.30
50		52.32	51.86	57.32
51	Out-sample	52.32	51.86	50.46
52		52.32	51.86	53.73
⋮	⋮	⋮	⋮	⋮
70	Out-sample	52.32	51.86	52.67

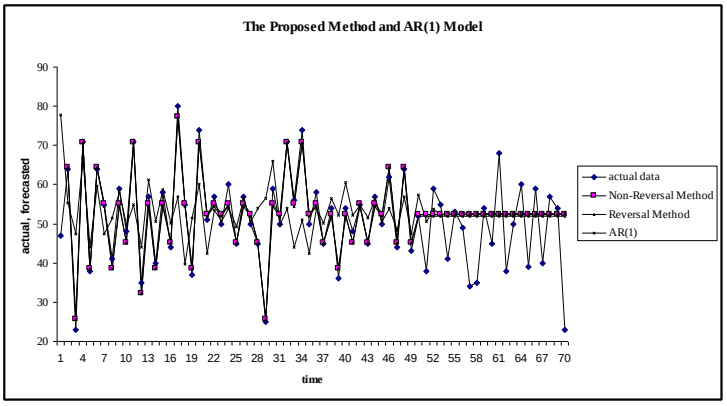


Figure 2. The forecast results based on the proposed method and AR(1) model.

Table 5. Comparison of MSE

Method	MSE	
	In-sample forecast (N = 49)	Out-sample forecast (N = 21)
Non-reversal	4.498	136.291
Reversal	5.227	132.309
AR(1)	115.780	137.907

5. CONCLUSIONS

In the paper, we proposed left and right (LAR) method to obtain an appropriate weight of fuzzy logical relationship for forecasting of AR(1) process. Its application is very useful to improve the forecasting accuracy. Moreover, the partition number is also affected factor to obtain a better forecast. The experimental results showed that MSE of proposed method are smaller as compared to AR(1). The performance of non-reversal and reversal methods are also better than AR(1) for forecasting of real data. Therefore, In future study, the appropriate weighted should be extended to reach a higher forecasting accuracy if the stationary time series follow MA(1) and ARMA(1, 1) process.

ACKNOWLEDGMENT

This research was supported by Research Management Centre, Universiti Teknologi Malaysia (UTM).

REFERENCES

Bintley, H. (1987). Time series analysis with reveal. *Fuzzy Sets and Systems*, 23, 97-118.

Bowerman, B. L. and O'Connel, R. T. (1987). *Time series forecasting; unified concepts and computer implementation*, Duxbury Press.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1994). *Time series analysis; forecasting and control*. Third edition, Prentice Hall.

Chen, S. M. (1996). Forecasting enrollments based on fuzzy time series. *Fuzzy Sets and Systems*, 81, 311-319.

Chen, S. M. (2000). Temperature prediction using fuzzy time series. *IEEE Transactions on Systems, Man, and Cybernetics*, 30, 263-275.

Chen, S. M. and Hsu, C. C. (2004). A new method to forecast enrollments using fuzzy time series, *International Journal of Applied Science and Engineering*, 3, 234-244.

Cheng, C. H., Chen, T. L. and Chiang, C. H. (2006). Trend-weighted fuzzy time series model for Taiex forecasting. *ICONIP*, Part III, LNNC 4234, 469-477.

Chu, H. H., Chen, T. L., Cheng, C. H. and Huang, C. C. (2009). Fuzzy dual-factor time series for stock index forecasting. *Expert Systems with Applications*, 36, 165-171.

Huarng, K. (2001). Heuristic models of fuzzy time series for forecasting. *Fuzzy Sets and Systems*, 123, 369-386.

Huarng, K., Tiffany Yu, H. K. and Yu, W. S. (2007). A multivariate heuristic model for fuzzy time series forecasting. *IEEE Transactions on Systems, Man, and Cybernetics*, 37, 263-275.

Jilani, T. A. and Burney, S. M. A. (2008). A refined fuzzy time series model for stock market forecasting. *Physica A*, 387, 2857-2862.

Kato, K. and Sakawa, M. (1998). An interactive fuzzy satisfying method for large scale multiobjective 0-1 programming problems with fuzzy parameters through genetic algorithms. *European Journal of Operation Research*, 107, 590-598.

Kleijnen, J.P.C. (1974). *Statistical techniques in simulation* - Part I, Dekker.

Kuo, I. H., Horng, S. J., Kao, T. W., Lin, T. L., Lee, C. L. and Pan, Y. (2009). An improved method for forecasting enrollments based on fuzzy time series and particle swarm optimization. *Expert Systems with Applications*, 36, 6108-6117.

Lee, M. H., Efendi, R. and Ismail, Z. (2009). Modified weighted for enrollment forecasting based on fuzzy time series. *Jurnal Matematika*, 25, 67-78.

Lee, C. H. L., Liu, A. and Chen, W. S. (2006). Pattern Discovery of Fuzzy time series for financial prediction. *IEEE Transactions on Knowledge and Data Engineering*, 18, 613-625.

Lee, D. H. and Park, D. (1997). An efficient algorithm for fuzzy weighted average. *Fuzzy Sets and Systems*, 87, 39-45.

Li, S. T. and Cheng, Y. C. (2009). An enhanced deterministic fuzzy time series forecasting model. *Cybernetics and Systems*, 40, 211-235.

Palit, A. K. and Popovic, D. (2005). *Computational intelligence in time series forecasting; theory and engineering applications*. Springer.

Sah, M. and Konstantin, Y. D. (2005). Forecasting enrollment model based on first-order fuzzy time series, *Proceeding of World Academy of Science, Eng and Tech*, 1, 375-378.

Singh, S.R. (2007a). A simple method of forecasting on fuzzy time series. *Applied Mathematics and Computation*, 186, 330-339.

Singh, S.R. (2007b). A robust method of forecasting based on fuzzy time series. *Applied Mathematics and Computation*, 188, 472-484.

Song, Q. and Chissom, B.S. (1993a). Forecasting enrollments with fuzzy time series – Part I. *Fuzzy Sets and Systems*, 54, 1-9.

Song, Q. and Chissom, B. S. (1993b). Fuzzy time series and its models. *Fuzzy Sets and Systems* 54, 269-277.

Song. Q. and Chissom, B.S. (1994). Forecasting enrollments with fuzzy time series – Part II. *Fuzzy Sets and Systems*, 64, 1-8.

Sugeno, M. and Tanaka, K. (1984). Successive identification of a fuzzy model and its application to prediction of a complex system. *Fuzzy Sets and Systems*, 13, 153-167.

Sullivan, J. and Woodall, W. H. (1994). A comparison of fuzzy forecasting and Markov modeling. *Fuzzy Sets and Systems*, 64, 279-293.

Yu, H. K. (2005). Weighted fuzzy time series models for Taiex forecasting. *Physica A* 349, 609-624.

Yu, T. H. K. and Huarng, K. H. (2008). A bivariate fuzzy time series model to forecast the Taiex. *Expert Systems with Application*, 34, 2945-2952.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 845–860.

MODELING THE IMPACT OF LAPINDO MUD FLOOD DISASTER AND NEW FUEL TARIFF TO VEHICLE VOLUME IN TOLL ROAD USING MULTI-INPUT INTERVENTION MODEL

[1]Muhammad Hisyam Lee and [2]Suhartono
[1]Department of Mathematics, Universiti Teknologi Malaysia, Malaysia
[2]Department of Statistics, Institut Teknologi Sepuluh Nopember, Indonesia
E-mail: [1]mhl@utm.my and [2]suhartono@statistika.its.ac.id

ABSTRACT

The purpose of this research is to study the impact of Lapindo mud flow disaster and new tariff of the oil fuel in Indonesia on transportation, particularly vehicle volume in toll-roads. The hot mud flow disaster from Lapindo Brantas, Inc. caused the Porong toll-road to be flooded and it had to be closed. The disaster very much impacted the traffic volume of the Surabaya-Gempol toll-road. This research focuses on developing a multi-input intervention model for explaining the impact of two interventions, namely the Lapindo mud flow disaster on May 29th 2006 and a new tariff for the oil fuel in October 2005. This model is used to analyze the decrease of the traffic volume in toll-road, specifically the magnitude and duration effects of both interventions. First, a theoretical study is carried out to derive the statistical inputs that could be used as a basic tool for determining the order of intervention model. These results are applied to construct a multi-input intervention model. This study shows that these two interventions significantly contributed to the decreasing of traffic volume of the toll-road. The decrease in traffic volume caused by the new tariff of the oil fuel is 388,512 vehicles since the month of the new tariff policy. The Lapindo mud flow resulted in a cumulative decrease in traffic volume of 387181, 553456 and 679485 vehicles for the first three months, fourth to sixth months and seventh until nineteenth months (the end of the duration), respectively. Thus, the multi input intervention model shows that the Lapindo mud flow disaster has a long lasting effect on the decrease of vehicle volume on this toll-road.

Keywords: Lapindo mud flow; new fuel tariff; traffic volume; multi input; intervention model.

1. INTRODUCTION

Malang and Pasuruan are cities with high tourism activities in East Java, Indonesia. Tourists who go to Malang and Pasuruan via Surabaya by cars and other big vehicles usually pass through Surabaya-Gempol toll road. On May 29, 2006 hot mud blasted from a volcano in Porong, southern of Sidoarjo, together with thick smoke. This mud came from the mining area of Banjar Panji I well, which is owned by Lapindo Brantas Inc. The mud flooded the Porong-Gempol toll road between the 38th and 39th km and blocked the traffic. This caused the toll road to be closed down on November 26, 2006 and people who travel between Malang and Surabaya had to find alternate routes. This event provides the main background for this research.

To date, a mathematical model has not been made to evaluate the impact of Lapindo mud flood to the amount of vehicles passing Waru-Gempol toll road. The main objective of this

research is to find a mathematical model which is appropriate to explain the impact of Lapindo mud flood on the number of vehicles in Waru-Gempol toll road. This model is used to measure the loss due to Lapindo mud flood in the transportation sector, especially the decreasing number of vehicles in the Waru-Gempol toll road. In this research, the model that examined and developed is the multi step function intervention model. In general, intervention models are a special type of time series models which are usually used to evaluate the internal and/or external impact in time series data.

Previous researches related to intervention models caused by internal factors could be seen in Box and Tiao (1975) who evaluate the impact of machine design laws to the oxidant polution rate in Los Angeles, McSweeny (1978) who investigates the impact of new tariff in Cincinnati Bell Telephone to the number of local help calls, Leonard (2001) who studies the impact of product promotion and the price rising, and Suhartono and Wahyuni (2002) who analyze the effect of promotion and price rising on consumer pulse consumption. Examples of external factor intervention can be seen in Montgomery and Weatherby (1980) who studied the impact of Arabian petroleum embargo to the electricity consumption in United States, Suhartono and Hariroh (2003) who investigate the impact of New York WTC bombing to the fluctuation of some world stock prices, and Suhartono (2007) who investigate Bali bombing effect to hotel occupancy rate in Bali.

2. INTERVENTION MODEL

An intervention model is a model which could be used to evaluate the impact of an intervention event that is caused by internal or external factors on a time series dataset (Suhartono, 2007). Generally, there are two common types of intervention, namely step and pulse functions. Detailed explanations and applications of intervention analysis can be found in Bowerman and O'Connell (1993), Hamilton (1994), Brockwell and Davis (1996), Tsay (2005), Wei (2006), and Suhartono (2007). An intervention model can be written as

$$Y_t = \frac{\omega_s(B)B^b}{\delta_r(B)} X_t + \frac{\theta_q(B)}{\phi_p(B)(1-B)^d} a_t, \tag{1}$$

where Y_t is a response variable at time *t* and X_t is a binary indicator variable that shows the existence of an intervention at time *t*. X_t can be step function S_t or pulse function P_t. Then, $\omega_s(B)$ and $\delta_r(B)$ are defined as

$$\omega_s(B) = \omega_0 - \omega_1 B - \omega_2 B^2 - \cdots - \omega_s B^s,$$

and

$$\delta_r(B) = 1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_r B^r.$$

Equation (1) shows that the magnitude and period of intervention effect is given by *b*, *s*, and *r*. The delay time is shown by *b*, while *s* gives information about the time which is needed for an effect of intervention to be stable, and *r* shows the pattern of an intervention effect. The impact of an intervention model on a time series dataset (Y_t^*) is

$$Y_t^* = Y_t - \frac{\theta_q(B)}{\phi_p(B)(1-B)^d} a_t = \frac{\omega_s(B)B^b}{\delta_r(B)} X_t \tag{2}$$

A step function is an intervention type which occurs over the long term. For example, Valadkhani and Layton (2004) applied a step function intervention for analyzing the impact of new tax system in Australia since September 2000. The intervention step function is written as (Wei, 2006)

$$X_t = S_t = \begin{cases} 0, & t < T \\ 1, & t \geq T, \end{cases} \tag{3}$$

where the intervention starts at *T*. A step function single input intervention model with *b*=2, *s*=1, and *r*=1 can be obtained by substituting Equation (3) into (1),

$$Y_t = \frac{(\omega_0 - \omega_1 B)B^2}{1-\delta_1} S_t + \frac{\theta_q(B)}{\phi_p(B)(1-B)^d} a_t. \tag{4}$$

Therefore, the effect of a step function single input intervention is

$$Y_t^* = \frac{(\omega_0 - \omega_1 B)B^2}{1-\delta_1} S_t. \tag{5}$$

If $|\delta_1| < 1$, we have

$$Y_t^* = \omega_0 S_{t-2} + (\omega_0\delta_1 - \omega_1)S_{t-3} + (\omega_0\delta_1^2 - \omega_1\delta_1)S_{t-4} + \cdots \tag{6}$$

The effect of an intervention's effect in Equation (6) can also be written as

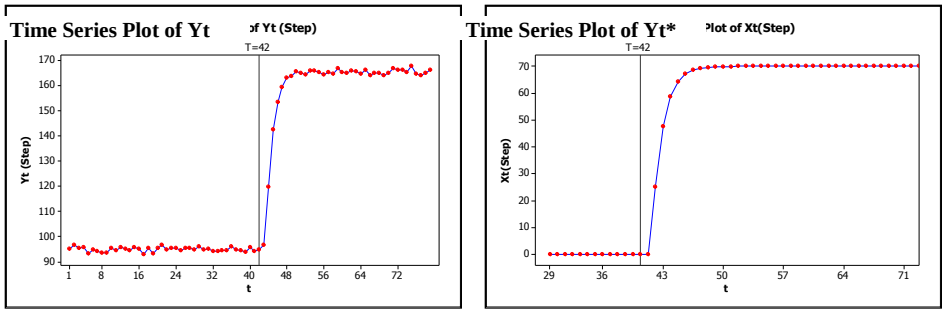
$$Y_t^* = \begin{cases} 0, & t < T+2 \\ \sum_{i=2}^{k} \omega_0\delta_1^{i-2} - \sum_{j=3}^{k} \omega_1\delta_1^{j-3}, & t = T+k,\ k \geq 2. \end{cases} \tag{7}$$

A simulation of this intervention, with $\omega_0 = 25$, $\omega_1 = -10$, $\delta_1 = 0.5$ occurring at $t = 42$ is drawn by Figure 1.

This intervention starts affecting Y_t two periods after intervention occurred $(b = 2)$, with a magnitude of 25. Three periods after intervention, the value of Y_t becomes 47.5 and reaches 64.4 in the fourth period. This increase becomes permanent effect and can be seen to extend at least as far as $t = 70$.

An intervention which occurs only at a certain time (*T*) is called pulse intervention. Examples of this intervention are public elections and the 11 September attacked in USA which affected the unemployment rate in USA (Dholakia, 2003). The pulse intervention function is written as

$$X_t = P_t = \begin{cases} 0, & t \neq T \\ 1, & t = T. \end{cases} \tag{8}$$



(a) (b)

Figure 1. (a) Simulation of an intervention model, (b) Intervention effect of Step Function Single Input $(b=2, s=1, r=1)$

An explanation of a single input intervention effect with pulse function can be done similarly to the step function intervention in Equation (4)-(7). A simulation of a pulse single input intervention model $(b=2, s=1, r=1)$ which the value of $\omega_0 = 25$, $\omega_1 = -10$, and $\delta_1 = 0.5$ is drawn in Figure 2. Figures 1 and 2 show the difference between step and pulse interventions and their effects. The effect of a step function is felt until $t > T$, where $T > 50$, while the pulse function has an impermanent effect, whereby for a certain *T*, the time series dataset will not affected by the intervention event.

A multi input intervention model, based on Equation (1), is (Wei, 2006)

$$Y_t = \frac{\omega_{s_1}(B)B^{b_1}}{\delta_{r_1}(B)} X_{1_t} + \frac{\omega_{s_2}(B)B^{b_2}}{\delta_{r_2}(B)} X_{2_t} + \cdots + \frac{\omega_{s_k}(B)B^{b_k}}{\delta_{r_k}(B)} X_{k_t} + \frac{\theta_q(B)}{\phi_p(B)(1-B)^d} a_t$$

which can be written

$$Y_t = \sum_{i=1}^{k} \frac{\omega_{s_i}(B)B^{b_i}}{\delta_{r_i}(B)} X_{i_t} + \frac{\theta_q(B)}{\phi_p(B)(1-B)^d} a_t. \qquad (9)$$

Equation (9) shows that there are *k* events affecting a time series dataset. For illustration, consider a multi input intervention with two events, pulse function (*b*=1, *s*=2, *r*=0) which is followed by step function $(b=1, s=1, r=1)$, i.e.

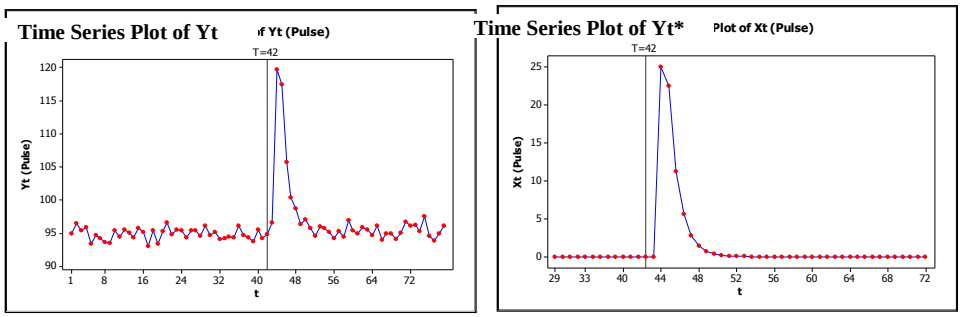
$$Y_t = [(\omega_{0_1} - \omega_{1_1} B + \omega_{2_1})B^1]P_t + \frac{(\omega_{0_2} - \omega_{1_2} B)B^1}{1-\delta_1(B)} S_t + \frac{\theta_q(B)}{\phi_p(B)(1-B)^d} a_t. \qquad (10)$$

The impact is

$$Y_t^* = \omega_{0_1} P_{t-1} - \omega_{1_1} P_{t-2} - \omega_{2_1} P_{t-3} + \omega_{0_2} S_{t-1} + (\omega_{0_2}\delta_1 - \omega_{1_2})S_{t-2} + (\omega_{0_2}\delta_1 - \omega_{1_2})\delta_1 S_{t-3} + \cdots \qquad (11)$$

which can also be written as

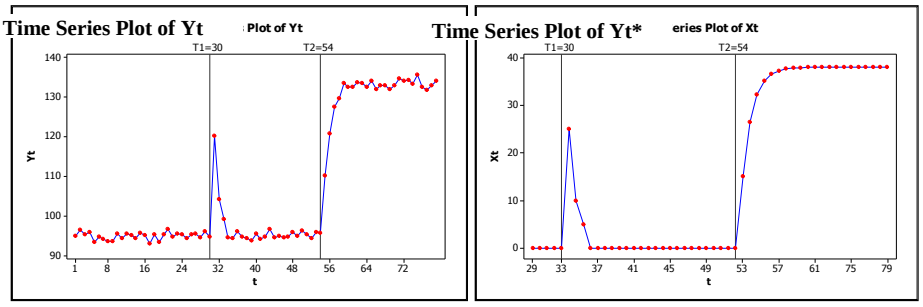
$$Y_t^* = \begin{cases} 0 & t \le T_1 \\ \omega_{0_1} & t = T_1 + 1 \\ -\omega_{1_1} & t = T_1 + 2 \\ -\omega_{2_1} & t = T_1 + 3 \\ 0 & t = T_1 + k,\ T_1 + k \le T_2 \\ \omega_{0_2} & t = T_2 + 1 \\ (\omega_{0_2} - \omega_{1_2})\left(\sum_{i=2}^{k} \delta_1^{i-2}\right) + \omega_{0_2}\delta_1^{k-1} & t \ge T_2 + m, m \ge 2. \end{cases}$$



(a) (b)

Figure 2. (a) Simulation of an intervention model, (b) Intervention effect of pulse Function Single Input $(b = 2, s = 1, r = 1)$

An illustration of Equation (10) and its impact are represented by Figure 3, for $\omega_{0_1} = 25, \omega_{1_1} = -10, \omega_{2_1} = -5, \omega_{0_2} = 15, \omega_{1_2} = -4,$ and $\delta_1 = 0.5.$ The first intervention occurs at $t = 30,$ with a magnitude of 25. The pulse function intervention has an effect that lasts for 4 periods beyond $t = T_1 = 30$ with the magnitude effects being 10 and 5 on the third and fourth after the intervention, respectively. The effect of this pulse intervention will be equal to zero. A second intervention begins at $t = T_2 = 54$. This step intervention is felt at $t = 55$ and its impact is 15. From $t = 56$ to $t = 59$ the impacts of this step intervention are 26.5, 32.25, 36.5, and 37.3, respectively. The impact doesn't increase beyond 38.



(a) (b)

Figure 3. (a) Simulation of an intervention model, (b) Intervention effect of Multi input intervention where Pulse Function $(b = 1, s = 2, r = 0)$ occurs at $t = 30$ and was followed by Step Function $(b = 1, s = 1, r = 1)$ at $t = 54$

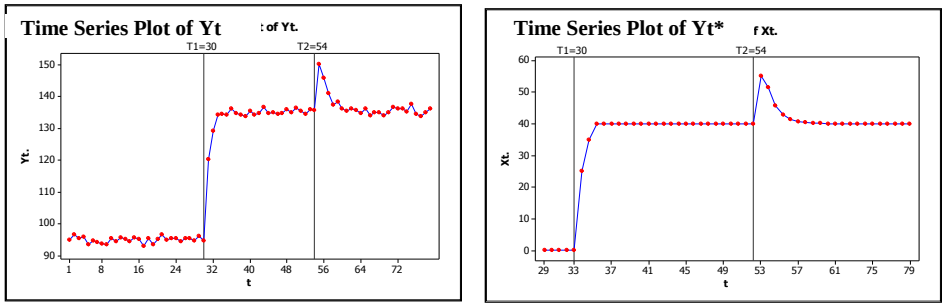
Now, we will show the other multi input intervention model, where the step function intervention $(b=1,s=2,r=0)$ is the first intervention and followed by a pulse function intervention $(b=1,s=1,r=1)$. The model is

$$Y_t=[(\omega_{0_1}-\omega_{1_1}B+\omega_{2_1})B^1]S_t+\frac{(\omega_{0_2}-\omega_{1_2}B)B^1}{1-\delta_1(B)}P_t+\frac{\theta_q(B)}{\phi_p(B)(1-B)^d}a_t,$$

and the impact is $Y_t^*=\omega_{0_1}S_{t-1}-\omega_{1_1}S_{t-2}-\omega_{2_1}S_{t-3}+\omega_{0_2}P_{t-1}+(\omega_{0_2}\delta_1-\omega_{1_2})P_{t-2}+(\omega_{0_2}\delta_1-\omega_{1_2})\delta_1P_{t-3}+\cdots$

The first intervention, namely the step function intervention, starts affecting the data at one period after the intervention event occurs, and its impact is ω_{0_1}. This impact will be $(\omega_{0_1}-\omega_{1_1})$ in the second period. From the third period until $t=T_2$, the impact is $(\omega_{0_1}-\omega_{1_1}-\omega_{2_1})$. One period after $t=T_2$, the second intervention, namely the pulse function intervention, gives additional impact to the time series dataset, ω_{0_2}. Therefore, the net impact will be $(\omega_{0_1}-\omega_{1_1}-\omega_{2_1}+\omega_{0_2})$. The second and third periods after $t=T_2$, the impacts are $(\omega_{0_1}-\omega_{1_1}-\omega_{2_1}+\omega_{0_2}\delta_1-\omega_{1_2})$ and $(\omega_{0_1}-\omega_{1_1}-\omega_{2_1}+\omega_{0_2}\delta_1^2-\omega_{1_2}\delta_1)$. Thereafter, the impact decreases gradually goes to zero. Consequently, the impact will be back to $(\omega_{0_1}-\omega_{1_1}-\omega_{2_1})$.

Figure 4 shows a simulation of a multi input intervention where the first intervention is the step function and the second intervention is the pulse function. The initial value for this simulation are $\omega_{0_1}=25,\omega_{1_1}=-10,\omega_{2_1}=-5,\omega_{0_2}=15,\omega_{1_2}=-4,$ and $\delta_1=0.5.$ The first intervention, which occurs at $t=T_1=30$, starts to affect the data on $t=31$, and the impact is 25. There is a rapid increase in the intervention effect (see Figure 4(b)) from $t=32$ to $t=35$, but the effect remain constant between $t=35$ and $t=54$. The second interventions occurs at $t=T_2=54$ and starts to affect the data at $t=55$. This effect becomes 0 from $t=59$ onwards.



(a) (b)

Figure 4. (a) Simulation of an intervention model, (b) Intervention effect of Multi input intervention where Step Function $(b=1,s=2,r=0)$ occurs at $t=30$ and was followed by Pulse Function $(b=1,s=1,r=1)$ at $t=54$

Let the intervention model be defined as

$$Y_t=\frac{\omega_s(B)}{\delta_r(B)}X_{t-b}+\frac{\theta_q(B)}{\phi_p(B)(1-B)^d}a_t. \qquad (12)$$

Equation (12) can be rewritten as

$$\delta_r(B)\phi_p(B)(1-B)^d Y_t = \omega_s(B)\phi_p(1-B)^d X_{t-b} + \delta_r(B)\theta_q(B)a_t, \quad (13)$$

or

$$c(B)Y_t = d(B)X_{t-b} + e(B)a_t$$

where

$$c(B) = \delta_r(B)\phi_p(B)(1-B)^d = (1-c_1B-c_2B^2-\cdots-c_{p+r}B^{p+r})(1-B)^d,$$
$$d(B) = \omega_s(B)\phi_p(1-B)^d = (d_0-d_1B-d_2B^2-\cdots-d_{p+s}B^{p+s})(1-B)^d,$$
$$e(B) = \delta_r(B)\theta_q(B) = 1-e_1B-e_2B^2-\cdots-e_{r+q}B^{r+q}.$$

Thus, we have

$$a_t = \frac{c(B)Y_t - d(B)X_{t-b}}{e(B)}. \quad (14)$$

The nonlinear least square estimation to estimate the unknown parameters can be found by minimizing

$$S(\delta,\omega,\phi,\theta \mid b) = \sum_{t=t_0}^{n} a_t^2, \quad (15)$$

where $t_0 = \max(p+r+1, b+p+s+1)$ and a_t are the residuals under the white noise assumption and Normal distribution. The parameters of the multi input intervention can be obtained by replacing Equation (12) with Equation (9) and following the same minimization procedure as Equation (13)-(15). As in Suhartono (2007), the intervention response or Y_t^* is easily formulate using the response values charts for determining the order of intervention model, i.e. *b*, *s*, and *r*. The intervention response which is denoted as Y_t^* is basically residual or error, i.e. the difference between actual data and ARIMA model forecasts from data before the intervention. The complete procedure of intervention model building which can be used to evaluate two step function interventions at time T_1 and T_2 (in this case, new fuel tariff and Lapindo mud flood) based on theoretic studies can be described as follows.

(1) **Dividing the dataset into 3 parts**,

- Data 1, which is the data before the first intervention, as many as n_0 time periods, i.e. $t = 1,2,\ldots,T_1-1$. Denoted as Y_{0_t}.
- Data 2, which is the data from the first intervention until just before the second intervention, as many as n_1 time periods, i.e. $t = T_1, T_1+1, T_1+2,\ldots,T_2-1$. Denoted as Y_{1_t}.
- Data 3, which is data from the second intervention until the end of data, as many as n_2 time periods, i.e. $t = T_2, T_2+1, T_2+2,\ldots,n$. Denoted as Y_{2_t}.

(2) **Modeling of the first intervention**

a. Step 1

- ARIMA model building for time series data before the first intervention occurs (Y_{0_t}), so we have

$$Y_{0_t} = \frac{\theta_q(B)}{\phi_p(B)(1-B)^d}a_t.$$

- Forecasting of Data 2 (Y_{1_t}) using the ARIMA model. In this step, we get the forecast data, i.e. $\hat{Y}_{T_1}, \hat{Y}_{T_1+1},\ldots,\hat{Y}_{T_1+n_1-1}$.

b. Step 2

- Calculate the response values of the first intervention or $Y_{1_t}^*$. These are the residuals of the data for $t = T_1, T_1+1, T_1+2, ..., T_2-1$, based on the forecasting of the ARIMA model in the first step. This step produces response values of the first intervention, i.e.

$$Y_{T_1}^*, Y_{T_1+1}^*, ..., Y_{T_2-1}^*.$$

- Determination of b_1, s_1, r_1 from the first intervention by using the plot of response values $Y_{T_1}^*, Y_{T_1+1}^*, ..., Y_{T_2-1}^*$ and a confidence interval of width, i.e. $\pm 3\hat{\sigma}_{a_0}$, where $\hat{\sigma}_{a_0}$ is Root Mean Square Error (MSE) of the previous ARIMA model. This interval is based on the determination of control chart bounds during statistical quality control for detecting outlier observations.

c. Step 3

- Parameter estimation and significance test for the first intervention model
- Diagnostic checking for examining the residual assumption, i.e. white noise and Normality distribution. In this step, we have the first input intervention model, i.e.

$$Y_t = \frac{\omega_{s_1}(B)B^{b_1}}{\delta_{r_1}(B)} X_{1_t} + \frac{\theta_q(B)}{\phi_p(B)(1-B)^d} a_t. \quad (16)$$

(3) **Modeling of the second intervention**

a. Step 1

- Forecasting of Data 3 (Y_{2_t}), based on the first intervention model. In this step, we obtain the forecasted values from the ARIMA model

$$\hat{Y}_{T_2}, \hat{Y}_{T_2+1}, ..., \hat{Y}_{T_2+n_2-1}.$$

b. Step 2

- Calculate the second intervention responses $(Y_{2_t}^*)$, i.e. residual of the data for $t = T_2, T_2+1, T_2+2, \ldots, n$, based on the forecasting of the first intervention model. These response values are denoted

$$Y_{T_2}^*, Y_{T_2+1}^*, \ldots, Y_n^*$$

- Identification of b_2, s_2, r_2 from the second intervention model from the plot of response values $Y_{T_2}^*, Y_{T_2+1}^*, \ldots, Y_n^*$, and the confidence interval of width $\pm 3\hat{\sigma}_{a_1}$.

c. Step 3

- Parameter estimation and significance test for the second intervention model
- Diagnostic checking for examining the residual assumption, i.e. white noise and Normality distribution. In this step, we have

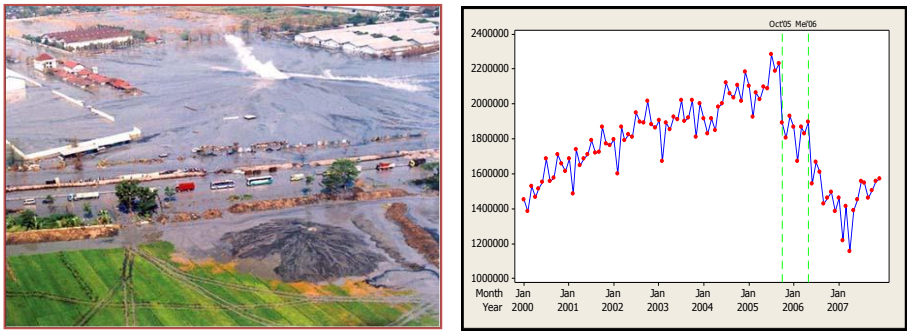
$$\sum_{j=1}^{2} \frac{\omega_{s_j}(B)B^{b_j}}{\delta_{r_j}(B)} X_{j_t} + \frac{\theta_q(B)}{\phi_p(B)(1-B)^d} a_t. \quad (17)$$

This procedure could be used iteratively for *k* interventions to build *k* multi input intervention models, i.e.

$$Y_t = \sum_{j=1}^{k} \frac{\omega_{s_j}(B)B^{b_j}}{\delta_{r_j}(B)} X_{j_t} + \frac{\theta_q(B)}{\phi_p(B)(1-B)^d} a_t.$$

3. THE DATA

As a sample case study, these techniques are illustrated via the analysis of the amount of monthly traffic on the Waru-Gempol toll road from January 2000 until December 2007. Hence, there is 96 observations. The time series plot of the data and photo of the intervention event are shown in Figure 5.



(a) (b)

Figure 5. (a) Photo of Lapindo mud flood, (b) Time series plot of the amount of monthly vehicles on the Waru-Gempoll toll road, January 2000 – December 2007.

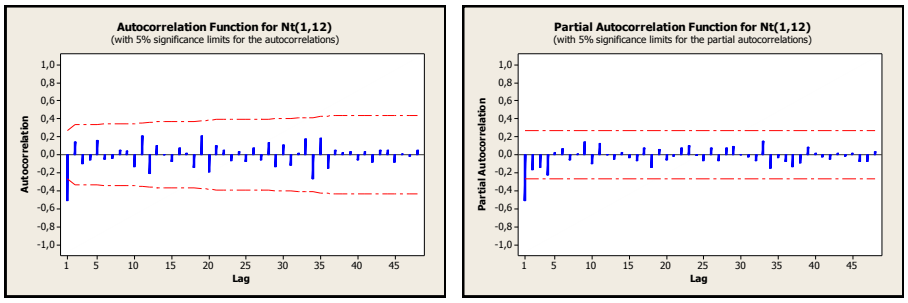
During this time period, there are two interventions which affect the amount of vehicles on the toll road. These interventions are the new fuel tariff applied since October 2005 ($t \geq 70$) and the Lapindo mud flood which occurred in May 2006 ($t \geq 77$). Both intervention variables are step function intervention variables.

4. RESULTS

The Box-Jenkins procedure (see Box et al., 1994) is utilized, including identification, parameter estimation, diagnostic checking, and forecasting to find the best ARIMA model before the first intervention, i.e. new fuel tariff since October 2005. The identification step shows that the data is stationary in variance, but not stationary in mean (that is, the data contain trend and seasonal pattern). Regular and seasonal order differencing is applied to get stationary data. Plot of ACF and PACF for the stationary data is shown in Figure 6.

Both ACF and PACF plots cut off after lag 1, so there are 2 possible ARIMA models, i.e. ARIMA$(0,1,1)(0,1,0)^{12}$ and ARIMA$(1,1,0)(0,1,0)^{12}$. The results of parameter estimation, parameter significance test, and diagnostic checking can be seen in Table 1. From Table 1, we know that both models are appropriate for forecasting the amount of vehicles before raising of the fuel tariff intervention. The comparison of mean square error (MSE) shows that ARIMA$(0,1,1)(0,1,0)^{12}$ model yields less MSE than ARIMA $(1,1,0)(0,1,0)^{12}$.

The results of the first step intervention modeling are calculated, namely the new fuel tariff since October 2005 which occurred at $t = T = 70$. The first step of modeling is to determine the order *b*, *s*, and *r* for the first step function intervention model. To determine the decrease due to new fuel tariff and to determine the first step function intervention model order, a residual chart is shown in Figure 7.



(a) (b)

Figure 6. (a) Plot of ACF and (b) PACF of stationary data after regular and seasonal differencing (d=1, dan D=1, S=12) before the first intervention.

Table 1. Results of parameter estimation, parameter significance test, and diagnostic checking for ARIMA$(0,1,1)(0,1,0)^{12}$ and ARIMA$(1,1,0)(0,1,0)^{12}$ models

ARIMA model	Parameter	Estimate	*P-value*	MSE
$(0,1,1)(0,1,0)^{12}$	$\hat{\theta}_1$	0.6305	< 0,0001	52296.39*
$(1,1,0)(0,1,0)^{12}$	ϕ_1	–0.5234	< 0,0001	5398704*

* = residual satisfies white noise and normal distribution assumptions

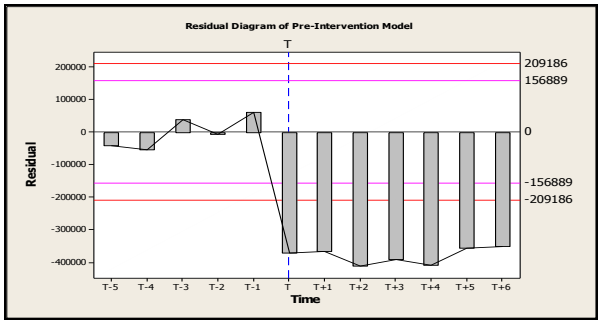


Figure 7. Chart of response values of the amount of vehicles after the first intervention and prior to the second intervention

Based on Figure 7, the appropriate order of the first step function intervention model is *b*=0, *s*=0, and *r*=0. It caused the residuals to remain constant, indicating a constant effect. The results of parameter estimation and significance test for the first intervention model show that all model parameters are significant (at the 5% significance level). Diagnostic checking of the model shows that the first step function intervention model has satisfied the assumptions of white noise and normally distributed residuals. The intervention model for the amount of vehicles on the toll road after the first step function intervention and prior to the second step function intervention can be written as

$$Y_t = -378189.9S1_t + \frac{(1-0.63307B)a_t}{(1-B)(1-B^{12})}. \quad (21)$$

After modeling the first intervention, i.e. the intervention model due to the new fuel tariff, analysis of the second step function intervention is conducted, namely the Lapindo mud flood since May 2006 which equates to $t = 77$. The first step is to determine the order of the second intervention model. Figure 8 shows a chart of residuals for determining the order *b*, *s*, and *r* for the intervention model, which will be used to model the decrease due to the Lapindo mud flood.

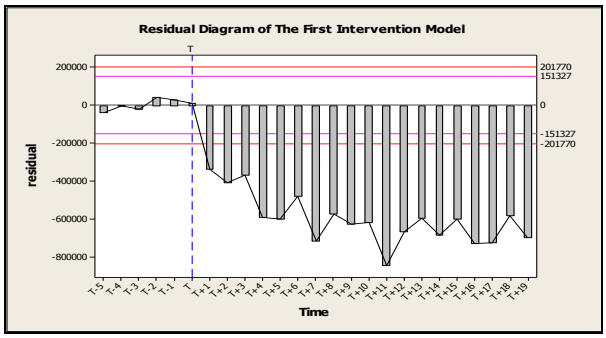


Figure 8. Chart of response values of the amount of vehicles after the second intervention

Based on Figure 8, there are two possible set orders of the second step function intervention model. The first set order is $b = 1$, $s = 1$, and $r = 1$ and the second is $b = 1$, $s = (3,6)$, and $r = 0$. Parameter estimation and significance tests for the first and the second set of model orders show that all parameters of the intervention model are significant. Diagnostic checking shows that only the second set of model order yields residuals that satisfy the residual assumptions. Thus, we use the second intervention model to explain the effects of the Lapindo mud flood on the decrease of amount of vehicles on the toll road, i.e.

$$Y_t = -381263.1S1_t + \frac{(-370316.1B + 233677.5B^2)}{(1 - 0.82344B)} S2_t + \frac{(1 - 0.73048B)}{(1 - B)(1 - B^{12})} a_t \qquad (22)$$

A residual plot from the second intervention model is shown in Figure 9. Based on Figure 9, there is no residual which is out of the interval ±4σ, although there is 1 residual out of the interval ±3σ (where σ is root of MSE). The next step is to evaluate the effects of the observation for which the residual is out of the ±3σ boudaries.

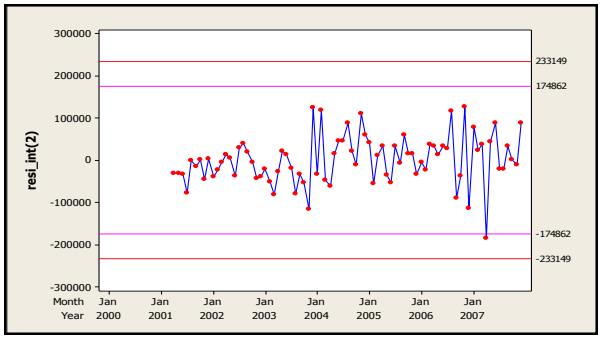


Figure 9. Time series plot of residuals from the second intervention model

We can explain the presence of the residual which is out of the ±3σ interval, i.e. observation for April 2007. The reason why the data at April 2007 is out of the ±3σ interval is because the mud flood embankment near the Porong highway was broken. Hence, the mud flows to the Porong highway and caused the Porong highway to be closed for several days. This is an outlier event that can be categorized as a pulse function intervention. To determine the decrease due to the broken embankment, a pulse function intervention model with order $b=0$, $s=0$, and $r=0$ is employed. Parameter estimation and a significance test for the pulse function intervention addition show that all parameters are significant. A diagnostic check shows that residual has satisfied the white noise and normal distribution assumptions.

In the previous section we showed that the predicted order of the second intervention model, i.e. $b=1$, $s=(3,6)$, and $r=0$, is the best model for explaining the effects of the second step function intervention. Therefore, the best multi input intervention model is a model with order $b=0$, $s=0$, and $r=0$ for explaining the impact of the first step function intervention ($S1_t$), the new fuel tariff, a model with order $b=1$, $s=(3,6)$, and $r=0$ for explaining the effect of the second step function intervention ($S2_t$), Lapindo mud flood, and a model with order $b=0$, $s=0$, and $r=0$ for explaining the impact of the pulse function intervention (P_t) in April 2007, and ARIMA(1,1,0)(0,1,0)^{12} as the noise model. Mathematically, the best intervention model is written as

$$Y_t = -388512\,S1_t + (-387180.8B - 166.275B^4 - 126029.3B^7)\,S2_t - 205549P_t + \frac{a_t}{(1+0.56475B)(1-B)(1-B^{12})}. \tag{23}$$

The calculation of the first step function intervention effect for the new fuel tariff, i.e. at period *T*, *T*+1, *T*+2, until *T*+*k*, where $k=8$ (October 2005, November 2005, December 2005, …, April 2006) is

$$Y^*_{T+k} = -388512\,S_{T+k} = -388512 \tag{24}$$

This implies that the new fuel tariff caused the decrease of the monthly amount of vehicles passing through the Waru-Gempol toll road by as many as 388512 vehicles. This decrease occurred after the new fuel tariff (October 2005) before the Lapindo mud flood happened in April 2006. Based on the best intervention model in Equation (23), the calculation of the Lapindo mud flood effect on the amount of vehicles is as follows.

- **Effect at period *t* = *T* (May 2006)**

  The amount of the second intervention effect at $t=T$ is

$$\begin{aligned} Y^*_t &= -388512\;S1_t + (-387180.8B - 166275B^4 - 126029.3B^7)S2_t \\ &= -388512\,S1_T - 387180.8S2_{T-1} - 166275\,S2_{T-4} - 126029.3S2_{T-7} \\ &= -388512 \end{aligned} \tag{25}$$

  Equation (25) implies that the Lapindo mud flood has not given any effect during the first month of the mud blast in May 2006. The explanation of this condition is because the mud blast occurred at the end of the month, i.e. on May 29, 2006. In this month, the decrease of 388512 vehicles is caused by the first step function intervention, i.e. the new fuel tariff.

- **Effects at periods $t = T+1$, $T+2$, and $T+3$ (June, July, and August 2006)**

  The calculation of the amount of the second intervention effects on these 3 periods is

$$\begin{aligned} Y_{T+k}^{*} &= -388512\,S1_{T+k} - 387180.8\,S2_{T+k-1} - 166275\,S2_{T+k-4} - 126029\,S2_{T+k-7} \\ &= -775692.8 \text{ where } k = 1,2,3. \end{aligned} \tag{26}$$

  The result in Equation (26) shows that the Lapindo mud flood has reduced the amount of vehicles traveling on the Waru-Gempol toll road as many as 387181 vehicles every month for 3 following months (June, July, and August 2006). This is because the Lapindo mud blast is near to the Porong toll road so that road users tend to pick alternative routes in order to reach Malang or Surabaya. Overall, the decrease of vehicle numbers due to the new fuel tariff and Lapindo mud flood is as many as 775693 vehicles in those months.

- **Effects at periods $t = T+4$, $T+5$, and $T+6$ (September, October, November 2006)**

  The amount of the second intervention effects at $t = T+4, T+5,$ and $T+6$ is

$$\begin{aligned} Y_{T+k}^{*} &= -388512\,S1_{T+k} - 387180.8\,S2_{T+k-1} - 166275\,S2_{T+k-4} - 126029\,S2_{T+k-7} \\ &= -941967.8 \text{ where } k = 4,5,6. \end{aligned} \tag{27}$$

  These calculations imply that the Lapindo mud flood has reduced the vehicle numbers as many as 553456 vehicles every month for the 4th, 5th, and 6th months following to the month of the Lapindo mud blast (that is, September, October, and November 2006). The Lapindo mud flow started to reach and swamp the Porong toll road so that a single direction traffic system had to be applied to this toll road. This event and the previous new fuel tariff event made the decrease as many as 941968 vehicles in those months.

- **Effects at periods $t = T+7$ until $T+k$, where k=10 until the last observation i.e. December 2007 (December 2006, January 2007, ... , December 2007)**

  Calculation of the second intervention effects in these periods is

$$\begin{aligned} Y_{T+k}^{*} &= -388512\,S1_{T+k} - 387180.8\,S2_{T+k-1} - 166275\,S2_{T+k-4} - 126029\,S2_{T+k-7} \\ &= -1067996.8 \text{ where } k = 7,8,9. \end{aligned} \tag{28}$$

  Equation (28) shows that the Lapindo mud flood reduced the number of vehicles in the Waru-Gempol toll road as many as 679485 vehicles in the 7th, 8th, and 9th months following to the month of the Lapindo mud blast (that is, December 2006, January and February 2007). This is because the permanent closure of the Porong-Gempol toll road since November 2006 led to the amount of vehicles passing through the Waru-Gempol toll road to be reduced. The reduction of vehicle numbers due to new fuel tariff and Lapindo mud flood in this period is as many as 1067997 vehicles.

Based on the intervention model given in Equation (23), the calculation of the broken Lapindo mud embankment effect in April 2007 is

$$\begin{aligned} Y_{T}^{*} &= -388512\,S1_{T} - 387180.8\,S2_{T-1} - 166275\,S2_{T-4} - 126029\,S2_{T-7} - 205549\,P_{t} \\ &= -1067996.8 - 205549 \\ &= -1273545.8 \text{ where } T = 88 \text{ (April 2007).} \end{aligned}$$

This calculation shows that the broken mud embankment in April 2007 has reduced the number of vehicles passing through the Waru-Gempol toll road to 205549 vehicles. This is because the condition of road in Porong was so poor that it could not be used by vehicles. Therefore, road users chose alternatives routes to Malang or Surabaya.

The reconstruction of the overall intervention model effects, including the first intervention $S1_t$ and the second intervention $S2_t$, and also the pulse function intervention P_t, can be seen in Figure 10. This graph shows that the theoretical effect reconstruction for the intervention model yields prediction data which is close to the actual data.

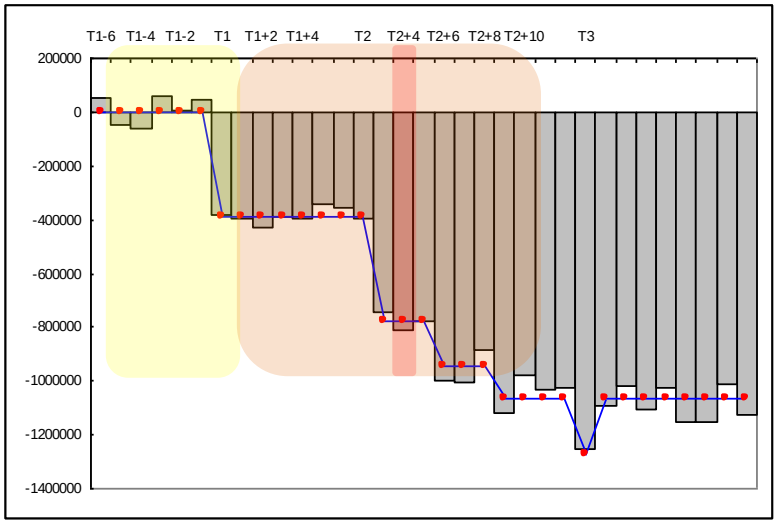


Figure 10. Theoretical effect reconstruction for the multi input intervention model

5. COMMENTS AND CONCLUSION

The appropriate intervention model for the number of vehicles in the Waru-Gempol toll road due to the new fuel tariff and Lapindo mud flood is an intervention model with order $b=0$, $s=0$, and $r=0$ for explaining the impact of the first step function intervention $(S1_t)$, namely the new fuel tariff, and an intervention model with order $b=1$, $s=[3,6]$, and $r=0$ for explaining the impact of the Lapindo mud flood as the second step function intervention $(S2_t)$, and a model with order $b=0$, $s=0$ and $r=0$ for explaining the effect of the pulse function intervention (P_t) in April 2007, with ARIMA $(1,1,0)(0,1,0)^{12}$ as the noise model, which can be mathematically written as

$$Y_t = -388512\,S1_T - 387180.8\,S2_{T-1} - 166275\,S2_{T-4} - 126029.3\,S2_{T-7} - 205549\,P_t + \frac{a_t}{(1+0.56475B)(1-B)(1-B^{12})}.$$

Calculation of the amount and the period of the effects for each intervention are as follows:

a. The application of the new fuel tariff has caused a decrease of vehicles in the Waru-Gempol toll road as many as 388512 vehicles every month since the first step function intervention occurred (October 2005) until the last observation (December 2007).

b. Generally, the Lapindo mud flood has caused a decrease of vehicles in the Waru-Gempol toll road. There are 3 phases of different reductions, i.e.
   (i). Decrease as many as 387181 vehicles on period 1-3 months since the mud blast (June – August 2006).
   (ii). Decrease as many as 553456 vehicles on period 4-6 months after the mud blast (September – November 2006).
   (iii). Decrease as many as 679485 vehicles since December 2006 until the last observation (December 2007).
c. The Broken Lapindo mud embankment in April 2007 has caused a decrease in the number of vehicles passing through the Waru-Gempol toll road by as many as 205549 vehicles.

This research shows that the proposed procedure for multi input intervention model building has been applied well at certain real data. Further research opportunities include the application of this multi input intervention model to other real data sets to further validate this new procedure.

ACKNOWLEDGMENT

This research was supported by Research Management Centre, Universiti Teknologi Malaysia (UTM), Malaysia and Institut Teknologi Sepuluh Nopember, Indonesia.

REFERENCES

Bowerman, B.L. and O'Connell, R.T. (1993). *Forecasting and Time Series: An Applied Approach, 3rd edition*. Belmont, California: Duxbury Press.

Box, G.E.P., Jenkins, G.M. and Reissel, G.C. (1994). *Time Series Analysis Forecasting and Control, 3rd edition*. Englewood Cliffs : Prentice Hall.

Box, G.E.P. and Tiao, G.C. (1975). Intervention Analysis with Applications to Economic and Environmental Problems. *Journal of the American Statistical Association*, 70, 70-79.

Brockwell, P.J. and Davis, R.A. (1996). *Introduction to Time Series and Forecasting*. Spinger-Verlag, New York.

Dholakia, K. (2003). What Has Affected The Unemployment Rates in The USA: Preliminary Analysis of The Last 12 Years–Elections and 9/11. *Midwestern Business and Economic Review*, 32, 34-43.

Hamilton, J.D. (1994). *Time Series Analysis*. Princeton University Press, New Jersey.

Leonard, M. (2001). *Promotional Analysis and Forecasting for Demand Planning: A Practical Time Series Approach*. Cary, NC, USA: SAS Institute Inc.

McSweeny, A.J. (1978). The Effects of Response Cost on the Behavior of a Million Persons: Charging for Directory Assistance in Cincinnati. *Journal of Applied Behavioral Analysis*, 11, 47-51.

Montgomery, D.C. and Weatherby, G. (1980). Modeling and Forecasting Time Series Using Transfer Function and Intervention Methods, *AIIE Transactions*, 289-307.

Parwitasari, D. (2006). *Modeling Transportation CPI in Surabaya which Contains of Structural Change.* Unpublished Bachelor Final Project, Department of Statistics, Institut Teknologi Sepuluh Nopember.

Suhartono dan Hariroh, E. (2003). Analyzing the impact of WTC New York bombing to the world exchange rate by using intervention model. *Proceeding of National Seminar on Mathematics and Statistics*. ITS Surabaya and Alumni PPS Matematika UGM.

Suhartono dan Wahyuni, W. (2002). Analysis of the promotion and price raising effects to the number of customer and pulse consumption at PT. Telkom Divre V. *Journal of Forum Statistika dan Komputasi*. Special Edition for National Seminar on Statistics, IPB, Bogor.

Suhartono. (2007). Theory and application of pulse function intervention model. *Jurnal Ilmiah MatStat*, 7(2), 191-214.

Tsay, R.S. (2002). *Analysis of Financial Time Series*. John Wiley & Sons, Inc., New Jersey.

Valadkhani, A. and Layton, A.P. (2004). Quantifying the effect of the GST on inflation in Australia's capital cities: an intervention analysis. *Australian Economic Review*, 37(2), 125-138.

Wei, W.W.S. (2006). *Time Series Analysis: Univariate and Multivariate Methods.* 2nd Edition California: Addison-Wesley Publishing Company, Inc.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 861–869.

A STRATEGIC FRAMEWORK FOR GAINING ECONOMIC LEADERSHIP IN THE NEW ECONOMY: THE PERSPECTIVE OF OIC MEMBER STATES

Suleman Aziz Lodhi [1], Abdul Majid Makki [2] and Munir Ahmed [3]
School of Business Administration
National College of Business Administration and Economics, Lahore, Pakistan
E-mail: [1] sulemanlodhi@yahoo.com, [2] Abdul7896@yahoo.com.au, [3] drmunir@brain.net.pk

ABSTRACT

Economic planners associated with the classical school of thought may be a little hesitant to recognize the changes taking place in the business and economic environment due to the emergence of the New Economy. Globalization of businesses and the resultant emergence of new business entities in the form of franchises, networked organizations, distributed organizations and EDI (Electronic Data Interchange) integration has completely changed the way trade is being conducted. The novel paradigm of the New Economy which is also referred to as the knowledge economy is much more complex than its predecessor system. Many of the critical factors controlling the dynamics in this system belong to an intangible domain and cannot be directly controlled by the governments .It has become very difficult to plan successful FDI (Foreign Direct Investment) , FTA (Free Trade Agreements) or any other integrated commerce activity without bringing into consideration the impact of intangibles affecting the regional economic system. We present a literature survey to provide a critical review of emerging tools for the strategic management and policy analysis domains. A conceptual framework is developed that after further in-depth study of the particular factors involved in a given country can be used as a strategic management tool by policy makers to take strategic decisions and visualize the nation's development policies in the knowledge economy. It is argued that if a nation is able to manage and improve its intangible assets, its economic value generation capacity would increase and result in the nation moving towards prosperity.

Keywords: Policy development framework, Economic Policy, Knowledge Economy, New Economy, Intangible Asset Monitor, Strategic Management

1. INTRODUCTION

The roots of the English word "wealth", which is used to denote economic status, originated from the old English word "Weal"; which was used as an adjective to describe the ownership of great qualities. It is unfortunate that present usage of word wealth is narrowly applied to represent monetary value alone. Adam (1776) while investigating the causes of the wealth of nations, states that the wealth of a nation consists in the well-being of the mass of ordinary citizens, arguing later that the best contribution governments can make to the wealth of their nations and to the progress of human society is to leave individuals free to follow their natural tendency to make exchanges. Criticizing the mercantile system of public policy, Adam (1776) comments that it was once thought that the wealth of a nation existed in money (gold and silver) and therefore the governments worked to make their countries wealthy by restricting the export

of gold and silver. It was perceived that since money can be translated easily and quickly into military power, countries must have high gold and silver available to buy arms and soldiers in time of war. However, the merchant class had a different opinion. They viewed this as inconvenient, because they needed to send gold and silver abroad for business activities. The business community persuaded governments to view success of a country on the bases of "the balance of trade", which is the relation between imports and exports of a country. The business community argued that what makes a country rich is when its exports are higher than its imports. The governments responded by making policies that encouraged local production and exports while making policies to hinder imports, what we call "protectionism" in modern terms.

Though it has been more than two hundred years, economists started to acknowledge the potential role of manufacturing in the economics of a society(UNDP, 2003). But with the rapid developments in IT and communication sciences, the world has entered a new phase - an era in which the wealth of nations is dependent on its ability to create, transform and capitalize knowledge. The era of knowledge-based industries have arrived, where employee know-how, innovative capabilities, and skills are the brainpower of an organization. These new factors are playing a predominant role in the productive power of the corporation.

Human resources now account for an increasing proportion of the capital generation in industries (Sveiby 1997). Empirical studies suggest that a major percent of the value created by a firm comes, not from management of traditional physical assets, but rather from the management of its intangible assets (Prusak 2001, Sveiby 2002). It is argued that the science and technology sectors are expanding faster than most of the other industries. This rapidly increasing demand for knowledge-based products and services is changing the structure of the global economy and transforming the economic infrastructures of many countries, including Islamic countries like Pakistan (Kalim and Lodhi 2002, 2004). There is a general consensus that in the new economic paradigm.3 factors have become evident

1) The wealth of a nation is no longer limited to its natural resources. Traditional national assets like oil, minerals, agricultural and manufactured products are now complemented with a new, nontraditional category of resources in the form of intangible asset, which include law and order situation in a country, quality and level of education, health services provided by government etc.

2) Knowledge is a primary competitive factor in modern economy.

3) The accumulation, transformation, and value creation from knowledge requires active and voluntary participation of intellectuals. Therefore, countries should adopt strategies to maximize an environment of collaboration.

This discussion of the new economic paradigm and “wealth” cannot be concluded without mentioning the opposite of “wealth”, which is “poverty”. The new paradigm has redefined the concept of poverty also. Conventionally governments reports Per Capita GDP as an indicator of economic wellbeing in a country. But many (for example, Saunders 2002) have argued that measuring poverty in financial terms is too narrow a view. Poverty means much more than just lack of financial earning and must be viewed in a holistic manner. The population in a region is not just poor; this is usually also correlated with low education, health care facilities, corruption and general unrest. It is argued that simply providing the poor with additional income would not solve anything. The socio-economic system in a society must be developed to improve the

performance of a community and make it sustainable in the region. Peter (SPRC 2004) provides a collection of alternative definitions of poverty arguing that not having opportunities for social achievement such as taking part in the community-life, having equal opportunity for intellectual development etc, health care etc, should be considered as social dimensions of poverty. This provides researchers with the prospect of using social indicators for measuring wealth.

In the light of the broader definitions of poverty UNDP (2007) reports the Human Development Index (HDI) of the countries in the world. Table 1 gives the HDI of selected countries for the year 2005. Angola and Tanzania have Human Development Index scores of 0.446 and 0.467 respectively, but their GDP per capita are significantly different. It can be seen that the population living in Tanzania earn much less than the population of Angola, but have much better living conditions than that of Angola. Thus, HDI presents the social aspect of poverty in a much better form than GDP alone . The GDP can only present the financial dimension of wealth and therefore should not be viewed or reported in isolation.

Table 1 Human Development Index and GDP Per Capita Selected Country Data for the Year 2005

Country	HDI	GDP Per Capita (PPP US$)
Iceland	0.968	36,510
Czech Republic	0.891	20,538
United Arab Emirates	0.868	25,514
Chile	0.867	12,027
Bahrain	0.866	21,482
Egypt	0.708	4,337
Pakistan	0.551	2,370
Bhutan	0.579	1,969
Tanzania	0.467	744
Angola	0.446	2,335
Sierra Leone	0.336	806

Source: UNDP Human Development Report 2007

Inequalities are not limited to income, although Gini coefficient and income-share inequality numbers provide a fair picture of ground realities, other inequalities are also important. For instance, the quality of basic services like health, education and rule of law provided to the majority population can provide a picture of inequalities in practice. The poor members of society may have a greater need for health services as they are at a higher risk of getting sick due to lack of clean drinking water and sanitation conditions. Similarly, poor population may have a higher mortality rate due to lack of medical care and low nutrition, but in practice they may not be getting their due share of these services from the governments. Education can play an important role in improving quality of life in a region, since as the level of education in a society is increased it becomes more productive. It also plays the role of catalyst in improving effectiveness of health awareness programmes. The condition of the law and order system of a

society and its practice plays a vital role in development of a society. If a law treats poor and rich differently, the society cannot develop.

All these societal characteristics form part of the intangible assets of a nation. These intangible assets may be very difficult to measure can be viewed as the true national wealth. If a nation is able to improve in these areas, its value creation capacity will increase. Viewing this from a systems perspectives, as these sub-processes improve the output of the system as a whole (nation) will improve, resulting in more value addition.

2. FRAMEWORK FOR MEASURING ABSOLUTE WEALTH OF NATIONS

Researchers have applied a number of frameworks for measuring these intangible assets or wealth of the nation. Skandia Navigator is used by some researchers (Edvinsson and Malone 1997) for measuring and developing national policy initiatives. Later, Rembe (1999) investigated the intangible assets of Sweden with the objective of increasing foreign investment in the country. This study used metrics to develop a strategic plan for the future directions in Sweden. Rembe (1999) defines a human capital index based on quality of life, life expectation, education, etc., a market capital index based on tourism, service balance, etc., a process capital index based on management quality, information and communication technologies, etc., and renewal capital index based on research and development, ratio of young to old people, etc. Further studies were initiated for evaluating the Intellectual Capital of Sweden and Organization for Economic Cooperation and Development O.E.C.D countries (O.E.C.D 2000),. Malaysia (Bontis et al. 2000), and the Arab Region (Bontis, 2002, Bontis 2004). . Extending these concepts Malhotra (2003) discussed methods for measuring intangible assets of nations extensively. He modified the Balance Scorecard (Kaplan and Norton 1992) to suggest a framework for measuring and managing knowledge assets of a nation. Bontis (2001, 2004) also discussed methods for measuring intangible assets.

After a careful review of methods for measuring tangible and intangible assets, we selected Intangible Asset Monitor (IAM) of Sveiby 1997 as base model, and extended it for assessing the wealth of nations. Figure 1 gives a suggested framework for viewing the wealth of nations in the new economy. The total wealth of a country may be viewed as sum of its tangible and intangible assets. The tangible assets of a nation would include traditional resources of nations like forests, mineral resources, oil, factories, roads network, communication infrastructure, agricultural and other material products. All assets that are tangible are viewed under this category. But as seen above the wealth of a nation is not limited to tangible assets only; intangible assets are also important. Intangible assets are further classed into three categories; namely (1) External Structure (2) Internal Structure and (3) Competence.

The External Structure includes a nation's external relations, its trade agreements like FTAs (Free Trade Agreement), RTAs (Regional Trade Agreement), defense pacts, participation in international bodies, the image of a country in the international community etc. The value of these assets would be illustrated in the form of influence the country is able to exercise internationally and solve its global issues. The Internal Structure of a country that adds to a national wealth includes all factors that may promote collaboration, knowledge sharing and innovation in a country. Therefore the internal structure of a country would include the functioning of the justice system, democracy, education system, health system, right to information and freedom to form associations, basic human rights etc.

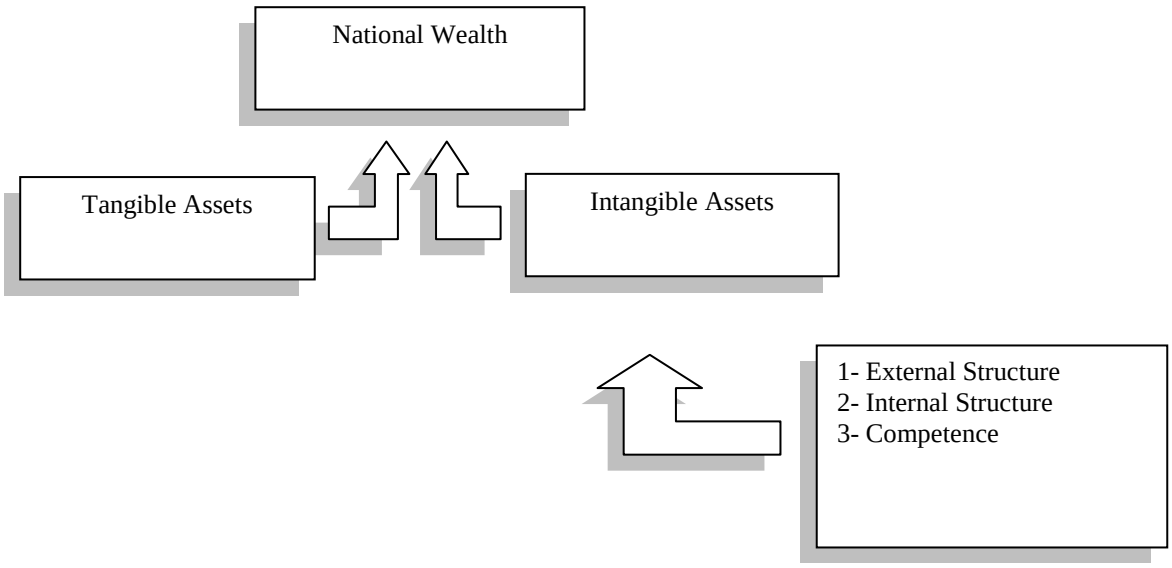


Figure 1 Framework for Measuring Wealth of Nations (Extended from IAM Sveiby 1997)

The third category of Intangible assets is "Competence", which would include factors such as the literacy rate in a country, the quality of research activity taking place, the effectiveness of professional associations, etc. The list of three categories for intangible assets is not limited and can be expanded. We therefore argue that the wealth of the nations should not be measured through financial resources alone, but should also include intangibles such as level of community services, maturity of processes in government and business (good governance), intellectual freedom etc. If a nation scores low on these intangibles, it will not be able to perform well in the knowledge economy. Presently there is no standard scale available to measure the intangible wealth of nations. Research in this critical domain is needed to develop indices for measuring and comparing the intangible assets of i countries. In order to explain this idea further, a three-dimensional plot between Human Development Index, Corruption Perception Index (CPI) and GDP per Capita is shown in figure 3.

The HDI is used to indicate the living standard of general population in a country which can also be seen as a measure of good governance. . CPI scores from Transparency International (2008) are used as an indicator for measuring maturity of processes. A country scoring low on CPI would mean that its processes and controls are not mature enough to prevent corruption. The purpose of this plot is to show a wider picture in which nations are not compared on a financial basis alone. It can be seen that a number of OIC (Organization of The Islamic Countries) countries are found with low CPI score, meaning that there is a critical need to improve business and management systems in these countries. The HDI in most OIC countries is also found to be low, an indication that the standard of living in these countries is not high. Referring to Table 1, it can be seen that UAE has an HDI score of 0.868 with GDP per Capita of approximately US $ 25,500 which is less than the HDI score of the Czech Republic which is 0.891. The Czech Republic is able to achieve this HDI score on GDP per Capita of approximately US $ 20,500, which is lower that of UAE. Similarly the comparison of Bahrain and Chile show that Chile is able to maintain the same level of governance but with a lower GDP than Bahrain.

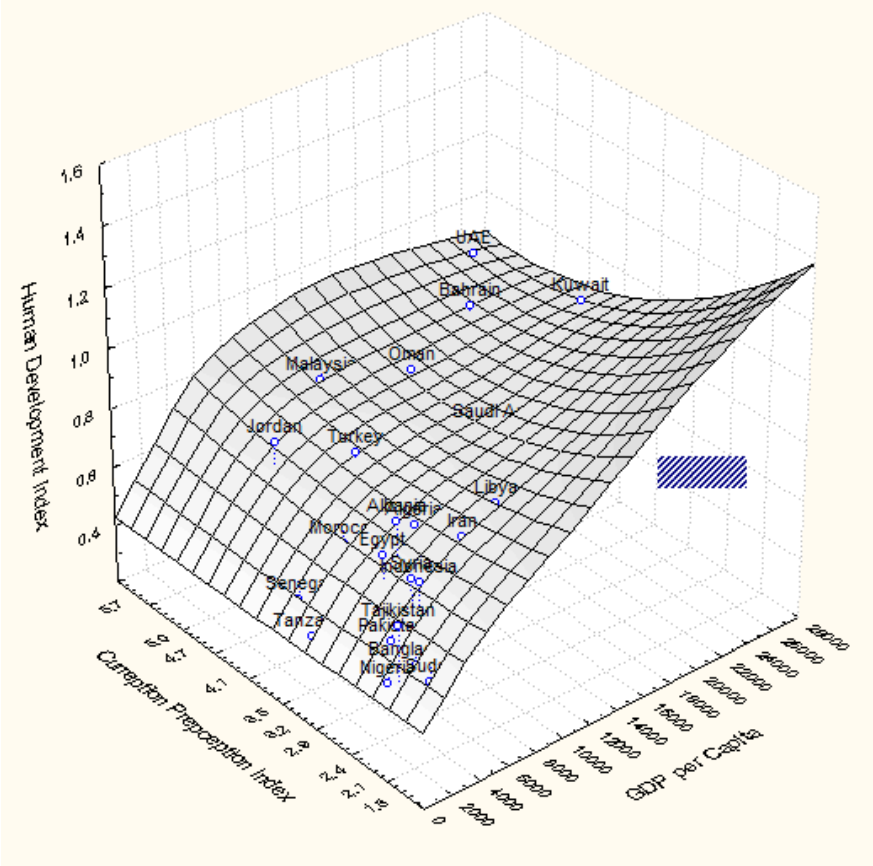


Figure 2 3D Plot between HDI, CPI and GDP per Capita Selected OIC Countries

3. DEVELOPING ECONOMIC POLICIES FOR THE NEW ECONOMY

Development economists, policy makers and social scientists are well aware of the interdependent nature of national policies. Initiatives taken to improve literacy under education policy are beneficial not only for health related initiatives, but they also strengthen economic activities in a region. Similarly, the foreign policy of a country cannot work in isolation; it will have a corresponding effect on the country's trade policy.

This framework for viewing national assets can be used for designing development policies of a nation in the New- Economy. Keeping in view the complexity and highly interdependence of national policies, a policy matrix may be developed; instead of following the traditional approach of developing a single policy for a specific purpose. The policy makers may view the assets of a nation using the framework and accordingly , develop policies to build up the tangible as well as intangible assets of a nation. Once a nation starts to progress by building its intangible assets, the human development index (HDI) of the country would also start to improve, meaning that quality of life in the region would increase. This would have a positive outcome on external structure, internal structure and competence of a country, finally resulting in an increase in business opportunities in the region.

A policy matrix developed in the light of the extended IAM for nations is given as figure 3. This figure presents only the framework, the details are intentionally left out for the purpose of simplicity. Developing guidelines for an economic policy using the policy matrix is beyond the

scope of this paper. Since this paper is limited to developing a framework and does not intend to develop indicators. A detailed study will be needed before deciding on indicators, as the policies must be aligned with strategic directions in which the nation intends to move.

In practice the policy makers would first have to develop indicators for assessing the status of each category of intangible asset of a country and then in the light of these indicators, the policy makers may propose initiatives for building these assets. Following this policy matrix frame work will also ensure that the policies developed should include efforts to renew, improve utilization and minimize risk of losing these national assets.

		Tangible Assets	Intangible Assets		
			External Structure	Internal Structure	Competence
Strategic Imperatives and Policy Priority	Grow Volume				
	Innovate/ Renew				
	Utilize Efficiently				
	Minimize Risk				

Figure 3 A Policy Matrix for Enhancing National Assets (Extended from IAM Sveiby 1997)

4. CONCLUSION

The wealth of a nation in the New-Economy is integrated with the complex socio-economic structure of the society, and cannot be measured by financial dimensions alone. It is more closely related with the quality of life of the people living in a region and includes much more than simply the earning capacity of the population. Low income in a region is the result of low performance of its social systems. In other words, low income is the effect of poverty and not a cause itself. Therefore, policies aim to increase financial earning alone would not be expected to be successful in raising a region out of poverty. The framework presented can be used to view areas in which the country has potential, but in which this potential is underutilized or to identify areas that may be of critical importance for the growth of the economy.

Indicators for measuring the present status and growth of intangible assets are developed and then taking these indicators as basis; policies for improving the intangible assets can be developed and placed in policy matrix framework. The improvement in the intangible assets due to policy initiatives will be reflected by an increase in the indicator value in successive year's status report. Policy initiatives which may be linked with improving intangible assets may be continued or improved, whereas initiatives which may be a cause of decrease in indicator value may be abandoned. The framework helps in understand the interdependence of Government policies within different sectors and ensure that the overall development policy is not in contradiction with any of its lower level (sector specific) policies.

Table 2: World Bank Development Indicators Selected OIC Countries

Country	GDP Per Capita (PPP US$) 2005	Global Competi-tiveness Index (GCI) (2008-09)	HDI 2005	CPI score 2007	MDG Net primary enrolment rate 2005	MDG Population Undernour-ished 2004	MDG Population using an improved water source 2004	MDG Popula-tion below income poverty line
Malaysia	10882	5.04	0.811	5.1	95	3	99	15.5
Qatar	27664	4.83	0.875	6.0	96	NA	100	NA
Saudi Arabia	15711	4.72	0.812	3.4	78	4	90	NA
UAE	25514	4.68	0.868	5.7	71	2.5	100	NA
Kuwait	26321	4.58	0.891	4.3	87	5	NA	NA
Bahrain	21482	4.57	0.866	5.0	87	NA	NA	NA
Oman	15602	4.55	0.814	4.7	76	NA	80	NA
Brunei	28161	4.54	0.894	NA	93	4	NA	NA
Jordan	5530	4.37	0.773	4.7	89	6	97	14.2
Indonesia	3843	4.25	0.728	2.3	96	6	77	27.1
Turkey	8407	4.15	0.775	4.1	89	3	96	27
Morocco	4555	4.08	0.646	3.5	86	6	81	19
Syria	3808	3.99	0.724	2.4	95	4	93	NA
Egypt	4337	3.98	0.708	2.9	94	4	98	16.7
Libya	10355	3.85	0.818	2.5	NA	2.5	71	NA
Nigeria	1128	3.81	0.470	2.2	68	9	48	34.1
Senegal	1792	3.73	0.499	3.6	69	20	76	33.4
Algeria	7062	3.71	0.733	3.0	97	4	85	22.6
Pakistan	2370	3.65	0.551	2.4	68	24	91	32.6
Albania	5316	3.55	0.801	2.9	94	6	96	NA
Bangladesh	2053	3.51	0.547	2.0	94	30	74	49.8
Tanzania	744	3.49	0.467	3.2	91	44	62	35.7
Tajikistan	1356	3.46	0.673	2.1	97	10	79	NA
Iran	7968	NA	0.759	2.5	95	4	94	NA
Sudan	2083	NA	0.526	1.8	43	26	70	NA

Legend: HDI = Human Development Index GCI = Competitiveness Index
CPI = Corruption Perceptions Index MDG = Millennium Development Goals

Policies should be designed to improve the social system as a whole. This is not something that can be attained in a short time or with a little effort. A continuous focus and growth in the right direction is needed for development of a sustainable socio-economic system.

REFERENCES

Adam, Smith (1776). "An Inquiry into the Nature and Causes of the Wealth of Nations". Adam Smith Institute London: 2001. pp. 10-11.

Bontis, Nick, Chua, W. and Richardson, S. (2000) "Intellectual Capital and the Nature of Business in Malaysia", Journal of Intellectual Capital, 1, 1, 85-100.

Bontis, Nick (2001). Assessing knowledge assets: A review of the models used to measure intellectual capital, *International Journal of Management Reviews*, Vol. 3: 1, 41-60.

Bontis, Nick. (2002): National Intellectual Capital Index: Intellectual Capital Development in the Arab Region, United Nations office for Project Services, New York.

Bontis, Nick (2004). National Intellectual Capital Index: A United Nations initiative for the Arab region, *Journal of Intellectual Capital*, Vol. 5:1, 13-39.

Edvinsson, Leif and Malone, Michael (1997). *Intellectual Capital - Realizing your company's true value by finding its hidden brainpower*, Harper Collins New York.

Kalim, Rukhsana and Lodhi, Suleman (2002). The Knowledge Based Economy: Trends and Implications for Pakistan, The Pakistan Development Review. 41:4. pp. 787-804. Available at www.pide.org.pk/PSDE/Papers.html

Kalim, Rukhsana and Lodhi, Suleman (2004). Digital Economy and its impact on Employment in the Developing Countries" The ICFAI Journal of Applied Economics – Institute of Chartered Financial Analysts of India, ICFAI Press.

Kaplan, Robert and Norton, David (1992). The balanced scorecard - measures that drive performance, *Harvard Business Review*, 70, 1, 71-9.

Malhotra, Yogesh (2003). *Measuring National Knowledge Assets: Conceptual Framework and Analytical Review*, United Nations Department of Economic and Social Affairs, Ad Hoc Expert Group Meeting on Knowledge Systems for Development. New York, 4-5 September 2003.

O.E.C.D (2000), "Knowledge Management in the Learning Society". OECD Publications Service. France.

Prusak, Laurence (2001). Where did knowledge management come from? IBM *Systems Journal*, Vol. 40, No 4. 1002-1007.

Rembe, A. (1999) Invest in Sweden: Report 1999, Halls Offset AB: Stockholm, Sweden.

Saunders, Peter (2002). *The Ends and Means of Welfare*. Coping with Economic and Social Change in Australia, Cambridge University Press, Melbourne.

SPRC (2004). *Social Policy Research Towards a Credible Poverty Framework*: From Income Poverty to Deprivation. 1-18.

Sveiby, Karl (1997). *The New Organizational Wealth: Managing and Measuring Knowledge Based Assets*. San Francisco: Berrett-Koehler.

Sveiby, Karl. and Roland, Simons (2002).Collaborative climate and effectiveness of knowledge work - an empirical study. *Journal of Knowledge Management*, Vol. 6: 5, 420-433.

Transparency International (2008). Global Corruption Report 2008. Cambridge University Press, The Edinburgh Building, Cambridge CB2 8RU, UK.

UNDP (2003). United Nations Development Programme. *Poverty Reduction and Human Rights*. A Practice Note, 2-14.

UNDP (2006). United Nations Development Programme, Human Development Report 2006, 1-24.

UNDP (2007). United Nations Development Programme, Human Development Report 2007.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 870–880.

STATISTICAL INFERENCE ON THE TYPE-II EXTREME VALUE DISTRIBUTION BASED ON THE KERNEL APPROACH

M. Maswadah
Department of Mathematics, Faculty of Science,
South Valley University, Aswan, Egypt.
E-mail: maswadah@hotmail.com

ABSTRACT

In this paper, a practical approach based on the adaptive kernel density estimation (AKDE) has been applied for deriving some characteristics for the confidence intervals (CIs) of the Type-II Extreme Value distribution parameters. The proposed approach utilized the non-parametric AKDE as a tool for estimating the density function of a pivotal that depend on the unknown parameter and thus the characteristics for the confidence intervals can be studied. The efficiency of this technique has been studied comparing to the conditional inference on the basis of the mean lengths, the covering percentages and the standard error for the covering percentage of the confidence intervals, via Monte Carlo simulations and some real data. From our results it appears that the kernel approach competes and outperforms the conditional approach and has a number of appealing features, it can perform quite well and attains a good level of accuracy even when the number of bootstraps is extremely small. Finally, a numerical example is given to illustrate the densities based on the inferential methods developed in this paper.

Keywords: Adaptive kernel density estimation; Conditional inference; Covering percentage; The standard error of the covering percentage .

1. INTRODUCTION

In statistical inference, the classical approach offers a consistent way for using the pivotal quantities in inference since Student's paper (1908) that derived exact treatment for the mean of a normal sample and since that time numerous exact solutions have become available. In general, it is not possible to evaluate probability points for the pivotal analytically, but unconditional probability points can be computed, via extensive Monte Carlo simulations, see Thomas et al. (1969). This work introduces a new unconditional approach in statistical inference for estimating the density function of a pivotal, via the non-parametric (AKDE), which is asymptotically converges to any density function depending only on a random sample, though the underlying distribution is not known. This approach has been applied recently on some distributions, see Maswadah (2006, 2007, 2009). As a continuation for these efforts, in this paper the statistical analysis of the proposed procedure has been studied and its performances compared, via Mote Carlo simulations and some real data, to the performances of the classical conditional inference when the experimental data are collected under complete samples from the Type-II Extreme Value distribution (EVD) or (Fre'chet distribution), which has probability density function given by:

$$f(x) = \alpha\beta^{\alpha} x^{-(\alpha+1)} exp(-(\beta/x)^{\alpha}), \quad x > 0, \tag{1.1}$$

where ($\alpha > 0$) and ($\beta > 0$) are the shape and scale parameters respectively. This distribution is a commonly applied as skewed distribution, however the traditional applications involving analyzing natural catastrophes such as wind gusts, drought, rainfall, flood, etc. and recently the most important application of the EVD is in molecular biology where the DNA protein sequences are aligned with those in a database. Thus for its importance, it has been studied by several authors, see Gumbel (1965) derived the estimations for the parameters based on the methods of moments, the reciprocal moments and the maximum likelihood, Harter and Moore (1968) derived the conditional maximum likelihood estimates (MLEs) for singly censored samples, Singh (1987) derived the MLEs based on the *m-th* extremes of several samples, Singh et al. (1989) derived the MLEs based on the joint distribution of largest *m* extremes of several samples, Eldusoky et al. (2003) derived the Bayesian estimation of the parameters and reliability based on complete and censored samples and Maswadah (2005) derived the conditional confidence intervals for the parameters based on censored generalized order statistics, for a detailed discussion on various properties and uses of this distribution, see Johnson et al. (1995).

2. KERNEL ESTIMATIONS

In this section, the basic elements associated with the adaptive kernel estimators of the density function are presented, which has been extensively studied see, for example Guillamon et al. (1998, 1999). Also a good discussion on kernel estimation techniques can be found in Scott (1992).

In the univariate case, the adaptive kernel density estimation based on a random sample of size n from the random variable X with unknown probability density function $f(x)$ and support on $(0,\infty)$ is given by

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i}K\left(\frac{x-x_i}{h_i}\right), \tag{2.1}$$

where $h_i = h\lambda_i$ and λ_i is a local bandwidth factor which narrows the bandwidth near the modes and widens it in the tails and can be defined as:

$$\lambda_i = \left(\frac{G}{\hat{f}(x_i)}\right)^{0.5}, \tag{2.2}$$

where G is the geometric mean of the $\hat{f}(x_i)$, $i = 1,2,...,n$ and h is a fixed (pilot) bandwidth. We can see that our estimate $\hat{f}(x)$ is bin-independent regardless of our choice of K, where the role of K is to spread out the contribution of each data point in our estimate of the parent distribution; that is, controls the shape. Though there are variety of kernel functions with different properties have been used in the literature, but an obvious and natural choice of K is the Gaussian kernel for its continuity, differentiability and locality properties.

The most important part in the kernel estimation method is to select the bandwidth (scaling) or the smoothing parameter. Its selection has been studied by many authors, see Abramson (1982), Terrell (1990) and Jones (1991), based on minimizing the mean square errors, thus the optimal choice in most cases is $h = 1.059 \cdot S \cdot n^{-0.2}$, where S is the sample standard deviations and we will considered it as the pilot bandwidth. However, it must be mentioned that the optimal choice

for h can not possibly be optimal in every application and its choice is really depend on the application under consideration, and in some situations it might be quite useful to have a set of estimators corresponding to different bandwidths.

To utilize the kernel function for estimating the probability density function (pdf) of a pivotal, we can summarize the method in the following algorithm:

1- Let X_i, $i = 1,2,...,n$ be a random sample of size *n* from the random variable X, whose pdf is $f(x;\theta)$, where θ represents the unknown parameter with support on $(0,\infty)$.
2- Bootstrapping with replacement *n* samples X_i^*, $i = 1,2,...,n$ of size *n* from the parent sample in step 1.
3- For each sample in step 2 calculate the pivotal quantity Z based on the parameter θ and its consistent estimator such as the MLE. Thus we have an objective and informative random sample from the pivotal quantities Z_i, $i = 1,2,...,n$ of size n, which constitute the informative sample for the pivotal Z.
4- Finally, based on the informative sample in step 3 we can use the AKDE for estimating $g(z)$ at any given value for Z and thus the confidence interval of the pivotal can be constructed and converted to the unknown parameter Fiducially.

Utilizing the above algorithm, the AKDE of the quantile z_p of order p, for Z can be derived as:

$$\hat{G}(z_p) = \int_{-\infty}^{z_p} \hat{g}(z)dz = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{z_p} \frac{1}{h_i} K(\frac{z - z_i}{h_i})dz = p \tag{2.3}$$

Thus

$$\sum_{i=1}^{n} \Pi(\frac{z_p - z_i}{h_i}) = np, \tag{2.4}$$

where

$$\Pi(x) = \int_{-\infty}^{x} K(y)dy. \tag{2.5}$$

For deriving the value of the quantile estimator z_p, equation (2.4) can be solved recurrently as the limit of the sequence $\{\tilde{z}_1, \tilde{z}_2, \tilde{z}_3, ...\}$ that defined by the formulas

$$\tilde{z}_1 = \tfrac{1}{n}\sum_{i=1}^{n} z_i, \tag{2.6}$$

$$\tilde{z}_{r+1} = \tilde{z}_r + C\left[np - \sum_{i=1}^{n} \Pi\left(\frac{\tilde{z}_r - z_i}{h_i}\right)\right], \qquad r = 1,2,3,.... \tag{2.7}$$

The convergence of (2.7) is guaranteed by the condition $0 < C \le \frac{2h_i}{nL_1}$, where $L_1 = K(0)$, see Kulczycki (1999).

3. CONDITIONAL INFERENCE

In this section we outlined the key ideas for constructing the confidence intervals of the extreme value parameters based on the conditional inference. For more details about this method see Lawless (1972, 1982) who applied this approach on some lifetime distributions. Let $Z_1 = \alpha/\hat{\alpha}$ and $Z_2 = (\hat{\beta}/\beta)^{\hat{\alpha}}$ be pivotal quantities for the parameters α and β, depending on their MLEs $\hat{\alpha}$ and $\hat{\beta}$ respectively, and define $a_i = (x_i/\hat{\beta})^{\hat{\alpha}}$, for $i = 1,2,....,n$ as the ancillary statistics. Make the change of variables from ($x_1, x_2, x_3, ..., x_n$), whose pdf is (1.1) to ($\hat{\alpha}, \hat{\beta}, A$), where $A = (a_1, a_2 ..., a_{n-2})$. This transformation can be written as

$$x_i = \hat{\beta} a_i^{1/\hat{\alpha}},\ i = 1,2,....,n-2.,\ x_{n-1} = \hat{\beta} a_{n-1}^{1/\hat{\alpha}} \text{ and } x_n = \hat{\beta} a_n^{1/\hat{\alpha}},$$

where a_n and a_{n-1} can be expressed in terms of A. The Jacobin of this transformation is independent of Z_1 and Z_2. Making further the change of variables from ($\hat{\alpha}, \hat{\beta}, A$) to ($Z_1, Z_2, A$), the Jacobin of this transformation is proportional to $1/Z_1 Z_2$. Finally the conditional pdf of Z_1 and Z_2 given A can be derived as:

$$f(z_1, z_2 \mid A) = C z_1^{n-1} z_2^{-nz_1-1} \prod_{i=1}^{n} a_i^{-z_1} \exp(-z_2^{-z_1} \sum_{i=1}^{n} a_i^{-z_1}), \tag{3.1}$$

where C is the normalizing constant independent of Z_1 and Z_2. The marginal densities of Z_1 and Z_2 conditional on A can be derived respectively as:

$$g_1(z_1 \mid A) = C\Gamma(n) z_1^{n-2} \prod_{i=1}^{n} a_i^{-z_1} (\sum_{i=1}^{n} a_i^{-z_1})^{-n}, \tag{3.2}$$

$$g_2(z_2 \mid A) = C \int_0^{\infty} z_1^{n-1} z_2^{-nz_1-1} \prod_{i=1}^{n} a_i^{-z_1} \exp(-z_2^{-z_1} \sum_{i=1}^{n} a_i^{-z_1}) dz_1, \tag{3.3}$$

where

$$C^{-1} = \Gamma(n) \int_0^{\infty} z_1^{n-2} \prod_{i=1}^{n} a_i^{-z_1} (\sum_{i=1}^{n} a_i^{-z_1})^{-n} dz_1. \tag{3.4}$$

Using equations (3.2) and (3.3) , we can find the desired probabilities for Z_1 and Z_2 and thus the confidence intervals for the parameters α and β can be derived Fiducially.

It is clear that both procedures depend on the MLEs of the parameters α and β based on the complete sample, which can be derived for the underlying distribution by taking the derivatives of the log likelihood function of (1.1) with respect to α and β respectively, and setting equal to zero yielding the two equations:

$$\frac{n}{\alpha}-\sum_{i=1}^{n}\ln(x_i)+\frac{n\sum_{i=1}^{n}x_i^{-\alpha}\ln(x_i)}{\sum_{i=1}^{n}x_i^{-\alpha}}=0 \tag{3.5}$$

$$\beta=\left(\frac{1}{n}\sum_{i=1}^{n}x_i^{-\alpha}\right)^{-1/\alpha}. \tag{3.6}$$

Thus using an iterative technique such as Newton-Raphson method for solving (3.5), we can derive the MLE for α and thus for β from (3.6).

4. SIMULATION STUDY AND COMPARISONS

The confidence intervals have become familiar in applied statistics, they combine point estimation and hypothesis testing into a single inferential statement of great intuitive appeal. Thus the characteristics of these intervals for the unknown parameters have been studied, via Monte Carlo simulations, based on the two approaches to measure their performances in terms of the following criteria:

1- The Covering percentage(*CP*), which is defined as the fraction of times the *CI* covers the true value of the parameter in repeated sampling. Thus if the *CP* is greater than (less than) the nominal level then the procedure is conservative (anti-conservative).
2- The mean length of the intervals (*MLI*), which is defined as the average length of the intervals in repeated sampling. If a short interval has high *CP* , the data allows us to estimate the parameter accurately. Though, higher *CP* generally requires a longer interval and short interval generally have lower *CP* . Therefore the procedures with the same *CPs* , the one that provides shorter interval is better.
3- The standard error of the covering percentage (*SDE*), which is defined for the nominal level $(1-\alpha)100\%$ by $SDE(\hat{\alpha})=\sqrt{\frac{\hat{\alpha}(1-\hat{\alpha})}{M}}$, where $(1-\hat{\alpha})100\%$ denote the corresponding Monte Carlo estimate and M is the number of Mote Carlo trials. Thus for the nominal level 95% and 1000 simulation trials, say, the standard error of the covering percentage is 0.0049, which is approximately $\pm 1\%$. Therefore, we say the procedure is adequate if the SDE is within $\pm 2\%$ error for the nominal level 95%.

The comparative results, based on 1000 Monte Carlo simulation trials are given for sample sizes $n = 20,\ 40,\ 60,\ 80$ and 100 which have been generated from the EVD for shape parameter values $\alpha = 0.5$, 1, 2 and 3 . The scale parameter β was set to 2 throughout, where all estimations are equivariant under scale chages of the data, so setting one value for β involves no loss of generality. The confidence intervals for the pivotals and the corresponding parameters are derived and their characteristics *CP* s, *MLI* s and *SDE* s are calculated and discussed in the following main points:

1- The results in Table 1 indicated that, the values of the *CP*, *MLI* and *SDE* for Z_1 have the same values for the different values of α and also they have the same values for Z_2 as expected because they are independent from the parameters α and β. Thus we have written their values for the sample size $n = 40$ only.

2- The results in Table 2 indicated that, as the sample size increases, the two approaches have values of *MLIs* and *CPs* getting decrease and the values of *SDEs* getting increase for Z_1 and Z_2 for all values of α .

3- The results in Table 2 indicated that, the values of the *MLI* based on the kernel approach are less than those based on the conditional approach. However the values of the *CP* are greater and thus the values of *SDE* are lesser than those based on the conditional approach.

4- The results in Table 3 indicated that, The values of *MLI* decrease as the sample size increases for both parameters α and β , however the values of *MLI* for α increase with the same average of increasing α as expected. Also the values of *MLI* for β decrease with increasing the values of the shape parameter α.

5- The kernel approach is conservative for estimating the parameters α and β because the covering percentages are much greater than the nominal level than the ones based on the conditional inference for all sample sizes. On the contrary the conditional approach is anti-conservative for estimating α and almost conservative for estimating β, when the sample size greater than 20.

6- Finally, both the two procedures are adequate because the *SDEs* are less than $\pm 2\%$ for the nominal level 95%.

Thus the simulation results indicated that the kernel intervals possess good statistical properties and they can perform quite well with reasonable accurate results even when the number of bootstrapping are extremly small.

Table (1) : The Kernel and conditional (MLIs), (CPs) and (SDEs) based on the nominal level 95% and n =40 for the Pivotals Z_1 and Z_2.

Pivotals	α	Kernel			Conditional		
		MLI	CP	SDE	MLI	CP	SDE
Z_1	0.5	0.4681	0.950	0.0069	0.4839	0.941	0.0075
	1.0	0.4681	0.950	0.0069	0.4839	0.941	0.0075
	2.0	0.4681	0.950	0.0069	0.4839	0.941	0.0075
	3.0	0.4681	0.950	0.0069	0.4839	0.941	0.0075
Z_2	0.5	0.7143	0.976	0.0048	0.6969	0.939	0.0076
	1.0	0.7143	0.976	0.0048	0.6969	0.939	0.0076
	2.0	0.7143	0.976	0.0048	0.6969	0.939	0.0076
	3.0	0.7143	0.976	0.0048	0.6969	0.939	0.0076

Table (2) : The Kernel and conditional (MLIs), (CPs) and (SDEs) based on the nominal level 95% for the Pivotals Z_1 and Z_2.

Pivotals	N	Kernel			Conditional		
		MLI	CP	SDE	MLI	CP	SDE
Z_1	20	0.6710	0.982	0.0042	0.6365	0.937	0.0077
	40	0.4681	0.950	0.0069	0.4839	0.941	0.0075
	60	0.3859	0.953	0.0067	0.3963	0.953	0.0067
	80	0.3309	0.953	0.0067	0.3449	0.949	0.0069
	100	0.2973	0.951	0.0068	0.3064	0.946	0.0071
Z_2	20	1.2450	0.980	0.0048	1.0637	0.945	0.0072
	40	0.7143	0.976	0.0048	0.6969	0.939	0.0076
	60	0.5551	0.951	0.0068	0.5562	0.963	0.0059
	80	0.4684	0.957	0.0067	0.4763	0.947	0.0071
	100	0.4133	0.953	0.0076	0.4233	0.955	0.0066

Table (3) : The Kernel and conditional (MLIs) based on the nominal level 95% and the values of $\alpha = 0.5,1,2,3.$ for the parameters α and β.

Parameter	N	Kernel				Conditional			
		$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 2.0$	$\alpha = 3.0$	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 2.0$	$\alpha = 3.0$
α	20	0.3746	0.7493	1.4985	2.2478	0.3435	0.6870	1.3740	2.0609
	40	0.2515	0.5029	1.0058	1.5087	0.2540	0.5081	1.0161	1.5242
	60	0.1983	0.3967	0.7934	1.0175	0.2021	0.4042	0.8084	1.2126
	80	0.1696	0.3392	0.6783	1.0175	0.1750	0.3501	0.7001	1.0502
	100	0.1521	0.3041	0.6082	0.9124	0.1553	0.3107	0.6214	0.9321
β	20	6.0234	2.6578	1.1379	0.7342	5.9361	2.2287	1.0243	0.6697
	40	3.2447	1.4139	0.6797	0.4491	3.2116	1.4096	0.6781	0.4480
	60	2.4388	1.1120	0.5405	0.3571	2.4529	1.1316	0.5523	0.3662
	80	1.9726	0.9297	0.4569	0.3035	2.0620	0.9672	0.4744	0.3149
	100	1.7072	0.8243	0.4064	0.2700	1.7718	0.8493	0.4197	0.2786

5. NUMERICAL EXAMPLE

In this section, via studying some real data, we will measure the performance of the bandwidth that selected in the kernel approach to see how well it performs in practice and how well the conclusion based on the real data will be consistent with the conclusion in the simulations. Thus considering the data given by Dumonceaux and Antle (1973), that represents the maximum flood levels (in millions of cubic feet per second) of the Susquehenna River at Harrisburg, Pennsylvenia over 20 four-year periods (1890-1969) as:

0.654, 0.613, 0.315, 0.449, 0.297, 0.402, 0.379, 0.423, 0.379, 0.324,
0.269, 0.740, 0.418, 0.412, 0.494, 0.416, 0.338, 0.392, 0.484, 0.265.

The *MLEs* for the parameters α and β based on this data are given respectively as 4.3143 and 0.3583. Thus for the purpose of comparisons, the 95% probability intervals for the pivotals Z_1 and Z_2 and their corresponding parameters α and β are derived based on the kernel and the conditional approaches. The results in Table 4 have been indicated that the length of intervals for the pivotals Z_1 and Z_2 based on the kernel approach are shorter than those based on the conditional inference and thus the conclusions are the same for their corresponding parameters α and β, and they contain their MLEs for the two approaches , which ensure the simulation results. Finally, in Figure 1 the probability densities of Z_1 based on the kernel and the conditional inferences are plotted in quite close symmetric shape, however in Figure 2 the probability densities of Z_2 based on the conditional approach is right skewed than the ones based on the kernel approach which ensure the results in Table 4 and the simulation results.

Table (4) : The Kernel and conditional Upper limit (UL), Lower limit (LL) and length of the intervals of the pivotals Z_1 and Z_2 and the corresponding parameters α and β respectively for the nominal level 95%.

Parameters	Kernel			Conditional		
	LL	UL	Length	LL	UL	Lenth
$Z1$	0.6401	1.2609	0.6209	0.6654	1.3841	0.7187
α	3.1313	6.1685	3.0373	2.8708	5.9713	3.1005
$Z2$	0.4727	1.2901	0.8174	0.5988	1.7173	1.1175
β	0.3355	0.4119	0.0764	0.3161	0.4034	0.0873

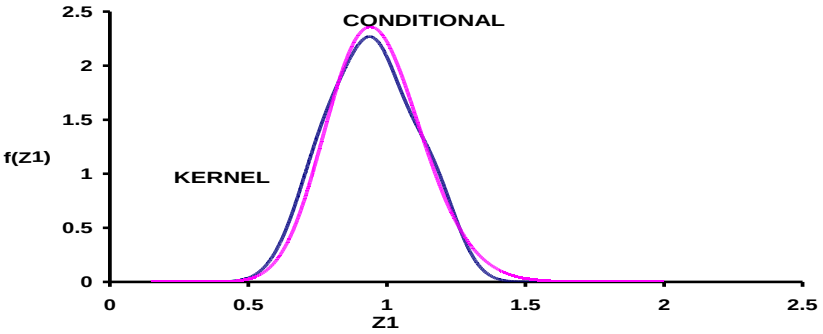


Fig. 1: THE PDF OF THE PIVOTAL Z1 BASED ON THE KERNEL AND THE CONDITIONAL INFERENCES

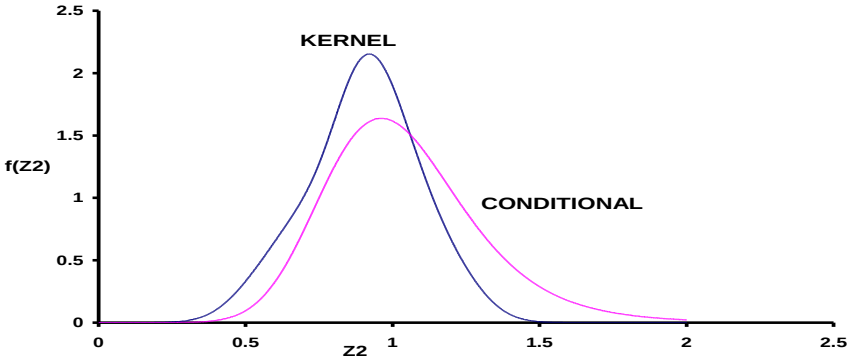


Fig. 2: THE PDF OF THE PIVOTAL Z2 BASED ON THE KERNEL AND THE CONDITINAL INFERENCES

6. CONCLUSION

The kernel estimation technique constitutes a strong basis for statistical inference and it has a number of benefits relative to the usual conditional procedure. First, it is easy to be used and it doesn't need tedious work as the conditional inference. Second, it can perform quite well even when the number of bootstraps is extremely small up to 20 replications. Finally, it is uniquely determined on the basis of the information content in the pivotal quantities and thus we can considering it as an alternative and reliable technique for estimation especially for problems with unknown parameters for which no sufficient statistics exist. Thus, from the results of this paper the kernel inference strengthens traditional inference statements and allows construction of alternative stronger types of inferences than the conditional inference.

REFERENCES

Abramson, I. (1982). On Bandwidth Variation in Kernel Estimates: A Square Root Law. Ann. Statist.10, 1217-1223.

Dumonceaux, R. and Antle, C. E. (1973). Discrimination between the Lognormal and Weibull

Distribution. Technometrics 15, 923-926.

Efron, B. and Tibshirani, R. J. (1993). An introduction to the bootstrap Monographs on statistics and Applied probability . NO. 57, Chapman and Hall, London.

Eldusoky, B., Maswadah, M. & El- Moasry, A. (2003). Bayesian Estimation and Prediction Based on Complete and Censored Samples from the Extreme Value Distribution, M. Sc. Dissertation. South Valley University.

Guillamon, A. Navarro, J. and Ruiz, J. M. (1998). Kernel density estimation using weighted data. Commun. Statist.-Theory Meth., 27(9), 2123-2135.

Guillamon, A. Navarro, J. and Ruiz, J. M. (1999). A note on kernel estimators for positive valued random variables. Sankhya: The Indian Journal of Statistics. Vol.(6) series (A) 276-281.

Gumbel, E. J. (1965). A quick estimation of the parameters of Frechet distribution. Rev. Int. Stat. Inst. No. 33(3), 349-363.

Harter, H. L. and Moore, A. H. (1968). Conditional Maximum likelihood estimation, from singly censored samples of the scale parameter of type-II extreme value distributions. Technometrics, No. 10, 349-359.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995). Continuous Univariate Distributions-II. 2-nd edition. John Wiley & Sons.

Jones, M. C. (1991). Kernel density estimation for length biased data. Biometerika 78, 511-519.

Kotz, S. and Johnson, N. L. (1992). Breakthroughs in statistics, Vol.I,II, New York: Springer-Verlag.

Kulczycki, P. (1999). Parameter Identification Using Bayes and Kernel Approaches. Proceedings of the National Science Council ROC(A), Vol.(23) No. 2, 205-213.

Lawless, J. F. (1972). Confidence interval estimation for the parameters of the Weibull distribution. Utilitas Mathematica Vol.2. , 71-87.

Lawless, J. F. (1973). On the estimation of safe life when the underling life distribution is Weibull. Technometrics, 15(4), 857-865.

Lawless, J. F. (1982). Statistical Models and Methods for Lifetime Data, John Wiley and Sons, New York.

Maswadah, M. (2005). Conditional Confidence Interval Estimation For The Type-II Extreme Value Distribution Based on Censored Generalized Order Statistics. Journal of Applied Statistical Science. Vol.14, No. 1/2, 71-84.

Maswadah, M. (2006). Kernel Inference On The Inverse Weibull Distribution". The Korean Communications in Statistics. Vol.13, No. 3, 503-512.

Maswadah, M. (2007). Kernel Inference On the Weibull Distribution. Proc. Third National Statistical Conference, Lahore, Pakistan. May 28-29, Vol.14, 77-86.

Maswadah, M. (2009). The Kernel and conditional Inferences On the Weibull Distribution parameters. Statistics (In press).

Scott, D. W. (1992). Multivariate Density Estimation. New York; Wiley inter-science.

Singh, N.P. (1987). Maximum Likelihood estimation of Fréchet distribution parameters. Journal of statistical studies. No.7, 11-22.

Singh, N. P., Singh, K.P. and Singh U. (1989). Estimation of distribution parameters by joint distributions of m extremes. Statistica 49.

Student. (1908).The probable Error of a Mean. Biometrika. 6, 1-25.

Terrell, G. R. (1990). The maximal smoothing principle in density estimation. Journal of the American Statistical Association, 85, 470-477.

Thoman, D. R., Bain, L. J. and Antle, C. E. (1969). Inferences on the parameters of the Weibull distribution. Technometrics 11, 445-460.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 881–897.

ON TESTS OF FIT BASED ON GROUPED DATA

Sherzod M. Mirakhmedov
Ghulam Ishaq Khan Institute of Engineering Sciences & Technology,
Topi-23460, Swabi ,NW.F.P. Pakistan
E-mail: shmirakhmedov@yahoo.com

and

Saidbek S. Mirakhmedov
Institute of Algoritm and Engineering,
Fayzulla Hodjaev-45, Tashkent -700149. Uzbekistan
E-mail: saeed_0810@yahoo.com

ABSTRACT

Let a sample of size *n* is grouped to *N* intervals and $\eta_1,...,\eta_N$ be a vector of frequencies of the intervals. The statistics of form $f(\eta_1)+...+f(\eta_N)$ is considered under assumption that $N\to\infty$ as $n\to\infty$. Goodness of fit test based on this class of statistics testing hypothesis of a continuous distribution versus a family of sequence of alternatives converging to hypothesis is considered. The problem of asymptotic efficiencies within the family of Pitman's and intermediate sequence of alternatives are studied. For the classical chi-square statistic the probability of large deviation result is presented.

Keywords: Grouped method, asymptotic efficiency, chi-square statistic, likelihood ration test, Pitman efficiency, Intermediate efficiency, Bahadur efficiency, large deviation.

1. INTRODUCTION

The classical goodness-of-fit problem of statistical inference is to test whether a sample has come from a given population. Specifically, we consider the problem of testing the goodness of fit of a continuous distribution P to a set of *n* observations grouped into *N* equal probability intervals. A large class of the tests based on statistics of the general form

$$R_N^f=\sum_{k=1}^{N} f(\eta_k), \tag{1.1}$$

where $\eta_1,...,\eta_N$ are the numbers of observations in the intervals. In the sequel the statistics R_N^f is called f- statistics, a test based on f-statistics is called f- tests. Three most popular cases of f-statistics are $X_N^2=\sum_{k=1}^{N}\eta_k^2$ is called the chi-square statistic, $\Lambda_N=\sum_{k=1}^{N}\eta_k\ln\eta_k$ is called the

likelihood ratio statistic, $\mu_r = \sum_{m=1}^{N} \Uparrow \{\eta_m = r\}$ is a number of intervals containing exactly r observations, here $\Uparrow\{A\}$ is the indicator of the event A. Particularly μ_0 is known empty boxes statistic (see, for instance, Kolchin et al (1978)).

The probability integral transformation $x \to F(x)$ reduces the problem of testing the goodness of fit of a continuous distribution P to that of Uniform [0,1]. So we consider the problem of testing the null hypothesis H_0: $p(x) = \mathrm{P}'(x) = 1,\ 0 < x < 1$, versus the family of sequence of alternatives

$$H_1 : p(x) = 1 + d\delta(n) l(x), \tag{1.2}$$

where constant $d > 0$, $l(x)$ is a function on [0,1] such that

$$\int_0^1 l(x)dx = 0, \quad \int_0^1 l^2(x)dx = 1,$$

$\delta(n) \to 0$ will be chosen so that the power for a f-test of size ω has a limit in $(\omega,1)$. The problem of testing H_0 against H_1 is called problem (H_0, H_1).

Holst (1972), Ivchenko and Medvedev (1978) and Gvanceladze and Chibisov (1979) have shown for similar to problem (H_0, H_1) that for $\delta(n) = n^{-1/2}$ the power of the f-tests tends to the significance level as $n \to \infty$ whenever $N \to \infty$; hence such of alternatives can not be detected by f-tests. In Mirakhmedov (1987), see also Quine and Robinson (1985), it was pointed out that if we let $N \to \infty$ then a sequence of alternatives that convergence to uniform must be in the form (1.2) with $\delta(n) = (n^2/N)^{-1/4}$, in order to keep the power bounded away from the significance level and unity, hence the f-tests do not discriminate alternatives (1.2) with $\delta(n) = o\left((n^2/N)^{-1/4}\right)$. This is a poor in comparison with other tests based on empiric distribution function, for example the Kolmogorov-Smirnov and Cramer -von Mises tests, who can detect similar alternatives at a distance $O\left(n^{-1/2}\right)$. On the other hand not always we need to consider the alternatives converging to hypothesis with the extreme rate of $O\left(n^{-1/2}\right)$. Moreover, concerning to the choice of the number of groups in chi-square test there is a well-known result by Mann and Wald (1942) stating that the optimal number N increase with n as $N = O\left(n^{2/5}\right)$. Hence it is unnatural to keep fixed number of groups when number of observations goes to infinity.

We are concerning with asymptotic results when $N = N(n) \to \infty$ as $n \to \infty$.This case intensively has been studied by many authors. We refer to Holst (1972), Morris (1975), Medvedev (1977), Borovkov (1978), Ivchenko and Medvedev (1978), Quine and Robinson (1984, 1985), Kallenberg (1985), Jammalamadaka and Tiwari (1985, 1987), Mirakhmedov (1985,1987), Jammalamadaka et al (1989). By Mirakhmedov (1990), Ivchenko and Mirakhmedov (1991, 1995) and Sirajdinov et al (1989) for R_N^f it is proved: the central limit theorem under mild condition together with Berry-Esseen bound, Edgeworth type asymptotic

expansion with exact formula for the first three terms, specified for statistics X_N^2, Λ_N and μ_r, and Cramer type large deviation result under Cramer condition $E\exp\{H|f(\xi)|\}<\infty$, $\exists\, H>0$, where ξ is the Poisson(α) r.v. , $\alpha_n = n/N \to \alpha, 0<\alpha<\infty$. We refer also to Ronzhin (1984) where under Cramer condition the Chernoff type large deviation result for R_N^f was proved. Note that the statistics Λ_N and μ_r are satisfy Cramer condition whereas chi-square statistic X_N^2 do not. Nevertheless by Quine and Robinson (1985) asymptotic result for Chernoff type large deviation probabilities of X_N^2 and Λ_N was obtained. These probabilistic results have been used to study asymptotic efficiencies (AE) of the f-tests. In detail corresponding results are as follows.

There are two basic ways of comparison of tests. One of them is in principle based on asymptotic analysis of the power of the tests. A test having maximal power within a class of tests under consideration is called asymptotic most powerful (AMP) test. AMP test may be not unique. In such a case an asymptotic behavior, as $n\to\infty$, of the difference in powers of two AMP tests is of interest; this situation gives rise to the concepts of second order efficient tests. The definition of the second order efficient test adapted to our problem is given below.

Another method of comparison of two tests of the same level is based on comparison of the number of observations needed to get same asymptotic power, when number of observations increases. If we have two tests with corresponding numbers of observations n_1 and n_2, then the limit of ratio n_2/n_1 is called the asymptotic relative efficiency of test 1 w.r.t. test 2. To investigate AE of a test we consider that ratio where n_1 and n_2 corresponds to that test and the AMP test respectively. AE of a test depends on three parameters: the level ω_n, the power β_n and the alternative H_1, which may depend on n. When sending n to infinity three concepts are: Pitman approach when $\omega_n \to \omega > 0$, $H_1 \to H_0$ (in some sense) in such rate that $\beta_n \to \beta \in (\omega,1)$; Bahadur approach when $\omega_n \to 0$, $\beta_n \to \beta \in (0,1)$, H_1 is fixed, i.e. does not approach the hypothesis; Kallenberg intermediate approach when $\omega_n \to 0$, $\beta_n \to \beta \in (0,1)$, $H_1 \to H_0$ but more slow than in Pitman case. Optimality of a test can be expressed by first order efficiency, which means that $n_2(\omega,\beta,H_1)/n_1(\omega,\beta,H_1)$ converges to 1, where the limit is taken according to the efficiency concept involved. The problem of finding of the limit of ratio $n_2(\omega,\beta,H_1)/n_1(\omega,\beta,H_1)$ being very difficult problem as usually reduces to finding of the ratio of what is called slopes of the tests under consideration, see, for instance, Fraser (1957), Nikitin (1995), Inglot (1999).

Let now ξ be the Poisson(α_n) r.v. with $\alpha_n = n/N$; P_i, E_i, Var_i stands for the probability, expectation and variance under H_i; $A_{i,N}$ and $\sigma_{i,N}^2$ stands for the asymptotic value of $\mathrm{E}_i R_N^f$ and $Var_i R_N^f$, $i=0,1$, respectively. Put

$$g(\xi) = f(\xi) - Ef(\xi) - \gamma(\xi-\alpha_n), \quad \gamma = \alpha^{-1}\operatorname{cov}(f(\xi),\xi) \tag{1.3}$$
$$\sigma^2(f) = Var\, g(\xi) = Var f(\xi)\left(1 - corr^2\left(f(\xi),\xi\right)\right).$$

From Theorem 2 of Mirakhmedov (1990) it follows that if

$$\kappa_N = \frac{E\left|g(\xi)\right|^3}{\sqrt{N}\sigma^3(f)} \to 0, \tag{1.4}$$

as n and $N \to \infty$, then

$$\nabla_N = \sup_x \left| P_i\left\{ R_N^f < x\sigma_{iN} + A_{iN} \right\} - \Phi(x) \right| = O\left(\kappa_N + \frac{1}{\sqrt{n}} \right), \tag{1.5}$$

where $\Phi(x)$ is the standard normal distribution function.

Remark 1.1. The statistics X_N^2 and Λ_N satisfy the condition (1.4) if and only if $n\alpha_n \to \infty$. But for the statistic μ_r (1.4) is valid under additionally conditions for r and α. Namely, if $\alpha_n \to 0$ and $n\alpha_n \to \infty$ then (1.4) satisfies for μ_r with $0 \le r \le 2$; if α_n is far away from zero then (1.4) is still true for μ_r, $r \ge 0$, if $\alpha_n - \ln N - r \ln\ln N \to \infty$. In what follows we assume that $n\alpha_n \to \infty$.

Asymptotical Most Powerful Test. It is known, see Holst (1972), Ivchenko and Mirakhmedov (1991, 1995), that under alternative (1.2) with $\delta(n) \to 0$

$$A_{0N} = NEf(\xi),\ \sigma_{0N}^2 = N\sigma^2(f),\ \sigma_{1N}^2 = \sigma_{0N}^2(1+o(1)), \tag{1.6}$$

$$x_N(f) \stackrel{def}{=} \left(A_{1N}(f) - A_{0N}(f)\right)/\sqrt{N}\sigma(f) = \sqrt{\frac{n\alpha_n}{2}}\delta^2(n)\rho(f,\alpha_n)d^2\left(1+o(1)\right), \tag{1.7}$$

with $\rho(f,\alpha_n) = corr\left(f(\xi) - \gamma\xi, \xi^2 - (2\alpha_n + 1)\xi \right)$.

Let $\beta(f)$ be the asymptotical power of the f- test of a size ω. If (1.4) is satisfied then (1.5) and (1.7) imply

$$\beta(f) = \Phi\left(\sqrt{\frac{n\alpha_n}{2}}\delta^2(n)\rho(f,\alpha_n)d^2 - u_\omega \right),\ u_\omega = \Phi^{-1}(1-\omega). \tag{1.8}$$

Hence functional $\rho(f,\alpha_n)$ plays the key rule in determining of the asymptotic quality of the f-test. Its meaning is clarified by the following (see Lemma 1 by Ivchenko and Mirakhmedov (1995))

$$\rho(f,\alpha_n) = corr_0\left(R_N^f, X_N^2\right)\left(1+o(1)\right). \tag{1.9}$$

Hence $\left|\rho(f,\alpha_n)\right| \le 1$, and $\left|\rho(f,\alpha_n)\right| = 1$ for any α_n only for chi-square test. The equality (1.8) means that f-test does not detect alternatives (1.2) with $\delta(n) = o\left((n\alpha_n)^{-1/4}\right)$.

Let in (1.2) $\delta(n) = (n\alpha_n)^{-1/4}$. Then for the problem (H_0, H_1) the chi-square test is AMP within class of f-tests for any α_n. Nevertheless if $\alpha_n \to 0$ or $\alpha_n \to \infty$ then there exist other

AMP tests also, because in these cases may $\rho(f,\alpha_n)\to 1$. For example, if $\alpha_n\to 0$ and $\Delta^2 f(0)\neq 0$, where operator $\Delta f(x)=f(x+1)-f(x)$, then

$$\rho(f,\alpha_n)=1-\frac{\alpha_n}{6}\frac{\Delta^3 f(0)}{\Delta^2 f(0)}+O\left(\alpha_n^2\right);$$

if $\alpha_n\to\infty$ for the statistic Λ_N $\rho(f,\alpha_n)=1-\frac{1}{6\alpha_n}\left(1+o(1)\right)$.

Remark 1.2. Let $M(n,p_1,...,p_N)$ be a multinomial distribution with parameters n, $p_m>0$, $p_1+...+p_N=1$. Above stated goodness of fit problem can be reduced by grouping of observations to that of testing of hypothesis $p_m=N^{-1}, m=1,...,N$, versus the sequence of alternatives $p_m=N^{-1}\left(1+d\delta(n)\ell_m\right)$, $m=1,...,N$, where $\sum_{m=1}^{N}\ell_m=0, \frac{1}{N}\sum_{m=1}^{N}\ell_m^2=1$. Such goodness of fit problem have been studied by Ivchenko and Medvedev (1978), Ivchenko and Mirakhmedov (1991,1995); here the case when $\alpha_n\to 0$ is of interest (it corresponds to so namely "small sample" situation), whereas for above stated problem $\left(H_0,H_1\right)$ it is not of interest because the quality of f-tests goes dawn.

Second Order Asymptotic Efficiency. Future comparison of the AMP tests, when $\alpha_n\to 0$ or $\alpha_n\to\infty$, based on a notion of the second order efficiency. Set $\vartheta=d^2/\sqrt{2}-u_\omega$, where ω is a size of f-test and u_ω from (1.8), $\beta_n(R_N^f;\vartheta)$ stands for the power of the f-test of a size ω. By Ivchenko and Mirakhmedov (1991) was shown the following asymptotic expansion of the power $\beta_n(X_N^2;\vartheta)$ of chi-square test $\beta_n(X_N^2;\vartheta)=\Phi\left(\vartheta\right)+\varepsilon_n(X_N^2;\vartheta)\left(1+o(1)\right)$, where

$$\varepsilon_n(X_N^2;\vartheta)=\frac{\exp\left\{-\vartheta^2/2\right\}}{\sqrt{2\pi n\alpha_n}}\left(\frac{1-\vartheta^2}{3\sqrt{2}}+\frac{\vartheta d^2}{2}+\sqrt{2}S_1\left(\vartheta\sqrt{\frac{n\alpha_n}{2}}+\frac{n}{2}\right)\right),\ \text{if}\ \alpha_n\to 0,$$

$$\varepsilon_n(X_N^2;\vartheta)=\frac{\exp\left\{-\vartheta^2/2\right\}}{\sqrt{2\pi N}}\left(1-\frac{\vartheta d^2}{\sqrt{2}}\right),\ \text{if}\ \alpha_n\to\infty.$$

Here $S_1(x)=-\{x\}+0.5$, $\{x\}$ denotes a fractional part of the x. The function $S_1(x)$ is well known in the theory of asymptotical expansion of the cumulative distribution function of lattice random variables. It is raised here because chi-square statistic X_N^2 is the lattice random variable with span equal to 2.

Definition. The AMP f-test is called second order asymptotic efficient (SOAE) with respect to chi-square test, if its power has asymptotic expansion $\beta_n(R_N^f,\vartheta)=\Phi\left(\vartheta\right)+\varepsilon_n(f;\vartheta)\left(1+o(1)\right)$, with $\varepsilon_n(f;\vartheta)\to 0$ and $\varepsilon_n(f;\vartheta)=\varepsilon_n(X_N^2;\vartheta)\left(1+o(1)\right)$.

The conclusion of Ivchenko and Mirakhmedov (1991) is as follows. Let $\alpha_n \to 0$, then there exist SOAE f-tests only if $\alpha = O\left((n\alpha_n)^{-1/2}\right)$, i.e. $n = O\left(N^{3/4}\right)$. For example empty cells test based on μ_0 is SOAE, but likelihood ratio test is not SOAE. If $\alpha_n \to \infty$, then SOAE does not exist.

Pitman Efficiency. Under certain regularity conditions (see, for example, Fraser (1957)), the efficacy of a test based on statistic, say V, is given by $e(V) = \mu_V^2 / \sigma_V^2$. Here μ_V and σ_V^2 are the mean and variance of the limiting normal distribution under the sequence of alternatives when the test statistics has been normalized to have limiting standard normal distribution under the hypothesis. In such a situation, the Pitman's asymptotic relative efficiency (in sense of comparison of the sample sizes, see above) of one test with respect to another is the ratio of their efficacies. Because of (1.6) and (1.8) the alternatives (1.2) with $\delta(n) = (n\alpha_n)^{-1/4}$ are form P_{alt} - the family of Pitman alternatives: f-test of a size $\omega_n(f) \to \omega > 0$ and $\rho(f, \alpha_n)$ is far away from zero has the power (see (1.8)) $\beta_n(f) \to \beta(f) = \Phi\left(2^{-1/2} d^2 \rho(f, \alpha_n) - u_\omega\right) \in (\omega, 1)$. For the efficacy $e(f)$ of the f -test we have $e(f) = x_N^2(f) = 2^{-1} d^4 n \alpha_n \delta^4(n) \rho^2(f, \alpha_n)$. Hence the Pitman asymptotic relative efficiency of f-test is determined by functional $\rho(f, \alpha_n)$;within class of f-tests the chi-square test is asymptotic most efficient (AME) in Pitman sense; the Pitman efficiency of f- tests goes down as number of intervals *N* increases for a given sample size *n*. These verdicts have been proved by Holst (1972), Ivchenko and Medvedev (1978), Mirakhmedov (1987) and Quine and Robinson (1985).

Bahadur Efficiency. Another extreme family of alternatives is B_{alt} - Bahadur (as well as Hodges-Lehman) family of alternatives when alternatives do not approach the hypothesis, i.e. $\delta(n)$ is constant. The Bahadur's AE of f-tests in the family B_{alt} have been developed by Ronzhin (1984) who showed, for a certain subclass of f-tests, that whenever $\beta_n(f) \to \beta(f) < 1$, $-n^{-1} \log \omega_n(f)$ converges to limit (which is called slope of the f-test to the alternative) of specifies the Bahadur AE of f-test. This limit is determined by the logarithmic rate deviations probabilities (for deviation of order $O(\sqrt{N})$) under H_0, which require restrictive Cramer's condition (see (2.1) below) on the test statistics. In particular, this condition excludes the chi-square statistic. A comparative analysis of chi-square test's Bahadur efficiency relative to the likelihood ratio test was carried out by Quine and Robinson (1985). They showed that likelihood ratio test is much more Bahadur efficient than that chi-square test, in contrast to their relative Pitman efficiency. In a similar setup, the AE of the chi-square and likelihood ratio tests when $N = o\left(\sqrt{n}\right)$ were studied by Kallenberg (1985).

Intermediate Efficiency. The situation when $\delta(n) \to 0$ but slower than that in the P_{alt} give rise to intermediate family of sequence of alternatives of three types: K_{all} -family of alternatives (1.2) with $\delta(n) \to 0$, $\delta(n)(n\alpha_n)^{1/4} \to \infty$; $K_{1/6}$ - subfamily of K_{all} with $\delta(n) = o\left((n\alpha_n^2)^{-1/6}\right)$ if $\alpha_n \geq 1$, and $\delta(n) = o\left(n^{-1/6}\right)$ if $\alpha_n < 1$; $K_{\sqrt{\log}}$ - subfamily of K_{all} with $\delta(n) = O\left(n\alpha_n\right)^{-1/4} \log^{1/4} N\Big)$. Actually such division of family of intermediate alternatives becomes from probability of large

deviations results presented below in Section 3 and because of relation (1.7). These families of alternatives are adapted (to the problem considering in the present paper) variant of that introduced by Kallenberg (1983).

Following to the logic of the Bahadur's approach, intermediate AE (between Pitman and Bahadur settings) of f-tests for the family of intermediate alternatives can be measured by the logarithmic rate of decrease of the test size when the power is fixed. Therefore by Ivchenko and Mirakhmedov (1995) as a measure of the performance of f-test was considered the asymptotic value of a slope

$$e_n^{\omega}(R_N^f) = -\log P_0\left\{R_N^f \geq NA_{1N}(f)\right\}. \tag{1.10}$$

The test having largest asymptotic value of $e_n^{\omega}(R_N^f)$ is called asymptotic most intermediate efficient test. We distinguish three types of intermediate efficiencies: weak intermediate, intermediate and strong intermediate for families $K_{\sqrt{\log}}$, $K_{1/6}$ and K_{all} respectively. Asymptotic relative efficiency of one test to another is defined as ratio of its asymptotic slopes. For Pitman's alternatives this is equal to the Pitman's asymptotic relative efficiency, whereas for intermediate alternatives it is related to the intermediate relative asymptotic efficiency in weak sense introduced by Inglot (1999). Ivchenko and Mirakhmedov (1995) has extended above said the Pitman's and the Bahadur's efficiencies properties of chi-square test: chi-square test is still optimal (in sense of asymptotic value of $e_n^{\omega}(R_N^f)$) within class of f-tests in the family $K_{\sqrt{\log}}$ but in the family K_{all} except $K_{1/6}$ it is much inferior to those statistics satisfying the Cramer condition, particularly to likelihood ratio test.

As it follows from above said the chi-square test is AMP, SOAE, and AME in the Pitman's and the weak intermediate senses, but it losses optimality property in terms of the Bahadur's and the strong intermediate efficiencies senses. For the f-tests satisfying Cramer condition AE for all range of alternatives (1.2), i.e. for family of alternatives P_{alt}, K_{all} and B_{alt} in the situation when α is far away from zero and infinity have been studied also. AE of the chi-square test in the family of alternatives $K_{1/6}$ was open problem.

Theorem 2.1 of Section 2 covers that existing gap in study of AE of the chi-square test; also it extends result of Ivchenko and Mirakhmedov (1995) on AE of chi-square test within family of strong intermediate alternatives for the cases $\alpha_n \to 0$ and $\alpha_n \to \infty$. In Section 3 the probability of large deviation result for chi-square statistic, likely to be own interest, is presented. The auxiliary assertions are collected in Section Appendix.

2. ASYMPTOTIC EFFICIENCY

We assume $n\alpha \to \infty$. Also we continue to use denotes (1.3), (1.6), (1.7) and (1.10).

Theorem 2.1.

1. If $0 < c_0 \leq \alpha_n \leq c_1 < \infty$ and

$$\mathrm{E}\exp\{H\,|f(\xi)|\} < \infty, \tag{2.1}$$

for some $H > 0$, then in the family of alternatives K_{all} $\dfrac{e_n^{\omega}(R_N^f)}{n\alpha_n\delta^4(n)} = \dfrac{d^4}{4}\rho^2(f,\alpha_n)(1+o(1))$.

2. In the family of alternatives $K_{1/6}$

$$\frac{e_n^{\omega}(X_N^2)}{n\alpha_n\delta^4(n)} = \frac{d^4}{4}(1+o(1)).$$

3. In the family of alternatives K_{all} if

$$\delta^3(n)(n\alpha_n)^{1/2}\log^{-1}N \to \infty, \tag{2.2}$$

then $\dfrac{e_n^{\omega}(X_N^2)}{n\alpha_n\delta^4(n)} = o(1).$

First assertion of Theorem 2.1 is case (ii) of Theorem 3 of Ivchenko and Mirakhmedov (1995). Theorem 2.1 together with results of Quine and Robinson (1985) implies that chi-square test is optimal within class of f-tests for the families of Pitman and intermediate sequence of alternatives but it is much inferior to those statistics satisfying the Cramer condition (particularly to likelihood ratio test and μ_r, $r \ge 0$, tests) for families of strong intermediate alternatives under condition (2.2) and Bahadur family of alternatives. It remains a gap in the study of the intermediate efficiency of the chi-square test in the family K_{all} with

$$c_2(n\alpha_n^2)^{-1/6} \le \delta(n) \le c_3(n\alpha_n)^{-1/6}\log^{1/3}N \quad \text{for} \quad \alpha_n \ge 1,$$

and

$$c_2 n^{-1/6} \le \delta(n) \le c_3(n\alpha_n)^{-1/6}\log^{1/3}N \quad \text{for} \quad \alpha_n < 1.$$

Remark 2.1. An alternative approach to testing of uniformity [0,1] is to construct tests based on spacings. Let $X_{1n} \le X_{2n} \le ... \le X_{n,n}$ be the order statistics of the sample $X_1, X_2, ..., X_n$, $W_{m,n}^{(s)} = X_{ms,n} - X_{(m-1)s,n}$, $m = 1, 2, ..., N'$, $W_{N'+1,n}^{(s)} = 1 - X_{N's,n}$, with notation $X_{o,n} = 0$ and $X_{n+1,n} = 1$, be their s-spacings; $N' = [(n+1)/s]$, $N = N'$ if $(n+1)/s$ is an integer and $N = N'+1$ otherwise; $W = (W_{1,n}^{(s)}, ..., W_{N,n}^{(s)})$. The step of the spacings s may increase together with n, but $s = o(n)$.

The order statistics $X_{0n}, X_{1n}, ..., X_{n,n}, X_{n+1,n}$ divide interval [0,1] to $s+1$ subintervals (groups), that is we again, actually, deal with method of grouping data. In contrast to above considered method here the ends of intervals are random and we are using, for a statistical procedure, the length of intervals instead of frequencies of intervals. Most common among tests based on spacings are tests based on statistics of the form $R_N^f(W) = \sum_{m=1}^{N} f\left((n+1)W_{m,n}^{(s)}\right)$.

AE properties here alike to those of f-tests (the step of spacings s plays the role of α). For example: such tests can detect alternatives (1.2) with $\delta(n)=(ns)^{-\varepsilon}, \varepsilon\in(0,1/4]$, AMP test for the alternatives (1.2) with $\delta(n)=(ns)^{-1/4}$ is the Greenwood's test based on statistic

$$G_N^2=\sum_{k=1}^{N}\left((n+1)W_{k,n}^{(s)}\right)^2.$$

While considerable attention has been devoted in literature to type $R_N^f(W)$ statistics, we are not in position to give here all the details of existing results. Reader can find detailed information, applications, and references, for instance, in papers by Pyke (1965), Rao and Sethuraman (1975), Jammalamadaka et al (1989), Zhou and Jammalamadaka (1989), Jammalamadaka and Goria (2004), Mirakhmedov (2005, 2006) and Mirakhmedov and Naem (2008). We wish only refer to Jammalamadaka and Tiwari (1985, 1987) and Jammalamadaka et al (1989) where the Pitman's ARE of chi-square test and Greenwood test with $s=[\alpha_n]$ were studied. They shown that if $s\to\infty$ then these two tests have same Pitman efficiency, but for fixed s (that corresponds to the case $1\le\alpha_n<c<\infty$) spacings tests are preferable to comparable chi-square procedure. From Theorem 2.1 and results of Mirakhmedov (2009) it follows that the same verdicts are still true for intermediate asymptotic efficiency of chi-square test and spacings based tests.

Proof of Theorem 2.1. *Part* 1 was proved by Ivchenko and Mirakhmedov (1995). To *prove of Part* 2 we note that

$$e_n^{\omega}(R_N^f)=-\log P_0\left\{R_N^f\ge x_N(f)\sigma(f)\sqrt{N}+Nf(\xi)\right\}$$
$$=-\log P_0\left\{\frac{R_N^f-Nf(\xi)}{\sigma(f)\sqrt{N}}\ge\sqrt{\frac{n\alpha_n}{2}}\delta^2(n)\rho(f,\alpha_n)d^2\left(1+o(1)\right)\right\}.$$

since (1.10). Therefore, Part 2 follows from Theorem 3.1 of Sec.3, the facts that $\rho(f,\alpha)=1$ for the chi-square statistic and that $-\log\Phi(-x)=2^{-1}x^2(1+o(1))$, as $x\to\infty$.

Proof of Part 3. We use similar to Quine and Robinson (1985) approach. Let $\eta\square B(k,p)$ means that a
r.v. η has the Binomial distribution with parameters n and p, $0<p<1$; and
$\nu(n)=\left[\alpha_n+\sqrt{\alpha_n+d^2n\alpha_n\delta^2(n)}\right]+1$, where $[a]$ is an integer part of *a*. We have

$$P\left\{X_N^2>NA_{1N}\right\}=P\left\{X_N^2-n(\alpha_n+1)>x_N(f)\sqrt{2n\alpha_n}\right\}$$
$$=P\left\{\sum_{m=1}^{N}\left((\eta_m-\alpha_n)^2-\alpha_n\right)>d^2n\alpha_n\delta^2(n)\left(1+o(1)\right)\right\}$$
$$=P\left\{\sum_{m=2}^{N}\left((\eta_m-\alpha_n)^2-\alpha_n\right)\ge 0\Big/\eta_1=\nu(n)\right\}P\left\{\eta_1=\nu(n)\right\}$$
$$=P\left\{\sum_{m=1}^{N-1}\left((\hat{\eta}_m-\alpha_n)^2-\alpha_n\right)\ge 0\right\}P\left\{\eta_1=\nu(n)\right\},\qquad(2.3)$$

where $\hat{\eta}_m \sim Bi\left(n-v(n),(N-1)^{-1}\right)$, because of fact that the conditional distribution of multinomial vector $(\eta_1,...,\eta_N)$ given one component is again multinomial, but N replaced by $N-1$. Put $\tilde{\alpha}_n=(n-v(n))/(N-1)$. It easy to see that $v(n))/n=\left(N^{-1}+d\delta(n)N^{-1/2}\right)\left(1+o(1)\right)$ and $\tilde{\alpha}_n=\alpha_n\left(1+O\left(N^{-1}+d\delta(n)N^{-1/2}\right)\right)$. Simple algebra shows that $\sum_{m=1}^{N}\left(\hat{\eta}_m-\alpha_n\right)^2\geq\sum_{m=1}^{N}\left(\hat{\eta}_m-\hat{\alpha}_n\right)^2$. Using this fact we have

$$P\left\{\sum_{m=1}^{N-1}\left((\hat{\eta}_m-\alpha_n)^2-\alpha_n\right)\geq 0\right\}\geq P\left\{\sum_{m=1}^{N-1}(\hat{\eta}_m-\hat{\alpha}_n)^2\geq(N-1)\alpha_n\right\}$$

$$=P\left\{\sum_{m=1}^{N-1}\left((\hat{\eta}_m-\hat{\alpha}_n)^2-\hat{\alpha}_n\right)\geq\left(v(n)-\alpha_n\right)\right\}$$

$$=P\left\{\frac{\sum_{m=1}^{N-1}\left((\hat{\eta}_m-\hat{\alpha}_n)^2-\hat{\alpha}_n\right)}{\sqrt{2(n-v(n))^2/(N-1)}}\geq d\delta(n)+o(1)\right\}\geq c>0, \qquad (2.4)$$

because of $(n-v(n))\tilde{\alpha}_n=n\alpha_n\left(1+o(1)\right)\to\infty$, and hence asymptotic normality of chi-square statistic $\sum_{m=1}^{N-1}(\hat{\eta}_m-\hat{\alpha}_n)^2$ is valid, see above Remark 1.1.

Set $g(x,p)=x\log\left(x/p\right)+(1-x)\log\left((1-x)/(1-p)\right)$, $x\in(0,1)$ and $p\in(0,1)$. The following Lemma 2.1 is presented by Quine and Robinson (1985).

Lemma 2.1. If $\xi\sim Bi\left(k,p\right)$ then for integer kx

$$P\left\{\xi=kx\right\}\geq 0.8\left(2\pi kx(1-x)\right)^{-1/2}\exp\left\{-kg(x,p)\right\}.$$

Note that $\eta_1\sim B(n,N^{-1})$. Therefore due to Lemma 2.1 we have

$$P\left\{\eta_1=v(n)\right\}$$

$$\geq c\left[v(n)\left(1-v(n)n^{-1}\right)\right]^{-1/2}\exp\left\{-v(n)\log(\alpha_n^{-1}v(n))-n(1-n^{-1}v(n))\log\frac{1-v(n)n^{-1}}{1-N^{-1}}\right\}$$

$$\geq c\left(v(n)\right)^{-1/2}\exp\left\{-v(n)\log(\alpha_n^{-1}v(n))\right\},$$

since $v(n)/n>N^{-1}$. Hence

$$-\frac{\log P\left\{\eta_1=v(n)\right\}}{n\alpha_n\delta^4(n)}\leq c\frac{\log v(n)+v(n)\log(\alpha_n^{-1}v(n))}{n\alpha_n\delta^4(n)}$$

$$\leq c\frac{\alpha_n+\delta(n)\sqrt{n\alpha_n}}{n\alpha_n\delta^4(n)}\log\frac{v(n)}{\alpha_n}\leq c\left[\frac{1}{n\delta^4(n)}+\frac{1}{\delta^3(n)\sqrt{n\alpha_n}}\right]\log N=o(1), \qquad (2.5)$$

because of (2.2). The Part 3 follows from (2.3), (2.4) and (2.5).

4. LARGE DEVIATION

We recall denotes from Sec. 2: ξ is Poisson (α_n) r.v., $\alpha_n = n/N$, $\gamma = \alpha_n^{-1}\operatorname{cov}\big(f(\xi),\xi\big)$,

$$g(\xi) = f(\xi) - Ef(\xi) - \gamma(\xi - \alpha_n),\ \sigma^2(f) = Varg(\xi) = Varf(\xi)\Big(1 - corr^2\big(f(\xi),\xi\big)\Big).$$

The following Theorem 3.1 is likely of own interest giving an asymptotic for the large deviation probability of chi-square statistic.

Theorem 3.1. If $\alpha_n \ge 1$ then for all $x \ge 0$ and $x = o\Big(N^{1/6}\Big)$, if $\alpha_n < 1$ then for all $x \ge 0$ and $x = o\Big((n\alpha_n^3)^{1/6}\Big)$ one has $P\Big\{X_N^2 > x\sqrt{2n\alpha_n} + n(\alpha_n + 1)\Big\} = \big(1 - \Phi(x)\big)\big(1 + o(1)\big)$.

Proof. Let $\xi_1, \xi_2, \ldots$ be independent copies of the r.v. ξ. Also let $C_k(\varsigma)$ be a cummulant of the *k* th order of a r.v. ς,

$$S_N = \sum_{m=1}^{N}\big(\xi_m - \alpha_n\big),\ X_{1,N}^2 = \sum_{m=1}^{N}\big(\eta_m - \alpha_n\big)^2,\ \tilde{X}_{1,N}^2 = \sum_{m=1}^{N}\big(\xi_m - \alpha_n\big)^2.$$

Lemma 3.1. For any fixed $k \ge 3$ and enough large *n*, $C_k(X_{1,N}^2) = C_k(\tilde{X}_{1,N}^2)\big(1 + o(1)\big)$.

Proof. Let fixed $k \ge 3$. It is well known that $L\,((\mu_1, \ldots, \eta_N)) = L\,((\xi_1, \ldots, \xi_N)/S_N = 0)$, where $L\,(X)$ denotes the distribution of the random vector X. Hence

$$E(X_{1,N}^2)^k = E\big((\tilde{X}_{1,N}^2)^k / S_N = 0\big). \tag{3.1}$$

On the other hand $E\big((X_{1,N}^2)^k e^{i\tau S_N}\big) = E\Big\{e^{i\tau S_N} E\big((X_{1,N}^2)^k / S_N\big)\Big\}$. Integrating w.r.t. τ both side of this equality over interval $[-\pi, \pi]$, and taking into account (3.1) we have

$$E\left(X_{1,N}^2\right)^k = d_n \int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} E\left(\tilde{X}_{1,N}^2\right)^k \exp\left\{i\tau\frac{S_N}{\sqrt{n}}\right\} d\tau,$$

where

$$d_n = \Big(2\pi\sqrt{n}P\{S_N = 0\}\Big)^{-1} = \frac{1}{2\pi\sqrt{n}}\frac{n!e^n}{n^n}.$$

Hence, putting $\tilde{\xi}_m = \xi_m - \alpha_n$ we have

$$E\left(X_{1,N}^2\right)^k = d_n \sum_{l=1}^{k}\sum_k'\sum_l'' \int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} E\left[\tilde{\xi}_{j_1}^{2k_1} \cdot \ldots \cdot \tilde{\xi}_{j_l}^{2k_l} \exp\left\{i\tau\frac{S_N}{\sqrt{n}}\right\}\right] d\tau, \tag{3.2}$$

where $\sum_k'$ is the summation over all l-tuples $(k_1,...,k_l)$ with positive integer components such that $k_1+...+k_l=k$; $\sum_l''$ is the summation over all l-tuples $(j_1,...,j_l)$ with components not equal of each others and such that $j_m=1,2,...,N$; $m=1,2,...,l$.

Putting $S_{N,l}=\sum_{i=1}^{l}\tilde{\xi}_{j_i}$ we have

$$\int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} E\left[\tilde{\xi}_{j_1}^{2k_1}\cdot...\cdot\tilde{\xi}_{j_l}^{2k_l}\exp\left\{i\tau\frac{S_N}{\sqrt{n}}\right\}\right]d\tau$$

$$=\int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} E\left[\tilde{\xi}_{j_1}^{2k_1}\cdot...\cdot\tilde{\xi}_{j_l}^{2k_l}\exp\left\{i\tau\frac{S_{N,l}}{\sqrt{n}}\right\}\right]\left[E\exp\left\{i\tau\frac{S_N-S_{N,l}}{\sqrt{n}}\right\}-\exp\left\{-\frac{\tau^2}{2}\left(1-\frac{l}{N}\right)\right\}\right]d\tau$$

$$+\int_{-\pi\sqrt{n}}^{\pi\sqrt{n}}\exp\left\{-\frac{\tau^2}{2}\left(1-\frac{l}{N}\right)\right\}E\left[\tilde{\xi}_{j_1}^{2k_1}\cdot...\cdot\tilde{\xi}_{j_l}^{2k_l}\left(\exp\left\{i\tau\frac{S_{N,l}}{\sqrt{n}}\right\}-1\right)\right]d\tau$$

$$+E\left[\tilde{\xi}_{j_1}^{2k_1}\cdot...\cdot\tilde{\xi}_{j_l}^{2k_l}\right]\int_{-\pi\sqrt{n}}^{\pi\sqrt{n}}\exp\left\{-\frac{\tau^2}{2}\left(1-\frac{l}{N}\right)\right\}d\tau\overset{def}{=}J_1+J_2+J_3. \tag{3.3}$$

We have

$$\left|E\exp\left\{\frac{i\tau\tilde{\xi}_m}{\sqrt{n}}\right\}\right|=\exp\left\{-2\alpha\sin^2\frac{\tau}{2\sqrt{n}}\right\}\le\exp\left\{-\frac{2\alpha}{n\pi^2}\tau^2\right\}, \tag{3.4}$$

since $\sin^2 u/2\ge u^2/\pi^2$, $|u|\le\pi$. Put $\Delta_n=3N^{-1/2}+(n\alpha_n)^{-1/2}$, then $\sqrt{n}Var^{3/2}\xi_m/E|\xi_m|^3\ge\Delta_n^{-1}$. Taking into account this inequality and using Assertion 4 and (3.4) by some algebra we have

$$|J_1|\le E\left[\tilde{\xi}_{j_1}^{2k_1}\cdot...\cdot\tilde{\xi}_{j_l}^{2k_l}\right]\left\{\int_{|\tau|\le(6\Delta_n)^{-1}}\left|E\exp\left\{i\tau\frac{S_N-S_{N,l}}{\sqrt{n}}\right\}-\exp\left\{-\frac{\tau^2}{2}\left(1-\frac{l}{N}\right)\right\}\right|d\tau\right.$$

$$\left.+\int_{\pi\sqrt{n}\ge|\tau|\ge(6\Delta_n)^{-1}}\left(\left|E\exp\left\{i\tau\frac{S_N-S_{N,l}}{\sqrt{n}}\right\}\right|+\exp\left\{-\frac{\tau^2}{2}\left(1-\frac{l}{N}\right)\right\}\right)d\tau\right\}$$

$$\le C_4\Delta_n E\left[\tilde{\xi}_{j_1}^{2k_1}\cdot...\cdot\tilde{\xi}_{j_l}^{2k_l}\right]. \tag{3.5}$$

Lemma 3.2. For any integer $s\ge2$ one has

$$E\tilde{\xi}_m^s=s!\sum_{l=1}^{[s/2]}c_{l,s}\alpha_n^l, \tag{3.6}$$

with

$$0 < c_{l,s} < 1 \quad \text{for all of} \quad l = 1, 2, ..., [s/2]. \tag{3.7}$$

Proof. We have $Ee^{i\tau\tilde{\xi}_m} = \exp\left\{\alpha_n\left(e^{i\tau} - 1 - i\tau\right)\right\}$. Applying here the Bruno's formula we find

$$E\tilde{\xi}_m^s = s!\sum \alpha_n^{k_2+...+k_s} \prod_{j=2}^{s} \frac{1}{k_j!(j!)^{k_j}} = s!\sum_{l=1}^{[s/2]} c_{l,s}\alpha_n^l$$

where $\sum$ is summation over all non-negative $k_2, ..., k_s$ such that $2k_2 + ... + sk_s = s$ and $l = k_2 + ... + k_s$, $c_{l,s} = \sum\prod_{j=2}^{s} \frac{1}{k_j!(j!)^{k_j}}$. Lemma 3.2 follows.

In particularly, from Lemma 3.2 it follows that $E\tilde{\xi}_m^{2k+1} \le C_1 E\tilde{\xi}_m^{2k}$ and $E\tilde{\xi}_m^{2k+2} \le C_2 k^2 \alpha_n E\tilde{\xi}_m^{2k}$. Using this fact after some algebra we obtain

$$|J_2| \le \frac{1}{2n} E\left[\tilde{\xi}_{j_1}^{2k_1} \cdot ... \cdot \tilde{\xi}_{j_l}^{2k_l} S_{N,l}^2\right] \int_{-\infty}^{\infty} \tau^2 \exp\left\{-\frac{\tau^2}{2}\left(1 - \frac{l}{N}\right)\right\} d\tau$$

$$\le C \frac{\alpha_n k^2 l + k^2 l^2}{n} E\left[\tilde{\xi}_{j_1}^{2k_1} \cdot ... \cdot \tilde{\xi}_{j_l}^{2k_l}\right]. \tag{3.8}$$

It is obvious that

$$J_3 = \sqrt{2\pi} E\left[\tilde{\xi}_{j_1}^{2k_1} \cdot ... \cdot \tilde{\xi}_{j_l}^{2k_l}\right]\left(1 + O\left(\frac{l}{N}\right)\right). \tag{3.9}$$

Apply (3.5), (3.8) and (3.9) in (3.3) to get

$$\int_{-\infty}^{\infty} E\left(\tilde{\xi}_{j_1}^{2k_1} \cdot ... \cdot \tilde{\xi}_{j_l}^{2k_l} \exp\left\{i\tau \frac{S_N}{\sqrt{n}}\right\}\right) d\tau$$

$$= \sqrt{2\pi}\left(1 + O\left(\frac{k^2}{\sqrt{n}} + \frac{k^3}{N} + \frac{1}{\sqrt{n\alpha_n}}\right)\right) E\left[\tilde{\xi}_{j_1}^{2k_1} \cdot ... \cdot \tilde{\xi}_{j_l}^{2k_l}\right]. \tag{3.10}$$

Using Stirling's formula it is easy to see that $d_n = (2\pi)^{-1/2}\left(1 + O(n^{-1})\right)$. Inserting this and (3.10) into relation (3.2) we have

$$E\left(X_{1,N}^2\right)^k = E\left(\tilde{X}_{1,N}^2\right)^k \left(1 + o(1)\right). \tag{3.11}$$

Lemma 3.1 follows from Assertion 3, see Appendix, and (3.11), because $C_k(X_N^2) = C_k(X_{1,N}^2)$.

Let $\alpha_n \ge 1$. Due to Lemma 3.2 and Stirling's formula we have: for $k \ge 3$

$$\left|E\left(\tilde{\xi}_m^2 - E\tilde{\xi}_m^2\right)^k\right| \le 2^k E\tilde{\xi}_m^{2k} \le (2k)!2^k \sum_{l=1}^{k} \alpha_n^l \le (k!)^2 \frac{2^k (2k)!}{(k!)^2} k\alpha_n^k$$

$$\le (k!)^2 2^{3k+1} k\alpha_n^k \le (k!)^2 \left(2^{10} k^{1/(k-2)} \alpha_n\right)^{k-2} Var\tilde{\xi}_m^2 \le (k!)^2 \left(2^{12}\alpha_n\right)^{k-2} Var\tilde{\xi}_m^2.$$

because of $Var\tilde{\xi}_m^2 = 2\alpha_n^2 + \alpha_n$. Therefore by the Assertion 1 with $\zeta = \tilde{\xi}_m^2 - E\tilde{\xi}_m^2$, we have

$$\left|C_k\left(\tilde{\xi}_m^2\right)\right| = \left|C_k\left(\tilde{\xi}_m^2 - E\tilde{\xi}_m^2\right)\right| \le (k!)^2 \left(2^{12}\alpha_n\right)^{k-2} Var\tilde{\xi}_m^2, \quad k \ge 3.$$

Hence

$$\left|C_k\left(\tilde{X}_{1,N}^2\right)\right| \le (k!)^2 \left(2^{12}\alpha_n\right)^{k-2} NVar\tilde{\xi}_1^2, \quad k \ge 3,$$

because as $\tilde{X}_{1,N}^2$ is a sum of i.i.d. r.v.'s. Apply this and Lemma 2.1 to get: for any fixed $k = 3,4,...$ and enough large *n*

$$\left|C_k\left(X_N^2\right)\right| = \left|C_k\left(X_{1,N}^2\right)\right| \le (k!)^2 \left(2^{10}\alpha_n\right)^{k-2} \frac{Var\tilde{\xi}_1^2}{2\alpha_n^2} 2n\alpha_n \le (k!)^2 \left(2^{12}\alpha_n\right)^{k-2} VarX_N^2,$$

Thus r.v. X_N^2 satisfy the Statulevicius condition (S_ν) with $\nu = 1$ and $\Delta = 2^{12}\alpha_n$, see Appendix. Theorem 2.3 follows from the Assertion 3 with $\zeta = X_N^2$, $\sigma = \sqrt{2n\alpha_n}$ and $\Delta = 2^{12}\alpha_n$.

Let $\alpha_n < 1$. Then

$$\left|E\left(\tilde{\xi}_m^2 - E\tilde{\xi}_m^2\right)^k\right| \le (2k)!2^k \sum_{l=1}^{k} \alpha_n^l \le (k!)^2 \frac{2^k (2k)!}{(k!)^2} k\alpha_n \le (k!)^2 2^{12(k-2)} Var\tilde{\xi}_m^2.$$

Hence in this case $\left|C_k\left(\tilde{X}_N^2\right)\right| \le (k!)^2 \left(2^{12}\alpha_n^{-1}\right)^{(k-2)} 2n\alpha_n$ and $\left|C_k\left(X_N^2\right)\right| \le (k!)^2 \left(2^{12}\alpha_n^{-1}\right)^{(k-2)} VarX_N^2$. So r.v. X_N^2 satisfy the Statulevicius condition (S_ν) with $\nu = 1$ and $\Delta = 2^{12}\alpha_n^{-1}$. Theorem 3.1 follows from the Assertion 3 with $\zeta = X_N^2$, $\sigma = \sqrt{2n\alpha_n}$ and $\Delta = 2^{12}\alpha_n^{-1}$.

APPENDIX

Let ζ be a r.v. with $E\zeta = 0$, $Var\,\zeta = \sigma^2 > 0$, $C_k(\zeta)$ and $\mu_k(\zeta)$ be, respectively, cumulant and moment of *k* th order of the r.v. ζ. The following two conditions plays essential role in the theory of large deviations, see Saulis and Statulevicius (1991).

Bernstein's condition (B_ν): there exists the constants $\nu \ge 0$ and $B > 0$ such that

$$\left|\alpha_k(\zeta)\right| \le (k!)^{\nu+1} B^{k-2}\sigma^2, \text{ for all } k = 3,4,\ldots.$$

Statulevicius condition (S_ν): there exists the constants $\nu \ge 0$ and $\Delta > 0$ such that

$$\left|C_k(\zeta)\right| \le (k!)^{\nu+1}\Delta^{(k-2)}\sigma^2 \text{, for all } k=3,4,\ldots.$$

Assertion 1 (Saulis &Statulevicius, 1991). If ζ satisfy condition (B_ν) then it also satisfy condition (S_ν) with $\Delta = 2B$.

Assertion 2. Suppose that r.v. ξ depending on a parameter Δ satisfy condition (S_1) (i.e. $\nu = 1$). Then $P\{\zeta > x\sigma\} = \Phi(-x)\left(1 + c\theta\frac{\Delta}{\sigma}(x+1)^3\right)$ for all $x:\ 0 \le x \le \frac{1}{12}\left(\frac{\sigma}{\Delta}\right)^{1/3}$, where $|\theta| \le 1$.

Assertion 2 is Lemma 2.3 with $\nu = 1$ by Saulis &Statulevicius (1991).

Remark. The condition (S_ν) presented by Saulis & Statulevicius (1991) has form $\left|C_k(\zeta/\sigma)\right| \le (k!)^{\nu+1}\Delta^{-(k-2)}$, in such variant of the condition (S_ν) everywhere above it should be written Δ/σ instead of Δ. It seems that presented here slightly different formulation of the condition (S_ν) is more convenient; for example, Assertion 1 now is easy to understand because of the following well known relation (see, for instance Saulis and Statulevicius, 1991, p.15).

Assertion 3. One has

$$C_k(\xi) = k!\sum(-1)^{m_1+m_2+\ldots+m_k-1}(m_1+m_2+\ldots+m_k-1)!\prod_{l=1}^{k}\frac{1}{m_l!}\left(\frac{\mu_l(\xi)}{l!}\right)^{m_l}$$

here $\sum$ is summation over all non-negative integer $m_1, m_2, \ldots, m_k$ such that $m_1 + 2m_2 + \ldots + km_k = k$

Assertion 4. (Petrov (1975)). Let $\zeta, \zeta_1, \zeta_2, \ldots,$ be i.i.d. r.v.s. and $Var\,\zeta = \sigma^2 > 0$, $\beta_3 = E|\zeta|^3$. For all t: $|t| \le \sigma^3\sqrt{n}/6\beta_3$ one has

$$\left|E\exp\left\{\frac{it}{\sigma\sqrt{n}}\sum_{m=1}^{n}\zeta_m\right\} - \exp\left\{-\frac{t^2}{2}\right\}\right| \le C\frac{\beta_3|t|^3}{\sigma^3\sqrt{n}}\exp\left\{-\frac{t^2}{4}\right\}.$$

REFERENCES

Borovkov, A. A., (1977). On the power of chi-square test with increasing number of groups. Theory Probabl. Appl., 22, 375-379.

Fraser, D. A. S.,(1957). Nonparametric Methods in Statistics, John Wiley. New-York.

Gvanceladze, L. G. and Chibisov, D.M. (1979). On tests of fit based on groped data. In Contributions to Statistics, J.Haek Memorial Valume. J.Jurechova, ed. 79-89. Academia, Prague.

Holst, L. (1972). Asymptotic normality and efficiency for certain goodness-of-fit tests. Biometrica, 59,137-145.

Inglot, T., (1999). Generalized intermediate efficiency of goodness-of-fit tests. Math. Methods Statist; 8, p.487-509.

Ivchenko, G. I., Medvedev, Y.I.(1978). Decomposable statistics and verifying of tests. Small sample case. Theory of Probability Appl., Vol.23, 796-806.

Ivchenko, G. I., Mirakhmedov, S. A.(1991).The limit theorems for divisible statistic and efficiency of correspondently tests. Discrete Math. Vol.3, 73-88.

Ivchenko, G. I., Mirakhmedov, S. A., (1995). Large deviations and intermediate efficiency of the decomposable statistics in multinomial scheme. Math. Methods in Statist., Vol.4, 294-311.

Jammalamadaka, S. R. and Tiwari, R. C. (1985), Asymptotic comparison of three tests for goodness of fit. J. Statist. Plan. Inf., 12, 295-304.

Jammalamadaka, S. R. and Tiwari, R. C. (1987), Efficiencies of some disjoint spacings tests relative to a χ^2 tests. In: Puri ML,Vilaplana J, Wertz W. (eds) New Perspectives in Theoretical and Applied Statistics. John Wiley, New York, p.311-318.

Jammalamadaka, S. R. , Zhou, X. and Tiwari, R. C.(1989), Asymptotic efficiencies of spacings tests for goodness of fit. Metrika, 36, 355-377.

Jammalamadaka, S. R and Gatto, R., (2006). Small sample asymptotics for higher-order spacings. Advances in Distributions, Order Statistics, and Inference (Eds. Balakrishnan, N., Castillo, E., And Sarabia J.M.), Birkhauser, 239-252.

Jammalamadaka, S. R and Goria, M. N., (2004). A test of goodness of fit based on Gini's index spacings. Statist.& Probab. Letters, 68**,** 177-187.

Kallenberg, W. C. M., (1983). Intermediate efficiency, theory and examples. Annals of Statistics, 11, 170-182.

Kallenberg, W. C. M. (1985). On moderate and large deviations in multinomial distributions. Ann.Stat.,13, 1554-1580.

Kolchin, V. F., Sevastyanov, B. A. and Chistyakov, V. P. (1978). Random Allocation. John Wiley, New-York-Toronto-London.

Mann, H. B. and Wald, A. (1942). On the choice of the number of intervals in the application of the chi-square test. Ann. Math. Statist. 13, 306-317.

Medvedev, Y. I. (1977), Decomposable statistics in the multinomial scheme. Theory Probabl.Appl., 22, 3-17.

Mirakhmedov, S. A. (1985). Estimations of the closeness to the normal law of the distribution of randomized decomposable statistics in a multinomial scheme. Theory Probabl.Appl.,v.30(1), p.175-178.

Mirakhmedov, S. A. (1987). Approximation of the distribution of multi-dimensional randomized divisible statistics by normal distribution.(multinomial case). Theory Probabl. Appl., 32,p.696-707.

Mirakhmedov, S. A. (1990). Randomized decomposable statistics in the scheme of independent allocations of particles into cells. Discretnaya Mathem. (In Russian) v.2, #2, p. 97-111.

Mirakhmedov, S. A.(1996). Limit theorems on decomposable statistics in generalized allocation scheme. Discrete Math. Appl. Vol.6, 379-404.

Mirakhmedov, S. M.[1] (2005). Lower estimation of the remainder term in the CLT for a sum of the functions of k- spacings. Statist.& Probab. Letters. 73, 411-424.

Mirakhmedov, S. M. (2006). Probability of large deviations for the sum of functions of spacings. Inter. J. Math. Math. Sci.. V.2006, Article ID 58738, 1-22.

Mirakhmedov, S. M. and Naeem, M., (2008). Asymptotic properties of the goodness-of-fit tests based on spacings. Pak. J. Statist., v.24, # 4, p.253-268.

Mirakhmedov, S. M. (2009). On the Greenwood Goodness-of-Fit Test. J. Statist. Plan. Inf.. Submitted.

Morris, C. (1975). Central limit theorems for multinomial sums. Ann. Statist. 3, p.165-188.

Nikitin, Y. Y.(1995). Asymptotic efficiency of Nonparametric Tests. Cambridge Univ.Press.

Petrov, V. V., (1975). Sum of independent random variables, 1 st. ed., Ergebnisse der Mathematik und There Grenegebiete, 82, Springer, New York.

Pyke, R. (1965). Spacings. J.Roy.Statist.Soc. ser. B 27, 395-449.

Quine, M. P. and Robinson, J. (1984). Normal approximations to sums of scores based on occupancy numbers. Ann. Probab. 13, 794-804.

Quine, M. P. and Robinson, J. (1985). Efficiencies of chi-square and likelihood ratio goodness-of-fit tests, The Annals of Statistics, 13 (1985), 727–742.

Rao, J. S. and Sethuraman, J. (1975). Weak convergence of empirical distribution functions of random variables subject to perturbations and scale factors. Ann.Statist., 3 (1975), 299-313.

Ronzhin, A. F. (1984), Theorem on the probability of large deviations for decomposable statistics and its statistical applications. Mathem. Notes. 36, 603-615.

Saulis, L. and Statulevicius, V. (1991). Limit theorems for large deviations. Dordrecht, Boston, London: Kluwer Academic Publishers, p. 232.

Sirajdinov, S. K., Mirakhmedov, S. A. and Ismatullaev, S. A. (1989), Large deviation probabilities for decomposable statistics in a multinomial scheme. Theory Probabl.Appl.,34 706-719.

Zhou, X and Jammalamadaka, S. R. (1989). Bahadur efficiencies of spacings tests for goodness of fit. Ann. Inst. Statist. Math, 28, 783-786.

[1] Former Mirakhmedov S.A.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 898–919.

SOCIO-ECONOMIC CHARACTERISTICS OF HOUSEHOLD HEADS IN EGYPT

Ghada Mostafa
Central Agency for Public Mobilization and Statistics, Salah Salem St.,
Nasr City, Cairo, Egypt
E-mail: ghadaabd@yahoo.com

ABSTRACT

Poverty eradication is the first of the Millennium developmental goals. Reducing poverty is a primary goal of policy makers. While poverty is not a gender concern, studies suggest that women, along with their children, tend to be more vulnerable to poverty than men and among children in poor households, girls are generally more vulnerable than boys. The study aims to identify the characteristics of household headed by women as well as their family members, especially illiteracy rates among children and child labor prevalence in these households. The study shows that, women heading households are older than men and are less likely to participate in the labor force. Most women heading households are widows and widowed women heading households with children are the most disadvantaged in terms of the incidence, depth and severity of poverty. Most women heading households are illiterate and children in these households are more likely to be illiterate and more likely to work at young ages.

1. INTRODUCTION

Poverty eradication is the first of the Millennium developmental goals. Reducing poverty is a primary goal of policy makers. In Egypt poverty is not gender concern, while studies suggest that women, along with their children tend to be more vulnerable to poverty than men. In this context, the continuous and increasing poverty burden on women, which is stated in different international and Arabic documents, is one of the fundamental issues of the world's concern. This situation has led to:

- Construct and implement national strategies to eradicate poverty;
- Ensure that national strategies should focus more on women;
- Establish institutions and national associations directed mainly to support women head of households.
- Assist women's small Projects and Loans.

According to the 2006 Census data, the number of households headed by women in Egypt is about 2.3 million, which constitutes 14% of the total households. The situation might be due to husband absence because of death, divorce, migration multiple marriage or many other reasons like being unmarried. Also it is important to note that the percentage of households headed by women is low due to the traditions and culture prevailing in the Egyptian society especially in rural areas and upper Egypt, where they consider men to be the head of household even in case of being children. While statistics in Egypt suggest that poverty is not particularly feminized, it is important to keep in mind that among the poor women headed households are particularly vulnerable and among children in poor households girls are more vulnerable than boys.

2. OBJECTIVES OF THE STUDY

The objectives of this study are as follows:

- Study the characteristics of household headed by women such as family size, household head age, place of residence, educational level, employment status of household head and members, marital status of household head and economic level of the household.
- Illiteracy rates among children in households headed by women.
- Child labor (6-17 year) prevalence in households headed by women.
- Highlight some recommendations with the goal of formulating policies to reduce poverty among women.

3. SOCIO-ECONOMIC CHARACTERISTICS OF HOUSEHOLD HEADS

This section of the study aims at examining socio-economic characteristics of women-headed households and men-headed households. Census data of 2006 showed that there were 14 % of total household headed by women (6.3 in urban and 7.6 % in rural). This percentage might not reflect the actual situation because of the traditions that household head should be a man even though he is a child and the actual person responsible for the household is a woman. That is why it is important to study socio-economic and demographic characteristics of such households. Table 1 displays the distribution of household heads by gender, place of residence, age, marital, educational, employment status according to 2006 census data (see also Figure 1).

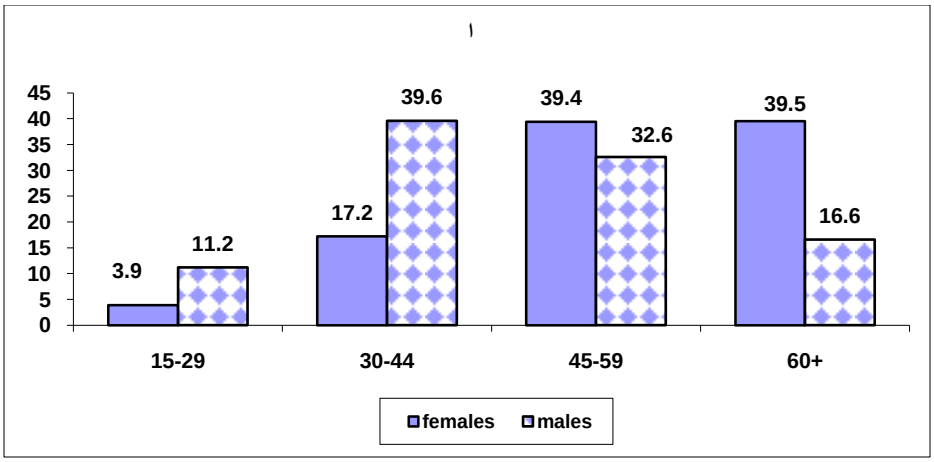
3.1 Age

A review of related literature indicates that women heads of households are more likely to be of old age, widowed, divorced, separated, or with ill husband. This observation was valid according to 2006 census data, where the highest percentage, about 36% of women heads of households, were in age group (60 years or over) while the highest percent (41%) of men heads of households were in the age group (30-44 years). The same pattern was observed in both urban and rural. The data also show higher percentage (12.2%) of young women-headed households (15-29 year) in rural areas compared to urban areas (3.9%).

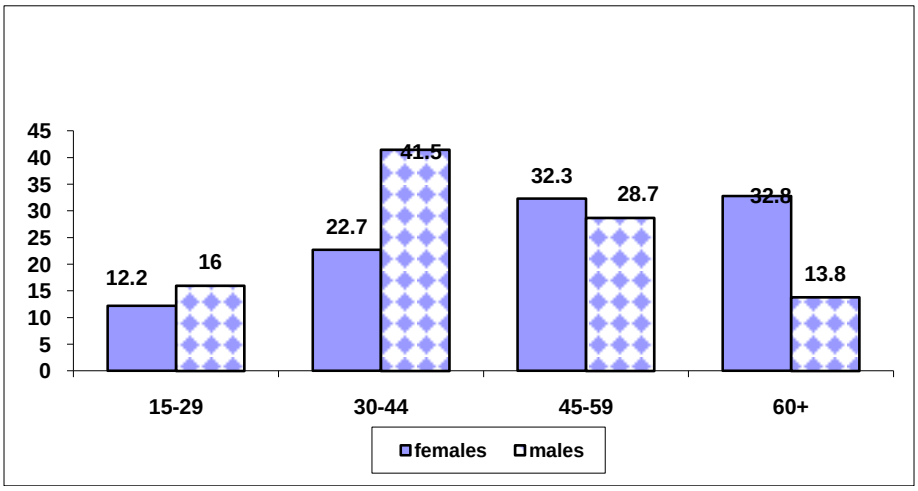
3.2 Marital Status

Previous Studies indicated that the majority of women heads of households are widowed while most of men heads of households are married. This observation is valid according to the 2006 census data, where being a widow is the main reason for this phenomenon. Figure 2 indicated that widowed women constitute about 74% of women headed households, while most of men heads of households (94.2%) are married and there was no significant difference between urban and rural in this pattern.

The data also showed that 25% of women heads of households in rural areas were married which might be due to illness, disability, unemployment or separation of their husbands and in all cases women were totally responsible for the whole family.



Urban Areas



Rural Areas

Figure 1. The distribution of household heads by gender, age and place of residence.

3.3 Educational Status

Household head's educational status is one of the main factors affecting socio-economic characteristics of all household members. Data in Figure 3 indicated that in general, educational status of men heads of household is better than that of women heads of households. More than two thirds of women headed households (69%) in total Egypt were illiterate compared to 34.2% of men. About 13% of women headed households have intermediate and less than university compared to 30% of men.

The same pattern was observed in both urban and rural areas, where illiterate women heads of households represent 56% and 80% in urban and rural areas, respectively, compared to 24% and 43% of men heads of households in urban and rural areas, respectively. Figure 3 also shows that about one third of men heads of households in urban have intermediate and less than university education compared to 16.5% of women heads of households.

Table 1. Percentage of household heads by gender, demographic, socio- economic characteristics, wealth index, 2006

Characteristics	Urban		Rural		Total	
	Men	Women	Men	Women	Men	Women
Age						
15-29	11.2	3.9	16	12.2	13.8	8.4
30-44	39.6	17.2	41.5	22.7	40.7	20.2
45-59	32.6	39.4	28.7	32.3	30.5	35.6
60+	16.6	29.5	13.8	32.8	15.1	35.9
Total	6744893	1099550	8146701	1296575	14891594	2396125
Marital status						
Never married	3.5	3.9	2.3	2.3	2.9	3.0
Contracted	0.2	0.1	0.1	0.1	0.1	0.1
Married	92.6	9.7	95.5	25.0	94.2	18.0
Divorced	0.6	6.5	0.2	4.0	0.4	5.2
Widowed	3.1	79.8	1.7	68.6	2.4	73.3
Total	6745189	1099663	8194681	1296766	14939870	2396429
Educational Status						
Illiterate	23.7	55.8	42.9	80.2	34.2	69.0
Read and write	10.2	9.7	14.4	5.5	12.5	7.5
Primary	4.8	4.4	3.7	1.6	4.2	2.9
Preparatory	5.7	3.7	4.1	1.5	4.8	2.5
Intermediate & less than Univ.	32.7	16.5	27.5	9.5	29.8	12.7
University+	22.9	9.9	7.3	1.6	14.4	5.4
Total	6745189	1099663	8147681	1296766	14892870	2396429
Employment status						
Wage earning	82.7	92.6	91.8	87.6	87.9	90.4
Employer	7.9	3.4	2.4	3.1	4.8	2.3
Self-employed	8	2.2	4.6	5.5	6.0	3.6
Unpaid family workers	0.1	0.1	0.1	1.9	0.1	0.9
Previously worked unemployed	0.7	0.6	0.3	0.2	0.4	0.4
Newly employed	0.6	1.2	0.8	1.8	0.7	1..5
Total	5473794	190546	7177272	147513	12651066	338059
Participation Rate	81.2	17.3	88.1	11.4	85	14.2
Wealth (Index)						
Lowest	9.6	13.9	28.2	38.8	20	27.4
Second	11.0	11.9	26.3	25.2	19.3	19.1
Middle	19.7	21.7	20.1	17.2	19.9	19.3
Fourth	27.1	26.2	14.7	11.5	20.3	18.3
Highest	32.5	26.3	10.8	7.3	20.6	16
Total	6745188	1099663	8147681	1296766	14892869	2396429

3.4 Employment Status

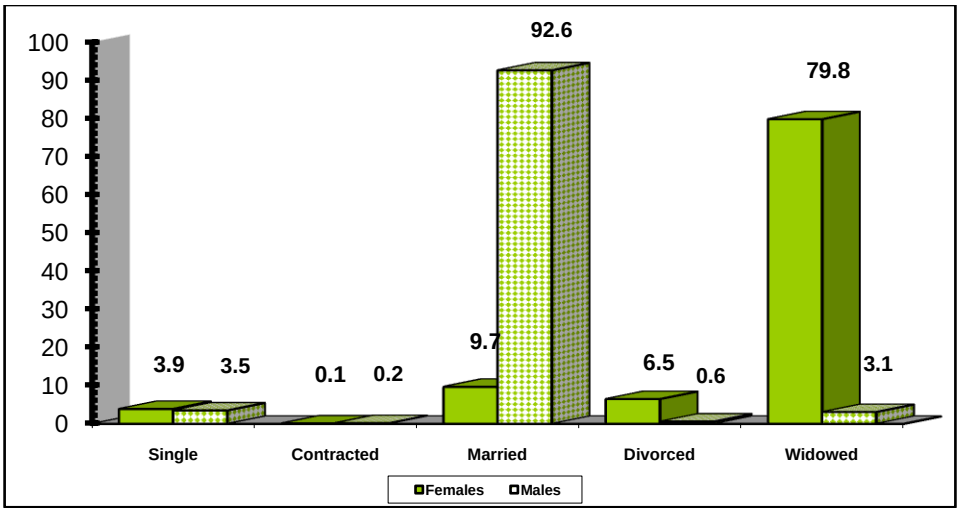
Previous studies showed a significant relationship between household head and household economic status. Figure 4 indicates that the highest percent of household heads, men or women, were wage earning workers with higher percent (90%) among women than among men (88%). It also indicates that in urban areas percentage of women heads of households who are wage-earning were higher (93%) than that of men (83%). A lower percentage (88%) of women heads of households who are wage-earning in rural areas relative to men (92%) was reported. The percentage of women's unpaid family workers constituted the lowest in urban, rural as well as total Egypt. Data also showed that women heads of households have low level of participation in the labor force in the nation as well as in urban and rural areas, where women participation constitutes one fourth of that of men, whereas they constitute one eighth of men participation in rural areas which might be due to the fact that women heads of households were of higher ages.

3.5 Wealth Index

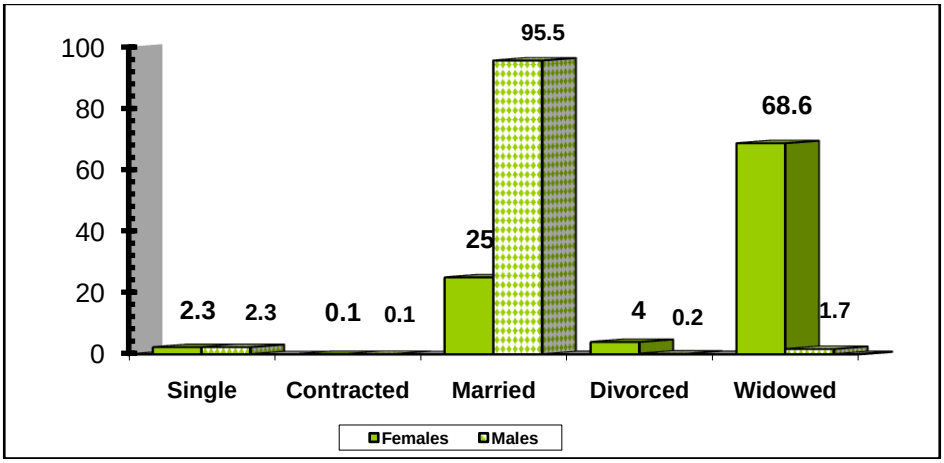
Wealth Index is a tool used to measure the economic level of households. Where this index was constructed based on household proprieties and housing conditions. Wealth index was divided into five levels each constitutes 20%. The Lowest level represents the poorest households, the Second represents the poor ones, the third represents the intermediate, and the fourth level represents the rich. The fifth one represents the richest households. Figure 5 shows significant differences between the economic level of households headed by women and those headed by men where about 27.4% of households headed by women were among poorest group which indicated that women headed households were poorer than men heads of households. Rural areas showed higher percentage of poorest women (39%) and men (28.2%) household heads. The percentage of women and men in the second level (poor) is higher in rural (25.2% for women and 26.3% for men household heads) than in urban (12% and 11% for the two groups respectively). Data also showed higher percentage of urban men household heads in the richest group (about 33% relative to that of women household heads 26.3%).

3.6 Region of Residence

Many statistical studies showed that poverty is not specifically related to women, but we should note that households headed by women are poor and women members of these households are more exposed to poverty than male members. Table 2 gives the distribution of the percentage of household heads by gender, wealth index and region of residence. The data also show that higher percentage of households headed by men that were reported among the poorest group in rural lower and upper Egypt (28.1 and 48.3 for the two groups respectively) compared to urban governorates and urban Lower and upper Egypt (9.3%, 3.4% and 8.3%, respectively. They also indicate that the percentage of households headed by women that is reported among the poorest group in rural Upper Egypt constituted 52% followed by those of rural Lower Egypt (24%). The data show the increase of the percentage of households headed by men that is reported among the most rich and rich group in all regions compared to that of women except for urban governorates (27% and 32% for men compared to 32% and 37% for women.



Urban Areas



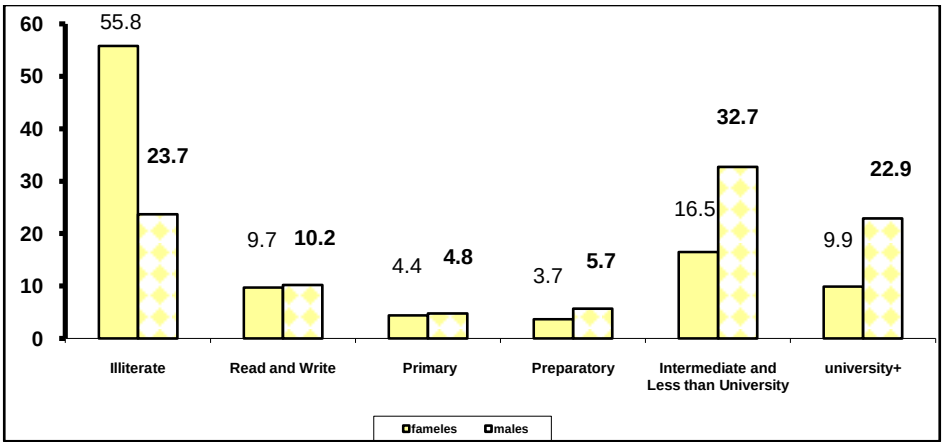
Rural Areas

Figure 2. The distribution of household heads by gender, marital status and place of residence, 2006.

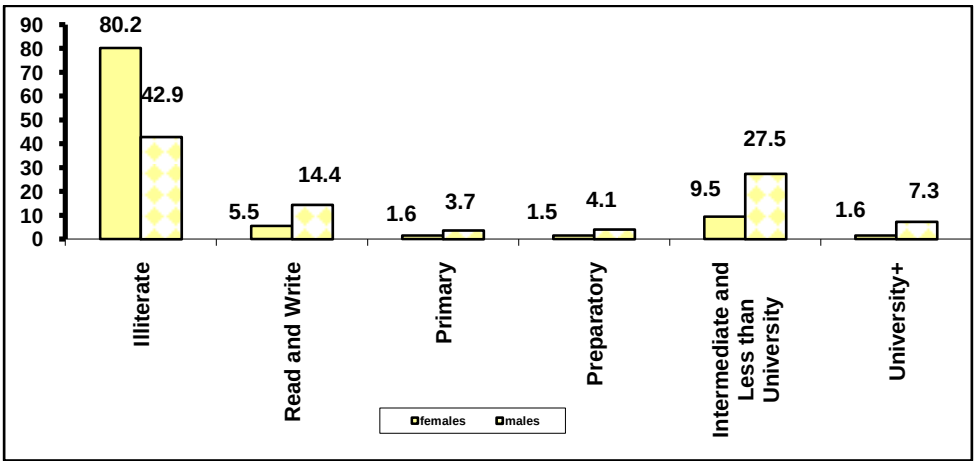
4. DEMOGRAPHIC CHARACTERISTICS OF HOUSEHOLD HEADS

4.1 Average Household Size

Table 3 shows the average household size according to gender and wealth index by place of residence in 2006. Note that there is no significant difference between the wealth index (the economic level of household), size and household composition. Note also that the average size of the households headed by women in urban areas is about three persons, compared to those headed by men, which is about four persons. Concerning rural area, the size of households headed by women is about three persons compared to about five persons in the households headed by men, mainly in the highest economic level. It is also noticed that when the household size increases the economic level increases. This might be because of the number of working persons in the household.



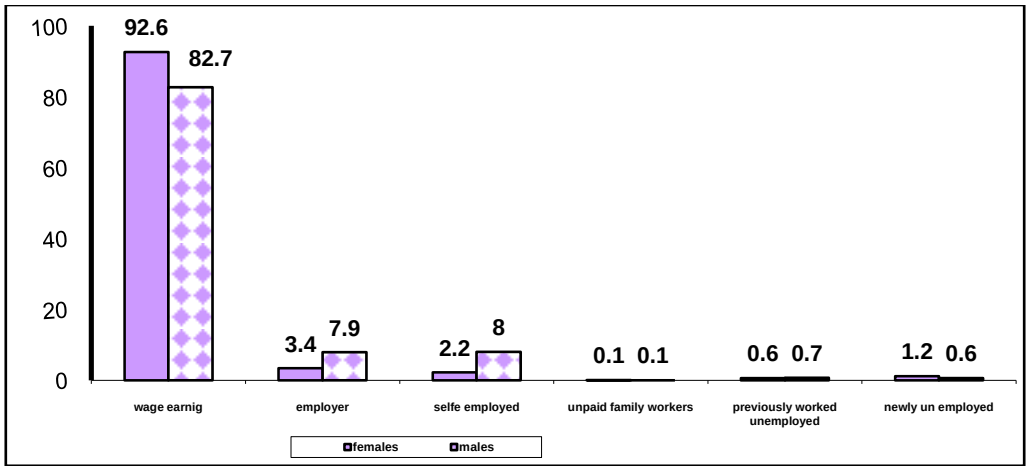
Urban Areas



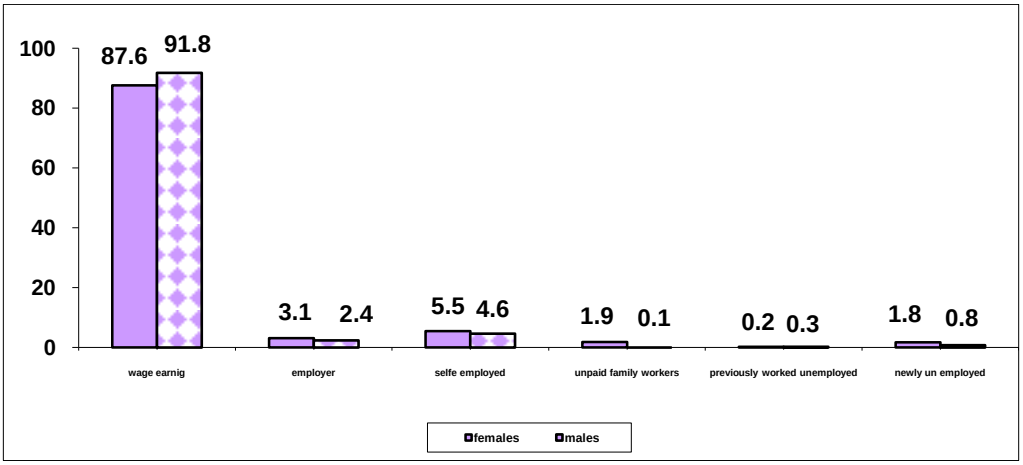
Rural Areas

Figure 3. The distribution of household heads by gender, educational status and place of residence, 2006.

The data in Table 3 give the average number of children (less than 14 years old) in the households headed by males and females in urban and rural areas for different categories of wealth indicators. Also, Table 3 shows that adult persons aged from (14-64) in households headed by women in urban areas reached two persons mainly in the richest households. There is no difference with respect to economic level and number of adults in household in the households headed by men. It is noticed that the average number of adults is about three persons. The same pattern appears in rural areas. Concerning number of ageing persons whether in households headed by men or women in urban areas is less than those in rural areas. This might be explained by the type of family, where extended families are prevailed in rural communities compared to urban areas regardless of the economic level of the household.



Urban Areas



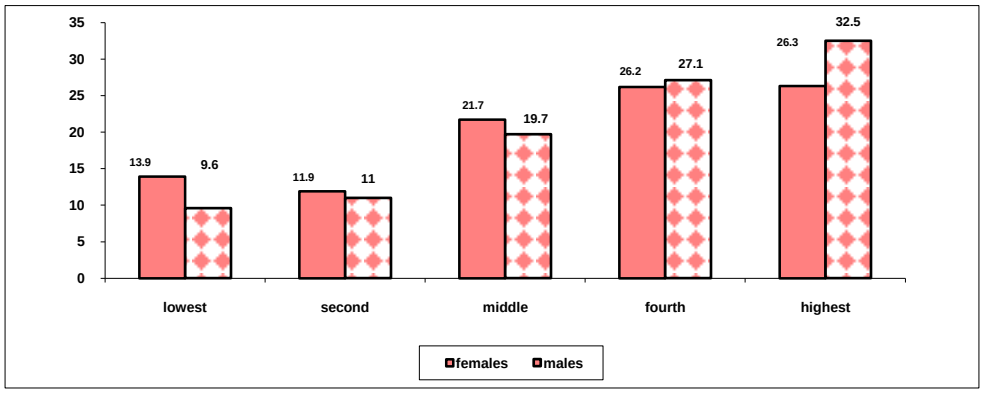
Rural Areas

Figure 4. The distribution of household heads by gender, employment status and place of residence, 2006.

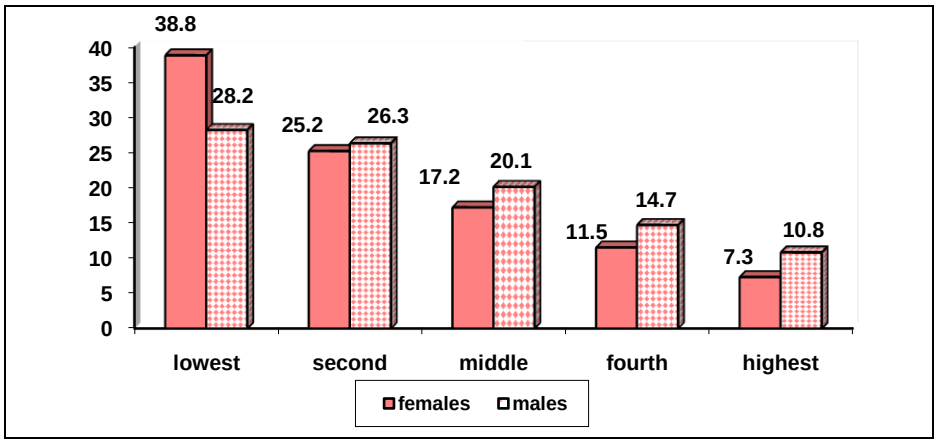
4.2 Household Structure

Household structure affects income distribution among household members and consequently affects the welfare of those members. Often, dependency ratio in poor households is higher than in rich households and in rural areas higher than the level in urban areas. Poor household feel the necessity to have more children as they are considered an additional source for household income. Table 4 shows the distribution of household heads by household structure and place of residence. It indicates that households headed by women alone without children are higher than that headed by men. Also, the highest percentage of women headed households with one to three children was more observed among widowed women in both urban and rural areas. The percentage is 42.9% in urban areas as compared to 32.1% in rural areas, and the percentage of widowed women without children in urban areas is 28.9% and in rural areas is 27.3%. The percentage of households headed by widowed women with more than three children in rural areas is 9.2% compared to 8.1% in urban areas. Concerning the households headed by men, the percentage of married men who have from one to three children reached (62.5%) and (16.7%)

among married men who have more than (3) children in urban areas. The same pattern was revealed in rural areas.



Urban Areas



Rural Areas

Figure 5. Distribution of household heads by gender, wealth index and place of residence, 2006.

5. THE RELATIONSHIP BETWEEN HOUSEHOLD ECONOMIC LEVEL AND CHARACTERISTICS OF HOUSEHOLD MEMBERS

The 1971 Constitution stipulates upon the complete equality between men and women in education. Basic education is compulsory (primary and preparatory stage), and tuition is free. The enrolment rate of girls is lower than that of boys. According to the Ministry of Education's data, the enrolment rate reached 94.3% and 96.5% among girls and boys respectively in 2007/2008.

Although there is a remarkable improvement in the field of education, there are a lot of factors that contributes in the reduction of woman benefit from these efforts. Also there are a lot of factors related to the basic infrastructure of the educational process as the location of the schools and the quality of the educational buildings, also the availability of the transportation, besides some characteristics of the household such as income and the prejudiced view against the education of women. The poor girls especially in the rural areas and in Upper Egypt fail to the

other boys and girls standard. From what is mentioned previously, it appears the importance of the studying the relation between the educational level of household members and the educational level of household head.

Table 2. The distribution of household heads by gender, region of residence and wealth index, 2006

Wealth Index	Urban Governorates	Urban Lower	Rural Lower	Urban upper	Rural upper
Men					
Lowest	9.3	3.4	28.1	8.3	48.3
Second	8.9	5.2	37.7	10.6	36.1
Middle	19.4	12.1	37	12.2	17.9
Fourth	27.1	18.7	31.2	13.7	8.1
Highest	32	23.7	23.9	14.7	4.4
Total	19.5	12.8	31.5	11.9	32.7
Women					
Lowest	9.7	4	24	9	51.9
Second	10.2	6.2	33.1	11.6	38
Middle	23	15.1	32.6	12.8	15.6
Fourth	31.9	20.2	27.5	13.4	6.4
Highest	37	23.5	21	14.3	3.7
Total	20.8	12.6	27.5	11.9	26.2

5.1 The Educational Status of the head of Household Chief and the Standard of Children's Educational level

Table 5 shows the distribution of the percentage of household heads according to gender, place of residence, and educational level of household heads and household members in 2006. The data indicate that there is a direct relationship between the educational status of household heads and the educational status of the household members especially in urban areas regardless of the gender of the household head.

About 39% of the men household heads who are university graduates or higher in urban areas, their household members have acquired the same educational level. While this percentage among household headed by women reached about 35%. On the other hand whenever the educational level of the household heads decreased, the possibility of the household members being illiterate increased. Data showed that 41% of illiterate men household heads in urban areas, are associated with illiterate household members, and this rate decreased among household headed by women in urban areas.

Table 3. Average size of households by the head of the household, wealth indictor and place of residence in 2006.

Wealth index-	Urban			Rural		
	M	F	T	M	F	T
Average size of households						
Lowest	**4**	**2.5**	**3.7**	**4.5**	**2.5**	**4.1**
Second	**4.2**	**2.7**	**4**	**4.6**	**3.1**	**4.4**
Middle	**4.1**	**2.7**	**3.9**	**4.7**	**3.1**	**4.5**
Fourth	**4.1**	**2.8**	**3.9**	**4.6**	**3.2**	**4.5**
Highest	**4.2**	**2.9**	**4**	**4.8**	**3.5**	**4.6**
Total	**4.1**	**2.7**	**3.9**	**4.6**	**2.9**	**4.4**
Average number of children						
Lowest	**1.3**	**0.4**	**1.1**	**1.6**	**0.7**	**1.5**
Second	**1.4**	**0.5**	**1.2**	**1.6**	**0.9**	**1.5**
Middle	**1.3**	**0.4**	**1.1**	**1.6**	**0.9**	**1.5**
Fourth	**1.2**	**0.4**	**1.1**	**1.5**	**0.9**	**1.5**
Highest	**1.2**	**0.5**	**1.1**	**1.5**	**1.1**	**1.5**
Total	**1.2**	**0.5**	**1.1**	**1.6**	**0.8**	**1.5**
Average number of adults						
Lowest	**2.6**	**1.7**	**2.4**	**2.7**	**1.5**	**2.4**
Second	**2.7**	**1.9**	**2.6**	**2.9**	**2**	**2.7**
Middle	**2.7**	**2**	**2.6**	**3**	**2**	**2.9**
Fourth	**2.8**	**2.1**	**2.7**	**3**	**2.1**	**2.9**
Highest	**2.9**	**2.2**	**2.8**	**3.1**	**2.3**	**3**
Total	**2.8**	**2**	**2.7**	**2.9**	**1.8**	**2.7**
Average number of aging						
Lowest	**0.1**	**0.3**	**0.2**	**0.2**	**0.3**	**0.2**
Second	**0.1**	**0.3**	**0.2**	**1**	**2**	**0.2**
Middle	**0.1**	**0.3**	**0.2**	**1**	**2**	**0.1**
Fourth	**0.1**	**0.3**	**0.2**	**1**	**2**	**0.1**
Highest	**0.1**	**0.2**	**0.1**	**1**	**2**	**0.1**
Total	**0.1**	**0.3**	**0.2**	**1**	**3**	**0.2**

Table 4. The distribution of household heads according to the household structure and place of residence, 2006

Household Structure	Urban		Rural		Total		Total
	F	M	F	M	F	M	
Married without children	1.4	13.4	2.8	12.8	2.1	13.1	1999290
Married and have from (1-3)child	0.7	62.5	1.53	56.5	8.6	59.2	9087796
Married and have more than 3 child	1.8	16.7	7	26.2	4.6	21.9	3372001
Widowed without children	28.9	1.7	27.3	0.8	28	1.2	847292
Widowed and have from (1-3)child	42.9	1.3	32.1	0.8	37	1	1035016
Widowed and have more than 3 children	8.1	0.2	9.2	0.2	8.7	0.2	236525
Other	10.5	4.3	6.4	2.8	8.3	3.4	711379
Total	1099663	6745189	1296766	8147681	2396429	14892870	17289299

It can also be noticed that among household headed by women, knowing how to read and write, have a positive influence on the education of their household members. Therefore, it was found that the percentage of women headed household who can read and write and their household members have acquired intermediate and less than university education is about 43%. The same percentage for the women household heads who acquired the primary education was reported, the same pattern was observed in rural areas. This indicates that women who acquire a simple extent of education are eager to support their household members towards education.

5.2 Illiteracy among Children aged 10 -17 years

Table 6 gives the percentage of illiteracy among children (10-17 years) by gender, household heads, wealth index, and place of residence. The data reveal that there is a strong negative relationship between the economic level of the household as measured by the wealth index and percentage of illiteracy among children in the family.

Table 6 also shows that the total percentage of illiterate boys in the household with low economic level reached about 39%, 44% compared to about 44%, and 49.3% among the girls in the households that is supported by men and women respectively. More than half of girls in the households that are supported by women in rural areas (52.8%) who live in low economic standards suffer from illiteracy as compared to 47.8% among boys in households supported by women and have the same economic standard.

Table 5. The Distribution of Household Heads According to Gender, Educational Status, Residence and the Educational Status of the Household Members, 2006

Educational status	Illiterate	Read & write	Primary	Preparatory	Intermediate & less than university	University & Higher	%	Total
Men in Urban Areas								
Illiterate	40.7	10.9	12.2	10.9	22.2	3.1	100	4184027
Read& write	20.5	21.4	12.4	12.4	28.1	5.2	100	1583146
Primary	14.1	11.4	20.8	14.8	30	5.8	100	795216
Preparatory	7.2	11.7	16	19.1	32.8	6.3	100	897650
Intermediate & less than university	3.3	11.2	11.5	13	46.8	10.1	100	4549542
University & higher		15.2	9.7	10.5	30.1	38.9	100	3290242
Women in Urban Areas								
Illiterate	19.8	9.3	11.5	13.1	37.1	9.1	100	906681
Read& write	4.7	10.6	10.0	13.4	43.3	17.9	100	151680
Primary	4.2	6.3	12.8	14.6	42.7	19.5	100	76203
Preparatory	3.6	7.2	10.9	15.9	40.6	21.8	100	63168
Intermediate & less than university	2.9	8.9	11.3	14.8	38.4	23.7	100	274699
University & Higher	2.3	7.4	10.3	13.4	31.7	34.8	100	154754

In general the percentage of illiteracy among children whether men or women decreases by the increase in the economic level of the household, so this percentage in the households with higher economic level and especially in rural areas about 3.8%, and 5.1% between the girls and boys, respectively in the households supported by men in rural areas and is about 2.4%, 3.5% between girls and boys, respectively in the households supported by women in rural areas at the higher economic standard.

5.3 Child Labor

Children in poor families are more exposed to labor market to face the hard economic circumstances of their families. Child labor is affected by the region (urban-rural) and also by the gender of household head.

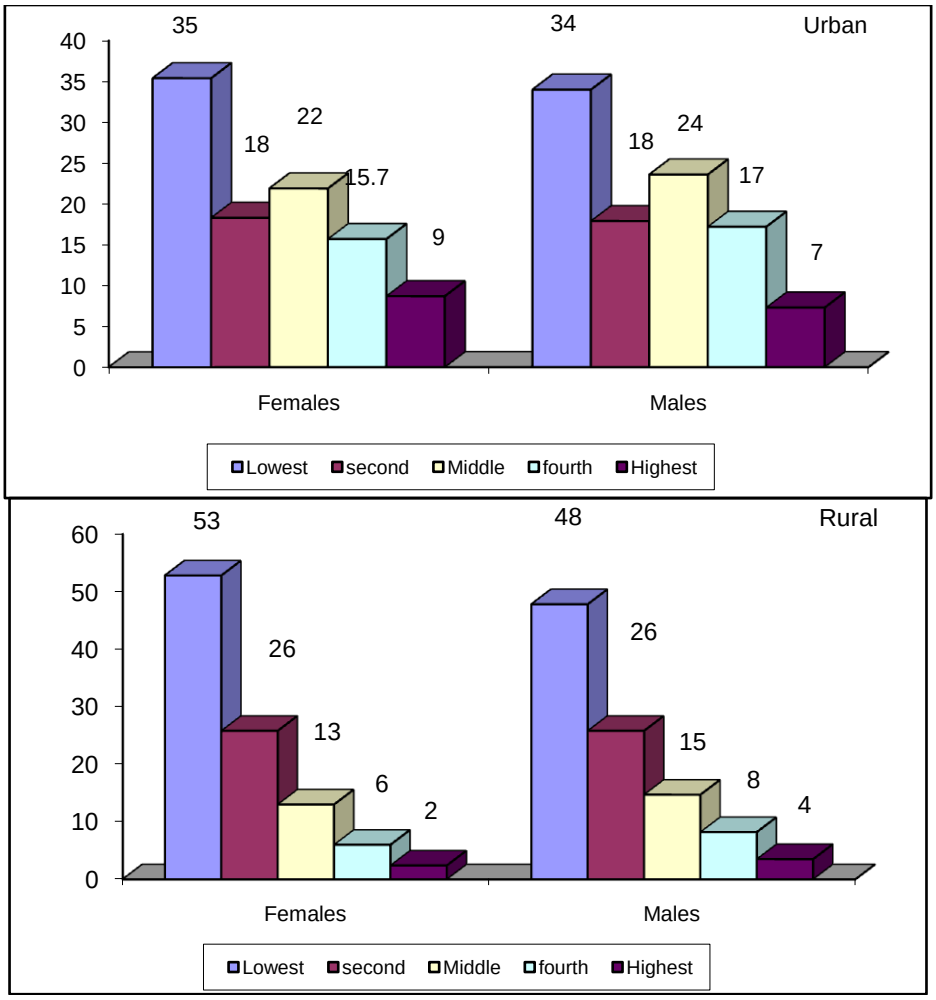
Although laws no. 127 of 1981 and child law of 1996 prevent children below age 14 from work or even training, there is a proportion of working children in households headed by women while this percentage is lower among households headed by men.

Table 5 (Cont.). The Distribution of Household Heads According to Gender, Educational Status, Residence and the Educational Status of the Household Members, 2006

Men in Rural Areas								
Illiterate	47.1	12.5	12	10.1	16.7	1.6	100	9916183
Read& write	30.8	19.1	12.9	12	22.3	2.9	100	1702098
Primary	28.1	14.1	18.6	13	23.4	2.7	100	723155
Preparatory	23.5	14.8	15.6	15.7	27.2	3.3	100	782831
Intermediate & less than university	16.4	15	12.2	12.2	40	4	100	4467229
University &higher	7.7	13.1	11.3	11.8	34.1	22	100	1252733
Women in Rural Areas								
Illiterate	23.7	14	14.4	14.6	29	4.2	100	1531204
Read& write	6.7	17.3	15.4	17	25.4	8.3	100	96550
Primary	5.5	18.4	19.6	18.1	30.5	7.8	100	29777
Preparatory	5.7	21.3	19.8	20.3	25.9	7	100	22626
Intermediate & less than university	5.5	21.4	18.4	19	27.6	8	100	116650
University & Higher	5.2	17.7	16.4	18.6	29.2	13	100	18864

Table 6. Percentage of Illiteracy among Children (10-17 Years) According to Gender, Household Heads, Wealth Index, and Place of Residence, 2006

Place of residence	Urban		Rural		Total	
Wealth Index	F	M	F	M	F	M
Men support households						
Lowest	**23.1**	**28.7**	**46.5**	**42.5**	**43.9**	**38.8**
Second	**19.5**	**18.6**	**26.6**	**26**	**25.3**	**24**
Middle	**21.6**	**23.1**	**15.3**	**16.6**	**16.5**	**18.4**
Fourth	**16.8**	**19**	**7.8**	**9.8**	**9.5**	**12.3**
Highest	**10**	**10.6**	**3.8**	**5.1**	**5**	**6.6**
%	**100**	**100**	**100**	**100**	**100**	**100**
Total	**79905**	**84102**	**353411**	**225471**	**433316**	**309573**
Women support households						
Lowest	**35.4**	**34**	**52.8**	**47.8**	**49.3**	**43.5**
Second	**18.3**	**17.9**	**25.8**	**25.8**	**24.3**	**23.3**
Middle	**21.9**	**23.6**	**13**	**14.7**	**14.8**	**17.5**
Fourth	**15.7**	**17.2**	**6**	**8.2**	**8**	**10.9**
Highest	**8.7**	**7.3**	**2.4**	**3.5**	**3.7**	**4.7**
%	**100**	**100**	**100**	**100**	**100**	**100**
Total	**11376**	**12853**	**44128**	**28709**	**55504**	**41562**

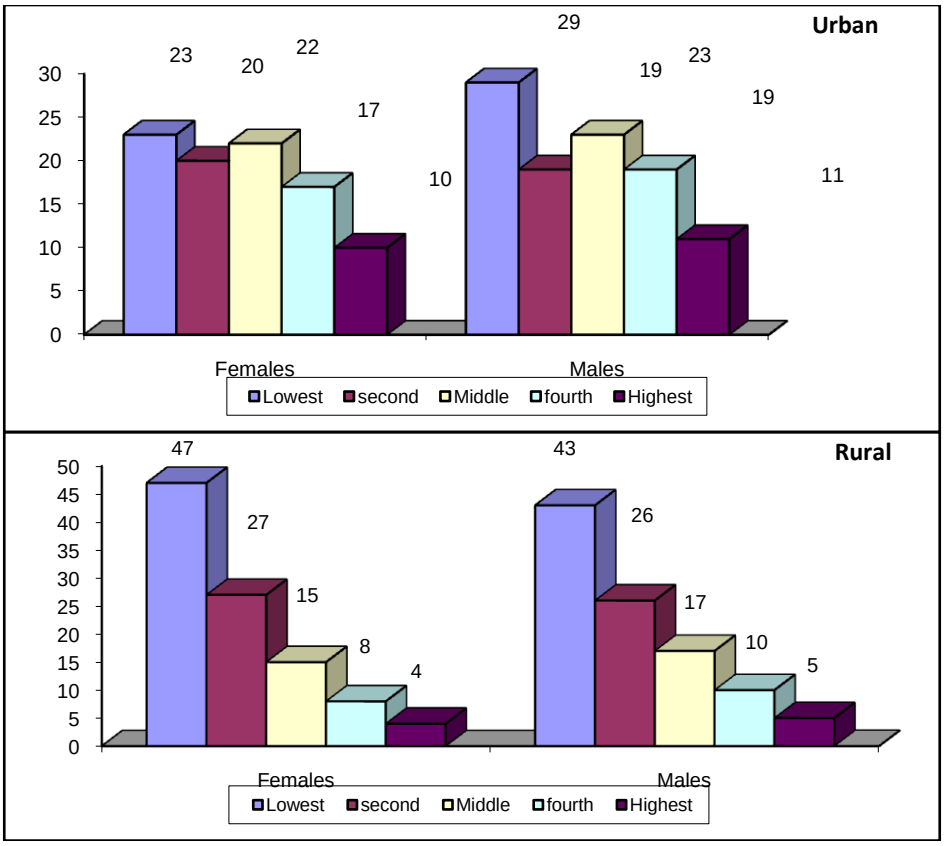


Women-Headed Households

Figure 6(a). Percentage of Illiteracy among Children from (10 -17) years, According to gender, wealth index, and place of residence (Women-Headed Households)

Table 7 reveals high percentage of child labor in the age group 6-13 in households headed by women in urban areas (3.1%) and rural areas (2.7%); this can also be seen in Figure 7. The table also indicates a decrease in the percentage of child labor among child aged 6-13 with increase in the economic level of households headed by women in both urban and rural areas.

The percentage of child labor increase with the increase in the age of the child where it is 20% among boys aged 14-17 in both urban and rural compared to 3.6% and 5% among girls in the same age group in urban and rural respectively. In general, data in Table 6 also show a decrease in the percentage of child labor with the increase in the economic level of the household in urban and rural as well as for men and women.



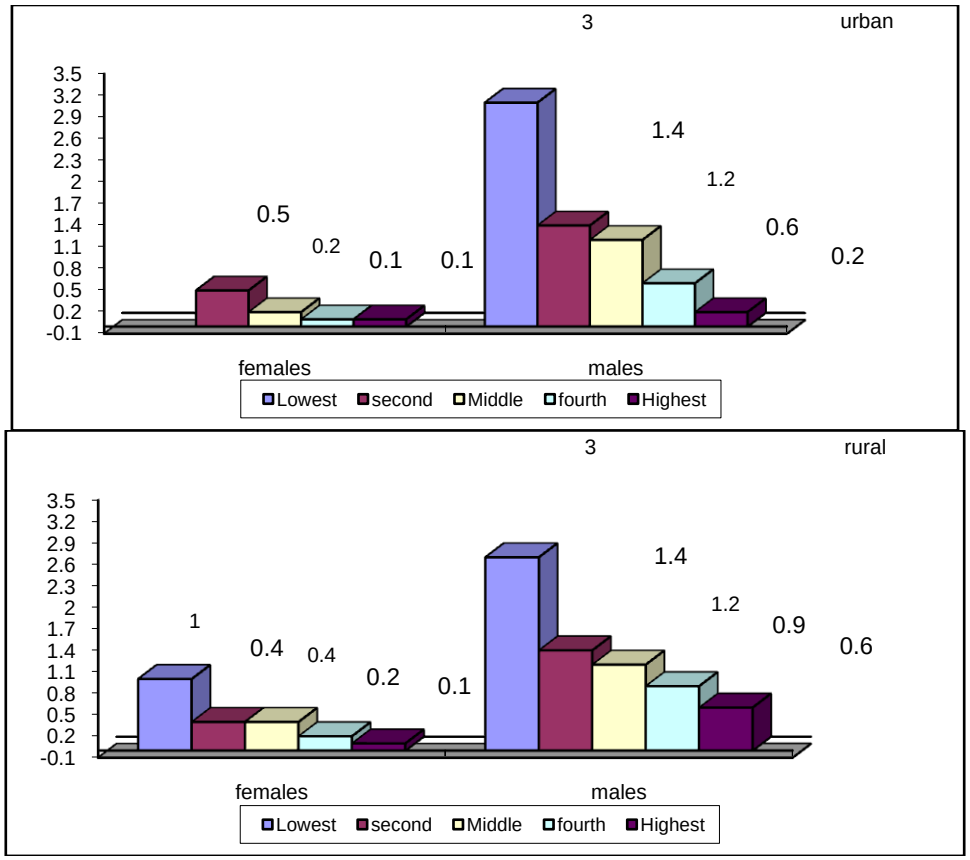
Men-Headed Households

Figure 6(b). Percentage of Illiteracy among Children from (10 -17) years, According to gender, wealth index, and place of residence (Men-Headed Households)

Table 7. The percentage of children employment in age (6 to 17) for families headed by women regarding to wealth index and place of residence in year 2006

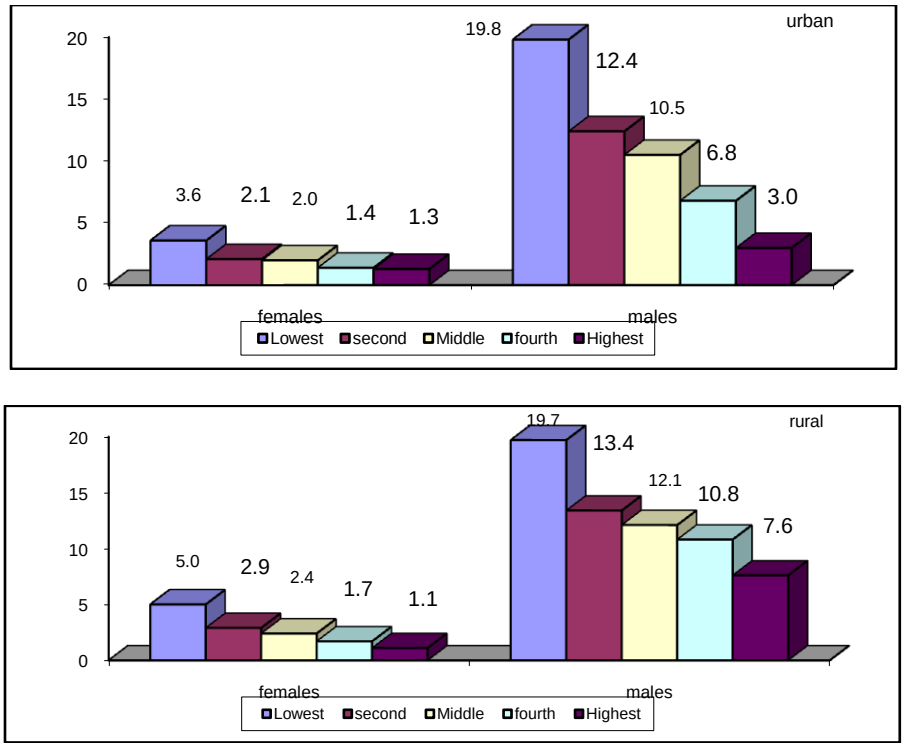
Place of Residence / Wealth Index	Urban		Rural		Total	
	M	F	M	F	M	F
Under age 14 (6-13)						
Lowest	3.1	0.5	2.7	1.0	2.7	0.9
Second	1.4	0.2	1.4	0.4	1.4	0.4
Middle	1.2	0.1	1.2	0.4	1.2	0.3
Fourth	0.6	0.1	0.9	0.2	0.8	0.1
Highest	0.2	0.3	0.6	0.1	0.4	0.3
Total	1.1	0.2	1.7	0.6	1.5	0.4
(14-17) years						
Lowest	19.8	3.6	19.7	5	19.7	4.7
Second	12.4	2.1	13.4	2.9	13.2	2.7
Middle	10.5	2	12.1	2.4	11.4	2.3
Fourth	6.8	1.4	10.8	1.7	8.6	1.6
Highest	3	1.3	7.6	1.1	4.6	1.2
Total	8.9	1.9	14.2	3.2	12.2	2.7

Figure 8 indicates an increase in the percentage of boys working in the households headed by men compared to that among households headed by women in both urban and rural areas. Data also showed an increase in the percentage of boys working in rural areas than urban areas for both genders. This percentage increase was associated with the increase in age of children. The data show an increase in the percentage of working children in such households in the age group (7-14) as census data showed where this percentage was 17.3% for men in urban and 19% in rural areas. For women, it was 3% in urban and about 5% in rural areas among households in the lowest economic level or those which are headed by men.



Age 6 to **13**

Figure 7(a). The percentage of children employment in age (6 to **13**) for families headed by women regarding to wealth index and place of residence in year 2006 (Age 6 to **13**).



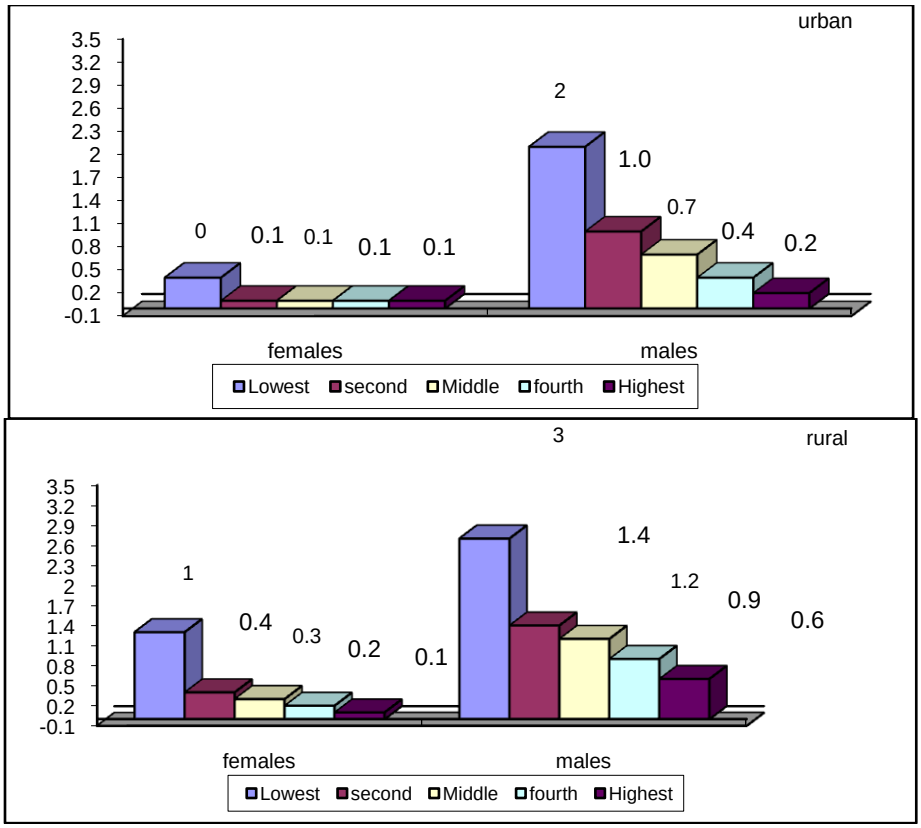
Age 14 to 17

Figure 7(b). The percentage of children employment in age (14 to 17) for families headed by women regarding to wealth index and place of residence in year 2006 (Age 14 to 17).

6. Summary of Results and Recommendations

6.1 Important Results

1. Households headed by women represent small proportion out of the total households. These households are characterized by different characteristics that differ from the characteristics of households headed by men. Women are older than men, so they have a lower ability to participate in labor market and they depend largely on pensions and remittances.
2. Husband's death is the main reason behind the households headed by women where 76.8% of the households headed by women are widowed in urban areas and 68.6% in rural areas.
3. Most households headed by women are widowed and have children. While the households headed by men are married and have children.
4. Educational level of the household heads is one of the main factors behind poverty. In urban areas 49.4% of households headed by women are poor and illiterate compared to 60.9% in rural areas.
5. Households with children are the most suffering from severe poverty.
6. Poverty rates increased significantly in Lower Egypt and rural areas more than other areas of Egypt.
7. Households headed by women are suffering low economic standards in urban areas while in rural areas they are suffering from severe poverty.



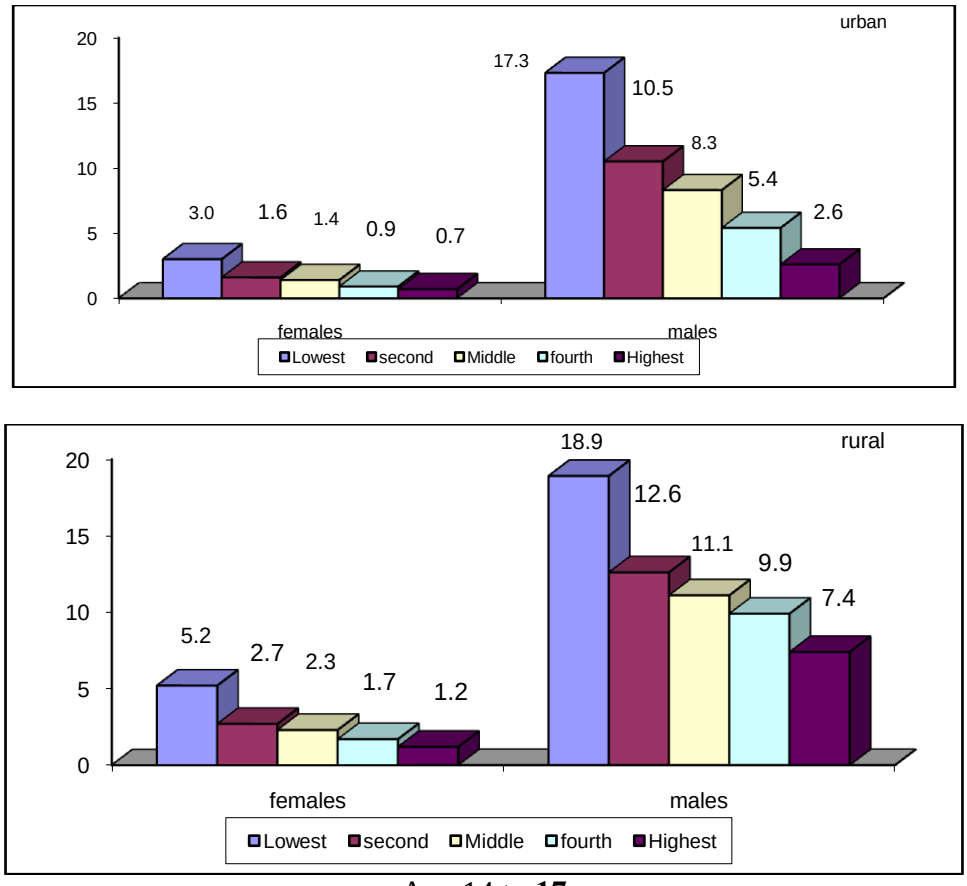
Age 6 to **13**

Figure 8(a). The percentage of children employment in age (6 to **13**) for families headed by women regarding to wealth index and place of residence in year 2006 (Age 6 to **13**).

8. Poverty leads to transmitting the low educational level from one generation to the other as illiterate women heading their households are usually associated by illiterate children causing them to enter labor market early.
9. The majority of households headed by women tend more to get children to labor market which leads to a low educational level and continued suffering from poverty. The high ratio of illiteracy and child labor are the main factors for transmitting poverty from one generation to the other.
10. The percentage of girls in labor market is less than boys where girls are kept for the housework.

6.2 Recommendations

1. Designing new strategies to eradicate illiteracy is a necessity– especially for women.
2. Raising the quality (rather than coverage) of education is a necessity. The educational processes have to include values and concepts that are capable of changing the women's status and role in the society.
3. Raising interest in the technical education for women and helping them acquire the skills required for labor market.



Age 14 to **17**

Figure 8(b). The percentage of children employment in age (6 to **13**) for families headed by women regarding to wealth index and place of residence in year 2006 (Age 14 to **17**).

4. Reducing illiteracy rates among women and developing informal education programs.
5. Providing training opportunities for women especially for household heads to help them find suitable job opportunities and to develop their productive and marketing capabilities.
6. Exerting more efforts by the state, national associations and civil sector to provide job opportunities for women heads of households to provide a source of income particularly in rural areas.
7. Setting objectives through insurance system to fulfill the needs of different poor groups. Assuring equal economic opportunities for both men and women who are heads of households such as the financial aids, employment and small projects.
8. Increasing development programs directed particularly to the poorest women in rural areas and particularly heads of households. These programs have to direct their inputs according to education, health and economic priorities defined by the poorest households in rural areas.

Table 8. The percentage of children employment in age (6 to 17) for families headed by men regarding to wealth index and place of residence in year 2006

Place of Residence / Wealth Index	Urban		Rural		Total	
	M	F	M	F	M	F
Under age 14 (6-13)						
Lowest	2.1	0.4	2.8	1.3	2.7	1.1
Second	1	0.1	1.4	0.4	1.3	0.4
Middle	0.7	0.1	1.1	0.3	1	0.2
Fourth	0.4	0.1	0.9	0.2	0.6	0.1
Highest	0.2	0.1	0.6	0.1	3.	0.1
Total	0.6	0.1	1.6	0.6	1.2	0.4
(14-17) years						
Lowest	17.3	3	18.9	5.2	18.6	4.8
Second	10.5	1.6	12.6	2.7	12.2	2.5
Middle	8.3	1.4	11.1	2.3	10.1	2.0
Fourth	5.4	0.9	9.9	1.7	7.6	1.3
Highest	2.6	0.7	7.4	1.2	4.3	0.9
Total	6.7	1.2	12.9	2.9	10.5	2.3

REFERENCES

Central Agency for Public Mobilization and Statistics, 2008: "Labor force Sample Survey 2008" Cairo, Egypt.

Central Agency for Public Mobilization and Statistics, 2007: "Population, Housing and Establishment Central, 2006:" Cairo, Egypt.

Cairo, Demographic Central, 2001:" Population and Sustainable Development and Challenges of Demographic in Third World Countries" The Thirty-Annual Conference, November 2000, Cairo, Egypt.

El Zanati, F.et al, 2005, "Demographic Health Survey 2005", Cairo, Egypt.

Handoussa, H. 1994: "Economic Framework for Policies Affecting Women Heads of Households in Egypt" presented in Central Agency for Public Mobilization and Statistics, Conference March 20, 1994.

National Council for Women 2005: "The Development of The Situation of Women in the Mubarak era From1981 to 2004", Cairo, Egypt.

National Council for Women 2004: the Fourth Conference of the National Council for Women "Egyptian Women and the Millennium Development Goals", Cairo, - March 2004.

National Council for Women 2003, "Statistical Report on the situation of Egyptian Women" Cairo, Egypt.

National Center for Social Criminological Research 2002: "Women Households in The slum" Cairo, Egypt.

National Council for Women 2000: "Egyptian Women and National Plan 2002-2007" the second conference of the National Council for Women, Cairo, Egypt.

Nasser, H., 2002:"Soial and Economic Policies to Reduce Poverty in Egypt", Center for Research and Economic Studies- Cairo- Egypt.

Population Studies and Research Center, 2009: semi annual book "The Status of Women and Men in Egypt", CAPMAS Cairo, Egypt.

The Ministry of Education, 2008:"Statistial Yearbook 2072008" Information Center, Cairo, Egypt.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 920–920.

ON THE NECESSARY CONDITIONS FOR ERGODICITY of SMOOTH THRESHOLD AUTOREGRESSIVE (1) PROCESSES WITH GENERAL DELAY PARAMETER

D. Nur
School of Mathematical and Physical Sciences
The University of Newcastle Callaghan, NSW 2308, AUSTRALIA
E-mail: Darfiana.Nur@newcastle.edu.au

and

G. M. Nair
School of Mathematics and Statistics
The University of Western Australia, Perth, WA 6009, AUSTRALIA
E-mail: gopal@maths.uwa.edu.au

ABSTRACT

This paper establishes a necessary condition for ergodicity for the general first-order Smooth Threshold Autoregressive processes with general delay parameter *d*. This is achieved by investigating the non-linear dynamic behavior generated by the delay parameter of a smooth threshold model. It turns out that the ergodic region depends on the delay parameter *d* for which the region is reduced as *d* increases.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 921–930.

FAVORABLE CLIENTELE EFFECT AND THE VALUATION MULTIPLES OF ISLAMIC FINANCIAL INSTITUTIONS IN THE UNITED ARAB EMIRATES

M. F. Omran
Business School, Nile University, Egypt
E-mail: mfomran@nileuniversity.edu.eg

ABSTRACT

We extend the study of Omran (2009) by employing different valuation measures to examine the effect of Islamic values and beliefs on the stock valuation of Islamic financial institutions in the stock markets of the United Arab Emirates (UAE) during the period from 2001 to 2005. The current study examines whether UAE investors favor Islamic financial institutions in comparison with traditional financial institutions and other companies in the economy. The difference between the current study and Omran (2009) is that we employ the price to sales and price to book value multiples instead of price to earnings multiple used in Omran (2009). The contribution of the current study is in confirming whether the results of Omran (2009) hold regardless of the valuation multiple used. It is found that there is a strong clientele preference for Islamic stocks in the UAE despite the modest financial performance achieved by these stocks. UAE investors were willing to give up the high return available on other stocks for the comfort of investing in stocks that closely follow Islamic laws.

Keywords: Islamic Financial Institutions, Valuation, Clientele Preference

1. INTRODUCTION

The interest in Islamic finance has been growing for several years. Al-Salem (2008) states that the average annual growth in the assets of Islamic financial institutions has been 23% since 1994. The interesting trend in the last few years is the growing interest of non-Muslims in Islamic financial products and institutions. The interests of non Muslims is obviously not due to religion but it is rooted in the fact that many of the high risk endeavors taken by traditional financial institutions are not allowed at all under Islamic finance. Many of the highly risky exposures that led to huge losses for traditional banks would not be allowed under Islamic laws which consider them pure forms of gambling. However, the variety of Islamic financial products failed to grow at the same rate as the growth in interest in Islamic finance. The fact remains that there are not many products that comply with Islamic laws that can absorb the massive flow of funds. That could have an impact on profitability of Islamic financial institutions since they have to accept deposits that they may not have a use for. There is very little written on how Islamic financial institutions perform in comparison with traditional financial institutions.

Omran (2009) was the first study to examine the financial performance of Islamic financial institutions compared with traditional financial institutions in the United Arab Emirates. His results indicate that the return on equity of Islamic financial institutions lagged behind the UAE stock markets for the period from 2001 to 2005. However, Islamic financial institutions price earnings multiples were the highest in the market despite the poor financial performance. This

can only be attributed to the clientele effect of the UAE investors. They prefer to pay a premium for their faith namely for the comfort of knowing that their money is in full compliance with Islamic laws and regulations. The most obvious reason for the low return on equity is the lack of financial products that comply with Islamic regulations. Therefore, most of the funds lie idle which in turn reduces the return on equity. Return on equity has three drivers. The first is the net profit margin which is net profit after taxes divided by sales. Sales in case of financial institutions are revenues from loans and revenues from other services. The second driver is assets turnover which is sales divided by total assets. High assets turnover will certainly lead to higher return on equity. The last driver is the financial leverage which is total assets divided by equity. It is the second driver which is the problem in Islamic financial institutions. Total assets are not limited since no financial institution can refuse deposits. However, the product mix available for Islamic financial institutions is still limited which leads to a very low assets turnover and subsequently lower return on equity. The limited mix of products is due mainly to the insufficient academic research done on how to create innovative products that comply with Islamic laws and still utilize the idle funds available to Islamic financial institutions.

The objective of the current study is rather modest. We aim to determine if the results of Omran (2009) remain valid if different measures of valuation are used. In other words, is the clientele effect still there if we use different measures of valuation? We employ the price to book value and price to sales multiples instead of the price to earnings multiple used in Omran (2009). The study confirms the results of Omran (2009) that there is a strong clientele preference for Islamic financial institutions in the UAE regardless of which measure of valuation is adopted.

The study is divided into 6 sections with the introduction in the first section. Section 2 discusses the composition and drivers of values for the price to book value and price to sales multiples. Section 3 describes the data set. Section 4 examines the determinants of the price book value multiple. Section 5 examines the determinants of the price sales multiple. Section 6 concludes the study.

2. VALUATION MODELS

The three most used multiples in valuation are the price to earnings (*PE*), the price to book value (*PBV*), and the price to sales (*PS*). The focus of the current study is on the *PBV* and *PS*. The book value of equity is the difference between the book value of assets minus the book value of liabilities. The price is the market value of the asset, which can deviate significantly from the book value in line with the expectation of future earnings power and cash flows. Lie and Lie (2002) studied 8,621 companies that were active in the Compustat database for the fiscal year 1998. They found that *PBV* estimates of per share value of equity are more precise and less biased than the *PS* and *PE* estimates. They also found that valuation seems to be more accurate for financial firms presumably due to the high liquidity of their assets.

The *PBV* for a stable firm can be derived as follows. The price at time zero is given by

$$P_0 = \frac{BV_0(ROE_1)(1-b)}{K-g}$$

where BV_0 is the book value of equity at time zero, ROE_1 is the expected return on equity at time one, $1-b$ is the payout ratio, b is the retention ratio, g is the growth rate, and K is the required rate of return. Omran (2003) shows that *PBV* is an increasing function of *ROE*, growth rate, and a decreasing function of the required rate of return (and therefore a decreasing function of the

firm's risk). The same economic fundamentals drive the *PBV* for companies experiencing supernormal growth rates.

The price sales (*PS*) multiple for a stable firm can be derived as follows. The price at time zero is given by

$$P_0 = \frac{S_0(NPM)(1-b)(1+g)}{K-g}$$

where S_0 is the sales at time zero, *NPM* is the net profit margin, *1-b* is the payout ratio, *g* is the growth rate and k is the required rate of return. Omran (2003) shows that *PS* is an increasing function of net profit margin, payout ratio, and growth rate, and a decreasing function of the firm's risk (and therefore a decreasing function of the required rate of return).

3. DATA

The study is a regression analysis of the panel data for the firms listed in the local share directories from 2001 to 2005. The local share directory is published by the national bank of Abu Dhabi. The sample has 88 companies that are traded in at least one of the two major stock markets in the UAE, the Abu Dhabi Stock Market or Dubai Financial Market. The appendix at the end of the paper contains the companies' names along with their stock market codes. The sample includes 8 financial services companies, 18 banks, 19 insurance companies, 6 Islamic financial institutions (3 banks and 3 insurance companies), and 37 companies in other sectors of the economy. The pooled data approach is used in the study. There are 308 observations in the sample. The data are end of year values and pooled in time series and cross section directions. For example, for each company in the sample the end of year results are included and pooled with the rest of the companies in the cross section sample. Dummies will be added in the methodology section to take into account different intercepts per industry.

4. THE DETERMINANTS OF THE PRICE BOOK VALUE (*PBV)* MULTIPLE

The analysis in section 2 indicated that *PBV* is positively related to return on equity (*ROE*), growth rate and a decreasing function of the required rate of return. The correlation coefficients between the *PBV and each of ROE* and growth rate are 0.40 and 0.15, respectively. The correlation coefficients have positive signs, as expected. Figure 1 plots *PBV* on the y-axis versus *ROE* on the x-axis together with the fitted OLS line. This figure confirms the previously mentioned economic theory that there is a positive relationship between the two variables.
The following equation is obtained using OLS based on the determinants that significantly drive the *PBV* multiple in the UAE between 2001 and 2005.

Equation 1

$$PBV_i = \beta_0 + \beta_1 ROE_i + \beta_2 G_i + \beta_3 DFServices + \beta_4 DBanks + \beta_5 DInsurance + \beta_6 DIslamic + u_i$$

Where PBV_i is the price to book value for company *i*.

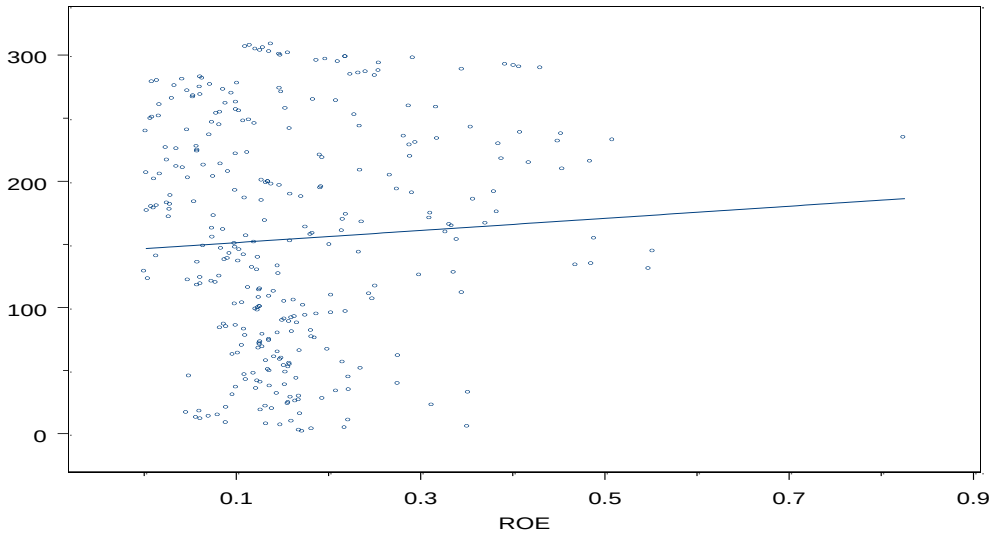


Figure 1: Scatter plot of *PBV* versus *ROE* along with the least squares fit

Table 1: The regression results for the *PBV* multiple. * refers to significance at 5% level and ** refers to significance at the 10% level.

Coefficients	Value	t-value	Pr(>|t|)
β_0	1.4104*	8.7509	0.0000
β_1 (*ROE*)	5.9173*	8.8611	0.0000
β_2 (Growth)	0.1326**	1.8172	0.0702
β_3 (Financial Services)	-0.3982	-1.2913	0.1976
β_4 (Banks)	0.5401*	2.7534	0.0063
β_5 (Insurance)	-0.8421*	-4.1962	0.0000
β_6 (Islamic)	1.0721*	3.2918	0.0011

Multiple R-Squared: 0.3078. F-statistic: 22.31 on 6 and 301 degrees of freedom, the p-value is 0 indicating significance at 1% level.

G is the growth rate in total assets from one year to the next for company *i*,
DFServices is a dummy variable that takes a value of one for a traditional financial services such as finance houses, and zero otherwise,
DBanks is a dummy variable that takes a value of one for a traditional bank, and zero otherwise.
DInsurance is a dummy variable that takes a value of one for a traditional insurance company, and zero otherwise.
DIslamic is a dummy variable that takes a value of one for an Islamic financial institution (whether it is a bank or insurance company), and zero otherwise.

The results indicate that the *PBV* is significantly positively related to *ROE* at the 1% level. The growth rate has the expected positive sign but it is only significant at the 10% level. The coefficients for banks, insurance and Islamic financial institutions are all significant at the 1% level. Insurance has a negative coefficient indicating low preference for the common stocks of insurance companies. Banks demand a higher premium but still half that of Islamic financial institutions. The premium coefficient for banks is 0.54 compared with 1.07 for Islamic financial institutions. The model explained 31% of the total variation in *PBV* and the overall model is significant at the 1% according to the *F* test. Table 2 has the average *PBV* multiples, average growth, average payout, and average return on equity (ROE) for the UAE stock markets as well as for each sector. The Islamic sector contains 3 Islamic banks and 3 Islamic insurance companies.

Table 2 Average *PBV* by sectors along with averages for growth rates, payouts, and return on equity (ROE) during the period from 2001 to 2005.

	Number	Average *PBV*	Average Growth	Average Payout	Average ROE
The Whole Sample	88	2.42	39.88%	43.67%	16.45%
Financial	8	2.33	90.24%	23.85%	24.13%
Banks	18	2.56	23.40%	33.63%	16.06%
Insurance	19	1.66	33.94%	40.45%	17.68%
Islamic	6	3.43	130.81%	43.21%	13.17%
The Rest	37	2.33	27.37%	57.46%	14.85%

This table shows that Islamic financial institutions have 42% more premium in *PBV* compared with the average *PBV* of the UAE stock markets. Islamic financial institutions have demanded the highest premium despite of the fact that they had achieved the lowest return on equity during the period. The growth in total assets of Islamic financial institutions averaged 130.81% during the period in contrast with an overall average of 39.88% for the whole market. The huge growth in the assets of Islamic financial institutions is a clear indication for the UAE clientele preference for companies closely following Islamic laws despite of the fact that those companies achieved the lowest return on equity compared with the rest of the companies in the market.

5. DETERMINANTS OF THE PRICE SALES *(PS)* MULTIPLE

The analysis in section 2 indicated that *PS* is an increasing function of net profit margin (*NPM*), growth rate and payout ratio. The correlation coefficients between *PS* on one hand and each of *NPM*, growth and payout ratio are -0.05328789, 0.02995263 and -0.01065787, respectively. Omran (2003) showed that the logarithm of *PS* should be modeled as a linear function of *NPM*. Figure 2 shows the logarithm to the base ten of *PS* on the Y-axis versus *NPM* on the X-axis.

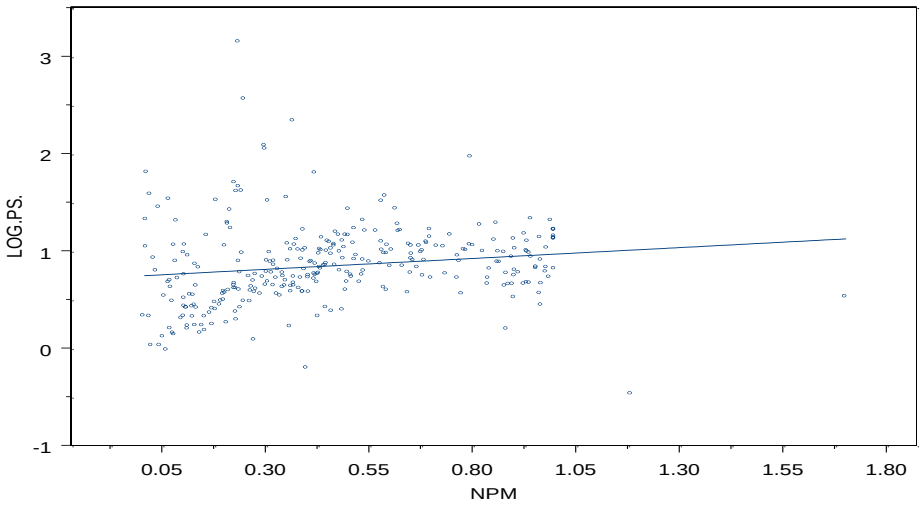


Figure 2: logarithm of price sales (*PS*) versus net profit margin (*NPM*)

The correlation coefficients between *log*(*PS)* and each of *NPM*, growth and payout ratio are 0.1596912 0.10363279 0.07811821, respectively. A regression model of *log(PS)* on *NPM*, growth, and payout yielded significant coefficient for NPM only. Accordingly, it was decided to run the following regression equation.

Equation 2

$$\log(PS_i) = \beta_0 + \beta_1 NPM_i + \beta_2 DFServices + \beta_3 DBanks + \beta_4 DInsurance + \beta_5 DIslamic + u_i$$

where *Log(PS$_i$)* is the log to the base 10 of the price sales multiple for company *i*.
NPM$_i$ is the net profit margin for company *i*,
DFService is a dummy variable that takes a value of one for traditional financial services such as finance houses, and zero otherwise,
DBanks is a dummy variable that takes a value of one for a traditional bank, and zero otherwise.
DInsurance is a dummy variable that takes a value of one for a traditional insurance company, and zero otherwise.
DIslamic is a dummy variable that takes a value of one for an Islamic financial institution (whether it is a bank or insurance company), and zero otherwise.

The results indicate that *NPM* is no longer significant when industrial sectors are included. Financial services are not significant in line with the results for *PBV* multiple in Equation 1. Insurance companies demand a positive significant premium equal to that of the banks which contradicts the results on PBV in equation 1. where insurance companies commanded a negative premium. However, in line with the results on *PBV* reported in equation 1 and *PE* reported in Omran (2009), Islamic financial institutions demand far more premium compared with the whole market. Their *log*(*PS)* premium is almost 3 times the premium charged by banks and insurance companies. Table 4 shows the average *log*(*PS)* multiples, average growth, average payout, and average return on equity (ROE) for the whole market as well as for each sector.

Table 3: The regression results for the *log*(*PS)* multiple. * refers to significance at 5% level and ** refers to significance at the 10% level.

Coefficients	Value	t-value	Pr(>|t|)
β_0	0.6535*	14.3718	0.0000
β_1 (*NPM*)	0.0859	0.7860	0.4325
β_3 (Financial Services)	0.0208	0.2292	0.8189
β_4 (Banks)	0.2245*	3.8882	0.0001
β_5 (Insurance)	0.2109*	2.5283	0.0120
β_6 (Islamic)	0.6407*	7.1057	0.0000

Multiple R-Squared: 0.1831. F-statistic: 13.53 on 5 and 302 degrees of freedom, the p-value is 0.

Table 4 Average *log*(*PS)* by sector along with averages for growth rates, payouts, and return on equity (ROE) during the period from 2001 to 2005.

	Number	Average *log*(*PS)*	Average Growth	Average Payout	Average ROE
The Whole Sample	88	0.85	39.88%	43.67%	16.45%
Financial	8	0.73	90.24%	23.85%	24.13%
Banks	18	0.92	23.40%	33.63%	16.06%
Insurance	19	0.94	33.94%	40.45%	17.68%
Islamic	6	1.34	130.81%	43.21%	13.17%
The Rest	37	0.68	27.37%	57.46%	14.85%

This table shows similar results to table 2. Islamic financial institutions demand a much higher premium than the whole market. The clientele effect seems to be robust regardless of which model is used. Diagnostic tests were conducted on the residuals from equations 1 and 2 which included the normal probability plots and heteroscedasticity tests that involve running the squared residuals on the independent variables. Both models have passed the two diagnostic tests. The residuals QQ plot is close to the line drawn by the normal distribution. Also, none of the coefficients in the regression of the squared residuals on the cross section variables is significant. The whole regression is also not significant at the 10% level suggesting that the residuals can be described as free of cross sectional heteroscedasticity.

6. COMMENTS AND CONCLUSION

This study examined if the results of Islamic clientele preference found by Omran (2009) were specific to the valuation model used in his study. There are three major valuation models used namely, the price earnings, price book value and price sales multiples. Omran (2009) used price earnings *PE* multiple and reached the conclusion that Islamic financial institutions seem to be over priced compared with the rest of the industry. The higher price premium was explained by

Omran (2005) as a preference by UAE investors to deal with Islamic financial institutions regardless of how much profit they achieve in relation to their assets base. Islamic financial institutions achieved the lowest return on equity during the period of study (2001-2005) and yet they witnessed a massive growth in their assets during the same period. It is argued that their low return on equity is because of lack of investment opportunities and products that comply with Islamic laws. This leads to Islamic banks having a large amount of idle cash that does not earn profit which impacts negatively their return on equity. This study found that the results of Omran (2009) hold true regardless of the valuation model. Both of the price book value *PBV* and price sales *PS* models show Islamic clientele preference for Islamic financial institutions regardless of low profitability. We recommend that more research be devoted for developing more Islamic products that can accommodate the massive flow of funds into Islamic financial institutions.

APPENDIX

1	NBAD	NATIONAL BANK OF ABU DHABI
2	ADCB	ABU DHABI COMMERCIAL BANK
3	ADIB	ABU DHABI ISLAMIC BANK
4	AEIB	ARAB EMIRATES INVESTMENT BANK
5	BOS	BANK OF SHARJAH
6	CBD	COMMERCIAL BANK OF DUBAI
7	CBI	COMMERCIAL BANK INTERNATIONAL COMMERCIAL INTERNATIONAL BANK
8	CBIE	EGYPT
9	DIB	DUBAI ISLAMIC BANK
10	EBI	EMIRATES BANK INTERNATIONAL
11	EIB	EMIRATES ISLAMIC BANK
12	FGB	FIRST GULF BANK
13	FH	FINANCE HOUSE
14	IB	INVEST BANK
15	MASHRAQ	MASHRAQ BANK
16	NBD	NATIONAL BANK OF DUBAI
17	NBF	NATIONAL BANK OF FUJAIRAH
18	NBQ	NATIONAL BANK OF UMM AL-QAIWAIN
19	RAK	NATIONAL BANK OF RAS AL-KHAIMAH
20	SIB	SHARJAH ISLAMIC BANK
21	UAB	UNITED ARAB BANK
22	UNB	UNION NATIONAL BANK
23	ADNIC	ABU DHABI NATIONAL INSURANCE
24	AIN AHLIA	ALAIN AHLIA INSURANCE
25	BUHAIRA	AL BUHAIRA NATIONAL INSURANCE
26	DHAFRA	AL DHAFRA INSURANCE
27	KHAZNA	AL KHAZNA INSURANCE
28	WATHBA	AL WATHBA NATIONAL INSURANCE
29	ALLIANCE	ALLIANCE INSURANCE
30	AMAN	DUBAI ISLAMIC INSURANCE (AMAN)

31	ARIG	ARAB INSURANCE GROUP
32	ASI	ARABIAN SCANDINAVIAN INSURNACE
33	SAQR	AL SAQR NATIONAL INSURANCE
34	DIC	DUBAI INSURANCE
35	DNI&RC	DUBAI NATIONAL INSURANCE
36	EIC	EMIRATES INSURANCE
37	NGI	NATIONAL GENERAL INSURANCE
38	OI	OMAN INSURANCE
	RAK	
39	INSURANCE	RAS AL KHAIMAH NATIONAL INSURANCE
40	SALAMA	ISLAMIC ARAB INSURANCE
41	SIC	SHARJAH INSURANCE
42	TAKAFUL	ABU DHABI NATIONAL TAKAFUL
43	UIC	UNITED INSURANCE
44	UNION	UNION INSURANCE
45	ASMAK	INTERNATIONAL FISH FARMING
46	RAK POULTRY	RAS AL KHAIMAH POULTRY & FEEDING
		ABU DHABI NATIONAL CO. FOR BUILDING
47	BILDCO	MATERIALS
48	FCIC	FUJAIRAH CEMENT INDUSTRIES
49	GC	GULF CEMENT
50	NC	NATIONAL CEMENT
51	UAQC	UMM AL QAIWAIN CEMENT INDUSTRIES
52	RAKCEMENT	RAS AL KHAIMA CEMENT
53	RAKCERAMICS	RAS AL KHAIMA CERAMICS
54	RAKWC	RAS AL KHAIMA WHITE CEMENT
		SHARJAH CEMENT & INDUSTRIAL
55	SCID	DEVELOPMENT
	UNION	
56	CEMENT	UNION CEMENT
57	AGTHIA	EMIRATES FOODSTUFF & MINERAL WATER
58	FOODCO	ABU DHABI NATIONAL FOOD STUFF
59	JEEMA	JEEMA MINERAL WATER
60	ABAR	ABAR FOR PETROLEUM INVESTMENT
61	TAQA	ABU DGABI NATIONAL ENERGY
62	ADSB	ABU DHABI SHIP BUILDING
63	AMLAK	AMLAK FINANCJSC
64	DI	DUBAI INVESTMENT
65	GGI	GULF GENERAL INVESTMENT
66	GLOBAL	GLOBAL INVESTMENT HOUSE
67	IFA	INTERNATIONAL FINANCIAL ADVISORS
68	O&E	OMAN & EMIRATES INVESTMENT HOLDING
69	SHUAA	SHUAA CAPITAL
70	ADA	ABU DHABI AVIATION
71	OASIS	OASIS INTERNATIONAL LEASING
72	GMP	GULF MEDICAL PROJECTS

73	JULPHAR	GULF PHARMACEUTICAL INDUSTRIES
74	ALDAR	AL DAR PROPERTIES
75	ARTC	ARAB TECHNICAL CONSTRUCTIONS
76	EMAAR	EMAAR PROPERTIES
77	RAKP	RAS AL KHAIMAH PROPERTIES
78	UP	UNION PROPERTIES
79	ARAMEX	ARAB INTERNATIONAL LOGISTICS
80	ED	EMIRATES DRIVING
81	NMD	NATIONAL MARINE DREDGING
82	TABREED	NATIONAL CENTRAL COOLING
83	ETISALAT	EMIRATES TELECOMMUNICATION
84	PALTEL	PALESTINE TELECOMMUNICATION
85	QTEL	QATAR TELECOM.
86	SUDATEL	SUDAN TELCOMMUNICATION
87	ADNH	ABU DHABI NATIONAL HOTELS
88	NCTH	NATIONAL CORPORATION FOR TOURISM & HOTELS

REFERENCES

Al-Salem, F., (2008). The size and scope of the Islamic finance industry: an analysis. *International Journal of Management*, 25(1), 124-130.

Lie, E., and Lie H., (2002). Multiples used to estimate corporate value. *Financial Analysts Journal*, 58(2), 44-54.

Omran, M.F. (2009). Examining the effects of Islamic Beliefs and Teachings on the Valuation of Financial Institutions in the United Arab Emirates, *Review of Middle East Economics and Finance*, Vol. 5, No. 1, article 4.

Omran, M.F., (2003). Equity valuation using multiples in the emerging market of the United Arab Emirates, *Review of Middle East Economics and Finance*, Vol. 1, No. 3, 267-283.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 931–949.

TRENDS AND PROSPECTS OF CONTRACEPTIVE USE ON FERTILITY DECLINE AMONG THE MUSLIM WOMEN IN NIGERIA

Osuafor, G.N.[1] and Stiegler, N.[2]
Department of Statistics and Demography, University of the Western Cape, X17 7535 Bellville, South Africa
E-mail: [1] gnosuafor@gmail.com, [2] nstiegler@uwc.ac.za

ABSTRACT

Background: The present study examines the trend and future prospects of contraceptive use in reducing fertility among Muslim women in Nigeria. Studies have shown that contraceptive use contributed significantly to below fertility replacement in most of the western countries and ultimately to population ageing. Governments of Nigeria and non-governmental organization have embarked on massive promotion of contraceptive use to reduce population growth rate to two percent. Hospitals have been equipped with health facilities to provide family planning services. Family planning program have been extended to males because of their prerogative in fertility decision making in Nigeria. It is expected that involvement of male in family planning program will pave way for the adoption of contraceptives use and consequently fertility decline. Low fertility may be relevant for socioeconomic development, although not always consistent.

Methods: Demographic and Health Survey (NDHS) conducted in Nigeria 1990, 1999 and 2003, which covered 4269, 3620 and 3601 respectively of women aged 15-49 were used in the study. We employed descriptive analysis to explain already existing conditions surrounding fertility of the Muslim women. Trends in the indicators of current use of contraceptive, intention to use by non users, husband approval of family planning (FP), Discussed FP with partner, ideal number of children and desire for more children were examined. Trends in average number of children ever born by background were examined in relation to contraceptive use.

Results: The use of modern contraceptives has increased by two percent from 1990 to 2003. The percentage of women who do not intend to use contraceptive has remained stable at seventy percent between 1990 and 2003. The percentage of women who want no more children increased with the number of living children. Fifty-four percent of men did not approve FP for their spouse and eighty-one percent of the women have never discussed FP with their partners. Average number of children ever born by background did not show meaningful changes over time. Increase in wealth, education and number of living children are associated with increasing contraceptive use.

Conclusion: Contraceptive use does not seem to play any decisive role or posit bright future for fertility decline among Muslim women in Nigeria. Trend in average number of children ever born shows that population ageing is not so imminent. Bright prospects for contraceptive use may depend on dynamics in education, economic advantage and living number of children.

Keywords: contraceptive use, socioeconomic development, fertility decline, population ageing, Family planning

1. INTRODUCTION

Nigeria comprises fifty percent Muslim, forty percent Christian and ten percent other religion has the potential of becoming the fifth largest population by 2050 (US Bureau of census 2001). This is of concern because the population growth is through natural increase. The tempo of population increase in Nigeria has therefore attracted attention of the government, international organization, and non-governmental organization (NGOs). Consequently, there have been aggressive campaigns promoting the use of contraceptives and establishment of health facilities for family planning. While the sole aim is to reduce fertility, it is important to note that excessive use of contraceptive is a prime cause of population ageing.

Religious leaders, husbands and the influential people within lineages are some of the critics of modern contraception in Nigeria (Pearce, 2001). The oppositions does not seem to be based on side effects of contraception, but rather children are gifts and preventing their coming into being is a crime against existence and violation of divine order. An average man in Nigeria holds a contrary view that high fertility is the cause of poverty in Nigeria. In addition, Nigeria as a country has enough natural resource that can sustain the population in the absence of greed. This has been the cause for the loss of confidence in the leadership of the country. The effect of this was that men did not endorse the distribution of contraceptives among wives (Olusanya's, 1969; Jinadu & Ajuwon, 1997). Traditionally, Nigeria nation is a pro-natalist society which advocates for large family size. In addition clinging to their religious beliefs which by no means consider contraception relevant has sustaining effects on high fertility. Although in the recent times men are involved in the family planning programme, it is still unknown to what extent the involvement has faltered their cultural and traditional belief in fertility decision making.

There are reports that fertility is declining in Nigeria (NPC, 2000; United Nations, 2000) and prospects for further reduction is bright (Feyisetan and Bankole, 2002; Oladosu, 2001). A recent report suggests that fertility decline that started in most Sub Saharan Africa has slowed down (Bongaarts, 2005). There has been consistent increase in the use of modern contraceptives, education, urbanization and involvement of men in family planning in the recent times in Nigeria. All these activities are known to reduce rapid population growth and fertility in Nigeria. Most importantly, the government and nongovernmental organization are now seemingly working in an orchestrated fashion to reduce population growth through contraceptive use and family planning. Thus far, it is not known to what extent the efforts of the government, nongovernmental organizations and other pressure groups have affected fertility among the Muslim women in Nigeria. Considering the patriarchal nature of Muslim society in Nigeria, the paper describes the trend in modern contraceptive use among the Muslim women and its future prospects on fertility decline in Nigeria. The indicators of fertility between 1990 and 2003 Nigeria Demographic and Health Surveys are used to evaluate the trend, future prospects of contraceptive use and fertility decline among Muslim women.

2. FACTORS AFFECTING FERTILITY AND FERTILITY REGULATION

The entire edifice of fertility in Nigeria is modulated by economical, cultural and religious background. In general, these setting may sustain high fertility in Nigeria from historical norm.

2.1 Economical Perspective

Economic status may have increasing or decreasing effects on fertility depending on the availability of resources or economic orientation of the people. From the economic perspective, sustained high fertility between 1970s and 1980s was as a result of booming oil revenues. Consequently, there was increase in salary of workers and food importation (Bankole and Bamisaye, 1985). Parents had the ability to cater for children with confidence. For instance in 1983-1986, fertility was 7.4 per woman (Feyisetan and Bankole 2002). Unfortunately, this era of affluence met its debacle by 1986. The introduction of structural adjustment program (SAP) in 1986 affected child bearing disposition. SAP introduced economic conditions that made child rearing exorbitant. Thus fall in fertility in the interval of 1986-1990 to 6.3 was motivated by economic hard times. In deed studies have shown that economic upheaval was the main reason for fertility decline in Nigeria (Orubuloye 1998; National Research Council 1993).

Political and economical turmoil motivated fertility transitions have been documented in Nigeria (Lesthaeghe 1989; National Research Council 1993). Of a particular importance while other regions are experiencing rapidly fertility decline due to economic downturn, Northwest and Northeast dominated by Muslims were lagging behind (Feyisetan and Bankole 2002). Surveys have shown that lack of access to and costs of obtaining contraceptives are extremely negligible to be the cause for high fertility in Nigeria (NPC, 2000; NDHS 2003). In the recent study conducted in Northern Nigeria, eighty-five percent of the male dissuade contraceptive use on the ground of poverty (Duze & Mohammed, 2006). Inability to cater for children does not deter having them among the Muslims. The only explanation is based on Islamic belief that every child comes into existence with his sustenance. However, studies and findings in Europe and other Muslim countries support that economic downturn accounted for fertility decline (Sundguist 2008; Caldwell 2008).

The unfavorable economic state is persisting in Nigeria. The government of Nigeria was unable to provide jobs for the graduates, high cost of schooling and poor health services. The failed development planning of the government lashed its impact on the poor. There was rise in corruption among the leaders and politicians leading to trouncing of confidence in the government by the populace. The poor and the middle class men turned to religion for certainty in uncertain world (Imam 2003). The religious groups provided the services which the state could no longer provide (Imam 2003). The price the poor masses paid was coercion to the tenets of the religious groups.

2.2 Religious Perspective

In Nigeria irrespective of creed, tribe or social status, the belief that children are gift from God is formidable enough to desire many children. This unifying belief makes it difficult to disentangle the roles of culture and religion on fertility matters. NDHS 2003 shows that nine percent of currently married women are prohibited from using contraceptives for religious reasons. In addition, fear of side effects of contraceptives is increasing over time among the women. Of a

particular importance, FP has been perceived as a subtle program of the western powers to impinge on the Muslim population (Renne 1996). This may have confounded the already existing negative attitude to FP. It is noteworthy that Muslim jurists accept contraception and abortion up to forty days (Imam 2003). However, Muslim religious right in Nigeria has disillusioned the knowledge about contraceptive techniques on grounds that it promotes immorality (Imam 2003). The result of this is that women's interests and rights are sacrificed on the altar of appeasement for religious right that favours male dominance. In the northern Nigeria, negative attitude to family planning is driven by religious belief, which is embedded in their culture and tradition (Duze & Mohammed 2006).

2.3 Socio-Cultural Perspective

Nigeria cultural settings in its entirety support high fertility through the medium of male dominance in fertility decision making. Surveys have shown that Nigerian men like any other patrilineal society in Sub Saharan Africa desire large family sizes (Isiugo-Abanihe 1994). The pro-natalist behaviour of men entrenched in cultural belief and anticipated benefit on children is still esteemed in Nigeria. Nigeria is a rural based nation with about eighty percent of the population engages in subsistence agriculture where children are productive agent. Northern Nigeria evidently is the food basket of the nation. In this regard, cost of regulating fertility is high among the Muslim. Irrespective of the number of living children, about forty-five percent and twenty-eight percent of currently married women aged 15-29 and 30-49 respectively do not intend to use contraceptives (NDHS 2003). The reason was that they want as many children as possible. In patriarch northern Nigeria, women owe allegiance to husbands' family in terms of labour and childbearing because bride wealth has been paid to her family. This therefore gives the husband unquestionable control over family issues, which may be exercised in a despotic manner. Even Sharia Penal Code permits husbands to beat wives in the 1960 Penal Codes (Imam 2003). Over seventy-five percent of Hausa and Kanuri spouses reported that wife opinion is negated on family size (Duze & Mohammed 2006).

Some social and public institutions are in favour of men even in fertility and family matters. For instance, women may be denied FP services in government hospitals without the consent of their spouses (Duze & Mohammed 2006). Patriarchal dominion of men is a serious barrier in adopting family planning among Muslim women. Studies show in patrilineal societies like Nigeria or elsewhere men influence the use of contraceptive (Khalfa 1988; Oni & McCarthy 1990; Mbizuo & Adamchack 1991).

Another practice that sustains high fertility in Nigeria is the practice of polygyny. The Muslim men are under obligation to practice polygyny in order to be like the Prophet (Imam 2003). Polygyny itself introduces an aspect of competition among the wives to gain greater assets of the family, which depends on the number of male children. There is high preference for male children since they will retain the family name. The parents equally value female children, although they are expected to leave their fathers house and name due to marital debut. However, parents still depend on them for fiscal remittance.

The crave for children especially male is not limited to Nigeria, but it cuts across the Sub Saharan Africa countries, India and China. For instance, low contraceptive use in Tanzania is attributed to desire for more children (Mwageni 2001). In India, failure to adopt contraceptive use is due to desire to have male children. In China, negative attitude to use contraceptive have been associated with unmet desire to have a male and female child (Whyte and Gu, 1987: 478).

It is noteworthy that common interest of economic and security at old age underlie the desire for children irrespective of races.

Taken together culture, religion and even adverse economic conditions among Nigeria Muslims are still antagonistic to the crusade for FP and contraceptive use. This may partly account for the declined trends in support of FP from respondent and spouse that were observed in Nigeria Demographic and Health Surveys (NDHS) 1990 and 1999 (Oladosu 2001). Unknown side effects of contraceptives may have also contributed in the decline for FP support. More important, seventy-three percent of married women in 1999 NDHS with four living children want to have another child (NPC, 2000:88). This may suggest that motivation for fertility regulation or low fertility is apparently weak and discouraging (Easterlin 1985).

2.4 Need for Fertility Regulation

It appears that the zeal to regulate fertility is low, but the need has been echoed by considerable number of women who desire to end childbearing (Dodoo 1993). The percentage of women with unmet needs of FP has increased by twenty-seven percent between 1999 and 2003 (NDHS 1999; 2003). Those who want to stop childbearing by number of living children increased by forty-six percent within the same interval. Apart from the prevailing socio-cultural, economic and religious impediment on contraception, FP facilities are located in remote places or urban areas. In Kano state, health facilities are sited in urban areas where about eighty percent of public and private hospitals are located (Mohammed and Khalid 1995). Hence, the majority of the population has no access to the services. The decisive role of men in fertility in Sub Saharan Africa is a severe vitiation to women's right. About eighty percent of men in Sudan were against limiting family size due to religious belief (Duze & Mohammed 2006). This may be a justified reason to discourage contraceptive use. However, refusal to approve contraceptive use at the detriment of women's life due to incessant pregnancies (Khalfa 1988; Mustapha and Mumford 1984) is astounding. Thus the resultant of patriarchal dominion of men is abuse of women and children. One of the things that promote abuse of women is the dependence on male as the provider in the house

3. ROLE OF GOVERNMENT AND NONGOVERNMENTAL ORGANISATIONS ON FERTILITY REGULATION

Nigerian government and some African countries like Ethiopia, Tanzania and Somalia were adamant to international concern over population growth (Pearce 2001). Until mid 1980s, Nigerian government became concerned on population growth (Pearce 2001), possibly due to economic downturn. Role of government has been very cynical on matters of fertility which NGOs have not admired. Dependency of women on male for economic support was upheld in the Nigeria population policy of 1988. Despite the fact that population control was motivated by economic crises, the government failed to stipulate her interest on economic development in the policy.

We hold the view that the population policy of 1988 was skeptical because firstly government was not controlling disbursement of fund for fertility control. Secondly, power brokers, religious and community leaders, were against contraceptive use (UNFPA 1996:23). Thirdly, men dominated the strategic places that could promote fertility regulation (Berer 1996). Finally, the population policy document stipulates "the patriarchal family system in the country

shall be recognized for the stability of the home" (FMOH 1988; 19). On 29 January 2004, Nigerian Minister of Health then, Eyitayo Lambo announced a new policy that would replace the 1988 national population policy under which each couple was encouraged to have four children or more (RedOrbit 2004). Lambo said that the new policy would also encourage Nigerians on the need to have the number of children they could cater for, since there was no ceiling on the number of children per couple in the new policy. He said the target for the 2004 policy was to ensure that Nigeria's population growth rate was reduced from the current three percent per annum by 2015. The government would check population surge, by promoting the use of modern contraceptives. Indeed, there is nothing spectacular in this new policy compared to that of 1988. The expectation that it would affect population growth and fertility is with little confidence.

Recall that the population policy of 1988 did not put any development plan down. However, it was initially sold to the populace as a health and development benefit (Pearce 2001) and population was propagandized as a developing state feature in the midst of stagnant economy. Thus, the state procured external funds, technical advice, technology and equipment (Pearce 2001). However, with the cession of external funding, future of the entire family planning program is in jeopardy, since the neither government nor the public can bear the cost of supplies (UNFPA 1996). The entire program is at the verge of collapsing; as agencies began withdrawal.

Succinctly there has been profound reliance on external funding from Britain, USA, UNFPA, UNICEF and NGOs like The Ford Foundation, *International Planned Parenthood Federation* (IPPF). With the raging economic problem and sensation that the government is not working in the interest of the masses, nongovernmental organizations multiplied to manage the community problems (Pearce 2001). All hope was that NGOs would rectify the problems of the marginalized, especially those in the rural areas in terms of health and family planning. However, the multiplication of NGOs gave birth to duplication of health and reproductive projects among the locally and externally driven groups. The erratic scene was difficult to resolve because NGOs were working in the interest of their funders. Hence, none took full control of policies and activities designed for women's health. However, NGOs are always at conflict with the government on issues of policy regarding women health. The coalition formed by 140 NGOs working on women's reproductive health insisted that Nigeria's agenda for the 1994 population conference in Cairo should included plans for economic development and other issues that affect women's health (Olukoya 1996).

Activities of NGOs in providing family planning sensitization, education, counseling and delivery services have increased over the years. NGOs especially Planned Parenthood Federation of Nigeria (PPFN) has extended their reproductive health services to adolescents. Adolescents were of secondary interest until recently (Makinwa-Adebusoye 1991; Jinandu & Ajuwon 1997). There was incorporation of family planning and maternal and child health service under the primary health care system to offer more opportunities to reach potential clients (PHC 1987). There was community based distribution program in many parts of Nigeria, including the North which has the greatest resistance to family planning services. High levels of male participation in family planning have been documented for the Southwest and Southeast (Feyisetan et al. 1998; Odimegwu 1999). The use of the mass media to promote family planning has been found to be effective in changing contraceptive behavior in Nigeria (Bankole et al. 1999; Odimegwu 1999).

These newer dimensions in the provision of family planning services are expected to increase access to family planning services, minimize adolescents' pregnancies and change men negative attitude to contraception. However, attempts have been made to deter non-governmental organizations to run workshops on sexuality education and FP. There is evidence of removal of

sex education from school curricula (Imam 2003) in the northern Nigeria. Indeed significant participation of men in family planning has been documented for the Southwest and Southeast. The involvement of men in the north is not yet clear. The use of the mass media for *Information, education* and *communication* (IEC) has been efficient in disseminating family planning.

4. PAST AND PRESENT DEMOGRAPHIC TRENDS

Trend in total fertility rate (TFR) in Nigeria shows that fertility has declined gradually and stabilized above 5 per woman. TFR in Nigeria was 7.4 per woman in 1983-1986 and 5.7 in 1999-2003 suggesting a decrease by twenty-three percent. United Nations statistics suggest a TFR of 3.4 between 2020-2025 and is expected to reach replacement between 2045-2050 (United Nations 2000). Currently, the Nigeria population growth rate has declined from three percent in 1988 to two and half percent in 2009 (Population Reference Bureau, 2009). While there are evidence of decrease in TFR and population growth rate, it is unknown if the factors that produce the decline are consistent.

The Ability of a woman to take charge of her fertility and contraceptive method partly depend on her empowerment status and self-image. Low status of Muslim women prevent them from obtaining education and good jobs (Sundquist 2008). Hence their lives are circumscribed with increased desire for large family size. The use of modern contraceptives has been increasing gradually from about four percent in early 1980, twelve percent in 1996 and seventeen percent in 2003 (WFS 1981/2; UNFPA 1996; NDHS 2003). There are regional differentials in the use of modern contraceptives. Data show that in 1990, use of any method of contraceptive was two percent in the North and twelve percent in the South with corresponding TFR of 6.6 in the North and 5.5 in the south (NDHS 1990). Level of education and economic independence encourage women's self-esteem. Studies have shown that the use of contraceptives correlates positively with educational attainment (NDHS 2003). For instance among women with no education and higher education the use of any modern method increased by two percent and twenty-two percent respectively (NDHS 2003). Even the traditional methods increased with level of education, slightly below two percent for women with no education to fifteen percent for those with higher education. Studies have shown that the tendency to use contraceptive increases with the number of living children (NDHS 2003). Most economically advantaged women are four times likely to use contraceptives compared to the least advantaged (NDHS 2003).

Before extending the past into the future, there is need to examine the present changes taking place in term of FP in the northern Nigeria. There is a recent evidence of increased contraceptive use in Muslim-dominated northern Nigeria. Community Participation for Action in the Social Sector Project (COMPASS) funded by USAID is helping women to avoid unwanted and often high-risk pregnancies. COMPASS is persistently breaking through cultural and religious beliefs which have long discouraged millions of men and women from accessing mainstream family planning services. Islamic leaders in conservative, Muslim-dominated northern Nigeria, particularly Kano, had long opposed the use of contraceptives, but the outreach groups like COMPASS is helping to break down the barrier. Evaluation conducted in all five COMPASS-supported states showed that contraceptive use increased from nine percent to thirty-two percent between 2005 and 2007; while in Nasarawa it increased from eight to twenty percent. (Costa 2008). In addition, unreceptive Muslim Religious leaders promote and practice Family Planning in Nasarawa State (Costa 2008). COMPASS is a five-year integrated community-driven project with nine implementing partners, including the Federation of Muslim Women's

Association and the Nigerian Medical Association. The activities of COMPASS commenced in Nigeria in 2004.

The unprecedented positive attitude to FP and contraception by the Muslim world is not limited to Nigeria. A compendium of findings captioned "The Muslim world's changing views towards family planning and contraception" by Sundquist (2008) reported that Islam was not an obstacle to family planning that led to reduced birth rates in the Muslim countries of Iran, Egypt, Morocco, Tunisia and Bangladesh. Furthermore, conservative Islamic nation of Pakistan, mosques distribute contraceptives and literature to spread the importance of family planning and safe sex. Muslim leaders of Indonesia have acceptable voluntary sterilization as a form of contraception. Muslim clerics in the Philippines and the Grand Mufti of Egypt have admitted the relevance of family planning. Demographic and Social Statistics unit of the U.N., Statistical Division of December 2007 found that Arab birth rates are dropping dramatically, and that the number of births among women under the age of 20 is dropping even more sharply. The Total Fertility Rates (TFRs) of some Arab countries, notably Tunisia, the United Arab Emirates, Bahrain, Kuwait and Lebanon are either below or very close to the stability level of 2.1 children per woman.

The fertility rate of below 2.1 is a caution on the use of contraception. In Europe, fertility has fallen below replacement levels. Unfortunately, the fertility declining did not stop even below the replacement levels. Below replacement fertility in Europe has resulted in an age structure with fewer children, leading to fewer women entering reproductive age in future (Lutz, O'Neill & Scherbov 2003). Contraception while it regulates fertility is the salient cause of population ageing. Institute for Family Policy says that Spain, with one of the western world's lowest birth rates and a high average life expectancy, is now the most rapidly aging country in the European Union (White 2009).The Institute's head, Eduardo Hertfelder, reported that the government's "dreadful" contraceptive policies are having a "catastrophic effect." This is unexpected development that raises a potential fear of family extinction and is of great concern to demographers. It is not clear what the demographic state in Europe would be in future, but demographers are increasingly warning that the prospect of population recovery is remote.

5. PROSPECTS OF CONTRACEPTIVE USE AND FERTILITY DECLINE

Trends in the indicators of current use of contraceptive, intention to use by non users, husband approval of family planning (FP), respondent's approval of family planning, Discussed FP with partner, ideal number of children and desire for more children were examined. Average number of children ever born and contraceptive use by background was examined.

5.1 Current Use of Modern Contraceptive

The use of modern contraceptives is shown in figure 1. Muslim women in Nigeria that use modern contraceptive methods have increased over time for the age groups 20-24 and 25-29. The trend among women aged 30-49 did not show consistent increase over time. The dramatic increase use of modern contraceptives among the younger women may have declining effect on fertility in future. This is consistent with the postulation of Caldwell et al. (1992) that the first group to adopt contraceptive use will be the young women in attempt to avoid pregnancy and expulsion from school.

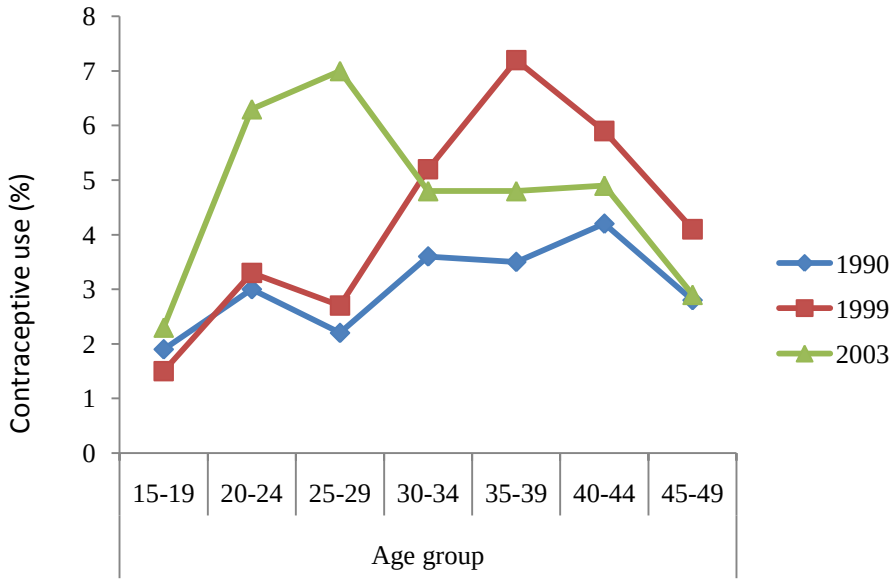


Figure 1: Use of modern contraceptives by Muslim women in Nigeria 1990-2003

5.2 Women with No Intention To Use Modern Contraceptive

The percentage of all women who do not intend to use modern contraceptive has remained stable at seventy percent from 1990 to 2003, figure 2. The percentage of women who do not intend to use contraceptives increases gradually from the age group 30-34. The trend of no intention to use modern contraceptive is an indication that no major changes have occurred in demand for FP over time.

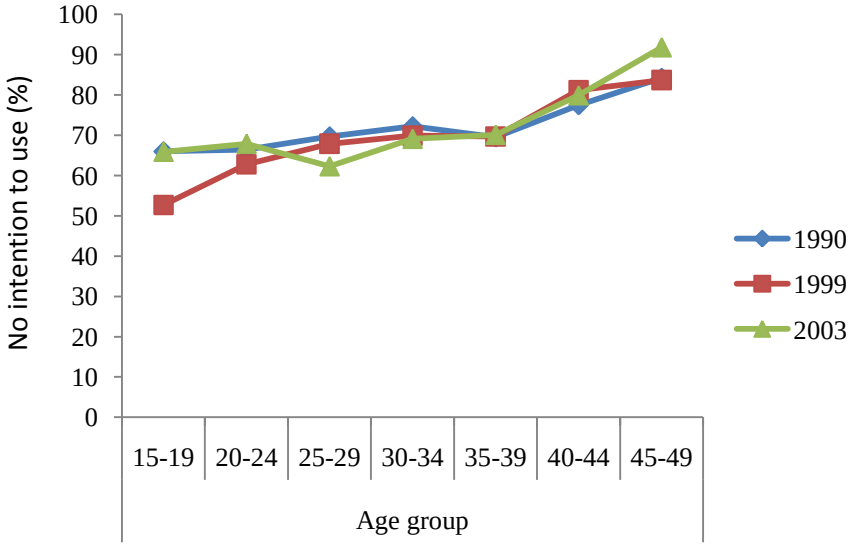


Figure 2: Muslim women who have no intention to use modern contraceptive in Nigeria 1990-2003

5.3 Husband's Approval of Family Planning (FP)

The approval of FP by the husband may promote the adoption by the wives. Figure 3 shows the trend in disapproval of FP by the husbands. There is no consistent increasing trend over time. However over fifty percent of men did not approve FP in 2003. This is a true reflection of the trend in the use of modern contraceptive. The husbands of women aged 20-24 and 25-29 show lowest resistance to FP in 2003. The lowest resistance corresponds to the higher use of modern contraceptives by the women. Husband disapproval of FP decreased by ten percent among all women from 1990-1999. However, the resistance increased by eighteen percent between 1999 and 2003.

5.4 Respondent's Approval of Family Planning (FP)

Over fifty percent all respondents disapprove FP in 2003, figure 4. Disapproval of FP by the women decreased between 1990 and 1999 by twenty-five percent, and increased by twenty-one percent from 1999 to 2003. Disapproval of FP increased progressively for women age 40-44 and 45-49. Trends in disapproval of FP are similar for the husband and the respondents.

5.5 Discussion of FP with Partner

Spousal discussion about FP may create a platform for negotiating the family Size and other things that will favour the woman reproductive health. However, over eight-one percent of all women have never discussed FP with their Partners since 1990 to 2003, figure 5. Women of age group 15-20 rank highest by eight-nine percent while 25-29 ages rank lowest by seventy-four percent in 2003. About eighty-five percent and eighty-four percent of women aged 40-44 and 45-49 respectively have never discussed FP with their partners. Absence of discussion may suggest absolute lack of interest FP.

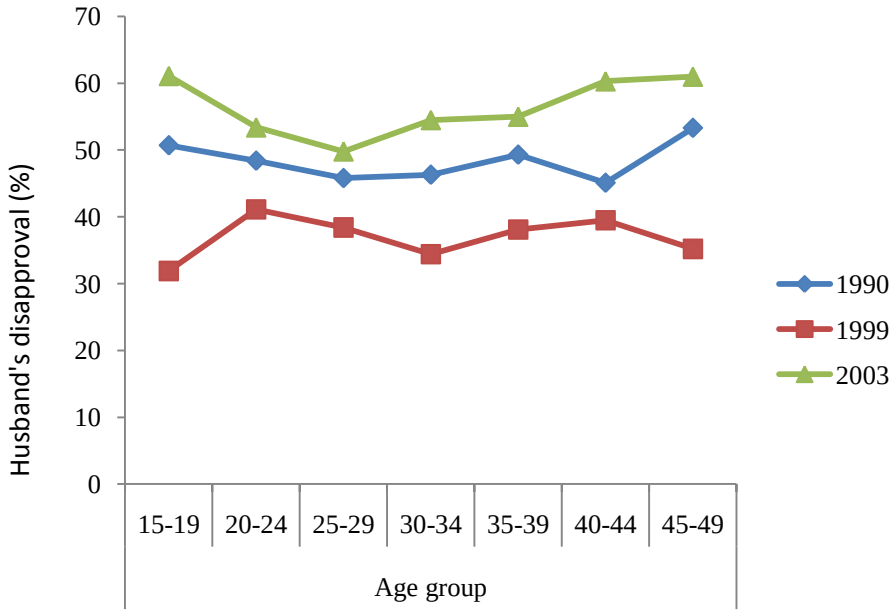


Figure 3: Husband of Muslim women who did not approve FP in Nigeria 1999-2003.

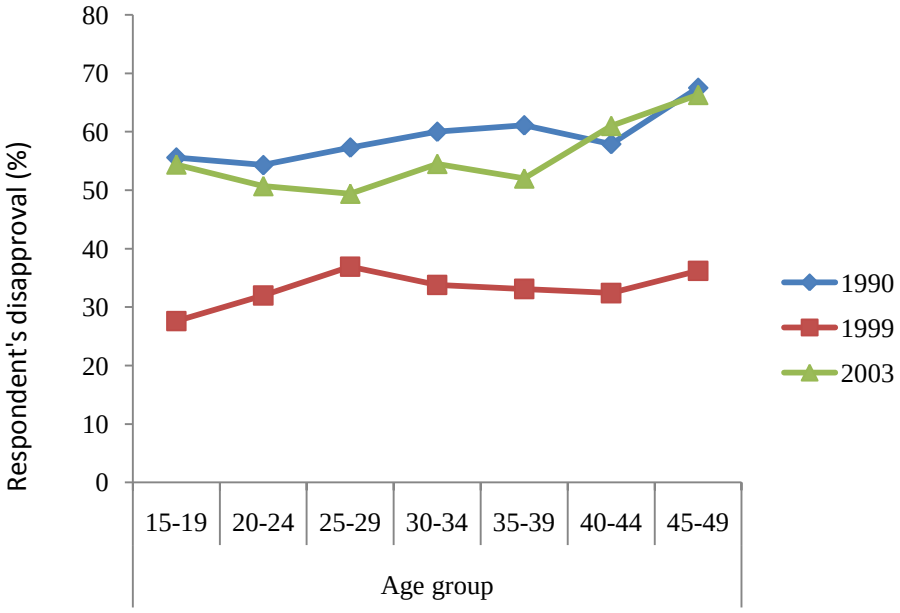


Figure 4: Muslim women who did not approve FP in Nigeria 1999-2003

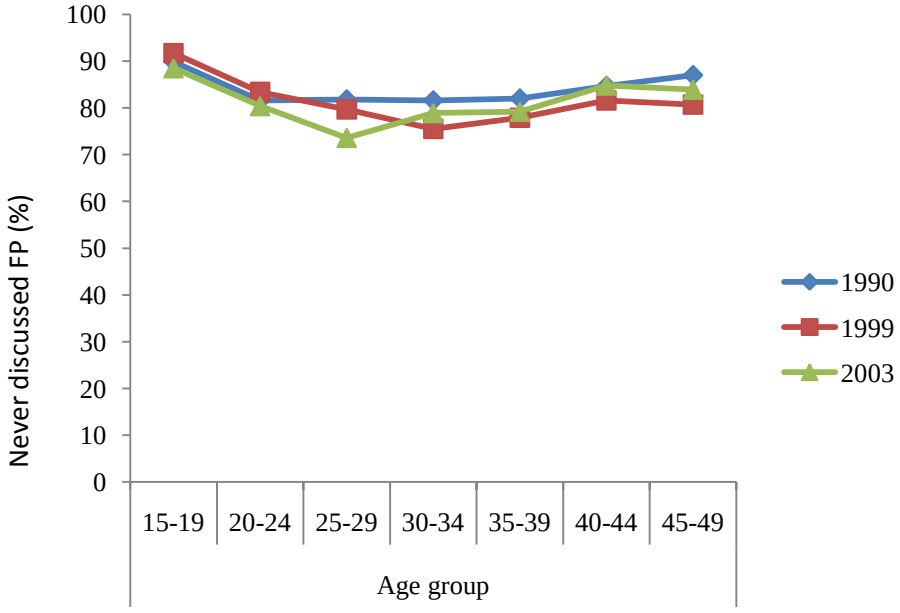


Figure 5: Muslim women who never discussed FP with their partners in Nigeria 1999-2003

5.6 Ideal Number of Children

Table 1 shows the response of Muslim women in Nigerian on what they consider to be ideal number of children. The percentage of women that said six and above increased over time from thirteen percent in 1990 to sixty-two percent in 2003. Non-numeric response has decreased

significantly over time. This show that women are becoming clearer on the number of children they want. The general trend supports large family size. The percentage of women that consider four children as Ideal is consistently greater than those that opted for five. This may suggest that women prefer equal number of male and female children.

Table 1: Ideal number of children by Muslim women in Nigeria 1999-2003

Ideal number of children	**1990**	**1999**	**2003**
1	0.1	0.2	0.1
2	0.6	1.2	1
3	1.3	2.3	2.3
4	7.3	11.3	9.3
5	4.3	9.1	8.1
6+	12.5	48.1	62
Population size (N)	4269	3620	3601

5.7 Desire For Children By The Number of Living Children

Table 2 shows the proportion of Muslim women that want no more children by the number of living children in Nigeria. The desire to have no more children increases with increasing number of living children. The percentage of women with eight children and above who wants no more children has stabilized at forty-three percent. The general trend suggests that Muslim women want more children. This may support negative attitude towards contraceptive use adoption.

Table 2: Distribution of Muslim women aged 15-49 who want no more children by number of living children

Number of Living children	1990	1999	2003
0	1.3	1.6	1.1
1	3.8	1.4	0.9
2	4.5	3.8	3.7
3	6.9	10.3	6.7
4	12.9	15	15.2
5	17.5	22	22.7
6	29.5	36.1	27.1
7	33.6	32	32.6
8+	43.2	41.7	42.9
Total	9.9	11.3	10.1
Population size (N)	4269	3620	3601

5.8 Fertility and Contraceptives' Use by Background Among the Muslim Women in Nigeria

Table 3 shows Fertility and contraceptive use by background. Average number of children born by the urban women did not change over time. However, contraceptive use increased from 6.2 percent in 1990 to 8.2 percent in 2003. Average number of children born in rural area was higher by one compared to urban women. The use of modern contraceptive has increased from 1.1 percent in 1990 to 2.7 percent in 2003 among the rural women. The difference in contraceptive use between urban and rural women has remained constant at five percent over time.

Average number of children ever born decreases with higher educational attainment. The average number of children born has remained consistent at four for women with no education and three for those with primary over time. Women with secondary have an average of one child while those with higher education have an average number of two children. The use of modern contraceptive has increased from 1.3 percent in 1990 to 2.1 percent in 2003 for women with no education. Contraceptive use has increased from 6.2 percent in 1990 to 7.8 percent in 2003 among women with primary. Contraceptive use has increased by three percent between 1990 and 2003 among women with secondary education. However increase in contraceptive use was associated with increase in average children among those with secondary education. The use of contraceptive has declined by three percent among women with higher education between 1990 and 2003. The cause of the decline is not clear from the data. The average number of children ever born decreases with increasing affluence. The poorest echelon has one child higher than the middle class women. The richest has an average of two children. Contraceptive use increases with the level of affluence. The trend shows that increased contraceptive use is associated with low fertility. In summary average number of children ever born increased by one while contraceptive use increased by two percent between 1990 and 2003. There is consistent increasing trend on average number of children ever born in some of the background categories. However contraceptive use did not show consistent trend in most of these categories.

Table 3: Fertility and contraceptive use by background

	Average children ever born			Modern Contraceptive use		
Background	1990	1999	2003	1990	1999	2003
Residence						
Urban	2.9	3	3.2	6.2	8.3	8.2
Rural	3.5	3.2	3.8	1.1	2	2.7
Education						
No education	3.7	3.7	4.3	1.3	1.7	2.1
Primary	2.7	2.7	3.1	6.6	8.4	7.8
Secondary	1.1	1.3	1.6	8.2	7.4	11.2
Higher	1.7	2	1.7	25	22.2	21.6
Wealth Index						
Poorest			4.2			1.2
Poorer			4.1			1.8
Middle			3.6			3.6
Richer			3.3			7.3
Richest			2.4			13.8
Total	3.3	3.2	3.6	2.9	3.9	4.9

6. COMMENTS AND CONCLUSION

The ethics of Nigerian Muslim in support of no fertility regulation is unequivocal which does not give bright future to contraceptives use. Although Nigeria in general value children for many reasons, but the pride the Muslims take on children seem to be exceptional. In Hausaland, children are viewed gift and mortals have no powers to limit the number. Although religion is playing critical role in negative attitude to family planning, the desire for children among the women cannot be ruled out. Consistent with the previous reports, positive attitude to family planning increases with the number of living children (Duze & Mohammed 2006). The infant mortality rate is 100 per 1000 in Nigeria (Population Reference Bureau 2007). This rate is among the highest in the Sub Saharan Africa. Nigerians take pride in children and at the same it is a highly volatile state. Tribal and ethnic clashes are high and contribute reasonably to depopulation. There is high internal and external migration in which lives are lost as a result of illegal crossing of border. Hence, families are always in dread of death and extinction of lineage because no one accounts for those that left. The study shows that tendency to stop childbearing increases with the number of living children. This may suggest that the low contraceptive use and increasing no intention to use contraceptives may be partly due to low infant survival rate. The trend in the indicators may suggest that no reasonable changes have occurred in overall health and socioeconomic states of Muslim women in Nigeria. We posit that opting for large number of children as a means of ensuring continuity of lineage in the midst of high mortality vices may hinder adoption of contraception.

Our findings show that women are following suit with their husbands in disapproving FP. This may imply that both spouse and respondent have no place for FP. Disapproval of FP by both Spouse and respondents is an indication of general ignorance and in particular vitiation of reproductive rights of the women. Studies have shown that women stayed away from FP to please their husbands (Jinadu and Ajuwon 1997). The decision never to use contraception may be self denial of the women to maintain peace in the family as the culture and tradition demands. It is not the intention of the present study to incite women against their husband. However obfuscation of reproductive and sexual rights on the ground that it promotes immorality is detrimental to the life of women. Devoted Muslim men reject family planning on the ground that Islamic tenets forbid it are unaware of the stand of Islam on contraceptives and liberal interpretations of reproductive rights. There is an allusion that FP can lead to infertility or even death. Thus, the overall disapproval of FP observed in the present study is attributed to ignorance and lack of reproductive right knowledge. We suggest that FP service provider should integrate correction of overwhelming negative impression on FP which has been inculcated over several decades. However, we cannot rule out patriarchal clout of men on the decision taken by the women on FP.

The extreme neglected spousal discussion observed in the present study is in agreement with the high level disapproval of FP. The trend would have been considered as an effect of perennial gender inequality if considerable number of women approves FP or have raised the discussion. The extent to which culture and tradition may have discouraged spousal discussion is not known. A previous study has reported increase in spousal discussion on FP (Oladosu 2001) which is in conflict with the present findings. However, the previous studies examined spousal discussion at national level whereas the present study considered spousal discussion among Muslim women. Spousal discussion is milestone to adoption of FP. However, chances of FP adoption through the spousal communication for the Muslim women are highly negligible.

Findings of the study show that residence, education and wealth index play role on contraceptive use and average number of children ever born. It is not unexpected that the percentage of women using modern contraceptive in urban is greater compared to rural areas. However, the differential is quite unexpected considering the long time preference given to the urban in terms of medical services (Pearce 1980; Duze & Mohammed 2006). The difference in average number of children ever born does not reflect a good impact of contraceptive use on fertility. Consistent with previous reports, women who spent more years schooling has lower fertility than those with no or less education (Ashurt et. al. 1984). In addition, attitude to adopt contraceptive use increases with level of education and urbanization (Rodriguez and Aravena 1991) as well as wealth index. However, increase in contraceptive use does not show consistence with the average number of children ever born. For instance, the contraceptive prevalence rate for higher education was over twenty percent while that of secondary education was eleven percent, but the average number of children was the same. When we examine what proportion each category represents in the entire population as shown in Appendix, women with secondary are constantly higher than those with higher education. It is expected that average number of children born by women with higher education should be lower than that of women with secondary education due to differentials in contraceptive prevalence rate. However, this is not the case. It implies average number of children ever born shows inverse association with higher education and contraceptive use. The increasing trend in contraceptive prevalence by wealth index may be attributed to the ability to afford compatible method of contraception. Studies have shown that women abstain from contraceptive due to side effects. Modern contraceptive method has been tagged with infertility and death in the northern Nigeria. It requires patience and practical evidence to convince the poor to use it. Unfortunately, the rural, uneducated, and economically disadvantaged women constitute the bulk of the reproductive women. Their states have made them complacent to have large family size as security and pride. Possibility of the disadvantaged women seeing contraception as a strategy of elimination may not be ruled out, since it has been purported to cause death and infertility. In general, the increasing trend of average number of children ever born may suggest that Muslim population is not at risk of population ageing in Nigeria.

Desire for more children observed among the rural and poor women is not unexpected. The edifice of survival for rural, poor, and uneducated women in Nigeria is through subsistence agriculture where labour intensive is mandatory. Nigeria is a rural based nation with about eighty percent of the population engages in subsistence agriculture where children are productive agent. Northern part of Nigeria is predominantly Muslim is the food basket on the nation, has been backward in terms of education. Hence, programs that centers on providing contraception without education and poverty alleviation and liberation of women may have difficulty in accomplishing the purpose.

From all indication, limited commitment to Contraception is still persisting among the Muslim women in Nigeria through the instruments of cultural and traditional values and desire for large family size. Therefore, population ageing among the Nigeria Muslims may not be very imminent. Education and affluence are relevant for the success of FP among Muslim women. However, the proportions of educated and affluent women are still negligible to elicit reasonable increase on contraceptive use. Since there is gradual transformation in education among the women, the future of FP appears promising.

Appendix Percentage of Muslim women in each age group by background in Nigeria 1999-2003

			Residence				
		Urban				Rural	
Age group	1990	1999	2003		1990	1999	2003
15-19	41.6	28.6	41.3		58.4	71.4	58.7
20-24	39.8	31.5	41		60.2	68.5	59
25-29	33.9	30.6	39.7		66.1	69.4	60.3
30-34	36.3	29.9	37.4		63.7	70.1	62.6
35-39	29.5	29.1	39.7		70.5	70.9	60.3
40-44	31.4	33.1	36.5		68.6	66.9	63.5
45-49	26	31.1	38.4		74	68.9	61.6
			Education				
No education						Primary	
Age group	1990	1999	2003		1990	1999	2003
15-19	55.3	56.7	54.3		20.7	17.4	20.5
20-24	61.3	62.4	57		17.8	15.5	16.8
25-29	77.3	67.8	63.8		13.1	16.9	16.6
30-34	83.9	74.4	69.3		10.8	14.1	14.4
35-39	88	81.5	75.1		8	9.3	15.3
40-44	90.7	87.2	82.5		7	8.1	12.8
45-49	94.1	87.2	98.2		4.5	9.6	7.5
Secondary					Higher		
Age group	1990	1999	2003		1990	1999	2003
15-19	23.7	25.6	24.3		0.3	0.3	0.9
20-24	19.9	19.1	22.1		1	3	4.1
25-29	8.7	12.9	15.1		1	2.4	4.5
30-34	3.7	9.5	12.9		1.6	2	3.3
35-39	3.8	4.8	7.1		0.2	4.3	2.5
40-44	2.1	2.2	3.9		0.2	2.5	0.9
45-49	1.4	1.8	1.8		0	1.4	1.4

Wealth Index					
Age group	Poorest	Poorer	Middle	Richer	Richest
15-19	19.4	18.2	24.4	26.6	11.4
20-24	19.6	19.3	22.2	21.3	17.6
25-29	19.2	23.1	19.5	21.7	16.6
30-34	24.2	23	18.8	19.4	14.6
35-39	22.2	22.7	20.9	19.7	14.5
40-44	27	24.9	19.3	21.4	7.4
45-49	24.4	26.2	23.3	20.8	5.4

REFERENCES

Ashurst, H. S. and Casterline, J. B. (1984). Socio-economic differentials in recent fertility, *World Fertility Survey Comparative Studies no. 42. Voorburg: ISI.*

Bankole, A. and Bamisaye, O. (1985). The impact of petroleum production on the political economy of food in Nigeria since independence. *AMAN, 4*(2), 125-132.

Berer, M. (1996). Men. *Reproductive Health Matters, 7*, 7- 10.

Bongaarts, J. (2005). The causes of stalling fertility transitions. *Studies in Family Planning, 37*(1), 1- 16.

Caldwell, J. C. (2008). Three fertility compromises and two transitions. *Population Res Policy Review, 27,* 427 - 446.

Costa, G. (2008). Contraceptive Use Increases in Muslim-Dominated Northern Nigeria http://www.compassnigeria.org/site/DocServer/Success_Stories-October_2008-Couples.pdf?docID=382 Accessed 2000/10/05.

Dodoo, N. N. (1993). A couple analysis of micro-level supply demand factors in fertility regulation. *Population Research and Policy Review, 12*, 93 - 101.

Duze, M. C., and Mohammed, I. Z. (2006). Males' knowledge, attitudes, and family planning practices in northern Nigeria. *African Journal of Reproductive Health, 10*(3), 53 - 65.

Easterlin, R. and Crimmins, E. M. (1985). The fertility revolution: A supply and demand analysis. *(Chicago: University of Chicago Press).*

Federal Ministry of Health. (1988). National policy on population for development, unity, progress and self-reliance. Lagos: Govt. printed.

Feyisetan, B., Oyediran, A. K. and Ishola, G.P. (1998). The role of men in family planning in Nigeria. *Report Submitted to Population Research Fund, NISER, Ibadan, Nigeria.*

Feyisetan, B. J. and Bankole, (2002). A. Fertility transition in Nigeria: Trends and prospects,. *http://www.Un.org/esa/population/publications/completingfertility/RevisedBANKOLEpaper. PDF, Accessed Jul. 03, 2009.*

Ihejiamaizu, E. C. (2002). Adolescent fertility behaviour in Nigeria: Trends and determinants. *Global Journal of Social Sciences, 1*(1), 67-74.

Imam, A. (2003). Women, Muslim laws and human rights in Nigeria. *Africa Program; Woodrow Wilson Plaza Washington, D.C. 20004-3027.*

Isiugo-abanihe, U. C. (1994). The socio-cultural context of high fertility among Igbo women. *Journals of International Sociological Association, 9*(2), 237-58.

Jinadu, M. and Ajuwan, B. (1997). Traditional fertility regulation methods among the Yoruba of southwestern Nigeria. *African Journal of Reproductive Health, 1*(1), 65 - 73.

Khalfa, M. A. (1988). Attitude of urban Sudanese men toward family planning. *Studies in Family Planning, 19*(4), 231-243.

Lesthaeghe, R. (1989). Reproduction and social organization in Sub-Saharan Africa. *Berkeley; University of California Press, Berkeley.*

Lutz, W., O'Neill, B.C., and Scherbov, S. (2003). Europe's population at a turning point. *Science, 299*, 1991-1992.

Makinwa-Adebusoye, P. (1991). Adolescent reproductive behaviour in Nigeria. *A Study of Five Cities. NISER, Monograph Series no. 3.*

Mbizuo, T. M. and Adamchack, J. D. (1991). Family planning knowledge, attitude and practices of men in Zimbabwe. *Studies in Family Planning, 22*(1), 31-38.

Mohammed , I. Z. and Khalid, S. (1995). Access to reproductive health services in Nigeria: A study of vesico vaginal fistulae (VVF) in Kano. *A Paper Presented at the Annual Conference of Population Association of Nigeria.*

Mustapha, M. B. and Mumford, S. D. (1984). Male attitudes to family planning in khartoum sudan. *Journal of Biosiocial Science, 16*(4), 437-449.

Mwageni, E. A., Ankomah, A. and Powell, R. A. (2001). Sex preference and contraceptive behaviour among men in mbeya region, tanzania. *Journal of Family Planning and Reproductive Health Care, 27*(2), 85-89.

National Population Commission [Nigeria] and ORC Macro. 2004. Nigeria Demographic and Health Survey 2003: Key Findings. Calverton, Maryland, USA: National Population Commission and ORC Macro.

National Research Council. (1993). Factors affecting contraceptives use in Sub-Saharan Africa. *National Academy Press, Washington, D.C.*

NPC 2000. (2000.). Nigeria demographic and health survey 1999. *National Population Commission, Abuja, Nigeria.*

NPC and ORC Macro 2004. (2004). Nigeria demographic and health survey 2003. Calverton, Maryland. *National Population Commission and ORC Macro.*

Odimegwu, C. O. (1999). Family planning attitudes and use in Nigeria: A factor analysis. *International Family Planning Perspectives, 25*(2), 86-91.

Oladosu, M. (2001). Prospects for fertility decline in Nigeria: Comparative analysis of the 1990 and 1999 NDHS data. *Population Division, Department of Economic and Social Affairs United Nations Secretariat.*

Olusanya, P. (1969). Nigeria: Cultural barrier to family planning among the Yoruba. *Studies in Family Planning, 33*, 13-16.

Oni, A.G. and McCarthy, J. (1990). Contraceptive knowledge and practices in Ilorin 1983-1988. *Studies in Family Planning, 21*(2), 104-109.

RedOrbit, (2004). Nigeria adopts new population policy to improve quality of life - Science News 29 January - redOrbit.mht Accessed, August, 17, 2009.

Orubuloye, I. O. (1998). Fertility transition in southwest Nigeria in the era of structural adjustment. *Paper Presented at the IUSSP Seminar on Reproductive Change in Sub-Sahara Africa, Nairobi, Kenya. November 2-4.*

Pearce, T. (1980). Political and economic changes in Nigeria and the organization of Medicare care. *Social Science and Medicine, 14B*, 91- 98.

Renne, E. P. (1996). Perception of population policy development and family planning programs in north Nigeria. *Studies in Family Planning, 27*(3), 127-136.

Rodriguez, G. and Aravena, R. (1991). Socio-economic factors and the transition to low fertility in less developed countries: A comparative analysis. *Demographic and Health Surveys World Conference, Washington, Volume 1. Columbia, MD: IRD/Macro International, pp. 39-72.*

Sundquist, B. (2008). The Muslim world's changing views toward family planning and contraception. Http://home.Windstream.net/bsundquist1

Tola, O. P. (2001). Women, the state and reproductive health in Nigeria. *Journal of Culture and African Women Studies.*

United Nations. (2000). World population prospects. *The 2000 Revision Highlights. United Nations. New York. DRAFT ESA/P/WP. 165, 28 Febuary 2001.*

United Nations Population Fund. (1996). Programme review and strategy development report: Nigeria.

United States of America, Bureau of the Census. The international data base (IDB). Http://www.Census.gov/ipc/www/idbnew.Html.

White, H. (2009). Spain's "Disastrous" Contraceptive Policies have resulted in the Oldest European Population www.lifesitenews.com/ Accessed 2009/10/06.

Whyte, M. and Gu, S. Z. (1987). Popular response to china's fertility transition. *Population and Development Review, 13*(3), 569-571.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 950–962.

CALCULATION OF RUIN PROBABILITY IN RISK PROCESS USING DE VYLDER'S METHOD

Hasih Pratiwi
Sebelas Maret University, Surakarta, Indonesia
Ph.D. student, Gadjah Mada University, Yogyakarta, Indonesia
E-mail: hasihpratiwi@ymail.com

Subanar
Gadjah Mada University, Yogyakarta, Indonesia
E-mail: subanar@ugm.ac.id

Danardono
Gadjah Mada University, Yogyakarta, Indonesia
E-mail: danardono@ugm.ac.id

J.A.M. van der Weide
Delft University of Technology, Delft, The Netherlands
E-mail: jamvanderweide@tudelft.nl

ABSTRACT

In risk theory, the risk process is a very important model for understanding how the capital or surplus of an insurance company evolves over time. By adding to the previous surplus the current premium flow and deducting the claims made during the period, the process gives the value of the capital that is available to the insurer at each point in time. Each period is tracked so that the surplus never gets below zero because if it does, it provides an indication of ruin, that is, the company is in a position of negative cash flow. In this paper we describe an approximation, known as De Vylder's method, to calculate the ruin probability. We find the solution for ruin probability with assuming Lundberg's inequality applies. All that is required to apply De Vylder's approximation is that the first three moments of the individual claim amount distribution exist. In situation when the adjustment coefficient exists, the method provides good approximations when ruin probabilities are small.

Keywords: risk process, ruin probability; Lundberg's inequality; survival probability; De Vylder's method.

1. INTRODUCTION

Insurance companies are in the business of risks. They exist to pool together risks faced by individuals or companies who in the event of a loss are compensated by the insurer to reduce the financial burden. In its simplest form, when certain events occur, an insurance contract will provide the policyholder the right to claim all or a portion of the loss. In exchange for this

entitlement, the policyholder pays a specified amount called the premium and the insurer is obligated to honor its promises when they come due (Valdez and Mo, 2002).

In this paper, we simplify a real life insurance operation by assuming that the insurer starts with some non-negative amount of money, collects premiums, and pays claims as they occur. Our model of an insurance surplus process is then deemed to have three components: initial surplus (or surplus at time zero), premiums received and claims paid. If the insurer's surplus falls at zero or below, we say that ruin occurs.

The aim of this work is to approximate a ruin probability in risk process. We start in Section 2 by describing a risk process and in Section 3 we give some definitions of ruin probability. We then consider the adjustment coefficient in Section 4 and prove Lundberg's inequality in Section 5. In Section 6 we derive a survival probability and finally in Section 7 we use De Vylder's method to approximate the ruin probability.

2. RISK PROCESS

In the classical risk process, an insurer's surplus at a fixed time $t>0$ is determined by three quantities: the amount of surplus at time 0, the amount of premium income received up to time *t*, and the amount of claims paid out up to time *t*. The only one of these which is random is claims outgo, so we start by describing the aggregate claim process, which we denote by $\{S(t)\}_{t\geq 0}$.

Let $\{N(t)\}_{t\geq 0}$ be a counting process for the number of claims, so that for a fixed value $t>0$, the random variable $N(t)$ denotes the number of claims that occur in the fixed time interval $[0,t]$. In the classical risk process, it is assumed that $\{N(t)\}_{t\geq 0}$ is a Poisson process (Dickson, 2005). Individual claim amounts are modelled as a sequence of independent and identically distributed random variables $\{X_i\}_{i=1}^{\infty}$, so that X_i denotes the amount of the *i*th claim. We can then say that the aggregate claim amount up to time $t, S(t)$, is

$$S(t)=\sum_{i=1}^{N(t)} X_i$$

with the understanding that $S(t)=0$ when $N(t)=0$. The aggregate claim process $\{S(t)\}_{t\geq 0}$ is then a compound Poisson process (Dickson, 2005).

We can now describe the risk process, denote by $\{U(t)\}_{t\geq 0}$, as

$$U(t)=u+ct-S(t) \tag{2.1}$$

where *u* is the insurer's surplus at time 0 and *c* is the insurer's rate of premium income per unit time, which we assume to be received continuously. A realization of this risk process is depicted in Figure 1.

Throughout this section we denote the distribution function of X_1 by *F* and we assume that $F(0)=0$ so that all claim amounts are positive. For simplicity, we assume that this distribution is continuous with density function f and the *k*th moment of X_1 is denoted by m_k. Whenever the

moment generating function of X_1 exists, there exists some quantity $\gamma, 0 < \gamma \le \infty$, such that $M_x(r)$ is finite for all $r < \gamma$ with

$$\lim_{r \to \gamma^-} M_x(r) = \infty$$

As an illustration, suppose that X_1 is exponentially distributed with parameter λ. Then $M_x(r) = \int_0^\infty e^{xx} \lambda e^{-\lambda x} dx = \int_0^\infty \lambda e^{-(\lambda - r)x} dx = \frac{\lambda}{\lambda - r}$ for $r < \lambda$ and

$$\lim_{r \to \lambda^-} M_x(r) = \infty$$

so that in this case, γ is λ.

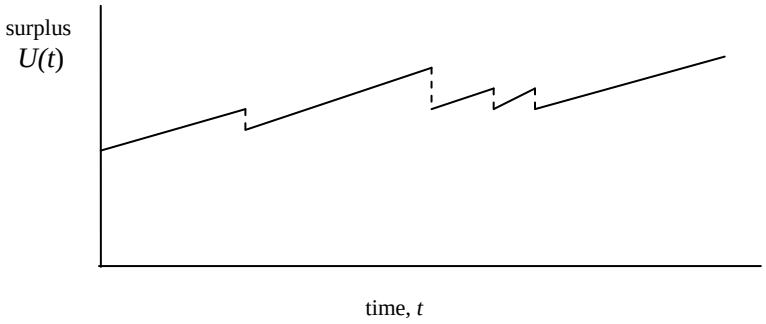


Figure 1. A realization of a risk process

3. RUIN PROBABILITY

The probability of ruin in infinite time, also known as the ultimate ruin probability, is defined as $\psi(u) = \Pr(U(t) < 0 \text{ for some } t > 0)$. In words, $\psi(u)$ is the probability that the insurer's surplus falls below zero at some time in the future, that is that claims outgo exceeds the initial surplus plus premium income. This is a probability of ruin in continuous time, and we can also define a discrete time ultimate ruin probability as $\psi_r(U) = \Pr(U(t) < 0 \text{ for some } t, t = r, 2r, 3r, ...)$. Thus, under this definition, ruin occurs only if the surplus is less than zero at one of the time pints $r, 2r, 3r, ...$ If ruin occurs under the discrete time definition, it must also occur under the continuous time definition. However, the opposite is not true. To see this, we consider a realization of a surplus process which, for some integer n, has $U(nr) > 0$ and $U((n+1)r) > 0$ with $U(\tau) < 0$ for some $\tau \in (nr, (n+1)r)$. If $U(t) > 0$ for all t outside the interval $(nr, (n+1)r)$, then ruin occurs under the continuous time definition. Thus $\psi_r(u) < \psi(u)$. However, as r becomes small, so that we are 'checking' the surplus level very frequently, then $\psi_r(u)$ should be a good approximation to $\psi(u)$.

We define the finite time ruin probability $\psi(u,t)$ by $\psi(u,t) = \Pr(U(s) < 0 \text{ for some } s, 0 < s \le t)$. Thus, $\psi(u,t)$ is the probability that the insurer's surplus falls below zero in the finite time interval $(0,t]$. We can also define a discrete time ruin probability in finite time as

$$\psi_r(u,t) = \Pr(U(s) < 0 \text{ for some } s, 0 < s \le t)$$

where t is an integer multiple of r. The arguments used above to explain why $\psi_r(U) < \psi(u)$ also apply in finite time to give $\psi_r(u,t) < \psi(u,t)$, and if r is small then $\psi_r(u,t)$ should be a good approximation to $\psi(u,t)$ (Dickson, 2005; Burnecki and Mista, 2007).

In this work we concentrate mostly on the ultimate ruin probability and we assume that $c > \lambda m_1$, so that the premium income exceeds the expected aggregate claim amount per unit of time. We can find the expected value of $S(t)$ as follows:

$$E[S(t)] = \sum_{k=0}^{\infty} E[S(t)|N(t)=k]\Pr(N(t)=k)$$

because $S(t)=0$ when $N(t)=0$, we have

$$E[S(t)] = \sum_{k=1}^{\infty} E\left[\sum_{i=1}^{k} X_i \,|\, N(t)=k\right]\Pr(N(t)=k) = \sum_{k=1}^{\infty} E\left[\sum_{i=1}^{k} X_i\right]\Pr(N(t)=k)$$

and

$$E[S(t)] = \sum_{k=1}^{\infty} k m_1 \Pr(N(t)=k) = m_1 \sum_{k=1}^{\infty} k \Pr(N(t)=k) = m_1 E[N(t)] = \lambda m_1 t$$

because $\{X_i\}$ and $\{N(t)\}$ are independent. From equation (2.1) we obtain $E[U(t)] = u + ct - \lambda m_1 t$. If the condition $c > \lambda m_1$, known as the net profit condition, does not hold, then $\psi(u) = 1$ for all $u \geq 0$. It is often convenient to write $c = (1+\theta)\lambda m_1$, so that θ is the premium loading factor.

4. THE ADJUSTMENT COEFFICIENT

The adjustment coefficient, which we denote by *R*, gives a measure of risk for a surplus process. It takes account of two factors in the surplus process: aggregate claims and premium income. For the classical risk process, the adjustment coefficient is defined to be the unique positive root of

$$\lambda M_X(r) - \lambda - cr = 0 \qquad (4.1)$$

so that *R* is given by

$$\lambda + cR = \lambda M_X(R). \qquad (4.2)$$

The root is illustrated in Figure 2. By writing c as $(1+\theta)\lambda m_1$, we can see that *R* is independent of the Poisson parameter λ, and we discuss this point further in the next section. To see that there is a unique positive root of equation (4.1) we consider the function

$$g(r) = \lambda M_X(r) - \lambda = cr.$$

First note that $g(0) = 0$. Second,

$$\frac{d}{dr} g(r) = \lambda \frac{d}{dr} M_X(r) - c$$

so that $\frac{d}{dr} g(r)\big|_{r=0} = \lambda m_1 - c$

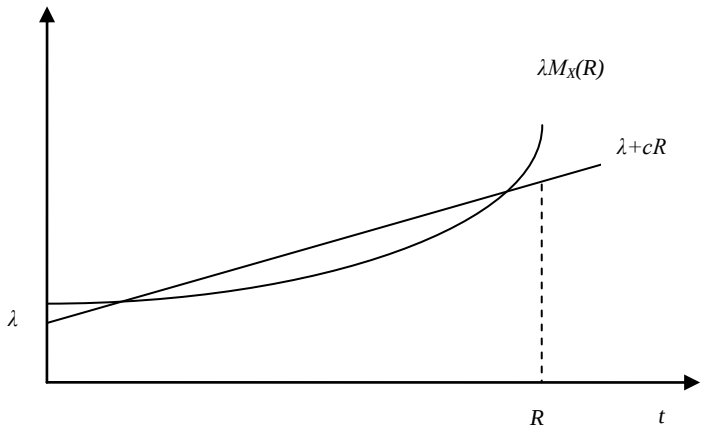


Figure 2. The adjustment coefficient

and hence g is a decreasing function at zero as we have assumed that $c > \lambda m_1$. Next we note that

$$\frac{d^2}{dr^2} g(r) = \lambda \frac{d^2}{dr^2} M_X(r) = \lambda \frac{d^2}{dr^2} E\left(e^{rx}\right) = \lambda \int_0^\infty x^2 e^{rx} f(x)\,dx > 0$$

so that if g has a turning point, then the function attains its minimum at that turning point. Finally, we note that

$$\lim_{r \to \gamma^-} g(r) = \infty \tag{4.3}$$

where γ is as defined in Section 2, so that as g is decreasing at zero, the function must have a unique turning point, and hence there is a unique positive number R such that $g(R) = 0$. To see that equation (4.3) is true, we consider separately the cases $\gamma < \infty$ and $\gamma = \infty$. In the former case, equation (4.3) clearly holds. In the latter case, we note that since all claim amounts are positive, there exist a positive number ε and a probability p such that

$$\Pr(X_1 > \varepsilon) = p > 0$$

so that

$$M_X(r) = \int_0^\infty e^{rx} f(x)\,dx \geq \int_\varepsilon^\infty e^{xx} f(x)\,dx \geq e^{r\varepsilon} p\,.$$

and hence

$$\lim_{r \to \infty} g(r) \geq \lim_{r \to \infty} \left(\lambda e^{r\varepsilon} p - \lambda - cr\right) = \infty$$

5. LUNDBERG'S INEQUALITY

For the classical risk process Lundberg's inequality states that

$$\psi(u) \leq \exp\{-Ru\} \tag{5.1}$$

where R is the adjustment coefficient (Dickson, 2005; Asmussen, 2000). In many cases the right-hand side of (5.1) is a good approximation of $\psi(u)$ (Mammitzsch, 1986). We can prove this result by an inductive argument. We define $\psi_n(u)$ to be the probability of ruin at or before the *n*th claim. It is then sufficient to show that $\psi_n(u) \le \exp\{-Ru\}$ for n =1, 2, 3,..., since $\psi(u) = \lim_{n\to\infty} \psi_n(u)$. We first must show that the result is true when n =1. We consider the time and the amount of the first claim. Suppose that the first ruin occurs at time $t > 0$ and that the amount of this claim is *x*. If ruin occurs at the first claim, then $x > u + ct$ or $u + ct - x < 0$ so that

$$e^{-R(u+ct-x)} \ge 1$$

and we have

$$\begin{aligned}
\psi_1(u) &= \int_0^\infty \lambda e^{-\lambda t} \int_{u+ct}^\infty f(x)\,dx\,dt \\
&\le \int_0^\infty \lambda e^{-\lambda t} \int_{u+ct}^\infty f(x) e^{-R(u+ct-x)}\,dx\,dt \\
&\le \int_0^\infty \lambda e^{-\lambda t} \int_0^\infty f(x) e^{-R(u+ct-x)}\,dx\,dt \\
&= e^{-Ru}
\end{aligned}$$

Next, we assume that for a fixed value of n, where $n \ge 1, \psi_n(u) \le \exp\{-Ru\}$. We establish an expression for $\psi_{n+1}(u)$ by considering the time and the amount of the first claim as previous step. If ruin occurs at or before the (n + 1)th claim, then either (1) ruin occurs at the first claim, so that $x > u + ct$, or (2) ruin does not occur at the first time, so that the surplus after payment of this claim, $u + ct - x$, is non-negative, and ruin occurs from this surplus level at one of the next n claims.

Since claims occur as a Poisson process (with parameter λ), the distribution of the time until the first claim is exponential with parameter λ. Hence, integrating over all possible times and amounts for the first claim we have

$$\psi_{n+1}(u) = \int_0^\infty \lambda e^{-\lambda t} \int_{u+ct}^\infty f(x)\,dx\,dt + \int_0^\infty \lambda e^{-\lambda t} \int_0^{u+ct} f(x)\psi_n(u+ct-x)\,dx\,dt\,.$$

The first integral represents the probability of ruin at the first claim and the second represents the probability that ruin does not occur at the first claim but does occur at one of the next n claims. Note that in probabilistic terms the surplus process 'starts over' again after payment of the first claim, and so the probability of ruin within n claims after payment of the first claim is just $\psi_n(u + ct - x)$.

We now apply our inductive hypothesis to write

$$\psi_{n+1}(u) \le \int_0^\infty \lambda e^{-\lambda t} \int_{u+ct}^\infty f(x)\,dx\,dt + \int_0^\infty \lambda e^{-\lambda t} \int_0^{u+ct} f(x) e^{-R(u+ct-x)}\,dx\,dt.$$

Next, we use the fact that $\exp\{-R(u+ct-x)\} \ge 1$ for $x \ge u + ct$, so that

$$\int_{u+ct}^\infty f(x)\,dx \le \int_{u+ct}^\infty e^{-R(u+ct-x)} f(x)\,dx$$

and hence

$$\begin{aligned}\psi_{n+1}(u) &\le \int_0^\infty \lambda e^{-\lambda t}\int_0^\infty f(x)e^{-R(u+ct-x)}dxdt\\ &= e^{-Ru}\int_0^\infty \lambda e^{-(\lambda+cR)t}\int_0^\infty e^{Rx}f(x)dxdt\\ &= e^{-Ru}\int_0^\infty \lambda e^{-(\lambda+cR)t}M_X(R)dt.\end{aligned}$$

Since $\lambda + cR = \lambda M_X(R)$,

$$\begin{aligned}\psi_{n+1}(u) &\le e^{-Ru}\int_0^\infty \lambda M_X(R)e^{-\lambda M_X(R)t}dt\\ &= \exp\{-Ru\}\end{aligned}$$

because the integral equals 1, and hence the proof is complete.

6. SURVIVAL PROBABILITY

We define $\phi(u) = 1-\psi(u)$ to be the probability that ruin never occurs starting from initial surplus u, a probability also known as the survival probability. An equation for ϕ can be established by adapting the reasoning used to prove Lundberg's inequality. By considering the time and the amount of the first claim, we have

$$\phi(u) = \int_0^\infty \lambda e^{-\lambda t}\int_0^{u+ct} f(x)\phi(u+ct-x)dxdt \tag{6.1}$$

noting that if the first claim occurs at time t, its amount must not exceed $u+ct$, or $x < u+ct$, since ruin otherwise occurs. Substituting $s = u+ct$ in equation (6.1) we have $t = (s-u)/c$ and we get

$$\phi(u) = \frac{1}{c}\int_u^\infty \lambda e^{-\lambda(s-u)/c}\int_0^s f(x)\phi(s-x)dxds$$

or

$$\phi(u) = \frac{\lambda}{c}e^{\lambda u/c}\int_u^\infty e^{-\lambda s/c}\int_0^s f(x)\phi(s-x)dxds \tag{6.2}$$

We can establish an equation for ϕ, known as an integro-differential equation, by diffrentiating equation (6.2), and the resulting equation can be used to derive explicit solutions for ϕ. Differentiation gives

$$\frac{d}{du}\phi(u) = \frac{\lambda^2}{c^2}e^{\lambda u/c}\int_u^\infty e^{-\lambda s/c}\int_0^s f(x)\phi(s-x)dxds - \frac{\lambda}{c}\int_0^u f(x)\phi(u-x)dx$$

and from equation (6.2) we have

$$\frac{d}{du}\phi(u) = \frac{\lambda}{c}\phi(u) - \frac{\lambda}{c}\int_0^u f(x)\phi(u-x)dx \tag{6.3}$$

As the first equation (6.3) does not appear to be a very promising route, since the function ϕ appears in three different places in this equation. However, by eliminating the integral term, a

differential equation can be created, and solved. To see how such an approach works, let us consider the situation when $F(x)=1-e^{-\alpha x}$. Then we have

$$\frac{d}{du}\phi(u)=\frac{\lambda}{c}\phi(u)-\frac{\lambda}{c}\int_0^u \alpha e^{-\alpha x}\phi(\pi u-x)dx$$
$$=\frac{\lambda}{c}\phi(u)-\frac{\alpha\lambda}{c}\int_0^u e^{-\alpha(u-x)}\phi(x)dx$$

or, equivalently,

$$\frac{d}{du}\phi(u)=\frac{\lambda}{c}\phi(u)-\frac{\alpha\lambda}{c}e^{-\alpha u}\int_0^u e^{\alpha x}\phi(x)dx. \tag{6.4}$$

Differentiation of equation (6.4) yields

$$\frac{d^2}{du^2}\phi(u)=\frac{\lambda}{c}\frac{d}{du}\phi(u)-\frac{\alpha\lambda}{c}\left(-\alpha e^{-\alpha u}\int_0^u e^{\alpha x}\phi(x)dx+e^{\alpha u}e^{\alpha u}\phi(u)\right)$$
$$=\frac{\lambda}{c}\frac{d}{du}\phi(u)+\frac{\alpha^2\lambda}{c}e^{-\alpha u}\int_0^u e^{\alpha x}\phi(x)dx-\frac{\alpha\lambda}{c}\phi(u)$$

and we can write as

$$\frac{d^2}{du^2}\phi(u)=\frac{\lambda}{c}\frac{d}{du}\phi(u)-\alpha\left(\frac{\lambda}{c}\phi(u)-\frac{\alpha\lambda}{c}e^{-\alpha u}\int_0^u e^{\alpha x}\phi(x)dx\right) \tag{6.5}$$

The integral term in equation (6.5) is simply the integral term ini equation (6.4) multiplied by $-\alpha$. Hence, if we multiply equation (6.4) by α we have

$$\alpha\frac{d}{du}\phi(u)=\alpha\left(\frac{\lambda}{c}\phi(u)-\frac{\alpha\lambda}{c}-e^{-\alpha u}\int_0^u e^{\alpha x}\phi(x)dx\right) \tag{6.6}$$

Adding the resulting equation (6.6) to equation (6.5) we find that

$$\frac{d^2}{du^2}\phi(u)+\alpha\frac{d}{du}\phi(u)=\frac{\lambda}{c}\frac{d}{du}\phi(u)$$

or

$$\frac{d^2}{du^2}\phi(u)+\left(\alpha-\frac{\lambda}{c}\right)\frac{d}{du}\phi(u)=0. \tag{6.7}$$

This is a second order differential equation and we can find its general solution as follows. Let $\phi(u)=e^{ru}$, we have $\frac{d}{du}\phi(u)=re^{ru}$ and $\frac{d^2}{du^2}\phi(u)=r^2e^{ru}$. Then we can write equation (6.7) as

$$r^2e^{ru}+\left(\alpha-\frac{\lambda}{c}\right)re^{ru}=0$$

or

$$r^2+\left(\alpha-\frac{\lambda}{c}\right)r=0 \tag{6.8}$$

The solution of equation (6.8) is $r=0$ or $r=-\left(\alpha-\frac{\lambda}{c}\right)$. Hence the general solution of equation (6.7) is

$$\phi(u)=c_1+c_2e^{-(\alpha-\lambda/c)u} \tag{6.9}$$

where c_1 and c_2 are constants. Since Lundberg's inequality (5.1) applies, we know that and $\lim_{u\to\infty}\psi(u)=0$ and $\lim_{u\to\infty}\phi(u)=1$. From equation (6.9) we have

$$\lim_{u\to\infty}\phi(u)=\lim_{u\to\infty}\left(c_1+c_2e^{-(\alpha-\lambda/c)u}\right)=1$$

which gives $c_1=1$. It then follows that $\phi(0)=1+c_2$, that is $c_2=\phi(0)-1=-\psi(0)$, so that

$$\phi(u)=1-\psi(0)e^{-(\alpha-\lambda/c)u} \tag{6.10}$$

All that remains is to solve for $\psi(0)$, and this can be done generally on the assumption that Lundberg's inequality applies. Writing $\phi=1-\psi$ in equation (6.3) it follows that

$$\begin{aligned}-\frac{d}{du}\psi(u)&=\frac{\lambda}{c}\left(1-\psi(u)\right)-\frac{\lambda}{c}\int_0^u f(x)\left(1-\psi(u-x)\right)dx\\&=\frac{\lambda}{c}-\frac{\lambda}{c}\psi(u)-\frac{\lambda}{c}\int_0^u f(x)dx+\frac{\lambda}{c}\int_0^u f(x)\psi(u-x)dx\end{aligned}$$

or

$$\begin{aligned}\frac{d}{du}\psi(u)&=-\frac{\lambda}{c}+\frac{\lambda}{c}\psi(u)+\frac{\lambda}{c}\int_0^u f(x)dx-\frac{\lambda}{c}\int_0^u f(x)\psi(u-\pi)dx\\&=\frac{\lambda}{c}\psi(u)-\frac{\lambda}{c}\int_0^u f(x)\psi(u-x)dx-\frac{\lambda}{c}\left(1-\frac{\lambda}{c}\int_0^u f(x)dx\right)\\&=\frac{\lambda}{c}\psi(u)-\frac{\lambda}{c}\int_0^u f(x)\psi(u-x)dx-\frac{\lambda}{c}\left(1-F(u)\right)\end{aligned}$$

Integrating this equation over $(0,\infty)$ we find that

$$-\psi(0)=\frac{\lambda}{c}\int_0^\infty\psi(u)du-\frac{\lambda}{c}\int_0^\infty\int_0^u f(x)\psi(u-x)dxdu-\frac{\lambda}{c}\int_0^\infty\left(1-F(u)\right)du. \tag{6.11}$$

Changing the order of integration in the double integral in equation (6.11), we have

$$\begin{aligned}\int_0^\infty\int_0^u f(x)\psi(u-x)dxdu&=\int_0^\infty\int_x^\infty\psi(u-x)duf(x)dx\\&=\int_0^\infty\int_0^\infty\psi(y)dyf(x)dx\\&=\int_0^\infty\psi(y)dy\end{aligned}$$

Thus, the first two terms on the right-hand side of equation (6.11) cancel, and we find that

$$\psi(0)=\frac{\lambda}{c}\int_0^\infty\left(1-F(u)\right)du=\frac{\lambda m_1}{c} \tag{6.12}$$

which holds generally. We did not have to specify the form of F to prove this result, but we did assume that Lundbergs's inequality applies.

For example, if $F(x)=1-e^{-\alpha x}, x\geq 0$, then

$$\psi(0)=\frac{\lambda}{c}\int_0^\infty\left(1-\left(1-e^{-\alpha u}\right)\right)du=\frac{\lambda}{c}\int_0^\infty -e^{-\alpha u}du=\frac{-\lambda}{\alpha c}e^{-\alpha u}\Big|_0^\infty=\frac{\lambda}{\alpha c}$$

From equation (6.10), the complete solution for ϕ is

$$\phi(u)=1-\frac{\lambda}{\alpha c}\exp\left\{-(\alpha-\lambda/c)u\right\} \tag{6.13}$$

Although this method of solution can be used for other forms of *F*, we do not pursue it further.

In section 4 we see that if the premium is written as $c=(1+\theta)\lambda m_1$, then the adjustment coefficient is independent of λ. If we write c in this way in equation (6.13) then

$$\begin{aligned}\phi(u)&=1-\frac{\lambda}{\alpha(1+\theta)\lambda m_1}\exp\left\{-\left(\alpha-\frac{\lambda}{(1+\theta)\lambda m_1}\right)u\right\}\\&=1-\frac{1}{\alpha(1+\theta)m_1}\exp\left\{-\left(\alpha-\frac{1}{(1+\theta)m_1}\right)u\right\}\end{aligned}$$

independent of λ. For $\alpha=1$we have

$$\begin{aligned}\phi(u)&=1-\frac{1}{1+\theta}\exp\left\{-\left(1-\frac{1}{1+\theta}\right)u\right\}\\&=1-\frac{1}{1+\theta}\exp\left\{-\left(\frac{\theta}{1+\theta}\right)u\right\}\end{aligned}$$

7. APPROXIMATE CALCULATION OF RUIN PROBABILITY

In this section we describe an approximation, known as De Vylder's method, to approximate the ruin probability. Suppose we have a classical risk process $\{U(t)\}_{t\geq 0}$ for which we wish to calculate the probability of ultimate ruin. We can approximate the risk process by a process $\{\tilde{U}(t)\}_{t\geq 0}$ given by

$$\tilde{U}(t)=u+\tilde{c}t-\tilde{S}(t),$$

where the aggregate claim process $\{S(t)\}_{t\geq 0}$ is a compound Poisson process with parameter $\tilde{\lambda}$ and individual claim amount distribution $\tilde{F}(x)=1-\exp(\tilde{\alpha}x), x\geq 0$. Thus, a process $\{\tilde{U}(t)\}_{t\geq 0}$ has the following characteristics (Dickson, 2005):

- $\tilde{U}(0)=u$
- the Poisson parameter is $\tilde{\lambda}$
- the premium income per unit time is $\tilde{c}$
- the individual claim amount distribution is $\tilde{F}(x)=1-\exp(\tilde{\alpha}x), x\geq 0.$

Since the individual claim amount distribution in the approximating risk process is exponential with parameter $\tilde{\alpha}$, it follows that the probability of ultimate ruin for the risk process $\{\tilde{U}(t)\}_{t\geq 0}$, is

$$\frac{\tilde{\lambda}}{\tilde{\alpha}\tilde{c}}\exp\left\{-\left(\tilde{\alpha}-\frac{\tilde{\lambda}}{\tilde{c}}\right)u\right\}.$$

This is De Vylder's approximation to the ultimate ruin probability for the risk process $\{U(t)\}_{t\geq 0}$. The parameters $\tilde{\lambda}$, $\tilde{c}$ and $\tilde{\alpha}$ are chosen by matching moments of the two risk processes. We set

$$E[U(t)]=E[\tilde{U}(t)]$$
$$E\left[\left(U(t)-E[U(t)]\right)^2\right]=E\left[\left(\tilde{U}(t)-E[\tilde{U}(t)]\right)^2\right]$$
$$E\left[\left(U(t)-E[U(t)]\right)^3\right]=E\left[\left(\tilde{U}(t)-E[\tilde{U}(t)]\right)^3\right]$$

which give

$$\tilde{c}=c-\lambda m_1+\tilde{\lambda}/\tilde{\alpha} \tag{7.1}$$

$$\tilde{\alpha}=3m_2/m_3 \tag{7.2}$$

and

$$\tilde{\lambda}=\frac{9\lambda m_2^3}{2m_3^2} \tag{7.3}$$

We give two examples to show the approximation. First, if $f(x) = 4xe^{-2x}$, $x>0$ and $c=1.2\lambda$, we have $m_1 = 1$, $m_2 = 3/2$ and $m_3 = 3$. Applying the Laplace transform, we have the exact solution for $\phi(u)$:

$$\phi(u) = 1 - 0.8518\, e^{-0.2268u} + 0.0185\, e^{-2.9399u}.$$

By equation (7.1) - (7.3) we obtain

$$\tilde{\lambda}=\frac{27\lambda}{16},\tilde{\alpha}=\frac{3}{2},\tilde{c}=\frac{53\lambda}{40}$$

and the approximation to $\psi(u)$ is

$$\frac{45}{53}\exp\left\{-\frac{12u}{53}\right\}$$

Table 1. Exact and approximate values of $\psi(u)$ for $f(x) = 4xe^{-2x}$, $x>0$

u	Exact	Approximate
0	0.8333	0.8491
1	0.6780	0.6770
2	0.5411	0.5399
3	0.4314	0.4305
4	0.3438	0.3433
5	0.2741	0.2737
6	0.2184	0.2182
7	0.1741	0.1740
8	0.1388	0.1388
9	0.1106	0.1107
10	0.0882	0.0882
11	0.0703	0.0704
12	0.0560	0.0561
13	0.0447	0.0447
14	0.0356	0.0357
15	0.0284	0.0284
16	0.0226	0.0227
17	0.0180	0.0181
18	0.0144	0.0144
19	0.0115	0.0115

Table 1 shows exact and approximate values of $\psi(u)$ for u = 0,1,2, …, 19. Second, for $f(x) = exp(-2x) + 1/3\ exp(-2/3\ x)$ and the loading factor is 10%, we have

$$\phi(u) = 1 - 0.8984\ e^{-0.0719u} + 0.0107\ e^{-1.6857u}.$$

The parameters in De Vylder's approximation are $\tilde{\alpha} = 5/7, \tilde{\lambda} = 125/196, \tilde{c} = 139/140$ and the numerical values are given in Table 2. We can see from these tables that the approximation is very good when ruin probabilities are small.

8. CONCLUDING REMARKS

We found the solution for ψ with assuming Lundberg's inequality applies. All that is required to apply De Vylder's approximation is that the first three moments of the individual claim amount distribution exist. In situation when the adjustment coefficient exists, the method provides good approximations when ruin probabilities are small. However the approximation is inaccurate for small values of u, especially u = 0. Generally, the method is not particularly accurate when the adjustment coefficient does not exist.

Table 2. Exact and approximate values of $\psi(u)$ for $f(x) = exp(-2x) + 1/3\ exp(-2/3\ x)$

u	Exact	Approximate
0	0.8877	0.8993
5	0.6271	0.6276
10	0.4377	0.4380
15	0.3056	0.3057
20	0.2133	0.2133
25	0.1489	0.1489
30	0.1039	0.1039
35	0.0725	0.0725
40	0.0506	0.0506
45	0.0353	0.0353
50	0.0247	0.0246
55	0.0172	0.0172
60	0.0120	0.0120

ACKNOWLEDGMENTS

This work would not have been possible without the support of The Indonesian Ministry of National Education, Sebelas Maret University and Gadjah Mada University. In this connection the authors would like to express appreciation to Sebelas Maret University and Gadjah Mada University for giving the opportunity to present this paper in The Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X) 2009. We also would like to thank to Directorate of Higher Education, The Indonesian Ministry of National Education, through Hibah Penelitian Sesuai Prioritas Nasional, for giving the financial support to finish our research.

REFERENCES

Asmussen S. (2000), *Ruin probabilities*, World Scientific Publishing Co. Pte. Ltd., Singapore.

Burnecki K. and Mista P. (2007), *Ruin Probability in Infinite Time,* Technical Report, Hugo Steinhaus Center, www.im.pwr.wroc.pl/hugo.html.

Dickson D.C. (2005), *Insurance risk and ruin*, Cambridge University Press, Cambridge.

Mammitzsch V. (1986), A note on the adjustment coefficient in ruin theory, *Insurance: Mathematics and Economics* 5, 147-149.

Valdez E.A. and Mo K. (2002), *Ruin probability with dependent claims*, Technical Report, University of New South Wales, Sydney, Australia.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 963–976.

PRESERVING SEMANTIC CONTENT IN TEXT MINING USING MULTIGRAMS

Yasmin H. Said and Edward J. Wegman

Department of Computational and Data Sciences
George Mason University, Fairfax, VA 22030 USA
E-mails: ysaid99@hotmail.com, ewegman@gmail.com

ABSTRACT

Text mining can be thought of as a synthesis of information retrieval, natural language processing, and statistical data mining. The set of documents being considered can scale to hundreds of thousands and the associated lexicon can be a million or more words. Information retrieval often focuses on the so-called vector space model. Clearly, the vector space model can involve very high-dimensional vector spaces. Analysis using the vector space model is done by consideration of a term-document matrix. However, the term-document matrix basically is a bag-of-words approach capturing little semantic content. We have been exploring bigrams and trigrams as a method for capturing some semantic content and generalizing the term-document matrix to a bigram-document matrix. The cardinality of the set of bigrams is in general not as big as $\binom{n}{2}$; it is nonetheless usually considerably larger than n, where n is the number of words in the lexicon.

1. INTRODUCTION

Text mining capabilities have dramatically improved in recent years, but have been principally focused on the English language. Manning and Schütze (1999), Berry (2003), Feldman and Sanger (2007), Weiss et al. (2005), Solka (2008), and Rao, Wegman and Solka (2006) are recent discussions of text mining and related methodologies. Text mining as a general field can be thought of as a synthesis of information retrieval, natural language processing, and traditional data mining.

We would like to demonstrate some ideas on text mining. Text mining operations often deal with very high dimensional vector spaces and we would like to provide some sense of the dimensionality involved. We would like explain by placing text mining as we understand it in the framework of related disciplines: information retrieval, natural language processing, and data mining. Then the focus of this paper is on a tool we have been using for several years, namely bigrams and related vector space ideas such as the term-document and bigram-document matrices. Finally, we would like to illustrate the dimensionality of the application to text mining with a corpus of nearly 15,863 documents.

The focus of information retrieval can generally be described as searching for documents. But, with somewhat more precision, it is fair to say that information retrieval also includes

searching for information within documents and searching for metadata that describe documents. The search can be in standalone relational databases or hypertext-related databases such as the World Wide Web. Common search engines function with either a set-theoretic Boolean model of the document corpus, or a vector space model, or within a probabilistic Bayesian framework. Although usually information retrieval deals with text documents, it can also deal with imagery, video, audio, and other multimedia types.

1.1 Natural Language Processing

Natural languages have four key elements: morphology, syntax, semantics, and lexicon. Morphology refers to the grammar of word forms, for example, how nouns are made plural, or masculine or feminine, how verbs are conjugated and so on. Syntax refers to the grammar of word combinations, for example, where adjectives are placed relative to the nouns they modify, where verbs are placed in the sentence, how a sentence is modified to become interrogative or imperative, and so on. Semantics refers to the meaning of a word or a sentence, and, of course, lexicon refers to the set of words used in a language. Generally, natural language processing is a difficult discipline because natural languages have many ambiguities. For example, the sentence "Time flies like an arrow." could be interpreted as "Time passes speedily like an arrow passes speedily." or it could be interpreted to mean "Measure the speed of a fly like you would measure the speed of an arrow." In the first interpretation, "time" is a noun and "flies" is a verb. In the second, "time" is a verb and "flies" is a noun. There is ambiguity in the sense that these two words could be either a noun or a verb depending on interpretation. There is an obvious ambiguity of meaning as well.

1.2 Text Mining Tasks

Within text mining, we can identify six major classes of tasks. Text classification focuses on assigning a new document to one of several predetermined classes. This is also known as supervised learning in the machine learning literature. Text clustering focuses on determining natural clusters of documents within a corpus and is also known as unsupervised learning. Text summarization is extracting a summary automatically from a document and uses elements of syntax and semantics. This is, by the way, an area of considerable interest to the military for discovering relevant documents in open source literature. Author identification focuses on determining the author of a document where the author is either unknown or authorship is disputed. This task not only depends on the syntax and semantics found in the document, but also the characterizations of style of the document. Perhaps the most difficult task is automatic translation, which includes the morphology, the syntax, the semantics, and the lexicon of two languages. One reason translation is so difficult is that idiomatic expressions are hard to recognize and a literal translation may not make sense. Using multi-grams or strings of words to recognize idioms is a fruitful way to augment traditional translation techniques and one reason we have focused on these structures. Cross corpus discovery refers to comparison of documents for two or more corpora usually with

the idea of finding similar or related documents.

2. TEXT MINING BASICS

2.1 Preprocessing

Before launching on processing documents within a corpus, it is usually advisable to do some preprocessing. These preprocessing steps usually reduce the size of the lexicon while preserving the semantic content of the documents. Two tasks are usually undertaken: denoising and stemming. Denoising usually refers to removal of stopper words that have little semantic content. Words such as the, an, and, of, by, that, and other articles, conjunctions, or prepositions are likely candidates for stopper words. These are often just taken from a predetermined list. However, the stopper words may be corpus dependent. For example, in a corpus consisting of mathematical documents, words like "theorem" and "proof" may be treated as stopper words. A way of automatically determining stopper words is to calculate the so-called Term Frequency Log Inverse Document Frequency. This measure down weights a word if it occurs infrequently in all of the documents of the corpus and also down weights a word that occurs in almost every document of the corpus. By thresholding on the so-called TFIDF measure, one can automatically remove stopper words that are corpus dependent. Stemming is the other preprocessing step. This procedure removes suffixes, prefixes and infixes and is an attempt to replace words with their root. The example here replaces wake, waking, awake, woke with wake. Of course, there are perils with automated procedures as well. For example, browse, browsing, browsed, could conceivably be replaced with brows, so that a leisurely afternoon in a bookstore could become facial hair.

2.2 Bigrams and Trigrams

We like to use bigrams and trigrams and more generally multi-grams rather than just single words. The reason is that multi-grams in general capture word combinations that involve syntactic and semantic structures that are not captured by single words. A bigram is a word pair where the order of the words is preserved. The first word is called the reference word and the second is the neighbor word. Interestingly, we adapted this language from image processing applications where pixel pairs are characterized as reference pixels and neighbor pixels. A trigram is analogously a word triple where the order is preserved.

Consider the sentence, "Hell hath no fury like a woman scorned." The "a" is a stopper word. It is also possible that we might consider "no" a stopper word, although in this case we will not. The denoised version is "Hell hath no fury like woman scorned." Of course "hath" is an archaic form of "has" and "scorned" stems to "scorn." It is also conceivable that "woman" could be stemmed to "man." If this is done, the meaning would change and so we would not want to stem this way. The stemmed and denoised version is "Hell has no fury like woman scorn." The bigrams of the stemmed and denoised sentence are "hell has," "has no," "no fury," "fury like," "like woman," "woman scorn," and "scorn ." Note also that

Table 1:

Sentence	Hell hath no fury like a woman scorned.
Denoised Sentence	Hell hath no fury like woman scorned.
Stemmed Sentence	Hell has no fury like woman scorn.
Bigrams	Hell has, has no, no fury, fury like, like woman, woman scorn, scorn .
Note	The "." and any other sentence ending punctuation is treated as a word.

"." "?" "!" ";" and other sentence ending punctuations are stemmed to "." and treated like a word in the bigram computation.

2.3 **Bigram Proximity Matrix**

The bigram proximity matrix of a document is a mathematical object representing the document. This representation has some claim to capturing the semantic content of the document because it captures noun-verb, adjective-noun, verb-adverb, and verb-object pairs as well as other word pairs that have semantic as well as syntactic meaning. Because the Bigram Proximity Matrices are mathematical objects in a vector space we can create metrics on them and thus measure similarity between bigram proximity matrices and infer how similar their corresponding documents are.

This idea was explored by Dr. Angel Martinez in his 2002 dissertation, Martinez (2002), and several subsequent papers, Martinez and Wegman (2002, 2003), Martinez et al. (2004, 2008). In a way similar to the Bigram Proximity Matrix, trigrams can be arranged in an analogous three-dimensional structure and distances also measured. If the cardinality of the stemmed and denoised lexicon is n, the number of bigrams is in general much less than $\binom{n}{2}$, but is usually considerably larger than n. The bigram proximity matrix is usually very sparse.

Table 2:

	.	fury	has	hell	like	no	scorn	woman
.								
fury					1			
has						1		
hell			1					
like								1
no		1						
scorn	1							
woman							1	

We can construct a bigram proximity matrix, usually abbreviated BPM. The rows correspond to the reference word and the columns correspond to the neighbor word. The rows and columns are usually arranged alphabetically according to the lexicon of the whole corpus and an entry is made corresponding to each bigram. The entries may be binary, one if the bigram appears and zero otherwise or they may be frequency counts of the bigram. Of course, we would compute the matrix for a whole document, not just a sentence. In the example given in Table 2, for simplicity, the zeros are omitted. The matrix is normally square with dimension in rows and columns equal to the cardinality of the stemmed and denoised lexicon. In the case of some document sets, the stemmed and denoised lexicon could be as large as 100,000 terms and the number of bigrams could easily approach 1,000,000 or more.

2.4 **Similarity and Distance Measure Complexities**

In order to define Similarity Measures and Pseudometrics on the BPM following Martinez (2002), we use the Ochiai-Cosine measure in the case of the BPM:

$$S(X,Y) = \frac{|X \text{ and } Y|}{\sqrt{(|X||Y|}}.$$

We consider here only binary matrices. Here X is the BPM for Article X and Y is the BPM for Article Y. $|X \,\text{and}\, Y|$ is the number of 1's that X and Y have in common and, of course, $|X|$ is the number of 1's in matrix X and $|Y|$ is the number of 1's in matrix Y. Clearly if $X = Y$, then $S(X,Y) = 1$ and if X and Y have no common bigrams, then $S(X,Y) = 0$. $S(X,Y)$ is the binary equivalent of the cosine similarity measure and is essentially a normalized version of the dot product of the angle between vectorized X and vectorized Y.

Consider the Euclidean distance between normalized X and Y.

$$\|X - Y\|_2 = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (1)$$

$$= \sqrt{\sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n} X_i Y_i + \sum_{i=1}^{n} Y_i^2} \quad (2)$$

$$= \sqrt{[2 - 2\sum_{i=1}^{n} X_i Y_i]}. \quad (3)$$

Because $\sum_{i=1}^{n} X_i^2$ and $\sum_{i=1}^{n} Y_i^2$ are normalized and binary, they both sum to 1. Thus, the Euclidean distance between normalized vectors X and Y is inversely related to the Ochiai-Cosine similarity measure. This motivates us to use $S(X, Y)$ to form a metric by:

$$d(X, Y) = \sqrt{[2 - 2S(X, Y)]}.$$

Let x be the set of word pairs or triplets in Article X. Let y be the set of word pairs or triplets in Article Y. Then Article X and Article Y can be described by the intersection of the sets x and y. The sets x and y are represented as hash tables where the key, word pair or triplet, maps to the number of occurrences in the article. The intersection of x and y can then be computed by the number of keys in x that are also in y. The contains Key function being used is close to $O(1)$ so the computation of X and Y should be close to $O(\text{size of } x)$ or for all keys in x check if y contains key. The value of $|\text{Article } X||\text{Article } Y|$ is $|\text{size of } x||\text{size of } y|$.

2.5 Interpoint Distance Complexity Issues

Let n be the number of documents in the corpus. The interpoint distance matrix involves $\frac{n(n-1)}{2}$ comparisons, which results in $O(n^2)$ operations. It will pay to make each of these "comparisons" as efficient as possible.

3. VECTOR SPACE METHODS

The classic structure in vector space text mining methods is a term-document matrix, where the rows correspond to terms and columns correspond to documents. The entries may be binary or frequency counts. A simple and obvious generalization is a bigram-document matrix where rows correspond to bigrams, columns to documents, and again entries are either binary or frequency counts. As with the bigram proximity matrix, these matrices are usually very sparse.

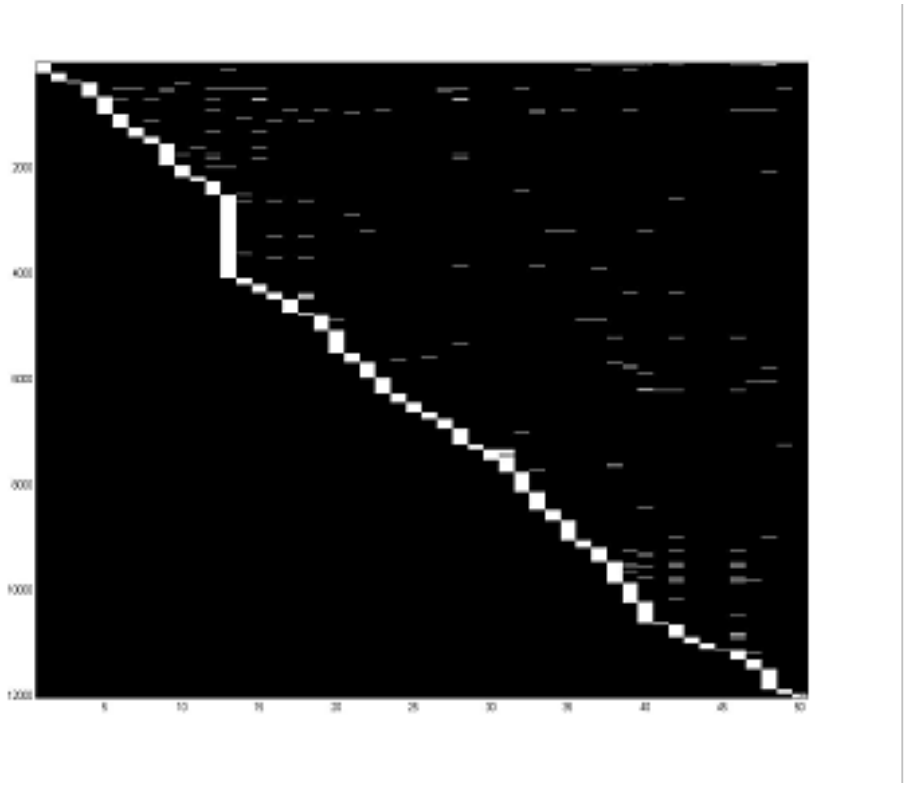


Figure 1: Bigram-Document Matrix for the First 50 documents.

The example that we report here is based on the full set of 15,863 documents. In order to illustrate the dimension and scale of the vector space methods, we use the text data that were collected by the Linguistic Data Consortium in 1997. These data were originally used in Martinez (2002). The data consisted of 15,863 news reports collected from Reuters and CNN from July 1, 1994 to June 30, 1995. The full lexicon for the text database included 68,354 distinct words. In all 313 stopper words are removed and after denoising and stemming, there remain 45,021 words in the lexicon.

This document corpus has 1,834,123 bigrams. Thus, the Term-Document Matrix, TDM is 45,021 by 15,863 and the Bigram-Document Matrix, BDM is 1,834,123 by 15,863. The term vector is 45,021 dimensional and the bigram vector is 1,834,123 dimensional. The BPM for each document is 1,834,123 by 1,834,123 and, of course, very sparse. A corpus can easily reach 20,000 documents or more and scaling is a significant issue in text processing.. The 15,863 document database proves to be challenging.

3.1 **Bigram-Document Matrices and Bigram-Bigram Matrices**

Term-document and bigram-document matrices resemble the so-called two-mode social network adjacency matrices. This idea is explored in Wegman and Said (2009). Although the

resemblance is interesting, the scale of the matrices involved in text mining is dramatically higher in dimension than almost any social network analysis would entail. Figure 1 illustrates the bigram-document matrix for just 50 documents. There are actually 1,834,123 bigrams, but this image was truncated at 12,000 bigrams. In this image, one is coded as white and zero is coded as black. As new documents are introduced new bigrams are also introduced. The vertical white bars represent new bigrams being introduced as a new document appears. The light horizontal bars in the upper-right part of the matrix represent bigrams that had been introduced from earlier articles and are being reused by later articles.

Using techniques from social network analysis (c.f. Wegman and Said, 2009), we can convert the bigram-document matrix into a bigram-bigram matrix. Figure 2 illustrates the Bigram-Bigram matrix clustered by the allegiance methodology for the 50 documents. This image represents only the first 9,000 bigrams of the 1,834,123 bigrams. Again, one is coded as white and zero is coded as black. The allegiance methodology is a way of clustering in social network analysis and was described in Said et al. (2008).

Obviously illustrating a 1,834,123×1,834,123 matrix is not feasible. However, by selecting the most frequently occurring bigrams, we are able to see some of the structure in the bigram-bigram matrix. The bigram-bigram matrix illustrated in Figure 3 is derived from the bigram-Document matrix and it shows a satisfying block structure. For example, the large central block contains the following bigrams: heir-apparent, general-luck, general-Shalikashvili, base-general, luck-recommend, first-summer, police-station, full-term, past-think, and courts-reflect. This cluster of words appears to be related to the presidential actions in 1994 and 1995.

3.2 **Text Clusters**

We would like to provide an example of how bigrams and trigrams can capture semantic content in a completely automated way. Based on this corpus, we used a text-based agglomerative clustering software called CLUTO. This clustering method uses a recursive splitting algorithm. In this example, we hypothesized 25 clusters. A portion of the hierarchical agglomerative tree for the clusters is given in Figure 4.

A slightly more elegant visualization is shown in Figure 5 using a freely available shareware called SPACETREE. One can download both CLUTO and SPACETREE from the web. Googling these names will quickly lead you to these two software products.

3.3 **Text Cluster Example**

Cluster 0, Size: 157, ISim: 0.142, ESim: 0.008

Descriptive: ireland 12.2%, ira 9.1%, northern.ireland 7.6%, irish 5.5%, fein 5.0%, sinn 5.0%, sinn.fein 5.0%, northern 3.2%, british 3.2%, adam 2.4%

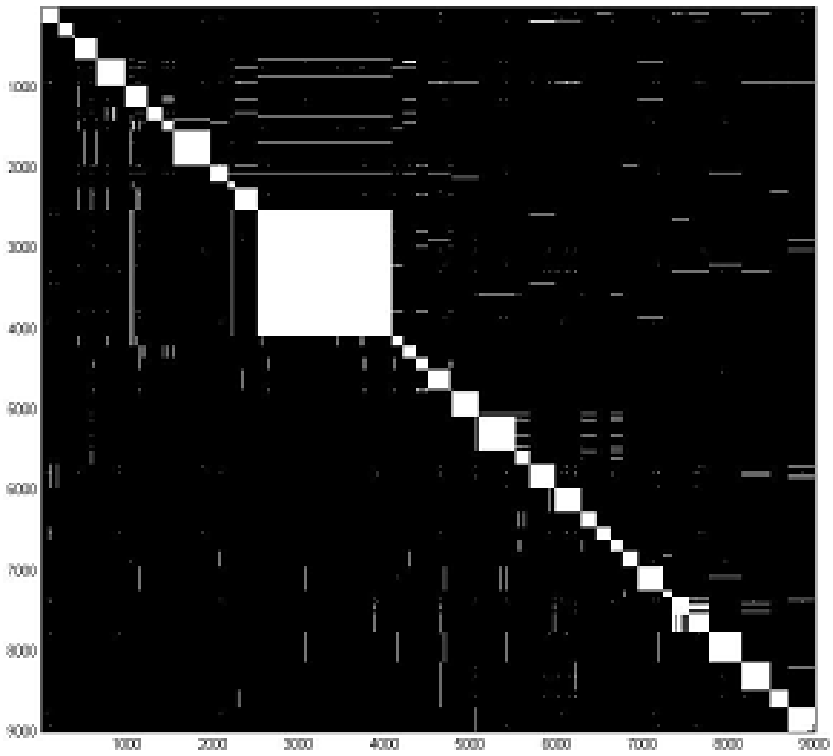


Figure 2: Bigram-Bigram Matrix for the Same 50 Documents illustrated in Figure 1.

Discriminating: ireland 7.7%, ira 5.9%, northern.ireland 4.9%, irish 3.5%, fein 3.2%, sinn 3.2%, sinn.fein 3.2%, northern 1.6%, british 1.5%, adam 1.5%

Terms: ireland 121, northern 119, british 116, irish 111, ira 110, peac 107, minist 104, govern 104, polit 104, talk 102

Bigrams: northern.ireland 115, sinn.fein 95, irish.republican 94, republican.armi 91, ceas.fire 87, polit.wing 76, prime.minist 71, peac.process 66, gerri.adam 59, british.govern 50

Trigrams: irish.republican.armi 91, prime.minist.john 47, minist.john.major 43, ira.ceas.fire 35, ira.polit.wing 34, british.prime.minist 34, sinn.fein.leader 30, rule.northern.ireland 27, british.rule.northern 27, declar.ceas.fire 26

Cluster 0 is the first cluster developed by the CLUTO software. There are 157 documents on Cluster 0. ISim is a measure of internal similarity and essentially measures the internal coherency of the cluster. ESim is a measure of the similarity of the Cluster 0 to the remaining documents. Of course we want ESim to be low and ISim to be relatively high, in this case, the ratio of ISim to Esim is 17.75. Phrases 2 and 3 are respectively the bigrams and trigrams.

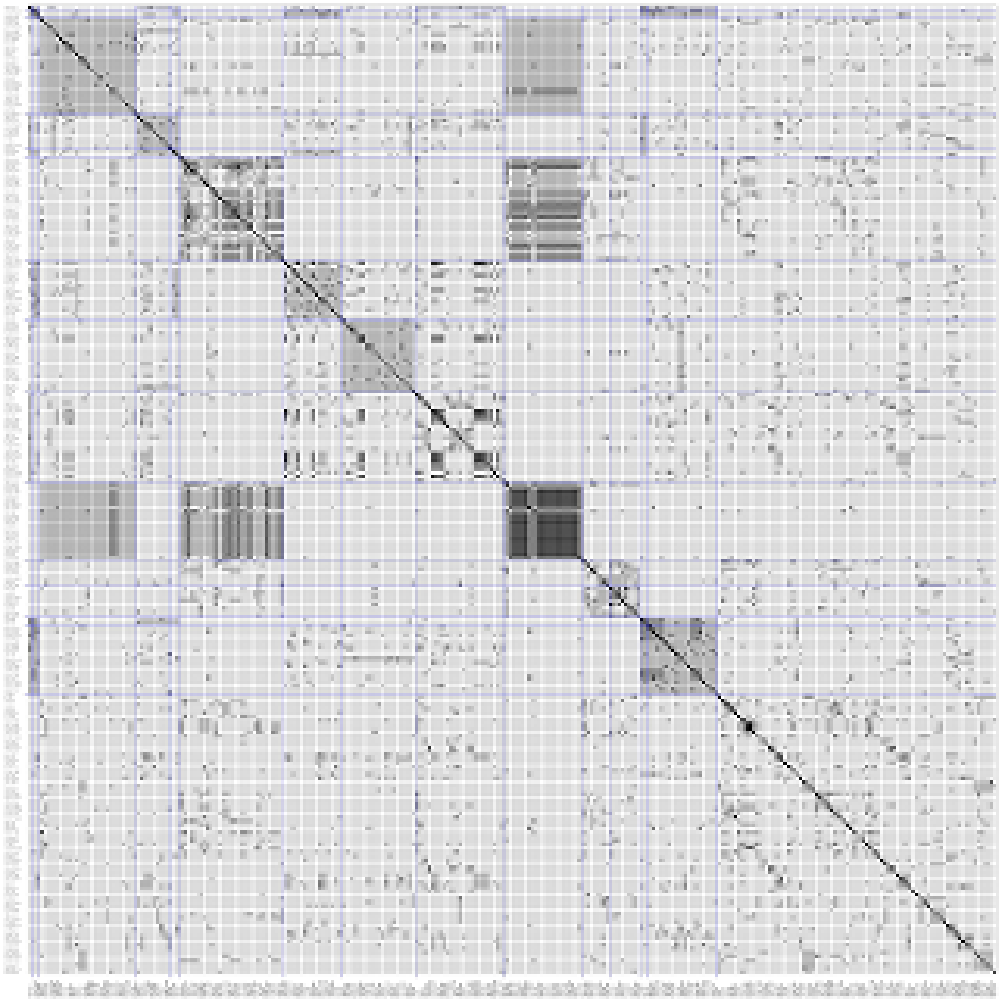


Figure 3: Bigram-Bigram Matrix Using the 253 Most Frequently Occurring Bigrams.

A quick perusal of these bigrams and trigrams immediately tells us that this cluster focuses on the sectarian conflict in Northern Ireland that was prominent in the middle 1990s. Please note that this clustering was done completely automatically with no human input.

Cluster 1, Size: 323, ISim: 0.128, ESim: 0.008

Descriptive: korea 19.8%, north 13.2%, korean 11.2%, north.korea 10.8%, kim 5.8%, north.korean 3.7%, nuclear 3.5%, pyongyang 2.0%, south 1.9%, south.korea 1.5%

Discriminating: korea 12.7%, north 7.4%, korean 7.2%, north.korea 7.0%, kim 3.8%, north.korean 2.4%, nuclear 1.7%, pyongyang 1.3%, south.korea 1.0%, simpson 0.8%

Terms: korea 305, north 303, korean 285, south 243, unit 215, nuclear 204, offici 196, pyongyang 179, presid 167, talk 165

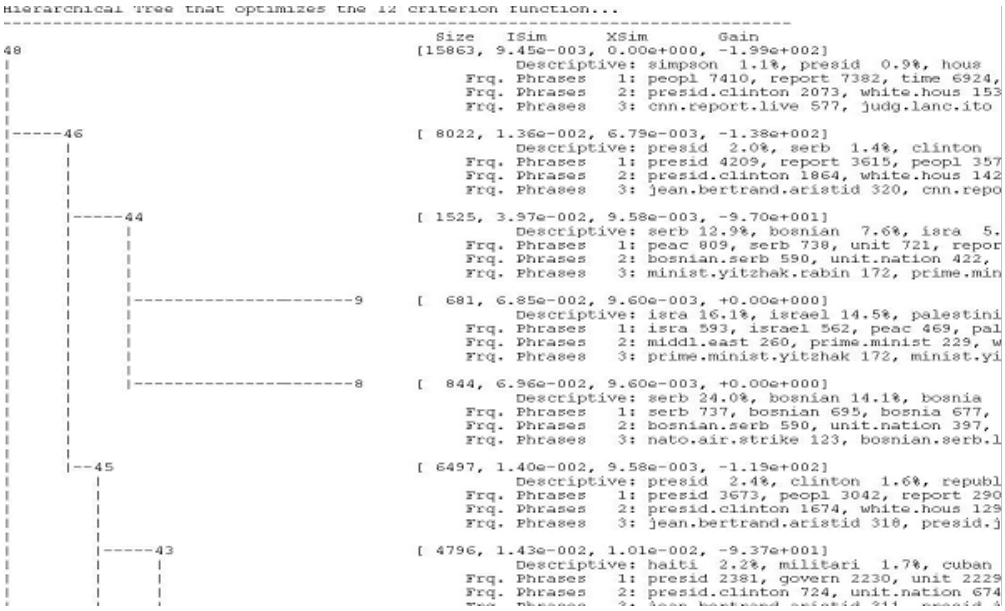


Figure 4: Snapshot of the recursive partitioning tree.

Bigrams: north.korea 291, north.korean 233, south.korea 204, south.korean 147, kim.sung 108, presid.kim 83, nuclear.program 79, kim.jong 74, light.water 71, presid.clinton 69

Trigrams: light.water.reactor 56, unit.north.korea 55, north.korea.nuclear 53, chief.warrant.offic 49, presid.kim.sung 46, leader.kim.sung 39, presid.kim.sam 37, north.korean.offici 36, warrant.offic.bobbi 35, bobbi.wayn.hall 29

Cluster 1 has 323 documents in it. Again there is a substantial difference between ISim and ESim. The ratio of ISim to ESim in this case is 16.00 meaning that this is a very coherent cluster as well. Again considering the bigrams and trigrams quickly convinces one that the topic of this cluster is related to the nuclear ambitions of North Korea, still a continuing topic of interest. Note that one of discriminating terms is "simpson." The 1994 to 1995 time frame was the time when O.J. Simpson was on trial for the murder of his wife and her friend. Because the coverage of this event was so extensive, "simpson" actually appears in many clusters as a single term, but does not usually make it into the list of bigrams and trigrams.

Cluster 24, Size: 1788, ISim: 0.012, ESim: 0.007

Descriptive: school 2.2%, film 1.3%, children 1.2%, student 1.0%, percent 0.8%, compani 0.7%, kid 0.7%, peopl 0.7%, movi 0.7%, music 0.6%

Discriminating: school 2.3%, simpson 1.8%, film 1.7%, student 1.1%, presid 1.0%, serb 0.9%, children 0.8%, clinton 0.8%, movi 0.8%, music 0.8%

Terms: cnn 1034, peopl 920, time 893, report 807, don 680, dai 650, look 630, call 588, live

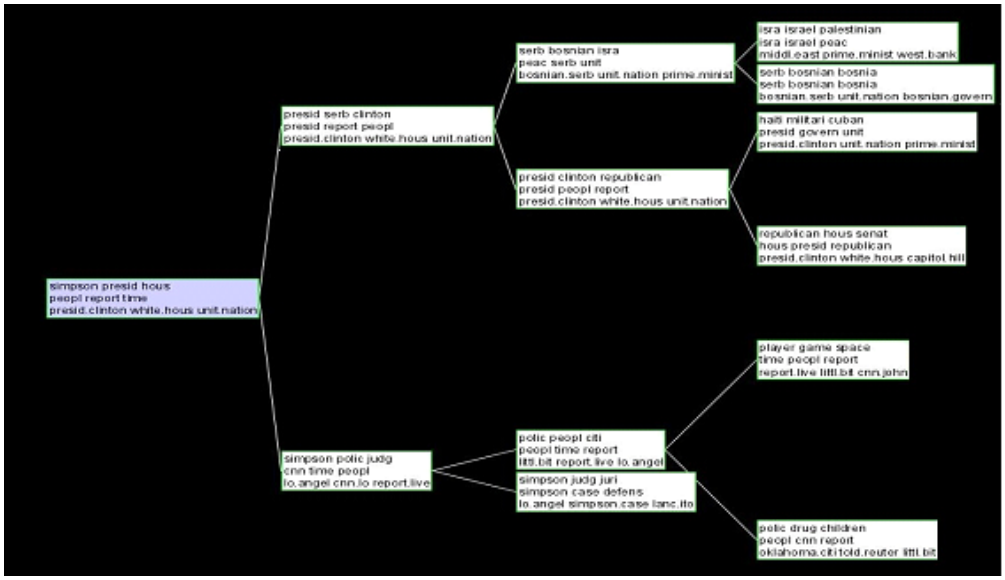


Figure 5: Snapshot of the recursive tree using SPACETREE.

535, lot 498

Bigrams: littl.bit 99, lot.peopl 90, lo.angel 85, world.war 71, thank.join 67, million.dollar 60, 000.peopl 54, york.citi 50, garsten.cnn 48, san.francisco 47

Trigrams: jeann.moo.cnn 41, cnn.entertain.new 36, cnn.jeann.moo 32, norma.quarl.cnn 30, cnn.norma.quarl 28, cnn.jeff.flock 28, jeff.flock.cnn 27, brian.cabel.cnn 26, pope.john.paul 25, lisa.price.cnn 25

This is the last and 25th cluster of our hypothesized 25 clusters. Note that the ISim and ESim are very close, the ratio ISim to ESim being only about 1.7, suggesting that this cluster is not very coherent. Considering the bigrams and trigrams in this case confirms that the cluster is not very coherent. Our guess that there are 25 clusters is probably much too small. It is hard to see how Los Angeles, San Francisco, World War, and Pope John Paul easily fit into a cluster along with the many references to CNN news people.

Just to illustrate the capability of the methodology, Figure 6 illustrates two articles from cluster one. The words under "intersects" are the bigrams these two documents share. One interesting feature that can be discovered from this methodology is to realize that news stories often evolve over a few day period and that as this happens, the original text is reused and augmented by additional material.

4. CLOSING REMARKS

To recap, Henry James said, "To read between the lines is easier than to follow the text." Text mining presents great challenges, but is amenable to statistical and mathematical ap-

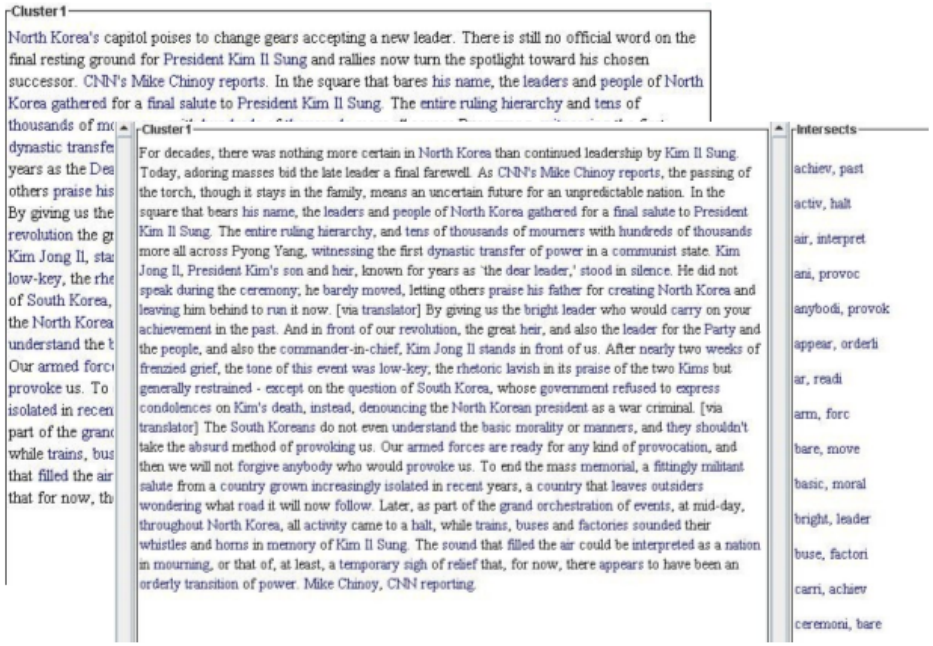


Figure 6: Two articles from Cluster 1.

proaches. Text mining using vector space methods challenges both mathematics and visualization especially in terms of dimensionality and sparsity. Our use of term-term, bigram-bigram, and document-document one-mode networks is just beginning and needs further exploration. Finally, Winston Churchill said that "The length of this document defends it well against the risk of its being read." We hope that this is not the case with this article.

ACKNOWLEDGEMENTS

We would like to acknowledge the contributions of several former students: Dr. Walid Sharabati who provided the examples for the bigram-document and bigram-bigram documents; Dr. Faleh Alshameri who provided the document clustering material; Dr. Jeffrey Solka who provided the CLUTO and SPACETREE illustrations; and Dr. Angel Martinez who provided the data for all of these examples. We are in debt to the support provided by the Isaac Newton Institute for Mathematical Science at the University of Cambridge in Cambridge, England, which has created the opportunity to formulate this work. We have filed a patent disclosure on certain aspects of this work.

REFERENCES

Berry, Michael W. (ed) (2004) *Survey of Text Mining: Clustering Classification and Retrieval*, New York: Springer.

CLUTO – Software for clustering high dimensional datasets, URL:http://glaros.dtc.umn.edu /gkhome/cluto/cluto/download.

Feldman, Ronen and Sanger, James (2007) *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge UK: Cambridge University Press.

Manning, Christopher D. and Schütze, Hinrich (1999) *Foundations of Statistical Natural Language Processing*, Cambridge MA: The MIT Press.

Martinez, Angel (2002) *A Framework for the Representation of Semantics*, Ph.D. Dissertation, School of Computational Science, George Mason University.

Martinez, A. and Wegman, E (2002) "A text stream transformation for semantic-based clustering," *Computing Science and Statistics*, 34, 184–203.

Martinez, A. and Wegman, E. (2003) "Encoding of Text to Preserve 'Meaning'," *Proceedings of the Eighth U.S. Army Conference on Applied Statistics*. ACAS02/MartinezAngel/MartinezAngel.pdf.

Martinez, A., Martinez, W., and Wegman, E.J. (2004) "Using weights with a text proximity matrix," in *COMPSTAT 2004*, (Antoch, J., ed.), Berlin: Physica-Verlag, 327–338.

Martinez, W., Martinez, A. and Wegman, E. (2008) "Classification and clustering using weighted text proximity matrices," *Computing Science and Statistics*, 36, 600–611.

Rao, C. R., Wegman, E. J., and Solka, J. L. (2005) *Handbook of Statistics: Data Mining and Data Visualization, Volume 24*, Amsterdam: Elsevier/North Holland.

Said, Y. H., Wegman, E. J., Sharabati, W. K., and Rigsby, J. T. (2008) "Style of author-coauthor social networks," *Computational Statistics and Data Analysis*, 52, 2177–2184; doi:10.1016/j.csda.2007.07.021, 2007.

Solka, J. L. (2008) "Text data mining: Theory and methods," *Statistics Surveys*, 2, 94–112.

SpaceTree: a novel node-link tree browser, URL: http://www.cs.umd.edu/hcil/spacetree/.

Wegman, Edward J. and Said, Yasmin H. (2010) "Text mining and social network analysis: Some unexpected connections," to appear in *Proceedings of HDM 2008*

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 977–988.

MEASUREMENT OF SCHOOLING DEVELOPMENT

Mamadou-Youry Sall
Professor, Unit of Formation and Research in
Economic Sciences and Management at
Gaston Berger University, Saint-Louis, Senegal, BP 234
E-mail: sallmy@ufr-seg.org

ABSTRACT

With the UNESCO's objective, Education for All (EFA) by 2015, it has become necessary, for the countries that have not yet reached the universal schooling, to conduct a continuous assessment to determine the level of educational development. This allows us to know the scale of the effort to be furnished for that purpose. In other words, it is an attempt to know how the schooling level evolves by generation in these countries. The Admission Gross Rate (AGR) and the Admission Net Rate (ANR) are still used as indicators to give an idea about the proportion of children in school per generation. However, these two indicators do not always reflect reality and can be misleading to policymakers or planners. These two indicators can be used when information can't be improved or for international comparability, but one must be aware that they are not statistically robust. When more complete data is available, better indicators can be found. In this article, we propose a less biased estimator of the number of children enrolled per generation, say the Generational Admission Rate (GAR). Considering the quality of the school data currently available in many countries in addition to the availability of computer software, it should not be difficult to calculate this indicator. In this paper we examine the construction of the indicator and its application using data from Senegal.

Keywords: Statistics; estimator; indicator; education; rate of education access

1. INTRODUCTION

In social science the indicators are often built empirically. That is, the quality and quantity of the data determines the construction of the indicators and their robustness. Because of insufficient information, one often proceeds by approximation in order to find the parameters of the theoretical distribution. This also holds for education. The results obtained in this way should be readjusted when one has more information. One cannot, for example, continue to use the gross rate of admission or schooling when the age distribution of pupils exists. It is now unacceptable to find in some scientific publications, a rate of schooling over hundred percent when we are sure that there are children not yet enrolled!

Such indicators, even if they are useful for an international comparability, are mathematically not very robust and might not correspond to local reality. That is, they would not meet the national needs for planning. In this paper, we propose another estimator to find the number of children registered at school per generation (children having the same age). Considering the quality of the educational data, currently available in a large number of countries, one can achieve this goal with more statistical robustness. We just need to find the statistical law, which

would generate these data in order to find their parameters. Once these are found, it will be easy to represent reality more accurately and to better plan educational policy.

With the UNESCO's objective, Education for All (EFA) by 2015, it has become necessary, for the countries that have not yet reached the universal schooling, to conduct a continuous assessment to determine the level of educational development. This allows us to know the scale of the effort to be furnished for that purpose. In other words, it is an attempt to know how the schooling level evolves by generation in these countries.

The Admission Gross Rate (AGR) and the Admission Net Rate (ANR), are still used as the indicators which give an idea about the proportion of the children in school per generation. However, these two indicators do not always reflect reality and can be misleading to policymakers or planners. The Admission Gross Rate which is a ratio between two incomparable populations on the basis of age, can be biased from the statistical point of view. Representing the number of all the children who are in the first year at elementary school, its use will overestimate the proportion of children admitted at school. Its value can be hundred percent even if the total population is not at school.

The Admission Net Rate, if the legislation on the minimum age for admission to the first school class was respected, would be a better estimate of the proportion of school children per generation. However, this is not the case in many countries concerned by UNESCO goal. The group of children can be scattered in different school classes. One can find them in the first, second, third or fourth year of study. Hence, if we limit ourselves to the pupils enrolled in the first school class with the required schooling age, we underestimate the genuine number. Hence, the value of this indicator will never reach hundred percent, even if the total population is at school.

These two indicators can be used when information can't be improved or for international comparability, but one must be aware that they are not statistically robust. When more complete data is available, better indicators can be found.

In this paper, we propose a less biased estimator of the number of children enrolled per generation, say the Generational Admission Rate (GAR). Considering the quality of the school data currently available in many countries in addition to the availability of computer software, it should not be difficult to calculate this indicator. Section 2 illustrates the construction of the indicator and Section 3 presents its application using data from Senegal.

2. CONSTRUCTION OF INDICATOR

To build an educational access indicator one must take into account four parameters: The level in which students are enrolled, the school year, the different ages of students and also the entry date into School, which gives formally

$$\boldsymbol{P_{t,\,k}} \tag{1}$$

i.e. the population of $\boldsymbol{k}$ years at time $\boldsymbol{t}$

$$\boldsymbol{P_{t,\,k,\,d}} \tag{2}$$

the number of schoolchildren among them at date $\boldsymbol{d}$, hence, the admission rate will be, by definition,

$$AR = \frac{\sum_{d}^{D} p_{t_o,k_o,d}}{p_{t_o,k_o}} \tag{3}$$

corresponding to the proportion of children of k_o years, during the school year t_0, found in the school at time D.

This includes all members of the generation getting into school at the normal age k_o, before this age or after. It is clear that, simple observations don't permit to get the accurate number. It will be necessary to organize a census after the enrolment of the last member in the group (generation), before we are able to count all the concerned children. That is not feasible as it might require many years which makes it impractical for policymakers. As a result of this difficulty, statistical estimators should be used as indicators.

To clarify the idea, let's consider now one cycle of school, with six levels of study and age of enrolment seven years. So, the Admission Gross Rate (AGR), which is the usual approximation of the admission rate defined above, is written as

$$AGR = \frac{\sum_{k=k_m}^{n} s_{1,t_o,k}}{p_{t_o,7}} \tag{4}$$

where, $s_{i,t,k}$ is the number of children of k years old, enrolled during the school year t, at level i, K_m is the youngest student's age and n the eldest one. We can rewrite the AGR as follows,

$$AGR = \frac{s_{1,t_o,k_n}}{p_{t_o,7}} + \frac{\sum_{\substack{k=k_m \\ k\neq k_n}}^{n} s_{1,t_o,k}}{p_{t_o,7}} \tag{5}$$

K_n, representing the normal age to access school. It is clear that, the value of the second term, corresponding to the number of students not belonging to the age group, can be of significant importance. Hence, this approximation is not the best one. Due to the fact that the legislation of the age of school access is not often respected in most African countries, the following relation is always true: $k_m < k_n \leq n$ knowing that empirically (see Figure 1) $5 \leq n \leq 15$.

The other estimator of the AR is the Admission Net Rate (ANR) noted as follows:

$$ANR = \frac{s_{1,t_0,7}}{P_{t_0,7}} \tag{6}$$

$$\Leftrightarrow ANR = AGR - \frac{\sum_{\substack{k=k_m \\ k\neq 7}}^{n} s_{1,t_0,k}}{P_{t_0,7}} \tag{7}$$

As we see, here the superfluous term is removed from the AGR. However, we do not take into account the members of the age group who came into school late, after the normal age. This means that, the value of this estimator is still less than the number to be found.

Taking into account these two flaws, we propose another estimator, **generational admission rate** (***GAR).*** This will be written as:

$$GAR = \frac{\sum_{i=1}^{6} s_{i,t_o,7}}{p_{t_o,7}} + \frac{\sum_{k>7}^{n} s_{1,t_o,k}}{p_{t_o,7}} \tag{8}$$

Here, we consider all members of the age group who are at school, at all levels (classes), as well as the pupils in the first school class who have exceeded the required age. The second part of this relation is the estimated number of the group of late comers, those who registered after the school year t_0. We try to take into account all those who are not registered at the normal access age. So we reduce the biases existing with the usual estimators. The difference between this indicator and the admission gross rate (***AGR***) is that part noted I_R.

$$GAR = AGR - I_R \tag{9}$$

where

$$I_R = \frac{\sum_{k<7} s_{1,t_o,k} - \sum_{i=2}^{6} s_{i,t_o,7}}{p_{t_o,7}} \tag{10}$$

This part enables us to see the evolution of early entries, the children admitted before the required age. $I_R > 0$, means that the number of children who have the preschool age continue to increase in primary school. In other words, there is a rejuvenation of the entire primary pupils. Thus

$$\mathbf{ANR < GAR < AGR} \tag{11}$$

$I_R < 0$, means the reverse of the above phenomenon, which gives

$$\boldsymbol{ANR < AGR < GAR} \tag{12}$$

Hence, this indicator will characterize the policy of school recruitment. It shows how one fills the deficit about the request for access at the legal age. This can be done by the recruitment of children who have not yet achieved the required age or children who have exceeded this age.

3. APPLICATION WITH DATA OF SENEGAL

3.1 Data Sources

The current information state, provided by the Department of Education, accessible through the Internet, makes the calculation of the proposed indicator possible. This possibility didn't exist a year ago, because of the non-availability of data on the school children age. The Ministry of Education in Senegal publishes an annual statistical yearbook. The accounts of school are crossed with those from the services having in charge censuses and surveys because the size of the population reaching the required age for school access is the base of all indicators constructed at this level.

Table 1: Estimation of the number of children aged seven years in October 2003.

Region of Senegal	Men	Rate of population growth	Women	Rate of population growth	Total
Dakar	305 81	2,53%	33 050	2,77%	63 632
Ziguinchor	9 380	2,55%	9 514	2,50%	18 894
Diourbel	22 262	5,20%	19 627	3,89%	41 889
St. louis	24 725	5,70%	23 687	5,15%	48 412
Tamba	9 940	3,20%	15 460	6,17%	25 400
Kaolack	19 607	2,47%	24 009	3,67%	43 616
Thiès	21 861	2,21%	26 117	3,17%	47 977
Louga	13 894	3,50%	14 466	3,30%	28 359
Fatick	12 333	0,97%	12 340	0,74%	24 672
Kolda	16 137	3,08%	17 634	3,51%	33 770
Sénégal	18 0719	3,14%	195 903	3,43%	376 622

The totals might not correspond to the sum of elements because of rounding

It should be noted that the documents issued by the Direction of Planning and Statistics from the Ministry of Economy and Finances, about the census of 1988 and the Senegalese Households Survey in 2001 (ESAM_II), do not allow one to know directly the size of studied group because the data are grouped per age. One has to look for this number by using the Sprague coefficients (See Appendix A: Tables A.2). These enables us to reconstruct the pyramid of the population at different ages. Table 1 is calculated from the sources listed above. We assumed, to realize this table, that the birth rate, infant mortality, as well as that of migration, has not decreased the growth rate of the target population. As a result, the number of children aged seven years, increases by 3.24% annually. This means that, they are 376.622 children in October 2003. On the basis of this number, one can calculate the different access rates to education.

3.2 Generational Admission Rate (GAR)

To build this indicator, we have taken into account all pupils of seven years, regardless of their scholar level (class), plus all those who have exceeded this age in the first year of elementary school (Class of Initiation = CI). It should be noted that, the recruitment of members of an age group at school can last for years. It can start as early as four years until the age of fourteen (See Figure I). There are late comers, who are older than 7 years in the first level (CI) and anticipants or preschool; those who entered before reaching the legal age, which is seven years in Senegal.

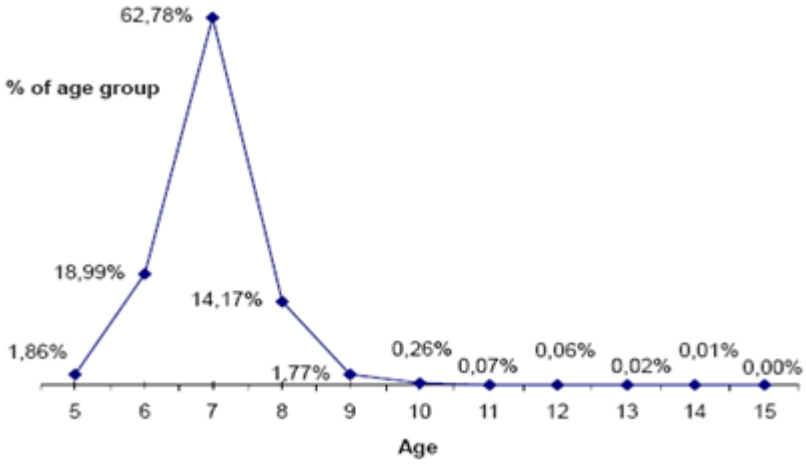


Figure 1. Pyramid of ages in the first year at elementary school (CI)

The number of elders in the first school class makes the estimation of the number of those who came into school late possible. Hence, we can say that the proportion of pupils among the children of seven years old is equal to 77.2% (Table 2) at October 2003. That is to say three children over four per generation. This means that, the school registration, from October 1997 to October 2003, exceeds by 2.13% that of the target population. This shows to what extent the gap between the growth of target population and the enrolled has now been reduced. If the system keeps this pace, the goal of education for all (EFA) will be achieved in twelve years, at 2016 (See Figure 2). To realize this goal in October 2015, the GAR must increase annually by 2.2% (Table 2). In other words, the mean growth of the number of schoolchildren aged 7 years must annually exceed by 2.2% the growth of population of the same age. That is to say, the enrollment in that group must grow by 5.61% annually.

Table 2: Admission Gross Rate, annual growth From 97 to 2003 and necessary rhythm to reach the EFA at 2015

Region of Senegal	Female Admission Rate	Annual Growth Rate	Global Admission Rate	Annual Growth Rate	Necessary Annual Rhythm for 2015	Necessary Registration Growth Rate for 2015
Ziguinchor	98,27%	2,89%	102,18%	1,83%	-0,2%	2,33%
Kolda	89,12%	7,05%	99,22%	5,47%	0,1%	3,58%
Fatick	91,66%	14,25%	94,26%	12,68%	0,5%	1,23%
Dakar	86,83%	-0,59%	92,00%	-0,47%	0,7%	3,46%
Thiès	76,09%	4,14%	85,02%	4,32%	1,4%	4,53%
Tamba	53,69%	-6,89%	68,78%	-4,87%	3,2%	9,34%
Kaolack	62,58%	6,32%	67,55%	5,82%	3,3%	6,99%
St-Louis	69,94%	-1,83%	64,88%	-3,75%	3,7%	8,82%
Louga	62,29%	5,81%	64,80%	3,28%	3,7%	6,99%
Diourbel	47,89%	6,37%	44,45%	2,52%	7,0%	10,88%
Matam						
Sénégal	73,12%	2,97%	77,20%	2,13%	2,2%	5,61%

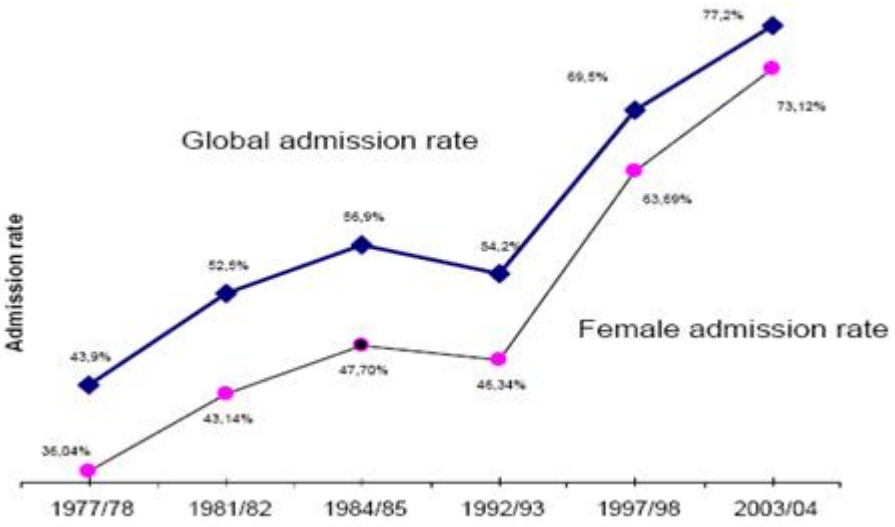


Figure 2. Evolution of admission rates in Senegal

The upper parts of the curve signify that the growth of the registered children is faster than that of the targeted population. It can be seen that Senegal maintains its recruitment policy. The number of children who have not yet achieved the school age among newcomers is constant. The GAR and AGR merge (see Figure 3).

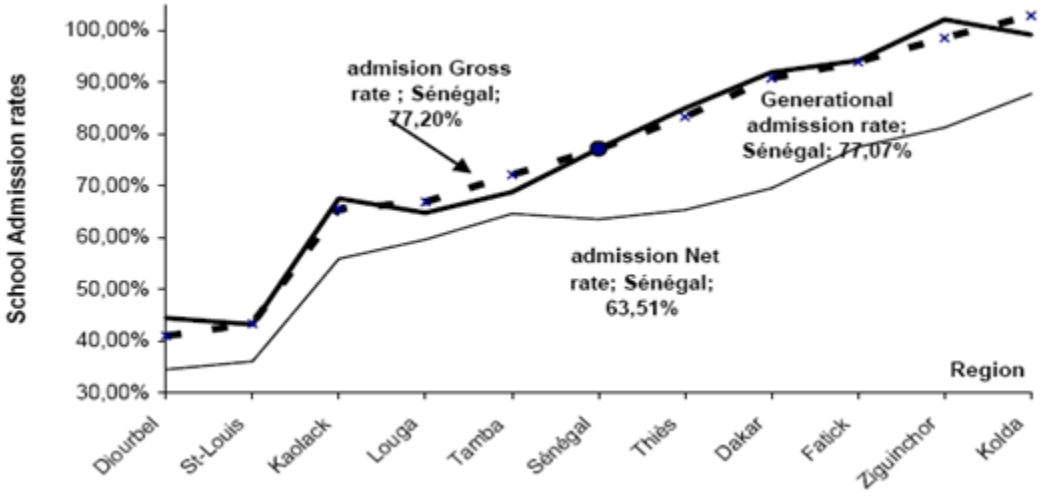


Figure 3. School Access Rates

In addition, one should also say that the admission rate varies from one region to another. It is geographically disparate, which obviously generates inequality. The analysis of Table 2, shows that the level of enrollment in the regions of Thies, Dakar, Fatick, Ziguinchor and Kolda, is higher than the national average. Children in these areas have, on average, 1.43 times more opportunities than the others. The overall disparity is 16.45%. (See Figure 4).

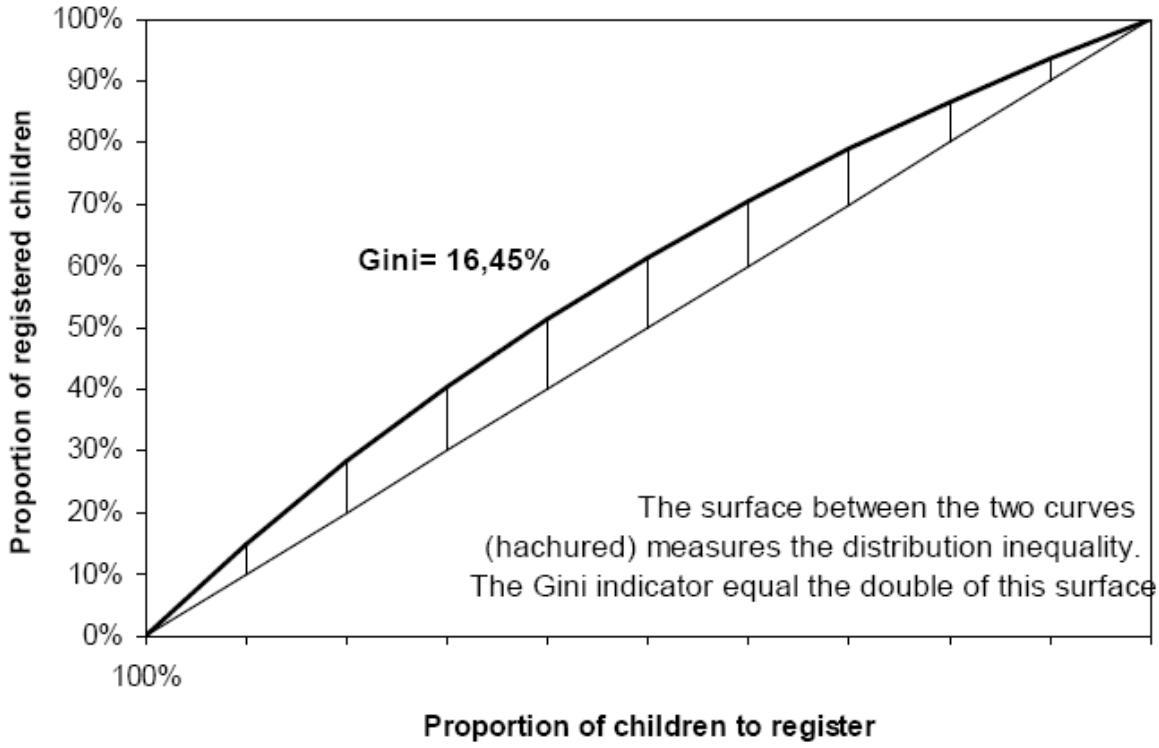


Figure 4. Regional inequality measure of admission in the first school class Year 2003

Usually, this curve is drawn below the first bisector curve. But, to get a more clear interpretation, one can draw this one above the bisector in order to show the concentration of the wealth compared to that of poverty, and not the opposite.

4. CONCLUSION

The usual indicators of school access, namely the gross rate or net rate of admission would give, in the absence of good information about school, an idea of educational development in a country. However, planning solely on this basis, one runs the risk of missing the target.
The generational admission rate provides a better approximation of the true measure of school access per generation, allowing us to know how the schooling grows in a society and to what extent the State respects its commitments to achieve education for all. In addition to this, it is sufficient to have the distribution of schoolchildren by age to be able to calculate this indicator and deduce other useful information such as the recruitment policy.

APPENDIX

A. Demographic Data

Table A.1: Second Senegalese survey about the households (ESAM_II) DPS, August 2001: Population of 7 years.

Region	Men	Growth rate 88-01	Women	Growth rate 88-01	Total	Growth rate 88-01
Dakar	29 030	2,53%	31 225	2,77%	60 255	2,65%
Ziguinchor	8 901	2,55%	9 036	2,50%	14 727	2,53%
Diourbel	20 034	5,20%	18 127	3,89%	38 161	4,55%
St. louis	22 032	5,70%	21 337	5,15%	43 368	5,42%
Tamba	9 309	3,20%	13 649	6,17%	2 958	4,69%
Kaolack	18 638	2,47%	22 274	3,67%	40 912	3,07%
Thiès	20 888	2,21%	24 476	3,17%	45 364	2,69%
louga	12 934	3,50%	13 519	3,30%	26 452	3,40%
Fatick	12 086	0,97%	12 152	0,74%	24 238	0,86%
kolda	15 148	3,08%	16 410	3,51%	31 558	3,30%
Senegal	169 000	3,14%	182 205	3,43%	347 993	3,29%

Table A.2 : Estimation, by the Sprague coefficients, of the seven year class age from the census of 30 may 1988

Region	Men	Growth rate	Women	Growth rate	Totale	Growth rate
Dakar	20974	3,67%	21900	3,90%	42874	3,78%
Ziguinchor	6417	2,77%	6551	3,55%	12968	3,16%
Diourbel	10369	4,25%	11033	5,36%	21401	4,81%
St-Louis	10721	1,00%	11109	1,76%	21829	1,38%
Tamba	6181	2,58%	6266	2,94%	12447	2,76%
Kaolack	13580	3,70%	13944	4,44%	27525	4,07%
Thiès	15722	3,59%	16323	4,03%	32044	3,81%
Louga	8273	1,23%	8859	2,92%	17131	2,07%
Fatick	10654	3,70%	11043	4,44%	21697	4,07%
Kolda	10206	2,77%	10473	3,55%	20679	3,16%
Senegal	113096	3,03%	117 501	3,74%	230597	3,38%

- The Sprague Coefficient has been used in estimating the number of children at the age of seven years from grouped data of a population ($-0{,}008*P_{0-4} + 0{,}216*P_{5-9} - 0{,}008*P_{10-14}$)
- To count the number of children of seven years between the end of a census and the school entry, we rectify the formula as follows: $P_7^k = P_7^n * (1+r)^{n-k-1} * (1+c*r)$ where ***r*** represents the growth between census, ***c*** the gap corrector between the two census dates and the size of 7 year class age at time ***k***

B. Scholar Data

Table A.3 : Estimation of the number of children aged 7 years at October 2003.
From the census of may 1988 and the Senegalese survey about the households (ESAM_II) in august 2001

Region	Male Population	Growth rate	Female Population	Growth rate	Total
Dakar	30581	2,53%	33050	2,77%	63 632
Ziguinchor	9380	2,55%	9514	2,50%	18 894
Diourbel	22262	5,20%	19627	3,89%	41 889
St. louis	24725	5,70%	23687	5,15%	48 412
Tamba	9940	3,20%	15460	6,17%	25 400
Kaolack	19607	2,47%	24009	3,67%	43 616
Thiès	21861	2,21%	26117	3,17%	47 977
louga	13894	3,50%	14466	3,30%	28 359
Fatick	12333	0,97%	12340	0,74%	24 672
kolda	16137	3,08%	17634	3,51%	33 770
Senegal	**180719**		**195903**		**376 622**

Table A.4: Number of scholar population aged 7 years and the children in the first year at elementary school (CI), October 2003.

Region	7 years in school	>7years in CI	% in the group	<7years in CI	% in the group
Dakar	44232	13587	17,93%	20348	26,85%
Diourbel	14439	2734	6,19%	2616	5,92%
Fatick	19133	4046	10,09%	5470	13,65%
Kaolack	24370	4235	8,33%	6258	12,31%
Kolda	29650	5093	15,25%	5325	15,94%
Louga	16903	2070	8,72%	3808	16,04%
Matam	9877	1935		2718	
St-Louis	17462	3546	20,30%	5625	30,89%
Tamba	16414	1912	10,11%	2359	12,47%
Thiès	31357	8632	15,16%	7144	12,55%
Ziguinchor	15353	3278	15,65%	3378	16,13%
Senegal	239 190	51 068	13,03%	65 049	16,60%

Table B.1 : Enrolled in the first year at elementary school (CI) in 2003/04

Enrolled in CI		Enrolled in Public CI in 2003/04		
Region	Enrolled	Global Enrollment	Repeaters	New scholars
Dakar	58540	40944	5106	35838
Diourbel	18618	16820	1735	15085
Fatick	23257	22750	2019	20731
Kaolack	29461	28335	2297	26038
Kolda	33508	33196	2742	30454
Louga	18377	16654	1489	15165
Matam	10506	10301	1167	9134
St Louis	20902	20498	2235	18263
Tamba	17470	16648	846	15802
Thiès	40792	38127	4342	33785
Ziguinchor	19305	18405	2089	16316
Senegal	290 736	262 678	26 067	236 611

Table B.2: Admission Gross Rate from 77 to 2003

Year	Enrollment in CI	Enrollment in CI female	Population	Female Population	Admission Gross Rate	Admission Gross Rate of women
1977/78	71 151	28 681	161969	79591	43,9%	36,04%
1981/82	94 659	38 689	180 398	89673	52,5%	43,14%
1984/85	113 369	47 757	199349	100118	56,9%	47,70%
1992/93	144 805	63 996	267226	138099	54,2%	46,34%
1997/98	215 979	102 268	310981	160561	69,5%	63,69%
2003/04	290 736	143 247	376 622	195903	77,2%	73,12%

REFERENCES

France (1998). *Note d'information.* N° 98.32 octobre; Ministère de l'éducation nationale, de la recherche et de la technologie.

Jean Emile Charlier (2004). Les écoles au Sénégal, de l'enseignement officiel au Daara, les modèles et leurs répliques. *Cahiers de la recherche sur l'éducation et les savoirs* n°3/2004, p35-53.

Mingat A., Rakotomolala R. and Ton J.-P. (2001). *Rapport d'État d'un système éducatif national (RESEN). Guide méthodologique pour sa préparation*. Washington, banque Mondiale, Équipe DH-PPTE, Afrique.

Sall M.-Y. (2003). *Evaluating* the cost of wastage rates: The case of the Gaston Berger University of Senegal. *Higher Education Policy* ; 2003, 16 (333-349), Unesco, Paris.

Sall M.-Y (1997). *Mesure de l'inégalité dans l'éducation : Le cas du Sénégal*. Atelier National de reproduction de thèses (ANRT.). Université lille3, Lille 1999, Université Mendès- France. Grenoble.

Senegal (1992/93 et 2003/04). *Statistiques de l'enseignement primaire*. Direction de la planification et de la reforme de l'éducation, Ministère de l'éducation.

Senegal (1993). *Présentation des résultats préliminaires du recensement de la population de Mai 1988*. Direction de la prévision et de la statistique, Ministère de l'économie, des finances et du plan.

Ta Ngoc Châu (1969). *Les aspects démographiques de la planification de l'enseignement.* Principes de la planification de l'éducation (Collection). UNESCO ; IIPE. 1969.

Tore Thonstad (1983). *Analyse et projection des effectifs scolaires dans les pays en développement : Manuel de méthodologie*. Unesco. Paris.

Unesco *(*1974). *Le Modèle de simulation de l'Unesco pour l'éducation.* Rapport et documents de sciences sociales, n° 29, Division des méthodes et de l'analyse, Département des sciences sociales, Paris.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 989–997.

REEMPHASIZING THE ROLE OF AGRULTURAL SECTOR IN MALAYSIA: ANALYSIS USING INPUT-OUTPUT APPROACH

Mohd Sahar Sauian[1] and Raja Halipah Raja Ahmad[2]
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Malaysia.
E-mail: [1]mshahar@tmsk.uitm.edu.my; [2]Raja.halipah@yahoo.com

ABSTRACT

In Malaysia, the manufacturing and the service sectors have higher contributions to the gross domestic products (GDP) compared to the agricultural sector. However, the agricultural sector still plays an important role in the Malaysian economy. This sector has been re-emphasized through the inclusion of biotechnology and modernization of the agro-based industries. Using the input-output approach of analysis, we could determine the impact of this sector using the inter-industry linkages effects. Analysis of the recent input-output table, reveals that the Malaysian agricultural sector has an above average linkage effect on the other sectors. This means that the other sectors are still dependent on the agricultural sector for the value-added expansion in the economy. With the rapid development in the economy, Malaysian agricultural sector could potentially have a large impact on the rest of the economy, either as a supplier or a purchaser, especially in the production and processing of agro-based products and food industries.

Keywords: contribution to GDP, inter-industry linkages, input-output analysis, value added generation, agro-based products

1. INTRODUCTION

In Malaysia, the agricultural sector was the main contributor to the national economy for the first three decades since independence (1957). It was the foundation and the driving force behind the economic growth of the economy (Chua, 2000). However, in the late eighties and the nineties, the manufacturing sector has surpassed the agricultural and the service sectors as the main contributor to the gross domestic product of Malaysia. This was due to the impact of "import substitution" as well the "export promotion" policy embarked on during the eighties (Wong, 1985).

The experience of the 1998 financial crisis in Malaysia has changed the mindset of its citizen in promoting back this sector. During the time of the crises, the recovery in agricultural output, underpinned by the significant improvement in palm oil yield, contributed towards the positive growth in the economy (Bank Negara Malaysia, 2000). Today, although this sector is the third highest sector after manufacturing and services, it has become an important economic growth catalyst for the nation.

2. MALAYSIAN AGRICULTURAL SECTOR

The Malaysian agricultural sector is dominated by the production of palm oil, rubber and forestry products. Other major agricultural production includes rice, poultry, fruits and vegetables (Wong, 2007). Oil palm and rubber currently account for 70 percent of all agricultural land use. Area expansion was the major source of agriculture production growth until recently. However, as modest expansion of agricultural land use in agriculture continues, crop and other plantation are also being lost due to urbanization. In retrospect, pressure on land and labour resources have caused a structural change in the agricultural sector.

Even though the agricultural sector contributed only 8.2% of the Malaysian gross national product, the total labour force engaged in this sector accounted for 13.1% of the overall workforce in 2006 (Statistics Department, 2007). It is therefore pertinent that this sector should not be neglected as it involves a large proportion of the Malaysian labour force. Moreover, with the increasing trend of global food prices, the increasing need for food security as well as the opportunity for large commercialisation of food products, the sector should be given greater attention by the government. Table 1 shows the profile of Malaysian agricultural sector as well as the agro-based industries production output in the year 2000.

The above table shows that within the agriculture sector itself, the largest contribution to the GDP are given by forestry and logging (10.7%), oil palm plantation (8.1%),other agriculture (which comprises of paddy, fruits, pepper, flower planting and veterinary services) (5.2%) and livestock breeding (4.8%). On the other hand, the agro-based industries that contribute large proportions to the agriculture sector as a whole are manufacturers of oils and fats (21.3%), sawmills (8.1%), rubber and related products (5.6%) and furniture (5.3%). It can also be seen that in terms of food production, if we aggregate meat and meat products, dairy production, seafood, preservation of fruits and vegetables, bakeries, confectionary and other food, they account for almost 10.3% of the overall contribution to the agriculture and agro-based sectors.

In view of the fact that the values of agricultural production and the related agro-based products are still significant in the Malaysian economy, the 9th Malaysian Plan (9MP) has given a new focus in this sector on the implementation of the plan (Ministry of Finance, 2005). One of the policies reflected in the 9MP is the revitalizing of the agricultural production, particularly in the rural areas. This is because priority was given to the rural populace with the goal of eradication of poverty level and at the same time uplifting the income level of the rural income as stipulated in the New Economic Policy (NEP).

In addition to that, the Third National Agricultural Policy (NAP3), which covers the period of 1998-2010, provides the policy framework for the future growth of the agricultural sector in the next decade of 2010 to 2020 (Ministry of Agriculture, 2000). The policy has been formulated to ensure that the agriculture sector's strategic role in the national development is sustained and enhanced. The overriding objective of the NAP3 is the maximization of farm income through optimal utilization of resources in the sector and to enhance the domestic food production.

3. ECONOMIC LINKAGES IN THE INPUT OUTPUT FRAMEWORK

The relationship between the flow of the various sectors in the economy can be traced using input-output analysis. Analysis can be done through the input-output table whereby the relationship between the producers and the consumers as well as the interdependence among industries can be shown. It tracks the commodity flow (goods and services) from one industry to

another industry. This flow of commodities supplied and used is compiled systematically in the form of input-output tables as mentioned in (Hj Ismail, 2007).

Table 1: Profile of Malaysian Agricultural and Agro-based Production 2000

No	Sector	Value (RM Million) Current Prices	Percentage
	Agriculture		

1	Other Agriculture	6,858,108	5.2
2	Rubber plantation	2,036,377	1.6
3	Oil Palm Plantation	10,643,095	8.1
4	Coconut	145,326	0.1
5	Tea estates	27,580	0.0
6	Livestock breeding	6,290,101	4.8
7	Forestry and Logging	13,928,279	10.7
8	Fishing	5,452,827	4.2
	Agro-based Industries		

9	Meat and meat production	1,389,961	1.1
10	Dairy production	2,236,793	1.7
11	Preservation of fruits & vegetables	607,143	0.5
12	Preservation of seafood	1,499,537	1.1
13	Manufactures of oils and fats	27,847,207	21.3
14	Grain mills	2,357,800	1.8
15	Bakeries	1,687,408	1.3
16	Manufactures of confectionary	1,164,080	0.9
17	Manufacture of other food	4,786,273	3.7
18	Manufacture of animal feed	2,469,741	1.9
19	Soft drinks	1,180,033	0.9
20	Tobacco	1,637,130	1.3
21	Sawmills	10,554,243	8.1
22	Manufacture of wood products	2,116,467	1.6
23	Furniture	6,884,801	5.3
24	Paper and board industries	6,517,466	5.0
25	Rubber processing	3,255,365	2.5
26	Rubber and related products	7,351,856	5.6
	Total	130,969,997	100.0

Source: Input-Output Table Malaysia 2000, (published 2005).

The input-output table consists of four quadrants. The first quadrant is the intermediate input quadrant, which is referred to as the heart of an input-output matrix (Jensen and West, 1986). The second quadrant represents the final demand where it is considered as the output of the producing sectors, i.e the sectoral distribution of household expenditure, government expenditure, fixed capital formation and exports (the destinations of output that do not flow to other sectors as inputs). On the other hand, the third quadrant shows the primary inputs quadrant, which consists of the sectoral distribution of wages, operating surplus, value added, indirect taxes, subsidies and depreciation, while the fourth quadrant represents the primary inputs directly

linked to final demand (O'Connors & Henry, 1975; Miller and Blair, 1985, Jensen and West, 1986; Sauian 2007).

The input-output tables describe the complex process of production, the use of goods and services and the way in which income and value-added products are generated within the various sectors of the economy, where the set of producers of similar goods and services forms a homogenous industry (Valadkhani, 2003). Through a set of tables during a period, the structural change in the economy and the specific sector's economic characteristics can be revealed (Wu & Zhang, 2005).

In essence, the symmetric input-output table is a product or industry matrix describing the domestic production processes and the transactions in products of the national economy in detail. For example, a two-sector input-output table allows us to understand the industrial relationship between agriculture and the rest of the economy, thus highlighting the implication for structural and policy analysis (Pizzoli, 2004).

Therefore, input-output analysis has multifarious applications. For instance, it offers a static view of the structural relationship among the different sectors in the economy (typically national or regional) for a certain period of time, generally a year. The relationship is expressed purely in monetary terms (Lee and Mokhtarian, 2004). Other applications of input-output tables are determining the technical capability of production, labour productivity and comparing the technological standard of one country compared vis-à-vis other countries. Similarly, the study of economic linkages among the various sectors of the economy can easily be discovered.

4. METHODOLOGICAL APPROACHES

4.1 Forward and Backward Linkages

In an interdependent economy, a sector is linked to its input and output sectors by its direct and indirect purchases and sales (Cai & Leong, 2002). A sector's linkage through its direct and indirect purchases is called the backward linkage. On the other hand, a sector is said to be forward-linked to other sectors through its direct and indirect sales to them. Hirschman (1958) stated the analysis of strengths of backward and forward linkages allows us to identify the most important sectors in the economy.

Backward linkage or input provision is defined as an activity that employs significant amount of intermediate inputs from other activities for production purposes. Output utilization or forward linkage, on the other hand, is defined as an activity that caters for final demand but also induces attempts to utilize its output as inputs in other new activities (Hirschman, 1958 and Linnemann, 1987). Linkages between agriculture and the rest of the economy had been used in many input-output based general equilibrium models. Examples are the works of Norton (1988) and Song (1998). The former analysed the incidence by sector and household income for the benefits of food aid programmes, while the latter examined the impact of agriculture and other industries. Van Zyl and Rooyen (1990), also used the concept of economic linkages to evaluate the contribution of agriculture to the economic growth of South Africa.

Two common approaches to measure the strength of backward and forward linkages are the works of Rasmussen (1956) and the Chenery & Watanabe (1958). Since the Chenery-Watanabe Approach of evaluating the impact of a sector to the overall economy is only confined to the direct linkage only, we resort only to the Rasmussen's approach as it considers both the direct and the indirect linkages.

For inter-industry comparison purposes, the linkage indices are normalized in such a way that their average value is unity. Based on Rasmussen's model, the measure of backward linkages is called the Power of Dispersion Index. It describes the relative extent to which an increase in final demand for a product of a given industry is dispersed throughout the total system of industries. The Power of Dispersion Index is defined as:

$$BL_i^{AW} = \frac{\sum_i B_{ij}^{AW}}{\left(\sum_i \sum_j B_{ij}^{AW}\right)/N} \quad (1)$$

where BL_j indicates backward linkage of agriculture in jth sector

A indicates agricultural sector

W indicates other sector

N indicates number of sectors

B_{ij} is the ij^{th} element of the Leontief's inverse matrix (see Miller & Blair 1985).

The forward linkage index is measured using The Sensitivity of Dispersion Index. The index describes the importance of a given industry as a supplier of resources to other industries. It is defined as in equation 2 below:

$$FL_i^{WA} = \frac{\sum_i B_{ij}^{WA}}{\left(\sum_i \sum_j B_{ij}^{WA}\right)/N} \quad (2)$$

Where FL_j indicates the extent of forward linkage of agriculture in jth sector and A, W, N and B_{ij} as defined in equation 1, above.

4.2 Impact Analysis

Another approach in looking at the impact of agriculture is through the use of impact analysis. This analysis is also known as multiplier analysis. Multiplier analysis measures the total change throughout the economy from one unit change for a given sector. For instance, for every one dollar of final demand for a product of a sector generates direct and indirect income to the economy as a whole. The relationship between the initial spending and the total effect generated by the spending is known as the impact of that sector to the economy as a whole.

In this analysis, we use only the income multiplier because it is the simplest form of various multipliers. It is considered useful because it is expressed as a ratio of the sum of direct and indirect income change resulting from a unit change of final demand in that sector. The calculation is obtained by multiplying the rows of technical coefficients of the income in each sector by the column of the interdependence coefficients (O'Cornor & Henry (1975)). It should be noted that the multiplier value is less than unity. The partial income multiplier is defined as:

$$\varepsilon = (I - A)^{-1} A_y^T \quad (3)$$

where $(I - A)^{-1}$ is the interdependence coefficient A_y^T is the technical coefficients of income arising.

5. RESULTS AND DISCUSSIONS

5.1 Results of the Economic Linkages Analysis

Using the Rasmussen's indices of backward and forward linkages we could determine the respective indices using equations (1) and (2) by adopting the Malaysian Input-Output Table 2000.For strong backward and backward linkages the values must be greater than 1. Table 2 shows the summary of the backward as well as the forward linkages of the agricultural and agro-based sectors where at least one of the linkages is greater than 1.

Table 2 shows that the index of dispersion and the index of sensitivity for livestock breeding, fishing, manufacture of oil and fats and manufacture of paper and board products are greater than one. This indicates that all these sectors have high backward as well as forward linkages, which implies that they are both important suppliers as well as important purchasers of materials from other industries. On the other hand, we have 13 agriculture and agro-based sectors with strong backward linkages but weak forward linkages. These include meat and meat products, dairy production, preservation of fruits and vegetables, preservation of seafood, grain mills, bakeries, other food, soft drinks, sawmills, wood products, furniture and rubber products. This imply they are important consumers or purchasers of inputs from other industries.

There are four agriculture and agro-based industries that show strong forward linkages but weak backward linkages. They are other agriculture, oil palm estates, forestry and logging and manufactures of animal feed. In this regards, these sectors are good suppliers of inputs or materials to other sectors in the economy.

Table 2: Summary of Backward and Forward Linkages of Selected Agriculture and Agro-based Sectors year 2000

No	Sectors	Sector 1	Sector 2	Sector 3	Sector 4	Sector 5	Sector 6	Sector 7	Sector 8
	Interdependence Coefficients								

	Oil Palm	1.00045	0.01024	0.00010	0.01072	0.00780	0.01195	0.00086	0.01239
	Livestock	0.00410	1.30296	0.00010	0.00026	0.59650	0.00622	0.00044	0.00218
1	breeding								
2	Forestry, Logging	0.00286	0.00368	1.02836	0.00214	0.00382	0.00009	0.04837	0.00702
	Fishing	0.00026	0.00640	0.00014	1.10434	0.02768	0.00040	0.00042	0.16520
3	Meat, meat	0.00011	0.00214	0.00005	0.00047	1.11316	0.02032	0.00023	0.00072
	products								
4	Oil and Fats	0.01195	0.02670	0.00029	0.00192	0.02032	1.98708	0.00225	0.05537
5	Other food	0.00030	0.00651	0.00013	0.00085	0.03029	0.05537	1.05939	0.04179
	Paper, wood	0.00135	0.00570	0.00059	0.00170	0.05136	0.00225	0.00054	1.17805
6	products								
7	Technical								
8	Coefficients								

	Compensation of Employees	0.12269	0.05205	0.13624	0.10680	0.05124	0.06092	0.08652	0.09200
	Income M'plier								

	Income Arising (ε)	0.12399	0.07308	0.14110	0.11993	0.09985	0.12906	0.09862	0.13564

* Calculated from Input-Output Table Malaysia 2000

5.2 Results of the Impact Analysis

We could determine the partial income or output multiplier using equation (3) utilizing the data of the input-output tables of Malaysia for the year 2000. It should be noted that in this analysis, we only considered relevant sectors with strong backward and/or forward linkage. Table 3 below depicts the partial income multipliers of the eight sectors with these properties.

The above table shows that an increase in one unit of final demand of oil palm production, will result in an increase 0.12239 unit of income to the entire economy. Similarly, an increase in one unit of final demand in the paper and wood products sector will result in an increase of 0.13564 unit of income as a whole. Overall, we can see that income multipliers of the relevant eight sectors have a decent multiplying impact between 7 to 14 percent to the entire economy.

6. CONCLUDING REMARKS

From this analysis, it was shown that the agriculture sectors with the agro-based industries had benevolent impact on the whole of the Malaysian economy. Using the economic linkages analysis, we showed that 4 sectors have strong backward as well as forward linkages to the other sectors of the economy. 13 sectors have strong backward linkages but weak forward linkages, while 4 other sectors have strong forward linkages but weak backward linkages. Generally, it appears that the agriculture and the agro-based sectors have above average linkages to the other sectors in the economy. It is therefore clear that they are still important sectors as they are either the supplier of inputs to the other sectors or consumers from other sectors in the economy.

Using the multiplier analysis, we showed that most of the agriculture and agro-based sectors have significant impacts on the economy as a whole. This was indicated by the modest values of partial income multipliers on at least 8 selected sectors.

This study also gives an affirmative note to laud the government efforts in enhancing these sectors as catalyst for growth. It justifies the greater focus by the government on agriculture in the New Agricultural Policy as well as in the 9th Malaysian Development Plan (9MP). With the new significant contributors to the GDP like tourism and halal products (products which are Syariah compliant), golden opportunities in agriculture lie ahead. Concerted efforts are now carried out to promote “agro-tourism” as well as the setting up of “halal hub” in Malaysia for importers and exporters. With these new embarkation and commitment by the government, the potentials of agriculture and agro-based sectors to emulate growth and contributing higher proportion in the economy are promising.

REFERENCES

Bank Negara Malaysia (2006), *Annual Report 2006.*

Cai, J; Leong P.S (2002); The Linkages of Agriculture to Hawaiian Economy,” *Cooperative Extension Service; College of Agriculture and Human Resources; University of Hawaii.*

Cai, J, Leong P.S et al (2005); Economic Linkage Impacts of Hawaii Longline Fishing Regulations*; Fisheries Research, 74-:232-242.*

Chua, P C (2000); Farm Management in Agricultural Extension- Malaysia ; *Food and Agriculture Organization (FAO), United Nations; Regional Office for Asia and Pacific, Bangkok.*

Chenery, H, B; Watanabe, T (1958); International Comparison of the Structure of Production; *Econometrica 26, pp 487-521.*

Department of Statistics Malaysia (2007); *DOSM Malaysia*.

Economic Planning Unit (EPU) (2006); Ninth Malaysia Plan*; Prme Minister's Deapartmaent, Malaysia.*

Hirshman, A.O (1958*), The Strategy of Economic Development*; New Haven; Yale University Press.

Hj Ismail, F;(2007); Structural Change of Agricultural Sector; *Department of Statistic*s *Malaysia.*

Jensen, R.C; West G.R (1986); Input and Output for Practitioners: Theory and Applications; *Australian Government Publishing Service, Canberra.*

Linnemann, M (1987*); Inter-industry Linkages and the Process of International Integration*; Department of Applied Economics, University of Cambridge.

Lee, T; Mokhtarian, P.L (*2004); An Input-Output Analysis of the Relationship between Communication and Travel Industry*; University of California, Davis.

Miller, R. E; Blair P.D (1985); *Input-Output Analysis: Foundation and Extensions*; New Jersey; Printice-Hall Inc.

Ministry of Agriculture (2000); Third National Agricultural Policy (NAP3) (1998-2010); *Ministry of Agriculture Malaysia.*

Norton, R.D (1988); Policy Analysis for Food and Agricultural Development: Basic Data Series and Their Uses*; FAO, United Nations.*

O'Cornor, R, Henry E.W (1975); *Input-Output Analysis and Its Applications*; London; Charles Griffin and Company Ltd.

Pizzoli, E (2004); Agricultural Sector in an input-output Matrix: Microdata Approach for the Italian Case*; Conference on Input-Output and General Equilibrium: Data, Modelling and Policy Analysis; Free University of Brussels.*

Raja Ahmad; R>H (2008); Measuring the Impact of Agriculture in the Malaysian Economy: Input Output Analysis*; Project Paper, UiTM, Malaysia.*

Rasmussen, P N (1956*); Studies in Inter-sectoral Relations*; North-Hoplland.

Sauian, M S (2007); Input-Output Analysis: An Enthusiastic Approach in Securing Sectoral and Productivity Planning; *Proceedings of the Ninth Islamic Countries Conference on Statistical Sciences, ICCS-IX, Shah Alam, Malaysia.*

Song B, Woods M.D;Doeken G.A et al (1998); Multiplier Analysis of Agriculture and Other Industries; *Multiplier Cooperative Extension Service; AGC-821, State University of Oklahoma.*

Valadkhani, A (2003); Using Input-Output Analysis to Identify Australia's High Employment Generating Industries; *University of Wallongong, Australia.*

Van Zyl, J; Van Rooyen C.J; (1990); Agricultural Production in South Africa: Harvest of Discontent; the Land Question in South Africa, *IDASA.*

Wong, J (1985); ASEAN Experience in Regional Economic Cooperation*; Asian Development Review, ADB Vol 30 No1*.

Wong, L.C.Y (2007); Development of Malaysian Agricultural Sector: Agriculture as an Engine for Growth; *Conference on Economics Development & Challenges, ISEAS, Singapore.*

Wu, X; Zhang, Z.(2005); Input-Output Analysis of the Chinese Construction Industry; *Construction Management and Economics; 905-912.*

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 998–1029.

HUMAN CAPABILITIES AND INCOME SECURITY
A NEW METHODOLOGICAL APPROACH

Hussein Abdel-Aziz Sayed
Faculty of Economics,
Cairo University
husseinsayed@hotmail.com

Ali Abdallah
Faculty of Commerce,
Assiut University
ali_statistics@yahoo.com

Zeinab Khadr
Faculty of Economics,
Cairo University
zeinabk@aucegypt.edu

ABSTRACT

As the World Bank moves toward a broader understanding of poverty reduction and the relationship of risk to poverty, the standard concepts and interventions of social protection are no longer sufficient. In its first strategy paper for the social protection sector published in 2001, the Bank highlights the need to expand the definition of social protection to encompass all public interventions that help individuals, households, and communities to manage risk or that provide support to the critically poor (World Bank, a, 2001). The paper recommends that social protection programs be embedded in an integrated approach to poverty reduction based on a new framework for social risk management. In the same regard, many researchers - among them Nobel Prize winners - have studied the causes of poverty and the means to eliminate this phenomenon introducing new approaches and innovative solutions. Theodore Schultz[1] and Gary Becker[2] pioneered new ideas focusing on human capital investment as a highly efficient tool in poverty reduction. Moving in the same direction, this study introduces an integrated approach based on social risk management to propose an integrated income security plan. The underlying approach relies on utilizing human capabilities during early stages of the lifecycle to increase the chance of not falling into poverty during the later stages. This paper first introduces reclassification of the human life cycle to improve income related risk management, and then introduces a comprehensive income security plan aimed at tackling poverty and its causes. The plan we suggest is called Human Capital – Income Security (HCIS), and it adopts a self-aid approach that allows students to borrow using on their future expected income as collateral, to pay for their education expenses with deferred repayment. It also helps individuals to manage retirement benefits during old age. Thus, the HCIS plan helps individuals to manage their lifetime income and expenses during education, employment and retirement stages of their lifecycle managing the changing needs and surplus from stage to stage. Three different risks are managed through the suggested plan; the risk of insufficient resources to finance investment in education during the pre-employment stage, the risk of not having enough income to payback study loans during the work stage, and the risk of insufficient income during old age. In this paper, we develop an actuarial model for the suggested HCIS plan and tests it using death and invalidity rates from the Social Insurance Fund for Government Employees in Egypt.

[1] Theodore Schultz was the 1979 winner of the Nobel Prize in economics; he promulgated the idea of educational capital, an offshoot of the concept of human capital.
[2] Gary Becker was awarded the Nobel Prize in economics in 1992; he is interested in social economics and among the foremost exponent of the study of human capital.

Keywords: Human lifecycle, Income security, Human capital, Education financing, Income contingent loans, Student loans, pension plan.

1. INTRODUCTION

Historically, the primary source of wealth has shifted over the last few hundred years from land (at the end of the 18th century) to physical capital (at the end of the 19th century) to human capital - education and cognitive ability - by the end of the 20th century (DeMuth, 1997). Along this line, the definition of poverty has evolved into a second stage. Defining poverty through primarily quantitative approach dominated the line of thought over the second half of the twentieth century, whereas utilizing the capability approach for the definition of poverty became widely accepted since the 1990s. The quantitative approach focuses on the insufficient cash income to maintain a minimally acceptable standard of living. However, the capability approach goes beyond the lack of income concept and is based on the performance of certain functions necessary for human welfare including social attributes such as knowledge, skills, health, security, and freedom, which enables a person to use her/his own capabilities to generate enough lifetime income over different stages of their lifecycle (Ali-Eldin, 2003).

Changing definitions and approaches affect strategies relating to fighting poverty and identifying ways in which poverty can be effectively tracked. Schultz and Becker in their work dismissed the pessimistic vision of Malthus which sees poverty as an inevitable catastrophe of the interaction between population growth and destruction of natural resources (Mursa, 1981). They conclude that state of underdevelopment in many countries does not originate from the rarity of physical capital but from insufficient resources allocated to raise population capabilities and to knowledge improvement through investment in education and other form of human capital (Lee, 1999).

On one hand, Gary Becker pointed out during the early sixties that "Education is an investment and it adds to our human capital just as other investments add to the physical capital" (Becker, 1964). On the other hand, Schultz proposed an approach to economic growth focusing on efficient management of expenses coupled with improvement of the quality of the working force, which dismisses the increase of physical capital approach that dominated economic growth since the 1960s (Mursa, 1981).

The fundamental idea is that education enhances individuals' abilities, qualifications and knowledge, which in turn increases their productivity. Thus the human potential, helped by the institutional arrangement, is capable of generating creativity, intelligence, and effort, and of providing adequate answers to the problem of resource scarcity (Mursa, 1981). These ideas were proven to be effective in achieving economic prosperity as clearly illustrated by countries such as Japan, Taiwan, Hong Kong, South Korea, and other fast-growing Asian economies. Although these countries are characterized by somewhat limited natural resources, they still managed to achieve significant and rapid development through building the capabilities of their populations (Becker, 1998).

In many countries, where student loans are not available, the current system in education is something similar to pay-as-you-go mechanism in pension plans, where current workers pay for current retirees with a promise that tomorrow's workers will pay for their retirement benefits. In education, current parents/ generation pay for education expenses of the current students, with the passive promise that current students will pay for educational expenses of the following

generation. Modigliani[3] argued that the pay-as-you-go system must be discarded since it has proved financial unsoundness (Modigliani et al, 1999). Contributory personal pension schemes introduce a reform option to overcome the pitfalls of pay-as-you-go scheme (they allow each person to pay for her/his own retirement benefits) (Palacios and Sluchynsky, 2006).

It is not only the pay-as-you-go mechanism involved with the retirement pension plan which has proved to be inefficient but also policies and approaches related to educational investment policies need to be revised to move from pay-as-you-go mechanism to self-aid plans. Such revision is about the financing mechanism and options available for students to finance their educational expenses.

Similar to the concept of the personal saving plan which introduces an alternative solution to avoid the disadvantages and pitfalls of the pay-as-you-go mechanism; this study suggests income security plan based on self-aid approach that helps an individual to use her/his own lifetime income to manage the risk of not having sufficient income to pay for education expenses.

This paper proposes a comprehensive approach to income security aimed at tackling poverty and its causes. It suggests a life course plan in which each individual has a personal saving account through which s/he adds or receives transfers during different stages of their lifecycle to achieve lifetime income smoothing. We refer to this plan as Human Capital – Income Security (HCIS). HCIS adopts a self-aid approach and enables students to borrow utilizing their future income as collateral to pay for their education expenses with deferred repayment. It also helps the subscriber to manage retirement benefits during old ages.

This plan benefits from the insurance criterion of pooling similar risks to deal with uncertainty concerning future income. It enables students to pool a fraction of their future earnings with others' earnings, in the same way as insurance companies do when they allow individuals to pool their risk with others who face similar risks. Thus reduction in uncertainty for the student translates into greater uncertainty for the insurer (Vandenberghe, 2004). However, the insurer is in a much better position than the student to diversify risk and if processing a large number of subscribers, can benefits from the law of large number so that there is a less uncertainty about future income. Thus, the suggested plan relies on two different pillars: the first concerns the introduction of a sustainable borrowing mechanism. The second pillar is to help the individual to obtain retirement benefits during old age.

2. HUMAN CAPITAL INVESTMENTS

Human capital refers to knowledge and skills of individuals. Previous schooling, computer training courses, for example, are accumulated capital that enhance an individual's personal values whether in the labor market or everyday life over much of her/his lifetime (Becker, 1998). Therefore, economists regard investment in education, training, and medical care to be similar in many ways to investment in physical assets. The underlying idea is that people possess skills, and knowledge which are viewed as a form of capital enabling them to increase productivity. Therefore, a person will invest in his self at the present time to achieve greater rewards later in the form of higher level of earning, greater job satisfaction over one's lifetime and a greater appreciation of nonmarket activities and interests (Becker, 1964).

[3] Franko Modigliani was awarded Nobel Prize in economics for his pioneering research in several fields of economic theory that had practical applications. One of these was his analysis of personal savings, termed the lifecycle theory.

Education is a fundamental component of human capital, and is recognized for its positive impact on alleviating poverty, reducing child labor, and, in the long run, contributing to the growth of the economy as well as to the whole social development process of any society. It is the most effective way to increase social mobility among young people with poor backgrounds (Becker, 1993).

Student's investment in higher education is made with the expectation that the future financial returns from acquired skills and increased income will outweigh the current costs, both direct and indirect (Becker, 1993 and Perkins, 2003).

Defining wealth in the form of human capital as present and future earnings due to education, training, knowledge, skills, and health, lead to estimates in developed countries that place the value of human capital at three to four times the combined value of stocks, bonds, housing, and all other physical assets (Becker, 1997). Such estimates should not be surprising since wages and salaries account for 75% of the national income in these countries.

Empirical studies aimed at determining the internal rate of return of educational investments are numerous. Psacharopoulos and Patinos (2002) found that the worldwide rates of return for investment in higher education are approximately 19% per year. In the developing countries, the rates of return are even higher. For example, in Sub-Saharan Africa, it is estimated to be 27.8% per year (Carver, 2004). Additionally, Becker (1964) found that the secondary school rate of return in USA is 28% and the college rate of return is 14.8%.

These empirical studies are descriptive in nature and faced by a large number of challenges and methodological disagreements. For example these studies use income as the only measure for return on investment, which neglects other benefits of educational investments, such as enhanced social opportunity or status. However, these studies often provide a major contribution namely policy enlightenment which promotes investment in education and ensures that low income families in particular make educational investments (Carver, 2004).

3. FINANCING INVESTMENT IN EDUCATION

Investment in education has usually been the responsibility of government and/or individual families. Financing and supporting investment in education are commonly carried out through two different approaches:

- Supply side support: in which government supports human capital investment in education through general investment in schools and universities. In this approach, education is theoretically an unconditional free service.

- Demand side support: every student pays for her/his education expenses. Individual financial aid is used to pay for students who have insufficient funds to pay for their capital investment. This approach is currently utilized in USA, and it is a vital component within demand side support that makes contributions to education for over half of all students entering colleges and universities (Perkin, 2003).

With the many problems facing supply side support, there is a shift away from societal/ supply side support, with more responsibility being handed to students themselves (Perkins, 2003). The demand side support is built on two main sources to finance education expenses (Perkins, 2003):

- Family contribution: parents and other family resources
- Financial aid packages which can be accomplished through:

- Gift aid: grants, scholarships, and fellowships
- Self-help aid: employments and loans

Many ideas and plans were experimented and discussed in USA to present a fair borrowing plan to finance human capital investment in education expenses (Perkin, 2003).

3.1 Student Loans (All Debt: Fixed Repayment Stream)

Student financial aid started with gifts and fixed student loans. The limited government budget and the high default rate of student loans put a heavy burden on the lenders. Accordingly, subsidized federal loans by which the U.S. government subsidizes almost all student loans became a popular means for financing higher education, and today are an important resource for most students. Such loans are carried out by either directly issuing them at a low interest rate, or guaranteeing the creditworthiness of the borrower. However, they are not the optimal solution since instead of addressing the problem; they transfer the risk of the investment to the taxpayer (Wirt et al, 2002).

During the 1999-2000 school year, 29% of the 16 million enrolled undergraduates in the U.S. took out student loans averaging $5100 each for total undergraduate loan amount of $24.4 billion (Perkins, 2003).

3.2 Income-Contingent Loans (Percentage Of Income Loan: Variable Repayment Stream)

Friedman first suggested his idea about future income loans in 1962 (Friedman, 1962). He proposed that students pay back a percentage of their income over a specified interval of time. Eventually, a hybrid between traditional fixed repayment loans and Friedman's proposal came into being in the form of income-contingent loans. This is a financing instrument that requires the borrower to pay back a percentage of income until the loan principal is paid off (Chapman, 2003).

3.3 The Human Capital Contract

Currently, Friedman's ideas are considered as a way to address increasing education costs and limited government resources. The human capital contract attempts to create financial instruments that allow equity-like investments in higher education that can attract private capital to the human capital investment market. This instrument is based on a contract by which an individual obtains resources to finance his or her education by giving back a percentage of her/his income over a predefined period of time after graduation. Such an instrument is referred to as a human capital contract. Palacios explains that human capital contracts are convenient for students and investors for at least four main reasons (Palacios, 2003):

1- They relieve the student from any uncertainty about being able to make fixed loan payments,
2- They virtually eliminate default due to financial distress,
3- They are needs blind,
4- They give a subsidy to those who most need it during the repayment period as an individual pays a percentage of her/his income with no obligation to repay the full loan in the case of low income career.

4. INCOME SECURITY DURING OLD AGE

Currently, there is a policy debate in many countries concerning the roles and mechanism of pension schemes. Modigliani et al (1999) argued that the pay-as-you-go system must be discarded since it is financially unsound. Their underlying line of thought is that since contributions (which are dependent solely on the active working age population and basically a type of mandated saving) are used entirely to finance pensions, the capacity to pay current pensions is sensitive to unforeseeable changes in the demographic structure and the growth trend of productivity.

A move towards funded pension schemes is currently under discussion in many developed and developing countries – as a pension reform option. The personal saving plan is a suggested alternative, in which each person pays forward for her\his retirement benefits. This move comes in response to the unsuccessful reforms of unfunded schemes (World Bank, 2002).

5. MODIFICATION OF THE HUMAN LIFE-CYCLE

The life-cycle approach is a potentially powerful tool that could be used to plan current/future outcomes/transfers of different age groups across different stages of an individual's lifecycle (Modigliani, and Jappeli, 2003). The previous discussion concerning investment in education during young ages and saving for retirement benefits during old ages illustrates the need for a modified classification of the human lifecycle that enables a new, more efficient mechanism.

5.1 Traditional Human Life-Cycle And The Pay-Get Transfer Mechanism

Economically, the life cycle of an average person is commonly divided into three successive stages, young age, work, and old age/ retirement, as Figure 1 shows. While people produce in excess of their consumption during the middle stage, the reverse is true during the first and last stages. Therefore, the working age population accumulates savings for their future consumption. These savings are called transferred wealth, which is defined as the present value of the difference between the transfers that an individual expects to receive in the future, and the transfers s/he expects to give (Lee et al, 2000). The Life-Cycle approach posits that the main motivation for saving is to accumulate resources for later expenditure and in particular, to support consumption at the same standard during retirement (Modigliani and Jappeli, 2003). Thus, saving should be positive for households during their working span and negative during retirement, so that wealth should be bell-shaped, as Figure 2 illustrates.

Figure (1): Traditional human life-cycle Pay-get transfer mechanism

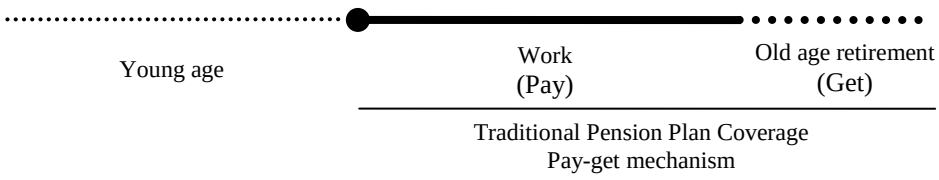
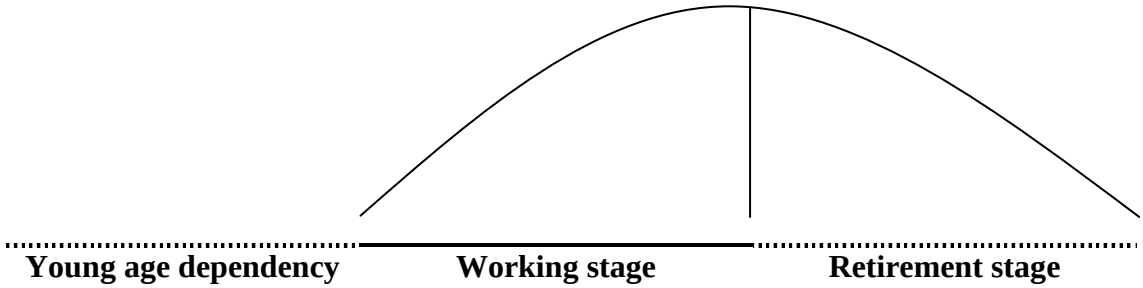


Figure (2): Wealth accumulation according to traditional life cycle assumptions

Pay-Get Mechanism



The traditional pension plan is a transfer mechanism that helps an individual to achieve income security during work and retirement. An individual pays contributions in advance during the employment stage and collects benefits during the retirement stage. Therefore, it is called a pay-get mechanism.

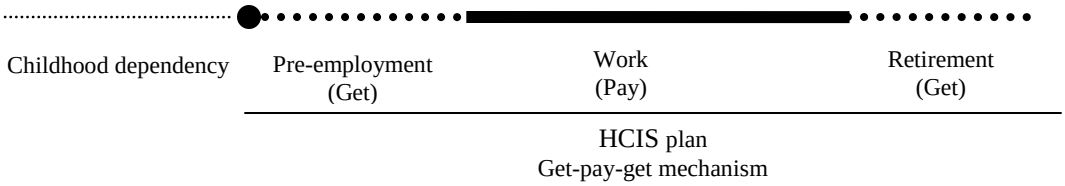
The redistribution of income through the work and retirement stages of the life cycle aims at achieving income security during old age. It is either enforced by law such as public pension plans, or reflects individual decisions (Abdallah, 2004).

5.2 Reclassification Of The Human Life-Cycle And The Get-Pay-Get Transfer Mechanism

Practically, an individual cannot join a pension plan or any other income security plans before s/he starts working. However, the young age dependency stage may extend until the age of 20 or 25 years or sometimes even later. This stage is crucial in building human capabilities, which in turn significantly improves the prospective chances for individuals to prosper during the following stages (Rank & Hershl, 2001).

This study suggests dividing the young age dependency stage presented above as part of the traditional life cycle into two stages (figure 3). An individual is supposed to receive educational support transfers during the latter stage. The first stage lasts up to the end of the preparatory school if an individual starts receiving educational transfers during the high school or it lasts up to the end of the high school if s\he starts receiving educational transfers during university. The second stage continues from this age and ends upon being employed in a paying job. The first stage is referred to as the childhood dependency stage while the second is referred to as the pre-employment stage. Thus, the human life cycle according to this new classification is divided into four stages, namely childhood dependency, pre-employment, working, and the retirement stages.

Figure (3): Modified classification of human life-cycle
Get-pay-get transfer mechanism



The new human life cycle assumptions are that the individual relies only on her/his own life time income starting in pre-employment stage, and that the main motivation for saving is to accumulate resources to finance previous and later expenditures. Therefore, saving should be positive during the working span and negative during the pre-employment and retirement stages. According to the new classification, wealth takes the pattern illustrated in figure 4.

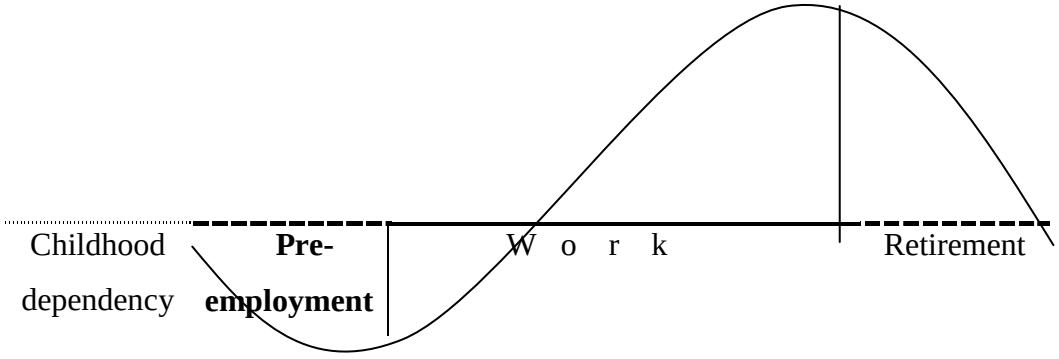
As shown in this figure, the wealth curve accumulates in the negative direction during the pre-employment stage, and then starts increasing during the working stage, reaching a peak at the retirement age, at which point it starts decreasing again.

The most important risk related to the pre-employment stage is insufficient human capital investment. This leads to many problems both in the short and long term. In the short term could lead to inadequate skills development which could result in poor prospects for employment. In addition, it could also lead to long-term unemployment or low wage employment which could result in inter-generational transmission of poverty, and subsequent negative consequences.

The risk of not having enough resources during the retirement stage is an insurable risk which can be covered by a public pension plan. However the risk related to not having sufficient qualifications to generate adequate income, has a very crucial impact on person's earning potential and needs more attention and risk management.

Figure (4): Wealth accumulation according to the modified life cycle stages

Get-Pay-Get Model



Preventive risk management is an effective option in this regard that consists of increasing human capital investment in education during the pre-employment stage, which prospective chances of obtaining a adequate work and income, and reduces the possibility of falling into poverty in the future. We regard this form of human capital contract as the most promising avenue of financing education expenses during young ages.

6. THE PROPOSED HCIS PLAN

The proposed HCIS plan introduces a comprehensive income security plan covering risks related to income during pre-employment, work and retirement stages. A borrower student receives a certain amount of money to invest in their education, and then repays part of their income during their work stage to pay back their study loan as well as save for the retirement stage. The repaid amounts for their study loan and retirement benefits are uncertain since they are dependent on the person's capacity to pay. If the career path of the borrower/student is less profitable than expected, s/he repays less than the amount of her/his study loan and benefits from the insurance of pooling similar risks to manage uncertainty concerning future income. Insurance organizations benefits from the law of large number, reducing the uncertainty of their income. High-income earners will cover the losses resulting from low-income earners.

The HCIS plan manages three risks relates to income security at different stages of the lifecycle including:

- the risk of insufficient resources to finance an investment in education during the pre-employment stage,
- the risk of not having sufficient income to pay back study loans during the work stage
- the risk of insufficient income during old age.

The HCIS plan has the following main features:

1. It targets support for higher educational levels,
2. It is self-aid type, i.e. each individual pays for her/his own educational expenses during the pre-employment stage,
3. It focuses on demand-side support,
4. It is an insurance plan not a welfare program, i.e. it relates benefits to contributions.

6.1 The HCIS Plan Targets Support For Higher Educational Level

Concerning support for education there are two different levels with different tools and objectives:

- Minimal educational support: the main objective here is to reduce illiteracy rate and enable basic education for all.
- High educational support: such plans aim at supporting those who have demonstrated willingness and ability to succeed at high school and university levels.

While the first approach has received and continues to receive attention, the second approach is crucial for the progress of any country. The second approach supports those who become top bureaucrats, planners and most importantly those who make great contributions to the development of any nation.

Providing fair chances in education by means of an even amount of support for everybody could be useful for the first objective (the basic educational level). However during the high school and university period it causes disappointment for those who pay more efforts in their study. The HCIS plan attempts to contribute to the second objective through supporting individuals according to their demonstrated willingness and ability to succeed.

6.2 The HCIS Plan Is Self-Aid Type

Transfers are a very important tool used in poverty alleviation and income security in particular. In this regard, there are three types of transfers to be considered:

- Inter-generational transfers: it is a type of income redistribution between generations, for example parental transfers to finance their children's education and health care. Transfers from young workers to retirees in the theoretical model of Pay-As-You-Go pension schemes and national investment in public education provides examples of such transfers.
- Between-group transfers in the same society: this relates to income redistribution among different groups in the same society. For examples taxes collected from rich people for the benefit of poor individuals or families.
- Self transfers from one lifecycle stage to another for the same person: This refers to transfers made by an individual through different stages of her/his life cycle.

Our suggested plan is a self-transfer strategy that depends mainly on an income redistribution mechanism through which an individual transfers funds from the working stage where there are surplus funds to other stages of financial needs thereby achieving better income smoothing during her/his different lifecycle stages.

6.3 The HCIS Plan Focuses On Demand-Side Support

As discussed in section 3, focusing on the supply-side support leads to low quality of services and a waste of resources. Such resources could be used more efficiently in helping poor people if demand-side approach were adopted.
The supply-side support is important for basic education during the young age dependency stage, whereas the demand-side support is more efficient during the pre-employment stage. Our study adopts the demand-side support.

6.4 The HCIS Plan Is An Insurance Plan Not A Welfare Program

Social security adopts two different approaches; welfare programs and insurance plans. The underlying notion of classification is whether the individual contributes to the program. Welfare programs assume that individuals do not pay contributions. These include many programs such as social assistance, social safety nets, social funds, and child labor reduction programs. Social insurance generally refers to systems in which workers themselves make contributions to fund the underlying programs. Such programs are designed to assist individuals, households, and communities to better manage certain contingencies related to incapacity for work due to illness, unemployment, old age, and death of the breadwinner, for example. Such contingencies are stated in ILO's 1952 International Labor Conference, at which the Social Security Minimum Standards Convention, (N° 102), was accepted (ILO, 2000).

Social insurance systems should ideally be designed as fully funded insurance schemes, without any welfare considerations. This is the basic idea of those who propose personal saving schemes as a reform option for retirement plan.

Insurance plans are contributory transfers that can be seen as a type of borrowing mechanism in which individuals pay for transfers backward or forward. A retirement pension plan is an example of a forward contributory transfers in which the insured person pays in advance for her/his retirement benefits. In the backward contributory transfer, an individual receives transfers and then later repays the loans, for examples student loans. Human capital contracts, future income loans, and collateral loans are recent forms of backward contributory transfers.

Our suggested HCIS plan can be considered a type of social security program. It is aims to support individuals during the pre-employment stage for high school and university education. It is also a type of insurance plan that adopts both backward and forward contributory transfers. The backward transfers are used to finance the human capital investment during the pre-employment stage, whereas the forward transfers are used to finance the retirement benefits during the retirement stage.

7. HUMAN CAPITALS – INCOME SECURITY (HCIS) PLAN: MECHANISM, ASSUMPTIONS AND MODEL DESCRIPTION

Pensions linked to contributions are currently the dominant strategy of old age income security policies and a move towards funded pension schemes is currently under discussion in many developed and developing countries, as a pension reform option (Palacios and Sluchynsky, 2006). The personal saving plan is a suggested alternative, in which each person pays forward her\his retirement benefits (Abdallah, 2004). This idea was the starting point for our suggested HCIS plan that helps an individual to manage their needs and surplus during different periods of their life cycle.

This section illustrates the mechanism of our suggested HCIS plan, then introduces assumptions, and develops the actuarial model.

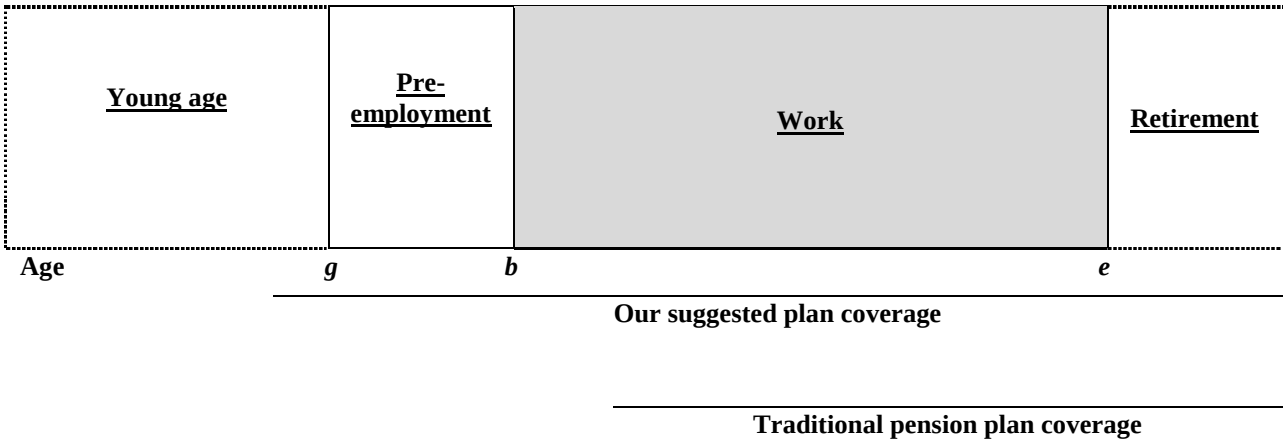
7.1 Human Capital - Income Security Plan (HCIS) Mechanism

Figure 5 represents lifetime income needs and surplus during different life cycles for a typical university graduate.

The above diagram shows that an individual starts with the young age dependency stage, which extends till age (g). During this stage an individual does not receive any transfers and s/he is mainly supported by her/his parents.

During the pre-employment stage, an individual receives educational transfers, which stop at age (b), when an individual starts the work stage and pays contributions to our suggested plan. Finally (e) represents the retirement age.

Figure (5): Lifetime income needs and surplus
General model



It can be seen that the young age dependency and retirement stages are all periods of financial need with no income generated. The work stage is the only period of life during which an individual generates income and incurs a surplus.

We will develop a generalized actuarial model in this section after introducing the underlying assumptions. The generalized Model allows implementing our suggested HCIS plan for different scenarios, i.e. to enable educational support during university study period only or to encompass the high school study period within the pre-employment stage.

7.2 Assumptions

The following assumptions are required for our suggested HCIS plan:

i. The human lifecycle is classified into four stages, namely, young age dependency, pre-employment, work, and retirement stages.

ii. A subscriber to our suggested HCIS plan is assumed to join and start receiving educational transfers at age *(g).*
The pre-employment stage extends to the age *(b)*, at which a subscriber starts to work. The value of *(b)* depends on the graduation age *(v)* and the unemployment period *(u)*. During an unemployment period, subscribers receive the same amount of transfers s/he receives during their education period.
$$b = v + u$$
$$v = g + \textit{educational support period}$$

iii. We consider retirement age to be *(e).*

iv. An employee's salary complies with a pre-specified salary scale (s_x) and the retirement benefits increased by annual increment scale (k_x).

v. During pre-employment stage a subscriber student should receives educational transfers during education years and extends to cover unemployment years. Such transfer is assumed to amount to a flat annual percentage (t %) of the starting salary. This assumption depends on the existence of a reliable expected salary scale for the future career path for the subscribing student in each field.

vi. In our plan, a subscriber individual can exit the work stage due to invalidity, death and reaching retirement age, according to given age related decrement rates; q_x^d, q_x^i and q_x^r consequently.

vii. Investing in human capital during the pre-employment stage results in an increase in individual's future lifetime income represented in a shift in the expected salary scale, this shift is referred to as (*m*%).

viii. Retirement benefits equals to *(p %)* of final salary (s_e).

ix. We assume the interest rate to be *(i %)*.

7.3 The Model

This section develops the generalized actuarial model for our suggested HCIS plan. The HCIS plan helps subscriber individuals to smooth lifetime income over their pre-employment, work, and retirement stages of their lifecycle. Each subscriber contributes a percentage of her/his lifetime income during the work stage. C_{TRAD} denotes to the required contribution rate for the traditional personal saving retirement plan and C_{HCIS} denotes to the required contribution rate for our suggested HCIS plan. The required model is developed through 6 steps:

7.3.1 Developing the main functions of the multiple-decrement table;

7.3.2 Computing the contingent present value of retirement benefits transfers that an individual receives during retirement stage (PVB);

7.3.3 Computing the contingent present value of contributions paid by workers during work stage (PBC);

7.3.4 Computing the contingent present value of educational transfers that a student receives during pre-employment stage (PVT);

7.3.5 Computing the required contribution rate for the traditional retirement pension plan (C_{TRAD});

7.3.6 Computing the required contribution rate for the suggested HCIS plan (C_{HCIS}).

7.3.1 Definitions, Notations and Basic Functions

This subsection presents definitions, notations, and functions needed to find the present values of contributions and benefits. Our model is a multiple-decrement model that is based on some probabilistic functions deduced from the multiple-decrement table with causes of decrement includes death, invalidity, and retirement. Our model uses the distribution of two random variables *T(x)* and *J(x)* and defines

$$ {}_z q_x^{(j)} = \Pr[T(x) \le z,\ J = j] \quad z \ge 0, \quad j = 1,2,3 \tag{1} $$

The symbol ${}_z q_x^{(j)}$ can be interpreted as the probability that a person aged x will exit within *t* years due to cause j. ${}_z q_x^{(j)}$ is the joint p.d.f. of $T(x)$ and $J(x)$.

If $z = 1$, that permits us to omit the prefix in the symbols defined above, and we have

$$ q_x^j = \Pr[a\ person\ aged\ x\ will\ die\ within\ 1\ year\ due\ to\ cause\ j] \tag{2} $$

Now consider the following notations**,**

l_x : number of active subscribers surviving at exact age *x*.

q_x^d : the probability of decrement during the next year due to death

d_x^d : number of exits due to death in the year of age *x* to *x* + 1.

$$d_x^d = l_x \; p_x^d \tag{3}$$

q_x^i : the probability of decrement during the next year due to invalidity

d_x^i : number of retirements due to invalidity in the year of age *x* to *x* + 1.

$$d_x^i = l_x \; p_x^i \tag{4}$$

q_x^r : the probability of decrement during the next year due to age retirement

d_x^r : number of retirements due to attainment of retirement age x.

$$d_x^r = l_x \; p_x^r \tag{5}$$

$$d_x^r = 0 \quad for \; x \neq 60,$$

and,
$$l_{x+1} = l_x - d_x^d - d_x^i - d_x^r \tag{6}$$

Let *v* denotes to the present value of one financial unit moved one year backward using a given discount rate (*i*), then

$$v = \frac{1}{1+i} \tag{7}$$

Therefore v^x is the present value of one financial unit received after *x* years.
Assuming that every surviving individual at age *x* during the work stage gets one financial unit as salary, and that there are l_x survivals, then the present value of LE l_x moved backward for x years (to the starting age of the life table) referred to as D_x, is given by:

$$D_x = v^x \times 1 \times l_x = v^x l_x \tag{8}$$

Since pension fund functions are mid-year funds, then $\overline{D}_x$ the mid-year present value, is defined as:

$$\overline{D}_x = \frac{1}{2}\left(D_x + D_{x+1}\right) \tag{9}$$

The sum of present value of all persons' (one financial unit) salaries over an age ranging from (*x*) to (*x* + *n*) and brought backward for *x* years (to the starting age of the life table), is given by:

$$_n\overline{N}_x = \sum_{t=0}^{n} \overline{D}_{x+t} \tag{10}$$

If we include the salary s_x in equation (9) and equation (10), we obtain $^s\overline{D}_x$ and $_y^sN_x$, as follows: $^s\overline{D}_x$ is the present value of the salaries of s_x received by all living persons at age *x*, brought to the starting age of the life table:

$$^s\overline{D}_x = s_x\overline{D}_x \tag{11}$$

$_y^sN_x$ is the sum of the present values of all salaries received by all living persons between age *x* and age *x* + *y*, brought to the starting age of the life table, and assuming a given salary scale s_x:

$$_y^s\overline{N}_x = \sum_{t=0}^{y} {}^s\overline{D}_{x+t} \tag{12}$$

7.3.2 Contingent Present Value Of Retirement Benefits (*PVB*)

Let *PVB* refer to the contingent present value of a retirement benefits of one financial unit per annum starts from the retirement age *(e)* and increase according to retirement benefit increment scale k_x, brought to age *x* (*x* <= *e*). *PVB* is given by the following formula.

$$PVB = v^{e-x}\ \frac{r_e}{l_x}\ \overline{a}_e^k \tag{13}$$

$\overline{a}^k{}_e$: is the contingent present value of whole life retirement benefit annuities of one financial unit that starts at age e and brought to age e.

$$\overline{a}_e^k = \frac{\overline{N}_e^k}{D_e} \tag{14}$$

and $$\overline{N}_e^k = \sum_{t=e}^{99} k_t\ \ \overline{D}_t$$

$\frac{r_e}{l_x}$: is the probability that a subscriber aged x attain retirement age (e) so s/he would be eligible to retirement benefits. Let $C_e^r = v^e\ r_e\ \overline{a}_e^k$

Then,

$$pvb = \frac{v^e \; r_e \; \bar{a}_e^k}{v^x l_x} = \frac{C_e^r}{D_x} \tag{15}$$

Since retirement benefits represent only a percentage p of the final salary at age *(e)*.

$$\beta(x) = p \; s_e \times PVB \tag{16}$$

$$\beta(x) = p \; s_e \; \frac{C_e^r}{D_x} \tag{17}$$

$$\beta(x) = p \; s_x \; \frac{s_e}{s_x} \frac{C_e^r}{D_x} \tag{18}$$

s_x is the salary at age x and $s_x = s_b \;\; for \; x \leq b.$ Let ${}^sC_e^r = s_e \; C_e^r$ and

$$\beta(x) = p \;\; s_x \; \frac{{}^sC_e^r}{{}^sD_x} \tag{19}$$

$\beta(x)$ Represents the contingent present value of annual whole life retirement benefits equals to p% of the final salary (s_e). The contingent retirement benefits starts upon retirement at age e, brought to age x.

7.3.3 Contingent Present Value Of Contributions (PVC)

The contingent present value of all salaries at age y brought to age x is

$$s_y v^{y-x} \frac{l_y}{l_x} \tag{20}$$

s_y refers to the salary scale at age y. On the assumption that individuals contribute *c*% of their salaries to the HCIS plan and these contributions are paid at mid years, then the contingent present value of contributions collected from covered persons at age y and brought to age *x* is *PVC*

$$PVC = c \; s_y \; v^{y-x} \frac{l_y}{l_x} \tag{21}$$

$$\therefore \qquad PVC = c \; s_y \; \frac{v^y}{v^x} \frac{l_y}{l_x} \tag{22}$$

Using the definition of D_x,

$$PVC = c \; s_y \; \frac{D_y}{D_x} \tag{23}$$

Taking the mid-year functions,

$$D_{y+0.5} = 0.5(D_y + D_{y+1}) = \overline{D}_y$$

$$PVC = c\, s_y \frac{\overline{D}_y}{D_x} \tag{24}$$

$$PVC = c\, s_x \frac{s_y}{s_x} \frac{\overline{D}_y}{D_x} \tag{25}$$

and based on the definition on the ${}^{s}\overline{D}_y$ (mentioned before), the required present value would be:

$$PVC = c\, s_x \frac{{}^{s}\overline{D}_y}{{}^{s}D_x} \tag{26}$$

Summing over y from the age of starting work (*b* years old) till the end of the work stage (*e* years old) gives the total value of the contributions paid by all working persons during the work stage and brought to age *x*:

$$\sum_{y=b}^{e-1} c\, s_x \frac{{}^{s}\overline{D}_y}{{}^{s}D_x} = c\, s_x \frac{\sum_{y=b}^{e-1} {}^{s}\overline{D}_y}{{}^{s}D_x} = c\, s_x \frac{{}^{s}\overline{N}_b}{{}^{s}D_x} \tag{27}$$

where ${}^{s}_{e-b-1}\overline{N}_b = \sum_{y=b}^{e-1} {}^{s}\overline{D}_y$.

s_x refers to the salary scale at age x, and $s_x = s_b \quad for\ x \le b$. Let $\tau(x)$ represents the contingent present value of contributions working individuals pay during work stage (from age b to age e), brought to age x.

$$\tau(x) = c\, s_x \frac{{}^{s}_{e-b-1}\overline{N}_y}{{}^{s}D_x} \tag{28}$$

and $s_x = s_b \ \ for\ x \le b$.

7.3.4 Present Value Of Educational Transfers During Pre-Employment Stage (*PVT*)

The present value of the transfers made to an individual when s/he is y years old, and brought to age x is computed as follows:

$$t\, s_b\, v^{y-x} \frac{l_y}{l_x} \quad \cdots y \ge x \tag{29}$$

Let ***(t %)*** denotes the percentage of the starting salary that is paid annually to individual as educational transfers, and assumes that an individual receives capabilities support transfers at mid-years. Thus *PVT* represent the contingent present value of the transfers paid to a covered person at age y and brought to age *x*.

$$PVT = t\, s_b\, v^{y-x} \frac{l_y}{l_x} \tag{30}$$

$$PVT = t\, s_b\, \frac{v^y}{v^x} \frac{l_y}{l_x} \tag{31}$$

Using the definition of D_x then,

$$PVT = t\, s_b\, \frac{D_y}{D_x} \tag{32}$$

Since $D_{y+0.5} = 0.5(D_y + D_{y+1}) = \overline{D}_y$, then, $PVT = t\, s_b \frac{\overline{D}_y}{D_x}$. Summing over y from age g (the starting age of receiving educational transfers) to the age b (the age of finishing pre-employment stage), and brought to age x (the comparison date). The contingent present value of educational transfers

$$\sum_{y=g}^{b-1} t\, s_b \frac{\overline{D}_y}{D_x} = t\, s_b \frac{\sum_{y=g}^{b-1} \overline{D}_y}{D_x} = t\, s_b \frac{{}_{b-g-1}\overline{N}_g}{D_x} \tag{33}$$

where, ${}_{b-g-1}\overline{N}_g = \sum_{y=g}^{b-1} \overline{D}_y$ and ${}_{b-g-1}\overline{N}_g$ refers to the sum of all transfers that are made to an individual during pre-employment stage, starting at age g and finishing at age b-1. Letting $T(x)$ represents the contingent present value of all transfers that are paid to living individuals during pre-employment stage starting at age g, and brought to age x assuming *t%* of the starting salary scale s_g as annual transfers.

$$T(x) = t\, s_b \frac{{}_{b-g-1}\overline{N}_g}{D_x} \tag{34}$$

Now, we will compute the required contribution rate for the current pension plan and the required contribution rate for the suggested HCIS plan.

7.3.5 The Required Contribution Rate For The Traditional Retirement Pension Plan

By equating the present values of total contributions and total benefits for the individual brought to age x, we get the required contribution rate for the traditional pension plan c_{Trad},

$$\tau(x) = \beta(x) \tag{35}$$

$$c_{Trad}\ s_x\ \frac{{}^{s}\overline{N}_{e-b-1}}{{}^{s}D_x} = p\ s_x\ \frac{{}^{s}C_e^r}{{}^{s}D_x} \tag{36}$$

$$c_{Trad} = p\ \frac{{}^{s}C_e^r}{{}^{s}\overline{N}_{e-b-1}} \tag{37}$$

For retirement benefits equal to p% of the final salary s_e, let $p = 1 - c_{Trad}$

$$c_{Trad} = (1-c_{Trad})\ \frac{{}^{s}C_e^r}{{}^{s}\overline{N}_{e-b-1}} \tag{38}$$

$$c_{Trad} = \frac{{}^{s}C_e^r}{{}^{s}\overline{N}_{e-b-1}} - c_{Trad}\left(\frac{{}^{s}C_e^r}{{}^{s}\overline{N}_{e-b-1}}\right) \tag{39}$$

$$c_{Trad} + c_{Trad}\left(\frac{{}^{s}C_e^r}{{}^{s}\overline{N}_{e-b-1}}\right) = \frac{{}^{s}C_e^r}{{}^{s}\overline{N}_{e-b-1}} \tag{40}$$

$$c_{Trad}\left(1+\frac{{}^{s}C_e^r}{{}^{s}\overline{N}_{e-b-1}}\right) = \frac{{}^{s}C_e^r}{{}^{s}\overline{N}_{e-b-1}} \tag{41}$$

$$c_{Trad}\left(\frac{{}^{s}\overline{N}_{e-b-1}+{}^{s}C_e^r}{{}^{s}\overline{N}_{e-b-1}}\right) = \frac{{}^{s}C_e^r}{{}^{s}\overline{N}_{e-b-1}} \tag{42}$$

$$c_{Trad} = \frac{{}^{s}C_e^r}{{}^{s}\overline{N}_{e-b-1}} \times \frac{{}^{s}\overline{N}_{e-b-1}}{{}^{s}\overline{N}_{e-b-1}+{}^{s}C_e^r} \tag{43}$$

$$c_{Trad} = \frac{{}^{s}C_e^r}{{}^{s}\overline{N}_{e-b-1}+{}^{s}C_e^r} \tag{44}$$

7.3.6 The Required Contribution Rate For The Suggested HCIS Plan

By equating the present values of total contributions and total benefits (educational and retirement transfers) for a covered person brought to age x, we get the required contribution rate. Also, to take expected shifts in salary scale into consideration, m% shift is assumed. This shift affects the annual salary s_x and the retirement benefits value which becomes p% of the shifted final salary. Moreover, shifting the salary scale does not affect the educational transfers.

$$\tau(x) = \beta(x) + T(x) \tag{45}$$

$$c_{HCIS}\ s_x\ \frac{(1+m)_{e-b-1}\,{}^{s}\overline{N}_b}{s_x D_x} = p(1+m)\ s_x\ \frac{{}^{s}C_e^r}{s_x D_x} + t\ s_b\ \frac{{}_{b-g-1}\overline{N}_g}{D_x} \tag{46}$$

$$c_{HCIS}\ \ (1+m)_{e-b-1}\,{}^{s}\overline{N}_b = p(1+m)\ \ {}^{s}C_e^r + t\ s_b\ \ {}_{b-g-1}\overline{N}_g \tag{47}$$

$$c_{HCIS} = p(1+m)\ \frac{{}^{s}C_{e}^{r}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}} + t\ s_{b}\ \frac{{}_{b-g-1}\overline{N}_{g}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}} \tag{48}$$

$$c_{HCIS} = \frac{p(1+m)\ {}^{s}C_{e}^{r} + t\ s_{b}\ {}_{b-g-1}\overline{N}_{g}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}} \tag{49}$$

$$c_{HCIS} = \frac{p(1+m)\ {}^{s}C_{e}^{r} + t\ s_{b}\ {}_{b-g-1}\overline{N}_{g}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}} \tag{50}$$

When there is no shifts in the salary scale then *m = zero*, and the required contribution rate becomes;

$$c_{HCIS} = \frac{p\ {}^{s}C_{e}^{r} + t\ s_{b}\ {}_{b-g-1}\overline{N}_{g}}{{}_{e-b-1}{}^{s}\overline{N}_{b}} \tag{50}$$

For retirement benefits equal to p% of the final salary, let $p = 1 - c_{HCIS}$

$$c_{HCIS} = (1-c_{HCIS})(1+m)\ \frac{{}^{s}C_{e}^{r}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}} + t\ s_{b}\ \frac{{}_{b-g-1}\overline{N}_{g}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}} \tag{51}$$

$$c_{HCIS} = \frac{(1+m)^{s}C_{e}^{r}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}} + t\ s_{b}\ \frac{{}_{b-g-1}\overline{N}_{g}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}} - c_{HCIS}\ \frac{(1+m)^{s}C_{e}^{r}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}} \tag{52}$$

$$c_{HCIS} + c_{HCIS}\ \frac{(1+m)^{s}C_{e}^{r}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}} = \frac{(1+m)^{s}C_{e}^{r}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}} + t\ s_{b}\ \frac{{}_{b-g-1}\overline{N}_{g}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}} \tag{53}$$

$$c_{HCIS}\left(1 + \frac{(1+m)^{s}C_{e}^{r}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}}\right) = \frac{(1+m)^{s}C_{e}^{r}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}} + t\ s_{b}\ \frac{{}_{b-g-1}\overline{N}_{g}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}} \tag{54}$$

$$c_{HCIS}\left(\frac{(1+m)_{e-b-1}{}^{s}\overline{N}_{b} + (1+m)^{s}C_{e}^{r}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}}\right) = \frac{(1+m)^{s}C_{e}^{r}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}} + t\ s_{b}\ \frac{{}_{b-g-1}\overline{N}_{g}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}} \tag{55}$$

$$c_{HCIS} = \left(\frac{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b} + (1+m)^{s}C_{e}^{r}}\right) \times \left(\frac{(1+m)^{s}C_{e}^{r}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}} + t\ s_{b}\ \frac{{}_{b-g-1}\overline{N}_{g}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b}}\right) \tag{56}$$

$$c_{HCIS} = \frac{(1+m)^{s}C_{e}^{r}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b} + (1+m)^{s}C_{e}^{r}} + t\ s_{b}\ \frac{{}_{b-g-1}\overline{N}_{g}}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b} + (1+m)^{s}C_{e}^{r}} \tag{57}$$

$$c_{HCIS} = \frac{(1+m)^{s}C_{e}^{r} + t\ s_{b}\left({}_{b-g-1}\overline{N}_{g}\right)}{(1+m)_{e-b-1}{}^{s}\overline{N}_{b} + (1+m)^{s}C_{e}^{r}} \tag{58}$$

Once again if there is no shifts in the salary scale then *m = zero*;

$$c_{HCIS} = \frac{{}^{s}C_{e}^{r} + t\ s_{b}\left({}_{b-g-1}\overline{N}_{g}\right)}{{}_{e-b-1}{}^{s}\overline{N}_{b} + {}^{s}C_{e}^{r}} \tag{59}$$

8. DATA

The general actuarial model presented in the previous section includes 5 age related functions and 9 parameter. Such model could be applied for two main alternative study path programs using suitable values for the involved parameters:

- Program One: 7 years support - University graduate model: this includes high school and university periods within the pre-employment stage.
- Program Two: 4 years support - University graduate model: this includes only the university period.

The following table shows the referred functions and parameters included in the model with its assumed values.

In dealing with mortality rates actuaries in the Social Insurance Fund for Government Employees (SIFGE) use the English Life Table A49/52 ULT to estimate the most suitable graduation for the Egyptian population using historical data from the fund. Periodic revisions are implemented to readjust the fitted rates, and the last revision was implemented in 2004 (SIFGE, 2009).

9. RESULTS OF THE SUGGESTED MODELS

In this section, the suggested models are tested through applying the male death rates, invalidity rates, and salary scales that are used in periodic actuarial pension plan evaluations by the Social Insurance Fund for Government Employees in Egypt.

According to the suggested HCIS plan, a preparatory school graduate will join high school at age 15 and start university education at age 18, get support during the pre-employment stage and get retirement benefits by age 60 years. Results are introduced for both the 7 years support program and the 4 years support program. Three different scenarios are examined for each educational program assuming three alternatives for unemployment period (zero, one, and two years). We show the required contribution rates for both the traditional and the suggested plans, after which we examine the required shift in salary scale that enables an individual to finance the HCIS program, without changing the disposable income received during the work stage and the retirement benefits received during the retirement stage.

Three important notes should be considered before going through these results. Firstly, the objective of both the traditional and our suggested income security plan is to smooth lifetime income during different covered life cycle stages. Secondly, an individual continues to receive

transfers during the unemployment period, and thirdly, the same salary scale is applied to both the traditional and suggested plan.

9.1 Program One: 7 Years- High School And University Support

Program One assumes a 7 years-high school and college support, and 60 years as retirement age. Table (2) shows results of Program One for three different unemployment scenarios assuming zero, one, and two years unemployment. The table also shows results for three salary scale shift scenarios assuming zero, 10%, and 10.68%. The 10.68% shift in the salary scale enables the subscriber person to finance the HCIS plan and keep the same disposable income during work and retirement he used to achieve under the traditional pension plan. The redistribution of lifetime income is shown as a percentage of the total lifetime income.

Table (1) Parameters' definitions and assumed values within the HCIS model

Serial	Paramete/ Function	Definition	Assumed values
1	g	The age of starting pre-employment stage	15 years old to get support during high school and university, and 18 years old to get support during university study only.
2	v	Graduation age	22 years old for university graduate
3	u	Unemployment period in years	Zero\ 1\ 2 years
4	b	The age of starting work $b = v + u$	22\23\24 for university graduate
5	e	Retirement age	60 years
6	p	% of the final salary paid as retirement benefits	It enhances smoothing mechanism
7	t	% of the starting salary paid as educational transfers	50%
8	m	Shift in salary scale	zero \ 10%\ the required shift
9	i	Interest rate	8.5%
10	s_x	Annual salary scale	$s_{x+1} = 1.10\, s_x$ *for* $b < x < e$
11	k_x	Retirement benefit annual increment scale	$k_{x+1} = 1.05\, k_x$ *for* $x > e$
12	q_x^d	Decrement rate at age x due to death	Based on the experience of the Social Insurance Fund for Government Employees in Egypt.
13	q_x^i	Decrement rate at age x due to invalidity	
14	q_x^r	Decrement rate at age x due to retirement	$q_x^r = 1$ *for* $x = e$

From column one in table (2) and under the traditional pension plan with no unemployment period and 60 years old as a retirement age, a university graduate is required to pay 24.02% of his lifetime income to save for his retirement benefits. This person achieves smoothing at

75.98%; meaning that he keeps 75.98% as disposable income. Moreover, he receives retirement benefits during old ages equals to 75.98% of his final salary that increases annually by 5%.
From column two, one year unemployment raises the required contribution rate to 24.42% and achieves smoothing at 75.58%. From column three, two years of unemployment raises the required contribution rate to 24.84% and achieves smoothing at 75.16%.

From column one and under Program One of our suggested HCIS plan with no unemployment, no salary scale shift, and 60 years retirement age assumption. A subscriber student receives 50% of his expected starting salary as educational transfers during high school and university and uses 10.68% of his lifetime income to pay back the educational transfers he has received, and uses 21.45% of his lifetime income to save for age retirement benefits amount to 67.87% of his final salary and increases by 5% annually.

Table (2): Result for Program One Lifetime income redistribution as percentage of lifetime income

Column's number		One	Two	Three	Four	Five
Unemployment period (u)		u = zero	u = 1	u = 2	u = zero	u = zero
Salary scale shift (m) For HCIS		m = zero	m = zero	m = zero	m = 10%	m = 10.68%
Traditional pension plan	c_{Trad}	24.02%	24.42%	24.84%		
	Disposable income $(1-c_{Trad})$	75.98%	75.58%	75.16%		
Suggested HCIS Contribution rate	Cost of retirement transfers	21.45%	21.19%	20.83%	20.96%	20.93%
	Cost of educational transfers	10.68%	13.22%	16.13%	10.43%	10.42%
	c_{HCIS}	32.13%	34.41%	36.96%	31.39%	31.35%
	Disposable income $(1-c_{HCIS})$	67.87%	65.59%	63.04%	68.61%	68.65%

A subscriber under Program One of the HCIS plan is required to pay 32.13% in total as a contribution rate to achieve smoothing at 67.87%. From column two, one year of unemployment under the HCIS plan raises the required contribution rate to 34.41% and achieves smoothing at 65.59%. Meanwhile, from column three under the HCIS plan, two years of unemployment raises the required contribution rate to 36.96% and achieves smoothing at 63.04%.

Better investment in human capital is expected to have its impact on individual's capabilities and his future income through shifting the salary scale. Therefore, column four in the above table shows the impact of a 10% shift in the salary scale on the required contribution rate for the suggested HCIS plan and the resulting smoothing level. With 10% shift in the salary scale under Program One, a subscriber person is required to use 31.39% of his shifted salary as a contribution rate and achieves smoothing at 68.61% of the shifted salary. Also, from column five, a 10.63% shift in the salary scale under Program One enables this subscriber to keep the same amount of money he used to receive under the traditional pension plan as disposable income during work and retirement stages. In this case the required contribution rate is 31.35% of the shifted salary and the smoothing level is 68.65% of the shifted salary. Table (3) shows the redistribution of an individual's salary in a given year represented in financial units assuming that this person receives 100 financial units for the referred year.

Table (3): Results for Program One

Distribution of annual income at age x when the annual salary = 100 financial units

Column's number		One	Two	Three	Four	Five
Unemployment period (u)		u = zero	u = 1	u = 2	u = zero	u = zero
Salary scale shift (m) For HCIS		m = zero	m = zero	m = zero	m = 10%	m = 10.68%
salary		100	100	100	110	110.68
Traditional pension plan	c_{Trad}	24.02	24.42	24.84		
	Disposable income $(1-c_{Trad})$	75.98	75.58	75.16		
Suggested HCIS Contribution rate	Cost of retirement transfers	21.45	21.19	20.83	23.06	23.17
	Cost of educational transfers	10.68	13.22	16.13	11.47	11.53
	c_{HCIS}	32.13	34.41	36.96	34.53	34.70
	Disposable income $(1-c_{HCIS})$	67.87	65.59	63.04	75.47	75.98

From column one in table (3) with no unemployment period, and 60 years old as retirement age, the disposable income under the traditional pension plan equals to 75.98 financial units, and the disposable income under the HCIS plan equals to 67.87 financial units. From column five, with 10.68% shift in the salary scale, the disposable income under the HCIS plan becomes 75.98 financial units (the same level of disposable income under the traditional pension plan).

The following graph shows the wealth accumulation curve according to Program One assuming no unemployment period for both traditional and suggested HCIS plans. For the 7 years high school and university support plan, the wealth curve starts with the age of joining high school and receiving education support transfers (age 15). The curve decreases until it reaches its minimum just before joining a paid job, at which point it increases until reaching its peak by the retirement age, after which the curve starts decreasing again. The graph also shows the annual

increasing salary scale during the work stage and the disposable income during the pre-employment, the work, and the retirement stage. Income increases during the work stage by 10% $(s_{x+1} = 1.10\, s_x)$ and retirement benefits increases annually by 5% $(k_{x+1} = 1.05\, k_x)$.

9.2 Program Two: 4 Years- University Support Model

This model assumes 4 years of university support program and a retirement age of 60 years. Table (4) show results for three different scenarios assuming zero, one, and two years of unemployment period. In addition to three different scenarios for the salary scale shift assuming zero, 10%, and 5.34%. The redistribution of lifetime income is shown as a percentage of the total lifetime income.

Figure (6): Salary scale, disposable income, and probabilistic wealth accumulation curve
Program One - No unemployment – 60 yrs retirement age

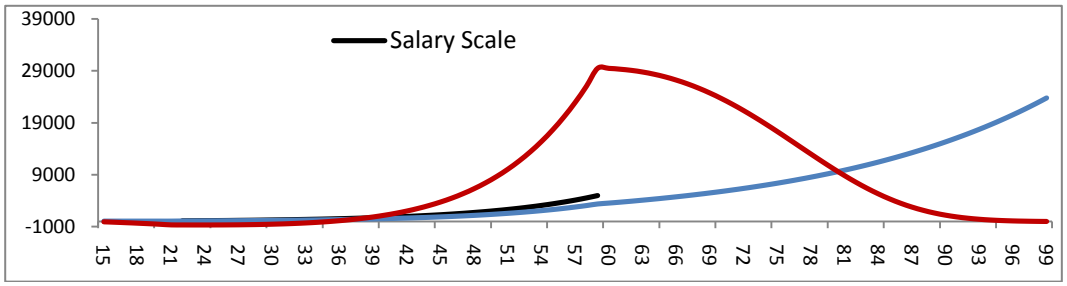
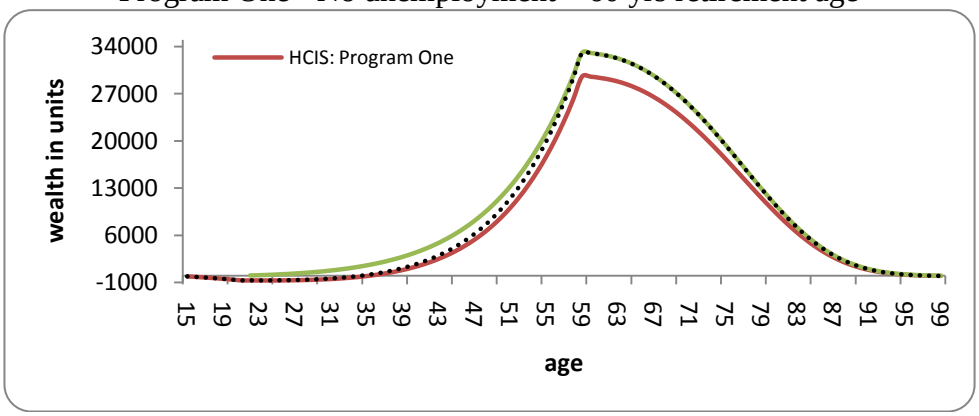


Figure (7) shows the required shift (10.68%) in the salary scale which enables an individual to finance education expenses and keep the same level of disposable income he used to receive during work and retirement stages under the traditional plan.

Figure (7): The probabilistic wealth accumulation curve and the required shift in salary scale
Program One - No unemployment – 60 yrs retirement age



From column one in table (4), a subscriber under Program One of the HCIS plan is required to pay 28.07% in total as a contribution rate to achieve smoothing at 71.93%.
From column two, one year of unemployment under the HCIS plan raises the required contribution rate to 29.88% and achieves smoothing at 70.12%. Meanwhile, from column three under the HCIS plan, two years of unemployment raises the required contribution rate to 31.89% and achieves smoothing at 68.11%.

Better investment in human capital is expected to have its impact on individual's capabilities and his future income through shifting the salary scale. Therefore, column four in the above table shows the impact of a 10% shift in the salary scale on the required contribution rate for the suggested HCIS plan and the resulting smoothing level. With 10% shift in the salary scale under Program Two, a subscriber person is required to use 27.70% of his shifted salary as a contribution rate and achieves smoothing at 72.30% of the shifted salary. Also, from column five, a 5.34% shift in the salary scale under Program Two enables this subscriber to keep the same amount of money he used to receive under the traditional pension plan as disposable income during work and retirement stages. In this case the required contribution rate is 27.87% of the shifted salary and the smoothing level is 72.13% of the shifted salary. Table (5) shows the redistribution of an individual's salary in a given year represented in financial units assuming that this person receives 100 financial units for the referred year.

Table (4): Result for Program Two
Lifetime income redistribution as percentage of lifetime income

Column's number		One	Two	Three	Four	Five
Unemployment period (u)		u = zero	u = 1	u = 2	u = zero	u = zero
Salary scale shift (m) For HCIS		m = zero	m = zero	m = zero	m = 10%	m = 5.34%
Traditional pension plan	c_{Trad}	24.02%	24.42%	24.84%		
	Disposable income $(1-c_{Trad})$	75.98%	75.58%	75.16%		
Suggested HCIS Contribution rate	Cost of retirement transfers	22.73%	22.66%	22.51%	22.43%	22.57%
	Cost of educational transfers	5.34%	7.22%	9.38%	5.27%	5.30%
	c_{HCIS}	28.07%	29.88%	31.89%	27.70%	27.87%
	Disposable income $(1-c_{HCIS})$	71.93%	70.12%	68.11%	72.30%	72.13%

From column one in table (5) with no unemployment period, and 60 years old as retirement age, the disposable income under the traditional pension plan equals to 75.98 financial units, and the disposable income under the HCIS plan equals to 71.93 financial units. From column five, with 5.34% shift in the salary scale, the disposable income under the HCIS plan becomes 75.98 financial units (the same level of disposable income under the traditional pension plan).

The following graph shows the wealth accumulation curve according to Program Two assuming no unemployment period for both traditional and suggested HCIS plans. For the 4 years university support program, the wealth curve starts with the age of joining high school and receiving education support transfers (age 18). The curve decreases until it reaches its minimum just before joining a paid job, at which point it increases until reaching its peak by the retirement age, after which the curve starts decreasing again. The graph also shows the annual increasing salary scale during the work stage and the disposable income during the pre-employment, the work, and the retirement stage. Income increases during the work stage by 10% ($s_{x+1} = 1.10\, s_x$) and retirement benefits increases annually by 5% ($k_{x+1} = 1.05\, k_x$).

Table (5): Results for Program Two

Distribution of annual income at age x when the annual salary = 100 financial units

Column's number		One	Two	Three	Four	Five
Unemployment period (u)		u = zero	u = 1	u = 2	u = zero	u = zero
Salary scale shift (m) For HCIS		m = zero	m = zero	m = zero	m = 10%	m = 5.34%
salary		100	100	100	110	105.34
Traditional pension plan	c_{Trad}	24.02	24.42	24.84		
	Disposable income $(1-c_{Trad})$	75.98	75.58	75.16		
Suggested HCIS Contribution rate	Cost of retirement transfers	22.73	22.66	22.51	24.67	23.78
	Cost of educational transfers	5.34	7.22	9.38	5.80	5.58
	c_{HCIS}	28.07	29.88	31.89	30.47	29.36
	Disposable income $(1-c_{HCIS})$	71.93	70.12	68.11	79.53	75.98

Figure (8): Salary scale, disposable income, and probabilistic wealth accumulation curve
Program Two - No unemployment – 60 yrs retirement age

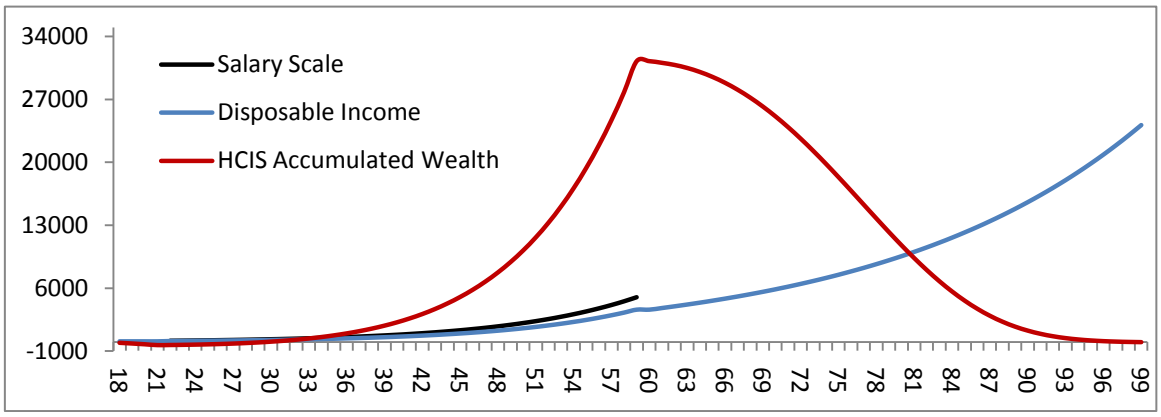
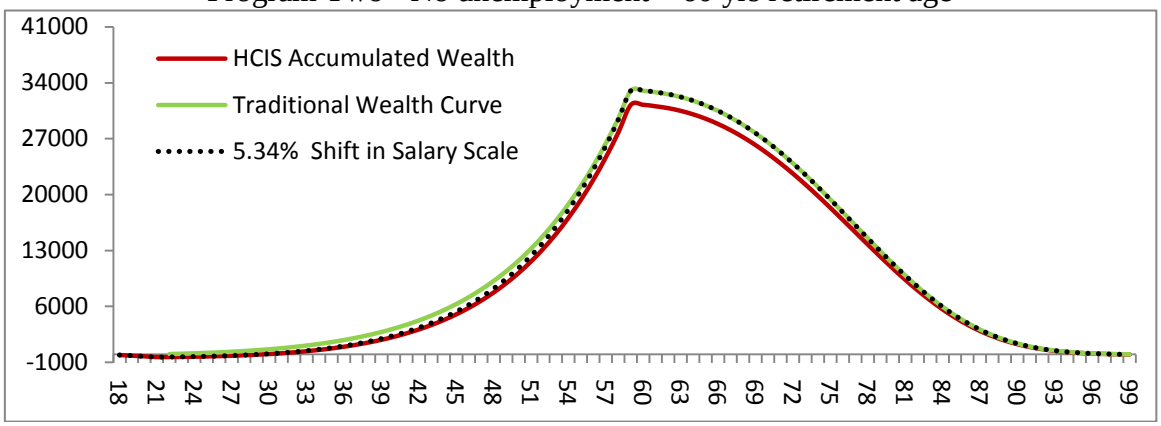


Figure (9) shows the required shift (5.34%) in the salary scale which enables an individual to finance education expenses and keep the same level of disposable income he used to receive during work and retirement stages under the traditional plan.

Figure (9): The probabilistic wealth accumulation curve and the required shift in salary scale
Program Two - No unemployment – 60 yrs retirement age



9.3 Conclusions and Recommendations

Over the past few decades, local governments and donor communities have exerted serious efforts to find ways in which poverty can be effectively reduced; however, poverty remains an unresolved global issue. Our study adopts the ideas of Schultz and Becker concerning the importance of human capital investment as a highly efficient tool for human well being, and introduces a new approach concerning financing human capital investment through education. This approach concludes the following:

- Reclassification of the human life cycle to four stages, namely childhood, preemployment, work, and retirement.
- Education expenses in our plan are considered as resources for investment rather than consumption as they encumber the current wealth with the purpose of increasing future wealth.
- Moving from the pay-as-you-go mechanism to a self-aid plan concerning financing human capital investment in education.

Therefore, we suggest a Human-Capital Income-Security plan (HCIS) that helps an individual to use lifetime income earned during the work stage to achieve income security during different stages of their lifecycle, based on the self-aid approach.

Testing the suggested plan using death and invalidity rates from the Social Insurance Fund in Egypt shows that an individual should pay 28.07% of his lifetime income during the work stage, and receive 50% of his starting salary during 4 years study at university, as well as receiving 71.93% of his final salary during retirement that increases by 5% annually.

Moreover, if this person wants to finance education expenses incurred during 3 years of high school in addition to the 4 years university study, he should pay 32.13% of his lifetime income during the work stage and receive 67.87% of his final salary as retirement benefits.

If a person agrees to pay 28.07% of lifetime income for 4 university years support, or 32.13% for 7 years high school and university support, this will be more convenient (for the parent) if we assume that children are no longer part of financial responsibility of their parents during their high school and/or university period, and we suggest as they pay for their own education expenses according to the self-help aid plan.

On the national scale, our HCIS plan aims at shifting at least part of the higher educational cost burden from government (taxpayers) to individuals, and minimizes the degree of dependency on parental altruism concerning human capital investment in education. It also enables children from poor families to pay for their education expenses during pre-employment stage. Thus the HCIS plan helps in breaking the poverty cycle by reducing the possibility of falling into poverty in the future stages of their lifecycle.

Participating in the proposed plan is suggested to be optional and selective in a way that encourages free competition among students, and education service providers. This may increase the education system efficiency and enable students from poor families to finance their capital investment in education, conditional upon evidence of satisfactory progress; academic records and/or follow up reports. With some accurate actuarial calculations, the suggested plan proposes an alternative approach which we hope will to change the current way of thinking concerning education financing, poverty fighting and human capabilities support.

REFERENCES

Abdallah A., (2004), *"Aging and pension plan in Egypt"*, Ms thesis, Faculty of Economics and Political Science, Cairo University. Egypt.

Ali-Eldin K., (2003), *"Understanding and combating poverty: A quest for conceptualization, measurement indicators, and empirical methodologies*", ERF, Working paper series, Working paper 0338.

Becker G., (1998), *"Human capital and poverty"*, Acton Institute series, Religion and Liberty vol 8 no 1, Italy.

Becker, G. S. (1964), *"Human Capital – A Theoretical and empirical analysis with special reference to education*", Chicago: The University of Chicago Press.

Becker, G. S. (1997), *"Why a crash would not cripple the economy"*, Hoover Digest No. 4.

Becker. G. S. (1993), *"Nobel lecture: The economic way of looking at behavior"*, Journal of Political Economy, 101, 385-409.

Bowers N., Gerber H., Hickman J., Jones D., and C. Nesbitt, (1986), *"Actuarial Mathematics"*, The Society of Actuaries, London, England.

Carver B. A, (2004), *"Income collateralized loans: market and policy exploration"*, PhD dissertation, Stanford university.

Chapman, Bruce (2005), *"Income contingent loans for higher education: International reform"* Center for Economic Policy Research, Discussion paper No. 491, ISBN 0731535618, Australian National University, Australia.

Cho, Y., (2005), *"Investment in Children's Human Capital: Implication of PROGRESA"*, Korea Development Institute, Seoul, Korea.

DeMuth, C. (1997), *"The new wealth of nations"*, Commentary, 104.

Eitelbreg C., (1999), *"Public Pension Design and Responses to a Changing Workforce"*, Pension Research Council, University of Pennsylvania, Philadelphia, USA.

El-Mahdi, A. and Abdallah, A. (2007), *"Gender and rights in the informal economy of Egypt"*, paper presented at the CAWTAR-ILO Conference on Gender Rights in the Informal Economies of Arab states, Tunisia.

Friedman M. (1962), *"Capitalism and Freedom"*, Chicago: The University of Chicago Press 19962. In (Carver, 2004).

Holzmann R. (1999), *"The World Bank Approach to Pension Reform"*, Social Protection Discussion Paper Series No. 9807, Social Protection Unit, Human Development Network.

ILO (2000), "*Learning from experience: A gendered approach to social protection for workers in the informal economy"*, ISBN 92-2-112107-0.

Lee R., and Edward R. (2001), *"The Fiscal Impact of Population Change"*, Federal Reserve Bank of Boston Conference Series No 46 pp. 270-237, Boston, USA.

Lee R., Mason A., and Miller T., (2000), *"Life Cycle Saving and the Demographic Transition in East Asia"*, Stanford University Press PP. 155-184, Stanford, USA.

Lee, R., and Yamagata H. (2000), *"Sustainable Social Security: What Would It Cost?"* Working paper, Center for Economics and Demography of Aging, Berkeley, USA.

Lee. J. B. (1999), *"How do students and families pay for college?"*, In (Pekins, 2003).

Loewe, M. (2000), *"Social Security in Egypt An Analysis and Agenda for Policy Reform"*, Economic Research Forum, Working paper 2024, Cairo.

Mitchell O. S. (1998), *"International Models for Pension Reform"*, Pension Research Council, University of Pennsylvania, Philadelphia, USA.

Mitchell O., and Bodie, Z. (2000), *"A Framework for Analyzing and Managing Retirement Risks"*, Pension Research Council, University of Pennsylvania, Philadelphia, USA.

Modigliani F. , Ceprini M., and Arun, S. (1999) *"A solution to the social security crisis"*, Center for studies in economic and finance CSEF, MIT press, Working Paper 4051 - November 1999 (Third Revision)

Modigliani F., and Jappeli, T. (2003), *"The age-saving profile and the lifecycle hypothesis"*, Center for studies in economic and finance CSEF, the collected papers of Franco Modigliani, Volume 6, MIT Press, Working paper No. 9.

Mursa C. G. (1981), *"Theodore W. Schultz, Investing in people"*, University of Chicago Press, USA.

Neill A., (1989), *"Life Contingencies"*, The Institute of Actuaries, Scotland.

Palacios R., and Sluchynsky O. March (2006), *"The role of social pensions"*, Pension Reform Primer.

Palacios, Miguel (2003), *"Financing human capital: A capital market approach to higher education funding manuscript*", The Batten Institute.

Palacios R. (2002), *"Human capital contracts: Equity like instruments for financing higher education"*, Policy analysis December 16, 2002, No. 462.

Perkins A. (2003), *"student loan debt: insights into economic well-being issues"*, PhD dissertation, Iowa university.

Psaharopoulos, G., and Patrinos, H., (2002), "*Returns to investment in education: A further update"*, World Bank policy research working paper 2881.

Rank M. & Hirshl T. (2001), *"The occurrence of poverty across the life cycle: Evidence from the PSID"*, Journal of policy analysis and management, vol. 20, No. 4, 737-755 (2001).

Rawling L., (2004), *"A new approach to social assistance: Latin America's experience with conditional cash transfer programs"*, Social Protection Unit, Human Development Network, World Bank.

Rofman R. P. (1993), *"Social security and income redistribution"*, PhD dissertation, Berkeley, USA.

Shinichi N. (2000), *"Altruism, lifetime uncertainty and intergenerational transfers"*, PhD dissertation, University of Pennsylvania, USA.

Takahashi k., and Otsuka K., (2007), "*Human Capital Investment and Poverty Reduction over Generations*", Institute of Developing Economies, Discussion paper No. 96

Unni J. and Rani U., (2003), *"Regional Overview of Social Protection of Informal Workers in Asia: Insecurities, Instruments and Institutional, Arrangements"*, Gender & Development Discussion Paper Series No. 14, United Nations Economic and Social Commission for Asia and Pacific, Bangkok.

Vandenberghe V. & Debande O. (2004), *"Financing higher education with student loans: the crucial role of income-contingency and risk pooling"*, IRES, University Catholiques de Louvain, Belgium.

Wirt, J., Choy, S.P., Gerald, D., Provasnik, S., Rooney, P., Watanabe, S., & Tobin R. (2002), *"The condition of Education 2002"*, National Center for Education Statistics, Statistical analysis report. NCES 20020-025. Washington D.C.: Office of educational research and improvement, US. Department of Education. In (Perkins, 2003).

World Bank, a, (2001), "*Social Protection Sector Strategy: from safety net to springboard"*, Washington, D.C., USA.

World Bank, b, (2001), "*World Development Report: Attacking poverty* ", Chapter 2: Causes of poverty and framework for action, New York: Oxford University Press.

World Bank, (2002), *"A user's guide to poverty and social impact analysis"*, Poverty reduction group and social development department.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 1030–1037.

MOMENTS OF ORDER STATISTICS OF A GENERALIZED BETA AND ARCSINE DISTRIBUTIONS WITH SOME SPECIAL CASES

Kamal Samy Selim [1] and Wafik Youssef Younan [2]
[1] Department of Computational Social Sciences
Faculty of Economics and Political Science
Cairo University
Cairo, Egypt
E-mail: kselim9@yahoo.com
[2] Department of Economics
The American University in Cairo
Cairo, Egypt
E-mail: wyounan@aucegypt.edu

ABSTRACT

In this paper, we derive explicit forms of the moments of order statistics arising from a generalized form of the two-parameter beta distribution with two shape parameters p and q as positive values not necessarily integers. The generalized form considered is a linear function in the two-parameter beta random variable X, namely, $Y = a + bX$, where $-\infty < a < \infty$ and $b > 0$. If the distribution of Y is denoted by beta (p, q, a, b), the moments of order statistics of some special cases can be easily obtained according to the values of the four parameters p, q, a and b. These special cases are the two-parameter beta distribution, a general form of the power distribution, the uniform distribution, generalized arcsine distribution, family of all arcsine distributions and the standard arcsine distribution. For these special cases, short tables are given for the moments of all order statistics in random samples of size 1,2,3,4, and 5.

Keywords: Beta distribution; Arcsine distribution; Moments of order statistics.

1. INTRODUCTION

The computations of single and product moments of order statistics arising from the beta distribution have been discussed in a number of recent articles. For example, Nadarajah (2008) has derived explicit closed form expressions for moments of order statistics from the normal, log normal, gamma and beta distributions. The expressions take the form of finite sums of well known special functions namely, the Lauricella function of type A and the generalized Kampe de Feriet function. Thomas and Samuel (2008) have obtained certain recurrence relations for the single and product moments of order statistics of a random sample of size n arising from a beta distribution when the two shape parameters are positive integers. Abdelkader (2008) has discussed the case of moments of order statistics of independent non-identically beta random variables. The independent identical case has been also considered in the same paper if the two shape parameters are positive integers.

In this article, we derive explicit forms for all single moments of order statistics arising from a generalized form of the two-parameter beta distribution with two shape parameters p and q as positive values not necessarily integers. The generalized form considered is a linear function in the two-parameter beta random variable X, namely, $Y = a + bX,$ where $-\infty < a < \infty$ and $b > 0.$ If the distribution of Y is denoted by Beta $(p,q,a,b),$ the moments of order statistics of some special cases can be directly obtained. These special cases are: (1) The two-parameter beta distribution Beta $(p,q,0,1)$, (2) A general form of the power distribution Beta $(p,1,a,b),$ (3) The uniform distribution Beta $(1,1,a,b-a),$ (4) Generalized arcsine distribution Beta $(p,1-p,0,1)$ as indicated by Feller (1971), (5) Family of all arcsine distributions Beta $(0.5,0.5,a,b)$ as defined by Balakrishnan and Nevzorov (2003) and (6) The standard arcsine distribution Beta $(0.5,0.5,0,1)$ as given in Balakrishnan and Nevzorov (2003) too.

The linear transformation with the two additional parameters a and b enhances the role of order statistics in fitting models in many real life applications. The computed moments of order statistics are crucial in robustness properties of modeling and inferential procedures

The paper is outlined in the following two sections: Section 2 derives an explicit expression for the k^{th} moment of the r^{th} order statistic $\mu_{r:n}^{k}$ when $Y_1, Y_2, ..., Y_n$ is a random sample of size n from a generalized beta distribution $Y = a + bX,$ where $-\infty < a < \infty$ and $b > 0$ and X is a two-parameter beta random variable. Section 3 gives short tables for the computed first and second moments of order statistics for the random variable Y and for the six special cases at chosen values of the four parameters p, q, a and b.

2. MOMENTS OF ORDER STATISTICS OF A GENERALIZED BETA DISTRIBUTION

Let X be a random variable having a two-parameter beta distribution

$$f(x) = \frac{1}{\beta(p,q)} x^{p-1}(1-x)^{q-1}, \qquad 0 < x < 1, \tag{1}$$

where p and q are positive values not necessarily integers. And let

$$Y = a + bX, \qquad a < y < a + b, \tag{2}$$

where $-\infty < a < \infty$ and $b > 0.$ If $Y_{(1)} < Y_{(2)} < ... < Y_{(n)}$ are the order statistics of a random sample of n observations from the distribution of the random variable Y, then the k^{th} moment of the r^{th} order statistic denoted by $\mu_{r:n}^{k}$ - is given by the following proposition.

Proposition

Let $Y_{(1)} < Y_{(2)} < ... < Y_{(n)}$ be the sample order statistics from the distribution of the random variable descried in (2), then the k^{th} moment of the r^{th} order statistic $Y_{(r)}$ is given by

$$\mu_{r:n}^{k} = C\sum_{t=0}^{n-r}\binom{n-r}{t}\left[\frac{-\Gamma(p+q)}{\Gamma(p)}\right]^{t}\sum_{m=0}^{k}\binom{k}{m}\left(\frac{a}{b}\right)^{m}\sum_{i=0}^{\infty}c_i\cdot\beta\left[k-m+i+p(t+r),q\right], \tag{3}$$

where

$$C = \frac{r\cdot b^{k}}{\Gamma(q)}\binom{n}{r}\left[\frac{\Gamma(p+q)}{\Gamma(p)}\right]^{r}, \tag{4}$$

$$c_0 = \left[\frac{1}{p\,\Gamma(p)}\right]^{t+r-1}, \tag{5}$$

$$c_i = p\Gamma(q)\cdot\sum_{j=1}^{i}\left[\frac{(t+r)j}{i}-1\right](a_j)(c_{i-j}) \qquad i\geq 1, \tag{6}$$

and

$$a_j = \frac{(-1)^{j}}{\Gamma(q-j)\cdot(p+j)\cdot j!}. \tag{7}$$

A detailed sketch for the proof of the above assertion is given in the Appendix. If $a=0$ and $b=1$ in (3), we get the k^{th} moment of the r^{th} order statistic of the beta distribution as follows

$$\mu_{r:n}^{k} = \frac{r}{\Gamma(q)}\binom{n}{r}\left[\frac{\Gamma(p+q)}{\Gamma(p)}\right]^{r}\sum_{t=0}^{n-r}\binom{n-r}{t}\left[\frac{-\Gamma(p+q)}{\Gamma(p)}\right]^{t}\sum_{i=0}^{\infty}c_i\cdot\beta\left[k+i+p(t+r),q\right], \tag{8}$$

where c_0, c_i and a_j are as defined in (5) – (7).

It is worthy to mention that the moment given in (8) is a different form of the same moment obtained by Nadarajah (2008) through an alternative approach. However, in his article, no computations have been indicated.

3. COMPUTATIONS OF SOME MOMENTS OF ORDER STATISTICS

In this section, we use the explicit expressions given in equations (3) – (7) to compute the first and second single moments of order statistics for the random variable Y and the special cases at the chosen values of the four parameters p, q, a and b. Table 1 summarizes the different special cases that could be derived from the general form. The last two columns of the table describe the selected specific instances of the indicated distributions considered for computation and the table numbers of the corresponding moments that follow. The special case of the two-parameter beta distribution is considered two times (Tables 3 and 4). In Table 3, the parameters p and q are assumed to be positive values, while they are restricted to positive integers in Table 4.

An Intel core 2 duo processor (2.4 GH) and Fortran 90 compiler are used for computations. The main difficulty in programming the proposed form for moments is due to the computation of the gamma function for values higher than 25. For this purpose, a module described in Zhang and Jin (1996) is adopted. The infinite series are computationally converged to a tolerance level less than 0.00001 in a reasonable number of steps (less than 25 terms in most of the cases).

For general checks on computations of the moments, the following known relations are applied

$$\sum_{r=1}^{n} \mu_{r:n} = n\,\mu \quad \text{and} \sum_{r=1}^{n} E(X_{r:n}^2) = n\,E(X^2)$$

where μ is the mean of the parent distribution (David, 1981, pp. 38). In other words, for any given sample size *n*, the average of first moments of all order statistics must equal to the first moment of the parent distribution, that is, the moment at $n = 1$ and $r = 1$. A similar check has been made for the second moments.

Table 1: Generalized and Special Cases Considered, and the Instance of Computations for Each*

Distribution	Restrictions to Parameters				Instance of Computation	Table Number
	p	q	a	b		
A Generalized Beta	$p \varepsilon R^+$	$q \varepsilon R^+$	$a \varepsilon R$	$b \varepsilon R^+$	(2.5, 1.5, 2, 3)	2
Two-Parameter Beta	$p \varepsilon R^+$	$q \varepsilon R^+$	$a = 0$	$b = 1$	(2.5, 1.5, 0, 1)	3
Two-Parameter Beta	$p \varepsilon N$	$q \varepsilon N$	$a = 0$	$b = 1$	(3, 2, 0, 1)	4
A General Form of Power	$p \varepsilon R^+$	$q = 1$	$a \varepsilon R$	$b \varepsilon R^+$	(2.5, 1, 0, 1)	5
Uniform	$p = 1$	$q = 1$	$a \varepsilon R$	$b > a$	(1, 1, 0, 1)	6
Generalized Arcsine	$0 < p < 1$	$q = 1- p$	$a \varepsilon R$	$b \varepsilon R^+$	(0.25, 0.75, 0, 1)	7
Family of Arcsine	$p =0.5$	$q = 0.5$	$a \varepsilon R$	$b \varepsilon R^+$	(0.5, 0.5, 2, 3)	8
Standard Arcsine	$p = 0.5$	$q = 0.5$	$a = 0$	$b = 1$	(0.5, 0.5, 0, 1)	9

* R is the set of all real values, R^+ is the set of positive values and N is the set of positive integers.

Table 2: Moments of Order Statistics of a Generalized Beta Distribution
Beta (2.5,1.5,2,3)

r / n	First moment					Second moment				
	1	2	3	4	5	1	2	3	4	5
1	3.8750					15.4375				
2	3.5039	4.2461				12.6174	18.2576			
3	3.3022	3.9073	4.4155			11.1815	15.4891	19.6418		
4	3.1701	3.6983	4.1163	4.5153		10.2832	13.8765	17.1018	20.4885	
5	3.0747	3.5516	3.9184	4.2482	4.5820	9.6561	12.7914	15.5041	18.1669	21.0689

Table 3: Moments of Order Statistics of a Two-Parameter Beta Distribution
Beta (2.5,1.5,0,1) (The Two Parameters are not Positive Integers)

r / n	First moment					Second moment				
	1	2	3	4	5	1	2	3	4	5
1	0.6250					0.4375				
2	0.5013	0.7487				0.2891	0.5859			
3	0.4341	0.6358	0.8052			0.2192	0.4289	0.6644		
4	0.3900	0.5661	0.7054	0.8384		0.1781	0.3426	0.5152	0.7142	
5	0.3582	0.5172	0.6395	0.7494	0.8607	0.1508	0.2872	0.4256	0.5749	0.7490

Table 4: Moments of Order Statistics of a Two-Parameter Beta Distribution
Beta (3,2,0,1) (The Two Parameters are Positive Integers)

r / n	First moment					Second moment				
	1	2	3	4	5	1	2	3	4	5
1	0.6000					0.4000				
2	0.4857	0.7143				0.2667	0.5333			
3	0.4250	0.6072	0.7678			0.2054	0.3892	0.6054		
4	0.3855	0.5434	0.6709	0.8001		0.1695	0.3131	0.4653	0.6521	
5	0.3570	0.4993	0.6096	0.7118	0.8222	0.1456	0.2650	0.3853	0.5186	0.6855

Table 5: Moments of Order Statistics of the Power Distribution
Beta (2.5,1,0,1)

r / n	First moment					Second moment				
	1	2	3	4	5	1	2	3	4	5
1	0.7143					0.5556				
2	0.5952	0.8333				0.3968	0.7143			
3	0.5252	0.7353	0.8824			0.3133	0.5639	0.7895		
4	0.4775	0.6684	0.8021	0.9091		0.2611	0.4699	0.6579	0.8333	
5	0.4421	0.6189	0.7427	0.8418	0.9259	0.2251	0.4051	0.5672	0.7184	0.8621

Table 6: Moments of Order Statistics of the Uniform Distribution. Beta (1,1,0,1)

r / n	First moment					Second moment				
	1	2	3	4	5	1	2	3	4	5
1	0.5000					0.3333				
2	0.3333	0.6667				0.1667	0.5000			
3	0.2500	0.5000	0.7500			0.1000	0.3000	0.6000		
4	0.2000	0.4000	0.6000	0.8000		0.0667	0.2000	0.4000	0.6667	
5	0.1667	0.3333	0.5000	0.6667	0.8333	0.0476	0.1429	0.2857	0.4762	0.7143

Table 7: Moments of Order Statistics of Generalized Arcsine Distribution. Beta $(0.25, 0.75, 0, 1)$

r / n	First moment					Second moment				
	1	2	3	4	5	1	2	3	4	5
1	0.2500					0.1563				
2	0.0910	0.4090				0.0371	0.2754			
3	0.0406	0.1918	0.5176			0.0111	0.0890	0.3687		
4	0.0208	0.1001	0.2835	0.5956		0.0039	0.0325	0.1454	0.4431	
5	0.0117	0.0571	0.1646	0.3628	0.6538	0.0016	0.0134	0.0613	0.2015	0.5035

Table 8: Moments of Order Statistics of Family of All Arcsine Distributions. Beta $(0.5, 0.5, 2, 3)$

r / n	First moment					Second moment				
	1	2	3	4	5	1	2	3	4	5
1	3.5000					13.3750				
2	2.9152	4.0848				9.2810	17.4690			
3	2.5894	3.5669	4.3437			7.1718	13.4995	19.4538		
4	2.4150	3.1128	4.0210	4.4513		6.1217	10.3220	16.6769	20.3794	
5	2.3082	2.8422	3.5187	4.3559	4.4751	5.5154	8.5467	12.9851	19.1380	20.6898

Table 9: Moments of Order Statistics of the Standard Arcsine Distribution. Beta $(0.5, 0.5, 0, 1)$

r / n	First moment					Second moment				
	1	2	3	4	5	1	2	3	4	5
1	0.5000					0.3750				
2	0.3051	0.6949				0.1800	0.5700			
3	0.1965	0.5223	0.7812			0.0905	0.3591	0.6754		
4	0.1383	0.3709	0.6737	0.8171		0.0513	0.2079	0.5103	0.7305	
5	0.1027	0.2807	0.5062	0.7853	0.8250	0.0314	0.1309	0.3234	0.6349	0.7544

APPENDIX

Proof of the proposition

If $\varphi(y)$ and $\Phi(y)$ are the density function and the distribution function respectively of the random variable Y, then

$$\varphi(y) = \frac{1}{b\beta(p,q)}\left(\frac{y-a}{b}\right)^{p-1}\left[1-\left(\frac{y-a}{b}\right)\right]^{q-1}, \qquad a < y < a+b,$$

and

$$\Phi(y) = F\left(\frac{y-a}{b}\right), \qquad a < y < a+b,$$

where the function F is the distribution function of the random variable X expressed in the following form (Andrews, 1985, pp. 70)

$$F(x)=\frac{1}{\beta(p,q)}x^{p}\Gamma(q)\sum_{i=0}^{\infty}\frac{(-1)^{i}x^{i}}{\Gamma(q-i)\cdot(p+i)\cdot i!},\qquad 0<x<1.$$

The k^{th} moment of the r^{th} order statistic Y_r could be then written as

$$\mu_r^k=r\binom{n}{r}\int_a^{a+b}y_r^k\left[\Phi(y_r)\right]^{r-1}\left[1-\Phi(y_r)\right]^{n-r}\varphi(y_r)\,dy_r,$$

$$\mu_r^k=r\binom{n}{r}\sum_{t=0}^{n-r}\binom{n-r}{t}(-1)^t\int_a^{a+b}y_r^k\left[\frac{1}{\beta(p,q)}\left(\frac{y_r-a}{b}\right)^{p}\Gamma(q)\sum_{i=0}^{\infty}a_i\left(\frac{y_r-a}{b}\right)^{i}\right]^{t+r-1}$$

$$\times\frac{1}{b\beta(p,q)}\left(\frac{y_r-a}{b}\right)^{p-1}\left[1-\left(\frac{y_r-a}{b}\right)\right]^{q-1}dy_r,$$

where a_i is as given by (7) with subscript *i* in place of *j*. The substitution $u=\frac{y_r-a}{b}$ yields to

$$\mu_r^k=\frac{r}{\Gamma(q)}\binom{n}{r}\left[\frac{\Gamma(p+q)}{\Gamma(p)}\right]^{r}\sum_{t=0}^{n-r}\binom{n-r}{t}\left[\frac{-\Gamma(p+q)}{\Gamma(p)}\right]^{t}$$

$$\times\int_0^1(bu+a)^k\cdot u^{p(t+r)-1}(1-u)^{q-1}\left[\sum_{i=0}^{\infty}a_iu^i\right]^{t+r-1}du. \qquad \text{(A-1)}$$

According to Gradshteyn and Ryzhik (1980, pp. 14), we can write

$$\left[\sum_{i=0}^{\infty}a_iu^i\right]^{t+r-1}=\sum_{i=0}^{\infty}c_iu^i, \qquad \text{(A-2)}$$

where

$$\left.\begin{aligned}c_0&=a_0^{t+r-1},\\ c_i&=\frac{1}{a_0}\sum_{j=1}^{i}\left[\frac{(t+r)j}{i}-1\right](a_j)(c_{i-j}),\quad i\geq 1,\\ a_j&=\frac{(-1)^j}{\Gamma(q-j)\cdot(p+j)\cdot j!}.\end{aligned}\right\} \qquad \text{(A-3)}$$

Expanding $(bu+a)^k$ in (A-1), substituting from (A-2) and (A-3) for the term $\left[\sum_{i=0}^{\infty}a_iu^i\right]^{t+r-1}$ into (A-1) and following some algebraic manipulation we finally get the moments μ_r^k as given in (3).

REFERENCES

Abdelkader Y.H. (To appear). "*Computing the Moments of Order Statistics from Independent Nonidentically Distributed Beta Random Variables*". Statistical Papers, available online, as of September 5, 2009 from SpringerLink: https://commerce.metapress.com/content/q87m43708m2k1j64/

Andrews C.L. (1985). *Special Functions for Engineers and Applied Mathematics*. Macmillan Publishing Company.

Balakrishnan N. and Nevzorov V.B. (2003). *A Primer on Statistical Distributions*. John Wiley & Sons, Inc.

David H.A. (1981). *Order Statistics*, Second Edition. John Wiley & Sons, Inc.

Feller. W. (1971). *An Introduction to Probability Theory and its Applications*, Volume 2, Second Edition. John Wiley & Sons, Inc.

Gradshteyn I.S. and Ryzhik I.M. (1980). *Tables of Integrals, Series and Products*. Corrected and Enlarged Edition. Academic Press.

Nadarajah S. (2008). "*Explicit Expressions for Moments of Order Statistics*". Statistics & Probability Letters 78, 196 – 205.

Thomas, P.Y. and Samuel P. (2008). "*Recurrence Relations for the Moments of Order Statistics from Beta Distribution*". Statistical Papers 49, 139 – 146.

Zhang S. and Jin J. (1996). *Computation of Special Functions*, John Wiley & Sons, Inc.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 1038–1049.

D-OPTIMAL DESIGNS PROFILE-BASED SENSITIVITY IN REGRESSION MODELS

H. Sulieman
Department of Mathematics and Statistics
American University of Sharjah, P.O.Box 26666, Sharjah, U.A.E.
E-mail: hsulieman@aus.ae

P. J. McLellan
Department of Chemical Engineering
Queen's University, Kingston, Ontario, Canada, K7L 3N6.
E-mail: mclellnj@chee.queensu.ca

ABSTRACT

Local D-optimal experimental designs for precise parameter estimation are designs which minimize the determinant of the variance-covariance matrix of the parameter estimates based on local sensitivity coefficients. For nonlinear models, this determinant may not give a true indication of the volume of the joint inference region for the parameters because of the underlying nonlinearity of the estimation problem. In this article, we investigate sequential D-optimal experimental designs using profile-based sensitivity coefficients developed by Sulieman et.al. (2001, 2004). Profile-based sensitivity coefficients account for both parameter estimate correlations and model nonlinearity and are, therefore, expected to yield better precision of parameter estimates when used in D-optimal design criteria. Some characteristics of the profile-based designs and related computational aspects are discussed. Applications of the new designs to linear and nonlinear model cases are also presented.

Keywords: Sequential D-optimal design, Local sensitivity coefficient, parameter estimation, profile-based sensitivity coefficient.

1. INTRODUCTION

Mathematical modeling, simulation and optimization are nowadays essential tools in understanding, explaining and exploiting the behavior of complex systems. There are two methods for acquiring information about the models representing these systems and their parameters: parameter identifiability and parameter sensitivity. This article focuses on parameter sensitivity which in general describes the impact of perturbations in the values of model input parameters on the model outputs. Sensitivity results are used to improve the quality of the model perhaps by reducing complexity or by guiding further experiments to reduce uncertainty or discriminate among rival models. Several design of experiment techniques have been developed in the literature and applied successfully to wide range of systems (Franceschini and Macchietto, 2008). The objectives of these techniques typically focus on model precision

or/and model discrimination. The most popular design criterion is the D-optimality which minimizes the determinant of the variance-covariance matrix of the parameter estimates. For linear models, the optimum design does not depend, at least in general, on the values of the model parameters and it is possible to arrive at a common criterion. For nonlinear models, however, the optimum experimental designs depend on the values of the unknown parameters and the problem of their construction is necessarily more complicated than that for linear models.

The primary goal of this article is to employ two different sensitivity measures in the construction of D-optimal designs and compare the results. The first measure is the conventional local sensitivity measure defined by the first-order partial derivatives of the regression model response function with respect to the parameters. To date, D-optimal designs are defined in terms of these local measures. In sensitivity assessment, the local sensitivity coefficients measure the marginal impact of the parameters on the model predictions due to their inability to incorporate simultaneous changes in parameter values. To surmount the drawbacks of the local sensitivity assessment, Sulieman *et al.* (2001, 2004) proposed an alternative assessment procedure in which simultaneous perturbations in the values of all model parameters are achieved using the profiling scheme introduced by Bates and Watts (1988) for nonlinearity assessment of regression models. The profile-based sensitivity measure; defined by the total derivative of the model function with respect to parameter of interest, was shown to account for both nonlinearity within the parameter estimation problem and parameter estimate co-dependencies. Like any derivative measure, profile-based sensitivity is inherently local, it provides, however, a more comprehensive picture of the prediction sensitivity in the presence of parameter co-dependencies and model nonlinearity. Sulieman *et al.* (2009) called profile-based sensitivity coefficients hybrid local-global sensitivity measure.

The present work provides preliminary results in the construction of D-optimal designs using profile-based sensitivity measures. The designs are constructed sequentially (Myers *et al.*, 1989) where only one additional experiment is generated for an existing design. The new design is evaluated through re-estimation of model parameter values and assessment of their accuracy. In section 2 we present profile-based sensitivity procedure using the notion of model re-parameterization in single-response nonlinear regression models. In particular, we adopt *predicted value* re-parameterization in which the expression for the predicted response at a selected design point is defined as one of the parameters in the new system. Such a re-parameterization allows the profiling-based sensitivity measure to be calculated automatically with the calculations of the profile vector when fitting the newly formulated model. In Section 3 we develop the profile-based D-optimal designs for linear and nonlinear regression models and discuss their properties. Illustrative model cases are presented in Section 4, and conclusions are summarized in Section 5.

2. PROFILE-BASED SENSITIVITY ANALYSIS

Consider the general mathematical form of a single response nonlinear regression model

$$\mathbf{y} = \mathbf{f}(\mathbf{X}, \Theta) + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{y}$ is an n-element vector of observed values of the response variable for particular values of the regressor variables $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, Θ is a p-element vector of unknown parameters, $\mathbf{f}$ is an n-element vector of predicted values of the response variable for given $\mathbf{X}$ and Θ, $\mathbf{f}(\mathbf{X}, \Theta) = \{f(\mathbf{x}_1, \Theta), f(\mathbf{x}_2, \Theta), \ldots$
$f(\mathbf{x}_n, \Theta)\}$, and $\boldsymbol{\epsilon}$ is an n-element vector of independent random errors with a specified joint distribution. In most cases, including the case here, $\boldsymbol{\epsilon}$ is assumed to have a spherical normal distribution, with $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $var(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2\mathbf{I}$.

Sulieman *et al.* (2001) developed a profile-based sensitivity measure to assess the sensitivity of the predicted responses from model (1). The motivation for the profile-based approach arises from the profiling algorithm developed by Bates and Watts (1988) for constructing likelihood intervals for individual parameters in single response nonlinear regression models. The profiling algorithm was implemented to a reformulated model function using special re-parameterization in which the predicted response is one of the parameters in the new formulation, namely:

$$\begin{aligned} \phi_1 = \eta_0(\Theta) &= f(\Theta, \mathbf{x} = \mathbf{x}_0) \\ \phi_2 &= \theta_2 \\ &\vdots \\ \phi_p &= \theta_p \end{aligned} \tag{2}$$

where $\eta_0(\Theta) = f(\Theta, \mathbf{x} = \mathbf{x}_0)$ is the predicted response at a selected prediction point $\mathbf{x}_0$. The re-parameterization in equation (2) is called the predicted response parametrization. It is not necessary that only ϕ_1 is used to define the predicted response at $\mathbf{x}_0$, nor that only θ_1 is used as its inverse ϕ_1^{-1}. Any other two parameters from Φ and Θ parameters can be chosen for these purposes. For reader's convenience, however, we will denote the predicted response parameter by ϕ_1 throughout the paper, whereas the inverse transformation ϕ_1^{-1} will be defined by one of the least nonlinear behaving parameters in $\boldsymbol{\eta}(\Theta)$.

Bates and Watts (1981) showed that when the regression model has only one nonlinear parameter, the transformation (2) will reduce parameter nonlinearity to zero for all Θ. Clarke (1987) referred to this class of transformation as "optimal parameter" transformations since it produces zero or small parameter nonlinearities in the model and minimum estimation bias. However, as shown by Clarke (1987), the effectiveness of such re-parameterizations depends substantially on the choice of the points $\mathbf{x}_0$ at which the predicted responses are to be estimated. In using this class of transformations, our purpose is not to reduce either the parameter nonlinearities in the model but rather to characterize the parameter sensitivities of the predicted response at a predetermined point $\mathbf{x}_0$. Therefore, the re-parameterization defined in (2) is more functional here since it defines the identity transformation for $(p-1)$ parameters, so that neither the nonlinearities of these parameters nor the biases in their

estimates are changed by the transformation. The only parameter for which the nonlinearity or the bias may change is the one defined by the inverse transformation ϕ_1^{-1}, say θ_1, where θ_1 is chosen to be one of the most linearly behaving parameters in the model.

Let $\boldsymbol{\eta}_{new}(\Phi)$ be the new formulation of the model function in terms of Φ and suppose that ϕ_i is the parameter of interest. In profiling algorithm, the vector of parameters Φ is partitioned as $\Phi = (\phi_i, \Phi_{-i})$ which is, in terms of Θ, equivalent to $\Theta = (\theta_i, \Theta_{-i})$. $\phi_i = \theta_i$ is then varied across its range of uncertainty and for each value of ϕ_i, let $\tilde{\Phi}_{-i}(\phi_i)$ be the conditional least squares estimates of Φ_{-i}. The joint behavior of the predicted response parameter and each of the remaining parameters in $\boldsymbol{\eta}_{new}$ can be understood through the profile traces, and so sensitivity information can be extracted from the profile trace plots. Using the studentized parameters, the profile trace of the predicted response parameter ϕ_1 versus ϕ_i, $i = 2, 3, \ldots, p$, is the curve consisting of the points $(\delta(\phi_i), \tilde{\delta}(\phi_1))$, where $\tilde{\delta}(\phi_1)$ is the studentized conditional estimate of ϕ_1 given a fixed value of ϕ_i, i.e,

$$\tilde{\delta}(\phi_1) = \frac{\tilde{\phi}_1 - \hat{\phi}_1}{se(\hat{\phi}_1)} \tag{3}$$

and

$$\delta(\phi_i) = \frac{\phi_i - \hat{\phi}_i}{se(\hat{\phi}_i)} \tag{4}$$

where the parameter estimates $\hat{\phi}_1$ and $\hat{\phi}_i$ and their associated standard errors $se(\hat{\phi}_1)$ & $se(\hat{\phi}_i)$ are obtained from the unconditional least squares fitting of $\boldsymbol{\eta}_{new}$ to the data.

Sulieman *et al.* (2001) defined the Profile-based Sensitivity Coefficient (PSC) as the slope of $\tilde{\delta}(\phi_1)$ with respect to $\delta(\phi_i)$. At a given design point x_0 PSC is derived as follows:

$$PSC_i(x_0) = \frac{\partial \tilde{\delta}(\phi_1)}{\partial \delta(\phi_i)} = \frac{se(\hat{\phi}_i)}{se(\hat{\phi}_1)} \frac{\partial(\tilde{\phi}_1)}{\partial(\phi_i)} \tag{5}$$

Expressing equation (5) in terms of Θ parameters yields:

$$PSC_i(x_0) = \frac{se(\hat{\theta}_i)}{se(\hat{\eta}_0)} \frac{\partial(\tilde{\eta}_0(\Theta))}{\partial(\theta_i)} = \frac{se(\hat{\theta}_i)}{se(\hat{\eta}_0)} \frac{D\eta_0(\theta_i, \tilde{\Theta}_{-i})}{D\theta_i} \tag{6}$$

where the operator $D\eta_0(\theta_i, \tilde{\Theta}_{-i})$ denotes the total derivative of $\eta_0(\theta_i, \tilde{\Theta}_{-i})$ with respect to θ_i, (Sulieman *et al.*, 2001). Expressing the total derivative in terms of partial derivatives gives the following result:

$$\frac{D\eta_0(\theta_i, \Theta_{-i}(\theta_i))}{D\theta_i} = \frac{\partial \eta_0}{\partial \theta_i} + \frac{\partial \eta_0}{\partial \Theta_{-i}}\Big|_{\tilde{\Theta}_{-i}} \frac{\partial \tilde{\Theta}_{-i}}{\partial \theta_i} \tag{7}$$

Using least squares estimation criterion, Sulieman *et al.* (2001) showed that the term $\frac{\partial \tilde{\Theta}_{-i}}{\partial \theta_i}$ is given by:

$$\frac{\partial \tilde{\Theta}_{-i}}{\partial \theta_i} = -\left\{ \left(\frac{\partial^2 S(\theta_i, \Theta_{-i}(\theta_i))}{\partial \Theta_{-i} \partial \Theta'_{-i}} \right)^{-1} \frac{\partial^2 S(\theta_i, \Theta_{-i}(\theta_i))}{\partial \theta_i \partial \Theta_{-i}} \right\}\Big|_{\tilde{\Theta}_{-i}} \tag{8}$$

where $S(\theta_i, \Theta_{-i}(\theta_i)) = \sum_{i=1}^{n}(y_i - \eta_i(\theta_i, \Theta_{-i}(\theta_i)))^2$ is the conditional least squares function. Substituting equations (7)and (8) into equation (6)yields:

$$PSC_i(x_0) = \frac{se(\hat{\theta}_i)}{se(\hat{\eta}_0)}\left\{\frac{\partial\eta_0}{\partial\theta_i} - \frac{\partial\eta_0}{\partial\Theta_{-i}}(\frac{\partial^2 S}{\partial\Theta_{-i}\partial\Theta'_{-i}})^{-1}\frac{\partial^2 S}{\partial\theta_i\partial\Theta_{-i}}\Big|_{\tilde{\Theta}_{-i}}\right\} \tag{9}$$

Using first and second order derivative information of the model function $\boldsymbol{\eta}(\Theta)$, equation (9) can be shown to equal:

$$PSC_i(x_0) = \frac{se(\hat{\theta}_i)}{se(\hat{\eta}_0)}\left\{v_{0_i} - \mathbf{v}'_{0_{-i}}(V'_{-i}V_{-i} - [e'][V_{-i-i}])^{-1}(V'_{-i}\mathbf{v}_i - \mathcal{D}'_{\cdot\cdot}e)\right\} \tag{10}$$

where v_{0_i} is the i-th component of the first derivative vector $\mathbf{v}_0$ evaluated at $\mathbf{x}_0$; V_{-i} is an $n \times (p-1)$ matrix consisting of first derivative vectors of $\boldsymbol{\eta}(\Theta)$ with respect to Θ_{-i}; $\mathbf{v}_{0_{-i}}$ is a $(p-1)$ dimensional vector consisting of the elements in the row of V_{-i} which corresponds to $\mathbf{x}_0$; V_{-i-i} is the $n \times (p-1) \times (p-1)$ array of the second derivatives of $\boldsymbol{\eta}(\Theta)$ with respect to Θ_{-i}; $\mathcal{D}_{\cdot\cdot}$ is the $n \times (p-1)$ matrix of the second derivatives of $\boldsymbol{\eta}(\Theta)$ with respect to Θ_{-i} and θ_i, and e is the n-element residuals vector. The quantities in equation (10) are evaluated at $(\hat{\theta}_i, \tilde{\Theta}_{-i}(\hat{\theta}_i))$.

The first term in equation (10), v_{0_i} gives a measure of the conventional Marginal Sensitivity Coefficient, MSC, (Sulieman *et al.*, 2001). The second term in the equation, $\left\{\mathbf{v}'_{0_{-i}}(V'_{-i}V_{-i} - [e'][V_{-i-i}])^{-1}(V'_{-i}\mathbf{v}_i - \mathcal{D}'_{\cdot\cdot}e)\right\}$, represents an adjustment term containing two components of information. The first is the marginal effects of Θ_{-i} on the predicted response at $\mathbf{x}_0$ through the derivative vector $\mathbf{v}_{0_{-i}}$. The second component of the information relates to the co-dependency structure of the parameters Θ. It includes two sets of co-dependencies: the pairwise correlations among the elements of $\tilde{\Theta}_{-i}$ via the term $(V'_{-i}V_{-i} - [e'][V_{-i-i}])^{-1}$ and correlations between $\hat{\theta}_i$ and $\tilde{\Theta}_{-i}$ via the term $(V'_{-i}\mathbf{v}_i - \mathcal{D}'_{\cdot\cdot}e)$. All terms in equation (10)are scaled by the factor $\dfrac{se(\hat{\theta}_i)}{se(\hat{\eta}_0)}$. Because second-order derivatives of $\boldsymbol{\eta}(\Theta)$ are included in the co-dependency terms, model nonlinearity is accounted for by the profile-based sensitivity coefficient up to second-order derivatives.

The adjustment term in equation (10) incorporates the simultaneous changes in the parameter values making $PSC_i(x_0)$ global sensitivity measure while it is inherently local since it is derivative-based. It is only when the parameter co-dependency structure and model nonlinearity are insignificant that $PSC_i(x_0)$ becomes equivalent measure to local sensitivity coefficient. $PSC_i(x_0)$ is called *hybrid local-global* sensitivity measure (Sulieman *et al.*, 2009).

For a linear model, $\mathbf{y} = \boldsymbol{\eta}(\boldsymbol{\beta}) + \boldsymbol{\epsilon}$, it can be shown that the profile-based sensitivity coefficient for the predicted response at $\mathbf{x} = \mathbf{x}_0$, $\eta_0(\boldsymbol{\beta})$, to the i-th parameter β_i evaluated at the least squares parameter estimates, reduces to

$$PSC_i(\mathbf{x}_0) = \frac{se(\hat{\beta}_i)}{se(\hat{\eta}_0)}\{x_{0_i} - \mathbf{x}_{0_{-i}}(\mathbf{X}'_{-i}\mathbf{X}_{-i})^{-1}\mathbf{X}'_{-i}\mathbf{x}_i\} \tag{11}$$

where $\mathbf{X}_{-i}$ is the $n \times (p-1)$ matrix of regressor variables formed by removing the i-th column $\mathbf{x}_i$ of the original matrix $\mathbf{X}$ of regressor variables, and $\mathbf{x}_{0_{-i}}$ is the row of $\mathbf{X}_{-i}$ corresponding to $\mathbf{x} = \mathbf{x}_0$.

The first term in equation (11) gives the scaled marginal sensitivity coefficient,

$$MSC_i(\mathbf{x}_0) = \frac{se(\hat{\beta}_i)}{se(\hat{\eta}_0)} x_{0_i}.$$

The adjustment term in the second term is proportional to the portion of $\mathbf{x}_i$ that is explained by the variables in $\mathbf{X}_{-i}$ obtained from linearly regressing $\mathbf{x}_i$ on $\mathbf{X}_{-i}$. This implies that PSC in the linear model measures the influence of β_i on the predicted response at $\mathbf{x}_0$ after the effects of the remaining parameters on β_i have been removed. In contrast, MSC gives the effect of β_i on the predicted response with the remaining parameters fixed at their conditional least squares values.

For a linear model, the profile-based sensitivity coefficients are independent of the values of the parameter estimates; they depend only on the matrix $\mathbf{X}$. For orthogonal matrix $\mathbf{X}$, the profile-based and marginal sensitivity assessments are equivalent.

3. PROFILE-BASED D-OPTIMAL DESIGN

D-optimal designs are most commonly used of the alphabet designs (Ryan, 2007; Myers *et al.*, 1989). A D-optimal design minimizes the volume of the parameter joint conference region or equivalently maximizing the determinant of the Fisher Information matrix. Box and Lucas (1959) gave the first formulation of the D-optimal design for nonlinear models. They defined the D-optimality objective function, using the *unscaled* local sensitivity coefficients, as:

$$D = |V_0' V_0| \tag{12}$$

where the matrix of local sensitivity coefficients V_0 is evaluated at an initial parameter estimates Θ_0. By maximizing D, the volume of the linear approximation to the exact confidence region of Θ is minimized at Θ_0. When model nonlinearity is pronounced, the *local* D-optimality can produce designs with poor performance and little information about parameters. Hamilton and Watts (1985) introduced quadratic designs based on second-order approximation to the inference region of Θ. Quadratic designs have the distinct advantage of taking into account the nonlinearity of response function.

The D-optimal designs are often constructed by sequential experimental strategies (Myers *et al.*, 1989). Sequential designs are appealing because they offer the chance to change strategy after the first round of experiment has been completed and new information is available. The unscaled marginal sensitivity coefficients for the prior experiments are included in the V_0 matrix along with the new row corresponding to the new experimental conditions to be selected. In sequential designs, the V_0 is constructed as follows:

$$V_0 = \begin{bmatrix} V_{0_{old}} \\ V_{0_{new}} \end{bmatrix} \tag{13}$$

where $V_{0_{old}}$ is the unscaled marginal sensitivity matrix for the pre-existing experimental settings and $V_{0_{new}}$ contains rows of sensitivity coefficients corresponding to the new experimental settings being selected. Many strategies have been developed for generation of D-optimal designs by sequentially adding runs to an existing design (Franceschini *et al.*, 2008).

In what follows, the *unscaled* profile-based sensitivity coefficients developed in the previous section are utilized in a sequential D-optimal design strategy. Let $\mathbf{P_i}$ denote the vector of the *unscaled* profile-based sensitivity coefficients corresponding to the parameter θ_i and evaluated at all prediction points of interest. If a pre-existing design consisting of n prediction points is available, $\mathbf{P_i}$ is an $n \times 1$ vector. From equation (10), $\mathbf{P_i}$ is expressed as:

$$\mathbf{P_i} = \mathbf{v_i} - \mathbf{V_{-i}}(\mathbf{V'_{-i}V_{-i}} - [\mathbf{e'}][\mathbf{V_{-i-i}}])^{-\mathbf{1}}(\mathbf{V'_{-i}v_i} - \mathcal{D}'\mathbf{..e}) \tag{14}$$

The corresponding D-optimality criterion is:

$$D_P = |\boldsymbol{\mathcal{P}}'_0\boldsymbol{\mathcal{P}}_0| \tag{15}$$

where the matrix $\boldsymbol{\mathcal{P}} = [\mathbf{P_1P_2}\ldots\mathbf{P_p}]$ is evaluated at Θ_0, i.e., each element $\mathbf{P_i}$ is evaluated at Θ_0 . In sequential design approach, Θ_0 is often taken as the least squares estimate of Θ from the current design.
For linear models, the $\mathbf{P_i}$ is independent of Θ and given by:

$$\mathbf{P_i} = \mathbf{x_i} - \mathbf{X_{-i}}(\mathbf{X'_{-i}X_{-i}})^{-\mathbf{1}}\mathbf{X'_{-i}x_i} \tag{16}$$

which can be expressed as:

$$\mathbf{P_i} = \mathbf{Lx_i} \tag{17}$$

where $\mathbf{L}$ is a linear transformation matrix given by:

$$\mathbf{L} = \mathbf{I_n} - \mathbf{X_{-i}}(\mathbf{X'_{-i}X_{-i}})^{-\mathbf{1}}\mathbf{X'_{-i}} \tag{18}$$

where $\mathbf{I_n}$ is the n-dimensional identity matrix. A glance at the matrix $\mathbf{L}$ reveals that in the linear model the profile-based sensitivity coefficient for β_i is a projection of the column $\mathbf{x}_i$ onto the column space orthogonal to $\mathbf{X}_{-i}$. In other words, $\mathbf{P_i}$ is the residuals from regressing $\mathbf{x}_i$ on $\mathbf{X}_{-i}$. In parameter sensitivity terms, $\mathbf{P_i}$ measures the influence that β_i exerts on the predicted response after the removal of its co-dependencies with the remaining parameters. The corresponding D-optimality criterion is independent of $\boldsymbol{\beta}$ parameters and is given by:

$$D_P = |\mathbf{X'L'LX}| \tag{19}$$

which reduces to:

$$D_P = |\mathbf{X'LX}| = |\mathbf{L}||\mathbf{X'X}| \approx |\mathbf{X'X}| \tag{20}$$

because $\mathbf{L}$ is symmetric ($\mathbf{L'} = \mathbf{L}$) and idempotent ($\mathbf{L^2} = \mathbf{L}$) and $|L|$ is constant.

Equation (20) implies that the design settings that maximize D_P are equal to the settings maximizing D in equation (12). This result stems from the invariance property of D-optimal designs to linear transformations of the design space that are independent of model parameters.

4. ILLUSTRATIVE EXAMPLES

In the following two examples, we use sequential design strategy to generate one additional design point using D and D_P optimality criteria given in equations (12) and (15), respectively. For each criterion the corresponding design matrix is an $(n+1) \times p$ consisting of the pre-existing design of size n and an additional row of sensitivity coefficients evaluated at the new design point to be generated. The optimization algorithm is implemented on MATLAB 7.6 using appropriate minimization solver where both D^{-1} and D_P^{-1} are minimized leading to the same optimum D and D_P, respectively, are maximized.

Example 4.1: Linear Model Case
Consider the following first-order plus two-factor interaction linear model:

$$f(\mathbf{x}, \boldsymbol{\beta}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \tag{21}$$

A data set on tire wear (response), temperature (x_1) and contact pressure (x_2) is used in Seber and Lee (2003) to fit the model. The data set is reported in Table 1. A 12^{th} design

Table 1: Linear regression model data reported in Seber and Lee (2003), Example 4.1.

Observation no.	Temperature (F°)	Pressure (psi)	Wear
1	500	15	0.1014
2	500	21	0.5009
3	1000	15	0.8152
4	1000	21	0.4026
5	750	13.5	0.7001
6	750	22.5	0.1995
7	375	18	0.5753
8	1125	18	0.8747
9	750	18	0.4893
10	750	18	0.5031
11	750	18	0.5118

point is generated by D^{-1} and D_P^{-1} optimality criteria. We begin by evaluating both criteria at the original design points and at sequential design points at the corners of the region. The corner point ($x_1 = 500$, $x_2 = 15$) is the design setting at which the optimum value of both criteria occur. This result is not a surprise. As discussed in Section 3, for linear models, the structure of the D_P optimality can be viewed as a linear transformation of the design matrix X and therefore, both D and D_P yield the same optimal value because of the invariance property of the D- optimality criterion.

Example 4.2: Michaelis-Menten Model
Michaelis-Menten model is commonly used in enzymatic kinetics with well-known formulation:

$$f(x, \Theta) = \frac{\theta_1 x}{\theta_2 + x} \tag{22}$$

where y is the measured initial velocity of an enzymatic reaction and x is the substrate concentration. The unknown parameters θ_1 and θ_2 represent maximum conversion rate and Michaelis-Menten constant, respectively. The data set used by Bates and Watts (1988) for fitting model (22) is used here. As shown in Table 2 below, the data set contains 6 different design settings, each with one replicate.

The given design is used to estimate model parameters. The results are given in Table 3:

Table 2: Michaelis-Menten equation data reported in Bates and Watts (1988), Example 4.2.

Observation no.	Substrate Concentration (ppm)	Velocity (counts/min^2)
1	0.02	76
2		47
3	0.06	97
4		107
5	0.11	123
6		139
7	0.22	159
8		152
9	0.56	191
10		201
11	1.10	207
12		200

Table 3: Summary of parameter estimates for the Michaelis-Menten model

Parameter	Estimate	St.error	
θ_1	212.68	6.94	$corr(\hat{\theta}_1, \hat{\theta}_2) = 0.77$, $s^2 = 119.5$
θ_2	0.06411	0.008	with 10 degrees of freedom

The 13^{th} concentration point is generated using MATLAB 7.6 minimization routine for restricted x in the domain $0 < x \leq x_{max} = 1.1$. The optimal value for the additional concentration point is $x = 0.0747$ when D^{-1} is minimized and $x = 0.05116$ when D_P^{-1} is minimized. In an attempt to evaluate the information content of the new design, formed by adding the optimal value of x to the existing design in Table 2, the parameters θ_1 and θ_2 are re-estimated using the 13 experimental runs in the combined design. The response variable

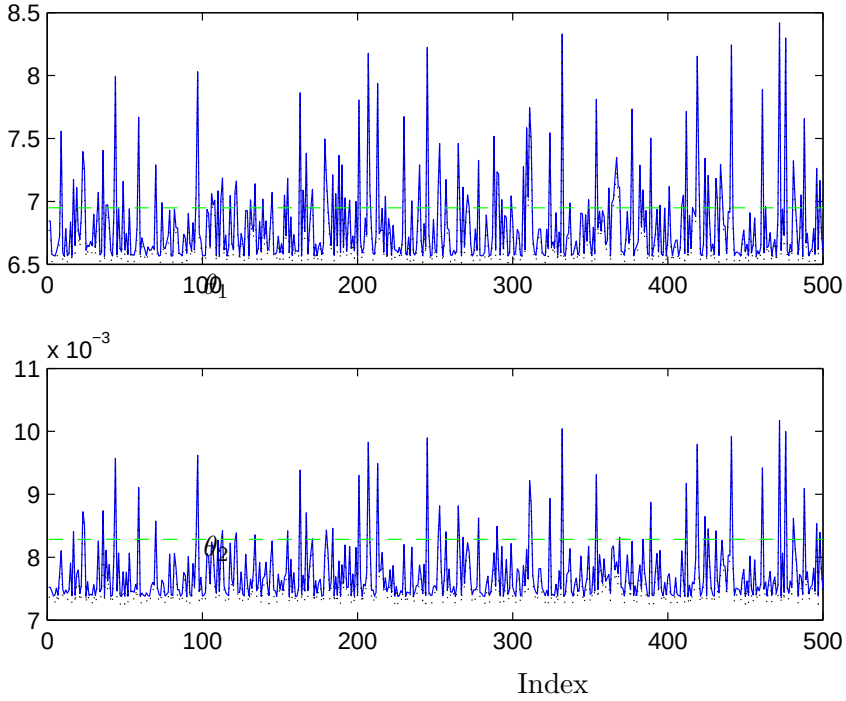


Figure 1: Standard errors of the re-estimated Michaelis-Menten model parameters θ_1 and θ_2 using additional D optimal design point. The solid and dotted lines join the values of $se(\hat{\theta}_i)$ resulting from the optimal D and D_P designs, respectively. The dashed line gives $se(\hat{\theta}_i)$ resulting from the original design

for the 13^{th} concentration point is simulated by using the fitted model in Table 3 and adding normally distributed random noise with variance equal to $s^2 = 119.5$ given in Table 3. For each of the 500 simulations that were carried out, the accuracy of the resulting parameter estimates expressed by their standard errors is assessed. The results are shown in the plots in Figure 1.

Despite the gain in degrees of freedom when the combined design is used, for a number of simulations, the standard errors of the two parameter estimates using the combined design exceed the corresponding ones using the existing design (dashed line). This behavior is seen more occuring in $\hat{\theta}_1$ than in $\hat{\theta}_2$ implying that the D and D_P optimal concentration points provide more information for estimating θ_2 than for estimating θ_1. For the majority of simulations, the dotted line representing $se(\hat{\theta}_i)$ using D_P design is lower than the solid line representing $se(\hat{\theta}_i)$ using D design for both parameter estimates. This is to say that the reduction in standard errors associated with D_P design where $x = 0.05116$ is more substantial than that associated with D design where $x = 0.0747$. The former design point yielded more significant improvement in the estimates of the two parameters than the latter design point.

Table 4 reports the average $se(\hat{\theta}_i)$ over the 500 simulations for both optimal design points. It is obvious from the table that the relative improvement in the precision of $\hat{\theta}_1$ and $\hat{\theta}_2$ is greater for D_P-optimal than for D-optimal design. The relative reduction in $se(\hat{\theta}_i)$ is greater for $\hat{\theta}_2$ than for $\hat{\theta}_1$ in both designs implying the information gained by the additional

Table 4: Average $se(\hat{\theta}_i)$ over the 500 simulations

Design	$se(\hat{\theta}_1)$	$se(\hat{\theta}_2)$
D-optimal ($x = 0.0747$)	6.82	0.0077
D_P-optimal $x = 0.05116$	6.74	0.0072
Existing	6.94	0.008

concentration point is mostly utilized in estimating θ_2. It should be noted that θ_2 in model (22) represents the half-concentration, i.e. the value of x such that when the concentration reaches that value the velocity y is one-half its ultimate value. From the observed velocity seen in Table 2, one-half of maximum y, $y_{max}/2$, is reached at a concentration of about 0.06 and so adding additional experimental run around this concentration provides most information about θ_2 estimate.

5. CONCLUSIONS

Whereas local D-optimal designs take account of neither the co-dependency structure among model parameters nor the model nonlinearity, the profile-based D-optimal designs take both characteristics into account. In turn, using profile-based D-optimal criterion in sequential design strategy provides more informative data for use in parameter estimation. For linear models, both local and profile-based optimal design criteria give same experimental data due to the invariance property of D-optimality criterion in general.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support of the American University of Sharjah, United Arab Emirates and the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Bates, D.M. and Watts, D.G. (1988), *Nonlinear Regression Analysis and Its Applications*. Wiley: New York.

Bates, D.M. and Watts, D.G. (1981), "Parameter transformation for improved approximate confidence regions in nonlinear least squares". *Annals of Statistics*, 9, , p 1152–1167.

Box, G.E.P. and Lucas, H.L. (1959). "Design of experiments in non-linear situations". *Biometrika*, 46, pp 77–90.

Clarke, G.P.Y. (1987), "Approximate confidence limits for a parameter function in Nonlinear Regression". *JASA*, 82(397), p 221–230.

Franceschini, G. and Macchietto, S. (2008),"Model-based design of experiments for parameter precision: State of the art". *Chemical Engineering Science*, 63, pp 4846–4872.

Hamilton, D.C. and Watts, D.G. (1985), "A quadratic design criterion for precise estimation in nonlinear regression models". *Technometrics*, 27(3), p 241–250.

Myers, R.H., Khuri, A.I. and Carter, W.H. (1989), "Response Surface Methodology: 1966–1988". *Technometrics*, 31(2), p 137–57.

Ryan, T.P. (2007), *Modern Experimental Design*. John Wiley & Sons, Nw York.

Seber, G.A.F and Lee A.J. (2003), *Linear Regression Analysis, 2nd ed.*, Wiley: New York.

Sulieman, H., Kucuk, I. and McLellan J.(2009) "Parametric sensitivity: A case study comparison". *Computational Statistics & Data Analysis*, 53(7), p 2640–2652.

Sulieman H. (2008) "Improved local sensitivity measures for regression models with correlated parameters". *Proceedings in Computational Statistics 18th Symposium*, Porto, Portugal.

Sulieman, H., McLellan, P.J. and Bacon, D.W. (2004)"A Profile-based approach to parametric sensitivity in multiresponse regression models". *Computational Statistics & Data Analysis*, 45, p 721–740.

Sulieman, H., McLellan, P.J. and Bacon, D.W. (2001) "A Profile-Based Approach to Parametric Sensitivity Analysis of Nonlinear Regression Models". *Technometrics*, 43(4), p 425–33.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 1050–1069.

GENERALIZED ORDER STATISTICS FROM SOME DISTRIBUTIONS

Khalaf S. Sultan and T. S. Al-Malki
Department of Statistics and Operations Research,
College of Science,
King Saud University,
P.O.Box 2455, Riyadh 11451,
Saudi Arabia
E-mail: ksultan@ksu.edu.sa

ABSTRACT

In this paper, we derive explicit forms for the moments of the generalized order statistics (GOS) from the power function and log-logistics distributions. Then, we deduce the moments of the ordinary order statistics (OOS) and record values (RV) as special cases. Also, we use these moments to develop the best linear unbiased estimate (BLUE) of the scale parameter. In addition, we compare the BLUE with the corresponding maximum likelihood estimate. Finally, we show the usefulness and performance of the BLUE and MLEs through Monte Carlo Simulations.

1. INTRODUCTION

Generalized order statistics (GOS) have been introduced and extensively studied in Kamps (1995a,b) as a unified theoretical set-up which contains a variety of models of ordered random variables with different interpretations. Examples of such models are: (i) The ordinary order statistics (OOS), (ii) Record values (RV), (iii) The k-th record values (k-RV), (iv) Progressive Type-II censored order statistics (POS) and Sequential order statistics (SOS).

These models can be effectively applied in different aspects in real life problems. Ordinary order statistics (OOS), k-records [record values (RV) when $k = 1$], sequential order statistics, ordering via truncated distributions and censoring schemes can be discussed as they are special cases of the GOS. Kamps's book (1995a) gave several applications in a variety of disciplines, recurrence relations of the moments of GOS and characterizations, (for a survey of the models contained and of the results obtained in the GOS, see Kamps 1995a, b, 1999). Ahsanullah (1996, 1997, 2000) has discussed the GOS from the uniform and exponential distributions, Keseling (1999) has characterized some continuous distributions based on conditional expectations of GOS. Ahsanullah (2000) has characterized the exponential distribution based on independence of functions of GOS, Ahmed and Fawzy (2003) have characterized some of the doubly truncated distributions based on the concept of the generalized order statistics. Some specific distributions have been characterized by using the

relationships of the expected values of record values see Ahsanullah (1982, 1990, 1991), Balakrishnan and Ahsanullah (1994), Gupta (1984), Lin (1987) and Nagaraja (1977), Ahsanullah and Kirmani (1991) have characterized the exponential distribution through lower records while Abu-Youssef (2003) has characterized general classes of distributions through record values. AL-Hussaini and Ahmad (2003a,b) have constructed Bayesian interval prediction of the generalized order statistics and record values, respectively. AL-Hussaini, Ahmad and El Kashif (2005) have established some new recurrence relations for the moment generating function of the generalized order statistics. Sultan and El-Mougod (2005) have characterized general classes of doubly truncated absolutely continuous distributions by considering the conditional expectation of functions of record values.

Let $X_{1,n,\bar{m},k}, X_{2,n,\bar{m},k}, \ldots, X_{n,n,\bar{m},k}$ represent n GOS from a continuous population whose pdf and cdf are $f(x)$ and F(x), where $k \geq 1$ and $\bar{m} = (m_1, m_2, \ldots, m_{n-1})$ are real numbers. Then the joint pdf of $X_{1,n,\bar{m},k}, X_{2,n,\bar{m},k}, \ldots, X_{n,n,\bar{m},k}$ is given by [see Kamps 1995a]

$$f_{1,2,\ldots,n}(x_1, \ldots, x_n) = k \left(\prod_{j=1}^{n-1} \gamma_j \right) \left(\prod_{i=1}^{n-1} f(x_i) [\bar{F}(x_i)]^{m_i} \right) [\bar{F}(x_n)]^{k-1} f(x_n),$$
$$F^{-1}(0) < x_1 < \cdots < x_n < F^{-1}(1), \ \bar{F}(.) = 1 - F(.), \tag{1.1}$$

where

$$\gamma_j = k + n - j + \sum_{i=j}^{n-1} m_i. \tag{1.2}$$

Different models can be deduced from (1.1) as follows:

1. Let $m_1 = m_2 = \cdots = m_{n-1} = 0$ and $k = 1$ in (1.1) and (1.2), then GOS $\rightarrow$ OOS [see Arnold et al. (1992)].

$$f_{1,2,\ldots,n}(x_1, x_2 \ldots, x_n) = n! \prod_{i=1}^{n} f(x_i). \tag{1.3}$$

2. Let $m_1 = m_2 = \cdots = m_{n-1} = -1$ and $k = 1$ in (1.1) and (1.2), then GOS $\rightarrow$ upper RV [see Arnold et al. (1998)].

$$f_{1,2,\ldots,n}(x_1, x_2 \ldots, x_n) = f(x_n) \prod_{i=1}^{n-1} \frac{f(x_i)}{\bar{F}(x_i)}. \tag{1.4}$$

3. Let $m_1 = m_2 = \cdots = m_{n-1} = -1$ and $k > 1$ in (1.1) and (1.2), then GOS $\rightarrow$ k-th RV.

4. Let $m_i = R_i, \ i = 1, 2, \ldots, m-1$ and $k = R_m + 1$ in (1.1) and (1.2), then GOS $\rightarrow$ Type-II progressive censoring with removal scheme $(R_1, R_2, \ldots, R_m)$ [see Balakrishnan and Aggarwala (2000)].

Throughout this paper, we consider the pdf of the r-th GOS $X_{r,n,m,k}$ when $m_1 = m_2 = \cdots = m_{n-1} = m$ which is given by: [see Kamps (1995)]

$$f_{X_{r,n,m,k}}(x) = \frac{c_{r-1}}{(r-1)!} f(x)[1-F(x)]^{\gamma_r - 1} g_m^{r-1} F(x), \quad -\infty < x < \infty, \tag{1.5}$$

where

$$g_m(F(x)) = h_m(F(x)) - h_m(0), \tag{1.6}$$

$$h_m(F(x)) = \begin{cases} \frac{-(1-F(x))^{m+1}}{m+1}, & m \neq -1, \\ -\ln(1-F(x)), & m = -1, \end{cases} \tag{1.7}$$

and

$$c_{r-1} = \prod_{i=1}^{r} \gamma_i, \quad \gamma_i = k + (n-i)(m+1). \tag{1.8}$$

The joint pdf of $X_{r,n,m,k}$ and $X_{s,n,m,k}$, $1 \leq r < s \leq n-1$ when $m_1 = m_2 = \cdots = m_{n-1} = m$ is given by

$$\begin{aligned} f_{X_{r;n,m,k},X_{s;n,m,k}}(x,y) &= \frac{c_{s-1}}{(r-1)!(s-r-1)!} \{1-F(x)\}^m g_m^{r-1} F(x) \\ &\times \{h_m[F(y) - h_m(F(x)]\}^{s-r-1} \{1-F(y)\}^{\gamma_s - 1} \\ &\times f(x) f(y), \quad -\infty < x < y < \infty. \end{aligned} \tag{1.9}$$

2. EXACT MOMENTS OF GOS

2.1 The doubly truncated power function distribution

A random variable X is said to have the doubly truncated power function distribution if its pdf is given by

$$f(x) = \frac{\theta}{P-Q} x^{\theta-1}, \quad 0 < Q_1 \leq x \leq P_1 < 1, \ \theta > 0, \tag{2.1}$$

where $P = P_1^\theta$ and $Q = Q_1^\theta$. The cdf is given by

$$F(x) = \frac{x^\theta}{P-Q} - Q_2, \quad Q_1 \leq x \leq P_1, \ \theta > 0, \tag{2.2}$$

where $Q_2 = \frac{Q}{P-Q}$ and $P_2 = \frac{P}{P-Q}$.

The complete pdf and cdf of the power function distribution can be obtained, respectively, from (2.1) and (2.2) when $Q_1 = 0$ and $P_1 = 1$. For more details of the power function and its properties, see Johnson, Kotz and Balakrishnan (1994).

Single moments of GOS The single moments of GOS from the doubly truncated power function distribution are given in the following theorem.

Theorem (1):

The moment generating function of the r-th GOS from the doubly truncated power function distribution is given by

$$M_{X_{r;n,m,k}}(t) = \begin{cases} \frac{\theta c_{r-1}}{(r-1)!(m+1)^{r-1}} \sum_{d=0}^{\infty} \sum_{j=0}^{r-1} \sum_{i=0}^{\gamma_{(r-j)}-1} \binom{r-1}{j} \binom{\gamma_{(r-j)}-1}{i} \\ \quad \times \frac{(-1)^{j+i} t^d P^{\gamma_{(r-j)}-1-i}}{d!(P-Q)^{\gamma_{(r-j)}}} \{ \frac{P_1^{\theta(i+1)+d} - Q_1^{\theta(i+1)+d}}{\theta(i+1)+d} \}, \quad m \neq -1, \\ \\ k^r \sum_{d=0}^{\infty} \sum_{j=0}^{d/\theta} \binom{d/\theta}{j} \frac{t^d (P)^{d/\theta} (-1)^j}{P_2^j d! (k+j)^r}, \quad m = -1. \end{cases} \tag{2.3}$$

Hence, the a-th single moment of the r-th GOS is

$$\mu^{(a)}_{r;n,k,m} = \begin{cases} \frac{\theta c_{r-1}}{(r-1)!(m+1)^{r-1}} \sum_{j=0}^{r-1} \sum_{i=0}^{\gamma_{(r-j)}-1} \binom{r-1}{j} \binom{\gamma_{(r-j)}-1}{i} \\ \quad \times \frac{(-1)^{j+i} P_2^{\gamma_{(r-j)}}}{P^{1+i}} \{ \frac{P_1^{\theta(i+1)+a} - Q_1^{\theta(i+1)+a}}{\theta(i+1)+a} \}, \quad m \neq -1, \\ \\ k^r \sum_{j=0}^{\frac{a}{\theta}} \binom{\frac{a}{\theta}}{j} \frac{(P)^{\frac{a}{\theta}} (-1)^j}{P_2^j (k+j)^r}, \quad m = -1. \end{cases} \tag{2.4}$$

The moment generating function of the r-th GOS from the complete power function distribution derived by Saran and Pandy (2003) can be obtained from our result in (2.3) as a special case. That is

$$M_{X_{r;n,m,k}}(t) = \frac{\theta c_{r-1} (-1)^{r-1} e^t}{(r-1)!(m+1)^{r-1}} \sum_{d=0}^{\infty} \sum_{j=0}^{r-1} \binom{r-1}{j} \frac{(-1)^{j+d} t^d}{d!(\theta(m+1)(r-1-j)+\gamma_r+d)}. \tag{2.5}$$

Special cases:

From Theorem (1), we deduce some special cases as follows:

1. Let $m = 0$ and $k = 1$ in (2.4), we get single moments of OOS from the doubly truncated power function distribution as

$$\begin{aligned}\mu_{r:n}^{(a)} &= \frac{\theta c_{r-1}}{(r-1)!}\sum_{j=0}^{r-1}\sum_{i=0}^{n-r+j}\binom{r-1}{j}\binom{n-r+j}{i}\frac{(-1)^{j+i}}{(P-Q)^{n-r+j+1}} \\ &\times P^{n-r-i}\{\frac{P_1^{\theta(i+1)+a}-Q_1^{\theta(i+1)+a}}{\theta(i+1)+a}\}. \end{aligned} \tag{2.6}$$

   (a) If $n = r = 1$, then from (2.6), we get

$$\mu_{1:1}^{(a)} = \frac{\theta}{\theta+a}[P_2P_1^a - Q_2Q_1^a]. \tag{2.7}$$

   (b) IF $r = n$, $P = 1$ and $Q = 0$, then from (2.6), we get

$$\mu_{n:n}^{(a)} = \frac{n\theta}{n\theta+a}. \tag{2.8}$$

   (c) If $n = r = 1$, we get

$$\mu_{1:1}^{(a)} = \frac{\theta}{\theta+a}. \tag{2.9}$$

   The results (2.7)-(2.9) are given in Balakrishnan and Sultan (1998).

2. If $Q = 0$ (right truncated power function), $1 \le r \le n-1$, $\frac{a}{\theta}+1 > 1$, then from (2.4), we have

$$\mu_{r,n,m,k}^{(a)} = \frac{c_{r-1}}{(r-1)!(m+1)^{(r-1)}}\sum_{j=0}^{r-1}\binom{r-1}{j}(-1)^j P^{a/\theta}\frac{\Gamma(a/\theta+1)\Gamma(\gamma_{(r-j)})}{\Gamma(a/\theta+1+\gamma_{(r-j)})}, \tag{2.10}$$

   hence, when $m = 0$ and $k = 1$, we get (OOS)

$$\begin{aligned}\mu_{r:n}^{(a)} &= \frac{n!}{(n-r)!(r-1)!}\sum_{j=0}^{r-1}\binom{r-1}{j}(-1)^j P^{a/\theta}\frac{\Gamma(a/\theta+1)\Gamma(n-r+j+1)}{\Gamma(a/\theta+n-r+j+2)} \\ &= P^{a/\theta}\frac{\Gamma(a/\theta+r)\Gamma(n+1)}{\Gamma(r)\Gamma(a/\theta+n+1)}. \end{aligned} \tag{2.11}$$

   The results in (2.10) and (2.11) are given in Kamps (1995a). If $P = 1$ (complete power function) and $1 \le r \le n-1$, then form (2.11), we get

$$\mu_{r:n}^{(a)} = E(X_{r;n,m,k}^a) = \frac{c_{r-1}}{(r-1)!(m+1)^{r-1}}\sum_{j=0}^{r-1}\binom{r-1}{j}(-1)^j\frac{\Gamma(a/\theta+1)\Gamma(\gamma_{(r-j)})}{\Gamma(a/\theta+1+\gamma_{(r-j)})}. \tag{2.12}$$

3. Setting $m = -1$ and $k > 1$ k-th (RV) in (2.4), we get

$$\mu_{r:n}^{(a)} = E(X_{r,n,-1,k}^{a}) = k^{r} \sum_{j=0}^{a/\theta} \binom{a/\theta}{j} \frac{P^{a/\theta}(-1)^{j}}{P_2^{j}(j+k)^{r}}. \tag{2.13}$$

This result is given by Kamps (1995a) when $Q = 0$ as

$$\mu_{r}^{(a)} = E(X_{U_{(r)}^{k}}^{a}) = k^{r} \sum_{j=0}^{a/\theta} \binom{a/\theta}{j} \frac{P^{a/\theta}(-1)^{j}}{(j+k)^{r}}. \tag{2.14}$$

4. Setting $m = -1$ and $k = 1$ (RV) in (2.15), we get

$$\mu_{r:n}^{(a)} = E(X_{(r;n,-1,1)}^{a}) = \sum_{j=0}^{a/\theta} \binom{a/\theta}{j} \frac{P^{a/\theta}(-1)^{j}}{P_2^{j}(j+1)^{r}}. \tag{2.15}$$

This result is given in see Ahsanullah (1988).

Double moments Let $X_{1,n,m,k}, X_{2,n,m,k}, \ldots, X_{n,n,m,k}$ represent the GOS from the doubly truncated power function distribution. Then the double moments of GOS are given in the following theorem

Theorem (2):

The joint moment generating function of the GOS $X_{r,n,m,k}$ and $X_{s;n,m,k}$ from the doubly truncated power function distribution is given by

$$
M_{X_{r;n,m,k},X_{s;n,m,k}}(t_1,t_2)=\begin{cases}
\frac{c_{s-1}\theta^2}{(r-1)!(s-r-1)!(m+1)^{s-2}(P-Q)^{(m+1)(s-r-1)+\gamma_s+1}} \\
\times\ \sum_{d_2=0}^{\infty}\sum_{j_2=0}^{s-r-1}\sum_{i_2=0}^{(m+1)j_2+\gamma_s-1}\binom{s-r-1}{j_2} \\
\times\ \binom{(m+1)j_2+\gamma_s-1}{i_2}\sum_{d_1=0}^{\infty}\sum_{j_1=0}^{r-1}\sum_{i_1=0}^{(m+1)(s-r-j_2+j_1)-1} \\
\times\ \binom{r-1}{j_1}\binom{(m+1)(s-r-j_2+j_1)-1}{i_1} \\
\times\ \frac{(-1)^{j_2+i_2}t_2^{d_2}P^{(m+1)(s-r-j_2+j_1)-1-i_1}(-1)^{j_1+i_1}t_1^{d_1}}{d_2!(P-Q)^{(m+1)j_1+m}(\theta(i_2+1)+d_2)d_1!} \\
\times\Bigg\{P_1^{\theta(i_2+1)+d_2}\big(\frac{P_1^{\theta(i_1+1)+d_1}-Q_1^{\theta(i_1+1)+d_1}}{\theta(i_1+1)+d_1}\big) \\
\quad-\frac{P_1^{\theta(i_2+i_1+2)+d_2+d_1}-Q_1^{\theta(i_2+i_1+2)+d_2+d_1}}{\theta(i_2+i_1+2)+d_2+d_1}\Bigg\}, m\neq -1, \\
\\
\frac{k^s}{(r-1)!(s-r-1)!}\sum_{i=0}^{n-1}\sum_{d_2=0}^{\infty}\sum_{j_2=0}^{s-r-1}\sum_{i_2=0}^{d_2/\theta}\binom{s-r-1}{j_2}\binom{d_2/\theta}{i_2} \\
\times\ \frac{(-1)^{j_2+i_2+1}(P-Q)^{d_2/\theta}P_2^{d_2/\theta-i_2}t_2^{d_2}\Gamma(s-r-j_2)}{d_2!(i_2+k)^{s-r-j_2}i!}\sum_{d_1=0}^{\infty}\sum_{i_1=0}^{d_1/\theta}\binom{d_1/\theta}{i_1} \\
\times\ \frac{t_1^{d_1}(P-Q)^{d_1/\theta}P_2^{d_1/\theta-i_1}(-1)^{i_1}}{d_1!}\frac{\Gamma(r+j_2+i_1)}{(i_1+1)^{r+j_2+i_1}}, m=-1.
\end{cases}
\tag{2.16}
$$

Then, the (a,b)-th moment $(a,b=0,1,2,...)$ of the r-th and s-th generalized order statistics

$(r,s=1,2,...),\ r<s$, is given by

$$\mu_{r;s,m,k}^{(a,b)}=\begin{cases}\frac{c_{s-1}\theta^2}{(r-1)!(s-r-1)!(m+1)^{s-2}(P-Q)^{(m+1)(s-r-1)+\gamma_s+1}}\sum_{j_2=0}^{s-r-1}\sum_{i_2=0}^{(m+1)j_2+\gamma_s-1}\\ \times\binom{s-r-1}{j_2}\binom{(m+1)j_2+\gamma_s-1}{i_2}\sum_{j_1=0}^{r-1}\sum_{i_1=0}^{(m+1)(s-r-j_2+j_1)-1}\\ \times\binom{r-1}{j_1}\binom{(m+1)(s-r-j_2+j_1)-1}{i_1}\frac{(-1)^{j_2+i_2}P^{(m+1)(s-r-j_2+j_1)-1-i_1}(-1)^{j_1+i_1}}{(P-Q)^{(m+1)j_1+m}(\theta(i_2+1)+b)}\\ \times\{P_1^{\theta(i_2+1)+b}(\frac{P_1^{\theta(i_1+1)+a}-Q_1^{\theta(i_1+1)+a}}{\theta(i_1+1)+a})-\frac{P_1^{\theta(i_2+i_1+2)+b+a}-Q_1^{\theta(i_2+i_1+2)+b+a}}{\theta(i_2+i_1+2)+b+a}\},\\ \qquad\qquad m\neq-1,\\ \frac{k^s}{(r-1)!(s-r-1)!}\sum_{i=0}^{n-1}\sum_{j_2=0}^{s-r-1}\sum_{i_2=0}^{b/\theta}\sum_{i_1=0}^{a/\theta}\binom{a/\theta}{i_1}\binom{s-r-1}{j_2}\binom{b/\theta}{i_2}\\ \times\frac{(-1)^{j_2+i_2+1i_1}(P-Q)^{b/\theta+a/\theta}P_2^{b/\theta-i_2+a/\theta-i_1}}{i!}\ \times\ \frac{\Gamma(s-r-j_2+k)\Gamma(r+j_2+i_1)}{(i_2+1)^{s-r-j_2+k}(i_1+1)^{r+j_2+i_1}},\\ \qquad\qquad m=-1.\end{cases}\tag{2.17}$$

Special cases:

Form Theorem (2), we deduce some special cases as follows:

1. Setting $m=0$ and $k=1$ (OOS), $a,b\in N$ and $1\leq r<s\leq n-1, s-r\geq 2$, $n=1,2,\ldots$. Then from (2.17), we get

$$\begin{aligned}\mu_{r,s:n}^{(a,b)} &= \frac{c_{s-1}\theta^2}{(r-1)!(s-r-1)!(P-Q)^{n-r+1}}\\ &\times\sum_{j_2=0}^{s-r-1}\sum_{i_2=0}^{j_2+n-s}\binom{s-r-1}{j_2}\binom{j_2+n-s}{i_2}(-1)^{j_2+i_2}\\ &\times\sum_{j_1=0}^{r-1}\sum_{i_1=0}^{s-r-j_2+j_1-1}\binom{r-1}{j_1}\binom{s-r-j_2+j_1-1}{i_1}\\ &\times\frac{P^{s-r-j_2+j_1-1-i_1}(-1)^{j_1+i_1}}{(P-Q)^{j_1}(\theta(i_2+1)+b)}\\ &\times\{P_1^{\theta(i_2+1)+b}(\frac{P_1^{\theta(i_1+1)+a}-Q_1^{\theta(i_1+1)+a}}{\theta(i_1+1)+a})\\ &-\frac{P_1^{\theta(i_2+i_1+2)+b+a}-Q_1^{\theta(i_2+i_1+2)+b+a}}{\theta(i_2+i_1+2)+b+a}\}\end{aligned}\tag{2.18}$$

As a check, we set $b=0$ in (2.18), we get (2.4) that is $\mu_{r,s:n}^{(a,0)}=\mu_{r:n}^{(a)}$.

2. Setting $Q = 0$ (right truncated power function), $a, b \in N$ and $1 \leq r < s \leq n-1, s-r \geq 2$, $\frac{a}{\theta} + 1 > 1$ and $\frac{b}{\theta} + 1 > 1$, Then from (2.18), we get

$$\mu_{r,s:n}^{(a,b)} = P^{a/\theta+b/\theta} \frac{\Gamma(a/\theta + r)\Gamma(n+1)\Gamma(s + a/\theta + b/\theta)}{\Gamma(r)\Gamma(a/\theta + s)\Gamma(a/\theta + b/\theta + n + 1)}, \tag{2.19}$$

when $P = 1$ (complete power function), we get

$$\mu_{r,s:n}^{(a,b)} = \frac{\Gamma(a/\theta + r)\Gamma(n+1)\Gamma(s + a/\theta + b/\theta)}{\Gamma(r)\Gamma(a/\theta + s)\Gamma(a/\theta + b/\theta + n + 1)}. \tag{2.20}$$

The results (2.19) and (2.20) are given by Balakrishnan and Sultan (1998).

3. Setting $m = -1$ and $k = 1$ (RV) in (2.18), we get

$$\begin{aligned} \mu_{r,s:n}^{(a,b)} &= \frac{1}{(r-1)!(s-r-1)!} \\ &\times \sum_{i=0}^{n-1} \sum_{j_2=0}^{s-r-1} \sum_{i_2=0}^{b/\theta} \sum_{i_1=0}^{a/\theta} \binom{a/\theta}{i_1} \binom{s-r-1}{j_2} \binom{b/\theta}{i_2} \\ &\times \frac{(-1)^{j_2+i_2+1i_1}(P-Q)^{b/\theta+a/\theta} P_2^{b/\theta-i_2+a/\theta-i_1}}{i!} \\ &\times \frac{\Gamma(s-r-j_2+1)\Gamma(r+j_2+i_1)}{(i_2+1)^{s-r-j_2+1}(i_1+1)^{r+j_2+i_1}}. \end{aligned} \tag{2.21}$$

As a check, we set $b = 0$ in (2.21), we get $\mu_{r,s:n}^{(a,0)} = \mu_r^{(a)}$.

2.2 The doubly truncated Log-logistic distribution

A random variable X is said to have a doubly truncated Log-logistic distribution if its pdf is of the form

$$f(x) = \frac{\theta x^{\theta-1}}{(P-Q)(1+x^\theta)^2},\ 0 < Q_1 \leq x \leq P_1 < 1, \theta > 0, \tag{2.22}$$

where $P = P_1^\theta/(1+P_1^\theta)$ and $Q = Q_1^\theta/(1+Q_1^\theta)$. The cdf is of the form

$$F(x) = Q_2 - \frac{1}{(P-Q)(1+x^\theta)},\ Q_1 \leq x \leq P_1, \theta > 0, \tag{2.23}$$

where $Q_2 = \frac{1-Q}{P-Q}$ and $P_2 = \frac{1-P}{P-Q}$. The pdf and cdf of the Log-logistic distribution can be obtained, respectively, from (2.22) and (2.23) by setting $Q_1 = 0$ and $P_1 = 1$ [see Johnson, Kotz and Balakrishnan (1995)].

Single moments In the following theorem, we derive the exact form of the single moment of GOS from the doubly truncated Log-logistic distribution, then we given the corresponding moments.

Theorem (3):

Let $X_{1,n,m,k}, X_{1,n,m,k}, \ldots, X_{n,n,m,k}$ represent the GOS from the doubly truncated Log-logistic distribution, then the single moment of the r-th GOS is given by

$$\mu^{(a)}_{X_{r;n,m,k}} = \begin{cases} \frac{c_{r-1}}{(r-1)!(m+1)^{r-1}} \sum_{j=0}^{r-1} \sum_{i=0}^{\gamma_{(r-j)}-1} \binom{r-1}{j} \binom{\gamma_{(r-j)}-1}{i} \\ \times \frac{(-1)^{j+i} P_2^i}{(P-Q)^{\gamma_{(r-j)}-i}} B\left(a/\theta+1, \gamma_{(r-j)}-i-a/\theta\right) \\ \times \{I_{1-Q}\left(a/\theta+1, \gamma_{(r-j)}-i-a/\theta\right) \\ I_{1-P}\left(a/\theta+1, \gamma_{(r-j)}-i-a/\theta\right)\}, m \neq -1, \\ \\ \sum_{j=0}^{[a/\theta]} \binom{[a/\theta]}{j} \frac{(-1)^{[a/\theta]-j}}{P^j} \left(\frac{1}{P_2^j} - \frac{1}{(1+P_2)^j}\right), m = -1. \end{cases} \tag{2.24}$$

where $I_\alpha(a,b)$ is the incomplete beta ratio defined by

$$I_\alpha(a,b) = \frac{1}{B(a,b)} \int_0^\alpha t^{a-1}(1-t)^{b-1} dt. \tag{2.25}$$

Special cases: From Theorem (3), we deduce some special cases as follows:

1. Setting $m = 0$ and $k = 1$ (OOS) in (2.25), we get $c_{r-1} = \prod_{j=1}^{r} \gamma_j$ where $\gamma_r = n - r + 1$, $1 \leq r \leq n-1$, $n = 1, 2, ...$ and

$$\begin{aligned} \mu^{(a)}_{r:n} &= \frac{n!}{(r-1)!(n-r)!} \sum_{j=0}^{r-1} \sum_{i=0}^{n-r+j} \binom{r-1}{j} \binom{n-r+j}{i} \\ &\times \frac{(-1)^{j+i} P_2^i}{(P-Q)^{n-r+j-i}} B\left(a/\theta+1, n-r+j-i-a/\theta+1\right) \\ &\times \{I_{1-Q}\left(a/\theta+1, n-r+j-i-a/\theta+1\right) \\ &- I_{1-P}\left(a/\theta+1, n-r+j-i-a/\theta+1\right)\}, \end{aligned} \tag{2.26}$$

hence hen $n = r = 1$ for $a = 1, 2, ...,$ we get

$$\begin{aligned} \mu^{(a)}_{1:1} &= \frac{1}{(P-Q)} B\left(a/\theta+1, 1-a/\theta\right) \\ &\times \{I_{1-Q}\left(a/\theta+1, 1-a/\theta\right) - I_{1-P}\left(a/\theta+1, 1-a/\theta\right)\}. \end{aligned} \tag{2.27}$$

2. Setting $Q = 0$ (right truncated Log-logistic) and for $a = 1, 2, \ldots,$ $1 \leq r \leq n-1$, and $\frac{a}{\theta} + 1 > 1$. Then from (2.24), we have

$$\mu_{r:n}^{(a)} = \begin{cases} \frac{c_{r-1}}{(r-1)!(m+1)^{r-1}} \sum_{j=0}^{r-1} \sum_{i=0}^{\gamma_{(r-j)}-1} \binom{r-1}{j} \binom{\gamma_{(r-j)}-1}{i} \\ \quad \times \frac{(-1)^{j+i}(1-P)^i}{P^{\gamma_{(r-j)}}} B\left(a/\theta + 1, \gamma_{(r-j)} - i - a/\theta\right) \\ \quad \times \{B\left(a/\theta + 1, \gamma_{(r-j)} - i - a/\theta\right) \\ \quad I_{1-P}\left(a/\theta + 1, \gamma_{(r-j)} - i - a/\theta\right)\}, \ m \neq -1, \\ \\ \sum_{j=0}^{[a/\theta]} \binom{[a/\theta]}{j} \frac{(-1)^{[a/\theta]-j}}{P^j} \left(\frac{1}{P_2^j} - \frac{1}{(1+P_2)^j}\right), m = -1. \end{cases} \tag{2.28}$$

Then, when $m = 0$ and $k = 1$ (OOS), we get

$$\begin{aligned} \mu_{r:n}^{(a)} = E(X_{r;n}^a) &= \frac{n!}{(r-1)!(n-r)!} \sum_{j=0}^{r-1} \sum_{i=0}^{n-r+j} \binom{r-1}{j} \binom{n-r+j}{i} \\ &\times \frac{(-1)^{j+i}(1-P)^i}{P^{n-r+j}} B\left(a/\theta + 1, n - r + j - i - a/\theta + 1\right) \\ &\times B\left(a/\theta + 1, n - r + j - i - a/\theta + 1\right) \\ &- I_{1-P}\left(a/\theta + 1, n - r + j - i - a/\theta + 1\right)\}. \end{aligned} \tag{2.29}$$

3. Setting $P = 1$ (Log-logistic) and for $a = 1, 2, ..,$ and $1 \leq r \leq n-1$, $n = 1, 2, ...$ Then form (2.28), we have

$$\mu_{r:n,m,k}^{(a)} = E(X_{(r;n,m,k)}^a) = \begin{cases} \frac{c_{r-1}}{(r-1)!(m+1)^{r-1}} \sum_{j=0}^{r-1} \binom{r-1}{j} (-1)^j \\ \times B\left(a/\theta + 1, \gamma_{(r-j)} - a/\theta\right), \ m \neq -1, \\ \\ \sum_{j=0}^{a/\theta} \binom{a/\theta}{j} (-1)^j \left(\frac{k}{k - a/\theta + j}\right)^r, \ m = -1. \end{cases} \tag{2.30}$$

Hence, when $m = 0$ and $k = 1$ (OOS), we get

$$\begin{aligned} \mu_{r:n}^{(a)} = E(X_{r;n}^a) &= \frac{n!}{(r-1)!(n-r)!} \sum_{j=0}^{r-1} \binom{r-1}{j} (-1)^j \\ &\times B\left(a/\theta + 1, n - r + j - a/\theta + 1\right). \end{aligned} \tag{2.31}$$

From (2.31), when $n = r = 1$ and $a = 1, 2, \ldots,$ we get

$$\mu_{1:1}^{(a)} = E(X^a) = \Gamma\left(a/\theta + 1\right) \Gamma\left(1 - a/\theta\right), \tag{2.32}$$

and hence when $m=-1$ and $k=1$ (ORV),we get

$$\mu_{r:n}^{(a)}=\sum_{j=0}^{a/\theta}\binom{a/\theta}{j}(-1)^j\left(\frac{1}{1-a/\theta+j}\right)^r. \tag{2.33}$$

Double moments In the following theorem, we derive the double moments of the GOS from the Log-logistic distribution.

Theorem (4):

Let $X_{r,n,m,k}$ and $X_{s,n,m,k}$, $r<s$ be the r-th and s-th GOS from the Log-logistics distribution, then the (a,b)-th double moment is given by

$$\mu_{r;s,m,k}^{(a,b)}=\begin{cases}\frac{c_{s-1}}{(r-1)!(s-r-1)!(m+1)^{s-2}}\sum_{j=0}^{s-r-1}\sum_{i=0}^{r-1}\binom{s-r-1}{j}\binom{r-1}{i}(-1)^{j+i}\\ \times\ \beta(b/\theta+1,\gamma_{s-j}-b/\theta)\beta(a/\theta+1,\gamma_{s-r-j+i}-a/\theta),\\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad m\neq -1,\\ k^s\sum_{i=0}^{a/\theta}\sum_{j=0}^{b/\theta}\binom{a/\theta}{i}\binom{b/\theta}{j}\frac{(-1)^{i+j}}{(k-b/\theta+i)^{s-r}(k-a/\theta+i)^r},\ m=-1.\end{cases} \tag{2.34}$$

Special cases:

From Theorem (4), we deduce some special cases as follows:

1. Setting $m=0$ and $k=1$ (OOS) in (2.34), we get

$$\begin{aligned}\mu_{r:n}^{(a,b)} &= \frac{n!}{(n-s)!(r-1)!(s-r-1)!}\sum_{j=0}^{s-r-1}\sum_{i=0}^{r-1}\binom{s-r-1}{j}\binom{r-1}{i}(-1)^{j+i}\\ &\quad B(b/\theta+1,n-s-b/\theta+1)B(a/\theta+1,n-s+r+j-i-a/\theta+1).\end{aligned} \tag{2.35}$$

2. Setting $m=-1$ and $k=1$ (RV) in (2.34), we get

$$\mu_{r;s}^{(a,b)} = \sum_{i=0}^{a/\theta}\sum_{j=0}^{b/\theta}\binom{a/\theta}{i}\binom{b/\theta}{j}\frac{(-1)^{i+j}}{(1-b/\theta+i)^{s-r}(1-a/\theta+i)^r}. \tag{2.36}$$

3. RECURRENCE RELATIONS BASED ON GOS FROM DOUBLY TRUNCATED POWER FUNCTION

In this section, we establish some new recurrence relation between the single moments of GOS from the doubly truncated power function distribution by using the hypergeometric

function. By using the relations between the pdf and cdf of the doubly truncated power function distribution we can write

$$x = (P-Q)^{1/\theta}[Q_2 + F_d(x)]^{1/\theta}. \tag{3.1}$$

Then from (3.1), the single moments of GOS from doubly truncated power function can be derived in terms of the hypergeometric function as

$$\mu^{(a)}_{r,n,m,k} = \frac{c_{r-1}P^{-A}}{(r-1)!(m+1)^{r-1}} \sum_{j=0}^{r-1} \binom{r-1}{j} (-1)^j \frac{1}{B}\ {}_2F_1(A, B; B+1; z), \tag{3.2}$$

where $z = \frac{1}{P_2}$, $A = -a/\theta$, $B = \gamma_{r-j} = k + (n-(r-j))(m+1)$ and ${}_2F_1(A, B; B+1; z)$ is the hypergeometric function.

By using the properties of the hypergeometric function ${}_2F_1(a, b, c, z)$ given in Rainville (1971), establish some new recurrence relations as given in the following theorem.

Theorem (5):

The single moments of GOS from the doubly truncated power function distribution stratify the following recurrence relations:

$$(\gamma_r + a/\theta)\mu^{(a)}_{r;n,m,k} = a/\theta P \mu^{(a-\theta)}_{r;n,m,k} + \gamma_r \mu^{(a)}_{r+1;n,m,k} \tag{3.3}$$

$$(\gamma_1 + a/\theta)\mu^{(a)}_{r;n,m,k} = a/\theta P, \mu^{(a-\theta)}_{r;n,m,k} + \gamma_1 \mu^{(a)}_{r-1;n-1,m,k}, \tag{3.4}$$

$$(\gamma_r + a/\theta)Q\mu^{(a)}_{r;n,m,k} = (k + n(m+1) + a/\theta)\mu^{(a+\theta)}_{r;n,m,k} - r(m+1)P\mu^{(a)}_{r+1;n,m,k}, \tag{3.5}$$

$$(k + n(m+1))Q\mu^{(a)}_{r;n,m,k} = (k + n(m+1))P\mu^{(a-\theta)}_{r;n,m,k} - \frac{rP}{P_2}\mu^{(a)}_{r+1;n+1,m,k}, \tag{3.6}$$

where $m \neq -1$.

4. ESTIMATION BASED ON GOS FROM LOG-LOGISTIC DISTRIBUTION

Let X be a random variable from the the tow-parameter Log-logistic distribution with scale parameter σ as

$$f(x;\sigma) = \frac{\theta(\frac{x}{\sigma})^{\theta-1}}{\sigma\left(1 + (\frac{x}{\sigma})^{\theta}\right)^2}, x \geq 0, \theta \geq 1, \sigma > 0, \tag{4.1}$$

and the cdf is

$$F(x;\sigma) = 1 - \frac{1}{1+(\frac{x}{\sigma})^{\theta}}, , x \geq 0, \theta \geq 1, \sigma > 0. \tag{4.2}$$

when $\sigma = 1$, then the pdf of Log-logistic distribution given in (4.1) reduces to the one-parameter Log-logistic distribution given in (3.23)

4.1 **Best linear unbiased estimation (BLUE)**

Let $X_{1,n,m,k} \leq X_{2;n,m,k} \leq ... \leq X_{n,n,m,k}$ denote the available GOS from the Log-logistic distribution in (4.1), and let $Z_{i,n,m,k} = X_{i,n,m,k}/\sigma, i = 1, 2, ..., n$ be the corresponding GOS from the one-parameter Log-logistic distribution. Let us denote $E(Z_{i,n,m,k})$ by $\mu_{i,n,m,k}$, $Var(Z_{i,n,m,k})$ by $\sigma_{i,i,n,m,k}$ and $Cov(Z_{i,n,m,k}, Z_{j,n,m,k})$ by $\sigma_{i,j,n,m,k}$. Further, let

$$\mathbf{X} = (X_{1,n,m,k}, X_{2,n,m,k}, ..., X_{n,n,m,k})^T,$$

$$\mu = (\mu_{1,n,m,k}, \mu_{2,n,m,k}, ..., \mu_{n,n,m,k})^T$$

$$\Omega = ((\sigma_{i,j,n,m,k})), 1 \leq i, j \leq n, \tag{4.3}$$

where Ω is a positive definite symmetric matrix of order n.
Then, the best linear unbiased estimate (BLUE) of σ is given by [see Balakrishnan and Cohen (1991) and Arnold, Balakrishnan and Nagarja (1998)]

$$\sigma^* = \{\frac{\mu^T\Omega^{-1}}{\mu^T\Omega^{-1}\mu}\}X = \sum_{i=1}^{n} a_i X_{i,n,m,k} \tag{4.4}$$

Furthermore, the variance of this BLUE is given by

$$Var\{\sigma^*\} = \frac{\sigma^2}{\mu^T\Omega^{-1}\mu}. \tag{4.5}$$

By using the exact explicit expressions of the moments of the GOS from the one-parameter Log-logistic distribution given in (2.30) and (2.35), we have

1. If $m \neq -1$ for $1 \leq r \leq n-1$ and $k = 1, 2, ...$

$$\mu_{i,n,m,k} = \frac{c_{i-1}}{(i-1)!(m+1)^{i-1}} \sum_{\ell=0}^{i-1} \binom{i-1}{\ell} (-1)^{\ell} \beta(\frac{1}{\theta} + 1, \gamma_{(i-\ell)} - \frac{1}{\theta})$$

$$\mu^{(2)}_{i;n,m,k} = \frac{c_{i-1}}{(i-1)!(m+1)^{i-1}} \sum_{\ell=0}^{i-1} \binom{i-1}{\ell} (-1)^{\ell} \beta(\frac{2}{\theta} + 1, \gamma_{(i-\ell)} - \frac{2}{\theta})$$

and

$$\begin{aligned}\mu_{i,j;n,m,k} &= \frac{c_{j-1}}{(i-1)!(j-i-1)!(m+1)^{j-2}} \\ &\times \sum_{\ell_1=0}^{j-i-1}\sum_{\ell_2=0}^{i-1}\binom{j-i-1}{\ell_2}\binom{i-1}{\ell_1}(-1)^{\ell_1+\ell_2} \\ &\times \beta(\frac{1}{\theta}+1,\gamma_{(j-\ell_2)}-\frac{1}{\theta})\beta(\frac{1}{\theta}+1,(m+1)(j-i+\ell_1-\ell_2)-\frac{1}{\theta}),\end{aligned}$$

where $\gamma_r = k+(n-r)(m+1)$, $c_{r-1} = \prod_{i=1}^{r}\gamma_i$ and $\beta(.,.)$ is the beta function. The moments of order statistics OOS can be obtained from (4.6), (4.7) and (4.8), then we use these moments to calculate the coefficients of BLUE from (4.4). Table (1) represents the BLUE of the scale parameter of the Log-logistic distribution when $n = 5, 15, 20, 25, 30$ and $\theta = 4$.

2. If $m = -1$, $k = 1$, we get the moments of the upper RV

$$\mu_{U(i)} = \sum_{\ell=0}^{\delta_1}\binom{\delta_1}{\ell}(-1)^{\ell}(\frac{1}{1+\ell-\delta_1})^i, \tag{4.6}$$

$$\mu^{(2)}_{U(i)} = \sum_{\ell=0}^{\delta_2}\binom{\delta_2}{\ell}(-1)^{\ell}(\frac{1}{1+\ell-\delta_2})^i, \tag{4.7}$$

and

$$\mu_{U(i,j)} = \sum_{\ell_2=0}^{\delta_1}\sum_{\ell_1=0}^{\delta_1}\binom{\delta_1}{\ell_1}\binom{\delta_1}{\ell_2}\frac{(-1)^{\ell_2+\ell_1}(\ell_1+1-\delta_1)^i}{(1+\ell_1-\delta_1)^j(1+\ell_1+\ell_2-\delta_2)^i}. \tag{4.8}$$

Example: In this example, sample of 5 order statistics is generated from the Log-logistic distribution with $\sigma = 1$, $\theta = 4$ as: 0.2817, 0.5489, 0.9496, 1.2863, 2.9003. Next, and use the entries of Table (1), we have

$$\begin{aligned}\sigma^* &= 0.2817\times 0.0729 + 0.1759\times 0.5489 + 0.2543\times 0.9496 \\ &\quad +0.2470\times 1.2863 + 0.1298\times 2.900 = 1.05.\end{aligned}$$

Table 1: The Coefficient of the BLUE $a_i,\ i = 1, 2, \ldots n$, when $\theta = 4$ and $\sigma = 1$.

$n = 5$	$n = 10$	$n = 15$	$n = 20$	$n = 25$	$n = 30$
0.0729	0.1298	0.0146	0.0010	0.0001	0.0000
0.1759	0.0073	0.0007	0.0001	0.0000	0.0000
0.2543	0.0291	0.0041	0.0005	0.0001	0.0000
0.2470	0.0695	0.0135	0.0021	0.0003	0.0002
0.1298	0.1230	0.0328	0.0065	0.0011	0.0006
	0.1717	0.0637	0.0159	0.0033	0.0017
	0.1903	0.1032	0.0325	0.0079	0.0040
	0.1643	0.1417	0.0572	0.0167	0.0086
	0.1071	0.1647	0.0882	0.0308	0.0165
	0.0503	0.1601	0.1199	0.0510	0.0287
		0.1283	0.1431	0.0763	0.0456
		0.0833	0.1490	0.1029	0.0666
		0.0429	0.1339	0.1250	0.0893
		0.0172	0.1028	0.1359	0.1098
		0.0051	0.0667	0.1314	0.1231
			0.0361	0.1122	0.1252
			0.0161	0.0839	0.1149
			0.0058	0.0546	0.0946
			0.0016	0.0306	0.0695
			0.0004	0.0146	0.0453
				0.0059	0.0260
				0.0020	0.0131
				0.0005	0.0057
				0.0001	0.0022
					0.0007
					0.0002
					0.0000
					0.0000
					0.0000
					0.0000

4.2 MLE

The likelihood function L based on the n GOS from the Log-logistic distribution can be written as

$$\begin{aligned} L(\sigma|\underline{X}) &= k\left(\prod_{j=1}^{n-1}\gamma_j\right)\left(\prod_{i=1}^{n-1}(1-F(x_i))^m f(x_i)\right)(1-F(x_n))^{k-1}f(x_n) \\ &= \left(\prod_{j=1}^{n}\gamma_j\right)\prod_{i=1}^{n-1}\left(\frac{\theta(\frac{x_i}{\sigma})^{\theta-1}}{\sigma\left(1+(\frac{x_i}{\sigma})^{\theta}\right)^{m+2}}\right)\frac{\theta(\frac{x_n}{\sigma})^{\theta-1}}{\sigma\left(1+(\frac{x_n}{\sigma})^{\theta}\right)^{k+1}} \end{aligned} \tag{4.9}$$

where $\underline{X} = (X_{(1,n,m,k)}, ..., X_{(n,n,m,k)})$ and $\gamma_n = k$. Hence

$$\begin{aligned} \log L(\sigma|\underline{X}) &= \sum_{j=1}^{n}\log(\gamma_j) + n(\log(\theta)-\log(\sigma)) + (\theta-1)\sum_{i=1}^{n}(ln(x_i)-ln(\sigma)) \\ &- (m+2)\sum_{i=1}^{n-1}\log\left(1+(\frac{x_i}{\sigma})^{\theta}\right) - (k+1)ln\left(1+(\frac{x_n}{\sigma})^{\theta}\right) \end{aligned} \tag{4.10}$$

Differentiating with respect to σ, then the MLE of σ can be obtained by solving the following nonlinear equation

$$-n + (m+2)\sum_{i=1}^{n-1}\left(\frac{(\frac{x_i}{\sigma})^{\theta}}{1+(\frac{x_i}{\sigma})^{\theta}}\right) + (k+1)\left(\frac{(\frac{x_n}{\sigma})^{\theta}}{1+(\frac{x_n}{\sigma})^{\theta}}\right) = 0. \tag{4.11}$$

From (4.14), we have

1. If $m = 0$ and $k = 1$ (OOS)

$$\sum_{i=1}^{n}\{\frac{(\frac{x_i}{\sigma})^{\theta}}{1+(\frac{x_i}{\sigma})^{\theta}}\} = \frac{n}{2} \tag{4.12}$$

2. If $m = -1$ and $k = 1$ (RV)

$$\sum_{i=1}^{n-1}\{\frac{(\frac{x_i}{\sigma})^{\theta}}{1+(\frac{x_i}{\sigma})^{\theta}}\} + 2\{\frac{(\frac{x_n}{\sigma})^{\theta}}{1+(\frac{x_n}{\sigma})^{\theta}}\} = n. \tag{4.13}$$

The MSE of the MLE and the BLUE of the scale parameter based on order statistics from the Log-logistic distribution are calculated in Table (2). Similar argument can be done based on RV from the Log-logistic distribution.

From the Table (2), we see both of the MLE and BLUE of the scale parameter are perform well in small and large sample of order statistics from the Log-logistic distribution. Also both estimates are consistent in mean squared error. The BLUE behave quite better than the MLE starting from $n = 10$.

Table 2: Table (2): The MSEs of BLUE and MLE of the scale parameter

n	$MSE(\hat{\sigma})$	$MSE(\sigma^*)$
5	0.04532	0.00810
10	0.02045	0.00019
15	0.01363	0.00001
20	0.00948	0.00001
25	0.00765	0.00001
30	0.00637	0.00001

REFERENCES

Abu-Youssef S.E. (2003). On characterization of certain distributions of record values, *Appl. Math. Comput.*, **145**, 443-450.

Ahmad, A.E.A and Fawzy, M.A.(2003). Recurrence relations for single moments of generalized order statistics from doubly truncated distributions, *J. Statist. Plann. Inf.*, **117**, 241-249.

Ahsanullah, M.(1982). Characterizations of the exponential distribution by some properties of the record values, *Statist. Hefte*, **23**, 326-332.

Ahsanullah, M. (1988). *Introduction to Record Values*, Ginn Press, Needham Heights, Massachusetts.

Ahsanullah, M. (1990). Some characterizations of the exponential distribution by first moment of record values, *Pakistan J. Statist.G.*, 183-188.

Ahsanullah, M.(1991). Some characteristic properties of the record values from exponential distribution, *Sankhyā Ser. B*, **53** , 403-408.

Ahsanullah, M., (1996). Generalized order statistics from two parameter uniform distribution, *Comm. Statist. Theor. Meth.*, **25**, 2311-2318.

Ahsanullah, M., (1997). Generalized order statistics from power function distribution, *J. Appl. Statist. Sci*, **5**, 283-290.

Ahsanullah, M., (2000). Generalized order statistics from exponential distribution, *J. Statist. Plann. Inf.*, **85**, 85-91.

Ahsanullah, M. and Kirmani, S.N.U.A. (1991). Characterizations of the exponential distribution through a lower record, *Comm. Statist. Theor. Meth.*, **20**, 1293-1299.

AL-Hussaini, E.K. and Ahmad, A.A. (2003a). On Bayesian predictive distribution of generalized order statistics, *Metrika*, **57**, 165-176.

AL-Hussaini, E.K. and Ahmad, A.A. (2003b). On Bayesian interval prediction of future records, *Test,* **12**, 79-99.

AL-Hussaini, E.K., Ahmad, A.A. and El-Kashif, M. (2005). Recurrence relations for moment generating functions of order statistics, *Metron*, **LXII**, 85-99.

Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1992). *A First Course in Order Statistics*, John Wiley & Sons, New York.

Balakrishnan, N. and Ahsanullah, M. (1994). Recurrence relations for single and product moments of record values from exponential generalized Pareto distribution, *Comm. Statist. Theor. Meth.*, **23 (10)**, 2841- 2852.

Balakrishnan, N. and Aggarwala, R. (2000). Progressive Censoring: Theory, Methods and Applications. Birkhauser, Boston.

Balakrishnan, N. and Cohen, A. C. (1991). *Order Statistics and Inference: Estimation Methods*, Academic Press, San Diego.

Balakrishnan, N. and Sultan, K. S. (1998). Recurrence relations and identities for moments of order statistics. In Balakrishnan, N., Rao, C.R. (Eds.), Handbook of Statistics, 16, *Order Statistics: Theory and Methods*. North-Holland, Amsterdam, pp. 149-228.

Gupta, R.C. (1984). Relationships between order statistics and record values and some characterization results, *J. Appl. Prob.*, **21**, 425- 430.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, **Vol. 1**, Second edition, John Wiley & Sons, New York.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, **Vol. 2**, Second edition, John Wiley & Sons, New York.

Kamps, U. (1995a). *A Concept of Generalized Order Statistics*, Teubner, Stuttgart.

Kamps, U. (1995b). A concept of generalized order statistics, *J. Statist. Plann. Inf.*, 48, 123.

Kamps, U. (1999). Order statistics, generalized, In: Kotz, S., Read, C.B., Banks, D.L. (Eds.), *Encyclopedia of Statistical Sciences, Update Vol* **3***, Wiley, New York*, pp. 553-557.

Keseling, C. (1999). Conditional distributions of generalized order statistics and some characterizations, *Metrika*, **49**, 2740.

Lin, G.D (1987). On characterization of distribution via moments of record values, *Prob. Theor. Relat. Fields*, **74**, 479- 483.

Nagaraja, H.N. (1977). On characterization based on record values, *Aust. J. Statist.*, **19**, 1, 70- 73.

Rainville, E. D. (1971). *Special Functions*, Chelsea Pub. Co., Bronx, New York.

Saran, J. and Pandey, A. (2003). Recurrence relations for marginal and joint moment generating functions of generalized order statistics from power function distribution. *Metron-Inter. J. Statist*, IXI, 27-33.

Sultan, K.S. and Abd El-Mougod G.A. (2006). Characterization of general classes of doubly truncated distributions based on record values, *J. Prob. Statist. Sc.*, **4(1)**, 65-72.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 1070–1078.

ESTIMATING THE EFFECT OF THE 1997 ECONOMIC CRISIS ON THE DEMAND FOR HOUSE CHARACTERISTICS IN RURAL INDONESIA

Yusep Suparman
Statistics Department, Padjadjaran University
Jl. Ir. H. Juanda no. 4, Bandung 40115 - Indonesia
E-mail: yusep.suparman@unpad.ac.id

ABSTRACT

In 1997, Indonesia suffered an economic crisis. This crisis had a great impact on the livelihoods in Indonesia. In this paper we estimate the effect of the crisis on demand for house characteristics. We express the demand on house characteristics in terms of willingness to pay, which is elicited from a hedonic price model. To estimate the crisis effect, we adopt a continuous time modeling approach, namely the exact discrete time– structural equation model (EDM-SEM), and apply it to a three-wave panel data set of Indonesia Family Life Survey (IFLS). We show that due to the crisis, the average valuation on certain house characteristics is reduced by 46.52%.

Keywords: Exact discrete time model, structural equation model, panel data, hedonic price model, economic crisis effect.

1. INTRODUCTION

Starting in September 1997, a crisis hit the Indonesian economy and lasted until the beginning of 1999. This crisis resulted in a high inflation rate. During 1998, inflation reached 77.63% (BPS, 2001). This high inflation rate had a great impact on Indonesian livelihoods. Particularly at the household level, the crisis decreased household real income significantly. Generally, a decrease of income will decrease the demand on goods and services, including the demand on house characteristics.

We observed that only one study of demand on house characteristics during an economic crisis, namely Suparman et al. (2008). They conducted a hedonic price study to estimate the demand of house characteristics, particularly in-house piped water service, which are expressed in the terms of willingness to pay (WTP). By means of a discrete time model, the estimation was based on three-wave Indonesia Family Life Survey (IFLS) panel data, whose time interval covered the 1997 crisis. Suparman et al. accommodated the crisis effect implicitly by allowing the intercepts in the model to be varied. No explicit crisis effect parameter was defined in their model.

In the current paper we aim to estimate the effect of the 1997 economic crisis on the demand on house characteristics by applying a continuous time model, namely the exact discrete time – structural equation model (EDM-SEM) (Oud and Jansen, 2000). We use the same data set as Suparman et al. (2008) i.e. the three-wave IFLS panel data set and extend their model by introducing continuous time parameters and constraints and crisis effect parameters.

The paper is organized as follow. In section 2, we discuss the EDM-SEM. The proposed hedonic price model for estimating the 1997 economic crisis effect is presented in section 3. We provide the empirical results and discussion in section 4. Section 5 concludes and summarizes our findings.

2. EXACT DISCRETE TIME MODEL - STRUCTURAL EQUATION MODEL

In econometrics, dynamic models are typically specified as discrete time (DT) models, although several authors including Koopmans (1950), Gandolfo (1981), Bergstrom (1988), and Phillips (1993) have strongly recommended continuous time (CT) models. The arguments presented in these sources to use CT models instead of DT models can be summarized as follows. First, real life socioeconomic processes evolve in CT, since they are the outcomes of large numbers of decisions concluded at different points in time. By their very nature, continuously evolving processes are more adequately represented by CT models than by DT models. Secondly, modeling results should not depend on the length of the observation interval. In DT modeling, this condition is not usually met, as the coefficients of a model estimated on the basis of, e.g. weekly data, will typically differ from those estimated on the basis of monthly or yearly data. This follows from the fact that the impact of an intervention will typically vary over the adjustment interval. For instance, if the adjustment process tails off, impacts measured at short intervals will be stronger than impacts measured at longer intervals. For adjustment processes that cut off, no impact may be measured for intervals longer than the adjustment interval. Furthermore, the sign of an effect may reverse when passing from one interval to another, which gives rise to what Oud (2002) refers to as the "paradox of DT modeling".

DT models which are made up of systems of difference equations are formulated in relation to the data available, for instance yearly or monthly models. In contrast, CT models depart from the assumption that there is no obvious time interval that can serve as a natural unit. CT models analyze the continuous nature of social processes by means of systems of differential equations.

A CT state space model describes the development or trajectory of an *n*-dimensional state vector $\mathbf{x}(t)$ over time. Oud and Jansen (2000) present the following system of stochastic differential equations to describe the trajectory of $\mathbf{x}(t)$:

$$\frac{d\mathbf{x}(t)}{d(t)} = \mathbf{A}(t)\mathbf{x}(t) + \boldsymbol{\gamma} + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{G}(t)\frac{d\mathbf{W}(t)}{dt}, \qquad (1)$$

where $\mathbf{A}(t)$, the drift matrix, models the changes in $\mathbf{x}(t)$ as functions of the state variables themselves while $\mathbf{B}(t)$ represents the impacts of fixed input variables $\mathbf{u}(t)$ on the state variables and accommodates for nonzero and nonconstant mean trajectories $\mathrm{E}[\mathbf{x}(t)]$. In addition to the unit variable (1 for all subjects and time points) $\mathbf{u}(t)$ may contain other constant or nonconstant exogenous variables, for example gender, or socioeconomic status. The trait variables $\boldsymbol{\gamma}$, which are unobserved and assumed to be constant over time, specify random subject effects which keep a subject-specific distance from $\mathrm{E}[\mathbf{x}(t)]$. The zero mean normally distributed trait variables can be viewed as a special kind of state variables, i.e. unobserved and constant over time. $\mathbf{W}(t)$ is a standard multivariate Wiener process with an identity covariance matrix. The standard Wiener

process is transformed into a general Wiener process with an arbitrary covariance matrix $\mathbf{G}(t)\mathbf{G}(t)'$ by means of a Cholesky factor $\mathbf{G}(t)$. $\mathbf{\Phi}_{\gamma}$ and $\mathbf{\Phi}_{\gamma,\mathbf{x}_{t_0}}$ are the trait variables covariance matrix and the trait and the initial state variables covariance matrix, respectively.

Since the time period in the application below is rather short, we assume time invariant parameter matrices in (1), i.e. $\mathbf{A}(t)=\mathbf{A}$, $\mathbf{B}(t)=\mathbf{B}$ and $\mathbf{G}(t)=\mathbf{G}$. This gives:

$$\frac{d\mathbf{x}(t)}{d(t)} = \mathbf{A}\mathbf{x}(t)+\boldsymbol{\gamma}+\mathbf{B}\mathbf{u}(t)+\mathbf{G}\frac{d\mathbf{W}(t)}{dt}. \tag{2}$$

To estimates the CT parameter in (2) based on DT data, we first derive the relations between the CT and DT parameters by means of the exact discrete time model (EDM) (Hamerle, Nagl & Singer, 1993; Oud & Jansen, 2000). For (2), with a discrete time interval Δt, the EDM yields:

$$\mathbf{x}_t = \mathbf{A}_{\Delta t}\mathbf{x}_{t-\Delta t}+\boldsymbol{\gamma}_{\Delta t}+\mathbf{B}_{\Delta t}\mathbf{u}_{t-\Delta t}+\mathbf{w}_{t-\Delta t}, \tag{3}$$

With $\mathrm{cov}(\mathbf{w}_{t-\Delta t})=\mathbf{Q}_{\Delta t}$ and

$$\mathbf{A}_{\Delta t} = \mathrm{e}^{\mathbf{A}\Delta t}, \tag{4a}$$

$$\mathbf{B}_{\Delta t} = \mathbf{A}^{-1}(\mathbf{A}_{\Delta t}-\mathbf{I})\mathbf{B}, \tag{4b}$$

$$\mathbf{H}_{\Delta t} = \mathbf{A}^{-1}(\mathbf{A}_{\Delta t}-\mathbf{I}), \tag{4c}$$

$$\mathbf{\Phi}_{\gamma\Delta t} = \mathbf{H}_{\Delta t}\mathbf{\Phi}_{\gamma}\mathbf{H}'_{\Delta t}, \tag{4d}$$

$$\mathbf{\Phi}_{\gamma_{\Delta t},\mathbf{x}_{t_0}} = \mathbf{H}_{\Delta t}\mathbf{\Phi}_{\gamma,\mathbf{x}_{t_0}}, \tag{4e}$$

$$\mathbf{Q}_{\Delta t} = \mathrm{irow}\left[(\mathbf{A}\otimes\mathbf{I}+\mathbf{I}\otimes\mathbf{A})^{-1}(\mathbf{A}_{\Delta t}\otimes\mathbf{A}_{\Delta t}-\mathbf{I}\otimes\mathbf{I})row(\mathbf{G}\mathbf{G}')\right]. \tag{4f}$$

$\boldsymbol{\gamma}_{\Delta t}=\mathbf{H}_{\Delta t}\boldsymbol{\gamma}$ is the DT trait vector with covariance matrix $\mathbf{\Phi}_{\gamma\Delta t}$. Moreover, $\mathbf{\Phi}_{\gamma_{\Delta t},\mathbf{x}_{t_0}}$ is the covariance matrix of the DT traits and the initial states. $\otimes$ is the Kronecker product operation and irow is the inverse operation of row which puts the elements of a matrix row-wise in a column vector. $\mathbf{B}_{\Delta t}$ is obtained on the assumption that the input variables $\mathbf{u}(t)$ are piecewise constant between measurements. Observe that all the expressions in (4) involve the nonlinear matrix restriction $\mathrm{e}^{\mathbf{A}\Delta t}$.

The discrete time point t of the EDM takes values in the set $\{t_0, t_0+\Delta t_1, \ldots, t_{T-2}+\Delta t_{T-1}\}$ for integers t_0 and $T\geq 2$, with t_0 the initial time point and T the total number of time points considered.

In many cases the state or input variables cannot be observed directly, i.e. they are latent. The state or input variables are represented by one or more observed variables or indicators. In that case, an output or measurement model that defines the relationships between latent state variables and their indicators have to be added to the EDM. For a point time t, let $\mathbf{y}_t$ be the vector of indicators, $\mathbf{x}_t$ be the vector of latent state variables, $\mathbf{u}_t$ be the vector of input variables, $\mathbf{C}_t$ and $\mathbf{D}_t$ be the matrices in which the relations between indicators and the latent state and input variables are defined respectively, and $\mathbf{v}_t$ the vector of measurement errors, then the measurement model reads:

$$\mathbf{y}_t = \mathbf{C}_t\mathbf{x}_t+\mathbf{D}_t\mathbf{u}_t+\mathbf{v}_t, \text{ with } \mathrm{cov}(\mathbf{v}_t)=\mathbf{R}_t \tag{5}$$

Measurement model (5) combined with the structural model (3) form a structural equation model (SEM). Following the SEM model, we apply the assumptions that the elements of $\mathbf{v}_t$ are normally distributed and they are uncorrelated with the state variables, $\mathrm{E}(\mathbf{v}_t\mathbf{x}_t') = \mathbf{0}$ for all t and t'.

For model estimation by means of SEM, we remove $\boldsymbol{\gamma}_{\Delta t}$ from (3). Instead, we add $\boldsymbol{\gamma}$ to the state vector and specify that its element to be identical at both sides of the equation. Accordingly, the parameter matrices should be extended. Respectively, the extended initial state mean and covariance matrix become:

$$\boldsymbol{\mu}_{t_0} = \begin{bmatrix} \mathbf{E}(\mathbf{x}_{t_0}) \\ \mathbf{0} \end{bmatrix}, \ \boldsymbol{\Phi}_{t_0} = \begin{bmatrix} \boldsymbol{\Phi}_{\mathbf{x}_{t_0}} & \boldsymbol{\Phi}_{\mathbf{x}_{t_0},\boldsymbol{\gamma}} \\ \boldsymbol{\Phi}'_{\mathbf{x}_{t_0},\boldsymbol{\gamma}} & \boldsymbol{\Phi}_{\boldsymbol{\gamma}} \end{bmatrix}. \tag{6}$$

Denoting the extended state vector $\bar{\mathbf{x}}_t' = (\mathbf{x}_t' \quad \boldsymbol{\gamma}')$ and $\mathbf{u}$ as the fixed input vector, in which all input variables over all time points are combined, except identical and linearly dependent input variables are specified only once. Next, we denote

$$\boldsymbol{\eta} = [\mathbf{u}' \quad \mathbf{x}']' \text{ with } \mathbf{x} = [\bar{\mathbf{x}}_{t_0}' \quad \bar{\mathbf{x}}_{t_1}' \quad \cdots \quad \bar{\mathbf{x}}_{t_{T-1}}']' \qquad \text{and} \qquad \mathbf{y} = [\mathbf{u}' \quad \mathbf{y}_0']' \text{ with } \mathbf{y}_0 = [\mathbf{y}_{t_0}' \quad \mathbf{y}_{t_1}' \quad \cdots \quad \mathbf{y}_{t_T}']'.$$

Hence, the state model (3) and (5) may be written as the SEM form

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta}, \text{ with } \boldsymbol{\Psi} = \mathrm{E}(\boldsymbol{\zeta}\boldsymbol{\zeta}'), \tag{7}$$

$$\mathbf{y} = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \text{ with } \boldsymbol{\Theta} = \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'). \tag{8}$$

All of the state model parameters are put in the SEM parameter matrices $\mathbf{B}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$ and $\boldsymbol{\Theta}$ (Oud & Jansen, 1996).

The related model implied covariance matrix of (7) and (8) is

$$\boldsymbol{\Sigma} = \mathrm{E}(\mathbf{y}\mathbf{y}') = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Psi}(\mathbf{I} - \mathbf{B}')^{-1} + \boldsymbol{\Theta}. \tag{9}$$

The corresponding sample moment matrix is $\mathbf{S}_{(q+p_0)\times(q+p_0)} = \frac{1}{N}\mathbf{Y}\mathbf{Y}'$ for the data in $\mathbf{Y} = [\mathbf{U}' \quad \mathbf{Y}_0']$. q is the number of fixed element in $\mathbf{u}$ and $p_0 = pT$ is the number of elements in observed random vector $\mathbf{y}_0$. The span of $\{\mathbf{u}_i\}$ for $i = 1,2,\ldots,N$ is to be not less than q-dimensional. If we use the maximum likelihood estimation method, the parameter estimates will be obtained by minimizing

$$F_{ML} = \log|\boldsymbol{\Sigma}| + \mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) - \log|\mathbf{S}| - (q + p_0). \tag{10}$$

3. HEDONIC PRICE MODEL

We apply the EDM-SEM procedure outlined above to estimate the WTP for house characteristics in rural areas in Indonesia (Suparman et al., 2008). There is only one directly observed state variable in the model, i.e. monthly house rent. This implies an identity relationship between the latent and the observed rent variable: $\eta_7 = y_9$. The input variables included in the model are the latent variable household characteristics (η_1) measured by two

observables, viz. household size (y_1) and household monthly expenditure (y_2); the latent variable house size (η_2) measured by the observables floor area (y_3) and number of rooms (y_4). The other latent explanatory variables are identical to their indicators. Specifically, they are house condition index $(\eta_3 = y_5)$, in house tap water $(\eta_4 = y_6)$, well water $(\eta_5 = y_7)$, and finally, the neighborhood characteristics represented by median household monthly expenditure: $(\eta_6 = y_8)$.

The data set is a three-wave panel dataset of 1315 unit observations, collected at $t_0 = 1993$, $t_1 = 1997$ and $t_2 = 2000$. Accordingly, we have $\Delta t_1 = 4$ and $\Delta t_2 = 3$. For a discrete time point t_i, Suparman et al. (2008) present the following hedonic price model which is formulated to account for omitted variables:

$$\eta_{7t_i} = \beta_{0t_{i-1}} + \sum_{j=1}^{6} \beta_{jt_i} \eta_{jt_i} + \sum_{j=1}^{7} \rho_{jt_{i-1}} \eta_{jt_{i-1}} + \zeta_{7t_{i-1}}, \text{ for } i = 1,2. \tag{11}$$

Since (11) has already accounted for omitted variables, the treat variables in (1) that is the individual specific component representing the effect of unobservable variables (Hamerle et al., 1993) or omitted variables, has to be excluded from (11). (11), which assumes that the omitted variables develop according to an autoregression model, accounts for omitted variables in a more flexible way than (1), which assumes the omitted variables to be constant over time.

In this EDM model, $\beta_{0t_{i-1}}$'s allow for non-linear mean trajectories over time. Their estimation is performed under the EDM constraint (4b). The second term of the right handed of (11) is the original right sided of hedonic price model. No EDM constraints apply to this model component. We assume constant expenditure preferences, i.e. the parameters that reflect the proportion of expenditure on rent are constant over time. The third term is the dynamic part of the model originating from the omitted variable bias removal. In this term, we impose the restriction $\rho_{jt_{i-1}} = \rho_{7t_{i-1}} \beta_{jt_{i-1}}$ for $j = 1,2,\ldots,6$. For further detail on constant preferences and omitted variables bias removal constraints, readers may refer to Suparman et al. (2008).

The autoregression parameters, $\rho_{7t_{i-1}}$'s, are the dynamic part of the model to which the EDM constraint (4a) is applied. The last terms, $\zeta_{7t_{i-1}}$'s, are the error terms of the models. We apply the EDM constraint (4f) to these error terms. Since, there is no latent trait variable in the model, (4c)-(4e) do not apply. To account for the 1997 economic crisis that hit Indonesia just after the second wave of data collection was finished, Suparman et al. (2008) specified the intercepts in (11) to be different for t_1 and t_1, without any further constraints. Here we assume that the crisis occurred right after t_1, say t_1^*. Hence for $i = 2$, we replace t_1 in (11) by t_1^*. We furthermore assume that the crisis, on average, reduced the household income by the proportion ω_m of the t_1 level. Given the constant preference assumption, the income decrease due to the crisis implies that the β_{jt_1}'s are reduced by the same proportion. Hence, at t_1^* and at t_1, the coefficients are related as follows:

$$\beta_{jt_1^*} = \omega_m \beta_{jt_1}, \text{ for } j = 1,2,\ldots,6. \tag{12}$$

The other crisis effects, which cannot be explained by the variables in the model, are aggregated in the parameter ω_a. We may interpret ω_a as the crisis shock to the mean of monthly rent. Thus, at t_1^*, the intercept is

$$\beta_{0t_1^*} = \beta_{0t_i} + \omega_a \,. \tag{13}$$

We also apply the multiplicative effect to the error terms which gives:

$$\zeta_{7t_1^*} = \omega_m \zeta_{7t_1} \text{, with } \operatorname{var}\left(\zeta_{7t_1^*}\right) = \omega_m^2 \operatorname{var}\left(\zeta_{7t_1}\right) \tag{14}$$

Accommodating the 1997 economic crisis effect by substituting (12)-(14) into (11) we obtain for $i = 2$,

$$\eta_{7t_2} = \beta_{0t_1^*} + \sum_{j=1}^{6} \beta_{jt_2} \eta_{jt_2} + \sum_{j=1}^{7} \rho_{jt_1^*} \eta_{jt_1} + \zeta_{7t_1^*} \,. \tag{11}$$

For further detail on the measurement equation, the data and the derivation of (11), we refer to Suparman et al. (2008).

4. EMPIRICAL RESULTS

We used the maximum likelihood procedure to estimate the parameters of the EDM-SEM WTP model. For this purpose we used the Mx software package (Neale et al., 2003) which allows highly nonlinear parameter matrix restrictions. Estimation results are presented in Table 1.

For each parameter in Table 1, the first, second and third row are the estimated coefficient, standard error and p-value, respectively. The standard errors and p-values are based on 2000 bootstrap samples of size 1315 (Efron and Tibshirani, 1993). The reason for bootstrapping instead of applying standard ML procedures is that the fitted covariance matrix is nearly non-definite positive, i.e. its determinant is very small. This property would affect the standard ML standard errors and p-values, which are functions of the inverse of the fitted covariance matrix (Jöreskog, 1973). (This is a similar situation to the multicollinearity problem in regression analysis.)

From the *R*-square values we can infer that the model provides a good fit. The sign of the WTP estimates are positive as expected and the values are equivalent to their respective values in Suparman et al. (2008). The parameter significances are also consistent with Suparman et al. (2008), with the exception of the WTP estimates for in-house piped water which was not significant in our results.

Now, we return to the crisis effect parameters. We obtain a value of 0.5348 for the multiplicative effect estimate. The value tells us that due to the crisis, a WTP for a house characteristic is decreased to 46.52% (1-0.5348) of its value just before the crisis. The crisis effect on house rent can be estimated as the total of the additive effect and the sum of variables in the second terms of (15) multiplied by the multiplicative effect. The results is 1.4425 which means that on average, the monthly house rent price in rural area of Indonesia decreased by

IDR144,250 due to the 1997 economic crisis which is equivalent to 74.24% of the average of house rent just before the crisis.

Table 1. Parameter Estimates

Variable / Parameter	Year	
	1997	2000
Constant term	0.5671 (0.1617) 0.00	-0.0790 (0.1385) 0.71
Household characteristics (η_1)	0.0561 (0.0106) 0.00	0.0529 (0.0100) 0.00
house size (η_2)	0.0675 (0.0236) 0.00	0.0636 (0.0222) 0.00
House conditions index (η_3)	0.1104 (0.0354) 0.00	0.1041 (0.0334) 0.00
Presence of in house tap water (η_4)	0.1178 (0.1069) 0.13	0.1110 (0.1008) 0.13
Presence of well water (η_5)	0.0047 (0.0713) 0.48	0.0044 (0.0671) 0.48
Neighborhood characteristics (η_6)	0.1401 (0.0415) 0.00	0.1320 (0.0391) 0.00
Lagged monthly house rent (η_7)	0.0656 (0.0208) 0.00	0.0693 (0.0168) 0.00
Multiplicative crisis effect (ω_m)	not applicable	0.5348 (0.0189) 0.00
Additive crisis effect (ω_a)	not applicable	-0.6438 (0.0548) 0.00
Drift (A)	-0.6812 0.1284 0.00	
Mean trajectory (B)	0.4134 0.1416 0.00	
R-square	0.63	0.71

In this paper we regard the crisis as a structured disturbance, which distorts house rent from its "normal behavior" over time. Hence, we need to define the house rent during "normal behavior" in order to estimate the crisis effect. In the case of a standard DT model, we set the intercepts to be different, since no further constraint concerning house rent behavior over time can be imposed, and thus no crisis constraint can be applied accordingly. In this case, the crisis effect is not well defined in the model and hence it is difficult to estimate its effect directly. However, it would be possible to estimate its effect by collecting additional data right after the crisis and assuming constant intercepts for right before and after the crisis model and defining a multiplicative and an additive crisis effect. It will be difficult to determine the time for data collection after the crisis though, since it is difficult to pinpoint the time of the crisis end. Different observation times may provide different results.

In contrast, in the EDM-SEM model, we apply the EDM constraint to describe house rent behavior over time. Under this assumption and the constant preference assumption, we can define the crisis effect as a structured disturbance expressed as two parameters i.e. the multiplicative and additive effect. These parameters respectively represent the crisis effect through the variables in the model and the one which cannot be explained by the variables in the model.

5. CONCLUSION

In this paper, we used an EDM-SEM model to show that the 1997 economic crisis decreased the average of household WTP for certain house characteristics by 46.52% and decreased the average WTP for renting a house by 74.24% from their respective values right before the crisis. In addition, we show that a CT model can provide more information on the effect of a structured disturbance than a DT model.

REFERENCES

Badan Pusat Statistik (2001). *Monthly Statistical Bulletin – Economic Indicators November-2001*, Jakarta: BPS.

Bergstrom, A. R. (1988). The history of continuous-time econometric models. *Econometric theory, 4*, 365-383.

Efron, B. & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. London: Chapman and Hall, 224-227.

Gondolfo, G. (1993). Continuous-time econometrics has come of age. In G. Gondolfo (ed.), *Continuous time econometrics*. London: Chapman Hall, 1-11.

Hamerle, A., Nagl, W., & Singer, H. (1993). Problems with the Estimation of Stochastic Differential Equation Using Structural Equation Models. *Journal of Mathematical Sociology, 16*, 201-220.

Jöreskog, K. (1973). A General Method for Estimating a Linear Structural Equation System. In A.S. Goldberger & O.D. Duncan (eds.), *Structural Equation Model in the Social Sciences*. London: Seminar Press, 85-112.

Neale, M.C., Boker, S.M., Xie, G., & Maes, H.H. (2003) *Mx: Statistical Modeling* (6th ed.). Richmond: Department of Psychiatry.

Oud, J.H.L & Jansen, R.A.R.G. (1996). Nonstationary Longitudinal LISREL Model Estimation From Incomplete Panel Data Using EM and the Kalman Smoother. In U. Engel & J Reinecke (Eds.). *Analysis of Change: Advanced Techniques in Panel Data analysis*. New York: de Gruyter, 135-159.

(2000). Continuous Time State Space Modeling of Panel Data by Means of SEM. *Psychometrika, 65,* 199-215.

Oud, J.H.L. (2002). Continuous Time Modeling of the Cross-Lagged Panel Design. *Kwantitatieve Methoden, 69,* 1-26.

Phillips, P.C.B. (1993). The ET Interview: A.R. Bergstrom. In P.C.B. Phillips (Ed.). *Models, Methods, and Applications of Econometrics*. Cambridge MA: Blackwell, 12-31.

Suparman, Y., Folmer, H., Oud, J.H.L., & Resosudarmo, B.P. (2008). Eliciting the Willingness to Pay for Piped Water from Self-Reported Rent Appraisals in Indonesia: A SEM Autoregressive Panel Approach. A paper presented at 16th *Annual Conference of the European Association of Environmental and Resource Economist*. Gothenburg University, Sweden.

World Bank (2009). *World Development Indicators*. http://ddp-xt.worldbank.org/ext/DDPQQ/member.do?method=getMembers&userid =1&queryId=135 (last accessed at 2.30 pm. 13-08-2009).

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 1079–1086.

MULTISCALE SEASONAL AUTOREGRESSIVE FOR FORECASTING TREND AND SEASONAL TIME SERIES

Umu Sa'adah[1], Subanar[2], Suryo Guritno[2] and Suhartono[3]

[1]Mathematics Department,
Gadjah Mada University, Yogyakarta, Indonesia;
Mathematics Department, Brawijaya University, Malang, Indonesia

E-mail: umusaadah@yahoo.com

[2]Mathematics Department, Gadjah Mada University, Yogyakarta, Indonesia
E-mail: subanar@yahoo.com

[3]Statistics Department, Institut Teknologi Sepuluh Nopember Surabaya, Indonesia
E-mail: suhartono@statistika.its.ac.id

ABSTRACT

The aim of this research is to study further some latest development about wavelet decomposition for modeling time series with both trend and seasonal patterns. This research focuses on the Maximal Overlap Discrete Wavelet Transform (MODWT) and Multiscale Autoregressive (MAR) model that firstly proposed by Renaud et al (2003). First, we develop new procedure for model building of MAR. This procedure accommodates lags of scale and wavelet coefficients proposed by Renaud et al (2003) and some appropriate lags addition. The main issue in MAR modeling for trend and seasonal time series is how to determine the best lags of scale and wavelet coefficients as predictors in MAR model. In this research, some filters from wavelet family are used and compared, i.e. Haar, Daubechies(4), Coiflet(6) and Least Asymmetric(8). Data about International Airline Passenger is used as case study. The results show that MAR model based on Haar filter yields better result than other filters and ARIMA model, both in testing and training data. It is shown by the smallest value of RMSE on both parts of data.

Keywords: Trend, seasonal, wavelet, filter, MODWT, MAR.

1. INTRODUCTION

Many business and economic time series are non-stationary time series that contain trend and seasonal variations. The trend is the long-term component that represents the growth or decline in the time series over an extended period of time. Seasonality is a periodic and recurrent pattern caused by factors such as weather, holidays, or repeating promotions. Accurate forecasting of trend and seasonal time series is very important for effective decisions in retail, marketing, production, inventory control, personnel, and many other business sectors (Makridakis and Wheelwright, 1987).

Wavelet transform is a multiresolution decomposition techniques that can produce a good local representation of the signal in both the time domain and the frequency domain, (Mallat,

1989; Ogden, 1997). It can be used to make model and estimate data that contain both the autocorrelation and the trend, (Nason and von Sachs, 1999). Wavelet analysis processes information effectively at different scales and can be very useful for feature detection from complex and chaotic time series, (Shin and Han, 2000). It automatically separates the trend from the signal or data. The wavelet coeffisients are calculated only from data obtained previously in time (Renaud et al, 2003).

The aim of this paper is to develop procedure from Renaud et al (2003) for forecasting trend and seasonal time series, especially in the case of the International Airline Passenger data. To determine the best lags of scale and wavelet coefficients as predictors in MAR model, the procedure accommodates some appropriate lags addition (i.e. seasonal lags or near-seasonal lags) of scale and wavelet coefficients beside lags input selection that proposed by Renaud et al (2003). This seasonal lags addition was motivated by the fact that Cross Correlation Functions (CCF) between a stationary seasonal time series data X_t with scale or wavelet coefficients form appropriate seasonal pattern. It means that the seasonal lags have influence either on the scale or wavelet coefficients or on the data X_t. We use Maximal Overlap Discrete Wavelet Transform (MODWT) as wavelet decomposition with level J =4 and some filters from wavelet families, that is Haar, Daubechies(4) or D4, Coiflet(6) or C6 and Least Asymetric(8) or LA8 to comparison. The use of MODWT is reasoned by the MODWT of level J is well defined for any sample size N. Whereas the *Discrete Wavelet Transform* (DWT) of level J restricts the sample size to a integer multiple of $N = 2^J$.

2. MULTISCALE PREDICTION BASED MODWT

The motivation for formulating the MODWT is essentially to define a transform that acts as much as possible like the DWT, but does not suffer from the DWT's sensitivity to the choice of a starting point for a time series. In Percival and Walden (2000), this sensitivity is entirely due to downsampling (subsampling) the outputs from the wavelet and scaling filters at each stage of the pyramid algorithm.

Let define $\tilde{\mathcal{A}}$ as the $N \times N$ matrix that contains the MODWT scaling filter $\tilde{g}$ and $\tilde{\mathcal{B}}$ as the $N \times N$ matrix that contains the MODWT wavelet filter $\tilde{h}$. Let $J = 1$, $\tilde{\mathcal{B}}_1$ be defined as the $N \times N$ matrix as in Equation (1), so that we have $\tilde{W}_1 = \tilde{\mathcal{B}}_1 X$. With an analogous definition for $\tilde{\mathcal{A}}_1$, we have $\tilde{V}_1 = \tilde{\mathcal{A}}_1 X$. Whereas $\tilde{\mathcal{A}}_1$ has a similar structure like $\tilde{\mathcal{B}}_1$ with each $\tilde{h}_l$ being replaced by $\tilde{g}_l$.

$$\tilde{\mathcal{B}}_1 = \begin{bmatrix} \tilde{h}_0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \tilde{h}_3 & \tilde{h}_2 & \tilde{h}_1 \\ \tilde{h}_1 & \tilde{h}_0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & \tilde{h}_3 & \tilde{h}_2 \\ \tilde{h}_2 & \tilde{h}_1 & \tilde{h}_0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \tilde{h}_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \tilde{h}_3 & \tilde{h}_2 & \tilde{h}_1 & \tilde{h}_0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \tilde{h}_3 & \tilde{h}_2 & \tilde{h}_1 & \tilde{h}_0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \tilde{h}_3 & \tilde{h}_2 & \tilde{h}_1 & \tilde{h}_0 \end{bmatrix} \quad (1)$$

The first stage of the MODWT pyramid algorithm can be represented as

$$\begin{bmatrix} \tilde{W}_1 \\ \tilde{V}_1 \end{bmatrix} = \begin{bmatrix} \tilde{\mathcal{B}}_1 \\ \tilde{\mathcal{A}}_1 \end{bmatrix} X = \tilde{\mathcal{P}}_1 X \tag{2}$$

with

$$\tilde{\mathcal{P}}_1 \equiv \begin{bmatrix} \tilde{\mathcal{B}}_1 \\ \tilde{\mathcal{A}}_1 \end{bmatrix} \tag{3}$$

and $\tilde{\mathcal{P}}_1^T$ is an orthonormal matrix. If $\tilde{\mathcal{P}}_1$ defined as in Equation (3), we can re-express X as

$$X = \tilde{\mathcal{P}}_1^{-1} \begin{bmatrix} \tilde{W}_1 \\ \tilde{V}_1 \end{bmatrix} \tag{4}$$

Since $\tilde{\mathcal{P}}$ is orthonormal matrix, so $\tilde{\mathcal{P}}_1^T = \tilde{\mathcal{P}}_1^{-1}$. It follows that

$$X = \tilde{\mathcal{P}}_1^T \begin{bmatrix} \tilde{W}_1 \\ \tilde{V}_1 \end{bmatrix} = \left[\tilde{\mathcal{B}}_1^T, \tilde{\mathcal{A}}_1^T \right] \begin{bmatrix} \tilde{W}_1 \\ \tilde{V}_1 \end{bmatrix} \tag{5}$$

Therefore

$$X = \tilde{\mathcal{B}}_1^T \tilde{W}_1 + \mathcal{A}_1^T \tilde{V}_1 \tag{6}$$

The MODWT of level J is defined in terms of a computationally efficient pyramid algorithm. In efficient pyramid algorithm, computating the *j*th level MODWT wavelet and scaling coefficients $\tilde{W}_j$ and $\tilde{V}_j$ is based upon the scaling coefficients $\tilde{V}_{j-1}$ of level *j*-1. In Mallat (1989); Percival and Walden (2000), the key to this algorithm is to note the relationship between the filters used to compute the coefficients of level *j*-1 and *j*.

The implementation of MODWT described in Percival and Walden (2000) is used Renaud et al. (2003) for MAR modeling. The basic idea is to use the coeffisients $w_{j,t-2^j(k-1)}$ for $k = 1,...,A_j$ dan $j = 1,...,J$ and $v_{J,t-2^J(k-1)}$ for $k = 1,...,A_{j+1}$ as predictors. The first point is to know how many and which wavelet coefficients will be used at each scale.

Assume a stationary signal $X = (X_1, X_2, ..., X_t)$ and assume we want to predict X_{t+1}. Type prediction used is simplest model for prediction i.e. autoregressive (AR). Recall that to minimise its mean square error, the one-step forward prediction of an AR(*p*) process is written

$$\hat{X}_{t+1} = \sum_{k=1}^{p} \hat{\phi}_k X_{t-(k-1)} = \hat{\phi}_1 X_t + \hat{\phi}_1 X_{t-1} + ... + \hat{\phi}_p X_{t-p+1} \tag{7}$$

$\hat{\phi}_k$ could be estimated by maximum likelihood estimation, Yule-Walker or least squares estimation, has the same asymptotic efficiency. In using the decomposition based MODWT, the AR prediction was modified by Renaud et al. (2003) to the AR Multiscale, that is:

$$\hat{X}_{t+1} = \sum_{j=1}^{J} \sum_{k=1}^{A_j} \hat{a}_{j,k} w_{j,t-2^j(k-1)} + \sum_{k=1}^{A_{j+1}} \hat{a}_{J+1,k} v_{J,t-2^J(k-1)}, \tag{8}$$

where j is number of levels ($j = 1, 2, ..., J$), A_j is order of MAR model ($k = 1, 2, ..., A_j$), $w_{j,t}$ is value of wavelet coeffisients from the Haar wavelet, $v_{j,t}$ is value of scale coeffisients from the Haar wavelet, and $\hat{a}_{j,k}$ is value of scale MAR coeffisients.

Let *t*=36, *J=2* and A_j *=2* $(k = 1,2)$, MAR(*2*) model base upon Equation (8) is

$$\hat{X}_{t+1} = \sum_{j=1}^{2}\sum_{k=1}^{2} \hat{a}_{j,k} w_{j,t-2^j(k-1)} + \sum_{k=1}^{2} \hat{a}_{J+1,k} v_{J,t-2^J(k-1)} \tag{9}$$
$$= a_{1,1}w_{1,t} + a_{1,2}w_{1,t-2} + a_{2,1}w_{2,t} + a_{2,2}w_{2,t-4} + a_{3,1}v_{2,t} + a_{3,1}v_{2,t-4},$$

or it can be written in matrix form as in the following Equation (10):

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ \vdots \\ X_{33} \\ X_{34} \\ X_{35} \\ X_{36} \end{bmatrix} = \begin{bmatrix} w_{1,0} & w_{1,-2} & w_{2,0} & w_{2,-4} & v_{2,0} & v_{2,-4} \\ w_{1,1} & w_{1,-1} & w_{2,1} & w_{2,-3} & v_{2,1} & v_{2,-3} \\ w_{1,2} & w_{1,0} & w_{2,2} & w_{2,-2} & v_{2,2} & v_{2,-2} \\ w_{1,3} & w_{1,1} & w_{2,3} & w_{2,-1} & v_{2,3} & v_{2,-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{1,32} & w_{1,30} & w_{2,32} & w_{2,28} & v_{2,32} & v_{2,28} \\ w_{1,33} & w_{1,31} & w_{2,33} & w_{2,29} & v_{2,33} & v_{2,29} \\ w_{1,34} & w_{1,32} & w_{2,34} & w_{2,30} & v_{2,34} & v_{2,30} \\ w_{1,35} & w_{1,33} & w_{2,35} & w_{2,31} & v_{2,35} & v_{2,31} \end{bmatrix} \begin{bmatrix} a_{1,1} \\ a_{1,2} \\ a_{2,1} \\ a_{2,2} \\ a_{3,1} \\ a_{3,2} \end{bmatrix} \tag{10}$$

Matrix equation in Equation (10) can also be written in form

$$\mathbf{s} = \mathbf{A}\boldsymbol{\alpha}, \tag{11}$$

where $\mathbf{s}$ is a vector of $(X_1, X_2, ..., X_{36})$ data, $\mathbf{A}$ is an matrix that contains the wavelet and scale coeffisients and becomes the input in MAR model, and $\boldsymbol{\alpha}$ is a vector of $(a_{1,1}, a_{1,2}, .. a_{3,2})$ parameters. To estimate parameters in $\boldsymbol{\alpha}$ vector, it is used Normal Equation, i.e. $\mathbf{A'A\alpha} = \mathbf{A's}$ which follows Least Squares Principle. So that, it will be obtained

$$\hat{\boldsymbol{\alpha}} = (\mathbf{A'A})^{-1}\mathbf{A's}. \tag{12}$$

Renaud et al. (2003) show that MAR Model in Equation (9) is a model that can be used to forecasting stationary time series data.

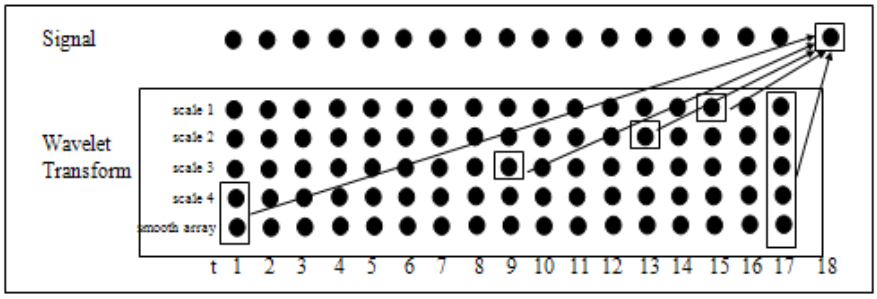


Figure 1. Illustration from the wavelet modeling process for J=4 and A_j =2.

An input process from the wavelet modeling is proposed by Renaud et al. (2003). To make it clear, which chosen input in forecast procedure for $(t+1)^{th}$ data in the wavelet modeling is described in Figure 1. The figure represents general form of the wavelet modeling with level J=4 and order A_j =2. Figure 1 show that if it will be made forecasting for 18th data by using the wavelet modeling for the 2nd order, i.e. MAR(2), then input variable for the MAR(2) is the wavelet coeffisients in the 1st level at t=17 and t=15, in the 2nd level at t=17 and t=13, in the 3th level at t=17 and t=9, in the 4th level at t=17 and t=1; and the smooth coeffisients in the 4th level at t=17 and t=1. Therefore, it can be concluded that the 2nd input in each level is $t-2^j$.

3. RESEARCH METHODOLOGY

The purpose of this research is to provide empirical evidence on the comparative study of many MAR models for forecasting trend and seasonal time series. To determine the best lags of scale and wavelet coefficients as predictors in MAR model, we investigate:

"how many and which wavelet coefficients will be used at each scale?"

We conduct empirical study with real data, the International Airline Passenger Numbers data in January 1949 - in December 1960, to address this question. This real data has been analyzed by many researchers, see for example Nam and Schaefer (1995), Hill *et al*. (1996), Faraway and Chatfield (1998), Suhartono *et al*. (2005). This data also has become one of two data to be competed in Neural Network Forecasting Competition on June 2005 (see www.neural-forecasting.com). The data contain 144 month observations. The first 120 data observations (called in sample or training data) are used for model selection and parameter estimation and the last 24 points (called out sample or testing data) are reserved as the test for forecasting evaluation and comparison. The time series of these data has an upward trend together with seasonal variations.

Training data is transformed become stationary seasonal data using Trend Analysis. Detrended data is decomposed based on MODWT with level J =4 and some filters from wavelet families, that is Haar, D4, C6 and LA8. MODWT result is four wavelet coefficients and four scale coefficients which each coefficient has the same size with training data. To build MAR model in this level is needed the four wavelet coefficients and the fourth level of scale coefficient.

To take choice the lags of scale and wavelet coefficients as predictors in MAR model, four methods are applied for each filter. The first 2 methods, selection of predictors in MAR model based on the lags of scale and wavelet coefficients that proposed by Renaud et al (2003) and applied on first and second order from MAR model. The last 2 methods, selection of predictors in MAR model based on the lags of scale and wavelet coefficients that proposed by Renaud et al (2003) and seasonal lags of scale and wavelet coefficients. The last 2 methods are also applied on first and second order from MAR model.

Stepwise method is applied to determine the best lags of scale and wavelet coefficients as predictors in MAR model. The stepwise Directed Search can select automatically the subset with the smallest Mallows' statistic (Broersen, 1986). Based on the best lags of scale and wavelet coefficients as predictors in MAR model, is determined forecast value for the training data and is counted errors value. The MAR model that Normality and White-noise assumption satisfied is chosen and is calculated mean square error value (MSE).

To forecast testing data is done to 24-step ahead forecasting based on the Trend model and the MAR model. MSE of testing data is calculated. The MAR model that has the smallest MSE of testing data is the best model.

4. EMPIRICAL RESULTS

Table 1 summarizes the result of the comparison MAR models based on lags input selection and report performance measures across training data and evaluation of model fit for residual from the International Airline Passenger Numbers data. The seasonal lags addition into the Renaud et al method reduces the MSE values in training data just on filter Haar, especially on MAR(1) model with lag 12 and lag 24 addition; MAR(2) model with lag 12 addition. The residuals that

satisfy both White-noise and Normality assumptions come from the combination MAR(1) model with lag 12, MAR(1) model with lag 12 and lag 24 addition and MAR(2) model with lag 12 addition. There is no residual that satisfies White-noise and Normality assumption for filter D4, C6 or LA8.

Table 1. The result of the comparison MAR models

Filter	Model	Lags Input Selection	MSE in-sample	Evaluation of model fit for residual	
				Normality	White-noise
Haar	MAR(1)	Renaud, et al. method	411.084	Yes	No
		Renaud, et al. method + lag 12[(1)]	110.733	Yes	Yes
		Renaud, et al. method + lag 12 + lag 24[(2), (3)]	78.956	Yes	Yes
			83.2022	Yes	Yes
	MAR(2)	Renaud, et al. method	309.247	No	No
		Renaud, et al. method + lag 12[(4)]	84.623	Yes	Yes
D4	MAR(1)	Renaud, et al. method	802.905	No	No
		Renaud, et al. method + lag 12	468.992	Yes	No
	MAR(2)	Renaud, et al. method	189.304	No	No
		Renaud, et al. method + lag 12	170.853	Yes	No
C6	MAR(1)	Renaud, et al. method	827.956	Yes	No
		Renaud, et al. method + lag 12+lag 24	413.109	Yes	No
	MAR(2)	Renaud, et al. method	176.261	Yes	No
		Renaud, et al. method + lag 12	198.721	Yes	No
LA8	MAR(1)	Renaud, et al. method	613.794	No	No
		Renaud, et al. method + lag 24	249.355	Yes	No
	MAR(2)	Renaud, et al. method	325.657	Yes	No
		Renaud, et al. method + lag 12+lag 24	238.033	Yes	No

(1), (2), (3) and (4) satisfy Normality and White-noise assumption

The results of the best model selection and report performance measures across training and testing samples for the International Airline Passenger Numbers data are summarized in Table 2. Numbers greater than one on column ratio indicates poorer forecast performance compare to the ARIMA models, and better for numbers less than one. We can clearly see on the ratio of testing samples that model (1), (2), (3) and (4) do not yield MSE less than ARIMA.

Table 2. The result of the comparison between ARIMA and MAR models that Normality and White-noise assumption satisfied.

Model	In-sample (training data)		Out-sample (testing data)	
	MSE	Ratio to ARIMA	MSE	Ratio to ARIMA
ARIMA	88.862	1.0000	1527.03	1.0000
Model (1)	110.733	1.2461	1699.489	1.1129
Model (2)	78.956	0.8885	1570.675	1.0286
Model (3)	83.202	0.9363	1688.072	1.1055
Model (4)	84.653	0.9526	1619.316	1.0605
Model (5)*	78.072	0.8786	1524.387	0.9983

* The best model both in training data and in testing data.

Based on model (2) is formed MAR model with near-seasonal lags addition. Modified model (2), called model (5) reduce the MSE values in testing data, that is 3.03% from model (2) or 0.17% from ARIMA model.

5. CONCLUSION

Based on the results we can conclude that the seasonal lags addition in Renaud et al method can be applied for forecasting seasonal time series (no trend). Filter D4, C6, LA8 are not appropriate to be used in MAR modeling. The results agree with reason of Renaud et al (2003).

ACKNOWLEDGEMENTS

This work is supported by Institute for Research and Community Service, Gadjah Mada University, Yogyakarta, Indonesia; Directorate General of Higher Education, Department of National Education, Indonesia Government and Brawijaya University, Malang, Indonesia.

REFERENCES

Broersen, P. M. T. (1986). Subset Regression with Stepwise Directed Search. *Applied Statistics*, 35, 2, 168-177.

Faraway, J. and Chatfield, C. (1998). Time series forecasting with neural network: a comparative study using the airline data. *Applied Statistics*, 47, 231–250.

Hill, T., M. O'Connor and Remus, W. (1996). Neural network models for time series forecasts. *Management Science*, 42, 1082–1092.

Makridakis, S. and Wheelwright, S.C. (1987). *The Handbook of Forecasting: A Manager's Guide*. 2nd Edition, John Wiley & Sons Inc., New York.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans, Pattern Anal. Machine Intelligence*, 11, 674–693.

Nam, K. and Schaefer, T. (1995). Forecasting international airline passenger traffic using neural networks. *Logistics and Transportation Review,* 31, 3, 239–251.

Nason, G.P. and R. von Sachs. (1999). Wavelets in time series analysis. *Phil. Trans. R. Soc. London,* A, 357, 2511-2526.

Ogden, R.T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*. Boston: Birkhauser.

Percival, D.B and Walden, A.T. (2000). *Wavelets methods for Time Series Analysis*. Cambridge: Cambridge University Press.

Renaud, O., Stark, J.L. and Murtagh, F. (2003). Prediction based on a multiscale decomposition. *Int. Journal of Wavelets, Multiresolution and Information Processing*, 1, 2, 217-232.

Shin, T. and I. Han. (2000). Optimal signal multi-resolution by Genetic Algorithms to support Artificial Neural Networks for exchange-rate forecasting. *Expert Systems with Applications*, 18, 257–269.

Suhartono, Subanar and Rezeki, S. (2005). Feedforward Neural Networks Model for Forecasting Trend and Seasonal Time Series. *Proceedings of the 1st IMT-GT Regional Conference on Mathematics, Statistics and Their applications*, Medan, North Sumatera, Indonesia.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 1087–1093.

CONTIGUOUS DISEASES OUTBREAK IN INDONESIA: THE APPLICATIONS OF SPATIAL SCAN STATISTICS METHOD

Yekti Widyaningsih[1], Asep Saefuddin[2]
Department of Statistics, Bogor Institute of Agriculture, Indonesia
E-mail: [1]yekti@ui.ac.id, [2]asaefuddin@gmail.com

ABSTRACT

The ability to detect disease outbreaks early is important in order to minimize morbidity and mortality through timely implementation of disease prevention and control measures. Many nationals, states, and local health departments are launching disease surveillance systems without statistical testing. Spatial scan statistic is a statistical tool to detect location of clusters (outbreak) of interest. This paper shows how scan statistic method detects diseases clusters. The applications are on detections of contiguous diseases (HIV/AIDS, Tuberculosis and Malaria) hotspot in 2001 - 2006. SaTScan software was used for the computation. As the result, for AIDS cases in Indonesia, there was a moving hotspot area of the mortality number from 2002 to 2006. In 2002, AIDS mortality hotspot is around east part, but in 2006, it was around the central part of Indonesia. The spreading of Tuberculosis in Indonesia was around east and central part of the country in 2001, whereas, in 2002 a reduction of the hotspot was in the central part. For malaria cases, the cluster regions of diseases cases were found, that the cluster regions of high malaria cases tend to decrease from year to year. The highest was in the east and central part of the country in 2001 and 2002.

Keywords: Spatial scan statistics, HIV- AIDS, tuberculosis, malaria, hotspot, SatScan

1. INTRODUCTION

Health officials are often asked to evaluate local disease clusters alarms. After the case definition is established, an early question is whether the cluster has occurred by chance or whether the outbreak is so great that it is probably due to some common elevated risk factor of limited geographical and/or temporal extension. Because of these needs, scan statistics and/or space-time scan statistics have become popular methods in disease surveillance for the detection of disease clusters. The standard approach is to look at a single disease, such as leukemia incidence, breast cancer mortality, HIV/AIDS mortality, bird flu, tuberculosis, and dengue fever. The aim of this paper is to identify the highest response areas of some contiguous diseases (HIV/AIDS, Tuberculosis and Malaria) occurred in Indonesia and to test whether those areas are significant statistically.

Kulldroff (2006), created the SaTScan software as a tool for data change detection within space and /or time. The properties of the data are the scan area geographic, probability distribution of the response under the null hypothesis, and the significance of the statistic test is evaluated with Monte Carlo simulation; Kulldorff (2006).

2. METHODS

2.1 Data

In the year 2005 the population of Indonesia was about 240 millions. As an archipelago, Indonesia consists of thousands small islands and five big islands; they are Sumatera, Kalimantan, Java, Sulawesi, and Papua. There are 33 provinces, 440 districts, and about 5900 sub-districts spread on the islands. Two seasons in Indonesia are the dry season and wet season with a transition period in September.

2.1.1 HIV/AIDS

HIV/AIDS disease is continuing to increase in number, and spreading through the regions. Recently, according to the Jakarta Post news paper, the number of AIDS cases was found in Yogyakarta and more number in Papua; Kulldorff (2007). Through the scan statistics calculation, we analyzed the cluster regions of the AIDS mortality cases in year 2002, which had the mortality number 379 of 1016 cases; and in year 2006, which had mortality number 1651 of 6987 cases; Depkes (2006). This application compared the two results, year 2002 and 2006 data. Based on the number of HIV/AIDS cases, assumed that the mortality number as Bernoulli distribution with possibilities of die or not die. The equation (2) is used as the model of the statistical test.

2.1.2 Ttuberculosis

Although tuberculosis had occurred in human since thousands years ago, this disease is still a big problem in the world and difficult to be combated. Nowadays, many people are still suffering of this disease. Estimated, one person contagious this disease in every second. According to the World Health Organization (WHO) data, more than 670,000 cases occurred each year with deadly number is 175,000 persons; Depkes (2007). Hotspot of TB cases in Indonesia was analyzed using Scan Statistics Method. The data was the number of TB cases of every province in 2001, 2002, 2004, and 2005.

2.1.3 Malaria

Malaria is presently endemic in a broad band around the equator, in areas of the Americas, many parts of Asia, and much more of Africa. The geographic distribution of malaria within large regions is complex, and malaria-free areas are often found close to each other. The global endemic levels of malaria have not been mapped since the 1960s. However, the Welcome Trust, UK, has funded the Malaria Atlas Project to rectify this, providing a more contemporary and robust means with which to assess current and future malaria disease burden; Oemijati (1992).

In Indonesia, malaria has spread to all provincial areas. Malaria often emerges as an outbreak with relatively high morbidity and mortality rates. As an archipelago nation, malaria condition in Indonesia is various for each island. Java and Bali Islands which populated of 70% from total Indonesian population is categorized as a hypo-endemic area. While in other islands that sparsely outer of Java and Bali consist of Sumatera, Kalimantan, Sulawesi, Nusa Tenggara, Maluku and Papua, malaria is found at much higher levels. These areas are categorized from hypo- to hyper endemic.

2.2 Scan Statistics

2.2.1 Purely Spatial Scan Statistics

Purely spatial scan statistics concern in two-dimensional space. Three basic properties of the scan statistic are the geometry of the area being scanned, the probability distribution generating events under the null hypothesis, and the shapes and sizes of the scanning window. Depending on the application, different models will be chosen, and depending on the model, the test statistics may be evaluated either through explicit mathematical derivations and approximations or through Monte Carlo sampling. In the latter case, random data sets are generated under the null hypothesis, and the scan statistics is calculated in each case, comparing the values from the real and random data sets to obtain a hypothesis test; Kulldorff (1999).

Bernoulli Model. Under the Bernoulli model, the null hypothesis is $H_o : p = q$, $N(A) \sim$ Binomial $(\mu(A), p)$ for all sets A. And the alternative hypothesis is $H_1 : p > q$, $N(A) \sim$ Binomial $(\mu(A), p)$ for set $A \subset Z$, and $N(A) \sim$ Binomial $(\mu(A), q)$ for set $A \subset Z'$. N(A) is the number of cases in A. Z' is hotspot areas. Probability density functions of an event is

$$f(x) = \begin{cases} p(1-p) & A \subset Z \\ q(1-q) & A \subset Z' \end{cases} \quad (1)$$

If n_Z is point (number of cases in zone Z), n_G is the number of observation, and G is the study area, Likelihood for Bernoulli model is

$$L(Z, p, q) = p^{n_Z}(1-p)^{\mu(Z)-n_Z} q^{n_G-n_Z}(1-q)^{(\mu(G)-\mu(Z))-(n_G-n_Z)} \quad (2)$$

The test statistic λ of the likelihood ratio test can be written as

$$\lambda = \frac{\sup_{Z\in\Omega, p>q} L(Z, p, q)}{\sup_{p=q} L(Z, p, q)} = \frac{L(\hat{Z})}{L_o} \quad (3)$$

L_0 is likelihood under null Hypothesis; Kulldorff (1997).

Poisson Model. Under the Poisson model, points are generated by an inhomogeneous Poisson process. There is exactly one zone $Z \subset G$ such that $N(A) \sim$ Poisson$(p\mu(A\cap Z) + q\mu(A\cap Z^c)) \forall A$. The null hypothesis is $H_o : p = q$, while the alternative hypothesis states that $H_1 : p > q$, $Z \in$ **Z.** Under H_o, $N(A) \sim$ Poisson $(p\mu(A)) \forall A$. Note that one of the parameters, Z disappears under the null hypothesis. The probability of n_G number of points in the study area is

$$\frac{e^{-p\mu(Z)-q(\mu(G)-\mu(Z))}[p\mu(Z)+q(\mu(G)-\mu(Z))]^{n_G}}{n_G!} \quad (4)$$

The likelihood function for the Poisson model is

$$L(Z, p, q) = \frac{e^{-p\mu(Z)-q(\mu(G)-\mu(Z))}}{n_G!} p^{n_Z} q^{(n_G-n_Z)} \prod_{a_i} \mu(a_i) \quad (5)$$

The test statistic λ of the likelihood ratio test can now be written as

$$\lambda = \frac{\sup_{Z\in\Omega,\, p>q} L(Z,p,q)}{\sup_{p=q} L(Z,p,q)} = \frac{L(\hat{Z})}{\frac{e^{-n_G}}{n_G!}\left(\frac{n_G}{\mu(G)}\right)^{n_G}\prod_{a_i}\mu(a_i)}$$

$$= \sup_{Z\in\Omega} \frac{\left(\frac{n_Z}{\mu(Z)}\right)^{n_Z}\left(\frac{n_G - n_Z}{\mu(G)-\mu(Z)}\right)^{n_G - n_Z}}{\left(\frac{n_G}{\mu(G)}\right)^{n_G}} I\left(\frac{n_Z}{\mu(Z)} > \frac{(n_G - n_Z)}{(\mu(G)-\mu(Z))}\right) \quad (6)$$

if there is at least one zone Z such that $\frac{n_Z}{\mu(Z)} > \frac{(n_G - n_Z)}{(\mu(G)-\mu(Z))}$, and λ=1 otherwise. I() is the indicator function. The ratio λ, is used as the test statistic, and its distribution is find through Monte Carlo repetition as described below; Kulldorff (1997).

In order to find the value of the test statistic, we need a way to calculate the likelihood ratio as it is maximized over the collection of zones in the alternative hypothesis. Once the value of the test statistic has been calculated, it is easy to do the inference. We cannot expect to find the distribution of the test statistic in closed analytical form. Instead we rely on Monte Carlo simulation. Because we know the underlying measure μ, we can obtain replications of the data set generated under the null hypothesis when we condition on the total number of points n_G. With 9999 such replications, the test is significant at the 5 percent level if the value of the test statistic for real data set is among 500 highest values of the test statistic coming from the replications; Kulldorff (1997).

In addition, the most likely cluster for the real data has the significantly test statistic at α = 0.05, that is, its likelihood ratio is on 5% highest among the values of replicated data. The p-value of the Monte Carlo hypothesis is defined as p-value = *r /(1 + sim)*; *r* is the ranking and *sim* is the number of repetitions of the data simulation under the null hypothesis; Kulldorff (2006).

3. APPLICATIONS AND THE RESULTS

The datasets; case and geographical datasets are appended to a master archive using SaTScan. The goal of data analysis is to detect the cluster region (outbreak) of some contiguous diseases. The performance of the purely spatial scan statistic evaluated HIV/AIDS mortality, Tuberculosis (TB) and Malaria data. TB and Malaria cases were assumed as Poisson distribution, whereas HIV/AIDS mortality was assumed as Bernoulli distribution with possibilities of die or not die as mentioned before.

Figure 1 shows the atlases of the spatial scan statistics results for HIV/AIDS mortality cases through SatScan software; ESRI (1996 – 2000). In 2002, the regions of the Most Likely Cluster of aids mortality were Maluku, Central Sulawesi, Papua, North Sulawesi, South Sulawesi, East Nusatenggara, and East Kalimantan. This cluster was statistically significant with relative risk = 1.305, Likelihood ratio (LLR) = 5.175177, and p-value= 0.020. Secondary clusters (Central Java and some parts of Sumatera areas) were not statistically significant.

In 2006, the regions of the Most Likely Cluster of aids mortality were the central part of the country. The provinces were East Nusatenggara, West Nusatenggara, South Sulawesi, Southeast Sulawesi, Bali, West Sulawesi, Central Sulawesi, South Kalimantan, East Java, Gorontalo, Yogyakarta, Central Kalimantan, Maluku, North Sulawesi, East Kalimantan, Central Java, and North Maluku. This cluster was statistically significant with relative risk = 1.578, LLR = 50.587, and p-value= 0.001. The regions of the first secondary cluster were some parts of Sumatera. The

provinces were West Sumatera, Riau, and Riau Islands. This secondary cluster was statistically significant with relative risk = 2.024, LLR = 41.15658, and p-value= 0.001.

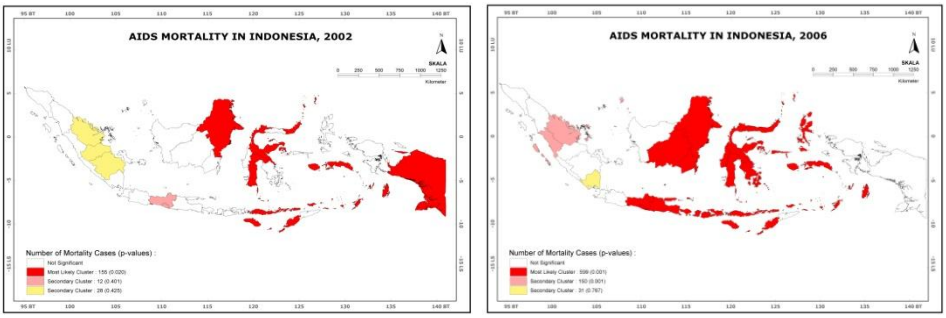


Figure 1. Hotspot of Aids Mortality cases, 2002 and 2006

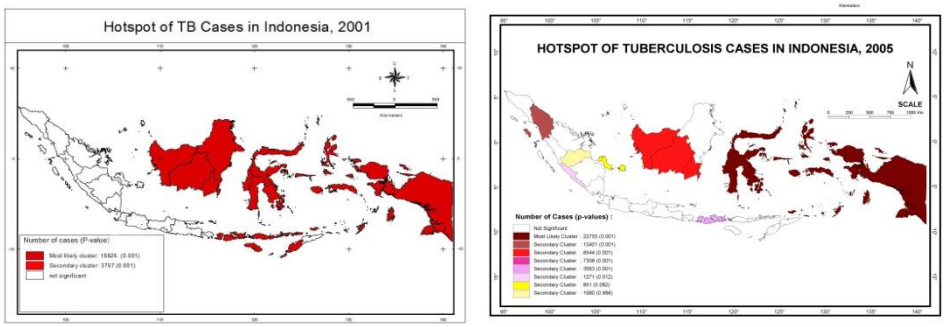


Figure 2. Hotspot of TB Cases, 2001 and 2005

Figure 2 shows the hotspot of Tuberculosis cases, 2001 and 2005. The cases was high in Kalimantan, Sulawesi, Maluku, Nusa Tenggara Islands, and Papua in year 2001. In 2002 these areas were still high, except West and Central Kalimantan (not showed). Followed by Jakarta and West Java, also had significantly high TB cases. Furthermore, in 2004, North Sulawesi, North Maluku, South East Sulawesi, Gorontalo, South Sulawesi, West Sulawesi, and South Kalimantan, the number of TB cases was high. There were 17,288 cases in those areas. Furthermore, in 2005, the highest cases were in Maluku, Sulawesi, and Papua. As a result, hotspots of TB cases were relatively moving just around East and central part of Indonesia and some areas in Sumatera. In addition, hotspot of TB cases also high in North Sumatra in 2004 and 2005, whereas it was not significant in 2001 and 2002. In 2005, the most likely cluster of TB consists provinces in *the east part* of Indonesia, they were Maluku, North Maluku, the whole Sulawesi, and Papua. This cluster was statistically significant as a hotspot with the number of cases was 22755. The secondary clusters were North Sumatera with 13401 cases, followed by Central Kalimantan, South Kalimantan, West Kalimantan with 8544 cases; Jakarta, 7308 cases; West Nusa Tenggara, 3563 cases; and Bengkulu, 1271 cases. All these clusters were statistically significant as the hotspots of Tuberculosis case in Indonesia. Whereas, the estimation number of tuberculosis cases in Indonesia were 296,381 cases spread around Indonesia Islands.

In 2001, hotspot areas of malaria were Kalimantan, Sulawesi, West Nusa Tenggara, East Nusa Tenggara, North Maluku, and Papua. In 2002, hotspot areas were still in the same areas,

except Central Kalimantan and West Kalimantan. In 2003 to 2004, hotspot area is held out in East Nusa Tenggara. In 2005, hotspots areas back to Gorontalo, Southeast Sulawesi, North Sulawesi, and South Sulawesi.

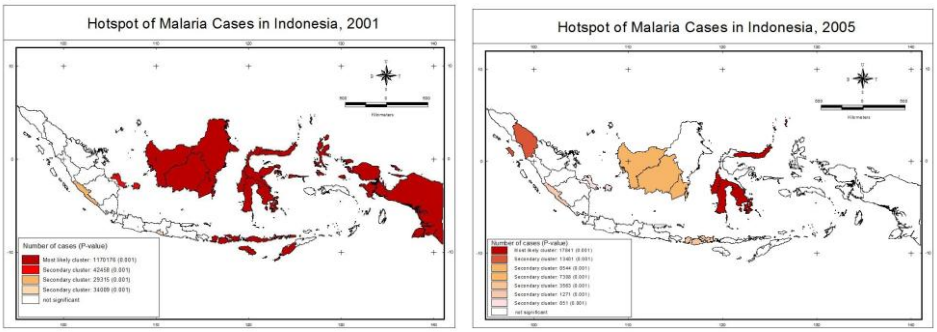


Figure 3. Hotspot of Malaria Cases, 2001 and 2005

There are different results about hotspot areas between the health department report and scan statistics results. The results were based on the reported data. It is believed that many cases are not reported to Health Department. It could be seen, as an example, in 2004 and 2005 malaria cases were high because the government received global fund to investigate malaria cases. It was possible that before 2004 the cases also high but they were not reported.

4. COMMENTS AND CONCLUSIONS

According to the result of calculation for HIV/AIDS mortality in Indonesia, there was a moving hotspot area of the mortality number from year 2002 to 2006. In 2002, AIDS mortality hotspot is around the east part of Indonesia, but in 2006, it was around the central part, include Central Java and Yogyakarta. The news about rising AIDS cases in Yogyakarta in 2007 is relevant with the analysis results. The second highest was in the Central part of Sumatera.

In 2001, the hotspot regions of TB cases were Papua, Sulawesi, and Kalimantan. Furthermore, in 2005, the highest TB cases were in Maluku, Sulawesi, and Papua. As a result, hotspots of TB cases were relatively moving just around east and central part of Indonesia and some areas in Sumatera. In addition, hotspot of TB cases also high in North Sumatra in 2004 and 2005, whereas it was not significant in 2001 and 2002.

Hotspot area of malaria cases was narrower from 2001 to 2005. In 2001, the hotspot areas of malaria cases were Kalimantan, Sulawesi, North Maluku, and Papua, while in 2005 hotspot areas were North Sulawesi, Gorontalo, Southeast Sulawesi, and South Sulawesi. The malaria cases was move from the east to some areas of Sumatera, Kalimantan, and Sulawesi.

As a conclussion, contiguos diseases tended to move from the east part to the central part in period 2001 to 2006 On the basis of study, prevention strategies are recommended that focus on these hotspot areas. The present study analyzed the association between human population and diseases cases. Gathering and including vector population data (including species, population

density, distribution, and infection prevalence rate) and environmental variables in the risk analysis of a disease in these areas provide a more comprehensive view of the disease risk.

ACKNOWLEDGMENTS

Our thanks are due to Prof. Martin Kulldorff for allowing us to use the software and methods. Also thanks are due to The American University in Cairo (AUC), New Cairo, Egypt for the conference facilities and to The University of Indonesia for the support and fund.

REFERENCES

Depkes. Departemen Kesehatan Republik Indonesia. (2006) Ditjen PP & PL. Jakarta.

Depkes. Departemen Kesehatan Republik Indonesia. (2007). *Indonesia Berada di Urutan Tiga Besar Kasus TBC,* http://www.Kapanlagi.com

ESRI. 1996 – 2000. ArcGIS V3.3, Using ArcView GIS.

Kulldorff, M. (1997). A Spatial Scan Statistic. *Communication in Statistics: Theory and Methods*, 26:1481—1496.

Kulldorff, M. (1999). Spatial Scan Statistics: Models, Calculations, and Application. National Cancer Institute, Bethesda, MD

Kulldorff, M. (2006). *SatScan User Guide version 6.1*.

Kulldorff, M., (2007). HIV/AIDS Cases on the Rise in Yogya. *The Jakarta Post*.

Oemijati, S. (1992). *Risk Behavior in Malaria Transmission in Indonesia*. Southeast Asian J Trop Med Public Health 1992; 23: 47-50.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 1094–1108.

Shrinkage Estimation for Cumulative Logit Models

Faisal Maqbool Zahid[1] & Christian Heumann [2]

Department of Statistics
Ludwig Strasse 33, 80539. Ludwig-Maximilians-University Munich, Germany

E-mail[1]: faisalmz99@yahoo.com
E-mail[2]: christian.heumann@stat.uni-muenchen.de

Abstract

A shrinkage estimation method for cumulative logit models is developed. The proposed method is based on shrinking the responses for each category towards the underlying probabilities. The method handles not only the problem of separation in the cumulative logit models but estimates also exist when the number of covariates is larger relative to the sample size. The estimates exist even when MLE does not exist. The computation of the parameter estimates for cumulative logit model with the proposed method is very simple and can easily be done with all commonly used statistical packages supporting fitting procedures using weights. Estimates are compared with the MLE in a simulation study and application.

KEY WORDS: Logistic regression, Shrinkage estimation, Pseudo data, Cumulative logit

1 INTRODUCTION

Regression models with ordinal responses are usually fitted using the maximum likelihood approach. The usual ML estimates face the problems and may not exist when the model under consideration has larger number of covariates relative to the sample size. The method of maximum likelihood estimation is also sensitive to outliers. The penalization methods are the alternative in such situations which use the penalized likelihood function for the estimation of parameters. Ridge regression, one of the oldest penalization methods for linear models, was extended to GLM type models by Nyquist [5], although a definition of a ridge estimator for the logistic regression model, which is a particular case of generalized linear models was suggested by Schaefer et al. [8] and Schaefer [7]. Segerstedt [9] discussed a generalization of ridge regression for ML estimation in GLM. Although many alternative penalization/shrinkage methods are proposed for univariate GLMs in the literature but according to knowledge of the authors, not so much literature is available for multicategory responses. Zhu and Hastie [11] use ridge type penalization, Krishnapuram et al. [3] consider multinomial logistic regression with lasso type estimates and Friedman et al. [2] use the elastic net with L1 and L2 penalties as its special cases. Tutz and Leitenstorfer [10] considered a shrinkage type

estimator for binary regression that has connection with the estimators of Rousseeuw and Christmann [6] which are based on the responses which are closely related but not equal to the unobservable true responses. Rousseeuw and Christmann [6] estimates are robust against separation and always exist.

Our technique for shrinking the parameter estimates is not using any penalty term as used in ridge regression or lasso but we are shrinking the discrete responses appearing as 0 or 1 in the same fashion followed by Tutz and Leitenstorfer [10] for binary responses. These discrete values are the exaggeration of the true unobservable probabilities, and we are shrinking these values towards the underlying probabilities by replacing or one can say that by transforming $0/1$ values to some corresponding higer/smaller values (probabilities). To do this we are making use of some pseudo data sets. We make $q = k-1$ pseudo observations for each response in the original data. For example, for a three-categories (ordered) response variable, the categories are labeled as $1,2$ and 3. Let for a particular response the category 1 appears, then for this response there will be two pseudo responses with categories labels 3 and 2 (with clockwise rotation of $1,2,3$) or 2 and 3 (with anti-clockwise rotation of $1,2,3$) respectively. So against the ith response with category j, rest of $k-1$ categories get the representation in $k-1$ pseudo responses respectively with identical value of the covariate $\mathbf{x}_i$ and as a result we are then downgrading $y_{.j} = 1$ using its counterpart categories by assigning different weights to the original and corresponding pseudo responses. In this case rather than using the log-likelihood for n observations we are working with a weighted log-likelihood for kn observations. This approach is not only simple and easy to implement but also robust in terms of existence of estimates and provides significant improvement over usual MLE. Any statistical software used for fitting the logit models with weights for ordinal responses can be used for implementing this technique.
In section 2 the basic idea of shrinking estimation is described. Section 3 descibes the way to decide about the weights for the original and pseudo observations. The performance of shrinkage methods is investigated and compared with the usual MLE in section 4. In section 5 the estimates are computed for a real data set. Section 6 completes the discussion with some concluding remarks.

2 SHRINKAGE ESTIMATION WITH DATA TRANSFORMATION

For ordinal responses several statistical models such as cumulative logit model, continuation-ratio model, constrained and unconstrained partial proportional odds model, adjacent-category logit model, polytomous logit model and stereotype logistic model, can be used. Ananth and Kleinbaum [1] describe these models and interpretation of the model parameters. Cumulative logit model (also called proportional odds model by [4]) is most commonly used model and is followed in this text. The shrinkage technique described here can be applied in the same way to the other models for ordinal responses. For the cumulative logit model, let for a given vector $\mathbf{x}$ of explanatory variables, there is an observable variable $Y \in \{1,\ldots,k\}$ connected with an unobservable latent variable Z as

$$Y = j \quad \Leftrightarrow \quad \theta_{j-1} < Z < \theta_j, \qquad j = 1,\ldots,k$$

where $-\infty = \theta_0 < \theta_1 < \ldots < \theta_k = \infty$. This indicates that Y is a categorized version of Z determined by $\theta_1,\ldots,\theta_{k-1}$. The cumulative logistic model has the form

$$P(Y \leq j) = P(Z \leq \theta_j) = \frac{exp(\theta_j - \boldsymbol{\beta}^T\mathbf{x})}{1+exp(\theta_j - \boldsymbol{\beta}^T\mathbf{x})} \qquad j = 1,\ldots,q = k-1. \tag{1}$$

The alternative form can be shown as

$$\log\left[\frac{\phi_j(\mathbf{x})}{1-\phi_j(\mathbf{x})}\right] = \theta_j - \boldsymbol{\beta}^T\mathbf{x} \qquad j = 1,\ldots,k-1, \tag{2}$$

where $\phi_j(\mathbf{x}) = P(Y \leq j|\mathbf{x})$ is the cumulative probability up to and including category j, when the covariate vector is $\mathbf{x}$. In case of cumulative logistic model one gets the link function as

$$g_j(\pi_{i1},\ldots,\pi_{iq}) = \log\left[\log\left(\frac{\phi_r}{1-\phi_r}\right) - \log\left(\frac{\phi_{r-1}}{1-\phi_{r-1}}\right)\right] \tag{3}$$

(2) is known as the proportional-odds model, where each cumulative logit has its own intercept θ_j $(j = 1,\ldots,q)$, and $\{\theta_j\}$ are increasing in j. The negative sign in (2) ensures that the probability is increasing for large values of $\boldsymbol{\beta}^T\mathbf{x}$ with increasing j. In (2) $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are unknown and $\boldsymbol{\theta}$ must satisfy $\theta_1 \leq \ldots \leq \theta_{k-1}$ to ensure that the probabilities are non-negative. In case of k responses (ordered) and p covariates, for the estimation of parameters, let the cumulative logit model has the form

$$\text{logit}(\phi_{ij}) = \mathbf{X}_i\boldsymbol{\beta}^*$$

where $\mathbf{X}_i$ (fixed coefficients), are the components of $nk \times p^*$ design matrix $\mathbf{X}$ and $\boldsymbol{\beta}^*$ is the vector of length $p^* = p+k-1$ for the model (2) with components $\boldsymbol{\beta}^* = (\theta_1,\ldots,\theta_q,\beta_1,\ldots,\beta_p)$. The design matrix $\mathbf{X}$ is given as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix}$$

with $\mathbf{X}_i$, a $q \times p^*$ matrix given by

$$\mathbf{X}_i = [\,\mathbf{I}_{k-1} : \mathbf{1}\mathbf{x}_i\,] = \begin{bmatrix} 1 & & & & \mathbf{x}_i^T \\ & 1 & & & 0 \\ & & \ddots & & \vdots \\ & & & 1 & 0 \end{bmatrix}$$

The original data set is increased by k times by generating q pseudo data sets of the original data. For the jth category against the ith response, each of the rest of q categories are given representation (with different weights) in the ith row of each of the shadow data with identical value of the design point. Since we have k categories numbered $1,2,\ldots,k$ and rotating these numbers clock wise (or anti-clock-wise) we get q arrangements as $(k,1,\ldots,k-1)$, $(k-1,k,1,\ldots,k-2)$, $\ldots$, $(2,3,\ldots,k,1)$. For the jth category in the original data, alternative category for the ith row of each of the shadow data sets can be chosen from these q arrangements respectively. For example, in the case of three categories, if we have a sample of size three with outcome category labels as $(1,2,3)$, then it has $k-1=2$ pseudo data sets with outcome category labels as $(3,1,2)$ and $(2,3,1)$ which can be then written in the form of a response matrix of dummy variables. In

this example

$$\text{pseudo data sets of the original response matrix } \begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} \text{ are } \begin{bmatrix}0&0&1\\1&0&0\\0&1&0\end{bmatrix} \text{ and } \begin{bmatrix}0&1&0\\0&0&1\\1&0&0\end{bmatrix}.$$

with last column as redundant. Each pseudo observation of the ith response has identical values for the design point $\mathbf{x}_i$ for $i = 1,2,\ldots,n$. As a result, the size of our data is increased from n to kn. If we assign $\delta_s \ \ (s = 1,\ldots,q)$ weights to the sth pseudo data and $\delta_0 = 1 - \sum_{s=1}^{k-1} \delta_s$ weights to the original data where $\delta_s \in [0, \frac{1}{k}]$. In this case the weighted log likelihood function for kn observations is given by

$$l_w(\boldsymbol{\beta}) = \sum_{i=1}^{kn} w_i l_i(\boldsymbol{\beta}), \tag{4}$$

where

$$l_i(\boldsymbol{\beta}) = \sum_{j=1}^{k} \log \pi_j(\mathbf{x}_i)^{y_{ij}} = \sum_{j=1}^{k} \log\left[\phi_j(\mathbf{x}_i) - \phi_{j-1}(\mathbf{x}_i)\right]^{y_{ij}}$$

and

$$w_i = \begin{cases} \delta_{i0} & i \le n \\ \delta_{is} & i > n. \end{cases}$$

For $\delta_{is} = 0$, we have usual (unweighted) likelihood function for cumulative logit model. As δ_{is} get larger values, the pseudo data will get more weights and original responses get low weights.
If the response variable has three (ordered) categories labeled $1,2$ and 3. There will be two pseudo data sets getting weights δ_{i1} and δ_{i2} respectively and the weights for original data are $\delta_{i0} = 1 - \sum_{s=1}^{2} \delta_{is}$. The weighted log-likelihood function in this case is given by

$$\begin{aligned}
l_w(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \Bigg[\delta_{i0} \Big\{ y_{i1}\log(\pi_{i1}) + y_{i2}\log(\pi_{i2}) + y_{i3}\log(\pi_{i3}) \Big\} \\
&\qquad + \delta_{i1} \Big\{ y_{i3}\log(\pi_{i1}) + y_{i1}\log(\pi_{i2}) + y_{i2}\log(\pi_{i3}) \Big\} \\
&\qquad + \delta_{i2} \Big\{ y_{i2}\log(\pi_{i1}) + y_{i3}\log(\pi_{i2}) + y_{i1}\log(\pi_{i3}) \Big\} \Bigg] \\
&= \sum_{i=1}^{n} \Bigg[\Big\{ y_{i1} + (y_{i3} - y_{i1})\delta_{i1} + (y_{i2} - y_{i1})\delta_{i2} \Big\} \log(\pi_{i1}) \\
&\qquad + \Big\{ y_{i2} + (y_{i1} - y_{i2})\delta_{i1} + (y_{i3} - y_{i2})\delta_{i2} \Big\} \log(\pi_{i2}) \\
&\qquad + \Big\{ y_{i3} + (y_{i2} - y_{i3})\delta_{i1} + (y_{i1} - y_{i3})\delta_{i2} \Big\} \log(\pi_{i3}) \Bigg] \\
&= \sum_{i=1}^{n} \tilde{y}_{i1}\log(\pi_{i1}) + \tilde{y}_{i2}\log(\pi_{i2}) + \tilde{y}_{i3}\log(\pi_{i3}) = \sum_{i=1}^{n}\sum_{j=1}^{3} \tilde{y}_{ij}\log(\pi_{ij})
\end{aligned}$$

This expression for $l_w(\boldsymbol{\beta})$ indicates that use of pseudo data sets with different weights lead to the transformed responses $\tilde{y}_{ij}$ and weighted log-likelihood with pseudo data simplifies to an un-weighted log-likelihood of transformed responses $\tilde{y}_{ij}$ with probabilites π_{ij}. The weights here are being used to process the exaggerated values of the original responses of the form 1 or 0 to transform them in a more realistic smaller/higher values respectively. This expression with three response categories can easily be extended for the k response categories and is given as

$$l_w(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{k} \tilde{y}_{ij}\log(\pi_{ij}) \tag{5}$$

with the ith transformed observation for the jth category

$$\tilde{y}_{ij} = \begin{cases} y_{i1} + \sum_{r=2}^{k}(y_{ir} - y_{i1})\delta_{i,k-(j-1)} & \text{if } \ j = 1, \\ y_{ik} + \sum_{r=1}^{k-1}(y_{ir} - y_{ik})\delta_{i,(j-r)} & \text{if } \ j = k, \\ y_{ij} + \sum_{r=1}^{j-1}(y_{ir} - y_{ij})\delta_{i,k-(r+1)} + \sum_{r=j+1}^{k}(y_{ir} - y_{ij})\delta_{i,k-(r-j)} & \text{otherwise.} \end{cases}$$

One can proceed with the log-likelihood function given in (5) rather than working with weighted log-likelihood given in (4), but in this text we follow the weighted version of log-likelihood for kn observations using the pseudo data sets. The score function for the weighted log-likelihood function is given by

$$s_w(\boldsymbol{\beta}^*) = \frac{\partial l_w(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*} = \sum_{i=1}^{kn} s_{wi}(\boldsymbol{\beta}^*),$$

with the components

$$s_{wi}(\boldsymbol{\beta}^*) = \mathbf{X}_i^T \operatorname{diag}(w_i)\mathbf{D}_i(\boldsymbol{\beta}^*)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}^*)[\mathbf{y}_i - h(\boldsymbol{\eta}_i)],$$

where w_i are the weights, $\mathbf{D}_i(\boldsymbol{\beta}^*) = \frac{\partial h(\boldsymbol{\eta}_i)}{\partial \boldsymbol{\eta}}$ is the derivative of $h(\boldsymbol{\eta})$ evaluated at $\boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta}^*$ and $\boldsymbol{\Sigma}_i(\boldsymbol{\beta}^*) = cov(\mathbf{y}_i)$ is the covariance matrix of ith observation of $\mathbf{y}$ given parameter vector $\boldsymbol{\beta}^*$. Alternatively

$$s_{wi}(\boldsymbol{\beta}^*) = \mathbf{X}_i^T \operatorname{diag}(w_i)\mathbf{W}_i(\boldsymbol{\beta}^*)\frac{\partial g(\boldsymbol{\mu}_i)}{\partial \boldsymbol{\mu}^T}[\mathbf{y}_i - h(\boldsymbol{\eta}_i)]$$

with

$$\mathbf{W}_i(\boldsymbol{\beta}^*) = \mathbf{D}_i(\boldsymbol{\beta}^*)\boldsymbol{\Sigma}_i^{-1}\mathbf{D}_i^T(\boldsymbol{\beta}^*) = \left\{\frac{\partial g(\boldsymbol{\mu}_i)}{\partial \boldsymbol{\mu}^T}\boldsymbol{\Sigma}_i(\boldsymbol{\beta})\frac{\partial g(\boldsymbol{\mu}_i)}{\partial \boldsymbol{\mu}}\right\}^{-1}$$

In matrix notation

$$\begin{aligned} s_w(\boldsymbol{\beta}^*) &= \sum_{i=1}^{n} \mathbf{X}_i^T \operatorname{diag}(w_i)\mathbf{D}_i(\boldsymbol{\beta}^*)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}^*)[\mathbf{y}_i - h(\boldsymbol{\eta}_i)] \\ &= \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{D}_i(\boldsymbol{\beta}^*)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}^*)[\tilde{\mathbf{y}}_i - h(\boldsymbol{\eta}_i)] \\ &= \mathbf{X}^T \mathbf{D}(\boldsymbol{\beta}^*)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)[\tilde{\mathbf{y}} - h(\boldsymbol{\eta})] \end{aligned} \tag{6}$$

where $\mathbf{y}$ and $h(\boldsymbol{\eta})$ are given by

$$\mathbf{y}^T = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T), \quad h(\boldsymbol{\eta}) = (h(\boldsymbol{\eta}_1), \ldots, h(\boldsymbol{\eta}_n))^T.$$

The matrices have block diagonal form

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}^*) = \operatorname{diag}(\boldsymbol{\Sigma}_i(\boldsymbol{\beta}^*)), \quad \mathbf{W}(\boldsymbol{\beta}^*) = \operatorname{diag}(\mathbf{W}_i(\boldsymbol{\beta}^*)), \quad \mathbf{D}(\boldsymbol{\beta}^*) = \operatorname{diag}(\mathbf{D}_i(\boldsymbol{\beta}^*)),$$

The simple form of the score equations is given by

$$\begin{aligned} \frac{\partial l_w(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*} &= \sum_{ij} x_{ij} w_i (y_{ij} - h(\eta_{ij})) \\ &= \sum_{ij} x_{ij} (\tilde{y}_{ij} - h(\eta_{ij})) = 0 \end{aligned} \tag{7}$$

or in matrix notation it can be written as

$$\mathbf{X}_{nk\times p^*}^T \operatorname{diag}(w_i)[\mathbf{y}_{nk\times 1} - h(\boldsymbol{\eta})_{nk\times 1}] = \mathbf{X}^T[\tilde{\mathbf{y}} - h(\boldsymbol{\eta})] = \mathbf{0}$$

Score function in (7) uses shrinked responses corresponding to the original responses $y_{\cdot j}$ (closing to $\frac{1}{k}$) such that for $y_{\cdot j} = 1$, $\tilde{y}_{\cdot j}$ assumes the value $1 - \sum_{s=1}^{q} \delta_s$, which is less than 1, and $y_{\cdot j} = \frac{1}{k}$ if we use the same weights $\delta_s = \frac{1}{k}$ $(s = 1, \ldots, q)$, in which case (7) leads to the solution $\boldsymbol{\beta}^* = \mathbf{0}$.
The use of weighted score function by assigning different weights to pseudo data sets and the original data, ensures the existence of estimates, also in the situations where the usual MLE fails to exist e.g., if we have large number of covariates relative to the sample size or if there is some problem of separation in the data. The estimation with weighted responses is very simple and any statistical software that allows the fitting of cumulative logit models with weights can be used to obtain the estimates.

3 WEIGHTS FOR SHRINKAGE TECHNIQUE (WMLE)

The basic idea is that instead of using the exaggerated observed responses in the form of dummy variables as 1 or 0, a weighted smoothed version of these responses should be used. To downgrade the responses $y_{ij} = 1$, the weights can be chosen in the interval $[0, \frac{1}{k}]$. In this section for the cumulative logit models with intercept

we exploit the property of MLE

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_{ij}=\bar{y}_{.j},\qquad j=1,2,\ldots,k \tag{8}$$

and decide about the weights based on the fulfillment of this property. Here $\bar{y}_{.j}$ is the mean of responses corresponding to the jth category. Different weights should be assigned to each of $y_j \quad (j=1,\ldots,k)$ to hold the property (8). From (7) it is clear that there will be $k-1$ score equations corresponding to the intercept terms and are of the form

$$\sum_{i=1}^{n}\tilde{\mathbf{y}}_{i.}=\sum_{i=1}^{n}\hat{\boldsymbol{\pi}}_{i.} \tag{9}$$

Let $\tilde{\delta}_j$ be the weight associated with $y_{ij} \quad j=1,\ldots,k$, then from (9) under the MLE property given in (8) we obtain $(k-1)$ equations of the form

$$\sum_{\substack{j=1\\ j\neq r}}^{k}\bar{y}_{.j}\tilde{\delta}_j-(k-1)\bar{y}_{.r}\tilde{\delta}_r=0 \qquad r=1,\ldots,k-1. \tag{10}$$

solving this system of $k-1$ equations for $\tilde{\delta}$'s, we get

$$\tilde{\delta}_j=\frac{\bar{y}_{.k}}{\bar{y}_{.j}}\,\tilde{\delta}_k,\qquad j=1,\ldots,k-1. \tag{11}$$

Here if $\tilde{\delta}_k=\bar{y}_{.j}$ we then have $\tilde{\delta}_j=\bar{y}_{.k} \ \ \forall j$, i.e., in this situation each response $y_{.j}=1$ is shrinking towards $\bar{y}_{.k}$, the mean of the responses for the kth category. The optimum value of the weight $\tilde{\delta}_k$ can be searched in the interval $[0,\bar{y}_{.j}]$. But in case if all weights are equal to $\frac{1}{k}$, the solution for the estimates will be $\hat{\boldsymbol{\beta}}_w=\mathbf{0}$. It is intuitive to search the optimum value of the weight $\tilde{\delta}_k$ in interval $[0,\frac{1}{k}]$ when $\bar{y}_{.j}>\frac{1}{k}$. Since for $\tilde{\delta}_k=\bar{y}_{.j}$ each of the response $\bar{y}_{.j}=1 \quad (j=1,\ldots,k-1)$, shrinks towards $\bar{y}_{.k}$, it is sensible to shrink the response $y_{.k}=1$ towards the mean of rest of $k-1$ responses i.e., shrinking $y_{.k}=1$ towards $\frac{1}{k-1}\sum_{j=1}^{k-1}\bar{y}_{.j}$. The weighting scheme for the original data and the lth $(l=1,\ldots,k-1)$ pseudo data set on the basis of (11) is given by

$$\alpha_i=\begin{cases}1-\sum_{l=1}^{k-1}\alpha_{il} & i\leq n\\ \alpha_{il} & i=1,\ldots,n \ \ \forall l\end{cases} \tag{12}$$

where

$$\alpha_{il}=\begin{cases}\left(\frac{\bar{y}_{.k}}{\bar{y}_{.j}}\right)\delta_j & \text{for } y_{j\in\{1,\ldots,k-1\}}=1\\ \frac{1}{k-1}\sum_{j=1}^{k-1}\delta_j & \text{for } y_k=1\end{cases}$$

for $i=1,\ldots,n$ and $l=1,\ldots,k-1$. The optimum values of the tuning parameters δ_j can be decided on the basis of cross-validation criteria. The statistical distances used in this text for the cross-validation purpose are:

averaged Kullback-Leibler discrepancy

$$\mathrm{L_{KL}} = \sum_{i=1}^{n}\sum_{j=1}^{k} \pi_{ij}\log\left(\frac{\pi_{ij}}{\hat{\pi}_{ij}}\right)$$

with a convention that $0.\log(0) = 0$, and the averaged squared error

$$\mathrm{ASE} = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}(\pi_{ij} - \hat{\pi}_{ij})^2,$$

In this text leave-one-out cross-validation is used i.e., for the *i*th observation the fit is computed using all the data except the *i*th observation. However to save the time and computational burden one can use the *k*-fold cross-validation for searching the optimum values of the tuning parameters.

4 SIMULATION STUDY

In a simulation study, we generated Gaussian data with n observations and p covariates. We used different number of combinations of n ($n = 30, 50$ and 100) and p ($p = 2, 5, 10, 15$ and 20). The values of the parameters used are $\beta_j = (-1)^j exp(-2(j-1)/20)$ for $j = 1, \ldots, p$ and for the intercept terms $\beta_{01} = -0.3$ and $\beta_{02} = 0.8$. The covariates are drawn from $N(0,1)$. In each combination of n and p, $S = 200$ data sets are generated. For the computation of usual MLE and the shrinkage estimates, the function `polr` of the R package `MASS` is used. Only those data sets are considered in the study for which the usual MLE exists. The combination of $n = 30$ and $p = 20$ is the exception where the ML estimates are not existing and therefore in Table 1 the results for MLE are not available for this combination. In Table 1, shrinkage estimates with weights given in (12), are compared with the usual MLE in terms of MSE($\hat{\boldsymbol{\beta}}$) and MSE($\hat{\boldsymbol{\pi}}$). For shrinkage estimates, optimal values of the tuning parameters are chosen by leave-one-out cross validation based on error measures described in section 3 and the corresponding results are denoted by CV(KL) and CV(SE) for Kullback-Leibler and squared error loss respectively. MSE($\hat{\boldsymbol{\beta}}$) and MSE($\hat{\boldsymbol{\pi}}$) are computed as:

$$\mathrm{MSE}(\hat{\boldsymbol{\pi}}) = \tfrac{1}{S}\textstyle\sum_s \mathrm{MSE}_s(\hat{\boldsymbol{\pi}}) \quad \text{with} \quad \mathrm{MSE}_s(\hat{\boldsymbol{\pi}}) = \tfrac{1}{kn}\displaystyle\sum_{i=1}^{n}\sum_{r=1}^{k}(\hat{\pi}_{ir} - \pi_{ir})^2 \text{ for the } s\text{th sample}$$

and

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}) = \tfrac{1}{S}\textstyle\sum_s ||\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}||^2$$

where $\hat{\boldsymbol{\pi}}$ is a vector of length kn and $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ are of length $p + k - 1$. If $\mathrm{MSE_s}$ is the MSE of $\hat{\pi}$ (or $\hat{\boldsymbol{\beta}}$) for shrinkage method and MSE_s^{ML} is the corresponding MSE for the maximum likelihood estimate, then the ratio $\mathrm{MSE}_s/\mathrm{MSE}_s^{ML}$ for the sth simulation will provide a measure of improvement of shrinkage method over MLE. The distribution of the ratios $\mathrm{MSE}_s/\mathrm{MSE}_s^{ML}$ is skewed and therefore the logarithms of these ratios are considered. In Table 1 along with the MSEs, the means of $\log(\mathrm{MSE}_s/\mathrm{MSE}_s^{ML})$ denoted by $lR_{ML}(\hat{\pi})$ and $lR_{ML}(\hat{\boldsymbol{\beta}})$ are considered for comparing the shrinkage estimates with the usual MLE. The negative values of these log-ratios refer to an improvement of shrinkage estimates over the usual ML estiamtes. Although the main focus was on the development of estimation method that is more robust than usual MLE in terms of

Table 1: Simulation results for comparison of MLE and shrinkage estimates in terms of MSE($\hat{\boldsymbol{\pi}}$) and MSE($\hat{\boldsymbol{\beta}}$)

		MLE		WMLE, CV(KL)				WMLE, CV(SE)			
p	n	MSE($\hat{\pi}$)	MSE($\hat{\boldsymbol{\beta}}$)	MSE($\hat{\pi}$)	$lR_{\mathrm{ML}}(\hat{\pi})$	MSE($\hat{\boldsymbol{\beta}}$)	$lR_{\mathrm{ML}}(\hat{\boldsymbol{\beta}})$	MSE($\hat{\pi}$)	$lR_{\mathrm{ML}}(\hat{\pi})$	MSE($\hat{\boldsymbol{\beta}}$)	$lR_{\mathrm{ML}}(\hat{\boldsymbol{\beta}})$
2	30	0.0368	0.9791	0.0366	−0.0333	0.5952	−0.4047	0.0355	−0.0592	0.5950	−0.3957
	50	0.0229	0.5691	0.0221	−0.0698	0.4180	−0.2410	0.0227	−0.0395	0.4486	−0.1964
	100	0.0096	0.2127	0.0099	−0.0207	0.1827	−0.1379	0.0101	−0.0043	0.1852	−0.1007
5	30	0.0692	3.3661	0.0682	−0.0312	1.2873	−0.6145	0.0667	−0.0720	1.3580	−0.6159
	50	0.0421	2.0174	0.0385	−0.0660	1.0261	−0.2778	0.0383	−0.0638	1.0371	−0.2498
	100	0.0203	0.5253	0.0187	−0.0834	0.4126	−0.1945	0.0190	−0.0658	0.4226	−0.1679
10	30	0.1042	5.6688	0.0992	−0.0241	2.5853	−0.4784	0.0993	−0.0241	2.5853	−0.4780
	50	0.0744	4.4952	0.0569	−0.2618	1.8216	−0.6804	0.0582	−0.2340	1.9481	−0.6085
	100	0.0387	1.5693	0.0313	−0.2185	0.8000	−0.5277	0.0315	−0.2126	0.8148	−0.5154
15	30	0.1508	13.7381	0.2149	0.3805	4.3969	−0.7083	0.2160	0.3867	4.3988	−0.7081
	50	0.0980	9.0355	0.0875	−0.0854	2.8129	−0.9129	0.0858	−0.1053	2.7818	−0.9265
	100	0.0569	3.0941	0.0416	−0.3010	1.1224	−0.8178	0.0420	−0.2957	1.1704	−0.8043
20	30	–	–	0.1544	–	10.6867	–	0.1674	–	16.8981	–
	50	0.1201	6.8241	0.1291	0.0897	3.7917	−0.4628	0.1295	0.0878	3.7792	−0.4674
	100	0.0747	5.4948	0.0520	−0.3548	1.5407	−1.0737	0.0520	−0.3601	1.5386	−1.0808

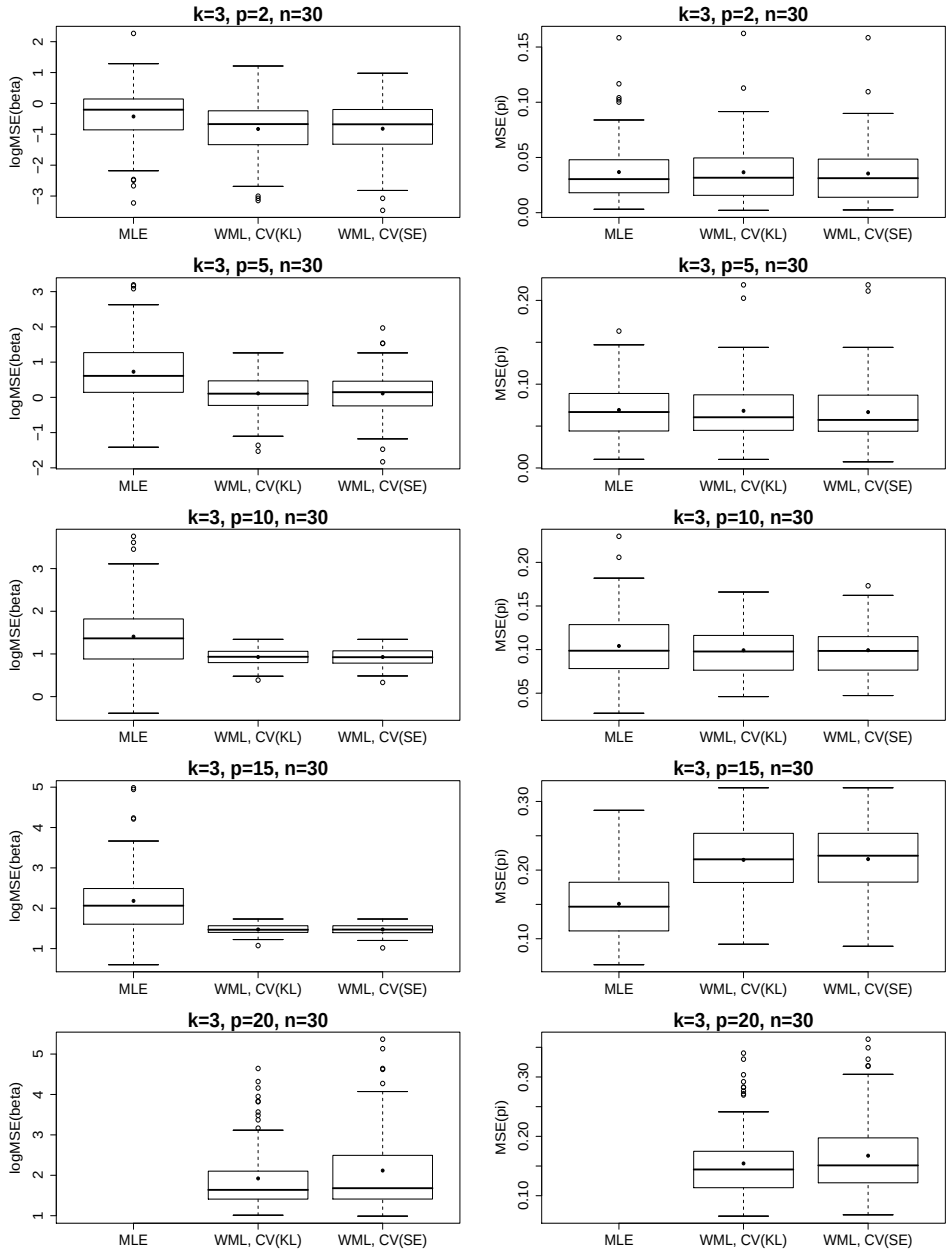


Figure 1: Illustration of the simulation study: Box plots of $\log(\mathrm{MSE}(\hat{\boldsymbol{\beta}}))$ (left column) and $\mathrm{MSE}(\hat{\boldsymbol{\pi}})$ (right column) for $n = 30$.

existence of parameter estimates especially in case of larger number of covariates with small samples or for no overlapping observations in the data, we also considered a simple case of only two covariates even with a large sample size to observe the behaivour of shrinkage estimates. We were expecting the MLE

with favourable asymptotic properties to perform better than shrinkage estimates with large samples in low dimension but results in Table 1 show better performance of our estimates in every situation not only with respect to the parameter estimates but also the fit. For the case $p = 20$ and $n = 30$, MLE is not existing but the shrinkage estimates do exist. In case of $p = 15$ and $n = 30$ our approach is showing some weakness in terms of $\hat{\boldsymbol{\pi}}$ but still giving excellent performance in terms of $\hat{\boldsymbol{\beta}}$. The shrinkage estimates and the ML estimates are compared with respect to MSE($\hat{\boldsymbol{\beta}}$) and MSE($\hat{\boldsymbol{\pi}}$) in terms of box plots in Fig.1 for the most interesting case of small samples, i.e., $n = 30$. The bullets (solid circles) within the boxes are the mean of 200 values for which the box plots are drawn. The shrinkage technique also showed good performance with increasing number of response categories in the simulation studies, the results of which are not shown here.

5 APPLICATION

In this section we are comparing our shrinkage estimates with the usual MLE using the "housing" data set provided by the R library `MASS`. The response variable in this data set is "Sat"(Satisfaction of householders with their present housing circumstances) with three ordered categories (1:Low(L), 2: Medium(M), 3: High(H)). The covariates are "Infl" (Perceived degree of influence householders have on the management of the property) with three categories (Low(L), Medium(M) and High(H)), "Type" (Type of rental accommodation) with four categories (Tower(Tw), Atrium(At), Apartment(Ap), Terrace(Tr)) and "Cont" (Contact residents are afforded with other residents) with two levels (Low(L) and High(H)). The sample size of actual data set is $n = 1681$. We draw a random sample of size $n = 50$ from the actual data and proceed with this data set of 50 responses. This data set can be accessed at `www.stat.uni-muenchen.de/~zahid/housing.txt`.

Table 2: Estimates and standard errors for "housing" data

Estimation Method	Intercept 1	Intercept 2	Infl(M)	Infl(H)	Type(At)	Type(Ap)	Type(Tr)	Cont(H)
MLE	−0.5109	1.1088	−0.2794	−1.7398	−0.2687	0.1178	−0.1470	−0.3219
	(0.7179)	(0.7375)	(0.6387)	(0.8178)	(0.7078)	(0.7782)	(0.9984)	(0.6132)
WMLE, CV(KL)[a]	−0.7729	1.0207	−0.0568	−0.3361	−0.0493	0.0169	−0.0216	−0.0685
	(0.1937)	(0.1956)	(0.1691)	(0.1839)	(0.1881)	(0.2073)	(0.2281)	(0.1464)
WMLE, CV(SE)[b]	−0.7311	1.0204	−0.0739	−0.4419	−0.0648	0.0234	−0.0288	−0.0891
	(0.2320)	(0.2346)	(0.2030)	(0.2219)	(0.2254)	(0.2487)	(0.2785)	(0.1777)

[a] Results are based on optimum values of tuning parameters $\boldsymbol{\delta}^T = (0.1322449, 0.2789116)$.
[b] Results are based on optimum values of tuning parameters $\boldsymbol{\delta}^T = (0.1224490, 0.2517007)$.

Table 3: MSPE and Mean deviances

Estimation Method	MSPE	Mean Deviance
MLE	7.4997	13.0398
WMLE, CV(KL)	6.6132	10.9193
WMLE, CV(SE)	6.6321	10.9602

We fit the proportional odds model on this data under the assumption that proportional odds assumption is fulfilled. The results for the parameter estimates and their standard errors (within brackets) for usual MLE

and the shrinkage approach are presented in Table 2. An expression for computing the standard errors of the parameter estimates for the shrinkage parameter estimates is derived in Appendix A. For shrinkage the optimum values of tuning parameters δ's used are decided on the basis of leave-one-out cross-validation. The information in Table 2 is reflecting the improved estimates of standard errors of parameter estimates with smaller values than those for ML estimates. The parameter estimates of shrinkage method are compared with usual ML estimates on the basis of prediction error. For this purpose we use two approaches, one is with the help of MSPE (mean squared prediction error) and second using the deviance function. We use 80% of 500 random splits of our data set with 50 observations as the training data set and rest of the 20% as test data set. The parameter estimates are obtained by fitting the model with training data sets and these estimates are used to get the fit and prediction error from the test data sets. The squared prediction error is computed using the formula

$$\text{SPE}_s = \frac{1}{kn}\sum_{i=1}^{n}\sum_{r=1}^{k}(\hat{\pi}_{ir}^{test} - y_{ir}^{test})^2,$$

where y's are the observed responses in the form of dummy variables 0 or 1. The MSPE for 500 random permutations computed as

$$\text{MSPE} = \frac{1}{500}\sum_{s=1}^{500}\text{SPE}_s.$$

and the deviances are calculated as

$$D = \sum_{i=1}^{n}\sum_{r=1}^{k} y_{ir}\log\left(\frac{y_{ir}}{\hat{\pi}_{ir}}\right)$$

if $y_{ir} = 0$, the term $y_{ir}\log\left(\frac{y_{ir}}{\hat{\pi}_{ir}}\right)$ is set to zero. The mean of these 500 deviance values is used to compare our method with MLE. The results of computed mean squared prediction error (MSPE) and mean deviances are presented in Table 3. Although our method mainly focus on the accuracy and existence of estimates but Table 3 shows that it has also better performance in terms of prediction error and reducing the prediction error with both measures i.e., the mean squared prediction error and the deviance as compared to usual MLE.

6 COMMENTS AND CONCLUSION

The shrinkage technique discussed in this text shrinks the parameter estimates without using any penalty term as in ridge or lasso but shrinking the responses towards the true unknown probabilities in a simple and easy way. However rather than choosing an optimum value of penalty term, one has to decide about the optimum values of the tuning parameters using some cross-validation criteria. Pseudo data sets are used here as a tool to compute the estimates and the way the pseudo data sets are used, assure the existence of estimates addressing the problem of separation (if any) and also in case where the number of parameter to be estimated are large relative to the sample size. Our shrinkage estimates are easy to compute with pseudo responses, have better performance than MLE in terms of MSE and have improved existence of estimates i.e., they are robust against separation and in the situations with large number of covariates relative to the number of response observations.

APPENDIX

A. STANDARD ERRORS FOR SHRINKAGE ESTIMATES

The score function of weighted log-likelihood function given in (6) can be written as

$$\begin{aligned} s_w(\boldsymbol{\beta}^*) &= \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{D}_i(\boldsymbol{\beta}^*) \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}^*)[\tilde{\mathbf{y}}_i - h(\boldsymbol{\eta}_i)] \\ &= \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{D}_i(\boldsymbol{\beta}^*) \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}^*)[\mathbf{y}_i - h(\boldsymbol{\eta}_i) + \mathbf{y}_i^*] \end{aligned} \tag{A.1}$$

with $\mathbf{y}_i^* = \tilde{\mathbf{y}}_i - \mathbf{y}_i$. From here the first order approximation yields

$$\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^* \approx \left(\frac{-\partial s_w(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^{*T}} \right)^{-1} s_w(\boldsymbol{\beta}^*).$$

From (A.1) the weighted score function can be written as

$$s_w(\boldsymbol{\beta}^*) = s(\boldsymbol{\beta}^*) + s_\alpha(\boldsymbol{\beta}^*),$$

where

$$\begin{aligned} s(\boldsymbol{\beta}^*) &= \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{D}_i(\boldsymbol{\beta}^*) \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}^*)[\mathbf{y}_i - h(\boldsymbol{\eta}_i)] \quad \text{and} \\ s_\alpha(\boldsymbol{\beta}^*) &= \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{D}_i(\boldsymbol{\beta}^*) \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}^*) \mathbf{y}_i^* \end{aligned}$$

The derivatives needed here are

$$-\frac{\partial s}{\partial \boldsymbol{\beta}^{*T}} = F + \sum_{i=1}^{n} X_i X_i^T \frac{\partial^2 h}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} \{\mathbf{y}_i - \boldsymbol{\pi}_i\},$$

where F is the weighted Fisher matrix,

$$F = \sum_{i=1}^{n} X_i X_i^T \boldsymbol{\Sigma}_i^{-1} \left(\frac{\partial h}{\partial \boldsymbol{\eta}} \right) \left(\frac{\partial h}{\partial \boldsymbol{\eta}} \right)^T, \tag{A.2}$$

and

$$-\frac{\partial s_w}{\partial \boldsymbol{\beta}^{*T}} = \sum_{i=1}^{n} X_i X_i^T \boldsymbol{\Sigma}_i^{-1} \frac{\partial^2 h}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} \mathbf{y}_i^*.$$

The property of score function that $E(s(\boldsymbol{\beta}^*)) = 0$ is not fulfilled for our weighted score function because

$E(s_w(\boldsymbol{\beta}^*)) \neq 0$. For the covariance of weighted score function, after some laborious derivation we get

$$cov(s_w(\boldsymbol{\beta}^*)) = \sum_{i=1}^{n} \boldsymbol{\Gamma}_i^T \boldsymbol{\Sigma}_i \boldsymbol{\Gamma}_i$$

where

$$\boldsymbol{\Gamma}_i = \mathbf{A}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_i^T \mathbf{X}_i$$

with a $q \times q$ matrix $\mathbf{A}_i$ given as

$$\mathbf{A}_i = \begin{bmatrix} \alpha_{i(1)} - \alpha_{i(2)} & \alpha_{i(k)} - \alpha_{i(2)} & \alpha_{i,(k-1)} - \alpha_{i(2)} & \cdots & \alpha_{i(3)} - \alpha_{i(2)} \\ \alpha_{i(2)} - \alpha_{i(3)} & \alpha_{i(1)} - \alpha_{i(3)} & \alpha_{i(k)} - \alpha_{i(3)} & \cdots & \alpha_{i(4)} - \alpha_{i(3)} \\ \alpha_{i(3)} - \alpha_{i(4)} & \alpha_{i(2)} - \alpha_{i(4)} & \alpha_{i(1)} - \alpha_{i(4)} & \cdots & \alpha_{i(5)} - \alpha_{i(4)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha_{i,(k-1)} - \alpha_{i(k)} & \alpha_{i,(k-2)} - \alpha_{i(k)} & \alpha_{i,(k-3)} - \alpha_{i(k)} & \cdots & \alpha_{i(1)} - \alpha_{i(k)} \end{bmatrix}$$

for $\alpha_{i(l)}$, the weight corresponding to the ith observation in the original data $(l = 1)$ and q pseudo data sets $(l = 2, \ldots, k)$. The term $\partial s_\alpha / \partial \boldsymbol{\beta}^{*T}$ in the expression for $cov(s_w(\boldsymbol{\beta}))$ may be neglected asymptotically as $\alpha_{i(l)} \to 0$, for $l = 2, \ldots, q, \forall i$ with increasing sample size and one obtains as approximation the sandwich matrix

$$cov(\hat{\boldsymbol{\beta}}) = F(\boldsymbol{\beta}^*)^{-1} \mathrm{cov}(s_w(\boldsymbol{\beta}^*)) F(\boldsymbol{\beta}^*)^{-1}. \tag{A.3}$$

With $\alpha_{i(l)} = 0$ for $l = 2, \ldots, q, \forall i$, the covariance coincides with that of usual MLE. The accuracy of approximation of the standard errors of such type of estimates has been investigated by Tutz and Leitenstorfer [10] for binary responses.

References

[1] Ananth, C. V., Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology*, 26, Number 6, 1323–1333.

[2] Friedman, J., Hastie, T., Tibshirani, R. (2008). Regularization paths for generalized linear models via coordinate descent.

[3] Krishnapuram, B., Carin, L., Figueiredo, M. A., Hartemink, A. J. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 957–968.

[4] McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society*, B 42, 109–142.

[5] Nyquist, H. (1991). Restricted estimation of generalized linear models. *Journal of Applied Statistics*, 40, 133–141.

[6] Rousseeuw, P., Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics and Data Analysis*, 43, 315–332.

[7] Schaefer, R. (1986). Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computation and Simulation*, 25, 75–91.

[8] Schaefer, R., Roi, L., Wolfe, R. (1984). A ridge logistic estimator. *Communications in Statistics: Theory and Methods*, 13, 99–113.

[9] Segerstedt, B. (1992). On ordinary ridge regression in generalized linear models. *Communications in Statistics: Theory and Methods*, 21, 2227–2246.

[10] Tutz, G., Leitenstorfer, F. (2006). Response shrinkage estimators in binary regression. *Computational Statistics & Data Analysis*, 50, Issue 10, 2878–2901.

[11] Zhu, J., Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5, 427–443.

Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume II,
The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010): 1109–1117.

MULTIPLE IMPUTATIONS FOR MULTIVARIATE NORMAL BIO-DATA

Lu Zou
Hicks Building, Sheffield University, Sheffield S3 7RF
E-mail: heron_20012003@hotmail.com

ABSTRACT

Missing value is a common issue in clinical trials, and they may cause difficulties of analyses when they cannot be simply deleted. This study introduces the modified multiple imputations using the classic EM algorithm with the Bootstrapping re-sampling, which considers the correlations between the variables, and allows investigating the uncertainty of the imputations. Study focuses on the multivariate normal distributed case, and the application is carried out on a real Bio-data. The current study starts with a review and the comparisons between two widely used Multiple Imputations methods: the Multiple Imputation using Additive Regression, bootstrapping and Predictive Mean Matching (PMM) and the EM algorithm. Instead of converging the parameters of the data distribution, the modification is made by converging the missing values using EM algorithm. The combination with Bootstrapping re-sampling enables the replication of the imputation so that the uncertainty of imputations can be investigated. The application on a real Bio-data identified that the modified multiple imputations have better control on the estimates of missing values.

Keywords: Multiple Imputations; Bootstrapping; Predictive Mean Matching; EM algorithm.

1. INTRODUCTION

Missing value is a common phenomenon in clinical trials. It may be caused by dropping out of patients, technique limits, none response of survey and so on. If the amount of missing values is minor, they can be removed. However, if ignoring the missing values causes the loss of valuable information, they need to be filled. The easy and straight way is to use the mean of a variable to estimate all the missing values in that variable. But this method does not allow investigating the uncertainty of the estimation, and also ignores the correlation between variables. To solve this, Rubin (1987) introduced the Multiple Imputations (MI), as well as how to use MI to obtain valid inferences of imputed data. Multiple Imputations estimate the same missing value sufficient times, say k times. Based on the k estimates, the imputing variation can be studied using the standard deviation, confidence intervals, etc. Two multiple imputations methods for continuous variables are studied in current paper: Expectation Maximization (EM) algorithm (Schafer, J.L., 1997) and PMM (Multiple Imputation using Additive Regression, Bootstrapping and Predictive Mean Matching). The conventional EM algorithm is modified to apply the Multiple Imputations in this study by combining the Bootstrapping re-sampling.

The data used here was collected by an experiment aiming to compare two paired treatments, and the patients were followed up at several time points. A set of biomarkers were recorded with non-ignorable amount of missing values randomly located. Thus the imputation needs to be done properly. Because biomarkers are believed to be related, a multivariate normal distribution is

assumed on the logarithm of biomarkers. A brief introduction of data will be given in section 2. Following this, section 3 introduces the Multiple Imputations using Additive Regression, Bootstrapping and Predictive Mean Matching (PMM). Section 4 and section 5 perform Multiple Imputation using EM algorithm, the existing R function *em.norm* and EM imputation with bootstrapping (*em.boots*) respectively.

2. DATA INTRODUCTION

The data documented the variables of experiment containing two paired treatments. Both of them were applied randomly on the right/left side of each subject. During the experiment, patients were examined at five chosen time points: Visit 2 when patients were enrolled and controlled for 21 days to reach an uniform condition, Visit 6 when two treatments began, and then Visit 7, 8 and 9 corresponding to 4, 11 and 18 days after treatments respectively. The time line of this clinical trial is described in Fig.1. The 16 biomarkers were recorded through the five time points, and 25 out of 368 records have missing values. Since this is a longitudinal data, removing one missing record also deletes the whole information on the corresponding patient. As a result, alternative approach, imputation, is required.

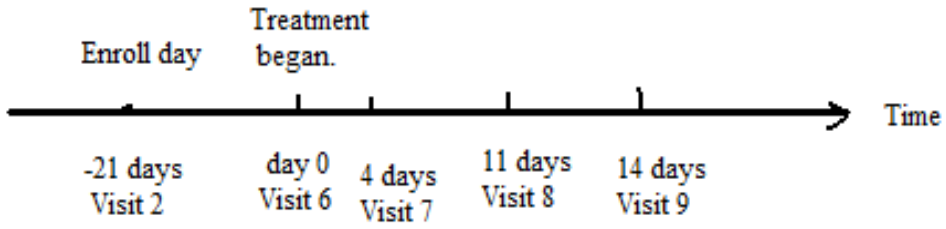


Figure 1: Time line of experiment with paired treatments.

The variables of the data are introduced in Table 1. Treatments and Side of mouth together indicated which treatment was applied on which side of patient, and they are completely recorded. Visit number indicated the time point that the patient has been examined. Both clinical score A and B are categorical variables with three levels: 0 for health patients and non-zero indicates the severity of disease, i.e. higher value means more serious. Another clinical score C has five levels and is also an ordinal variable. At the end, 16 biomarkers were measured and they are non-negative continuous.

A complete study is made from the original 16 biomarkers, ignoring other variables, and one single missing value is located randomly among them. This enables a comparison between estimated value and the original value later on. As a result, one of Bio1 value was chosen to be missing, and the true value is 20.45. In following sections, three imputing methods will be applied to estimate this missing value.

3. PMM IMPUTATION

PMM in current context stands for the Multiple Imputations using Additive Regression, Bootstrapping and Predictive Mean Matching. It can be performed by an existing R function

"aregImpute" in the library "Hmisc". It is widely used to deal with the missing values in medical research. Rouxel et al (2004) and Eijkemans et al (2008) both adopted the "aregImpute" function in Cox-based analyses to handle the missing information, with 10 iterations and a single iteration respectively. Koopman et al (2008) applied "aregImpute" function in different ways to fill the missing values in Individual Patient Data Meta-analyses.

Table 1: Structure of dataset '16Bio'

Variable name	***Explanation***	***Missing values***
Subject ID number	38 subjects in total	0
Gender	Female or Male	0
Treatments	L77=normal product G70=product with active	0
Side of mouth	Left or Right	0
Visit number	2 for 21 days before the treatments 6 for the day treatments start 7 for 4 days after treatments 8 for 11days after treatments 9 for 18 days after treatments	2
Clinical score A	Three levels indicated by 0, 1 and 2	35
Clinical score B	Three levels indicated by 0, 1 and 2	153
Clinical score C	Five levels indicated by 1, 2, 3, 4, and 5	83
Bio1-Bio16	16 biomarkers	25

Basically, the function "aregImpute" applies an additive regression with bootstrapping re-sampling and the predictive mean. It uses the Predictive mean matching, which avoids computing residuals or constraining imputed values to be in the range of observed values. Predictive mean matching imputes each missing value of the target variable with the observed value whose predicted transformed value is closest to the predicted transformed value of the missing value. Instead of using the estimated value by the regression model, it uses the observed value from dataset which has the closest predicted value to that of missing value.

The process of PMM method can be summarized in three steps: **1)** complete the data, fill the missing values with initial values randomly sampled from the non-missing part; **2)** for each variable carrying missing values, a random sample is drawn with replacement from the complete new data in step one (Bootstrapping). And a flexible additive model is fitted; the missing values are estimated based on this model with the predictive mean matching; **3)** the current completed target variable is used as a predictor of other missing variables. Step 2 and 3 can be repeated n times as required.

Apply 100 iterations using PMM method on the data selected from the original Bio-data in Section 2, so the single missing value of Biomarker one is estimated 100 times. The results are displayed in Fig.2: black circles on the graph are the 100 estimates of this missing value, and the average of 100 estimates is 24.67 while the true value is 20.45.

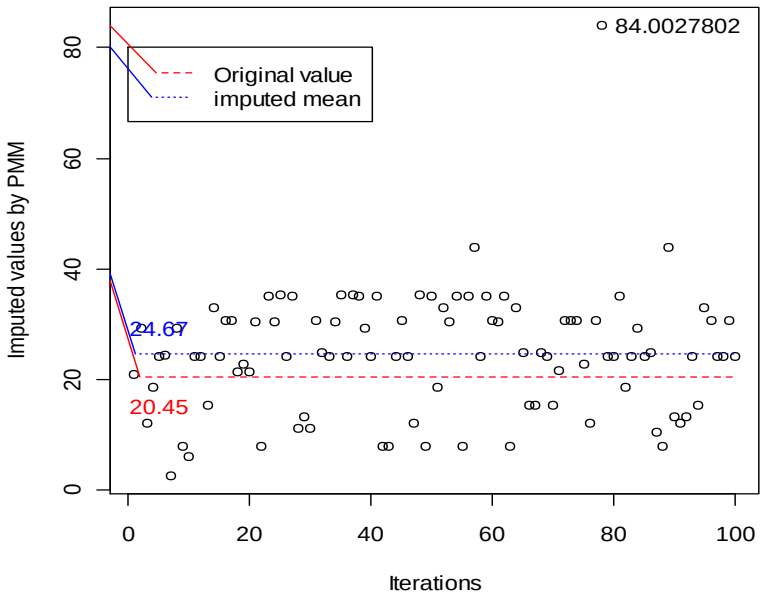


Figure 2: 100 imputations by aregImpute {Himcs}

Because that the predictive mean matching takes the observation with the best-fitted estimate as estimates, all the points on Fig.2 are actually observations. To explain this clearer, Fig.3 plots all the observations in Bio1, and the chosen 'missing' value is marked in black bold. The red triangles points out the observations used at least once as an estimate, and three of them with the highest frequencies are printed in red bold. They are the subject 51, 54 and 58 with frequencies 10/100, 9/100, and 20/100, respectively. Now one question comes up that as increasing the number of iterations, will all the observations be used as estimates? So larger number of iterations was tried in the study, and Fig.4 shows the attempt with 100000 iterations. It is found that more observations have been used as estimates, but the three highest frequency observations are the same: Subject 51 has been used 7525 times, Subject 54 has been used 9972 times and Subject 58 has been used 14829 times. Subject 12 has also been used 7331 times. However, Subject 55 which is further away from others has been chosen once. This pulls the average value of estimates higher than it is supposed to be. Therefore, rather than using the arithmetic mean, the weighted mean is more suitable in this case; the weights are the frequencies the observation has been chosen as estimate. The weighted mean here is 25.48, while the arithmetic mean is 32.07.

4. EM IMPUTATION

This section introduces another imputation method, EM algorithm, which is widely used in various areas of researches. The earliest EM algorithm was introduced by Little and Rubin (1987) and normally applied as an imputation in the following steps: 1). Replace missing values by estimated values (initial values); 2). Estimate the parameters by maximizing the likelihood function; 3). Re-estimate the missing values under these new parameters, and replace the missing cells with these new estimates. 4). Repeat step 2 and 3 until converged.

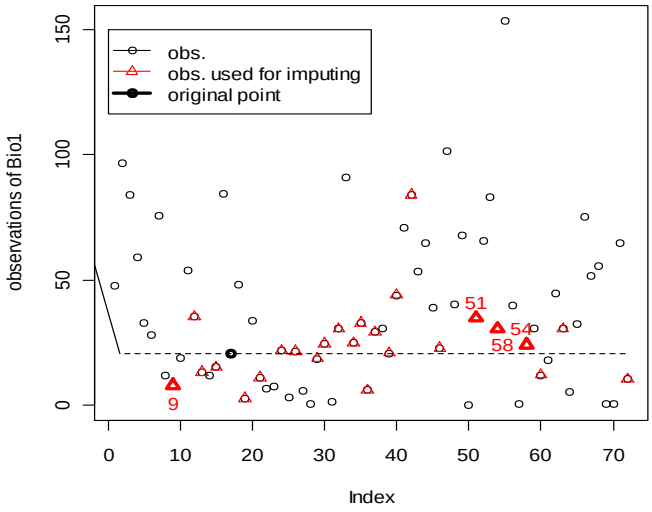


Figure 3: observations used as estimates in 100 iterations.

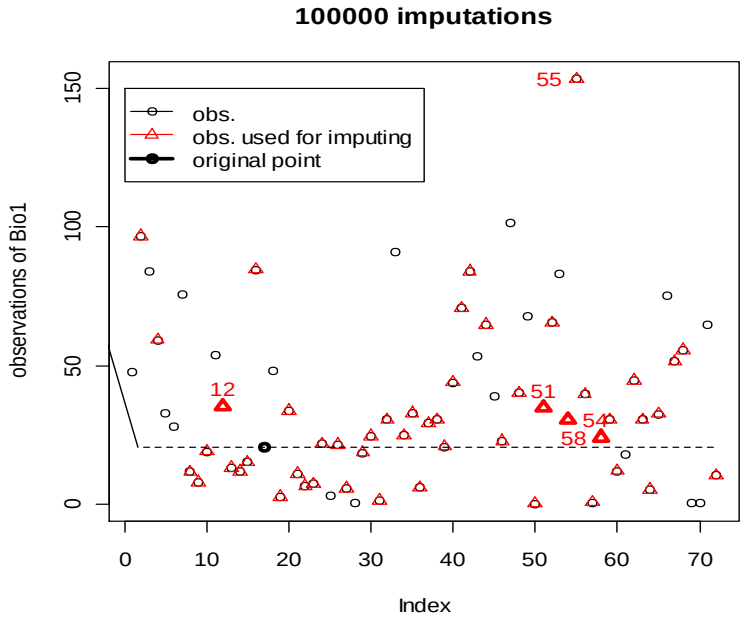


Figure 4: observations used as estimates in 100000 iterations.

This process can be implemented by an existing R function *em.norm{norm}*. It simulates an estimate by a random drown from its predictive distribution given the observations and the converged parameters of the data distribution. The replication of imputations can be realized by repeating the random drowns.

The same missing value is imputed by this EM method 100 times, and the estimates are revealed on Fig.5. The grey points are the observations and the black circles are the 100 imputed values. Taking the average of 100 estimates gives 15.43, highlighted by the red dashed line, while the true value is 20.45.

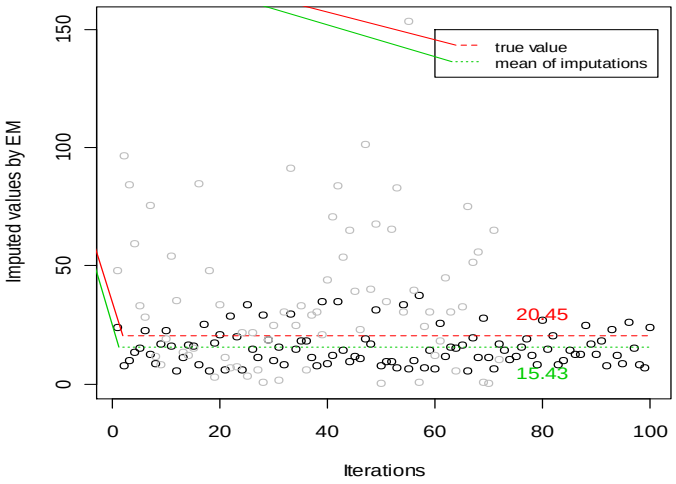


Figure 5: 100 imputations by em.norm{norm}.

Based Little and Rubin's theorem (1987), a converged estimate for missing value is expected, i.e. the conditional mean of the predictive distribution given the latest estimated parameters and the observations. However, the function *em.norm{norm}* practices the multiple imputations by random estimates from the corresponding predictive distribution. It allows the consideration of uncertainty, but adds another source of uncertainty by random sampling.

5. EM IMPUTAION BY SELF-WRITTEN FUNCTION

To achieve a desirable control on the estimates of missing values, instead of using the existing R function, My function applies the EM imputation, specific for the Bio-data. This assumes multivariate normal distribution on the logarithms of 16 biomarkers. The convergence focuses on the estimate of missing value rather than on the parameters of data distribution.

Precisely, suppose there are three variables following a jointly multivariate normal distribution $X = (X_1 \quad X_2 \quad X_3)' \sim N(\mu, \Sigma)$ with mean $\mu = (\mu_1 \quad \mu_2 \quad \mu_3)^T$ and covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix}$.

The probability function can be written as

$$f_X(X) = \frac{1}{(2\pi)^{\frac{3}{2}} |\Sigma|^{\frac{1}{2}}} \cdot \exp\{-\frac{1}{2}(X-\mu)^T \Sigma^{-1} (X-\mu)\}$$

Considering one missing value case, suppose the k_{th} value is missing in X_1, and then the data is arranged as $\begin{pmatrix} * & x_{2k} & x_{3k} \\ X_{1obs} & X_{2obs} & X_{3obs} \end{pmatrix}$. Given x_{2k} and x_{3k}, we want to estimate the missing value x_{1k} using the conditional distribution of x_{1k} given x_{2k} and x_{3k}

$$f(x_{1k} \mid x_{2k}, x_{3k}) = \frac{f_{x_1,x_2,x_3}(x_1, x_2, x_3)}{f_{x_2,x_3}(x_2, x_3)}$$

where $f_{x_2,x_3}(x_2, x_3)$ follows a bi-normal distribution. So the estimate of x_{1k} uses its conditional mean given x_{2k} and x_{3k}, i.e.

$$\hat{x}_{1k} = E(x_1 \mid x_2, x_3) = \mu_1 + \Sigma_{12}\Sigma_{23}^{-1}\left[\begin{pmatrix} x_2 \\ x_3 \end{pmatrix} - \begin{pmatrix} \mu_2 \\ \mu_3 \end{pmatrix}\right],$$

where $\Sigma_{12} = (\sigma_{12} \quad \sigma_{13})$ and $\Sigma_{23} = \begin{pmatrix} \sigma_2^2 & \sigma_{23} \\ \sigma_{32} & \sigma_3^2 \end{pmatrix}$.

The imputing process converges the conditional mean of each missing value. Firstly, it fills x_{1k} with the mean of X_1 ignoring x_{1k}; estimate $\hat{x}_{1k}^{(t)}$ as $E(x_{1k} \mid x_{2k}, x_{3k})$; update x_{1k} by $\hat{x}_{1k}^{(t)}$, and repeat the imputation, calculate $\hat{x}_{1k}^{(t+1)}$ based on $\hat{x}_{1k}^{(t)}$ until reached the convergence.

To explain how this function works, suppose we have three biomarkers following a multivariate normal distribution and this data has three random missing values, see below. To complete the data, initial values are given to fill the missing values; for *a*, use the conditional mean given $x_{27} = -0.168 \quad x_{37} = 6.550$; for *b*, use $E(x_{22} \mid x_{12}, \bar{X}_{3obs})$, where $\bar{X}_{3obs}$ is the mean of variable 'b3' ignoring the missing value; for *c*, use the conditional mean given

$$x_{12} = 3.566 \quad x_{22} = E(x_{22} \mid x_{12}, \bar{X}_{3obs}).$$

	b1	b2	b3
[1,]	3.987529	1.44348489	6.775554
[2,]	3.565562	NA(b)	NA(c)
[3,]	2.586146	0.18357229	5.245294
[4,]	2.479586	-0.78245757	4.702854
[5,]	2.727846	0.06172303	5.924735
[6,]	4.437846	1.71982918	6.559854
[7,]	NA(a)	-0.16840910	6.549635
[8,]	3.874542	1.74304080	7.433668
[9,]	1.002142	-0.49336389	4.907529
[10,]	3.518252	0.67625807	6.651102

In the case when there are two missing values (2nd row), $\hat{b}^{(t+1)}$ is based on $\hat{c}^{(t)}$, and then $\hat{c}^{(t+1)}$ is based on $\hat{b}^{(t+1)}$. Additionally, the convergence is reached when the maximum difference between two consecutive iterations,

$$\begin{pmatrix} \hat{a}^{(t+1)} - \hat{a}^{(t)} \\ \hat{b}^{(t+1)} - \hat{b}^{(t)} \\ \hat{c}^{(t+1)} - \hat{c}^{(t)} \end{pmatrix},$$

is smaller than given value C. Table 2 compares the results using $em.norm\{norm\}$ and using the proposed function. Both functions ran 100 iterations. The $em.norm\{norm\}$ randomly drew 100 estimates from the conditional predictive distribution, and the mean of 100 iterations was taken; the proposed function converged the estimates at 75^{th} iteration with a criterion of 1×10^{-50}. This criterion is much stricter compared with that of $em.norm\{norm\}$ 1×10^{-4}. If loose this value to 1×10^{-6}, the convergence will be achieved at the iteration 27. As you can see from Tab.2, the results by $em.norm\{norm\}$ are close to those by the proposed function.

Table 2 Comparison between my function and 'em.norm'

Methods	Estimates
em.norm(10000 draws)	2.558 0.913 6.457
em.zl(critzl=1E-50, converged.it=75)	2.525 0.930 6.463

This process can be repeated by the re-sampling technique. İn current study, the Bootstrapping is used. Bootstrapping is used on the completed cases to make a new complete data and then the same missing value is imputed based on this new data. Doing this we achieve the multiple imputations (MI).

Use the same data in Section 3 and 4, where a single missing value was randomly located in Bio1. Our proposed function (*em.boots*) ran with different numbers of iterations. Table 3 compares the results with those by the R function *em.norm*. The first noticeable thing is that the standard errors of estimates by *em.boots* are much smaller than those by *em.norm*. Another advantage of *em.boots* is that the estimate is converged with less interations: the estimate is 14.35 (s.e.=0.018) with 10000 iterations, while the estimate does not show evidence of convergence by *em.norm* with 10000 iterations.

Table 3: comparison between em.boots and em.norm

Number of iterations	Estimates (s.e.)
100	em.norm: 15.47 (0.82) em.boots: 14.54 (0.17)
1000	em.norm: 15.75 (0.264) em.boots: 14.38 (0.056)
100000	em.norm: 15.83 (0.027) em.boots: 14.35 (0.006)

6. CONCLUSION

This paper studies three methods to impute the missing values in the case where the data is assumed to be multivariate normal distributed. Two popular methods are reviewed and investigated: Multiple Imputations using Additive Regression, Bootstrapping and Predictive Mean Matching (PMM) and the Expectation Maximization (EM) algorithm. The former method

estimates the missing values by selecting the observations based on the additive regression. This may bias the imputation by picking the extreme values or outliers. The weighted mean is suggested to use with PMM instead of the arithmetic mean. The EM algorithm applied by *em.norm* is modified to converge the estimate of missing value rather than to converge the parameters of distribution. In addition, our proposed EM imputation together with Bootstrapping provides the multiple imputations (MI). The application on the Bio-data showed a better control of estimation and a quicker convergence.

REFERENCES

Eijkemans, M. J. C., Lintsen, A. M. E., Hunault, C. C., Bouwmans, C. A. M., Hakkaart, L., Braat, D. D. M. and Habbema, J. D. F., 2008. Pregnancy chances on an IVF/ICSI Waiting List: A Natinal Prospective Cohort Study. *Human Reproduction*, **23**: 1627.

Koopman, L., Van der Heijden, Geert, J. M. G., Grobbee, E. D., Rovers, M. M., 2008. Comparison of Methods of Handling Missing Data in Individual Patient Data Meta-analyses: An Empirical Example on Antibiotics in Children with Acute Otitis Media. *American Journal of Epidemiology*, **167**: 540.

Little, J. A. R. and Rubin, B. D. (1987). *Statistical Analysis with Missing Data.* John Wiley & Sons.

Rouxel, A., Hejblum, G., Bernier, M. O., Bo*ë*lle, P. Y., Menegaux, F., Mansour, G., Hoang, C., Aurengo, A. and Leenhardt, L. (2004). Prognostic Factors Associated with the Survival of Patients Developing Loco-Regional Recurrences of Differentiated Thyroid Carcinomas. *The Journal of Clinical Endocrinology & Metabolism*, **89**: 5362.

Rubin, D. B. (1987). *Multiple Imputations for Nonresponse in Surveys*. J. Wiley & Sons, New York.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

List of Contributors

Ali Abdallah 998
Faculty of Commerce, Assiut University
ali_statistics@yahoo.com

Eman A. Abd El-Aziz 349
Department of Statistics
Faculty of Commerce,
Al-Azhar University (Girls' Branch)
be123en@hotmail.com

Samar M. M. Abdelmageed 323
Statistical Researcher, Egyptian Cabinet's Information and Decision Support Center

Amina I. Abo-Hussien 349
Department of Statistics
Faculty of Commerce,
Al-Azhar University (Girls' Branch)
aeabohussien@yahoo.com

Moertiningsih Adioetomo 70
Center for Ageing Studies Universitas Indonesia
Demographic Institute Universitas Indonesia
toening@Idfeui.org

Munir Ahmad 317, 357,646, 664, 861
National College of Business Administration and Economics,
Lahore, Pakistan
drmunir@brain.net.pk

Raja Halipah Raja Ahmad 989
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA (UiTM),
40450 Shah Alam, Malaysia.
Raja.halipah@yahoo.com

Rashid Ahmed 367
Department of Statistics,
The Islamia University of Bahawalpur, Pakistan
rashid701@hotmail.com

Sheikh Bilal Ahmad 693
Department of Statistics
Amar Singh College, Srinagar,
Kashmir, India
sbilal_sbilal@yahoo.com

Zahoor Ahmad 357
University of Gujrat, Gujrat.
zahoor_ahmed_stat@yahoo.com

Munir Akhtar 367
COMSATS Institute of Information Technology, Attock, Pakistan
munir_stat@yahoo.com,
dir-attock@comsats.edu.pk,

Sibel Al 375
Hacettepe University, Department of Statistics, Ankara, Turkey
sibelal@hacettepe.edu.tr

Cagdas Hakan Aladag 384
Hacettepe University, Department of Statistics, Ankara, Turkey
aladag@hacettepe.edu.tr

Ibrahim M. Abdalla Al-Faki 392
College of Business and Economics
United Arab Emirates University
Al-Ain, UAE, P. O. Box 1755
i.abdalla@uaeu.ac.ae

Zalila Ali 483
School of Mathematical Sciences,
Universiti Sains Malaysia

M. A. Al-Jebrini 401
Department of Statistics, Yarmouk University, Irbid, Jordan
mjebrini@hotmail.com

T. S. Al-Malki 1050
Department of Statistics and Operations Research, College of Science,
King Saud University,
P.O.Box 2455, Riyadh 11451, Saudi Arabia

Hafez Al Mirazi 4
Kamal Adham Center for Journalism Training and Research, The American University in Cairo, P.O.Box 74, New Cairo 11835, Egypt
mirazi@aucegypt.edu

M. T. Alodat 401, 759
Department of Statistics,
Yarmouk University, Irbid, Jordan
malodat@yu.edu.jo, alodatmts@yahoo.com

Hessah Faihan AlQahtani 483
School of Mathematical Sciences,
Universiti Sains Malaysia

M. Y. Al-Rawwash 401
Department of Statistics,
Yarmouk University,
Irbid, Jordan
rawwash@yu.edu.jo

S. A. Al-Subh 759
School of Mathematical Sciences,
Universiti Kebangsaan Malaysia,
Selangor, Malaysia
salsubh@yahoo.com

Emad-Eldin A. A. Aly 80
Department of Statistics & Operations Research, Faculty of Science
Kuwait University, P.O. Box 5969, Safat 13060, Kuwait
emadeldinaly@yahoo.com

Raed Alzghool 409
Department of Applied Science
Faculty of Prince Abdullah Ben Ghazi for Science and Information Technology
Al-Balqa' Applied University,
Al-Salt, Jordan
raedalzghool@bau.edu.jo

Ayman A. Amin 771
Statistics & Insurance Department,
Menoufia University,
Menoufia, Egypt, and
Information and Decision Support Center,
The Egyptian Cabinet, Cairo, Egypt
aymanamin2008@gmail.com

Zeinab Amin 424
Department of Mathematics and Actuarial Science, The American University in Cairo, Egypt, and Faculty of Economics and Political Science,
Cairo University, Egypt.
zeinabha@aucegypt.edu

Lisa Anderson 6
The American University in Cairo,
P.O.Box 74, New Cairo 11835, Egypt

Jayanthi Arasan 451
Department of Mathematics,
Faculty of Science, Universiti Putra Malaysia,
jayanthi@math.upm.edu.my

V.N. Arief 1
The University of Queensland,
School of Land, Crop and Food Sciences,
Brisbane 4072, Australia

Abdu M. A. Atta 452, 817
School of Mathematical Sciences,
Universiti Sains Malaysia, Penang, Malaysia
abduatta@yahoo.com

Natal Ayiga 7
Department of Population Studies

University of Botswana
Natal.Ayiga@mopipi.ub.bw

Afzalina.Azmee 255
Department of Probability & Statistics
University of Sheffield
Afzalina.Azmee@sheffield.ac.uk

Ahmed Badr 461
Economic Issues Program (EIP),
Information and Decision Support Center
Egyptian Cabinet.
amabadr@idsc.net.eg

Adam Baharum 483
School of Mathematical Sciences,
Universiti Sains Malaysia
adam@cs.usm.my

K.E. Basford 1
The University of Queensland, School of Land, Crop and Food Sciences,
Australian Centre for Plant Functional Genomics, Brisbane 4072, Australia
k.e.basford@uq.edu.au

Jan Beirlant 2
University Center of Statistics,
Katholieke Universiteit Leuven
Goedele Dierckx, HUBrussel
jan.beirlant@wis.kuleuven.be

Abdelhafid Belarbi 805
Faculty of Economics and Administrative Sciences,
Al-Zaytoonah University of Jordan,
P.O.Box 130, Amman 11733, Jordan

Anil K. Bera 207
Department of Economics, University of Illinois, 1407 W. Gregory Drive,
Urbana, IL 61801
abera@illinois.edu

Jim Berger 2
Department of Statistical Science
Duke University, Durham,
NC 27708-0251, USA
Statistical and Applied Mathematical Sciences Institute , P.O. Box 14006,
Research Triangle Park,
Durham, NC 27709-4006
berger@samsi.info, berger@stat.duke.edu

Paul Bigala 7
Population Studies and Demography
North West University (Mafikeng Campus)
South Africa
paulgigs@yahoo.com

Atanu Biswas 205
Applied Statistics Unit, Indian Statistical Institute
203 B. T. Road, Kolkata 700 108, India
atanu@isical.ac.in

Chafik Bouhaddioui 494
Department of Statistics,
United Arab Emirates University,
Al Ain, UAE.
ChafikB@uaeu.ac.ae

Jennifer Bremer 4
Public Policy and Administration Department, School of Public Affairs,
The American University in Cairo,
P.O.Box 74, New Cairo 11835, Egypt
jbremer@aucegypt.edu

Matthew John Burstow 98, 106
Department of Surgery, Ipswich Hospital,
Queensland, Australia
mjburstow@gmail.com

Manisha Chakrabarty 246
Indian Institute of Management,
Calcutta, India

Asis Kumar Chattopadhyay 264
Department of Statistics, Calcutta University, India
akcstat@caluniv.ac.in

Tanuka Chattopadhyay 264, 266
Department of Applied Mathematics,
Calcutta University, 92 A.P.C. Road,
Calcutta 700009, India
tanuka@iucaa.ernet.in

Mohammad Ashraf Chaudhary 506
Mail Stop UG1C-60, Merck & Co., Inc.
351 North Sumneytown Pike
North Wales PA19454 USA
Mohammad_Chaudhary@Merck.Com

Sanjay Chaudhuri 157
Department of Statistics and Applied
probability, National University of
Singapore, Singapore 117546
stasc@nus.edu.sg

Ching-Shui Cheng 80
University of California at Berkeley
cheng@stat.berkeley.edu

Sooyoung Cheon 823
Department of Statistics, Duksung Women's
University,
Seoul 132-714, South Korea
KU Industry-Academy Cooperation Group
Team of Economics and Statistics,
Korea University, Jochiwon 339-700,
South Korea.
s7cheon@gmail.com

Hulya Cingi 375
Hacettepe University, Department of
Statistics, Ankara, Turkey
hcingi@hacettepe.edu.tr

J. Crossa 1
International Maize and Wheat
Improvement Center (CIMMYT),
APDo. Postal 6-641, 06600 México, D.F.,
Mexico

Danardono 950
Gadjah Mada University, Yogyakarta,
Indonesia
danardono@ugm.ac.id

Samarjit Das 246
Indian Statistical Institute
samarjit@isical.ac.in

G. S. Datta 159
University of Georgia

Emmanuel Davoust 264
Laboratoire d'Astrophysique de Toulouse-
Tarbes,
Universite de Toulouse,France
davoust@obs-mip.fr

I. H. Delacy 1
The University of Queensland, School of
Land, Crop and Food Sciences,
Australian Centre for Plant Functional
Genomics,
Brisbane 4072, Australia

Vita Priantina Dewi 70
Center for Ageing Studies Universitas
Indonesia
vitapriantinadewi@yahoo.com

M. J. Dieters 1
The University of Queensland, School of
Land, Crop and Food Sciences, Brisbane
4072, Australia

Jean-Marie Dufour 494
Department of Statistics,
United Arab Emirates University,
Al Ain, UAE
jean-marie.dufour@mcgill.ca

Riswan Efendi 831
Department of Mathematics,
Universiti Teknologi Malaysia, 81310
Skudai, Johor, Malaysia
wanchaniago@gmail.com

Erol Egrioglu 384
Ondokuz Mayis University, Department of Statistics, Samsun, Turkey
erole@omu.edu.tr

Samira Ehsani 451
Department of Mathematics,
Faculty of Science, Universiti Putra Malaysia
samira_p_ehsani@yahoo.com

Abdulhakeem Abdulhay Eideh 507
Department of Mathematics
Faculty of Science and Technology
Al-Quds University, Abu-Dies Campus
P.O. Box 20002, Jerusalem, Palestine
msabdul@science.alquds.edu

Elamin H. Elbasha 506
Mail Stop UG1C-60, Merck & Co., Inc.,
351 North Sumneytown Pike,
North Wales PA19454 USA
Elamin_Elbasha@Merck.Com

Wisame H. Elbouishi 572
Statistics Department, Faculty of Science,
El-Fateh University, Libya.

Fadlalla G. Elfadaly 537
Department of Mathematics and Statistics,
The Open University, Walton Hall,
Milton Keynes, MK7 6AA, UK
f.elfadaly@open.ac.uk

Ali El Hefnawy 652
Faculty of Economics and Political Science,
Cairo University, Cairo, Egypt
ahefnawy@aucegypt.edu

Abeer A. El-Helbawy 349
Department of Statistics
Faculty of Commerce,
Al-Azhar University (Girls' Branch)
a_elhelbawy@hotmail.com

Dina El Khawaja 4
Ford Foundation

Remah El-Sawee 708
Department of Statistics,
Faculty of Commerce,
Alexandria University, Egypt
remah-elsawee@hotmail.com

Mostafa Kamel El Sayed 5
Department of Political Science, Faculty of Economics and Political Science, Cairo University

Yousef M. Emhemmed 572
Statistics Department, Faculty of Science,
El-Fateh University, Libya.
emhemmedy@yahoo.co.uk

Stephen Everhart 6
School of Business, The American University in Cairo,
P.O.Box 74, New Cairo 11835, Egypt
severhart@aucegypt.edu

Nabil Fahmy 6
School of Public Affairs, The American University in Cairo,
P.O.Box 74, New Cairo 11835, Egypt
nfahmy@aucegypt.edu

ZA Siti Farra 17
Institute of Gerontology
Universiti Putra Malaysia, Malaysia

Nick Fieller 247, 255, 583
Department of Probability & Statistics
University of Sheffield
Sheffield, S3 7RH, U.K.
n.fieller@sheffield.ac.uk

Didier Fraix-Burnet 280
Université Joseph Fourier - Grenoble 1 / CNRS, Laboratoire d'Astrophysique de Grenoble (LAOG) UMR 5571
BP 53, F-38041 GRENOBLE

Cedex 09, France
fraix@obs.ujf-grenoble.fr

May Gadallah 584
Department of Statistics,
Faculty of Economics and Political Science,
Cairo University, Cairo, Egypt
mayabaza@hotmail.com

Hesham F. Gadelrab 607
Mansoura University, Faculty of Education,
Psychology Department
Mansoura, Egypt 35516
The British University in Egypt (BUE),
Business Administration Department
Sherouk City, Cairo,
Postal No. 11837, P.O. Box 43
heshfm@mans.edu.eg,
hesham.gadelrab@bue.edu.eg

Antonio F. Galvao Jr. 207
Department of Economics,
University of Wisconsin-Milwaukee,
Bolton Hall 852, 3210 N. Maryland Ave.,
Milwaukee, WI 53201
agalvao@uwm.edu

Paul H. Garthwaite 537
Department of Mathematics and Statistics,
The Open University, Walton Hall,
Milton Keynes, MK7 6AA, UK
p.h.garthwaite@open.ac.uk

Ronald Geskus 247
Department of Clinical Epidemiology,
Biostatistics and Bioinformatics
Academic Medical Center
Meibergdreef 15
1105 AZ, Amsterdam, The Netherlands
R.B.Geskus@amc.uva.nl
E.Hogervorst@lboro.ac.uk

Malay Ghosh 157, 159
Department of Statistics, University of
Florida, Gainesville, FL 32611
ghoshm2000@yahoo.com

Pulak Ghosh 119
Department of Quantitative sciences, Indian
Institute of Management,
Bangalore, India
pulakghosh@gmail.com

Suryo Guritno 1079
Mathematics Department,
Gadjah Mada University, Yogyakarta,
Indonesia
Guritno0@mailcity.com

Hyung-Tae Ha 634
Department of Applied Statistics,
Kyungwon University, Sungnam-ci,
Kyunggi-do South Korea, 461-701
htha@kyungwon.ac.kr

Saleha Naghmi Habibullah 646
Kinnaird College for Women, Lahore,
Pakistan
salehahabibullah@hotmail.com

Ali S. Hadi 6, 424
Department of Mathematics and Actuarial
Science, The American University in Cairo,
Egypt, and Department of Statistical
Science, Cornell University, USA.
ahadi@aucegypt.edu

Ramadan Hamed 652
Faculty of Economics and Political Science,
Cairo University, Cairo, Egypt
ramadanh@aucegypt.edu

Tengku-Aizan Hamid 17
Institute of Gerontology
Universiti Putra Malaysia, Malaysia
tengkuaizan06@gmail.com

Muhammad Hanif 357, 664, 676
Lahore University of Management Sciences,
Lahore, Pakistan
hanif@lums.edu.pk

Muna F. Hanoon 805
Faculty of Economics and Administrative Sciences,
Al-Zaytoonah University of Jordan,
P.O.Box 130, Amman 11733, Jordan

Toni Hartono 70
National Commission for Older Persons
Indonesia

Siti Rahayu Mohd. Hashim 685
Department of Probability and Statistics,
Hicks Building,
Hounsfield Road, S3 7RY, University of Sheffield, UK.
stp08sm@sheffield.ac.uk

Anwar Hassan 693
PG Department of Statistics
University of Kashmir, Srinagar-India
anwar.hassan5@gmail.com,
anwar.hassan2007@gmail.com

Siti Fatimah Hassan 305
Centre for Foundation Studies in Science,
Universiti of Malaya,
50603 Kuala Lumpur, Malaysia.

Christian Heumann 1094
Department of Statistics,
Ludwig Strasse 33, 80539.
Ludwig-Maximilians University Munich, Germany
christian.heumann@stat.uni-muenchen.de

Rafiq H. Hijazi 701
Department of Statistics
United Arab Emirates University
P. O. Box 17555, Al-Ain, UAE
rhijazi@uaeu.ac.ae

Eef Hogervorst 70
Department of Human Sciences
Loughborough University, UK

Zahirul Hoque 205, 206
Department of Statistics
College of Business and Economics
United Arab Emirates University
zahirul.hoque@uaeu.ac.ae

Md Belal Hossain 132, 140, 148
Department of Mathematics and Computing,
Australian Centre for Sustainable Catchments,
University of Southern Queensland,
Toowoomba, Queensland, Australia
hossainm@usq.edu.au

Shereen Hussein 25
Social Care Workforce Research Unit King's College London Melbourne House, 5th Floor Strand, London, UK, WC2R 2LS
shereen.hussein@kcl.ac.uk

Osama Abdelaziz Hussien 708, 733
Department of Statistics,
Faculty of Commerce,
Alexandria University, Egypt
osama52@gmail.com,
ossama.abdelaziz@alexcommerce.edu.eg

Abdul Ghapor Hussin 303, 305
Centre for Foundation Studies in Science,
Universiti of Malaya,
50603 Kuala Lumpur, Malaysia.
ghapor@um.edu.my

K. Ibrahim 759
School of Mathematical Sciences,
Universiti Kebangsaan Malaysia,
Selangor, Malaysia
Kamarulz@ukm.my

Noor Akma Ibrahim 747
Institute for Mathematical Research
Universiti Putra Malaysia
43400 UPM, Serdang, Selangor
Malaysia
nakma@putra.upm.edu.my

B. Ismail 786
Department of Statistics, Mangalore University,
Mangalagangothri, Mangalore-574199 India
ismailbn@yahoo.com

Mohamed A. Ismail 323, 771
Statistics Professor, Cairo University, and Consultant at Egyptian Cabinet's Information and Decision Support Center
m.ismail@idsc.net.eg

Mohd Tahir Ismail 796
School of Mathematical Sciences,
Universiti Sains Malaysia,
11800 USM, Penang
mtahir@cs.usm.my

Zuhaimy Ismail 831
Department of Mathematics,
Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
zuhaimy@utm.my, zhi@fs.utm.my

S. Rao Jammalamadaka 303
Department of Statistics and Applied Probability, University of California, Santa Barbara, CA. 93106 USA
rao@pstat.ucsb.edu

A. A. Jemain 759
School of Mathematical Sciences,
Universiti Kebangsaan Malaysia,
Selangor, Malaysia
kpsm@ukm.my

Bing-Yi Jing 81
Department of Math, HKUST,
Clear Water Bay, Kowloon, Hong Kong
majing@ust.hk

Zeinab Khadr 998
Faculty of Economics, Cairo University
zeinabk@aucegypt.edu

Faisal G. Khamis 805
Faculty of Economics and Administrative Sciences,
Al-Zaytoonah University of Jordan,
P.O. Box 130, Amman 11733, Jordan
faisal_alshamari@yahoo.com

Anjum Khan 786
Department of Statistics, Mangalore University, Mangalagangothri,
Mangalore-574199 India

Hafiz T. A. Khan 45
Business School, Middlesex University
London NW4 4BT, UK
h.khan@mdx.ac.uk

Shahjahan Khan 98, 106, 121, 132, 140, 148, 236, 317, 333
Department of Mathematics and Computing,
Australian Centre for Sustainable Catchments,
University of Southern Queensland,
Toowoomba, Queensland, Australia
khans@usq.edu.au

Michael B. C. Khoo 452, 817
School of Mathematical Sciences,
Universiti Sains Malaysia, Penang, Malaysia
mkbc@usm.my

Kaveh Kiani 451
Applied & Computational Statistics Laboratory,
Institute for Mathematical Research,
Universiti Putra Malaysia,
kamakish@yahoo.com

Jaehee Kim 823
Department of Statistics, Duksung Women's University, Seoul 132-714,
KU Industry-Academy Cooperation Group
Team of Economics and Statistics,
Korea University, Jochiwon 339-700,
South Korea
jaehee@duksung.ac.kr

Xinbing Kong 81
Department of Math, HKUST, Clear Water Bay, Kowloon, Hong Kong

K. Krishnamoorthy 119
University of Louisiana
Lafayette, LA, USA
krishna@louisiana.edu

P. M. Kroonenberg 1
Department of Education and Child Studies, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands
kroonenb@fsw.leidenuniv.nl

Parthasarathi Lahiri 158
JPSM, 1218 Lefrak Hall, University of Maryland, College Park, MD 20742, USA
plahiri@survey.umd.edu

Habibah Lateh 483
School of Mathematical Sciences,
Universiti Sains Malaysia
habibah@usm.my

Muhammad Hisyam Lee 831, 845
Department of Mathematics,
Universiti Teknologi Malaysia, Malaysia
mhl@utm.my

George W. Leeson 45
Oxford Institute of Ageing
University of Oxford
Oxford OX2 6PR, UK
george.leeson@ageing.ox.ac.uk

Huilin Li 158
Division of Cancer Epidemiology and Genetics, National Cancer Institute, USA
lih5@mail.nih.gov

S. K. Lim 452
School of Mathematical Sciences,
Universiti Sains Malaysia, Penang, Malaysia

Yan-Xia Lin 409
School of Mathematics and Applied Statistics
University of Wollongong
Wollongong, NSW 2500
Australia
yanxia@uow.edu.au

Zhi Liu 81
Department of Math, HKUST, Clear Water Bay, Kowloon, Hong Kong

Birgit Loch 333
Department of Mathematics and Computing
University of Southern Queensland
Toowoomba, Qld 4350, AUSTRALIA
Birgit.Loch@usq.edu.au

Suleman Aziz Lodhi 317, 861
National College of Business Administration & Economics,
Lahore, Pakistan.
sulemanlodhi@yahoo.com

Nadia Makary 5
Department of Statistics,
Faculty of Economics and Political Science,
Cairo University
nmakary@aucegypt.edu

Abdul Majid Makki 861
Abdul7896@yahoo.com.au

Saumen Mandal 205
Department of Statistics, University of Manitoba, Winnipeg, MB, R3T 2N2, Canada
saumen_mandal@umanitoba.ca

J. Maples 159
US Bureau of the Census

M. Maswadah 870
Department of Mathematics, Faculty of Science,
South Valley University, Aswan, Egypt
maswadah@hotmail.com

Thomas Mathew 120
Department of Mathematics and Statistics,
University of Maryland Baltimore County,
Baltimore, Maryland 21250, USA
mathew@umbc.edu

Christine McDonald 333
Department of Mathematics and Computing
University of Southern Queensland
Toowoomba, Qld 4350, AUSTRALIA
Christine.McDonald@usq.edu.au

Dean E. McLaughlin 265
Keele University, UK
dem@astro.keele.ac.uk

P. J. McLellan 1038
Department of Chemical Engineering
Queen's University, Kingston, Ontario,
Canada, K7L 3N6
mclellnj@chee.queensu.ca

Ahmed Zogo Memon 646
National College of Business Administration
& Economics
Lahore, Pakistan

Breda Memon 98, 106, 140, 148
Department of Surgery, Ipswich Hospital,
Queensland, Australia
bmemon@yahoo.com

Muhammed Ashraf Memon 98, 106, 121,
132, 140, 148
Department of Surgery, Ipswich Hospital,
Queensland, Australia
Department of Surgery, University of
Queensland, Herston, Queensland, Australia
Faculty of Medicine and Health Sciences,
Bond University, Gold Coast, Queensland,
Australia, Faculty of Health Science, Bolton
University, Bolton, Lancashire, UK
mmemon@yahoo.com

Sherzod M. Mirakhmedov 881
Ghulam Ishaq Khan Institute of Engineering
Sciences & Technology,
Topi-23460,Swabi ,NW.F.P. Pakistan
shmirakhmedov@yahoo.com

Saidbek S. Mirakhmedov 881
Institute of Algoritm and Engineering,
Fayzulla Hodjaev-45, Tashkent -700149.
Uzbekistan
saeed_0810@yahoo.com

Ibrahim Mohamed 303
Institute of Mathematical Sciences,
University of Malaya, 50603 Kuala Lumpur,
Malaysia
imohamed@um.edu.my

Saptarshi Mondal 264
Department of Statistics, Calcutta
University, India

Gabriel V. Montes-Rojas 207
Department of Economics, City University
of London, 10 Northampton Square, London
EC1V 0HB, U.K
Gabriel.Montes-Rohas.1@city.ac.uk

Ghada Mostafa 898
Central Agency For Public Mobilization and
Statistics
Salah Salem St. Nasr City
ghadaabd@yahoo.com

G. M. Nair 920
School of Mathematics and Statistics,
The University of Western Australia,
Perth, WA 6009,
Australia
gopal@maths.uwa.edu.au

Kenneth Nordstrom 120
Department of Mathematical Sciences,
University of Oulu, Finland

D. Nur 920
School of Mathematical and Physical Sciences,
The University of Newcastle
Callaghan, NSW 2308, AUSTRALIA
Darfiana.Nur@newcastle.edu.au

Y. Nurizan 17
Institute of Gerontology
Universiti Putra Malaysia, Malaysia

Teresa Azinheira Oliveira 82
CEAUL and DCeT, Universidade Aberta,
rua Fernão Lopes nº9, 2ºdto, 1000-132
Lisboa, Portugal
toliveir@univ-ab.pt
nurizan@putra.upm.edu.my

M. F. Omran 921
Business School, Nile University, Egypt
mfomran@nileuniversity.edu.eg

Emma Osland 121, 132
Dept of Nutrition and Dietetics, Ipswich Hospital, Ipswich, Queensland, Australia
Department of Mathematics and Computing, Australian Centre for Sustainable Catchments, University of Southern Queensland, Toowoomba, Queensland, Australia
Emma_Osland@health.qld.gov.au

Magued Osman 4
Chairman, Information and Decision Support Center
The Egyptian Cabinet
magued_osman@idsc.net.eg

G. N. Osuafor 931
Department of Statistics and Demography,
University of the Western Cape,
X17 7535 Bellville, South Africa
gnosuafor@gmail.com

Pinakpani Pal 206
Indian Statistical Institute, Calcutta, India
pinak@isical.ac.in

Sung Y. Park 207
Department of Economics, University of Illinois,
1407 W. Gregory Drive, Urbana, IL 61801,
and The Wang Yanan Institute for Studies in Economics,
Xiamen University, Xiamen, Fujian 361005, China
sungpark@sungpark.net

Ajoy Paul 246
Bidhan Nagar Govt. College

Koay Swee Peng 483
School of Mathematical Sciences,
Universiti Sains Malaysia

Danny Pfeffermann 186
Southampton Statistical Sciences Research Institute, University of Southampton, SO17 1BJ UK, Department of Statistics, Hebrew University of Jerusalem, 91905, Israel
msdanny@soton.ac.il

H. N. Phua 817
School of Mathematical Sciences,
Universiti Sains Malaysia, Penang, Malaysia

Hasih Pratiwi 950
Sebelas Maret University, Surakarta, Indonesia
Gadjah Mada University, Yogyakarta, Indonesia
hasihpratiwi@ymail.com

Tri Budi W. Rahardjo 70
Center for Ageing Studies Universitas Indonesia
Center for Health Research Universitas Indonesia

National Commission for Older Persons Indonesia
tri.budi.wr@gmail.com

Mohamed Ramadan 652
Population Council WANA Regional Office,
59 Misr Agricultural Road, Maadi, Cairo, Egypt
mramadan@popcouncil.org

Omar Rouan 341
GREDIM, Ecole Normale Supérieure Marrakech- Maroco
omarrouan@gmail.com

Umu Sa'adah 1079
Mathematics Department,
Gadjah Mada University, Yogyakarta, Indonesia,
Mathematics Department,
Brawijaya University, Malang, Indonesia
umusaadah@yahoo.com

Ishmael Kalule-Sabiti 7
North West University (Mafikeng Campus) South Africa
Ishmael.KaluleSabiti@nwu.ac.za

Asep Saefuddin 248, 1087
Department of Statistics, Faculty of Mathematics and Science,
IPB, 16680, Darmaga, Bogor, Indonesia
asaefuddin@gmail.com

Yasmin H. Said 963
Department of Computational and Data Sciences, George Mason University,
Fairfax, VA, USA
ysaid99@hotmail.com

Mohamed Saleh 461
Cairo University, Egypt
University of Bergen, Norway
saleh@salehsite.info

Mamadou-Youry Sall 977
Unit of Formation and Research in Economic Sciences and Management at Gaston Berger University, Saint-Louis, Senegal, BP 234
sallmy@ufr-seg.org

Mohd Sahar Sauian 989
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA (UiTM),
40450 Shah Alam, Malaysia
mshahar@tmsk.uitm.edu.my

Hussein Abdel-Aziz Sayed 998
Faculty of Economics, Cairo University
husseinsayed@hotmail.com

Kamal Samy Selim 1030
Department of Computational Social Sciences,
Faculty of Economics and Political Science
Cairo University, Cairo, Egypt
kselim9@yahoo.com

Muhammad Qaiser Shahbaz 676
Department of Mathematics,
COMSATS Institute of Information Technology,
Lahore, Pakistan
qshahbaz@gmail.com

Qi-Man Shao 80, 97
Department of Mathematics
Hong Kong University of Science and technology
Clear Water Bay, Kowloon
Hong Kong, China
maqmshao@ust.hk

M. E. Sharina 264, 293
Special Astrophysical Observatory, Nizhnij Arkhyz,
Zelenchukskiy region, Karachai-Cherkessian Republic, Russia 369167
sme@sao.ru

Furrukh Shehzad 367
National College of Business Administration & Economics, Lahore, Pakistan
fshehzad.stat@gmail.com

R. Steorts 159
US Bureau of the Census

N. Stiegler 931
Department of Statistics and Demography, University of the Western Cape,
X17 7535 Bellville, South Africa
nstiegler@uwc.ac.za

Andrew Stone 5
World Bank

John Stufken 97
Professor and Head, University of Georgia
jstufken@uga.edu

Subanar 950, 1079
Mathematics Department,
Gadjah Mada University, Yogyakarta, Indonesia
subanar@yahoo.com, subanar@ugm.ac.id

Subarkah 70
Center for Health Research Universitas Indonesia

Manjunath S Subramanya 140, 148
Department of Surgery, Mount Isa Base Hospital, Mount Isa, Queensland, Australia
manjunathbss9@yahoo.com

Etih Sudarnika 248
Laboratory of Epidemiology, Faculty of Veterinary Medicine, IPB, 16680, Darmaga, Bogor, Indonesia
etih23@yahoo.com

Suhartono 845, 1079
Perum ITS U-71, Jl. Teknik Komputer II, Keputih Sukolilo, Surabaya, Indonesia
suhartono@ statistika.its.ac.id

Suliadi 747
Dept. of Statistics, Bandung Islamic University
Jl. Tamansari No. 1 Bandung
Indonesia
suliadi@gmail.com

H. Sulieman 1038
Department of Mathematics and Statistics
American University of Sharjah,
P.O.Box 26666, Sharjah, U.A.E.
hsulieman@aus.ae

Khalaf S. Sultan 1050
Department of Statistics and Operations Research,
College of Science, King Saud Universit,
P.O.Box 2455, Riyadh 11451, Saudi Arabia
ksultan@ksu.edu.sa

Yusep Suparman 1070
Statistics Department, Padjadjaran University
Jl. Ir. H. Juanda no. 4, Bandung 40115
Indonesia
yusep.suparman@unpad.ac.id

Jef L. Teugels 2
Katholieke Universiteit Leuven & EURANDOM, Eindhoven
Jan Beirlant, University Center of Statistics, Katholieke Universiteit Leuven
Goedele Dierckx, HUBrussel
jef.teugels@wis.kuleuven.be
jan.beirlant@wis.kuleuven.be

Inam-Ul-Haq 664, 676
National College of Business Administration & Economics, Lahore, Pakistan
inam-ul-haq786@hotmail.com

J. A. M. Van der Weide 950
Delft University of Technology, Delft, The Netherlands
jamvanderweide@tudelft.nl

Edward J. Wegman 3, 963
Center for Computational Statistics
George Mason University
368 Research I, Ffx, MSN: 6A2
ewegman@gmu.edu

Yekti Widyaningsih 1087
Department of Statistics, Bogor Institute of Agriculture, Indonesia
yekti@ui.ac.id

Wafik Youssef Younan 1030
Department of Economics
The American University in Cairo
Cairo, Egypt
wyounan@aucegypt.edu

Yudarini 70
Center for Health Research Universitas Indonesia

Rossita Mohamad Yunus 98, 106. 121, 236
Department of Mathematics and Computing,
Australian Centre for Sustainable Catchments,
University of Southern Queensland,
Toowoomba, Queensland, Australia
Institute of Mathematical Sciences,
University of Malaya, Kuala Lumpur, Malaysia
Rossita.MuhamadYunus@usq.edu.au

Anis Y. Yusoff 5
Institute of Ethnic Studies (KITA),

National University of Malaysia
anis.yusoff@gmail.com

Faisal Maqbool Zahid 1094
Department of Statistics, Ludwig Strasse 33, 80539.
Ludwig-Maximilians University Munich, Germany
faisalmz99@yahoo.com

Enas Zakareya 461
Economic Issues Program (EIP),
Information and Decision Support Center (IDSC),
Egyptian Cabinet.
enabd@idsc.net.eg

Lu Zou 1109
Hicks Building,
Sheffield University,
Sheffield S3 7RF
heron_20012003@hotmail.com

Yong Zulina Zubairi 304, 305
Centre for Foundation Studies in Science,
University of Malaya, 50603 Kuala Lumpur, Malaysia
yzulina@um.edu.my

Organized by:

الجمعية الاحصائية لدول العالم الاسلامي
Islamic Countries Society of Statistical Sciences

The Egyptian Cabinet
Information and Decision Support Center
(IDSC)

Cosponsored by:

American University in Cairo

Islamic Development Bank
Together We Build a Better Future

ISBN: 978-977-416-365-8

The American University in Cairo
P.O.box 74, New Cairo 11835, Egypt
tel. 20.2.2615.2720
Fax 20.2.2795.7565
iccs-x@aucegypt.edu
www.iccs-x.org.eg/
www.isoss.com.pk/

The Islamic Countries Society of Statistical Sciences (ISOSS)
Address: Plot No. 44-A, Civic Centre, Liaqat Chowk, Sabzazar Scheme, Multan Road, Lahore, PAKISTAN.
Tel: +92 -42- 3784 0065, +92 -42- 3587 8583 Fax: +92 -42- 575 2547 E-Mail: secretary@isoss.com.pk