รายงานการวิจัยฉบับสมบูรณ์

โครงการ

เกณฑ์ในการคัดเลือกตัวแบบบนพื้นฐานของความแตกต่างแบบสมมาตรของ Kullback-Leibler สำหรับตัวแบบสมการหลายชั้น

Model Selection Criterion Based on Kullback-Leibler's Symmetric Divergence for Simultaneous Equations Model

โดย

ผู้ช่วยศาสตราจารย์ ดร.วรางคณา กีรติวิบูลย์
มหาวิทยาลัยทักษิณ วิทยาเขตพัทลุง

มิถุนายน 2556

สัญญาเลขที่ MRG5480044

รายงานการวิจัยฉบับสมบูรณ์

โครงการ

เกณฑ์ในการคัดเลือกตัวแบบบนพื้นฐานของความแตกต่างแบบสมมาตรของ Kullback-Leibler สำหรับตัวแบบสมการหลายชั้น

Model Selection Criterion Based on Kullback-Leibler's Symmetric Divergence for Simultaneous Equations Model

โดย

ผู้ช่วยศาสตราจารย์ ดร.วรางคณา กีรติวิบูลย์
มหาวิทยาลัยทักษิณ วิทยาเขตพัทลุง

สนับสนุนโดยสำนักงานคณะกรรมการการอุดมศึกษา
สำนักงานกองทุนสนับสนุนการวิจัย
และมหาวิทยาลัยทักษิณ

(ความเห็นในรายงานนี้เป็นของผู้วิจัย สกอ. และ สกว. ไม่จำเป็นต้องเห็นด้วยเสมอไป)

EXECUTIVE SUMMARY

Background

Most problems in the errors of simultaneous equations model (SEM) are the autocorrelated error (AR) and moving average (MA) error. When these problems occur, the ordinary least squares (OLS) estimators can not be used because they are not efficient. In this research, we will propose a transformation matrix to correct the first-order moving average, MA(1), which generated in the fitted model and to recover the one lost observation in a SEM. The MA(1) is in the form,

$$v_{tj} = \varepsilon_{tj} - \theta_j \varepsilon_{t-1,j},\ t = 1,\ 2,\ \ldots,\ T,\ j = 1,\ 2,\ \ldots,\ M, \tag{1}$$

where T is the number of observations in each equation, M is the number of equations, the error $\varepsilon_{t-1,j}$ is called the first-lag of error ε_{tj}, the moving average parameter θ_j of the model must satisfy the following condition to ensure the invertibility of the error terms (Box et al., 1994),

$$\left|\theta_j\right| < 1. \tag{2}$$

The error ε_{tj} in (1) is an independent identically distributed random variable, obeying

$$\varepsilon_{tj} \sim N\left(0, \sigma_{jj}\right), \tag{3}$$

so that

$$\boldsymbol{\varepsilon}_t' = \begin{bmatrix} \varepsilon_{t1} & \varepsilon_{t2} & \cdots & \varepsilon_{tM} \end{bmatrix} \sim N_M\left(\mathbf{0}, \boldsymbol{\Sigma}\right),$$

where $M \times M$ contemporaneous covariance matrix of the error terms,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1M} & \sigma_{2M} & \cdots & \sigma_{MM} \end{bmatrix},$$

is nonsingular and is of positive symmetric definite matrix. It is noteworthy that the values of v_{1j} in the MA(1) model in (1) depend on the values of ε_{0j}, which is unknown. The recovery of v_{1j} will be find by extend the knowledge of Keerativibool (2010).

After the errors are transformed to be independent, we consider the problem of fitting a parametric model to an observed data set. This problem requires two tasks, determination of the order of the model and estimation of these parameters. In real life, we may not know what the true model is, but we hope to find a model that is a reasonably accurate representation. The crucial part of this fitting problem is to determine the order of the model. Such determination is often facilitated by the use of a model selection criterion where one only has to evaluate two simple terms that trade-off quality of fit to the data and model's complexity. The widespread criterion for choosing the best model in univariate and multivariate regression analysis is the Akaike information criterion (AIC) (Akaike, 1973, 1974; Bedrick and Tsai, 1994). The corrected version of the AIC (AIC$_c$) (Hurvich and Tsai, 1989) is extended for the case of small sample. AIC and AIC$_c$ were designed, respectively, to be asymptotically and exactly unbiased estimator of a variant of Kullback-Leibler's directed divergence between the true model and a fitted candidate model. The development of a new family of selection criteria, Kullback information criterion (KIC) and the corrected version of the KIC (KIC$_c$), are the criteria constructed to target a symmetric divergence. This divergence is an alternate of directed divergence, obtained by sum of the two directed divergences, which arguably more sensitive than either of its individual components (Cavanaugh, 1999, 2004; Seghouane and Bekara, 2004; Hafidi and Mkhadri, 2006). With this motivation, we will propose a model selection criterion, called Kullback information criterion for a system of SEM (SKIC), which serves as an asymptotically unbiased estimator of a variant of Kullback-Leibler's symmetric divergence between the true model and the fitted candidate model. Next, we will examine the performance of the proposed criterion, SKIC, relative to SAIC proposed by Keerativibool (2009).

Objectives of the Research

The objectives of this research are to develop a model selection method, to separate the most fitting SEM when the errors are both MA(1) and contemporaneously correlated for analyzing a specific system, by applying the Kullback information criterion (KIC) (Cavanaugh, 1999). The topics covered in this research comprise:

1) To derive a transformation matrix in order to correct the MA(1) problem of errors in a SEM.

2) To derive the Kullback information criterion for a system of SEM when errors are both MA(1) and contemporaneously correlated, called SKIC.

3) To examine the performance of the proposed criterion, SKIC, relative to SAIC proposed by Keerativibool (2009).

Methodologies

The methodologies of this research are as follows:

1) Derive a transformation matrix in order to correct the MA(1) problem of errors in a SEM.

2) Simulate the SEM when errors are both MA(1) and contemporaneously correlated by the SAS programming.

3) Carry out a proposed transformation matrix to correct the MA(1) problem of errors in a SEM.

4) Examine the errors of SEM after transformation.

5) Derive the Kullback information criterion for a system of SEM (SKIC) when errors are both MA(1) and contemporaneously correlated.

6) Examine the performance of the proposed criterion, SKIC, relative to SAIC proposed by Keerativibool (2009).

Plans of the Research

The plans of the research are as follows:

Activity	Jun 15, 2011 – Dec 14, 2011	Dec 15, 2011 – Jun 14, 2012	Jun 15, 2012 – Dec 14, 2012	Dec 15, 2012 – Jun 14, 2013
1. Derive a transformation matrix in order to correct the MA(1) problem of errors in a SEM.	▨			
2. Simulate the SEM when errors are both MA(1) and contemporaneously correlated by the SAS programming.	▨			
3. Carry out a proposed transformation matrix to correct the MA(1) problem of errors in a SEM.		▨		
4. Examine the errors of SEM after transformation.		▨		
5. Derive the Kullback information criterion for a system of SEM (SKIC) when errors are both MA(1) and contemporaneously correlated.			▨	
6. Examine the performance of the proposed criterion, SKIC, relative to SAIC proposed by Keerativibool (2009).				▨

บทคัดย่อ

รหัสโครงการ	:	MRG5480044
ชื่อโครงการ	:	เกณฑ์ในการคัดเลือกตัวแบบบนพื้นฐานของความแตกต่างแบบสมมาตรของ Kullback-Leibler สำหรับตัวแบบสมการหลายชั้น
ชื่อนักวิจัย และสถาบัน	:	ผู้ช่วยศาสตราจารย์ ดร.วรางคณา กีรติวิบูลย์ มหาวิทยาลัยทักษิณ วิทยาเขตพัทลุง
อีเมล์	:	warang27@gmail.com
ระยะเวลาโครงการ	:	15 มิ.ย. 2554 – 14 มิ.ย. 2556
บทคัดย่อ	:	

ค่าเฉลี่ยเคลื่อนที่ในความคลาดเคลื่อนของตัวแบบสมการหลายชั้นนับเป็นปัญหาสำคัญที่ทำให้ตัวประมาณค่าจากวิธีกำลังสองน้อยที่สุดไม่มีประสิทธิภาพ ด้วยเหตุนี้ ผู้วิจัยจึงต้องการขยายเมทริกซ์การแปลงที่ได้เสนอโดยวรางคณา กีรติวิบูลย์ (2553) เพื่อแก้ไขปัญหาค่าเฉลี่ยเคลื่อนที่อันดับที่ 1 ซึ่งมักเกิดขึ้นในตัวแบบที่ได้จากการประมาณ และเพื่อทำการกู้คืนค่าสังเกตค่าหนึ่งที่หายไปในตัวแบบสมการหลายชั้น หลังจากที่ความคลาดเคลื่อนของตัวแบบถูกแปลงให้เป็นอิสระ เกณฑ์สารสนเทศคูลแบค (Kullback) สำหรับการคัดเลือกตัวแบบสมการหลายชั้น เรียก ณ ที่นี้ว่า SKIC จะได้นำเสนอขึ้น เกณฑ์นี้ถูกสร้างบนพื้นฐานของความแตกต่างแบบสมมาตร โดยได้มาจากการรวมกันของ 2 ความแตกต่างแบบโดยตรง ความแตกต่างแบบสมมาตรนี้ได้รับการยืนยันว่ามีความไวกว่าความแตกต่างแบบโดยตรงแต่ละตัว ประสิทธิภาพของ SKIC ซึ่งเป็นเกณฑ์ที่ได้นำเสนอในการวิจัยครั้งนี้จะถูกตรวจสอบเทียบกับ SAIC ที่ได้เสนอโดยวรางคณา กีรติวิบูลย์ (2552) ผลที่ได้จากการจำลอง พบว่า ความคลาดเคลื่อนของตัวแบบสมการหลายชั้นหลังการแปลงมีความเป็นอิสระ และ SKIC มีประสิทธิภาพดีกว่า SAIC เพราะ SAIC มีแนวโน้มที่จะได้ตัวแบบที่มีตัวแปรอิสระมากกว่า SKIC

คำสำคัญ: เกณฑ์สารสนเทศคูลแบคสำหรับการคัดเลือกตัวแบบสมการหลายชั้น; ค่าเฉลี่ยเคลื่อนที่อันดับที่ 1; ตัวแบบสมการหลายชั้น; เมทริกซ์การแปลง

ABSTRACT

PROJECT CODE	**:**	MRG5480044
PROJECT TITLE	**:**	Model Selection Criterion Based on Kullback-Leibler's Symmetric Divergence for Simultaneous Equations Model
INVESTIGATOR	**:**	Asst.Prof.Dr. Warangkhana Keerativibool Thaksin University, Phatthalung Campus
E-MAIL ADDRESS	**:**	warang27@gmail.com
PROJECT PERIOD	**:**	June 15, 2011 – June 14, 2013
ABSTRACT	**:**	

Moving average in the error of simultaneous equations model (SEM) is a crucial problem to make the estimators from the ordinary least squares (OLS) method are not efficient. For this reason, we extend the transformation matrix which proposed by Keerativibool (2010) in order to correct the first-order moving average, MA(1), that generated in the fitted model and to recover the one lost observation in a SEM. After the errors are transformed to be independent, the Kullback information criterion for select the appropriate SEM, called SKIC, to be going to derive. This criterion is constructed based on the symmetric divergence which obtained by sum of the two directed divergences. The symmetric divergence is arguably more sensitive than either of its individual components. The performance of the proposed criterion, SKIC, is examined relative to SAIC proposed by Keerativibool (2009). The results of simulation study show that the errors of the model after transformation are independent and SKIC convincingly outperformed SAIC, because SAIC has a tendency to overfit the order of the model than SKIC.

Keywords: First-order moving average MA(1); Kullback information criterion for a system of SEM (SKIC); Simultaneous equations model (SEM); Transformation matrix.

ACKNOWLEDGEMENTS

This research would not have been completed without the help and support of several people and my organization, Thaksin University. Thank is also dedicated to the research fund by The Thailand Research Fund, Office of The Higher Education Commission, and Thaksin University under grant No. MRG5480044.

I would like to express my sincere gratitude and appreciation to Assoc.Prof.Dr. Jirawan Jitthavech, Assoc.Prof.Dr. Vichit Lorchirachoonkul, and Assoc.Prof.Dr. Pachitjanut Siripanich, National Institute of Development Administration, for their valuable and continual advice, encouragement, and constructive criticisms throughout this research, which enabled me to complete this research successfully. Appreciation is also tendered to all teachers in the School of Applied Statistics, National Institute of Development Administration for the knowledge provided.

Finally, I would particularly like to thank my parents and my younger sister for their patience, understanding, and encouragement throughout the duration of this work. They had more faith in me than could ever be justified by logical argument.

Asst.Prof.Dr. Warangkhana Keerativibool
June 2013

TABLE OF CONTENTS

		Page
EXECUTIVE SUMMARY		
บทคัดย่อ		
ABSTRACT		
ACKNOWLEDGEMENTS		
TABLE OF CONTENTS		
CHAPTER 1	**INTRODUCTION**	**1**
	1.1 Background	1
	1.2 Objectives of the Research	5
	1.3 Scope of the Research	5
CHAPTER 2	**LITERATURE REVIEW**	**6**
	2.1 Transformation Matrix to Correct the Autocorrelation and/or Moving Average Problems	6
	2.2 Model Selection Criteria	7
CHAPTER 3	**METHODOLOGY**	**10**
	3.1 Derivation of a Proposed Transformation Matrix	10
	3.2 Derivation of a Proposed Model Selection Criterion	11
CHAPTER 4	**SIMULATION STUDY**	**13**
CHAPTER 5	**CONCLUSIONS AND FUTURE WORKS**	**31**
	5.1 Conclusions	31
	5.2 Future Works	31
BIBLIOGRAPHY		**32**
APPENDICES		**37**
	Appendix A Proofs	38
	Appendix B Outputs of this Research	44

CHAPTER 1

INTRODUCTION

1.1 Background

In the application of statistics, the statistical modeling is considered as a major task of study. The statistical processes which can help us to guide a good model with the properties are that parsimony, goodness-of-fit, and generalizability, can be classified into three ways; hypothesis testing of parameters, variable selection algorithms, and model selection criteria (Cavanaugh, 2010). Model selection criterion is a popular tool for selecting the appropriate model, by assessing whether it offers an optimal balance between goodness of fit and parsimony, which are the attributes of the best model (Keerativibool, 2011b). There are many model selection criteria for choosing the appropriate model. The Akaike information criterion, AIC (Akaike, 1973, 1974) was the first model selection criterion to gain widespread acceptance. The later criterion which equally popular was the Kullback information criterion, KIC (Cavanaugh, 1999). One of the primary focuses of this research is to compare the performance of selection the appropriate model from the model selection criterion based on AIC proposed by Keerativibool (2009) relative to the model selection criterion based on KIC proposed in this research.

The model to consider in this research is called a simultaneous equations model (SEM). It is a model that contains variables with two-way flows of influence characteristics. As a consequence, the endogenous variable will become stochastic or the explanatory variable and will correlate with the error terms of the equation. The structural-form of a SEM may be represented as a set of linear simultaneous equations as follows: (Greene, 2008)

$$
\begin{aligned}
y_{t1} = \quad & \gamma_{21} y_{t2} + \gamma_{31} y_{t3} + \ldots + \gamma_{M-1,1} y_{t,M-1} + \gamma_{M1} y_{tM} \\
& \qquad + \beta_{11} x_{t1} + \beta_{21} x_{t2} + \ldots + \beta_{K1} x_{tK} + u_{t1}, \\
\\
y_{t2} = \gamma_{12} y_{t1} + \quad & + \gamma_{32} y_{t3} + \ldots + \gamma_{M-1,2} y_{t,M-1} + \gamma_{M2} y_{tM} \\
& \qquad + \beta_{12} x_{t1} + \beta_{22} x_{t2} + \ldots + \beta_{K2} x_{tK} + u_{t2}, \\
& \qquad \vdots \\
y_{tM} = \gamma_{1M} y_{t1} & + \gamma_{2M} y_{t2} + \gamma_{3M} y_{t3} + \ldots + \gamma_{M-1,M} y_{t,M-1} \\
& \qquad + \beta_{1M} x_{t1} + \beta_{2M} x_{t2} + \ldots + \beta_{KM} x_{tK} + u_{tM}.
\end{aligned}
\qquad \ldots\ldots\ldots (1.1)
$$

In (1.1), there are M equations and M endogenous variables, denoted by $y_{t1}, y_{t2}, \ldots, y_{tM}$ and K predetermined variables, denoted by $x_{t1}, x_{t2}, \ldots, x_{tK}$. The first element of predetermined variables, x_{t1}, will usually be a constant, 1, to allow for the intercept term in each equation. The γ's and β's are denoted as the coefficients of endogenous and predetermined variables, respectively, and u_{t1}, $u_{t2}, \ldots, u_{tM}$ denote the structural errors that are in the form of the first-order moving average, MA(1), and contemporaneously correlated with zero means.

In matrix terms, the system in (1.1) can be written as

$$
\mathbf{Y\Gamma} + \mathbf{XB} = \mathbf{U}, \qquad \ldots\ldots\ldots (1.2)
$$

where $\mathbf{Y}$ is a $T \times M$ matrix of endogenous variables, $\mathbf{\Gamma}$ is a $M \times M$ matrix of coefficients of endogenous variables, and assumed nonsingular, $\mathbf{X}$ is a $T \times K$ matrix of predetermined variables, and assumed full-column rank, $\mathbf{B}$ is a $K \times M$ matrix of coefficients of predetermined variables, and $\mathbf{U}$ is a $T \times M$ matrix of MA(1) and contemporaneously correlated errors; i.e.,

$$
\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1M} \\ y_{21} & y_{22} & \cdots & y_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ y_{T1} & y_{T2} & \cdots & y_{TM} \end{bmatrix}, \quad
\mathbf{\Gamma} = \begin{bmatrix} 1 & -\gamma_{12} & \cdots & -\gamma_{1M} \\ -\gamma_{21} & 1 & \cdots & -\gamma_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ -\gamma_{M1} & -\gamma_{M2} & \cdots & 1 \end{bmatrix},
$$

$$
\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1K} \\ x_{21} & x_{22} & \ldots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T1} & x_{T2} & \ldots & x_{TK} \end{bmatrix}, \quad
\mathbf{B} = \begin{bmatrix} -\beta_{11} & -\beta_{12} & \cdots & -\beta_{1M} \\ -\beta_{21} & -\beta_{22} & \cdots & -\beta_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ -\beta_{K1} & -\beta_{K2} & \cdots & -\beta_{KM} \end{bmatrix}, \quad
\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & \ldots & u_{1M} \\ u_{21} & u_{22} & \ldots & u_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ u_{T1} & u_{T2} & \ldots & u_{TM} \end{bmatrix}.
$$

The reduced-form model described by the structural-form of the model in (1.2) can be written as

$$\mathbf{Y} = \mathbf{X\Pi} + \mathbf{V}, \qquad \text{……… (1.3)}$$

where $\mathbf{\Pi} = -\mathbf{B\Gamma}^{-1}$ is a $\mathrm{K} \times \mathrm{M}$ matrix of unknown parameters and $\mathbf{V} = \mathbf{U\Gamma}^{-1}$ is a $\mathrm{T} \times \mathrm{M}$ matrix of MA(1) and contemporaneously correlated errors; i.e.,

$$\mathbf{\Pi} = \begin{bmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1M} \\ \pi_{21} & \pi_{22} & \cdots & \pi_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{K1} & \pi_{K2} & \cdots & \pi_{KM} \end{bmatrix}, \ \mathbf{V} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1M} \\ v_{21} & v_{22} & \cdots & v_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ v_{T1} & v_{T2} & \cdots & v_{TM} \end{bmatrix}.$$

The j^{th} equation vector of reduced-form model in (1.3) is

$$\mathbf{y}_j = \mathbf{X}\boldsymbol{\pi}_j + \mathbf{v}_j, \ \mathrm{j} = 1, 2, \ldots, \mathrm{M}, \qquad \text{……… (1.4)}$$

where $\mathbf{y}_j$ is a $\mathrm{T} \times 1$ observation vector, $\boldsymbol{\pi}_j$ is a $\mathrm{K} \times 1$ parameter vector, and $\mathbf{v}_j$ is a $\mathrm{T} \times 1$ vector of MA(1) and contemporaneously correlated errors. For all M equations, the models in (1.4) can be represented as a stacked model as follows:

$$\mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\pi} + \mathbf{v}, \qquad \text{……… (1.5)}$$

where $\mathbf{y}$ is a $\mathrm{TM} \times 1$ observation vector consisting of M $(\mathrm{T} \times 1)$ $\mathbf{y}_j$ vectors, $\tilde{\mathbf{X}}$ is a $\mathrm{TM} \times \mathrm{KM}$ diagonal matrix of rank KM consisting of M $(\mathrm{T} \times \mathrm{K})$ identical $\mathbf{X}$ matrices, $\boldsymbol{\pi}$ is a $\mathrm{KM} \times 1$ unknown parameter vector consisting of M $(\mathrm{K} \times 1)$ $\boldsymbol{\pi}_j$ vectors, and $\mathbf{v}$ is a $\mathrm{TM} \times 1$ MA(1) and contemporaneously correlated error vector consisting of M $(\mathrm{T} \times 1)$ $\mathbf{v}_j$ vectors; i.e.,

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{bmatrix}, \ \tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{X} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{X} \end{bmatrix}, \ \boldsymbol{\pi} = \begin{bmatrix} \boldsymbol{\pi}_1 \\ \boldsymbol{\pi}_2 \\ \vdots \\ \boldsymbol{\pi}_M \end{bmatrix}, \ \mathbf{v} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_M \end{bmatrix}.$$

After we estimate the parameters of the reduced-form model in (1.5), we may plot the residuals obtained from the fitted model and may observe a systematic pattern. These residuals may suggest that some essential predetermined variables have not been included in the model. Exclusion could be due to the analyst’s inadequate knowledge of the problem. In this research, each element v_{tj} of the error vector $\mathbf{v}$ in (1.5) is assumed in the form of MA(1),

$$v_{tj} = \varepsilon_{tj} - \theta_j \varepsilon_{t-1,j}, \; t = 1, 2, \ldots, T, \; j = 1, 2, \ldots, M, \qquad \ldots\ldots\ldots (1.6)$$

where T is the number of observations in each equation, M is the number of equations, the error $\varepsilon_{t-1,j}$ is called the first-lag of error ε_{tj}, the MA(1) parameter θ_j of the model must satisfy the following condition to ensure the invertibility of the error terms (Box et al., 1994),

$$\left| \theta_j \right| < 1. \qquad \ldots\ldots\ldots (1.7)$$

The error ε_{tj} in (1.6) is an independent identically distributed random variable, obeying

$$\varepsilon_{tj} \sim N\left(0, \sigma_{jj}\right), \qquad \ldots\ldots\ldots (1.8)$$

so that

$$\boldsymbol{\varepsilon}'_t = \left[\varepsilon_{t1} \;\; \varepsilon_{t2} \;\; \cdots \;\; \varepsilon_{tM}\right] \sim N_M\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \qquad \ldots\ldots\ldots (1.9)$$

where $M \times M$ contemporaneous covariance matrix of the error terms,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1M} & \sigma_{2M} & \cdots & \sigma_{MM} \end{bmatrix},$$

is nonsingular and is of positive symmetric definite matrix. It is noteworthy that the values of v_{1j} in the MA(1) model in (1.6) depend on the values of ε_{0j}, which is unknown. The recovery of v_{1j} will be shown in Chapter 3.

The major consequences of the MA(1) problem are summarized as follows. Although, the ordinary least squares (OLS) estimators are still linear unbiased, they are not efficient; i.e., they do not have minimum variance compared the variance in the model that takes into account correlation. In short, if MA(1) exists in the errors, the OLS estimators are not the best linear unbiased estimators (BLUE). The estimated variances of OLS estimators are biased. Therefore, the usual t and F tests are not generally reliable, and if applied, are likely to give seriously misleading conclusions about the statistical significance of the estimated regression coefficients. Consequently, conventionally computed R^2 becomes an unreliable measure of true R^2. Finally, the conventionally computed variances and standard errors of forecast may be inefficient (Gujarati, 2006). Encouraged by the preceding finding, this research attempts to find a transformation in order to correct the MA(1) problem by extend the knowledge of Keerativibool (2010). After the errors are transformed to be independent, the Kullback information criterion for select the appropriate SEM, called SKIC, to be going to derive. The performance of the proposed criterion, SKIC, is compared and discussed relative to SAIC proposed by Keerativibool (2009).

1.2 Objectives of the Research

The objectives of this research are to develop a model selection method, to separate the most fitting SEM when the errors are both MA(1) and contemporaneously correlated for analyzing a specific system, by applying the Kullback information criterion (KIC) (Cavanaugh, 1999). The topics covered in this research comprise:

1) To derive a transformation matrix in order to correct the MA(1) problem of errors in a SEM.

2) To derive the Kullback information criterion for a system of SEM when errors are both MA(1) and contemporaneously correlated, called SKIC.

3) To examine the performance of the proposed criterion, SKIC, relative to SAIC proposed by Keerativibool (2009).

1.3 Scope of the Research

In this research, the model selection criterion focuses on the M equations of the SEM, with an emphasis on whether the equations are exactly identified or over-identified. In addition, the kinds of correlation among these equations are analyzed, to distinguish moving average (correlation of the error terms across periods of time in the j^{th} equation) and contemporaneous correlation (correlation across equations at time t). The problem of simultaneity is also considered, which is related to some repressors' acting as simultaneous endogenous variables likely to correlate with the error. It is assumed that all residuals from the error terms are normally distributed with conditional means zero vector.

CHAPTER 2

LITERATURE REVIEW

This chapter reviews the literature on the following two topics. Firstly, Section 2.1 is shown the reviews of the transformation matrix to correct the autocorrelation and/or moving average problems. Secondly, Section 2.2 is shown the reviews of the model selection criteria in various types of the model.

2.1 Transformation Matrix to Correct the Autocorrelation and/or Moving Average Problems

Occasionally, when we construct the forecasting model, a common problem in the fitted model is the discovery of autocorrelation (AR) and/or moving average (MA) problems in the residuals. This problem may occur when we start the plausible structural-form of a mis-specified model. A lot of literatures attention to this issue from the past to the present, such as Cochrane and Orcutt (1949) constructed an algorithm for estimating a time series linear regression in the presence of the first-order autocorrelation, AR(1), problem by eliminating the first observation. Prais and Winsten (1954) improved the original Cochrane and Orcutt algorithm by recovering the first observation for the parameter estimation. Ullah et al. (1983) derived a large sample asymptotic approximation for the covariance matrix of the two stage Prais-Winsten estimator of the regression coefficients and then analyzed numerically the efficiency properties of this estimator with respect to OLS and generalized least squares (GLS) with a known autocorrelation coefficient. Choudhury and Power (1995) constructed a new approximate GLS estimator for the linear regression model with AR and MA errors which this estimator consists of the two-step procedure followed by OLS estimation of the transformed model: the first step eliminates the AR component of the error and the second step addresses the MA component. Galbraith and Zinde-Walsh (1995) gave a transformation of the general ARMA error-components in the panel model to yield spherical disturbances. Marazzi and Yohai

(2006) proposed new estimators to transform the response variables which are based on the minimization of a robust measure of residual autocorrelation. These estimators are robust and consistent even if the assumptions of normality and homoscedasticity do not hold. Hwang et al. (2007) constructed a GLS estimator for explosive the AR(1) processes with conditionally heteroscedastic errors. The model under this consideration accommodates diverse conditionally heteroscedastic processes including Engle's, threshold-, and beta-autoregressive conditionally heteroscedastic (ARCH) processes. Vougas (2008) proposed the approximations of the usual GLS transformation matrices for estimation with the AR(1) and AR(2) errors that remove boundary discontinuities. This method avoids constrained optimization that unnecessarily enforces estimated parameters to be in the interior. Keerativibool (2009) and Keerativibool et al. (2009a, 2009b, 2011) proposed a transformation matrix to correct the AR(2) problem in a SEM by extended the Prais-Winsten transformation. Keerativibool (2010) proposed a transformation matrix to correct the MA(1) problem in a regression model. Keerativibool (2011a) proposed a transformation matrix to correct the AR(2) problem in a SEM by using the Cholesky decomposition.

From the past literatures review, we find that there is none of the transformation matrix to correct the MA(1) problem in a SEM. With this motivation, this research attempts to construct a transformation matrix to correct the MA(1) problem along with the consideration of contemporaneous correlation. The transformation is constructed by extend the knowledge of Keerativibool (2010).

2.2 Model Selection Criteria

As mentioned in Chapter 1, the model selection criterion is a popular way to get an appropriate model. The first model selection criterion to gain widespread acceptance was the Akaike information criterion, AIC (Akaike, 1973, 1974). Many other criteria have been then introduced and studied are Bayesian information criterion, BIC (Schwarz, 1978), Hannan and Quinn criterion, HQ (Hannan and Quinn, 1979), corrected version of AIC, AIC$_c$ (Hurvich and Tsai, 1989), multivariate AIC and multivariate AIC$_c$ (Bedrick and Tsai, 1994), modification of AIC, MAIC (Fujikoshi and Satoh, 1997), variants BIC; BIC, Fisher BIC, prior BIC, and Fisher prior BIC (Neath and Cavanaugh, 1997), corrected version of HQ, HQ$_c$ (McQuarrie and Tsai, 1998), Kullback information criterion, KIC (Cavanaugh, 1999), corrected version of BIC, BIC$_c$ (McQuarrie, 1999), an estimation rule of variable selection and parameter estimation in a linear statistical model based on generalized maximum entropy formalism (Golan, 2001), information complexity (ICOMP) criterion for determining influential observations in multivariate time series data based on an intelligent data mining and knowledge discovery technique (Bozdogan and Bearse, 2003), corrected version of KIC, KIC$_c$, (Cavanaugh, 2004; Seghouane and Bekara, 2004; Hafidi and Mkhadri, 2006), modification of KIC, MKIC (Cavanaugh, 2004), KIC$_c$, improved AIC, and improved KIC for nonlinear regression (Kim and Cavanaugh, 2005), incomplete data based on KIC (Seghouane et al., 2005), surface selection criterion, SSC (Bab-Hadiashar and Gheissari, 2006; Gheissari and Bab-Hadiashar, 2008), KIC$_c$ for vector autoregressive modeling (Hafidi, 2006), Multivariate KIC for small sample (Seghouane, 2006), quasi Akaike and quasi Schwarz criteria (Giombini and Szroeter, 2007), mixture regression criterion based on Kullback asymmetric and symmetric divergences (Naik et al., 2007; Hafidi and Mkhadri, 2010), information criterion for probabilistic principal component analysis, ICPPCA (Seghouane and Cichocki, 2007), predictive local asymptotic mixed normality information criterion, PMIC (Sei and Komaki, 2007), system of simultaneous equations AIC, SAIC (Keerativibool, 2009), system of simultaneous equations BIC, SBIC (Keerativibool, 20012).

All of the model selection criteria as reviewed above, AIC and KIC are two well-known measures. Although AIC remains arguably the most widely used of model selection criterion, KIC is a popular competitor. In fact, KIC is often preferred over AIC because its tendency to choose more parsimonious models than AIC. Since KIC is the criterion constructed to target a symmetric divergence, whereas AIC is based on a directed divergence. Symmetric divergence is an alternate of directed divergence, obtained by sum of the two directed divergences, which arguably more sensitive than either of its individual components (Cavanaugh, 1999). Unfortunately, all of the model selection criteria are stated above can not be used in a SEM when the AR and/or MA problems have been occurred, except SAIC and SBIC can be used in a SEM, when there exists the AR(2) problem. Keerativibool et al. (2009a, 2009b, and

2011) and Keerativibool (2010, 2011a, and 2011c) concluded that the AR and MA problems made the overestimated of the errors whether the models were regression or SEM. Consequently, the values of all model selection criteria are incorrect, because they depend on the sum of squared error (SSE) and the mean squared error (MSE). With this motivation, this research attempts to construct a model selection criterion, called the Kullback information criterion for a system of SEM (SKIC), in order to select the appropriate system of the model where the model's errors are considered both MA(1) and contemporaneously correlated. A comparison of performance from SAIC, proposed by Keerativibool (2009) relative to SKIC, proposed in this research, will be shown and discussed in Chapter 4.

CHAPTER 3

METHODOLOGY

This research attempted to determine a model selection criterion, called the Kullback information criterion for a system of SEM (SKIC), in order to enable the selection of the most appropriate system for the model, when model's errors are both MA(1) and contemporaneously correlated. The approach adopted consisted of two stages. First, the knowledge of transformation from Keerativibool (2010) is extended to correct the MA(1) problem and to recover the one lost observation in a SEM. Second, the log-likelihood function of the multivariate model is able to be applied directly to construct the SKIC for the transformed model.

3.1 Derivation of a Proposed Transformation Matrix

Theorem 1: The $\mathrm{TM}\times\mathrm{TM}$ transformation matrix $\mathbf{P}$, used to correct the MA(1) problem in a SEM, is defined by

$$\mathbf{P}=\begin{bmatrix}\mathbf{P}_1 & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{P}_2 & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P}_\mathrm{M}\end{bmatrix}, \qquad \ldots\ldots\ldots\ (3.1)$$

where the $\mathrm{T}\times\mathrm{T}$ transformation matrix $\mathbf{P}_\mathrm{j}$ for the j^{th} equation is

$$\mathbf{P}_\mathrm{j}=\begin{bmatrix}\frac{1}{\sqrt{1+\theta_\mathrm{j}^2}} & 0 & 0 & 0 & \cdots & 0\\ \theta_\mathrm{j} & 1 & 0 & 0 & \cdots & 0\\ \theta_\mathrm{j}^2 & \theta_\mathrm{j} & 1 & 0 & \cdots & 0\\ \theta_\mathrm{j}^3 & \theta_\mathrm{j}^2 & \theta_\mathrm{j} & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\ \theta_\mathrm{j}^{\mathrm{T}-1} & \theta_\mathrm{j}^{\mathrm{T}-2} & \theta_\mathrm{j}^{\mathrm{T}-3} & \theta_\mathrm{j}^{\mathrm{T}-4} & \cdots & 1\end{bmatrix}. \qquad \ldots\ldots\ldots\ (3.2)$$

The transformation matrix $\mathbf{P}$ in (3.1) is used to transform $\mathbf{y}$ and $\tilde{\mathbf{X}}$ in (1.5) to be $\mathbf{y}^*$ and $\tilde{\mathbf{X}}^*$, respectively, such that the errors of the model are independent, insignificantly different from zero, and the MA(1) problem does not exist, but contemporaneously correlated errors still exist. The transformed model can be written as

$$\mathbf{y}^* = \tilde{\mathbf{X}}^* \boldsymbol{\pi} + \boldsymbol{\varepsilon}, \qquad \text{.........} (3.3)$$

where $\mathbf{y}^* = \mathbf{P}\mathbf{y}$, $\tilde{\mathbf{X}}^* = \mathbf{P}\tilde{\mathbf{X}}$, $\mathrm{E}\left(\boldsymbol{\varepsilon} \,|\, \tilde{\mathbf{X}}^*\right) = \mathbf{0}$, and

$$\mathrm{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \,|\, \tilde{\mathbf{X}}^*\right) = \boldsymbol{\Sigma} \otimes \mathbf{I}_{\mathrm{T}} = \begin{bmatrix} \sigma_{11}\mathbf{I}_{\mathrm{T}} & \sigma_{12}\mathbf{I}_{\mathrm{T}} & \cdots & \sigma_{1\mathrm{M}}\mathbf{I}_{\mathrm{T}} \\ \sigma_{21}\mathbf{I}_{\mathrm{T}} & \sigma_{22}\mathbf{I}_{\mathrm{T}} & \cdots & \sigma_{2\mathrm{M}}\mathbf{I}_{\mathrm{T}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\mathrm{M}1}\mathbf{I}_{\mathrm{T}} & \sigma_{\mathrm{M}2}\mathbf{I}_{\mathrm{T}} & \cdots & \sigma_{\mathrm{MM}}\mathbf{I}_{\mathrm{T}} \end{bmatrix}.$$

3.2 Derivation of a Proposed Model Selection Criterion

Suppose that the transformed model in (3.3) is called the candidate model, then the true model can be given as

$$\mathbf{y}^* = \tilde{\mathbf{X}}^* \boldsymbol{\pi}_0 + \boldsymbol{\varepsilon}_0. \qquad \text{.........} (3.4)$$

The notations in (3.3) and (3.4) are defined as follows: $\mathbf{y}^*$ is a $\mathrm{TM}\times 1$ observation vector consisting of M $(\mathrm{T}\times 1)$ $\mathbf{y}_{\mathrm{j}}^*$ (or $\mathbf{P}_{\mathrm{j}}\mathbf{y}_{\mathrm{j}}$) vectors, $\tilde{\mathbf{X}}^*$ is a $\mathrm{TM}\times\mathrm{KM}$ diagonal matrix consisting of M $(\mathrm{T}\times\mathrm{K})$ $\mathbf{X}_{\mathrm{j}}^*$ (or $\mathbf{P}_{\mathrm{j}}\mathbf{X}$) matrices, $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_0$ are the $\mathrm{KM}\times 1$ unknown parameter vectors, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}_0$ are the $\mathrm{TM}\times 1$ independent identically distributed normal random vectors.

For the derivations of the criterion, the true model is assumed to be correctly specified or overfitted by all the candidate models. This means that $\boldsymbol{\pi}_0$ has $\mathrm{K}_0\mathrm{M}$ nonzero entries with $0 < \mathrm{K}_0\mathrm{M} \le \mathrm{KM}$ and the rest of $(\mathrm{K}-\mathrm{K}_0)\mathrm{M}$ entries are equal to zero. The Kullback information criterion for a system of SEM (SKIC) is proposed in Theorem 2.

Theorem 2. When the MA(1) problem is adjusted by the transformation matrix **P**, the Kullback information criterion for a system of SEM defined by

$$\mathrm{SKIC} = \mathrm{T}\log\left|\hat{\mathbf{\Sigma}}\right| + \frac{\mathrm{TM}\left(2\mathrm{K}+\mathrm{M}+1\right)}{\mathrm{T}-\mathrm{K}-\mathrm{M}-1} + \mathrm{TM}\log\left(\frac{2\mathrm{T}}{2\mathrm{T}-2\mathrm{K}-\mathrm{M}+1}\right) + \frac{2\mathrm{TM}}{2\mathrm{T}-2\mathrm{K}-\mathrm{M}+1} \qquad \ldots\ldots\ldots (3.5)$$

is called an asymptotically unbiased estimator of the Kullback-Leibler's symmetric divergence.

CHAPTER 4

SIMULATION STUDY

The model to consider in this research is a system of three SEM (M = 3) and the errors of the model appear the MA(1) problem,

$$\begin{aligned} y_{t1} &= 1 + 2x_{t2} + 3x_{t3} + 4x_{t4} + v_{t1} \\ y_{t2} &= 1 - 0.5x_{t2} - 5x_{t3} - 1.5x_{t4} + v_{t2} \\ y_{t3} &= 1 + x_{t2} + x_{t3} + x_{t4} + v_{t3}, \end{aligned} \qquad \ldots\ldots\ldots (4.1)$$

where $t = 1, 2, \ldots, T = 15$ for the small sample size, $t = 1, 2, \ldots, T = 30$ for the medium sample size, and $t = 1, 2, \ldots, T = 100$ for the large sample size. The steps for simulation and all results are as follows.

Step 1 Using the IML procedure of SAS programming to generate 100,000 vectors of the 3×1 multivariate normal $\boldsymbol{\varepsilon}_t$ in (1.9) as shown the SAS code in Figure 4.1, given zero mean vector, the correlation coefficients of the errors between the equations are

$$\rho_{12} = 0.9,\ \rho_{13} = 0.7,\ \rho_{23} = 0.8,$$

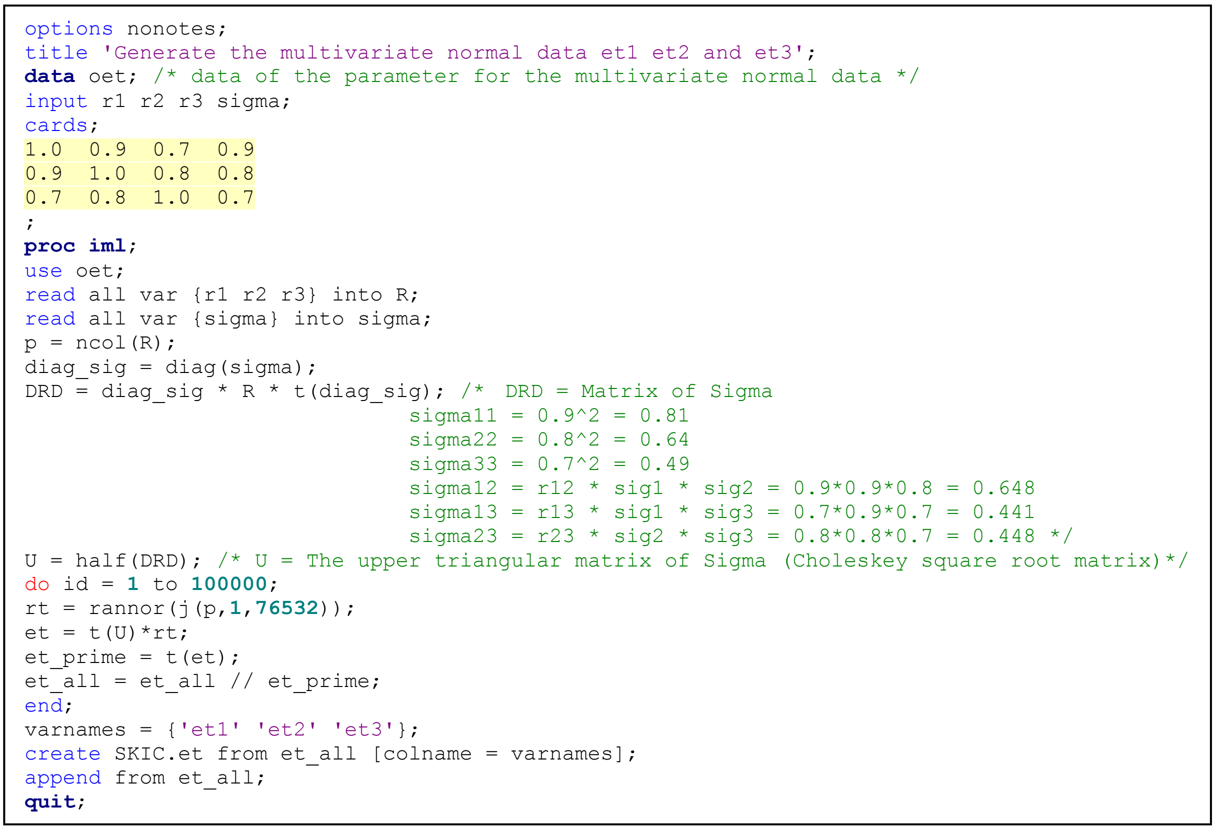
and the variances-covariances of the errors are

$$\sigma_{11} = 0.9^2 = 0.81,\ \sigma_{22} = 0.8^2 = 0.64,\ \sigma_{33} = 0.7^2 = 0.49,$$

$$\sigma_{12} = \rho_{12}\sqrt{\sigma_{11}\sigma_{22}} = 0.648,\ \sigma_{13} = \rho_{13}\sqrt{\sigma_{11}\sigma_{33}} = 0.441,\ \sigma_{23} = \rho_{23}\sqrt{\sigma_{22}\sigma_{33}} = 0.448,$$

then, the form to generate $\boldsymbol{\varepsilon}_t$ in (1.9) is represented by

$$\boldsymbol{\varepsilon}_t = \begin{bmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \\ \varepsilon_{t3} \end{bmatrix} \sim N_3\left(\mathbf{0},\ \boldsymbol{\Sigma} = \begin{bmatrix} 0.81 & 0.648 & 0.441 \\ 0.648 & 0.64 & 0.448 \\ 0.441 & 0.448 & 0.49 \end{bmatrix} \right). \qquad \ldots\ldots\ldots (4.2)$$



```
options nonotes;
title 'Generate the multivariate normal data et1 et2 and et3';
data oet; /* data of the parameter for the multivariate normal data */
input r1 r2 r3 sigma;
cards;
1.0  0.9  0.7  0.9
0.9  1.0  0.8  0.8
0.7  0.8  1.0  0.7
;
proc iml;
use oet;
read all var {r1 r2 r3} into R;
read all var {sigma} into sigma;
p = ncol(R);
diag_sig = diag(sigma);
DRD = diag_sig * R * t(diag_sig); /*  DRD = Matrix of Sigma
                              sigma11 = 0.9^2 = 0.81
                              sigma22 = 0.8^2 = 0.64
                              sigma33 = 0.7^2 = 0.49
                              sigma12 = r12 * sig1 * sig2 = 0.9*0.9*0.8 = 0.648
                              sigma13 = r13 * sig1 * sig3 = 0.7*0.9*0.7 = 0.441
                              sigma23 = r23 * sig2 * sig3 = 0.8*0.8*0.7 = 0.448 */
U = half(DRD); /* U = The upper triangular matrix of Sigma (Choleskey square root matrix)*/
do id = 1 to 100000;
rt = rannor(j(p,1,76532));
et = t(U)*rt;
et_prime = t(et);
et_all = et_all // et_prime;
end;
varnames = {'et1' 'et2' 'et3'};
create SKIC.et from et_all [colname = varnames];
append from et_all;
quit;
```

Figure 4.1 IML procedure to generate 100,000 vectors of the 3x1 multivariate normal $\boldsymbol{\varepsilon}_t$

Step 2 Using the multivariate normal error ε_{t1}, ε_{t2}, and ε_{t3} in Step 1 to construct two series of the MA(1) and contemporaneously correlated errors, v_{t1}, v_{t2}, and v_{t3}, as follows:

$$v_{t1} = \varepsilon_{t1} - 0.5\varepsilon_{t-1,1},\ v_{t2} = \varepsilon_{t2} - 0.6\varepsilon_{t-1,2},\text{ and } v_{t3} = \varepsilon_{t3} - 0.7\varepsilon_{t-1,3},\quad (1^{st}\text{ series}) \qquad \text{..........}\ (4.3)$$

$$v_{t1} = \varepsilon_{t1} + 0.6\varepsilon_{t-1,1},\ v_{t2} = \varepsilon_{t2} + 0.7\varepsilon_{t-1,2},\text{ and } v_{t3} = \varepsilon_{t3} + 0.8\varepsilon_{t-1,3},\quad (2^{nd}\text{ series}) \qquad \text{..........}\ (4.4)$$

for $t = 1, 2, \ldots, 100,000$ and ε_{0j} is arbitrarily given to be zero for all $j = 1, 2, 3$. Split the series of errors v_{t1}, v_{t2}, and v_{t3} in sequence to preserve the MA(1) problem into 1,000 samples, each of which consists of three levels of sample sizes, T = 15, 30, 100 observations. Estimate the MA(1) parameters and test the properties of MA(1) by the

MODEL and ARIMA procedures as shown the SAS code in Figure 4.2. The test confirm that the error of 1,000 samples satisfy the property of MA(1).

```
options nonotes;
/* Generate Macro 1,000 series of vt1, vt2, vt3 */
PROC IMPORT OUT= vt
            DATAFILE= "C:\My Documents\Thaksin\My Paper-
TSU\Statistics\SKIC\Excel\et_150000.csv"
            DBMS=CSV REPLACE;
RUN;
proc iml;
use vt; read point (2:100001) var {id et1 et2 et3 vt1_s1 vt2_s1 vt3_s1 vt1_s2 vt2_s2
vt3_s2} into vt;
varnames = {'id' 'et1' 'et2' 'et3' 'vt1_s1' 'vt2_s1' 'vt3_s1' 'vt1_s2' 'vt2_s2' 'vt3_s2'};
create SKIC.vt from vt [colname = varnames];
append from vt;
quit;
/*************** T15 ***************/
proc iml;
use SKIC.vt; read all into temp;
%macro split;
%local i;
ii = 0;
%do i = 1 %to 1000;
ii = ii+1;
vt_gp_temp = j(15,10,0);
k1 = 1+(ii-1)*15;
k2 = k1+14;
vt_gp_temp = temp[k1:k2,2:10];
varnames = {'et1' 'et2' 'et3' 'vt1_s1' 'vt2_s1' 'vt3_s1' 'vt1_s2' 'vt2_s2' 'vt3_s2'};
create SKIC.vt_T15_gp&i from vt_gp_temp [colname = varnames];
append from vt_gp_temp;
%end;
%mend;
%split;
quit;
title 'Estimate and Test MA(1) First series Theta1 = 0.5, Theta2 = 0.6, Theta3 = 0.7';
%macro esttabs1;
%local i;
%do i = 1 %to 1000;
proc model data = SKIC.vt_T15_gp&i;
 endo vt1_s1 vt2_s1 vt3_s1;
 parms theta1 theta2 theta3;
 vt1_s1 = -theta1*lag1(et1);
 vt2_s1 = -theta2*lag1(et2);
 vt3_s1 = -theta3*lag1(et3);
 fit vt1_s1 vt2_s1 vt3_s1 / outest = SKIC.esttabs1_T15_gp&i sur normal covout;
run; quit;
proc arima data = SKIC.vt_T15_gp&i;
 identify var = vt1_s1 nlag = 6;  estimate q = 1 noint;
 identify var = vt2_s1 nlag = 6;  estimate q = 1 noint;
 identify var = vt3_s1 nlag = 6;  estimate q = 1 noint;
run; quit;
%end;
%mend;
%esttabs1;
quit;
```

Figure 4.2 IML procedure to split 100,000 vectors of $\mathbf{v}_t$ into 1,000 samples and MODEL and ARIMA procedures to estimate and test the MA(1) parameters

```
title 'Estimate and Test MA(1) Second series Theta1 = -0.6, Theta2 = -0.7, Theta3 = -0.8';
%macro esttabs2;
%local i;
%do i = 1 %to 1000;
proc model data = SKIC.vt_T15_gp&i;
 endo vt1_s2 vt2_s2 vt3_s2;
 parms theta1 theta2 theta3;
 vt1_s2 = -theta1*lag1(et1);
 vt2_s2 = -theta2*lag1(et2);
 vt3_s2 = -theta3*lag1(et3);
 fit vt1_s2 vt2_s2 vt3_s2 / outest = SKIC.esttabs2_T15_gp&i sur normal covout;
run; quit;
proc arima data = SKIC.vt_T15_gp&i;
 identify var = vt1_s2 nlag = 6;  estimate q = 1 noint;
 identify var = vt2_s2 nlag = 6;  estimate q = 1 noint;
 identify var = vt3_s2 nlag = 6;  estimate q = 1 noint;
run; quit;
%end;
%mend;
%esttabs2;
quit;
/*************** T30 ***************/
proc iml;
use SKIC.vt; read all into temp;
%macro split;
%local i;
ii = 0;
%do i = 1 %to 1000;
ii = ii+1;
vt_gp_temp = j(30,10,0);
k1 = 1+(ii-1)*30;
k2 = k1+29;
vt_gp_temp = temp[k1:k2,2:10];
varnames = {'et1' 'et2' 'et3' 'vt1_s1' 'vt2_s1' 'vt3_s1' 'vt1_s2' 'vt2_s2' 'vt3_s2'};
create SKIC.vt_T30_gp&i from vt_gp_temp [colname = varnames];
append from vt_gp_temp;
%end;
%mend;
%split;
quit;
title 'Estimate and Test MA(1) First series Theta1 = 0.5, Theta2 = 0.6, Theta3 = 0.7';
%macro esttabs1;
%local i;
%do i = 1 %to 1000;
proc model data = SKIC.vt_T30_gp&i;
 endo vt1_s1 vt2_s1 vt3_s1;
 parms theta1 theta2 theta3;
 vt1_s1 = -theta1*lag1(et1);
 vt2_s1 = -theta2*lag1(et2);
 vt3_s1 = -theta3*lag1(et3);
 fit vt1_s1 vt2_s1 vt3_s1 / outest = SKIC.esttabs1_T30_gp&i sur normal covout;
run; quit;
proc arima data = SKIC.vt_T30_gp&i;
 identify var = vt1_s1 nlag = 6;  estimate q = 1 noint;
 identify var = vt2_s1 nlag = 6;  estimate q = 1 noint;
 identify var = vt3_s1 nlag = 6;  estimate q = 1 noint;
run; quit;
%end;
%mend;
%esttabs1;
quit;
```

Figure 4.2 (Continued)

```
title 'Estimate and Test MA(1) Second series Theta1 = -0.6, Theta2 = -0.7, Theta3 = -0.8';
%macro esttabs2;
%local i;
%do i = 1 %to 1000;
proc model data = SKIC.vt_T30_gp&i;
 endo vt1_s2 vt2_s2 vt3_s2;
 parms theta1 theta2 theta3;
 vt1_s2 = -theta1*lag1(et1);
 vt2_s2 = -theta2*lag1(et2);
 vt3_s2 = -theta3*lag1(et3);
 fit vt1_s2 vt2_s2 vt3_s2 / outest = SKIC.esttabs2_T30_gp&i sur normal covout;
run; quit;
proc arima data = SKIC.vt_T30_gp&i;
 identify var = vt1_s2 nlag = 6;  estimate q = 1 noint;
 identify var = vt2_s2 nlag = 6;  estimate q = 1 noint;
 identify var = vt3_s2 nlag = 6;  estimate q = 1 noint;
run; quit;
%end;
%mend;
%esttabs2;
quit;
/*************** T100 ***************/
proc iml;
use SKIC.vt; read all into temp;
%macro split;
%local i;
ii = 0;
%do i = 1 %to 1000;
ii = ii+1;
vt_gp_temp = j(100,10,0);
k1 = 1+(ii-1)*100;
k2 = k1+99;
vt_gp_temp = temp[k1:k2,2:10];
varnames = {'et1' 'et2' 'et3' 'vt1_s1' 'vt2_s1' 'vt3_s1' 'vt1_s2' 'vt2_s2' 'vt3_s2'};
create SKIC.vt_T100_gp&i from vt_gp_temp [colname = varnames];
append from vt_gp_temp;
%end;
%mend;
%split;
quit;
title 'Estimate and Test MA(1) First series Theta1 = 0.5, Theta2 = 0.6, Theta3 = 0.7';
%macro esttabs1;
%local i;
%do i = 1 %to 1000;
proc model data = SKIC.vt_T100_gp&i;
 endo vt1_s1 vt2_s1 vt3_s1;
 parms theta1 theta2 theta3;
 vt1_s1 = -theta1*lag1(et1);
 vt2_s1 = -theta2*lag1(et2);
 vt3_s1 = -theta3*lag1(et3);
 fit vt1_s1 vt2_s1 vt3_s1 / outest = SKIC.esttabs1_T100_gp&i sur normal covout;
run; quit;
proc arima data = SKIC.vt_T100_gp&i;
 identify var = vt1_s1 nlag = 6;  estimate q = 1 noint;
 identify var = vt2_s1 nlag = 6;  estimate q = 1 noint;
 identify var = vt3_s1 nlag = 6;  estimate q = 1 noint;
run; quit;
%end;
%mend;
%esttabs1;
quit;
```

Figure 4.2 (Continued)

```
title 'Estimate and Test MA(1) Second series Theta1 = -0.6, Theta2 = -0.7, Theta3 = -0.8';
%macro esttabs2;
%local i;
%do i = 1 %to 1000;
proc model data = SKIC.vt_T100_gp&i;
 endo vt1_s2 vt2_s2 vt3_s2;
 parms theta1 theta2 theta3;
 vt1_s2 = -theta1*lag1(et1);
 vt2_s2 = -theta2*lag1(et2);
 vt3_s2 = -theta3*lag1(et3);
 fit vt1_s2 vt2_s2 vt3_s2 / outest = SKIC.esttabs2_T100_gp&i sur normal covout;
run; quit;
proc arima data = SKIC.vt_T100_gp&i;
 identify var = vt1_s2 nlag = 6;  estimate q = 1 noint;
 identify var = vt2_s2 nlag = 6;  estimate q = 1 noint;
 identify var = vt3_s2 nlag = 6;  estimate q = 1 noint;
run; quit;
%end;
%mend;
%esttabs2;
quit;
```

Figure 4.2 (Continued)

Step 3 Using the RANNOR function of SAS programming to generate the independent variables x_{t2} until $x_{t,10}$ about 100,000 observations to be the normal random variables with zero mean and variance equal to one as shown the SAS code in Figure 4.3 where the relevant independent variables are x_{t2}, x_{t3}, and x_{t4} and irrelevant independent variables are x_{t5} until $x_{t,10}$. Again, split the series of independent variables x_{t2} until $x_{t,10}$ in sequence into 1,000 samples, each of which consists of 15, 30, 100 observations. For this research, x_{t1} is given as a constant which equals one.

```
options nonotes;
title 'Generate 100,000 Dataset NID(0,1) of xt2 - xt10’;
data SKIC.xt;
do id = -50 to 100000;
xt2 = rannor(5466666);
xt3 = rannor(2442111);
xt4 = rannor(1753365);
xt5 = rannor(9750004);
xt6 = rannor(2545654);
xt7 = rannor(6533777);
xt8 = rannor(6643221);
xt9 = rannor(6699044);
xt10 = rannor(1235566);
```

Figure 4.3 RANNOR function to generate 100,000 observations of the series of independent variables and IML procedure to split them into 1,000 samples

```
if id > 0
then output;
end;
run;
/*************** T15 ***************/
title 'Generate Macro 1,000 series of xt2 - xt10’;
proc iml;
use SKIC.xt; read all into temp;
%macro split;
%local i;
ii = 0;
%do i = 1 %to 1000;
ii = ii+1;
xt_gp_temp = j(15,10,0);
k1 = 1+(ii-1)*15;
k2 = k1+14;
xt_gp_temp = temp[k1:k2,2:10];
varnames = {'xt2' 'xt3' 'xt4' 'xt5' 'xt6' 'xt7' 'xt8' 'xt9' 'xt10'};
create SKIC.xt_T15_gp&i from xt_gp_temp [colname = varnames];
append from xt_gp_temp;
%end;
%mend;
%split;
quit;
/*************** T30 ***************/
title 'Generate Macro 1,000 series of xt2 - xt10’;
proc iml;
use SKIC.xt; read all into temp;
%macro split;
%local i;
ii = 0;
%do i = 1 %to 1000;
ii = ii+1;
xt_gp_temp = j(30,10,0);
k1 = 1+(ii-1)*29;
k2 = k1+29;
xt_gp_temp = temp[k1:k2,2:10];
varnames = {'xt2' 'xt3' 'xt4' 'xt5' 'xt6' 'xt7' 'xt8' 'xt9' 'xt10'};
create SKIC.xt_T30_gp&i from xt_gp_temp [colname = varnames];
append from xt_gp_temp;
%end;
%mend;
%split;
quit;
/*************** T100 ***************/
title 'Generate Macro 1,000 series of xt2 - xt10’;
proc iml;
use SKIC.xt; read all into temp;
%macro split;
%local i;
ii = 0;
%do i = 1 %to 1000;
ii = ii+1;
xt_gp_temp = j(100,10,0);
k1 = 1+(ii-1)*100;
k2 = k1+99;
xt_gp_temp = temp[k1:k2,2:10];
varnames = {'xt2' 'xt3' 'xt4' 'xt5' 'xt6' 'xt7' 'xt8' 'xt9' 'xt10'};
create SKIC.xt_T100_gp&i from xt_gp_temp [colname = varnames];
append from xt_gp_temp;
%end;
%mend;
%split;
quit;
```

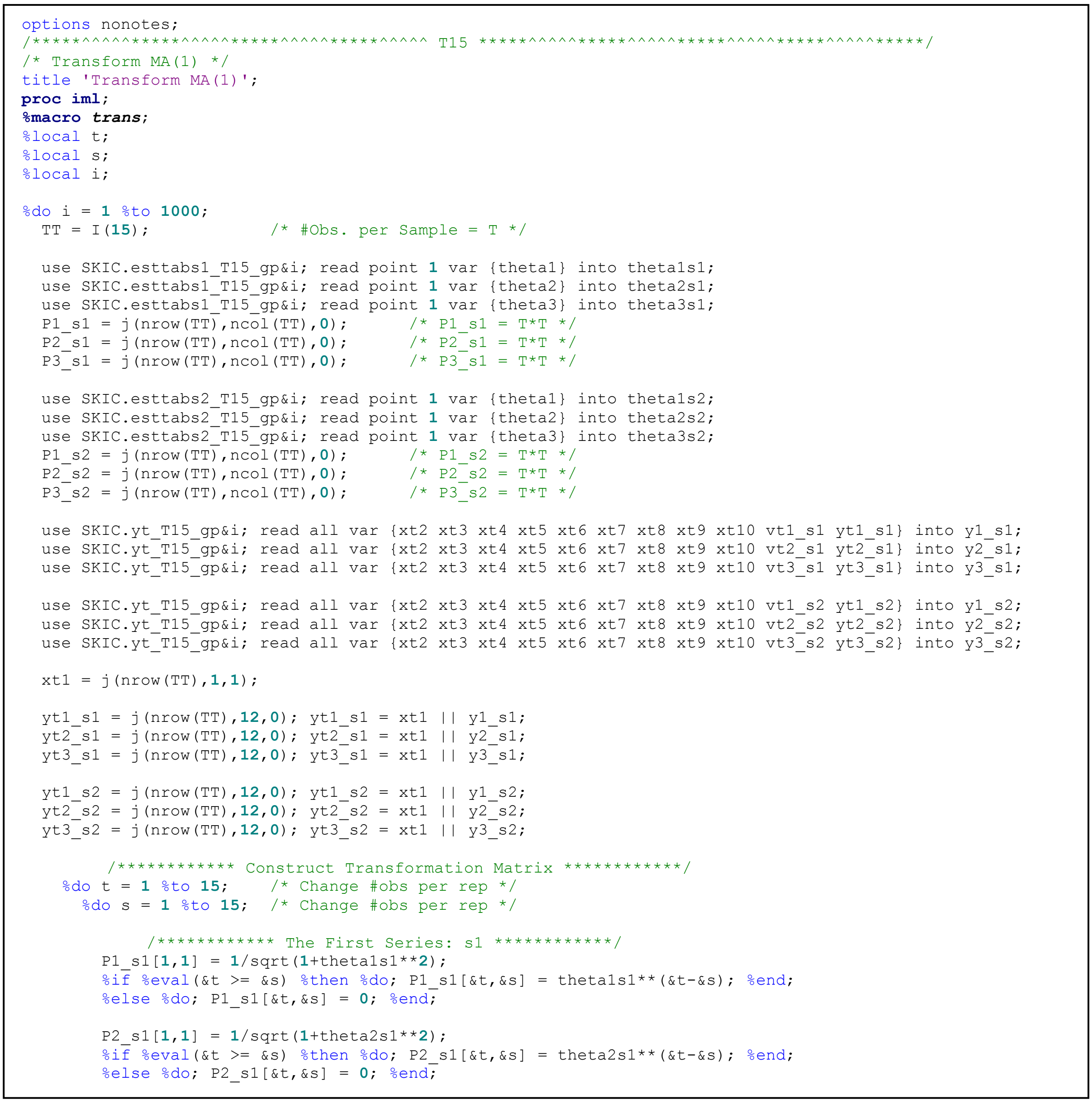
Figure 4.3 (Continued)

Step 4 Using the corresponding relevant independent variables x_{t2}, x_{t3}, and x_{t4} obtained in Step 3 and two series of the MA(1) errors obtained in Step 2 to construct the dependent variables described in (4.1). The SAS code is shown in Figure 4.4.

```
/*************** T15 ***************/
title 'Construct Macro 1,000 series of yt1_st1 - yt3_s1 and yt1_st2 - yt3_s2';
%macro cons_yt;
%local i;
%do i = 1 %to 1000;
data SKIC.yt_T15_gp&i;
set SKIC.xt_T15_gp&i;
set SKIC.vt_T15_gp&i;
yt1_s1 = 1+2*xt2+3*xt3+4*xt4+vt1_s1;
yt2_s1 = 1-0.5*xt2-5*xt3-1.5*xt4+vt2_s1;
yt3_s1 = 1+xt2+xt3+xt4+vt3_s1;

yt1_s2 = 1+2*xt2+3*xt3+4*xt4+vt1_s2;
yt2_s2 = 1-0.5*xt2-5*xt3-1.5*xt4+vt2_s2;
yt3_s2 = 1+xt2+xt3+xt4+vt3_s2;
%end;
%mend;
%cons_yt;
run;
quit;
/*************** T30 ***************/
title 'Construct Macro 1,000 series of yt1_st1 - yt3_s1 and yt1_st2 - yt3_s2';
%macro cons_yt;
%local i;
%do i = 1 %to 1000;
data SKIC.yt_T30_gp&i;
set SKIC.xt_T30_gp&i;
set SKIC.vt_T30_gp&i;
yt1_s1 = 1+2*xt2+3*xt3+4*xt4+vt1_s1;
yt2_s1 = 1-0.5*xt2-5*xt3-1.5*xt4+vt2_s1;
yt3_s1 = 1+xt2+xt3+xt4+vt3_s1;

yt1_s2 = 1+2*xt2+3*xt3+4*xt4+vt1_s2;
yt2_s2 = 1-0.5*xt2-5*xt3-1.5*xt4+vt2_s2;
yt3_s2 = 1+xt2+xt3+xt4+vt3_s2;
%end;
%mend;
%cons_yt;
run;
quit;
/*************** T100 ***************/
title 'Construct Macro 1,000 series of yt1_st1 - yt3_s1 and yt1_st2 - yt3_s2';
%macro cons_yt;
%local i;
%do i = 1 %to 1000;
data SKIC.yt_T100_gp&i;
set SKIC.xt_T100_gp&i;
set SKIC.vt_T100_gp&i;
yt1_s1 = 1+2*xt2+3*xt3+4*xt4+vt1_s1;
yt2_s1 = 1-0.5*xt2-5*xt3-1.5*xt4+vt2_s1;
yt3_s1 = 1+xt2+xt3+xt4+vt3_s1;

yt1_s2 = 1+2*xt2+3*xt3+4*xt4+vt1_s2;
yt2_s2 = 1-0.5*xt2-5*xt3-1.5*xt4+vt2_s2;
yt3_s2 = 1+xt2+xt3+xt4+vt3_s2;
%end;
%mend;
%cons_yt;
run;
quit;
```

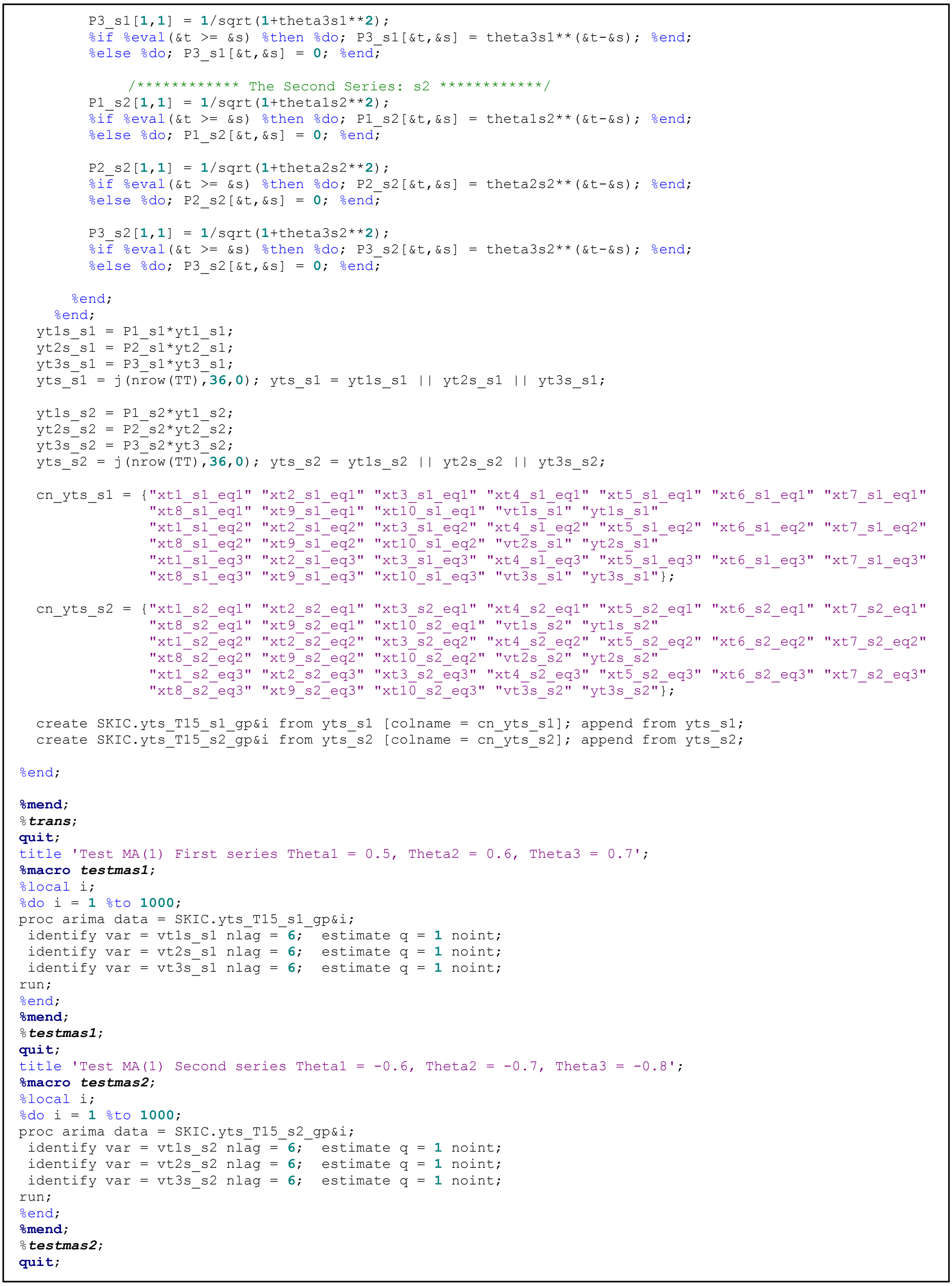
Figure 4.4 Macro facility to construct 1,000 samples of the SEM in (4.1)

Step 5 Using the estimated values of MA(1) parameters obtained in Step 2 to construct the estimate of transformation matrix $\mathbf{P}_j$ in (3.2) for each sample. Apply this transformation matrix to transform the SEM in Step 4 to give the stack of transformed model as shown in (3.3). Test the MA(1) problem in the errors by the ARIMA procedure. The SAS code is shown in Figure 4.5. The test shows that the errors of all transformed samples are independent. Therefore, we can say that the transformation matrix **P** in (3.1) has the power of transformation equal to 100%.



```
options nonotes;
/*****^^^^^*****^^^^^*****^^^^^*****^^^^^ T15 *****^^^^^*****^^^^^*****^^^^^*****^^^^^*****/
/* Transform MA(1) */
title 'Transform MA(1)';
proc iml;
%macro trans;
%local t;
%local s;
%local i;

%do i = 1 %to 1000;
  TT = I(15);            /* #Obs. per Sample = T */

  use SKIC.esttabs1_T15_gp&i; read point 1 var {theta1} into theta1s1;
  use SKIC.esttabs1_T15_gp&i; read point 1 var {theta2} into theta2s1;
  use SKIC.esttabs1_T15_gp&i; read point 1 var {theta3} into theta3s1;
  P1_s1 = j(nrow(TT),ncol(TT),0);      /* P1_s1 = T*T */
  P2_s1 = j(nrow(TT),ncol(TT),0);      /* P2_s1 = T*T */
  P3_s1 = j(nrow(TT),ncol(TT),0);      /* P3_s1 = T*T */

  use SKIC.esttabs2_T15_gp&i; read point 1 var {theta1} into theta1s2;
  use SKIC.esttabs2_T15_gp&i; read point 1 var {theta2} into theta2s2;
  use SKIC.esttabs2_T15_gp&i; read point 1 var {theta3} into theta3s2;
  P1_s2 = j(nrow(TT),ncol(TT),0);      /* P1_s2 = T*T */
  P2_s2 = j(nrow(TT),ncol(TT),0);      /* P2_s2 = T*T */
  P3_s2 = j(nrow(TT),ncol(TT),0);      /* P3_s2 = T*T */

  use SKIC.yt_T15_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt1_s1 yt1_s1} into y1_s1;
  use SKIC.yt_T15_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt2_s1 yt2_s1} into y2_s1;
  use SKIC.yt_T15_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt3_s1 yt3_s1} into y3_s1;

  use SKIC.yt_T15_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt1_s2 yt1_s2} into y1_s2;
  use SKIC.yt_T15_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt2_s2 yt2_s2} into y2_s2;
  use SKIC.yt_T15_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt3_s2 yt3_s2} into y3_s2;

  xt1 = j(nrow(TT),1,1);

  yt1_s1 = j(nrow(TT),12,0); yt1_s1 = xt1 || y1_s1;
  yt2_s1 = j(nrow(TT),12,0); yt2_s1 = xt1 || y2_s1;
  yt3_s1 = j(nrow(TT),12,0); yt3_s1 = xt1 || y3_s1;

  yt1_s2 = j(nrow(TT),12,0); yt1_s2 = xt1 || y1_s2;
  yt2_s2 = j(nrow(TT),12,0); yt2_s2 = xt1 || y2_s2;
  yt3_s2 = j(nrow(TT),12,0); yt3_s2 = xt1 || y3_s2;

        /************ Construct Transformation Matrix ************/
    %do t = 1 %to 15;     /* Change #obs per rep */
      %do s = 1 %to 15;   /* Change #obs per rep */

            /************ The First Series: s1 ************/
        P1_s1[1,1] = 1/sqrt(1+theta1s1**2);
        %if %eval(&t >= &s) %then %do; P1_s1[&t,&s] = theta1s1**(&t-&s); %end;
        %else %do; P1_s1[&t,&s] = 0; %end;

        P2_s1[1,1] = 1/sqrt(1+theta2s1**2);
        %if %eval(&t >= &s) %then %do; P2_s1[&t,&s] = theta2s1**(&t-&s); %end;
        %else %do; P2_s1[&t,&s] = 0; %end;
```

Figure 4.5 IML procedure to construct the estimate of transformation matrix



```
        P3_s1[1,1] = 1/sqrt(1+theta3s1**2);
        %if %eval(&t >= &s) %then %do; P3_s1[&t,&s] = theta3s1**(&t-&s); %end;
        %else %do; P3_s1[&t,&s] = 0; %end;

             /************ The Second Series: s2 ************/
        P1_s2[1,1] = 1/sqrt(1+theta1s2**2);
        %if %eval(&t >= &s) %then %do; P1_s2[&t,&s] = theta1s2**(&t-&s); %end;
        %else %do; P1_s2[&t,&s] = 0; %end;

        P2_s2[1,1] = 1/sqrt(1+theta2s2**2);
        %if %eval(&t >= &s) %then %do; P2_s2[&t,&s] = theta2s2**(&t-&s); %end;
        %else %do; P2_s2[&t,&s] = 0; %end;

        P3_s2[1,1] = 1/sqrt(1+theta3s2**2);
        %if %eval(&t >= &s) %then %do; P3_s2[&t,&s] = theta3s2**(&t-&s); %end;
        %else %do; P3_s2[&t,&s] = 0; %end;

      %end;
    %end;
  yt1s_s1 = P1_s1*yt1_s1;
  yt2s_s1 = P2_s1*yt2_s1;
  yt3s_s1 = P3_s1*yt3_s1;
  yts_s1 = j(nrow(TT),36,0); yts_s1 = yt1s_s1 || yt2s_s1 || yt3s_s1;

  yt1s_s2 = P1_s2*yt1_s2;
  yt2s_s2 = P2_s2*yt2_s2;
  yt3s_s2 = P3_s2*yt3_s2;
  yts_s2 = j(nrow(TT),36,0); yts_s2 = yt1s_s2 || yt2s_s2 || yt3s_s2;

  cn_yts_s1 = {"xt1_s1_eq1" "xt2_s1_eq1" "xt3_s1_eq1" "xt4_s1_eq1" "xt5_s1_eq1" "xt6_s1_eq1" "xt7_s1_eq1"
               "xt8_s1_eq1" "xt9_s1_eq1" "xt10_s1_eq1" "vt1s_s1" "yt1s_s1"
               "xt1_s1_eq2" "xt2_s1_eq2" "xt3_s1_eq2" "xt4_s1_eq2" "xt5_s1_eq2" "xt6_s1_eq2" "xt7_s1_eq2"
               "xt8_s1_eq2" "xt9_s1_eq2" "xt10_s1_eq2" "vt2s_s1" "yt2s_s1"
               "xt1_s1_eq3" "xt2_s1_eq3" "xt3_s1_eq3" "xt4_s1_eq3" "xt5_s1_eq3" "xt6_s1_eq3" "xt7_s1_eq3"
               "xt8_s1_eq3" "xt9_s1_eq3" "xt10_s1_eq3" "vt3s_s1" "yt3s_s1"};

  cn_yts_s2 = {"xt1_s2_eq1" "xt2_s2_eq1" "xt3_s2_eq1" "xt4_s2_eq1" "xt5_s2_eq1" "xt6_s2_eq1" "xt7_s2_eq1"
               "xt8_s2_eq1" "xt9_s2_eq1" "xt10_s2_eq1" "vt1s_s2" "yt1s_s2"
               "xt1_s2_eq2" "xt2_s2_eq2" "xt3_s2_eq2" "xt4_s2_eq2" "xt5_s2_eq2" "xt6_s2_eq2" "xt7_s2_eq2"
               "xt8_s2_eq2" "xt9_s2_eq2" "xt10_s2_eq2" "vt2s_s2" "yt2s_s2"
               "xt1_s2_eq3" "xt2_s2_eq3" "xt3_s2_eq3" "xt4_s2_eq3" "xt5_s2_eq3" "xt6_s2_eq3" "xt7_s2_eq3"
               "xt8_s2_eq3" "xt9_s2_eq3" "xt10_s2_eq3" "vt3s_s2" "yt3s_s2"};

  create SKIC.yts_T15_s1_gp&i from yts_s1 [colname = cn_yts_s1]; append from yts_s1;
  create SKIC.yts_T15_s2_gp&i from yts_s2 [colname = cn_yts_s2]; append from yts_s2;

%end;

%mend;
%trans;
quit;
title 'Test MA(1) First series Theta1 = 0.5, Theta2 = 0.6, Theta3 = 0.7';
%macro testmas1;
%local i;
%do i = 1 %to 1000;
proc arima data = SKIC.yts_T15_s1_gp&i;
 identify var = vt1s_s1 nlag = 6;  estimate q = 1 noint;
 identify var = vt2s_s1 nlag = 6;  estimate q = 1 noint;
 identify var = vt3s_s1 nlag = 6;  estimate q = 1 noint;
run;
%end;
%mend;
%testmas1;
quit;
title 'Test MA(1) Second series Theta1 = -0.6, Theta2 = -0.7, Theta3 = -0.8';
%macro testmas2;
%local i;
%do i = 1 %to 1000;
proc arima data = SKIC.yts_T15_s2_gp&i;
 identify var = vt1s_s2 nlag = 6;  estimate q = 1 noint;
 identify var = vt2s_s2 nlag = 6;  estimate q = 1 noint;
 identify var = vt3s_s2 nlag = 6;  estimate q = 1 noint;
run;
%end;
%mend;
%testmas2;
quit;
```

Figure 4.5 (Continued)

```
/*****^^^^^*****^^^^^*****^^^^^*****^^^^^ T30 *****^^^^^*****^^^^^*****^^^^^*****^^^^^*****/
/* Transform MA(1) */
title 'Transform MA(1)';
proc iml;
%macro trans;
%local t;
%local s;
%local i;

%do i = 1 %to 1000;
  TT = I(30);            /* #Obs. per Sample = T */

  use SKIC.esttabs1_T30_gp&i; read point 1 var {theta1} into theta1s1;
  use SKIC.esttabs1_T30_gp&i; read point 1 var {theta2} into theta2s1;
  use SKIC.esttabs1_T30_gp&i; read point 1 var {theta3} into theta3s1;
  P1_s1 = j(nrow(TT),ncol(TT),0);      /* P1_s1 = T*T */
  P2_s1 = j(nrow(TT),ncol(TT),0);      /* P2_s1 = T*T */
  P3_s1 = j(nrow(TT),ncol(TT),0);      /* P3_s1 = T*T */

  use SKIC.esttabs2_T30_gp&i; read point 1 var {theta1} into theta1s2;
  use SKIC.esttabs2_T30_gp&i; read point 1 var {theta2} into theta2s2;
  use SKIC.esttabs2_T30_gp&i; read point 1 var {theta3} into theta3s2;
  P1_s2 = j(nrow(TT),ncol(TT),0);      /* P1_s2 = T*T */
  P2_s2 = j(nrow(TT),ncol(TT),0);      /* P2_s2 = T*T */
  P3_s2 = j(nrow(TT),ncol(TT),0);      /* P3_s2 = T*T */

  use SKIC.yt_T30_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt1_s1 yt1_s1} into y1_s1;
  use SKIC.yt_T30_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt2_s1 yt2_s1} into y2_s1;
  use SKIC.yt_T30_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt3_s1 yt3_s1} into y3_s1;

  use SKIC.yt_T30_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt1_s2 yt1_s2} into y1_s2;
  use SKIC.yt_T30_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt2_s2 yt2_s2} into y2_s2;
  use SKIC.yt_T30_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt3_s2 yt3_s2} into y3_s2;

  xt1 = j(nrow(TT),1,1);

  yt1_s1 = j(nrow(TT),12,0); yt1_s1 = xt1 || y1_s1;
  yt2_s1 = j(nrow(TT),12,0); yt2_s1 = xt1 || y2_s1;
  yt3_s1 = j(nrow(TT),12,0); yt3_s1 = xt1 || y3_s1;

  yt1_s2 = j(nrow(TT),12,0); yt1_s2 = xt1 || y1_s2;
  yt2_s2 = j(nrow(TT),12,0); yt2_s2 = xt1 || y2_s2;
  yt3_s2 = j(nrow(TT),12,0); yt3_s2 = xt1 || y3_s2;

         /************ Construct Transformation Matrix ************/
    %do t = 1 %to 30;    /* Change #obs per rep */
      %do s = 1 %to 30;  /* Change #obs per rep */

             /************ The First Series: s1 ************/
        P1_s1[1,1] = 1/sqrt(1+theta1s1**2);
        %if %eval(&t >= &s) %then %do; P1_s1[&t,&s] = theta1s1**(&t-&s); %end;
        %else %do; P1_s1[&t,&s] = 0; %end;

        P2_s1[1,1] = 1/sqrt(1+theta2s1**2);
        %if %eval(&t >= &s) %then %do; P2_s1[&t,&s] = theta2s1**(&t-&s); %end;
        %else %do; P2_s1[&t,&s] = 0; %end;

        P3_s1[1,1] = 1/sqrt(1+theta3s1**2);
        %if %eval(&t >= &s) %then %do; P3_s1[&t,&s] = theta3s1**(&t-&s); %end;
        %else %do; P3_s1[&t,&s] = 0; %end;

             /************ The Second Series: s2 ************/
        P1_s2[1,1] = 1/sqrt(1+theta1s2**2);
        %if %eval(&t >= &s) %then %do; P1_s2[&t,&s] = theta1s2**(&t-&s); %end;
        %else %do; P1_s2[&t,&s] = 0; %end;

        P2_s2[1,1] = 1/sqrt(1+theta2s2**2);
        %if %eval(&t >= &s) %then %do; P2_s2[&t,&s] = theta2s2**(&t-&s); %end;
        %else %do; P2_s2[&t,&s] = 0; %end;

        P3_s2[1,1] = 1/sqrt(1+theta3s2**2);
        %if %eval(&t >= &s) %then %do; P3_s2[&t,&s] = theta3s2**(&t-&s); %end;
        %else %do; P3_s2[&t,&s] = 0; %end;

      %end;
    %end;
```

Figure 4.5 (Continued)

```
  yt1s_s1 = P1_s1*yt1_s1;
  yt2s_s1 = P2_s1*yt2_s1;
  yt3s_s1 = P3_s1*yt3_s1;
  yts_s1 = j(nrow(TT),36,0); yts_s1 = yt1s_s1 || yt2s_s1 || yt3s_s1;

  yt1s_s2 = P1_s2*yt1_s2;
  yt2s_s2 = P2_s2*yt2_s2;
  yt3s_s2 = P3_s2*yt3_s2;
  yts_s2 = j(nrow(TT),36,0); yts_s2 = yt1s_s2 || yt2s_s2 || yt3s_s2;

  cn_yts_s1 = {"xt1_s1_eq1" "xt2_s1_eq1" "xt3_s1_eq1" "xt4_s1_eq1" "xt5_s1_eq1" "xt6_s1_eq1" "xt7_s1_eq1"
               "xt8_s1_eq1" "xt9_s1_eq1" "xt10_s1_eq1" "vt1s_s1" "yt1s_s1"
               "xt1_s1_eq2" "xt2_s1_eq2" "xt3_s1_eq2" "xt4_s1_eq2" "xt5_s1_eq2" "xt6_s1_eq2" "xt7_s1_eq2"
               "xt8_s1_eq2" "xt9_s1_eq2" "xt10_s1_eq2" "vt2s_s1" "yt2s_s1"
               "xt1_s1_eq3" "xt2_s1_eq3" "xt3_s1_eq3" "xt4_s1_eq3" "xt5_s1_eq3" "xt6_s1_eq3" "xt7_s1_eq3"
               "xt8_s1_eq3" "xt9_s1_eq3" "xt10_s1_eq3" "vt3s_s1" "yt3s_s1"};

  cn_yts_s2 = {"xt1_s2_eq1" "xt2_s2_eq1" "xt3_s2_eq1" "xt4_s2_eq1" "xt5_s2_eq1" "xt6_s2_eq1" "xt7_s2_eq1"
               "xt8_s2_eq1" "xt9_s2_eq1" "xt10_s2_eq1" "vt1s_s2" "yt1s_s2"
               "xt1_s2_eq2" "xt2_s2_eq2" "xt3_s2_eq2" "xt4_s2_eq2" "xt5_s2_eq2" "xt6_s2_eq2" "xt7_s2_eq2"
               "xt8_s2_eq2" "xt9_s2_eq2" "xt10_s2_eq2" "vt2s_s2" "yt2s_s2"
               "xt1_s2_eq3" "xt2_s2_eq3" "xt3_s2_eq3" "xt4_s2_eq3" "xt5_s2_eq3" "xt6_s2_eq3" "xt7_s2_eq3"
               "xt8_s2_eq3" "xt9_s2_eq3" "xt10_s2_eq3" "vt3s_s2" "yt3s_s2"};

  create SKIC.yts_T30_s1_gp&i from yts_s1 [colname = cn_yts_s1]; append from yts_s1;
  create SKIC.yts_T30_s2_gp&i from yts_s2 [colname = cn_yts_s2]; append from yts_s2;

%end;

%mend;
%trans;
quit;
title 'Test MA(1) First series Theta1 = 0.5, Theta2 = 0.6, Theta3 = 0.7';
%macro testmas1;
%local i;
%do i = 1 %to 1000;
proc arima data = SKIC.yts_T30_s1_gp&i;
 identify var = vt1s_s1 nlag = 6;  estimate q = 1 noint;
 identify var = vt2s_s1 nlag = 6;  estimate q = 1 noint;
 identify var = vt3s_s1 nlag = 6;  estimate q = 1 noint;
run;
%end;
%mend;
%testmas1;
quit;
title 'Test MA(1) Second series Theta1 = -0.6, Theta2 = -0.7, Theta3 = -0.8';
%macro testmas2;
%local i;
%do i = 1 %to 1000;
proc arima data = SKIC.yts_T30_s2_gp&i;
 identify var = vt1s_s2 nlag = 6;  estimate q = 1 noint;
 identify var = vt2s_s2 nlag = 6;  estimate q = 1 noint;
 identify var = vt3s_s2 nlag = 6;  estimate q = 1 noint;
run;
%end;
%mend;
%testmas2;
quit;
/*****^^^^^*****^^^^^*****^^^^^*****^^^^^ T100 *****^^^^^*****^^^^^*****^^^^^*****^^^^^*****/
/* Transform MA(1) */
title 'Transform MA(1)';
proc iml;
%macro trans;
%local t;
%local s;
%local i;

%do i = 1 %to 1000;
  TT = I(100);             /* #Obs. per Sample = T */

  use SKIC.esttabs1_T100_gp&i; read point 1 var {theta1} into theta1s1;
  use SKIC.esttabs1_T100_gp&i; read point 1 var {theta2} into theta2s1;
  use SKIC.esttabs1_T100_gp&i; read point 1 var {theta3} into theta3s1;
  P1_s1 = j(nrow(TT),ncol(TT),0);      /* P1_s1 = T*T */
  P2_s1 = j(nrow(TT),ncol(TT),0);      /* P2_s1 = T*T */
  P3_s1 = j(nrow(TT),ncol(TT),0);      /* P3_s1 = T*T */
```

Figure 4.5 (Continued)

```
use SKIC.esttabs2_T100_gp&i; read point 1 var {theta1} into theta1s2;
use SKIC.esttabs2_T100_gp&i; read point 1 var {theta2} into theta2s2;
use SKIC.esttabs2_T100_gp&i; read point 1 var {theta3} into theta3s2;
P1_s2 = j(nrow(TT),ncol(TT),0);      /* P1_s2 = T*T */
P2_s2 = j(nrow(TT),ncol(TT),0);      /* P2_s2 = T*T */
P3_s2 = j(nrow(TT),ncol(TT),0);      /* P3_s2 = T*T */

use SKIC.yt_T100_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt1_s1 yt1_s1} into y1_s1;
use SKIC.yt_T100_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt2_s1 yt2_s1} into y2_s1;
use SKIC.yt_T100_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt3_s1 yt3_s1} into y3_s1;

use SKIC.yt_T100_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt1_s2 yt1_s2} into y1_s2;
use SKIC.yt_T100_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt2_s2 yt2_s2} into y2_s2;
use SKIC.yt_T100_gp&i; read all var {xt2 xt3 xt4 xt5 xt6 xt7 xt8 xt9 xt10 vt3_s2 yt3_s2} into y3_s2;

xt1 = j(nrow(TT),1,1);

yt1_s1 = j(nrow(TT),12,0); yt1_s1 = xt1 || y1_s1;
yt2_s1 = j(nrow(TT),12,0); yt2_s1 = xt1 || y2_s1;
yt3_s1 = j(nrow(TT),12,0); yt3_s1 = xt1 || y3_s1;

yt1_s2 = j(nrow(TT),12,0); yt1_s2 = xt1 || y1_s2;
yt2_s2 = j(nrow(TT),12,0); yt2_s2 = xt1 || y2_s2;
yt3_s2 = j(nrow(TT),12,0); yt3_s2 = xt1 || y3_s2;

      /************ Construct Transformation Matrix ************/
  %do t = 1 %to 100;    /* Change #obs per rep */
    %do s = 1 %to 100;  /* Change #obs per rep */

          /************ The First Series: s1 ************/
      P1_s1[1,1] = 1/sqrt(1+theta1s1**2);
      %if %eval(&t >= &s) %then %do; P1_s1[&t,&s] = theta1s1**(&t-&s); %end;
      %else %do; P1_s1[&t,&s] = 0; %end;

      P2_s1[1,1] = 1/sqrt(1+theta2s1**2);
      %if %eval(&t >= &s) %then %do; P2_s1[&t,&s] = theta2s1**(&t-&s); %end;
      %else %do; P2_s1[&t,&s] = 0; %end;

      P3_s1[1,1] = 1/sqrt(1+theta3s1**2);
      %if %eval(&t >= &s) %then %do; P3_s1[&t,&s] = theta3s1**(&t-&s); %end;
      %else %do; P3_s1[&t,&s] = 0; %end;

          /************ The Second Series: s2 ************/
      P1_s2[1,1] = 1/sqrt(1+theta1s2**2);
      %if %eval(&t >= &s) %then %do; P1_s2[&t,&s] = theta1s2**(&t-&s); %end;
      %else %do; P1_s2[&t,&s] = 0; %end;

      P2_s2[1,1] = 1/sqrt(1+theta2s2**2);
      %if %eval(&t >= &s) %then %do; P2_s2[&t,&s] = theta2s2**(&t-&s); %end;
      %else %do; P2_s2[&t,&s] = 0; %end;

      P3_s2[1,1] = 1/sqrt(1+theta3s2**2);
      %if %eval(&t >= &s) %then %do; P3_s2[&t,&s] = theta3s2**(&t-&s); %end;
      %else %do; P3_s2[&t,&s] = 0; %end;

    %end;
  %end;

yt1s_s1 = P1_s1*yt1_s1;
yt2s_s1 = P2_s1*yt2_s1;
yt3s_s1 = P3_s1*yt3_s1;
yts_s1 = j(nrow(TT),36,0); yts_s1 = yt1s_s1 || yt2s_s1 || yt3s_s1;

yt1s_s2 = P1_s2*yt1_s2;
yt2s_s2 = P2_s2*yt2_s2;
yt3s_s2 = P3_s2*yt3_s2;
yts_s2 = j(nrow(TT),36,0); yts_s2 = yt1s_s2 || yt2s_s2 || yt3s_s2;

cn_yts_s1 = {"xt1_s1_eq1" "xt2_s1_eq1" "xt3_s1_eq1" "xt4_s1_eq1" "xt5_s1_eq1" "xt6_s1_eq1" "xt7_s1_eq1"
             "xt8_s1_eq1" "xt9_s1_eq1" "xt10_s1_eq1" "vt1s_s1" "yt1s_s1"
             "xt1_s1_eq2" "xt2_s1_eq2" "xt3_s1_eq2" "xt4_s1_eq2" "xt5_s1_eq2" "xt6_s1_eq2" "xt7_s1_eq2"
             "xt8_s1_eq2" "xt9_s1_eq2" "xt10_s1_eq2" "vt2s_s1" "yt2s_s1"
             "xt1_s1_eq3" "xt2_s1_eq3" "xt3_s1_eq3" "xt4_s1_eq3" "xt5_s1_eq3" "xt6_s1_eq3" "xt7_s1_eq3"
             "xt8_s1_eq3" "xt9_s1_eq3" "xt10_s1_eq3" "vt3s_s1" "yt3s_s1"};
```

Figure 4.5 (Continued)

```
  cn_yts_s2 = {"xt1_s2_eq1" "xt2_s2_eq1" "xt3_s2_eq1" "xt4_s2_eq1" "xt5_s2_eq1" "xt6_s2_eq1" "xt7_s2_eq1"
               "xt8_s2_eq1" "xt9_s2_eq1" "xt10_s2_eq1" "vt1s_s2" "yt1s_s2"
               "xt1_s2_eq2" "xt2_s2_eq2" "xt3_s2_eq2" "xt4_s2_eq2" "xt5_s2_eq2" "xt6_s2_eq2" "xt7_s2_eq2"
               "xt8_s2_eq2" "xt9_s2_eq2" "xt10_s2_eq2" "vt2s_s2" "yt2s_s2"
               "xt1_s2_eq3" "xt2_s2_eq3" "xt3_s2_eq3" "xt4_s2_eq3" "xt5_s2_eq3" "xt6_s2_eq3" "xt7_s2_eq3"
               "xt8_s2_eq3" "xt9_s2_eq3" "xt10_s2_eq3" "vt3s_s2" "yt3s_s2"};

  create SKIC.yts_T100_s1_gp&i from yts_s1 [colname = cn_yts_s1]; append from yts_s1;
  create SKIC.yts_T100_s2_gp&i from yts_s2 [colname = cn_yts_s2]; append from yts_s2;

%end;

%mend;
%trans;
quit;
```

Figure 4.5 (Continued)

Step 6 Using the assumption of nested model to construct the candidate models which are the models include the columns of independent variables in a sequentially nested fashion; i.e., columns 1 to K define the design matrix for the candidate model with dimension K. For 1,000 transformed samples, we estimate the parameters of the transformed model by the GLS method. Then calculate SKIC in (3.5) and SAIC proposed by Keerativibool (2009),

$$\mathrm{SAIC} = \mathrm{T}\log\left|\,\hat{\boldsymbol{\Sigma}}_{\mathrm{UE}}\,\right| + \mathrm{M}\left(\mathrm{K}+\mathrm{M}+3\right), \qquad \ldots\ldots\ldots\ (4.5)$$

where $\hat{\boldsymbol{\Sigma}}_{\mathrm{UE}} = \dfrac{\mathrm{T}}{\mathrm{T}-\mathrm{K}}\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\Sigma}}_{\mathrm{UE}}$ and $\hat{\boldsymbol{\Sigma}}$ represent the estimated contemporaneous covariance matrix of the error terms by the methods of unbiased estimator and maximum likelihood estimator, respectively. Therefore SAIC in (4.5) can be rewritten as

$$\mathrm{SAIC} = \mathrm{T}\log\left|\,\hat{\boldsymbol{\Sigma}}\,\right| + \mathrm{TM}\log\left(\frac{\mathrm{T}}{\mathrm{T}-\mathrm{K}}\right) + \mathrm{M}\left(\mathrm{K}+\mathrm{M}+3\right). \quad \ldots\ldots\ldots\ (4.6)$$

The candidate model that has the minimum value of model selection criterion is called the best model. Model selection criterion performance is examined by a measure of counting the frequency of order being selected. The results of comparing are shown in Table 1.

Table 1. Frequency of the model order being selected by SAIC and SKIC for 1,000 samples

T	Series of Errors v_{tj}	Criteria	K								
			2	3	4	5	6	7	8	9	10
15	(4.3)	SAIC	0	0	832	75	30	15	16	2	30
		SKIC	0	0	**1000**	0	0	0	0	0	0
15	(4.4)	SAIC	0	0	809	98	32	13	18	2	28
		SKIC	0	0	**1000**	0	0	0	0	0	0
30	(4.3)	SAIC	0	0	919	60	13	6	2	0	0
		SKIC	0	0	**999**	1	0	0	0	0	0
30	(4.4)	SAIC	0	0	886	86	20	6	2	0	0
		SKIC	0	0	**994**	6	0	0	0	0	0
100	(4.3)	SAIC	0	0	952	39	9	0	0	0	0
		SKIC	0	0	**1000**	0	0	0	0	0	0
100	(4.4)	SAIC	0	0	910	55	20	7	5	0	3
		SKIC	0	0	**982**	12	5	0	0	0	1

Note: Boldface type indicates the maximum frequency of correct order being selected.

Step 7 Calculate the observed L_2 distance, scaled by $1/T$, between the true model in (3.4) and the candidate model in (3.3) which was defined by McQuarrie et al. (1997) and McQuarrie (1999),

$$L_2(k) = \frac{1}{T}(\boldsymbol{\pi}_0 - \hat{\boldsymbol{\pi}})' \tilde{\mathbf{X}}^{*\prime}\left(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_T\right)\tilde{\mathbf{X}}^{*}(\boldsymbol{\pi}_0 - \hat{\boldsymbol{\pi}}),$$

and calculate the observed L_2 efficiency which defined as

$$\text{Observed } L_2 \text{ efficiency } = \frac{\min_{1\le k\le K} L_2(k)}{L_2(k_s)},$$

where K is the class of all possible candidate models, k is the rank of fitted candidate model, and k_s is the model selected by specific model selection criterion. The closer the selected model is to the true model, the higher the efficiency. Therefore, the best model selection criterion will select a model which yields high efficiency even in small samples. For 1,000 transformed samples, the results of comparing the observed L_2 efficiency are shown in Table 2.

Table 2. Average and standard deviation of the observed L_2 efficiency over 1,000 samples

T	Series of Errors v_{tj}	Criteria	Statistics	
			Ave. L_2 eff.	S.D. L_2 eff.
15	(4.3)	SAIC	0.7762	0.3170
		SKIC	**0.8843**	**0.2060**
15	(4.4)	SAIC	0.7213	0.3486
		SKIC	**0.8293**	**0.2749**
30	(4.3)	SAIC	0.9436	0.1718
		SKIC	**0.9860**	**0.0868**
30	(4.4)	SAIC	0.8999	0.2341
		SKIC	**0.9487**	**0.1822**
100	(4.3)	SAIC	0.9757	0.1113
		SKIC	**1.0000**	**0.0005**
100	(4.4)	SAIC	0.9527	0.1581
		SKIC	**0.9894**	**0.0810**

Note: Boldface type indicates the best performance.

Step 8 The results of the frequency of correct order being selected from Steps 6 in Table 1 can be concluded that the performance of SKIC in (3.5) convincingly outperformed SAIC in (4.6) for all three levels of the sample sizes (T = 15, 30, 100) and two series of the MA(1) and contemporaneously correlated errors v_{tj} in (4.3) and (4.4), because SAIC has a tendency to overfit the order of the model than SKIC. The results of the observed L_2 efficiency from Steps 7 in Table 2 also confirm that SKIC has a large observed L_2 efficiency and small standard deviation of the observed L_2 efficiency than SAIC, then SKIC is likely better than SAIC. In Table 3, we show the average and standard deviation of SAIC and SKIC for 1,000 transformed samples. In this table we found that SAIC presents a large negative bias than SKIC that maybe the main reason for the number of correct model order being selected is less.

Table 3. Average and standard deviation of SAIC and SKIC for 1,000 samples of the sample size *T* and the series of errors v_{tj} in (4.3) and (4.4)

T = 15 and errors v_{tj} in (4.3)					T = 15 and errors v_{tj} in (4.4)				
K	SAIC		SKIC		K	SAIC		SKIC	
	Ave.	S.D.	Mean	S.D.		Ave.	S.D.	Mean	S.D.
2	6.295	0.977	8.281	0.977	2	7.215	1.310	9.201	1.311
3	3.290	0.998	6.443	0.998	3	3.888	1.232	7.041	1.232
4	**-2.351**	0.862	**2.309**	0.862	4	**-2.300**	0.903	**2.359**	0.903
5	-1.934	0.904	4.732	0.904	5	-1.919	0.949	4.747	0.949
6	-1.507	0.964	7.954	0.964	6	-1.493	1.006	7.968	1.006
7	-1.075	1.023	12.541	1.023	7	-1.066	1.072	12.549	1.072
8	-0.649	1.160	19.800	1.160	8	-0.648	1.174	19.801	1.174
9	1.434	1.375	35.330	1.375	9	1.577	1.406	35.473	1.406
10	0.185	1.529	73.700	1.529	10	0.143	1.481	73.659	1.481

T = 30 and errors v_{tj} in (4.3)					T = 30 and errors v_{tj} in (4.4)				
K	SAIC		SKIC		K	SAIC		SKIC	
	Ave.	S.D.	Mean	S.D.		Ave.	S.D.	Mean	S.D.
2	6.197	0.875	6.824	0.875	2	7.103	1.259	7.730	1.259
3	2.967	0.859	3.916	0.859	3	3.617	1.109	4.566	1.109
4	**-3.131**	0.494	**-1.827**	0.494	4	**-3.065**	0.522	**-1.762**	0.522
5	-2.938	0.501	-1.243	0.501	5	-2.885	0.529	-1.191	0.529
6	-2.734	0.509	-0.606	0.509	6	-2.685	0.533	-0.557	0.533
7	-2.528	0.527	0.081	0.527	7	-2.485	0.545	0.124	0.545
8	-2.306	0.543	0.840	0.543	8	-2.275	0.555	0.872	0.555
9	-0.309	0.656	3.440	0.656	9	-0.168	0.704	3.581	0.704
10	-1.846	0.559	2.582	0.559	10	-1.834	0.585	2.594	0.585

Table 3. (Continued)

	T = 100 and errors v_{tj} in (4.3)					T = 100 and errors v_{tj} in (4.4)			
K	SAIC		SKIC		K	SAIC		SKIC	
	Ave.	S.D.	Mean	S.D.		Ave.	S.D.	Mean	S.D.
2	6.104	0.617	6.241	0.617	2	7.028	1.034	7.166	1.034
3	2.721	0.570	2.927	0.570	3	3.453	0.898	3.659	0.898
4	**-3.752**	0.265	**-3.476**	0.265	4	**-3.718**	0.293	**-3.442**	0.293
5	-3.693	0.266	-3.344	0.266	5	-3.664	0.289	-3.315	0.289
6	-3.634	0.267	-3.210	0.267	6	-3.610	0.288	-3.187	0.288
7	-3.574	0.267	-3.074	0.267	7	-3.552	0.288	-3.053	0.288
8	-3.514	0.267	-2.936	0.267	8	-3.496	0.284	-2.918	0.284
9	-1.369	0.383	-0.711	0.383	9	-1.041	0.448	-0.383	0.448
10	-3.392	0.271	-2.652	0.271	10	-3.379	0.279	-2.638	0.279

Note: Boldface type indicates the minimum average value of SAIC and SKIC.

CHAPTER 5

CONCLUSIONS AND FUTURE WORKS

5.1 Conclusions

In this research, the transformation matrix in order to correct the MA(1) problem and to recover the one lost observation along with the consideration of contemporaneous correlation in a SEM is proposed. Then, the Kullback information criterion for a system of SEM, called SKIC, is proposed for selecting the most appropriate system of the models. SKIC is compared the performance of selection the order of the model, relative to SAIC proposed by Keerativibool (2009). The results of simulation study show that the proposed transformation matrix **P** can transform the MA(1) errors for both forms of (4.3) and (4.4) to be independent. For all situations of the sample sizes; small (T = 15), medium (T = 30), and large (T = 100), including two series of errors generated in the SEM, SKIC convincingly outperformed SAIC, because SAIC has a tendency to overfit the order of the model than SKIC. The results of the observed L_2 efficiency also confirm that SKIC has a large observed L_2 efficiency and small standard deviation of the observed L_2 efficiency than SAIC, then SKIC is likely better than SAIC. The average and standard deviation of SAIC and SKIC for 1,000 transformed samples show that SAIC presents a large negative bias than SKIC, which maybe the main reason of selecting the correct order of the model from SAIC is less than SKIC.

5.2 Future Works

Nowadays, there is not much the criterion to select the appropriate SEM. Therefore, it should be studied and established the other criteria. Including, other schema of the error-generation might also be considered, such as the autoregressive and moving average (ARMA) scheme instead of only the moving average (MA) scheme.

BIBLIOGRAPHY

Akaike, H. 1973. Information Theory and an Extension of the Maximum Likelihood Principle. In **2nd International Symposium on Information Theory.** B.N. Petrov and F. Csaki, eds. Akademiai Kiado, Budapest. Pp. 267 – 281.

Akaike, H. 1974. A New Look at the Statistical Model Identification. **IEEE Transactions on Automatic Control.** 19: 716 – 723.

Anderson, T. W. 2003. **An Introduction to Multivariate Statistical Analysis.** 3rd ed. Hoboken, New Jersey: Wiley.

Bab-Hadiashar, A. and Gheissari, N. 2006. Range Image Segmentation Using Surface Selection Criterion. **IEEE Transactions on Image Processing.** 15: 2006 – 2018.

Bedrick, E. J. and Tsai, C. L. 1994. Model Selection for Multivariate Regression in Small Samples. **Biometrics.** 50: 226 – 231.

Box, G. E. P.; Jenkins, G. M. and Reinsel, G. C. 1994. **Time Series Analysis: Forecasting and Control.** 3rd ed. Englewood Cliffs, New Jersey: Prentice Hall.

Bozdogan, H. and Bearse, P. 2003. Information Complexity Criteria for Detecting Influential Observations in Dynamic Multivariate Linear Models Using the Genetic Algorithm. **Journal of Statistical Planning and Inference.** 114: 31 – 44.

Cavanaugh, J. E. 1999. A Large-Sample Model Selection Criterion Based on Kullback's Symmetric Divergence. **Statistics & Probability Letters.** 42: 333 – 343.

Cavanaugh, J. E. 2004. Criteria for Linear Model Selection Based on Kullback's Symmetric Divergence. **Australian & New Zealand Journal of Statistics.** 46: 257 – 274.

Cavanaugh, J. E. 2010. Lecture I: Introductory Principles, Concepts, and Procedures. **Course Notes 171:290 Advanced Biostatistics Seminar: Model Selection.** Retrieved February 17, 2011 from http://myweb.uiowa.edu/cavaaugh/ms_lec_1_ho.pdf

Choudhury A. H. and Power S. 1995. A New Approximate GLS Estimator for the Linear Regression Model with ARMA(p,q) Disturbances. **Economics Letters.** 48: 119 – 127

Cochrane, D. and Orcutt, G. 1949. Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms. **Journal of the American Statistical Association.** 44: 32 – 61.

Fujikoshi, Y. and Satoh, K. 1997. Modified AIC and C_p in Multivariate Linear Regression. **Biometrika.** 84: 707 – 716.

Galbraith, J. W. and Zinde-Walsh, V. 1995. Transforming the Error-Components Model for Estimation with General ARMA Disturbances. **Journal of Econometrics.** 66: 349 – 355.

Gheissari, N. and Bab-Hadiashar, A. 2008. A Comparative Study of Model Selection Criteria for Computer Vision Applications. **Image and Vision Computing.** 26: 1636 – 1649.

Giombini, G. and Szroeter, J. 2007. Quasi Akaike and Quasi Schwarz Criteria for Model Selection: A Surprising Consistency Result. **Economics Letters.** 95: 259 – 266.

Golan, A. 2001. A Simultaneous Estimation and Variable Selection Rule. **Journal of Econometrics.** 101: 165 – 193.

Greene, W. 2008. **Econometric Analysis.** 6th ed. Upper Saddle River, New Jersey: Prentice Hall.

Gujarati, D. N. 2006. **Essentials of Econometrics.** 3rd ed. Singapore: McGraw – Hill.

Hafidi, B. and Mkhadri, A. 2006. A Corrected Akaike Criterion Based on Kullback's Symmetric Divergence: Applications in Time Series, Multiple and Multivariate Regression. **Computational Statistics & Data Analysis.** 50: 1524 – 1550.

Hafidi, B. 2006. A Small-Sample Criterion Based on Kullback's Symmetric Divergence for Vector Autoregressive Modeling. **Statistics & Probability Letters.** 76: 1647 – 1654.

Hafidi, B. and Mkhadri, A. 2010. The Kullback Information Criterion for Mixture Regression Models. **Statistics & Probability Letters.** 80: 807 – 815.

Hannan, E. J. and Quinn, B. G. 1979. The Determination of the Order of an Autoregression. **Journal of the Royal Statistical Society. Series B. Statistical Methodology.** 41: 190 – 195.

Hurvich, C. M. and Tsai, C. L. 1989. Regression and Time Series Model Selection in Small Samples. **Biometrika.** 76: 297 – 307.

Hwang, S. Y.; Kim, S.; Lee, S. D. and Basawa, I.V. 2007. Generalized Least Squares Estimation for Explosive AR(1) Processes with Conditionally Heteroscedastic Errors. **Statistics & Probability Letters.** 77: 1439 – 1448.

Keerativibool, W. 2009. Selection of a System of Simultaneous Equations Model. **Doctoral Dissertation, School of Applied Statistics, National Institute of Development Administration.**

Keerativibool, W.; Jitthavech, J. and Lorchirachoonkul, V. 2009a. Autocorrelation Correction in a Regression Model. In **Proceedings Applied Statistics Conference 2009 (ASCONF2009).** Pattaya, Chon Buri, Thailand. Pp. 519 – 533.

Keerativibool, W.; Jitthavech, J. and Lorchirachoonkul, V. 2009b. Autocorrelation Correction in a Simultaneous Equations Model. In **Proceedings the 5th International Conference on Mathematics, Statistics and their Applications (ICMSA2009).** West Sumatra, Indonesia. Pp. 192 – 197.

Keerativibool, W. 2010. Moving Average Correction in a Regression Model. **Thailand Statistician: Journal of Thai Statistical Association.** 8: 63 – 80.

Keerativibool, W.; Jitthavech, J. and Lorchirachoonkul, V. 2011. Model Selection in a System of Simultaneous Equations Model. **Communications in Statistics – Theory and Methods.** 40: 373 – 393.

Keerativibool, W. 2011a. Generalized Least Squares Transformation with the Second-order Autoregressive Error. **Thailand Statistician: Journal of Thai Statistical Association.** 9: 77 – 92.

Keerativibool, W. 2011b. The Advantage of Using Model Selection Criterion over the Multicollinearity Problem. In **Proceedings the 12th National Statistics and Applied Statistics Conference (StatConf2011).** Hat Yai, Songkhla, Thailand. Pp. 91 – 96.

Keerativibool, W. 2011c. Simultaneous Equations Model Selection Using Variants of the Akaike Information Criterion. In **Proceedings the 14th International Conference on Applied Stochastic Models and Data Analysis (ASMDA2011).** Rome, Italy. Pp. 669 – 676.

Keerativibool, W. 2012. New Criteria for Selection in Simultaneous Equations Model. **Thailand Statistician: Journal of Thai Statistical Association.** 10: 163 – 181.

Kim, H. J. and Cavanaugh, J. E. 2005. Model Selection Criteria Based on Kullback Information Measures for Nonlinear Regression. **Journal of Statistical Planning and Inference.** 134: 332 – 349.

Marazzi, A. and Yohai, V. J. 2006. Robust Box-Cox Transformations Based on Minimum Residual Autocorrelation. **Computational Statistics & Data Analysis.** 50: 2752 – 2768.

McQuarrie, A. D. R.; Shumway, R. and Tsai, C. L. 1997. The Model Selection Criterion AICu. **Statistics & Probability Letters.** 34: 285 – 292.

McQuarrie, A. D. R. and Tsai, C. L. 1998. **Regression and Time Series Model Selection.** Singapore: World Scientific.

McQuarrie, A. D. R. 1999. A Small-Sample Correction for the Schwarz SIC Model Selection Criterion. **Statistics & Probability Letters.** 44: 79 – 86.

Naik, P. A.; Shi, P. and Tsai, C. L. 2007. Extending the Akaike Information Criterion to Mixture Regression Models. **Journal of the American Statistical Association.** 102: 244 – 254.

Neath, A. and Cavanaugh, J. E. 1997. Regression and Time Series Model Selection Using Variants of the Schwarz Information Criterion. **Communications in Statistics-Theory and Methods.** 26: 559 – 580.

Prais, S. and Winsten, C. B. 1954. Trend Estimators and Serial Correlation. **Cowles Commission Discussion Paper.** 383: 1 – 26.

Schwarz, G. 1978. Estimating the Dimension of a Model. **The Annals of Statistics.** 6: 461 – 464.

Seghouane, A. K. and Bekara, M. 2004. A Small Sample Model Selection Criterion Based on Kullback's Symmetric Divergence. **IEEE Transactions on Signal Processing.** 52: 3314 – 3323.

Seghouane, A. K.; Bekara, M. and Fleury, G. 2005. A Criterion for Model Selection in the Presence of Incomplete Data Based on Kullback's Symmetric Divergence. **Signal Processing.** 85: 1405 – 1417.

Seghouane, A. K. 2006. Multivariate Regression Model Selection from Small Samples Using Kullback's Symmetric Divergence. **Signal Processing.** 86: 2074 – 2084.

Seghouane, A. K. and Cichocki, A. 2007. Bayesian Estimation of the Number of Principal Components. **Signal Processing.** 87: 562 – 568.

Sei, T. and Komaki, F. 2007. Bayesian Prediction and Model Selection for Locally Asymptotically Mixed Normal Models. **Journal of Statistical Planning and Inference.** 137: 2523 – 2534.

Ullah, A.; Srivastava, V. K.; Magee, L. and Srivastava, A. 1983. Estimation of Linear Regression Model with Autocorrelated Disturbances. **Journal of Time Series Analysis.** 4: 127 – 135.

Vougas, D. V. 2008. Generalized Least Squares Transformation and Estimation with Autoregressive Error. **Statistics & Probability Letters.** 78: 402 – 404.

APPENDICES

APPENDIX A
PROOFS

Proof of Theorem 1.

The reduced-form model in (1.5) at the t^{th} observation and the j^{th} equation can be written as follows:

$$y_{tj} = \mathbf{x}_t'\boldsymbol{\pi}_j + v_{tj}, \ t = 1, 2, \ldots, T, \ j = 1, 2, \ldots, M, \tag{A1.1}$$

where

$$\mathbf{x}_t' = \begin{bmatrix} x_{t1} & x_{t2} & \cdots & x_{tK} \end{bmatrix}, \ v_{tj} = \varepsilon_{tj} - \theta_j\varepsilon_{t-1,j}, \ t = 2, 3, \ldots, T, \ j = 1, 2, \ldots, M. \tag{A1.2}$$

Replacing v_{tj} in (A1.2) into (A1.1) and rearrange it into the term of ε_{tj},

$$\varepsilon_{tj} = y_{tj} - \mathbf{x}_t'\boldsymbol{\pi}_j + \theta_j\varepsilon_{t-1,j}, \ t = 2, 3, \ldots, T, \ j = 1, 2, \ldots, M. \tag{A1.3}$$

The i^{th} lag of ε_{tj} in (A1.3) can be written as

$$\varepsilon_{t-i,j} = y_{t-i,j} - \mathbf{x}_{t-i}'\boldsymbol{\pi}_j + \theta_j\varepsilon_{t-(i+1),j}. \tag{A1.4}$$

Using the knowledge of (A1.4), the equation in (A1.1) becomes

$$\begin{aligned}
y_{tj} &= \mathbf{x}_t'\boldsymbol{\pi}_j + \varepsilon_{tj} - \theta_j\left(y_{t-1,j} - \mathbf{x}_{t-1}'\boldsymbol{\pi}_j + \theta_j\varepsilon_{t-2,j}\right) \\
y_{tj} + \theta_j y_{t-1,j} &= \left(\mathbf{x}_t' + \theta_j\mathbf{x}_{t-1}'\right)\boldsymbol{\pi}_j + \varepsilon_{tj} - \theta_j^2\varepsilon_{t-2,j} \\
&= \left(\mathbf{x}_t' + \theta_j\mathbf{x}_{t-1}'\right)\boldsymbol{\pi}_j + \varepsilon_{tj} - \theta_j^2\left(y_{t-2,j} - \mathbf{x}_{t-2}'\boldsymbol{\pi}_j + \theta_j\varepsilon_{t-3,j}\right) \\
y_{tj} + \theta_j y_{t-1,j} + \theta_j^2 y_{t-2,j} &= \left(\mathbf{x}_t' + \theta_j\mathbf{x}_{t-1}' + \theta_j^2\mathbf{x}_{t-2}'\right)\boldsymbol{\pi}_j + \varepsilon_{tj} - \theta_j^3\varepsilon_{t-3,j} \\
&\vdots \\
\sum_{i=0}^{T}\theta_j^i y_{t-i,j} &= \sum_{i=0}^{T}\theta_j^i\mathbf{x}_{t-i}'\boldsymbol{\pi}_j + \varepsilon_{tj} - \theta_j^{T+1}\varepsilon_{t-(T+1),j}.
\end{aligned} \tag{A1.5}$$

As T becomes large and θ_j satisfies the invertibility condition, the value of θ_j^{T+1} in (A1.5) approach zero. Therefore, (A1.5) can be rewritten as

$$y_{tj}^* = \mathbf{x}_t^{*\prime}\boldsymbol{\pi}_j + \varepsilon_{tj}, \tag{A1.6}$$

where $y_{tj}^* = \sum_{i=0}^{T}\theta_j^i y_{t-i,j}$ and $\mathbf{x}_t^{*\prime} = \sum_{i=0}^{T}\theta_j^i\mathbf{x}_{t-i}'$ for $t = 2, 3, \ldots, T, \ j = 1, 2, \ldots, M.$

From (A1.6) we found that $\mathrm{Var}\left(y_{tj}^{*} \mid \mathbf{x}_{t}^{*}\right)=\mathrm{Var}\left(\varepsilon_{tj}\right)=\sigma_{jj}$, then we can argue that the MA(1) problem at $t=2, 3, \ldots, T$ and $j=1, 2, \ldots, M$ has been corrected. However, the transformation in (A1.6) does not include the first observation in (A1.1). The heteroskedasticity remains unsolved unless the first observation is eliminated, but if the first observation is included in the analysis, the transformation must be extended by the following steps. Firstly, we take the expectation to v_{tj} in (A1.2),

$$E\left(v_{tj}\right)=E\left(\varepsilon_{tj}\right)-\theta_{j}E\left(\varepsilon_{t-1,j}\right)=E\left(\varepsilon_{tj}\right)-\theta_{j}E\left(\varepsilon_{tj}\right)=\left(1-\theta_{j}\right)E\left(\varepsilon_{tj}\right).$$

Using the assumption in (1.8), we have the expectation of v_{tj} is equal to zero. Therefore, from (A1.1) the variance of y_{tj} given $\mathbf{x}_{t}$ for $t=1, 2, \ldots, T$ and $j=1, 2, \ldots, M$ can be written as

$$\mathrm{Var}\left(v_{tj}\right)=E\left[\left(\varepsilon_{tj}-\theta_{j}\varepsilon_{t-1,j}\right)^{2}\right]=E\left(\varepsilon_{tj}^{2}\right)+\theta_{j}^{2}E\left(\varepsilon_{tj}^{2}\right)=\left(1+\theta_{j}^{2}\right)E\left(\varepsilon_{tj}^{2}\right)=\left(1+\theta_{j}^{2}\right)\sigma_{jj}.$$

Hence, the first observation should weighted by $\sqrt{\dfrac{1}{1+\theta_{j}^{2}}}$, yields the model

$$y_{1j}^{*}=\mathbf{x}_{1}^{*\prime}\boldsymbol{\pi}_{j}+\varepsilon_{1j}, \tag{A1.7}$$

where $y_{1j}^{*}=\sqrt{\dfrac{1}{1+\theta_{j}^{2}}}\,y_{1j}$ and $\mathbf{x}_{1}^{*\prime}=\sqrt{\dfrac{1}{1+\theta_{j}^{2}}}\,\mathbf{x}_{1}^{\prime}$ for $j=1, 2, \ldots, M$.

It can be shown that the MA(1) problem at t = 1 has been corrected,

$$\mathrm{Var}\left(y_{1j}^{*}\left|\mathbf{x}_{1}^{*}\right.\right)=\frac{1}{1+\theta_{j}^{2}}\mathrm{Var}\left(y_{1j}\left|\mathbf{x}_{1}\right.\right)=\frac{1}{1+\theta_{j}^{2}}\cdot\left(1+\theta_{j}^{2}\right)\sigma_{jj}=\sigma_{jj}.$$

Combining the results in (A1.6) and (A1.7), we get the $T\times T$ transformation matrix $\mathbf{P}_{j}$ which was exhibited in (3.2).

Proof of Theorem 2.

The Kullback-Leibler's symmetric divergence is a measure that used to separate the discrepancy between the candidate model in (3.3) and the true model in (3.4), defined by

$$2\mathrm{J}\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}\right)=\mathrm{d}\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}\right)-\mathrm{d}\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0\right)+\mathrm{d}\left(\boldsymbol{\theta},\boldsymbol{\theta}_0\right)-\mathrm{d}\left(\boldsymbol{\theta},\boldsymbol{\theta}\right), \tag{A2.1}$$

where $\mathrm{d}\left(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j\right)=\mathrm{E}_{\boldsymbol{\theta}_i}\left\{-2\log\mathrm{L}\left(\boldsymbol{\theta}_j\,|\,\mathbf{y}^*\right)\right\}$.

Dropping $\mathrm{d}\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0\right)$ in (A2.1) since it does not depend on $\boldsymbol{\theta}$. The ranking of the candidate models according to $2\mathrm{J}\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}\right)$ in (A2.1) is then identical to ranking them according to

$$\mathrm{K}\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}\right)=\mathrm{d}\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}\right)+\mathrm{d}\left(\boldsymbol{\theta},\boldsymbol{\theta}_0\right)-\mathrm{d}\left(\boldsymbol{\theta},\boldsymbol{\theta}\right). \tag{A2.2}$$

Given a set of GLS estimators $\hat{\boldsymbol{\theta}}=\left(\hat{\boldsymbol{\pi}},\,\hat{\boldsymbol{\Sigma}},\,\hat{\mathbf{P}}\right)$ where $\hat{\mathbf{P}}$ is the estimate of the transformation matrix $\mathbf{P}$ in (3.1),

$$\hat{\boldsymbol{\pi}}=\left[\tilde{\mathbf{X}}^{*\prime}\left(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_{\mathrm{T}}\right)\tilde{\mathbf{X}}^*\right]^{-1}\tilde{\mathbf{X}}^{*\prime}\left(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_{\mathrm{T}}\right)\mathbf{y}^*,$$

and

$$\hat{\boldsymbol{\Sigma}}\otimes\mathbf{I}_{\mathrm{T}}=\frac{1}{\mathrm{T}}\left(\mathbf{y}^*-\tilde{\mathbf{X}}^*\boldsymbol{\pi}\right)\left(\mathbf{y}^*-\tilde{\mathbf{X}}^*\boldsymbol{\pi}\right)',$$

we have therefore the estimate of the symmetric measure in (A2.2) as

$$\mathrm{K}\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)=\mathrm{d}\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)+\mathrm{d}\left(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0\right)-\mathrm{d}\left(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}\right), \tag{A2.3}$$

where $\mathrm{d}\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)=\mathrm{E}_{\boldsymbol{\theta}_0}\left\{-2\log\mathrm{L}\left(\boldsymbol{\theta}\,|\,\mathbf{y}^*\right)\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$, $\mathrm{d}\left(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0\right)=\mathrm{E}_{\boldsymbol{\theta}}\left\{-2\log\mathrm{L}\left(\boldsymbol{\theta}_0\,|\,\mathbf{y}^*\right)\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$, and $\mathrm{d}\left(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}\right)=\mathrm{E}_{\boldsymbol{\theta}}\left\{-2\log\mathrm{L}\left(\boldsymbol{\theta}\,|\,\mathbf{y}^*\right)\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$.

From the minus twice log likelihood of the candidate model in (3.3),

$$-2\log\mathrm{L}\left(\boldsymbol{\theta}\,|\,\mathbf{y}^*\right)=\mathrm{TM}\log(2\pi)+\mathrm{T}\log|\,\boldsymbol{\Sigma}\,|+\left(\mathbf{y}^*-\tilde{\mathbf{X}}^*\boldsymbol{\pi}\right)'\left(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_{\mathrm{T}}\right)\left(\mathbf{y}^*-\tilde{\mathbf{X}}^*\boldsymbol{\pi}\right),$$

we have each term of the estimated symmetric measure in (A2.3) as follows:

$$\begin{aligned}\mathrm{d}\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right) &=\mathrm{TM}\log(2\pi)+\mathrm{T}\log\left|\,\hat{\boldsymbol{\Sigma}}\,\right|+\left(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}}\right)'\tilde{\mathbf{X}}^{*\prime}\left(\hat{\boldsymbol{\Sigma}}^{-1}\otimes\mathbf{I}_{\mathrm{T}}\right)\tilde{\mathbf{X}}^*\left(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}}\right)\\ &\quad+\mathrm{T}\,\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}_0\right),\end{aligned}$$

$$
\begin{aligned}
\mathrm{d}\left(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0\right) &= \mathrm{TM}\log(2\pi)+\mathrm{T}\log|\boldsymbol{\Sigma}_0|+(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0)'\,\tilde{\mathbf{X}}^{*\prime}\left(\boldsymbol{\Sigma}_0^{-1}\otimes\mathbf{I}_\mathrm{T}\right)\tilde{\mathbf{X}}^{*}(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0)\\
&\quad +\mathrm{T}\,\mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{\Sigma}}\right),\\
\mathrm{d}\left(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}\right) &= \mathrm{TM}\log(2\pi)+\mathrm{T}\log\left|\hat{\boldsymbol{\Sigma}}\right|+\mathrm{TM}.
\end{aligned}
$$

Therefore, the expected of the estimated symmetric measure in (A2.3) becomes

$$
\begin{aligned}
\Omega(\boldsymbol{\theta}_0,\mathrm{K}) &= \mathrm{E}_{\boldsymbol{\theta}_0}\left\{\mathrm{K}\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)\right\}=\mathrm{E}_{\boldsymbol{\theta}_0}\left\{\mathrm{d}\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)+\mathrm{d}\left(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0\right)-\mathrm{d}\left(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}\right)\right\}\\
&= \mathrm{TM}\left[\log(2\pi)+1\right]+\mathrm{E}_{\boldsymbol{\theta}_0}\left\{\mathrm{T}\log\left|\hat{\boldsymbol{\Sigma}}\right|\right\}\\
&\quad +\mathrm{E}_{\boldsymbol{\theta}_0}\left\{(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}})'\,\tilde{\mathbf{X}}^{*\prime}\left(\hat{\boldsymbol{\Sigma}}^{-1}\otimes\mathbf{I}_\mathrm{T}\right)\tilde{\mathbf{X}}^{*}(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}})\right\}\\
&\quad +\mathrm{E}_{\boldsymbol{\theta}_0}\left\{\mathrm{T}\,\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}_0\right)\right\}+\mathrm{E}_{\boldsymbol{\theta}_0}\left\{(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0)'\,\tilde{\mathbf{X}}^{*\prime}\left(\boldsymbol{\Sigma}_0^{-1}\otimes\mathbf{I}_\mathrm{T}\right)\tilde{\mathbf{X}}^{*}(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0)\right\}\\
&\quad +\mathrm{E}_{\boldsymbol{\theta}_0}\left\{\mathrm{T}\,\mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{\Sigma}}\right)\right\}-\mathrm{E}_{\boldsymbol{\theta}_0}\left\{\mathrm{T}\log\left(\left|\hat{\boldsymbol{\Sigma}}\right|/|\boldsymbol{\Sigma}_0|\right)\right\}-2\mathrm{TM}. \qquad \text{(A2.4)}
\end{aligned}
$$

From the facts that, $\hat{\boldsymbol{\pi}}$ and $\mathrm{T}\hat{\boldsymbol{\Sigma}}$ are asymptotically independent where $\hat{\boldsymbol{\pi}}$ is asymptotically distributed as a Gaussian distribution with mean vector $\boldsymbol{\pi}$ and variance-covariance matrix $\left[\tilde{\mathbf{X}}^{*\prime}\left(\boldsymbol{\Sigma}_0^{-1}\otimes\mathbf{I}_\mathrm{T}\right)\tilde{\mathbf{X}}^{*}\right]^{-1}$, and $\mathrm{T}\hat{\boldsymbol{\Sigma}}$ is asymptotically distributed as the Wishart distribution with $\mathrm{T}-\mathrm{K}$ degrees of freedom, $\mathrm{W}_{\mathrm{KM}}(\boldsymbol{\Sigma}_0,\mathrm{T}-\mathrm{K})$, then (Anderson, 2003)

$$
\mathrm{E}_{\boldsymbol{\theta}_0}\left\{\mathrm{T}\hat{\boldsymbol{\Sigma}}\right\}=(\mathrm{T}-\mathrm{K})\boldsymbol{\Sigma}_0 \text{ and } \mathrm{E}_{\boldsymbol{\theta}_0}\left\{\hat{\boldsymbol{\Sigma}}^{-1}\right\}=\frac{\mathrm{T}}{\mathrm{T}-\mathrm{K}-\mathrm{M}-1}\boldsymbol{\Sigma}_0^{-1}.
$$

Using the above results, we have

$$
\begin{aligned}
&\mathrm{E}_{\boldsymbol{\theta}_0}\left\{\mathrm{T}\,\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}_0\right)\right\}=\mathrm{T}\,\mathrm{tr}\left\{\mathrm{E}_{\boldsymbol{\theta}_0}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)\boldsymbol{\Sigma}_0\right\}=\mathrm{T}\,\mathrm{tr}\left\{\frac{\mathrm{T}}{\mathrm{T}-\mathrm{K}-\mathrm{M}-1}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_0\right\}=\frac{\mathrm{T}^2\mathrm{M}}{\mathrm{T}-\mathrm{K}-\mathrm{M}-1},\\
&\mathrm{E}_{\boldsymbol{\theta}_0}\left\{\mathrm{T}\,\mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{\Sigma}}\right)\right\}=\mathrm{tr}\left\{\boldsymbol{\Sigma}_0^{-1}\mathrm{E}_{\boldsymbol{\theta}_0}\left(\mathrm{T}\hat{\boldsymbol{\Sigma}}\right)\right\}=\mathrm{tr}\left\{\boldsymbol{\Sigma}_0^{-1}(\mathrm{T}-\mathrm{K})\boldsymbol{\Sigma}_0\right\}=(\mathrm{T}-\mathrm{K})\mathrm{M},\\
&\mathrm{E}_{\boldsymbol{\theta}_0}\left\{(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}})'\,\tilde{\mathbf{X}}^{*\prime}\left(\hat{\boldsymbol{\Sigma}}^{-1}\otimes\mathbf{I}_\mathrm{T}\right)\tilde{\mathbf{X}}^{*}(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}})\right\}\\
&\quad =\mathrm{E}_{\boldsymbol{\theta}_0}\left\{\mathrm{tr}\left[\left(\hat{\boldsymbol{\Sigma}}^{-1}\otimes\mathbf{I}_\mathrm{T}\right)\tilde{\mathbf{X}}^{*}(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}})(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}})'\,\tilde{\mathbf{X}}^{*\prime}\right]\right\}
\end{aligned}
$$

$$= \operatorname{tr}\left\{ \left[\mathrm{E}_{\boldsymbol{\theta}_0}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right) \otimes \mathbf{I}_{\mathrm{T}} \right] \mathrm{E}_{\boldsymbol{\theta}_0}\left[\tilde{\mathbf{X}}^{*}\left(\boldsymbol{\pi}_0 - \hat{\boldsymbol{\pi}}\right)\left(\boldsymbol{\pi}_0 - \hat{\boldsymbol{\pi}}\right)' \tilde{\mathbf{X}}^{*\prime} \right] \right\}$$

$$= \frac{\mathrm{T}}{\mathrm{T}-\mathrm{K}-\mathrm{M}-1} \operatorname{tr}\left\{ \left(\boldsymbol{\pi}_0 - \hat{\boldsymbol{\pi}}\right)' \tilde{\mathbf{X}}^{*\prime} \left(\boldsymbol{\Sigma}_0^{-1} \otimes \mathbf{I}_{\mathrm{T}}\right) \tilde{\mathbf{X}}^{*} \left(\boldsymbol{\pi}_0 - \hat{\boldsymbol{\pi}}\right) \right\} = \frac{\mathrm{TKM}}{\mathrm{T}-\mathrm{K}-\mathrm{M}-1},$$

$$\mathrm{E}_{\boldsymbol{\theta}_0}\left\{ \left(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0\right)' \tilde{\mathbf{X}}^{*\prime} \left(\boldsymbol{\Sigma}_0^{-1} \otimes \mathbf{I}_{\mathrm{T}}\right) \tilde{\mathbf{X}}^{*} \left(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0\right) \right\} = \mathrm{KM},$$

then $\Omega(\boldsymbol{\theta}_0, \mathrm{K})$ in (A2.4) can be written as

$$\begin{aligned} \Omega(\boldsymbol{\theta}_0, \mathrm{K}) &= \mathrm{TM}\left[\log(2\pi)+1\right] + \mathrm{E}_{\boldsymbol{\theta}_0}\left\{ \mathrm{T}\log\left|\hat{\boldsymbol{\Sigma}}\right| \right\} + \frac{\mathrm{TKM}}{\mathrm{T}-\mathrm{K}-\mathrm{M}-1} + \frac{\mathrm{T}^2\mathrm{M}}{\mathrm{T}-\mathrm{K}-\mathrm{M}-1} \\ &\quad + \mathrm{KM} + (\mathrm{T}-\mathrm{K})\mathrm{M} - \mathrm{E}_{\boldsymbol{\theta}_0}\left\{ \mathrm{T}\log\left(\left|\hat{\boldsymbol{\Sigma}}\right| / \left|\boldsymbol{\Sigma}_0\right| \right) \right\} - 2\mathrm{TM} \\ &= \mathrm{TM}\left[\log(2\pi)+1\right] + \mathrm{E}_{\boldsymbol{\theta}_0}\left\{ \mathrm{T}\log\left|\hat{\boldsymbol{\Sigma}}\right| \right\} + \frac{\mathrm{TM}(2\mathrm{K}+\mathrm{M}+1)}{\mathrm{T}-\mathrm{K}-\mathrm{M}-1} \\ &\quad - \mathrm{T}\,\mathrm{E}_{\boldsymbol{\theta}_0}\left\{ \log\left(\left|\mathrm{T}\hat{\boldsymbol{\Sigma}}\right| / \left|\boldsymbol{\Sigma}_0\right| \right) \right\} + \mathrm{TM}\log\mathrm{T}. \end{aligned} \qquad \text{(A2.5)}$$

Because $\left|\mathrm{T}\hat{\boldsymbol{\Sigma}}\right| / \left|\boldsymbol{\Sigma}_0\right|$ in (A2.5) is the distribution of a product of independent χ^2 random variables, $\prod_{i=1}^{M} \chi^2_{T-K-M+i}$, then we have

$$\log\left(\left|\mathrm{T}\hat{\boldsymbol{\Sigma}}\right| / \left|\boldsymbol{\Sigma}_0\right| \right) \sim \sum_{i=1}^{M} \log \chi^2_{T-K-M+i}.$$

Using the second-order of Taylor's series expansions to expand the function of $\log\left(\chi^2_p\right)$ about the mean p, we have

$$\log\left(\chi^2_p\right) \doteq \log(p) + \frac{1}{p}\left(\chi^2_p - p\right) - \frac{1}{2p^2}\left(\chi^2_p - p\right)^2 \text{ and } \mathrm{E}\left[\log\left(\chi^2_p\right)\right] \doteq \log(p) - \frac{1}{p}.$$

Then, the last two terms of the right-hand side in (A2.5) is

$$\begin{aligned} &-\mathrm{TE}_{\boldsymbol{\theta}_0}\left\{ \log\left(\left|\mathrm{T}\hat{\boldsymbol{\Sigma}}\right| / \left|\boldsymbol{\Sigma}_0\right| \right) \right\} + \mathrm{TM}\log\mathrm{T} \\ &\doteq -\mathrm{T}\sum_{i=1}^{M}\left[\log(\mathrm{T}-\mathrm{K}-\mathrm{M}+\mathrm{i}) - \frac{1}{\mathrm{T}-\mathrm{K}-\mathrm{M}+\mathrm{i}} \right] + \mathrm{TM}\log\mathrm{T}. \end{aligned} \qquad \text{(A2.6)}$$

McQuarrie and Tsai (1998) gave the simplification formulae for any T, K, M and assume $\mathrm{T}-\mathrm{K}-\mathrm{M}$ is much larger than M as follows:

$$\sum_{i=1}^{M} \log(\mathrm{T}-\mathrm{K}-\mathrm{M}+\mathrm{i}) = \mathrm{M}\log\left(\mathrm{T}-\mathrm{K}-\frac{\mathrm{M}-1}{2} \right) = \mathrm{M}\log\left(\frac{2\mathrm{T}-2\mathrm{K}-\mathrm{M}+1}{2} \right), \qquad \text{(A2.7)}$$

and

$$\sum_{i=1}^{M}\frac{1}{T-K-M+i}\doteq\frac{M}{T-K-\frac{M-1}{2}}=\frac{2M}{2T-2K-M+1}. \qquad \text{(A2.8)}$$

Replacing the results in (A2.7) and (A2.8) into (A2.6), we have

$$-TE_{\boldsymbol{\theta}_0}\left\{\log\left(\left|T\hat{\boldsymbol{\Sigma}}\right|/\left|\boldsymbol{\Sigma}_0\right|\right)\right\}+TM\log T$$

$$\doteq -TM\log\left(\frac{2T-2K-M+1}{2}\right)+\frac{2TM}{2T-2K-M+1}+TM\log T$$

$$=TM\log\left(\frac{2T}{2T-2K-M+1}\right)+\frac{2TM}{2T-2K-M+1}. \qquad \text{(A2.9)}$$

Replacing the results in (A2.9) into (A2.5), we have

$$\Omega(\boldsymbol{\theta}_0,K)\doteq TM\left[\log(2\pi)+1\right]+E_{\boldsymbol{\theta}_0}\{\text{SKIC}\},$$

where SKIC was exhibited in (3.5).

APPENDIX B
OUTPUTS OF THIS RESEARCH

Submitted paper

Author(s)	Title	Journal	Vol	Page	Year	Data Base	Impact factor/year
Warangkhana Keerativibool	Unifying the derivations of Kullback information criterion and corrected versions	Thailand Statistician: Journal of Thai Statistical Association				MathScinet	-
Warangkhana Keerativibool	Study on the penalty functions of model selection criteria	Thailand Statistician: Journal of Thai Statistical Association				MathScinet	-
Warangkhana Keerativibool and Jirawan Jitthavech	Model selection criterion based on Kullback-Leibler's symmetric divergence for simultaneous equations model	Statistics and Probability Letters				ISI	0.531/2012

Unifying the derivations of Kullback information criterion and corrected versions

Warangkhana Keerativibool

Department of Mathematics and Statistics, Faculty of Science,
Thaksin University, Phatthalung, Thailand.

E-Mail Address: warang27@gmail.com

Abstract

The Kullback information criterion (KIC) was proposed by Cavanaugh (1999) to serve as an asymptotically unbiased estimator of a variant of Kullback's symmetric divergence between the true and fitted candidate models. It was arguably more sensitive than the criterion based on the directed divergence. However, for a small sample size or if the dimension of candidate model is large relative to the sample size, it displayed a large negative bias. Many authors, Cavanaugh (2004), Seghouane and Bekara (2004), Hafidi and Mkhadri (2006), proposed the criteria to correct this bias, i.e., the corrected versions of KIC called, respectively, in this paper $\mathrm{KICc}_{\mathrm{C}}$, $\mathrm{KICc}_{\mathrm{SB}}$, and $\mathrm{KICc}_{\mathrm{HM}}$. Because they have multiple formulas, the aims of this paper are to unify and examine the performance of them relative to the AIC family of criteria, using theoretical and extensive simulation study methods.

Keywords: KIC; KICc; Kullback's directed divergence; Kullback's symmetric divergence; model selection.

1. Introduction

The Kulback information criterion (KIC) by Cavanaugh (1999) and the corrected versions (KICc) by Cavanaugh (2004) called $\mathrm{KICc}_{\mathrm{C}}$, by Seghouane and Bekara (2004) called $\mathrm{KICc}_{\mathrm{SB}}$, and by Hafidi and Mkhadri (2006) called $\mathrm{KICc}_{\mathrm{HM}}$ were designed based on Kullback's symmetric divergence, also known as the J-divergence, in order to assess the dissimilarity between the model generating the data and a fitted candidate model. However, when the dimension of candidate model increases compared to the sample size, the corrected version of the model selection criterion was better than the original version because it produced a bias reduction and strongly improved model selection (Hurvich and Tsai, 1989; Bedrick and Tsai, 1994; Cavanaugh, 1997, 2004; McQuarrie, 1999; Seghouane and Bekara, 2004; Hafidi, 2006; Hafidi and Mkhadri, 2006). Although KIC, $\mathrm{KICc}_{\mathrm{C}}$, $\mathrm{KICc}_{\mathrm{SB}}$, and $\mathrm{KICc}_{\mathrm{HM}}$ share the same fundamental objective, the justifications of the criteria proceed along different directions, making it difficult to reconcile how the different corrected versions of KIC refine the approximations used to establish KIC in the setting of linear regression model. With this motivation, the aims of this paper are to unify the derivations of KIC and the corrected versions in order to link the justifications of these criteria and the performance of them is then examined by the extensive simulation study. The remainder of this paper is organized as follows. In Section 2, we review the model selection criteria based on Kullback's directed and symmetric divergences. In Section 3, we show the unifications for the derivations of KIC and the corrected versions. Simulation study for 1,000 realizations of multiple regression models to examine the performance of the AIC and KIC families of criteria is shown in Section 4. Finally, Section 5 is the conclusions, discussion, and further study.

2. A review of model selection criteria based on Kullback's directed and symmetric divergences

Suppose that the true and the candidate models are, respectively, given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_0,\ \boldsymbol{\varepsilon}_0 \sim \mathrm{N}_n\left(\mathbf{0},\ \sigma_0^2\mathbf{I}_n\right), \tag{1}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathrm{N}_n\left(\mathbf{0},\ \sigma^2\mathbf{I}_n\right), \tag{2}$$

where $\mathbf{y}$ is an $n\times 1$ dependent random vector of observations, $\mathbf{X}$ is an $n\times p$ matrix of independent variables with full-column rank, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$ are $p\times 1$ parameter vectors of regression coefficients, $\boldsymbol{\varepsilon}_0$ and $\boldsymbol{\varepsilon}$ are $n\times 1$ noise vectors. The true model is assumed to be correctly specified or overfitted by all the candidate models. This means that $\boldsymbol{\beta}_0$ has p_0 nonzero entries with $0 < p_0 \le p$ and the rest of the $\left(p - p_0\right)$ entries are equal to zero. The $\left(p+1\right)\times 1$ vector of parameters is $\boldsymbol{\theta}_0 = \left[\boldsymbol{\beta}_0'\ \ \sigma_0^2\right]'$ and the maximum likelihood estimator of $\boldsymbol{\theta}_0$ is $\hat{\boldsymbol{\theta}} = \left[\hat{\boldsymbol{\beta}}'\ \ \hat{\sigma}^2\right]'$ where $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$ and $\hat{\sigma}^2 = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)/n$. (3)

The minus twice log likelihood of the candidate model in (2) when replacing the dependent vector $\mathbf{y}$ in (1) is defined by

$$-2\log L\left(\boldsymbol{\theta}\,|\,\mathbf{y}\right) = n\log 2\pi + n\log\sigma^2 + \frac{1}{\sigma^2}\boldsymbol{\varepsilon}_0'\boldsymbol{\varepsilon}_0 + \frac{1}{\sigma^2}\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right)'\mathbf{X}'\mathbf{X}\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right) + \frac{2}{\sigma^2}\boldsymbol{\varepsilon}_0'\mathbf{X}\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right). \tag{4}$$

A well-known measure to separate the discrepancy between two models is given by Kullback's directed divergence or I-divergence (Kullback, 1968),

$$2I\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}\right) = E_{\boldsymbol{\theta}_0}\left\{2\log\frac{L\left(\boldsymbol{\theta}_0\,|\,\mathbf{y}\right)}{L\left(\boldsymbol{\theta}\,|\,\mathbf{y}\right)}\right\} = d\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}\right) - d\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0\right),$$

where

$$d\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}\right) = E_{\boldsymbol{\theta}_0}\left\{-2\log L\left(\boldsymbol{\theta}\,|\,\mathbf{y}\right)\right\},\ d\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0\right) = E_{\boldsymbol{\theta}_0}\left\{-2\log L\left(\boldsymbol{\theta}_0\,|\,\mathbf{y}\right)\right\},$$

and the expectation $E_{\boldsymbol{\theta}_0}$ is taken with respect to the true model. Because $d\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0\right)$ does not depend on $\boldsymbol{\theta}$, any ranking of the candidate models according to $2I\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}\right)$ would be identical to ranking them according to $d\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}\right)$. Given a set of maximum likelihood estimators $\hat{\boldsymbol{\theta}}$ in (3), the estimated directed measure $d\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}\right)$ is

$$\begin{aligned} d\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right) &= E_{\boldsymbol{\theta}_0}\left\{-2\log L\left(\boldsymbol{\theta}\,|\,\mathbf{y}\right)\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &= n\log 2\pi + n\log\hat{\sigma}^2 + \frac{n\sigma_0^2}{\hat{\sigma}^2} + \frac{1}{\hat{\sigma}^2}\left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{X}\left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\right). \end{aligned} \tag{5}$$

However, the evaluation in (5) is not possible because it requires the knowledge of $\boldsymbol{\theta}_0$, Akaike (1973, 1974) proposed an asymptotically unbiased estimator of

$$\Delta\left(\boldsymbol{\theta}_0, p\right) = E_{\boldsymbol{\theta}_0}\left\{d\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)\right\} \tag{6}$$

as

$$\mathrm{AIC} = n\log\hat{\sigma}^2 + 2\left(p+1\right), \tag{7}$$

i.e., $E_{\boldsymbol{\theta}_0}\left\{\mathrm{AIC}\right\} + o\left(1\right) = \Delta\left(\boldsymbol{\theta}_0, p\right)$.

Because of a large negative bias of AIC when the sample size is small or the dimension of candidate model is large relative to the sample size, Hurvich and Tsai (1989) proposed an exactly unbiased estimator of (6) as

$$\text{AICc} = n\log\hat{\sigma}^2 + \frac{2n(p+1)}{n-p-2}, \tag{8}$$

i.e., $E_{\boldsymbol{\theta}_0}\{\text{AICc}\} = \Delta(\boldsymbol{\theta}_0, p)$.

Cavanaugh (1999), Seghouane and Bekara (2004), Seghouane (2006b) summarized that the directed divergence produced too underfitted value of model selection, and then it tended to be large for overparameterized models. An alternate measure to prevent both overfitting and underfitting problems is obtained by reversing the roles of two models in the definition of the measure, called Kullback's symmetric divergence or J-divergence,

$$2J(\boldsymbol{\theta}_0,\boldsymbol{\theta}) = 2I(\boldsymbol{\theta}_0,\boldsymbol{\theta}) + 2I(\boldsymbol{\theta},\boldsymbol{\theta}_0) = \left[d(\boldsymbol{\theta}_0,\boldsymbol{\theta}) - d(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0)\right] + \left[d(\boldsymbol{\theta},\boldsymbol{\theta}_0) - d(\boldsymbol{\theta},\boldsymbol{\theta})\right],$$

where $d(\boldsymbol{\theta},\boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}}\{-2\log L(\boldsymbol{\theta}_0|\mathbf{y})\}$ and $d(\boldsymbol{\theta},\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\{-2\log L(\boldsymbol{\theta}|\mathbf{y})\}$.

Dropping $d(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0)$, the ranking of the candidate models according to $2J(\boldsymbol{\theta}_0,\boldsymbol{\theta})$ is identical to rank

$$K(\boldsymbol{\theta}_0,\boldsymbol{\theta}) = d(\boldsymbol{\theta}_0,\boldsymbol{\theta}) + d(\boldsymbol{\theta},\boldsymbol{\theta}_0) - d(\boldsymbol{\theta},\boldsymbol{\theta}).$$

Given a set of maximum likelihood estimators $\hat{\boldsymbol{\theta}}$ in (3), the estimated symmetric measure $K(\boldsymbol{\theta}_0,\boldsymbol{\theta})$ is

$$K(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}) = d(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}) + d(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0) - d(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}), \tag{9}$$

where $d(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}})$ is exhibited in (5),

$$\begin{aligned} d(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0) &= E_{\boldsymbol{\theta}}\{-2\log L(\boldsymbol{\theta}_0|\mathbf{y})\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &= n\log 2\pi + n\log\sigma_0^2 + \frac{n\hat{\sigma}^2}{\sigma_0^2} + \frac{1}{\sigma_0^2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0), \end{aligned} \tag{10}$$

and

$$d(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}\{-2\log L(\boldsymbol{\theta}|\mathbf{y})\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = n\log 2\pi + n\log\hat{\sigma}^2 + n. \tag{11}$$

Yet, evaluating $K(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}})$ in (9) requires $\boldsymbol{\theta}_0$, Cavanaugh (1999) proposed an asymptotically unbiased estimator of

$$\Omega(\boldsymbol{\theta}_0,p) = E_{\boldsymbol{\theta}_0}\left\{K(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}})\right\} \tag{12}$$

as

$$\text{KIC} = n\log\hat{\sigma}^2 + 3(p+1), \tag{13}$$

i.e., $E_{\boldsymbol{\theta}_0}\{\text{KIC}\} + o(1) = \Omega(\boldsymbol{\theta}_0,p)$.

Seghouane and Bekara (2004) proposed an exactly unbiased estimator of (12) in order to correct a large negative bias of KIC in (13) as follows:

$$\text{KICc} = n\log\hat{\sigma}^2 + \frac{2n(p+1)}{n-p-2} - n\psi\left(\frac{n-p}{2}\right) + n\log\left(\frac{n}{2}\right),$$

i.e., $E_{\boldsymbol{\theta}_0}\{\text{KICc}\} = \Omega(\boldsymbol{\theta}_0,p)$.

Because the phi (ψ) or digamma function in KICc has no closed-form solution, Cavanaugh (2004), Seghouane and Bekara (2004), Hafidi and Mkhadri (2006) gave the asymptotically unbiased estimators of (12) called, respectively, in this paper KICc$_{\text{C}}$, KICc$_{\text{SB}}$, and KICc$_{\text{HM}}$,

$$\text{KICc}_{\text{C}} \quad = n\log\hat{\sigma}^2 + n\log\left(\frac{n}{n-p}\right) + \frac{n\left[(n-p)(2p+3)-2\right]}{(n-p-2)(n-p)}, \tag{14}$$

$$\text{KICc}_{\text{SB}} \quad = n\log\hat{\sigma}^2 + \frac{(p+1)(3n-p-2)}{n-p-2} + \frac{p}{n-p}, \tag{15}$$

$$\text{KICc}_{\text{HM}} \quad = n\log\hat{\sigma}^2 + \frac{(p+1)(3n-p-2)}{n-p-2}. \tag{16}$$

3. The unified derivations of KIC and KICc

To begin the unification of the derivations KIC, KICc$_{\text{C}}$, KICc$_{\text{SB}}$, and KICc$_{\text{HM}}$, we consider the expectation of the discrepancies in (5), (10), and (11) with respect to the true model (Seghouane and Bekara, 2004),

$$E_{\boldsymbol{\theta}_0}\left\{d\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)\right\} = n\log 2\pi + E_{\boldsymbol{\theta}_0}\left\{n\log\hat{\sigma}^2\right\} + E_{\boldsymbol{\theta}_0}\left\{\frac{n\sigma_0^2}{\hat{\sigma}^2}\right\} + E_{\boldsymbol{\theta}_0}\left\{\frac{1}{\hat{\sigma}^2}\left(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{X}\left(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}\right)\right\}, \tag{17}$$

$$E_{\boldsymbol{\theta}_0}\left\{d\left(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0\right)\right\} = n\log 2\pi + n\log\sigma_0^2 + E_{\boldsymbol{\theta}_0}\left\{\frac{n\hat{\sigma}^2}{\sigma_0^2}\right\} + E_{\boldsymbol{\theta}_0}\left\{\frac{1}{\sigma_0^2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)'\mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\right\}, \tag{18}$$

$$E_{\boldsymbol{\theta}_0}\left\{d\left(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}\right)\right\} = n\log 2\pi + E_{\boldsymbol{\theta}_0}\left\{n\log\hat{\sigma}^2\right\} + n. \tag{19}$$

From the fact that the terms $\frac{n\hat{\sigma}^2}{\sigma_0^2}$ and $\frac{1}{\sigma_0^2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)'\mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)$ are the independent χ^2 distributions with the degrees of freedom which are, respectively, $n-p$ and p, we have

$$E_{\boldsymbol{\theta}_0}\left\{\frac{n\hat{\sigma}^2}{\sigma_0^2}\right\} = n-p \text{ and } E_{\boldsymbol{\theta}_0}\left\{\frac{1}{\sigma_0^2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)'\mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\right\} = p. \tag{20}$$

Using the facts in (20), we have

$$E_{\boldsymbol{\theta}_0}\left\{\frac{n\sigma_0^2}{\hat{\sigma}^2}\right\} = E_{\boldsymbol{\theta}_0}\left\{\frac{n^2}{n\hat{\sigma}^2/\sigma_0^2}\right\} = \frac{n^2}{n-p-2}$$

and

$$E_{\boldsymbol{\theta}_0}\left\{\frac{1}{\hat{\sigma}^2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)'\mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\right\} = \frac{1}{n}E_{\boldsymbol{\theta}_0}\left\{\frac{n\sigma_0^2}{\hat{\sigma}^2}\right\}E_{\boldsymbol{\theta}_0}\left\{\frac{1}{\sigma_0^2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)'\mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\right\} = \frac{np}{n-p-2}. \tag{21}$$

Substituting the results in (21) into the expected discrepancy in (17) leads to

$$\begin{aligned}\Delta(\boldsymbol{\theta}_0,p) &= E_{\boldsymbol{\theta}_0}\left\{d\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)\right\} = n\log 2\pi + E_{\boldsymbol{\theta}_0}\left\{n\log\hat{\sigma}^2\right\} + \frac{n^2}{n-p-2} + \frac{np}{n-p-2} \\ &= n\left(\log 2\pi + 1\right) + E_{\boldsymbol{\theta}_0}\left\{\text{AICc}\right\},\end{aligned} \tag{22}$$

where AICc is the corrected version of AIC that was exhibited in (8).

Replacing the facts in (20) into the expected discrepancy in (18) yields

$$E_{\boldsymbol{\theta}_0}\left\{d\left(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0\right)\right\} = n\left(\log 2\pi + 1\right) + n\log\sigma_0^2. \tag{23}$$

Using the results in (19), (22), and (23), the expected value of $K\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)$ in (9) becomes

$$\Omega\left(\boldsymbol{\theta}_0,p\right)=E_{\boldsymbol{\theta}_0}\left\{K\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)\right\}=n\left(\log 2\pi+1\right)+E_{\boldsymbol{\theta}_0}\left\{\text{AICc}\right\}-E_{\boldsymbol{\theta}_0}\left\{n\log\frac{\hat{\sigma}^2}{\sigma_0^2}\right\}, \qquad (24)$$

where AICc is the corrected version of AIC that was exhibited in (8).

It is noteworthy that, in KIC and various corrected versions derived from $K\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)$ in (9), the differences in all formulas come from the last term of the right-hand side in (24). Therefore, in order to show the connections of KIC, KICc$_{\text{C}}$, KICc$_{\text{SB}}$, and KICc$_{\text{HM}}$, we give the following lemmas.

Lemma 1. $-E_{\boldsymbol{\theta}_0}\left\{n\log\frac{\hat{\sigma}^2}{\sigma_0^2}\right\}=-n\log\left(\frac{n-p}{2}\right)+\frac{n}{n-p}+n\log\left(\frac{n}{2}\right)+o\left(\frac{n}{\left(n-p\right)^2}\right).$ (25)

Proof. From McQuarrie and Tsai (1998) and Bernardo (1976) we have, respectively,

$$E_{\boldsymbol{\theta}_0}\left\{\log\chi_{df}^2\right\}=\psi\left(\frac{df}{2}\right)+\log 2 \text{ and } \psi\left(x\right)=\log x-\frac{1}{2x}+\text{o}\left(\frac{1}{x^2}\right) \text{ as } x\to\infty. \qquad (26)$$

Applying the facts $E_{\boldsymbol{\theta}_0}\left\{n\hat{\sigma}^2/\sigma_0^2\right\}=n-p$ in (20) and (26), we have

$$\begin{aligned}-E_{\boldsymbol{\theta}_0}\left\{n\log\frac{\hat{\sigma}^2}{\sigma_0^2}\right\} &=-E_{\boldsymbol{\theta}_0}\left\{n\log\frac{n\hat{\sigma}^2}{\sigma_0^2}\right\}+n\log n=-n\left[\psi\left(\frac{n-p}{2}\right)+\log 2\right]+n\log n\\ &=-n\left[\log\left(\frac{n-p}{2}\right)-\frac{1}{n-p}+o\left(\frac{1}{\left(n-p\right)^2}\right)\right]-n\log 2+n\log n\\ &=-n\log\left(\frac{n-p}{2}\right)+\frac{n}{n-p}+n\log\left(\frac{n}{2}\right)+o\left(\frac{n}{\left(n-p\right)^2}\right).\end{aligned}$$

Lemma 2.

$$-n\log\left(\frac{n-p}{2}\right)+\frac{n}{n-p}+n\log\left(\frac{n}{2}\right)+o\left(\frac{n}{\left(n-p\right)^2}\right)=p+\frac{n}{n-p}+o\left(\frac{p^2}{n}\right)+o\left(\frac{n}{\left(n-p\right)^2}\right). \qquad (27)$$

Proof. Applying the first-order Taylor's series expansion to expand the term $\log\left(\left(n-p\right)/2\right)$ about $n/2$, i.e.,

$$\log\left(\frac{n-p}{2}\right)=\log\left(\frac{n}{2}\right)-\frac{p}{n}+o\left(\left(\frac{p}{n}\right)^2\right),$$

to obtain the approximation in (27).

Lemma 3. $p+\frac{n}{n-p}+o\left(\frac{p^2}{n}\right)+o\left(\frac{n}{\left(n-p\right)^2}\right)=\left(p+1\right)+o\left(1\right).$ (28)

Proof. Rearrange $p+n/\left(n-p\right)$ to be $\left(p+1\right)+p/\left(n-p\right)$. As $n\to\infty$ and p is held constant, the term

$$\frac{p}{n-p}+o\left(\frac{p^2}{n}\right)+o\left(\frac{n}{\left(n-p\right)^2}\right)$$

is $o\left(1\right)$ which yields the approximation in (28).

Appling Lemma 1 into $\Omega\left(\boldsymbol{\theta}_0,p\right)$ in (24), we obtain

$$\begin{aligned}\Omega(\boldsymbol{\theta}_0, p) &= n(\log 2\pi + 1) + E_{\boldsymbol{\theta}_0}\{\text{AICc}\} - n\log\left(\frac{n-p}{2}\right) + \frac{n}{n-p} + n\log\left(\frac{n}{2}\right) + o\left(\frac{n}{(n-p)^2}\right) \\ &= n(\log 2\pi + 1) + E_{\boldsymbol{\theta}_0}\left\{\text{KICc}_{\text{C}} + o\left(\frac{n}{(n-p)^2}\right)\right\},\end{aligned}$$

where KICc_{C} is the corrected version of KIC from Cavanaugh (2004) that was exhibited in (14).

Appling Lemmas 1 and 2 into $\Omega(\boldsymbol{\theta}_0, p)$ in (24), we obtain

$$\begin{aligned}\Omega(\boldsymbol{\theta}_0, p) &= n(\log 2\pi + 1) + E_{\boldsymbol{\theta}_0}\{\text{AICc}\} + p + \frac{n}{n-p} + o\left(\frac{p^2}{n}\right) + o\left(\frac{n}{(n-p)^2}\right) \\ &= n(\log 2\pi + 1) + E_{\boldsymbol{\theta}_0}\left\{\text{KICc}_{\text{SB}} + o\left(\frac{p^2}{n}\right) + o\left(\frac{n}{(n-p)^2}\right)\right\},\end{aligned}$$

where KICc_{SB} is the corrected version of KIC from Seghouane and Bekara (2004) that was exhibited in (15).

Appling Lemmas 1, 2, and 3 into $\Omega(\boldsymbol{\theta}_0, p)$ in (24), we obtain

$$\begin{aligned}\Omega(\boldsymbol{\theta}_0, p) &= n(\log 2\pi + 1) + E_{\boldsymbol{\theta}_0}\{\text{AICc}\} + (p+1) + o(1) \\ &= n(\log 2\pi + 1) + E_{\boldsymbol{\theta}_0}\{\text{KICc}_{\text{HM}} + o(1)\},\end{aligned}$$

where KICc_{HM} is the corrected version of KIC from Hafidi and Mkhadri (2006) that was exhibited in (16).

The connections of KIC, KICc_{HM}, KICc_{SB}, and KICc_{C} are given by

$$\begin{aligned}\text{KICc}_{\text{HM}} &= \text{KIC} + \frac{2(p+1)(p+2)}{n-p-2}, \\ \text{KICc}_{\text{SB}} &= \text{KICc}_{\text{HM}} + \frac{p}{n-p}, \\ \text{KICc}_{\text{C}} &= \text{KICc}_{\text{SB}} + n\log\left(\frac{n}{n-p}\right) - p. \qquad (29)\end{aligned}$$

From the connections in (29), we found that the terms

$$\frac{2(p+1)(p+2)}{n-p-2} \text{ and } \frac{p}{n-p}$$

are not greater than zero if and only if $n-p>2$ and p belong to the sets of $[-2, -1]$ and $(-\infty, 0]$, respectively. Therefore, we can argue that these two terms have values of at least zero because p represents the number of regression coefficients which has the value of at least one and both terms are very close to zero if the ratio of p/n tends to zero. This conclusion links to $\text{KICc}_{\text{SB}} \geq \text{KICc}_{\text{HM}} \geq$ KIC. While the term

$$n\log\left(\frac{n}{n-p}\right) - p \qquad (30)$$

has the value in the range $[-p, \infty)$ where it is close to the lower bound $-p$ if the ratio of p/n tends to zero. If the value of p is fixed, this term is the decreasing function of n, whereas when the value of n is fixed, it is the increasing function of p. Whenever $n-p>0$ and the condition

$$\left(1-p/n\right)\exp\left(p/n\right)<1 \tag{31}$$

is true, we have the term in (30) being greater than zero. This means that the penalty function of KICc$_C$ is stronger than other criteria, KICc$_{SB}$, KICc$_{HM}$, and KIC, under the condition in (31). The strong penalty may cause KICc$_C$ to have the maximum frequency of the correct order being selected. However, occasionally it causes the model selection criterion to select underparameterized models (McQuarrie and Tsai, 1998). This confusion is studied by the extensive simulation in the next section.

4. Simulation study

To examine the model selection criteria performance, we generated 1,000 realizations of true multiple regression models in (1) for four cases as follows.

Model I represents a very weakly identifiable true model with large dimension of the model: $y_1=1+0.5X_2+0.1X_3+0.05X_4+0.01X_5+0.005X_6+0.001X_7+0.0005X_8+\varepsilon_1$.

Model II represents a weakly identifiable true model with small dimension of the model: $y_2=1+0.5X_2+0.25X_3+\varepsilon_2$.

Model III represents a very strongly identifiable true model with small dimension of the model: $y_3=1+2X_2+3X_3+4X_4+\varepsilon_3$.

Model IV represents a strongly identifiable true model with large dimension of the model: $y_4=1+X_2+X_3+X_4+X_5+X_6+X_7+X_8+\varepsilon_4$.

Model I and Model II represent the weakly identifiable true models which mean they are not easily identified compared to the strongly identifiable true models such as Model III and Model IV. From a previous study, Kundu and Murali (1996) concluded that the criteria performance did not change much when the true variance σ_0^2 and the distributions of $\boldsymbol{\varepsilon}_0$ in (1) were changed. As a result, we have taken the error random variables to be normally distributed with zero mean and the true variance σ_0^2 is assumed to be equal to 1. For each model, four different sample sizes are split into two categories: small sample (n = 15, 25) and large sample (n = 100, 500). Ten candidate variables, X_1 until X_{10}, are stored in an $n\times 10$ matrix **X** of the candidate model in (2), with a column of ones, followed by nine independent identically distributed normal random variables with zero mean and variance-covariance matrix equal to identity matrix $\mathbf{I}_{10}$. The candidate models include the columns of **X** in a sequentially nested fashion; i.e., columns 1 to p define the design matrix for the candidate model with dimension p. The criteria considered in this simulation are divided into two families. Firstly, is the criteria based on Kullback's directed divergence: AIC in (7) and AICc in (8). Secondly, is the criteria based on Kullback's symmetric divergence: KIC in (13), KICc$_C$ in (14), KICc$_{SB}$ in (15), and KICc$_{HM}$ in (16). Model selection criteria performance is examined by a measure of counting the frequency of order being selected. Particularly for the case of true model being weakly identifiable, we use an additional measure which is the observed L_2 efficiency. Observed L_2 distance, scaled by $1/n$, between the true model in (1) and the fitted candidate model in (2) is defined as (McQuarrie et al., 1997; McQuarrie, 1999)

$$L_2\left(p\right)=\frac{1}{n}\left(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{X}\left(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}\right).$$

Observed L_2 efficiency is defined by the ratio

$$\text{Observed } L_2 \text{ efficiency } = \frac{\min_{1\le p\le P} L_2\left(p\right)}{L_2\left(p_s\right)},$$

where P is the class of all possible candidate models, p is the rank of fitted candidate model, and p_s is the model selected by specific model selection criterion. The closer the selected model is to the true model, the higher the efficiency. Therefore, the best model selection criterion will select a model which yields high efficiency even in small samples or if the true model is weakly identifiable. For 1,000 realizations, the results of comparing the model selection criteria performance are shown in Table 1 and 2. Columns "d" and "K" in Table 1 stand for the estimated measures in (5) and (9), respectively. The conclusions of this simulation are as follows. In Table 1, for the small sample size and the true model is somewhat difficult to identify, such as Model I, Model II for n = 15, 25, and Model IV for n = 15, the original criteria AIC and KIC perform better than their corrected versions. When the sample size is still small but the true model is easily to identify, such as Model III for n = 15, 25 and Model IV for n = 25, the corrected versions work better. For the large sample size but the true model is very difficult to detect, such as Model I for n = 100, 500, the AIC family of criteria performs better than the KIC family. When the sample size is still large and the true model can be specified more easily, such as Model II, Model III, and Model IV for n = 100, 500, the KIC family performs the best. This simulation also found that when the true model is very difficult to detect, such as Model I and the sample size is small n = 15, 25, the estimated symmetric measure in (9) has the opportunity to cause more underfitted order being selected than the estimated directed measure in (5). This result contributes the criteria in KIC family to having a low frequency of choosing the correct model. In Table 2, the observed L_2 efficiency suggests that KICc$_C$ in KIC family is the best criterion for all sample sizes of a weakly identifiable true model.

Table 1. Frequency of the model order being selected by each criterion for 1,000 realizations

Model	n	Order	Criteria							
			AIC	AICc	KIC	KICcHM	KICcSB	KICcC	*d*	*K*
I very weakly identifiable (true order p_0 = 8)	15	Underfitted	596	1000	837	1000	1000	1000	982	986
		Correct	**54**	0	26	0	0	0	0	0
		Overfitted	350	0	137	0	0	0	18	14
	25	Underfitted	859	998	972	1000	1000	1000	987	992
		Correct	**39**	1	11	0	0	0	0	0
		Overfitted	102	1	17	0	0	0	13	8
	100	Underfitted	944	974	993	998	999	999	998	998
		Correct	**23**	14	5	2	1	1	0	0
		Overfitted	33	12	2	0	0	0	2	2
	500	Underfitted	958	962	998	998	998	999	1000	1000
		Correct	**21**	**21**	1	1	1	0	0	0
		Overfitted	21	17	1	1	1	1	0	0
II weakly identifiable (true order p_0 = 3)	15	Underfitted	284	820	542	859	864	875	577	547
		Correct	132	123	**148**	111	109	105	423	453
		Overfitted	584	57	310	30	27	20	0	0
	25	Underfitted	374	653	575	716	720	727	368	344
		Correct	244	235	**264**	235	231	226	631	655
		Overfitted	382	112	161	49	49	47	1	1
	100	Underfitted	109	133	214	230	231	232	34	26
		Correct	609	642	676	677	678	**680**	966	974
		Overfitted	282	225	110	93	91	88	0	0
	500	Underfitted	0	0	0	0	0	0	0	0
		Correct	737	751	890	895	895	**896**	1000	1000
		Overfitted	263	249	110	105	105	104	0	0

Table 1. (Continued)

Model	n	Order	Criteria							
			AIC	AICc	KIC	KICcHM	KICcSB	KICcC	d	K
III very strongly identifiable (true order $p_0 = 4$)	15	Underfitted	0	0	0	0	0	0	30	0
		Correct	325	939	583	964	964	**968**	970	1000
		Overfitted	675	61	417	36	36	32	0	0
	25	Underfitted	0	0	0	0	0	0	0	0
		Correct	549	855	749	899	904	**920**	1000	1000
		Overfitted	451	145	251	101	96	80	0	0
	100	Underfitted	0	0	0	0	0	0	0	0
		Correct	687	756	855	874	874	**878**	1000	1000
		Overfitted	313	244	145	126	126	122	0	0
	500	Underfitted	0	0	0	0	0	0	0	0
		Correct	713	731	885	**889**	**889**	**889**	1000	1000
		Overfitted	287	269	115	111	111	111	0	0
IV strongly identifiable (true order $p_0 = 8$)	15	Underfitted	36	887	94	943	955	969	724	554
		Correct	444	113	**532**	57	45	31	214	377
		Overfitted	520	0	374	0	0	0	62	69
	25	Underfitted	5	31	9	60	62	67	281	133
		Correct	710	**950**	840	928	928	925	663	846
		Overfitted	285	19	151	12	10	8	56	21
	100	Underfitted	0	0	0	0	0	0	0	0
		Correct	815	882	925	950	950	**953**	1000	1000
		Overfitted	185	118	75	50	50	47	0	0
	500	Underfitted	0	0	0	0	0	0	0	0
		Correct	854	864	951	**956**	**956**	**956**	1000	1000
		Overfitted	146	136	49	44	44	44	0	0

Note: Boldface type indicates the maximum frequency of correct order being selected.

Table 2. Average and standard deviation of the observed L_2 efficiency over 1,000 realizations

Circumstance	Stat.	Criteria					
		AIC	AICc	KIC	KICc_{HM}	KICc_{SB}	KICc_{C}
weakly identifiable (Model I and Model II), small sample size (15, 25)	Ave. L_2 eff.	0.5332	0.7791	0.6826	0.8048	0.8062	**0.8098**
	Rank	6	4	5	3	2	**1**
	S.D. L_2 eff.	0.3598	0.2765	0.3343	0.2480	0.2462	**0.2420**
	Rank	6	4	5	3	2	**1**
weakly identifiable (Model I and Model II), large sample size (100, 500)	Ave. L_2 eff.	0.7239	0.7418	0.7771	0.7808	0.7810	**0.7817**
	Rank	6	5	4	3	2	**1**
	S.D. L_2 eff.	0.3096	0.3001	0.2601	0.2563	0.2562	**0.2554**
	Rank	6	5	4	3	2	**1**
weakly identifiable (Model I and Model II)	Ave. L_2 eff.	0.6286	0.7604	0.7299	0.7928	0.7936	**0.7958**
	Rank	6	4	5	3	2	**1**
	S.D. L_2 eff.	0.3347	0.2883	0.2972	0.2522	0.2512	**0.2487**
	Rank	6	4	5	3	2	**1**

Note: Boldface type indicates the best performance.

5. Conclusions, discussion, and further study

This paper presents the derivations to unify the justifications of the criteria based on Kullback's symmetric divergence; the Kulback information criterion (KIC) by Cavanaugh (1999) and the corrected versions; KICc_{C} by Cavanaugh (2004), KICc_{SB} by Seghouane and Bekara (2004), and KICc_{HM} by Hafidi and Mkhadri (2006). The results show that KICc_{C} has the strongest penalty function under the condition in (31), followed, respectively, by KICc_{SB}, KICc_{HM}, and KIC. The performance of them is examined by the extensive simulation study relative to the criteria based on Kullback's directed divergence, AIC and AICc. Our simulation study indicates that, for the small sample size and the true model is somewhat difficult to identify, the performance of the original criteria AIC and KIC is better than their corrected versions. When the sample size is still small but the true model is easily to identify, the corrected versions perform the best. For the large sample size but the true model is very difficult to detect, the AIC family of criteria performs better than the KIC family. When the sample size is still large and the true model can be specified more easily, the KIC family performs the best. This simulation also found that, although the proofs in this study show that the criteria based on Kullback's symmetric divergence are stronger than the criteria based on the directed divergence, sometimes the performance of them is worse. This result may be because the estimated symmetric measure in (9) contributes to all criteria in KIC family usually having stronger penalty functions than the AIC family. This problem causes a greater number of underfitted orders being selected, which then contributes to a low frequency of choosing the correct model. However, when the true model is very difficult to detect, such as Model 1; none of the criteria correctly identify the true model more than 6% of the time. As a result, the frequency of correct order being selected may not be meaningful. For this reason, we have also used the observed L_2 efficiency to assess model selection criteria performance. This measure suggests that, in a weakly identifiable true model, whether the sample size is small or large, KICc$_C$ is the best criterion because it has highest average value of the observed L_2 efficiency and lowest standard deviation. The better performance of KICc$_C$ may be because its formula is closer to the expected estimated symmetric discrepancy than others. However, KICc$_C$ is more likely to select an underfitted model than other criterion which is because its penalty function is strong. Although KICc$_C$ tends to select underfitted models, these selected models are close to the true model which is weak.

In future work, we hope to extend KICc$_C$ from Cavanaugh (2004) to construct a model selection criterion to serve as an asymptotically unbiased estimator of a variant of Kullback's symmetric divergence for multivariate regression and seemingly unrelated regression models. Because, at this time, there exists the multivariate model selection based on the extensions of KICc$_{SB}$ (Seghouane, 2006a) and KICc$_{HM}$ (Hafidi and Mkhadri, 2006).

Acknowledgements

This project is financial supported by The Thailand Research Fund, Office of The Higher Education Commission, and Thaksin University under grant No. MRG5480044. I would like to give special thanks to Prof. Dr. Joseph E. Cavanaugh, Department of Biostatistics, College of Public Health, University of Iowa, USA, Assoc. Prof. Dr. Jirawan Jitthavech, and Assoc. Prof. Dr. Vichit Lorchirachoonkul, National Institute of Development Administration (NIDA), THAILAND for all comments and suggestions.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Proc. *2nd Int. Symp. on Information Theory*. Akademia Kiado, Budapest, pp. 267–281.

Akaike, H., 1974. A new look at the statistical model identification. *IEEE T. Automat. Contr.* **19**, 716–723.

Bedrick, E.J., Tsai, C.L., 1994. Model selection for multivariate regression in small samples. *Biometrics*. **50**, 226–231.

Bernardo, J.M., 1976. Psi (digamma) function. *J. Roy. Stat. Soc. C-App.* **25**, 315–317.

Cavanaugh, J.E., 1997. Unifying the derivation of Akaike and corrected information criteria. *Stat. Probabil. Lett.* **33**, 201–208.

Cavanaugh, J.E., 1999. A large-sample model selection criterion based on Kullback's symmetric divergence. *Stat. Probabil. Lett.* **42**, 333–343.

Cavanaugh, J.E., 2004. Criteria for linear model selection based on Kullback's symmetric divergence. *Aust. NZ. J. Stat.* **46**, 257–274.

Hafidi, B., 2006. A small-sample criterion based on Kullback's symmetric divergence for vector autoregressive modeling. *Stat. Probabil. Lett.* **76**, 1647–1654.

Hafidi, B., Mkhadri, A., 2006. A corrected Akaike criterion based on Kullback's symmetric divergence: applications in time series, multiple and multivariate regression. *Comput. Stat. Data. An.* **50**, 1524–1550.

Hurvich, C.M., Tsai, C.L., 1989. Regression and time series model selection in small samples. *Biometrika.* **76**, 297–307.

Kullback, S., 1968. *Information theory and statistics*. Dover, New York.

Kundu, D., Murali G. 1996. Model selection in linear regression. *Comput. Stat. Data. An*. **22**, 461–469.

McQuarrie, A.D., Shumway, R., Tsai, C.L., 1997. The model selection criterion AICu. *Stat. Probabil. Lett.* **34**, 285–292.

McQuarrie, A.D., Tsai, C.L., 1998. *Regression and time series model selection*. Singapore: World Scientific.

McQuarrie, A.D., 1999. A small-sample correction for the Schwarz SIC model selection criterion. *Stat. Probabil. Lett.* **44**, 79–86.

Seghouane, A.K., Bekara M., 2004. A small sample model selection criterion based on Kullback's symmetric divergence. *IEEE T. Signal Proces*. **52**, 3314–3323.

Seghouane, A.K., 2006a. Multivariate regression model selection from small samples using Kullback's symmetric divergence. *Signal Process*. **86**, 2074–2084.

Seghouane, A.K., 2006b. A note on overfitting properties of KIC and KICc. *Signal Process* **86**, 3055–3060.

Study on the penalty functions of model selection criteria

Warangkhana Keerativibool

Department of Mathematics and Statistics, Faculty of Science,
Thaksin University, Phatthalung, Thailand.

Email: warang27@gmail.com

Abstract

The aim of this paper is to study the penalty functions of the well-known model selection criteria, *AIC*, *BIC*, and *KIC*, which can unify their formulas as

$$APIC\alpha = \log\left(\hat{\sigma}^2\right) + \alpha\left(p+1\right)/n,$$

called Adjusted Penalty Information Criterion. The appropriate value of α for $APIC\alpha$ has been found to reduce the probabilities of over- and underfitting and also to overcome the weak signal-to-noise ratio. The value of α is selected based on four measurements: the probability of over- and underfitting, the signal-to-noise ratio, the probability of order selected, and the observed L_2 efficiency. Performance of $APIC\alpha$ is examined by theoretical and extensive simulation study.

Keywords: model selection; penalty function; probability of overfitting; signal-to-noise ratio; observed L_2 efficiency

1. Introduction

In regression analysis, the choice of an appropriate model from a class of candidate models to characterize the study data is a key issue. In real life, we may not know what the true model is, but we hope to find a model that is a reasonably accurate representation. A model selection criterion represents a useful tool to judge the propriety of a fitted model by assessing whether it offers an optimal balance between goodness of fit and parsimony. The first model selection criterion to gain widespread acceptance was the Akaike information criterion, AIC (Akaike, 1973). This serves as an asymptotically unbiased estimator of a variant of Kullback's directed divergence between the true model and a fitted approximating model. The directed divergence, also known as the Kullback-Leibler information, the I-divergence, or the relative entropy, assesses the dissimilarity between two statistical models. Other well-known criteria were subsequently introduced and studied such as Bayesian information criterion, BIC (Schwarz, 1978), and Kullback information criterion, KIC (Cavanaugh, 1999). BIC is an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model (Neath and Cavanaugh, 1997). KIC is a symmetric measure, meaning that an alternate directed divergence may be obtained by reversing the roles of the two models in the definition of the measure. The sum of the two directed divergences is Kullback's symmetric divergence, also known as the J-divergence (Cavanaugh, 1999; Cavanaugh, 2004). Although AIC remains arguably the most widely used model selection criterion, BIC and KIC are popular competitors. In fact, BIC is often preferred over AIC by practitioners who find appeal in either its Bayesian justification or its tendency to choose more parsimonious models than AIC (Neath and Cavanaugh, 1997). Likewise, KIC is a symmetric measure which combines the information in two related, though distinct measures; it functions as a gauge of model disparity that is arguably more sensitive than AIC that corresponds to only individual components (Cavanaugh, 1999; Cavanaugh, 2004). However, AIC, BIC, and KIC still have the problems of overfitting and weak signal-to-noise ratio due to the weak penalty functions. With this motivation, the aim of this paper is to study the penalty functions based on these criteria for the case of univariate regression model in order to find the appropriate value of penalty to reduce the probabilities of over- and underfitting and also to overcome the weak signal-to-noise ratio. The remainder of this paper is organized as follows. In Section 2, we unify AIC, BIC, and KIC in one form, called Adjusted Penalty Information Criterion $(APIC\alpha)$. The studies on the probability of overfitting and signal-to-noise ratio are also considered in this section. In Section 3, we simulate 1,000 realizations of multiple regression models in order to study the probability of the order selected and the observed L_2 efficiency of $APIC\alpha$ where the values of α range from 1 to 14. Finally, Section 4 is the conclusions, discussion, and further study.

2. Model selection criteria, probability of overfitting, and signal-to-noise ratio

Suppose data are generated by the operating model, i.e., true model

$$\mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_0,\ \boldsymbol{\varepsilon}_0 \sim \mathrm{N}_n\left(\mathbf{0},\ \sigma_0^2\mathbf{I}_n\right), \tag{1}$$

and the candidate or approximating model is in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathrm{N}_n\left(\mathbf{0},\ \sigma^2\mathbf{I}_n\right), \tag{2}$$

where $\mathbf{y}$ is an $n\times 1$ dependent random vector of observations, $\mathbf{X}_0$ and $\mathbf{X}$ are $n\times p_0$ and $n\times p$ matrices of independent variables with full-column rank, respectively, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$ are $p_0\times 1$ and $p\times 1$ parameter vectors of regression coefficients, respectively, $\boldsymbol{\varepsilon}_0$ and $\boldsymbol{\varepsilon}$ are $n\times 1$ noise vectors. The $(p+1)\times 1$ vector of parameters is $\boldsymbol{\theta}_0 = \left[\boldsymbol{\beta}_0'\ \ \sigma_0^2\right]'$ and the maximum likelihood estimator of $\boldsymbol{\theta}_0$ is $\hat{\boldsymbol{\theta}} = \left[\hat{\boldsymbol{\beta}}'\ \ \hat{\sigma}^2\right]'$ where

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \text{ and } \hat{\sigma}^2 = \left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)/n.$$

For each data set, we can construct many fitted candidate models. Nevertheless, we cannot know which model is the best. Criterion for model selection is a way to solve this problem. AIC, BIC, and KIC are three well-known criteria to consider in this study. We scale these criteria by $1/n$ in order to express them as a rate per observation. The formulas for them are based on the form of the log of the likelihood function of the maximum likelihood estimator of σ^2 plus a penalty function, called Adjusted Penalty Information Criterion,

$$APIC\alpha = \log\left(\hat{\sigma}^2\right) + \frac{\alpha(p+1)}{n}. \tag{3}$$

When the values of α in (3) are equal to 2, $\log(n)$, and 3, $APIC\alpha$ becomes AIC, BIC, and KIC, respectively. The appropriate value of α has been found to reduce the probabilities of over- and underfitting and also to overcome the weak signal-to-noise ratio. The value of α is selected by four measurements: the probability of over- and underfitting, the signal-to-noise ratio, the probability of order selected, and the observed L_2 efficiency. Theoretical and empirical methods are used to examine the performance of $APIC\alpha$.

The terms over- and underfitting can be defined in two ways. Under consistency, when a true model is itself a candidate model, overfitting is a situation when the model has extra variables with more parameters than the optimal model and underfitting is defined as choosing a model that either has too few variables or is incomplete. In view of efficiency, overfitting can be defined as choosing a model that has more variables than the model identified as closest to the true model, thereby reducing efficiency. Underfitting is defined as choosing a model with too few variables compared to the closest model, also reducing efficiency. Both over- and underfitting can lead to problems with the predictive abilities of a model. An underfitted model may have poor predictive ability due to a lack of detail in the model, while an overfitted model may be unstable in the sense that repeated samples from the same process can lead to widely differing predictions due to variability in the extraneous variables. A criterion that can balance the tendencies of over- and underfitted is preferable. (McQuarrie and Tsai, 1998; Seghouane, 2006).

The probability of model selection criterion preferring the overfitted model is analyzed here by comparing the true model of order p_0 to a more complex model or overfitted model of order $p_0 + l$, $l > 0$. Hence for finite n, the probability that $APIC\alpha$ prefers the overfitted model is defined by

$$P\left\{APIC\alpha_{p_0+l} < APIC\alpha_{p_0}\right\} = P\left\{\log\left(\hat{\sigma}^2_{p_0+l}\right) + \frac{\alpha\left(p_0+l+1\right)}{n} < \log\left(\hat{\sigma}^2_{p_0}\right) + \frac{\alpha\left(p_0+1\right)}{n}\right\}$$

$$= P\left\{\log\left(\frac{\hat{\sigma}^2_{p_0}}{\hat{\sigma}^2_{p_0+l}}\right) > \frac{\alpha l}{n}\right\} = P\left\{\frac{\hat{\sigma}^2_{p_0}}{\hat{\sigma}^2_{p_0+l}} > \exp\left(\frac{\alpha l}{n}\right)\right\} = P\left\{\frac{\hat{\sigma}^2_{p_0} - \hat{\sigma}^2_{p_0+l}}{\hat{\sigma}^2_{p_0+l}} > \exp\left(\frac{\alpha l}{n}\right) - 1\right\}. \quad (4)$$

Under the assumption of nested models; $p \geq p_0$ and $l > 0$, we have $n\left(\hat{\sigma}^2_p - \hat{\sigma}^2_{p+l}\right) \sim \sigma^2_0 \chi^2_l$, $n\hat{\sigma}^2_p \sim \sigma^2_0 \chi^2_{n-p}$, where χ^2_k represents the chi-square distribution with k degrees of freedom, and $\hat{\sigma}^2_p - \hat{\sigma}^2_{p+l}$ is independent of $\hat{\sigma}^2_{p+l}$ (McQuarrie and Tsai, 1998). (5)

Then the probability of overfitting by l extra variables of $APIC\alpha$ in (4) becomes

$$P\left\{APIC\alpha_{p_0+l} < APIC\alpha_{p_0}\right\} = P\left\{F_{l,\, n-p_0-l} > \frac{n-p_0-l}{l}\left[\exp\left(\frac{\alpha l}{n}\right) - 1\right]\right\}. \quad (6)$$

In (6), we found that $APIC\alpha$'s probability of overfitting depends on the value of α in (3). If the value of α tends to infinity under the same values of the sample size (n), the order of true model (p_0), and the additional variable (l), $APIC\alpha$ tends to less overfitting. When we replace the values of α in (6) by $2, \log(n),$ and $3,$ we get the probabilities of overfitting of $AIC, BIC,$ and $KIC,$ respectively. The proof of the probability of overfitting can be confirmed numerically in Table 1. The explanation of the result in Table 1 is that, e.g. for $n = 15,$ $p_0 = 3,$ and $l = 1,$ the probability of overfitting of $APIC1$ is 0.4025, this means that this criterion would select the model whose order is higher by one order than true model with a probability of 0.4025. Although the large value of α resulted in $APIC\alpha$ having the low probability of overfitting, sometimes it will be prone to underfitting. This result will be shown in the simulation study.

Table 1. Probability of overfitting by l extra variables of $APIC\alpha$ for different values of n, p_0, and l

n	p_0	l	Criteria													
			*APIC*1	*APIC*2	*APIC*3	*APIC*4	*APIC*5	*APIC*6	*APIC*7	*APIC*8	*APIC*9	*APIC*10	*APIC*11	*APIC*12	*APIC*13	*APIC*14
15	3	1	0.4025	0.2363	0.1469	0.0939	0.0611	0.0402	0.0266	0.0178	0.0119	0.0080	0.0054	0.0037	0.0025	0.0017
15	3	2	0.5134	0.2636	0.1353	0.0695	0.0357	0.0183	0.0094	0.0048	0.0025	0.0013	0.0007	0.0003	0.0002	0.0001
15	3	3	0.5947	0.2857	0.1287	0.0561	0.0240	0.0101	0.0042	0.0018	0.0007	0.0003	0.0001	0.0001	0.0000	0.0000
15	3	4	0.6664	0.3143	0.1305	0.0508	0.0190	0.0070	0.0025	0.0009	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000
15	4	1	0.4257	0.2599	0.1676	0.1110	0.0747	0.0509	0.0349	0.0241	0.0167	0.0117	0.0082	0.0057	0.0040	0.0028
15	4	2	0.5488	0.3012	0.1653	0.0907	0.0498	0.0273	0.0150	0.0082	0.0045	0.0025	0.0014	0.0007	0.0004	0.0002
15	4	3	0.6384	0.3362	0.1667	0.0802	0.0378	0.0176	0.0082	0.0037	0.0017	0.0008	0.0004	0.0002	0.0001	0.0000
15	4	4	0.7154	0.3784	0.1780	0.0788	0.0336	0.0140	0.0058	0.0023	0.0009	0.0004	0.0002	0.0001	0.0000	0.0000
30	3	1	0.3565	0.1922	0.1102	0.0651	0.0392	0.0239	0.0147	0.0091	0.0057	0.0035	0.0022	0.0014	0.0009	0.0006
30	3	2	0.4346	0.1889	0.0821	0.0357	0.0155	0.0067	0.0029	0.0013	0.0006	0.0002	0.0001	0.0000	0.0000	0.0000
30	3	3	0.4846	0.1795	0.0617	0.0204	0.0066	0.0021	0.0007	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
30	3	4	0.5256	0.1720	0.0482	0.0125	0.0031	0.0007	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
30	4	1	0.3661	0.2012	0.1175	0.0706	0.0433	0.0268	0.0168	0.0106	0.0067	0.0043	0.0027	0.0017	0.0011	0.0007
30	4	2	0.4493	0.2019	0.0907	0.0408	0.0183	0.0082	0.0037	0.0017	0.0007	0.0003	0.0002	0.0001	0.0000	0.0000
30	4	3	0.5033	0.1954	0.0705	0.0245	0.0083	0.0028	0.0009	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
30	4	4	0.5475	0.1902	0.0568	0.0157	0.0042	0.0011	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	3	1	0.3284	0.1670	0.0905	0.0506	0.0289	0.0167	0.0097	0.0057	0.0034	0.0020	0.0012	0.0007	0.0004	0.0003
100	3	2	0.3867	0.1496	0.0578	0.0224	0.0087	0.0033	0.0013	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
100	3	3	0.4178	0.1288	0.0367	0.0100	0.0027	0.0007	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	3	4	0.4395	0.1109	0.0236	0.0046	0.0009	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	4	1	0.3310	0.1692	0.0922	0.0519	0.0297	0.0173	0.0101	0.0060	0.0035	0.0021	0.0013	0.0008	0.0005	0.0003
100	4	2	0.3906	0.1526	0.0596	0.0233	0.0091	0.0036	0.0014	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
100	4	3	0.4227	0.1322	0.0382	0.0106	0.0029	0.0008	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	4	4	0.4453	0.1144	0.0248	0.0050	0.0009	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

The signal-to-noise ratio is the second measure used to study the property of $APIC\alpha$. McQuarrie and Tsai (1998) defined the signal-to-noise ratio as a measurement that is basically a ratio of the expectation to the standard deviation of the difference in criterion values for two models. The ratio tends to assess whether the penalty function is sufficiently strong in relation to the goodness of fit term. From the true model order p_0 and a candidate model order p_0+l where $l>0$, the true model is considered better than a candidate model if the criterion value of a model of order p_0 is less than that of order p_0+l, $APIC\alpha_{p_0} < APIC\alpha_{p_0+l}$. Then, the signal-to-noise ratio that the true model has selected compared to a candidate model is defined by

$$\frac{signal}{noise}=\frac{E\lfloor APIC\alpha_{p_0+l}-APIC\alpha_{p_0}\rfloor}{sd\left[APIC\alpha_{p_0+l}-APIC\alpha_{p_0}\right]}$$

$$=\frac{E\left[\log\left(\hat{\sigma}^2_{p_0+l}\right)+\frac{\alpha\left(p_0+l+1\right)}{n}-\log\left(\hat{\sigma}^2_{p_0}\right)-\frac{\alpha\left(p_0+1\right)}{n}\right]}{sd\left[\log\left(\hat{\sigma}^2_{p_0+l}\right)+\frac{\alpha\left(p_0+l+1\right)}{n}-\log\left(\hat{\sigma}^2_{p_0}\right)-\frac{\alpha\left(p_0+1\right)}{n}\right]}=\frac{E\left[\log\left(\frac{\hat{\sigma}^2_{p_0+l}}{\hat{\sigma}^2_{p_0}}\right)+\frac{\alpha l}{n}\right]}{sd\left[\log\left(\frac{\hat{\sigma}^2_{p_0+l}}{\hat{\sigma}^2_{p_0}}\right)+\frac{\alpha l}{n}\right]}. \tag{7}$$

Applying the second-order of Taylor's series expansions in order to find the signal in (7) is as follows: suppose $X\sim\chi^2_p$, expanding $\log\left(X\right)$ about $E\left(X\right)=p$, we have

$$\log\left(X\right)\doteq\log\left(p\right)+\frac{1}{p}\left(X-p\right)-\frac{1}{2p^2}\left(X-p\right)^2 \text{ and } E\left[\log\left(X\right)\right]\doteq\log\left(p\right)-\frac{1}{p}. \tag{8}$$

Using the results in (8) and the assumption in (5), the approximate signal in (7) is

$$E\left[APIC\alpha_{\mathrm{U}\,p_0+l}-APIC\alpha_{\mathrm{U}\,p_0}\right]=E\left[\log\left(n\hat{\sigma}^2_{p_0+l}\right)\right]-E\left[\log\left(n\hat{\sigma}^2_{p_0}\right)\right]+\frac{\alpha l}{n}$$

$$\doteq\left\{\log\left(\sigma_0^2\right)+\log\left(n-p_0-l\right)-\frac{1}{n-p_0-l}\right\}-\left\{\log\left(\sigma_0^2\right)+\log\left(n-p_0\right)-\frac{1}{n-p_0}\right\}+\frac{\alpha l}{n}$$

$$=\log\left(\frac{n-p_0-l}{n-p_0}\right)-\frac{l}{\left(n-p_0-l\right)\left(n-p_0\right)}+\frac{\alpha l}{n}. \tag{9}$$

Using the assumption in (5) to find the noise in (7) by the *Q*-statistic which has the Beta distribution as follows:

$$Q=\frac{n\hat{\sigma}^2_{p_0+l}}{n\hat{\sigma}^2_{p_0}}\sim Beta\left(\frac{n-p_0-l}{2},\frac{l}{2}\right), \tag{10}$$

and the log-distribution of *Q*-statistic is

$$\log\left(Q\right)=\log\left(\frac{n\hat{\sigma}^2_{p_0+l}}{n\hat{\sigma}^2_{p_0}}\right)\sim\log\text{-}Beta\left(\frac{n-p_0-l}{2},\frac{l}{2}\right). \tag{11}$$

Applying the first-order of Taylor's series expansions when $X\sim\chi^2_p$, we expand $\log\left(X\right)$ about $E\left(X\right)=p$ as follows:

$$\log\left(X\right)\doteq\log\left(p\right)+\frac{1}{p}\left(X-p\right). \tag{12}$$

Using (12) to expand $\log\left(Q\right)$ in (11) about $E\left(Q\right)=\frac{\left(n-p_0-l\right)/2}{\left(n-p_0-l\right)/2+l/2}=\frac{n-p_0-l}{n-p_0}$, we have

$$\log(Q) \doteq \log\left(\frac{n-p_0-l}{n-p_0}\right)+\frac{n-p_0}{n-p_0-l}\left(Q-\frac{n-p_0-l}{n-p_0}\right). \tag{13}$$

The variance of $\log(Q)$ in (11) is approximated by the variance of $\log(Q)$ in (13) as follows:

$$\operatorname{var}\left[\log\left(\frac{n\hat{\sigma}^2_{p_0+l}}{n\hat{\sigma}^2_{p_0}}\right)\right]=\operatorname{var}\left[\log(Q)\right]\doteq\operatorname{var}\left[\log\left(\frac{n-p_0-l}{n-p_0}\right)+\frac{n-p_0}{n-p_0-l}\left(Q-\frac{n-p_0-l}{n-p_0}\right)\right]$$

$$=\left(\frac{n-p_0}{n-p_0-l}\right)^2\operatorname{var}(Q)\ =\frac{(n-p_0)^2}{(n-p_0-l)^2}\left[\frac{\left[(n-p_0-l)/2\right](l/2)}{\left((n-p_0-l)/2+l/2\right)^2\left((n-p_0-l)/2+l/2+1\right)}\right]$$

$$=\frac{2l}{(n-p_0-l)(n-p_0+2)}\,. \tag{14}$$

Therefore, the standard deviation of $\log(Q)$ in (14) or the approximate noise in (7) is

$$sd\left[\log\left(\frac{\hat{\sigma}^2_{p_0+l}}{\hat{\sigma}^2_{p_0}}\right)+\frac{\alpha l}{n}\right]=sd\left[\log\left(\frac{n\hat{\sigma}^2_{p_0+l}}{n\hat{\sigma}^2_{p_0}}\right)\right]=sd\left[\log(Q)\right]\doteq\frac{\sqrt{2l}}{\sqrt{(n-p_0-l)(n-p_0+2)}}\,. \tag{15}$$

Combined, the approximations of signal in (9) and noise in (15) to be the approximate signal-to-noise ratio in (7) is as follows:

$$\frac{signal}{noise}\doteq\frac{\sqrt{(n-p_0-l)(n-p_0+2)}}{\sqrt{2l}}\left[\log\left(\frac{n-p_0-l}{n-p_0}\right)-\frac{l}{(n-p_0-l)(n-p_0)}+\frac{\alpha l}{n}\right]. \tag{16}$$

In (16), we found that the signal-to-noise ratio of $APIC\alpha$ depends on the value of α in (3). This conclusion is similar to the probability of overfitting, that is if the value of α tends to infinity under the same values of n, p_0, and l, $APIC\alpha$ has a strong signal-to-noise ratio. When we replace the values of α in (16) by 2, $\log(n)$, and 3, we have the approximate signal-to-noise ratios for $AIC, BIC,$ and $KIC,$ respectively. The proof of the signal-to-noise ratio can be confirmed numerically in Table 2. McQuarrie and Tsai (1998) concluded that the signal-to-noise ratios are strong or weak as follows. A strong signal-to-noise ratio refers to a large positive value (often greater than 2) and leads to small probability of overfitting. A weak signal-to-noise ratio usually refers to one that is small (less than 0.5) or negative and results in high probability of overfitting. The model selection criterion that has strong signal-to-noise ratio and lowest probability of overfitting is preferable.

Table 2. Signal-to-noise ratio of $APIC\alpha$ for different values of n, p_0, and l

n	p_0	l	Criteria													
			*APIC*1	*APIC*2	*APIC*3	*APIC*4	*APIC*5	*APIC*6	*APIC*7	*APIC*8	*APIC*9	*APIC*10	*APIC*11	*APIC*12	*APIC*13	*APIC*14
15	3	1	-0.2450	0.3400	0.9250	1.5100	2.0950	2.6800	3.2650	3.8500	4.4350	5.0200	5.6050	6.1900	6.7750	7.3600
15	3	2	-0.3884	0.4004	1.1892	1.9780	2.7668	3.5556	4.3444	5.1333	5.9221	6.7109	7.4997	8.2885	9.0773	9.8661
15	3	3	-0.5291	0.3874	1.3039	2.2204	3.1370	4.0535	4.9700	5.8865	6.8030	7.7195	8.6360	9.5526	10.4691	11.3856
15	3	4	-0.6752	0.3225	1.3203	2.3181	3.3159	4.3136	5.3114	6.3092	7.3070	8.3047	9.3025	10.3003	11.2981	12.2958
15	4	1	-0.3042	0.2333	0.7707	1.3082	1.8457	2.3832	2.9207	3.4582	3.9956	4.5331	5.0706	5.6081	6.1456	6.6831
15	4	2	-0.4734	0.2477	0.9688	1.6899	2.4110	3.1321	3.8532	4.5743	5.2954	6.0166	6.7377	7.4588	8.1799	8.9010
15	4	3	-0.6351	0.1976	1.0302	1.8629	2.6956	3.5282	4.3609	5.1936	6.0262	6.8589	7.6916	8.5242	9.3569	10.1896
15	4	4	-0.8002	0.0992	0.9985	1.8979	2.7973	3.6967	4.5961	5.4955	6.3948	7.2942	8.1936	9.0930	9.9924	10.8917
30	3	1	-0.1132	0.5340	1.1812	1.8284	2.4756	3.1229	3.7701	4.4173	5.0645	5.7117	6.3589	7.0062	7.6534	8.3006
30	3	2	-0.1785	0.7190	1.6166	2.5141	3.4116	4.3092	5.2067	6.1042	7.0017	7.8993	8.7968	9.6943	10.5918	11.4894
30	3	3	-0.2414	0.8356	1.9127	2.9897	4.0667	5.1438	6.2208	7.2978	8.3749	9.4519	10.5289	11.6060	12.6830	13.7600
30	3	4	-0.3054	0.9120	2.1295	3.3470	4.5644	5.7819	6.9994	8.2168	9.4343	10.6518	11.8692	13.0867	14.3041	15.5216
30	4	1	-0.1389	0.4847	1.1083	1.7319	2.3555	2.9791	3.6027	4.2263	4.8500	5.4736	6.0972	6.7208	7.3444	7.9680
30	4	2	-0.2149	0.6492	1.5133	2.3774	3.2415	4.1056	4.9697	5.8338	6.6979	7.5620	8.4261	9.2902	10.1543	11.0184
30	4	3	-0.2861	0.7499	1.7859	2.8219	3.8579	4.8940	5.9300	6.9660	8.0020	9.0380	10.0740	11.1101	12.1461	13.1821
30	4	4	-0.3573	0.8127	1.9827	3.1527	4.3227	5.4927	6.6627	7.8327	9.0027	10.1727	11.3427	12.5127	13.6827	14.8527
100	3	1	-0.0324	0.6569	1.3463	2.0356	2.7250	3.4143	4.1037	4.7930	5.4824	6.1717	6.8611	7.5504	8.2398	8.9291
100	3	2	-0.0510	0.9188	1.8886	2.8584	3.8282	4.7980	5.7678	6.7376	7.7074	8.6772	9.6470	10.6168	11.5866	12.5564
100	3	3	-0.0687	1.1128	2.2942	3.4757	4.6572	5.8387	7.0202	8.2016	9.3831	10.5646	11.7461	12.9276	14.1091	15.2905
100	3	4	-0.0867	1.2703	2.6273	3.9843	5.3413	6.6982	8.0552	9.4122	10.7692	12.1262	13.4831	14.8401	16.1971	17.5541
100	4	1	-0.0396	0.6426	1.3249	2.0072	2.6895	3.3717	4.0540	4.7363	5.4186	6.1008	6.7831	7.4654	8.1477	8.8299
100	4	2	-0.0612	0.8986	1.8584	2.8182	3.7780	4.7378	5.6976	6.6574	7.6171	8.5769	9.5367	10.4965	11.4563	12.4161
100	4	3	-0.0813	1.0880	2.2572	3.4264	4.5957	5.7649	6.9341	8.1034	9.2726	10.4418	11.6111	12.7803	13.9495	15.1187
100	4	4	-0.1011	1.2417	2.5845	3.9274	5.2702	6.6130	7.9559	9.2987	10.6415	11.9844	13.3272	14.6700	16.0129	17.3557

3. Simulation study

In addition to the proofs of probability of overfitting in (6) and the approximate signal-to-noise ratio in (16), we use the simulation data to find the appropriate value of α for $APIC\alpha$ in (3). Four cases of the true multiple regression models in (1) are constructed as follows.

Model 1 (very weakly identifiable true model with the true order $p_0 = 7$):

$y_1 = X_1 + 0.5X_2 + 0.1X_3 + 0.05X_4 + 0.01X_5 + 0.005X_6 + 0.001X_7 + \varepsilon_1$,

Model 2 (weakly identifiable true model with the true order $p_0 = 3$):

$y_2 = X_1 + 0.5X_2 + 0.25X_3 + \varepsilon_2$,

Model 3 (very strongly identifiable true model with the true order $p_0 = 4$):

$y_3 = X_1 + 2X_2 + 2X_3 + 2X_4 + \varepsilon_3$,

Model 4 (strongly identifiable true model with the true order $p_0 = 8$):

$y_4 = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + \varepsilon_4$.

Model 1 and Model 2 represent the weakly identifiable true models which mean they are not easily identified compared to the strongly identifiable true models such as Model 3 and Model 4. In this study, the true variance σ_0^2 in (1) is assumed equal to one. For each model, we consider 1,000 realizations for three levels of the sample sizes which are $n = 15$ (small), $n = 30$ (moderate), and $n = 100$ (large). Ten candidate variables, X_1 to X_{10}, are stored in an $n \times 10$ matrix $\mathbf{X}$ of the candidate model in (2), where X_1 is given as a constant which equals one, followed by nine independent identically distributed normal random variables with zero mean and equal variance-covariance matrix to identity matrix $\mathbf{I}_{10}$. The candidate models include the columns of **X** in a sequentially nested fashion; i.e., columns 1 to *p* define the design matrix for the candidate model with dimension *p*. Over 1,000 realizations, we apply $APIC\alpha$ in (3) with the values of α ranging from 1 to 14 on the datasets y of four models constructed. The probability of order selected by $APIC\alpha$ is a measure used to examine the effects of weak or strong penalty function in the proposed criterion. In addition to above measure, many authors (McQuarrie et al., 1997; McQuarrie, 1999) use the observed L_2 efficiency to assess model selection criterion performance, especially when the true model is very difficult to detect. The observed L_2 distance, scaled by $1/n$, between the true model in (1) and the fitted candidate model in (2) is defined as

$$L_2(p) = \left(\mathbf{X}_0\boldsymbol{\beta}_0 - \mathbf{X}\hat{\boldsymbol{\beta}}\right)' \left(\mathbf{X}_0\boldsymbol{\beta}_0 - \mathbf{X}\hat{\boldsymbol{\beta}}\right)/n.$$

Observed L_2 efficiency is defined by the ratio

$$\text{Observed } L_2 \text{ efficiency} = \frac{\min_{1 \le p \le P} L_2(p)}{L_2(p_s)},$$

where *P* is the class of all possible candidate models, *p* is the rank of fitted candidate model, and p_s is the model selected by specific model selection criterion. The closer the selected model is to the true model, the higher the efficiency. Therefore, the best model selection criterion will select a model which yields high efficiency even in small samples or the true model is weakly identifiable. In order to summarize the results in this study, the average observed L_2 efficiencies over the 1,000 realizations are ranked for $APIC\alpha$ where the values of α range from 1 to 14. The first rank of average observed L_2 efficiencies goes to the highest efficiency criterion and denotes better relative performance. Results of comparing the probability of order selected by $APIC\alpha$ and average observed L_2 efficiencies are shown in Table 3.

From the results of simulation in Table 3 we found that, for Model 1 and Model 2 which are the situations where the true model cannot be easily identified, $APIC\alpha$ with the small value of α (about 1 to 3) gives the greater probability of correct order being selected than the case of large value and also prevents the probability of underfitting. While, the observed L_2 efficiency suggests the large value of α causes the high efficiency of $APIC\alpha$, except when the true model can be specified more easily, such as Model 2, and sample sizes are moderate to large, the small value of α (about 3 to 4) is preferable. For Model 3 and Model 4 which are the situations where the true model is strongly identifiable, the value of α should be large (at least 8), except when the regression coefficients are not large enough, such as Model 4, and the sample sizes are small to moderate, the value of α should be moderate (about 4 to 6).

For all models, if the value of α tends to infinity, the probability of overfitted tends to decrease whereas the probability of underfitting tends to increase. The point that has the optimal probability of over- and underfitting always presents the maximum probability of correct order being selected.

4. Conclusions, discussion, and further study

In this paper, we study the penalty functions based on the well-known model selection criteria, $AIC, BIC,$ and $KIC,$ which can be unified in the form of the log likelihood function of the maximum likelihood estimator of σ^2 plus a penalty function, called Adjusted Penalty Information Criterion, i.e.,

$$APIC\alpha = \log\left(\hat{\sigma}^2\right) + \frac{\alpha\left(p+1\right)}{n}$$

when the values of α are equal to 2, $\log\left(n\right)$, and 3, $APIC\alpha$ becomes $AIC, BIC,$ and $KIC,$ respectively. Each criterion has a different value due to its penalty function, the differences in strong or weak penalty affecting the probabilities of over- and underfitting, including the problem of signal-to-noise ratio being weak.

The theoretical results show that, when the value of α tends to infinity, the probability of overfitting tends to zero and the signal-to-noise ratio tends to strong. At the same time, the results of simulation based on values of α for $APIC\alpha$ ranging from 1 to 14 suggest that, when the true model is weakly identifiable, the value of α should be small to give a high probability of correct order being selected and to prevent the probability of underfitting. However in the case of the true model is very difficult to detect, such as Model 1; none of the criteria correctly identify the true model more than 8% of the time. As a result, the probability of correct order being selected may not be meaningful. For this reason, we used the observed L_2 efficiency to assess the appropriate value of α. This measure suggests the large value of α causes the high efficiency of $APIC\alpha$ which indicates that the correct model is also the closet model, except when the true model can be specified more easily, such as Model 2, and sample sizes are moderate to large, then the small value of α is preferable. For the strongly identifiable true model, the large value of α performs well. Because the problem of underfitting does not occur in this situation, the underfitted order often gives the maximum value of the estimated mean squared error and hence, under the model selection criterion, it is not possible to select the underfitted model. In the situation where the regression coefficients are not large enough, such as Model 4, and the sample sizes are small to moderate, the value of α should be moderate.

In further work, we attempt to construct the model selection criteria to overcome the probability of over- and underfitting in the multivariate regression and simultaneous equations models.

Acknowledgements

This project is financial supported by The Thailand Research Fund, Office of The Higher Education Commission, and Thaksin University under grant No. MRG5480044. I would like to give special thanks to Prof. Dr. Joseph E. Cavanaugh, Department of Biostatistics, College of Public Health, University of Iowa, USA, Assoc. Prof. Dr. Jirawan Jitthavech, and Assoc. Prof. Dr. Vichit Lorchirachoonkul, National Institute of Development Administration (NIDA), THAILAND for all comments and suggestions. The author wishes to express his gratitude to an anonymous referee for carefully reading and constructively critiquing this manuscript.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Proc. *2nd Int. Symp. on Information Theory*. Akademia Kiado, Budapest, pp. 267–281.

Cavanaugh, J.E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Stat. Probabil. Lett.* **42**, 333–343.

Cavanaugh, J.E. (2004). Criteria for linear model selection based on Kullback's symmetric divergence. *Aust. NZ. J. Stat.* **46**, 257–274.

McQuarrie, A.D., Shumway, R., Tsai, C.L. (1997). The model selection criterion AICu. *Stat. Probabil. Lett.* **34**, 285–292.

McQuarrie, A.D., Tsai, C.L. (1998). *Regression and time series model selection*. Singapore: World Scientific.

McQuarrie, A.D. (1999). A small-sample correction for the Schwarz SIC model selection criterion. *Stat. Probabil. Lett.* **44**, 79–86.

Neath, A., Cavanaugh, J.E. (1997). Regression and time series model selection using variants of the Schwarz information criterion. *Commun. Stat-Theor M.* **26**, 559–580.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464.

Seghouane, A.K. (2006). A note on overfitting properties of *KIC* and *KIC*c. *Signal Process* **86**, 3055–3060.

Table 3. Probability of the order selected by $APIC\alpha$ and average observed L_2 efficiencies over 1,000 realizations

Model	n	Order	Criteria													
			APIC1	*APIC2*	*APIC3*	*APIC4*	*APIC5*	*APIC6*	*APIC7*	*APIC8*	*APIC9*	*APIC10*	*APIC11*	*APIC12*	*APIC13*	*APIC14*
1	15	Underfitted	0.191	0.560	0.809	0.931	0.980	0.997	0.997	0.999	1.000	1.000	1.000	1.000	1.000	1.000
very		Correct	**0.055**	0.044	0.018	0.006	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
weakly		Overfitted	0.754	0.396	0.173	0.063	0.019	0.003	0.003	0.001	0.000	0.000	0.000	0.000	0.000	0.000
identifiable		Ave. L_2 eff.	0.266	0.483	0.687	0.811	0.890	0.922	0.937	0.952	0.960	0.961	0.962	0.964	0.965	**0.966**
(true order		Rank	14	13	12	11	10	9	8	7	6	5	4	3	2	**1**
$p_0 = 7$)	30	Underfitted	0.441	0.853	0.982	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		Correct	**0.067**	0.029	0.006	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		Overfitted	0.492	0.118	0.012	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		Ave. L_2 eff.	0.386	0.646	0.795	0.858	0.885	0.913	0.923	0.929	0.934	0.935	0.939	0.941	**0.942**	**0.942**
		Rank	14	13	12	11	10	9	8	7	6	5	4	3	**1.5**	**1.5**
	100	Underfitted	0.588	0.927	0.996	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		Correct	**0.079**	0.022	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		Overfitted	0.333	0.051	0.003	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		Ave. L_2 eff.	0.470	0.642	0.703	0.723	0.735	0.748	0.756	0.765	0.772	0.778	0.782	0.784	0.784	**0.786**
		Rank	14	13	12	11	10	9	8	7	6	5	4	3	2	**1**
2	15	Underfitted	0.058	0.288	0.545	0.721	0.826	0.890	0.930	0.955	0.970	0.978	0.986	0.990	0.990	0.992
weakly		Correct	0.038	0.136	**0.167**	0.158	0.123	0.090	0.061	0.042	0.030	0.022	0.014	0.010	0.010	0.008
identifiable		Overfitted	0.904	0.576	0.288	0.121	0.051	0.020	0.009	0.003	0.000	0.000	0.000	0.000	0.000	0.000
(true order		Ave. L_2 eff.	0.301	0.469	0.615	0.703	0.746	0.771	0.786	0.797	0.802	0.805	0.808	0.810	0.810	**0.811**
$p_0 = 3$)		Rank	14	13	12	11	10	9	8	7	6	5	4	2.5	2.5	**1**
	30	Underfitted	0.102	0.376	0.584	0.712	0.799	0.857	0.900	0.927	0.941	0.959	0.972	0.978	0.982	0.990
		Correct	0.124	**0.282**	0.271	0.234	0.183	0.135	0.096	0.069	0.057	0.039	0.028	0.022	0.018	0.010
		Overfitted	0.774	0.342	0.145	0.054	0.018	0.008	0.004	0.004	0.002	0.002	0.000	0.000	0.000	0.000
		Ave. L_2 eff.	0.402	0.602	0.663	**0.670**	0.659	0.648	0.642	0.643	0.643	0.646	0.650	0.650	0.652	0.656
		Rank	14	13	2	**1**	3	8	12	11	10	9	7	6	5	4
	100	Underfitted	0.029	0.118	0.223	0.333	0.417	0.499	0.582	0.652	0.704	0.768	0.814	0.847	0.876	0.892
		Correct	0.271	0.575	**0.663**	0.628	0.565	0.496	0.415	0.346	0.295	0.231	0.186	0.153	0.124	0.108
		Overfitted	0.700	0.307	0.114	0.039	0.018	0.005	0.003	0.002	0.001	0.001	0.000	0.000	0.000	0.000
		Ave. L_2 eff.	0.515	0.748	**0.806**	0.782	0.732	0.679	0.616	0.562	0.524	0.479	0.449	0.427	0.407	0.397
		Rank	9	3	**1**	2	4	5	6	7	8	10	11	12	13	14

Table 3. (Continued)

Model	*n*	Order	Criteria													
			APIC1	*APIC2*	*APIC3*	*APIC4*	*APIC5*	*APIC6*	*APIC7*	*APIC8*	*APIC9*	*APIC10*	*APIC11*	*APIC12*	*APIC13*	*APIC14*
3 very strongly identifiable (true order $p_0 = 4$)	15	Underfitted	0.000	0.000	0.000	0.002	0.005	0.008	0.010	0.020	0.030	0.042	0.062	0.099	0.144	0.192
		Correct	0.091	0.312	0.558	0.728	0.851	0.909	0.944	**0.948**	0.946	0.942	0.929	0.895	0.851	0.805
		Overfitted	0.909	0.688	0.442	0.270	0.144	0.083	0.046	0.032	0.024	0.016	0.009	0.006	0.005	0.003
		Ave. L_2 eff.	0.435	0.568	0.719	0.828	0.906	0.942	0.964	**0.964**	0.961	0.955	0.941	0.909	0.867	0.823
		Rank	14	13	12	10	8	5	2	**1**	3	4	6	7	9	11
	30	Underfitted	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		Correct	0.223	0.602	0.789	0.890	0.937	0.961	0.978	0.984	0.989	0.991	0.996	**1.000**	**1.000**	**1.000**
		Overfitted	0.777	0.398	0.211	0.110	0.063	0.039	0.022	0.016	0.011	0.009	0.004	0.000	0.000	0.000
		Ave. L_2 eff.	0.525	0.753	0.868	0.928	0.958	0.973	0.984	0.988	0.993	0.994	0.997	**1.000**	**1.000**	**1.000**
		Rank	14	13	12	11	10	9	8	7	6	5	4	**2**	**2**	**2**
	100	Underfitted	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		Correct	0.307	0.684	0.855	0.932	0.961	0.982	0.989	0.993	0.997	0.997	0.999	0.999	**1.000**	**1.000**
		Overfitted	0.693	0.316	0.145	0.068	0.039	0.018	0.011	0.007	0.003	0.003	0.001	0.001	0.000	0.000
		Ave. L_2 eff.	0.577	0.805	0.910	0.955	0.974	0.988	0.992	0.995	0.998	0.998	0.999	0.999	**1.000**	**1.000**
		Rank	14	13	12	11	10	9	8	7	5.5	5.5	3.5	3.5	**1.5**	**1.5**
4 strongly identifiable (true order $p_0 = 8$)	15	Underfitted	0.011	0.036	0.094	0.171	0.300	0.503	0.680	0.834	0.922	0.968	0.995	0.997	0.998	0.999
		Correct	0.253	0.444	0.532	**0.555**	0.517	0.384	0.251	0.140	0.069	0.028	0.003	0.002	0.001	0.000
		Overfitted	0.736	0.520	0.374	0.274	0.183	0.113	0.069	0.026	0.009	0.004	0.002	0.001	0.001	0.001
		Ave. L_2 eff.	0.788	0.815	**0.830**	0.812	0.746	0.602	0.449	0.311	0.224	0.171	0.134	0.129	0.124	0.121
		Rank	4	2	**1**	3	5	6	7	8	9	10	11	12	13	14
	30	Underfitted	0.001	0.001	0.003	0.006	0.011	0.019	0.047	0.104	0.209	0.350	0.560	0.736	0.871	0.947
		Correct	0.489	0.759	0.875	0.932	0.964	**0.967**	0.944	0.895	0.790	0.649	0.440	0.264	0.129	0.053
		Overfitted	0.510	0.240	0.122	0.062	0.025	0.014	0.009	0.001	0.001	0.001	0.000	0.000	0.000	0.000
		Ave. L_2 eff.	0.848	0.912	0.950	0.969	**0.982**	0.981	0.962	0.917	0.820	0.688	0.485	0.317	0.185	0.109
		Rank	8	7	5	3	**1**	2	4	6	9	10	11	12	13	14
	100	Underfitted	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		Correct	0.593	0.815	0.925	0.966	0.985	0.995	0.999	0.999	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		Overfitted	0.407	0.185	0.075	0.034	0.015	0.005	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
		Ave. L_2 eff.	0.857	0.919	0.960	0.980	0.991	0.997	0.999	0.999	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		Rank	14	13	12	11	10	9	7.5	7.5	**3.5**	**3.5**	**3.5**	**3.5**	**3.5**	**3.5**

Note: Boldface type indicates the maximum value.

Model selection criterion based on Kullback-Leibler's symmetric divergence for simultaneous equations model

Warangkhana Keerativibool [a*] and Jirawan Jitthavech [b]

[a] Department of Mathematics and Statistics, Faculty of Science,
Thaksin University, Phatthalung, Thailand.
[b] School of Applied Statistics, National Institute of Development Administration,
Bangkok, Thailand

[*] Corresponding E-Mail Address: warang27@gmail.com

Abstract

Moving average in the error of simultaneous equations model (SEM) is a crucial problem to make the estimators from the ordinary least squares (OLS) method are not efficient. For this reason, we proposed the transformation matrix in order to correct the first-order moving average, MA(1), that generated in the fitted model and to recover the one lost observation in a SEM. After the errors are transformed to be independent, the Kullback information criterion for select the appropriate SEM, called SKIC, to be going to derive. This criterion is constructed based on the symmetric divergence which obtained by sum of the two directed divergences. The symmetric divergence is arguably more sensitive than either of its individual components. The performance of the proposed criterion, SKIC, is examined relative to SAIC proposed by Keerativibool (2009). The results of simulation study show that the errors of the model after transformation are independent and SKIC convincingly outperformed SAIC, because SAIC has a tendency to overfit the order of the model than SKIC.

Keywords: First-order moving average MA(1); Transformation matrix; Simultaneous equations model (SEM); Kullback information criterion for a system of SEM (SKIC).

1. Introduction

A system of simultaneous equations model (SEM) is a model that contains variables with two way flows of influence characteristics which most common and straightforward methods for modelling the economic data. The endogenous explanatory variable will become stochastic and will correlate with the error terms of the equation in which it appears as an explanatory variable. Most problems in the errors of SEM are the autocorrelated (AR) error or moving average (MA) error or both (ARMA). When these problems occur, the ordinary least squares (OLS) estimators cannot be used because they are not efficient (Gujarati, 2006). Therefore, in this paper we will propose a transformation matrix to correct the first-order moving average, MA(1), which generated in the fitted model and to recover the one lost observation in a SEM. After the errors are transformed to be independent, we consider the problem of fitting a parametric model to an observed data set. This problem requires two tasks, determination of the order of the model and estimation of these parameters. In real life, we may not know what the true model is, but we hope to find a model that is a reasonably accurate representation. The crucial part of this fitting problem is to determine the order of the model. Such determination is often facilitated by the use of a model selection criterion where one only has to evaluate two simple terms that trade-off quality of fit to the data and model's complexity. A lot of previous literary attention to the issue of model selection, the widespread criterion for choosing the best model in univariate and multivariate regression analysis is the Akaike information criterion (AIC) (Akaike, 1973, 1974; Bedrick and Tsai, 1994). The corrected version of the AIC (AIC$_c$) (Hurvich and Tsai, 1989) is extended for the case of small sample. AIC and AIC$_c$ were designed, respectively, to be asymptotically and exactly unbiased estimator of a variant of Kullback-Leibler's directed divergence between the true model and a fitted candidate model. The development of a new family of selection criteria, Kullback information criterion (KIC) and the corrected version of the KIC (KIC$_c$), are the criteria constructed to target a symmetric divergence. This divergence is an alternate of directed divergence, obtained by sum of the two directed divergences, which arguably more sensitive than either of its individual components (Cavanaugh, 1999, 2004; Seghouane and Bekara, 2004; Hafidi and Mkhadri, 2006). Recently it has developed the KICc more in the case of vector autoregressive and multivariate regression (Hafidi, 2006; Seghouane, 2006). Unfortunately, as of now, there is only one criterion, Akaike information criterion for a system of SEM (SAIC), for selecting a workable system of SEM (Keerativibool, 2009). With this motivation, we will propose the model selection criterion, called Kullback information criterion for a system of SEM (SKIC), which serves as an asymptotically unbiased estimator of a variant of Kullback-Leibler's symmetric divergence between the true model and the fitted candidate model. The remainder of this paper is organized as follows. In Section 2, we propose a transformation matrix in order to correct the MA(1) problem in the errors of a SEM. The criterion, SKIC, for selecting the best system of SEM is also proposed in this section. In Section 3, we simulate 1,000 samples of SEM in order to study the frequency of order being selected and the observed L_2 efficiency of the proposed criterion, SKIC, relative to SAIC proposed by Keerativibool (2009). Finally, Section 4 is the conclusions, discussion, and future works.

2. Methodology

The structural-form and reduced-form of the SEM (Greene, 2008) may be represented, respectively, as follows:

$$\mathbf{Y\Gamma} + \mathbf{XB} = \mathbf{U} \text{ and } \mathbf{Y} = \mathbf{X\Pi} + \mathbf{V}, \tag{1}$$

where $\mathbf{Y}$ is a $T \times M$ matrix of observations, $\mathbf{X}$ is a $T \times K$ design matrix of full-column rank, $\mathbf{\Gamma}$ is an $M \times M$ nonsingular matrix of coefficients of endogenous variables, $\mathbf{B}$ is a $K \times M$ matrix of coefficients of predetermined variables, $\mathbf{\Pi} = -\mathbf{B\Gamma}^{-1}$ is a $K \times M$ matrix of unknown parameters, $\mathbf{U}$ and $\mathbf{V} = \mathbf{U\Gamma}^{-1}$ are the $T \times M$ matrices of MA(1) and contemporaneously correlated errors. The j^{th} equation vector of reduced-form model in (1) is

$$\mathbf{y}_j = \mathbf{X}\boldsymbol{\pi}_j + \mathbf{v}_j, \; j = 1, 2, \dots, M, \tag{2}$$

where $\mathbf{y}_j$ is a $T \times 1$ observation vector, $\boldsymbol{\pi}_j$ is a $K \times 1$ parameter vector, and $\mathbf{v}_j$ is a $T \times 1$ vector of MA(1) and contemporaneously correlated errors. Each element v_{tj} in the vector $\mathbf{v}_j$ is in the form of MA(1),

$$v_{tj} = \varepsilon_{tj} - \theta_j \varepsilon_{t-1,j}, \; t = 1, 2, \dots, T, \; j = 1, 2, \dots, M, \tag{3}$$

where T is the number of observations in each equation, M is the number of equations, the error $\varepsilon_{t-1,j}$ is called the first-lag of error ε_{tj}, the MA(1) parameter θ_j of the model must satisfy the following condition to ensure the invertibility of the error terms (Box et al., 1994),

$$\left|\theta_j\right| < 1. \tag{4}$$

The error ε_{tj} in (3) is an independent identically distributed random variable, obeying

$$\varepsilon_{tj} \sim N\left(0, \sigma_{jj}\right), \tag{5}$$

so that

$$\boldsymbol{\varepsilon}_t' = \begin{bmatrix} \varepsilon_{t1} & \varepsilon_{t2} & \cdots & \varepsilon_{tM} \end{bmatrix} \sim N_M(\mathbf{0}, \boldsymbol{\Sigma}), \tag{6}$$

where $\boldsymbol{\Sigma}$ is the $M \times M$ contemporaneous covariance matrix of the error terms which is nonsingular and is of positive symmetric definite matrix. It is noteworthy that the values of v_{1j} in the MA(1) model in (3) depend on the values of ε_{0j}, which is unknown. The recovery of v_{1j} will be shown in Theorem 1.

For all M equations, the models in (2) can be represented as a stacked model as follows:

$$\mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\pi} + \mathbf{v}, \tag{7}$$

where $\mathbf{y}$ is a $TM \times 1$ observation vector consisting of M $(T \times 1)$ $\mathbf{y}_j$ vectors, $\tilde{\mathbf{X}}$ is a $TM \times KM$ diagonal matrix of rank KM consisting of M $(T \times K)$ identical $\mathbf{X}$ matrices, $\boldsymbol{\pi}$ is a $KM \times 1$ unknown parameter vector consisting of M $(K \times 1)$ $\boldsymbol{\pi}_j$ vectors, and $\mathbf{v}$ is a $TM \times 1$ MA(1) and contemporaneously correlated error vector consisting of M $(T \times 1)$ $\mathbf{v}_j$ vectors. The transformation matrix to correct the MA(1) correlated error vector is given in Theorem 1.

Theorem 1: The $TM \times TM$ transformation matrix $\mathbf{P}$, used to correct the MA(1) problem in a SEM, is defined by

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P}_M \end{bmatrix}, \tag{8}$$

where the $T \times T$ transformation matrix $\mathbf{P}_j$ for the j^{th} equation is

$$\mathbf{P}_j = \begin{bmatrix} \frac{1}{\sqrt{1+\theta_j^2}} & 0 & 0 & 0 & \cdots & 0 \\ \theta_j & 1 & 0 & 0 & \cdots & 0 \\ \theta_j^2 & \theta_j & 1 & 0 & \cdots & 0 \\ \theta_j^3 & \theta_j^2 & \theta_j & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_j^{T-1} & \theta_j^{T-2} & \theta_j^{T-3} & \theta_j^{T-4} & \cdots & 1 \end{bmatrix}. \tag{9}$$

The transformation matrix $\mathbf{P}$ in (8) is used to transform $\mathbf{y}$ and $\tilde{\mathbf{X}}$ in (7) to be $\mathbf{y}^*$ and $\tilde{\mathbf{X}}^*$, respectively, such that the MA(1) of the errors $\mathbf{v}$ in (7) is eliminated, to give the model

$$\mathbf{y}^* = \tilde{\mathbf{X}}^*\boldsymbol{\pi} + \boldsymbol{\varepsilon}, \tag{10}$$

where $\mathbf{y}^* = \mathbf{P}\mathbf{y}$, $\tilde{\mathbf{X}}^* = \mathbf{P}\tilde{\mathbf{X}}$, $E\left(\boldsymbol{\varepsilon} \middle| \tilde{\mathbf{X}}^*\right) = \mathbf{0}$, and $E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \middle| \tilde{\mathbf{X}}^*\right) = \boldsymbol{\Sigma} \otimes \mathbf{I}_T$.

Suppose that the transformed model in (10) is called the candidate model, then the true model can be given as

$$\mathbf{y}^* = \tilde{\mathbf{X}}^*\boldsymbol{\pi}_0 + \boldsymbol{\varepsilon}_0. \tag{11}$$

The notations in (10) and (11) are defined as follows: $\mathbf{y}^*$ is a $TM \times 1$ observation vector consisting of M $(T \times 1)$ $\mathbf{y}_j^*$ (or $\mathbf{P}_j\mathbf{y}_j$) vectors, $\tilde{\mathbf{X}}^*$ is a $TM \times KM$ diagonal matrix consisting of M $(T \times K)$ $\mathbf{X}_j^*$ (or $\mathbf{P}_j\mathbf{X}$) matrices, $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_0$ are the $KM \times 1$ unknown parameter vectors, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}_0$ are the $TM \times 1$ independent identically distributed normal random vectors.

For the derivations of the criterion, the true model is assumed to be correctly specified or overfitted by all the candidate models. This means that $\boldsymbol{\pi}_0$ has K_0M nonzero entries with $0 < K_0M \le KM$ and the rest of $(K - K_0)M$ entries are equal to zero. The Kullback information criterion for a system of SEM (SKIC) is given in Theorem 2.

Theorem 2. When the MA(1) problem is adjusted by the transformation matrix **P**, the Kullback information criterion for a system of SEM defined by

$$\text{SKIC} = T\log\left|\hat{\boldsymbol{\Sigma}}\right| + \frac{TM(2K+M+1)}{T-K-M-1} + TM\log\left(\frac{2T}{2T-2K-M+1}\right) + \frac{2TM}{2T-2K-M+1} \quad (12)$$

is called an asymptotically unbiased estimator of the Kullback-Leibler's symmetric divergence.

3. Simulation study

The model to consider in this study is a system of three SEM (M = 3) and the errors of the model appear the MA(1) problem,

$$\begin{aligned} y_{t1} &= 1 + 2x_{t2} + 3x_{t3} + 4x_{t4} + v_{t1} \\ y_{t2} &= 1 - 0.5x_{t2} - 5x_{t3} - 1.5x_{t4} + v_{t2} \\ y_{t3} &= 1 + x_{t2} + x_{t3} + x_{t4} + v_{t3}, \end{aligned} \quad (13)$$

where $t = 1, 2, \ldots, T = 15$ for the small sample size, $t = 1, 2, \ldots, T = 30$ for the medium sample size, and $t = 1, 2, \ldots, T = 100$ for the large sample size. The steps for simulation and all results are as follows.

1. Using the IML procedure of SAS programming to generate 150,000 vectors of the 3×1 multivariate normal $\boldsymbol{\varepsilon}_t$ in (6), given zero mean vector, the correlation coefficients of the errors between the equations are

$$\rho_{12} = 0.9,\ \rho_{13} = 0.7,\ \rho_{23} = 0.8,$$

and the variances-covariances of the errors are

$$\sigma_{11} = 0.9^2 = 0.81,\ \sigma_{22} = 0.8^2 = 0.64,\ \sigma_{33} = 0.7^2 = 0.49,$$

$$\sigma_{12} = \rho_{12}\sqrt{\sigma_{11}\sigma_{22}} = 0.648,\ \sigma_{13} = \rho_{13}\sqrt{\sigma_{11}\sigma_{33}} = 0.441,\ \sigma_{23} = \rho_{23}\sqrt{\sigma_{22}\sigma_{33}} = 0.448,$$

then, the form to generate $\boldsymbol{\varepsilon}_t$ in (6) is represented by

$$\boldsymbol{\varepsilon}_t = \begin{bmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \\ \varepsilon_{t3} \end{bmatrix} \sim N_3\left(\mathbf{0},\ \boldsymbol{\Sigma} = \begin{bmatrix} 0.81 & 0.648 & 0.441 \\ 0.648 & 0.64 & 0.448 \\ 0.441 & 0.448 & 0.49 \end{bmatrix}\right).$$

2. Using the multivariate normal error $\varepsilon_{t1}, \varepsilon_{t2}$, and ε_{t3} in Step 1 to construct two series of the MA(1) and contemporaneously correlated errors, v_{t1}, v_{t2}, and v_{t3}, as follows:

$v_{t1} = \varepsilon_{t1} - 0.5\varepsilon_{t-1,1}$, $v_{t2} = \varepsilon_{t2} - 0.6\varepsilon_{t-1,2}$, and $v_{t3} = \varepsilon_{t3} - 0.7\varepsilon_{t-1,3}$, (1st series) (14a)

$v_{t1} = \varepsilon_{t1} + 0.6\varepsilon_{t-1,1}$, $v_{t2} = \varepsilon_{t2} + 0.7\varepsilon_{t-1,2}$, and $v_{t3} = \varepsilon_{t3} + 0.8\varepsilon_{t-1,3}$, (2nd series) (14b)

for $t = 1, 2, \ldots, 150{,}000$ and ε_{0j} is arbitrarily given to be zero for all $j = 1, 2, 3$. Split the series of errors v_{t1}, v_{t2}, and v_{t3} in sequence to preserve the MA(1) problem into 1,000 samples, each of which consists of three levels of sample sizes, T = 15, 30, 100 observations. Estimate the MA(1) parameters and test the properties of MA(1) by the MODEL and ARIMA procedures in SAS version 9.1. Discard the samples that fail the test, and retain only 1,000 samples for further study.

3. Using the RANNOR function of SAS programming to generate the independent variables x_{t2} until $x_{t,10}$ about 150,000 observations to be the normal random variables with zero mean and variance equal to one where the relevant independent variables are x_{t2}, x_{t3}, and x_{t4} and irrelevant independent variables are x_{t5} until $x_{t,10}$. Again, split the series of independent variables x_{t2} until $x_{t,10}$ in sequence into 1,000 samples, each of which consists of 15, 30, 100 observations. For this study, x_{t1} is given as a constant which equals one. Test the multicollinearity problem for the series of independent variables and then discard the samples that fail the test, retain only 1,000 samples for further study.

4. Using the corresponding relevant independent variables x_{t2}, x_{t3}, and x_{t4} obtained in Step 3 and two series of the MA(1) errors obtained in Step 2 to construct the dependent variables described in (13).

5. Using the estimated values of MA(1) parameters obtained in Step 2 to construct the estimate of transformation matrix $\mathbf{P}_j$ in (9) for each sample. Apply this transformation matrix to transform the SEM in Step 4 to give the stack of transformed model as shown in (10). Test the MA(1) problem and the multivariate normality for the errors of the model by the ARIMA and MODEL procedures, respectively. The test shows that the errors of all transformed samples are independent. Therefore, we can say that the transformation matrix **P** in (8) has the power of transformation equal to 100%.

6. Using the assumption of nested model to construct the candidate models which are the models include the columns of independent variables in a sequentially nested fashion; i.e., columns 1 to *K* define the design matrix for the candidate model with dimension *K*. For 1,000 transformed samples, we estimate the parameters of the transformed model by the GLS method. Then calculate SKIC in (3.5) and SAIC proposed by Keerativibool (2009),

$$\text{SAIC} = T\log\left|\hat{\mathbf{\Sigma}}_{UE}\right| + M\left(K+M+3\right), \tag{15a}$$

where $\hat{\mathbf{\Sigma}}_{UE} = \frac{T}{T-K}\hat{\mathbf{\Sigma}}$, $\hat{\mathbf{\Sigma}}_{UE}$ and $\hat{\mathbf{\Sigma}}$ represent the estimated contemporaneous covariance matrix of the error terms by the methods of unbiased estimator and maximum likelihood estimator, respectively. Therefore SAIC in (15a) can be rewritten as

$$\text{SAIC} = T\log\left|\hat{\mathbf{\Sigma}}\right| + TM\log\left(\frac{T}{T-K}\right) + M\left(K+M+3\right). \tag{15b}$$

The candidate model that has the minimum value of model selection criterion is called the best model. Model selection criterion performance is examined by a measure of counting the frequency of order being selected. The results of comparing are shown in Table 1.

7. Calculate the observed L_2 distance, scaled by $1/T$, between the true model in (11) and the candidate model in (10) which was defined by McQuarrie et al. (1997) and McQuarrie (1999),

$$L_2\left(k\right) = \frac{1}{T}\left(\boldsymbol{\pi}_0 - \hat{\boldsymbol{\pi}}\right)' \tilde{\mathbf{X}}^{*\prime}\left(\hat{\mathbf{\Sigma}}^{-1}\otimes\mathbf{I}_{\mathrm{T}}\right)\tilde{\mathbf{X}}^{*}\left(\boldsymbol{\pi}_0 - \hat{\boldsymbol{\pi}}\right),$$

and calculate the observed L_2 efficiency which defined as

$$\text{Observed } L_2 \text{ efficiency } = \frac{\min_{1\le k\le K} L_2\left(k\right)}{L_2\left(k_s\right)},$$

where *K* is the class of all possible candidate models, *k* is the rank of fitted candidate model, and k_s is the model selected by specific model selection criterion. The closer the selected model is to the true model, the higher the efficiency. Therefore, the best model selection criterion will select a model which yields high efficiency even in small samples. For 1,000 transformed samples, the results of comparing the observed L_2 efficiency are shown in Table 2.

Table 1. Frequency of the model order being selected by SAIC and SKIC for 1,000 samples

T	Series of Errors v_{tj}	Criteria	*K*								
			2	3	4	5	6	7	8	9	10
15	(14a)	SAIC	0	0	832	75	30	15	16	2	30
		SKIC	0	0	**1000**	0	0	0	0	0	0
15	(14b)	SAIC	0	0	809	98	32	13	18	2	28
		SKIC	0	0	**1000**	0	0	0	0	0	0
30	(14a)	SAIC	0	0	919	60	13	6	2	0	0
		SKIC	0	0	**999**	1	0	0	0	0	0
30	(14b)	SAIC	0	0	886	86	20	6	2	0	0
		SKIC	0	0	**994**	6	0	0	0	0	0
100	(14a)	SAIC	0	0	952	39	9	0	0	0	0
		SKIC	0	0	**1000**	0	0	0	0	0	0
100	(14b)	SAIC	0	0	910	55	20	7	5	0	3
		SKIC	0	0	**982**	12	5	0	0	0	1

Note: Boldface type indicates the maximum frequency of correct order being selected.

Table 2. Average and standard deviation of the observed L_2 efficiency over 1,000 samples

T	Series of Errors v_{tj}	Criteria	Statistics	
			Ave. L_2 eff.	S.D. L_2 eff.
15	(14a)	SAIC	0.7762	0.3170
		SKIC	**0.8843**	**0.2060**
15	(14b)	SAIC	0.7213	0.3486
		SKIC	**0.8293**	**0.2749**
30	(14a)	SAIC	0.9436	0.1718
		SKIC	**0.9860**	**0.0868**
30	(14b)	SAIC	0.8999	0.2341
		SKIC	**0.9487**	**0.1822**
100	(14a)	SAIC	0.9757	0.1113
		SKIC	**1.0000**	**0.0005**
100	(14b)	SAIC	0.9527	0.1581
		SKIC	**0.9894**	**0.0810**

Note: Boldface type indicates the best performance.

8. The results of the frequency of correct order being selected from Steps 6 in Table 1 can be concluded that the performance of SKIC in (12) convincingly outperformed SAIC in (15b) for all three levels of the sample sizes (*T* = 15, 30, 100) and two series of the MA(1) and contemporaneously correlated errors v_{tj} in (14a) and (14b), because SAIC has a tendency to

overfit the order of the model than SKIC. The results of the observed L_2 efficiency from Steps 7 in Table 2 also confirm that SKIC has a large observed L_2 efficiency and small standard deviation of the observed L_2 efficiency than SAIC, then SKIC is likely better than SAIC. In Table 3, we show the average and standard deviation of SAIC and SKIC for 1,000 transformed samples. In this table we found that SAIC presents a large negative bias than SKIC that maybe the main reason for the number of correct model order being selected is less.

Table 3. Average and standard deviation of SAIC and SKIC for 1,000 samples of the sample size *T* and the series of errors v_{tj} in (14a) and (14b)

T = 15 and errors v_{tj} in (14a)					T = 15 and errors v_{tj} in (14b)				
K	SAIC		SKIC		K	SAIC		SKIC	
	Ave.	S.D.	Mean	S.D.		Ave.	S.D.	Mean	S.D.
2	6.295	0.977	8.281	0.977	2	7.215	1.310	9.201	1.311
3	3.290	0.998	6.443	0.998	3	3.888	1.232	7.041	1.232
4	**-2.351**	0.862	**2.309**	0.862	4	**-2.300**	0.903	**2.359**	0.903
5	-1.934	0.904	4.732	0.904	5	-1.919	0.949	4.747	0.949
6	-1.507	0.964	7.954	0.964	6	-1.493	1.006	7.968	1.006
7	-1.075	1.023	12.541	1.023	7	-1.066	1.072	12.549	1.072
8	-0.649	1.160	19.800	1.160	8	-0.648	1.174	19.801	1.174
9	1.434	1.375	35.330	1.375	9	1.577	1.406	35.473	1.406
10	0.185	1.529	73.700	1.529	10	0.143	1.481	73.659	1.481

T = 30 and errors v_{tj} in (14a)					T = 30 and errors v_{tj} in (14b)				
K	SAIC		SKIC		K	SAIC		SKIC	
	Ave.	S.D.	Mean	S.D.		Ave.	S.D.	Mean	S.D.
2	6.197	0.875	6.824	0.875	2	7.103	1.259	7.730	1.259
3	2.967	0.859	3.916	0.859	3	3.617	1.109	4.566	1.109
4	**-3.131**	0.494	**-1.827**	0.494	4	**-3.065**	0.522	**-1.762**	0.522
5	-2.938	0.501	-1.243	0.501	5	-2.885	0.529	-1.191	0.529
6	-2.734	0.509	-0.606	0.509	6	-2.685	0.533	-0.557	0.533
7	-2.528	0.527	0.081	0.527	7	-2.485	0.545	0.124	0.545
8	-2.306	0.543	0.840	0.543	8	-2.275	0.555	0.872	0.555
9	-0.309	0.656	3.440	0.656	9	-0.168	0.704	3.581	0.704
10	-1.846	0.559	2.582	0.559	10	-1.834	0.585	2.594	0.585

T = 100 and errors v_{tj} in (14a)					T = 100 and errors v_{tj} in (14b)				
K	SAIC		SKIC		K	SAIC		SKIC	
	Ave.	S.D.	Mean	S.D.		Ave.	S.D.	Mean	S.D.
2	6.104	0.617	6.241	0.617	2	7.028	1.034	7.166	1.034
3	2.721	0.570	2.927	0.570	3	3.453	0.898	3.659	0.898
4	**-3.752**	0.265	**-3.476**	0.265	4	**-3.718**	0.293	**-3.442**	0.293
5	-3.693	0.266	-3.344	0.266	5	-3.664	0.289	-3.315	0.289
6	-3.634	0.267	-3.210	0.267	6	-3.610	0.288	-3.187	0.288
7	-3.574	0.267	-3.074	0.267	7	-3.552	0.288	-3.053	0.288
8	-3.514	0.267	-2.936	0.267	8	-3.496	0.284	-2.918	0.284
9	-1.369	0.383	-0.711	0.383	9	-1.041	0.448	-0.383	0.448
10	-3.392	0.271	-2.652	0.271	10	-3.379	0.279	-2.638	0.279

Note: Boldface type indicates the minimum average value of SAIC and SKIC.

4. Conclusions, discussion, and future works

In this paper, the transformation matrix in order to correct the MA(1) problem and to recover the one lost observation along with the consideration of contemporaneous correlation in a SEM is proposed. Then, the Kullback information criterion for a system of SEM, called SKIC, is proposed for selecting the most appropriate system of the models. SKIC is compared the performance of selection the order of the model, relative to SAIC proposed by Keerativibool (2009). The results of simulation study show that the proposed transformation matrix **P** can transform the MA(1) errors for both forms of (14a) and (14b) to be independent. For all situations of the sample sizes; small (T = 15), medium (T = 30), and large (T = 100), including two series of errors generated in the SEM, SKIC convincingly outperformed SAIC, because SAIC has a tendency to overfit the order of the model than SKIC. The results of the observed L_2 efficiency also confirm that SKIC has a large observed L_2 efficiency and small standard deviation of the observed L_2 efficiency than SAIC, then SKIC is likely better than SAIC. The average and standard deviation of SAIC and SKIC for 1,000 transformed samples show that SAIC presents a large negative bias than SKIC, which maybe the main reason of selecting the correct order of the model from SAIC is less than SKIC.

Nowadays, there is not much the criterion to select the appropriate SEM. Therefore, it should be studied and established the other criteria. Including, other schema of the error-generation might also be considered, such as the autoregressive and moving average (ARMA) scheme instead of only the moving average (MA) scheme.

Acknowledgements

This project is financial supported by The Thailand Research Fund, Office of The Higher Education Commission, and Thaksin University under grant No. MRG5480044.

Appendix
Proofs

Proof of Theorem 1.

The reduced-form model in (7) at the t^{th} observation and the j^{th} equation can be written as follows:

$$y_{tj} = \mathbf{x}_t'\boldsymbol{\pi}_j + v_{tj}\,,\ t = 1,\ 2,\ \ldots,\ T,\ j = 1,\ 2,\ \ldots,\ M, \tag{A1}$$

where

$$\mathbf{x}_t' = \begin{bmatrix} x_{t1} & x_{t2} & \cdots & x_{tK} \end{bmatrix},\ v_{tj} = \varepsilon_{tj} - \theta_j\varepsilon_{t-1,j}\,,\ t = 2,\ 3,\ \ldots,\ T,\ j = 1,\ 2,\ \ldots,\ M. \tag{A2}$$

Replacing v_{tj} in (A2) into (A1) and rearrange it into the term of ε_{tj},

$$\varepsilon_{tj} = y_{tj} - \mathbf{x}_t'\boldsymbol{\pi}_j + \theta_j\varepsilon_{t-1,j}\,,\ t = 2,\ 3,\ \ldots,\ T,\ j = 1,\ 2,\ \ldots,\ M. \tag{A3}$$

The i^{th} lag of .. in (A3) can be written as

$$\varepsilon_{t-i,j} = y_{t-i,j} - \mathbf{x}_{t-i}'\boldsymbol{\pi}_j + \theta_j\varepsilon_{t-(i+1),j}\,. \tag{A4}$$

Using the knowledge of (A4), the equation in (A1) becomes

$$\begin{aligned}
y_{tj} &= \mathbf{x}_t'\boldsymbol{\pi}_j + \varepsilon_{tj} - \theta_j\left(y_{t-1,j} - \mathbf{x}_{t-1}'\boldsymbol{\pi}_j + \theta_j\varepsilon_{t-2,j}\right) \\
y_{tj} + \theta_j y_{t-1,j} &= \left(\mathbf{x}_t' + \theta_j\mathbf{x}_{t-1}'\right)\boldsymbol{\pi}_j + \varepsilon_{tj} - \theta_j^2\varepsilon_{t-2,j} \\
&= \left(\mathbf{x}_t' + \theta_j\mathbf{x}_{t-1}'\right)\boldsymbol{\pi}_j + \varepsilon_{tj} - \theta_j^2\left(y_{t-2,j} - \mathbf{x}_{t-2}'\boldsymbol{\pi}_j + \theta_j\varepsilon_{t-3,j}\right) \\
y_{tj} + \theta_j y_{t-1,j} + \theta_j^2 y_{t-2,j} &= \left(\mathbf{x}_t' + \theta_j\mathbf{x}_{t-1}' + \theta_j^2\mathbf{x}_{t-2}'\right)\boldsymbol{\pi}_j + \varepsilon_{tj} - \theta_j^3\varepsilon_{t-3,j} \\
&\vdots
\end{aligned}$$

$$\sum_{i=0}^{T}\theta_j^i y_{t-i,j} \quad =\sum_{i=0}^{T}\theta_j^i \mathbf{x}'_{t-i}\boldsymbol{\pi}_j+\varepsilon_{tj}-\theta_j^{T+1}\varepsilon_{t-(T+1),j}. \tag{A5}$$

As T becomes large and θ_j satisfies the invertibility condition, the value of θ_j^{T+1} in (A5) approach zero. Therefore, (A5) can be rewritten as

$$y_{tj}^* = \mathbf{x}_t^{*\prime}\boldsymbol{\pi}_j+\varepsilon_{tj}\,, \tag{A6}$$

where $y_{tj}^*=\sum_{i=0}^{T}\theta_j^i y_{t-i,j}$ and $\mathbf{x}_t^{*\prime}=\sum_{i=0}^{T}\theta_j^i\mathbf{x}'_{t-i}$ for $t=2, 3, \ldots, T$, $j=1, 2, \ldots, M$.

From (A6) we found that $\operatorname{Var}\left(y_{tj}^* \mid \mathbf{x}_t^*\right)=\operatorname{Var}\left(\varepsilon_{tj}\right)=\sigma_{jj}$, then we can argue that the MA(1) problem at $t=2, 3, \ldots, T$ and $j=1, 2, \ldots, M$ has been corrected. However, the transformation in (A6) does not include the first observation in (A1). The heteroskedasticity remains unsolved unless the first observation is eliminated, but if the first observation is included in the analysis, the transformation must be extended by the following steps. Firstly, we take the expectation to v_{tj} in (A2),

$$E\left(v_{tj}\right)=E\left(\varepsilon_{tj}\right)-\theta_j E\left(\varepsilon_{t-1,j}\right)=E\left(\varepsilon_{tj}\right)-\theta_j E\left(\varepsilon_{tj}\right)=\left(1-\theta_j\right)E\left(\varepsilon_{tj}\right).$$

Using the assumption in (5), we have the expectation of v_{tj} is equal to zero. Therefore, from (A1) the variance of y_{tj} given $\mathbf{x}_t$ for $t=1, 2, \ldots, T$ and $j=1, 2, \ldots, M$ can be written as

$$\operatorname{Var}\left(v_{tj}\right)=E\left[\left(\varepsilon_{tj}-\theta_j\varepsilon_{t-1,j}\right)^2\right]=E\left(\varepsilon_{tj}^2\right)+\theta_j^2 E\left(\varepsilon_{tj}^2\right)=\left(1+\theta_j^2\right)E\left(\varepsilon_{tj}^2\right)=\left(1+\theta_j^2\right)\sigma_{jj}.$$

Hence, the first observation should weighted by $\sqrt{\dfrac{1}{1+\theta_j^2}}$, yields the model

$$y_{1j}^* = \mathbf{x}_1^{*\prime}\boldsymbol{\pi}_j+\varepsilon_{1j}, \tag{A7}$$

where $y_{1j}^*=\sqrt{\dfrac{1}{1+\theta_j^2}}\,y_{1j}$ and $\mathbf{x}_1^{*\prime}=\sqrt{\dfrac{1}{1+\theta_j^2}}\,\mathbf{x}'_1$ for $j=1, 2, \ldots, M$.

It can be shown that the MA(1) problem at t = 1 has been corrected,

$$\operatorname{Var}\left(y_{1j}^* \left|\mathbf{x}_1^*\right.\right)=\frac{1}{1+\theta_j^2}\operatorname{Var}\left(y_{1j}\left|\mathbf{x}_1\right.\right)=\frac{1}{1+\theta_j^2}\cdot\left(1+\theta_j^2\right)\sigma_{jj}=\sigma_{jj}.$$

Combining the results in (A6) and (A7), we get the $T\times T$ transformation matrix $\mathbf{P}_j$ which was exhibited in (9).

Proof of Theorem 2.

The Kullback-Leibler's symmetric divergence is a measure that used to separate the discrepancy between the candidate model in (10) and the true model in (11), defined by

$$2J\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}\right)=d\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}\right)-d\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0\right)+d\left(\boldsymbol{\theta},\boldsymbol{\theta}_0\right)-d\left(\boldsymbol{\theta},\boldsymbol{\theta}\right), \tag{B1}$$

where $d\left(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j\right)=E_{\boldsymbol{\theta}_i}\left\{-2\log L\left(\boldsymbol{\theta}_j\left|\mathbf{y}^*\right.\right)\right\}$.

Dropping $d\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0\right)$ in (B1) since it does not depend on $\boldsymbol{\theta}$. The ranking of the candidate models according to $2J\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}\right)$ in (B1) is then identical to ranking them according to

$$K\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}\right)=d\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}\right)+d\left(\boldsymbol{\theta},\boldsymbol{\theta}_0\right)-d\left(\boldsymbol{\theta},\boldsymbol{\theta}\right). \tag{B2}$$

Given a set of GLS estimators $\hat{\boldsymbol{\theta}}=\left(\hat{\boldsymbol{\pi}},\,\hat{\boldsymbol{\Sigma}},\,\hat{\mathbf{P}}\right)$ where $\hat{\mathbf{P}}$ is the estimate of the transformation matrix $\mathbf{P}$ in (8),

$$\hat{\boldsymbol{\pi}}=\left[\tilde{\mathbf{X}}^{*\prime}\left(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_T\right)\tilde{\mathbf{X}}^{*}\right]^{-1}\tilde{\mathbf{X}}^{*\prime}\left(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_T\right)\mathbf{y}^{*},$$

and

$$\hat{\boldsymbol{\Sigma}}\otimes\mathbf{I}_T=\frac{1}{T}\left(\mathbf{y}^{*}-\tilde{\mathbf{X}}^{*}\boldsymbol{\pi}\right)\left(\mathbf{y}^{*}-\tilde{\mathbf{X}}^{*}\boldsymbol{\pi}\right)^{\prime},$$

we have therefore the estimate of the symmetric measure in (B2) as

$$K\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)=d\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)+d\left(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0\right)-d\left(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}\right),\qquad\text{(B3)}$$

where $d\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)=E_{\boldsymbol{\theta}_0}\left\{-2\log L\left(\boldsymbol{\theta}\,|\,\mathbf{y}^{*}\right)\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$, $d\left(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0\right)=E_{\boldsymbol{\theta}}\left\{-2\log L\left(\boldsymbol{\theta}_0\,|\,\mathbf{y}^{*}\right)\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$, and $d\left(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}\right)=E_{\boldsymbol{\theta}}\left\{-2\log L\left(\boldsymbol{\theta}\,|\,\mathbf{y}^{*}\right)\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$.

From the minus twice log likelihood of the candidate model in (10),

$$-2\log L\left(\boldsymbol{\theta}\,|\,\mathbf{y}^{*}\right)=TM\log(2\pi)+T\log|\boldsymbol{\Sigma}|+\left(\mathbf{y}^{*}-\tilde{\mathbf{X}}^{*}\boldsymbol{\pi}\right)^{\prime}\left(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_T\right)\left(\mathbf{y}^{*}-\tilde{\mathbf{X}}^{*}\boldsymbol{\pi}\right),$$

we have each term of the estimated symmetric measure in (B3) as follows:

$$\begin{aligned}
d\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)&=TM\log(2\pi)+T\log\left|\hat{\boldsymbol{\Sigma}}\right|+\left(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}}\right)^{\prime}\tilde{\mathbf{X}}^{*\prime}\left(\hat{\boldsymbol{\Sigma}}^{-1}\otimes\mathbf{I}_T\right)\tilde{\mathbf{X}}^{*}\left(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}}\right)+T\,tr\left(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}_0\right),\\
d\left(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0\right)&=TM\log(2\pi)+T\log|\boldsymbol{\Sigma}_0|+\left(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0\right)^{\prime}\tilde{\mathbf{X}}^{*\prime}\left(\boldsymbol{\Sigma}_0^{-1}\otimes\mathbf{I}_T\right)\tilde{\mathbf{X}}^{*}\left(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0\right)+T\,tr\left(\boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{\Sigma}}\right),\\
d\left(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}\right)&=TM\log(2\pi)+T\log\left|\hat{\boldsymbol{\Sigma}}\right|+TM.
\end{aligned}$$

Therefore, the expected of the estimated symmetric measure in (B3) becomes

$$\begin{aligned}
\Omega(\boldsymbol{\theta}_0,K)&=E_{\boldsymbol{\theta}_0}\left\{K\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)\right\}=E_{\boldsymbol{\theta}_0}\left\{d\left(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}\right)+d\left(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0\right)-d\left(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}\right)\right\}\\
&=TM\left[\log(2\pi)+1\right]+E_{\boldsymbol{\theta}_0}\left\{T\log\left|\hat{\boldsymbol{\Sigma}}\right|\right\}+E_{\boldsymbol{\theta}_0}\left\{\left(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}}\right)^{\prime}\tilde{\mathbf{X}}^{*\prime}\left(\hat{\boldsymbol{\Sigma}}^{-1}\otimes\mathbf{I}_T\right)\tilde{\mathbf{X}}^{*}\left(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}}\right)\right\}\\
&\quad+E_{\boldsymbol{\theta}_0}\left\{T\,tr\left(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}_0\right)\right\}+E_{\boldsymbol{\theta}_0}\left\{\left(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0\right)^{\prime}\tilde{\mathbf{X}}^{*\prime}\left(\boldsymbol{\Sigma}_0^{-1}\otimes\mathbf{I}_T\right)\tilde{\mathbf{X}}^{*}\left(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0\right)\right\}\\
&\quad+E_{\boldsymbol{\theta}_0}\left\{T\,tr\left(\boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{\Sigma}}\right)\right\}-E_{\boldsymbol{\theta}_0}\left\{T\log\left(\left|\hat{\boldsymbol{\Sigma}}\right|/|\boldsymbol{\Sigma}_0|\right)\right\}-2TM.\qquad\text{(B4)}
\end{aligned}$$

From the facts that, $\hat{\boldsymbol{\pi}}$ and $T\hat{\boldsymbol{\Sigma}}$ are asymptotically independent where $\hat{\boldsymbol{\pi}}$ is asymptotically distributed as a Gaussian distribution with mean vector $\boldsymbol{\pi}$ and variance-covariance matrix $\left[\tilde{\mathbf{X}}^{*\prime}\left(\boldsymbol{\Sigma}_0^{-1}\otimes\mathbf{I}_T\right)\tilde{\mathbf{X}}^{*}\right]^{-1}$, and $T\hat{\boldsymbol{\Sigma}}$ is asymptotically distributed as the Wishart distribution with $T-K$ degrees of freedom, $\mathrm{W}_{KM}\left(\boldsymbol{\Sigma}_0,\,T-K\right)$, then (Anderson, 2003)

$$E_{\boldsymbol{\theta}_0}\left\{T\hat{\boldsymbol{\Sigma}}\right\}=(T-K)\boldsymbol{\Sigma}_0\text{ and }E_{\boldsymbol{\theta}_0}\left\{\hat{\boldsymbol{\Sigma}}^{-1}\right\}=\frac{T}{T-K-M-1}\boldsymbol{\Sigma}_0^{-1}.$$

Using the above results, we have

$$\begin{aligned}
&E_{\boldsymbol{\theta}_0}\left\{T\,tr\left(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}_0\right)\right\}=T\,tr\left\{E_{\boldsymbol{\theta}_0}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)\boldsymbol{\Sigma}_0\right\}=T\,tr\left\{\frac{T}{T-K-M-1}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_0\right\}=\frac{T^2M}{T-K-M-1},\\
&E_{\boldsymbol{\theta}_0}\left\{T\,tr\left(\boldsymbol{\Sigma}_0^{-1}\hat{\boldsymbol{\Sigma}}\right)\right\}=tr\left\{\boldsymbol{\Sigma}_0^{-1}E_{\boldsymbol{\theta}_0}\left(T\hat{\boldsymbol{\Sigma}}\right)\right\}=tr\left\{\boldsymbol{\Sigma}_0^{-1}(T-K)\boldsymbol{\Sigma}_0\right\}=(T-K)M,\\
&E_{\boldsymbol{\theta}_0}\left\{\left(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}}\right)^{\prime}\tilde{\mathbf{X}}^{*\prime}\left(\hat{\boldsymbol{\Sigma}}^{-1}\otimes\mathbf{I}_T\right)\tilde{\mathbf{X}}^{*}\left(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}}\right)\right\}=E_{\boldsymbol{\theta}_0}\left\{tr\left[\left(\hat{\boldsymbol{\Sigma}}^{-1}\otimes\mathbf{I}_T\right)\tilde{\mathbf{X}}^{*}\left(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}}\right)\left(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}}\right)^{\prime}\tilde{\mathbf{X}}^{*\prime}\right]\right\}
\end{aligned}$$

$$= tr\left\{\left[E_{\boldsymbol{\theta}_0}\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)\otimes\mathbf{I}_T\right]E_{\boldsymbol{\theta}_0}\left[\tilde{\mathbf{X}}^{*}\left(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}}\right)\left(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}}\right)'\tilde{\mathbf{X}}^{*\prime}\right]\right\}$$

$$= \frac{T}{T-K-M-1}tr\left\{\left(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}}\right)'\tilde{\mathbf{X}}^{*\prime}\left(\boldsymbol{\Sigma}_0^{-1}\otimes\mathbf{I}_T\right)\tilde{\mathbf{X}}^{*}\left(\boldsymbol{\pi}_0-\hat{\boldsymbol{\pi}}\right)\right\} = \frac{TKM}{T-K-M-1},$$

$$E_{\boldsymbol{\theta}_0}\left\{\left(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0\right)'\tilde{\mathbf{X}}^{*\prime}\left(\boldsymbol{\Sigma}_0^{-1}\otimes\mathbf{I}_T\right)\tilde{\mathbf{X}}^{*}\left(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0\right)\right\} = KM,$$

then $\Omega\left(\boldsymbol{\theta}_0, K\right)$ in (B4) can be written as

$$\begin{aligned}\Omega\left(\boldsymbol{\theta}_0,K\right) &= TM\left[\log\left(2\pi\right)+1\right]+E_{\boldsymbol{\theta}_0}\left\{T\log\left|\hat{\boldsymbol{\Sigma}}\right|\right\}+\frac{TKM}{T-K-M-1}+\frac{T^2M}{T-K-M-1}\\ &\quad +KM+\left(T-K\right)M-E_{\boldsymbol{\theta}_0}\left\{T\log\left(\left|\hat{\boldsymbol{\Sigma}}\right|/\left|\boldsymbol{\Sigma}_0\right|\right)\right\}-2TM\\ &= TM\left[\log\left(2\pi\right)+1\right]+E_{\boldsymbol{\theta}_0}\left\{T\log\left|\hat{\boldsymbol{\Sigma}}\right|\right\}+\frac{TM\left(2K+M+1\right)}{T-K-M-1}\\ &\quad -T\,E_{\boldsymbol{\theta}_0}\left\{\log\left(\left|T\hat{\boldsymbol{\Sigma}}\right|/\left|\boldsymbol{\Sigma}_0\right|\right)\right\}+TM\log T. \end{aligned} \tag{B5}$$

Because $\left|T\hat{\boldsymbol{\Sigma}}\right|/\left|\boldsymbol{\Sigma}_0\right|$ in (B5) is the distribution of a product of independent χ^2 random variables, $\prod_{i=1}^{M}\chi^2_{T-K-M+i}$, then we have

$$\log\left(\left|T\hat{\boldsymbol{\Sigma}}\right|/\left|\boldsymbol{\Sigma}_0\right|\right)\sim\sum_{i=1}^{M}\log\chi^2_{T-K-M+i}.$$

Using the second-order of Taylor's series expansions to expand the function of $\log\left(\chi^2_p\right)$ about the mean p, we have

$$\log\left(\chi^2_p\right)\doteq\log\left(p\right)+\frac{1}{p}\left(\chi^2_p-p\right)-\frac{1}{2p^2}\left(\chi^2_p-p\right)^2 \text{ and } E\left[\log\left(\chi^2_p\right)\right]\doteq\log\left(p\right)-\frac{1}{p}.$$

Then, the last two terms of the right-hand side in (B5) is

$$-TE_{\boldsymbol{\theta}_0}\left\{\log\left(\left|T\hat{\boldsymbol{\Sigma}}\right|/\left|\boldsymbol{\Sigma}_0\right|\right)\right\}+TM\log T\doteq -T\sum_{i=1}^{M}\left[\log\left(T-K-M+i\right)-\frac{1}{T-K-M+i}\right]+TM\log T. \tag{B6}$$

McQuarrie and Tsai (1998) gave the simplification formulae for any *T*, *K*, *M* and assume $T-K-M$ is much larger than *M* as follows:

$$\sum_{i=1}^{M}\log\left(T-K-M+i\right)=M\log\left(T-K-\frac{M-1}{2}\right)=M\log\left(\frac{2T-2K-M+1}{2}\right), \tag{B7}$$

and

$$\sum_{i=1}^{M}\frac{1}{T-K-M+i}\doteq\frac{M}{T-K-\dfrac{M-1}{2}}=\frac{2M}{2T-2K-M+1}. \tag{B8}$$

Replacing the results in (B7) and (B8) into (B6), we have

$$-TE_{\boldsymbol{\theta}_0}\left\{\log\left(\left|T\hat{\boldsymbol{\Sigma}}\right|/\left|\boldsymbol{\Sigma}_0\right|\right)\right\}+TM\log T\doteq TM\log\left(\frac{2T}{2T-2K-M+1}\right)+\frac{2TM}{2T-2K-M+1}. \tag{B9}$$

Replacing the results in (B9) into (B5), we have

$$\Omega\left(\boldsymbol{\theta}_0,K\right)\doteq TM\left[\log\left(2\pi\right)+1\right]+E_{\boldsymbol{\theta}_0}\left\{\text{SKIC}\right\},$$

where SKIC was exhibited in (12).

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Proc. *2nd Int. Symp. on Information Theory*. Akademia Kiado, Budapest, pp. 267–281.

Akaike, H., 1974. A new look at the statistical model identification. *IEEE T. Automat. Contr.* **19**, 716–723.

Anderson, T.W. 2003. *An introduction to multivariate statistical analysis.* 3rd ed. Hoboken, New Jersey: Wiley.

Bedrick, E.J., Tsai, C.L., 1994. Model selection for multivariate regression in small samples. *Biometrics*. **50**, 226–231.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C. 1994. *Time series analysis: forecasting and control.* 3rd ed. Englewood Cliffs, New Jersey: Prentice Hall.

Cavanaugh, J.E., 1999. A large-sample model selection criterion based on Kullback's symmetric divergence. *Stat. Probabil. Lett.* **42**, 333–343.

Cavanaugh, J.E., 2004. Criteria for linear model selection based on Kullback's symmetric divergence. *Aust. NZ. J. Stat.* **46**, 257–274.

Greene, W. 2008. *Econometric analysis.* 6th ed. Upper Saddle River, New Jersey: Prentice-Hall.

Gujarati, D.N. 2006. *Essentials of econometrics.* 3rd ed. Singapore: McGraw-Hill.

Hafidi, B., 2006. A small-sample criterion based on Kullback's symmetric divergence for vector autoregressive modeling. *Stat. Probabil. Lett.* **76**, 1647–1654.

Hafidi, B., Mkhadri, A., 2006. A corrected Akaike criterion based on Kullback's symmetric divergence: applications in time series, multiple and multivariate regression. *Comput. Stat. Data. An.* **50**, 1524–1550.

Hurvich, C.M., Tsai, C.L., 1989. Regression and time series model selection in small samples. *Biometrika.* **76**, 297–307.

Keerativibool, W. 2009. *Selection of a system of simultaneous equations model.* Dissertation, School of Applied Statistics, National Institute of Development Administration, THIALAND.

McQuarrie, A.D., Shumway, R., Tsai, C.L., 1997. The model selection criterion AICu. *Stat. Probabil. Lett.* **34**, 285–292.

McQuarrie, A.D., Tsai, C.L., 1998. *Regression and time series model selection*. Singapore: World Scientific.

McQuarrie, A.D., 1999. A small-sample correction for the Schwarz SIC model selection criterion. *Stat. Probabil. Lett.* **44**, 79–86.

Seghouane, A.K., Bekara M., 2004. A small sample model selection criterion based on Kullback's symmetric divergence. *IEEE T. Signal Proces*. **52**, 3314–3323.

Seghouane, A.K., 2006. Multivariate regression model selection from small samples using Kullback's symmetric divergence. *Signal Process*. **86**, 2074–2084.