ISSN 1927-7032 (Print)
ISSN 1927-7040 (Online)

# International Journal of Statistics and Probability

Vol. 5, No. 6 November 2016

CANADIAN CENTER OF SCIENCE AND EDUCATION

# Contents

International Journal of Statistics and Probability; Vol. 5, No. 6; November 2016
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

# Time Series Chaos Detection and Assessment via Scale Dependent Lyapunov Exponent

Livio Fenga[1,2]

[1] ISTAT (National Institute of Statistics)

[2] UCSD (University of California San Diego)

Correspondence: Livio Fenga, ISTAT (National Institute of Statistics) Via Cesare Balbo, 16 - 00184, Rome, Italy. E-mail: fenga@istat.it

Received: August 2, 2016 Accepted: August 22, 2016 Online Published: October 17, 2016

doi:10.5539/ijsp.v5n6p1 URL: http://dx.doi.org/10.5539/ijsp.v5n6p1

**Abstract**

Many dynamical systems in a wide range of disciplines – such as engineering, economy and biology – exhibit complex behaviors generated by nonlinear components which might result in deterministic chaos. While in lab–controlled setups its detection and level estimation is in general a doable task, usually the same does not hold for many practical applications. This is because experimental conditions imply facts like low signal–to–noise ratios, small sample sizes and not–repeatability of the experiment, so that the performances of the tools commonly employed for chaos detection can be seriously affected. To tackle this problem, a combined approach based on wavelet and chaos theory is proposed. This is a procedure designed to provide the analyst with qualitative and quantitative information, hopefully conducive to a better understanding of the dynamical system the time series under investigation is generated from. The chaos detector considered is the well known Lyapunov Exponent. A real life application, using the Italian Electric Market price index, is employed to corroborate the validity of the proposed approach.

**Keywords:** deterministic chaos, economic time series, Lyapunov Exponent, multiresolution analysis

## 1. Introduction

Nature is per se' a non-linear entity, whose measurable expressions are non-linear as well. Real–world Data Generating Processes (DGPs) – except for trivial cases or lab–controlled experiments – hardly ever do possess features compatible with a linear framework. In particular, for time series produced by non linear dynamical systems, non linearity can unfold in many different ways, giving rise to various, mainly complex phenomena as a result, such as deterministic chaos, long memory, non stationarity, fractal and multi–fractal behavior. Deterministic chaos is among the most interesting features that can arise in such a framework. In particular, two topics have become crucial in several research fields – such as engineering, meteorology, cosmology and medicine – i.e. its discrimination from random noise and the estimation of its level. In fact, dealing with complicated structures – as in the case of weak RADAR signals (see for example Harman et al., 2006; Liu et al., 2007), noisy satellite images (Olsen et al., 2009) or complex biological data Freeman, 1992 and reference therein) – requires the analyst to have a picture as clear as possible on whether chaotic components are present in the dynamical system at hand as well as on their extent. However, it should be emphasized that, in its short life (it dates back to the early 70s), chaos theory has impacted a wide range of scientific activities, including the non physical ones, such as economics (see e.g. Peters, 1994; Brock et al., 1989; Puu, 2013), psychology (Guastello, 2013), sociology (Eve et al., 1997) and archaeology (Chadwick, 1998).

Deterministic chaos is characterized by sensitivity to initial conditions, which can be detected by measuring the rate of divergence between two trajectories starting from nearby states. Lyapunov Exponent (LE) – is one popular choice routinely employed to this end. This mathematical tool has been conceived and designed to perform in the field of physical science, where the full knowledge of the underlying stochastic process – in terms of both signal–to–noise ratio (SNR) and determining equations – guarantees the control of the system investigated. However, many times evidences of chaos in a given experimental system has to be evaluated on empirical basis, namely by direct observation of the empirical data. Regardless of the field of application, two typical features of such a framework are the lack of knowledge of the set of rules governing the system as well as of the quantity of noise present (and its exact nature). While the former makes system control procedures in general difficult to apply, the latter might have even more dangerous consequences. As an ubiquitous and unavoidable entity inherent with the experimental nature of the data, noise can have a catastrophic impact on virtually all the stages of chaos detection and assessment. Either in the form of environmental fluctuations or limited experimental resolution, noise can introduce significant amount of uncertainty in the system and chaos can go undetected as a result. Another serious problem is related to the fact that the estimation of the Lyapunov exponent, being performed on trajectories of finite length, can be severely biased due to their insufficient duration.

In the present paper, it is shown how chaos detection and assessment can be pursued in empirical setups, i.e. characterized by sensitivity to external noise and poor data resolution, through the estimation of the Lyapunov Exponent within the framework of multi–resolution approximation (MRA). The rest of the paper is organized as follows: in Section 1.1 the goals pursued are more precisely articulated and in Section 1.2 wavelet theory is introduced. Its particularization within Chaos theory is discussed in Section 2, where also the analytic tools employed for chaos detection are introduced. The employed variance decomposition technique is illustrated in Section 2.2. Finally, the empirical experiment is illustrated in Sections 3.

### *1.1 The Goal*

As already pointed out, the objective of the present paper is to undertake the analysis of a given time series to check for the presence of deterministic chaos and, if so, to provide an estimate – either quantitative and qualitative – of its degree. In more details, the proposed approach combines chaos and wavelet theories with the three-fold purposes of *a*) detecting deterministic chaos *b*) discriminating it from random noise and *c*) quantifying chaos at different time scales. By having a clearer picture of the embedded structures of the time series at hand, one is more likely to assess the predictive capabilities of a given forecasting tool or to conclude that a portion of it (either in time or frequency), or even the whole time series is unpredictable. In particular, wavelet decomposition can help the analyst choose the more predictable components and to design ad hoc forecasting procedures. Finally, a variance decomposition algorithm is employed as a proxy of the amount of chaos present at a given decomposition level. In more details, point a) is dealt with by repeating the LE estimation at each scale: by probing small length resolutions, chaos (and possibly other non-linear phenomena) detection as well as better understanding of its nature can be gained. Regarding the issue sub b), MRA low–pass filtering capabilities can be useful to discriminate pure random noise and chaotic components whereas point c) is pretty straightforward, as it consists in applying variance decomposition (W–ANOVA, i.e. Wavelet Analysis of Variance) procedures at each scale.

### *1.2 Signal Decomposing Procedure*

Dynamical decomposition of the observed time series is performed through a mathematical transform of the type MRA, which is induced by well localized functions: the wavelets, formalized in (1) and (2). *MRA* consists of a hierarchical sequence of nested subspaces $\left\{V_j\right\}_{j\in\mathbb{Z}}$ progressively approximating the Hilbert space $L^2(\mathbb{R})$ of all the squared integrable functions satisfying the following properties:

a) $\ldots V_{-1} \subset V_0 \subset V_1 \subset L^2(\mathbb{R})$ for all $j \in \mathbb{Z}$;

b) $\cup_{j=-\infty}^{+\infty} V_j$ is dense in $L^2(\mathbb{R})$ and $\cap_{j=-\infty}^{+\infty} V_j = \{0\}$;

c) $f(x) \in V_j \iff f(2x) \in V_{j+1}$, for all $j \in \mathbb{Z}$;

d) $f(x) \in V_j \to f(x-k) \in V_0 \ \forall k \in \mathbb{Z}$

e) $\exists\ \phi(x) \in V_0$, s.t. the set $\{\phi(x-k) | k \in \mathbb{Z}\}$ is a Riesz basis for $V_0$.

The scaling function $\phi$ (the father wavelet), constitutes the scale function for a given *MRA*. In other words, scaling by $2^j$ generates basis functions for the space $V_j$ so that, given that $(...V_{j-1} \subset V_j \subset V_{j+1}...)$, the scaling equation can be expressed as (1), under the condition of a proper choice of the coefficients $\{c_k; k \in \mathbb{Z}\}$; by linearly combining the scaled father wavelet with a suitably chosen set of the coefficients $\{b_k; k \in \mathbb{Z}\}$, i.e.

$$\langle \phi(x-k), \psi(x-l) \rangle = 0; \quad l, k \in \mathbb{Z},$$

the mother wavelet $\psi$ is obtained (2).

$$\phi(x) = \sum_{k\in\mathbb{Z}} c_k \phi(2x-k), \qquad (1) \qquad\qquad \psi(x) = \sum_{k\in\mathbb{Z}} b_k \phi(2x-k). \qquad (2)$$

By dilations and translations of $\psi(x)$, the space of functions $\Psi(x)$ is generated , i.e. $\Psi = \left\{\psi_{c,b}(x) = |c|^{-1/2} \psi \frac{x-b}{c}, c, b \in \mathbb{R}\text{ç} \neq 0\right\}$, with $\psi$ obeying to the following conditions: *a)* $\int_{-\infty}^{\infty} \psi(t)dt = 0$; *b)* $\int_{-\infty}^{\infty} |\psi(t)|dt < \infty$ ; *c)* $\int_{-\infty}^{\infty} \frac{|\hat{\psi}(\xi)|}{|\xi|} dt\xi < \infty$, $\hat{\psi}(\xi)$ being the Fourier transform of $\psi(t)$; *d)* $\int_{-\infty}^{\infty} t^j \psi(t)dt = 0, \quad j = 0, 1, ..., r-1$, under the conditions that it must exist at least a $r \geq 1$ and $\int_{-\infty}^{\infty} t^r \psi(t)dt < \infty$. By projecting the data onto shifted and translated transformations of the mother and father wavelets, the sets of wavelet and scaling coefficients are respectively obtained, i.e.

$$d_{j,k} = \int_{\mathbb{R}} \phi_{j,k} x(t) dt \qquad (3) \qquad\qquad s_{J,k} = \int_{\mathbb{R}} \psi_{J,k} x(t) dt. \qquad (4)$$

Here, each set of coefficients, usually referred to as crystal, is linked to a spacial scale $j$, whereas every single coefficient, called atom, accounts for a particular location. In practice, wavelet coefficients $\left\{d_{jk}; \quad j = 1, 2, ..., J\right\}$ in (3) account for and represent progressively finer and finer details whereas smooth dynamics at the coarsest scale are captured by the crystal $\{S_{Jk}\}$ (4). $MODWT$ is the filtering approach achieved used to perform MRA. Here the wavelet and scale coefficients are given by:

$$d_{j,t} = \frac{1}{2^{j/2}} \sum_{l=0}^{L_j-1} \tilde{h}_{j,l}, X_{t-l \mod N}, \qquad\qquad S_{J,t} = \frac{1}{2^{j/2}} \sum_{l=0}^{L_j-1} \tilde{g}_{j,l}, X_{t-l \mod N}, \qquad (5)$$

where $\left\{\tilde{h}_{j,l}\right\}$ and $\left\{\tilde{g}_{j,l}\right\}$ are the length $L$, level $j$, wavelet and scaling filters, obtained by rescaling their $DWT$ counterparts, i.e. $\left\{h_{j,l}\right\}$ and $\left\{g_{j,l}\right\}$, as follows: $\tilde{h}_{j,l} = \frac{h_{j,l}}{2^{j/2}}$ and $\tilde{g}_{j,l} = \frac{g_{j,l}}{2^{j/2}}$. Here, the sequences of coefficients $\left\{h_{j,l}\right\}$ and $\left\{g_{j,l}\right\}$ are approximate filters: the former, of the type band–pass, with nominal pass–band $f \in [\frac{1}{4\,\mu_j}, \frac{1}{2}\mu_j]$, and the latter of the type low-pass, with a nominal pass-band $f \in [0, \frac{1}{4}\mu_j]$, with $\mu_j$ denoting the scale. Considering all the $J = J^{max}$ sustainable scales, $MRA$ wavelet representation of the given time series $x_t$, in the $L^2(R)$ space, can be expressed as follows:

$$\begin{aligned} x(t) = \sum_k s_{J,k}\phi_{J,k}(t) + \sum_k d_{J,k}\psi_{J,k}(t) + \sum_k d_{J-1,k} + \\ + \psi_{J-1,k}(t) + ... + \sum_k d_{j,k}\psi_{j,k}(t) ... + \sum_k d_{1,k}\psi_{1,k}(t), \end{aligned} \qquad (6)$$

with $k$ taking integer values from 1 to the length of the vector of wavelet coefficients related to the component $j$.

### 1.2.1 Practicalities

In an empirical set up, boundary conditions can be selected on the basis of the characteristics of the time series and/or as a part of a preliminary ad hoc investigation on a guess and check basis. On the contrary, the choice of the wavelet function and its length $L$ is in general critical and strictly system specific. In what follows, the $4^{th}$ order Daubechies least asymmetric wavelet filter (known also as symmlets) of length $L = 8$, usually denoted by $LA(8)$, has been made use of. This type of filter, commonly adopted in several applications, has been chosen for its salient properties, that is: near symmetry in the midpoint and (approximately) linear phase. MRA has been performed using Maximum Overlapping Discrete Wavelet Transform (MODWT) algorithm. Technically, it is a filtering approach aimed at modifying the observed series $x_t$, by artificially introducing an extension of it, so that the unobserved samples $\{x\}_{t \in Z^-}$ are assigned the observed values $X_{T-1}, X_{T-2}, ..., X_0$. This method, considers the series as it were periodic[1] , and is known as using circular boundary conditions. For more details on MODWT, the reader is referred to Percival and Walden (2006) and Percival (2002).

## 2. The Wavelet Phase–space Reconstruction and the Scale Dependent Chaos Detector

The Lyapunov–based investigation is defined by considering an ensemble of trajectories, which requires a suitable reconstruction of the phase space of the underlying dynamical system, say $\boldsymbol{x_{n+1}} = \mathcal{F}(\boldsymbol{x_n})$. This is observed through a function $y = \mathcal{H}(\boldsymbol{x})$, whose elements are the scalar measurements $S(n)$ at a different time, i.e.

$$\boldsymbol{y_n} = \left[S(n), S(n+\tau), S(n+2\tau), \ldots, S(n+(m-1)\tau)\right], \qquad (7)$$

with $m$ being the embedding dimension and $\tau$ the time delay. Takens embedding theorem allows us to use the delay coordinates (7), in virtue of the two dynamical systems arising from this setup, i.e. $y_{n+k} = \mathcal{H}(\boldsymbol{x_{n+k}}) = \mathcal{H}(\mathcal{F}^k(\boldsymbol{x_n}))$ and $y_{n+1} = \mathcal{G}(\boldsymbol{y_n})$. Considering the latter, $\boldsymbol{x} \leftrightarrow \boldsymbol{y}$, one–to–one type relationship is guaranteed only for a "sufficiently" large $m$. Takens theorem particularizes to time dependent processes, the results obtained by Whitney in the case of dimension $D$ smooth manifold, say $\boldsymbol{M}$, whose dynamic can be embedded in $\mathbb{R}^{2D+1}$ and immersed in $\mathbb{R}^{2D}$. In essence, Takens

[1] However, it should be emphasized that, even though such an approach suffices in many cases, it shows weakness when non–periodic signals are affected by discontinuities, as in the case, for example, of certain deseasonalized economic indicators. To alleviate the problem, a common strategy is to introduce an artificial extension of the time series, by doubling up its original sample size through boundary conditions of the type reflection, i.e. the unobserved values $x_{-1}, x_{-2}, ..., x_{-T}$ are assigned the values observed at $x_0$, $x_1$,...$x_{T-1}$. Let $\{x_t^{\mathfrak{e}}\}$ the artificially extended time series generated according to

$$\begin{cases} \{X_t^{\mathfrak{e}}\} = \{X_t\}; \quad t = 0, ..., T-1 \\ \{X_t^{\mathfrak{e}}\} = \{X_{2T-1-t}\}; \quad t = T, ..., 2T-1, \end{cases}$$

the circular boundary conditions, are then re–expressed on $\{X_t^{\mathfrak{e}}\}$ to obtain new wavelet and scaling filters, respectively expressed by:

$$d_{j,t} = \frac{1}{2^{j/2}} \sum_{l=0}^{L-1} \tilde{h}_{j,l}, X^{\mathfrak{e}}_{t-l \mod 2N}, \qquad\qquad S_{j,t} = \frac{1}{2^{j/2}} \sum_{l=0}^{L-1} \tilde{g}_{j,l}, X^{\mathfrak{e}}_{t-l \mod 2N}.$$

proved that – given a diffeophorism $\mathfrak{M}$ defining the set of trajectories on $\boldsymbol{M}$, a smooth observation function $\underline{y} : \boldsymbol{M} \rightarrow \mathbb{R}$ generates an embedding of $\boldsymbol{M}$ in $2m + 1$ dimensions under the transformation $\widetilde{\mathfrak{M}}_{\mathfrak{M},y} : \boldsymbol{M} \rightarrow \mathbb{R}^{2m+1}$, where $\widetilde{\mathfrak{M}}_{\mathfrak{M},y}(\boldsymbol{x}) = < y(\boldsymbol{x}), y(\mathfrak{M}(\boldsymbol{x})), y(\mathfrak{M}^{\mathbf{2}}(\boldsymbol{x})), \ldots, y(\mathfrak{M}^{2m}(\boldsymbol{x})) >$. Here, each of the elements $< y(\boldsymbol{x}), y(\mathfrak{M}(\boldsymbol{x})), y(\mathfrak{M}^{\mathbf{2}}(\boldsymbol{x})), \ldots, y(\mathfrak{M}^{2m}(\boldsymbol{x})) >$ represents the time–shifted observations of the dynamics induced by $\mathfrak{M}$ on $\boldsymbol{M}$. An in-depth and detailed discussion of this subjected can be found in Deyle and Sugihara (2011) and Noakes (1991).

Relevant for the present analysis is the result from Sauer and Yorke (1993) on whether filtering procedure could affect proper embedding. Their central result is that by applying filters of the type Finite Impulse Response (FIR) embedding procedures, in general, would not be compromised, under the assumption that a sufficient number of independent observables is available. This setup is consistent with the adopted MRA approach – other than the assumption, previously stated, of time series of "sufficient" length. Regarding the first condition, a basis of the type Riesz can be transformed into an orthogonal basis through the transfer function $\left\{\tilde{g(\cdot)}\right\}$ (5), which entirely determines the scaling function $S_{J,t}$ (5). The g–induced scaling function $S_t$ is compactly supported if and only if $\left\{\tilde{g(\cdot)}\right\}$ has a finite number of non zero coefficients, that is $g(\cdot)$ is a FIR filter.

However, in order for the function $g(\cdot)$ to generate multiresolution approximations, some conditions need to be satisfied (see, for example Merry & Steinbuch, 2005). They are:

$\phi \in \boldsymbol{L}^{\mathbf{2}}(\mathbb{R})$ is an integrable scaling function *iff* the Fourier series

$$g[n] = \langle \frac{1}{\sqrt{2}} \phi(t/2), \phi(t-n) \rangle$$

satisfies

(i) $\forall \omega \in \mathbb{R}, |\hat{g}(\omega)|^2 + |\hat{g}(\omega + \pi)|^2 = 2;$

(ii) $\hat{g}(0) = \sqrt{2}.$

On the other hand, if

(a) $\hat{g}(\omega)$ is periodic (period = $2\pi$);

(b) $\hat{g}(\omega)$ of the type $\mathbf{C}^1$ in a neighborhood of $\omega = 0$;

(c) conditions (i, ii) above hold;

(d) $\inf_{\omega \in [-\frac{\pi}{2}; \frac{\pi}{2}]} |\hat{g}(\omega)| > 0,$

then $\hat{S}(\omega) = \prod_{p=1}^{+\infty} \frac{\hat{g}(2-p\omega)}{\sqrt{2}}$ is the Fourier transform of a scaling function $S \in \boldsymbol{L}^{\mathbf{2}}(\mathbb{R})$.

Finally, the application of Lyapunov exponent at different resolution levels finds its theoretical justification in the fact that (see e.g. Lamarque & Malasoma, 1996 ) given a scalar function $f : \mathbb{R} \rightarrow \mathbb{R}$, and a wavelet function $\phi(x)$ (1), the $f$–wavelet transformation $Q$ is given by

$$Q_{\phi,a,b}(f) = \frac{1}{a} \int_{-\infty}^{+\infty} f(t) \phi(\frac{x-b}{a}) dx, \tag{8}$$

with $b$ being the focal point of the mathematical microscope $\phi(\cdot)$ and $\frac{1}{a}$ the magnifying factor. The quantification of the separation ($\delta$) over time ($\Delta_t$) of couple of neighbor trajectories starting in $\delta_0 = ||\boldsymbol{x}_{t_1} - \boldsymbol{x}_{t_2}||$, i.e. $\delta_{\Delta_t} \approx ||\boldsymbol{x}_{t_1} + \boldsymbol{\Delta}_t - \boldsymbol{x}_{t_2} + \boldsymbol{\Delta}_t||$, giving rise to the Lyapunov exponent $\lambda$, i.e.

$$\delta_{\Delta_t} \approx \delta_0 e^{\lambda \Delta_t}, \tag{9}$$

is done with reference to the function $Q(\cdot)$.

Therefore, the Lyapunov exponent is now expressed as

$$Q(\cdot) = S(\Delta_t) = \frac{1}{N} \sum_{t_0=t_1}^{t_N} \ln \Big( \frac{1}{|U(\boldsymbol{x}_{t_0})|} \sum_{x_t \in U(x_{t0})} |\boldsymbol{x}_{t_0} + \boldsymbol{\Delta}_t - \boldsymbol{x}_t + \boldsymbol{\Delta}_t| \Big), \tag{10}$$

where $U(\boldsymbol{x}_{t_0})$ defines the neighborhood centered on $\boldsymbol{x}_{t0}$.

In the empirical analysis presented in Section 3, $x_0$ counts the data point falling within a radius $\epsilon = \frac{1}{10}\sqrt{\sigma^2}(Q(\cdot))$.

### *2.1 Embedding Dimension and Time Delay Estimation*

In what follows, the analytic tools employed for the optimal LE estimation are presented. They are designed to capture patterns in the probabilistic structure in the linear (autocorrelation function) and non linear (mutual information) case, whereas the estimation of the time delay is provided by the false nearest neighbor method.

#### 2.1.1 Autocorrelation and Mutual Information

The usual autocorrelation function (11–12) and the time delayed mutual information (13–14), as well as visual inspection of delay representations with various lags provide important information about reasonable delay times.

$$\rho(\tau) = \mathbb{E}\Big[\Big(X_t - \mu\Big)\Big(X_{t+\tau} - \mu\Big)\Big], \tag{11}$$

$$\hat{\rho}^j(\tau)\frac{1}{\Big(T - \tau\Big)\sigma^2_{X^j}}\sum_{t=1}^{T-\tau}\Big(x^j_t - \bar{x}^j\Big)\Big(x^j_{t-\tau} - \bar{x}^j\Big). \tag{12}$$

Here (12) – whose asymptotic confidence intervals are $\frac{1}{T} \pm 2\sqrt{T} \approx \pm 2\sqrt{T}$ – represents the estimator for (11) at a given resolution level $j$. However, more guidance has been gained by means of the mutual information function. This function is able to capture and account for linear and nonlinear correlations; in essence, it expresses the entropy–related concept of measuring the ”amount of information” obtained about one random variable, through the other random variable. It is given by:

$$\mathcal{I}^j(\mathcal{X}^j; \mathcal{Y}^j) = \int_{\mathcal{Y}^j}\int_{\mathcal{X}^j} p\Big(\mathbf{x^j}, \mathbf{y^j}\Big)\ \log_2 \frac{p(\mathbf{x^j}, \mathbf{y^j})}{p(\mathbf{x^j})p(\mathbf{y^j})}\, dy^j\, dx^j. \tag{13}$$

Its estimator, for the resolution level $j$, is given by

$$M^j = -\sum_{i,r} p_r(\tau)\log_2 p_{ir}(\tau)[(p_i \cdot p_r)]^{-1}. \tag{14}$$

#### 2.1.2 Time Delay Estimation

Determination of the proper time delay $d$ has been done with regard to nearest neighbor of each of the vectors (7) using the $L_2$ norm. This is a critical and in general a non trivial task: its underestimation is particularly dangerous, as two time delay vectors might show a small distance not as a result of the peculiar system dynamics but due to projection. The employed procedure, conducted at the $j$–level, declares a given neighbor of $\boldsymbol{y}_{j,k}$, denoted as $\boldsymbol{y}^\circ_{j,k}$, a false neighbor according to the rate of false nearest neighbors in the reconstructed phase space, i.e.:

$$\frac{\|\boldsymbol{y}_{j,k} - \boldsymbol{y}^\circ_{j,k}\|^2 + \Big[\boldsymbol{x}_{j,k+d} - \boldsymbol{x}^\circ_{j,k+d}\Big]^2}{\mathbf{B}^2(\mathbf{A})} > \mathbf{A}^2_{thr}.$$

Here, $\mathbf{A}_{thr} \approx 2$ (see e.g. Aittokallio et al., 1999) is the radius of the attractor whereas

$$\mathbf{B}^2(\mathbf{A}) = \frac{1}{N}\sum_{k=1}^{N}\Big[\boldsymbol{x}^j - \mu(\boldsymbol{x}^j)\Big]^2,$$

with $\mu(\cdot)$ denoting the mean value computed on all the points.

However, this procedure – designed to test progressively higher dimensionalies until a sufficiently small number is obtained (possibly 0) – is able to provide reliable outcomes in case of noise–free data. On the contrary, when noisy components are embedded in the time series, the likelihood of finding false nearest neighbors increase with the sample size. The use of wavelet theory, in the form of MRA algorithm, is a coping strategy for such a circumstance, as it is able to provide information at different levels which can be less affected by noise.

Spurious temporal correlation, which can seriously bias the estimation of the system embedding dimension, are dealt with by ruling out the set of points closer than some threshold time, i.e. the Theiler window. This quantity has been estimated using the method introduced by Provenzale et al. (1992), which is basically based on a sequence of space-time separation plots employed for the detection of temporal structures in the data. MODWT's translation-invariant property allows the effective alignment of the different events at different resolution levels so that the integrity of the dynamics induced by transient events is preserved.

Finally, being Lyapunov exponent invariant to the embedding dimension $m$, under $m > d$, $S(\Delta_t)$ has been calculated for $m = 2 \cdot (d), \ldots, m + 9$.

### *2.2 Wavelet ANOVA*

MODWT is a transformation of the type energy conserving, i.e. $||\boldsymbol{X}||^2 = \sum_{j=1}^{J_0} ||\widetilde{\boldsymbol{W}_j}||^2 + ||\widetilde{\boldsymbol{V}_{j_0}}||^2$, therefore a scale dependent analysis of variance can be derived, based on the set of wavelet and scaling coefficients, i.e.:

$$\hat{\sigma}_X^2 = ||\boldsymbol{X}||^2 - \bar{X}^2 = \frac{1}{N} \sum_{j=1}^{J_0} ||\widetilde{\boldsymbol{W}_j}||^2 + \frac{1}{N} ||\widetilde{\boldsymbol{V}_{j_0}}||^2 - \bar{X}^2. \tag{15}$$

Equation (15) enable us to quantify the allocation of the energy roughly attributable to chaotic components across the J scales

Ruling out the boundary coefficients sets, we have that the wavelet variance

$$\hat{\nu}_X^2(\tau_j) = \mathbb{E}(\widetilde{W}_{j,t})^2 = \frac{1}{N} \sum_{t=0}^{N-1} \widetilde{W}_{j,t}^2 \tag{16}$$

is time independent under second order stationarity or difference stationary assumptions. The former implies that the extracted sequence (signal) $\left\{\tau_{j,t}\right\}_{t \in \mathbb{Z}^+}$, that $\mathbb{E}(X) = \mu$ and $cov(\tau_{j,t}, \tau_{j,t+k}) = \gamma_{j,k}$, being $\mu$ constant and $\gamma_k$ time independent. Defining the back-shift operator $L$, i.e. $L\tau_t = \tau_{t-1}$ (therefore $L^n \tau = \tau_{t-n}$) and the difference operator (the subscript $j$ is omitted) $\nabla^d \tau_t = (1 - L)^d \tau_t \quad d = 0, 1, \ldots D$, the latter implies that the transformed time series $\nabla^d \tau_t$ = stationary. However, these conditions are usually met at coarser levels, where

$${}^{u}\hat{\nu}_X^2(\tau_j) = \mathbb{E}(\widetilde{W}_{j,t})^2 = \frac{1}{M_j} \sum_{t=L_j-1}^{N-1} \widetilde{W}_{j,t}^2, \tag{17}$$

with $M - j = N - L_j + 1$ being the size of the set of non–boundary coefficients for the $j$–level, $L_j$ the wavefilter length, and $u$ stands for unbiased. However, by using MODWT, one is forced to use either reflection, or circular boundary conditions. Therefore, the proper estimator is given by:

$${}^{bias}\hat{\nu}_X^2(\tau_j) = \mathbb{E}(\widetilde{W}_{j,t})^2 = \frac{1}{2N} \sum_{t=0}^{2N-1} \widetilde{W}_{j,t}^2. \tag{18}$$

Confidence intervals for the true wavelet variance can be built for both (17) and (18) on the basis of their asymptotic approximation to a scaled $\chi^2$ distribution with $\delta_j$ EDOF (Equivalent Degree Of Freedom), which reflects the correlation structure at different resolution levels $j$. Following Percival and Walden (2006), $100\%(1 - 2p)$ confidence intervals are approximated by

$$CI(\hat{\nu}_X^2(\tau_j)) \approx \Big[\frac{\delta_j \hat{\nu}_X^2(\tau_j)}{Q_{\delta_j}(1-p)}, \frac{\delta_j \hat{\nu}_X^2(\tau_j)}{Q_{\delta_j}(p)}\Big], \tag{19}$$

with

$$\delta_j = \max\left\{\frac{M_j}{2^j}, 1\right\} \qquad \text{or} \qquad \delta_j = \max\left\{\frac{N}{2^j}, 1\right\},$$

according to whether the case (17) or (18) is considered.

## 3. Empirical Experiment

In order to corroborate the validity of the presented approach, an empirical analysis has been conducted on the electricity market. Commonly denominated Italian Power Exchange (IPEX), the Electricity Market, is where producers, consumers and wholesale customers enter into hourly electricity purchase and sale contracts. Here, accepted demand bids are remunerated at the National Single Price (PUN), which therefore represents the purchase price.

The employed data set consists in the time series of PUNs, denominated henceforth $X_t$, which is freely and publicly available at the website

http://www.mercatoelettrico.org/En/download/DatiStorici.aspx.

The reported values are expressed in Euro currency (Euro per Megawatt), whereas the sampling frequency is one hour. The span of time considered in the present study is January $1^{st}$ 2005 –February $29^{th}$ 2016, for a total of 97.848 data. The time series has been differenciated twice, with difference of order 1 respectively at lag 24 and 168, therefore the effective sample size is T= 97656.

In Italy, the electric market has been fully liberalized on July $1^{st}$ 2007, therefore, in order to gain a better insight on the performances of the proposed approach, the whole analysis has been executed having in mind this reference time. In practice, PUN time series has been split in three non–overlapping sub-periods (denominated $Z_1, Z_2, Z_3$), covering the time before and after the liberalization, plus an arbitrary chosen warming up period located in between, that is:

- $Z_1 \equiv [01/01/2005; 06/30/2007]$, (21671 data);
- $Z_2 \equiv [07/01/2007; 12/31/2012]$, (48265 data);
- $Z_3 \equiv [01/01/2013; 12/29/2016]$, (27720 data).

Finally, with $Z_{tot}$ the full length window will be denoted, i.e. $Z_{tot} \equiv [01/01/2005; 12/29/2016]$.

## 4. Outcomes of the Experiment

In Table 1 the LE, computed for the three sub-series with span $Z_{1,2,3}$ and $Z_{tot}$ are reported. The presence of chaos is confirmed, for the windows $Z_{2,3}$, by a positive LE. The increasing pattern noticeable in the LE values, probably reflect the system behavior transitioning from a non chaotic to a chaotic status, as a result of the greater and greater degrees of freedom introduced into the electric market over the years. However, by inspecting Tables 2 – 4, additional information can be obtained. These are organized according to the six resolution levels – depicted in Figure 1 – the time series $X_t$ has been broken into. For each of them, other than the LEs, the following parameters are considered: variance estimation ($\sigma^2(d_j)$) as a value and as a ratio to the total variability ($\frac{\sigma^2(d_j)}{\sigma_X^2}$), equivalent degrees of freedom ($\delta_j$) and confidence intervals ($CI_{low}, CI_{high}$) for $\sigma^2(d_j)$. Deterministic chaos seems to be attributable in greater part to coarser resolution levels (> 2 days ), i.e. $d4 - d6$. In fact, in the case of $Z_1$ these components (weighing approx 47% of the total variability) are characterized by a negative LE whereas they show a positive LE for $Z_3$, where their relative weight is .70. By restricting our attention only on the resolution levels $j = 5, 6$, we are able to pinpoint with more precision the resolution levels at which deterministic chaos is generated. In fact, while d5 and d6 in $Z_1$ (i.e. 34% of total variability) show in both of the cases a negative LE, in $Z_2$ (47% of the total variability) they become positive as well as in $Z_3$. In the last case, however, they represent 54% of the total variation and, all the more so, LE in both the cases show greater magnitudes. Also, it is worth mentioning that the finest MRA levels, i.e. $j = 1, 2, 3$, exhibit values and relative importance consistently decreasing as the sub-series progress in time. In particular, the relative variability is 51% for $Z_1$ and 39% and .29% for $Z_2$ and $Z_3$ respectively. This last evidence gives account of the fact that the negative LEs found at levels 2 and 3 are related to a small quota of the overall variability (< $\frac{1}{4}\sigma_X^2$ ), and therefore cannot counteract the general chaotic behavior of this segment of $X_t$. The same reasoning applies to $Z_1$, where the same components – representing the 35% of the total variability – even if showing positive LE values, are embedded in the sub-series $Z_1$ whose overall dynamic is non chaotic.

Table 1. Series $X_t$, Lyapunov Exponent estimates for the time windows $Z_1, Z_2, Z_3, Z_{tot}$

| series | $Z_1$ | $Z_2$ | $Z_3$ | $X_t$ |
|---|---|---|---|---|
| Lyapunov Exponent Estimates | -.0005527541 | .0005230929 | .00476779 | .001668556 |

Table 2. Series $X_t$, MRA outcomes for the time frame $Z_1$

| Parameters<br>Crystals | Lyapunov exponent | $\sigma^2$ | $\frac{\sigma^2(d_j)}{\sigma^2_X}$ | $\delta_j$ | $CI_{low}$ | $CI_{high}$ |
|---|---|---|---|---|---|---|
| d1 | -.0003111276 | 39.43 | .17 | 7770.66 | 38.56 | 40.32 |
| d2 | .0001053132 | 43.16 | .17 | 4781.97 | 41.83 | 44.56 |
| d3 | .0003320897 | 44.73 | .18 | 2441.29 | 42.79 | 46.80 |
| d4 | -.000549612 | 35.91 | .14 | 1473.63 | 33.74 | 38.30 |
| d5 | -.001045508 | 55.99 | .22 | 505.57 | 51.30 | 61.40 |
| d6 | -003244198 | 28.33 | .11 | 366.81 | 25.05 | 32.35 |

Table 3. Series $X_t$, MRA outcomes for the time frame $Z_2$

| Parameters<br>Crystals | Lyapunov exponent | $\sigma^2$ | $\frac{\sigma^2(d_j)}{\sigma^2_X}$ | $\delta_j$ | $CI_{low}$ | $CI_{high}$ |
|---|---|---|---|---|---|---|
| d1 | .0002995352 | 32.92 | .11 | 16640.04 | 32.43 | 33.41 |
| d2 | .0004067459 | 40.80 | .14 | 9568.94 | 39.95 | 41.68 |
| d3 | .0003309755 | 41.50 | .14 | 5167.68 | 41.50 | 44.07 |
| d4 | -.0004087144 | 39.02 | .14 | 3313.77 | 39.02 | 42.48 |
| d5 | .0002110077 | 75.04 | .27 | 1428.65 | 75.04 | 84.62 |
| d6 | .00087144 | 55.49 | .20 | 852.20 | 55.49 | 65.79 |

Table 4. Series $X_t$, MRA outcomes for the time frame $Z_3$

| Parameters<br>Crystals | Lyapunov exponent | $\sigma^2$ | $\frac{\sigma^2(d_j)}{\sigma^2_X}$ | $\delta_j$ | $CI_{low}$ | $CI_{high}$ |
|---|---|---|---|---|---|---|
| d1 | .0001502846 | 8.38 | .066 | 9611.07 | 8.22 | 8.55 |
| d2 | -.001889886 | 11.18 | .088 | 5641.95 | 10.87 | 11.50 |
| d3 | -.001456725 | 18.31 | .14 | 2631.98 | 17.60 | 19.05 |
| d4 | .003439542 | 20.30 | .16 | 1718.19 | 19.87 | 22.23 |
| d5 | .01866641 | 38.90 | .31 | 685.46 | 36.00 | 42.20 |
| d6 | .01247388 | 29.27 | .23 | 471.07 | 26.24 | 32.88 |

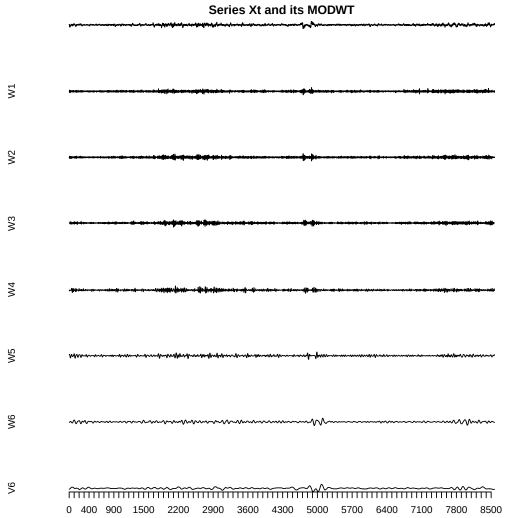


Figure 1. Time series $X_t$ and its $MODWT$ coefficient sequence $d_{j,t}$; $j = 1, \dots, 6$

## References

Aittokallio, T., Gyllenberg, M., Hietarinta, J., Kuusela, T., & Multamaäki, T. (1999). Improving the false nearest neighbors method with graphical analysis. *Physical Review E 60*(1), 416. http://dx.doi.org/10.1103/PhysRevE.60.416

Brock, W. A., Brock, A., & Malliaris, A. (1989). Differential equations, stability and chaos in dynamic economics. No. 90A16 BROd

Chadwick, A. (1998). Archaeology at the edge of chaos: further towards reflexive excavation methodologies. *Assemblage 3*, 97-117.

Deyle, E. R., & Sugihara, G. (2011). Generalized theorems for nonlinear state space reconstruction. *PLoS One 6*(3), e18295. http://dx.doi.org/10.1371/journal.pone.0018295

Eve, R. A., Horsfall, S., & Lee, M. E. (1997). *Chaos, complexity, and sociology: Myths, models, and theories*. Sage.

Freeman, W. J. (1992). Tutorial on neurobiology: from single neurons to brain chaos. *International journal of bifurcation and chaos 2*(03), 451-482. http://dx.doi.org/10.1142/S0218127492000653

Guastello, S. J. (2013). *Chaos, catastrophe, and human affairs: Applications of nonlinear dynamics to work, organizations, and social evolution*. Psychology Press.

Harman, S., Fenwick, A., & Williams, C. (2006). Chaotic signals in radar? *In: 2006 European Radar Conference*( pp. 49-52). IEEE. http://dx.doi.org/10.1109/EURAD.2006.280270

Lamarque, C. H. , & Malasoma, J. M. (1996). Analysis of nonlinear oscillations by wavelet transform: Lyapunov exponents. *Nonlinear Dynamics 9*(4), 333-347. http://dx.doi.org/10.1007/BF01833360

Liu, Z., Zhu, X., Hu, W., & Jiang, F. (2007). Principles of chaotic signal radar. *International Journal of Bifurcation and Chaos 17*(05), 1735-1739. http://dx.doi.org/10.1142/S0218127407018038

Merry, R. , & Steinbuch, M. (2005). Wavelet theory and applications. Literature Study, Eindhoven University of Technology, Department of Mechanical Engineering, Control Systems Technology Group.

Noakes, L. (1991). The takens embedding theorem. *International Journal of Bifurcation and Chaos 1*(04), 867-872. http://dx.doi.org/10.1142/S0218127491000634

Olsen, N., Mandea, M., Sabaka, T. J., & Tøffner-Clausen, L. (2009). Chaos-2a geomagnetic field model derived from one decade of continuous satellite data. *Geophysical Journal International 179*(3), 1477-1487. http://dx.doi.org/10.1111/j.1365-246X.2009.04386.x

Percival, D. B. (2002). *Wavelets*. Encyclopedia of environmetrics.

Percival, D. B., & Walden, A. T. (2006). *Wavelet methods for time series analysis*, vol. 4. Cambridge university press.

Peters, E. E. (1994). *Fractal market analysis: applying chaos theory to investment and economics*, vol. 24. John Wiley & Sons.

Provenzale, A., Smith, L. A., Vio, R., & Murante, G. (1992). Distinguishing between low-dimensional dynamics and randomness in measured time series. *Physica D: nonlinear phenomena 58*(1), 31-49. http://dx.doi.org/10.1016/0167-2789(92)90100-2

Puu, T. (2013). *Attractors, bifurcations, & chaos: Nonlinear phenomena in economics*. Springer Science & Business Media.

Sauer, T., & Yorke, J. A. (1993). How many delay coordinates do you need? *International Journal of Bifurcation and Chaos 3*(03), 737-744. http://dx.doi.org/10.1142/S0218127493000647

## Copyrights

ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

# Determination of Support Vector Boundaries in Generalized Maximum Entropy for Multilevel Models

Serpil Kılıç Depren[1] & Özer Depren[2]

[1] Department of Statistics, Yıldız Technical University, İstanbul, Turkey

[2] Customer Experience and Idea Management Department, Yapı Kredi Bank, İstanbul, Turkey

Correspondence: Serpil Kılıç Depren, Department of Statistics, Faculty of Arts and Science, Yıldız Technical University, Davutpasa Campus, 34220, Esenler-Istanbul, Turkey. E-mail: serkilic@yildiz.edu.tr

Received: August 9, 2016 Accepted: August 28, 2016 Online Published: October 17, 2016

doi:10.5539/ijsp.v5n6p10 URL: http://dx.doi.org/10.5539/ijsp.v5n6p10

**Abstract**

Generalized Maximum Entropy (GME) approach is one of the alternative estimation methods for Regression Analysis. GME approach is superior to other classical approaches in terms of parameter estimation accuracy when some or none of the assumptions of classical approaches are violated. However, determining bounds of parameter support vectors is one of the open parts of this approach when researchers have no prior information about the parameters. If support vectors cannot be determined correctly, parameters estimations will not be obtained correctly. There are some theoretical studies about GME for different datasets in the literature, but there are fewer studies about how to determine parameter support vectors. To obtain robust parameter estimations in GME, we introduced a new iterative procedure for determining parameter support vectors bounds for multilevel dataset. In this study, the new iterative procedure was applied for multi-level random intercept model and the new procedure was tested both simulation study and the real life data. The Classical and the new procedures of GME estimations were compared to Generalized Least Square Estimations in terms of Root Mean Square Error (RMSE) statistics. As a result, the estimations of the new approach provided lower RMSE values than classical methods.

**Keywords:** generalized maximum entropy, multilevel models, support vector bounds

## 1. Introduction

Shannon was defined the term "Entropy" as a measure of uncertainty in communication theory in 1948 and the basic principle of Generalized Maximum Entropy (GME) is based on Jaynes' Maximum Entropy Principle (Shannon, 1948; Jaynes, 1957). Golan, Judge and Miller generalized this principle for regression framework (Golan, Judge, & Miller, 1996). In this approach, Golan et al. maximized Shannon's entropy formula under model consistency constraints. GME approach requires fewer assumptions than the classical methods and it has been used an alternative approach for both classical linear and nonlinear estimation models (Golan, Judge, & Miller, 1996). The most important part of GME is re-parameterization of regression coefficients and the error term processes. The main topic of GME approach is determination of support vector boundaries when researchers have no prior information about the parameters. Thus, this approach has been receiving increasing attention in the statistics literature.

In this study, Random Intercept Model and GME Approach were combined. The main purpose of this study is to determine parameter support vector boundaries which allow to obtain more consistent parameters than classical methods for multilevel data. Therefore, the determination of error vector boundaries will not be taken into consideration in this study. Thus, the error vector boundaries were determined according to the literature by Pukelsheim (3σ rule). (Pukelsheim, 1994).

The outline of this paper organized as follows. The literature review was given in Section 2. We introduced Multilevel Modeling and briefly describe the GME estimation process in Section 3. In Section 4, we suggested the new procedure for determination of parameter support vector boundaries in Multilevel Random Intercept Models and we presented both real-life and simulation study. Also we compared the results obtained from different methods. Section 5 is a brief conclusion of the study.

## 2. Literature Review

In the literature, the term “Entropy” (and also known as “Maximum Entropy”) has been widely used in many disciplines such as thermodynamics, communication, education and statistics. In previous studies, researches were proposed new approaches which could be used rather than classical approaches and compared their results to different methods with simulation studies (Al-Nasser, 2014; Al-Rawwash & Al-Nasser, 2011). Also, there were several attempts to answer the question of “How parameter and the error term support vectors could be determined in the GME approach?” (Henderson, Golan & Seabold, 2015; Ciavolino & Calcagni, 2014; Golan & Gzly, 2012; Fernandez-Vazquez, Mayor-Fernandez & Rodriguez-Valez, 2008; Caputo & Paris, 2008), especially when researchers have no prior information about distribution of the parameters. The Maximum Entropy principle was also combined with Longitudinal Analysis, Data Envelopment Analysis, Multilevel Models, Regression Analysis and Logistic Regression Analysis in statistical area in order to compare GME results to the results of classical approach (Al-Nasser & Al-Atrash, 2011; Al-Nasser et al., 2010; Gastón & García-Vinas, 2011; Donoso et al., 2011). Furthermore, alternative solutions were proposed when data have the problem of multicollinearity and heteroscedasticity (Akdeniz et al., 2011).

Abellan, Baker and Coolen studied on the Nonparametric Predictive Inference (NPI) model for the multilevel data and proved that the lower and upper probabilities of the NPI could be obtained only by using the singleton probabilities (Abellan, Baker, & Coolen, 2011).

The performance of the GME estimator in both large and small samples was studied to assess the efficiency of the results in terms of estimation accuracy (Mittelhammer, Cardell, & Marsh, 2013; Gastón & García, 2011).

GME Approach was combined with Hierarchical Cumulative Logit Model in the study of Donoso, Grange and González in 2011. They compared the results of Hierarchical Cumulative Logit Model with Maximum Likelihood estimation using Monte Carlo simulations (Donoso, Grange, & González, 2011). In conclusion, they showed that the simulations produced reduced-bias in the estimates of the subjective value of time.

An alternative solution for the problem of multicollinearity was suggested by Akdeniz, Çabuk and Güler. Appropriate constraints were added to the classical Generalized Maximum Entropy Approach according to the characteristics of the relationship among independent variables and the results were compared with OLS (Akdeniz, Çubuk, & Güler, 2011).

The mathematical properties of Entropy, Maximum Entropy, Minimum Cross Entropy and Maximum Entropy Leuven Method were also explained in detail in Altaylıgil's study in 2008. The performance of the parameter estimations of this method was tested with OLS, Generalized OLS and Ridge by using Monte Carlo simulations (Altaylıgil, 2008).

In all of these studies, new algorithm or methodology comparison between classical methods and GME were provided, but there were not many study about how to determine parameter support vectors (especially bounds of support vectors), which is the missing part of this approach especially when researchers have no prior information about parameters distributions.

## 3. Methods

Multilevel Modelling is a generalization of linear regression models and such can be used for a variety of purposes, including prediction, data reduction and causal inference from clustered or hierarchical datasets (Gelman, 2006; Raudenbush et al., 2005; Raudenbush & Bryk, 2002). Also, GME, which is based on optimization technique, is one of the alternative solution approaches of Linear and Non-Linear Models. In this approach, all of the unknown parameters (β) and the error term (e) are re-parameterized by using finite-dimensional known support vectors z and v (Golan, Judge & Miller, 1996). Similar to the coefficient re-parameterization process, e could be re-parameterized as a finite and discrete random variable with $2 \leq J \leq \infty$ possible outcomes (Golan, Judge, & Miller, 1996). Adaptation of the re-parameterization process to Random Intercept Models is given in the next section.

### *3.1 Adaptation of GME Approach to Random Intercept Models*

The study of Al-Nasser in 2010, GME approach was adapted to Multilevel Random Coefficient Model (Al-Nasser et al., 2010). In this study, GME approach was adapted to Random Intercept Model which was illustrated by Raudenbush and Bryk (2002) and Raudenbush et al. (2005) and a new iterative approach was proposed for determination of parameter support vectors for Random Intercept Model.

Two-level Random Intercept Model can be expressed as two equations.

$$Level-1: y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \qquad j = 1,2, \dots n \qquad i = 1,2, \dots, n \tag{1}$$

In Equation 1, i refer to level-1 units while j refers to level-2 units. $Y_{ij}$ refers to response variable for level-1 unit i within level-2 unit j. $\beta_{0j}$ is the random intercept for level-2 unit j while $\beta_{1j}$ is the random slope of $X_i$ of unit j. $r_{ij}$ is the residual term for unit i within unit j.

$$\begin{aligned} Level-2: \beta_{0j} &= \gamma_{00} + \gamma_{01}W_j + U_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}W_j \end{aligned} \tag{2}$$

In Equation 2, $\gamma_{00}$ and $\gamma_{10}$ are the intercepts. $\gamma_{01}$ and $\gamma_{11}$ represent slopes predicting $\beta_{0j}$ and $\beta_{1j}$ respectively. Furthermore, $U_{0j}$ is the level-2 random errors. Finally, Equation 1 and 2 could be written as in Equation 3.

$$y_{ij} = \gamma_{00} + \gamma_{01}W_j + U_{0j} + (\gamma_{10} + \gamma_{11}W_j)X_{ij} + r_{ij} \tag{3}$$

In Equation 3, there are four unknown parameters and two error terms. They will be re-parameterized in order to use Generalized Maximum Entropy Approach (Al-Nasser et al. 2010).

$$\begin{aligned}
\gamma_{00} &= \textstyle\sum_{r=1}^{R} a_r p_r && where\ p_r \in (0,1)\ \ and\ \textstyle\sum_{r=1}^{R} p_r = 1\\
\gamma_{10} &= \textstyle\sum_{b=1}^{B} z_b q_b && where\ q_b \in (0,1)\ \ and\ \textstyle\sum_{b=1}^{B} q_b = 1\\
\gamma_{01} &= \textstyle\sum_{k=1}^{K} c_k N_k && where\ N_k \in (0,1)\ \ and\ \textstyle\sum_{k=1}^{K} N_k = 1\\
\gamma_{11} &= \textstyle\sum_{s=1}^{S} d_s G_s && where\ G_s \in (0,1)\ \ and\ \textstyle\sum_{s=1}^{S} G_s = 1\\
U_{0j} &= \textstyle\sum_{i=1}^{E} v_{ij} T_{ij} && where\ T_{ij} \in (0,1)\ \ and\ \textstyle\sum_{i=1}^{E}\ T_{ij} = 1\\
r_{ij} &= \textstyle\sum_{l=1}^{M} v_{lij}^{*} O_{lij} && where\ O_{lij} \in (0,1)\ \ and\ \textstyle\sum_{i=1}^{M}\ O_{lij} = 1
\end{aligned} \tag{4}$$

where j=1,2,…,J and I=1,2,…,$n_j$.

After re-parameterization process, the model could be written as below.

$$\begin{gathered}
y_{ij} = \textstyle\sum_{r=1}^{R} a_r p_r + (\sum_{b=1}^{B} z_b q_b)x_{ij} + (\sum_{k=1}^{K} c_k N_k)w_j +\\
+(\textstyle\sum_{s=1}^{S} d_s G_s)w_j x_{ij} + \sum_{i=1}^{E} v_{ij} T_{ij} + \sum_{l=1}^{M} v_{lij}^{*} O_{lij}
\end{gathered} \tag{5}$$

Therefore, the GME model for the Two-Level Random Intercept Model can be expressed by the following nonlinear programming system:

$$\begin{gathered}
Maximize\ H(p,q,N,G,T,O) = -\textstyle\sum pln(p) - \sum qln(q) - \sum Nln(N) - \sum Gln(G) - \sum Tln(T) -\\
\textstyle\sum Oln(O)
\end{gathered} \tag{6}$$

Subject to:

$$\begin{gathered}
y_{ij} = \textstyle\sum_{r=1}^{R} a_r p_r + (\sum_{b=1}^{B} z_b q_b)x_{ij} + (\sum_{k=1}^{K} c_k N_k)w_j + (\sum_{s=1}^{S} d_s G_s)w_j x_{ij} +\\
+\textstyle\sum_{i=1}^{E} v_{ij} T_{ij} + \sum_{l=1}^{M} v_{lij}^{*} O_{lij}\\
\textstyle\sum_{r=1}^{R} p_r = 1\\
\textstyle\sum_{b=1}^{B} q_b = 1\\
\textstyle\sum_{k=1}^{K} N_k = 1\\
\textstyle\sum_{s=1}^{S} G_s = 1\\
\textstyle\sum_{i=1}^{E}\ T_{ij} = 1\\
\textstyle\sum_{i=1}^{M}\ O_{lij} = 1
\end{gathered} \tag{7}$$

The new optimization problem for Random Intercept Model is solved by using Lagrangian Method.

### *3.2 The Procedure for the Determination of Parameter Support Boundaries*

In the literature, there are two main rules for determining support vector boundaries which are based on prior knowledge about parameters and the error term (Golan, Judge, & Miller, 1996).

- Support vector boundaries might be determined according to the prior information of parameters and the error term.
- They might also be determined extensively enough to include population parameters around zero and Pukelsheim's 3σ rule for the error term when the information about the parameters and the error term does not exist.

In the first rule, for example, if it is assumed that the mean, minimum and maximum values of $\beta_1$ are 1, 0.5 and 2, respectively; the parameter support vector should be $z_1$=[0.5 1 2] for M=3 and $z_1$=[0.5 0.75 1 1.5 2] for M=5. However, in the literature, it was not clearly identified how accurate support vector boundaries were determined when the researchers had no prior information about the parameters and the error term. As a result of these literature findings, the number of discrete points of parameters and the error term support vectors were selected as five (Golan, Judge, & Miller, 1996; Al-Nasser, 2011).

This study aimed to obtain prior information using the information of the current dataset. For this aim, the following steps, which were called as repeated sampling with replacement in Depren’s study, were adopted for multilevel dataset (Depren, 2014):

1. A new sample (n and 2n sample size) was created by using repeated sampling with replacement (n is the total number of observation).

2. Restricted Maximum Likelihood approach was used to obtain prior parameters without checking whether the data met the required assumptions, such as multicollinearity, autocorrelation and homoscedasticity or not.
3. The first two steps were repeated (t) 50, 100, 500, 1000, 1500 and 2000 times to obtain 50, 100, 500, 1000, 1500 and 2000 different values (parameter matrix) in each repetition.
4. All the obtained parameter estimations were sorted in an ascending order.
5. %5 of the top and the bottom values of the parameters (outliers) were extracted from the parameter matrix.
6. The support vector boundaries of each parameter were determined according to Equation (8).

$$z' = \left[TrimMean - \frac{\sigma}{4} \quad TrimMean - \frac{\sigma}{8} \quad TrimMean \quad TrimMean + \frac{\sigma}{8} \quad TrimMean + \frac{\sigma}{4}\right] \tag{8}$$

Since sample size (n), repeats (t) and parameter support vectors (z) are the parameters of the procedure to be identified, the procedure is much more complex than standard GME or other estimation techniques. However, simple code can be written in SAS, R or in other package programs to run the procedure. Thus, researchers can overcome this complexity in business life.

The results obtained from classical and new approaches were compared by using RMSE (Miller, 2002; Timm, 2002).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2} \tag{9}$$

## 4. Application

The new procedure was tested in both simulated and real-life datasets. In this study, the following questions to be answered were which sample size should be chosen (n or 2n) and how many repeats (t) should be run for sampling with replacement procedure.

### *4.1 Background of Simulated Data*

The simulation study was performed under following assumptions:

1. Generate 1000, 2000, 3000, 5000 and 10000 random sample of size n and 2n with repeated sampling with replacement technique.
2. The errors are distributed as $\boldsymbol{r_{ij}}$*~Normal(0,1)* and $\boldsymbol{U_{0j}}$*~Normal(0,1)*
3. The variables are distributed as $\boldsymbol{X_{ij}}$*~Normal(7,1)* and $\boldsymbol{W_j}$*~Uniform(0,1)*
4. The coefficients were set $\boldsymbol{\gamma_{00} = 1}$, $\boldsymbol{\gamma_{01} = 2}$, $\boldsymbol{\gamma_{10} = 1.5}$ and $\boldsymbol{\gamma_{11} = 1}$.
5. Five-point support vectors for parameters and the error term were used for GME estimator (Al-Nasser 2011).
6. For determination of parameter support vector boundaries, two different alternatives were tested.
   i. Parameter support vector bounds were determined extensively enough to include the population parameters around zero and these bounds were narrowed down in every iteration for each sample.

      1.iteration z'=[-1000 -500 0 500 1000]

      2.iteration z'=[-100 -50 0 50 100]

      3.iteration z'=[-10 -5 0 5 10]

   ii. Parameter support vector boundaries were determined according process which was explained in Section 3.2.
7. The error support bounds were determined by Pukelsheim's 3σ rule. e = [-3s -1.5s 0 1.5s 3s] where s is the standard deviation of the dependent variable.
8. Simulation results were compared using RMSE (Miller 2002; Timm 2002).

### *4.2 Simulation Study*

In order to identify n and t for sampling with the replacement procedure, Table 1 was prepared. Model outputs are shown by n and t values for all datasets.

Table 1. Outputs of the alternative approach for simulated dataset

| | Intercept | Coefficient of X | Coefficient of W | RMSE |
|---|---|---|---|---|
| Sample Size: n | -3.83602 | 2.1975 | 9.0183 | 1.5374 |
| t: 1000 | (3.23E-06) | (4.46E-07) | (1.80E-07) | |
| Sample Size: n | -3.83586 | 2.19752 | 9.0178 | 1.5374 |
| t: 2000 | (3.02E-06) | (4.11E-07) | (1.83E-07) | |
| Sample Size: n | -3.83502 | 2.19736 | 9.0191 | 1.5374 |
| t: 3000 | (2.99E-06) | (4.13E-07) | (1.80E-07) | |
| Sample Size: n | -3.8361 | 2.1976 | 9.01768 | 1.5374 |
| t: 5000 | (3.08E-06) | (4.23E-07) | (1.85E-07) | |
| Sample Size: n | -3.83704 | 2.19772 | 9.01822 | 1.5374 |
| t: 10000 | (3.16E-06) | (4.27E-07) | (1.82E-07) | |
| Sample Size: 2n | -3.83604 | 2.19746 | 9.01914 | 1.5374 |
| t: 1000 | (1.49E-06) | (2.06E-07) | (9.33E-08) | |
| Sample Size: 2n | -3.83772 | 2.19772 | 9.01894 | 1.5374 |
| t: 2000 | (1.53E-06) | (2.08E-07) | (9.10E-08) | |
| Sample Size: 2n | -3.83542 | 2.19744 | 9.0178 | 1.5374 |
| t: 3000 | (1.49E-06) | (2.06E-07) | (8.94E-08) | |
| Sample Size: 2n | -3.83598 | 2.1975 | 9.01826 | 1.5374 |
| t: 5000 | (1.53E-06) | (2.10E-07) | (9.01E-08) | |
| Sample Size: 2n | -3.83736 | 2.1977 | 9.01804 | 1.5374 |
| t: 10000 | (1.58E-06) | (2.14E-07) | (8.78E-08) | |

( ):Standard Deviation

Once the results of the different scenarios were compared according to RMSE, there was no significant difference for all scenarios in terms of β coefficients. For this reason, identifying n and t is not important for simulated dataset. In this study, first scenario (2n sample size and 1000 repeats) was chosen for further analysis.

The parameter support vectors, used for all datasets, were shown in Table 2. In this section, Support Vector I, II and III were used for Classical GME estimations and Support Vector IV was used for proposed GME estimation.

Table 2. Parameter support vectors used in analysis

| Support Vectors | Boundaries |
|---|---|
| Support Vector I<br>(Classical GME) | $\begin{bmatrix} Constant \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} -10 & -5 & 0 & 5 & 10 \\ -10 & -5 & 0 & 5 & 10 \\ -10 & -5 & 0 & 5 & 10 \\ -10 & -5 & 0 & 5 & 10 \end{bmatrix}$ |
| Support Vector II<br>(Classical GME) | $\begin{bmatrix} Constant \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} -100 & -50 & 0 & 50 & 100 \\ -100 & -50 & 0 & 50 & 100 \\ -100 & -50 & 0 & 50 & 100 \\ -100 & -50 & 0 & 50 & 100 \end{bmatrix}$ |
| Support Vector III<br>(Classical GME) | $\begin{bmatrix} Constant \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} -1000 & -500 & 0 & 500 & 1000 \\ -1000 & -500 & 0 & 500 & 1000 \\ -1000 & -500 & 0 & 500 & 1000 \\ -1000 & -500 & 0 & 500 & 1000 \end{bmatrix}$ |
| Support Vector IV<br>(Proposed GME) | $z' = \left[ TrimMean - \frac{\sigma}{4} \quad TrimMean - \frac{\sigma}{8} \quad TrimMean \quad TrimMean + \frac{\sigma}{8} \quad TrimMean + \frac{\sigma}{4} \right]$ |

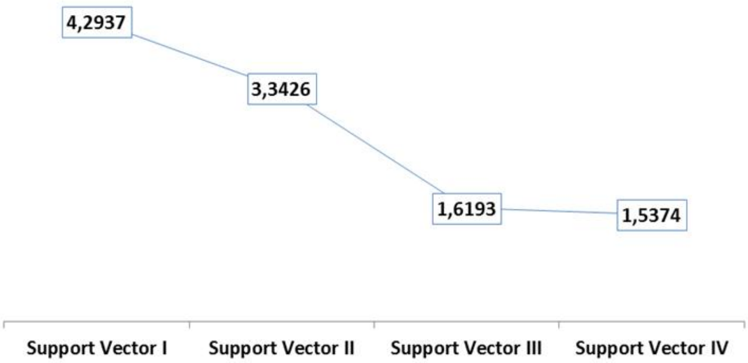


Figure 1. RMSE statistics of the classical and alternative approaches for simulated dataset

As shown in the Figure 1, as the parameter support vector bounds were widened, RMSE decreased. The new estimation (iterative) technique (Support Vector IV) produced better RMSE statistics.

Coefficients and standard errors were given in Table 3. In the new iterative approach (Support Vector IV), standard errors were estimated relatively small.

Table 3. Coefficients and standard errors of different parameter support vectors

| | Support Vector I | Support Vector II | Support Vector III | Support Vector IV |
|---|---|---|---|---|
| Intercept | 0.35571 | 0.44843 | -0.23014 | -3.83604 |
| | *(7.25E-04)* | *(1.08E-03)* | *(2.06E-02)* | *(1.49E-06)* |
| X | 2.45960 | 2.23619 | 1.73687 | 2.19746 |
| | *(4.78E-03)* | *(4.10E-03)* | *(4.39E-03)* | *(2.06E-07)* |
| W | 0.28050 | 2.53905 | 8.56674 | 9.01914 |
| | *(5.76E-04)* | *(1.56E-02)* | *(2.75E-02)* | *(9.33E-08)* |

( ):Standard Deviation

### *4.3 An Application to Real Life Data*

New procedure was tested with the data of the Programme for International Study Assessment (PISA) conducted in 2009. Data consists of a total sample of 515.985 students who are nested within 1.535 schools (OECD, 2009). Similar to the study of Kılıç et al. (2012), Turkey and neighbouring countries of Turkey, which are Bulgaria, Greece, Azerbaijan, Russian Federation, Israel, Serbia, Romania and Jordan, were included in this study in order to examine learning strategies accounted for mathematics achievement. Thus, the results of two studies could be compared in terms of estimation accuracy. In this study, the study of Kılıç et al. (2012) is named as referenced study.

Table 4. Mathematics scores of countries

| Countries | Mathematics Score |
|---|---|
| Russian Federation | 468 |
| Greece | 466 |
| Israel | 447 |
| Turkey | 445 |
| Serbia | 442 |
| Azerbaijan | 431 |
| Bulgaria | 428 |
| Romania | 427 |
| Jordan | 387 |
| OECD Average | 496 |

PISA mathematics test score of 42.417 15-year-old students were analysed. Three-level random coefficient model was used to model differences across countries and across schools. In this study, mathematics achievement was considered as a dependent variable. At the first level, gender, socio–economic status, elaboration, memorization, control strategy, home educational resources and cultural possession were considered. School size and student–teacher ratio were considered at the second level and gross domestic product (GDP) was considered at the third level variables. Mathematics achievement score of countries are given in Table 4.

The best performer country is Russian Federation while the worst performer country is Jordan in PISA 2009 study in terms of mathematics achievement. Turkey is at rank 4 among these countries.

First-Level Variables;

1. Gender: Male coded as 1 and Female coded as 0.
2. Socio–Economic and Cultural Status (ESCS): The index of ESCS was derived from three indices: home possessions, higher parental occupation (HISEI) and higher parental education expressed as years of schooling.
3. Cultural Possession (CULTPOSS): It was derived from students' responses to the three items listed below.
    a) Classic literature, b) Books of property, c) Works of art
4. Home Educational Resources (HEDRES): The PISA 2006 index of home educational resources was derived from students' responses to the some items.
    a) Desk for study, b) A quiet place to study, c) Your own calculator, d) Books to help with your school work, e) A dictionary
5. Memorization (MEMOR): It was derived from students' responses to the four items measuring preference for memorisation/rehearsal as a learning strategy for mathematics as listed below.
    a) I go over some problems in mathematics so often that I feel as if I could solve them in my sleep, b) When I study for mathematics, I try to learn the answers to problems off by heart, c) In order to remember the method for solving a mathematics problem, I go through examples again and again, d) To learn mathematics, I try to remember every step in a procedure.
6. Elaboration (ELAB): It was derived from students' responses to the five items measuring preference for elaboration strategy as listed below.
    a) When I am solving mathematics problems, I often think of new ways to get the answer, b) I thing how the mathematics I have learnt can be used in everyday life, c) I try to understand new concepts in mathematics by relating them to things I already know, d) When I am solving mathematics problems, I often think about how the solution might be applied to other interesting questions, e) When learning mathematics, I try to relate the work to things I have learnt in other subjects
7. Control Strategy (CSTRAT): Control learning strategies was derived from students' responses to the five items measuring preference for control as a learning strategy as listed below.
    a) When I study for a mathematics test, I try to work out what are the most important parts to learn, b) When I study mathematics, I make myself check to see if I remember the work I have already done, c) When I study mathematics, I try to figure out which concepts I still have not understood properly, d) When I cannot understand something in mathematics, I always search for more information to clarify the problem, e) When I study mathematics, I start by working out exactly what I need to learn.

Second-Level Variables;

1. School Size (SCSIZE): Total number of male and female students in a school.
2. Student–Teacher Ratio (STRATIO): The number of students per teacher in a school.

Third-Level Variable;

1. Gross Domestic Product (GDP): Gross Domestics Product of a country. Since GDP has right skewed distribution, Log(GDP) is used in this study.

Descriptive statistics of these variables are given in Table 5. Male-Female ratio was 50%-50%.

Table 5. Descriptive statistics of the variables

| | Min | Max | Mean | Standard Deviation |
|---|---|---|---|---|
| Socio–Economic and Cultural Status | -4.79 | 3.09 | -0.31 | 1.05 |
| Home Educational Resources | -4.52 | 2.64 | 0.11 | 1.25 |
| Cultural Possession | -2.24 | 1.82 | 0.16 | 1.03 |
| Memorization | -3.02 | 2.69 | 0.30 | 1.03 |
| Elaboration | -2.41 | 2.76 | 0.34 | 1.02 |
| Control Strategy | -3.45 | 2.50 | 0.18 | 1.03 |
| School Size | 9.00 | 3140.00 | 713.05 | 448.25 |
| Student–Teacher Ratio | 0.51 | 79.63 | 13.62 | 6.47 |
| *Log*(Gross Domestic Product) | 10.40 | 12.09 | 11.13 | 0.57 |

Output of Restricted Maximum Likelihood (REML) estimation method is given in Table 6.

Table 6. Restricted maximum likelihood estimation

| | Coefficient | Standard Deviation |
|---|---|---|
| Intercept | 123.66 | 81,63 |
| Gender (Male) | 18.21 | 0,65 |
| Socio–Economic and Cultural Status | 10.71 | 0,39 |
| Home Educational Resources | 4.16 | 0,35 |
| Cultural Possession | 4.15 | 0,35 |
| Memorization | -11.59 | 0,34 |
| Elaboration | 2.76 | 0,36 |
| Control Strategy | 11.56 | 0,39 |
| School Size | -1.45 | 0,25 |
| Student–Teacher Ratio | 0.02 | 0,003 |
| *Log*(Gross Domestic Product) | 28.42 | 7,31 |
| *RMSE* | 1219.5 | |

Table 7 was prepared to identify n and t for sampling with replacement procedure. Model outputs are shown by n and t values for PISA dataset.

Table 7. Coefficient estimates and standard deviation for different sample size and repeats

| | Intercept | Gender | Escs | Memor | Elab | Cstrat | Cultposs | Hedres | Stratio | Schsize | LogGDP | RMSE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample Size:n<br>t: 1000 | 131.56 | 10.29 | 23.96 | -15.10 | 0.29 | 17.47 | 5.49 | 4.00 | -1.43 | 0.01 | 29.24 | 733.68 |
| Sample Size:n<br>t: 2000 | 131.54 | 10.30 | 23.94 | -15.13 | 0.28 | 17.47 | 5.51 | 3.98 | -1.43 | 0.01 | 29.24 | 733.68 |
| Sample Size:n<br>t: 3000 | 131.54 | 10.29 | 23.95 | -15.11 | 0.29 | 17.48 | 5.50 | 3.99 | -1.43 | 0.01 | 29.24 | 131.54 |
| Sample Size:n<br>t: 5000 | 131.28 | 10.32 | 23.95 | -15.12 | 0.29 | 17.47 | 5.49 | 3.99 | -1.43 | 0.01 | 29.26 | 733.68 |
| Sample Size:n<br>t: 10000 | 131.60 | 10.29 | 23.95 | -15.12 | 0.29 | 17.47 | 5.50 | 3.99 | -1.43 | 0.01 | 29.23 | 733.68 |
| Sample Size:2n<br>t: 1000 | 131.23 | 10.31 | 23.95 | -15.12 | 0.28 | 17.48 | 5.50 | 3.99 | -1.43 | 0.01 | 29.26 | 733.68 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample Size:2n t: 2000 | 131.26 | 10.33 | 23.96 | -15.12 | 0.29 | 17.48 | 5.49 | 3.99 | -1.43 | 0.01 | 29.26 | 733.68 |
| Sample Size:2n t: 3000 | 131.58 | 10.29 | 23.95 | -15.13 | 0.29 | 17.47 | 5.50 | 4.00 | -1.43 | 0.01 | 29.23 | 733.68 |
| Sample Size:2n t: 5000 | 131.39 | 10.31 | 23.95 | -15.12 | 0.29 | 17.47 | 5.50 | 3.99 | -1.43 | 0.01 | 29.25 | 733.68 |
| Sample Size:2n t: 10000 | 131.39 | 10.31 | 23.95 | -15.12 | 0.29 | 17.47 | 5.50 | 3.99 | -1.43 | 0.01 | 29.25 | 733.68 |

Similar to simulation study, the results of the different scenarios were compared and there was no significant difference for all scenarios in terms of β coefficients and RMSE statistics. For this reason, similar to simulation study, determination of n and t is not important for PISA dataset. In this study, first scenario (2n sample size and 1000 repeats) was chosen for further analysis.

Different support vectors were used for each regression model. Thus, the efficiency of the new procedure could be compared by using RMSE statistics. The parameter support vectors are given in Table 8 for PISA dataset.

Table 8. Different support vectors of the new and classical approaches

| Support Vectors | Boundaries |
|---|---|
| Support Vector I for all coefficients | $[-10 \quad -5 \quad 0 \quad 5 \quad 10]$ |
| Support Vector II for all coefficients | $[-100 \quad -50 \quad 0 \quad 50 \quad 100]$ |
| Support Vector III for all coefficients | $[-1000 \quad -500 \quad 0 \quad 500 \quad 1000]$ |
| Support Vector IV for all coefficients | $TrimMean - \frac{\sigma}{4} \quad TrimMean - \frac{\sigma}{8} \quad TrimMean \quad TrimMean + \frac{\sigma}{8} \quad TrimMean + \frac{\sigma}{4}$<br>$k = 1,2,\dots,K$ |

( ):Standard Deviation

In Figure 2, the RMSE statistics of the new iterative approach and classical approach were shown.

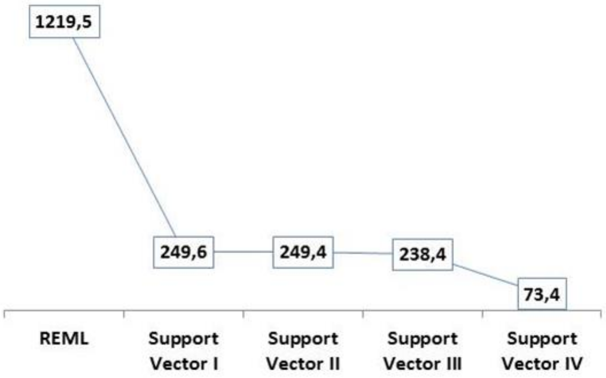


Figure 2. RMSE statistics of the classical and alternative approaches for PISA dataset

As shown in Figure 2, creation of prior information was important to obtain a more robust model in GME. As the parameter support vector bounds were widened, RMSE decreased. The new estimation (iterative) technique (Support

Vector IV) produced lower RMSE and it was significant that the iterative technique required no assumptions regarding distribution or parameters.

Coefficients and standard errors are given in Table 9.

Table 9. Coefficient estimates and standard deviation of the alternative approach for PISA dataset

| | REML | Support Vector I | Support Vector II | Support Vector III | Support Vector IV |
|---|---|---|---|---|---|
| Intercept | 123.66 | 0.00046 | 0.00139 | 0.08998 | 131.2301 |
| | (81.63) | (1.45E-06) | (5.81E-06) | (4.98E-04) | (4.70E-05) |
| Gender | 18.21 | 0.00022 | 0.67800 | 0.043324 | 10.31170 |
| | (0.65) | (7.09E-07) | (2.79E-06) | (2.39E-04) | (2.32E-07) |
| ESCS | 10.71 | -0.00014 | -0.00057 | -0.04168 | 23.94882 |
| | (0.39) | (4.31E-07) | (3.00E-06) | (2.63E-04) | (4.51E-08) |
| MEMOR | -11.59 | 0.00015 | 0.00036 | 0.020285 | -15.11750 |
| | (0.34) | (4.66E-07) | (1.39E-06) | (1.17E-04) | (5.38E-08) |
| ELAB | 2.76 | 0.00016 | 0.00045 | 0.027699 | 0.28162 |
| | (0.36) | (5.23E-07) | (1.89E-06) | (1.61E-04) | (6.17E-08) |
| CSTRAT | 11.56 | 0.00009 | 0.00019 | 0.008873 | 17.48090 |
| | (0.39) | (3.01E-07) | (7.03E-07) | (5.70E-05) | (4.70E-08) |
| CULTPOSS | 4.15 | 0.00008 | 0.00021 | 0.012142 | 5.49562 |
| | (0.35) | (2.56E-07) | (7.89E-07) | (6.70E-05) | (2.53E-08) |
| HEDRES | 4.16 | 0.00007 | 0.00009 | 0.001621 | 3.99492 |
| | (0.35) | (2.16E-07) | (4.90E-07) | (4.20E-05) | (1.59E-08) |
| STRATIO | 0.02 | 0.00698 | 0.02054 | 1.288434 | -1.43140 |
| | (0.003) | (2.20E-05) | (9.50E-05) | (0.00814) | (7.05E-08) |
| SCHSIZE | -1.45 | 0.44340 | 0.44260 | 0.41276 | 0.00884 |
| | (0.25) | (1.44E-03) | (1.44E-03) | (0.00139) | (5.90E-10) |
| LogGDP | 28.42 | 0.00507 | 0.15671 | 1.011791 | 29.26376 |
| | (7.31) | (1.60E-05) | (6.50E-05) | (0.00559) | (4.04E-06) |

( ):Standard Deviation

In the new iterative approach (Support Vector IV), standard errors were estimated relatively small. Furthermore, RMSE obtained from new approach was smaller than other alternatives.

## 5. Conclusion

In the literature, there are many different approaches in order to make a robust estimation. The GME estimator is a one of the robust estimators resistant to multicollinearity, heteroscedasticity and the existence of outliers. Although it was not necessary to make strict assumptions about parameters or population distributions, the most important point was to specify discrete supports for the coefficients and the error term in this approach, which had a significant effect on the results obtained.

In this study, a new approach based on iterative process was presented for Random Intercept Models. The Classical GME Approach, using wide parameter support boundaries, was mostly suggested in the literature in case the researcher had no prior information. As opposed to this technique, a new iterative approach is suggested in this paper which will help the researcher to obtain consistent parameters without prior information. Furthermore it was adopted multilevel datasets.

The new approach was tested on both simulated and real-life datasets. The obtained results proved that the suggested approach provides better parameter estimates and lowest RMSE than classical methods. The results produced with support vectors suggested in the literature (Support Vector I, II and III) were not closer to the true regression parameters. However, parameter estimations of proposed approach were relatively closer to the true regression parameters.

## References

Abellan, J., Baker, R. M., & Coolen, F. P. A. (2011). Maximising Entropy on the Nonparametric Predictive Inference Model for Multinomial Data. *European Journal of Operational Research*, *212*(1), 112-122. http://dx.doi.org/10.1016/j.ejor.2011.01.020

Akdeniz, F., Çubuk, A., & Güler, H. (2011). Generalized Maximum Entropy Estimators: Applications to the Portland Cement Dataset. *The Open Statistics & Probability Journal*, *3*, 13-20. http://dx.doi.org/10.2174/1876527001103010013

Al-Nasser, A. D. (2014). Two Steps Generalized Maximum Entropy estimation procedure for fitting linear regression when both covariates are subject to error. *Journal of Applied Statistics*, *41*(8), 1708-1720. http://dx.doi.org/10.1080/02664763.2014.888544

Al-Nasser, A. D. (2011). An Information-Theoretic Alternative to Maximum Likelihood Estimation Method in Ultrastructural Measurement Error Model. *Hacettepe University Bulletin of Natural Sciences and Engineering Series B: Mathematics and Statistics*, *40*(3), 469-481.

Al-Rawwash, M. Y., & Al-Nasser, A. D. (2011). Longitudinal Data Analysis Using Generalized Maximum Entropy Approach. *Jordan Journal of Mathematics and Statistics*, *4*(1), 47-60.

Al-Nasser, A. D., Eidous, O. M., & Mohaidat, L. M. (2010). Multilevel Linear Models Analysis using Generalized Maximum Entropy. *Asian Journal of Mathematics and Statistics*, *3*(2), 111-118. http://dx.doi.org/10.3923/ajms.2010.111.118

Al-Nasser, A. D., & Al-Atrash, A. R. (2011). Information Theoretic Approach for constructing a super Data Envelopment Analysis. *Asian Journal on Quality*, *12*(1), 54-66. http://dx.doi.org/10.1108/15982681111140543

Altaylıgil, B. (2008). *Entropi Ölçüsü ve Bazı Ekonometri Uygulamaları (Entropy Measure and Some of Its Econometric Applications)(Doctorate Thesis)*. İstanbul: İstanbul Üniversitesi.

Caputo, M. R., & Paris, Q. (2008). Comparative statics of the generalized maximum entropy estimator of the general linear model. *European Journal of Operational Research*, *185*(1), 195-203. http://dx.doi.org/10.1016/j.ejor.2006.12.031

Ciavolino, E., & Calcagni, A. (2014). A generalized maximum entropy (GME) approach for crisp-input/fuzzy-output regression model. *Quality & Quantity*, *48*(6), 3401-3414. http://dx.doi.org/10.1007/s11135-013-9963-9

Depren, Ö. (2014). *Regresyonda Maksimum Entropi Modellemesi (Maximum Entropy Modelling in Regression) (Doctorate Thesis)*. İstanbul: Marmara Üniversitesi.

Donoso, P., Grange, L. D., & González, F. (2011). A Maximum Entropy Estimator for the Aggregate Hierarchical Logit Model. *Entropy*, *13*(12), 1425-1445. http://dx.doi.org/10.3390/e13081425

Fernandez-Vazquez, E., Mayor-Fernandez, M., & Rodriguez-Valez, J. (2008). Estimating spatial autoregressive models by GME-GCE techniques. *International Regional Science Review*, *32*(2), 148-172. http://dx.doi.org/10.1177/0160017608326600

Gastón, A., & García-Vinas, J. I. (2011). Modelling Species Distributions with Penalised Logistic Regressions: A Comparison With Maximum Entropy Models. *Ecological Modelling*, *222*(13), 2037-2041. http://dx.doi.org/10.1016/j.ecolmodel.2011.04.015

Gelman, A. (2006). Multilevel (Hierarchical) Modeling: What It Can and Cannot Do?. *Technometrics*, *48*(3), 432-435. http://dx.doi.org/10.1198/004017005000000661

Golan, A., & Gzyl, H. (2012). An entropic estimator for linear inverse problems. *Entropy*, *14*(5), 892-923. http://dx.doi.org/10.3390/e14050892

Golan, A., Judge, G. G., & Miller, D. (1996). *Maximum Entropy Econometrics: Robust Estimation with limited data*. New York: John Wiley & Sons.

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review*, *4*(106), 620-630. http://dx.doi.org/10.1103/physrev.106.620

Henderson, H., Golan, A., & Seabold, S. (2015). Incorporating prior information when true priors are unknown: An Information-Theoretic approach for increasing efficiency in estimation. *Economics Letters*, *127*, 1-5. http://dx.doi.org/10.1016/j.econlet.2014.12.014

Kılıç, S., Çene, E., Demir, İ. (2012). Comparison of Learning Strategies for Mathematics Achievement in Turkey with Eight Countries. *Kuram ve Uygulamada Eğitim Bilimleri - Educational Sciences: Theory & Practice*, *12*(4), 2585-2598.

Miller, A. J. (2002). Subset Selection in Regression. London: Chapman&Hall/CRC. http://dx.doi.org/10.1201/9781420035933

Mittelhammer, R., Cardell, N. S., & Marsh, T. L. (2013). The Data-Constrained Generalized Maximum Entropy Estimator of the GLM: Asymptotic Theory and Inference. *Entropy, 15*(5), 1756-1775. http://dx.doi.org/10.3390/e15051756

Organization for Economic Cooperation and Development (OECD). (2009c). *PISA 2006 Technical Report*, OECD, Paris.

Pukelsheim, F. (1994). The Three Sigma Rule. *The American Statistician, 48*(2), 88-91. http://dx.doi.org/10.1080/00031305.1994.10476030

Raudenbush, S. W., Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods. London: Sage Publications.

Raudenbush, S. W., Bryk, A. S., & Condon, R. (2005). HLM: Hierarchical Linear and Non-Linear Modeling. Chicago: Scientific Software International.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*(3), 379–423. http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x

Timm, N. H. (2002). Applied Multivariate Analysis. New York: Springer-Verlag. http://dx.doi.org/10.1007/b98963

## Copyrights

International Journal of Statistics and Probability; Vol. 5, No. 6; November 2016
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

# Convolution Based Unit Root Processes: a Simulation Approach

Fabio Gobbi[1]

[1] Department of Statistics, University of Bologna

Correspondence: Fabio Gobbi, Department of Statistics, University of Bologna. E-mail: fabio.gobbi@unibo.it

Received: August 4, 2016 Accepted: August 22, 2016 Online Published: October 17, 2016

doi:10.5539/ijsp.v5n6p22 URL: http://dx.doi.org/10.5539/ijsp.v5n6p22

**Abstract**

We propose a convolution based approach to the simulation of a modified version of a unit root process where the state variable $Y_{t-1}$ is dependent on the innovation $\varepsilon_t$. The dependence structure is given by a copula function $C$. We study by simulation the effect of a negative correlation on the properties of unit roots. We call this process C-UR(1).

**Keywords:** unit root processes, $C$-convolution, stationarity.

## 1. Introduction

In this paper we present a modified version of a unit root processes using a convolution-based technique. This methodology exploits the properties of copula functions. The application of copula functions to stochastic processes (more in particular to Markov processes) was recently described in the book by Cherubini, Gobbi, Mulinacci and Romagnoli (2012). Our contribution relies on the application of a particular family of copulas, which are generated by the convolution operator, to the design of time series processes. From this point of view, the paper contributes to the literature modeling time series with copulas (Chen & Fan, 2006; Chen, Wu, & Yi, 2009; Cherubini & Gobbi, 2013). While this literature builds on the pioneering paper by Darsow, Nguyen and Olsen (DNO, 1992) on the link between copula functions and Markov processes, our paper exploits the concept of convolution based copulas to define a new version of the unit root process. Beyond the Markov property, there is a long standing and extremely vast literature on the fact that most of the changes of the processes, those that are called innovations, are not predictable on the basis of past information (Samuelson, 1963; 1973; Fama, 1965). In financial markets the natural representation of this concept is to assess that log-prices of assets follow a random walk, which is, in fact, a unit root process. Technically, this process is characterized by innovations that are permanent and independent of the level of the process. The same random walk hypothesis spread into the literature in the field macroeconomics in the 1980s, starting with the seminal paper by Nelson and Plosser (1982). Based on the first unit root tests, due to Dickey-Fuller (1979; 1981), Nelson and Plosser found that most of the US macroeconomic time series included a random walk component, that is a shock, independent and persistent. In this paper we propose an extension to this approach, which allows for dependent innovations, and for non linear dependence between the innovation and the value of the process of the previous period. This is our modified version of the unit root process. The dependence structure is modelled by a copula function and the distribution of the process for all $t$ is obtained by applying the $C$-convolution technique (Cherubini, Mulinacci, & Romagnoli, 2011) as it will be described in section 3. The choice of the family of copulas changes the probabilistic properties of the new process. In order to simplify the computational aspects, in this paper we concentrate on gaussian copulas for which a closed form of the $C$-convolution is available. In this framework, we propose a *C-convolution-based unit root* process, *C-UR(1)*, characterized by a negative correlation between the innovation and the value of the process of the previous period. We investigate the stationarity property of this new process by a simulation experiment.

The plan of the paper is as follows. In section 2 we present the standard linear autoregressive model and the unit root case. In section 3 we introduce our modified version of the unit root process based on the concept of *C-convolution*. In section 4 we describe the simulation algorithm and we discuss the results. Section 5 concludes.

## 2. The Standard AR(1) Process

We begin by describing briefly the property of the celebrate autoregressive process of order 1, AR(1). The definition is the following.

**Definition: 1.** *AR(1). The discrete time stochastic process $(Y_t)_t$ is a first order autoregressive process, AR(1), if*

$$Y_t = \phi Y_{t-1} + \varepsilon_t,$$

*where $\phi$ is a real number and $(\varepsilon_t)_t$ is a sequence of i.i.d. random variables, i.e., $(\varepsilon_t)_t$ is a white noise process. Moreover, $Y_{t-1}$ is independent of $\varepsilon_t$.*

In other words, a stochastic process $Y_t$ is an autoregressive process if the value at the time $t$ depends linearly on its own

previous values and on a stochastic term (a stochastican imperfectly predictable term); thus the model is in the form of a stochastic difference equation. The notation AR(1) indicates an autoregressive model of order 1.
It is well known the constraints on the autoregressive parameter for the model to remain wide-sense stationary. In particular, the process is wide-sense stationary if $|\phi| < 1$ since it is obtained as the output of a stable filter whose input is white noise. Conversely, the condition $|\phi| \geq 1$ identifies the case where the process is not stationary.

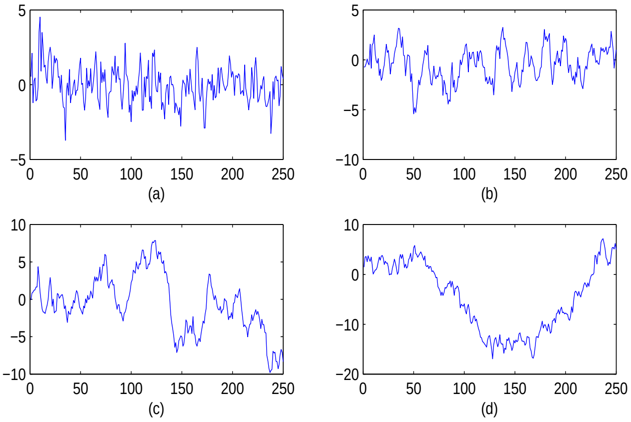


Figure 1. Example of trajectories of a AR(1) process with: (a) $\phi = 0.50$, (b) $\phi = 0.90$, (d) $\phi = 0.95$, (d) $\phi = 0.99$.

Figure 1 displays some examples of trajectories of a stationary AR(1) process for some values of the parameter. The mean-reverting property appear clear from the figure. Furthermore, the absence of any kind of trends is more evident for small values of the autoregressive parameter. Stationarity assures that the mean, $\mu = \mathbb{E}[Y_t]$, and the variance, $V_t^2 = Var(Y_t)$, are constants for all $t$. In particular, it is known that $\mu = \mathbb{E}[Y_t] = 0$ and $V_t^2 = \frac{\sigma_\varepsilon^2}{1-\phi^2}$ (see Hamilton (1994) for more details). The autocovariance function $\gamma_k = \mathbb{E}[(Y_t - \mu)(Y_{t-k} - \mu)]$, $k = 1, 2, ...$ depends only on the lag $k$ and it is given by

$$\gamma_k = \frac{\sigma_\varepsilon^2}{1 - \phi^2}\phi^k,$$

whereas the autocorrelation function (ACF), $\rho_k$, has the form

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \phi^k.$$

Notice that the ACF of a weakly stationary AR(1) process decays exponentially with rate $\phi$. Figures 2 shows the theoretical autocorrelation function for some values of the parameter. If the parameter assumes values close to 1 the decline of the ACF is much slower. For a detailed discussion on autoregressive processes we refer the reader to the manuals of Hamilton (1994) and of Brockwell and Davis (1991).

### *2.1 The Unit Root Case*

In this paper we are particularly interested in the unit root case, i.e., when the autoregressive parameter $\phi = 1$. The definition of a unit root process is the following.

**Definition: 2.** *I(1). The discrete time stochastic process $(Y_t)_t$ is called a unit root process, also known as integrated process, if*

$$Y_t = Y_{t-1} + \varepsilon_t,$$

*where $(\varepsilon_t)_t$ is a white noise process. We denote such a process by I(1). Moreover, $Y_{t-1}$ is independent of $\varepsilon_t$.*

Observe that a unit root process is a random walk. As mentioned in the previous section, since the autoregressive parameter is equal to 1 the I(1) process is not stationary, as we can also infer by observing figure 3 which reports some simulated examples of paths of a unit root process. We can observe that the trajectories are not stationary in their means as we would expect if they were constant over time. As regards the variance, and more in general all higher-order moments, it depends on $t$. In particular, by repeated substitutions, we can write $Y_t = Y_0 + \sum_{j=1}^{t} \varepsilon_j$. Then the variance of $Y_t$, say $V_t^2$, changes

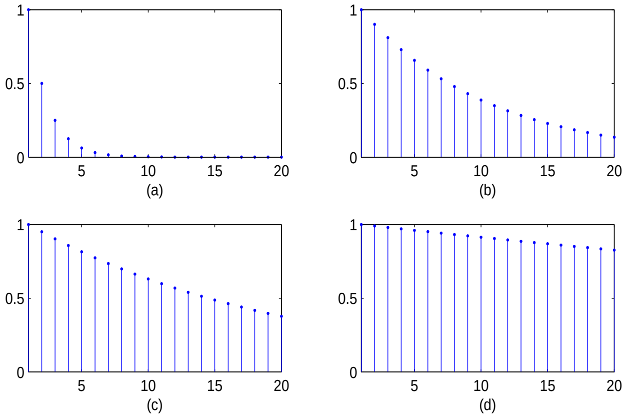


Figure 2. Autocorrelation function of a AR(1) process with: (a) $\phi = 0.50$, (b) $\phi = 0.90$, (d) $\phi = 0.95$, (d) $\phi = 0.99$.

linearly with $t$

$$V_t^2 = \sum_{j=1}^{t} \sigma_\varepsilon^2 = t\sigma_\varepsilon^2,$$

and it approaches to infinity when $t$ tends to infinity.

We can investigate the behavior of the state variable $Y_t$ and of its standard deviation $V_t$ with a Monte Carlo simulation. In particular, we generate 5000 trajectories of 250 points of an I(1) process with initial condition $Y_0 = 0$ with the assumption that $\varepsilon_t$ are i.i.d. $N(0, \sigma_\varepsilon)$; each simulated path $(\tilde{y}_t)_{t=1,\dots,250}$ is a realization of the I(1) process, whereas if we fix $t = t_0$ we have 5000 realizations of the state variable at the time $t_0$, $(\tilde{y}_{t_0}^{(i)})^{i=1,\dots,5000}$.

Figure 4 (panel (a)) reports the estimated probability density function relatively to $(\tilde{y}_t^{(i)})^{i=1,\dots,5000}$ for increasing value of $t$. We see that the dispersion of the distribution of $\tilde{y}_t$ increases as $t$ increases, signalling that the process is not stationary in variance. Moreover, figure 4 (panel (b)) displays the standard deviation, say $\tilde{V}_t$, of a realization $(\tilde{y}_t^{(i)})^{i=1,\dots,5000}$ for increasing values of $t$. As expected, $\tilde{V}_t$ is monotone increasing.

The theoretical autocorrelations of a I(1) process tend to one asymptotically for any lag $k$ but the sample autocorrelations may decline rather fast even with large sample (Hassler, 1994). The average ACF up to the lag $k = 20$ over the 5000 simulated trajectories of our random experiment is reported in figure 5. Clearly, the inspection of this ACF is not sufficient to find out the presence of a unit root. A several test for the presence of unit roots are available in literature (Dickey, 1976; Dickey & Fuller, 1979; 1981).

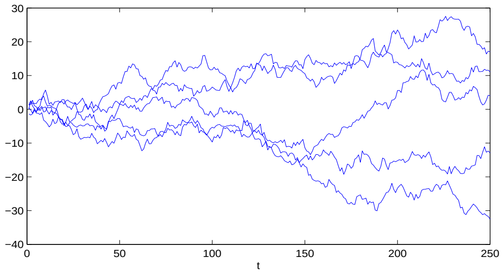


Figure 3. Examples of trajectories of a unit root process.

## 3. The Convolution Based Unit Root Process

In this section we introduce a modified version of the standard unit root process based on the notion of *C-convolution* introduced by Cherubini, Mulinacci and Romagnoli (2011). The $C$-convolution was originally introduced to determine the distribution function of a sum of two dependent and continuous random variables $X$ and $Y$. The dependence structure

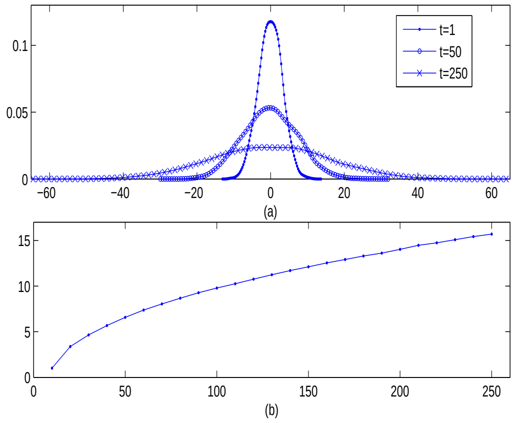


Figure 4. (a) Probability density function of the state variable of a simulated unit root process; (b) Standard deviation of the state variable of a simulated unit root process

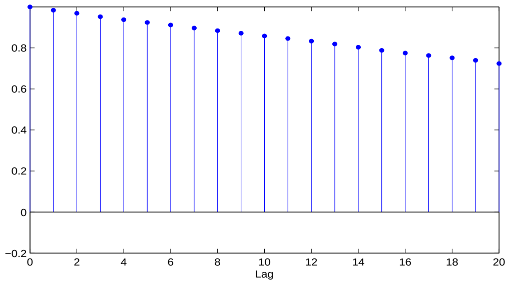


Figure 5. ACF of an I(1) process.

between $X$ and $Y$ is modeled by a *copula function*. The copula technique allows to write every joint distribution as a function of the marginal distributions. In other words, we can represent the joint distribution of $X$ and $Y$, say $\Pr(X \leq a, Y \leq b)$, with $a, b \in \mathbb{R}$ as a function of $F_X(a) \equiv \Pr(X \leq a)$ and $F_Y(a) \equiv \Pr(Y \leq b)$. More formally, there exists a function $C_{X,Y}(u, v)$ such that

$$Pr(X \leq a, Y \leq b) = C_{X,Y}(F_X(a), F_Y(b)). \tag{1}$$

Conversely, given two distribution functions $F_X$ and $F_Y$ and a suitable bivariate function $C_{X,Y}$ we may build joint distribution for $(X, Y)$. The requirements to be met by this function are that: i) it must be grounded ($C(u, 0) = C(0, v) = 0$); ii) it must have uniform marginals ($C(1, v) = v$ and $C(u, 1) = u$); iii) it must be 2-increasing (meaning that the volume $C(u_1, v_1) - C(u_1, v_2) - C(u_2, v_1) + C(u_2, v_2)$ for $u_1 > u_2$ and $v_1 > v_2$ cannot be negative).
The one to one relationship that results between copula functions and joint distributions is known as Sklar theorem. See Nelsen (2006) and Joe (1997) for a detailed discussion on copulas.

The $C$-convolution technique links the marginal distributions of $X$ and $Y$ and their dependence structure given by a copula so as to determine the probability distribution of the sum $X + Y$. The seminal paper is that of Cherubini, Mulinacci & Romagnoli (2011) where we may find the concept of *convolution-based copulas*. If $X$ e $Y$ be two real-valued random variables with corresponding copula $C_{X,Y}$ and continuous marginals $F_X$ and $F_Y$, then the distribution function of the sum $X + Y$, denoted by $F_X \overset{C}{*} F_Y$, is given by

$$F_{X+Y}(z) = (F_X \overset{C}{*} F_Y)(z) = \int_0^1 D_1 C_{X,Y}\left(w, F_Y(z - F_X^{-1}(w))\right) dw, \tag{2}$$

where $D_1 C_{X,Y}(u, v)$ denotes $\frac{\partial C_{X,Y}(u,v)}{\partial u}$.
The choice of the copula function affects the probabilistic behavior of the distribution of the sum (for a detailed discussion on this topic see the book of Cherubini, Gobbi, Mulinacci & Romagnoli (2012). Some of the most used copula functions

are the Gaussian copula, the Clayton copula, the Frank copula and the Gumbel copula. The Gaussian copula is constructed from a bivariate normal distribution over $\mathbb{R}^2$ by using the probability integral transform. For a given correlation coefficient $\rho$, the Gaussian copula with parameter $\rho$ can be written as

$$C(u, v; \rho) = \Phi_2\left(\Phi^{-1}(u), \Phi^{-1}(v)\right),$$

where $\Phi_2$ is the bivariate standard normal distribution with correlation coefficient $\rho$ and $\Phi$ is the standard normal distribution. The Clayton copula is an asymmetric Archimedean copula, exhibiting greater dependence in the negative tail than in the positive. Its functional form is given by

$$C(u, v; \theta) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta},$$

where $\theta$ is the parameter which assumes positive values, $\theta \in (0, +\infty)$. The Frank copula is a symmetric copula defined as

$$C(u, v; \theta) = -\ln\left(1 + \frac{(\exp(-\theta u) - 1)(\exp(-\theta v) - 1)}{\exp(-\theta) - 1}\right),$$

where the parameter $\theta$ is a real number, $\theta \in \mathbb{R}$. The Gumbel copula is an asymmetric archimedean copula, exhibiting greater dependence in the positive tail than in the negative. This copula is given by:

$$C(u, v; \theta) = \exp\left(-\left((-\ln u)^{\theta} + (-\ln v)^{\theta}\right)^{1/\theta}\right),$$

where $\theta \in [1, +\infty)$. It is important to notice that the $C$-convolution has a closed form if and only if the marginal distributions are gaussian and the copula linking them is the gaussian copula (see Cherubini, Gobbi, Mulinacci & Romagnoli, 2012). For computational purposes in this paper we only consider that case. The reader can find some examples of $C$-convolution with Clayton and Frank copulas in the book of Cherubini, Gobbi, Mulinacci & Romagnoli (2012).
Here, we are interested in how to use the $C$-convolution to modelling stochastic processes. As shown in Cherubini, Gobbi & Mulinacci (2016) we can construct a dependent increments Markov processes by a repeated application of the $C$-convolution technique. More precisely, given a stochastic process $(Y_t)_t$, let $Y_{t-1}$ with marginal distribution $F_{t-1}$ and $\Delta Y_t = Y_t - Y_{t-1}$ with distribution $H_t$. Moreover let $C$ be the copula associated to $(Y_{t-1}, \Delta Y_t)$. Then, we may recover the distribution of $Y_t = Y_{t-1} + \Delta Y_t$ iterating the $C$-convolution (2) for all $t$

$$F_t(y) = (F_{t-1} \overset{C}{*} H_t)(y) = \int_0^1 D_1 C\left(w, H_t(y - F_{t-1}^{-1}(w))\right) dw. \tag{3}$$

The process $(Y_t)_t$ is called the $C$-convolution based process. This methodology may be applied to define a new version of the unit root process I(1), $Y_t = Y_{t-1} + \varepsilon_t$, when $Y_{t-1}$ and $\varepsilon_t$ are not independent as in the standard case but linked by some copula $C$. This is our modified version of a I(1) process.

Notice that if the copula $C$ is the independent copula, that is $C(u, v) = uv$, the $C$-convolution coincides with the standard convolution and we obtain the standard I(1) process. In this section we consider a *C-convolution-based unit root process, C-UR(1)*, by imposing a dependence structure between $Y_{t-1}$ and $\varepsilon_t$. The distribution of $Y_t = Y_{t-1} + \varepsilon_t$ is given by the $C$-convolution between the distribution of $Y_{t-1}$, $F_{t-1}$, and the distribution of $\varepsilon_t$, $H_t$. Suppose that $Y_0$ has distribution $F_0$. Then, the distribution of $Y_t$ is

$$F_t(y_t) = (F_{t-1} \overset{C}{*} H_t)(y_t) = \int_0^1 D_1 C(w, H_t(y_t - F_{t-1}^{-1}(w)))dw, \quad t = 1, 2, ... \tag{4}$$

We are now ready to introduce the definition of our modified version of a I(1) process. In particular we have the following

**Definition: 3.** *C-UR(1). The discrete time stochastic process $(Y_t)_t$ is a C-convolution based unit root process, C-UR(1), if*

- *the functional form is that of a unit root process, $Y_t = Y_{t-1} + \varepsilon_t$;*
- *there exists a dependence structure between the state variable at the time $t-1$, $Y_{t-1}$, and the innovation $\varepsilon_t$. Moreover, this dependence structure is described by a copula function, C, with a time-invariant parameter.*

***Remark 1.*** Patton (2005; 2006) introduced the notion of conditional copulas that allows us to define a time-varying dependence structure. In other words, the parameter of the copula function depends on the time $t$ while remaining within the same family of copulas.

### *3.1 The Gaussian C-UR(1) Process*

As mentioned before, if $C$ is not the gaussian copula the above integral (4) cannot be expressed in closed form and they have to be evaluated numerically. In those cases the simulation of a C-UR(1) process may not be simple. However, since the gaussian family is closed under $C$-convolution, we can perform a simulation design under some restrictions. In particular, suppose that the following conditions hold.

1. The initial distribution is gaussian:
$$Y_0 \sim N(0, \sigma_0),$$
and the distribution of innovations is gaussian and stationary
$$\varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma_\varepsilon).$$

2. The copula function linking $Y_{t-1}$ and $\varepsilon_t$ is gaussian and stationary, i.e., $C(u, v) = G(u, v; \rho)$, where $\rho$ is the correlation coefficient.

Under this framework, as shown in Cherubini, Gobbi & Mulinacci (2016), by iterating the $C$-convolution technique (4) we recover the distribution of the state variable for all $t$

$$Y_t \sim N(0, V_t^2), \tag{5}$$

where the variance $V_t^2$ has a closed functional form

$$V_t^2 = Var(Y_t) = V_1^2 + (t-1)\sigma_\varepsilon^2 + 2\rho\sigma_\varepsilon \sum_{i=1}^{t-1} V_{t-i}, \quad t = 1, ... \tag{6}$$

Moreover, it is shown that the copula between two consecutive state variables, $Y_{t-1}$ and $Y_t$, is Gaussian with parameters

$$\rho_{X_{t-1},X_t} = \frac{V_{t-1} + \rho\sigma_\varepsilon}{V_t}, \quad t = 1, ...,$$

where $V_1 = \sigma_1$.

Furthermore, we can prove that the behavior of $V_t^2$ when $t \to +\infty$ depends on the correlation coefficient $\rho$. As shown in Cherubini, Gobbi & Mulinacci (2016) the limiting behavior of the standard deviation $V_t$ is

$$V_t \longrightarrow \begin{cases} -\frac{\sigma_\varepsilon}{2\rho}, & \text{if } \rho \in (-1, 0); \\ +\infty, & \text{otherwise.} \end{cases} \tag{7}$$

It is very significant to notice that the standard deviation of a C-UR(1) process does not grow indefinitely as $t$ tends to infinity only in the case of negative correlation between $Y_{t-1}$ and $\varepsilon_t$. Even more significant is the fact that the sufficient condition is $\rho < 0$.

## 4. Simulation Design

The simulation of a gaussian C-UR(1) process is based on the *conditional sampling* technique which allows to generate random pairs $(u, v)$ from a given family of copulas (Nelsen, 2006 and Cherubini, Luciano & Vecchiato, 2004). The method is based on the property that if $(U, V)$ are $U(0, 1)$ distributed r.vs. whose joint distribution is given by $C$ the conditional distribution of $V$ given $U = u$ is the first partial derivative of $C$

$$\mathbb{P}(V \leq v | U = u) = D_1 C(u, v) = c_u(v),$$

which is a non-decreasing function of $v$. With this result in mind the simulation of a pair $(u, v)$ from $C$ is obtained in the following two steps. We call the following algorithm **Alg1**.

1. Generate two independent r.vs. $(u, z)$ from a $U(0, 1)$ distribution: $(u, z) \overset{i.i.d.}{\sim} U(0, 1)$.

2. Compute $v = c_u^{-1}(z)$, where $c_u^{-1}(\cdot)$ is the quasi-inverse function of the first partial derivative of the copula.

3. $(u, v)$ is the desired pair.

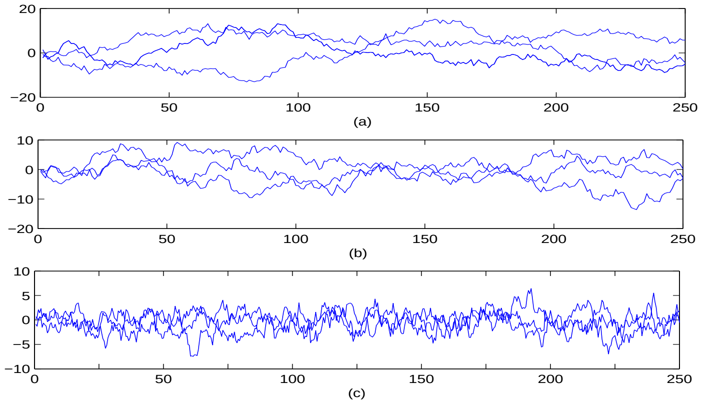


Figure 6. Example of trajectories of a C-UR(1) process. (a) $\rho = -5\%$; (b) $\rho = -10\%$; (c) $\rho = -25\%$.

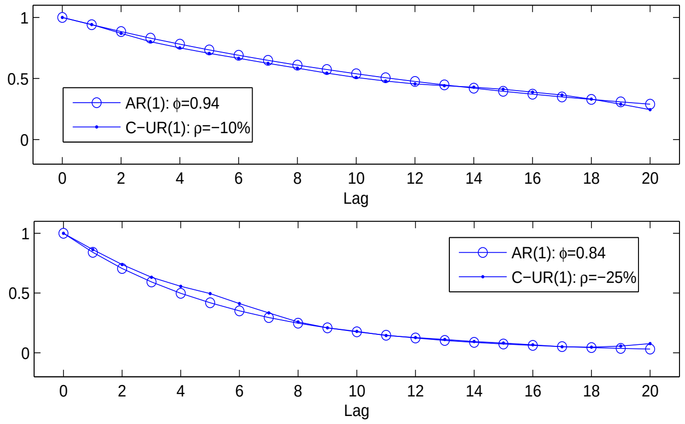


Figure 7. Comparison among autocorrelation functions of a standard AR(1) process and a C-UR(1) process with $\rho = -10\%$ (top) and $\rho = -25\%$ (bottom).

Now, we propose our algorithm to simulate trajectories from a C-UR(1) process using **Alg1**. The input is given by a sequence of distributions of innovations, $\varepsilon_t$, that for the sake of simplicity we assume stationary $H_t = H$ and gaussian: $H \sim N(0, \sigma_\varepsilon)$. Moreover we assume a dependence structure stationary and gaussian, $C_{Y_{t-1},\varepsilon_t}(u, v) = G(u, v; \rho)$, We also assume $Y_0 = 0$. We describe a procedure to generate a iteration of a $n$-step trajectory.

1. Generate $u$ from the uniform distribution.

2. Compute $\tilde{y}_t = H^{-1}(u)$ with $t = 1$.

3. Use **Alg1** to generate $v$ from $G(u, v; \rho)$.

4. Compute $\tilde{\varepsilon}_{t+1} = H^{-1}(v)$.

5. $\tilde{y}_{t+1} = \tilde{y}_t + \tilde{\varepsilon}_{t+1}$.

6. Compute the distribution $F_{t+1}$ by $C$-convolution given by equation (4).

7. Compute $u = F_{t+1}(\tilde{y}_{t+1})$.

8. Repeat steps 4-7 with $t = 2, 3, ..., n - 1$.

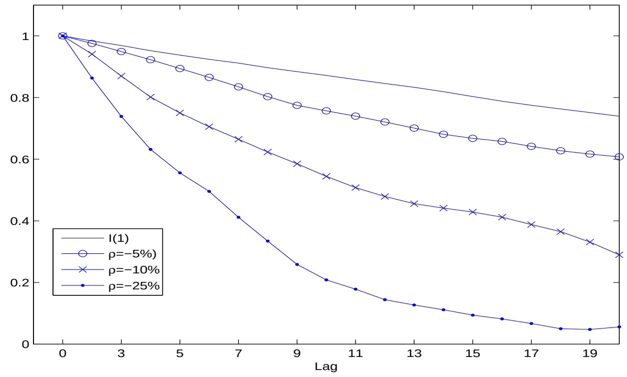


Figure 8. Comparison among autocorrelation functions.

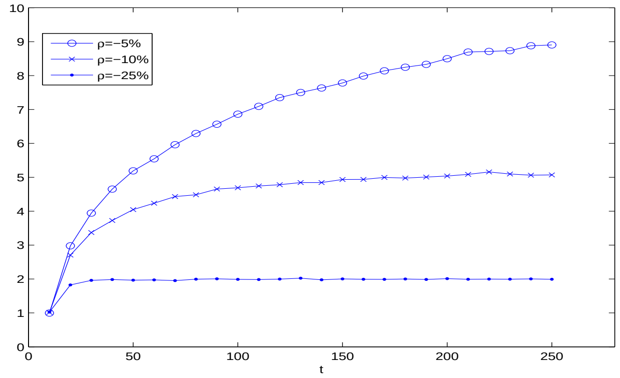


Figure 9. Standard deviation of the state variable of a simulated C-UR(1) process for three different values of the correlation coefficient.

We generate 5000 trajectories of 250 points. We can think of daily trajectories therefore 250 points refer to a calendar year. The number of trajectories has been chosen according to the computational resources available. Without loss of generality we set $\sigma_\varepsilon = 1$ and we select three different levels of negative correlation: $\rho = -5\%$, $\rho = -10\%$ and $\rho = -25\%$.

### *4.1 Results*

We now describe the results of our simulations. Figure 6 shows some simulated trajectories for each level of correlation. We can observe that as the correlation increases in absolute value the dynamics of trajectories appears more stationary both in mean and in variance. If we compare this figure with figure 3 the effect is even more clear. We notice a mean reverting effect which becomes stronger as the negative correlation increases. If we consider the autocorrelations the behavior of our C-UR(1) is also interesting. Table 1 reports the autocorrelation function for the first 20 lags for a I(1) process and for our C-UR(1) process with correlation level from -3% to -25%. The impact of negative correlation is clear. If in the case of low negative correlation the decline of autocorrelations is in fact identical (with $\rho = -3\%$) or very similar (with $\rho = -5\%$) to that of a I(1) process, when $\rho$ is -10% or -25% the situation drastically changes. In particular, a negative correlation greater than -20% virtually eliminates serial correlation while being in the presence of a unit root. For example, in the case of $\rho = -10\%$ autocorrelations are very close to those of a standard AR(1) process with autoregressive parameter $\phi$ around 0.94 whereas in the case of $\rho = -25\%$ are very similar to those of a standard AR(1) process with $\phi$ around 0.84 as we can see in figure 7. Figure 8 compares the dynamics of autocorrelations of a I(1) process with those of a C-UR(1) process with negative correlation from -5% to -25%. As regards the variance of the state variable, figures 9 and 10 show the impact of the negative correlation. More precisely, figure 9 reports the behavior of the standard deviation $V_t$ as a function of the time $t$. The convergence towards a constant level (given by equation 7) is faster as the negative correlation increases. If $\rho = -25\%$ the convergence to the limit value is immediate and the variance of the state variable is constant over time as in stationary processes. Figure 10 emphasizes this aspect showing that the dispersion is the same for

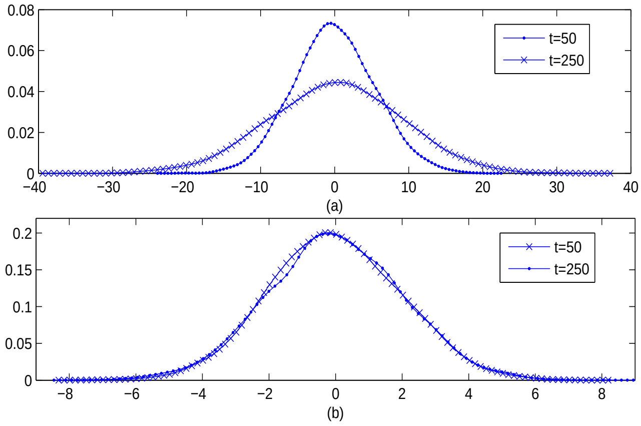


Figure 10. Panel (a). Probability density function of the state variable of a simulated C-UR(1) process with $\rho = -10\%$. Panel (b). Probability density function of the state variable of a simulated C-UR(1) process with $\rho = -25\%$.

both the instants of time considered if $\rho = -25\%$. The linear relationship between the time $t$ and the variance disappears.

Table 1. Comparison among autocorrelation values for different choices of the correlation coefficient.

| Lag | I(1) | $\rho = -3\%$ | $\rho = -5\%$ | $\rho = -10\%$ | $\rho = -25\%$ |
|---|---|---|---|---|---|
| 1 | 0.9836 | 0.9817 | 0.9756 | 0.9410 | 0.8633 |
| 2 | 0.9689 | 0.9637 | 0.9493 | 0.8699 | 0.7388 |
| 3 | 0.9520 | 0.9487 | 0.9227 | 0.8015 | 0.6317 |
| 4 | 0.9377 | 0.9373 | 0.8943 | 0.7502 | 0.5557 |
| 5 | 0.9241 | 0.9227 | 0.8653 | 0.7056 | 0.4955 |
| 6 | 0.9118 | 0.9096 | 0.8348 | 0.6649 | 0.4115 |
| 7 | 0.8970 | 0.8956 | 0.8030 | 0.6235 | 0.3342 |
| 8 | 0.8841 | 0.8813 | 0.7748 | 0.5851 | 0.2584 |
| 9 | 0.8718 | 0.8705 | 0.7569 | 0.5443 | 0.2084 |
| 10 | 0.8583 | 0.8589 | 0.7397 | 0.5079 | 0.1784 |
| 11 | 0.8456 | 0.8459 | 0.7208 | 0.4790 | 0.1443 |
| 12 | 0.8332 | 0.8321 | 0.7009 | 0.4556 | 0.1271 |
| 13 | 0.8190 | 0.8164 | 0.6807 | 0.4415 | 0.1115 |
| 14 | 0.8034 | 0.8022 | 0.6675 | 0.4284 | 0.0941 |
| 15 | 0.7882 | 0.7899 | 0.6576 | 0.4118 | 0.0818 |
| 16 | 0.7750 | 0.7788 | 0.6417 | 0.3879 | 0.0668 |
| 17 | 0.7630 | 0.7661 | 0.6276 | 0.3651 | 0.0500 |
| 18 | 0.7515 | 0.7548 | 0.6166 | 0.3311 | 0.0476 |
| 19 | 0.7395 | 0.7420 | 0.6075 | 0.2898 | 0.0558 |
| 20 | 0.7240 | 0.7261 | 0.5953 | 0.2449 | 0.0766 |

## 5. Conclusion

In this paper we propose a convolution based approach to the simulation of a modified version of a unit root process which we called $C$-convolution-based unit root process, C-UR(1). The idea is that once the distribution of innovations is specified, and the dependence structure between innovations and levels of the process is chosen, the distribution of the process can be automatically recovered. The variance of this new process converges to a constant level and this convergence is faster as the correlation becomes more negative. The autocorrelation function rapidly decay towards zero as soon as the correlation is around -20%. For these reasons, the model is well suited to address problems of persistent and unpredictable shocks, beyond the standard paradigm of linear models.

## References

Andrews, D. W. K. (1988). Law of Large Numbers for dependent non-identically distributed random variables. *Econo-*

*metric Theory, 4*, 458-467. http://dx.doi.org/10.1017/S0266466600013396

Brockwell, P. J., & Davis, R. A. (1991). *Time Series*. Theory and Methods, Springer Series in Statistics. Springer-Verlag. http://dx.doi.org/10.1007/978-1-4419-0320-4

Chen, X., & Fan, Y. (2006). Estimation of copula-based semiparametric time series models. *Journal of Econometrics, 130*, 307-335. http://dx.doi.org/10.1016/j.jeconom.2005.03.004

Chen, X., Wu, W. B., & Yi, Y. (2009). Efficient Estimation of Copula-Based semiparametric Markov models. *Annals of Statistics, 37*(6B), 4214-4253. http://dx.doi.org/10.1214/09-AOS719

Cherubini, U., & Gobbi, F. (2013). A Convolution-based Autoregressive Process, in F. Durante, W. Haerdle, P. Jaworski editors. *Workshop on Copula in Mathematics and Quantitative Finance*. Lecture Notes in Statistics-Proceedings. Springer, Berlin/Heidelberg. http://dx.doi.org/10.1007/978-3-642-35407-6_1

Cherubini, U., Luciano E., & Vecchiato, W. (2004). *Copula Methods in Finance*. London, John Wiley, (John Wiley Series in Finance). http://dx.doi.org/10.1002/9781118673331

Cherubini, U., Gobbi, F., & Mulinacci, S. (2016). *Convolution Copula Econometrics*. SpringerBriefs in Statistics.

Cherubini, U., Mulinacci, S., & Romagnoli, S. (2011). A Copula-based Model of Speculative Price Dynamics in Discrete Time, *Journal of Multivariate Analysis, 102*, 1047-1063. http://dx.doi.org/10.1016/j.jmva.2011.02.004

Cherubini U., Gobbi F., Mulinacci S., & Romagnoli S. (2012). *Dynamic Copula Methods in Finance*. John Wiley & Sons.

Darsow, W. F., Nguyen, B., & Olsen, E. T. (1992). Copulas and Markov Processes, *Illinois Journal of Mathematics, 36*(4).

Dickey, D. A. (1976). *Estimation and hypothesis testing in non-stationary time series*. (Unpublished doctoral dissertation). Iowa State University, Ames, IA.

Dickey, D. A., & 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.

Dickey, D. A., & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica, 49*, 1057-1072. http://dx.doi.org/10.2307/1912517

Fama, E. F. (1965). Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance, 25*(2), 383-417. http://dx.doi.org/10.2307/2325486

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hassler, U. (1994). The sample autocorrelation function of I(1) processes. *Statistical Papers, 35*, 1-16. http://dx.doi.org/10.1007/BF02926395

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall, London. http://dx.doi.org/10.1201/b13150

Nelsen, R.(2006). *An Introduction to Copulas*. Springer.

Nelson, C., & Plosser, C. (1982). Trend and Randon Walk in Macroeconomic Time Series. *Journal of Monetary Economics, 10*, 139-169. http://dx.doi.org/10.1016/0304-3932(82)90012-5

Patton, A. (2005). Modelling time-varying exchange rate dependence. *International Economic Review, 47*.

Patton, A. (2006). Estimation of multivariate models for time series of possibly different lengths. *Journal of Applied Econometrics, 21*, 147-173. http://dx.doi.org/10.1002/jae.865

Samuelson, P. A. R. (1963). Proof That Properly Anticipated Prices Fluctuate Randomly. *Industrial Management Review, 6*, 41-50.

Samuelson, P. A. R. (1973). Mathematics of Speculative Price. *SIAM Review, 15(1)*, 1-42. http://dx.doi.org/10.1137/1015001

## Copyrights

International Journal of Statistics and Probability; Vol. 5, No. 6; November 2016
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

# Linear Hybrid Deterministic Dynamic Modeling for Time-to-Event Processes: State and Parameter Estimations

E. A. Appiah[1] & G. S. Ladde[1]

[1] Department of Mathematics and Statistics, University of South Florida, Tampa, FL, USA.

Correspondence: G.S. Ladde, Department of Mathematics and Statistics, University of South Florida, 4202 East Fowler Avenue, CMC 342, Tampa, FL 33620-5700, USA. E-mail: gladde@usf.edu

Received: August 22, 2016 Accepted: September 18, 2016 Online Published: October 17, 2016

doi:10.5539/ijsp.v5n6p32 URL: http://dx.doi.org/10.5539/ijsp.v5n6p32

## Abstract

In this work, we initiate an innovative alternative modeling approach for time-to-event dynamic processes. The proposed approach is composed of the following basic components: (1) development of continuous-time state of dynamic process, (2) introduction of discrete-time dynamic intervention process, (3) formulation of continuous and discrete-time interconnected dynamic system, (4) utilizing Euler-type discretized schemes, and (5) introduction of conceptual and computational state and parameter estimation procedures. The presented approach is motivated by state and parameter estimation of time-to-event processes in biological, chemical, engineering, epidemiological, medical, military, multiple-markets and social dynamic processes under the influence of discrete-time intervention processes. The role and scope of our approach is exhibited by presenting several well-known hazard/risk rate and survival function estimates as special cases. Moreover, conceptual algorithms are illustrated by time-series data sets under the influence of intervention processes.

**Keywords:** Kaplan-Meier estimator, hazard/risk rate function, piecewise exponential estimator, time-to-event closed process, totally discrete-time hybrid system

## 1. Introduction

In the survival and reliability data analysis, the main interest is focused on a nonnegative random variable, say $T$ which describes a time-to-event process characterizing an occurrence of time until a certain event. Historically well-known time-to-event processes are deaths in population dynamic and component failures in mechanical systems (Kalbfleisch & Prentice, 2011). The human mobility, electronic communications, technological changes, advancements in engineering, medical, and social sciences have diversified the role and scope of time-to-event processes in cultural, epidemiological, financial, military and social sciences (Ladde, 2015; Chandra & Ladde, 2014; Ladde & Ladde, 2012; Wanduku & Ladde, 2011; Anis, 2009).

The study of survival analysis rests on the concept of time-to-event. The mathematical statistics development of time-to-event analysis is based on the probabilistic approach and the concept of hazard rate. Moreover, the time-to-event is described by the closed form expressions of survival function that is determined by the concept of hazard rate (Kalbfleisch & Prentice, 2011; Lawless, 2011; Miller, 2011). We note that in general, hazard rate is unknown. This leads to a problem of determining hazard rate function. This is based on a feasible approach of collecting data set for the time-to-event processes in biological, chemical, engineering, epidemiological, medical, multiple-markets and social sciences. The hazard/risk rate and survival function estimation problems in the survival and reliability analysis are centered around the idea of “right censored data” (Miller, 2011). In fact, the common conventional understanding for resolving ties between censored and uncensored observations is adopted by shifting the censored observations slightly to the left of uncensored observations (Whittemore & Keller, 1983). In short, the items/individuals/objects in a given sample are decomposed into two mutually exclusive groups, namely, (a) deaths/failure /removal/non-operational/inactive, and (b) censored/losses/ withdrawals.

In the survival and reliability data analysis, parametric and nonparametric methods are applied to estimate the hazard/risk rate and survival functions (Kalbfleisch & Prentice, 2011; Lawless, 2011). A parametric approach is based on the assumption that the underlying survival distribution belongs to some specific family of distributions (e.g. normal, Weibull, exponential). On the other hand, a nonparametric approach is centered around the best-fitting member of a class of survival distribution functions (Kaplan & Meier, 1958). Moreover, Kaplan-Meier(KME) (Kaplan & Meier, 1958) and Nelson-Aalen (Aalen, 1978; Nelson, 1969) type nonparametric approach do not assume neither distribution class, nor closed-form distributions. In fact, it just depends on a data. The Kaplan-Meier and Nelson-Aalen type nonparametric estimation approaches are systematically analyzed by our totally discrete-time hybrid dynamic modeling process.

In the existing literature (Kalbfleisch & Prentice, 2011; Lawless, 2011), the closed-form expression for a survival function is based on the usage of probabilistic analysis approach. The closed-form representation of the survival function coupled with mathematical statistics method (parametric approach) is used to estimate both survival and hazard/risk rate functions. In fact, the parametric approach/model has advantages of simplicity, the availability of likelihood based inference procedures and the ease of use for a description, comparison, prediction, or decision (Lawless, 2011). In this work, we initiate an innovative alternative approach for modeling time-to-event dynamic processes. This approach leads to the development for estimating survival and hazard/risk rate functions. The presented approach is motivated by a simple observation regarding the probabilistic definition of the survival function (Kalbfleisch & Prentice, 2002). Moreover, this approach does not require a knowledge of either a closed-form solution distribution or a class of distributions.

Historically, exponential distributions have been widely used in analyzing survival/reliability data (Lawless, 2011; Davis, 1952). This was partly due to the mathematical simplicity and the availability of simple statistical methods. An application of the exponential model with covariates to medical survival data was initiated in Feigl and Zelen (1965). The assumption of a constant hazard/risk rate function is very restrictive. In fact, it is often violated. This is due to the fact that in some real life applications, sudden changes in the hazard rate at unknown times can be encountered due to a major maintenance in a mechanical system or a new treatment procedure in medical sciences (Anis, 2009). For example, usually a machine component functions with a constant hazard/risk rate function $\lambda_1$, until it suffers a shock. After this shock, the component may continue to operate but with a different constant hazard/risk rate function $\lambda_2$. In the medical field, there is usually a high initial risk after a major operation which settles down to a lower constant long-term risk rate (Anis, 2009). This type of change could occur in multiple times. In view of this, one is often interested in detecting the locations of such changes and estimating the sizes of the detected changes. Recently, several authors (Han, Schell & Kim 2014; He & Su, 2013; Fang & Su, 2011, Goodman, Li, & Tiwari, 2011) have proposed estimators based on change point hazard models. A Bayesian approach for estimating the piecewise exponential distribution (Gamerman, 1994) and estimating the grid of time-points (Demarqui, Loschi, & Colosimo, 2008) for the piecewise exponential model are also available in the literature. In order to incorporate these types of sudden changes (intervention process) in the hazard rate function, we modify the developed continuous state dynamic model to an interconnected hybrid dynamic model that is composed of both continuous time state and discrete time state (intervention process) dynamic processes.

Employing the total time on test (TTT) for undefined censored data beyond the last observation, the idea of Piecewise Exponential Estimator (PEXE) of a survival function was introduced by (Kitchin, Langberg, & Proschan, 1980) and applied for estimating life distribution from incomplete data. The PEXE has been modified to address the issues regarding the presence of ties in the data by Whittemore and Keller (1983).

The comparison of the PEXE with the KME (Kim & Proschan, 1991) exhibits the advantage of the PEXE over the KME. For example, the PEXE is a continuous survival function. Moreover, it exhibits the complete information that is coming from the censored data. Using a total time test and the PEXE based approach, the estimators of the hazard/risk rate and cumulative distribution functions on the left closed pairwise consecutive failure time intervals are determined in Kulasekera and White (1996). The PEXE is further extended by Malla and Mukerjee (2010) with an exponential tail extension in the framework of the Kaplan and Meier (1958) nonparametric estimator approach. Under the presented dynamic framework, we develop the PEXE and new PEXE of Malla and Mukerjee (2010) types in a systematic and unified way. In short, the presented novel approach incorporates all the existing features such as: incomplete data, issues regarding the ties, exponential tail extensions in the framework of Kaplan and Meier (1958), and so on in a coherent manner.

The organization of the presented work is as follows. In Section 2, recognizing the classical probabilistic analysis model of time-to-event as a dynamic process, we initiate a linear hybrid deterministic dynamic model for time-to-event processes. Moreover, a fundamental mathematical result that provides a basis for interconnected continuous-discrete-time and totally discrete-time dynamic processes, is developed. Utilizing the dynamic model and the main result developed in Section 2, basic conceptual analytic algorithms and its special cases for interconnected continuous-discrete-time and totally discrete-time linear hybrid dynamic models for time-to-event processes are presented in Section 3. In Section 4, we outline conceptual computational schemes. In Section 5, we present a very general conceptual and computational algorithm for estimating a hazard/risk rate function for multiple censoring times between consecutive failure times. These general results include the presented results in Section 4 as special cases. In Section 6, conceptual computational and simulation algorithms are developed. The developed computational schemes are applied to estimate hazard/risk rate and survival functions in a systematic and unified way. Moreover, several well-known results are exhibited as special cases. A few conclusions are drawn in Section 7 to exhibit the role and scope of linear hybrid deterministic modeling for time-to-event processes. Moreover, further extensions and generalizations to both deterministic and stochastic nonlinear and non-stationary hybrid modeling for time-to-event processes are currently underway. In addition, currently, a complex time-to-event dynamic analysis is also undertaken by the authors. These results will appear elsewhere. Finally, proofs of theorems and corollaries in Sections 2, 3, 4 and 5 are outlined in supplementary Section 8.

## 2. Linear Hybrid Dynamic Modeling of Time-to-event Process

In this section, based on the probabilistic definition of the survival function, we develop a model for time-to-event dynamic processes. From the probabilistic definition of the survival function (Kalbfleisch & Prentice, 2011; Lawless, 2011; Miller, 2011) and differential calculus (Apostol, 1967), we recognize that

$$\lambda(t)\Delta t \approx \frac{S(t) - S(t+\Delta t)}{S(t)}\,, \tag{1}$$

where $S$ and $\lambda$ are survival and hazard/risk rate functions, respectively. Moreover, from (1) and differential calculus (Apostol, 1976), we have

$$\mathrm{d}S = -\lambda(t)S\,\mathrm{d}t\,, \quad S(t_0) = S_0\,, \quad t \in [t_0, \infty)\,, \tag{2}$$

where d$S$ is a differential of a survival function $S$. In fact, (2) is a differential equation, and it is an initial value problem (IVP) (Ladde & Ladde, 2012). Based on continuous-time dynamic modeling (Ladde & Ladde, 2012), (2) represents a continuous-time linear dynamic model of time-to-event processes. In fact, we consider time-to-event processes to be probabilistic dynamic processes. The state of the process is represented by survival/infective/operational/radical and its complementary state, failure/removal/death/non-operational/normal, and it is measured by a probability distribution function. Employing Newtonian modeling approach, the instantaneous rate of change of survival state is directly proportional to the magnitude of the survival. The negative sign in (2) signifies that the state of survival is decaying/diminishing/decreasing. $\lambda$ is a positive constant of proportionality. In general, it is a function of time. This is because of the fact that in general, the time-to-event processes are non-stationary. The solution of (2) on the interval $[t_0, \infty)$ is given by

$$S(t) = S_0 \exp\left[-\Lambda(t)\right]\,, \tag{3}$$

where

$$\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u\,, \tag{4}$$

and it is the cumulative hazard/risk rate function.

**Remark 2.1.** If $\lambda(t) = \lambda$ for $t \geq 0$, $t_0 = 0$, $S(0) = 1$, then (3) reduces to the following well-known exponential distribution function:

$$S(t) = \exp[-\lambda t]\,, \quad t \in [0, \infty)\,, \tag{5}$$

and a complementary state of the survival state of time-to-event process is represented by

$$F(t) = 1 - S(t) = 1 - \exp[-\lambda t]\,, \quad t \in [0, \infty)\,,$$

and it is referred as a failure distribution function. Furthermore, we note that survival state dynamic model (2) signifies that the time-to-event process is closed (Rosen, 1970), that is, $S(t) + F(t) = 1$. It is analogous to epidemiological dynamic modeling process without removal (Ladde & Ladde, 2012; Wanduku & Ladde, 2011).

The presented motivational observation coupled with the introduction of the idea of continuous-time state dynamic process (2) operating under the discrete-time intervention processes further leads to a development of a linear hybrid dynamic model (Ladde & Ladde, 2012) for time-to-event processes. It is known (Ladde & Ladde, 2012) that many real world time-to-event dynamic processes are subject to intervention processes (internal or external). Therefore, it is natural that time-to-event dynamic processes undergo state adjustment processes. This causes a modification of the presented state dynamic processes that are described by simple state dynamic model (2). We note that the dynamic state adjustment processes are caused by periodic changes in science, technology, medicine, culture, socio-economic, environmental conditions and general behavior.

In the following, we introduce a type of hazard/risk rate function. Moreover, using dynamic approach, we present a development of PEXE (Kitchin et al., 1980; Kim & Proschan, 1991) in a systematic and unified way.

**Definition 2.1.** Let $\tau_0 < \tau_1 < \tau_2 < \ldots < \tau_k < \tau_{k+1}$ be a given partition of a time interval $[\tau_0, \mathscr{T}]$, with $\tau_0 = 0$ and $\tau_{k+1} = \infty$. Let $\lambda_1, \lambda_2, \ldots, \lambda_{k+1}$ be model parameters. A hazard/risk rate function for a nonnegative random variable $T$ that characterizes time-to-event processes, is of the following form:

$$\lambda(t) = \sum_{i=1}^{k+1} = \lambda_j I_{[\tau_{j-1},\tau_j)}(t)\,, \quad t \in \mathbf{R}_+ = [0,\infty)\,, \tag{6}$$

where $\lambda_j$ are positive real numbers for $j \in I(1,k+1)$, ($I(1,l) = \{1,2,\ldots,l\}$); $I_{[\tau_{j-1},\tau_j)}$ is the characteristic function with respect to $[\tau_{j-1},\tau_j)$. Moreover, $T$ is said to have a piecewise constant hazard function.

**Definition 2.2.** $\prod\limits_{i|\tau_j\leq t}$ denotes the symbol for a product of objects for all positive integers $i \in I(1,\infty)$ that satisfy the conditions $\tau_i \leq \tau_j$ and $\tau_j \leq t < \tau_{j+1}$ for some $j \in I(1,n)$ and for $\tau_i, \tau_{j-1}, \tau_{j+1}, t \in [\tau_0, \mathscr{T}]$.

From Definition 2.1, we recognize that the sudden changes in the hazard/risk rate function are encountered due to various types of intervention processes (internal or external) (Ladde & Ladde, 2012). This causes to interrupt the current continuous-time state dynamic process (2). Following the linear hybrid dynamic model (Ladde & Ladde, 2012), a modified version of time-to-event dynamic model (2) is represented by:

$$\begin{cases} \mathrm{d}S = -\lambda(t)S\,\mathrm{d}t\,, & S(\tau_{j-1}) = S_{j-1}\,, \quad t \in [\tau_{j-1},\tau_j)\,, \\ S_j = S(\tau_j^-,\tau_{j-1},S_{j-1})\,, & S(\tau_0) = S_0\,, \quad j \in I(1,k+1)\,, \end{cases} \tag{7}$$

where $S(\tau_j^-) = S(\tau_j^-|\lambda,\tau_{j-1},S_{j-1})$ describes a very simpler form of intervention process generated at an intervention time $\tau_j$; $\tau_j^-$ stands for $t \in [\tau_{j-1},\tau_j)$, that is less than $\tau_j$ and very close to $\tau_j$. We note that system (7) is interconnected hybrid dynamic system composed of both continuous and discrete time state dynamic systems. Imitating the procedure described in Ladde and Ladde (2012), the solution process of the IVP (7) is as follows:

$$S(t,\tau_{j-1},S_{j-1}|\lambda) = S_{j-1}\exp\left[-\int_{\tau_{j-1}}^{t}\lambda(u)\mathrm{d}u\right]\,, \text{ for all } t \in [\tau_{j-1},\tau_j)\,. \tag{8}$$

Furthermore, the solution process of the overall time-to-event dynamic process (7) on $[\tau_0,\mathscr{T})$ is

$$S(t,\tau_{j-1},S_0|\lambda) = S_0\prod_{m=1}^{j-1}\exp\left[-\int_{\tau_{m-1}}^{\tau_m}\lambda(u)\mathrm{d}u\right]\exp\left[-\int_{\tau_{j-1}}^{t}\lambda(u)\mathrm{d}u\right],\ t \in [\tau_0,\mathscr{T})\ , j \in I(1,k+1)\,. \tag{9}$$

**Remark 2.2.** From (7) and (8), we note that the solution process (8) is indeed PEXE (Kitchin et al., 1980; Kim & Proschan, 1991).

In the following, we present a very simple fundamental auxiliary result that would be used, subsequently. Moreover, it exhibits an analytic unified bridge and basis for (7) and its complete discrete-time version.

**Theorem 2.1.** *Let $\{\tau_j\}_0^n$ be a partition of $[0,\mathscr{T}]$ and let $\beta$ be a monotonic nondecreasing function defined by*

$$\beta(t) = \begin{cases} 0, & t \in [\tau_{j-1},\tau_j)\,, \\ 1, & t = \tau_j\,, \end{cases} \tag{10}$$

*for each $j \in I(1,n)$. Let $x$ be a state dynamic process in biological, engineering, epidemiological, human, medical, military, physical and social sciences under the influence of time-to-event processes. Let $x$ be described by:*

$$\begin{cases} \mathrm{d}x = [-\alpha(t)\,x + \gamma(t)]\,\mathrm{d}\beta(t), & t \in [\tau_{j-1},\tau_j)\,, \\ x_j = (1-\alpha_j)x(\tau_j^-,\tau_{j-1},x_{j-1}) + \gamma_j\,, & x(\tau_0) = x_0\,, \end{cases} \tag{11}$$

*where $\alpha$ and $\gamma$ are real-valued continuous functions defined on $[0,\infty)$; $\alpha_j = \alpha(\tau_j)$ and $\gamma_j = \gamma(\tau_j)$. Then*

$$x(t) = \prod_{k|\tau_j\leq t}(1-\alpha_k)x_0 + \sum_{i=1}^{j-1}\Phi(t,\tau_i)\gamma_i + \gamma_j\,, \quad \textit{for} \quad t \geq \tau_0\,, \tag{12}$$

*where $j$ is the largest integer so that $\tau_j \leq t < \tau_{j+1}$, $\tau_k \leq \tau_j$ and*

$$\Phi(t,\tau_i) = \prod_{\tau_i\leq\tau_j\leq t}(1-\alpha_i)\,, \quad \Phi(\tau_i,\tau_i) = 1 \quad \text{for} \quad i \in I(0,n)\,.$$

*Proof.* The proof of Theorem 2.1 is given in the supplementary Section 8.

**Remark 2.3.** From (10), the hybrid dynamic system (11), is equivalent to the hybrid dynamic system

$$\begin{cases} \mathrm{d}x = 0\,\mathrm{d}t\,, \quad x(\tau_{j-1}) = x_{j-1}\,, \quad t \in [\tau_{j-1}, \tau_i)\,, \\ x_j = (1-\alpha_j)x(\tau_j^-, \tau_{j-1}, x_{j-1}) + \gamma_j\,, \quad x(\tau_0) = x_0\,, \end{cases} \tag{13}$$

for $j \in I(1,n)$. The solution process of (13) is represented in (12).

In the following, we present a couple of special cases of Theorem 2.1. These special cases illustrate a systematic way for exhibiting the existing results in Kaplan and Meier (1958), Nelson (1969), Aalen (1978) and Malla and Mukerjee (2010) in the framework of presented innovative dynamic approach.

**Corollary 2.1.** *If functions $\alpha$ and $\gamma$ in Theorem 2.1 are replaced by functions $\lambda$ and $\gamma = 0$, then (12) reduces to*

$$x(t) = \prod_{j|\tau_j \le t} (1-\lambda_j)x_0\,, \quad t \ge \tau_0\,. \tag{14}$$

**Corollary 2.2.** *If $\alpha = 0$ and $x_0 = 0$ in Theorem 2.1, then the conclusion of Theorem 2.1 reduces to*

$$x(t) = \sum_{i|\tau_{j-1} \le t} \gamma_i\,, \quad t \ge \tau_0 \quad \text{and} \quad t \in [\tau_{j-1}, \tau_j)\,. \tag{15}$$

In the following, we present a definition of cumulative jump process (Malla & Mukerjee, 2010) in the framework of hybrid dynamic model.

**Example 2.1.** Let $T_1, T_2, \ldots, T_n$ be discrete failure times for the discrete-time event process, and $0 = a_0 < a_1 \le a_2 \le \ldots \le a_m$ be jumps of a survival function in magnitude. Then the dynamic for the cumulative jump process is as described in Corollary 2.2, and its solution process is exhibited in (15).
In this example, applying Corollary 2.2 in the context of $\gamma_0 = 0$, $\gamma_i = a_i$, the cumulative jump process is represented by

$$x(t) = \begin{cases} A_{j-1} = \sum\limits_{i=1}^{j-1} a_i\,, \quad \text{for} \quad t \in [\tau_{j-1}, \tau_j)\,, \\ A_j = \sum\limits_{i=1}^{j} a_i\,, \quad t = \tau_j\,. \end{cases} \tag{16}$$

From (16), we recognize that the cumulative jump defined in Malla and Mukerjee (2010) is indeed recast as the discrete time intervention process described by the hybrid dynamic system illustrated in Corollary 2.2 at the discrete time $\tau_j$ for $j \in I(1,m)$ with $\gamma_0 = a_0 = 0$ and $\gamma_i = a_i$.

**Example 2.2.** Under the conditions of Example 2.1, the magnitude of the survival function at the failure times is represented by

$$S(t) = \begin{cases} 1 - A_{j-1}\,, \quad \text{for} \quad t \in [\tau_{j-1}, \tau_j)\,, \\ 1 - A_j\,, \quad t = \tau_j, \quad j \in I(1,m)\,, \end{cases} \tag{17}$$

where $\gamma_0 = 1$ and $x(\tau_j) = A_j$. The $S(t)$ in (17) is the magnitude of the survival function determined by the cumulative jump (Malla & Mukerjee, 2010) process described in Example 2.1.

**Remark 2.4.** We remark that the continuous-time dynamic model can be exhibited by the cumulative hazard/risk rate function. In fact, from (2), we have

$$\mathrm{d}\ln S = -\lambda(t)\mathrm{d}t\,, \quad \ln S(\tau_0) = S_0\,. \tag{18}$$

Based on the solution processes of (2) and (7), the solution process of (18) can be represented as:

$$-\ln\left[\frac{S(t)}{S(\tau_0)}\right] = \Lambda(t, \tau_0, S_0|\lambda) = \int_{\tau_0}^{t} \lambda(u)\mathrm{d}u\,. \tag{19}$$

and

$$-\ln\left[\frac{S(t)}{S(\tau_0)}\right] = \Lambda(t, \tau_0|\lambda) = \sum_{m=1}^{j-1} \int_{\tau_{m-1}}^{\tau_m} \lambda(u)\mathrm{d}u + \int_{\tau_{j-1}}^{t} \lambda(u)\mathrm{d}u\,,\ t \in [\tau_{j-1}, \tau_j)\,. \tag{20}$$

respectively. Furthermore, we set $x = \ln S$ , $S_0 = 1$ and $\gamma(t) = -\lambda(t)$ where $S$ and $\lambda$ are defined in (18). From Corollary 2.2, we have

$$\ln S(t) = -\Lambda(t)\,, \tag{21}$$

where $\Lambda(t) = \sum_{i|\tau_i \le t} \lambda_i$ is a cumulative hazard function.

**Remark 2.5.** We remark that if $x$ is replaced by survival function, $S$ in Corollary 2.1, and $x$ and $\gamma$ are replaced by $S$ and $\lambda$ in Corollary 2.2, then (14) and (15) are replaced by:

$$S(t) = \prod_{j|\tau_j \le t} (1 - \lambda_j) S_0\,, \quad t \ge \tau_0 \tag{22}$$

and

$$S(t) = \sum_{i|\tau_i \le t} \lambda_i\,, \quad t \ge \tau_0\,, \tag{23}$$

respectively. Moreover, (22) is the solution process of the discrete-time dynamic system described by Corollary 2.1. Furthermore, dynamic system outlined in Corollary 2.1 provides an innovative alternative approach for finding the discrete-time survival function (Kaplan & Meier, 1958) in a systematic manner.

We utilize the above presented concepts and results in subsequent sections in a systematic and unified way.

## 3. Fundamental Results for Continuous and Discrete-Time to Event Dynamic Processes

In this section, we utilize hybrid dynamic model (7) and fundamental analytic Theorem 2.1 for time-to-event process to develop a general fundamental result. The developed result provides basic analytic and computational tools for estimating survival state and parameters. The presented approach also provides a systematic and unified way of estimating the parameters and survival functions.

Let $x(t)$ be the total number of units/individuals operating/alive (or survivals) at time $t$, for $t \in [\tau_0, \mathscr{T}]$. It is described by (11). Let $\lambda$ and $S$ be hazard/risk rate and survival functions of the units/patients/infectives/species/individuals, respectively. Employing a dynamic model for number of units/species/ individuals coupled with survival state dynamic model (2) or (7), we present an interconnected hybrid dynamic model below.

Following the argument used in developing dynamic models (Ladde & Ladde, 2012), we introduce the following interconnected system of differential equations:

$$\begin{cases} \mathrm{d}S = -\lambda(t) S\,\mathrm{d}t\,, \quad t \in [\tau_{j-1}, \tau_j)\,, \\ S_j = (1 - \beta_j) S(\tau_j^-, \tau_{j-1}, S_{j-1})\,, \quad S(\tau_0) = 1\,, \\ dx = (-\alpha(t)x + \gamma(t))\mathrm{d}\beta(t)\,, \quad x(\tau_0) = x_0\,, \quad t \in [\tau_{j-1}, \tau_j)\,, \\ x_j = (1 - \alpha_j) x(\tau_j^-, \tau_{j-1}, x_{j-1}) + \gamma_j\,, \end{cases} \tag{24}$$

**Remark 3.1.** We outline a few important observations that exhibit the role and scope of dynamic approach to illustrate the existing results (Han et al., 2014; Kim & Proschan, 1991; Thaler, 1984; Kitchin et al., 1980; Kaplan & Meier, 1958) as special cases.

(i) Dynamic system (24) in the context of (13) (Remark 2.3) is reduced to

$$\begin{cases} \mathrm{d}S = -\lambda(t) S\,\mathrm{d}t\,, \quad t \in [\tau_{j-1}, \tau_j)\,, \\ S_j = (1 - \beta_j) S(\tau_j^-, \tau_{j-1}, S_{j-1})\,, \quad S(\tau_0) = 1\,, \\ dx = 0\,\mathrm{d}t\,, \quad x(\tau_0) = x_0\,, \quad t \in [\tau_{j-1}, \tau_j)\,, \\ x_j = (1 - \alpha_j) x(\tau_j^-, \tau_{j-1}, x_{j-1}) + \gamma_j\ . \end{cases} \tag{25}$$

(ii) From Corollary 2.1 in the context of Remark 2.5, in particular (22), system (24) becomes:

$$\begin{cases} \mathrm{d}S = 0\,\mathrm{d}t\,, \quad t \in [\tau_{j-1}, \tau_j)\,, \\ S_j = (1 - \lambda_j) S_{j-1}\,, \\ \mathrm{d}x = 0\,\mathrm{d}t\,, \quad x(\tau_0) = x_0\,, \\ x_j = (1 - \alpha_j) x_{j-1} + \gamma_j\ . \end{cases} \tag{26}$$

We note that (26) is a special version of (24). In addition, we refer to system (26) as a totally discrete-time hybrid dynamic system.

Now, we are ready to present a basic result regarding continuous and discrete time interconnected dynamic of survival species or objects or thoughts operating under the time-to-event intervention processes. Prior to the formulation of the fundamental result, we introduce a concept of number of survivals.

**Definition 3.1.** Let $z$ be a function defied by $z(t) = x(t)S(t)$, where $S$ and $x$ are solution process of (24) for $t \in [\tau_0, \mathscr{T}]$. Moreover, for each $t \in [\tau_0, \mathscr{T}]$, $z(t)$ stands for the number of survivals at $t$ under an influence of time-to-event process.

**Theorem 3.1.** *Let $(x, S)$ be a solution process of (24). Then the interconnected hybrid dynamic population model for time-to-event process (24) and corresponding intervention iterative process are described by:*

$$\begin{cases} \mathrm{d}z = -\lambda(t)z\mathrm{d}t\,, \quad z(\tau_{j-1}) = z_{j-1}\,, \quad \textit{for} \quad t \in [\tau_{j-1}, \tau_j)\,, \quad j \in I(1,k)\,, \\ z(\tau_j) = (1-\alpha_j)(1-\beta_j)z(\tau_j^-) + \gamma_j(1-\beta_j)\,, \end{cases} \tag{27}$$

*and*

$$z(\tau_j) = (1-\lambda(\tau_j)\Delta\tau_j)(1-\alpha_j)(1-\beta_j)z(\tau_{j-1}) + \gamma_j(1-\beta_j)\,. \tag{28}$$

*respectively, where $z$ is defined in Definition 3.1 and $\Delta\tau_j = \tau_j - \tau_{j-1}$ for $j \in I(1,k)$.*

*Proof.* For the detailed proof of Theorem 3.1, the readers are encouraged to read the supplementary Section 8.

In the following, we present a few special/trivial cases that exhibit existing results in the framework of hybrid dynamic of time-to-event interconnected system.

**Corollary 3.1.** *Let us consider a very special/trivial case of Theorem 3.1 as follows:*

$$\begin{cases} \mathrm{d}S = -\lambda(t)S\,\mathrm{d}t\,, \quad t \geq \tau_0\,, \\ \mathrm{d}x = 0\,\mathrm{d}t\,, \quad t \geq \tau_0\,, \\ x(\tau_j) = x(\tau_j^-, \tau_{j-1}, x_{j-1})\,, \quad x(\tau_0) = x_0\,, \quad j \in I(1,k)\,. \end{cases} \tag{29}$$

*Applying Theorem 3.1 and using (27) and (28), (29) reduces to*

$$\begin{cases} \mathrm{d}z = -\lambda(t)z\mathrm{d}t\,, \quad z(\tau_{j-1}) = z_{j-1}\,, \quad t \in [\tau_{j-1}, \tau_j)\,, \\ z(\tau_j) = z(\tau_j^-, \tau_{j-1}, z_{j-1}) = z(\tau_{j-1})\,, \quad j \in I(1,k)\,, \end{cases} \tag{30}$$

*and*

$$z(\tau_j) = \left(1 - \lambda(\tau_j)\Delta\tau_j\right)z(\tau_{j-1})\,. \tag{31}$$

**Corollary 3.2.** *Let us consider a special case of (24) as follows:*

$$\begin{cases} \mathrm{d}S = -\lambda(t)S\,\mathrm{d}t\,, \quad S(\tau_{j-1}) = S_{j-1}\,, \quad t \in [\tau_{j-1}, \tau_j)\,, \\ S(\tau_j) = S(\tau_j^-, \tau_{j-1}, S_{j-1})\,, \end{cases} \tag{32}$$

where $a_j$ is defined in Example 2.1. Then applying Euler-type discretization scheme (Atkinson, 2008) on $[\tau_{j-1}, \tau_j^-]$, yields

$$S(\tau_j^-) - S(\tau_{j-1}) = -\lambda(\tau_{j-1})\Delta\tau_j S(\tau_{j-1})\,. \tag{33}$$

Moreover, from (32) and (33), we have

$$S(\tau_j) - S(\tau_{j-1}) = -\lambda(\tau_j)\Delta\tau_j S(\tau_{j-1})\,. \tag{34}$$

**Corollary 3.3.** *Under the assumptions of Theorem 3.1 in the context of Remark 3.1(ii), (26) becomes:*

$$\begin{cases} \mathrm{d}z = 0\,\mathrm{d}t\,, \quad z(\tau_{j-1}) = z_{j-1}\,, \quad t \in [\tau_{j-1}, \tau_j)\,, \\ z(\tau_j) = (1-\lambda_j)(1-\alpha_j)z_{j-1} + \gamma_j\,, \end{cases} \tag{35}$$

*and*

$$z(\tau_j) = (1-\lambda_j)(1-\alpha_j)z(\tau_{j-1}) + \gamma_j\,. \tag{36}$$

This corollary is indeed a totally discrete-time version of hybrid dynamic system operating under discrete-time intervention process.

Using Definition 3.1 and the discrete-time iterative process (28), we introduce a couple of definitions.

**Definition 3.2.** Let $\tau_{j-1}$ and $\tau_j$ be a pair of consecutive observation times belonging to $[0, \mathscr{T}]$. $z(\tau_{j-1})$ stands for the number of survivals at the time $\tau_{j-1}$ for each $j \in I(1, k)$. Moreover, $z(\tau_{j-1})$ is the number of survivals under observation over the sub-interval of time $[\tau_{j-1}, \tau_j)$. $z(\tau_{j-1})\Delta\tau_j$ is the amount of time spent under observation/testing/evaluation by $z(\tau_{j-1})$ survivals over the length $\Delta\tau_j$ of time interval $[\tau_{j-1}, \tau_j)$.

**Definition 3.3.** For $j \in I(1, k)$, $z(\tau_{j-1}) - z(\tau_j)$ stands for the change in number of survivals over the interval of time $[\tau_{j-1}, \tau_j]$ of length $\Delta\tau_j$.

**Remark 3.2.** The discrete-time processes (28), (31), (34) and (36) are referred as our numerical schemes with respect to interconnected hybrid dynamic models for a survival population dynamic processes. Moreover, from (28), we will introduce three more special numerical schemes, namely, time-to-event: (i) failure/death/removal/infective, (ii) censored/withdrawn, and (iii) admission/joining/susceptible/relapsed processes. We further note that the presented numerical schemes allow "ties" with deaths/failure or censored/quiting process. In addition, the population under the presented observation/-supervision process includes the patient/objects population as a special case.

(i) For each $j \in I(1, k)$, let us assume that either $\tau_{j-1}$ and $\tau_j$ are consecutive failure/death/removal/infective times of individual/machine/species, or $\tau_{j-1}$ and $\tau_j$ are censored and failure times, respectively. For $\alpha_j = \gamma_j = \beta_j = 0$, the numerical scheme (28) for failure/death/removal/infective/etc process data set is described by

$$z(\tau_j) = (1 - \lambda(\tau_j)\Delta\tau_j)z(\tau_{j-1})\,, \tag{37}$$

and hence

$$z(\tau_j) - z(\tau_{j-1}) = -\lambda(\tau_j)z(\tau_{j-1})\Delta\tau_j\,, \tag{38}$$

where $\tau_{j-1}$ is either the failure or censored time.

Moreover, $\alpha_j = \gamma_j = \beta_j = 0$ in (28) coupled with (94) is equivalent to the Kaplan and Meier (1958) assumption, namely,

$$x(\tau_j^-) - x(\tau_j) = \text{the number of deaths at } \tau_j\,.$$

That is

$$z(\tau_{j-1}) - z(\tau_j^-) = 0 \quad \text{and} \quad z(\tau_j) = z(\tau_j^+)\,.$$

This implies that $z(t)$ is left discontinuous and right continuous at $\tau_j$.

(ii) Let us assume that either $\tau_{j-1}$ and $\tau_j$ are consecutive censored times, or $\tau_{j-1}$ and $\tau_j$ are failure and censored times, respectively. For $\alpha_j = \beta_j = 0$, and $\gamma_j^c$ stands for the number of censored objects/infectives/etc at a time $\tau_j$. The numerical scheme (28) for censored/listed/identified process data set is described by

$$z(\tau_j) = \left(1 - \lambda(\tau_j)\Delta\tau_j\right)z(\tau_{j-1}) - \gamma_j^c\,, \tag{39}$$

where $\tau_{j-1}$ is either a failure or censored time.

Thus

$$z(\tau_j) - z(\tau_{j-1}) = -\lambda(\tau_j)z(\tau_{j-1})\Delta\tau_j - \gamma_j^c \tag{40}$$

Again, we note that $\alpha_j = \beta_j = 0, \gamma_j^c$, in the context of (94) is equivalent to the Kaplan and Meier (1958) assumption, namely,

$$z(\tau_j) = z(\tau_j^-) \quad \text{and} \quad z(\tau_j) - z(\tau_j^+) = \gamma_j^c\,.$$

This implies that $z(t)$ is left continuous and right discontinuous at $\tau_j$.

(iii) Let us assume that $\tau_{j-1}$ is either failure or censored time, and $\tau_j$ is a joining/admitting/relapsing time. For $\alpha_j = 0$ and $\gamma_j^a$ denoting the number of objects/infectives that joined the observation process at time $\tau_j$. The numerical scheme (28) for admission/joining/sustainable/recruiting/relapsing process is

$$z(\tau_j) = \left(1 - \lambda(\tau_j)\Delta\tau_j\right)z(\tau_{j-1}) + \gamma_j^a\,. \tag{41}$$

The scheme determined by $\alpha_j = 0$ in (28) with (94) and the addition $\gamma_j^a$ in (41) is equivalent to $z(\tau_j) - z(\tau_j^-) = \gamma_j^a$ and $z(\tau_j) = z(\tau_j^+)$.

(iv) Remarks (i), (ii) and (iii) remain valid for the iterative processes (28), (31) and (36).

(I) For $\alpha_j = 0 = \beta_j = \gamma_j$ in (28), (34) reduces to (38); for $\alpha_j = 0 = \beta_j = \gamma_j$, (36) reduces to $z(\tau_j) = (1-\lambda_j)z(\tau_{j-1})$.

(II) For $\alpha_j = 0 = \beta_j$ and $\gamma_j = -\gamma_j^c$ in (28), (28) reduces to (40); for $\alpha_j = 0 = \lambda_j$ and $\gamma_j = -\gamma_j^c$, (36) becomes

$$z(\tau_j) - z(\tau_{j-1}) = (1-\lambda_j)z(\tau_{j-1}) - \gamma_j^c\,. \tag{42}$$

(III) For $\alpha_j = 0 = \beta_j$ and $\gamma_j = \gamma_j^a$ in (28), and $\alpha_j = 0 = \lambda_j$ and $\gamma_j = \gamma_j^a$ in (36), (28) reduces to (41), and (36) reduces to

$$z(\tau_j) - z(\tau_{j-1}) = (1-\lambda_j)z(\tau_{j-1}) + \gamma_j^a. \tag{43}$$

## 4. Estimations of Risk Rate and Survival Functions

Now, we are ready to find an estimate for the hazard/risk rate and survival functions for interconnected continuous and discrete-time survival state dynamic processes. For the sake of completeness and clarity, we first introduce a couple of definitions.

**Definition 4.1.** For $j \in I(1,k)$, let $\tau_{j-1}$ and $\tau_j$ be consecutive change times under continuous-time state survival dynamic process. The parameter estimate at $\tau_j$ is defined by the quotient of change of objects over the consecutive time change interval $[\tau_{j-1}, \tau_j)$ and the total time spent by the objects under observation over the time interval of length $\Delta\tau_j$.

**Definition 4.2.** For $j \in I(1,k)$, let $\tau_{j-1}$ and $\tau_j$ be consecutive change times for discrete-time state survival dynamic process. The parameter estimate at $\tau_j$ is defined by the quotient of the change in the number of survival state over the consecutive time change interval $[\tau_{j-1}, \tau_j)$ and the number of objects at the immediate past time, that is, either the change time or the censored time.

**Remark 4.1.** We observe that the Definitions 4.1 and 4.2 are consistent with each other. This statement can be justified in the context of discrete-time iterative scheme (95) and the continuous and discrete-time hybrid-type descriptions of survival state dynamic model (25) and totally discrete-time hybrid dynamic system (26).

Now, we are ready to present a main result regarding parameter and survival state estimation problems. This result includes several existing results as special cases. In the following, we simply state a conceptual computational algorithm. The detailed proof is given in the supplementary section.

**Theorem 4.1.** *Let us assume that the conditions of Theorem 3.1 in the context of Remarks 3.1 and 3.2(i),(ii) are satisfied.*

*(a) For $j \in I(1,k)$, if $\tau_{j-1}$ and $\tau_j$ are consecutive risk/failure/removal/death/non-operational times in $[\tau_0, \mathscr{T}]$ then an estimate for the hazard/risk rate function at $\tau_j$ is determined by:*

$$\hat{\lambda}(\tau_j) = \frac{z(\tau_{j-1}) - z(\tau_j)}{z(\tau_{j-1})\Delta\tau_j}\,, \tag{44}$$

*and an estimate for the hazard/risk rate function is*

$$\hat{\lambda}(t) = \hat{\lambda}(\tau_j)\,, \quad \textit{for} \quad t \in [\tau_{j-1}, \tau_j) \quad \textit{and} \quad j \in I(1,k)\,. \tag{45}$$

*(b) For $j \in I(1,k)$, if $\tau_{j-1} < \tau_j^c < \tau_j$, and $\tau_j^c$ is censored time between a pair of consecutive failure times $\tau_{j-1}$ and $\tau_j$ in $[\tau_0, \mathscr{T})$, then,*

*(i) a change in the number of items/subjects/thoughts that are under observation over the subinterval $[\tau_{j-1}, \tau_j)$ of the time interval of study $[\tau_0, \mathscr{T}]$ is*

$$z(\tau_{j-1}) - z(\tau_j) - \gamma_j^c\,. \tag{46}$$

*(ii) a total amount of time spent under the observation/testing/evaluation of $z(\tau_{j-1}) - z(\tau_j) - \gamma_j^c$ items/patients/infectives/radicals/subjects over the time interval $[\tau_{j-1}, \tau_j)$ is*

$$z(\tau_{j-1})\Delta\tau_j^c + z(\tau_j^c)\Delta\tau_{jc}\,, \quad \Delta\tau_{jc} = \tau_j - \tau_j^c. \tag{47}$$

(*iii*) *an estimate for the hazard/risk rate function at* $\tau_j$ *is defined as:*

$$\hat{\lambda}(\tau_j) = \frac{z(\tau_{j-1}) - z(\tau_j) - \gamma_j^c}{z(\tau_{j-1})\Delta\tau_j^c + z(\tau_j^c)\Delta\tau_{jc}}\,, \tag{48}$$

*and an estimate for the hazard/risk rate function is*

$$\hat{\lambda}(t) = \hat{\lambda}(\tau_j)\,, \quad \textit{for} \quad t \in [\tau_{j-1}, \tau_j) \quad \textit{and} \quad j \in I(1,k)\,. \tag{49}$$

(*iv*) *Moreover, an estimate for the survival function in* (24) *is*

$$\hat{S}(t) = S_0 \exp\left[\sum_{m=1}^{j-1} \hat{\lambda}_m(\tau_m - \tau_{m-1}) + \hat{\lambda}_j\left(t - \tau_{j-1}\right)\right]\,,\ t \in [\tau_{j-1}, \tau_j). \tag{50}$$

**Remark 4.2.** We note that if $\tau_j^c = \tau_j$ in Theorem 4.1(b), then we have "ties" between censored and failure times. In this case, $\Delta\tau_j^c = \Delta\tau_j$ and $\Delta\tau_{jc} = 0$. From this, (47) and (48) reduce to

$$z(\tau_{j-1})\Delta\tau_j\,, \tag{51}$$

and

$$\hat{\lambda}(\tau_j) = \frac{z(\tau_{j-1}) - z(\tau_j) - \gamma_j^c}{z(\tau_{j-1})\Delta\tau_j} \quad \text{for} \quad j \in I(1,k)\,. \tag{52}$$

This observation justifies Remark 3.2 regarding the mixed "ties."

In the following, we exhibit the role and scope of Theorem 4.1. This is achieved by presenting the well-known hazard/risk rate and survival functions as special cases.

**Corollary 4.1.** *Let us assume that conditions of Corollary 3.3 in the context of Remark 3.2(iv)(I) and (II) are satisfied.*

(*a*) *For* $j \in I(1,k)$, *if* $\tau_{j-1}$ *and* $\tau_j$ *are consecutive risk/failure times in* $[\tau_0, \mathscr{T}]$, *then employing Remark 3.2(iv)(I) and Definitions 3.2, 3.3 and 4.2, an estimate for the risk/hazard rate function at* $\tau_j$ *is determined by:*

$$\hat{\lambda}(\tau_j) = \frac{z(\tau_{j-1}) - z(\tau_j)}{z(\tau_{j-1})}\,, \tag{53}$$

*and*

$$\lambda(t) = \hat{\lambda}(\tau_j)\,, \quad t \in [\tau_{j-1}, \tau_j)\,. \tag{54}$$

*Substituting (53) into (22), an estimate for the survival function is obtained as:*

$$\begin{aligned} S(t) &= \prod_{i|\tau_{j-1} \le t} \left(1 - \hat{\lambda}_i\right) = \prod_{i|\tau_{j-1} \le t} \left(1 - \frac{z(\tau_{i-1}) - z(\tau_i)}{z(\tau_{i-1})}\right) \\ &= \prod_{i|\tau_{j-1} \le t} \left(1 - \frac{d_i}{z(\tau_{i-1})}\right)\,, \quad t \ge \tau_0\,, \end{aligned} \tag{55}$$

*where* $d_i = z(t_{i-1}) - z(\tau_i)$ *is the number of deaths over the consecutive risk/failure time interval* $[\tau_{i-1}, \tau_i)$, $\tau_i \le \tau_{j-1} \le t < \tau_j$ *for some* $j \in I(1,k)$.

(*b*) *For* $j \in I(1,k)$, *if* $\tau_{j-1} < \tau_j^c < \tau_j$, *and* $\tau_j^c$ *is censored time between a pair of consecutive risk/failure times* $\tau_{j-1}$ *and* $\tau_j$ *in* $[\tau_0, \mathscr{T})$, *then, employing Remark 3.2(iv)(II) and Definitions 3.2, 3.3 and 4.2, an estimate for the risk/hazard rate function at* $\tau_j$ *is determined by:*

$$\hat{\lambda}(\tau_j) = \frac{z(\tau_{j-1}) - z(\tau_j) - \gamma_j^c}{z(\tau_j^c)}\,, \tag{56}$$

*and*

$$\lambda(t) = \hat{\lambda}(\tau_j)\,, \quad t \in [\tau_{j-1}, \tau_j)\,. \tag{57}$$

*Substituting (56) into (22), an estimate for the survival function when $\tau_j^c$ is a censored time between consecutive failure times, $\tau_{j-1}$ and $\tau_j$ is given by:*

$$S(t) = \prod_{i|\tau_{j-1}\leq t}\left(1-\lambda_i\right) = \prod_{i|\tau_{j-1}\leq t}\left(1-\frac{z(\tau_{i-1})-z(\tau_i)-\gamma_i^c}{z(\tau_i^c)}\right)$$
$$= \prod_{i|\tau_{j-1}\leq t}\left(1-\frac{d_i}{z(\tau_i^c)}\right), \quad t \geq \tau_0\,, \tag{58}$$

*where i runs over the positive integers for which $\tau_i \leq \tau_{j-1}$, $\tau_{j-1} \leq t < t$ for some $j \in I(1,k)$; $\tau_{i-1}, \tau_i$ are consecutive failure times for $i \in I(1, j)$, and $d_i = z(t_{i-1}) - z(\tau_i) - \gamma_i^c$ is the number of deaths over the consecutive failure time interval $[\tau_{j-1}, \tau_j)$.*

**Remark 4.3.** (a) We remark that (55) and (58) are indeed the Kaplan and Meier (1958)-type survival estimate functions.

(b) In the literature (Kalbfleisch & Prentice, 2011; Lawless, 2011), the numbers in the denominator of (55) and (58) are referred to as the number of individuals at rist at $\tau_{j-1}$ and $\tau_j^c$ respectively. Denoting this by $n_j$, we can write both (55) and (58) as

$$S(t) = \prod_{i|\tau_{j-1}\leq t}\left(\frac{n_i-d_i}{n_i}\right)\,. \tag{59}$$

This is the well-known formula cited in the literature (Kalbfleisch & Prentice, 2011; Lawless, 2011).

(c) From Remark 2.4, we obtain

$$\hat{\Lambda}(t) = \sum_{\tau_j\leq t}\hat{\lambda}_j = \sum_{\tau_j\leq t}\frac{d_j}{n_j}, \quad t \geq \tau_0\,, \tag{60}$$

where

$$n_j = \begin{cases} z(\tau_{j-1}) & \text{if there are no censors in} \quad [\tau_{j-1}, \tau_j)\,, \\ z(\tau_j^c) & \text{if} \quad \tau_j^c \quad \text{is a censored time in} \quad [\tau_{j-1}, \tau_j)\,. \end{cases} \tag{61}$$

This is the estimator introduced by Nelson (1969) and Aalen (1978). These special cases exhibit the role and scope of the presented innovative alternative dynamic approach.

In the following, we state a corollary that further illustrates the role and scope of our dynamic approach. Further details regarding the proof is outlined in the supplementary section.

**Corollary 4.2.** *Let us assume that the conditions of Corollary 3.2 and Example 2.1 in the context of Remark 3.2(iii) are satisfied. For $j \in I(1,n)$, if $\tau_{j-1}$ and $\tau_j$ are consecutive risk/failure times in $[\tau_0, \mathscr{T}]$, then employing Definitions 3.2, 3.3 and 4.2, an estimate for the risk/hazard rate function at $\tau_j$ is determined by:*

$$\hat{\lambda}(\tau_j) = \frac{a_j}{(1-A_{j-1})\Delta\tau_j}\,, \tag{62}$$

*and*

$$\hat{\lambda}(t) = \hat{\lambda}(\tau_j)\,, \quad t \in [\tau_{j-1}, \tau_j)\,, \tag{63}$$

*where $a_j$ and $A_{j-1}$ are defined in Example 2.1.*

*Moreover, an estimate for the survival function is represented by*

$$\hat{S}(t) = S_{j-1}\exp\left[-\hat{\lambda}_j(t-\tau_{j-1})\right]\,, \quad for \quad t \in [\tau_{j-1}, \tau_j)\,. \tag{64}$$

**Remark 4.4.** The PEXE of Kitchin et al. (1980), as well as Kim and Proschan (1991) is undefined beyond the last observed failure time. To rectify that, Malla and Mukerjee (2010) provided the following exponential tail hazard/risk rate estimate:

$$\hat{\lambda}_{\text{tail}} = \frac{\exp(-\hat{\Lambda}_m)}{\sum\limits_{i=1}^{m}(I_j - J_j)} \tag{65}$$

where

$$I_j = \int_{\tau_{j-1}}^{\tau_j} \hat{S}^{KM}(t)\mathrm{d}t = (1 - A_{j-1})(\tau_j - \tau_{j-1})$$

and

$$J_j = \int_{\tau_{j-1}}^{\tau_j} \hat{S}^{MN}(t) = \exp(-\hat{\Lambda}_{j-1})\frac{(1 - A_{j-1})(\tau_j - \tau_{j-1})}{a_j}\left[1 - \exp\left(-\frac{a_j}{1 - A_{j-1}}\right)\right] .$$

Thus, under the following assumptions: (i) no ties among the failure times, (ii) the last observation is uncensored, a new PEXE of Malla and Mukerjee (2010) is given by

$$S(t) = \begin{cases} \exp(-\Lambda_{j-1})\exp\left(\frac{-a_j(t-\tau_{j-1})}{(1-A_{j-1})(\tau_j-\tau_{j-1})}\right), & \tau_{j-1} \le t < \tau_j\,, \quad j \in I(1,m) \\ \exp(-\hat{\Lambda}_m)\exp(-\hat{\lambda}_{\text{tail}}(t-\tau_m))\,, & \tau_m \le t < \infty\,. \end{cases} \tag{66}$$

We further note that the presented dynamic approach does not require the failure function to be invertible.

## 5. Multiple Censored Times between Consecutive Failure Times

In this section, we further apply the conceptual dynamic results developed in Sections 2 and 3 to multiple censored times between consecutive failure times. We present a result that provides a very general algorithm for estimating a hazard rate function for multiple censoring times between consecutive failure times $\tau_{j-1}$ and $\tau_j$ with $\tau_{j-1}, \tau_j \in [\tau_0, \mathscr{T})$. We further note that the presented results in this section extend the results of Section 4 in a systematic and unified manner.

**Theorem 5.1.** *Let the hypotheses of Theorem 3.1 in the context of Remarks 3.1, 3.2(i) and 3.2(ii) be satisfied. For each $j \in I(1,m)$, let $\tau_{j-1}$ and $\tau_j$ be consecutive failure times. Let $\{\tau_{j-1l}\}_{l=1}^{k_j}$ be a finite sequence of censored time observations over a time interval $[\tau_{j-1}, \tau_j]$. Let $\gamma_j^l$ be the number of objects censored at time $\tau_{j-1l}$, for $l \in I(1,k_j)$ and $\{\gamma_j^l\}_{l=1}^{k_j}$ be a corresponding sequence of observed number of objects/species/patients/etc. Then*

1. *$z(\tau_{j-1}) - z(\tau_j) - \sum_{l=1}^{k_j}\gamma_j^l$ is a change in the number of items/subjects that is under the observation over the sub-interval $[\tau_{j-1}, \tau_j]$ of the time interval of study $[\tau_0, \mathscr{T}]$*
2. *$\sum_{l=1}^{k_j+1} z(\tau_{j-1l-1})\Delta(\tau_{j-1l})$ is a total amount of time spent under the observation/testing/evaluation/monitoring of $z(\tau_{j-1l-1})$ items/patients/ infectives/subjects on the interval $[\tau_{j-1l-1}, \tau_{j-1l})$ for $l \in I(1,k_j))$ and $j \in I(1,n)$.*
3. *an estimate for the hazard rate function at $\tau_j$ is determined by*

$$\hat{\lambda}(\tau_j) = \frac{z(\tau_{j-1}) - z(\tau_j) - \sum_{l=1}^{k_j}\gamma_j^l}{\sum_{l=1}^{k_j+1} z(\tau_{j-1l-1})\Delta(\tau_{j-1l})}\,, \tag{67}$$

*and an estimate for the hazard rate function is*

$$\hat{\lambda}(t) = \hat{\lambda}(\tau_j)\,, \quad \textit{for} \quad t \in [\tau_{j-1}, \tau_j) \quad \textit{and} \quad j \in I(1,n) \tag{68}$$

*Proof.* The detailed proof of Theorem 5.1 is given in the supplementary section 8.

**Corollary 5.1.** *Under the conditions of Theorem 5.1 and assumptions of Corollary 3.3 in the context of Remark 3.2(iv), an estimate for the hazard rate function at $\tau_j$ is determined by*

$$\hat{\lambda}(\tau_j) = \frac{z(\tau_{j-1}) - z(\tau_j) - \sum_{l=1}^{k_j}\gamma_j^l}{z(\tau_{j-1k_j})}\,, \tag{69}$$

*and an estimate for the hazard rate function is $\hat{\lambda}(t) = \hat{\lambda}(\tau_j)$, for $t \in [\tau_{j-1}, \tau_j)$ and $j \in I(1,n)$. An estimate for the survival function is thus given by*

$$\hat{S}(t) = \prod_{i|\tau_{j-1}<t}(1 - \hat{\lambda}(\tau_i)),\ t \ge \tau_0,\ \tau_i \le \tau_{j-1} \le t < \tau_j \textit{ for some } j \in I(1,n). \tag{70}$$

**Corollary 5.2.** *Under the conditions of Theorem 5.1 and estimate for the cumulative hazard/risk rate and survival functions are represented by:*

$$\hat{\Lambda}(t, \tau_0) = \sum_{m=1}^{j-1} \hat{\lambda}_m(\tau_m - \tau_{m-1}) + \hat{\lambda}_j \left(t - \tau_{j-1}\right) \ , \ t \in [\tau_{j-1}, \tau_j)$$

*and*

$$\hat{S}(t, \tau_0) = S_0 \exp\left[\sum_{m=1}^{j-1} \hat{\lambda}_m(\tau_m - \tau_{m-1}) + \hat{\lambda}_j \left(t - \tau_{j-1}\right)\right] , \ t \in [\tau_{j-1}, \tau_j)$$

*for $t \geq \tau_0$, $\tau_{j-1} \leq t < \tau_j$ for some $j \in I(1, n)$.*

**Remark 5.1.** (a) We remark that the innovative dynamic approach for the development of computational parameter estimation algorithm (67) is an alternative approach for the algorithm proposed by Kim and Proschan (1991).

(b) The estimates (67) in the context of (20) yields the estimate obtained by Kulasekera and White (1996) as special cases.

(c) For continuous-time interconnected hybrid state survival dynamic process, if $k_j = 0$, for some $j \in I(1, n)$, then $l = 0$ and $\gamma_j^0 = 0$ and (67) reduces to (44). On the other hand, if $k_j = 1$ for some $j \in I(1, n)$, then $l = 0$ and $\gamma_j^1 = \gamma_j^c$ and (67) implies (48).

(d) For discrete-time interconnected hybrid state survival dynamic process, if $k_j = 0$, for some $j \in I(1, n)$, then $l = 0$ and $\gamma_j^0 = 0$ and (69) reduces to (53). On the other hand, if $k_j = 1$, for some $j \in I(1, n)$, then $l = 0$ and $\gamma_j^1 = \gamma_j^c$ and (69) implies (56).

The presented innovative approach of parameter and state estimation includes the Thaler (1984)-type hazard rate estimation problem as a particular case. To justify this statement, we first introduce a concept of hazard/risk rate function for responder and non-responder states. In addition, we state a corollary of Theorem 5.1 without its proof. The proof is outlined in the supplementary section.

**Definition 5.1.** For $i \in I(0, 1)$, Let $\lambda_0(t)$ and $\lambda_1(t)$ represent the hazard/risk rate functions in the non-responder and responder states, respectively, at time $t$ (Thaler, 1984).

**Corollary 5.3.** *Let us assume that the conditions of Corollary 3.1 in the context of Remark 3.2(i) are satisfied. For $j \in I(1, n_0)$, let $\tau_{j-1}$ and $\tau_j$ be consecutive risk/failure times in state 0. For $j' \in (1, n_1)$, let $\tau_{j'-1}$ and $\tau_{j'}$ be consecutive failure times in state 1. Let $z_0(\tau_j)$ be the number of survivals at $\tau_j$ in state* 0. *Let $z_1(\tau_{j'})$ be the number of survivals at $\tau_{j'}$ in state* 1. *Then an estimate for the hazard/risk rate function at $\tau_j$ is determined by:*

$$\hat{\lambda}_0(\tau_j) = \frac{\sum\limits_{m=1}^{j} [z_0(\tau_{m-1}) - z_0(\tau_m)]}{\sum\limits_{m=1}^{j} z_0(\tau_{m-1})\Delta\tau_m} = \frac{\sum\limits_{m=1}^{j} d_{0j}}{\sum\limits_{m=1}^{j} z_0(\tau_{m-1})\Delta\tau_m} , \tag{71}$$

*where $d_{0j}$ is the number of deaths/failures at the jth distinct failure time in state i, and an estimate for the hazard rate function is*

$$\hat{\lambda}_0(t) = \hat{\lambda}_0(\tau_j) , \quad \textit{for} \quad t \in [\tau_{j-1}, \tau_j) \quad \textit{and} \quad j \in I(1, n_0) . \tag{72}$$

*An estimate for the hazard/risk rate function at $\tau_{j'}$ is determined by:*

$$\hat{\lambda}_1(\tau_{j'}) = \frac{\sum\limits_{m=1}^{j'} [z_1(\tau_{m-1}) - z_1(\tau_m)]}{\sum\limits_{m=1}^{j'} z_1(\tau_{m-1})\Delta\tau_m} = \frac{\sum\limits_{m=1}^{j} d_{1j'}}{\sum\limits_{m=1}^{j} z_1(\tau_{m-1})\Delta\tau_m} , \tag{73}$$

*where $d_{1j'}$ is the number of deaths/failures at the $j'$th distinct failure time in state* 1, *and an estimate for the hazard rate function is*

$$\hat{\lambda}_1(t) = \hat{\lambda}_1(\tau_{j'}) , \quad \textit{for} \quad t \in [\tau_{j'-1}, \tau_{j'}) \quad \textit{and} \quad j' \in I(1, n_1) . \tag{74}$$

*The hazard/risk ratio rate function estimate is given by:*

*The corresponding estimate of the* log *hazard/risk rate ratio function for patients currently in a response compared to a nonresponse state is given by:*

$$\hat{\rho}(t) = \ln\left[\frac{\hat{\lambda}_0(\tau_j)}{\hat{\lambda}_1(\tau_{j'})}\right] \text{ for}\,, \tau_{j-1} < t \leq \tau_j \,\text{and}\, \tau_{j'-1} \leq t < \tau_{j'}\ . \tag{75}$$

**Remark 5.2.** We remark that (71), (73) and (75) are identical to the result obtained in Thaler (1984). Moreover, the estimates in (71), (73) and (75) were obtained in the framework of an innovative dynamic approach.

In the following, we state a general theorem that provides a theoretical estimate for the hazard/risk rate function between two successive change point times, $\tau_{j-1}$ and $\tau_j$.

**Theorem 5.2.** *Let the hypothesis of Theorem 5.1 be satisfied. Let $\{T_i^j\}_{i=1}^n$ be a sequence of times(failure/ censor/arrival) that fall between the change point times $\tau_{j-1}$ and $\tau_j$ for $j = I(1,k)$. Then an estimate for the hazard rate function at $\tau_j$ is determined by*

$$\hat{\lambda}(\tau_j) = \hat{\lambda}(\tau_j) = \frac{z(\tau_{j-1}) - z(\tau_j) - \sum\limits_{m=1}^{l} \eta_m^j}{\sum\limits_{m=1}^{l+1} z(T_m^j)\Delta(T_m^j)}\ ,\ \ j \in I(1, k+1)\ . \tag{76}$$

*where*

$$\eta_m^j = \begin{cases} 0 & \textit{if} \quad T_m^j \textit{ is failure time} \\ \gamma_m^{jc} & \textit{if} \quad T_m^j \textit{ is censored time}\ ; \\ -\gamma_m^{ja} & \textit{if} \quad T_m^j \textit{ is arrival time} \end{cases} \tag{77}$$

*$\gamma_m^{jc}$ is the number of objects/items/individuals censored at time $T_m^j$; $\gamma_m^{ja}$ is the number of objects/items/individuals joining/arriving at time $T_m^j$, and an estimate for the hazard rate function is $\lambda(t) = \hat{\tau}_j$ for $t \in [\tau_{j-1}, \tau_j)$.*

*Proof.* The proof of Theorem 5.2 is outlined in the supplementary section.

## 6. Computational Algorithms

In this section, we outline very general conceptual computational, data organizational and simulation schemes. The computational and simulation algorithms are based on fundamental theoretical result (Theorem 5.1) developed in Section 5.

### 6.1 Conceptual Computational Parameter and State Estimation Scheme

The theoretical computational algorithm for interconnected continuous-time hybrid dynamic process (24), is as follows:

$$z(\tau_{j-1}) - z(\tau_j) - \sum_{l=1}^{k_j} \gamma_j^l = \hat{\lambda}(\tau_j) \sum_{l=1}^{k_j+1} z(\tau_{j-1l-1})\Delta(\tau_{j-1l}), \tag{78}$$

and the conceptual computational algorithm for totally discrete-time hybrid dynamic process (26) is

$$z(\tau_{j-1}) - z(\tau_j) - \sum_{l=1}^{k_j} \gamma_j^l = \hat{\lambda}(\tau_j) z(\tau_{j-1k_j})\,. \tag{79}$$

Here $\mathscr{P}_0^{\mathscr{T}} : \tau_0 < \tau_1 < \ldots < \tau_{j-1} < \tau_j < \ldots < \tau_n$ is a partition of failure times over the time interval $[0, \mathscr{T})$. Let $\mathscr{P}_j$ be a partition corresponding to a given finite sequence of censored times over the failure time interval $[\tau_{j-1}, \tau_j)$, and let it be represented by

$$\mathscr{P}_j : \tau_{j-1} = \tau_{j-10} < \tau_{j-11} < \ldots < \tau_{j-1l-1} < \tau_{j-1l} < \ldots < \tau_{j-1k_{j-1}} < \tau_{j-1k_j}\ . \tag{80}$$

For $j \in I(1,n), \lambda$ is the hazard rate function; $z(t)$ stands for the number of survivals at time $t$; $\gamma_j^l$ denotes the number of objects censored at the time $\tau_{j-1l}$, $j \in I(1,m)$ and $l \in I(0,k_j)$, $k_j \in I(0,\infty)$.

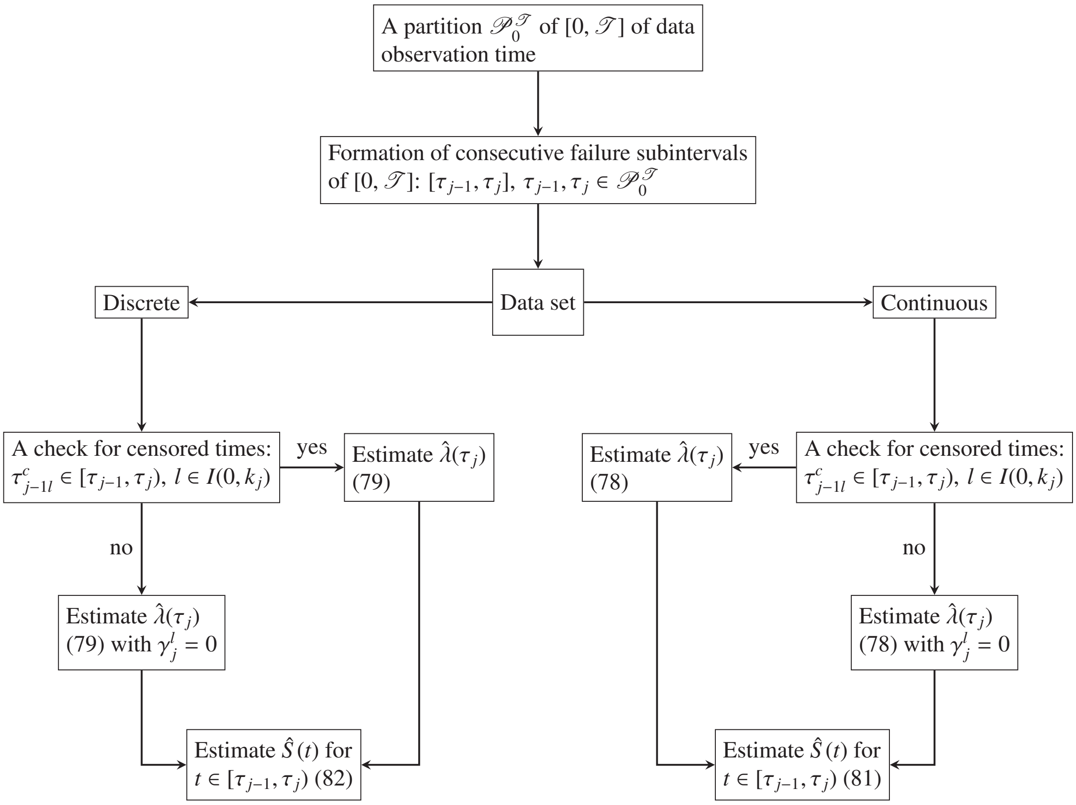
For the continuous-time hybrid dynamic process (24), an estimate of the survival function is represented by

$$\hat{S}(t,\tau_0) = S_0 \exp\left[\sum_{m=1}^{j-1} \hat{\lambda}_m(\tau_m - \tau_{m-1}) + \hat{\lambda}_j\left(t - \tau_{j-1}\right)\right]\ ,\ t \in [\tau_{j-1}, \tau_j) \,\text{for}\ t \geq \tau_0\ . \tag{81}$$

For the totally discrete-time hybrid dynamic process (26), an estimate of the survival function is represented by

$$\hat{S}(t) = \prod_{i|\tau_{j-1}<t} (1 - \hat{\lambda}(\tau_i)),\ t \geq \tau_0. \tag{82}$$

First, we construct a detailed flowchart for the general conceptual computational algorithm developed in Section 5.


A partition $\mathscr{P}_0^{\mathscr{T}}$ of $[0, \mathscr{T}]$ of data observation time
Formation of consecutive failure subintervals of $[0, \mathscr{T}]$: $[\tau_{j-1}, \tau_j]$, $\tau_{j-1}, \tau_j \in \mathscr{P}_0^{\mathscr{T}}$
Data set
Discrete
Continuous
A check for censored times: $\tau^c_{j-1l} \in [\tau_{j-1}, \tau_j)$, $l \in I(0, k_j)$
yes
Estimate $\hat{\lambda}(\tau_j)$ (79)
Estimate $\hat{\lambda}(\tau_j)$ (78)
yes
A check for censored times: $\tau^c_{j-1l} \in [\tau_{j-1}, \tau_j)$, $l \in I(0, k_j)$
no
no
Estimate $\hat{\lambda}(\tau_j)$ (79) with $\gamma^l_j = 0$
Estimate $\hat{\lambda}(\tau_j)$ (78) with $\gamma^l_j = 0$
Estimate $\hat{S}(t)$ for $t \in [\tau_{j-1}, \tau_j)$ (82)
Estimate $\hat{S}(t)$ for $t \in [\tau_{j-1}, \tau_j)$ (81)

Flowchart 1. Conceptual Computational Algorithm

We observe that the conceptual computational algorithm (Flowchart 1) is composed of two sub-conceptual computational algorithms, namely, continuous-time and discrete-time hybrid dynamic processes.

*6.2 Conceptual and Computational Simulation Algorithms*

A pseudocode for a simulation scheme for both interconnected continuous-time and totally discrete-time hybrid dynamic processes are outlined below:

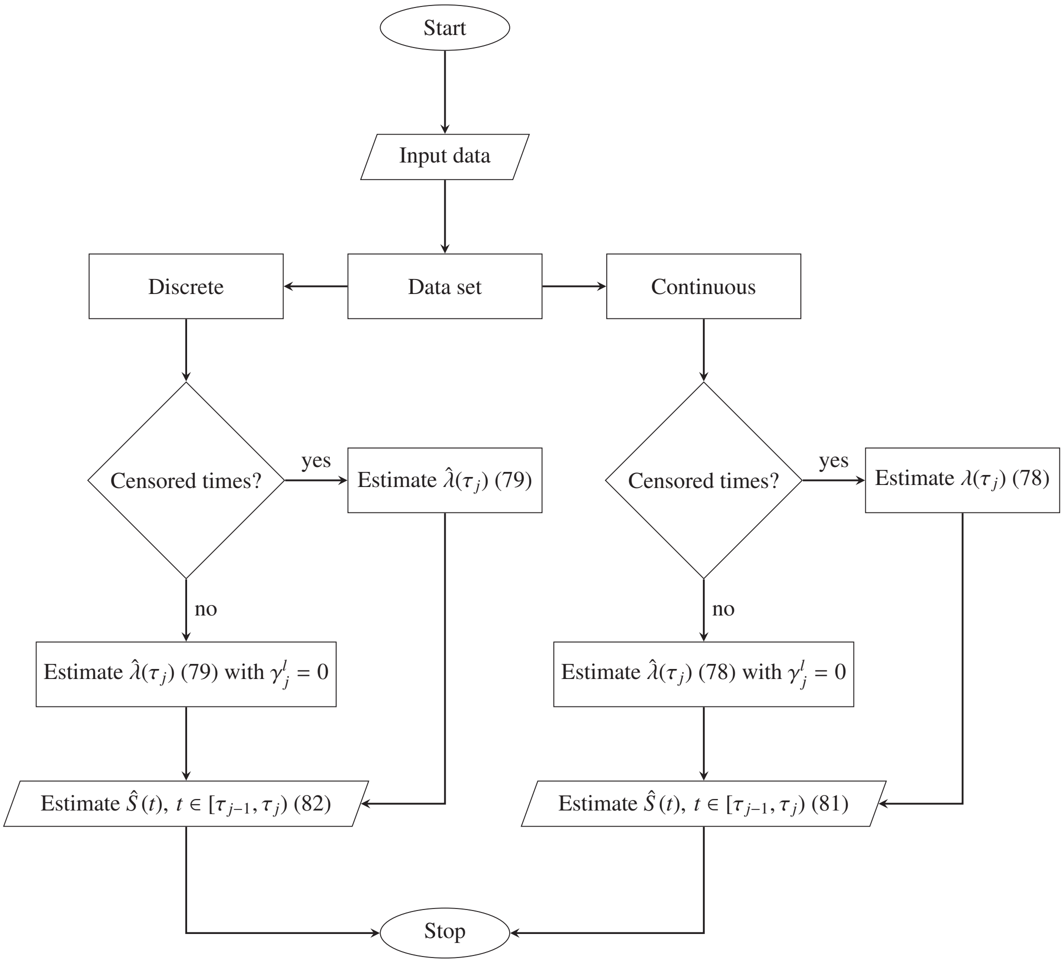
```
for $j = 1$ to $N$ do
Compute $k_j$, $z(\tau_{j-1}), z(\tau_j)$
    if $k_j = 0$ then
Compute $z(\tau_{j-1})\Delta\tau_j$
    else
Compute $\sum_{l=1}^{k_j} \gamma^l_j$, $\sum_{l=1}^{k_j+1} z(\tau_{j-1l-1})\Delta(\tau_{j-1l})$
    end if
Compute $\hat{\lambda}(\tau_j)$, $\hat{S}(t)$
end for
```

Simulation Scheme 1a. Pseudocode for interconnected continuous-time hybrid dynamic process

```
for $j = 1$ to $N$ do
Compute $k_j$, $z(\tau_{j-1}), z(\tau_j)$
    if $k_j = 0$ then
Compute $z(\tau_{j-1})$
    else
Compute $\sum_{l=1}^{k_j} \gamma^l_j$, $z(\tau_{j-1k_j})$
    end if
Compute $\hat{\lambda}(\tau_j)$, $\hat{S}(t)$
end for
```

Simulation Scheme 1b. Pseudocode for totally discrete-time hybrid dynamic process

Moreover, a flowchart for the simulation algorithm for parameter and state estimation problems for interconnected continuous-time (24) and discrete-time (26) hybrid dynamic processes are provided in Flowchart 2.


Start
Input data
Data set
Discrete
Continuous
Censored times?
yes
no
Estimate $\hat{\lambda}(\tau_j)$ (79)
Estimate $\lambda(\tau_j)$ (78)
Estimate $\hat{\lambda}(\tau_j)$ (79) with $\gamma_j^l = 0$
Estimate $\hat{\lambda}(\tau_j)$ (78) with $\gamma_j^l = 0$
Estimate $\hat{S}(t)$, $t \in [\tau_{j-1}, \tau_j)$ (82)
Estimate $\hat{S}(t)$, $t \in [\tau_{j-1}, \tau_j)$ (81)
Stop

Flowchart 2. Simulation Algorithm for interconnected hybrid dynamic processes

We note that flowchart for simulation algorithm (Flowchart 2) is composed of two sub-simulation algorithms, namely, continuous-time and totally discrete-time hybrid dynamic processes.

In the following, using the conceptual computational algorithm, we exemplify our theoretical procedure by estimating hazard rate and survival functions of two data sets in a systematic and unified way. The first data set can be found in Kaplan and Meier (1958).

**Illustration 6.1.** Suppose that out of a sample of 8 items the following are observed:

Table 1. Dataset used by Kaplan and Meier (1958)

| Order of Observation | Time of Cessation of Observation | Cause of Cessation | Time Notation |
|---|---|---|---|
| 1 | 0.8 | Failure | $\tau_1$ |
| 2 | 1.0 | Censored | $\tau_{11}$ |
| 3 | 2.7 | Censored | $\tau_{12}$ |
| 4 | 3.1 | Failure | $\tau_2$ |
| 5 | 5.4 | Failure | $\tau_3$ |
| 6 | 7.0 | Censored | $\tau_{31}$ |
| 7 | 9.2 | Failure | $\tau_4$ |
| 8 | 12.1 | Censored | |

We note that the data set in Table 1 is for the totally discrete-time hybrid time-to-event dynamic process (26). In view of this, we apply the totally discrete-time parameter and state estimation schemes (79) and (82). In short, we utilize the discrete-time conceptual computational sub-algorithm (Simulation Scheme 1b) "pseudocode" and simulation sub-algorithm (Flowchart 2).

For $t \in [\tau_0, \tau_1)$, there are no censored times between $[\tau_0, \tau_1)$. Therefore, $k_j = 0$, and from Remark 5.1(d) and hence using (79) we have

$$\hat{\lambda}(\tau_1) = \hat{\lambda}_1 = \frac{z(\tau_0) - z(\tau_1)}{z(\tau_0)} = \frac{1}{8}\,.$$

Utilizing (82), the corresponding survival function is given by

$$\hat{S}(t) = \begin{cases} 1\,, & \text{for} \quad t \in [\tau_0, \tau_1)\,, \\ 1 - \lambda_1 = \frac{7}{8}\,, & \text{for} \quad t = \tau_1\,. \end{cases}$$

For $t \in [\tau_1, \tau_2)$, we note that there are two censored times between $\tau_1$ and $\tau_2$. So, $k_j = k_2 = 2$. Hence

$$\sum_{l=1}^{2} \gamma_2^l = \gamma_2^1 + \gamma_2^2 = 1 + 1 = 2\,.$$

Also, $z(\tau_{j-1k_j}) = z(\tau_{12}) = 5$. Thus, from Remark 5.1(d) and hence applying (79), we have

$$\hat{\lambda}(\tau_2) = \hat{\lambda}_2 = \frac{z(\tau_1) - z(\tau_2) - \sum\limits_{l=1}^{2} \gamma_2^l}{z(\tau_{12})} = \frac{1}{5}\,.$$

Utilizing (82), the corresponding survival function is thus given by

$$\hat{S}(t) = \begin{cases} \frac{7}{8}\,, & \text{for} \quad t \in [\tau_1, \tau_2)\,, \\ \prod\limits_{k|\tau_j \le t} (1 - \hat{\lambda}_j) = \prod\limits_{j=1}^{2} (1 - \hat{\lambda}_j) = \frac{7}{10}\,, & \text{for} \quad t = \tau_2\,. \end{cases}$$

There is no censoring time between the interval $[\tau_2, \tau_3) = [3.1, 5.4)$. Therefore, $k_j = 0$, and from Remark 5.1(d) and hence using (79) we obtain

$$\hat{\lambda}(\tau_3) = \frac{z(\tau_2) - z(\tau_3)}{z(\tau_2)} = \frac{1}{4}\,.$$

Once again, utilizing (82), the corresponding survival function is thus given by

$$\hat{S}(t) = \begin{cases} \frac{7}{10}\,, & \text{for} \quad t \in [\tau_2, \tau_3)\,, \\ \prod\limits_{j=1}^{3} (1 - \hat{\lambda}_j) = \frac{21}{40}\,, & \text{for} \quad t = \tau_3\,. \end{cases}$$

Continuing in this manner, we record the estimates for hazard rate and survival functions in the following table with the last column exhibiting the survival function estimate as obtained by Kaplan and Meier (1958).

Table 2. Kaplan and Meier Survival estimates for data set given in Kaplan and Meier (1958).

| Failure Times<br>$\tau_j$ | Survivals<br>$z(\tau_j)$ | Hazard Rate Function<br>$\hat{\lambda}(\tau_j)$ | Survival Function<br>$\hat{S}(\tau_j)$ |
|---|---|---|---|
| 0.8 | 7 | 1/8 | 7/8 |
| 3.1 | 4 | 1/5 | 7/10 |
| 5.4 | 3 | 1/4 | 21/40 |
| 9.2 | 1 | 1/2 | 21/80 |
| (12.1) | 0 | 1/2 | 21/80 |

Using the dataset in Kim and Proschan (1991) and theoretical computational algorithm, Theorem 5.1, we illustrate the estimation of hazard rate and survival functions, systematically.

**Illustration 6.2.** Suppose that seven items (new) are put on test at time 0. Each item is observed until it fails or until it is withdrawn, whichever occurs first. The resulting set of observation (Kim & Proschan, 1991) is shown in Table 3 in order of occurrence.

Table 3. Data from Kim and Proschan (1991)

| Order of Observation | Time of Cessation of Observation | Cause of Cessation | Time Notation | Finite sequence of censored Time | Size of sequence | Number of Censored |
|---|---|---|---|---|---|---|
| 0 | 0 | | | | | |
| 1 | 2.0 | Failure | $\tau_1 = \tau_{01} = \tau_{10}$ | | | |
| 2 | 3.5 | Censored | $\tau_{11}$ | $\{\tau_{j-1l}\}_{l=1}^{2}$ | $k_2 = 2$ | $\{\gamma_2^l\}_{l=1}^{2}$ |
| 3 | 4.5 | Censored | $\tau_{12}$ | | | |
| 4 | 6.2 | Failure | $\tau_2 = \tau_{13} = \tau_{20}$ | | | |
| 5 | 8.0 | Censored | $\tau_{21}$ | $\{\tau_{j-1l}\}_{l=1}^{1}$ | $k_3 = 1$ | $\{\gamma_3^l\}_{l=1}^{2}$ |
| 6 | 8.8 | Failure | $\tau_3 = \tau_{22}$ | | | |
| 7 | 11.3 | Failure | $\tau_4$ | | | |

The data set in Table 3 is for the interconnected continuous-time hybrid dynamic time-to-event dynamic process (24). In view of this, we apply the continuous-time parameter and state estimation schemes (78) and (81). In short, we utilize the continuous-time conceptual computational sub-algorithm (Simulation Scheme 1a) "pseudocode" and simulation sub-algorithm (Flowchart 2).

For $[0, \tau_1)$ , since there are no censored times in between $[0, \tau_1)$, $k_j = k_1 = 0$. Thus from Remark 5.1(c) and using (78) we have

$$\hat{\lambda}(\tau_1) = \frac{z(\tau_0) - z(\tau_1)}{z(\tau_0)(\tau_{01} - \tau_0)} = \frac{1}{14}.$$

Thus $\hat{\lambda}(t) = \frac{1}{14} \approx 0.0714$ for $t \in [\tau_0, \tau_1) = [0, 2.0)$.

For the estimate on $[\tau_1, \tau_2) = [2.0, 6.2)$, we note that there are two censoring times between $[\tau_1, \tau_2)$, hence $k_j = k_2 = 2$ and

$$\sum_{l=1}^{2} \gamma_2^l = \gamma_2^1 + \gamma_2^2 = 1 + 1 = 2.$$

Thus from Remark 5.1(c) and thus applying (78), we have

$$\hat{\lambda}(\tau_2) = \frac{z(\tau_1) - z(\tau_2) - \sum\limits_{l=1}^{k_2} \gamma_2^l}{\sum\limits_{l=1}^{k_2+1} z(\tau_{1l-1})\Delta\tau_{1l}} = \frac{z(\tau_1) - z(\tau_2) - \sum\limits_{l=1}^{2} \gamma_2^l}{\sum\limits_{l=1}^{3} z(\tau_{1l-1})\Delta\tau_{1l}} = \frac{1}{20.8}.$$

Thus, $\hat{\lambda}(t) = \frac{1}{20.8}$, for $t \in [2.0, 6.2)$.

On the interval $[\tau_2, \tau_3) = [6.2, 8.8)$, we have only one censoring time in between the two failure times. So, $k_j = k_3 = 1$. Thus from Remark 5.1(c) and hence, using (67), we obtain

$$\hat{\lambda}(\tau_3) = \frac{z(\tau_2) - z(\tau_3) - \sum\limits_{l=1}^{1} \gamma_3^l}{\sum\limits_{l=1}^{2} z(\tau_{2l-1})\Delta\tau_{2l}} = \frac{3 - 1 - 1}{z(\tau_{20})\Delta\tau_{21} + z(\tau_{21})\Delta\tau_{22}} = \frac{1}{7}.$$

Hence, $\hat{\lambda}(t) = \frac{1}{7}$, for $t \in [6.2, 8.0)$.

There is no censoring in the interval $[\tau_3, \tau_4)$. Thus,

$$\hat{\lambda}(\tau_4) = \frac{z(\tau_3) - z(\tau_4)}{z(\tau_3)\Delta\tau_4} = \frac{1}{2.5},$$

which implies that $\hat{\lambda}(t) = \frac{1}{2.5} = 0.4$, for $t \in [8.0, 11.3)$.

Following this estimation procedure we have

$$\hat{\lambda}(t) = \begin{cases} 0.0714 & 0 \leq t < \tau_1 = 2 \\ 0.0481 & \tau_1 \leq t < \tau_2 = 6.8 \\ 0.1429 & \tau_2 \leq t < \tau_3 = 8.8 \\ 0.4 & \tau_3 \leq t < \tau_4 = 11.3\ . \end{cases} \tag{83}$$

To obtain the estimate of survival function, we use (81) or we apply the solution process described in Section 2 regarding (7) and obtain exponential pieces on successive intervals between failure times that are joined to form a continuous function. Thus,

$$\hat{S}(t) = \begin{cases} \exp(-0.0714t)\ , & 0 \leq t < 2 \\ \exp\left[-0.1429 - 0.0481(t-2)\right]\ , & 2 \leq t < 6.2 \\ \exp\left[0.3448 - 0.1429(t-6.2)\right]\ , & 6.2 \leq t < 8.8 \\ \exp\left[0.4591 - 0.4(t-8.8)\right]\ , & 8.8 \leq t < 11.3 \\ \text{no estimator,} & t \geq 11.3 \end{cases} \tag{84}$$

**Remark 6.1.** These are the same results obtained by using the method proposed Kim and Proschan (1991).

### 7. Conclusions

Most of the research work in the area of survival and reliability analysis is centered around the probabilistic analysis approach. In general, a closed-form solution is not feasible. In addition, a hazard rate function is nonlinear in covariate state processes and non-stationary. The presented linear hybrid deterministic dynamic modeling is more suitable for a complex time-to-event processes. This innovative approach does not require a closed-form solution distribution. The influence of both continuous and discrete-time states can be easily incorporated as an interconnected hybrid dynamic model for time-to-event processes. In fact, it allows to have a time-varying covariate state influence on the dynamic of a complex survival/reliability of systems. The influence of human mobility, electronic communications, rapid technological changes, advancements in biological, engineering, medical, military, physical and social sciences is motivated to initiate, formulate and to develop an innovative interconnected alternative modeling approach for time-to-event processes in biological, chemical, engineering, epidemiological, medical, multiple-markets and social dynamic processes through discrete-time intervention processes. The presented innovative modeling approach further enhanced our motivation to develop state and parameter estimation procedures. Moreover, the parameter and state estimation approach is dynamic. The dynamic nature rather than the existing algebraic approach plays a very significant role in state and parameter estimation problems in systematic and unifying way. The discrete-time dynamic is exhibited by the two flowcharts and Simulation algorithms 1(a) and 1(b). Furthermore, the significance of the conceptual computational algorithms are also exhibited by illustrations. At the initial level of our objective, we began with a very simple observation of the probabilistic definition of the survival function. This has led to the development of this approach. The role and scope of the presented dynamic approach is exhibited through several existing results (Han et al., 2014; Malla & Mukerjee, 2010; Kim & Proschan, 1991; Thaler, 1984; Aalen, 1978; Nelson, 1969; Kaplan & Meier, 1958) as corollaries, illustrations and remarks. In fact, the full force of the role and scope of hybrid deterministic modeling for time-to-event processes is currently being explored (Appiah E. A. *Time-To-Event Dynamic Processes: Modeling, Methods and Estimations*-Ph.D Dissertation, 2017) for both deterministic and stochastic nonlinear and non-stationary hybrid modeling for time-to-event processes. Furthermore, a complex time-to-event dynamic study is also currently undertaken by Ladde and his team. These developed results will be reported elsewhere.

### 8. Supplements: Proofs of Theorems

In this supplementary section, proofs of a few theorems and corollaries stated in sections 2, 3, 4 and 5 are presented.

***Proof of Theorem 2.1*:** The theorem is proved by the principle of mathematical induction (PMI) (Ladde & Ladde, 2012). From (11), for $j = 1$, we have

$$\mathrm{d}x = [-\alpha(t)\,x + \gamma(t)]\mathrm{d}\beta(t),\ \ x(t_0) = x_0,\ \ t \in [\tau_0, \tau_1)\ .$$

From (10) and the definition of Riemann-Stieltjes integral (Apostol, 1974), we have

$$x(t) - x(\tau_0) = \int_{\tau_0}^{t} [-\alpha(s)\,x(s) + \gamma(s)]\mathrm{d}\beta(s) = 0, \text{ for } t \in [\tau_0, \tau_1)\ . \tag{85}$$

We define

$$x(t) = x(t, \tau_0, x_0) = x_0(t, \tau_0, x_0)\,, \quad x_0(\tau_0) = x_0, \text{ for } t \in [\tau_0, \tau_1)\,. \tag{86}$$

From (10), (11), (85), and $x_0(t, \tau_0, x_0) = x_0(\tau_1^-, \tau_0, x_0)$ for $t \in [\tau_0, \tau^-]$, we have

$$x_0(\tau_1) - x_0(\tau_0) = 0 + \int_{\tau_1^-}^{t} [-\alpha(s)\, x(s) + \gamma(s)]\, \mathrm{d}\beta(s), \text{ for } t \in [\tau_0, \tau_1]\,.$$

From this, the continuity of $\alpha$ and $\gamma$, the definitions of Riemann-Stieltjes integral (Apostol, 1974) and the initial value problem (Ladde & Ladde, 2012), we have

$$\begin{aligned} x_0(\tau_1, \tau_0, x_0) &= x_0(\tau_0) + \beta(\tau_1)[-\alpha(t_1^*)x(t_1^*) + \gamma(t_1^*)] - \beta(t_1^*)[-\alpha(t_1^*)x(t_1^*) + \gamma(t_1^*)] \\ &= x_0(\tau_0) - \alpha_1 x_0(\tau_1^-, \tau_0, x_0) + \gamma_1\,, \end{aligned} \tag{87}$$

for $t_1^* \in [\tau_1^-, \tau_1]$. From (87) and setting $x_0(\tau_1, \tau_0, x_0) = x(\tau_1) = x_1$ and again $x(\tau_1^-, \tau_0, x_0) = x_0$, we obtain

$$\begin{aligned} x_1 &= x(\tau_1^-, \tau_0, x_0) - \alpha_1 x(\tau_1^-, \tau_0, x_0) + \gamma_1 \\ &= (1 - \alpha_1)x_0 + \gamma_1\,. \end{aligned} \tag{88}$$

Continuing the above argument, we can establish the induction hypothesis (Ladde & Ladde, 2012) as:

$$x_j = \Phi(\tau_j, \tau_0)x_0 + \sum_{i=1}^{j} \Phi(\tau_j, \tau_i)\gamma_i\,, \quad \text{for} \quad x(\tau_j) = x_j\,,$$

where

$$\Phi(\tau_j, \tau_i) = \prod_{k=i}^{j}(1 - \alpha_k)\,, \Phi(\tau_i, \tau_i) = 1 \quad \text{for} \quad i \in I(0, n)\,.$$

Now, we consider

$$\mathrm{d}x = [-\alpha(t)\, x + \gamma(t)]\, \mathrm{d}\beta(t), \quad x(\tau_j) = x_j,\ t \in [\tau_j, \tau_{j+1})\,.$$

From the definitions of $x_j$ and $\Phi$, and using the above argument, one can establish the following:

$$x_j(t) = x(t, \tau_j, x_j) = \prod_{k=1}^{j}(1 - \alpha_k)x_0 + \sum_{i=1}^{j-1} \Phi(\tau_j, \tau_i)\gamma_i + \gamma_j \quad \text{for } t \in [\tau_j, \tau_{j+1})\,. \tag{89}$$

Hence

$$\begin{cases} x(\tau_{j+1}^-, \tau_j, x_j) = \prod_{k=1}^{j}(1 - \alpha_k)x_0 + \sum_{i=1}^{j} \Phi(\tau_j, \tau_i)\gamma_i\,, \\ x_{j+1}(\tau_{j+1}, \tau_j, x_j) = (1 - \alpha_{j+1})x_j + \gamma_{j+1}\,. \end{cases} \tag{90}$$

Therefore, from (89) and (90), we have

$$\begin{aligned} x_{j+1} &= (1 - \alpha_{j+1})x_j + \gamma_{j+1} \\ &= \prod_{k=1}^{j+1}(1 - \alpha_k)x_0 + \sum_{i=1}^{j+1} \Phi(\tau_{j+1}, \tau_i)\gamma_i\,. \end{aligned}$$

By the application of PMI and the definition of the IVP regarding hybrid dynamic system (Ladde & Ladde, 2012), we have

$$x(t) = \prod_{k|\tau_j \le t}(1 - \alpha_k)x_0 + \sum_{i=1}^{j-1} \Phi(t, \tau_i)\gamma_i + \gamma_j\,,$$

for $t \ge \tau_0$ and $t \in [\tau_{j-1}, \tau_{j+1})$ . This establishes the proof of the theorem.

***Proof of Theorem 3.1*:** For $t \in [\tau_{j-1}, \tau_j),\ j \ge 1$, from Definition 3.1, Remark 3.1 and the nature of $S$, we have

$$\mathrm{d}z(t) = -\lambda(t)z(t)\mathrm{d}t\,. \tag{91}$$

This establishes the continuous-time dynamic equation in (27). The proof of the discrete-time dynamic part in (27) and iterative process in (28) are outlined below.

Multiplying the discrete-time iterative process in (24) by $S(\tau_j^-)$ and noting the fact that $S(\tau_j) = S(\tau_j^-)$, we obtain

$$x(\tau_j)S(\tau_j) = (1 - \alpha_j)(1 - \beta_j)x(\tau_j^-)S(\tau_j^-) + \gamma_j(1 - \beta_j)S(\tau_j^-) . \tag{92}$$

Moreover, using the definition of $z$, (92) reduces to

$$z(\tau_j) = (1 - \alpha_j)(1 - \beta_j)z(\tau_j^-) + \gamma_j(1 - \beta_j) . \tag{93}$$

This establishes (27).

Applying the Euler-type numerical scheme (Atkinson, 2008) to (91) over an interval $[\tau_{j-1}, \tau_j^-]$, we obtain

$$z(\tau_j^-) - z(\tau_{j-1}) = -\lambda(\tau_{j-1})z(\tau_{j-1})\Delta\tau_j . \tag{94}$$

From (93) and (94) , we have

$$z(\tau_j) = (1 - \lambda(\tau_j)\Delta\tau_j)(1 - \alpha_j)(1 - \beta_j)z(\tau_{j-1}) + \gamma_j(1 - \beta_j) . \tag{95}$$

(95) exhibits the discrete time dynamic for survival process corresponding to the continuous-time dynamic process described in (27) and the discrete-time intervention process. Moreover, (95) exhibits the validity of (28). This establishes proof of Theorem 3.1.

***Proof of Theorem 4.1*:**

(a) Using the discrete-time iterative scheme (28), Remark 3.2(i)(38) and Definitions 3.2, 3.3 and 4.1, we have

$$\lambda(t) = \hat{\lambda}(\tau_j) = \frac{z(\tau_{j-1}) - z(\tau_j)}{z(\tau_{j-1})\Delta\tau_j}$$

for $t \in [\tau_{j-1}, \tau_j)$ and $j \in I(1, k)$. This establishes (a).

(b) Let $\tau_j^c$ be a censoring time between two consecutive risk/failure times, $\tau_{j-1}$ and $\tau_j$. We consider a partition of $[\tau_{j-1}, \tau_j]$ : $\tau_{j-1} < \tau_j^c < \tau_j$.

Employing iterative processes in (40) and (38) on respective subintervals $[\tau_{j-1}, \tau_j^c]$ and $[\tau_j^c, \tau_j]$, we have

$$\begin{aligned} z(\tau_j) - z(\tau_{j-1}) &= z(\tau_j^c) - z(\tau_{j-1}) + z(\tau_j) - z(\tau_j^c) \\ &= -\lambda(\tau_{j-1})\Delta\tau_j^c - \gamma_j^c - \lambda(\tau_j)z(\tau_j^c)\Delta\tau_{jc} \\ &= -\lambda(\tau_j)\left[z(\tau_{j-1})\Delta\tau_j^c + z(\tau_j^c)\Delta\tau_{jc}\right] - \gamma_j^c . \end{aligned} \tag{96}$$

From (96), we obtain:

$$z(\tau_{j-1}) - z(\tau_j) - \gamma_j^c = \lambda(\tau_j)\left[z(\tau_{j-1})\Delta\tau_j^c + z(\tau_j^c)\Delta\tau_{jc}\right] . \tag{97}$$

From (97) and knowing that $\lambda(\tau_j)$ is the hazard/risk rate of change per unit time per unit object/subject, we conclude that $z(\tau_{j-1}) - z(\tau_j) - \gamma_j^c$ is the number of failure/non-operating objects and $z(\tau_{j-1})\Delta\tau_j^c + z(\tau_j^c)\Delta\tau_{jc}$ denotes the total amount of time spent by $z(\tau_{j-1}) - z(\tau_j) - \gamma_j^c$ over the the interval $[\tau_{j-1}, \tau_j)$. This establishes (i) and (ii).

To complete the proofs of (iii) and (iv), we utilize Definition 4.1 and (97), and obtain

$$\hat{\lambda}(\tau_j) = \frac{z(\tau_{j-1}) - z(\tau_j) - \gamma_j^c}{z(\tau_{j-1})\Delta\tau_j^c + z(\tau_j^c)\Delta\tau_{jc}} \quad \text{for} \quad j \in I(1, k) .$$

and hence

$$\lambda(t) = \hat{\lambda}(\tau_j), \quad t \in [\tau_{j-1}, \tau_j) , \quad j \in I(1, k) .$$

This establishes proof of the theorem.

***Proof of Corollary 4.2***: Under the conditions of Example 2.1 and using the relationship between $S$, the cumulative jumps in Example 2.2, Corollary 3.2(in particular (34)), an estimate for the risk/hazard rate function at $\tau_j$ is obtained as:

$$\hat{\lambda}(\tau_j) = \frac{a_j}{(1 - A_{j-1})\Delta\tau_j}\,, \tag{98}$$

and an estimate for the risk/hazard rate function is

$$\hat{\lambda}(t) = \hat{\lambda}(\tau_j)\,, \quad \text{for} \quad t \in [\tau_{j-1}, \tau_j) \quad \text{and} \quad j \in I(1, m) \tag{99}$$

From (32), using (8) and (99), an estimate for the survival function is given by:

$$\hat{S}(t) = \exp(-\Lambda_{j-1}) \exp\left(\frac{-a_j(t - \tau_{j-1})}{(1 - A_{j-1})(\tau_j - \tau_{j-1})}\right), \quad \tau_{j-1} \le t < \tau_j\,, \tag{100}$$

where

$$\Lambda_j = \sum_{i=1}^{j} \frac{a_i}{1 - A_{i-1}},\ 1 \le j \le m,\ \Lambda_0 := 0,$$

and $\Lambda_j$ is the cumulative hazard function. This establishes the proof of the corollary.

***Proof of Theorem 5.1***: For each $j \in I(1, n)$ and $\tau_{j-1}, \tau_j \in \mathscr{P}_0^{\mathscr{T}}$, objects/subjects are censored $k_j$ times over a partition of $[\tau_{j-1}, \tau_j]$ of consecutive failure times. Let $\mathscr{P}_j$ be a partition corresponding to a given finite sequence of censored times over the failure time interval $[\tau_{j-1}, \tau_j)$, and let it be represented by

$$\mathscr{P}_j : \tau_{j-1} = \tau_{j-10} < \tau_{j-11} < \ldots < \tau_{j-1l-1} < \tau_{j-1l} < \ldots < \tau_{j-1k_{j-1}} < \tau_{j-1k_j}\,. \tag{101}$$

where $\mathscr{P}_j$ is a partition of $[\tau_{j-1}, \tau_j]$.

For each $j \in I(1, n)$, using the iterative schemes (38) and (40) we have

$$\begin{aligned} z(\tau_j) - z(\tau_{j-1}) &= \sum_{l=1}^{k_j} \left[z(\tau_{j-1l}) - z(\tau_{j-1l-1})\right] + [z(\tau_j) - z(\tau_{j-1k_j})] \\ &= -\lambda(\tau_j)\left[\sum_{l=1}^{k_j+1} z(\tau_{j-1l-1})\Delta\tau_{j-1l}\right] - \sum_{l=1}^{k_j} \gamma_j^l\,, \end{aligned} \tag{102}$$

and hence

$$z(\tau_{j-1}) - z(\tau_j) - \sum_{l=1}^{k_j} \gamma_j^l = \lambda(\tau_j) \sum_{l=1}^{k_j+1} z(\tau_{j-1l-1})\Delta(\tau_{j-1l})\,. \tag{103}$$

Thus, $z(\tau_{j-1}) - z(\tau_j) - \sum\limits_{l=1}^{k_j} \gamma_j^l$ is a change in the number of items/subjects that are under observation over the subinterval $[\tau_{j-1}, \tau_j]$, and $\sum\limits_{l=1}^{k_j+1} z(\tau_{j-1l-1})\Delta(\tau_{j-1l})$ is a total amount of time spent under the observation/testing/evaluation/monitoring of $z(\tau_{j-1l})$ items/patients/infectives/subjects on the interval $[\tau_{j-1l-1}, \tau_{j-1l})$ for $l \in I(1, k_j))$ and $j \in I(1, n)$. These statements establish conclusions 1 and 2 of Theorem 5.1.

Finally, from Definition 4.1, we obtain an estimate for a hazard rate function at $\tau_j \in [\tau_0, \mathscr{T})$ as:

$$\hat{\lambda}(\tau_j) = \frac{z(\tau_{j-1}) - z(\tau_j) - \sum\limits_{l=1}^{k_j} \gamma_j^l}{\sum\limits_{l=1}^{k_{j+1}} z(\tau_{j-1l-1})\Delta(\tau_{j-1l})}\,.$$

This establishes (67).

Moreover,

$$\hat{\lambda}(t) = \hat{\lambda}(\tau_j)\,, \quad \text{for} \quad t \in [\tau_{j-1}, \tau_j) \quad \text{and} \quad j \in I(1, n)\,. \tag{104}$$

This completes the proof of the theorem.

***Proof of Theorem 5.2*:**

Let $0 = \tau_0 < \tau_1 < \tau_2 < \ldots < \tau_{j-1} < \tau_j < \ldots < \tau_k$ be the partition of $[\tau_0, \mathscr{T})$ corresponding to change point times. For $j = 1, 2, \ldots, k$, we consider a partition of $[\tau_{j-1}, \tau_j]$ as follows:

$$\mathscr{P}_j^\tau : \tau_{j-1} = T_0^j < T_1^j < T_2^j < T_3^j < \ldots < T_{l-1}^j < T_l^j < \ldots < T_{n-1}^j < T_n^j < T_{n+1}^j = \tau_j \,. \tag{105}$$

Imitating the proof of Theorem 5.1, we have

$$\begin{aligned}
z(\tau_j) - z(\tau_{j-1}) &= \sum_{m=1}^{l} \left[ z(T_m^j) - z(T_{m-1}^j) \right] + [z(\tau_j) - z(T_l^j)] \\
&= \sum_{m=1}^{l} \left[ -\lambda(T_{m-1}^j) z(T_{m-1}^j) \Delta T_m^j - \eta_m^j \right] + [-\lambda(T_l^j) z(T_l^j) \Delta \tau_j] \\
&- \lambda(\tau_j) \left[ \sum_{m=1}^{l} z(T_{m-1}^j) \Delta T_m^j \right] - \sum_{m=1}^{l} \eta_m^j - \lambda(\tau_j) z(t_l^j) \Delta \tau_j \\
&= -\lambda(\tau_j) \left[ \sum_{m=1}^{l+1} z(T_{m-1}^j) \Delta T_m^j \right] - \sum_{m=1}^{l} \eta_m^j \,,
\end{aligned} \tag{106}$$

and hence

$$z(\tau_{j-1}) - z(\tau_j) - \sum_{m=1}^{l} \eta_m^j = \lambda(\tau_j) \sum_{m=1}^{l+1} z(T_{m-1}^j) \Delta T_m^j \tag{107}$$

Thus, $z(\tau_{j-1}) - z(\tau_j) - \sum_{m=1}^{l} \eta_m^j$ is a change in the number of items/subjects that is under the observation over the subinterval $[\tau_{j-1}, \tau_j]$ of the time interval of study $[\tau_0, \mathscr{T}]$ and $\sum_{m=1}^{l+1} z(T_m^j) \Delta T_m^j$ is a total amount of time spent under the observation/testing/evaluation of $z(T_m^j)$ items/patients/infectives/subjects on the interval $[T_{m-1}^j, T_m^j)$ for $m \in I(1, l))$ and $j \in I(1, k)$. These statements establish conclusions 1 and 2 of Theorem 5.1.

Finally, from Definition 4.1, we obtain an estimate for a hazard rate function at $\tau_j \in [\tau_0, \mathscr{T})$ as:

$$\hat{\lambda}(\tau_j) = \frac{z(\tau_{j-1}) - z(\tau_j) - \sum\limits_{m=1}^{l} \eta_m^j}{\sum\limits_{m=1}^{l+1} z(T_{m-1}^j) \Delta T_m^j} \,,$$

Moreover,

$$\hat{\lambda}(t) = \hat{\lambda}(\tau_j) \,, \quad \text{for} \quad t \in [\tau_{j-1}, \tau_j) \quad \text{and} \quad j \in I(1, k) \,. \tag{108}$$

This establishes proof of the theorem.

***Proof of Corollary 5.3***: Let $\tau_0 < \tau_1 < \ldots < \tau_{m-1} < \tau_m < \ldots < \tau_{j-1} < \tau_j < \ldots < \tau_n = \mathscr{T}$ be a partition of $[\tau_0, \mathscr{T}]$. Using (31), for fixed $i = 0$ and $j \in I(1, n_0)$, we have

$$z_0(\tau_m) - z_0(\tau_{m-1}) = -\lambda_0(\tau_m) z_0(\tau_{m-1}) \Delta \tau_m \,. \tag{109}$$

Summing (109) from $m = 1$ to $j$, we obtain

$$\begin{aligned}
\sum_{m=1}^{j} [z_0(\tau_m) - z_0(\tau_{m-1})] &= \sum_{m=1}^{j} -\lambda_0(\tau_m) z_0(\tau_{m-1}) \Delta_m \\
&= -\lambda_0(\tau_j) \sum_{m=1}^{j} z_0(\tau_{m-1}) \Delta \tau_m \,.
\end{aligned} \tag{110}$$

Rearranging (110) establishes (71). The proof of (73) is similar to the proof of (71). (75) is obtained by taking the natural log of the ratio of (71) and (73) . This establishes the proof of the corollary.

## Acknowledgments

This research is supported by the Mathematical Sciences Division, U.S. Army Research Office, Grant No: W911NF-15-1-0182.

## References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 701-726. http://dx.doi.org/10.1214/aos/1176344247

Anis, M. Z. (2009). Inference on a sharp jump in hazard rate: a review. *Economic Quality control, 24*(2), 213-229. http://dx.doi.org/10.1515/EQC.2009.213

Apostol, T. M. (1967). *Calculus, vol 1: one-variable calculus, with an introduction to linear algebra.* New York, NY: John Wiley & Sons.

Apostol, T. M. (1974). *Mathematical Analysis*. Reading, MA: Addison Wesley.

Atkinson, K. E. (2008). *An introduction to numerical anlaysis*. New York, NY: John Wiley & Sons.

Chandra, J., & Ladde, G. S. (2014). Multi-cultural dynamics on social networks under external random perturbations. *International Journal of Communications, Network and System Sciences, 7*(06), 181-195. http://dx.doi.org/10.4236/ijcns.2014.76020

Davis, D. J. (1952). An analysis of some failure data. *Journal of the American Statistical Association, 47*(258), 113-150. http://dx.doi.org/10.1080/01621459.1952.10501160

Demarqui, F. N., Loschi, R. H., & Colosimo, E. A. (2008). Estimating the grid of time-points for the piecewise exponential model. *Lifetime Data Analysis, 14*(3), 333-356. http://dx.doi.org/10.1007/s10985-008-9086-0

Fang, L., & Su, Z. (2011). A hybrid approach to predicting events in clinical trials with time-to-event outcomes. *Contemporary Clinical Trials, 32*(5), 755-759. http://dx.doi.org/10.1016/j.cct.2011.05.013

Feigl, P., & Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics*, 826-838. http://dx.doi.org/10.2307/2528247

Gamerman, D. (1994). Bayes estimation of the piece-wise exponential distribution. *IEEE Transactions on Reliability, 43*(1), 128-131. http://dx.doi.org/10.1109/24.285126

Goodman, M. S., Li, Y., & Tiwari, R. C. (2011). Detecting Multiple change points in piecewise constant hazard functions. *Journal of Applied Statistics, 33*(11), 2523-2532. http://dx.doi.org/10.1080/02664763.2011.559209

Han, G., Schell, M. J., & Kim, J. (2014). Improved survival modeling in cancer research using a reduced piecewise exponential approach. *Statistics in Medicine, 33*(1), 59-73. http://dx.doi.org/10.1002/sim.5915

He, P., Kong, G., & Su, Z. (2013). Estimating the survival functions for right-censored and interval-censored data with piecewise constant hazard functions. *contemporary Clinical Trials, 35*(2), 122-127. http://dx.doi.org/10.1016/j.cct.2013.04.009

Kalbfleish, J. D., & Prentice, R. L. (2011). *The statistical analysis of failure time data* (Vol. 360). New Jersey: John Wiley & Sons.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457-481. http://dx.doi.org/10.1080/01621459.1958.10501452

Kim, J. S., & Proschan, F. (1991). Piecewise exponential estimator of the survivor function. *IEEE Transactions on Reliability, 40*(2), 134-139. http://dx.doi.org/10.1109/24.87112

Kitchin, J., Langberg, N. A., & Proschan, F. (1980). *A new method for estimating life distributions from incomplete data* (No. FSU - Statistics-M548). Florida State Uni Tallahassee, Dept of Statistics.

Kulasekera, K., & White, W. H. (1996). Estimation of the survival function from censored data: a method based on the total time on test. *Commnications in Statistics-Simulation and Computation, 25*(1), 189-200. http://dx.doi.org/10.1080/03610919608813306

Ladde, A. G., & Ladde, G. S. (2012). *An introduction to differential equations. Deterministic Modeling, Methods and Anlaysis*(Vol. 1). Singapore: World Scientific

Ladde, G. S. (2015). *Network dynamic processes under stochastic perturbations*. Technical report, U.S. Army Research Office, Mathematical Sciences Division, Research Triangle Park, NC.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data* (Vol. 362). New Jersey: John Wiley & Sons.

Malla, G., & Mukerjee, H. (2010). A new piecewise exponential estimator of a survival function. *Statistics and Probability Letters, 80*(23), 1911-1917. http://dx.doi.org/10.1016/j.spl.2010.08.019

Miller R, G. (2011). *Survival Analysis* (Vol 66). New York, NY: John Wiley & Sons.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Control, 1*(1), 27-52.

Rosen, R. (1970). *Dynamical system theory biology: stability theory and its applications*. New York, NY: John Wiley & Sons.

Thaler, H. (1984). Nonparametric estimation of the hazard ratio. *Journal of the American Statistical Association, 79*(386), 290-293 http://dx.doi.org/10.1080/01621459.1984.10478043

Wanduku, D., & Ladde, G. S. (2011). A two-scale network dynamic model for human mobility process. *Mathematical Biosciences, 229*(1), 1-15. http://dx.doi.org/10.1016/j.mbs.2010.11.003

Whittemore, A. S., & Keller, J. B. (1983). *Survival estimation with censored data*. Department of Statistics, Stanford University, Stanford, CA.

**Copyrights**

# Identification of Biomarkers for Predicting the Overall Survival of Ovarian Cancer Patients: a Sparse Group Lasso Approach

Kristi Mai[1] & Qingyang Zhang[1]

[1] Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR, USA

Correspondence: Qingyang Zhang, Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701, USA. E-mail: qz008@uark.edu

Received: September 12, 2016 Accepted: October 11, 2016 Online Published: October 26, 2016

doi:10.5539/ijsp.v5n6p57 URL: http://dx.doi.org/10.5539/ijsp.v5n6p57

**Abstract**

Next-generation sequencing has been routinely applied to cancer biology, making it possible for researchers to elucidate the molecular mechanisms underlying cancer initiation and progression. However, how to identify oncomarkers from massive complex genomic data poses a great challenge for both modeling and computing. In this paper, we propose a novel computational pipeline to identify genes related to the overall survival of ovarian cancer patients from the rich Cancer Genome Atlas data. Different from the existing studies, we incorporate dependence structure among genes and pathway information into the variable selection. Firstly, the dimensionality of the ovarian cancer data is reduced by a novel stepwise feature screening which mimics the hierarchy of the underlying causal network. The second step of the pipeline is to divide genes into clusters with distinct cellular functions by k-means, x-means and PAMSAM learning algorithms. In the final step, we fit a cox proportional hazard model with a sparse group lasso penalty for further variable selection. Of the 115 genes in the final list, many were reported to be associated with cancer initiation or progression in the literature. In addition, we find several gene families including the NEK family and RNF family, which are closely associated with the survival of ovarian cancer patients.

**Keywords:** The Cancer Genome Atlas, ovarian cancer, k-means clustering, stepwise feature selection, sparse group lasso.

## 1. Introduction

Ovarian cancer is one of the most malignant gynecologic cancers, ranking fifth as the cause of cancer-related deaths among women in the United States. According to American Cancer Society, about 22, 280 women will receive a new diagnosis of ovarian cancer and about 14, 240 women will die from this disease in 2016. The latest data shows that about 70% of deaths occur in patients with high-grade serous epithelial ovarian cancer. The standard treatment for these patients is usually debulking surgery, followed by platinum-taxane chemotherapy. Platinum resistant cancer recurs within six months in about 25% of patients and the overall five-year survival rate is about 31%. Approximately 13% of high-grade serous ovarian cancer can be attributed to germline mutations in *BRCA1* and *BRCA2* and a smaller percentage can be accounted for by other germline mutations (The Cancer Genome Atlas Research Network, 2011).

With the rapid advances in high-throughput sequencing technology, it is now possible to investigate a large number of genetic and epigenetic features simultaneously (The Cancer Genome Atlas Research Network, 2011; Zhang, Burdette, & Wang, 2014; Zhang & Wang, 2016; Zhang, 2015; Kumar, Breen, & Ranganathan, 2013; Konstantinopoulos, Spentzos, & Cannistra, 2008; Popovic et al., 2014). The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) project provides the most comprehensive genomic data resource for more than 20 cancer types and subtypes including ovarian serous cystadenocarcinoma (OV), breast invasive carcinoma (BRCA), glioblastoma multiforme (GBM), lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). For instance, TCGA ovarian cancer data contains both clinical information and molecular profile of 568 tumor samples. The clinical information includes records on age, race, survival, outcome of debulking surgery, and treatment resistance etc. The molecular profile includes copy number variation (CNV), DNA methylation, exon expression, gene expression (both microarray and RNA-seq), genotype (SNP), MicroRNA expression (microarray), protein expression, and somatic mutation. These massive complex datasets have driven enthusiasm to elucidate molecular mechanisms of cancer through computational approaches (Zhang et al., 2014; Chen et al., 2012; Xu et al., 2012; Xi et al., 2014; Matveeva et al., 2016).

In this paper, we aim to identify prognostic genes which play crucial roles in the survival of the ovarian cancer patients. Relevant works include but are not limited to McLaughlin et al. (McLaughlin et al., 2013), Nagle, Chenevixtrench, Webb, & Spurdle (Nagle, Chenevixtrench, Webb, & Spurdle, 2007), and Konstantinopoulos et al. (Konstantinopoulos et al., 2008). However, the methods used in current studies tend to be over simplistic and inaccurate, mostly the single-round independent screening based on the t test or proportional hazard model. As pointed out by many researchers, these naive methods might result in poor selection of important features by overlooking the complex dependence structure among them. For instance, the independent test may suffer from spurious correlations in high dimensional data and fail to identify important features presented in the underlying causal network. Due to several major difficulties, the association between cancer survival and different signaling pathways has been much less studied and numerous questions remain unanswered in this field. To fill this gap, we develop an efficient and general pipeline which achieves higher accuracy in the biomarker/pathway identification. The advantage of the proposed methods is twofold. First, the feature selection is conducted in account of the dependence structure among genes, resulting in a more accurate selection of biomarkers that may directly or indirectly affect cancer survival. Second, the sparse group lasso method offers a further refinement for the candidate set of biomarkers, by encouraging genes in the same pathway to be selected and balancing gene-wise and group-wise selection in the meanwhile.

The rest of paper is organized as follows. In Section 2, we briefly summarize the TCGA ovarian cancer data and elaborate the three steps in the computational pipeline: (1) stepwise feature selection for initial screening; (2) k-means clustering along with a ”elbow method” to determine the number of clusters; (3) Cox proportional hazards model with sparse group lasso penalty for further variable selection. We present and discuss the main results from the analysis in Section 3, and conclude the article in Section 4.

## 2. Material and Method

### *2.1 Data Integration and Preprocessing*

Using ”data matrix” tool on the TCGA website, we extracted the level-3 microarray data containing the expression level of 17, 814 genes in 568 tumor samples, as well as the clinical information. Table 1 summarizes our data set. The overall survival time of each patient is defined as the time between diagnosis and death. The censoring indicator was set to be 1 if death event occurred and 0 otherwise. Throughout this study, we assume that censoring mechanism is independent of survival mechanism.

Table 1. Data types, platforms and sample size in the analysis

| Data type | Platform | Cases |
|---|---|---|
| Gene expression | Agilent 244K | 572 (8 organ-specific controls) |
| Clinical information | N/A | 583 |

The gene expression data were normalized using a quantile normalization method by Balstad et al. (Balstad et al., 2002) to correct the bias due to non-biological causes. We applied an existing method by Hsu et al. (Hsu et al., 2012) to remove age and batch effects (three age groups are defined as < 40 y.o., [40, 70] y.o., and > 70 y.o.). This method is based on a median-matching and variance-matching strategy. For example, the batch-effect-adjusted gene expression value can be obtained as follows:

$$g^*_{ijk} = M_i + (g_{ijk} - M_{ij})\frac{\hat{\sigma}_{g_i}}{\hat{\sigma}_{g_{ij}}},$$

where $g_{ijk}$ represents the gene expression value for gene i from batch j and sample k, $M_{ij}$ refers to the median of $g_{ij} = (g_{ij1}, ..., g_{ijn})$, $M_i$ refers to the median of $g_i = (g_{i1}, ..., g_{iJ})$, $\hat{\sigma}_{g_i}$ and $\hat{\sigma}_{g_{ij}}$ are the sample standard deviation of $g_i$ and $g_{ij}$, respectively.

### *2.2 Stepwise Feature Selection*

A necessary and crucial step for genome-wide association study is feature screening, i.e., to filter out irrelevant or redundant features. A refined variable set helps improve computing efficiency and estimation accuracy (Zhang et al., 2014). Existing feature selection methods can be classified into either wrapper approach (Kohavi & John, 1997; Leng, Valli, & Armstrong, 2010) or filter approach (Haindl, Somol, Ververidis, & Kotropoulos, 1999; Jouve & Nicoloyannis, 2010). The filter approach using independent test for two conditions is more commonly used due to its efficiency and simplicity. However, it tends to filter out many related features in high-dimensional settings. To this end, Zhang et al. (Zhang et al., 2014) proposed a novel stepwise correlation-based selector (SCBS) to select features from TCGA data for further Bayesian network inference. Assume there is a causal chain X→Y→cancer. Though X to Y or Y to cancer has directed association, the association between X and cancer could greatly decay so that it cannot be detected by independent test. The SCBS procedure starts with detection of features strongly correlated with the phenotype and then progressively selects features that correlate with features selected in previous step. This procedure is a natural mimic of sparse network structure and is capable of identifying nodes that are indirectly associated with the phenotype. In practice, the method can be implemented as follows:

- Step 1: Calculate the Spearman’s correlation coefficients between the current variable $X_i$ and all the other variables, denoted by $\rho_{ij}$, $j \neq i$. Keep $k$ most correlated variables with $X_i$ based on $\rho_{ij}$ for further filtering.

- Step 2: Calculate the p-value of correlation coefficient for each of the $k$ variables from step 1, select the variable if the p-value is significant under Benjamini-Hochberg (BH) procedure with FDR$\leq 0.05$.

- Step 3: Repeat step 1 and 2 until $p$ variables are selected.

In practice, the total number of selected variables $p$ is subject to the scale of the model to build. For the TCGA data, we run the SCBS for 4 rounds, in order to select more than 500 but less than 1000 genes. The computing time of SCBS is sublinear to $p$. The choice of $k$ is essential for SCBS which partially depends on the network density. Based on an extensive simulation study, a $k$ of 4 or 5 is recommended by Zhang et al. (Zhang et al., 2014) to attain moderate complexity or sparsity of the model. We set $k = 4$ in our analysis and obtained a set of 603 genes.

### *2.3 K-means Clustering*

The unsupervised k-means clustering is applied to cluster the 603 selected genes based on a correlation metric defined as follows:

$$||\mathbf{g}_i, \mathbf{g}_j||_\rho = 1 - |\rho_{\mathbf{rg}_i, \mathbf{rg}_j}|,$$

where $\mathbf{g}_i$ and $\mathbf{g}_j$ represent expression level of gene i and gene j, $i, j = 1, 2, ..., p$, $\mathbf{g}_i$ is n-dimensional vector where n is number of samples. The Spearman's correlation between gene i and gene j is denoted by $\rho_{\mathbf{rg}_i, \mathbf{rg}_j}$, where $\mathbf{rg}_i$ and $\mathbf{rg}_j$ represent the ranks of $\mathbf{g}_i$ and $\mathbf{g}_j$. An immediate consequence by this definition is $0 \leq ||\mathbf{g}_i, \mathbf{g}_j||_\rho \leq 1$, where the two equalities hold when $\rho_{\mathbf{rg}_i, \mathbf{rg}_j} = 0$ and $|\rho_{\mathbf{rg}_i, \mathbf{rg}_j}| = 1$, respectively. The k-means clustering algorithm aims to partition $p$ variables into K clusters $\mathbf{C} = (C_1, C_2, ..., C_K)$, where K is a predefined number of clusters. Its objective is to find:

$$\arg\min_{\mathbf{C}} \sum_{k=1}^{K} \sum_{\mathbf{g} \in C_k} ||\mathbf{g} - \boldsymbol{\mu}_k||_\rho .$$

An "elbow method" was used for the choice of optimal number of clusters. Figure 1a shows the percentage of variance explained by the clusters against the number of clusters. At K=4 or 5, the marginal gain began to drop substantially, giving an angle in the graph. The number of clusters was chosen at the "elbow" K=4. The multi-dimensional Scaling (MDS) plot is shown in Figure 1b where clusters were highlighted by different colors. The x-means clustering (Pelleg & Moore, 2000), an alternative and variation of k-means, and PAMSAM algorithm were also applied to our data set. However, in terms of clustering, three methods did not give a significant difference.

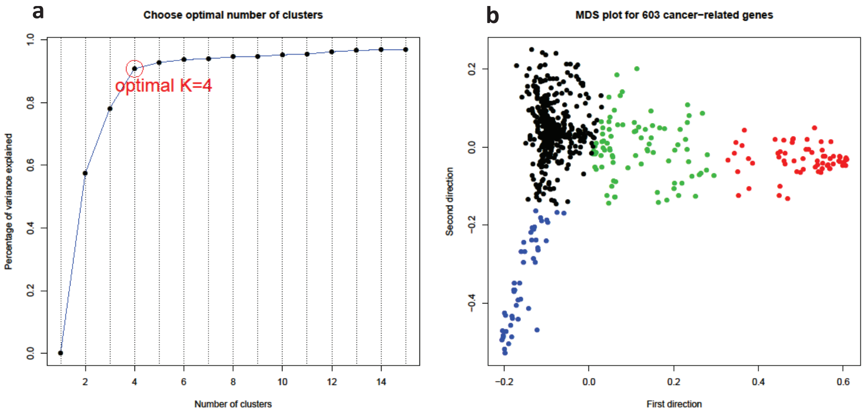


Figure 1. Four gene clusters. (a) The proportion of variance that can be explained by clustering (y-axis) against the number of clusters (x-axis) based on different values of k (k=1,2,...,15) by k-means clustering method. From this plot, the most likely number of clusters is four. (b) Multidimensional scaling (MDS) plots based on correlation dissimilarity metric among 603 genes, where genes in different clusters were highlighted by different colors.

### *2.4 Cox Proportional Hazard Model with Sparse Group Lasso Penalty*

The last step of the pipeline conducts pathway level selection of prognostic genes. A natural way is to fit a regression model with group lasso penalty where each group represents a pathway. The group lasso, however, only works for large number of groups and gives a sparse set of groups. We therefore turned to a sparse group lasso (SGL), which generates a solution balancing both between-group and within-group sparsity. A Cox proportional hazards model and sparse group lasso regularization were then pieced together for further variable selection.

Let $p$ be the number of genes, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)^T$ be the $n \times p$ data matrix, where $\mathbf{X}_i = (X_{i1}, X_{i2}, ..., X_{ip})^T$. Let $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)^T$ denote an n-dimensional vector which corresponds to failure/censor times. Let $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$ be the vector of coefficients, and $\boldsymbol{\delta} = (\delta_1, \delta_2, ..., \delta_n)$ be the censoring indices, where $\delta_i = 1$ indicates event (death) occurred for subject i and $\delta_i = 0$ indicates censoring. The Cox proportional hazards model can be written as follows:

$$\log \frac{\lambda(t|\mathbf{X}_i)}{\lambda_0(t)} = \mathbf{X}_i^T\boldsymbol{\beta}.$$

where $\lambda_0(t)$ stands for the baseline hazard function. The loglikelihood can be written as follows:

$$\ell(\boldsymbol{\beta}) = \frac{1}{n}\{\log(\sum_{i:\delta_i=1}(\sum_{j:Y_j\geq Y_i}\exp(\mathbf{X}_j^T\boldsymbol{\beta}) - \mathbf{X}_j^T\boldsymbol{\beta}))\}$$

With the assumption of sparsity, the parameters $\boldsymbol{\beta}$ can be estimated through a SGL penalized likelihood:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n}\{\log(\sum_{i:\delta_i=1}(\sum_{j:Y_j\geq Y_i}\exp(\mathbf{X}_j^T\boldsymbol{\beta}) - \mathbf{X}_j^T\boldsymbol{\beta}))\} + (1-\alpha)\lambda\sum_{k=1}^{K}\sqrt{p_k}\|\boldsymbol{\beta}^k\|_2 + \alpha\lambda\|\boldsymbol{\beta}\|_1,$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ denote $\ell_1$-norm and $\ell_2$-norm respectively, and $p_k$ represents the size of group k and $\boldsymbol{\beta}^k$ represents the coefficients of genes in group k. The SGL fit is simply a combination of the lasso and group lasso penalties ($\alpha = 0$ gives the group lasso fit, $\alpha = 1$ gives the lasso fit). In practice, one should choose $\alpha$ before the parameter estimation. In our problem, we expect a strong overall sparsity but encourage grouping, therefore a $\alpha = 0.8$ was used. Here the choice of $\alpha$ is different from the choice of $\lambda$, which can be determined by data-driven method. In practice, the mixing rate $\alpha$ need to be predefined depending on the expected overall sparsity and group sparsity. Given two tuning parameters $\alpha$ and $\lambda$, a routine blockwise coordinate descent (BCD) approach can solve the optimization problem and we implemented the BCD algorithm using R package *SGL* (Simon, Friedman, Hastie, & Tibshirani, 2013). A sequence of ten candidate $\lambda$'s with $\lambda_{\min} = 0.05\lambda_{\max}$ in the regularization path was used, as suggested by R package *SGL*.

In the lasso-type problems, the common method for selecting the tuning parameter $\lambda$ is cross-validation. However, it tends to yields a large number of false positives in the sparse network problem, as pointed out by Fu and Zhou in their seminal paper (Fu & Zhou, 2013). Fu and Zhou proposed an "elbow method" that outperforms the cross-validation method, where the optimal tuning parameter corresponds to the change point at which an increase of $\lambda$ does not yield a substantial decrease of log-likelihood. In our Cox model with SGL regularization, the optimal lambda selected by this rule is $\lambda = 0.000492$ as shown in Figure 2 and 115 genes were identified in the final list.

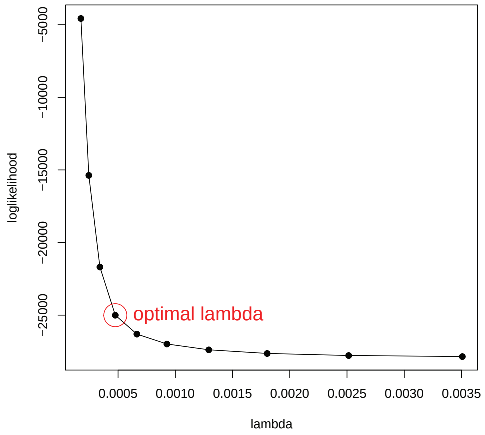


Figure 2. Selection of tuning parameter for the penalty term is sparse group lasso regression. The log-likelihood (y-axis) against tuning parameter (x-axis) for the sparse group lasso penalty, where the optimal $\lambda$ is circled.

## 3. Results and Discussion

### *3.1 Gene Clusters*

Using the k-means approach, the set of 603 genes from initial screening were further clustered into four subgroups. Gene functions in each cluster were investigated. Interestingly, we found that genes within the same cluster tend to have similar/related cellular functions. For instance, cluster 1 (black dots in Figure 1b), the core cluster containing 407 genes including *CENPJ* and *CDK5RAP2*, is functionally related to cell cycle, spindle formation, and mitosis etc. Cluster 2 (blue dots in Figure 1b) contains 48 genes including *MYOG* and *CDK5R2*, mostly related to protein binding and transmembrane activity etc. Cluster 3 (green dots in Figure 1b), containing 85 genes including *COL5A2* and *COL8A2*, corresponds to the pathways related to collagen biosynthesis and enzymes modification etc. Cluster 4 (red dots in Figure 1b), containing 63 genes including *CD48* and *CD53*, is related to immune response and T-cell and B-cell development. This finding indicates that certain cellular pathways/functions may play crucial roles in the progression of the serous ovarian cancer, which may provide new clues for the cancer prevention and treatment.

Table 2. List of 115 identified prognostic genes and corresponding coefficients in the Cox model.

| Gene | Group | $\hat{\beta}$ | Gene | Group | $\hat{\beta}$ | Gene | Group | $\hat{\beta}$ |
|---|---|---|---|---|---|---|---|---|
| ADCK1 | 1 | -0.150 | LOC389458 | 1 | -0.139 | TIPRL | 1 | 0.118 |
| ADH4 | 1 | 0.139 | MAP9 | 1 | 0.255 | TOP2B | 1 | 0.162 |
| ADORA2A | 1 | 0.107 | MBL2 | 1 | -0.096 | TTBK2 | 1 | -0.094 |
| ARHGEF12 | 1 | -0.134 | MGC27348 | 1 | 0.140 | VAMP4 | 1 | 0.201 |
| ATP4B | 1 | -0.188 | MRGPRX4 | 1 | -0.227 | VPS29 | 1 | 0.094 |
| BOLA3 | 1 | 0.111 | MRPS22 | 1 | 0.199 | WDFY3 | 1 | 0.120 |
| C1orf75 | 1 | 0.107 | NARF | 1 | 0.204 | WTAP | 1 | -0.213 |
| C4orf27 | 1 | 0.127 | NDUFB4 | 1 | -0.165 | YSK4 | 1 | 0.129 |
| C6orf115 | 1 | 0.090 | NEK1 | 1 | 0.175 | YY1AP1 | 1 | -0.082 |
| CACNA1S | 1 | -0.084 | NEK2 | 1 | -0.091 | ZFHX2 | 1 | -0.094 |
| CDK5RAP2 | 1 | -0.096 | NEK9 | 1 | -0.107 | ZNF167 | 1 | -0.080 |
| CNGB1 | 1 | -0.124 | NLRX1 | 1 | 0.090 | ZNF197 | 1 | -0.164 |
| COX17 | 1 | 0.109 | NRAS | 1 | -0.136 | ZNF621 | 1 | -0.117 |
| CPA2 | 1 | -0.085 | NSL1 | 1 | 0.119 | ZNF782 | 1 | -0.146 |
| CRIPT | 1 | 0.120 | OR7D4 | 1 | 0.123 | CTRB2 | 2 | -0.103 |
| CRY1 | 1 | 0.170 | OR9Q1 | 1 | 0.176 | DRD3 | 2 | -0.144 |
| DEDD2 | 1 | -0.133 | OS9 | 1 | 0.153 | LCE3A | 2 | -0.092 |
| DLG3 | 1 | -0.117 | PALB2 | 1 | -0.169 | LMAN1L | 2 | -0.105 |
| DNAH7 | 1 | -0.087 | PLEKHH1 | 1 | 0.191 | OR2G2 | 2 | 0.102 |
| DNAI1 | 1 | -0.118 | POMP | 1 | -0.094 | TAAR8 | 2 | 0.208 |
| DNAJC19 | 1 | 0.164 | PPP2R2B | 1 | 0.128 | CLEC4A | 3 | 0.105 |
| DNAJC5 | 1 | 0.087 | RNF12 | 1 | -0.104 | CTSS | 3 | 0.086 |
| DNASE1 | 1 | -0.105 | RNF181 | 1 | 0.093 | EBI3 | 3 | 0.099 |
| ELA2A | 1 | 0.154 | RNF20 | 1 | 0.175 | LY86 | 3 | 0.080 |
| EPB41 | 1 | -0.082 | RNF31 | 1 | -0.144 | RNASE6 | 3 | 0.096 |
| EWSR1 | 1 | -0.197 | RNF7 | 1 | 0.248 | SRGN | 3 | 0.099 |
| EXOSC8 | 1 | 0.152 | RPL21 | 1 | -0.100 | ABCG5 | 4 | 0.137 |
| FAM86B1 | 1 | -0.083 | SCYL1BP1 | 1 | 0.178 | AP1B1 | 4 | -0.224 |
| GAS2L2 | 1 | -0.231 | SEBOX | 1 | 0.093 | CD248 | 4 | -0.080 |
| GHRH | 1 | -0.108 | SEC22B | 1 | 0.083 | CIDEA | 4 | -0.131 |
| GRM6 | 1 | -0.090 | SFT2D1 | 1 | 0.087 | COL8A2 | 4 | -0.109 |
| GSTA3 | 1 | 0.192 | SLC8A2 | 1 | 0.168 | FABP4 | 4 | -0.098 |
| HEXDC | 1 | -0.097 | SNRPG | 1 | -0.143 | GPBAR1 | 4 | -0.137 |
| HMOX2 | 1 | -0.127 | SPN | 1 | 0.228 | GRN | 4 | -0.087 |
| JUB | 1 | -0.095 | SRrp35 | 1 | 0.135 | OAS1 | 4 | 0.082 |
| KIAA0323 | 1 | -0.108 | STAT2 | 1 | 0.124 | OASL | 4 | 0.091 |
| KIF27 | 1 | -0.099 | TBPL1 | 1 | 0.094 | TIMP4 | 4 | 0.201 |
| KIF4B | 1 | -0.097 | KLHL22 | 1 | -0.100 | ZNF660 | 4 | -0.120 |

### *3.2 Prognostic Gene Identification*

Using the Cox model with sparse group lasso penalty, we obtain a final list of 115 prognostic genes (in Table 2), of which many were reported to be involved in cancer initiation and progression. To name a few, gene *CTSS* is closely related to gastric cancer and silencing *CTSS* expression suppressed the migration and invasion of gastric cancer cells (Yang et al., 2010). Gene *CD248* can facilitate tumor growth via its cytoplasmic domain and multiple pathways regulated by the cytoplasmic domain of *CD248* highlight its potential as a therapeutic target to treat cancer (Maia et al., 2011). Gene

*DRD3* is a dopamine receptor, whose expression can change as stress factors associated with breast cancer (Pornour et al., 2014). Gene *CDK5RAP2* is required for spindle checkpoint function and is a common target in paclitaxel and doxorubicin. Cancer cells cultured in the presence of paclitaxel or doxorubicin exhibit a dramatic decrease in *CDK5RAP2* levels (Zhang et al., 2009).

We also identify several subgroups (families) of genes whose associations with cancer have been reported. For instance, three genes in the NEK family, *NEK1*, *NEK2* and *NEK9*, were identified in our final list. Mutations of NEK family members have also been identified as drivers behind the development of ciliopathies and cancer. Recent emergence of comprehensive cancer genomes is highlighting certain members of the NEK family as targets of frequent mutations (Moniz, Dutt, Haider, & Stambolic, 2011). We also identified five genes in the RNF family: *RNF12, RNF181, RNF20, RNF31,RNF7*. A recent study reported that the RNF family such as *RNF20* drives histone *H2B* monoubiquitylation and modulates inflammation and inflammation-associated cancer in mice and humans (Tarcic et al., 2016).

The predictive power of our 115-gene signature was illustrated using Kaplan-Meier curves in Figure 3, where the samples were equally divided into two groups based on the hazard risk. The moderate separation of two groups demonstrates the effectiveness of our method.

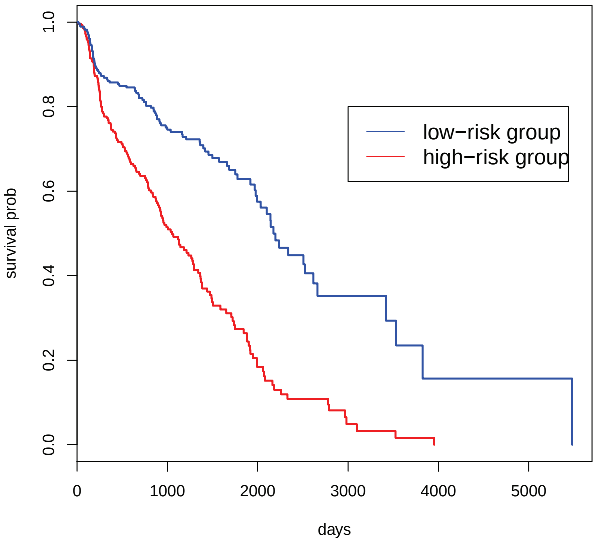


Figure 3. Kaplan-Meier curves. Survival probability against time (in days) of different groups by the hazard risk based on the 115-gene signature. The red and blue lines are based on high-risk group and low-risk group, respectively.

## 4. Conclusion

In this paper, we developed a flexible three-step computational pipeline for identifying prognostic biomarkers related to the overall survival of serous ovarian cancer patients using the rich TCGA data set. This pipeline facilitates the pathway level analysis of the biomarkers associated with cancer survival. The proposed methods are computationally efficient and can be generally applied to many large-scale genomic cancer data sets including the TCGA data. We applied this pipeline to TCGA ovarian cancer data and identified a list of 115 genes, as well as several gene families including the NEK family and RNF family, which may greatly affect the overall survival of ovarian cancer patients. Some of these findings are well supported by literature.

## Acknowledgements

Support has been provided in part by the Arkansas Biosciences Institute, the major research component of the Arkansas Tobacco Settlement Proceeds Act of 2000.

## References

Bolstad, B., Irizarry, R., Astrand, M., & Speed, T. (2002). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics, 19*(2). http://dx.doi.org/10.1093/bioinformatics/btg202

Chen, L., Xuan, J., Gu, J., Wang, Y., Zhang, Z., Wang, T., & Shih, L. (2012). Integrative network analysis to identify aberrant pathway networks in ovarian cancer. *Pacific Symposium Biocomputing, 31*. http://dx.doi.org/10.1142/978981436

6496-0004

Fu, F., & Zhou, Q. (2013). Learning Sparse Causal Gaussian Networks With Experimental Intervention: Regularization and Coordinate Descent. *Journal of American Statistical Association, 108*(501), 288-300. http://dx.doi.org/10.1080/01621459.2012.754359

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology, 7* (3), 601-20. http://dx.doi.org/10.1145/332306.332355

Haindl, M., Somol, P., Ververidis, D., & Kotropoulos, C. (1999). Feature Selection Based on Mutual Correlation. Technical Report

Hsu, F., Serpedin, E., Hsiao, T., Bishop, A., Dougherty, E., & Chen, Y. (2012). Reducing confounding and suppression effects in TCGA data: an integrated analysis of chemotherapy response in ovarian cancer. *BMC Genomics, 13*. http://dx.doi.org/10.1186/1471-2164-13-s6-s13

Jouve, P. E., & Nicoloyannis, N. (2010). A Filter Feature Selection. Technical Report

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*, 273-324.

Kumar, G., Breen, E. J., & Ranganathan, S. (2013). Identification of ovarian cancer associated genes using an integrated approach in a Boolean framework. *BMC System Biology, 7* (12). http://dx.doi.org/10.1186/1752-0509-7-12

Konstantinopoulos, P. A., Spentzos, D., & Cannistra, S. A. (2008). Gene-expression profiling in epithelial ovarian cancer. *Nature Clinical Practice Oncology, 5*, 577-87. http://dx.doi.org/10.2174/138920206777304641

Leng, J., Valli, C., & Armstrong, L. (2010). A Wrapper-based Feature Selection for Analysis of Large Data. Technical Report

Moniz, L., Dutt, P., Haider, N., & Stambolic, V. (2011). Nek family of kinases in cell cycle, checkpoint control and cancer. *Cell Division, 6*(18). http://dx.doi.org/10.1186/1747-1028-6-18

Matveeva, E., Maiorano, J., Zhang, Q., Wang, J. P., & Fondufe-Mittendorf, Y. (2016). Involvement of PARP1 in the regulation of alternative splicing. *Cell Discovery, 2*(15046). http://dx.doi.org/10.1038/celldisc.2015.46

McLaughlin, J., Rosen, B., Moody, J., Pal, T., Fan, I., Shaw, R., Narod, S. (2013). Long-term ovarian cancer survival associated with mutation in BRCA1 or BRCA2. *Journal of National Cancer Institute, 105*(2). http://dx.doi.org/10.1093/jnci/djs494

Nagle, C., ChenevixTrench, G., Webb, P., & Spurdle, A. (2007). Ovarian cancer survival and polymorphisms in hormone and DNA repair pathway genes. *Cancer Letters, 251*(1). http://dx.doi.org/10.1016/j.canlet.2006.11.011

Pelleg, D., & Moore, A. (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *ICML Proceedings of the Seventeenth International Conference on Machine Learning*. http://dx.doi.org/10.1007/3-540-44491-2-3

Popovic, R., Martinez-Garcia, E., Giannopoulou, E. G., Zhang, Q., Zhang, Q., Ezponda, T., & Licht, J. (2014). Histone Methyltransferase MMSET/NSD2 Alters EZH2 Binding and Reprograms the Myeloma Epigenome through Global and Focal Changes in H3K36 and H3K27 Methylation. *PLoS Genetics, 10*(9). http://dx.doi.org/10.1371/journal.pgen.1004566

Pornour, M., Ahangari, G., Hejazi, S., Ahmadkhaniha, H., & Akbari, M. (2014). Dopamine receptor gene (DRD1-DRD5) expression changes as stress factors associated with breast cancer. *Asian Pacific Journal of Cancer Prevention, 15*(23). http://dx.doi.org/10.7314/apjcp.2014.15.23.10339

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics, 22*. http://dx.doi.org/10.1080/10618600.2012.681250

Tarcic, O., Paterals, I., Cooks, T., Shema, E., Kanterman, J., & Oren, M. (2016). RNF20 Links Histone H2B Ubiquitylation with Inflammation and Inflammation-Associated Cancer. *Cell Reports, 14*(6). http://dx.doi.org/10.1016/j.celrep.2016.01.020

The Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature, 474*. http://dx.doi.org/10.1158/2159-8290.cd-rw042711-14

Xu, Y., Zhang, J., Yuan, Y., Mitra, R., Muller, P., & Ji, Y. (2012). A Bayesian graphical model for integrative analysis of TCGA data. *2012 IEEE International Workshop on Genomic Signal Processing and Statistics, 31*. http://dx.doi.org/10.1017/cbo9781107706484.010

Xi, L., Brogaard, K., Zhang, Q., Lindsay, B., Widom, J., & Wang, J. P. (2014). A locally convoluted cluster model for nucleosome positioning signals in chemical maps. *Journal of the American Statistical Association, 109*(505). http://dx.doi.org/10.1080/01621459.2013.862169

Yang, Y., Lim, S., Choong, L., Lee, H., Chen, Y., Chong, P., & Lim, Y. (2010). Cathepsin S mediates gastric cancer cell migration and invasion via a putative network of metastasis-associated proteins. *Journal of Proteome Research, 9* (9). http://dx.doi.org/10.1021/pr100492x

Zhang, X., Liu, D., Lv, S., Wang, H., Zhong, X., Liu, B., & Xu, X. (2009). CDK5RAP2 is required for spindle checkpoint function. *Cell Cycle, 8*(8) http://dx.doi.org/10.4161/cc.8.8.8205

Zhang, Q., Burdette, J. E., & Wang, J. P. (2014). Integrative network analysis of TCGA data for ovarian cancer. *BMC Systems Biology, 8* (1338), 1-18. http://dx.doi.org/10.1186/s12918-014-0136-9

Zhang, Q., & Wang, J. P. (2016). A Bayesian network approach for modeling mixed features in TCGA ovarian cancer data. *Handbook of Mathematical Methods in Cancer Biology, 1.*

Zhang, Q. (2015). Learning Sparse Bayesian Network with Mixed Variables and its Application to Cancer Systems Biology. Unpublished PhD Dissertation, Northwestern University.

**Copyrights**

International Journal of Statistics and Probability; Vol. 5, No. 6; November 2016
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

# Seasonal Modelling of Fourier Series with Linear Trend

Iberedem A. Iwok[1]

[1] Department of Mathematics/Statistics, University of Port-Harcourt, P.M.B.5323, Port-Harcourt, Rivers State, Nigeria.

Correspondence: Iberedem A. Iwok, Department of Mathematics/Statistics, University of Port-Harcourt, P.M.B.5323, Port-Harcourt, Rivers State, Nigeria. E-mail: ibywok@yahoo.com

Received: July 28, 2016 Accepted: September 6, 2016 Online Published: October 26, 2016

doi:10.5539/ijsp.v5n6p65 URL: http://dx.doi.org/10.5539/ijsp.v5n6p65

## Abstract

This work was motivated by the need to model a periodic time series function with linear trend. A Fourier series representation with detrended linear function was proposed. In this representation, the time series $X_t$ is expressed as a combination of the linear trend component and a linear combination of $s$ orthogonal trigonometric functions; where $s$ is the number of seasons. The method was applied to a rainfall data and the proposed model was found to give a good fit. Comparative study was carried out with the complete Fourier representation. Diagnostic checks revealed that the proposed method performs better the pure Fourier approach.

**Keywords:** Fourier series, seasonal model, linear trend, periodogram, spectral density and white noise process.

## 1. Introduction

The usual autoregressive integrated moving average (ARIMA) models developed by Box and Jenkins (1970) has been extensively used in modelling linear time series. The ARIMA models assume that the current observation depends on weighted previous observations, weighted previous random shocks and the current shock. However, most time series arising in nature do not assume linearity but rather, periodic or seasonal with linear trend. Seasonal time series contain a seasonal phenomenon that repeats itself after a regular period of time. Such phenomena stem from factors such as weather, which affects many business and economic activities, cultural events and graduation ceremonies. Series with seasonal pattern cannot be adequately represented by ARIMA models. To analyze such series, Wold (1974) arranged the series in a two dimensional table according to the season; and the totals and averages were computed. In Wold (1974) representation, a time series is thought to consist of trend-cycle, seasonal and irregular components. To estimate these components, several decompositions are usually involved. Box, Jenkins and Reinsel (2008) made an extension of the Box and Jenkins (1970) ARIMA models to include the seasonal part and is called the seasonal autoregressive integrated moving average (SARIMA) models. Despite these efforts, the models do not adequately represent most periodic series.

A better procedure extensively used for modelling periodic time series is the Fourier analysis. This method represent the time series by a set of elementary functions called basis such that all functions under study can be written as linear combinations of the elementary functions in the basis. These elementary functions involve the sine and cosine functions or complex exponentials. The Fourier series approach describes the fluctuation of time series in terms of sinusoidal behaviour at various frequencies. Despite the wider acceptability of the method, however, Fourier approach still suffers some set backs. One major problem associated with it is the cumbersomeness in Fourier representation and non inclusion of trend component. As will be seen in the methodology, the inconveniences in representing the time series is enormous if we are to include all the terms required in Fourier series. This cannot go well with a series of large sample size because representing all the terms will consume several pages and can be boring to both the researcher and the reader. Hence, there is need to shorten the number of terms in the Fourier expression and give a summarized representation that adequately describe the time series. This is the intent of this work. As earlier stated, seasonal variations in time series can be caused by climatic factor and we are going to use rainfall data in our illustration.

## 2. Literature Review

The need for accurate rainfall prediction is necessary when considering the importance in which such information would give for river control, reservoir operation, forestry interests, flood mitigation, etc. Due to the numerous benefits of rainfall modelling and prediction, studies on rainfall analysis have been on the fore front in the research world.

Afshar, Joshua, Buckman, and Samuel (2014) modelled rainfall data using ARIMA model and artificial neural networks (ANNs). The ARIMA was found to give discouraging result. However, with application of artificial neural network to the

ARIMA component, the combined model was found to be adequate and forecast were generated.

Cenis (1989) studied temperature in solarized soil using Fourier analysis. He obtained the daily maximum and minimum temperatures at two dept on daily basis for three summer periods. He used the values to fit sinusoidal equations which accounted for 93% variation. The variation and the hourly mean differences between the measured temperatures were calculated. The analysis gave an overall encouraging result.

Serangelo, Ferrari and De Luca (2011) applied non-homogeneous Poisson process to examine the seasonal effects of daily rainfall. The modelling process involved the partitioning of observed daily rainfall data into calibration periods for the estimation of parameters. Though the validation period for checking the occurrence process changed; the model which was applied to the set of rain gauges placed at different geographical areas was shown to provide good fit.

Falahah and Suorapto (2010) carried out research on rainfall data using analytic factor method. The data was obtained from 50 weather stations for a period of 30 years. The result was plotted on pattern factors to reveal dominant factor for each region and inspection period. The method explained factors that influence rainfall in Indonesia and the reasons for having relatively high humidity in one area than the other.

Necholas, Mahmood and Hazan (2013) modelled rainfall data amounts for agriculture planning using gamma distribution models. Daily rainfall data of two stations having two different mean annual rainfalls were analyzed. Generalized linear models were used to fit smooth regression curves. The mean amount of rain per rainy day was computed using the estimates of parameters of the model for each day of the season. The adequacy of the fitted model was check by the analysis of deviance residuals and was found to be satisfactory. Fourier approach was employed for comparative study. It was discovered that though reasonable results were obtained, Fourier analysis was time consuming and boring. However, Fourier series was found suitable in fitting gamma distribution for the determination of mean rain per rainy day.

Zakaria (2013) conducted a study on periodic and stochastic modelling of monthly rainfall and the periodicities were determined. Stochastic components were estimated using the auto-regressive model approach. Residuals obtained from the model were shown to follow a white noise process; thus indicating the adequacy of the fitted model.

Beatrice, Nasser, Afshar, Selaman and Fahmi (2014) analyzed data from eight rain gauge stations. Annual rainfall data for 27 years were computed with the Fourier series equation. The result was compared with that obtained from harmonic series models. It was discovered that both models were capable of describing rainfall pattern and were able to provide reasonable relationship between the simulated and the observed data.

Akpanta, Okorie and Okoye (2015) adopted SARIMA modelling of the frequency approach in analyzing monthly rainfall data in Umuahia. Probability time series approach was considered. The original data plotted showed seasonality which was removed by differencing. After subjecting the model to diagnostic checks, SARIMA $(0,0,0)(0,1,1)_{12}$ was found to fit the data well and was used for prediction.

## 3. Methodology

In this method, a periodic time series is first observed whether it contains a linear trend or not. Visual inspection of the raw data plot can reveal this pattern. Assuming a linear trend is detected, a linear regression model of the form

$$Y_t = \beta_0 + \beta_1 t + e_t \qquad (1)$$

is first fitted to the data;

where $Y_t$ is the observed time series, $t$ is the time points $(t = 1,2,\dots,n)$, $n$ is the number of observations, $\beta_0$ and $\beta_1$ are the regression parameters, and $e_t$ is the error component.

Fitting the above model (1) to the data $Y_t$ , we can obtain the estimate of the error component

$$\widehat{e_t} = \widehat{Y_t} - \widehat{\beta_0} - \widehat{\beta_1}t$$

which can be tested for randomness or white noise.

After obtaining the trend equation ( i.e. $\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}t$ ), the main series $Y_t$ is detrended by the expression

$$y_t = Y_t - \widehat{Y_t} = Y_t - \widehat{\beta_0} - \widehat{\beta_1}t \qquad (2)$$

The resulting series $y_t$ is then used to fit seasonal model using Fourier representation.

### 3.1 *Fourier Series Representation of the Time Series* $\boldsymbol{y_t}$

Given a time series of $n$ observations, the Fourier representation is the set of $q$ orthogonal trigonometric functions shown below:

$$y_t = \sum_{i=1}^{q}(\alpha_i \cos 2\pi f_i t \ + \beta_i \sin 2\pi f_i t \ ) + e_t \qquad (3)$$

estimated by

$$\widehat{y_t} = \sum_{i=1}^{q}(a_i \cos 2\pi f_i t \ + b_i \sin 2\pi f_i t \ ) \tag{4}$$

where $q = {}^{n}\!/_{2}$, $a_i = \frac{2}{n}\sum_{t=1}^{n} y_t \cos 2\pi f_i t$, $b_i = \frac{2}{n}\sum_{t=1}^{n} y_t \sin 2\pi f_i t$,

$e_t \sim NIID(0, \sigma^2)$; period $= p_i = {}^{n}\!/_{i}$ and $f_i = {}^{i}\!/_{n}$ is the $i^{th}$ harmonic of the fundamental frequency ${}^{1}\!/_{n}$.

*3.2 The Peridogram*

The periodogram is defined as the function of intensities $I(f_i)$ at frequency $f_i = i/n$ and is given as

$$I(f_i) = \frac{n}{2}(a_i^2 + b_i^2) \qquad ; \qquad i = 1,2,\dots,q.$$

Periodogram is the plot of the intensities against the frequencies or periods. The periodogram $I(f_i)$ is simply the sum of squares associated with the pair of coefficients $(a_i, b_i)$ and hence with the frequency $f_i$ or period $p_i$. That is,

$$\sum_{t=1}^{q}(y_t - \overline{y})^2 = \sum_{t=1}^{n/2} I(f_i).$$

In the context at hand, the periodogram is used to determine the seasonality or periodicity of a time series. This is usually indicated by the largest peak in the periodogram plot.

*3.3 The Spectrum*

The sample spectrum is obtained by allowing the frequency $f$ to vary continuously in the range 0 to 0.5 cycle so that the periodogram can be re-defined as

$$I(f) = \frac{n}{2}\left(a_j^2 + b_j^2\right) \qquad ; \qquad 0 \le f \le 0.5.$$

The function $I(f)$ is called the spectrum.

*3.4 Autocorrelation Function*

This is the plot of autocorrelation at lag *k* ($\rho_k$) versus *k*.

*3.5 Spectral Density Function*

Spectral density is the Fourier transform of the auto-correlation function and is estimated by

$$g(f) = 2[1 + \sum_{k=1}^{\infty} \rho_k \, cos(2\pi f k)] \qquad ; \qquad 0 \le f \le 0.5$$

where $\rho_k$ is the autocorrelation at lag *k*. The spectral density performs the same function as the periodogram. The period or seasonality of a time series is obtained at where the spectral density is maximum.

*3.6 White Noise Process*

A process $\{\varepsilon_t\}$ is said to be a white noise process with mean 0 and variance $\sigma_\varepsilon^2$ written $\{\varepsilon_t\} \sim WN(0, \sigma_\varepsilon^2)$, if it is a sequence of uncorrelated random variables from a fixed distribution.

*3.7 The Seasonal Fourier Representation*

Rather than fitting the entire Fourier series expression in equation (3), we fit only the Fourier terms up to the season detected by the periodogram. That is, suppose the season determined by the periodogram in the detrended series $y_t$ is $s$, then equation (3) reduces to

$$y_t = \sum_{i=1}^{s}(\alpha_i \cos 2\pi f_i t \ + \beta_i \sin 2\pi f_i t) + \varepsilon_t \tag{5}$$

$$\Rightarrow \widehat{Y_t} = \widehat{\beta_0} + \widehat{\beta_1} t + \sum_{i=1}^{s}(a_i \cos 2\pi f_i t \ + b_i \sin 2\pi f_i t) \tag{6}$$

and

$$\widehat{\varepsilon_t} = Y_t - \widehat{Y_t} \tag{7}$$

Comparatively, the expression (5) is less cumbersome in carrying out analysis than the complete Fourier form expressed in equation (3). The model (5) can be fitted to any periodic or seasonal data and the estimated residuals $\widehat{\varepsilon_t}$ obtained from (7) can be tested for white noise to determine whether the model is adequate or not.

## 4. Data Analysis and Result

The data used for this work is the average monthly rainfall data ($Y_t$) in Calabar, Nigeria between 2005-2015 (Source: www.cbn.gov.ng); and the analysis is carried out using Minitab and gretl softeware.

### *4.1 Complete Fourier Series Model*

Fitting the full Fourier series in equation (3) where $q = \frac{120}{2} = 60$ result in a residual variance of 11.23 and the Fourier coefficients are displayed in Appendix C. The residual autocorrelation function is displayed in figure 5. Clearly, there is a significant spike at lag 12 ( $\rho_k = -0.36$). This shows that the residuals are correlated (at lag 12) and hence do not follow a white noise process. The actual and estimate values plots displayed in figure 6 shows a low correlation between these values. Thus, the full Fourier series, despite it cumbersome nature does not fit adequately to the data.

### *4.2 The Proposed Approach*

#### 4.2.1 Seasonality and the Estimated Trend

The raw data plot in figure 1 clearly shows the existence of seasonality and trend. This is indicated by the periodic pattern and upward movement of the graph. Fitting the trend equation gives the Minitab output in table 1 below.

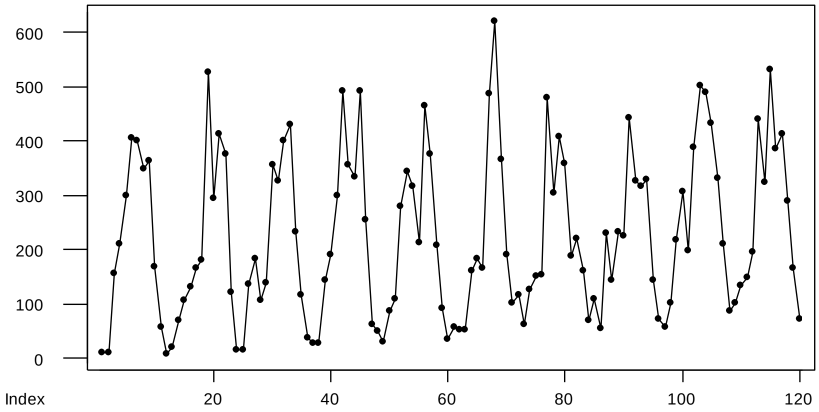


Figure 1. Raw data plot of the series $Y_t$

Table 1. Minitab Output for the Trend equation

| Predictor | Coef | StDev | T | P |
|---|---|---|---|---|
| Constant | 185.28 | 26.96 | 6.87 | 0.000 |
| t | 0.6639 | 0.3898 | 1.70 | 0.031 |

The regression equation is Yt = 185 + 0.6639 t

The *p*–values in table 1 shows that both the constant term and the coefficient of $t$ are significant. Thus, the estimated trend equation is

$$\widehat{Y_t} = 185.28 + 0.6639t$$

Next, we obtain the new series from equation (2) as

$$y_t = Y_t - \widehat{Y}_t = Y_t - 185.28 - 0.6639t$$

The new series $y_t$ now becomes our working data.

#### 4.2.2 The Peridogram and Spectral Density of $\boldsymbol{y_t}$

The periodogram analysis was conducted using the gretl software. The periodogram is displayed in figure 2. The values for the spectral densities, periods and frequencies are equally displayed in appendix A. From the table of appendix A and figure 2; it is observed that the periodogram is dominated by a very large peak at scaled frequency, $f_i^* = i = 10$ ($f_i = i/n = 10/120 = 0.0833$). This is indicated by the largest spectral density of 16.397 shown in bold figures. This density and frequency correspond to a *period* or *season*, $s = n/i = 120/10 = 12$ months. This indicates a 12 – month cycle.

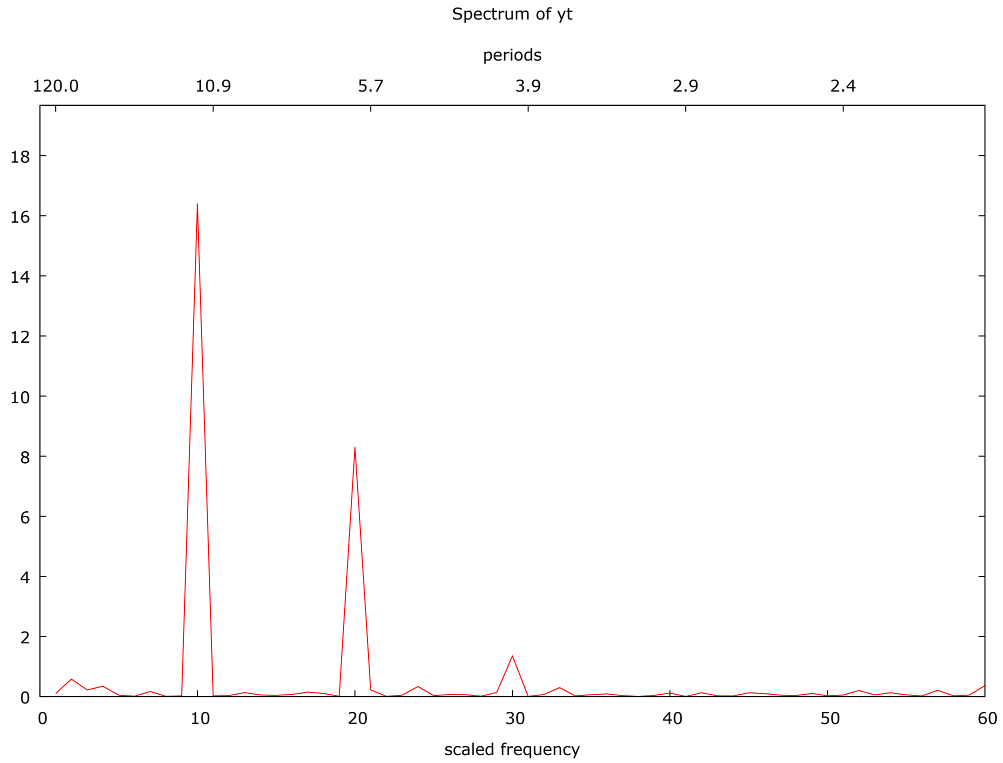


Figure 2. Periodogram plot of $y_t$

4.2.3 The Seasonal Fourier Representation of $\boldsymbol{y_t}$

Rather than use $q = 60$ in equation (3), we reduce the Fourier components to the number of seasons, $s = q = 12$ and apply equation (5). That is,

$$y_t = \sum_{i=1}^{12}(\alpha_i \cos 2\pi f_i t + \beta_i \sin 2\pi f_i t) + \varepsilon_t \tag{8}$$

Equivalently, equation (8) can be expressed as

$$y_t = \sum_{i=1}^{12}(\alpha_i \cos\omega_i t + \beta_i \sin \omega_i t) + \varepsilon_t \tag{9}$$

where, $\omega_i = 2\pi f_i$.

Subjecting equation (9) to regression analysis give the parameter estimates displayed in Appendix B. In appendix B, some coefficients of the variables are not statistically significant (i.e. their $p$ values are less than 0.05) and are therefore excluded in the overall equation.

Hence, from (6), the resulting estimated model is

$$\begin{aligned}\widehat{Y}_t = 185.28 + 0.6639t - 0.621 \sin \omega_1 t - 0.314 \sin \omega_2 t - 0.129 \sin \omega_3 t - 0.105 \sin \omega_4 t \\ -0.085 \sin \omega_5 t - 0.081 \cos \omega_8 t - 0.155 \cos \omega_9 t - 0.074 \sin \omega_{10} t \\ + 0.142 \cos \omega_{11} t - 0.206 \cos \omega_{11} t + 0.094 \cos \omega_{12} t\end{aligned} \tag{10}$$

The residual $\widehat{\varepsilon}_t$ of the fitted is obtained from

$$\widehat{\varepsilon}_t = Y_t - \widehat{Y}_t$$

## 5. Diagnosis

We present here two diagnostic checks to ensure that the proposed model (10) fitted to the data is adequate.

### *5.1 Actual and Estimate Plots*

The overlaid plots of the actual values ($Y_t$) and the estimated values ($\widehat{Y}_t$) is displayed in figure 3. The two superimposed plots move together in the same direction indicating closeness and strong correlation between the values of the two variables. This shows that the model is adequate.

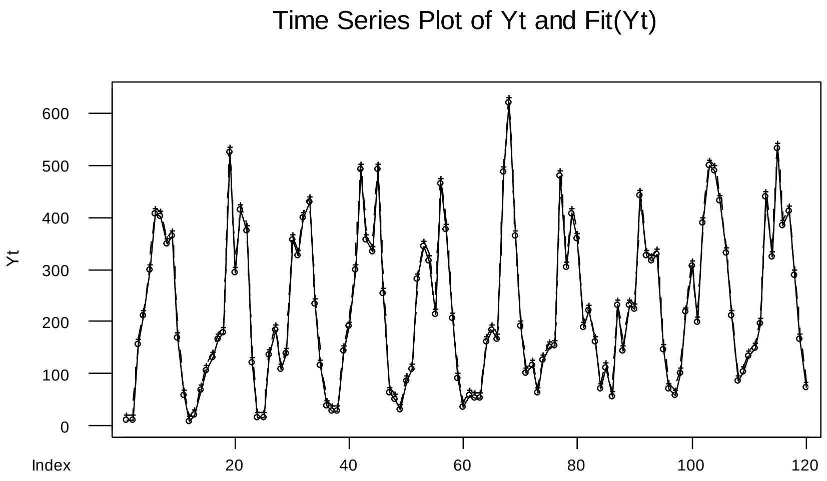
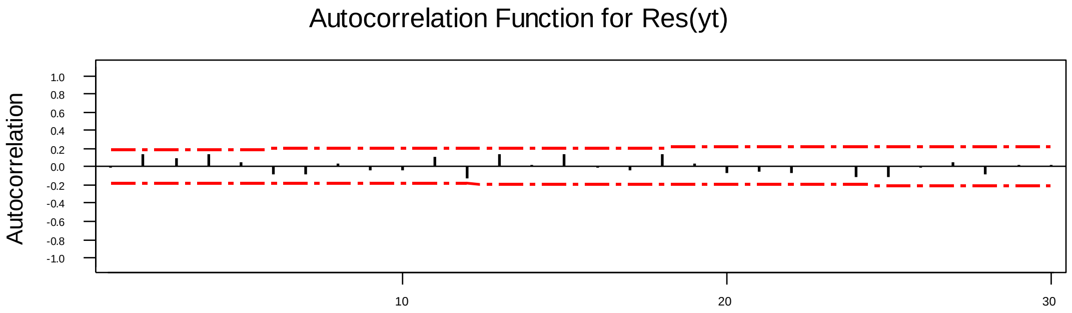


Figure 3. Actual and estimate plot of $Y_t$

*5.2 Residual Variance, Autocorrelation and White Noise*

The residual variance of the fitted model is 4.78. This is significantly smaller than the 11.23 obtained from fitting the full Fourier series. The residual autocorrelation function is displayed in figure 4 and it shows that there is no significant autocorrelation of the residuals. That is, all autocorrelations at all lags are within the range $\pm 2/\sqrt{n}$ (as indicated by the two red lines in figure 4). This means the residuals of the fitted model are not serially correlated. In more precise terms, the residuals follow a white noise process. Hence, the model is adequate.

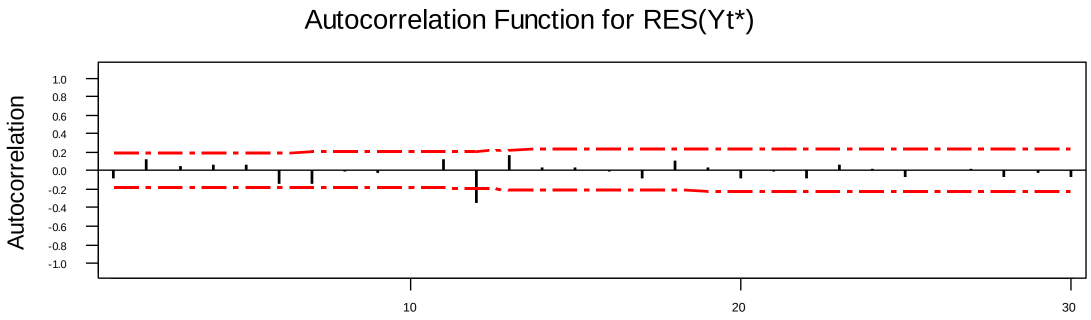


| Lag | Corr | T | LBQ | Lag | Corr | T | LBQ | Lag | Corr | T | LBQ | Lag | Corr | T | LBQ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -0.01 | -0.08 | 0.01 | 10 | -0.05 | -0.47 | 8.44 | 19 | 0.03 | 0.29 | 20.11 | 28 | -0.09 | -0.87 | 27.79 |
| 2 | 0.14 | 1.52 | 2.41 | 11 | 0.11 | 1.11 | 9.99 | 20 | -0.07 | -0.65 | 20.79 | 29 | 0.02 | 0.15 | 27.83 |
| 3 | 0.09 | 0.96 | 3.41 | 12 | -0.14 | -1.39 | 12.51 | 21 | -0.06 | -0.54 | 21.27 | 30 | 0.01 | 0.09 | 27.85 |
| 4 | 0.13 | 1.36 | 5.46 | 13 | 0.13 | 1.30 | 14.81 | 22 | -0.08 | -0.73 | 22.13 | | | | |
| 5 | 0.05 | 0.54 | 5.80 | 14 | 0.01 | 0.11 | 14.83 | 23 | 0.01 | 0.07 | 22.14 | | | | |
| 6 | -0.09 | -0.92 | 6.79 | 15 | 0.13 | 1.30 | 17.25 | 24 | -0.11 | -1.06 | 24.04 | | | | |
| 7 | -0.09 | -0.93 | 7.83 | 16 | -0.01 | -0.08 | 17.26 | 25 | -0.11 | -1.07 | 26.02 | | | | |
| 8 | 0.03 | 0.28 | 7.93 | 17 | -0.04 | -0.40 | 17.50 | 26 | -0.02 | -0.18 | 26.08 | | | | |
| 9 | -0.04 | -0.43 | 8.16 | 18 | 0.13 | 1.28 | 19.97 | 27 | 0.05 | 0.43 | 26.41 | | | | |

Figure 4. Residual Autocorrelation function plot of $\widehat{\varepsilon_t}$

## 6. Discussion and Conclusion

It has been noted that the Fourier series model can only be applied to periodic series that are stationary in mean. If the series contain trend, however, special technique is required for the modelling process. Perhaps, this constituted the problem of Afshar *et al* (2014) that made them to obtain a discouraging result by applying ARIMA model to a periodic data without considering the trend. Besides, as noted by Necholas *et al* (2013), Fourier series modelling is time consuming and boring because of the large number of Fourier coefficients involved. In this work, however, these problems have been addressed. As clearly demonstrated in this method, the trend component of a periodic series is taken care of. Also, setting $q = s$ has reduced the computational burden by 80% and has given adequate Fourier representation as confirmed by the diagnostic checks. It is believed that this work has opened another possibility of addressing periodic functions.



| Lag | Corr | T | LBQ | Lag | Corr | T | LBQ | Lag | Corr | T | LBQ | Lag | Corr | T | LBQ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -0.09 | -1.00 | 1.02 | 10 | 0.00 | 0.01 | 9.71 | 19 | 0.02 | 0.21 | 36.28 | 28 | -0.07 | -0.60 | 40.84 |
| 2 | 0.12 | 1.30 | 2.79 | 11 | 0.11 | 1.16 | 11.46 | 20 | -0.09 | -0.79 | 37.45 | 29 | -0.02 | -0.20 | 40.92 |
| 3 | 0.04 | 0.48 | 3.04 | 12 | -0.36 | -3.67 | 29.42 | 21 | -0.02 | -0.18 | 37.52 | 30 | -0.08 | -0.67 | 41.89 |
| 4 | 0.06 | 0.67 | 3.53 | 13 | 0.17 | 1.51 | 33.18 | 22 | -0.09 | -0.78 | 38.70 | | | | |
| 5 | 0.05 | 0.56 | 3.89 | 14 | 0.03 | 0.25 | 33.29 | 23 | 0.06 | 0.52 | 39.25 | | | | |
| 6 | -0.15 | -1.62 | 6.88 | 15 | 0.03 | 0.29 | 33.44 | 24 | 0.02 | 0.15 | 39.29 | | | | |
| 7 | -0.14 | -1.50 | 9.56 | 16 | -0.02 | -0.15 | 33.48 | 25 | -0.07 | -0.60 | 40.02 | | | | |
| 8 | -0.01 | -0.15 | 9.59 | 17 | -0.09 | -0.82 | 34.66 | 26 | -0.00 | -0.00 | 40.02 | | | | |
| 9 | -0.03 | -0.31 | 9.71 | 18 | 0.10 | 0.92 | 36.20 | 27 | 0.02 | 0.17 | 40.08 | | | | |

Figure 5. Residual Autocorrelation function plot of $\widehat{e_t}$

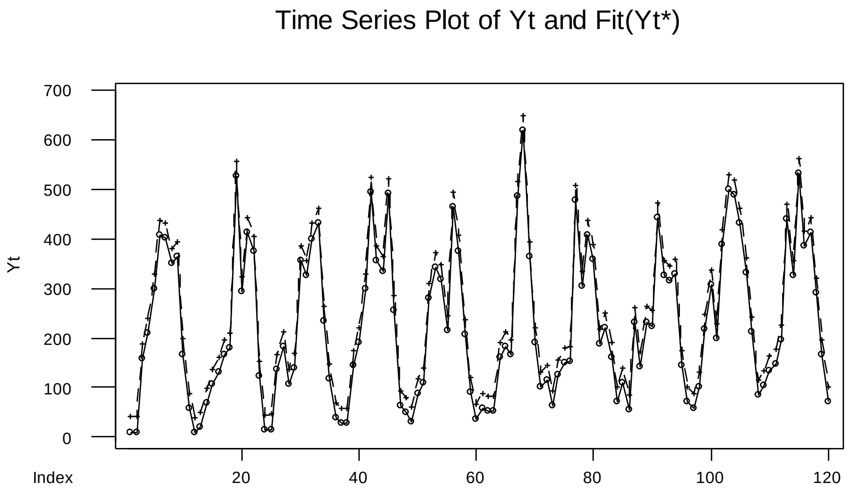


Figure 6. Actual and estimate plot of $Y_t^*$

## References

Afshar, B. N., Joshua, M. A, Buckman, A., & Samuel, T. (2014). Time series modeling of rainfall in new Juaben municipality of the Eastern region of Ghana. *Special issue on contemporoary research in business and social science*, *4*(7), 21-28.

Akpanta, C. A., Okorie, I. E., & Okoye, N. N. (2015). SARIMA modelling of the frequency of monthly rainfall in Umuahia, Abia state of Nigeria. *American journal of mathematics and statistics*, *5*, 82-87.

Beatrice, C. B., Nasser, R., Afshar, K., Selaman, O., & Fahmi, H. (2014). Application of mathematical modelling in rainfall forecast: A case study in SGS. *Sarawak basin. International Journal of Research in Engineering and Technology*, *3*, 2321-7308.

Box, G. E. P., & Jenkins, G. M. (1970). *Time Series Analysis; forecasting and control*. Holden Day. San Francisco, California.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). *Time Series Analysis; Forecasting and Control*. John Willey & Sons.

Cenis, J. L. (1989). Temperature evaluation in solarized soils by Fourier analysis. *Photopathology*, *79*, 506-510. http://dx.doi.org/10.1094/Phyto-79-506

Falahah, M., & Suprapto, S. (2010). Interpretation of rainfall data using analysis factor method. *Proceeding of third* *international conference on mathematics and natural sciences, Indonesia* (pp. 1288-1293).

Necholas, Z., Mahmood, Z., & Hazan, Y. (2013). Modeling the daily rainfall amounts of north-west Pakistan for agricultural planning. *Sarhad J. Agric*, *30*, 5-23**.**

Serangelo, B. A., Ferrari, H. K., & DeLuca, T. (2011), Occurrence analysis of daily rainfall through non-homogeneous Poisson process. *Nat. Hazards Earth Syst. Sci.*, *11*,1657-2011. http://dx.doi.org/10.5194/nhess-11-1657-2011

Wold, H. O. A. (1974). *A study in the analysis of stationary time series*. Almqvist and Wiksell, Uppsala.

Zakaria, A. (2013). A study of periodic and stochastic modelling of monthly rainfall from Purajaya station. *Asian Transactions on Engineering*, *1*(3), 1-7.

## Copyrights

International Journal of Statistics and Probability; Vol. 5, No. 6; November 2016
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

# Estimating Three-way Latent Interaction Effects in Structural Equation Modeling

Birhanu Worku Urge[1], Kepher Makambi[2] & Anthony Wanjoya[3]

[1] Pan African University Institute for Basic Sciences,Technology and Innovation, Nairobi, Kenya

[2] Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University, Washington, DC, USA

[3] Department of Statistics and Actuarial sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Correspondence: Birhanu Worku Urge, Pan African University Institute for Basic Sciences,Technology and Innovation, Nairobi, Kenya. E-mail: biree2016@gmail.com

Received: September 7, 2016 Accepted: September 18, 2016 Online Published: October 26, 2016

doi:10.5539/ijsp.v5n6p73 URL: http://dx.doi.org/10.5539/ijsp.v5n6p73

**Abstract**

A Monte Carlo simulation was performed for estimating and testing hypotheses of three-way interaction effect in latent variable regression models. A considerable amount of research has been done on estimation of simple interaction and quadratic effect in nonlinear structural equation. The present study extended to three-way continuous latent interaction in structural equation model. The latent moderated structural equation (LMS) approach was used to estimate the parameters of the three-way interaction in structural equation model and investigate the properties of the method under different conditions though simulations. The approach showed least bias, standard error,and root mean square error as indicator reliability and sample size increased. The power to detect interaction effect and type I error control were also manipulated showing that power increased as interaction effect size, sample size and latent covariance increased.

**Keywords:** Interaction effect size, latent variable, latent interaction effect, LMS, nonlinear

## 1. Introduction

Structural Equation Modeling (SEM) is a statistical method used for building models, making inference and quantify the relationship among latent variables that are not observable or cannot be measured precisely. But, measurement on the indicator variable related to those unobservable variables are available. This relationship began its bases as a method for modeling linear relationship. However, because of many of the models for observable variables in the social and behavioral sciences involves nonlinearity, its unlikely that linear models are always enough to describe the relationship between latent variables.

Extending SEM to include nonlinear functions allows researchers meaningfully and accurately model the relationship underlying their data. (Kenny & Judd, 1984) introduced the first statistical method aimed at producing estimates of parameters in a nonlinear structural equation model (specifically a quadratic or cross-product structural model with a linear measurement model). Their method attracted methodological discussions and alterations by a number of papers. For instances, (Hayduck, 1987) demonstrated how the Kenny-Judd model could be implemented in LISREL.

Generally, most of the available literature were only for the specific quadratic and simple cross-product structural model. Hence, this study extended the simple cross-product to three-way interaction effect in nonlinear SEM and estimate its effects using latent moderated structural(LMS) equation method. By Montecarlo simulation,the statistical properties of the approach(LMS) were discussed.

In empirical research, models such as (1) can be very useful. It covers the situation in which there are two moderator variables which jointly influence the regression of the dependent variable on an independent variable. In other words, a regression model that has a significant three-way interaction of continuous variables. For instance, to study the moderating effect of social support, hardiness on the relationship between stress and depression,one hypothesizes that the effect of stress on depression was moderated by hardiness and social support. In such cases Model (1) gives a direct test of this hypothesis.

## 2. Model

A model with three latent variables with three observed indicators each for both endogenous and exogenous latent variables was used. For the identification purposes we chose to set a single factor loading to 1 for $\eta, \xi_1, \xi_2$ and $\xi_3$. By three-way interaction, we mean the interaction of three continuous exogenous latent variables $(\xi_1, \xi_2, \xi_3)$. Following the LISREL specification, the structural equation considered was:

$$\eta = \gamma_1\xi_1 + \gamma_2\xi_2 + \gamma_3\xi_3 + \gamma_4\xi_1\xi_2 + \gamma_5\xi_1\xi_3 + \gamma_6\xi_2\xi_3 + \gamma_7\xi_1\xi_2\xi_3 + \zeta \tag{1}$$

The measurement equations for each models were given by

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \lambda_{2y} & 0 & 0 & 0 \\ \lambda_{3y} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \lambda_{x2} & 0 & 0 \\ 0 & \lambda_{x3} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \lambda_{x5} & 0 \\ 0 & 0 & \lambda_{x6} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & \lambda_{x8} \\ 0 & 0 & 0 & \lambda_{x9} \end{bmatrix} \begin{bmatrix} \eta \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \\ \delta_7 \\ \delta_8 \\ \delta_9 \end{bmatrix}$$

Where x is a qx1 vector of independent indicator variables,and y is a px1 vector of dependent indicator variables. $\lambda_x$ is a regression coefficients predicting x by $\xi$ and $\lambda_y$ is a regression coefficients predicting y by $\eta$. $\tau_x$ is vector of x-intercept and $\tau_y$ is a px1 vector of y-intercept. $\delta$ is a q x 1 vector of measurements errors of x and $\varepsilon$ is a p x 1 vector of measurements error of y.

$\eta$ is mx1 vector of latent endogenous variables and $\xi$ is n x 1 vector of exogenous variables. $\Gamma_i$ is regression coefficients predicting $\eta$ by $\xi$ and $\zeta$ is a vector of disturbance.

It was assumed that

- $\xi_1, \xi_2, \xi_3, \delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6, \delta_7, \delta_8, \delta_9, \varepsilon_1, \varepsilon_2$, and $\varepsilon_3$ are multivariate normally distributed.
- $\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6, \delta_7, \delta_8, \delta_9, \varepsilon_1, \varepsilon_2$, and $\varepsilon_3$ have expected values of zero and are uncorrelated with $\xi_1, \xi_2$ and $\xi_3$.
- Finally, $\zeta$ has an expected value of zero and assumed to be uncorrelated with $\xi_1, \xi_2, \xi_3, \delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6, \delta_7, \delta_8, \delta_9, \varepsilon_1$, $\varepsilon_2$, and $\varepsilon_3$.

Based on these assumptions the mean vector and covariance matrix of $(\xi_1, \xi_2, \xi_3, \xi_1\xi_2, \xi_1\xi_3, \xi_2\xi_3, \xi_1\xi_2\xi_3)$ were derived as follows:

$cov(\xi_1, \xi_1\xi_2) = E(\xi_1)cov(\xi_1, \xi_2) + E(\xi_2)cov(\xi_1, \xi_1) + E[(\xi_1 - E(\xi_1))(\xi_1 - E(\xi_1))(\xi_2 - E(\xi_2))]$

By centering $\xi_1$ and $\xi_2$, the expected value of both becomes ,$E(\xi_1) = 0$ and $E(\xi_2) = 0$. Hence $cov(\xi_1, \xi_1\xi_2) = E[(\xi_1 - E(\xi_1))(\xi_1 - E(\xi_1))(\xi_2 - E(\xi_2))]$. Under multivariate normality all third moments vanish (see Bohnstedt and Goldberger 1969). This indicates that $cov(\xi_1, \xi_1\xi_2) = E(\xi_1\xi_1\xi_2) = 0$. Accordingly, all the covariance of the main effects with their two-way interaction is zero under normality condition and the given assumptions.

Following the same procedure, $cov(\xi_1\xi_2, \xi_1\xi_3) = E(\xi_1)E(\xi_1)cov(\xi_2, \xi_3) + E(\xi_1)E(\xi_3)cov(\xi_2, \xi_1) + E(\xi_2)E(\xi_1)cov(\xi_1, \xi_3) + E(\xi_2)E(\xi_3)cov(\xi_1, \xi_1) + cov(\xi_1, \xi_1)cov(\xi_2, \xi_3) + cov(\xi_1, \xi_3)cov(\xi_2, \xi_1)$.

Centering $\xi_1, \xi_2, \xi_3$ the first four terms are zero and we have

$cov(\xi_1\xi_2, \xi_1\xi_3) = cov(\xi_1, \xi_1)cov(\xi_2, \xi_3) + cov(\xi_1, \xi_3)cov(\xi_2, \xi_1) = \phi_{11}\phi_{23} + \phi_{13}\phi_{12}$ and

$cov(\xi_1\xi_2, \xi_2\xi_3) = cov(\xi_1, \xi_2)cov(\xi_2, \xi_3) + cov(\xi_1, \xi_3)cov(\xi_2, \xi_2) = \phi_{12}\phi_{23} + \phi_{13}\phi_{22}$

The covariance of the main effects with the product of the three exogenous variables can be found in the similar manner

$$\begin{aligned} cov(\xi_1, \xi_1\xi_2\xi_3) &= cov(\xi_1, \xi_1)cov(\xi_2, \xi_3) + cov(\xi_1, \xi_2)cov(\xi_1, \xi_3) + cov(\xi_1, \xi_3)cov(\xi_1\xi_2) \\ &= \phi_{11}\phi_{23} + \phi_{12}\phi_{13} + \phi_{13}\phi_{12} \end{aligned}$$

$$\begin{aligned} cov(\xi_2, \xi_1\xi_2\xi_3) &= cov(\xi_2, \xi_1)cov(\xi_2, \xi_3) + cov(\xi_2, \xi_2)cov(\xi_1, \xi_3) + cov(\xi_2, \xi_3)cov(\xi_1, \xi_2) \\ &= \phi_{21}\phi_{23} + \phi_{22}\phi_{13} + \phi_{23}\phi_{12} \end{aligned}$$

$$\begin{aligned} cov(\xi_3, \xi_1\xi_2\xi_3) &= cov(\xi_3, \xi_1)cov(\xi_2, \xi_3) + cov(\xi_3, \xi_2)cov(\xi_1, \xi_3) + cov(\xi_3, \xi_3)cov(\xi_1, \xi_2) \\ &= \phi_{31}\phi_{23} + \phi_{32}\phi_{13} + \phi_{33}\phi_{12} \end{aligned}$$

The covariance between two and three product is zero since the covariances involving five variables, for example $cov(\xi_1\xi_2, \xi_1\xi_2\xi_3) = 0$ under normality.

Following (Bohnstedt & Goldberger, 1969) and under normality, the variance of the latent product is

$$var(\xi_1\xi_2) = E^2(\xi_1)var(\xi_2) + E^2(\xi_2)var(\xi_1) + 2E(\xi_1)E(\xi_2)cov(\xi_1, \xi_2) + var(\xi_1)var(\xi_2) + cov(\xi_1, \xi_2)^2$$

And under the given assumptions it reduces to

$$\begin{aligned} var(\xi_1\xi_2) &= var(\xi_1)var(\xi_2) + cov(\xi_1, \xi_2)^2 \\ &= \phi_{11}\phi_{22} + \phi_{12}^2 \end{aligned}$$

Similarly,

$$\begin{aligned} var(\xi_1\xi_3) &= var(\xi_1)var(\xi_3) + cov(\xi_1, \xi_3)^2 \\ &= \phi_{11}\phi_{33} + \phi_{13}^2 \end{aligned}$$

$$\begin{aligned} var(\xi_2\xi_3) &= var(\xi_2)var(\xi_3) + cov(\xi_2, \xi_3)^2 \\ &= \phi_{22}\phi_{33} + \phi_{23}^2 \end{aligned}$$

For a normally distributed random variables $\xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6$ with mean zero, the fourth and six moment is $E(\xi_1\xi_2\xi_3\xi_4) = cov(\xi_1\xi_2, \xi_3\xi_4) + cov(\xi_1\xi_3, \xi_2\xi_4) + cov(\xi_1\xi_4, \xi_2\xi_3)$ and

$E(\xi_1\xi_2\xi_3\xi_4\xi_5\xi_6) = cov(\xi_1, \xi_2)E(\xi_3\xi_4\xi_5\xi_6) + cov(\xi_1, \xi_3)E(\xi_2\xi_4\xi_5\xi_6) + cov(\xi_1, \xi_4)E(\xi_2\xi_3\xi_4\xi_5) + cov(\xi_1, \xi_5)E(\xi_2\xi_3\xi_4\xi_6) + cov(\xi_1, \xi_6)E(\xi_2\xi_3\xi_4\xi_5)$ (Kendall & Stuart, 1958).

Then we can find $var(\xi_1\xi_2\xi_3)$. That is ;

$$\begin{aligned} var(\xi_1\xi_2\xi_3) &= E(\xi_1^2\xi_2^2\xi_3^2) - E(\xi_1\xi_2\xi_3) \\ &= E(\xi_1^2\xi_2^2\xi_3^2) \\ &= var(\xi_1)E(\xi_2^2\xi_3^2) + cov(\xi_1, \xi_2)E(\xi_1\xi_2\xi_3^2) + cov(\xi_1, \xi_2)E(\xi_1\xi_2\xi_3^2) + cov(\xi_1, \xi_3)E(\xi_1\xi_2^2\xi_3) + cov(\xi_1, \xi_3)E(\xi_1\xi_2^2\xi_3) \\ &= var(\xi_1)E(\xi_2^2\xi_3^2) + 2cov(\xi_1, \xi_2)E(\xi_1\xi_2\xi_3^2) + 2cov(\xi_1, \xi_3)E(\xi_1\xi_2^2\xi_3) \end{aligned}$$

Using the fourth moment, it can be shown that;

$$\begin{aligned} E(\xi_2^2\xi_3^2) &= var(\xi_2)var(\xi_3) + 2cov((\xi_2, \xi_3) \\ &= \phi_{22}\phi_{33} + 2\phi_{23}\phi_{23} \end{aligned}$$

$$\begin{aligned} E(\xi_1\xi_2\xi_3^2) &= cov(\xi_1, \xi_2)var(\xi_3) + 2cov(\xi_1, \xi_3)cov(\xi_2, \xi_3) \\ &= \phi_{12}\phi_{33} + 2\phi_{13}\phi_{23} \end{aligned}$$

and

$$E(\xi_1\xi_2^2\xi_3) = 2\phi_{12}\phi_{23} + \phi_{13}\phi_{22}$$

Hence,

$$var(\xi_1\xi_2\xi_3) = \phi_{11}(\phi_{22}\phi_{33} + 2\phi_{23}\phi_{23}) + 2\phi_{12}(\phi_{12}\phi_{33} + 2\phi_{13}\phi_{23}) + 2\phi_{13}(2\phi_{12}\phi_{23} + \phi_{13}\phi_{22})$$

The mean vectors for $(\xi_1\xi_2, \xi_1\xi_3, \xi_2\xi_3, \xi_1\xi_2\xi_3)$

Centering $\xi_1, \xi_2, \xi_3$, the mean for the two products is

$$\begin{aligned}
E(\xi_1\xi_2) &= cov(\xi_1, \xi_2) = \phi_{12} \\
E(\xi_1\xi_3) &= cov(\xi_1, \xi_3) = \phi_{13} \\
E(\xi_2\xi_3) &= cov(\xi_2, \xi_3) = \phi_{23} \\
E(\xi_1\xi_2\xi_3) &= 0
\end{aligned}$$

Then the mean and variance of the endogenous latent variable in equation (1) is

$$E(\eta) = \gamma_4\phi_{12} + \gamma_5\phi_{13} + \gamma_6\phi_{23}$$

and

$$\begin{aligned}
var(\eta) = {} & \gamma_1^2 var(\xi_1) + \gamma_2^2 var(\xi_2) + \gamma_3^2 var(\xi_3) + \gamma_4^2 var(\xi_1\xi_2) + \gamma_5^2 var(\xi_1\xi_3) + \gamma_6^2 var(\xi_2\xi_3) \\
& + \gamma_7^2 var(\xi_1\xi_2\xi_3) + 2[cov(\gamma_1\xi_1, \gamma_2\xi_2) + cov(\gamma_1\xi_1, \gamma_3\xi_3) \\
& + cov(\gamma_2\xi_2, \gamma_3\xi_3) + cov(\gamma_1\xi_1, \gamma_7\xi_1\xi_2\xi_3) \\
& + cov(\gamma_2\xi_2, \gamma_7\xi_1\xi_2\xi_3) + cov(\gamma_3\xi_3, \gamma_7\xi_1\xi_2\xi_3)] + var(\zeta)
\end{aligned}$$

That is

$$\begin{aligned}
var(\eta) = \sigma_\eta^2 = {} & \gamma_1^2\phi_{11} + \gamma_2^2\phi_{22} + \gamma_3^2\phi_{33} + \gamma_4^2(\phi_{11}\phi_{22} + \phi_{12}^2) + \gamma_5^2(\phi_{11}\phi_{33} + \phi_{13}^2) \\
& + \gamma_6^2(\phi_{22}\phi_{33} + \phi_{23}^2) + \gamma_7^2[\phi_{11}(\phi_{22}\phi_{33} + 2\phi_{23}\phi_{23}) \\
& + 2\phi_{12}(\phi_{12}\phi_{33} + 2\phi_{13}\phi_{23}) + 2\phi_{13}(2\phi_{12}\phi_{23} \\
& + \phi_{13}\phi_{22})] + 2[\gamma_1\gamma_2\phi_{12} + \gamma_1\gamma_3\phi_{13} + \gamma_2\gamma_3\phi_{23} \\
& + \gamma_1\gamma_7(\phi_{11}\phi_{23} + \phi_{12}\phi_{13} + \phi_{13}\phi_{12}) \\
& + \gamma_2\gamma_7(\phi_{21}\phi_{23} + \phi_{22}\phi_{13} + \phi_{23}\phi_{12}) \\
& + \gamma_3\gamma_7(\phi_{31}\phi_{23} + \phi_{23}\phi_{13} + \phi_{33}\phi_{12})] + \psi
\end{aligned}$$

see (Gerry Gray, 1999)

Hence the variance covarince matrix for the latent variables in the model become:

$$\Phi = \begin{pmatrix}
\phi 11 & & & & & & \\
\phi_{21} & \phi_{22} & & & & & \\
\phi_{31} & \phi_{32} & \phi_{33} & & & & \\
0 & 0 & 0 & \phi_{11}\phi_{22} + \phi_{12}^2 & & & \\
0 & 0 & 0 & \phi_{11}\phi_{23} + \phi_{13}\phi_{12} & \phi_{11}\phi_{33} + \phi_{13}^2 & & \\
0 & 0 & 0 & \phi_{12}\phi_{23} + \phi_{13}\phi_{22} & \phi_{12}\phi_{33} + \phi_{13}\phi_{23} & \phi_{22}\phi_{33} + \phi_{23}^2 & \\
\phi_{11}\phi_{23} + 2\phi_{12}\phi_{13} & \phi_{22}\phi_{13} + 2\phi_{23}\phi_{12} & \phi_{33}\phi_{12} + 2\phi_{32}\phi_{13} & 0 & 0 & 0 & \omega
\end{pmatrix}$$

where $\omega$ stands for

$$var(\xi_1\xi_2\xi_3) = \phi_{11}(\phi_{22}\phi_{33} + 2\phi_{23}\phi_{23}) + 2\phi_{12}(\phi_{12}\phi_{33} + 2\phi_{13}\phi_{23}) + 2\phi_{13}(2\phi_{12}\phi_{23} + \phi_{13}\phi_{22})$$

### *2.1 Estimation Method*

Let $f_i = (\eta, \xi_1, \xi_2, \xi_3)', Z_i = (y_1, y_2, y_3, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9)', \epsilon_i = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6, \delta_7, \delta_8, \delta_9)'$. Then the full nonlinear structural equation model can be specified as follows

$$Z_i = \Lambda f_i + \epsilon_i \tag{2}$$

Following the notation in (Wall, 2009), let $\theta_m$ represent the measurement model parameters (i.e., parameters in $\Lambda, \Theta$ and $\theta_s$ denote the nonlinear structural parameters (i.e.,$\gamma_1$ to $\gamma_7, \Psi$). Where $\Theta$ is varance- covariance matrix for $\epsilon_i$ in equation

(2). Note that $\theta = ((\theta m)', (\theta_s)')'$.

For individual i, the joint distribution of the observed data and the latent variables conditional on the parameter vector $\theta$ can be written under the nonlinear structural equation model in equation (1) and (2) as follows.

$$\begin{aligned} P(Z_i, f_i; \theta) &= P(Z_i|f_i, \theta_m)P(f_i, \theta_s) \\ &= P(Z_i|\eta_i, \xi_{1i}, \xi_{2i}, \xi_{3i}; \theta_m)P(\eta_i, \xi_{1i}, \xi_{2i}, \xi_{3i}; \theta_s) \\ &= P(Z_i|\eta_i, \xi_{1i}, \xi_{2i}, \xi_{3i}; \theta_m)P(\eta_i|\xi_{1i}, \xi_{2i}, \xi_{3i}; \theta_s)P(\xi_{1i}, \xi_{2i}, \xi_{3i}; \theta_\xi) \end{aligned} \tag{3}$$

Where $\theta_\xi$ is describing the distribution of $\xi_i$. However, the latent variables are not observable. Therefore, one must integrate the latent variables out of the joint distribution to obtain the marginal density of $Z_i$. That is:

$P(Z_i; \theta_m, \theta_s, \theta_\xi) = \int P(Z_i|\eta_i, \xi_{1i}, \xi_{2i}, \xi_{3i}; \theta_m)P(\eta_i|\xi_{1i}, \xi_{2i}, \xi_{3i}; \theta_s)P(\xi_{1i}, \xi_{2i}, \xi_{3i}; \theta_\xi)d\xi_i$

Hence, the likelihood function is

$$L(\Theta) = \prod \int P(Z_i|\eta_i, \xi_{1i}, \xi_{2i}, \xi_{3i}; \theta_m)P(\eta_i|\xi_{1i}, \xi_{2i}, \xi_{3i}; \theta_s)P(\xi_{1i}, \xi_{2i}, \xi_{3i}; \theta_\xi)d\xi_i \tag{4}$$

Rather than directly approximate the integral in equation (4) (Klein & Moosbrugger, 2000) proposed the latent moderated structural equation method, which does not require the creation of indicators for the interaction of latent variable. LMS uses numerical integration methods for approximating the integrals in Equation (4) and uses a finite mixture of normal distributions to approximate the nonnormal distribution.Then they develop an EM algorithm to find the MLEs of this distribution(see Klein & Moosbrugger, 2000).

### 2.2 Simulation Design

The simulation study was designed to examine the performance of the estimation method in terms of parameter bias, root-mean-square error (RMSE), and standard error. There are twelve observed variables in the model. Nine indicators,$x_1, \cdots, x_6$, for the three latent exogenous variables, $\xi_1, \xi_2$ and $\xi_3$. Three observed indicators , $y_1, \cdots, y_3$, for the latent endogenous variable $\eta$. The observed variable covariance matrix contains ($\frac{12(12+1)}{2} = 78$) unique elements. The model contains 44 parameters to be estimated : eight of the the twelve factor loading , twelve error variances, eight factor variances , three covariance between main effects, three covariance betweeen main effects and the product of the latent variables three way interaction term ,and three covariance between two way interactions term. All variables were simulated to come the following population parameters

$$\eta = 0.3\xi_1 + 0.4\xi_2 + 0.5\xi_3 + 0.1\xi_1\xi_2 + 0.2\xi_1\xi_3 + 0.2\xi_2\xi_3 + \gamma_7\xi_1\xi_2\xi_3 + \zeta \tag{5}$$

Where $\xi_1, \xi_2$ and $\xi_3$ are standard normal variables. The values of $\gamma_1$ to $\gamma_6$ paths were chosen based on values used by Klein and Muthn (2007). The values of $\gamma_7$ varied depending on the magnitude of the interaction effect size.

The errors for the 12 indicators in the measurement model were generated with the variances of the errors chosen so that the reliability of each indicator is 0.64. These population values are chosen so that the variances of the factor indicators are one which makes the parameter values more easily interpretable. Reliability is calculated as the ratio of the variance of the factor indicator explained by the factor to the total variance of the factor indicator using the following formula,

$$\frac{\lambda^2 * \psi}{\lambda^2 * \psi + \theta}$$

where $\lambda$ is the factor loading, $\psi$ is the factor variance, and $\theta$ is the residual variance of the factor indicators. We have used indicator reliability because of it has been shown to affect power to detect interaction effect in a latent variable interaction model (Harring et al., 2012). We chose the indicator reliabilities to be equal across the 12 indicator variables. The latent factor, $\xi_1, \xi_2, \xi_3$, were generated under the distributional study conditions with mean 0 and variance 1.

The error term $\zeta$ was generated from a normal distribution with mean 0 and variance 0.4 which is the same value used by (Klein & Muthn, 2007).

Sample size (n=50 n=100, n=250,n=500) were used in the current study. Past simulation studies investigating interactions between two latent variables have used similar sample sizes (Klein & Muthn, 2007; Marsh et al., 2004). The loading of 0.8 was selected to represent adequate loading size and is comparable to what has been used in previous studies ( Klein & Muthn, 2007; Little et al., 2006; Marsh et al., 2004).

In the first simulation study,the correlation between the two first-order latent variables $\xi_1, \xi_2$ and $\xi_3$ were set equal to the values used by (Klein & Muthn, 2007):$\phi_{11} = \phi_{22} = \phi_{33} = 1, \phi_{12} = 0.3, \phi_{13} = 0.1, \phi_{32} = 0.2$ . When first-order latent variables are strongly related, the standard errors associated with the gamma estimates will become very large (Cohen et al., 2003). Thus, for the current study a larger value for $\phi_{12}, \phi_{13}, \phi_{32}$ were selected to investigate the robustness of the standard errors when the covariance of the latent exogenous factors were high.

The effect size represents the additional variance that the three way interaction effect term explains in $\eta$ above and beyond that which can be explained by the first-order effects and the other three two way interaction term (Marsh et al., 2004) as shown below.

$$R^2_{\gamma 7} = \gamma_7^2[\frac{\phi_{11}(\phi_{22}\phi_{33} + 2\phi_{23}\phi_{23}) + 2\phi_{12}(\phi_{12}\phi_{33} + 2\phi_{13}\phi_{23}) + 2\phi_{13}(2\phi_{12}\phi_{32} + \phi_{13}\phi_{22})}{\sigma^2_\eta}]$$

(Jaccard & Wan, 1995) did a review of the social science literature and found that interaction effect sizes typically accounted for 0.05 and 0.1 of the variance in the dependent variable in the case of two-way latent interaction effects. In the case of three-way interaction effects, the current study chose similar effect sizes for interaction effects in which the proportion of variance in $\eta$ accounted for by the interaction effect was set equal to .0 (to investigate Type I error rates), .05, and .10 (to investigate power)

The squared multiple correlation $R^2$ is

$$\begin{aligned} R^2 = \gamma_1^2\phi_{11} + \gamma_2^2\phi_{22} + \gamma_3^2\phi_{33} + \gamma_4^2(\phi_{11}\phi_{22} + \phi_{12}^2) + \gamma_5^2(\phi_{11}\phi_{33} + \phi_{13}^2) \\ + \gamma_6^2(\phi_{22}\phi_{33} + \phi_{23}^2) + \gamma_7^2[\phi_{11}(\phi_{22}\phi_{33} + 2\phi_{32}) \\ + 2\phi_{12}(\phi_{12}\phi_{33} + 2\phi_{13}\phi_{23}) + 2\phi_{13}(2\phi_{12}\phi_{32} \\ + \phi_{13}\phi_{22})] + 2[\gamma_1\gamma_2\phi_{12} + \gamma_1\gamma_3\phi_{13} + \gamma_2\gamma_3\phi_{23} \\ + \gamma_1\gamma_7(\phi_{11}\phi_{23} + \phi_{12}\phi_{13} + \phi_{13}\phi_{12}) \\ + \gamma_2\gamma_7(\phi_{21}\phi_{23} + \phi_{22}\phi_{13} + \phi_{23}\phi_{12}) \\ + \gamma_3\gamma_7(\phi_{31}\phi_{23} + \phi_{32}\phi_{13} + \phi_{33}\phi_{12})]/\sigma^2_\eta \end{aligned}$$

For the interaction effect size 0,0.05,0.1 and the population variance covariance matrix defined above, squared multiple correlation is ,65.95%, 71.61%, 74.65%,72.86%,79.93%,and 83.01 respectively.

The design of study is 3(effect size)x 4 (sample size)x 2(indicator reliability)x 2(latent covariance) completely crossed factorial design resulting in 48 possible combinations (Table 1). Once the data were generated, they were analyzed with Mplus 7.4.

For each of the 48 possible condition combinations, 500 data sets were generated with Mplus version 7.4. This decision was based on the number of replications used in previous studies for latent interaction, and factors that are known to influence the number of necessary replications for Monte Carlo simulations. For instance, (Powell & Schafer, 2001) conducted a meta analysis of 219 simulation studies in structural equation modeling and reported that the number of replications used in these studies ranged from 20 to 1,000, with the median number of replications being 200. Similarly, (Bandalos, 2006) suggested that 500 replications were large for SEM Monte Carlo simulation studies. She argued that this number of replications would provide stable standard error estimates even when data were generated to come from a non-normal distribution. To check the stability of the model estimation, we have used different seeds to implement the same Monte Carlo simulations, and the model results basically remain unchanged. Thus we conclude that Monte Carlo simulation results are stable.

## 3. Results

### *3.1 Bias, Standard Error and RMSE for Main Effects' Regression Coefficients*

While the bias of the $\gamma_7$ parameter was the primary interest, bias was also examined for the main effects. Bias of the main effects,$\gamma_1, \gamma_2$, and $\gamma_3$, were examined across different conditions (see table 2). With small sample size(i.e, n=50)and moderate reliability (reliability=0.64), this bias was very high.The resulting overestimation decreased as reliability of the indicators and sample size increased, but kept increasing as the interaction effect size for the three-way interaction term $((R_\gamma^2 7))$ and co-variance between latent exogenous variables increased. That is, bias decreased as $\phi_{12}, \phi_{13}, \phi_{23}$ decreased. In reference to the criterion of .05, the estimation method in this study (LMS) produced unbiased estimates for sample size 500 and also for n=250 with high reliability (0.84). Therefore. with moderate reliability and small sample size(i.e, n=50), the bias estimates for $\gamma_1, \gamma_2$, and $\gamma_3$ resulting from LMS approach cannot be trusted.

The column labeled SE-Bias in table 2 stands for standard error bias for the estimates of $\gamma_1, \gamma_2$, and $\gamma_3$. It was found that, this bias is very large (in absolute value) with small sample size(n=50) and moderate reliability indicators, indicating that the LMS approach underestimated standard errors. With the same sample size(n=50), the standard error bias for the estimate of first-order effects decreased as $((R^2_\gamma 7))$ and reliability increased. For all sample size under study,this bias increased as the covariance of latent exogenous increased which is consistent with result of (Cohen et al., 2003). In reference to the criterion of 0.1, the LMS produced unbiased estimates for main effects with sample size 100 and greater but underestimated standard errors because most values were negatives. However, the standard error estimates were fairly accurate when the sample size was 500.

### *3.2 Bias, Standard Error and RMSE for Three-way Interaction Term Regression Coefficients*

Table 3 shows the latent moderated structural equations (LMS) approach parameter estimates of $\gamma_7$ in the all conditions understudy. Perhaps not too surprisingly,bias was greatest for sample size 50 coupled with moderate indicator reliability, but reduced almost by 10% when reliability of the indicators was good (e.g., reliability = 0.84). In the same conditions the standard error and root mean square error reduced by 8% and 10 % respectively for the increment of indicator reliability from 0.64 to 0.84. As anticipated, bias across conditions decreased as sample size increased, but there was a pattern indicative of diminishing returns for sample sizes larger than 500. The increment of interaction effects size resulted increased bias, standard errors and RMSE at small sample size(i.e,n=50), but showed inconstant pattern for the sample size greater than 50. Similarly, for the increase of the covariance between latent factors, the bias, standard error and RMSE reduced for small sample. However, this properties showed inconstant pattern for the others samples size in study.

### *3.3 Type I Error Rates and Empirical Power*

As previously stated, the proportion of variance in $\eta$ accounted for by the three-way interaction effect was set equal to .00 (to investigate Type I error rates), .05, and .10 (to investigate power).The empirical Type I error rates of the nominal size = .05 two-sided tests (under the null hypothesis, H0 : $\gamma_7 = 0$) when using the LMS procedure are given in Table 4. The Type I error rate was computed as the proportion of converged solutions that had a statistically significant three-way interaction effect (at the .05 level) in the simulated data when H0 was true. In addition, empirical power (probability of rejecting a false null hypothesis, H0 : $\gamma_7 = 0$) was represented by the proportion of converged solutions that have a statistically significant interaction effect in the simulated data when H0 was false (Marsh et al., 2004)and tabulated in table 5 under the 5% and 10% effect size conditions.

**Type I error rates**.

When the sample size was 50,100 and indicator reliability 0.64, type I error rates closest to the desired level, but increased as the covariance between latent factors and indicator reliability increased. Moreover, when sample size 100 and reliability was 0.84 the approach in this study (LMS)had very high Type I error rates, rejecting 10% of true models. In this condition( indicator reliability 0.84), the approach under study rejected the null hypothesis (with all the samples) more frequently than the nominal level would predict, except when coupled with moderate reliability. In general, in this study the type I error increased as latent factor covariance and indicators reliability increased(see table 4).

**Empirical power**

Empirical power is represented by the proportion of converged solutions that have a significant interaction effect in the simulated data when the population interaction effect is not equal to zero. Empirical power rates for effect size $R^2_\gamma 7 = 0.05$ and $R^2_\gamma 7 = 0.1$ were computed using an level of .05, and are shown in table 5. As anticipated, empirical power increased as the size of the effect increased from 5% to 10% across methods and conditions. That is, when medium to large three-way interaction effects exist in the population, the methods were able to detect them with a great deal of certainty for moderate sample sizes under high reliability. This was the case even when the sample size was extremely small (n = 50) and the indicators were moderate(reliability = .64). Predictably, power increased as reliability and sample size increased. Power for the LMS approach under study increased as $R^2_\gamma 7$ increased, sample size increased, and $\phi_{12}, \phi_{13}, \phi_{23}$ increased.

## 4. Discussion

Although many simulation studies have been conducted to study latent interaction effects in nonlinear SEM, majority of these studies has focused on two-way latent interactions and quadratic effects. In current study an examination of three-way continuous latent interaction effects was conducted via monte carlo simulation using latent moderated structural method. The simulated data were varied as a function of the size of the three-way interaction term effect, sample size, indicator reliability and the size of the relation between first-order latent variables.

The findings in the Monte Carlo simulation study indicated that, when indicator reliability was moderate and three-way interaction effect present in the generating population-generating model(i.e, $R^2_{\gamma 7} \neq 0.00$), the LMS method led to biased estimate of interaction effect. As with past simulation studies in two-way interaction, indicator variable reliability tended to have the greatest impact on the ability of the LMS to accurately and precisely estimate the three-way interaction effect with size of the relation between the first-order latent variable exerting less influence. Moreover,Parameter estimates for the LMS approach became less biased as the size of the interaction effect and the correlation between the first-order latent variables decreased. We observed that this result for three-way interaction was similar to previous findings of two -way interaction in which the LMS approach was found to result in unbiased estimates of the interaction effect across all sizes of the interaction effect (Klein & Moosbrugger, 2000; Klein & Muthn, 2007).

This finding suggest that the method appeared to control Type I error fairly by reducing the size of the relation between first-order latent variables. Hence moderate indicator reliability, and small sample sizes appear to have the greatest negative impact on the estimation accuracy, precision, and deflation of standard errors of the three-way interaction parameter. Because the method investigated here performed poorly under these circumstances, if data exhibit these characteristics in practice, statistical conclusions should be made cautiously.

## 5. Conclusion and Recommendations

It is the conclusion of the authors that the latent moderated structural approach can be used to study three-way continuous latent interaction in nonlinear structural equation modeling using Mplus software. The approach had no model convergence problems across the conditions in the study and did not produced unrealistic estimates. However, because of the complexity of the model, it took along time to get monte carlo simulation output.

In the conditions considered in the current study, the method led to the least biased estimates of the interaction effect, and accurate standard error estimates, particularly when the sample size was 250 or greater and the indicator reliability was high. Additionally, the latent moderated structural approach accurately estimated first-order effects provided that the sample size was 250 or greater. For the small size(i.e, n=50), the bias for interaction effects and exogenous regression coefficients was high. But for the same sample size, the method had less bias (approximately less than 2%) in estimating the exogenous covariances. This bias increased as the interaction effect size ($R^2_{\gamma 7}$) increased and decrease when sample size increased.

Type I error rates were close to the desired alpha level, particularly when the sample size was 250 or greater. Comparing to other conditions in the study, when indicator reliability were low and the sample size was 50, the method had low power to detect true three-way interaction effects and a sample size of at least 250 was necessary to have acceptable power(greater than 0.8).

Based on these findings,high indicator reliability and a sample size of 250 or more is recommended for use with the latent moderated structural method, although it performs fairly well with sample sizes of 100. It also recommended that,under small sample (i.e,n=50),the method provided sufficient power to detect the three-way interaction effects when high indicator reliability and the covariance of the exogenous latent variable was increased.

## Acknowledgment

I give glory to God for everything. My heartfelt gratitude goes to my respective Supervisor Prof. Kepher H. Makambi for his unlimited constructive advice, suggestions, ideas and comments in all phases of the thesis. I cannot forget to appreciate my second supervisor Dr. Anthony K. Wanjoya for his instruction and guidance. Finally, I would like to extend my appreciation and thanks to Pan African University and Jomo Kenyata University for kind assistance in many ways.

## List of Tables

Table 1. Summary of Manipulated Features

| Factor | 1 | 2 | 3 | 4 | | | |
|---|---|---|---|---|---|---|---|
| Sample size | 50 | 100 | 250 | 500 | | | |
| Indicator reliability | 0.64 | 0.84 | | | | | |
| Effect size(($R^2_\gamma 7$)) | 0.00 | 0.05 | 0.10 | | | | |
| $\phi_{12}, \phi_{13}, \phi_{23}$ | 0.3 | 0.1 | 0.2 | | 0.6 | 0.4 | 0.5 |
| Distribution of $\xi_1, \xi_2, \xi_3$ | Normal | | | | | | |
| Factor loading | 0.8 | | | | | | |
| Estimation method | LMS | | | | | | |

Table 2. Parameter Estimates for the first-order main effects $\gamma_1$ to $\gamma_3$ when $R^2_{\gamma 7} = 0.05, 0.1$ and with different covarince of exogenous latent variables

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\phi_{12} = 0.3$ | | | | | | $\phi_{12} = 0.6$ | | | |
| | $\phi_{13} = 0.1$ | | | | | | $\phi_{13} = 0.4$ | | | |
| | $\phi_{23} = 0.2$ | | | | | | $\phi_{23} = 0.5$ | | | |
| | | | $R^2_{\gamma 7} = 0.05$ | | $R^2_{\gamma 7} = 0.1$ | | $R^2_{\gamma 7} = 0.05$ | | $R^2_{\gamma 7} = 0.1$ | |
| Rel. | N | Par. | Bias | SE-Bias | Bias | SE-Bias | Bias | SE-Bias | Bias | SE-Bias |
| 0.64 | 50 | $\gamma_1$ | 33.32 | -0.929 | 55.146 | -0.90 | 95.33 | -0.968 | 142.297 | -0.969 |
| | | $\gamma_2$ | 44.115 | -0.944 | 48.110 | -0.932 | 3.482 | -0.958 | -11.977 | -0.962 |
| | | $\gamma_3$ | 40.259 | -0.948 | 56.899 | -0.940 | 58.602 | -0.943 | 78.215 | -0.939 |
| | 100 | $\gamma_1$ | 0.182 | -0.049 | 0.183 | -0.047 | 0.207 | -0.051 | 0.212 | -0.045 |
| | | $\gamma_2$ | 0.141 | -0.029 | 0.143 | -0.028 | 0.148 | -0.007 | 0.155 | -0.008 |
| | | $\gamma_3$ | 0.172 | -0.009 | 0.176 | -0.026 | 0.198 | -0.026 | 0.2 | -0.056 |
| | 250 | $\gamma_1$ | 0.069 | -0.035 | 0.069 | -0.027 | 0.081 | -0.029 | 0.083 | -0.026 |
| | | $\gamma_2$ | 0.051 | -0.055 | 0.049 | -0.058 | 0.054 | -0.059 | 0.054 | -0.063 |
| | | $\gamma_3$ | 0.062 | -0.015 | 0.062 | -0.017 | 0.007 | -0.029 | 0.075 | -0.025 |
| | 500 | $\gamma_1$ | 0.039 | -0.024 | 0.039 | -0.025 | 0.045 | -0.033 | 0.046 | -0.037 |
| | | $\gamma_2$ | 0.023 | -0.017 | 0.024 | -0.016 | 0.022 | -0.023 | 0.023 | -0.024 |
| | | $\gamma_3$ | 0.031 | -0.008 | 0.030 | -0.009 | 0.036 | -0.019 | 0.036 | -0.018 |
| 0.84 | 50 | $\gamma_1$ | 0.193 | -0.126 | 0.197 | -0.126 | 0.200 | -0.138 | 0.209 | -0.142 |
| | | $\gamma_2$ | 0.171 | -0.109 | 0.174 | -0.106 | 0.163 | -0.109 | 0.169 | -0.108 |
| | | $\gamma_3$ | 0.204 | -0.074 | 0.209 | -0.015 | 0.209 | -0.089 | 0.217 | -0.097 |
| | 100 | $\gamma_1$ | 0.087 | -0.052 | 0.088 | -0.049 | 0.092 | -0.057 | 0.094 | -0.058 |
| | | $\gamma_2$ | 0.076 | -0.097 | 0.077 | -0.096 | 0.073 | -0.066 | 0.075 | -0.067 |
| | | $\gamma_3$ | 0.084 | -0.006 | 0.085 | -0.005 | 0.089 | -0.024 | 0.092 | -0.026 |
| | 250 | $\gamma_1$ | 0.038 | -0.045 | 0.038 | -0.043 | 0.042 | -0.036 | 0.043 | -0.036 |
| | | $\gamma_2$ | 0.031 | -0.053 | 0.031 | -0.057 | 0.031 | -0.055 | 0.032 | -0.056 |
| | | $\gamma_3$ | 0.034 | -0.008 | 0.034 | -0.010 | 0.038 | -0.015 | 0.039 | -0.015 |
| | 500 | $\gamma_1$ | 0.023 | -0.035 | 0.024 | -0.035 | 0.025 | -0.029 | 0.026 | -0.030 |
| | | $\gamma_2$ | 0.014 | -0.027 | 0.014 | -0.027 | 0.013 | -0.034 | 0.014 | -0.032 |
| | | $\gamma_3$ | 0.018 | 0.011 | 0.018 | 0.011 | 0.021 | 0.000 | 0.021 | 0.000 |

Table 3. Parameter Estimates for nonlinear effects $\gamma_7$ across the study conditions

| | $\phi_{12} = 0.3$ | | | | | | $\phi_{12} = 0.6$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\phi_{13} = 0.1$ | | | | | | $\phi_{13} = 0.4$ | | |
| | $\phi_{23} = 0.2$ | | | | | | $\phi_{23} = 0.5$ | | |
| Rel. | $R^2_{\gamma 7}$ | N | Parameter | Bias | SE-Bias | RMSE | Bias | SE-Bias | RMSE |
| 0.64 | 0.0 | 50 | $\gamma_7$ | 2.0698 | -0.924 | 84.206 | -0.471 | -0.853 | 53.911 |
| | | 100 | $\gamma_7$ | -0.0154 | -0.089 | 0.150 | -0.014 | -0.120 | 0.117 |
| | | 250 | $\gamma_7$ | -0.007 | -0.029 | 0.069 | -0.007 | -0.079 | 0.055 |
| | | 500 | $\gamma_7$ | -0.0003 | -0.026 | 0.047 | -0.003 | -0.092 | 0.037 |
| | 0.05 | 50 | $\gamma_7$ | 58.87 | -0.96 | 138.74 | 35.704 | -0.901 | 62.404 |
| | | 100 | $\gamma_7$ | -0.025 | -0.158 | 0.173 | -0.004 | -0.208 | 0.152 |
| | | 250 | $\gamma_7$ | -0.033 | -0.068 | 0.079 | -0.029 | -0.096 | 0.069 |
| | | 500 | $\gamma_7$ | 0.002 | -0.019 | 0.053 | -0.006 | -0.060 | 0.045 |
| | 0.1 | 50 | $\gamma_7$ | 73.838 | -0.965 | 249.624 | 46.579 | -0.925 | 107.664 |
| | | 100 | $\gamma_7$ | 0.004 | -0.186 | 0.202 | 0.036 | -0.323 | 0.216 |
| | | 250 | $\gamma_7$ | -0.02 | -0.079 | 0.092 | -0.013 | -0.102 | 0.081 |
| | | 500 | $\gamma_7$ | 0.003 | -0.017 | 0.059 | 0.000 | -0.032 | 0.052 |
| 0.84 | 0.0 | 50 | $\gamma_7$ | 0.003 | -0.171 | 0.182 | 0.002 | -0.189 | 0.151 |
| | | 100 | $\gamma_7$ | -0.004 | -0.187 | 0.104 | -0.003 | -0.158 | 0.081 |
| | | 250 | $\gamma_7$ | -0.003 | -0.080 | 0.053 | -0.004 | 0.088 | 0.041 |
| | | 500 | $\gamma_7$ | 0.000 | -0.065 | 0.036 | -0.0016 | -0.077 | 0.028 |
| | 0.05 | 50 | $\gamma_7$ | 0.088 | -0.174 | 0.198 | 0.102 | -0.179 | 0.171 |
| | | 100 | $\gamma_7$ | -0.005 | -0.176 | 0.114 | 0.000 | -0.126 | 0.093 |
| | | 250 | $\gamma_7$ | -0.012 | -0.089 | 0.060 | -0.016 | -0.080 | 0.051 |
| | | 500 | $\gamma_7$ | 0.004 | -0.042 | 0.040 | -0.002 | -0.041 | 0.035 |
| | 0.1 | 50 | $\gamma_7$ | 0.085 | -0.175 | 0.214 | 0.102 | -0.176 | 0.193 |
| | | 100 | $\gamma_7$ | 0.002 | -0.169 | 0.125 | 0.007 | -0.114 | 0.106 |
| | | 250 | $\gamma_7$ | -0.006 | -0.091 | 0.068 | -0.008 | -0.075 | 0.060 |
| | | 500 | $\gamma_7$ | 0.005 | -0.029 | 0.045 | 0.002 | -0.020 | 0.040 |

Table 4. Type I error rates for $R^2_{\gamma 7} = 0$ and with different covarince of exogenous latent variables and reliability of the latent indicators

| | $\phi_{12} = 0.3$ | | | $\phi_{12} = 0.6$ |
|---|---|---|---|---|
| | $\phi_{13} = 0.1$ | | | $\phi_{13} = 0.4$ |
| | $\phi_{23} = 0.2$ | | | $\phi_{23} = 0.5$ |
| Reliability | N | Type I error | type I error | |
| 0.64 | 50 | 0.029 | 0.034 | |
| | 100 | 0.064 | 0.068 | |
| | 250 | 0.070 | 0.070 | |
| | 500 | 0.066 | 0.082 | |
| 0.84 | 50 | 0.080 | 0.092 | |
| | 100 | 0.100 | 0.108 | |
| | 250 | 0.082 | 0.076 | |
| | 500 | 0.080 | 0.080 | |

Table 5. Type I error rates for $R^2_{\gamma 7} = 0$ and with different covarince of exogenous latent variables and reliability of the latent indicators

| | | $\phi_{12} = 0.3$<br>$\phi_{13} = 0.1$<br>$\phi_{23} = 0.2$ | | $\phi_{12} = 0.6$<br>$\phi_{13} = 0.4$<br>$\phi_{23} = 0.5$ | |
|---|---|---|---|---|---|
| | | Power | | Power | |
| Reliability | N | $R^2_{\gamma 7} = 0.05$ | $R^2_{\gamma 7} = 0.1$ | $R^2_{\gamma 7} = 0.05$ | $R^2_{\gamma 7} = 0.1$ |
| 0.64 | 50 | 0.099 | 0.162 | 0.151 | 0.238 |
| | 100 | 0.344 | 0.526 | 0.468 | 0.666 |
| | 250 | 0.746 | 0.942 | 0.884 | 0.982 |
| | 500 | 0.972 | 1.000 | 0.996 | 1.000 |
| 0.84 | 50 | 0.302 | 0.507 | 0.414 | 0.612 |
| | 100 | 0.602 | 0.832 | 0.704 | 0.908 |
| | 250 | 0.940 | 0.994 | 0.982 | 1.000 |
| | 500 | 1.000 | 1.000 | 1.000 | 1.000 |

## References

Bandalos, D. L. (2006). The use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course (pp. 385-426)*. Greenwich, CT: Information Age Publishing

Bohnstedt, G. W., & Goldberger, A. S. (1969). On the exact covariance of products of random variables. *Journal of the American Statistical Association, 64*, 325-328.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gerry Gray (1999). Covariances in Multiplicative Estimates. *Transactions of the American Fisheries Society,128*(3), 475-482. http://dx.doi.org/10.1577/1548-8659(1999)128<0475:CIME>2.0.CO;2

Harring, J. R., Weiss, B. A., Hsu, J. C. (2012). A comparison of methods for estimating quadratic effects in nonlinear structural equation models. *Psychological Methods, 17*(2), 193-214. http://dx.doi.org/10.1037/a0027539

Hayduk, L. A. (1987). Structural equation modeling with LISREL. *Essentials and advances*. Johns Hopkins University; Baltimore, MD.

Jaccard, J., & Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological Bulletin, 117*, 348-357. http://dx.doi.org/10.1037/0033-2909.117.2.348

Kendall, M. G., & Stuart, A. (1958). *The advanced Theory of Statistics(vol.1)*. London:Griffin

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin, 96*, 201-210. http://dx.doi.org/10.1037/0033-2909.96.1.201

Klein, A. G., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika, 65*, 457-474. http://dx.doi.org/10.1007/BF02296338

Klein, A. G., & Muthn, B. (2007). Quasi-maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. *Multivariate Behavioral Research, 42*, 647-673. http://dx.doi.org/10.1080/00273170701710205

Little, T. D., Bovaird, J. A., & Widaman, K. F. (2006). On the merits of orthogonalizing powered and product terms: Implications for modeling interactions among latent variables. *Structural Equation Modeling: A Multidisciplinary Journal, 13*, 497-519. http://dx.doi.org/10.1207/s15328007sem1304_1

Marsh, H. W., Wen, Z., & Hau, K. T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods, 9*, 275-300. http://dx.doi.org/10.1037/1082-989X.9.3.275

Powell, D. A., & Schafer, W. D. (2001). The robustness of the likelihood ratio chi?square test for structural equation models: A meta-analysis. *Journal of Educational and Behavioral Statistics, 26*, 105-132. http://dx.doi.org/10.3102/10769986026001105

Weiss, B. A. (2010). A comparison of methods for testing interaction effects in structural equation modeling. University of Maryland; College Park. (Unpublished doctoral dissertation)

## Appendix

Simulation code using Mplus version 7.4

```
TITLE: Monte Carlo simulation for three-way continuous latent interaction
    MONTECARLO: NAMES=x1-x9 y1-y3;
    NOBSERVATIONS = 250; ! for sample size 250
    NREPS = 500;
    SEED = 12345;
    ANALYSIS: ESTIMATOR = MLR;
    TYPE = RANDOM;
    ALGORITHM = INTEGRATION;
    MODEL POPULATION:
    [x1 − x9@0y1 − y3@0];
    xi1 BY x1-x3@0.8;
    xi2 BY x4-x6@0.8;
    xi3 BY x7-x9@0.8;
    eta BY y1-y3@0.8;
    xi1@1;
    xi2@1;
    xi3@1;
    eta@0.4;! we set the var(zeta)=0.4
    D — xi1 XWITH xi2;
    E — xi1 XWITH xi3;
    F — xi2 XWITH xi3;
    G — xi1 XWITH F;
    eta ON xi1@0.3 xi2@0.4 xi3@0.5 D@0.1 E@0.2 F@0.2 G@0.3;

x1-x9@0.36; y1-y3@0.36;
    xi1 WITH xi2@0.3 xi3@0.1;! for the first variance covariance condition
    xi2 WITH xi3@0.2;
    MODEL:
    [x1 − x9 ∗ 0y1 − y3 ∗ 0];
    xi1 BY x1-x3*0.8;
    xi2 BY x4-x6*0.8;
    xi3 BY x7-x9*0.8;
    eta BY y1-y3*0.8;
    xi1@1;
    xi2@1;
    xi3@1;
    eta@0.4;
    D — xi1 XWITH xi2;
    E — xi1 XWITH xi3;
    F — xi2 XWITH xi3;
    G — xi1 XWITH F;
    eta ON xi1*0.3 xi2*0.4 xi3*0.5 D*0.1 E*0.2 F*0.2 G*0.3;

x1-x9*0.36; y1-y3*0.36;
    xi1 WITH xi2*0.3 xi3*0.1;
    xi2 WITH xi3*0.2;
    OUTPUT: TECH9;
```

## Copyrights

International Journal of Statistics and Probability; Vol. 5, No. 6; November 2016
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

# Estimating Dependence Structure and Risk of Financial Market Crash

Ogunyiola Ayorinde Joshua[1], Peter N. Mwita[2] & Carolyn N. Ngenja[3]

[1] Pan African University Institute for Basic Science, Technology and Innovation, Kenya

[2] Jomo Kenyatta University of Agriculture and Technology, Kenya

[3]Srathmore University, Kenya

Correspondence: Ogunyiola Ayorinde Joshua, Pan African University Institute for Basic Sciences, Technology and Innovation, Kenya. E-mail: ayoogunyiola@yahoo.com

Received: September 14, 2016 Accepted: September 26, 2016 Online Published: October 26, 2016

doi:10.5539/ijsp.v5n6p85 URL: http://dx.doi.org/10.5539/ijsp.v5n6p85

## Abstract

In this paper, we estimate the dependence structure between international stock markets using copulas. Different relationships that exist in normal and extreme periods were estimated using Clayton copula. The Inference Functions for Margins method was used in estimating the clayton copula parameter thereby obtaining dependence estimates used in estimating Value-at-Risk. Extreme events are likely to alter the dependence structure of financial markets. This could have implications for investment decisions and ability to estimate the risk of financial markets crash. Results reveal that during the crisis period (2007-2009), maximum possible loss of market value is 75.9% and 77.6% with a confidence interval of 90% for the Kenya-Nigeria and Kenya-South Africa portfolios respectively. This implies that the Kenya-South Africa portfolio has the highest risk.

**Keywords:** copula, asymmetric dependence, Value-at-Risk

## 1. Introduction

Dependence structure between random variables is crucial in multivariate analysis. In finance, dependence structure between financial markets are critical for investors, policymakers and researchers to make informed decision about investing their resources and making correct investment strategies (Ling, 2006). This is because, interest rates and equity prices move in opposite directions in normal periods but in periods of financial turmoil, they tend to co-move.

The extent of interconnectedness and interdependence of the financial system was highlighted during the 2007-2009 financial crisis (Aloui et al., 2011). The studies of Dennis (2013); Nguyen and Nguyen (2014) suggest and further confirm that financial markets are likely to be more correlated in period of burst than periods of booms. When dependence structure of financial markets (stock, bond, exchange rates and money markets) are closely related, they tend to be faced with a possibility of a market crash.

Until recently, extreme events (financial crisis) were regarded as outliers and often excluded from statistical analysis of financial market (Wu et al., 2012). The financial turmoil has highlighted the importance of analysing extreme events in investment decision, pricing of financial assets as well as risk management (Aloui et.al, 2011). Therefore, extreme events are likely to alter the dependence structure of financial markets. This could have implications for investment decisions and ability to estimate the risk of financial markets crash.

Market crashes are considered catastrophic events when the values of equity market suddenly decline, exchange rates depreciate rapidly and there is a credit default. These rare events could lead to instability of the financial system and exposure to systemic risk. In recent times, extreme events in the financial market are no longer considered as outlier with negligible probability.

Analysing extreme events with a normality assumption might be misleading (Chen et al., 2004). Restriction to elliptical distributions, implies dealing with measures which only captures dependence in the linear sense (Aloui et.al., 2011). Linear correlation models however, from empirical literature have been found not to be appropriate for measuring non normal distributions (Nelsen, 2006). Since financial market data exhibit heavy tails as a result, linear correlation models cannot capture the structure of dependence (Embrechts et al., 2002). Financial decision based on linear correlation models may be misleading as it is not robust in modelling of nonlinear dependence (Boyer et al., 1999).

Linear correlation models can lead to underestimation of the risks that could be associated with a financial market crash; thus it is pertinent to use models which capture financial market non-linearity. The aim of this paper is to estimate dependence structure and estimate the risk of market crash.

The remainder of the paper is organized as follows: section 2 gives the theoretical framework. Empirical results are analysed and discussed in Section 3. Conclusion and policy implications are provided in Section 4.

## 2. Theoretical Framework

Assume a portfolio of two financial market index. The initial value of the portfolio from Carmona (2004) is given as

$$P_0 = n_1 A_1 + n_2 A_2 \tag{1}$$

where $n_1\ and\ n_2$ are the number of units of the two financial markets index, which are valued at $A_1\ and\ A_2$ at the beginning of a period. We denote $A_1'\ and\ A_2'$ their values at the end of the period. The new value of this portfolio at the end of the new period is given by

$$P_1 = n_1 A_1' + n_2 A_2' \tag{2}$$

to get log returns on the individual financial market index, we denote

$$X = log\left(\frac{A_1'}{A_1}\right) \tag{3}$$

and

$$Y = log\left(\frac{A_2'}{A_2}\right) \tag{4}$$

The log return of the portfolio is given from equation 2 is given by

$$= n_1 A_1 e^x \ + n_2 A_2 e^y$$

$$R = log\left(\frac{A}{A_0}\right) = \log\left(\frac{n_1 A_1}{n_1 A_1 + n_2 A_2} e^X + \frac{n_1 A_1}{n_1 A_1 + n_2 A_2} e^Y\right)$$

$$= log(\lambda_1 e^X + \lambda_2 e^Y) \tag{5}$$

where $\lambda_1\ and\ \lambda_2$ are the individual market index.

We apply copula model to estimate the structure of dependence since linear correlation models are not capable to model dependence structure. In order to apply copula model to estimate the risk of financial market crash, we filter our log returns to obtain independent and identically distributed (i.i.d) data using the GARCH (1,1) model. The GARCH filtering provides us with a standardized residual of returns series which is use in estimating the marginal distribution. The standard GARCH (p,q) model by Bollerslev (1986) is given by

$$\sigma_t^2 = \omega + \textstyle\sum_{i=1}^{q} \alpha_i \quad \varepsilon_{t-i}^2 \ \sum_{j=1}^{p} \beta_j \, \sigma_{t-j}^2 \tag{6}$$

${\sigma_t}^2$is conditional variance and $\varepsilon_t$ is the innovation or residual returns defined as $\varepsilon_t$= $\sigma_t e_t,\ e_t \sim N(0,1)$are standardized residual returns.

To measure dependence among the financial market returns, the filtered residuals are joined together applying copula function modelling. The joint distribution function of the random variable $X\ and\ Y$ is given by

$$F_{XY}(x,y) = P_r\ (X \le x, Y \le y) \tag{7}$$

Using the theorem of Sklar (1959), gives us a connection between marginal distribution and copulas to the joint distribution. In this case, let $F_{XY}$ represent a bivariate cumulative distribution function with marginal distribution $F_X$ and $F_Y$, then there exist a two dimensional copula cumulative distribution function $C$ on $[0,1]^2$, such that for all $(x,y)\ \epsilon\ \mathbb{R}^2$ $F_{XY}(x,y) = C\big(F_X(x), F_Y(y)\big)$ holds. For continuous $F_X$ and $F_Y$, $C$ is uniquely determined by

$$C(u,v) = \quad F_{XY}\big(F_X^{-1}(u), F_Y^{-1}(v)\big) \tag{8}$$

the random variables $u = F_X(x)\ and\ v = F_Y(y)$, are obtained by the probability integral transformation uniformly distributed on $[0,1]$, where $F_Y^{-1}(u)\ and\ F_X^{-1}(v)$ are the generalised inverse distribution functions of marginal.

The Joint density function of (X Y) from Sklar (1959) is given by

$$f(x,y) = C\big(F_X(x), F_Y(y)\big)\, f_X(x), f_Y(y) \tag{9}$$

Literature offers several copulas such as Gaussian, Gumbel and Clayton which can be used in modelling dependence structure of a relationship described above (Kjersti, 2004). However, of interest to this study is the Clayton Copula. The Clayton copula is an asymmetric copula, exhibiting greater dependence in the negative tail than in the positive. Mathematically, the bivariate Clayton copula is expressed as

$$C_\theta(u,v) = (u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}} \tag{10}$$

its generator is

$$\varphi_\theta(X) = \frac{1}{\theta}\left(x^{-\theta} - 1\right) \tag{11}$$

Where $u\ and\ v$ are random variables, $0 < \theta < \infty$ is a parameter controlling the dependence. Perfect dependence is obtained if $\theta \to \infty, while\ \theta \to 0$ implies independence. Therefore, markets crashing jointly can be modelled using the Clayton approach which is the main focus area of this research as it can tell about market risk in periods of extreme financial events.

**Estimating Copula Parameters**

We use the inference function for margins method (IFM) to fit copula and estimate the structure of dependence. The IFM is based on the pioneering work of (Joe and Xu, 1996). The estimation method of IFM is presented below:

Assume we observe $n$ independent observations $X_t = \left(x_{t1}, x_{t2}, \dots, x_{tp}\right)$ from a multivariate distribution, which can be constructed with $p$ marginal distributions and a copula function $C(F_1(x), \dots, F_n(x); \theta)$ with parameter $\theta$. The probability distribution function (PDF) of the marginal distributions is defined as $f_i(x;\theta_i)$ with a cumulative density distribution (CDF) as $F_i(x;\theta_i)$, where $\theta_i$ is the parameter of marginal distributions. The IFM method estimates the parameters of the marginal distribution in the first step.

The log-likelihood function of the first step can be written as

$$Logl(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{p} log\ \ f_i\left(x_{ij};\theta_i\right) \tag{12}$$

The estimation of the parameter $\theta = (\theta_1, \dots, \theta_n)$ of marginal distribution can be made through maximizing the log-likelihood function (Joe and Xu, 1996).

$$\hat{\theta}_i = argmax \sum_{i=1}^{n} \sum_{j=1}^{p} log\ \ f_i\left(x_{ij};\theta_i\right) \tag{13}$$

The parameter $\theta$ of the copula function is estimated in the second step of IFM, with the parameter $\hat{\theta}$ of the p marginal distributions.

$$\hat{\theta} = argmax \sum_{t=1}^{n} log\ C\ \left(F_1\left(x_{i1};\hat{\theta}_i\right), \dots, F_p\left(x_{ip};\hat{\theta}_p\right);\theta\right) \tag{14}$$

The IFM is given by a vector $\theta^{IFM} = \left(\hat{\theta}, \hat{\theta}_{IFM}\right)$

where $\hat{\theta}_{IFM} = (\hat{\theta}_i, \hat{\theta}_p, \theta)$

**Asymptotic Properties of Inference Function for Margin Estimator**

We use theorem 1 (Joe, 2005) to show consistency and asymptotic normality of the IFM estimator $\hat{\theta}_{IFM}$.

**Theorem 1.** Let $X_1, \dots, X_n$ be independent and identically distributed random vectors with density $f_\theta$. Let $\theta \epsilon \Theta$ and $x \epsilon S\ :=\ supp(f_0) \subseteq R^2$ where $supp(f_0)$ is the support of $(f_0)$. Assuming the following conditions hold:

(a) The parameter space $\Theta \subseteq$ R is an open interval
(b) The Support $S$ is independent of $\theta$
(c) $f(x;\theta)$ is three times continuously differentiable with respect to $\theta$
(d) $El_\theta\,(X{:}\,\theta)^2 + El_{\theta\theta}\,(X{:}\,\theta) = 0\ and\ \int_s \frac{\partial}{\partial\theta} f(x{:}\,\theta)dx = \frac{\partial}{\partial\theta}\int_s\ f(x{:}\,\theta)\,dx = 0)$
(e) The fisher Information $I(\theta) = El_\theta(X{:}\,\theta)^2 = -El_{\theta\theta}(X{:}\,\theta)$ is positive and finite
(f) For all $\theta_0\,\epsilon\,\Theta$ and $\theta\,\epsilon\,\Theta$ and $\theta\,\epsilon\,U_\delta(\theta_0)$ there exists a measurable function $M_{\theta_0}$ with $E_{\theta_0}\big(M(X{:}\,\theta_0)\big) < \infty$ such that $|l_{\theta\theta\theta}(y{:}\,\theta)| \le M(x{:}\,\theta_0)$ for all $x\ \in$.

Imposing the regularity conditions from (White 1994 and Patton 2006b) to the marginal likelihood in equation (12), and the copula likelihood function, equation (14), a joint normality condition holds such that as $n \to \infty$,

$$\sqrt{n}\left(\left(\ \hat{\theta}_{IFM}\right) - \theta\right) \to N\left(0, \hat{G}\right)$$

where $\hat{G}$ is the estimator of the Godambe Information matrix (Joe, 1997).

G is defined as $G = ({D_g}^{-1}M{D_g}^{-1})^t$, where $D_g = E\left(\frac{\partial g^t(X,\eta)}{\partial\eta}\right)$ and $M_g = E\big(g^t(X,\eta)g\ (X,\eta)\big)$.

**Estimating Market Risk**

In estimating financial market risk, the literature offers several risk measures. However, we reviewed Value at Risk (VaR).

Value-at-Risk measures the minimum loss we would expect over a given time horizon. Carmona (2004) defined $VaR$ as the 100th percentile of the loss distribution given as:

$$q = \mathbb{P}\{-R \geq r\} = \mathbb{P}\{R \leq -r\} = F_R(-r) \tag{15}$$

where $q$ is the percentile, $\mathbb{P}$ is the probability $R$ is the random variable and $r$ is the losses.

From our earlier setting of a portfolio with two financial index and log returns denoted by $X\, and\, Y$, we solve for $r$ in equation (15) to get our VaR with our copula parameter inputted in the VAR (Copula –VAR) by computing the CDF of the log return $R.$ The latter can be expressed analytically as

$$= \iint_{\{(x,y);\lambda_1 e^x + \lambda_2 e^y\} \leq e^{-r}} f(X,Y)^{(x,y)dxdy} \tag{16}$$

where $\lambda_1 e^x + \lambda_2 e^y$ are log returns of the portfolio from equation (5) and $f(X,Y)$ is the CDF of the returns.

$$= \int_{-\infty}^{-r-log\lambda_1} dx \int_{-\infty}^{\log(e^{-r}/\lambda_2 - \lambda_1/\lambda_2 e^x)} c\big(F_X(x), F_Y(y)\big) f_X(x)\, f_Y(y) dy \tag{17}$$

$$= \int_0^{F_X(-r-log\lambda_1)} du \int_0^{F_Y(\log(e^{-r}/\lambda_2 - \lambda_1/\lambda_2 e_X^{F^{-1}(u)}))} dv\; c(u,v) \tag{18}$$

$$= \int_0^{F_X(-r-log\lambda_1)} du\, \frac{\partial}{\partial u} C(u,v)\Big|_{v=F_Y(\log(e^{-r}/\lambda_2 - \lambda_1/\lambda_2 e_X^{F^{-1}(u)}))} \tag{19}$$

The steps for inputting copula parameter into VaR is given in Appendix I.

In order to estimate the VaR in the given framework, we estimate the copula dependence parameter from a sample of pairs of log returns. The estimated parameter is then used to compute the VaR quantiles at different confidence levels which gives estimates for VaR.

## 3. Empirical Results

In the analysis of financial market dependence, monthly data were collected from January, 2000 to March, 2016 from stock exchange websites of countries namely; Nigeria, Kenya and South Africa. Consideration for proxy is All Share Index. Data was divided into three periods; Pre-crisis, crisis and Post crisis periods which captured financial market extreme events (2007-2009). We used negative log returns obtained as:

$$r_t = -In\left(\frac{A_t}{A_{t-1}}\right) \tag{20}$$

where $A_t$ is today's index and $A_{t-1}$is the previous day's index.

The summary statistics for each log return index are reported in Table 1. A glance at the results from the Jarque-Bera test, reveals none of series returns are normally distributed. Figure 1 gives the trend and log returns of each index.

Table 1. Monthly Summary Statistics for Each Index

| Index | Mean | Standard Deviation | Skewness | Kurtosis | Jarque-Bera |
|---|---|---|---|---|---|
| NSE20 | 0.0026 | 0.0404 | -0.474 | 2.502 | *p < 0.00005* |
| NASI | 0.0076 | 0.0719 | -0.509 | 5.360 | *p < 0.00005* |
| JSE | 0.0097 | 0.0490 | -0.3149 | 0.563 | *p < 0.00005* |

Table 1 gives a summary statistic for the negative log returns of Kenya, Nigeria and South Africa stock markets. The p-values of Jarque-Bera normality test are shown in the last column. The sample period on a monthly basis covers from January 2000 to March 2016. Observations are 195, collected from various stock exchange websites. NSE20-Kenya Stock Market Index, NASI; Nigerian Stock Market Index and JSE: South African Stock Index.

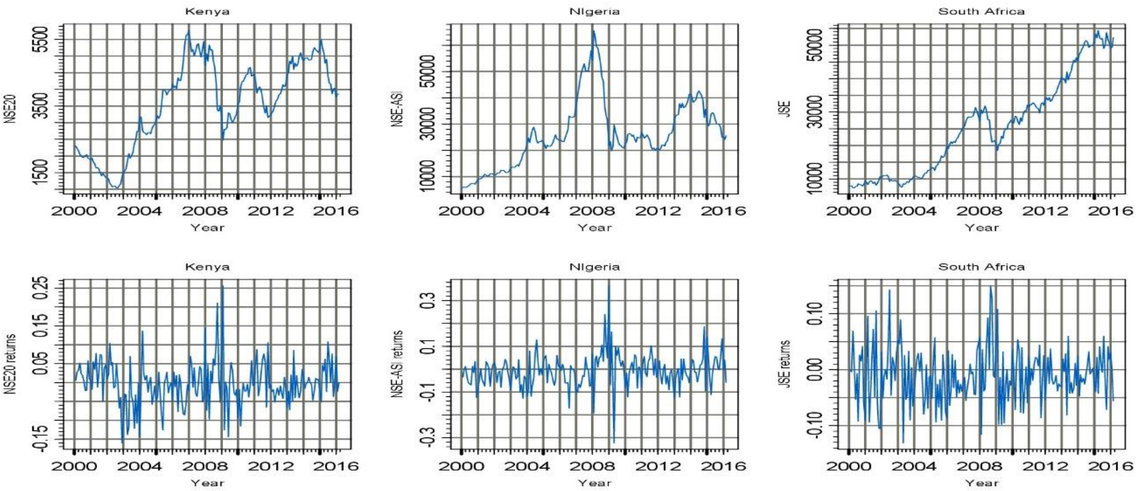


Figure 1. Monthly Stock Market Evolution and Log Returns

Figure 1 reveals that stock market index for Kenya, Nigeria and South Africa was trending upwards which can be described as a pre-crisis period from around 2003 to 2006 for Kenya, 2000 -2006 for Nigeria and 2000-2006 for South Africa. Between 2007-2009 a downward trend is seen for all index (crisis) and from 2010-2016 a post crisis period.

**Copula Fitting**

We report results from fitting Clayton copula. Figure 2 gives a visual display of the dependence between market pairs.

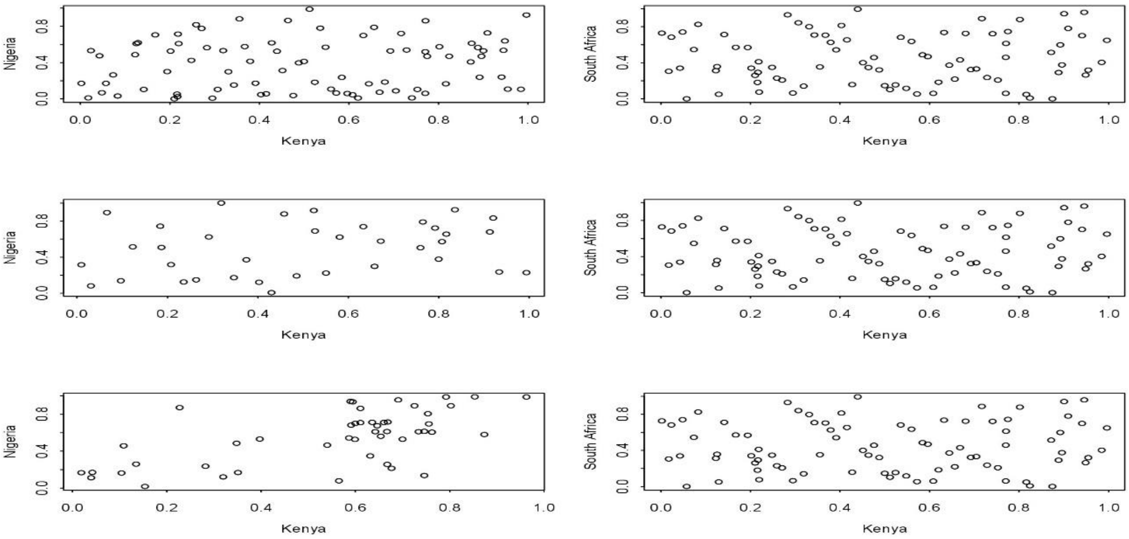


Figure 2. Scatter plots of Stock Market Pairs

The first row gives a scattered plot for the fitted copula, showing patterns of dependence between Kenya and Nigeria stock market between 2000-2006 and Kenya and South Africa in the same period. The second row in a similar manner shows the scattered plot for Kenya and Nigeria between 2007-2009 and Kenya South Africa Between 2007-2009. The observation shows a dependence originating from the low side of the plot. Third row gives the scatter plot for Kenya-Nigeria and Kenya-South Africa for the period 2010-2016.

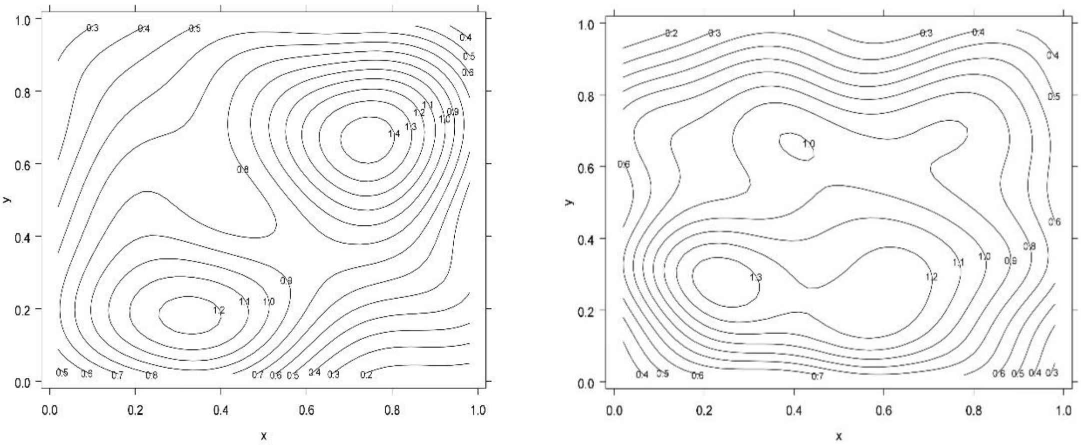


Figure 3. Contour Plot of Clayton Copula (Crisis Period)

Figure 3 gives us a contour plot for Kenya-Nigeria pair on the left and for Kenya-South African pair on the right.

Table 2 gives estimates from fitting clayton copulas to stock market pairs for a period considered as pre crisis between 2000-2006. The table reports the parameter of the copula and also the linear parameters estimates of the relationship between these pairs.

Table 2. Empirical Fitting of Copula 2000-2006 (Pre Crisis Period)

| | **Kenya -Nigeria (Market Pair)** | | | | |
|---|---|---|---|---|---|
| | **Parameter** | | | | |
| Copula | $\theta$ | LogLike | AIC | Kendall’s | Spearman |
| **Clayton** | 0.1574<br>(0.1374) | 0.8471 | 0.3048 | 0.0729 | 0.1670 |
| | **Kenya - South Africa (Market Pair)** | | | | |
| | **Parameter** | | | | |
| Copula | $\theta$ | LogLike | AIC | Kendall’s | Spearman |
| **Clayton** | 0.09943<br>(0.1259) | 0.36285 | 0.1673 | 0.0473 | 0.0709 |

Table 2 gives a summary of copula fit for Kenya-Nigeria stock market index pair and Kenya-South Africa stock market index pair. $\theta$ is the dependence parameter for clayton copula during the pre-crisis period. The Kendall’s tau reports a dependence of 0.072 and Spearman rho of 0.167 for Kenya-Nigeria stock market pair, while a dependence of 0.0473 and 0.070 for Kendall’s tau and Spearman are reported for Kenya-South Africa market pair.

Table 3. Copula Fitting for 2007-2009 (Crisis Period)

| | **Kenya -Nigeria (Market Pair)** | | | | |
|---|---|---|---|---|---|
| Copula | Parameter | | | | |
| | $\theta$ | LogLike | AIC | Kendall’s | Spearman |
| **Clayton** | 0.2863 (0.2486) | 0.9114 | 0.1771 | 0.1252 | 0.1865 |
| | **Kenya - South Africa (Market Pair)** | | | | |
| Copula | Parameter | | | | |
| | $\theta$ | LogLike | AIC | Kendall’s | Spearman |
| **Clayton** | 0.9201 (0.3189) | 6.6607 | -11.3215 | 0.3151 | 0.4542 |

Table 3 gives an overview of fitting clayton copula to stock market pairs in an extreme period considering the 2007-2009 period. In this period, the dependence parameter $\theta$ between Kenya –Nigeria stock market index is 0.286 and for Kenya–South Africa market index is 0.9201. The Kendall's tau reports a dependence of 0.125 and Spearman rho of 0.186 for Kenya-Nigeria stock market pair, while a dependence of 0.315 and 0.452 for Kendall's tau and Spearman are reported for Kenya-South Africa market pair.

Table 4. Copula Fitting for 2010-2016 (Post Crisis Period)

| | **Kenya -Nigeria (Market Index Pair)** | | | | |
|---|---|---|---|---|---|
| Copula | Parameter | | | | |
| | $\theta$ | LogLike | AIC | Kendall's | Spearman |
| **Clayton** | 0.4307(0.1847) | 3.328 | -4.657 | 0.2536 | 0.3727 |
| | **Kenya - South Africa (Market Index Pair)** | | | | |
| Copula | Parameter | | | | |
| | $\theta$ | LogLike | AIC | Kendall's | Spearman |
| **Clayton** | 0.3469 (0.159) | 3.1707 | -4.3414 | 0.1478 | 0.2196 |

In table 4, we report the fitting of clayton copula to stock market pairs in the post crisis period. In this period, the dependence between Kenya–Nigeria stock market pair is 0.430 and for Kenya –South Africa is 0.346. The Kendall's tau reports a dependence of 0.253 and Spearman rho of 0.37 for Kenya-Nigeria stock market pairs, while a dependence of 0.147 and 0.219 for Kendall's tau and Spearman are reported for Kenya-South Africa.

### Estimating Market Risk

After estimation of copula parameter $\theta$, we substituted the dependence parameter into the VaR functions to estimate risk measures at 90%, 95% and 99% using the corresponding quantiles 0.10. 0.05 and 0.01 respectively. The VaR estimates are used in measuring the maximum possible loss of market value over a holding period. The VaR is computed as 0.759 at 90%, 0.784 at 95% and 0.837 at 99% confidence interval which implies the losses from holding the Kenya-Nigeria portfolio is 75.9%, 78.4% and 83.7% at the respective confidence interval as shown in table 5. Simultaneously, VaR estimates computed for Kenya – South Africa portfolio reveal a 0.776 at 90%, 0.795 at 95% and 0.855 at 99% confident interval which implies the losses from holding the Kenya-South Africa portfolio is 77.6%, 79.5% and 85.5% at the respective confidence interval. Since a higher VaR value implies a higher risk, this indicates that holding a Kenya-South Africa portfolio is riskier than holding a Kenya-Nigeria portfolio.

Table 5. Value at Risk Estimates Using Clayton Copula

| | **Crisis Period: Clayton Copula** | | |
|---|---|---|---|
| | 90% VaR | VaR 95% | VaR99% |
| NSE20-NASI | 0.759 | 0.784 | 0.837 |
| | 90% VaR | VaR 95% | VaR 99% |
| NSE20-JSE | 0.776 | 0.795 | 0.855 |

## 4. Discussion and Conclusion

The study focused on stock markets (Kenya, Nigeria and South Africa) using copula technique which estimates dependence structure of stock markets during pre-crisis, crisis and post-crisis periods. Clayton copula dependence parameters have been estimated using the Inference function for margins method into the VaR framework. The results revealed that during the crisis period, the maximum possible loss of market value is 75.9% and 77.6% with a confident interval of 90% for the Kenya-Nigeria and Kenya-South Africa portfolios respectively. This implies that the Kenya-South Africa portfolio has the highest risk. A further implication is that dependence during crisis period imply that opportunities for portfolio diversification are reduced than at periods of booms.

## References

Aloui, R., A ïssa, M. S. B., & Nguyen, D. K. (2011). Global financial crisis, extreme interdependences, and contagion effects: The role of economic structure?. *Journal of Banking & Finance, 35*(1), 130-141. http://dx.doi.org/10.1016/j.jbankfin.2010.07.021

Boyer, B. H., Gibson, M. S., & Loretan, M. (1997). *Pitfalls in tests for changes in correlations* (Vol. 597). Board of

Governors of the Federal Reserve System.

Carmona. (2004). *Statistical Analysis of Financial Data in S-Plus*. Springer Verlag. http://www.princeton.edu/rcaroma/safd

Chen, Fan, Y., & Patton, A. (2004). Simple Test for Models of Dependence between Multiple Financial Time Series, with Application to U.S. Equity Returns and Exchange Rates). *London Economics Financial market group, 43*(483).

Dennis, E. (2013). Developing country vulnerability in light of the global financial crisis: Shock therapy. *Review of Development Finance, 3,* 61-83. http://dx.doi.org/10.1016/j.rdf.2013.02.001

Embrechts, P., McNeil, A., & Straumann, D. (2002). *Correlation and dependence in risk management: Properties and pitfalls*. Cambridge University Press, 176-223. http://dx.doi.org/10.1017/cbo9780511615337.008

Joe, H., & Xu, J. (1996). *The Estimation Method of Inference Functions for Margins for Multivariate Models*. Department of Statistics, University of British Columbia, (166).

Joe, H., (1997). Multivariate Models and Dependence Concepts. Chapman & Hall, London. http://dx.doi.org/10.1201/b13150

Joe, H., (2005). Asympotic Efficiency of the Two Stage estimation method for Copula based models. *Journal of multivariate Analysis 94* (2005), 401-419. http://dx.doi.org/10.1016/j.jmva.2004.06.003

Kjersti, A. (2004). Modelling the dependence structure of financial assets: a survey of four copulas. *Applied research and development.*

Ling, H. (2006). Dependence patterns across financial markets: A mixed copula approach. *Applied Financial Economics*, 717-729.

Nelsen, R. (2006). An introduction to copulas. *Computational Statistics and Data Analysis, 6*, 272.

Nguyen, C., & Nguyen, T. (2014). Analysing Dependence Structure of Equity, Bond and money Market by using time varying Copula. *International Journal of Economics and Finance*, 6(3), 787-815. http://dx.doi.org/10.5539/ijef.v6n3p37

Patton, A. (2006b). Modelling asymmetric exchange rate dependence. International Economic Review, 47(2), 527-556. http://dx.doi.org/10.1111/j.1468-2354.2006.00387.x

Poon, S., M, M. R., & Tawn, J. (2004). Extreme Value Dependence in Financial Markets: Diagnostics, Models, and Financial Implications. *Review of Financial Studies, 17,* 581-610. http://dx.doi.org/10.1093/rfs/hhg058

Sklar, A. (1959). *Fonctions de repartitiona n dimensions et leurs marges*. Publication de l'Institute de Statistique de l'Universite de Paris, 8.

Vaart A. W., (2000). *Asymptotic Statistics*. Cambridge University Press.

White H. (1994). *Estimation, Inference and Specification Analysis*. Econometric Society Monographs No. 22. Cambridge University Press: Cambridge. http://dx.doi.org/10.1017/CCOL0521252806

Wu, Zhang, Z., & Zhao, Y. (2012). Study of the Tail Dependence Structure in Global Financial Markets Using Extreme Value Theory. *Journal of Reviews on Global Economics, 1*, 62-81.

## Appendix A

### Copula VaR

We show how we input $\theta,$ the copula parameter into VaR. From the definition of VaR. Carmona (2004) defined $VaR_q$ as the $100_q$-th percentile of the loss distribution given as:

$$q = \mathbb{P}\{-R \geq r\} = \mathbb{P}\{R \leq -r\} = F_R(-r) \tag{15}$$

We solve for $r$ in equation (15) to get our VaR, we computing the CDF of the log return $R.$ The latter can be expressed analytically as

$$= \iint_{\{(x,y);\lambda_1 e^x+\lambda_2 e^y\}\leq e^{-r}} f(X,Y)^{(x,y)dxdy} \tag{16}$$

Now for a continuous case, the continuous distribution function (CDF) is given by

$$F(X,Y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(a,b)da\, db$$

**Note**: For double integral, we assume that one variable is constant and the other is varying. So here we assume $X$ to be the only variable and $Y$ as a constant. The above generates the upper limit of $X$ as follows.

$$log(\lambda e^{x} + \lambda e^{Y}) \leq -r \Rightarrow \lambda e^{x} + \lambda e^{Y} \leq e^{-r}$$

When $Y$ is considered a constant, we then have

$$\lambda_1 e^{X} \leq\Rightarrow \lambda_1 \leq \frac{e^{-r}}{e^{x}} \Rightarrow \lambda_1 \leq e^{-r-X} \Longrightarrow log(\lambda_1) \leq -r - X \leq -r - log\,\lambda_1$$

Hence, the upper limit for $X$ is $-r - log\lambda_1$

Then for $Y,$ we vary both $x\, and\, Y.$ This is because we cannot assume $X$ to be constant as we have already assumed $Y$ to be one but now we integrate with respect to X.

So we have,

$$\lambda_1 e^{X} + \lambda e^{Y} \leq e^{-r} \Rightarrow \frac{\lambda_1 e^{X}}{\lambda_2} + \frac{\lambda_2 e^{X}}{\lambda_2} \leq \frac{e^{-r}}{\lambda_2} - \frac{\lambda_1 e^{X}}{\lambda_2} \Rightarrow Y \leq \log\left(\frac{e^{-r}}{\lambda_2} - \frac{\lambda_1}{\lambda_2}\right)$$

As the upper limit for $Y.$ Which gives of the equation (17 ) below

$$= \int_{-\infty}^{-r-log\lambda_1} dx \int_{-\infty}^{\log(e^{-r}/\lambda_2 - \lambda_1/\lambda_2 e^{x})} c\big(F_X(x), F_Y(y)\big) f_X(x)\, f_Y(y) dy \qquad (17)$$

### Change of variables and integration

Now, we do a change of variables so as to integrate in a mathematically correct way. So recall that for copulas, $F(x_1, x_2, \ldots, x_n) = C\big(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)\big)$ from Sklar Theorem (1959),

Recall also that

To bring in $u$, we do a transformation,

$$u = f(X)$$

Such that when $X = -\infty,$ $u$=0 and when $X = -r - log\lambda_1$,$u = F_X(-r - log\lambda_1)$

For $v, v = f(Y) = F_Y(y)$ such that when $Y = -\infty, v = 0$ and when

$$Y = \frac{e^{-r}}{\lambda_2} - \frac{\lambda_1 e^{X}}{\lambda_2},$$

$$v = F_Y\left(\frac{e^{-r}}{\lambda_2} - \frac{\lambda_1 e^{F_X^{-1}(u)}}{\lambda_2}\right)$$

$$= \int_0^{F_X(-r-log\lambda_1)} du \int_0^{F_Y(\log(e^{-r}/\lambda_2 - \lambda_1/\lambda_2 e_X^{F^{-1}(u)}))} dv\; c(u,v) \qquad (18)$$

By integrating equation (18) we

$$= \int_0^{F_X(-r-log\lambda_1)} du\; \frac{\partial}{\partial u} C(u,v)\Big|_{v=F_Y(\log(e^{-r}/\lambda_2 - \lambda_1/\lambda_2 e_X^{F^{-1}(u)}))} \qquad (19)$$

### Copyrights

International Journal of Statistics and Probability; Vol. 5, No. 6; November 2016
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

# Reliability Estimates for Three Factor Score Estimators

André Beauducel[1], Christopher Harms[1] & Norbert Hilger[1]

[1] University of Bonn, Institute of Psychology, Germany

Correspondence: André Beauducel, University of Bonn, Institute of Psychology, Kaiser-Karl-Ring 9, 53111 Bonn, Germany. E-mail: beauducel@uni-bonn.de

Received: September 8, 2016 Accepted: October 11, 2016 Online Published: October 26, 2016

doi:10.5539/ijsp.v5n6p94 URL: http://dx.doi.org/10.5539/ijsp.v5n6p94

## Abstract

Estimates for the reliability of Thurstone's regression factor score estimator, Bartlett's factor score estimator, and McDonald's factor score estimator were proposed. Moreover, conditions for equal reliability of the factor score estimators were presented and the reliability estimates were compared by means of simulation studies. Under conditions inducing unequal reliabilities, reliability estimates were largest for the regression score estimator and lowest for McDonald's factor score estimator. We provide an R-script and an SPSS-script for the computation of the respective reliability estimates.

**Keywords:** factor analysis, reliability, factor score estimator

## 1. Introduction

Factor score estimators are computed when individual scores on the factors are of interest. If, for example, decisions are made on the individual level (e.g., in personnel selection) an individual score is needed. It should, however, be noted that the 'estimation' of factor scores does not refer to the estimation of population parameters from a sample. Even in the population, the individual scores on the factors cannot be computed because the number of common and unique factors exceeds the number of observed variables (McDonald & Burr, 1967). In this sense, the factor scores are indeterminate. Therefore, linear composites of the observed variables (e.g., sum scales) are often formed in order to provide factor score estimates. Meanwhile, several factor score predictors with different properties have been proposed (Thurstone, 1935; Bartlett, 1937; McDonald, 1981).

When factor score estimators are computed, their reliability and validity is relevant. The coefficient of determinacy, i.e., the correlation of the factor score estimator with the factor (Grice, 2001) has been related to the validity of factor score estimators (Gorsuch, 1983). However, the reliability of factor score estimators has rarely been investigated. Indexes for the reliability of scores in the context of factor analysis have been proposed, but these coefficients have not been related to the available factor score predictors (McDonald, 1985, 1999; Revelle, 1979; Revelle & Zinbarg, 2009; Zinbarg, Revelle, Yovel, & Li, 2005). A reliability estimate for Harman's ideal variable factor score estimator (Harman, 1976) has already been proposed (Beauducel, 2013). The present paper aims at proposing reliability estimates for Thurstone's regression factor score estimator (Thurstone, 1935), Bartlett's factor score estimator (Bartlett, 1937), and McDonald's correlation-preserving factor score estimator (McDonald, 1981).

Moreover, the effect of the size of loadings, the number of variables, the inter-correlation of the factors, and sampling error on the reliability estimates for the three factor score estimators will be investigated by means of a simulation study. It is, however, possible that –at the population level– the factor model does not perfectly represent the real-world relations between the measured variables, which is typically referred to as model error (MacCallum, 2003; MacCallum & Tucker, 1991). Accordingly, the effect of model error on the reliability estimates of the factor score estimators is also investigated by means of a simulation study. Finally, an R-script and an SPSS-script are presented that allows for the computation of the reliability estimates for the factor score estimators starting from the loading pattern, the factor inter-correlations, and the item covariances.

## 2. Method

In this section, we provide the definition of the factor model and the relevant reliability estimators.

### *2.1 Definitions*

In the population, the common factor model can be defined as

$$\mathbf{x} = \mathbf{\Lambda f} + \mathbf{e}, \tag{1}$$

where **x** is the random vector of observations or items of order *p*. There are *p* observed variables and **f** is the random vector of common factor scores of order *q*, **e** is the random error vector or unique vector of order *p*, and **Λ** is the factor pattern matrix of order *p* by *q*. The factors **f**, and the unique or error vectors **e** are assumed to have an expectation zero (ε[**x**] = 0, ε[**f**] = 0, ε[**e**] = 0). The covariance between the factors and the error scores is assumed to be zero (Cov[**f**, **e**] = ε[**fe** ′] = 0). The covariance matrix of observed variables **Σ** can be decomposed into

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Psi}^2, \tag{2}$$

where **Φ** represents the *q* by *q* factor correlation matrix and $\boldsymbol{\Psi}^2$ is a *p* by *p* diagonal matrix representing the expected covariance of the error scores **e** (Cov[**e**,**e**] = ε[**ee** ′] = $\boldsymbol{\Psi}^2$). It is assumed that $\boldsymbol{\Psi}^2$ is positive definite and that the expectation of the non-diagonal elements is zero.

*2.2 Reliability of Factor Score Estimators*

Cliff (1988, p. 277; Eq. 4) presented a formula for the reliability of weighted composites which is based on two sets of parallel observed variables. For the population of individuals, this formula can be written as

$$\mathbf{R}_{\mathbf{ttc}} = \mathrm{diag}(\mathrm{diag}(\mathbf{B}'\boldsymbol{\Sigma}_{11}\mathbf{B})^{-1/2}\mathbf{B}'\boldsymbol{\Sigma}_{12}\mathbf{B}\,\mathrm{diag}(\mathbf{B}'\boldsymbol{\Sigma}_{22}\mathbf{B})^{-1/2}). \tag{3}$$

where matrix **B** contains the weights, $\boldsymbol{\Sigma}_{11}$ is the covariance matrix of the first set of observed ariables, $\boldsymbol{\Sigma}_{22}$ is the covariance matrix of the second set of observed variables, and $\boldsymbol{\Sigma}_{12}$ is the covariance matrix of the first with the second set of observed variables.

2.2.1 Thurstone’s Regression Factor Score Estimator

For Thurstone’s regression factor score estimator the weights are $\mathbf{B}_{\mathbf{r}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}$. Entering these weights into Equation 4 and adding subscripts indicating the two sets of observed variables yields

$$\begin{aligned}\mathbf{R}_{\mathbf{ttr}} &= \mathrm{Cor}(\hat{\mathbf{f}}_{\mathbf{1r}}, \hat{\mathbf{f}}_{\mathbf{2r}}) \\ &= \mathrm{diag}(\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)^{-1/2}\,\mathrm{diag}(\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2)\,\mathrm{diag}(\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2)^{-1/2}\end{aligned} \tag{4}$$

Inserting $\boldsymbol{\Sigma}_{12} = \mathbf{x}_1\mathbf{x}_2'$, $\mathbf{x}_1 = \boldsymbol{\Lambda}_1\mathbf{f}_1 + \boldsymbol{\Psi}_1\mathbf{e}_1$, and $\mathbf{x}_2 = \boldsymbol{\Lambda}_2\mathbf{f}_2 + \boldsymbol{\Psi}_2\mathbf{e}_2$ into Equation 4 and some transformation yields

$$\begin{aligned}\mathbf{R}_{\mathbf{ttr}} = \mathrm{diag}(\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)^{-1/2}\,&\mathrm{diag}(\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1'\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\Lambda}_1\mathbf{f}_1\mathbf{f}_2'\boldsymbol{\Lambda}_2' + \boldsymbol{\Lambda}_1\mathbf{f}_1\mathbf{e}_2'\boldsymbol{\Psi}_2 \\ &+ \boldsymbol{\Psi}_1\mathbf{e}_1\mathbf{f}_2'\boldsymbol{\Lambda}_2' + \boldsymbol{\Psi}_1\mathbf{e}_1\mathbf{e}_2'\boldsymbol{\Psi}_2)'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2)\,\mathrm{diag}(\boldsymbol{\Phi}_2\boldsymbol{\Lambda}_2'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Lambda}_2\boldsymbol{\Phi}_2)^{-1/2}\end{aligned} \tag{5}$$

It is assumed that the same factors are measured ($\mathbf{f}_1 = \mathbf{f}_2$) and it is assumed that the same factor model holds in the population of the two sets of observed variables $(\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2, \boldsymbol{\Phi}_1 = \boldsymbol{\Phi}_2, \boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_2, \boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22})$. This also implies $\mathbf{f}_1\mathbf{e}_2 = \mathbf{0}$ and $\mathbf{f}_2\mathbf{e}_1 = \mathbf{0}$. When these conditions hold and when there is no systematic unique or error variance, i.e., when there is a zero covariance of the error scores across measurement occasions ( $\varepsilon[\mathbf{e}_1\mathbf{e}_2'] = \mathbf{0}$ ), Equation 5 can be transformed into

$$\mathbf{R}_{\mathbf{ttr}} = \mathrm{diag}(\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)^{-1/2}\,\mathrm{diag}(\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)\,\mathrm{diag}(\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)^{-1/2}. \tag{6}$$

2.2.2 Bartlett’s Factor Score Estimator

Entering $\mathbf{B}_{\mathbf{b}} = \boldsymbol{\Psi}^{-2}\boldsymbol{\Lambda}(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-2}\boldsymbol{\Lambda})^{-1}$ for Bartlett’s factor score estimator into Equation 4 and introducing the subscripts yields

$$\begin{aligned}\mathbf{R}_{\mathbf{ttb}} &= \mathrm{Cor}(\hat{\mathbf{f}}_{\mathbf{1b}}, \hat{\mathbf{f}}_{\mathbf{2b}}) \\ &= \mathrm{diag}((\boldsymbol{\Lambda}_1'\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1)^{-1}\boldsymbol{\Lambda}_1'\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Sigma}_{11}\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1(\boldsymbol{\Lambda}_1'\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1)^{-1})^{-1/2} \\ &\quad\ \mathrm{diag}((\boldsymbol{\Lambda}_1'\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1)^{-1}\boldsymbol{\Lambda}_1^{?}\boldsymbol{\Psi}_1^{-2}\mathbf{x}_1\mathbf{x}_2'\boldsymbol{\Psi}_2^{-2}\boldsymbol{\Lambda}_2(\boldsymbol{\Lambda}_2'\boldsymbol{\Psi}_2^{-2}\boldsymbol{\Lambda}_2)^{-1}) \\ &\quad\ \mathrm{diag}((\boldsymbol{\Lambda}_2'\boldsymbol{\Psi}_2^{-2}\boldsymbol{\Lambda}_2)^{-1}\boldsymbol{\Lambda}_2'\boldsymbol{\Psi}_2^{-2}\boldsymbol{\Sigma}_{22}\boldsymbol{\Psi}_2^{-2}\boldsymbol{\Lambda}_2(\boldsymbol{\Lambda}_2'\boldsymbol{\Psi}_2^{-2}\boldsymbol{\Lambda}_2)^{-1})^{-1/2}\end{aligned} \tag{7}$$

According to $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2, \boldsymbol{\Phi}_1 = \boldsymbol{\Phi}_2, \boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_2, \boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22}, \mathbf{f}_1\mathbf{e}_2 = \mathbf{0}, \mathbf{f}_2\mathbf{e}_1 = \mathbf{0}$, and $\mathbf{e}_1\mathbf{e}_2' = \mathbf{0}$ Equation 7 can be transformed into

$$\begin{aligned}\mathbf{R}_{\mathbf{ttb}} &= \mathrm{diag}((\boldsymbol{\Lambda}_1'\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1)^{-1}\boldsymbol{\Lambda}_1'\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Sigma}_{11}\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1(\boldsymbol{\Lambda}_1'\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1)^{-1})^{-1/2} \\ &\quad\ \mathrm{diag}(\boldsymbol{\Phi}_1)\,\mathrm{diag}((\boldsymbol{\Lambda}_1'\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1)^{-1}\boldsymbol{\Lambda}_1'\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Sigma}_{11}\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1(\boldsymbol{\Lambda}_1'\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1)^{-1})^{-1/2}\end{aligned} \tag{8}$$

It follows from $\mathrm{diag}(\boldsymbol{\Phi}_1) = \mathbf{I}$ that Equation 8 can be transformed into

$$\mathbf{R}_{\mathbf{ttb}} = \mathrm{diag}\mathbf{((\Lambda_1^{'}\Psi_1^{-2}\Lambda_1)^{-1}\Lambda_1^{'}\Psi_1^{-2}\Sigma_{11}\Psi_1^{-2}\Lambda_1(\Lambda_1^{'}\Psi_1^{-2}\Lambda_1)^{-1})}^{-1}. \tag{9}$$

Entering $\mathbf{\Lambda_1\Phi_1\Lambda_1^{'}} + \mathbf{\Psi_1^2}$ for $\mathbf{\Sigma_{11}}$ into Equation 9 yields

$$\mathbf{R}_{\mathbf{ttb}} = \mathrm{diag}\mathbf{((\Lambda_1^{'}\Psi_1^{-2}\Lambda_1)^{-1}\Lambda_1^{'}\Psi_1^{-2}}(\mathbf{\Lambda_1\Phi_1\Lambda_1^{'}} + \mathbf{\Psi_1^2})\mathbf{\Psi_1^{-2}\Lambda_1(\Lambda_1^{'}\Psi_1^{-2}\Lambda_1)^{-1})}^{-1}, \tag{10}$$

and, after some transformation,

$$\mathbf{R}_{\mathbf{ttb}} = \mathrm{diag}\mathbf{((\Lambda_1^{'}\Psi_1^{-2}\Lambda_1)^{-1}} + \mathbf{\Phi_1)}^{-1}. \tag{11}$$

2.2.3 McDonald’s Factor Score Estimator

Entering $\mathbf{B_m} = \mathbf{\Psi^{-2}\Lambda N(N^{'}\Lambda^{'}\Psi^{-2}\Sigma\Psi^{-2}\Lambda N)}^{-1/2}$ where $\mathbf{N}$ is a $q \times q$ matrix with $\mathbf{NN^{'}} = \mathbf{\Phi}$ for McDonald‘s correlation preserving factor score estimator into Equation 3 and introducing subscripts and assuming,

$\mathbf{\Lambda_1 = \Lambda_2}, \mathbf{\Phi_1} = \mathbf{\Phi_2}, \mathbf{\Psi_1 = \Psi_2}, \mathbf{\Sigma_{11} = \Sigma_{22}}, \mathbf{f_1e_2 = 0}, \mathbf{f_2e_1 = 0}$, and $\mathbf{e_1e_2^{'}} = \mathbf{0}$ yields

$$\begin{aligned}\mathbf{R}_{\mathbf{ttm}} &= \mathrm{Cor}(\hat{\mathbf{f}}_{\mathbf{1m}}, \hat{\mathbf{f}}_{\mathbf{2m}}) \\ &= \mathrm{diag}\,((\mathbf{N}_1^{'}\mathbf{\Lambda}_1^{'}\mathbf{\Psi}_1^{-2}\mathbf{\Sigma}_1\mathbf{\Psi}_1^{-2}\mathbf{\Lambda}_1\mathbf{N}_1)^{-1/2}\mathbf{N}_1^{'}\mathbf{\Lambda}_1^{'}\mathbf{\Psi}_1^{-2}\mathbf{\Lambda}_1\mathbf{\Phi}_1\mathbf{\Lambda}_1^{'}\mathbf{\Psi}_1^{-2}\mathbf{\Lambda}_1\mathbf{N}_1\mathbf{(N}_1^{'}\mathbf{\Lambda}_1^{'}\mathbf{\Psi}_1^{-2}\mathbf{\Sigma}_1\mathbf{\Psi}_1^{-2}\mathbf{\Lambda}_1\mathbf{N}_1\mathbf{)}^{-1/2})\end{aligned} \tag{12}$$

Thus, only the parameters of the factor model are necessary in order to calculate the reliabilities, when the hypothetical item set is equivalent.

### *2.3 Comparing Reliability Estimates For Different Factor Score Estimators*

Since reliability estimates are based on $\mathbf{\Lambda_1 = \Lambda_2}, \mathbf{\Phi_1} = \mathbf{\Phi_2}, \mathbf{\Psi_1 = \Psi_2}, \mathbf{\Sigma_{11} = \Sigma_{22}}, \mathbf{f_1e_2 = 0}, \mathbf{f_2e_1 = 0}$, and $\mathbf{e_1e_2^{'}} = \mathbf{0}$, all true variance and all reliability is due to the amount of variance of $\mathbf{f_1}$. Therefore, the factor score estimator with the highest correlation with $\mathbf{f_1}$ has the highest reliability. Thurstone’s regression factor score estimator has the highest correlation with $\mathbf{f_1}$ (Krijnen, et al., 1996), and $\mathrm{Cor}(\hat{\mathbf{f}}_{\mathbf{1r}}, \mathbf{f}_1) \geq \mathrm{Cor}(\hat{\mathbf{f}}_{\mathbf{1b}}, \mathbf{f}_1)$ implies $\mathbf{R}_{\mathbf{ttr}} \geq \mathbf{R}_{\mathbf{ttb}}$ and $\mathrm{Cor}(\hat{\mathbf{f}}_{\mathbf{1r}}, \mathbf{f}_1) \geq \mathrm{Cor}(\hat{\mathbf{f}}_{\mathbf{1m}}, \mathbf{f}_1)$ implies $\mathbf{R}_{\mathbf{ttr}} \geq \mathbf{R}_{\mathbf{ttm}}$ Although the regression factor score estimator has the same or a larger reliability than the other two factor score estimators, the conditions for having an equal reliability are also of interest.

Theorem 1 shows that the reliabilities of the regression factor score estimator and the Bartlett factor score estimator are equal when the condition $\mathbf{\Lambda_1^{'}\Sigma_{11}^{-1}\Lambda_1} = \mathrm{diag}\mathbf{(\Lambda_1^{'}\Sigma_{11}^{-1}\Lambda_1)}$ holds for orthogonal factor models (i.e., $\mathbf{\Phi_1} = \mathbf{I}$ ). The conditions $\mathbf{\Lambda_1^{'}\Sigma_{11}^{-1}\Lambda_1} = \mathrm{diag}\mathbf{(\Lambda_1^{'}\Sigma_{11}^{-1}\Lambda_1)}$ and $\mathbf{\Phi_1} = \mathbf{I}$ hold for one-factor models, since $q = 1$ implies $\mathbf{\Phi_1} = 1$ and that there is only one resulting number for $\mathbf{\Lambda_1^{'}\Sigma_{11}^{-1}\Lambda_1}$ .Moreover, the conditions $\mathbf{\Lambda_1^{'}\Sigma_{11}^{-1}\Lambda_1} = \mathrm{diag}\mathbf{(\Lambda_1^{'}\Sigma_{11}^{-1}\Lambda_1)}$ and $\mathbf{\Phi_1} = 1$ hold for orthogonal factor models with only one non-zero factorloading of each variable.

**Theorem 1.** *If,* $\mathbf{\Lambda_1 = \Lambda_2}, \mathbf{\Phi_1} = \mathbf{\Phi_2} = \mathbf{I}, \mathbf{\Psi_1 = \Psi_2}, \mathbf{\Sigma_{11} = \Sigma_{22}}$ *and* $\mathbf{\Lambda_1^{'}\Sigma_{11}^{-1}\Lambda_1} = \mathrm{diag}\mathbf{(\Lambda_1^{'}\Sigma_{11}^{-1}\Lambda_1)}$ *then* $\mathbf{R_{ttr} = R_{ttb}}$.

*Proof.* From Jöreskog (1969; Equation 10) we get

$$\mathbf{\Sigma_{11}^{-1}\Lambda_1} = \mathbf{\Psi_1^{-2}\Lambda_1}(\mathbf{I} + \mathbf{\Phi_1\Lambda_1^{'}\Psi_1^{-2}\Lambda_1})^{-1}. \tag{13}$$

Premultiplication with $\mathbf{\Lambda_1^{'}}$ and some transformation yields $\mathbf{\Lambda_1^{'}\Sigma_{11}^{-1}\Lambda_1} = ((\mathbf{\Lambda_1^{'}\Psi_1^{-2}\Lambda_1})^{-1} + \mathbf{\Phi_1})^{-1}$ which is entered into Equation 6. This yields

$$\begin{aligned}\mathbf{R}_{\mathbf{ttr}} = \mathrm{diag}\mathbf{(\Phi_1}((\mathbf{\Lambda_1^{'}\Psi_1^{-2}\Lambda_1})^{-1} + \mathbf{\Phi_1})^{-1}\mathbf{\Phi_1)}^{-1/2}\,\mathrm{diag}\mathbf{(\Phi_1}((\mathbf{\Lambda_1^{'}\Psi_1^{-2}\Lambda_1})^{-1} + \mathbf{\Phi_1})^{-1} \\ \mathbf{\Phi_1}((\mathbf{\Lambda_1^{'}\Psi_1^{-2}\Lambda_1})^{-1} + \mathbf{\Phi_1})^{-1}\mathbf{\Phi_1)}\,\mathrm{diag}\mathbf{(\Phi_1}((\mathbf{\Lambda_1^{'}\Psi_1^{-2}\Lambda_1})^{-1} + \mathbf{\Phi_1})^{-1}\mathbf{\Phi_1)}^{-1/2}\end{aligned} \tag{14}$$

According to the conditions of Theorem 1, Equation 14 can be transformed into

$$\mathbf{R}_{\mathbf{ttr}} = \mathrm{diag}\mathbf{(}((\mathbf{\Lambda_1^{'}\Psi_1^{-2}\Lambda_1})^{-1} + \mathbf{I})^{-1}\mathbf{)}^{-1/2}\,\mathrm{diag}\mathbf{(}((\mathbf{\Lambda_1^{'}\Psi_1^{-2}\Lambda_1})^{-1} + \mathbf{I})^{-2}\mathbf{)}\,\mathrm{diag}\mathbf{(}((\mathbf{\Lambda_1^{'}\Psi_1^{-2}\Lambda_1})^{-1} + \mathbf{I})^{-1}\mathbf{)}^{-1/2} \tag{15}$$

Since $\mathbf{\Lambda_1^{'}\Sigma_{11}^{-1}\Lambda_1} = \mathrm{diag}\mathbf{(\Lambda_1^{'}\Sigma_{11}^{-1}\Lambda_1)}$ and $\mathbf{\Phi_1} = \mathbf{I}$ implies $((\mathbf{\Lambda_1^{'}\Psi_1^{-2}\Lambda_1})^{-1} + \mathbf{I})^{-1} = \mathrm{diag}((\mathbf{\Lambda_1^{'}\Psi_1^{-2}\Lambda_1})^{-1} + \mathbf{I})^{-1}$ Equation 15 can be transformed into

$$\mathbf{R}_{\mathbf{ttr}} = \mathrm{diag}\mathbf{(}(\mathbf{\Lambda_1^{'}\Psi_1^{-2}\Lambda_1})^{-1} + \mathbf{I})^{-1} \tag{16}$$

This completes the proof. ■

Theorem 2 shows that the reliabilities of the regression factor score estimator and the McDonald factor score estimator are equal when the condition $\mathbf{\Lambda_1^{'}\Sigma_{11}^{-1}\Lambda_1} = \mathrm{diag}\mathbf{(\Lambda_1^{'}\Sigma_{11}^{-1}\Lambda_1)}$ holds for orthogonal factor models ( $\mathbf{\Phi_1} = \mathbf{I}$ ).

**Theorem 2.** *If* $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2, \boldsymbol{\Phi}_1 = \boldsymbol{\Phi}_2 = \mathbf{I}, \boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_2, \boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22}$, *and* $\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1 = \mathrm{diag}(\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1)$ *then* $\mathbf{R}_{\mathbf{ttr}} = \mathbf{R}_{\mathbf{ttm}}$.

*Proof.* For $\boldsymbol{\Phi}_1 = \boldsymbol{\Phi}_2 = \mathbf{I}$ Equation 14 can be written as

$$\mathbf{R}_{\mathbf{ttm}} = \mathrm{diag}\,((\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Sigma}_{11}\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1)^{-1/2}\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1(\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Sigma}_{11}\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1)^{-1/2}) \tag{17}$$

Entering $\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_1^{'} + \boldsymbol{\Psi}_1^2$ for $\boldsymbol{\Sigma}_{11}$ into Equation 17 and some transformation yields

$$\begin{aligned}\mathbf{R}_{\mathbf{ttm}} &= \mathrm{diag}\,((\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1(\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1)^{-2} + \boldsymbol{\Lambda}_1^{'}\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1(\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1)^{-2})^{-1}) \\ &= \mathrm{diag}\,(((\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1)^{-1} + \mathbf{I})^{-1}) \\ &= \mathrm{diag}\,((\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Psi}_1^{-2}\boldsymbol{\Lambda}_1)^{-1} + \mathbf{I})^{-1}.\end{aligned} \tag{18}$$

This completes the proof. ■

Thus, the three factor score estimators considered here have the same reliability for *q* = 1 and for orthogonal models with *q* > 1 and only one non-zero factor loading of each variable.

*2.4 Reliability of the Regression Score Estimator and the Coefficient of Determinacy*

In the following, the reliability estimate for the regression score estimator is compared with the determinacy coefficient (Grice, 2001), which represents the validity of the factor score predictor. The covariances of the regression factor score estimator with the corresponding common factor are the diagonal elements of

$$\mathrm{diag}(\varepsilon[\mathbf{f}_{\mathbf{r}}\mathbf{f}\,']) \;=\; \mathrm{diag}(\varepsilon[\boldsymbol{\Phi\Lambda}\,\boldsymbol{\Sigma}^{-1}\mathbf{x}\mathbf{f}\,']) \;=\; \mathrm{diag}(\boldsymbol{\Phi\Lambda}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}). \tag{19}$$

The standard deviation of the factor is one and the standard deviation of the regression factor score estimator is diag($\boldsymbol{\Phi\Lambda}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi}$)$^{-1/2}$. Accordingly, the factor score determinacy, i.e., the correlation of the regression score estimator with the corresponding common factors (Grice, 2001) is

$$\mathrm{diag}(\mathrm{cor}[\mathbf{f}_{\mathbf{r},}\ \mathbf{f}]) \;=\; \mathrm{diag}(\boldsymbol{\Phi\Lambda}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi})\ \mathrm{diag}(\boldsymbol{\Phi\Lambda}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi})^{-1/2} \;=\; \mathrm{diag}(\boldsymbol{\Phi\Lambda}\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda\Phi})^{1/2}\,. \tag{20}$$

When the common variance of the factor and the regression factor score estimator is computed for the factor models considered above, this yields

$$\mathrm{diag}\left(\mathrm{cor}\left[\mathbf{f}_{\mathbf{r}},\mathbf{f}\right]\right)^2 = \mathrm{diag}(\boldsymbol{\Phi}_1\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_1)\,. \tag{21}$$

For orthogonal factor models with $\boldsymbol{\Phi}_1 = \mathbf{I}$ and $\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1 = \mathrm{diag}(\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1)$ Equation 6 can be transformed into

$$\begin{aligned}\mathbf{R}_{\mathbf{ttr}} &= \mathrm{diag}(\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1)^{-1/2}\,\mathrm{diag}(\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1)\,\mathrm{diag}(\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1)^{-1/2} \\ &= \mathrm{diag}(\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}) = \mathrm{diag}\left(\mathrm{cor}\left[\mathbf{f}_{\mathbf{r}},\mathbf{f}\right]\right)^2.\end{aligned} \tag{22}$$

Thus, for orthogonal factor models with only one loading of each variable on one factor, the reliability estimate of the regression score estimator corresponds to the coefficient of determinacy. Since it has been shown that the reliability estimates of the regression score estimator, Bartlett’s factor score estimator, and McDonald’s factor score estimator are equal under these conditions, it follows that the abovementioned reliability estimates of the factor score estimators are equal to the (squared) determinacy coefficient for $\boldsymbol{\Phi}_1 = \mathbf{I}$ and $\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1 = \mathrm{diag}(\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1)$.

Theorem 3 describes the relation between the reliability estimate of the regression factor score estimator and factor score determinacy for orthogonal factor models that are identical across measurement occasions when $\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1 \neq \mathrm{diag}(\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1)$.

**Theorem 3.** *If* $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2, \boldsymbol{\Phi}_1 = \boldsymbol{\Phi}_2 = \mathbf{I}, \boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_2, \boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22}$, *and* $\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1 \neq \mathrm{diag}(\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1)$ *then* $\mathbf{R}_{\mathbf{ttr}} \geq \mathrm{diag}\left(\mathrm{cor}\left[\mathbf{f}_{\mathbf{r}},\mathbf{f}\right]\right)^2$.

*Proof.* For simplification we introduce $\mathrm{diag}(\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1) = \mathbf{D}$ and $\boldsymbol{\Lambda}_1^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Lambda}_1 - \mathbf{D} = \mathbf{H}$.

Accordingly, Equation 6 can be written as

$$\mathbf{R}_{\mathbf{ttr}} = \mathbf{D}^{-1/2}\mathrm{diag}(\mathbf{HH} + \mathbf{HD} + \mathbf{DH} + \mathbf{DD})\,\mathbf{D}^{-1/2} \tag{23}$$

Since **H** has a zero-diagonal, pre- and post-multiplication of **H** with the diagonal matrix **D** does not alter the diagonal elements, so that the diagonal elements in **HD** and **DH** are zero. Therefore, Equation 23 can be written as

$$\mathbf{R}_{\mathbf{ttr}} = \mathbf{D}^{-1/2}\mathrm{diag}(\mathbf{HH} + \mathbf{DD})\,\mathbf{D}^{-1/2} \tag{24}$$

Since these diagonal elements are squared elements, it follows that

$$\mathrm{diag}(\mathbf{HH}) \;\geq\; 0 \;\text{ and }\; \mathrm{diag}(\mathbf{DD}) \;\geq\; 0. \tag{25}$$

For orthogonal models Equation 21 can be written as

$$\operatorname{diag}\left(\operatorname{cor}\left[\mathbf{f}_{\mathbf{r}},\mathbf{f}\right]\right)^{2}=\mathbf{D}=\mathbf{D}^{-1/2}\mathbf{D}\mathbf{D}\ \mathbf{D}^{-1/2}. \qquad (26)$$

It follows from diag(**HH**) ≥ 0 that $\mathbf{R}_{\mathbf{ttr}} \geq \operatorname{diag}\left(\operatorname{cor}\left[\mathbf{f}_{\mathbf{r}},\mathbf{f}\right]\right)^{2}$ .

This completes the proof. ■

To summarize, the determinacy coefficient corresponds to the reliability of the three factor score estimators for orthogonal factor models with only one loading of each variable on one factor and the determinacy coefficient is a lower-bound estimate of the reliability of the regression score estimator for orthogonal factor models when $\mathbf{\Lambda}_{1}^{'}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Lambda}_{1} \neq \operatorname{diag}(\mathbf{\Lambda}_{1}^{'}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Lambda}_{1})$ , which can occur when there are non-zero secondary loadings.

## 3. Results

However, the abovementioned considerations do not allow for a quantification of the relative differences of the reliability estimates of the factor score estimators. Therefore, three simulation studies were performed in order to give an account of the reliabilities of the three factor score estimators under different conditions. First, a simulation study was performed at the level of the population for sets of observed variables for which the factor model holds in the population.

### *3.1 Simulation Study 1*

The first short simulation study describes the effects of different population parameters on the reliability estimates. The simulation study was performed with IBM SPSS Version 22 and gives an account of the reliability estimates for the three factor score estimators for *q* = 6, depending on the number of main loadings per factor *p/q* (5, 10), the size of main loadings *l* (.40, .50, .60, .70, .80), the size of secondary loadings *sl* (.00, .10), and the size of the factor inter-correlations *r* (.00, .30). This results in (2 levels of *p/q* × 5 levels of *l* × 2 levels of *sl* × 2 levels of *r*) 40 population models, for which population correlation matrices of observed variables were generated according to Equation 2. The models with *p/q* = 5 were based on 30 observed variables and the models with *p/q* = 10 were based on 60 observed variables.

The reliability estimates for the factor score estimators were computed from the population parameters of the factor model ( $\mathbf{\Lambda},\mathbf{\Phi},\mathbf{\Psi}$ ) and the corresponding item covariances (**Σ**) by means of Equations 6, 11, and 12. The results are summarized in Figure 1. No pronounced reliability differences occurred when the secondary loadings (*sl*) were zero, especially, when only reliabilities greater than .70 are considered. For *sl* = .10 and factor inter-correlations of .30, the regression score estimator had the largest reliability estimates. For factor inter-correlations of .30, McDonald’s factor score estimator had the lowest reliability estimates.

### *3.2 Simulation Study 2*

The next simulation was based on samples that were drawn from populations with the same model parameters as in the previous simulation. The simulation study was again performed with IBM SPSS Version 22. For each of the 40 population models of the previous simulation study 1,000 samples with *n* = 500 cases and 1,000 samples with *n* = 1,000 cases were drawn. Random numbers for the samples of factor scores were generated by means of the SPSS Mersenne Twister pseudo-random number generator. The corresponding samples of observed variables were generated from the common and unique factor scores by means of Equation 2. Maximum-likelihood factor analysis with subsequent Varimax-rotation for orthogonal population factor models and with Promax-rotation (kappa=4) for correlated factor models was performed in each sample of observed variables and the corresponding factor score reliabilities were computed from Equations 8, 13, and 14. The results can be found in Figure 2.

The results of the simulation study for the samples were essentially the same as the results for the population parameters with the highest reliability of the regression factor score estimator. The main difference to the results of the simulation study for the population was that the Bartlett factor score estimator was substantially more reliable than the McDonald factor score estimator when the factor inter-correlations were substantial and when there were non-zero secondary loadings.

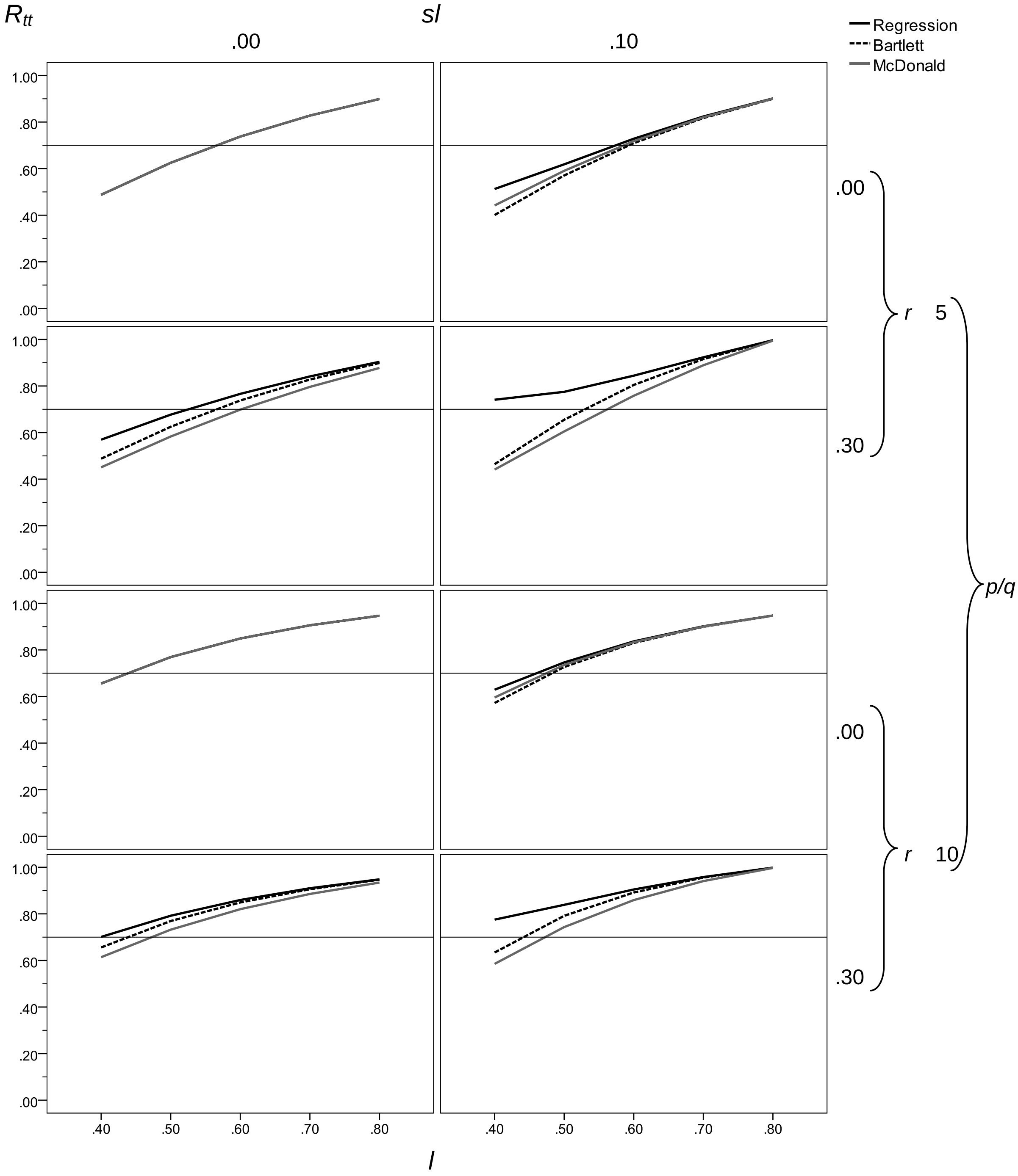


Figure 1. Reliability estimates for the regression factor score estimator, Bartlett's factor score estimator, and McDonalds' factor score estimator for population models with $q$ = 6. The horizontal line marks a reliability of .70 ($R_{tt}$ = Reliability estimate, $l$ = salient loadings, $sl$ = secondary loadings, $r$ = factor inter-correlations).

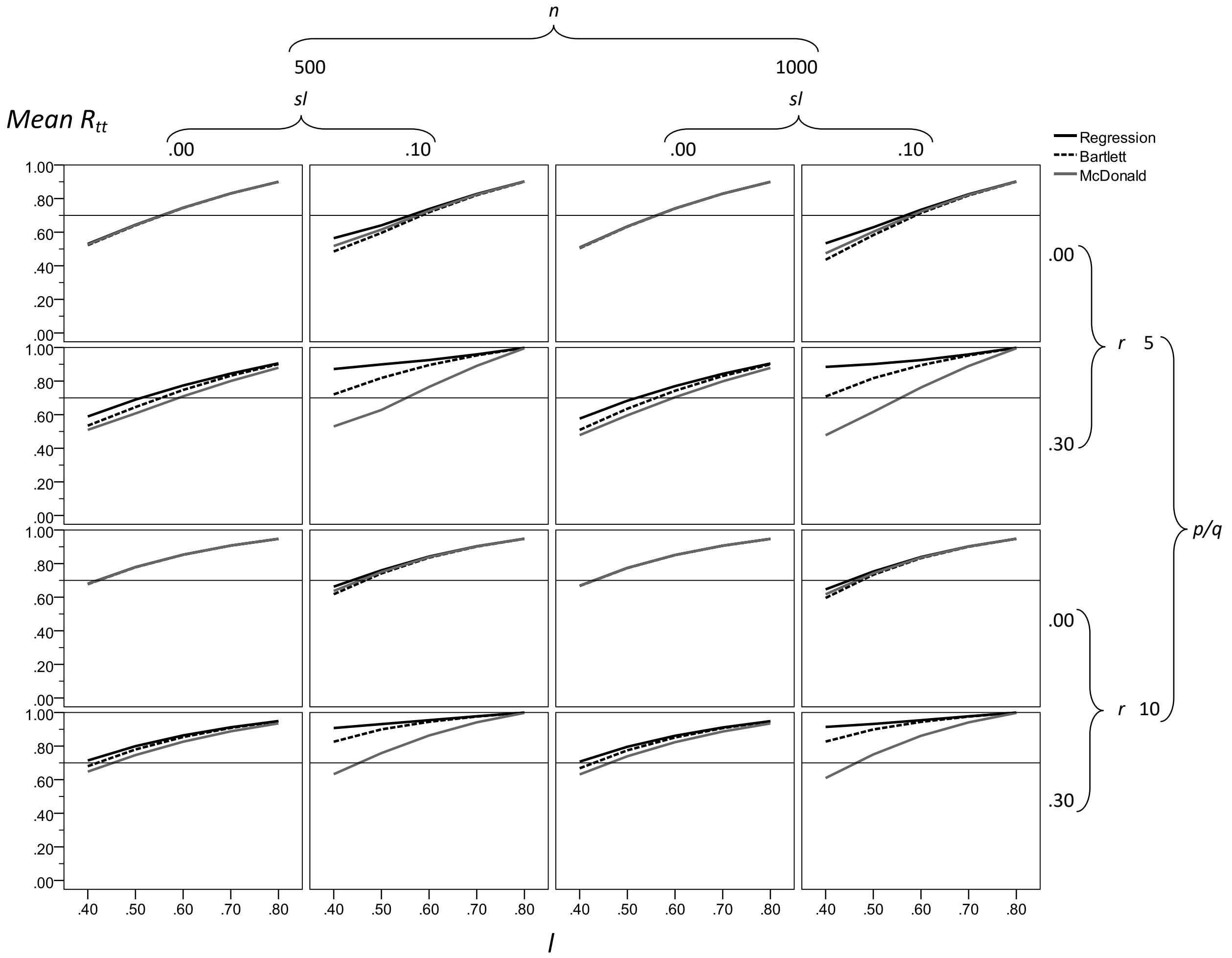


Figure 2. Reliability estimates for the regression factor score estimator, Bartlett's factor score estimator, and McDonalds' factor score estimator for samples based on population models with $q$ = 6. The horizontal line marks a reliability of .70 ($R_{tt}$ = Reliability estimate, $l$ = salient loadings, $sl$ = secondary loadings, $r$ = factor inter-correlations).

### *3.3 Simulation Study 3*

The third simulation study was again based on the population parameters of the first and second simulation study. The only difference was that the simulation study was based on imperfect models, thus, on population models that were hypothesized according to the common factor model, but that do not fit exactly to the population covariance matrix (MacCallum & Tucker, 1991; MacCallum, 2003). Imperfect models were generated as proposed by MacCallum and Tucker (1991). The population correlation matrices were generated from the loadings of the major factors corresponding to the factors in the simulation studies 1 and 2 as well as from the loadings of 100 'minor factors' and from the corresponding uniquenesses. Minor factors have very small nonzero population loadings and represent the 'many minor influences', which are thought to affect the values of the observed scores in the real world. Again, maximum-likelihood factor analysis with subsequent Varimax-rotation for orthogonal population factor models and with Promax-rotation (kappa=4) for correlated factor models was performed in each sample of observed variables and the corresponding factor score reliabilities were computed. The results for the imperfect models were extremely similar to those presented in simulation study 2, so that an additional figure was not necessary. Thus, imperfect models did not affect the reliability estimates substantially.

### *3.4 Standard Deviations and Effects of Factor Rotation*

Overall, the standard deviations of the reliability estimates were extremely small in simulation studies 2 and 3. The mean standard deviations were between .004 and .008 for $n$ = 500 and they were between .002 and .007 for $n$ = 1,000. However, the standard deviations were slightly larger for the correlated factor condition with non-zero secondary loadings. Therefore, the simulation study 2 was repeated for n = 250 for the correlated factor condition with non-zero secondary loadings. The mean standard deviations were greater than .01 for models with small salient loadings, especially for the Bartlett and McDonald factor score predictor (see Table 1). Overall, the mean standard deviations for the regression factor score estimator were smaller than the mean standard deviations of the remaining factor score estimators. The results indicate that sampling error may substantially affect reliability estimates when the sample size and salient loadings are small.

Table 1. Standard deviations of $R_{tt}$ (no model error, *q* = 6, *sl* = .10, *r* = .30, 1,000 samples per condition)

| | | *N* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 250 | | | 500 | | | 1000 | | |
| *p/q* | *l* | Regression | Bartlett | McDonald | Regression | Bartlett | McDonald | Regression | Bartlett | McDonald |
| 5 | .40 | .027 | .043 | .054 | .015 | .027 | .032 | .004 | .010 | .010 |
| | .50 | .008 | .013 | .016 | .005 | .007 | .011 | .003 | .005 | .008 |
| | .60 | .005 | .006 | .010 | .003 | .004 | .007 | .002 | .003 | .006 |
| | .70 | .003 | .003 | .005 | .002 | .002 | .003 | .001 | .001 | .002 |
| | .80 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |
| 10 | .40 | .012 | .014 | .011 | .005 | .008 | .009 | .003 | .005 | .008 |
| | .50 | .004 | .005 | .008 | .003 | .004 | .007 | .002 | .003 | .005 |
| | .60 | .003 | .003 | .005 | .002 | .002 | .004 | .001 | .001 | .003 |
| | .70 | .001 | .001 | .002 | .001 | .001 | .002 | .001 | .001 | .001 |
| | .80 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |

*Note. p* = number of variables, *q* = number of factors, *l* = salient loadings, *sl* = secondary loadings, *r* = factor inter-correlations.

In order to investigate the effects of factor rotation on the reliabilities of the factor score estimates, the simulation studies 2 and 3 were repeated with Equamax-rotation for the orthogonal factor models and with Oblimin-rotation (delta=0) for correlated factor models. Again, the reliability estimates were similar for the orthogonal models whereas the reliability estimates were highest for the regression factor score estimator when the analyses were performed for correlated factor models with non-zero secondary loadings. The results were very similar to those presented for the simulation studies 2 and 3 so that an additional figure was not necessary.

## 4. Discussion

Based on the reliability definition for weighted composites (Cliff, 1988), reliability estimates for Thurstone’s regression factor score estimator, Bartlett’s factor score estimator, and McDonald’s factor score estimator were proposed. It was shown that the reliability estimates are equal for the three factor score estimators when they are based on a one-factor model or when there are orthogonal factors with only one non-zero factor loading of each observed variable (Theorem 1 and 2). Moreover, it was shown that the reliability estimates for the regression factor score estimator are equal to the (squared) determinacy coefficient for the one-factor model or when there are orthogonal factors with only one non-zero loading of the items on a factor (Theorem 3). Thus, for one-factor models as well as for orthogonal models with only one non-zero loading of each variable, the determinacy coefficient represents a reliability estimate for the regression score estimator, for Bartlett’s factor score estimator, as well as for McDonald’s factor score estimator. Since some software (e.g. Mplus 7; Muthén & Muthén, 2012) calculates the determinacy coefficient, it might be interesting to use these coefficients as reliability estimates for these models. Mplus users should be aware that they should square the determinacy coefficient computed by Mplus in order to get the reliability coefficient. For orthogonal factor models with more than one non-zero loading of the items on a factor, the determinacy coefficient is a lower-bound estimate of the reliability of the regression factor score estimator.

The reliability estimates of the three factor score estimators were compared by means of a simulation study for the population and by means of a simulation study for samples drawn from a population in which the factor model holds as well as for samples drawn from a population in which the factor model does not hold. The population based simulation study revealed that –under conditions where different reliability estimates can occur– the reliability estimates were largest for the regression factor score estimator and that they were typically lowest for McDonald’s factor score estimator, especially when the factor inter-correlations were substantial. In contrast, for orthogonal factors and when only substantial reliabilities (>.70) were considered, the differences between the reliability estimates for all three factor score estimators were small. The results of the simulation studies for the samples were very similar to the results for the population based simulation study. Thus, computing the regression score predictor results in the highest reliability estimates and computing McDonald's factor score estimator typically results in considerably larger losses of reliability than computing Bartlett's factor score estimator, especially when the factor inter-correlations are substantial. It follows that McDonald's factor score estimator should only be computed when the salient factor loadings are very large. From a practical point of view it might be reasonable to compute the reliability estimates of the three factor score predictors and to only use Bartlett's or McDonald's factor score predictor when the corresponding losses of reliability can be neglected. This might be of interest because Bartlett's and McDonald's factor score predictor have properties that might convince applied researchers. For example, McDonald's factor score predictor is correlation-preserving. This means that the inter-correlations between McDonald's factor score predictors are the same as the inter- correlations between the factors. This property may facilitate the interpretation of McDonald's factor score predictor, but this property is lost with Bartlett's score predictor and with the regression score predictor.

Moreover, we found that using imperfect factor models for the simulation study did not affect the results. This implies that the reliability estimates can also be used when the corresponding factor model does not fit perfectly to the data. Finally, we found that the standard deviations of the reliability estimates can be substantial in correlated factor models when the sample size is about n = 250 and when there are small salient loadings. Thus, caution with reliability estimates is warranted when sample sizes are about 250 cases or below and when salient loadings are small. Factor rotations (Equamax versus Varimax for orthogonal models and Promax versus Oblimin for correlated factor models) did not alter the results of the simulation study.

From a practical point of view, it should be noted that whenever a specific score is computed and interpreted, the reliability estimate of the respective score should be known and substantial. Thus, when unit-weighted scales are computed, it might be reasonable to compute Cronbach's alpha as a reliability estimate. When a factor score estimate is computed and used in the context of assessment, the reliability of the factor score estimate should be evaluated. Accordingly, an R-script (Appendix A) as well as an SPSS-script (Appendix B) is presented that allows for the respective calculations of the reliability estimates from the loading pattern and factor inter-correlations. The R and SPSS scripts are also available in an online repository at https://github.com/neurotroph/reliability-factor-score-estimators.

## References

Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology, 28,* 97-104. http://dx.doi.org/j.2044-8295.1937.tb00863.x

Beauducel, A. (2013). Taking the error-term of the factor model into account: The factor score estimator interval. *Applied Psychological Measurement, 37,* 289-303. http://dx.doi.org/10.1177/0146621613475358

Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin, 103,* 276-279. http://dx.doi.org/10.1037/0033-2909.103.2.276

Gorsuch, R. L. (1983). *Factor analysis (2$^{nd}$ ed.).* Hillsdale, NJ: Lawrence Erlbaum.

Grice, J. W. (2001). Computing and evaluation factor scores. *Psychological Methods, 6,* 430-450. http://dx.doi.org/10.1037/1082-989X.6.4.430

Harman, H. H. (1976). *Modern factor analysis (3rd ed.).* Chicago, IL: University of Chicago Press.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183-202. http://dx.doi.org/10.1002/j.2333-8504.1967.tb00991.x

Krijnen, W. P., Wansbeek, T. J., & Ten Berge, J. M. F. (1996). Best linear predictors for factor scores. *Communications in Statistics: Theory and Methods, 25,* 3013-3025. http://dx.doi.org/10.1080/03610929608831883

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research,* 38, 113-139. http://dx.doi.org/10.1207/S15327906MBR3801_5

MacCallum, R. C. & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin,* 109, 502-511. http://dx.doi.org/10.1037/0033-2909.109.3.502

McDonald, R. P. (1981). Constrained least squares estimators of oblique common factors. *Psychometrika, 46,* 337-341. http://dx.doi.org/10.1007/BF02293740

McDonald, R. P. (1985). *Factor Analysis and Related Methods.* Hillsdale, NJ: Erlbaum.

McDonald, R. P. (1999). *Test Theory: A Unified Treatment.* Mahwah, NJ: Erlbaum.

McDonald, R. P., Burr, E. J. (1967). A comparison of four methods of constructing factor scores. *Psychometrika, 32,* 381-401. http://dx.doi.org/10.1007/BF02289653

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus User´s Guide* (Seventh Edition). Los Angeles, CA: Muthén & Muthén.

Revelle, W. (1979). Hierarchical cluster-analysis and the internal structure of tests. *Multivariate Behavioral Research, 14,* 57-74. http://dx.doi.org/10.1207/s15327906mbr1401_4

Revelle, W. & Zinbarg, R. E. (2009). Coefficients Alpha, Beta, Omega, and the GLB: Comments on Sijtsma. Psychometrika, 74, 145-154. http://dx.doi.org/10.1007/S11336-008-9102-Z

Thurstone, L. L. (1935). *The vectors of mind.* Chicago, IL: University of Chicago Press.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's $\alpha$, Revelle's $\beta$, McDonald's $\omega_H$: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70,* 123-133. http://dx.doi.org/10.1007/s11336-003-0974-7

## Appendix A

### R-script for reliabilities of factor score estimators

```
##' This function computes and returns reliability estimates for three commonly used
##' Factor Score Estimators in Factor Analyses.
##'
##' Explanations of the algebraic formulas are presented in the manuscript.
##'
##' @title Function for calculating reliability estimates for factor score estimators
##' @param Lambda a \code{matrix} containing the loadings of items on the factors
##' @param Phi a \code{matrix} containing the factor intercorrelations
##' @param Estimators a \code{vector} to select the estimators for which the reliability
##'    estimates should be calculated. Available values: \code{Regression}, \code{Bartlett},
##'    \code{McDonald}
##' @return Returns a two-dimensional list containing the reliability estimates for each
##'    factor. Depending on the \code{Estimators} parameter, the list contains the values
##'           only for the selected Estimators.
##' @export
##' @author André Beauducel (\email{beauducel@uni-bonn.de})
##' @author Christopher Harms (\email{christopher.harms@uni-bonn.de})
##' @author Norbert Hilger (\email{nhilger@uni-bonn.de})
##'
factor.score.reliability <- function(Lambda, Phi, Estimators=c("Regression", "Bartlett", "McDonald")) {
    # Helper functions for frequently used matrix operations
    Mdiag <- function(x) return(diag(diag(x)))
    inv <- function(x) return(solve(x))

    # If a 'loadings' class is provided for lambda, we can easily convert it
    if (is(Lambda, "loadings"))
        Lambda <- Lambda[,]

    # Perform several validity checks of the provided arguments
    if (any(missing(Lambda), missing(Phi), is.null(Lambda), is.null(Phi)))
        stop("Missing argument(s).")
    if (any(nrow(Phi) == 0, nrow(Lambda) == 0, ncol(Phi) == 0, ncol(Lambda) == 0))
```

```
        stop("Some diemension(s) of Phi or Lambda seem to be empty.")
    if (nrow(Phi) != ncol(Phi))
        stop("Phi has to be a q x q matrix.")
    if (ncol(Lambda) != nrow(Phi))
        stop("Phi and Lambda have a different count of factors.")
    if (any(round(min(Phi)) < 0, round(max(Phi)) > 1))
        stop("Phi contains invalid values (outside [0; 1]).")
    Estimators.Allowed <- c("Regression", "Bartlett", "McDonald")
    if (is.null(Estimators)) {
        message("No 'Estimators' defined, use 'Regression' as default.")
        Estimators <- c("Regression")
    }
    Estimators <- match.arg(Estimators, Estimators.Allowed, several.ok = TRUE)

    # Regenerate covariance matrix from factor loadings matrix
    Sigma <- (Lambda %*% Phi %*% t(Lambda))
    Sigma <- Sigma - Mdiag(Sigma) + diag(nrow(Lambda))

    # Calculate uniqueness/error of items
    Psi <- Mdiag(Sigma - Lambda %*% Phi %*% t(Lambda))^0.5
    if (round(min(diag(Psi))) < 0)
        stop("The diagonal of Psi contains negative values.")

    ret <- list()
if ("Regression" %in% Estimators) {
        # Reliability of Thurstone's Regression Factor Score Estimators
        # cf. Equation 6 in manuscript
Rtt.Regression <-
inv( Mdiag( Phi %*% t(Lambda) %*% inv(Sigma) %*% Lambda %*% Phi ) )^0.5 %*%
Mdiag( Phi %*% t(Lambda) %*% inv(Sigma) %*% Lambda %*% Phi %*% t(Lambda) %*% inv(Sigma) %*%
Lambda %*% Phi) %*%
inv( Mdiag( Phi %*% t(Lambda) %*% inv(Sigma) %*% Lambda %*% Phi ) )^0.5
        ret$Regression <- diag(Rtt.Regression)
    }
if ("Bartlett" %in% Estimators) {
        # Reliability of Bartlett's Factor Score Estimators
        # cf. Equation 11 in manuscript
Rtt.Bartlett <- inv( Mdiag( inv(t(Lambda) %*% inv(Psi)^2 %*% Lambda) + Phi ) )
        ret$Bartlett <- diag(Rtt.Bartlett)
    }
if ("McDonald" %in% Estimators) {
        # Reliability of McDonald's correlation preserving factor score estimators
        # cf. Equation 12 in manuscript
```

```
                Decomp <- svd(Phi)
                N <- Decomp$u %*% abs(diag(Decomp$d))^0.5
                sub.term <-
t(N) %*% t(Lambda) %*% inv(Psi)^2 %*% Sigma %*% inv(Psi)^2 %*% Lambda %*% N
                Decomp <- svd(sub.term)
                sub.term <- Decomp$u %*% (diag(Decomp$d)^0.5) %*% t(Decomp$u)
Rtt.McDonald <-
Mdiag( inv(sub.term) %*% t(N) %*% t(Lambda) %*% inv(Psi)^2 %*% Lambda %*% Phi %*% t(Lambda) %*%
inv(Psi)^2 %*% Lambda %*% N %*% inv(sub.term))
                ret$McDonald <- diag(Rtt.McDonald)
        }

# Return reliabilities as list, so it can be accessed via e.g. factor.score.reliability(L, P)$Regression
        return(ret)
}

## Example 1:
## Users may just enter their respective values for Loadings and InterCorr.
Loadings <- matrix(c(
         0.50,-0.10, 0.10,
         0.50, 0.10, 0.10,
         0.50, 0.10,-0.10,
        -0.10, 0.50, 0.15,
         0.15, 0.50, 0.10,
        -0.15, 0.50, 0.10,
         0.10, 0.10, 0.60,
         0.10,-0.10, 0.60,
         0.10, 0.10, 0.60
        ),
        nrow=9, ncol=3,
        byrow=TRUE)
InterCorr <- matrix(c(
        1.00, 0.30, 0.20,
        0.30, 1.00, 0.10,
        0.20, 0.10, 1.00
        ),
        nrow=3, ncol=3,
        byrow=TRUE)

reliabilities <- factor.score.reliability(Lambda = Loadings, Phi = InterCorr, Estimators = c("Regression", "Bartlett",
"McDonald"))
lapply(reliabilities, round, 3)
```

## Appendix B

**SPSS-script for reliabilities of factor score estimators**

```
* ' This function computes and returns reliability estimates for three commonly used
  ' Factor Score Estimators in Factor Analyses,
  '
  ' Explanations of the algebraic formulas are presented in the manuscript
  '
  ' André Beauducel (\email{beauducel@uni-bonn.de})
  ' Christopher Harms (\email{christopher.harms@uni-bonn.de})
  ' Norbert Hilger (\email{nhilger@uni-bonn.de})
/*.
MATRIX.
* Users may enter their respective numbers into the loading matrix:.
compute L={
 0.50,-0.10, 0.10;
 0.50, 0.10, 0.10;
 0.50, 0.10,-0.10;
-0.10, 0.50, 0.15;
 0.15, 0.50, 0.10;
-0.15, 0.50, 0.10;
 0.10, 0.10, 0.60;
 0.10,-0.10, 0.60;
 0.10, 0.10, 0.60
}.
print L/format=F5.2.

* Enter respective numbers into factor inter-correlations.
compute Phi={
 1.00, 0.30, 0.20;
 0.30, 1.00, 0.10;
 0.20, 0.10, 1.00
}.
print Phi/format=F5.2.

* Reproduce the observed covariances from the parameters of the factor model.
compute Sig=L*Phi*T(L).
compute Sig=Sig-Mdiag(diag(Sig))+ident(nrow(L),nrow(L)).

* Calculate specificity/uniqueness/error of items.
compute Psi=Mdiag(diag(Sig-L*Phi*T(L)))&**0.5.

* Equation 9.
```

```
compute Rtt_r = INV( Mdiag(diag( Phi*T(L)*INV(Sig)*L*Phi )) )&**0.5 *
Mdiag(diag(Phi*T(L)*INV(Sig)*L*Phi*T(L)*INV(Sig)*L*Phi)) *
INV(Mdiag(diag(Phi*T(L)*INV(Sig)*L*Phi)))&**0.5 .

* Equation 11.
compute Rtt_b=INV( Mdiag(diag(INV(T(L)*INV(Psi)&**2*L) + Phi)) ).

* Equation 12.
CALL svd(phi, QQ, eig, QQQ).
compute N=QQ*abs(eig)&**0.5.
compute help=T(N)*T(L)*INV(Psi)&**2*Sig*INV(Psi)&**2*L*N.
CALL svd(help, QQ, eig, QQQ).
compute help12=QQ*((eig)&**0.5)*T(QQ).

compute Rtt_m=Mdiag(diag(
INV(help12)*T(N)*T(L)*INV(Psi)&**2*L*Phi*T(L)*INV(Psi)&**2*L*N*INV(help12)
)).
print/Title "Reliabilities for Regression factor score estimators:".
print {T(diag(rtt_r))}/Format=F6.3.
print/Title "Reliabilities for Bartlett factor score estimators:".
print {T(diag(rtt_b))}/Format=F6.3.
print/Title "Reliabilities for McDonald factor score estimators:".
print {T(diag(rtt_m))}/Format=F6.3.
END MATRIX.
```

### Copyrights

International Journal of Statistics and Probability; Vol. 5, No. 6; November 2016
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

# A Comprehensive Analysis of the Determinants of Swap Problem in the Supply Chain of the Petroleum Industry

Raed Al-Hussain[1] & Reza Khorramshahgol[1]

[1] College of Business Administration, Dept. of Quantitative Methods and Information Systems, Kuwait University, P.O. Box 5486, Safat 13055, Kuwait.

Correspondence: Raed Al-Hussain, College of Business Administration, Dept. of Quantitative Methods and Information Systems, Kuwait University, P.O. Box 5486, Safat 13055, Kuwait. E-mail: raed@cba.edu.kw

Received: October 1, 2016 Accepted: October 21, 2016 Online Published: October 28, 2016

doi:10.5539/ijsp.v5n6p108 URL: http://dx.doi.org/10.5539/ijsp.v5n6p108

**Abstract**

Applying mathematical modeling to solve swap problems, specifically in the petroleum industry, have proven to help the decision makers to better determine what, where, and how much to swap in order to reduce supply chain (SC) costs and improve its surplus. However, for a better determination of the alternatives and a more profound evaluation of the tradeoffs among them, a comprehensive analysis of the results and a thorough investigation of their impact on the parties involved in swap are crucial. This research performs a detailed sensitivity analysis of the swap problem to examine the effect of different operational parameters on the cost savings realized along the supply chain of the organizations involved in swap. Findings of this study suggest that, if performed properly, swap can significantly reduce supply chain costs and may result in substantial savings, creating a win-win situation for all parties engaged in swap.

**Keywords:** Supply chain management, the swap practice, mathematical modeling, sensitivity analysis, petrochemicals industry.

## 1. Introduction

Swap transactions among same-level supply chain partners, specifically in the petroleum industry (where suppliers are scattered in distant geographical locations), can offer companies great cost savings. However, making decisions in such collaborations can be very complicated. Swapping commodities with other organizations (even with competitors) can drastically shrink transportation costs and reduce risks in comparison to shipping goods long distances to internationally remote locations. Despite the massive advantages, companies which have adopted the swap practice are still not reaping the complete savings that can be had through swap practices.

Al-Hussain et al. (2006) argue that one reason companies are not getting full savings benefits is because decisions surrounding swaps between two companies are often solely made using judgmental approaches and spreadsheets. Al-Hussain et al. (2008) proposed a mixed integer programming (MIP) model to provide a comprehensive analysis of the swap practice. The model included its limitations and strengths. Two earlier studies used linear programming (Khorramshahgol et al., 2010) and Goal Programming (Khorramshahgol et al., 2014) to help SC designers and managers better decide on what, where, and how much to swap with competitors. Results of these latter two studies show that the use of systematic approaches (such as mathematical modeling) to swap practices outperform the judgmental approaches currently in use. This can result in an increased savings of about 20% for companies involved in swap.

Most practitioners completely ignore sensitivity analysis (SA) when using mathematical programming models. Often, mathematical optimization methods are constructed that incorporate unchanging (rigid) constraints and which use the results of the static models for solving dynamic problems (such as the ones presented in any supply chain management (SCM). The basic idea in sensitivity analysis is to change the model and observe its results (Kleijnen, 1992). In practice the decision maker determines what to vary and what to observe (Eschenbach, et al. 1989; Eschenbach et al. 1990; French 1992).

The dynamic nature of SCM mandates performing SA as the first step in post optimality analysis (Belvardi, et al, 2012, Li, et al, 2016). Several very recent studies have incorporated sensitivity analysis into SCM decision analysis (Kim et al., 2014; Ameda, 2014; Zheng et al., 2016).

In this paper, a comprehensive investigation of swap problem in the oil industry is provided, considering factors such as demand pattern, production capacity, volume owed, sharing periodicity, price pattern, and their interactions. This research highlights some managerial aspects of the swap practice in order to help SC designers and managers fully utilize the potential benefits that swap practices can offer.

## 2. Methodology

Two earlier studies (Khorramshahgol et al., 2010; Khorramshahgol et al., 2014), offered models based on linear programming (LP) and goal programming (GP) to solve the swap problems for two petrochemicals companies (named A and B). This paper investigates the sensitivity of the results of the swap analysis (from the two earlier research studies) due to changes in demand pattern, production capacity, volume owed, sharing periodicity (τ), and rice pattern. Table 1 shows the variation in parameters and their relative values.

Table 1. Parameters to be varied in the Sensitivity Analysis and their Relative Values

| Parameters | Variation of Parameters | Num. of Levels |
|---|---|---|
| Demand | Base Case<br>Increasing<br>Decreasing<br>Cyclic | 4 |
| Production Capacity | Unconstrained<br>Constrained:<br>Average of Max. demand +<br>20% capacity cushion<br>10% capacity cushion | 3 |
| Volume Owed (VO) | Unconstrained<br>6000 MT<br>3000 MT | 3 |
| Sharing Periodicity τ | Every period<br>Quarterly<br>Twice a year | 3 |
| Price of Commodity | $450 (at the time of this study)<br>Increasing<br>Decreasing<br>Cyclic | 4 |

The parameters shown in Table 1 were determined by the authors and the SC directors in two companies from petrochemical industry (for anonymity, we refer to them as Company A and Company B). These parameters are those believed either to have the greatest potential for variation and are uncontrolled by firms, (e.g., demand and price patterns of commodity), or those that can possibly be controlled by firms (e.g., production capacity, volume owed, and τ). Nonetheless, both groups when varied can have significant effects on cost savings.

The demand for each company (A and B) is used to generate three different demand patterns (increasing, decreasing, and cyclic) for that company. For example, the increasing demand patterns for both companies are generated such that demand starts from the minimum demand value detected in the case, then increases for 20 periods until the sum of the generated demand is equal to the sum of the actual demand in the case. The decreasing and cyclic demand patterns are generated in a similar fashion.

In addition, information such as the minimum, maximum, average, and the range of the actual price of the commodity under study (i.e., Mono Ethylene Glycol—MEG) during the same time period in the case is used to generate three different price patterns (increasing, decreasing, and cyclic).

Since the current model is designed to handle one type of commodity at a time, it is reasonable to assume that the demand and price patterns are the same for both companies during any swap period. For example, if Company A's demand pattern increases while the price pattern is cyclic, then the same is assumed to be true for Company B's demand and price patterns.

It is important to mention, however, that although prices and demand are closely related in most industries, oil and petrochemicals have a very low price elasticity of demand. In other words, prices have to soar considerably to affect demand even a little (Anon, 2002). The short-run demand is therefore inelastic because petroleum plays a critical role in today's economy. Thus, the price elasticity of demand's effect is ignored in this study and price variation is assumed to have no influence on demand values.

In order to test the effect of production capacity on supply chain savings, production capacity is constrained to two levels:

(1) The average of both companies' maximum demand plus 20% of the capacity cushion:

(16,413MT x 0.2) + 16,413MT = 19,695MT, and

(2) The average of both companies' maximum demand plus 10% of the capacity cushion:

(16,413MT x 0.1) + 16,413MT = 18,054MT.

Since the model does not allow lost sales or backorders, assuring that all demands are met is necessary in order to obtain a feasible solution. As the production capacity of both companies is assumed to be equal, an average of the maximum demand of both companies is used.

Based on Table 1, there are (4)(3)(3)(3)(4) = 432 different scenarios in the sensitivity analysis case. Since the result of each scenario represents an optimal solution based on the given values of the parameters, it is not only desired to find which scenario will result in the best optimal solution and maximum savings, but also to observe the behavior of the swap model under the interaction of the variation of the operational parameters provided in Table 1. Accordingly, regression analysis is applied to explore the relationship between the parameters and their effect on supply chain savings. It is recommended that prior to conducting a regression analysis, unit root test of the variables be conducted to determine the stationary property of the variables (Sahin, et al. 2008).

Applying sensitivity analysis to the case is expected to provide further insight into the swap practice. It can help managers make better decisions regarding a swap agreement when parameters are expected to change. It can also help managers avoid negative consequences when market performance is poor.

## 3. Results

In order to evaluate the impact of operational parameters on supply chain savings, regression analysis was performed. Supply chain savings represents the dependent variable, and the independent variables are represented by the following operational parameters and interaction terms:

Demand pattern; Capacity constraints; Volume owed (VO) constraints; Sharing periodicity (τ);

Price pattern; Interaction of demand pattern and capacity constraints; Interaction of demand pattern and VO; Interaction of demand pattern and τ; Interaction of demand pattern and price pattern; Interaction of capacity constraints and VO; Interaction of capacity constraints and τ; Interaction of capacity constraints and price pattern; Interaction of VO and τ; Interaction of VO and price pattern;

Interaction τ and price pattern

The estimated regression equation, along with the standard error of estimation, the coefficient of determination R-square, and the adjusted R-square are:

Savings = 802639 + 45735 Demand + 0.034 Capacity + 13.7 VO – 48367 τ- 2379 Price - 0.0108 Dem.Cap – 6.09 Dem.VO + 5496 Dem. τ+ 406 Dem.Price +0.000002 Cap.VO – 0.0017 Cap. τ– 0.0001 Cap.Price + 3.48 VO. τ+ 0.184 VO.Price – 250 τ.Price

S = 37314 R-Sq = 72.1% R-Sq(adj) = 71.1%

The regression coefficients, their related standard error, t-value, and p-value are shown in Table 2 and Table 3 provides the analysis of variance.

Table 2. The Regression Analysis Coefficients

| Predictor | Coefficient | SE Coefficient | T | P |
|---|---|---|---|---|
| Constant | 802639 | 22575 | 35.55 | 0.000 |
| Demand | 45735 | 6209 | 7.37 | 0.000 |
| Capacity | 0.034 | 0.2019 | 0.17 | 0.866 |
| VO | 13.677 | 2.431 | 5.63 | 0.000 |
| τ | -48367 | 3636 | -13.3 | 0.000 |
| Price | -2379 | 6209 | -0.38 | 0.702 |
| Dem.Cap | -0.01077 | 0.04198 | -0.26 | 0.798 |
| Dem.VO | -6.0947 | 0.56 | -10.88 | 0.000 |
| Dem.τ | 5495.7 | 781.4 | 7.03 | 0.000 |
| Dem.Price | 406 | 1436 | 0.28 | 0.778 |
| Cap.VO | 0.00000151 | 0.00001637 | 0.09 | 0.926 |
| Cap.τ | -0.00169 | 0.02284 | -0.07 | 0.941 |
| Cap.Price | -0.00006 | 0.04198 | 0 | 0.999 |
| VO.τ | 3.4826 | 0.3047 | 11.43 | 0.000 |
| VO.Price | 0.1842 | 0.56 | 0.33 | 0.742 |
| τ.Price | -249.9 | 781.4 | -0.32 | 0.749 |

Table 3. Analysis of Variance

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 15 | 1.5001E+12 | 1.00E+11 | 71.83 | 0.000 |
| Residual Error | 416 | 5.79199E+11 | 1392305284 | | |
| Total | 431 | 2.0793E+12 | | | |

The coefficients of determination (R2), the adjusted R2, the F-statistics, and its associated p-value all suggest a very good fit in the swap model. The adjusted R2 measure indicates that the independent variables accounted for 71% of the total variation in the supply chain cost savings. Moreover, the F-statistic of 71.83 and its associated p-value of 0.000 imply that the swap model is highly significant. This means that the null hypothesis which states that there is no relationship between the dependent and any independent variables" can be rejected. Moreover, Table 2 indicates that not all operational parameters are statistically significant and that some operational parameters' impact is overruled by their significant interaction effect with other parameters.

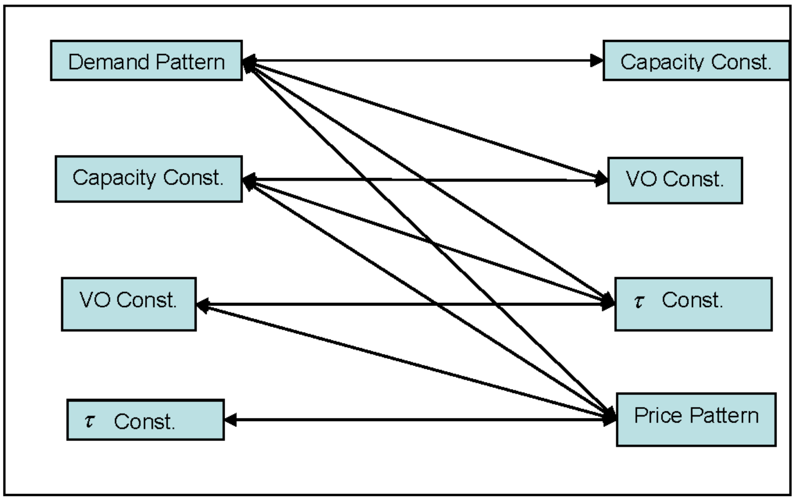


Figure 1. Interaction Terms Used in the Regression Analysis

### *3.1 Impact of the Interaction Terms*

All possible two-way interactions between the independent variables are tested in the regression analysis, as indicated in Figure 1. The inclusion of such interaction terms helps explain more about the behavior of the swap practice. However, not every interaction term proved to be significant in the model. Therefore, only the significant interaction terms will be discussed next.

#### 3.1.1 The Interaction Effect of the Demand Pattern and the Volume Owed

The coefficient of the demand pattern and VO interaction term in the regression equation, along with its t-value and p-value in Table 2 suggest that such interaction is statistically significant. This is evident in the example shown in Figure 2.

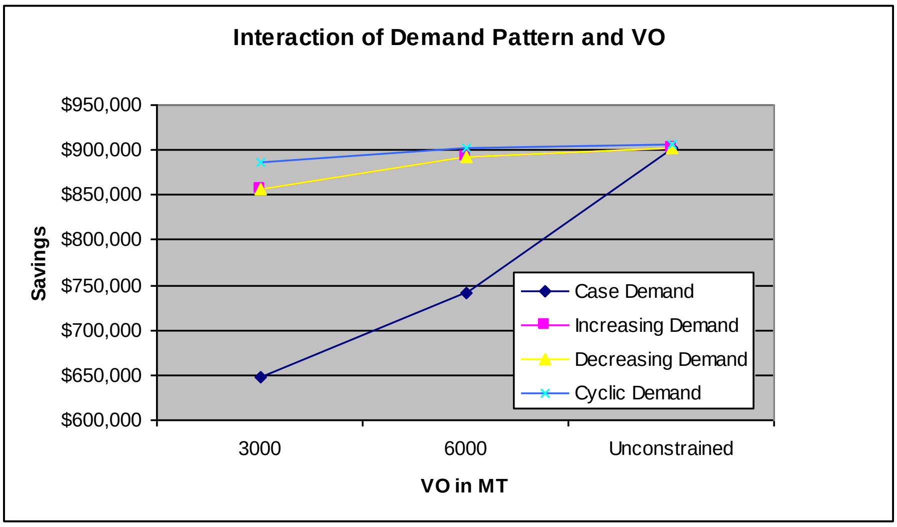


Figure 2. Interaction of Demand Pattern and Volume Owed

Results indicate that different demand patterns generate different savings under the same value of VO. Hence, constraining the value of VO before considering the demand pattern can eliminate chances of higher savings. For example, constraining the value of VO to 3000 Metric Tons (MT) per swap when the demand pattern is increasing generates $856,426 of savings. On the other hand, if the demand pattern is as given in the case, then constraining the value of VO to 3000 MT per swap reduces savings to $648,268. Therefore, before constraining the value of VO to any specified volume per swap, decision makers should first take into consideration the demand pattern.

It is also important to note that as the value of VO increases, savings also increase for all demand patterns. This is due to the fact that when the value of VO is constrained; the opportunities to swap are reduced, which means there are fewer chances to increase savings. The demand pattern of the case appears to have the lowest savings among all demand patterns under all values of VO. However, as the value of VO increases, a percentage increase in the supply chain savings of the demand pattern is the highest among all demand patterns. For example, a supply chain savings of the demand pattern increased by 14% when the value of VO increased from 3000 MT to 6000 MT per swap. On the other hand, the average percentage in supply chain savings increased by about 3.4% for all other demand patterns when the value of VO increased from 3000 MT to 6000 MT per swap.

At the time of this research, Companies A and B were pushing toward constraining the VO to 3000 MT per period. According to the output of the model, this is the lowest attainable savings among options ($648,268), and hence not the best decision to make. Therefore, companies planning to participate in a swap agreement and seeking maximum savings are urged to relax the value of VO as much as possible in order to fully utilize the savings opportunities offered by the swap practice.

#### 3.1.2 The Interaction Effect of the Demand Pattern and the Sharing Periodicity τ

The coefficient of the demand pattern and the sharing periodicity interaction term in the regression equation, along with its t-value and p-value (Table 2), suggest that such an interaction is statistically significant. Figure 3 shows the effect of the interaction of demand patterns and τ on supply chain savings.

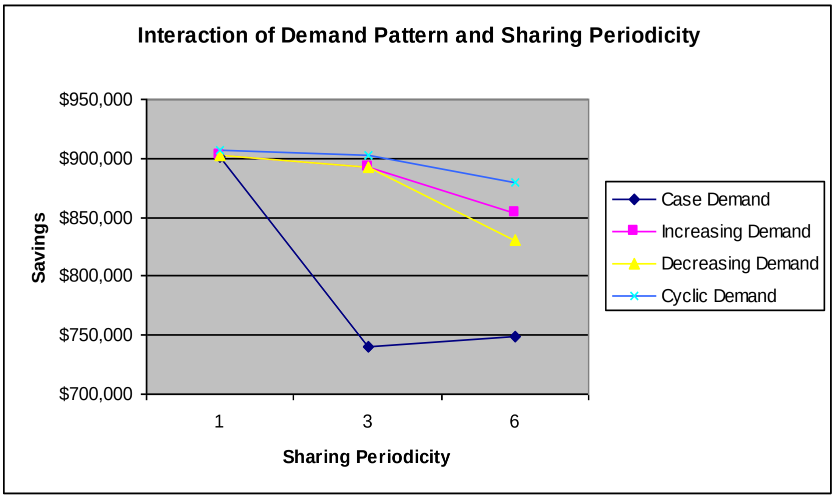


Figure 3. Interaction Effect of Demand Pattern and Sharing Periodicity

When the value of τ is set to one period, supply chain savings are at a maximum and relatively equal for all demand patterns. Then the supply chain savings are reduced as the value of τ is increased to three and six periods. Again, the demand pattern of the case is the most sensitive among all other demand patterns when τ is increased from 1 to 3 periods. The demand pattern of this case resulted in an 18% decrease in supply chain savings while the average decrease of supply chain savings for all other demand patterns was about 0.09%.

It is first expected that supply chain savings will always decrease as τ increases for all demand patterns. In other words, a higher sharing periodicity will always result in lower savings, regardless of the demand pattern. However, the results show that this is not generally the case. While the generated demand patterns (increasing, decreasing, and cyclic) all resulted in decreased supply chain savings as τ increased, the demand pattern of the case exhibited a unique phenomenon. Unlike the rest of the demand patterns, the demand pattern generated higher supply chain savings when τ was set to six periods than the case when τ was set to three periods. In fact, the supply chain savings actually tended to alternate depending on the demand pattern as the value of τ increased. This phenomenon, when tested on the demand pattern of the case, is shown in Figure 4. From the data we can conclude that a higher sharing periodicity does not always generate lower supply chain savings for demand patterns.

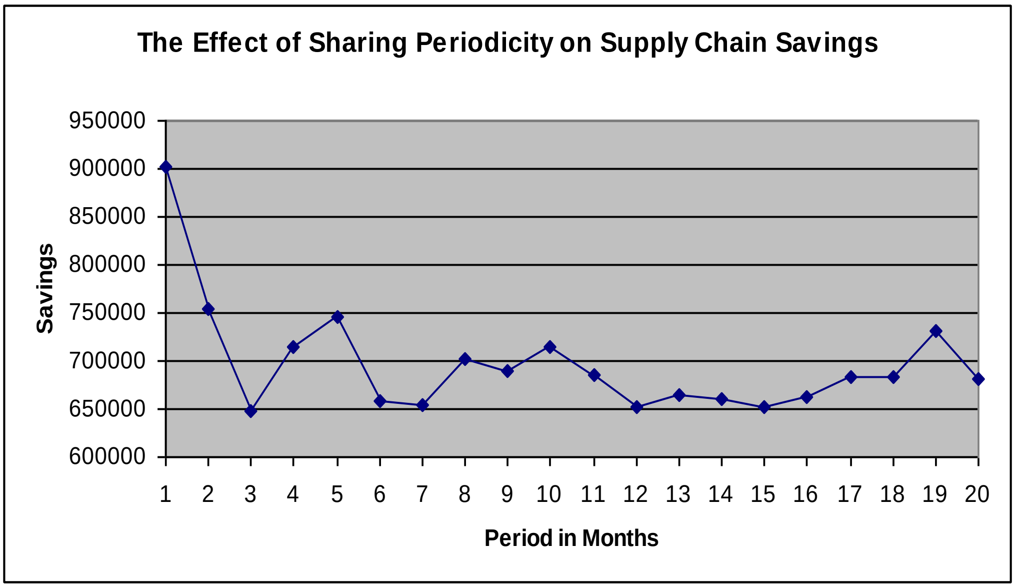


Figure 4. The Effect of Sharing Periodicity on Supply Chain Savings

Currently, companies involved in the study share their supply chain savings every quarter. According to the results, this value of τ generates the lowest supply chain savings versus the cases when τ is set to 1 or 6 periods. Therefore, companies interested in undergoing a swap agreement are urged to share their supply chain savings every period. If this is not possible, then based on their demand patterns, companies should first test the effect of all possible scenarios of τ on supply chain savings, as in Figure 4, then implement the best τ possible.

3.1.3 The Interaction Effect of the Volume Owed and the Sharing Periodicity τ

The regression analysis output implies that the effect of the interaction of the VO and τ on cost savings is statistically significant. Figure 5 illustrates the interaction effect and its impact on supply chain savings.

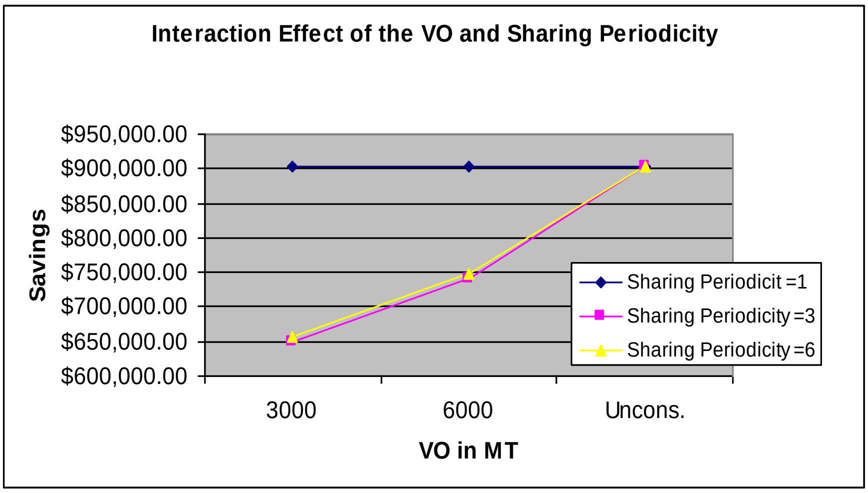


Figure 5. Interaction Effect of the VO and Sharing Periodicity

When τ is set to 1 period, supply chain savings are unaffected by the changes in the value of VO. This is a rational finding, considering the fact that no matter how much volume the companies owe each other, the opportunity costs of the volume owed by one company is neutralized by the opportunity of savings gained by the other, and hence should have no impact on savings. This remains true as long as the opportunities of costs/savings are instantly shared. When τ is set to 3 or 6 periods, supply chain savings are dramatically affected with changes in the value of VO. However, there seems to be a small difference in supply chain savings when τ is set to 3 and 6 periods for the VO values of 3000 and 6000 MT per swap. In addition, supply chain savings converge at the same value when VO is unconstrained. Thus, the value of τ has no impact on supply chain savings when VO is unconstrained.

Therefore, companies interested in undergoing a swap agreement can experiment with the sharing periodicity and volume owed parameters' values to reach their best option. For example, if companies must constrain the value of VO to 3000 MT per swap, then τ should be set to 1 period to gain maximum savings. In contrast, if τ must be set to other values, such as 3 periods, then the value of VO should be unconstrained, or as relaxed as possible in order to maximize savings.

*3.2 Impact of Individual Parameters*

3.2.1 Impact of the Production Capacity Constraints

The regression analysis output suggests that the effect of production capacity constraints has no statistical significance on supply chain savings. The value of the capacity coefficient of 0.034 is very small with a t-value of 0.17 and an associated p-value of 0.866. Figure 6 shows the average supply chain savings under different capacity constraints values.

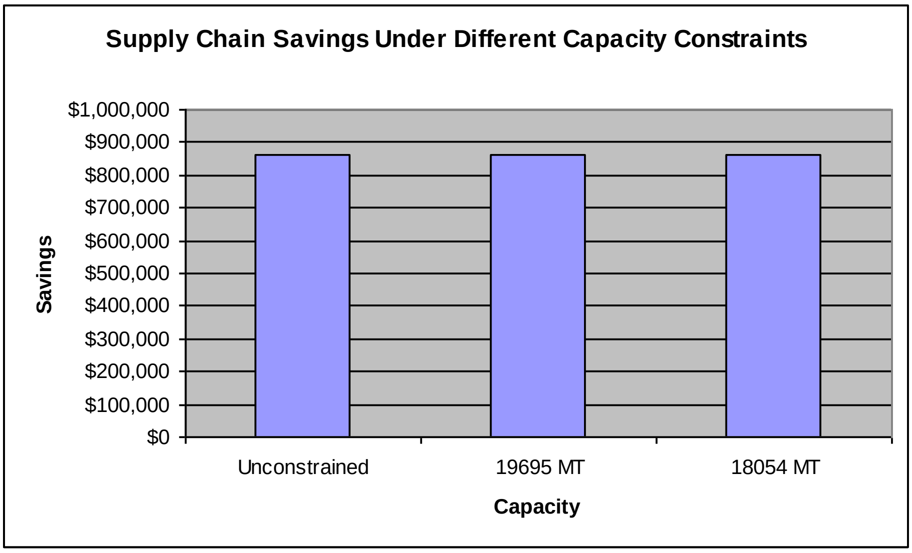


Figure 6. Supply Chain Savings under Different Values of Capacity Constraints

Nevertheless, the result on the significance of capacity constraints on supply chain savings is not conclusive. In general, the less production capacity available, the fewer commodities there are to meet the supply chain partner’s demand, and hence opportunities to swap and save in transportation costs are eliminated. However, because the swap model does not take into account lost sales or backorders, the levels of production capacity used in the sensitivity analysis are always enough to meet customer demands. Otherwise an infeasible solution is obtained. In order to obtain further insight into the significance of production capacity, a model is required that can take into account lost sales or backorders.

### 3.2.2 Impact of the Price Pattern

Although results suggest that in some scenarios different price patterns result in different supply chain savings, the regression analysis output implies that the general effect of a price pattern on supply chain savings is statistically insignificant (p-value = 0.702). This is because the differences in supply chain savings (as a result of different price patterns) are relatively small when patterns averaged over all of the supply chain savings with all different prices. The formulation of the model also takes into account the price value when the amount to be paid back to each company is calculated. In fact, the swap agreement between companies is solely based on volume, ignoring the effect of prices variation. Hence, this result supports their strategy. Figure 7 shows the average supply chain savings when different price patterns are applied to the model.

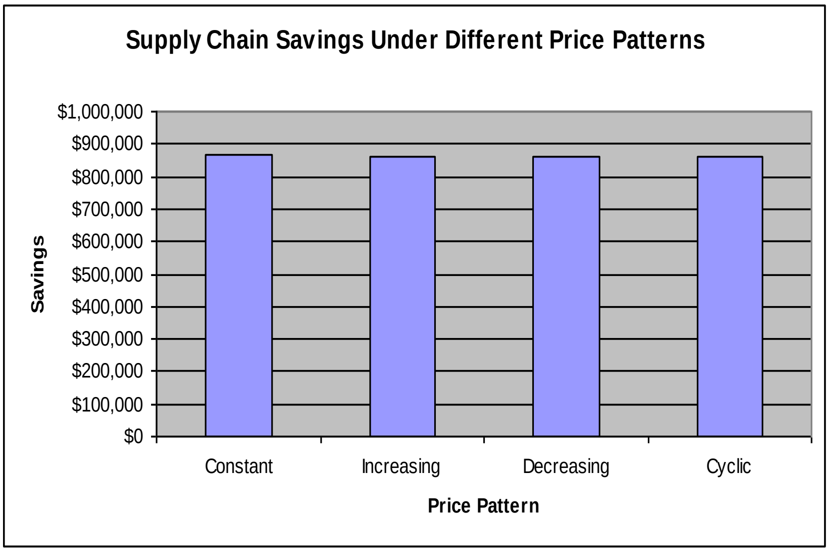


Figure7. Supply Chain Savings under Different Price Patterns

### 3.2.3 Impact of the Demand Pattern

According to the regression analysis output, the demand pattern is statistically significant. The demand pattern has the second largest coefficient in magnitude ($45,735). The t-value for the demand pattern is 7.37 with an associated p-value of 0.000. Hence, the demand pattern has a significant effect on supply chain savings in the swap model. Figure 8 shows the effect of different demand patterns on supply chain savings.

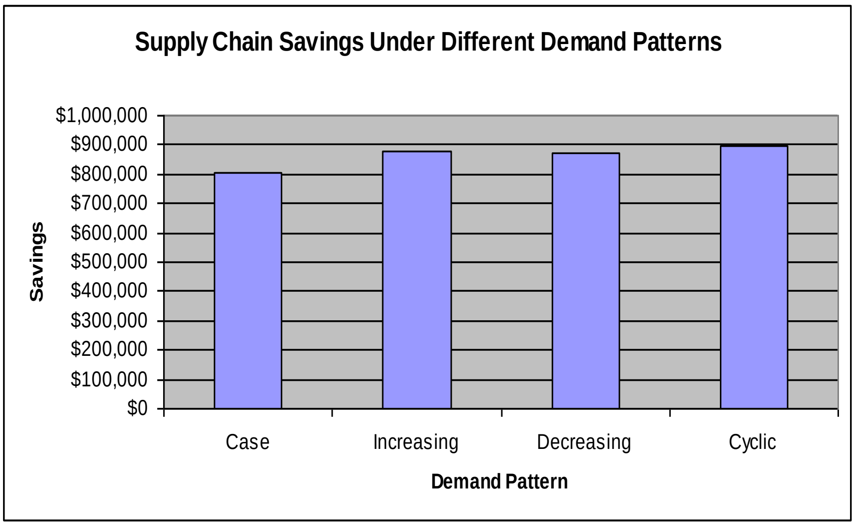


Figure 8. Supply Chain Savings under Different Demand Patterns

The swap model generates the greatest savings under the cyclic demand pattern when all other parameters are constant across all other demand patterns. However, although the swap model generates the most savings when the demand pattern is cyclic, other operational parameters, such as VO and sharing periodicity, influence the magnitude of the savings. This is also due to the significant interaction effect that the demand pattern has on these parameters. Therefore, for an accurate interpretation of the effects of demand pattern on supply chain savings, the interaction effects of a demand pattern with the VO and sharing periodicity need to be taken into consideration.

### 3.2.4 Impact of the Volume Owed Constraints

The regression analysis coefficient, t-value, and p-value for VO are 13.677, 5.63, and 0.000 respectively. The t-value and the p-value suggest that the effect of such parameters on supply chain savings is significant indeed. Figure 9 shows the average supply chain savings under different levels of the volume owed per swap between supply chain partners.

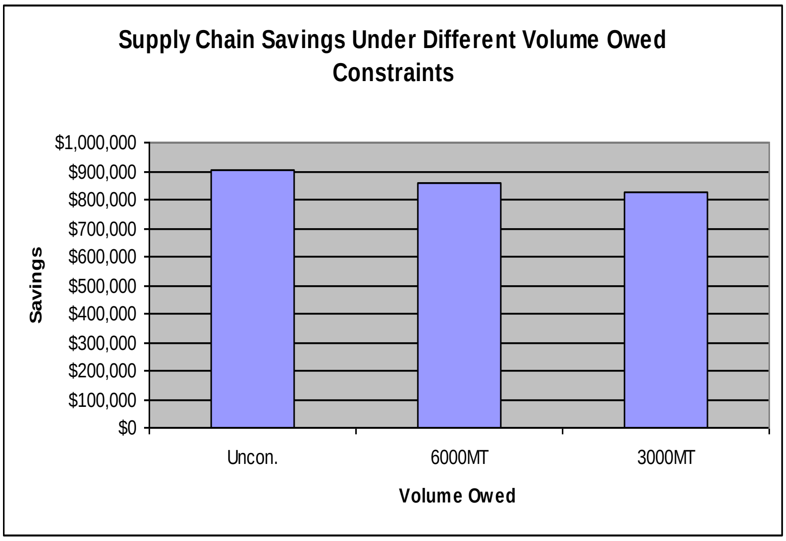


Figure 9. Supply Chain Savings under Different Volume Owed Constraints

The VO is the difference of total expenditures between supply chain partners when serving each other’s customers in terms of volume. Hence, the higher the VO value, the more volumes may be swapped, and the more savings are generated.

On average, the differences in supply chain savings when using different levels of VO constraints is negligible relative to the overall savings. But, VO constraints possess a significant effect on supply chain savings due to their significant interaction effect with demand pattern and sharing periodicity, as discussed earlier. An accurate interpretation of the VO constraints need to be taken into account when considering the interaction effect.

#### 3.2.5 Impact of the Sharing Periodicity (τ)

The variation of the value of sharing periodicity (τ) has the highest impact on supply chain savings. This is evident in the regression analysis output in Table 2. The coefficient of the sharing periodicity is the highest in magnitude among all other coefficients (-$48,367) for the values of τ tested (1, 3, and 6 periods). The t-value and the p-value of -13.30 and 0.000 respectively suggest that the effect is statistically significant. Figure 10 shows the average supply chain savings under different sharing periodicity values.

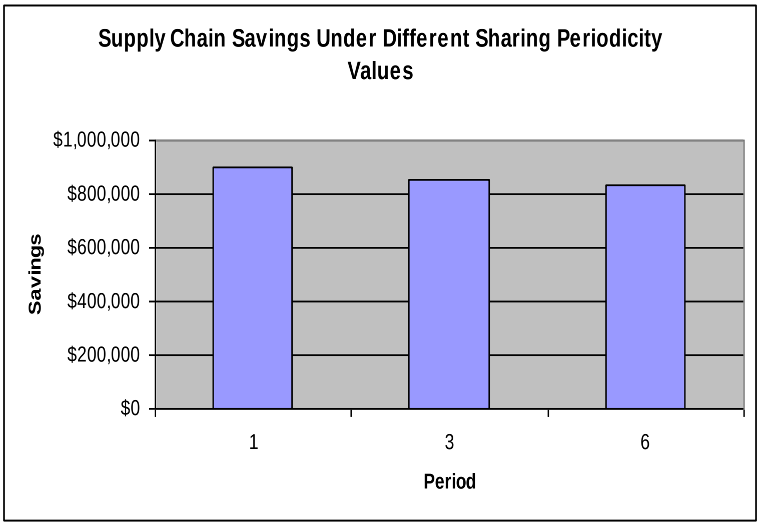


Figure 10. Supply Chain Savings under Different Levels of Sharing Periodicity

Although the values of τ tested in the sensitivity analysis show that on average, supply chain savings decrease as τ increases, this is not a conclusive result. Figure 11 shows that a trend of supply chain savings when τ increases or decreases does not exist. Hence, companies interested in undergoing a swap agreement are urged to test their supply chain savings under different values of τ.

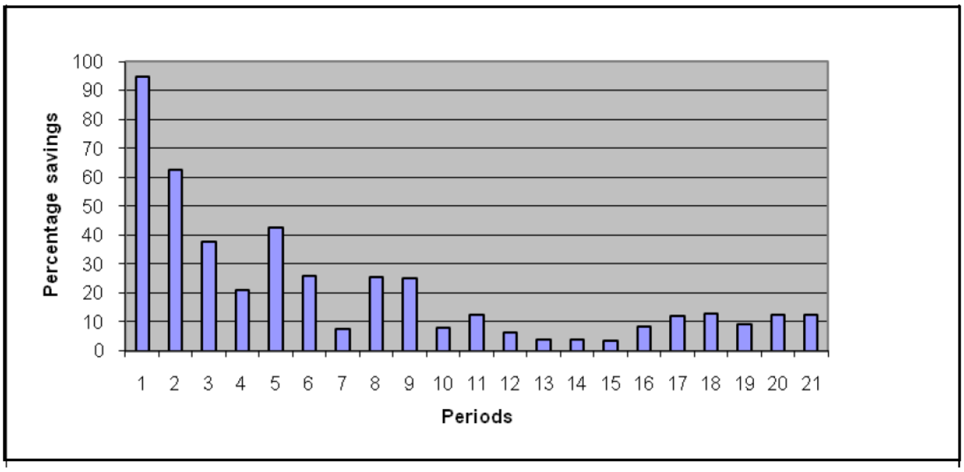


Figure 11. Cumulative Percentage Savings for Company A throughout the Entire Swap Period Before Sharing

### *3.3 Other Sensitivity Analysis Related Issues*

When interested in swapping with competitors, companies might want to consider their gain/loss process throughout the swap period in addition to the overall supply chain savings. In scenario 5 of the base case, where demand is as given in the case, capacity is unconstrained, VO is unconstrained, τ is set to three periods, and the price is set constant at $450/MT, while both companies saved in the overall supply chain cost at the end of the two-year period, Company B did most of the work and was suffering more costs than Company A throughout the swap period. Figures 11 and 12, show how much more each company made throughout the swap period relative to the No-Swap scenario in order to achieve the desired savings.

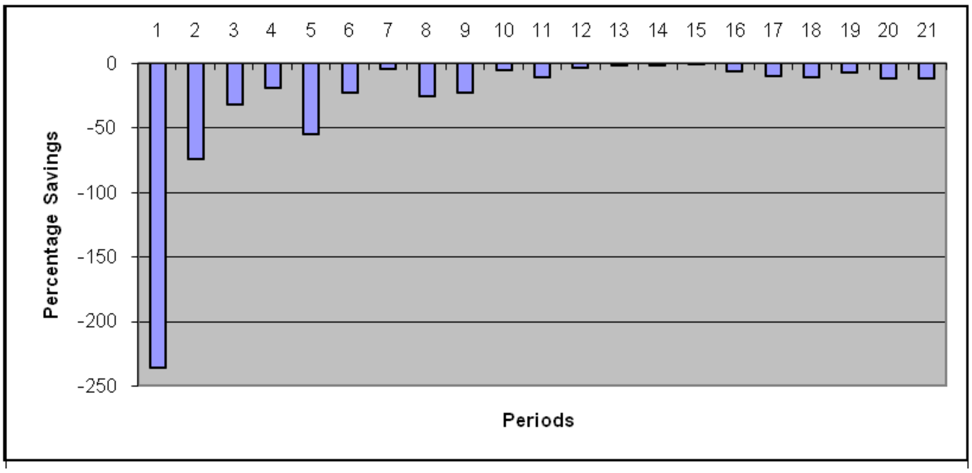


Figure 12. Cumulative Percentage Savings for Company B throughout the Entire Swap Period Before Sharing

If results of the base case, shown in Figures 11 and 12, is of concern to a company that is interested in undergoing a swap agreement with a competitor, then some restrictions should be enforced and agreed on in the swap contract. For example, restrictions on the VO between supply chain partners during a swap or on the production capacity availability for a swap can reduce the percentage of gain/loss for a company throughout the swap period. However, this comes with the expense of reducing the overall supply chain savings.

Figures 13 and 14 show the impact on both companies when applying constraints on the value of VO between supply chain partners, where demand pattern is as given in the case, capacity is set to be unconstrained, VO is set to 6000 MT only, τ is et to 3 periods, and the price pattern is set constant at $450/MT. While Company B has reduced its loss during the swap period (sharing a greater load with Company A), the overall supply chain saving has been reduced. In the scenario where the value of VO was unconstrained, the overall supply chain savings for each company was $902,198.94. On the other had, when the value of VO was constrained in order to share the load during the swap period, the overall supply chain savings per company came to $740,894.68. Accordingly, different swap strategies have different outcomes. Some offer great savings, but they come at the expense of a heavier workload, while others offer fewer savings and a lighter workload. Hence, the type of strategy to settle on remains the choice of companies interested in joining into a swap agreement.

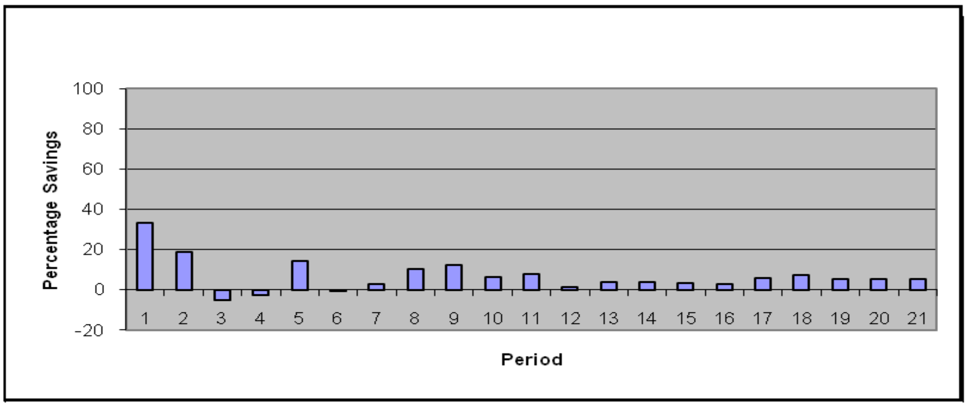


Figure 13. Cumulative Percentage Savings for Company A throughout the Entire Swap Period Before Sharing

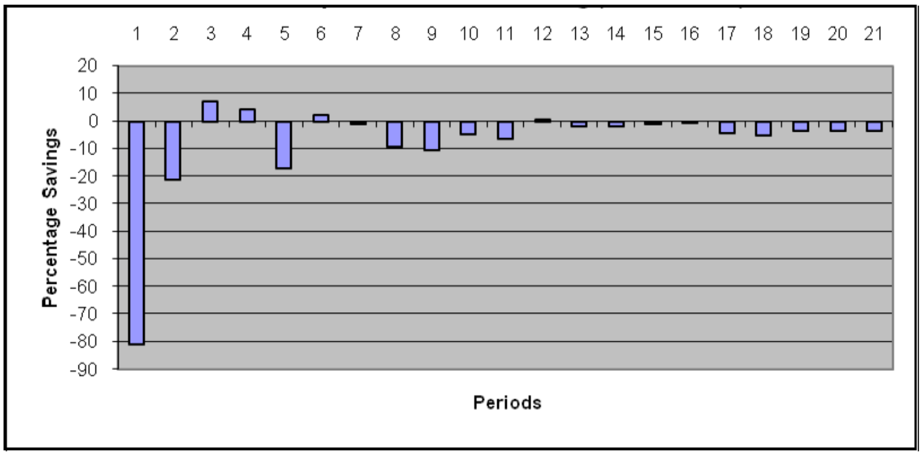


Figure 14. Cumulative Percentage Savings for Company B throughout the Entire Swap Period Before Sharing

## 4. Conclusion

The overriding purpose for this study was to perform a detailed sensitivity analysis of the swap problem to examine the effect of different operational parameters on the cost savings realized by parties involved in swap. Two companies from the petrochemical industry were chosen. These two companies (named company A and company B for anonymity) had been engaged in swapping their products for several years. A set of 432 different scenarios were developed by varying the values of demand pattern, capacity constraints, VO, sharing periodicity (τ), and price pattern. These scenarios were examined on swap practices of Companies A and B.

Results of the sensitivity analysis suggest that swapping in general can reduce supply chain costs. It was suggested that when swap decisions are made, the behavior of the parameters mentioned earlier (e.g., demand pattern, capacity constraints, VO, sharing periodicity) should be taken into consideration and various scenarios be evaluated to better determine *what, where, and how much to swap* in order to reduce supply chain (SC) costs. It was also shown that the interaction among these parameters can have a substantial impact on supply chain savings.

## References

Al-Husain, R., Assavapokee, T., & Khumawala, B. (2008). Modeling the Supply Chain Swap Problem in the Petroleum Industry. *International Journal of Applied Decision Sciences*, *1*, 261-281. http://dx.doi.org/10.1504/IJADS.2008.021223

Anon (2002). Biofuel use promoted by European Parliament. *Lipid Technology*, *14*, 75. 2.

Belvardi, G., Kiraly, A., Varga, T., Gyozsan, Z., & Abonyi, J. (2012). Monte Carlo Simulation Based Performance Analysis of Supply Chains. *International Journals of Managing Value and Supply Chains*, *3*, 1-14. http://dx.doi.org/10.5121/ijmvsc.2012.3201

Eschenbach, T. G., & Gimpel, R. J. (1990). Stochastic sensitivity analysis, *The Engineering Economist, 35*, 305-321. http://dx.doi.org/10.1080/00137919008903024

Eschenbach, T. G., & McKeague, L. S. (1989). Exposition on using graphs for sensitivity analysis. *The Engineering Economist*, *34*, 315-333. http://dx.doi.org/10.1080/00137918908902996

French, S. (1992). Mathematical programming approaches to sensitivity calculations in decision analysis. *Journal of the Operational Research Society*, *43*, 813-819. http://dx.doi.org/10.1057/jors.1992.120

Hussain, R., Assavapokee, T., & Khumawala, B. (2006). Supply Chain Management in the Petroleum Industry: Challenges and Opportunities. *International Journal of Global Logistics & Supply chain Management, 1*, 90-97.

Khorramshahgol, R., Al-Hussain, R., & Tamiz, M. (2010). Application of Linear Programming to Swap Analysis in Supply Chain Management of Oil Industry. *Journal of Information and Optimization Sciences, 31*, 1375-1388. http://dx.doi.org/10.1080/02522667.2010.10700033

Khorramshahgol, R., Tamiz, M., & Al-Hussain, R. (2014). Application of Goal Programming to Swap Analysis in Oil Industry. *Journal of Information and Optimization Sciences*, *35*, 73-91. http://dx.doi.org/10.1080/02522667.2014.894300

Kim, J., Realff, M., & Lee., J. (2011). Optimal Design and Global Sensitivity Analysis of Biomass Supply Chain

Networks for Biofuels under Uncertainty. *Computers and Chemical Engineering*, *35*, 1738-1751. http://dx.doi.org/10.1016/j.compchemeng.2011.02.008

Kleijnen, J. P. C. (1995) Sensitivity analysis and optimization of system dynamics models: regression analysis and statistical design of experiments. *System Dynamics Review*, *11*, 1-14. http://dx.doi.org/10.1002/sdr.4260110403

Li, C., Xiang, X., & Qu, Y. (2016). Product Quality Dynamics in Closed-Loop Supply Chains and its Sensitivity Analysis. *The Journal of Grey Systems*, *28*, 80-190.

Sahin, A., Yilmaz, A., & Atakan, C. (2008). An Investigation on the Shuttle Trade Dynamics of a Small-Open-Economy. *International Journal of Economic Sciences and Applied Research*, *1*, 1-12.

Umeda, S. (2014). Sensitivity Analysis of Reverse Supply Chain System Performance by using Simulation. *IFIP International Federation for information Processing*. B. Grabot (editor) 326-333. http://dx.doi.org/10.1007/978-3-662-44736-9_40

Zheng, Y., Liao, H., & Yang, X. (2016). Stochastic Pricing and Order Model with Transportation Mode Selection for Low-Carbon Retailers. *Sustainability*, *8*, 14-23. http://dx.doi.org/10.3390/su8010048

## Copyrights

International Journal of Statistics and Probability; Vol. 5, No. 6; November 2016
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

# Assessing the Risk of Road Traffic Fatalities Across Sub-Populations of a Given Geographical Zone, Using a Modified Smeed's Model

Christian A. Hesse[1], Francis T. Oduro[2], John B. Ofosu[1] & Emmanuel D. Kpeglo[1]

[1] Department of Mathematical Sciences, Faculty of Informatics and Mathematical Sciences, Methodist University College Ghana. P. O. Box DC 940, Dansoman – Accra, Ghana

[2] Department of Mathematics, College of Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

Correspondence: Thomas C. A. Hesse, Department of Mathematical Sciences, Methodist University College, P. O. Box DC 940, Dansoman – Accra, Ghana. E-mail: akrongh@yahoo.com

Received: September 29, 2016 Accepted: October 24, 2016 Online Published: October 31, 2016

doi:10.5539/ijsp.v5n6p121 URL: http://dx.doi.org/10.5539/ijsp.v5n6p121

**Abstract**

Smeed (1949) provided a regression model for estimating road traffic fatalities (RTFs). In this paper, a modified form of Smeed's (1949) model is proposed for which it is shown that the multiplicative error term is less than that of Smeed's original model for most situations. Based on this Modified Smeed's model, Bayesian and multilevel methods are developed to assess the risk of road traffic fatalities across sub populations of a given geographical zone. These methods consider the parameters of the Smeed's model to be random variables and therefore make it possible to compute variances across space provided there is significant intercept variation of the regression equation across such regions. Using data from Ghana, the robustness of the Bayesian estimates was indicated at low sample sizes with respect to the Normal, Laplace and Cauchy prior distributions. Thus the Bayesian and Multilevel methods performed at least as well as the traditional method of estimating parameters and beyond this were able to assess risk differences through variability of these parameters in space.

**Keywords:** risk, Bayesian, multilevel, road traffic fatalities

## 1. Introduction

Smeed (1949) proposed a model for estimating road traffic fatalities (RTFs) in his paper. He showed that the formula

$$\frac{D}{N} = 0.0003\left(\frac{N}{P}\right)^{-\frac{2}{3}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

(were $D$ = Number of RTFs, $P$ = population size and $N$ = number of vehicles in use) gave a fairly good fit to the data from 20 countries, including European countries, USA, Canada, Australia and New Zealand.

Ponnaluri (2012) used data from all states in India to develop seven different models for predicting RTFs and also examined if the individual models were more relevant for application. The seven models, including that of Smeed's, were tested for fitness with the actual data. Smeed's model was found to give the best fit. He showed that the original Smeed formulation cannot simply be discounted due to reasons cited by many researchers. This is because Smeed's model is ***parsimonious in parameter usage.*** According to Ponnaluri (2012), Smeed's model appears to be observation-driven, evidence-based, and logically valid in measuring the ***per vehicle fatality rate.***

The predominant factors affecting RTFs are not the same as those of road traffic accidents (RTAs). Exposures to risk of RTFs (such as human error, environmental/weather, nature of the road and condition of vehicle) are predominant factors influencing road traffic accidents within a geographical region. However, the rate of RTFs is determined by vulnerability to risk (such as insufficient ambulance and emergency medical services, improper pre-hospital care for RTA trauma patients, inadequate safety mechanism in vehicles).

Exposure to risk of RTFs and vulnerability to risk of RTFs are not correlated. Thus, high exposure does not necessarily imply high vulnerability. For instance, Greater Accra Region in Ghana, with the highest exposure to the risk of RTF (due to high population and vehicular densities), has the lowest RTF rate among all the other 9 regions in Ghana. Whilst the three Northern regions of Ghana, with the lowest population density have the highest rate of RTFs (Hesse and Ofosu, 2015). Nigeria and Ghana have almost the same vehicular density. However, inhabitants of Nigeria are more vulnerable to die as result of road traffic accidents. Developing countries, with only about 10% of the world motorization, account for about 85% of annual RTFs in the world (WHO, 2004, 2009). Thus, developed countries, though have ***greater exposure to risk of RTF*****s** due to high vehicular density, however less vulnerable to RTFs compared to developing countries.

Two predominant factors that determine risk of RTFs in a geographical region are

(1) Safety mechanism in vehicles (such as anti-lock braking systems (ABS), air bags and seatbelts),

(2) Emergency medical services (such as Ambulance service).

One reason why developing countries are more vulnerable to risk of RTF is due to the fact that a large proportion of road traffic accident trauma patients in these regions do not have access to formal emergency medical services (Tiska, et al., 2002). Secondly, the ages of vehicles and availability of modern safety mechanisms in vehicles plying the roads in these regions have significant effect on the consequences of road traffic accidents. It is obvious that if greater attention is paid on improving road safety mechanisms (such as anti-lock braking systems (ABS), air bags, better design of cars and increased wearing of seatbelts in cars) there could be substantial benefits in reducing injuries and fatalities with respect to road traffic accidents in developing countries (Hesse, et al., 2014).

Smeed's model is of the form

$$\frac{D}{N} = \alpha\left(\frac{N}{P}\right)^{\beta} e, \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

where $D$ = Number of RTFs, $P$ = population size, $N$ = number of vehicles in use, $e$ = multiplicative error term, and $\alpha$ & $\beta$ are parameters to be estimated. Equation (2) can be expressed as

$$Y = \alpha X^{\beta} e, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

where, the predictor variable is $X = N/P = \textit{vehicular density}$ and the dependent variable is $Y = D/N =$ *per vehicle fatality rate*.

The factors affecting RTAs correspond to exposure $X$ while the factors affecting RTFs correspond to vulnerability given the same exposure. In Smeed's model exposure is measured by the variable $X$ whereas vulnerability for a given $X$ is captured by the parameters $\alpha$ and $\beta$.

Let $X_1$ (with $Y = Y_1$) and $X_2$ (with $Y = Y_2$) be two predictor variables of two geographical regions such that $X_1 = X_2$. If $Y_1 \neq Y_2$, then the different values of $Y$ is not based on $X$ but is due to the fact that $\alpha$ and $\beta$ vary across the two geographical regions. It therefore follows that, the parameters of Smeed's model vary from one geographical region to another. Thus, one could use these parameters to assess variability of the risk of RTFs across geographical regions.

Smeed (1949) and other related studies by Ponnaluri (2012), Ghee *et al*., (1997), Bener and Ofosu (1991), Jacobs and Bardsley (1977), Fouracre and Jacobs (1977) used least squares regression (LSR) method to estimate the parameters. However, the LSR approach:

- does not allow the variability of the parameters,
- is *very sensitive* to violation of the normality assumption.

Thus, we need an estimation method that:

(1) is robust with respect to the assumptions of the model,

(2) could be used to estimate the variance of the parameters across geographical regions,

(3) enables us compare the risk of RTFs across the geographical regions.

As a general objective, therefore, this study aims at developing statistical methodology, based on Smeed's model, for assessing the risk of RTFs across sub-populations of a given geographical zone. The first specific objective is to develop a modified Smeed's model. Secondly, based on the modified Smeed model, the study seeks to develop and use

- the Bayesian analysis approach to derive an estimator, based on a prior distribution that is robust with respect to the normality assumption,
- the multilevel analysis approach to compare the risk of RTFs across geographical regions.

Finally, the study seeks to use data from Ghana to validate the developed method and to assess the robustness of the model.

## 2. Method

### *2.1 A Modified Smeed's Model*

Smeed's model in Equation (2) measures per vehicle fatality rate. Multiplying both sides of (2) by $N/P$, we obtain

$$\tfrac{D}{P}=\alpha\left(\tfrac{N}{P}\right)^{\beta}\left(\tfrac{N}{P}e\right). \quad \text{………………………………………………………………}(4)$$

The modified Smeed's model of this study, which estimates the ***per capita fatality rate*** (also called [1]***public health risk indicator***), is of the form

$$\tfrac{D}{P}=\alpha\left(\tfrac{N}{P}\right)^{\beta}u, \quad \text{…………………………………………………………}(5)$$

where $u=\left(\tfrac{N}{P}e\right)<e$ provided $N<P.$

Table A1, in the Appendix, is an extract from the list of countries with ranks based on the number of road motor vehicles per 1,000 inhabitants. For every country in the world, except San Marino, the number of registered vehicles in use, *N*, is less than the population size, *P*. Since $N<P$ for most situations, it follows that the multiplicative error term *u* in the modified Smeed's model of this study is less than that of Smeed's original model, making the modified Smeed's model preferred.

The modified Smeed's model is ***intrinsically linear***. Thus, Equation (5) can be transformed to a linear model by a logarithmic transformation of the form

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\ldots+\beta_kx_{ik}+\varepsilon_i,\quad i=1,\ldots,n. \quad \text{……………………………………………}(6)$$

For example, Equation (5) can be written in the form

$$\left.\begin{array}{l}\ln D=\ln\alpha+\beta\ln N+(1-\beta)\ln P+\ln u.\\ \text{or}\\ y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\varepsilon_i,\quad i=1,\ 2,\ \ldots,\ n,\end{array}\right\} \quad \text{…………………………………………}(7)$$

where $y_i=\ln D,\ x_{i1}=\ln N,\ x_{i2}=\ln P,\ \beta_0=\ln\alpha,\ \beta_1=\beta,\ \beta_2=\ln(1-\beta)$ and $\varepsilon_i=\ln u_i.$ Another possible linear transformation of Equation (5) is of the form

$$\left.\begin{array}{l}\ln(D/P)\ =\ \ln\alpha\ +\ \beta\ln(N/P)\ +\ \ln u,\\ \text{or}\\ y_i=\beta_0+\beta_1x_i+\varepsilon_i,\quad i=1,\ 2,\ \ldots,\ n,\end{array}\right\} \quad \text{… ………..…………………………...}(8)$$

where $\beta_0=\ln\alpha,\ \beta_1=\beta,\ x_i=\ln(N/P),\ y_i=\ln(D/P)$ and $\varepsilon_i=\ln u_i,\ i=1,\ 2,\ \ldots,n.$

The linear transformation in Equation (8) is preferred to that of Equation (7) because of the following reason. Since $D/P$ is a risk indicator (known as ***Public Health Risk indicator***) used in epidemiological studies, it follows that any one-to-one relation of this indicator, such as $Y=\ln(D/P)$, can also be used as risk indicator of RTF. This is in sync with the general objective of this study (see Hesse & Ofosu, 2014).

### *2.2 Bayesian Approach to Estimation of Regression Parameters*

In this Section, we develop, using the modified Smeed model, a Bayesian approach to derive an estimator, based on a given prior distribution, that is robust with respect to the normality assumption of the model.

The multiple linear regression model in (6), with *k* predictor variables, can be expressed as

$$y_i=\boldsymbol{\beta}'\boldsymbol{x}_i+\ \boldsymbol{\varepsilon}_i,\quad i=1,\ 2,\ \ldots,n \quad \text{………………………………………………………}(9)$$

where $\boldsymbol{x}_i'=(1,\ x_{1i},\ x_{2i},\ldots,\ x_{ki})$. It is assumed that the unknown parameter vector $\boldsymbol{\beta}'=(\beta_0,\ \beta_1,\ \ldots\ \beta_k)$ is a value of some multivariate random variable with a multivariate prior distribution. The range of possible values that the regression coefficients $\beta_0,\ \beta_1,\ \ldots\ \beta_k$ can take is –∞ to +∞. Thus, the largest possible domain of the prior distribution is the set of all real numbers. This limits us to distribution which can take both negative and positive values. Therefore, the most suitable prior distributions are the bivariate Normal, Laplace and Cauchy distributions.

Two Bayesian methods were used in estimating the parameters in Equation (9). These are the 'conjugate prior' method

[1] National Road Safety Commission of Ghana (2011). Building and Road Research Institute (BRRI), *Road Traffic Crashes in Ghana*, Statistics

and the maximum a posteriori method which are discussed in the following sequel.

*Conjugate Prior*

In this section, we assume that the random variable $Y$, with components $y_i$, in Equation (9), has the normal distribution with mean $\boldsymbol{\beta}'\boldsymbol{x}$ and variance $\sigma^2$. Thus, the likelihood function will also follow a normal distribution. Since the normal distribution is conjugate to itself (or *self-conjugate*) with respect to a normal likelihood function, choosing a bivariate normal prior over $\boldsymbol{\beta}$ will ensure that the posterior distribution is also normal. The conditional p.d.f. of $Y$ is then given by

$$f_Y\left(y_i\,|\boldsymbol{\beta}\right)=\tfrac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\tfrac{1}{2\sigma^2}\left(y_i-\boldsymbol{\beta}'\boldsymbol{x}\right)^2\right\},\quad |y_i|\geq 0. \quad \ldots\ldots(10)$$

The likelihood function is given by (see Mettle et al., 2016)

$$f_{\boldsymbol{Y}}\left(\boldsymbol{y}|\boldsymbol{\beta}\right)=\left(\tfrac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\exp\left\{-\tfrac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\boldsymbol{\beta}'\boldsymbol{x}_i\right)^2\right\},\ \boldsymbol{y}=(y_1,\,y_2,\,\ldots,\,y_n). \ldots\ldots \quad \ldots\ldots(11)$$

It is assumed that $\boldsymbol{\beta}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}=\left(\mu_0,\,\mu_1,\ldots,\,\mu_k\right)$ and covariance matrix $\boldsymbol{\Sigma}$. Thus, the p.d.f. of $\boldsymbol{\beta}$ is

$$p\left(\boldsymbol{\beta}\right)=\tfrac{1}{2\pi}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\mu}\right)\right\} \quad \ldots\ldots(12)$$

where $\boldsymbol{\Sigma}^{-1}=\begin{pmatrix} a_{00} & a_{01} & a_{02} & \cdots & a_{0k} \\ a_{10} & a_{11} & a_{12} & \cdots & a_{1k} \\ a_{20} & a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{k0} & a_{k1} & a_{k2} & \cdots & a_{kk} \end{pmatrix}$. The posterior distribution can therefore be expressed as

$$p\left(\boldsymbol{\beta}|\,\boldsymbol{y}\right)=k\,f\left(\boldsymbol{y}|\boldsymbol{\beta}\right)p\left(\boldsymbol{\beta}\right)=k\exp\left\{-\tfrac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\boldsymbol{\beta}'\boldsymbol{x}_i\right)^2-\tfrac{1}{2}\left[\left(\boldsymbol{\beta}-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\mu}\right)\right]\right\}. \quad \ldots\ldots(13)$$

The function under the exponent in Equation (13) can be written as

$$\begin{aligned} Q(\boldsymbol{\beta}) \;=\; & \left(\tfrac{n}{\sigma^2}+a_{00}\right)\beta_0^2+\sum_{j=1}^{k}\left(\tfrac{1}{\sigma^2}\sum_{i=1}^{n}x_{ji}^2+a_{jj}\right)\beta_j^2-2\left(\tfrac{1}{\sigma^2}\sum_{i=1}^{n}y_i+\sum_{l=1}^{k}a_{1l}\mu_l\right)\beta_0 \\ & -2\sum_{j=1}^{k}\left(\tfrac{1}{\sigma^2}\sum_{i=1}^{n}x_{ji}y_i+\sum_{l=1}^{k}a_{jl}\mu_l\right)\beta_j+2\sum_{j=0}^{k-1}\sum_{s=j+1}^{k}\left(\tfrac{1}{\sigma^2}\sum_{i=1}^{n}x_{ji}x_{si}+a_{js}\right)\beta_j\beta_s+v \quad \ldots\ldots(14) \end{aligned}$$

where $v$ is the constant term, independent of $\beta_j$. Therefore the posterior p.d.f. of $\boldsymbol{\beta}$ can be written as

$$p\left(\boldsymbol{\beta}|\,\boldsymbol{y}\right) \;=\; ke^{-\frac{1}{2}Q(\boldsymbol{\beta})}. \quad \ldots\ldots(15)$$

Hence Equation (13) follows the multivariate normal distribution with mean vector given by

$$\boldsymbol{\mu}_{\boldsymbol{\beta}}=-\tfrac{1}{2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{C}, \quad \ldots\ldots(16)$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is a $(k+1)\times(k+1)$ matrix with inverse $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}=\left(m_{ij}\right)$ whose elements are given as

$$\left.\begin{aligned} & m_{00}=\tfrac{n}{\sigma^2}+a_{00}, \\ & m_{j0}=\tfrac{1}{\sigma^2}\sum_{i=1}^{n}x_{ji}+a_{j0}, \quad j=1,2,\ldots k, \\ & m_{0j}=\tfrac{1}{\sigma^2}\sum_{i=1}^{n}x_{ji}+a_{0j}, \quad j=1,2,\ldots k, \\ & m_{ij}=\tfrac{1}{\sigma^2}\sum_{l=1}^{n}x_{il}x_{j_l}+a_{ij}, \quad i\neq j, \\ & m_{ii}=\tfrac{1}{\sigma^2}\sum_{i=1}^{n}x_{ij}^2+a_{ii}, \quad j=1,2,\ldots k. \end{aligned}\right\} \quad \ldots\ldots(17)$$

and $\boldsymbol{C}$ is a column vector of order ($k$ + 1) with elements given as

$$\left.\begin{aligned} C_0 &= -2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n} y_i + \sum_{j=1}^{k} a_{0j}\mu_j\right) \\ C_i &= -2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n} y_i x_{li} + \sum_{j=1}^{k} a_{ij}\mu_j\right), \qquad l = 1, 2, ..., k. \end{aligned}\right\} \text{.......................................(18)}$$

Let $\hat{\boldsymbol{\beta}}_l = \left(\hat{\beta}_{0l},\ \hat{\beta}_{1l},\ ...,\ \hat{\beta}_{kl}\right);\ l = 1,\ 2,\ 3,\ ...,\ n$ be the $l^{th}$ jackknife estimate of the regression. Then the estimate of the mean vector $\boldsymbol{\mu}$ of the random vector $\boldsymbol{\beta} = \left(\beta_0,\ \beta_1,\ ...,\ \beta_k\right)$ is given as $\hat{\boldsymbol{\mu}} = \left(\hat{\mu}_0,\ \hat{\mu}_1, ..., \hat{\mu}_k\right)'$, where

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n}\beta_{ji}, \quad j = 0,\ 1,\ ...,\ k. \text{ ..........................................................................(19)}$$

and an estimate of the covariance matrix of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{j=1}^{n}\left(\hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\mu}}_j\right)\left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\mu}}_j\right)' = \left(\hat{a}_{ij}\right). \text{...............................................................(20)}$$

The estimate of the standard error of the $i^{\text{th}}$ coefficient, based on the Bayesian estimate is the square root of the $i^{\text{th}}$ diagonal elements of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$.

*Maximum a Posteriori Method*

The goal here is to find the parameter estimates that maximizes the posterior probability of the parameters given the data. This corresponds to

$$\boldsymbol{\beta}_{MAP} = \arg\max_{\boldsymbol{\beta}}\ p\left(\boldsymbol{\beta}|\boldsymbol{y}\right) \text{ ..................................................................(21)}$$

We resort to sampling techniques, such as Markov chain Monte Carlo (MCMC), to get samples from the posterior distribution. The following algorithm is the description for the multivariate Metropolis Hastings procedure (Steyvers, 2011):

```
1.  Set $t = 1$
2.  Generate an initial value for $\beta_j \sim U(u_{1j},\ \mu_{2j}),\quad j = 0,\ 1,\ ...,\ k.$
3.  Repeat
     $t = t + 1$
     Do a MH step on $\beta_j$,
          Generate a proposal $\beta_j^* \sim N(\beta_j,\ \sigma_j^2)$;
          Evaluate the acceptance probability $a = \min\left[1,\ \frac{p^*(\boldsymbol{\beta}|\boldsymbol{y})}{p(\boldsymbol{\beta}|\boldsymbol{y})}\right]$;
          Generate a $u$ from a Uniform(0, 1) distribution
          If $u \le a$, accept the proposal and set $\beta_j = \beta_j^*,\quad j = 0,\ 1,\ ...,\ k.$
4.  Until $t = T$.
```

*2.3 Multilevel Random Coefficient (MRC) Model*

In this Section, we develop a Multilevel Analysis approach to estimate the regional distribution of parameters based on the modified Smeed's model and use them to compare the risk of RTFs across geographical regions.

Assuming the population is stratified into $J$ geographical regions with $n_j$ observations in each class, Equation (6) becomes

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + ... + \beta_{kj}x_{kij} + \varepsilon_{ij}, \\ &= \beta_{0j} + \sum_{l=1}^{k}\beta_{lj}x_{lij} + \varepsilon_{ij}, \qquad \begin{matrix} i = 1,\ 2,\ ..., n_j \\ j = 1,\ 2,\ ...,\ J \end{matrix} \end{aligned} \text{ ...............................................(22)}$$

Across all geographical regions, $\beta_j = (\beta_{0j},\ \beta_{1j},\ ...,\ \beta_{kj})$ are assumed to have multivariate normal distribution (Hox, 2010). Thus, each $\beta_{lj}$ $(l = 0, 1, 2, ..., k)$ can be modeled as

$$\beta_{0j} = \gamma_{00} + \gamma_{01} z_j + u_{0j} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(23)$$

$$\beta_{lj} = \gamma_{l0} + \gamma_{l1} z_j + u_{lj}, \qquad l = 1,\ ...,k \text{ and } j = 1,\ 2,\ ...,\ J\ . \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(24)$$

From Equations (22), (23) and (24), we have

$$y_{ij} = \gamma_{00} + \gamma_{01} z_j + u_{0j} + \sum_{l=1}^{k} \left(\gamma_{l0} + \gamma_{l1} z_j + u_{lj}\right) x_{lij} + \varepsilon_{ij}, \qquad \begin{matrix} i = 1,\ 2,\ ...,n_j \\ j = 1,\ 2,\ ...,\ J \end{matrix} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(25)$$

$u_{lj} \sim N(0,\ \tau_l),\ l = 0,1,...,k$ and $\varepsilon_{ij} \sim N(0,\ \sigma^2)$. $Y$ has the normal distribution with mean

$$\mu = \gamma_{00} + \gamma_{01} z_j + \sum_{l=1}^{k} \left(\gamma_{l0} + \gamma_{l1} z_j\right) x_{lij} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(26)$$

and variance

$$v = \tau_0 + \sum_{l=1}^{k} x_{lij}^2 \tau_l + 2 \sum_{l \neq r} x_{lij} x_{rij} \tau_{lr} + \sigma^2 . \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(27)$$

The parameters to be estimated are $\gamma_{l0},\ \gamma_{l1},\ \tau_l,\ \tau_{lr} (l \neq r)$ and $\sigma^2$, $l = 0,\ 1,\ \ldots,\ k$, where $\tau_l = \text{var}(u_{lj})$, $\tau_{lr} = \text{cov}(u_{lj},\ u_{rj})$, and $\text{var}(\varepsilon_{ij}) = \sigma^2$.

If $\tau_0$ differs significantly from 0, then the parameters of the modified Smeed's model can be used to compare the risk of RTFs across the $J$ geographical regions.

Equating the partial derivatives of the likelihood function to zero, we obtain the maximum likelihood estimators of the parameters $\gamma_{l0},\ \gamma_{l1},\ \tau_l,\ \tau_{lr} (l \neq r)$ and $\sigma^2$ as $\hat{\gamma}_{l0},\ \hat{\gamma}_{l1},\ \hat{\tau}_l,\ \hat{\tau}_{lr} (l \neq r)$ and $\hat{\sigma}^2$ respectively.

## 3. Validation of Method Using Data from Ghana

In this section the study seeks to use data from Ghana to validate the

(1) Bayesian method and to assess the robustness of the model

(2) multilevel method and to compare the risk of RTFs across the 10 geographical regions.

### *3.1 Validation of Bayesian Method*

*(i) Conjugate Prior Method*

Table A2, in the Appendix, gives the estimated population size and the number of motor vehicles and road traffic fatalities in Ghana (1991 – 2012). It can be seen that, the distribution of $\ln(D/P)$, with a Shapiro-Wilks normality $p$-value of 0.201, is closer to the normal distribution compared to that of $\ln(D)$ with a corresponding $p$-value of 0.086. This confirms that the logarithmic transformation in Equation (8) is preferred.

The 19 jackknife sample estimates of $\beta_0$ and $\beta_1$, based on the national data, derived from the values of $y_i$ and $x_i$ in Table A2 are given in Table A3. Based on Equations (19) and (20), jackknife estimate of the mean vector and covariance of the random vector $\boldsymbol{\beta}$ is computed as follows

$$\hat{\boldsymbol{\mu}} = (-8.3105,\ 0.3192) \qquad \text{and} \qquad \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.001860 & 0.000504 \\ 0.000504 & 0.000139 \end{pmatrix}.$$

Based on Equations (17) and (18),

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = \begin{pmatrix} 0.0017421 & 0.0004712 \\ 0.0004712 & 0.0001297 \end{pmatrix} \text{ and } \hat{\boldsymbol{C}} = \begin{pmatrix} 646278.208 \\ -2353969.324 \end{pmatrix}.$$

Thus, the posterior Bayes estimate of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\beta}} = -\tfrac{1}{2} \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \boldsymbol{C} = \begin{pmatrix} -8.31048 \\ 0.319162 \end{pmatrix}. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(28)$$

Table 1 shows the coefficients estimates and the corresponding standard errors for the least square and the conjugate prior methods.

Table 1. Comparison of Coefficients of Least Square and Conjugate Prior Methods

| | **Methods** | | | |
|---|---|---|---|---|
| | **Least Squares** | | **Conjugate Prior** | |
| | Coefficient | Standard Error | Coefficient | Standard Error |
| **$\beta_0$ = intercept** | –8.31179 | 0.17386 | –8.31048 | 0.04174 |
| **$\beta_1$ = coefficient of *x*** | 0.31879 | 0.04555 | 0.31916 | 0.01139 |
| **Coefficient of determination** | 0.7423 | | 0.7423 | |

It can be seen from Table 2, that the estimated coefficients $\beta_0$ and $\beta_1$, are almost the same for the least squares and the conjugate prior methods. Both methods also reported the same coefficient of determination $R^2$. The conjugate prior estimates recorded comparatively very small standard errors; making the conjugate prior method preferred.

*(ii) Maximum a posteriori method*

Our objective here is to determine the parameter estimates that maximize the posterior distribution given the data with respect to the bivariate Normal, Laplace and Cauchy prior distributions.

*Bivariate Normal prior distribution*

The prior distribution in Equation (12) can be written in terms of ρ as

$$p\left(\beta_0,\ \beta_1 \middle| \boldsymbol{y}\right) = k\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{19}\left(y_i-\beta_0-\beta_1 x_i\right)^2-\frac{1}{2}q\right\}, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(29)$$

where $q=\frac{1}{1-\rho^2}\left\{\left(\frac{\beta_0-\mu_0}{\sigma_0}\right)^2-2\rho\left(\frac{\beta_0-\mu_0}{\sigma_0}\right)\left(\frac{\beta_1-\mu_1}{\sigma_1}\right)+\left(\frac{\beta_1-\mu_1}{\sigma_1}\right)^2\right\},\ -\infty<\beta_0<\infty,\ -\infty<\beta_1<\infty,$

$\sigma_0^2=\mathrm{var}(\beta_0),\ \sigma_1^2=\mathrm{var}(\beta_1).$

The Metropolis Hastings algorithm, above, is used to estimate the values of $\beta_0$ and $\beta_1$. The MATLAB code for the implementation of component-wise Metropolis sampler for the posterior distribution is as given in Listings 1 and 2 in the appendix.

Table 2 shows estimated values of $\beta_0$ and $\beta_1$ based on least squares, conjugate prior and maximum a posteriori methods. The results show that the estimated coefficients of $\beta_0$ and $\beta_1$ are almost the same for the least squares, conjugate prior and maximum a posteriori methods of estimates.

Table 2. Comparison of Coefficients of Least Squares, Conjugate Prior and Maximum a Posteriori Methods

| | **Methods** | | |
|---|---|---|---|
| | **Least Square** | **Conjugate prior** | **Maximum a posteriori** |
| $\beta_0$ | –8.31179 | –8.31048 | –8.29094 |
| (**Standard error**) | (**0.17386**) | (**0.04174**) | (**0.03978**) |
| $\beta_1$ | 0.31879 | 0.31916 | 0.32460 |
| (**Standard error**) | (**0.04555**) | (**0.01139**) | (**0.01098**) |

**Laplace Prior Distribution**

It is assumed that $\boldsymbol{\beta}=\left(\beta_0,\beta_1\right)$ has a bivariate Laplace distribution with mean vector $\boldsymbol{\mu}=\left(\mu_0,\ \mu_1\right)$. The joint p.d.f. is given by

$$f(\beta_0,\beta_1)\ =\ \frac{1}{4b_0b_1}e^{-\left[\frac{1}{b_0}\left|\beta_0-\mu_0\right|+\frac{1}{b_1}\left|\beta_1-\mu_1\right|\right]}, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(30)$$

$-\infty<\alpha<\infty,\ -\infty<\beta<\infty,\ b_0>0,\ b_1>0.$ Thus, the posterior distribution can be expressed as

$$p\left(\beta_0,\ \beta_1\middle|\boldsymbol{y}\right)=k\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)^2-\frac{1}{\hat{b}_0}\left|\alpha-\hat{\mu}_0\right|-\frac{1}{\hat{b}_1}\left|\beta-\hat{\mu}_1\right|\right\}. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(31)$$

Using the above algorithm, the maximum a posteriori estimates of $\beta_0$ and $\beta_1$ to be –8.320085 and 0.317051, respectively, with standard errors of 0.039047 and 0.010450.

*Cauchy Prior Distribution*

The bivariate random variable $\boldsymbol{\beta} = (\beta_0, \beta_1)$ has the Cauchy distribution if the p.d.f. can be expressed in the form given in the following form

$$f(\beta_0, \beta_1) = \frac{1}{2\pi}\left[\frac{1}{(\beta_0 - a)^2 + (\beta_1 - b)^2 + 1}\right]. \quad \text{.......................................(32)}$$

Thus, the posterior distribution can be expressed as

$$p\left(\beta_0, \beta_1 \middle| \boldsymbol{y}\right) = k\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{19}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right\}\left\{\frac{1}{(\beta_0 - \hat{a})^2 + (\beta_1 - \hat{b})^2 + 1}\right\}. \quad \text{.........................(33)}$$

The component-wise Metropolis-Hastings sampler for the posterior distribution based on the MATLAB codes, gave maximum a posteriori estimates of $\beta_0$ and $\beta_1$ to be –8.312857 and 0.317400, respectively.

The resulting posterior Bayesian estimates for the Normal, Laplace and Cauchy prior distributions are summarized in the Table 3. Given a sample size 19, the posterior Bayes estimate is reasonably consistent for the Normal, Laplace and Cauchy prior distributions.

Table 3. Posterior Bayesian estimates for different priors with a sample size of 19

| | **Prior distribution** | | | | |
|---|---|---|---|---|---|
| | **Normal** | | **Laplace** | | **Cauchy** |
| | Estimate | Standard Error | Estimate | Standard Error | |
| $\beta_0$ | –8.31048 | 0.04174 | –8.32009 | 0.039047 | –8.31286 |
| $\beta_1$ | 0.31916 | 0.01139 | 0.31705 | 0.010450 | 0.31740 |

Table 4 shows the posterior Bayesian estimates of $\beta_0$ and $\beta_1$ at four different sample sizes (5, 10, 15 and 19) using the Normal, Laplace and Cauchy prior distributions. It can be seen that, at sample sizes of 5 and 10, the posterior Bayesian estimates of $\beta_0$ and $\beta_1$ are not consistent across the three prior distributions used. Thus, the estimated values of $\beta_0$ and $\beta_1$ are said to be sensitive with respect to the prior distribution. At a sample size of 15 or more, the model becomes insensitive to the prior distribution. The relative influence of the prior distribution decreases while that of the data increases with a sample size of 15 or more. It can also be seen that the posterior Bayesian estimate is reasonably consistent for the Laplace prior distribution across all four sample sizes used. Even at a sample size of 5 where the normality assumption was violated, the estimates based on the Laplace prior distribution was robust. Thus, the Laplace prior distribution is preferred when the sample size is small.

Table 4. Bayesian estimates with respect to sample size and prior distribution

| | **Prior distribution** | | | | | |
|---|---|---|---|---|---|---|
| **Sample size** | **Normal** | | **Laplace** | | **Cauchy** | |
| ***n*** | $\beta_0$ | $\beta_1$ | $\beta_0$ | $\beta_1$ | $\beta_0$ | $\beta_1$ |
| 5 | –5.99608 | 0. 99041 | –8.30608 | 0. 31923 | –5.13317 | 1.01961 |
| (**Standard error**) | (**0.67355**) | (**0.02767**) | (**0.61978**) | (**0.02195**) | | |
| 10 | –8.29381 | 0.32272 | –8.29637 | 0.32863 | –7.72230 | 0.46478 |
| (**Standard error**) | (**0.44057**) | (**0.01629**) | (**0.43884**) | (**0.01596**) | | |
| 15 | –8.31195 | 0.31647 | –8.29288 | 0.32266 | –8.31034 | 0.31694 |
| (**Standard error**) | (**0.36057**) | (**0.01328**) | (**0.35747**) | (**0.01298**) | | |
| 19 | –8.31048 | 0.31916 | –8.32009 | 0.31705 | –8.31286 | 0.31740 |
| (**Standard error**) | (**0.31916**) | (**0.01139**) | (**0.31705**) | (**0.01045**) | | |

### *3.2 Validation of Multilevel Method*

Table A4, in the Appendix, shows the value of $x_{ij} = \ln\left(N_{ij}/P_{ij}\right)$ and the corresponding values of $y_{ij} = \ln\left(D_{ij}/P_{ij}\right)$ for the ten regions of Ghana. Instead of estimating a separate regression equation for each of the 10 regions in Ghana, we wish to determine a single model for estimating regional distribution of RTFs. The collection of the regression parameters {$\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, …, $\boldsymbol{\beta}_{10}$} is assumed to be a random sample of size 10 taken from a population whose distribution depends on the parameters $\gamma_1$, $\gamma_2$, $\delta_0$, $\tau_0$, $\tau_1$, $\tau_{01}$ and $\sigma^2$, where $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j})$, $j = 1, 2, \ldots, 10$.

Equations (21), (22) and (23) can be written as

$$\left.\begin{array}{l} y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij}, \\ \beta_{0j} = \gamma_{00} + \gamma_{01}\overline{x}_j + u_{0j}, \\ \beta_{1j} = \gamma_{10} + u_{1j}. \end{array}\right\} \qquad \begin{array}{l} i = 1,\ 2,\ \ldots,19 \\ j = 1,\ 2,\ \ldots,\ 10 \end{array} \qquad \ldots\ldots\ldots\ldots\ldots(34)$$

Combining the three equations, we obtain

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}\overline{x}_j + u_{1j}x_{ij} + u_{0j} + \varepsilon_{ij}, \quad j = 1,\ 2,\ \ldots,\ 10. \quad \ldots\ldots\ldots\ldots(35)$$

$u_{0j} \sim N(0, \tau_0)$, $u_{1j} \sim N(0, \tau_1)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$. $Y$ has the normal distribution with mean $\gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}\overline{x}_j$ and variance $v = \tau_0 + 2\tau_{01}x_{ij} + \tau_1 x_{ij}^2 + \sigma^2$. Thus, the pdf of $Y$ given $X = x_{ij}$ is

$$f_Y\left(y_{ij}\left|X = x_{ij}\right.\right) = \frac{1}{\sqrt{2v\pi}}\exp\left[-\frac{1}{2v}\left(y_{ij} - \gamma_{00} - \gamma_{10}x_{ij} - \gamma_{01}\overline{x}_j\right)^2\right] \ldots\ldots\ldots\ldots(36)$$

Three models are considered in the next section.

*(i) The Unconditional Means Model, $M_0$*

An unconditional means model does not contain any predictors, but includes a random intercept variance term for groups. In this section, we examine if there will be significant intercept variation $(\tau_0)$. If $\tau_0$ does not differ significantly from 0, there may be little reason to use random coefficient modeling since simpler Ordinary Least Squares (OLS) modeling will suffice. Equation (34) therefore becomes

$$\left.\begin{array}{l} Y_{ij} = \beta_{0j} + \varepsilon_{ij}, \\ \beta_{0j} = \gamma_{00} + u_{0j} \end{array}\right\} \ldots\ldots\ldots\ldots\ldots(37)$$

Therefore

$$Y_{ij} = \gamma_{00} + u_{0j} + \varepsilon_{ij}. \ldots\ldots\ldots\ldots\ldots(38)$$

Application of the nlme package in *R*, using data in Table A3, shows that there is significant intercept variation in terms of *y* scores across the 10 regions.

*(ii) Random Intercept Model, $M_1$*

In this model, it is assumed that the intercept $\beta_{0j}$ vary across the 10 geographical regions whilst the slope $\beta_{1j}$ remain constant. Equation (34), therefore, becomes

$$\left.\begin{array}{l} y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij}, \\ \beta_{0j} = \gamma_{00} + \gamma_{01}\overline{x}_j + u_{0j} \\ \beta_{1j} = \gamma_{10}, \end{array}\right\} \ldots\ldots\ldots\ldots\ldots(39)$$

Combine the three rows into a single equation,

$$Y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}\overline{x}_j + u_{0j} + \varepsilon_{ij}, \quad j = 1,\ 2,\ \ldots,10. \ \ldots \quad \ldots\ldots\ldots\ldots(40)$$

The maximum likelihood estimates of the parameters, using data from Table A3 and nlme package in *R*, are given in Table 5.

*(iii) Random slope model $M_2$*

In section, we continue our analysis by trying to explain the third source of variation, namely, variation in the slope, $\tau_1$. The model that we test is:

$$\left.\begin{array}{l} y_{ij} = \alpha_j + \beta_j x_{ij} + \varepsilon_{ij}, \\ \alpha_j = \gamma_0 + \gamma_1\overline{x}_j + e_{\alpha j} \\ \beta_j = \delta_0 + e_{\beta j} \end{array}\right\} \ldots\ldots\ldots\ldots\ldots(41)$$

When we combine the three rows into a single equation in the form

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}\overline{x}_j + u_{1j}x_{ij} + u_{0j} + \varepsilon_{ij}, \quad j = 1,\ 2,\ \ldots,\ 10. \ \ldots\ldots\ldots\ldots(42)$$

Table 5 presents the parameter estimate and standard errors for the models $M_0$, $M_1$ and $M_2$. All the standard errors of the estimated parameters in model $M_2$ are smaller than the corresponding values of model $M_1$. Moreover, the deviance, which measures the model misfit, is much lower in $M_2$ as compare to that of $M_1$ (Hesse, et al., 2014b) Thus, estimate parameters based on model $M_2$ is preferred.

Table 5. Comparison of models $M_0$, $M_1$ and $M_2$

| Model | $M_0$: intercept only | | $M_1$: with predictor | | $M_2$: with predictor | |
|---|---|---|---|---|---|---|
| **Fixed effect** | Coefficient | Standard Error | Coefficient | Standard Error | Coefficient | Standard Error |
| $\gamma_{00}$ = Intercept | -9.6888 | 0.1401 | -10.0756 | 0.7426 | -9.2341 | 0.2065 |
| $\gamma_{10}$ = coffiecient of $x_{ij}$ | | | 0.4591 | 0.0374 | 0.4459 | 0.0707 |
| $\gamma_{01}$ = coefficient of $\overline{x}_j$ | | | -0.5448 | 0.1658 | -0.3384 | 0.0516 |
| **Random part** | Parameter | Standard Error | Parameter | Standard Error | Parameter | Standard Error |
| $\tau_0 = \text{var}(u_{0j})$ | 0.1891 | 0.2085 | 0.2094 | 0.1447 | 0.1545 | 0.1243 |
| $\tau_1 = \text{var}(u_{1j})$ | | | | | 0.0382 | 0.0618 |
| $\tau_{01} = \text{cov}(u_{0j},\ u_{1j})$ | | | | | 0.0766 | |
| $\sigma^2 = \text{var}(\varepsilon_{ij})$ | 0.1389 | 0.0855 | 0.0759 | 0.0632 | 0.0630 | 0.0576 |
| **Deviance** | 198.201 | | 94.554 | | 64.749 | |

The estimate of regional-level residuals $\hat{u}_{0j}$ and $\hat{u}_{1j}$ and the corresponding values of $\alpha$ and $\beta$ for each region are given in Table 6.

Table 6. Estimate of regional-level residuals and the values of $\alpha$ and $\beta$

| Regions | $\hat{u}_{0j}$ | $\hat{u}_{1j}$ | $\hat{\beta}_0$ | $\hat{\beta}_j$ | $\hat{\alpha}_j = e^{\hat{\beta}_0}$ |
|---|---|---|---|---|---|
| Greater Accra | -0.273 | -0.138 | -8.709877 | 0.3083572 | 0.0001649 |
| Ashanti | -0.168 | -0.084 | -8.073562 | 0.3614688 | 0.0003117 |
| Western | -0.085 | -0.041 | -7.677551 | 0.4053849 | 0.0004631 |
| Eastern | -0.470 | -0.235 | -7.930339 | 0.2109577 | 0.0003597 |
| Central | -0.342 | -0.170 | -7.743066 | 0.2758323 | 0.0004337 |
| Volta | -0.037 | -0.020 | -7.397244 | 0.4259363 | 0.0006129 |
| Northern | 0.427 | 0.214 | -7.251897 | 0.6594775 | 0.0007088 |
| Upper East | 0.395 | 0.198 | -7.400873 | 0.6439825 | 0.0006107 |
| Upper West | 0.703 | 0.353 | -7.206664 | 0.7993004 | 0.0007416 |
| Brong Ahafo | -0.152 | -0.077 | -7.694218 | 0.3686119 | 0.0004555 |

According to National Road Safety Commission (NRSC)[2] of Ghana 2011 report, two key national road traffic fatality indices required for characterization and comparison of the extent and risk of traffic fatality across the ten geographical regions of Ghana are *RTF*s per 100 accidents and *RTF per 100* casualties.

The last two columns of Table 7 give the means of RTFs per 100 accidents and RTFs per 100 casualties for each region from 1991 – 2009. This implies that the risk of dying as a result of road traffic fatality in Greater Accra is relatively low, recording an average rate of 5.7 road traffic fatalities per 100 accidents. Thus, out of every 100 road traffic accidents in the

[2] National Road Safety Commission of Ghana (2011). Building and Road Research Institute (BRRI), *Road Traffic Crashes in Ghana*, Statistics

Greater Accra, about 6 of the victims are likely to die (Hesse and Ofosu, 2015).

Table 7. Parameter estimates and Fatality indices

| Regions | $\hat{\alpha}\times 10^5$ | $\hat{\beta}\times 10^2$ | *RTF per 100* Accident | *RTF per 100* Casualties |
|---|---|---|---|---|
| Greater Accra | 16.5 | 30.836 | 5.7 | 7.7 |
| Ashanti | 31.2 | 36.147 | 17.8 | 12.2 |
| Western | 46.3 | 40.538 | 16.9 | 10.7 |
| Eastern | 36.0 | 21.096 | 19.9 | 9.7 |
| Central | 43.4 | 27.583 | 21.8 | 11.4 |
| Volta | 61.3 | 42.594 | 23.6 | 11.2 |
| Northern | 70.9 | 65.948 | 40.9 | 18.1 |
| Upper East | 61.1 | 64.398 | 27.3 | 17.0 |
| Upper West | 74.2 | 79.930 | 28.3 | 14.6 |
| Brong-Ahafo | 45.6 | 36.861 | 28.6 | 14.5 |

We wish to determine if strong positive correlation exist between the parameter estimates of the modified Smeed's model and the fatality indices based on NRSC definition of risk. The p-values in Table 8 show that there is strong positive correlation between the parameter estimates of the modified Smeed's model and the fatality indices. Thus, the parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ of the modified Smeed's model can be used as risk indicators of RTFs in Ghana.

Table 8. Correlations coefficients

| | $\hat{\alpha}$ | $\hat{\beta}$ | RTF per 100 Accident | RTF per 100 Casualties |
|---|---|---|---|---|
| $\hat{\alpha}$ | 1 | | | |
| $\hat{\beta}$ | 0.8312 **(0.003)** | 1 | | |
| RTF per 100 Accident | 0.8424 **(0.002)** | 0.6341 **(0.049)** | 1 | |
| RTF per 100 Casualties | 0.7708 **(0.009)** | 0.7610 **(0.010)** | 0.9011 **(0.000)** | 1 |

## 4. Conclusion

A modified Smeed's model,

$$\frac{D}{P} = \alpha\left(N/P\right)^{\beta} u,$$

has been developed. The multiplicative error term *u* in the modified Smeed's model of this study was found to be less than that of Smeed's, making the modified Smeed's model preferred. Using data from Ghana, it was confirmed that the modified Smeed's model for this studies, is relatively more accurate in estimating RTFs in Ghana than the Smeed equation.

Based on the modified Smeed's model of this study, the developed Bayesian method with respect to the Laplace prior distribution was found to be robust to violation of the normality assumption of the model. Using data from Ghana, the sensitivity of the Bayesian estimates at different sample sizes with respect to the Normal, Laplace and Cauchy prior distributions was assessed. At a sample size of 15 or more, the model becomes insensitive to the prior distribution. The posterior Bayesian estimate is consistent for the Laplace prior distribution across all four sample sizes. At a sample size of 5, the estimates based on Laplace prior distribution were robust with respect to violation of the normality assumption of the model.

The parameter estimates of modified Smeed's model can be used as risk indicator of RTFs across geographical regions provided there is significant intercept variation $\tau_0$ of the regression equation across geographical regions. Using data from Ghana, it was shown that the parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ across the 10 geographical regions can be used as *risk indicators of RTFs in Ghana*. Thus, the three Northern regions and the Brong-Ahafo region have the highest risk of RTFs.

## References

Bener, A., & Ofosu, J. B. (1991). Road traffic fatalities in Saudi Arabia. *Journal of the International Association of Traffic and Safety Sciences, 15,* 35-8.

Fouracre, P., & Jacobs, G. D. (1977). *Further research on road accident rate in developing countries*. TRRL report LR 270. Transport and Road Research Laboratory, Crowthorne, Berkshire.

Ghee, C. Silcock, D. Astrop, A., & Jacobs, G. (1997). *So cio-economic aspects of road accidents in developing countries.* TRL Report 247. Crowthorne: Transport Research Laboratory.

Hesse, C. A., & Ofosu, J. B. (2014). Epidemiology of road traffic accidents in Ghana. *European Scientific Journal, 10*(9), 370-381.

Hesse, C. A., & Ofosu, J. B. (2015). The Effect of Road Traffic Fatality Rate on Road Users in Ghana. *Research Journal of Mathematics and Statistics, 7*(4), 53-59. http://dx.doi.org/10.19026/rjms.7.2207

Hesse, C. A., Ofosu, J. B., & Oduro, F. T. (2014). A Bayesian Model for Predicting Road Traffic Fatalities in Ghana. *Mathematical Theory and Modeling, 4*(8), 1-9.

Hesse, C. A., Ofosu, J. B., & Lamptey, B. L. (2014). A Regression Model for Predicting Road Traffic Fatalities in Ghana. *Open Science Repository Mathematics*, Online(open-access), e23050497. http://dx.doi.org/10.7392/openaccess.23050497.

Hox, J. J. (2010). Multilevel analysis: Techniques and applications 2nd edition. Routledge Taylor and Francis Group.

Jacobs, G., & Bardsley, M. (1977). *Research on road accidents in developing countries*. Traffic engineering & control.

Mettle, F. O., Asiedu, L. Quaye, E. N. B., & Asare-Kumi, A. A. (2016). Comparison of Least Squares Method and Bayesian with Multivariate Normal Prior in Estimating Multiple Regression Parameters. *British Journal of Mathematics and Computer Science, 15*(1), 1-9. http://dx.doi.org/10.9734/BJMCS/2016/23145

Ponnaluri, R. V. (2012). Modeling road traffic fatalities in India: Smeed's law, time invariance and regional specificity. *International Association of Traffic and Safety Sciences,* 36, 75-82. http://dx.doi.org/10.1016/j.iatssr.2012.05.001

Smeed, R. (1949). Some statistical aspects of road safety research. *J. Roy Stats. Soc. Series-A, 12*(1), 1-23. http://dx.doi.org/10.2307/2984177

Steyvers, M. (2011). *Computational statistics with MATLAB.* University of California, Irvine, psiexp.ss.uci.edu/research/teachingP205C/205C.pd.

Tiska, M. A., Adu-Ampofo, M., Boakye, G., Tuuli, L., & Mock, C. N. (2002). A model of prehospital trauma training for lay persons devised in Africa. *Emergency Medical Journal*.

## Copyrights

ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

# Reviewer Acknowledgements

*International Journal of Statistics and Probability* wishes to acknowledge the following individuals for their assistance with peer review of manuscripts for this issue. Their help and contributions in maintaining the quality of the journal is greatly appreciated.

Many authors, regardless of whether *International Journal of Statistics and Probability* publishes their work, appreciate the helpful feedback provided by the reviewers.

**Reviewers for Volume 5, Number 6**

Afsin Sahin

Sajid Ali

Farida Kachapova

Carla J. Thompson

Anna Grana'

Bibi Abdelouahab

Hui Zhang

Philip Westgate

Douglas Lorenz

Samah M. Abo-El-Hadid

Qingyang Zhang

Chin-Shang Li(Editor-in-Chief)

Gane Samb Lo

Dr. Sohair F. Higazi

Wei Zhang

Vyacheslav Abramov

Abdullah A. SMADI

Jacek Białek

Ivair R. Silva

Milind Phadnis

Marcelo Bourguignon

Zaixing Li

Wojciech Gamrot

Vilda Purutcuoglu

Ali Reza Fotouhi

Nahid Sanjari Farsipour

Wendy Smith

On behalf of,

The Editorial Board of *International Journal of Statistics and Probability*

Canadian Center of Science and Education

The journal is peer-reviewed
The journal is open-access to the full text
The journal is included in:

- BASE
- Google Scholar
- JournalTOCs
- Library and Archives Canada
- LOCKSS
- PKP Open Archives Harvester
- SHERPA/RoMEO
- Standard Periodical Directory
- Ulrich's

**International Journal of Statistics and Probability**
Bimonthly

| | |
|---|---|
| Publisher | Canadian Center of Science and Education |
| Address | 1120 Finch Avenue West, Suite 701-309, Toronto, ON., M3J 3H7, Canada |
| Telephone | 1-416-642-2606 |
| Fax | 1-416-642-2608 |
| E-mail | ijsp@ccsenet.org |
| Website | http://ijsp.ccsenet.org |

ISSN 1927-7032

9 771927 703169 06>