GLS Discrepancy Based Information Criteria for Selecting Covariance Structure Models

(Last Modified: October 3, 2007)

Hirokazu YANAGIHARA[1] , Tetsuto HIMENO[2] AND Ke-Hai YUAN[3]

[1]*Department of Mathematics, Graduate School of Science, Hiroshima University*
1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8626, Japan

[2]*Graduate School of Mathematics, Kyushu University*
6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan

[3]*Department of Psychology, University of Notre Dame*
Notre Dame, Indiana 46556, USA

Abstract

This paper studies information criteria for selecting covariance structure models using the generalized least squares (GLS) procedure. A risk based on the predictive GLS discrepancy function is introduced and used to determine the quality of a model. By correcting the biases in the sample GLS discrepancy function, four information criteria are proposed. Monte Carlo results illustrate the merits of each criterion in model selection and in minimizing the risk.

AMS 2000 *subject classifications*: Primary 62H99; Secondary 62F07.
Key words: Bias correction; covariance structure; information criterion; generalized least squares discrepancy; Mallows' C_p criterion; model selection; nonnormality.

1. Introduction

Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be a random sample from a p dimensional population represented by $\boldsymbol{y}$. We are interested in modeling the covariance matrix of $\boldsymbol{y}$. Denote the interesting model by $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$. The null hypothesis or model is represented by

$$M: \ E[\boldsymbol{y}] = \boldsymbol{\mu}, \quad \mathrm{Cov}[\boldsymbol{y}] = \boldsymbol{\Sigma}(\boldsymbol{\theta}). \tag{1.1}$$

[1] Corresponding author, E-mail: *yanagi@math.sci.hiroshima-u.ac.jp*

For simplicity, we write $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ as $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. Of course, the interesting model is likely misspecified. The alternative hypothesis or true model is

$$M_* : \ E[\boldsymbol{y}] = \boldsymbol{\mu}_*, \quad \mathrm{Cov}[\boldsymbol{y}] = \boldsymbol{\Sigma}_*. \tag{1.2}$$

In practice, we may simultaneously consider multiple models. Among all the candidate models, the one with the fewest number of parameters that fits the data well is regarded as a good model. The idea can be executed through the so-called risk, commonly defined as the predictive discrepancy between the candidate model M and the true model M_*. The one having the smallest risk is regarded as the best model among all the candidate models. In covariance structure analysis, the normal distribution based maximum likelihood (ML) procedure is commonly used in practice. The risk associated with the ML procedure in model selection has been extensively studied (e.g., Cudeck & Browne, 1983; Akaike, 1987; Browne & Cudeck, 1989; De Gooijer, 1995; Yanagihara, 2005; Yanagihara et al., 2007). In addition to the ML procedure, the normal distribution based generalized least squares (GLS) procedure is systematically introduced in essentially every textbook on structural equation modeling (SEM) (e.g., Bollen, 1989, pp. 113–114; Kano & Miura, 2002, p. 136) and also available in all standard SEM software (AMOS, EQS, LISREL, MPLUS, SAS CALIS). Although the GLS procedure has been shown to be equivalent to the ML procedure under certain conditions, they are not equivalent with typically misspecified covariance structure models in practice (e.g., Yuan & Chan, 2005). Thus, it is necessary to study the risk associated with the GLS procedure in model selections. This paper conducts such a study with covariance structure analysis. We will show that, for the GLS procedure, the target risk can be defined through a mean square error (MSE) of the estimated covariance matrix. This directly implies that the best model enjoys the smallest MSE. Like the development in the ML case, we will show that the GLS discrepancy estimated at the current sample is a biased estimator of the target risk. By adding the estimated bias to the sample GLS discrepancy, we propose four information criteria according to model specification and the distribution of the data. The proposed information criteria are more close to Mallows' C_p criterion (Mallows, 1973; 1995), we call them the C_p-type criteria for covariance structure models.

In Section 2, we propose four criteria based on the predictive GLS discrepancy. In Section 3, we verify performance of our criteria. Technical details are provided in an appendix.

2. Information Criteria Based on Predictive GLS Discrepancy

Let $\bar{\boldsymbol{y}}$ be the sample mean of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, and $\boldsymbol{S}$ be the unbiased sample covariance matrix. We use $d(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}, \boldsymbol{S})$ to denote a discrepancy between $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ and $\boldsymbol{S}$. The GLS discrepancy used in this paper is defined as

$$d(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}, \boldsymbol{S}) = \frac{1}{2}\mathrm{tr}\left\{(\boldsymbol{\Sigma}_{\boldsymbol{\theta}} - \boldsymbol{S})\boldsymbol{\Sigma}_*^{-1}\right\}^2. \tag{2.1}$$

When replacing $\boldsymbol{\Sigma}_*$ in equation (2.1) by $\boldsymbol{S}$, the estimated GLS discrepancy function becomes

$$\hat{d}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}, \boldsymbol{S}) = \frac{1}{2}\mathrm{tr}\left\{(\boldsymbol{\Sigma}_{\boldsymbol{\theta}} - \boldsymbol{S})\boldsymbol{S}^{-1}\right\}^2 = \frac{1}{2}\left\{p - 2\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{S}^{-1}) + \mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{S}^{-1})^2\right\},$$

which is the commonly used GLS discrepancy function in covariance structure analysis. The GLS estimator is defined by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \hat{d}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}, \boldsymbol{S}). \tag{2.2}$$

Let $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)'$ be an $n \times p$ future observation matrix, which is independent and has the same distribution with $\boldsymbol{Y}$, and $\boldsymbol{W}$ be the unbiased sample covariance matrix corresponding to $\boldsymbol{U}$. We define the risk based on the predictive GLS discrepancy as

$$R_p = E_{\boldsymbol{U}}^* E_{\boldsymbol{Y}}^*[d(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}, \boldsymbol{W})], \tag{2.3}$$

where $E_{\boldsymbol{U}}^*$ and $E_{\boldsymbol{Y}}^*$ denote expectations under the true model (1.2) with respect to $\boldsymbol{U}$ and $\boldsymbol{Y}$, respectively. We regard the model having the smallest R_p as the best model, which is typically different from the true model. Notice that R_p in (2.3) can be rewritten as $E_{\boldsymbol{Y}}^*[d(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}, \boldsymbol{\Sigma}_*)] + E_{\boldsymbol{Y}}^*[d(\boldsymbol{\Sigma}_*, \boldsymbol{S})]$. Since $E_{\boldsymbol{Y}}^*[d(\boldsymbol{\Sigma}_*, \boldsymbol{S})]$ is the same for all the candidate models, the best model only minimizes $E_{\boldsymbol{Y}}^*[d(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}, \boldsymbol{\Sigma}_*)] = E_{\boldsymbol{Y}}^*[\mathrm{tr}\{(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\Sigma}_*)\boldsymbol{\Sigma}_*^{-1}\}^2]/2$. It implies that the best model leads to the smallest MSE for $\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}\boldsymbol{\Sigma}_*^{-1/2}$ as an estimator of $\boldsymbol{I}_p$.

In covariance structure analysis and many other contexts of statistical modeling, the aim is to determine the best model. Obtaining an unbiased estimator of R_p will enable us to correctly evaluate the discrepancy between data and model, which will further facilitate the selection of the best model. The simplest estimator of R_p is the sample GLS discrepancy function $\hat{d}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}, \boldsymbol{S})$. However, as when estimating the risk by directly using the sample quantities in other context of statistical modeling, $\hat{d}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}, \boldsymbol{S})$ generally

underestimates the R_p. An information criterion can be defined as $\hat{d}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}, \boldsymbol{S}) + \hat{B}$, where $\hat{B}$ is a n-consistent estimator of the bias

$$B = R_p - E^*_{\boldsymbol{Y}}[\hat{d}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}, \boldsymbol{S})]. \tag{2.4}$$

The following technical development is around obtaining B and its estimates.

In order to simplify the formula of the bias, we consider a certain fourth cumulant of the true distribution of $\boldsymbol{y}$. Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)' = \boldsymbol{\Sigma}_*^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu}_*)$. Then, for a $p^2 \times p^2$ matrix $\boldsymbol{A}$, we denote

$$\kappa(\boldsymbol{A}) = E^*_{\boldsymbol{Y}}\left[\mathrm{vec}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')'\boldsymbol{A}\mathrm{vec}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\right] - \mathrm{tr}\{(\boldsymbol{I}_{p^2} + \boldsymbol{K}_p)\boldsymbol{A}\} - \mathrm{vec}(\boldsymbol{I}_p)'\boldsymbol{A}\mathrm{vec}(\boldsymbol{I}_p), \tag{2.5}$$

where $\boldsymbol{K}_p$ is the commutation matrix (see Magnus & Neudecker, 1999, p. 48). Notice that $\kappa_4^{(1)} = \kappa(\boldsymbol{I}_{p^2})$ is just the multivariate kurtosis considered by Mardia (1970). It follows from $\mathrm{vec}(\boldsymbol{I}_p)'(\boldsymbol{B} \otimes \boldsymbol{C})\mathrm{vec}(\boldsymbol{I}_p) = \mathrm{tr}(\boldsymbol{B}\boldsymbol{C})$ and $\mathrm{tr}\{(\boldsymbol{B} \otimes \boldsymbol{C})\boldsymbol{K}_p\} = \mathrm{tr}(\boldsymbol{B}\boldsymbol{C})$ that, for any $p \times p$ symmetric matrices $\boldsymbol{B}$ and $\boldsymbol{C}$,

$$\kappa(\boldsymbol{B} \otimes \boldsymbol{C}) = E^*_{\boldsymbol{Y}}\left[(\boldsymbol{\varepsilon}'\boldsymbol{B}\boldsymbol{\varepsilon})(\boldsymbol{\varepsilon}'\boldsymbol{C}\boldsymbol{\varepsilon})\right] - \mathrm{tr}\boldsymbol{B}\mathrm{tr}\boldsymbol{C} - 2\mathrm{tr}(\boldsymbol{B}\boldsymbol{C}).$$

The above kurtosis was used by Yanagihara (2007b) for simplifying coefficients in an asymptotic expansion of a test statistic in the context of generalized MANOVA.

To simplify the following presentation, we denote

$$\boldsymbol{\Lambda}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{\Sigma}_*^{-1/2}, \quad \boldsymbol{\Omega}_{\boldsymbol{\theta}} = \boldsymbol{\Lambda}_{\boldsymbol{\theta}} - \boldsymbol{I}_p, \quad \boldsymbol{\Gamma}_{\boldsymbol{\theta}} = (\boldsymbol{\Lambda}_{\boldsymbol{\theta}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{\theta}}) - (\boldsymbol{\Omega}_{\boldsymbol{\theta}} \otimes \boldsymbol{\Omega}_{\boldsymbol{\theta}}). \tag{2.6}$$

Suppose that $\boldsymbol{\theta}_0$ is a $q \times 1$ vector such that $\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}_0$ as $n \to \infty$. Under proper conditions, as specified in Swain (1975) and White (1982), $\boldsymbol{\theta}_0$ satisfies

$$\boldsymbol{\Delta}'_{\boldsymbol{\theta}_0}\mathrm{vec}(\boldsymbol{\Omega}_{\boldsymbol{\theta}_0}) = \boldsymbol{0}_q,$$

where $\boldsymbol{0}_q$ is a vector of q zeros, and

$$\boldsymbol{\Delta}_{\boldsymbol{\theta}_0} = \left.\frac{\partial}{\partial \boldsymbol{\theta}'}\mathrm{vec}(\boldsymbol{\Lambda}_{\boldsymbol{\theta}})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = (\boldsymbol{\Sigma}_*^{-1/2} \otimes \boldsymbol{\Sigma}_*^{-1/2})\left.\left(\frac{\partial}{\partial \theta_1}\mathrm{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}), \ldots, \frac{\partial}{\partial \theta_q}\mathrm{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})\right)\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}. \tag{2.7}$$

We also use

$$\boldsymbol{H}_{\boldsymbol{\theta}_0} = \boldsymbol{\Delta}'_{\boldsymbol{\theta}_0}\boldsymbol{\Delta}_{\boldsymbol{\theta}_0} + \left.\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\mathrm{tr}(\boldsymbol{\Lambda}_{\boldsymbol{\theta}}\boldsymbol{\Omega}_{\boldsymbol{\theta}_0})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

and

$$\boldsymbol{\Pi}_{\boldsymbol{\theta}_0} = \boldsymbol{\Gamma}_{\boldsymbol{\theta}_0}\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}\boldsymbol{H}^{-1}_{\boldsymbol{\theta}_0}\boldsymbol{\Delta}'_{\boldsymbol{\theta}_0}\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0}. \tag{2.8}$$

With the above notation, we have the following result.

THEOREM 1. *Under standard regularity conditions, the bias of* $\hat{d}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}, \boldsymbol{S})$ *is expanded as*

$$B = \frac{1}{n}\Big\{\kappa(\boldsymbol{\Pi}_{\boldsymbol{\theta}_0}) - \frac{1}{2}\kappa(\boldsymbol{\Omega}_{\boldsymbol{\theta}_0} \otimes \boldsymbol{\Omega}_{\boldsymbol{\theta}_0}) - \kappa(\boldsymbol{\Omega}^2_{\boldsymbol{\theta}_0} \otimes \boldsymbol{I}_p) - 2\kappa(\boldsymbol{\Omega}_{\boldsymbol{\theta}_0} \otimes \boldsymbol{I}_p) \\ +2\mathrm{tr}\boldsymbol{\Pi}_{\boldsymbol{\theta}_0} - \frac{1}{2}(\mathrm{tr}\boldsymbol{\Omega}_{\boldsymbol{\theta}_0})^2 - \frac{(2p+3)}{2}\mathrm{tr}\boldsymbol{\Omega}^2_{\boldsymbol{\theta}_0} - 2(p+1)\mathrm{tr}\boldsymbol{\Omega}_{\boldsymbol{\theta}_0}\Big\} + o(n^{-1}). \quad (2.9)$$

The proof of the theorem is given in the Appendix. From now on, the candidate model in (1.1) will be called overspecified if $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0} = \boldsymbol{\Sigma}_*$; otherwise, it is an underspecified model. If the model is overspecified, $\boldsymbol{\Omega}_{\boldsymbol{\theta}_0} = \boldsymbol{O}_{p,p}$, a $p \times p$ matrix of 0's. Then $\boldsymbol{\Pi}_{\boldsymbol{\theta}_0} = \boldsymbol{P}_{\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}}$, where $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$; and the bias B can be simplified for overspecified models.

COROLLARY 1. *Under standard regularity conditions, the bias of* $\hat{d}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}, \boldsymbol{S})$ *for an overspecified model is*

$$B = \frac{1}{n}\Big\{\kappa(\boldsymbol{P}_{\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}}) + 2q\Big\} + o(n^{-1}).$$

Thus, when M is an overspecified model, B generally increases with the tails of $\boldsymbol{y}$ and the number of parameters; the bias B is also identical to that of the ML procedure (see e.g., Yanagihara, 2005). So, for overspecified models, ML and GLS discrepancy functions are equivalent up to the order of $O(n^{-1})$. When $\boldsymbol{y}$ is normally distributed, all kurtoses vanish, which further leads to the following result.

COROLLARY 2. *If* $\boldsymbol{y}$ *follows a multivariate normal distribution, then*

$$B = \frac{1}{n}\left\{2\mathrm{tr}\boldsymbol{\Pi}_{\boldsymbol{\theta}_0} - \frac{1}{2}(\mathrm{tr}\boldsymbol{\Omega}_{\boldsymbol{\theta}_0})^2 - \frac{(2p+3)}{2}\mathrm{tr}\boldsymbol{\Omega}^2_{\boldsymbol{\theta}_0} - 2(p+1)\mathrm{tr}\boldsymbol{\Omega}_{\boldsymbol{\theta}_0}\right\} + o(n^{-1}).$$

It follows from the two corollaries that $B = 2q/n + o(n^{-1})$ when the candidate model is overspecified and $\boldsymbol{y}$ is normally distributed. Combining Theorem 1, and Corollaries 1 and 2, we proposed the following information criteria:

(*i*) For normally distributed data and overspecified models:

$$C_p = \hat{d}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}, \boldsymbol{S}) + \frac{2q}{n}.$$

(*ii*) For overspecified models with nonnormally distributed $\boldsymbol{y}$ (corrected C_p):

$$CC_p = C_p + \frac{1}{n}\mathrm{tr}(\hat{\boldsymbol{\Psi}}_\kappa \hat{\boldsymbol{\Pi}}_{\hat{\boldsymbol{\theta}}}),$$

where $\hat{\mathbf{\Pi}}_{\hat{\theta}}$ is obtained when replacing $\mathbf{\Sigma}_*$ by $\boldsymbol{S}$ in $\mathbf{\Pi}_{\hat{\theta}}$, and

$$\hat{\mathbf{\Psi}}_\kappa = \frac{(n+1)}{n(n-1)} \sum_{i=1}^{n} \mathrm{vec}(\hat{\boldsymbol{\varepsilon}}_i \hat{\boldsymbol{\varepsilon}}_i') \mathrm{vec}(\hat{\boldsymbol{\varepsilon}}_i \hat{\boldsymbol{\varepsilon}}_i')' - \boldsymbol{I}_{p^2} - \mathrm{vec}(\boldsymbol{I}_p)\mathrm{vec}(\boldsymbol{I}_p)' - \boldsymbol{K}_p$$

with $\hat{\boldsymbol{\varepsilon}}_i = \boldsymbol{S}^{-1/2}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})$.

(*iii*) Normally distributed data with underspecified models (modified C_p under normal assumption):

$$MC_{p,\mathrm{N}} = \hat{d}(\mathbf{\Sigma}_{\hat{\theta}}, \boldsymbol{S}) + \frac{1}{n}\left\{ 2\mathrm{tr}\hat{\mathbf{\Pi}}_{\hat{\theta}} - \frac{1}{2}(\mathrm{tr}\hat{\mathbf{\Omega}}_{\hat{\theta}})^2 - \frac{(2p+3)}{2}\mathrm{tr}\hat{\mathbf{\Omega}}_{\hat{\theta}}^2 - 2(p+1)\mathrm{tr}\hat{\mathbf{\Omega}}_{\hat{\theta}} \right\},$$

where $\hat{\mathbf{\Omega}}_{\hat{\theta}}$ is defined by replacing $\mathbf{\Sigma}_*$ with $\boldsymbol{S}$ in $\mathbf{\Omega}_{\hat{\theta}}$.

(*iv*) No assumptions on either the model or data (modified C_p):

$$MC_p = MC_{p,\mathrm{N}} + \frac{1}{2n}\mathrm{tr}\left(\hat{\mathbf{\Psi}}_\kappa \{ 2\hat{\mathbf{\Pi}}_{\hat{\theta}} - (\hat{\mathbf{\Omega}}_{\hat{\theta}} \otimes \hat{\mathbf{\Omega}}_{\hat{\theta}}) - 2(\hat{\mathbf{\Omega}}_{\hat{\theta}}^2 \otimes \boldsymbol{I}_p) - 4(\hat{\mathbf{\Omega}}_{\hat{\theta}} \otimes \boldsymbol{I}_p) \} \right).$$

From above definitions, we can see that the bias correction terms in C_p and CC_p are always positive. The terminologies "corrected" and "modified" were used in Fujikoshi and Satoh (1997), implying that the bias is reduced after the correction or modification. Notice that the coefficient in the definition of $\hat{\mathbf{\Psi}}_\kappa$ is not $1/n$ but $(n+1)/\{n(n-1)\}$, which makes $\mathrm{tr}\hat{\mathbf{\Psi}}_\kappa$ an unbiased estimator of $\kappa_4^{(1)}$. Such an estimator was proposed Mardia (1970) in estimating the population kurtosis.

3. Numerical Study

In this section, we verify the performance of the proposed information criteria C_p, CC_p, MC_p and $MC_{p,\mathrm{N}}$ using Monte Carlo. We will mainly consider linear structure models due to its wide applications (Anderson, 1969; Siotani, Hayakawa & Fujikoshi, 1985 pp. 375–378).

Let $\boldsymbol{y} = \mathbf{\Sigma}_*^{1/2}\boldsymbol{\varepsilon}$ with $p = 6$. Two population covariance matrices are given by $\mathbf{\Sigma}_{*1} = \boldsymbol{I}_6 + \mathbf{1}_6\mathbf{1}_6'$ and $\mathbf{\Sigma}_{*2} = \mathrm{diag}(1,1,2,2,3,3)$. For each of the covariance matrices, six distributions are created when the elements ε_j of $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_6)'$ are independently and identically distributed standardized variables from each of the following six distributions:

1. *Normal distribution*: $\varepsilon_j \sim N(0,1)$, $(\kappa_{3,3}^{(1)} = \kappa_{3,3}^{(2)} = 0$ and $\kappa_4^{(1)} = 0)$.

2. *Laplace Distribution*: ε_j is generated from the Laplace distribution with mean 0 and standard deviation 1 ($\kappa_{3,3}^{(1)} = \kappa_{3,3}^{(2)} = 0$ and $\kappa_4^{(1)} = 18$).

3. *Uniform distribution*: ε_j is generated from the uniform distribution on $(-1, 1)$, divided by the standard deviation $1/\sqrt{3}$ ($\kappa_{3,3}^{(1)} = \kappa_{3,3}^{(2)} = 0$ and $\kappa_4^{(1)} = -7.2$).

4. *Skew-Laplace distribution*: ε_j is generated from the skew-Laplace distribution with location parameter 0, dispersion parameter 1 and skew parameter 1, standardized by mean 3/4 and standard deviation $\sqrt{23}/4$ ($\kappa_{3,3}^{(1)} = \kappa_{3,3}^{(2)} \approx 7.32$ and $\kappa_4^{(1)} \approx 19.56$).

5. *Chi-Square distribution*: ε_j is generated from the chi-square distribution with 2 degrees of freedom, standardized by mean 2 and standard deviation 2 ($\kappa_{3,3}^{(1)} = \kappa_{3,3}^{(2)} = 12$ and $\kappa_4^{(1)} = 36$).

6. *Log-Normal distribution*: ε_j is generated from the lognormal distribution such that $\log \varepsilon_j \sim N(0, 1/2)$, standardized by mean $e^{1/4}$ and standard deviation $\sqrt{e^{1/2}(e^{1/2} - 1)}$ ($\kappa_{3,3}^{(1)} = \kappa_{3,3}^{(2)} \approx 17.64$ and $\kappa_4^{(1)} \approx 111.06$).

Here, $\kappa_{3,3}^{(1)}$ and $\kappa_{3,3}^{(2)}$ are multivariate skewnesses (see e.g., Mardia, 1970) defined respectively by

$$\kappa_{3,3}^{(1)} = E[(\boldsymbol{\varepsilon}_1'\boldsymbol{\varepsilon}_2)^3] \quad \text{and} \quad \kappa_{3,3}^{(2)} = E[(\boldsymbol{\varepsilon}_1'\boldsymbol{\varepsilon}_1)(\boldsymbol{\varepsilon}_1'\boldsymbol{\varepsilon}_2)(\boldsymbol{\varepsilon}_2'\boldsymbol{\varepsilon}_2)],$$

where $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ are independent random vectors having the same distribution as $\boldsymbol{\varepsilon}$. The skew-Laplace distribution was proposed by Balakrishnan and Ambagaspitiya (1994) (for the probability density function, see e.g., Yanagihara & Yuan, 2005). The distributions in 1, 2, and 3 are symmetric, and distributions in 4, 5, and 6 are skewed.

A sample of size 50 is generated from each of the six $\boldsymbol{y}$'s. For each population, we consider the following five candidate models:

$$\begin{aligned}
M_1 &: \boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \theta_1 \boldsymbol{I}_6, \\
M_2 &: \boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \theta_1 \boldsymbol{I}_6 + \theta_2(\mathbf{1}_6\mathbf{1}_6' - \boldsymbol{I}_6), \\
M_3 &: \boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \mathrm{diag}(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6), \\
M_4 &: \boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \mathrm{diag}(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) + \theta_7(\mathbf{1}_6\mathbf{1}_6' - \boldsymbol{I}_6), \\
M_5 &: \boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma} \quad \text{(saturated model)}.
\end{aligned}$$

When $\boldsymbol{\Sigma}_* = \boldsymbol{\Sigma}_{*1}$, models M_2, M_4 and M_5 are all overspecified, but M_2 has the fewest number of parameters. When $\boldsymbol{\Sigma}_* = \boldsymbol{\Sigma}_{*2}$, models M_3, M_4 and M_5 are all correctly specified,

but M_3 has the fewest number of parameters. In addition to models, the four criteria are also affected by the population distributions, which will be evaluated next by Monte Carlo.

The number of replications is chosen as $N_r = 10,000$. The four criteria C_p, CC_p, MC_p and $MC_{p,\mathrm{N}}$ are evaluated at each replication. Because $\boldsymbol{\Sigma}_*$ is known, $R = d(\boldsymbol{\Sigma}_*, \boldsymbol{S}) + d(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}, \boldsymbol{\Sigma}_*)$ can also be obtained at each replication. The average of R across the N_r replications, $\bar{R}$, is regarded as the risk R_p. Let $\overline{\mathrm{IC}}$ be the average of any of the four criteria, the bias of the criterion is evaluated as $\bar{R} - \overline{\mathrm{IC}}$. The smallest IC at each replication for a given model is recorded, so is its frequency among the 10,000 replications. Let $\hat{\boldsymbol{\Sigma}}_{\mathrm{best},l}$ be the estimated covariance matrix for the "best model"(the model with the smallest IC) at the lth replication. Then, the MSE of the estimated covariance matrix is evaluated by $N_r^{-1} \sum_{l=1}^{N_r} d(\hat{\boldsymbol{\Sigma}}_{\mathrm{best},l}, \boldsymbol{\Sigma}_*)$.

Tables 1 and 2 contain the risks, the biases, the frequency of each of the models being selected by the information criteria and MSEs when $\boldsymbol{\Sigma}_* = \boldsymbol{\Sigma}_{*1}$ and $\boldsymbol{\Sigma}_* = \boldsymbol{\Sigma}_{*2}$, respectively. As $\kappa_4^{(1)}$ increases, the biases and MSEs become large and the frequencies of choosing the best model become low. By comparing the results of distributions in 2 and 4, we can see that the size of skewness of the distribution hardly influences the results, which parallels the analytical results of inference for correlations (Yuan & Bentler, 2000). Both tables clearly show that MC_p has the smallest bias among all the criteria. However, biases still remain even in MC_p. This is mainly because the correction terms in the criteria CC_p and MC_p contain the sample estimates of kurtosis, and population kurtosis tends to be grossly underestimated by the sample kurtosis even when the sample size is relatively large (see e.g., Yanagihara, 2007a). Although biases were not reduced sufficiently, comparing with $MC_{p,\mathrm{N}}$, we will notice that MSEs of MC_p are smaller even when $\kappa_4^{(1)}$ is large. Improved estimation of kurtosis may further improve the MSE. Among all the criteria, CC_p enjoys the smallest MSE. The criterion attaining the highest frequency changes with the true covariance matrix. When $\boldsymbol{\Sigma}_* = \boldsymbol{\Sigma}_{*1}$, CC_p works best in choosing the true model; when $\boldsymbol{\Sigma}_* = \boldsymbol{\Sigma}_{*2}$, C_p works the best in choosing the true model. However, when $\boldsymbol{\Sigma}_* = \boldsymbol{\Sigma}_{*1}$ for distribution conditions 5 and 6, C_p performs poorly.

Insert Tables 1 and 2 around here

Based on the empirical results, we recommend CC_p and MC_p for selecting covariance structure models with the GLS procedure.

Appendix

This appendix contains the proof of Theorem 2.1. The bias of $\hat{d}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}, \boldsymbol{S})$, defined in (2.4), can be written as

$$B = E^*_{\boldsymbol{U}} E^*_{\boldsymbol{Y}}[d(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}, \boldsymbol{W}) - \hat{d}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{W}, \boldsymbol{S} - \boldsymbol{W})] = -\frac{1}{2}\beta_1 + \beta_2 + \frac{1}{2}\beta_3, \tag{A.1}$$

where

$$\begin{aligned}
\beta_1 &= E^*_{\boldsymbol{U}} E^*_{\boldsymbol{Y}} \left[\operatorname{tr}\left\{ (\boldsymbol{S} - \boldsymbol{W})\boldsymbol{S}^{-1} \right\}^2 \right], \\
\beta_2 &= E^*_{\boldsymbol{U}} E^*_{\boldsymbol{Y}} \left[\operatorname{tr}\left\{ (\boldsymbol{S} - \boldsymbol{W})\boldsymbol{S}^{-1}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{W})\boldsymbol{S}^{-1} \right\} \right], \\
\beta_3 &= E^*_{\boldsymbol{U}} E^*_{\boldsymbol{Y}} \left[\operatorname{tr}\left\{ (\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{W})(\boldsymbol{\Sigma}_*^{-1} + \boldsymbol{S}^{-1})(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{W})(\boldsymbol{\Sigma}_*^{-1} - \boldsymbol{S}^{-1}) \right\} \right].
\end{aligned}$$

Because $E^*_{\boldsymbol{U}}[\boldsymbol{W}] = \boldsymbol{\Sigma}_*$, for any $p \times p$ constant matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$, we have

$$\begin{aligned}
&E^*_{\boldsymbol{U}} \left[\operatorname{tr}\left\{ (\boldsymbol{A} - \boldsymbol{W})\boldsymbol{B}(\boldsymbol{C} - \boldsymbol{W})\boldsymbol{D} \right\} \right] \\
&= E^*_{\boldsymbol{U}} \left[\operatorname{tr}\left\{ (\boldsymbol{A} - \boldsymbol{\Sigma}_* + \boldsymbol{\Sigma}_* - \boldsymbol{W})\boldsymbol{B}(\boldsymbol{C} - \boldsymbol{\Sigma}_* + \boldsymbol{\Sigma}_* - \boldsymbol{W})\boldsymbol{D} \right\} \right] \\
&= \operatorname{tr}\left\{ (\boldsymbol{A} - \boldsymbol{\Sigma}_*)\boldsymbol{B}(\boldsymbol{C} - \boldsymbol{\Sigma}_*)\boldsymbol{D} \right\} + E^*_{\boldsymbol{U}} \left[\operatorname{tr}\left\{ (\boldsymbol{W} - \boldsymbol{\Sigma}_*)\boldsymbol{B}(\boldsymbol{W} - \boldsymbol{\Sigma}_*)\boldsymbol{D} \right\} \right].
\end{aligned} \tag{A.2}$$

Applying (A.2) to β_1, β_2 and β_3 in (A.1) leads to

$$\beta_1 = \alpha_1 + \alpha_2, \quad \beta_2 = \alpha_2 + \alpha_3, \quad \beta_3 = \alpha_4 + \alpha_5,$$

where

$$\begin{aligned}
\alpha_1 &= E^*_{\boldsymbol{Y}} \left[\operatorname{tr}\left\{ (\boldsymbol{S} - \boldsymbol{\Sigma}_*)\boldsymbol{S}^{-1} \right\}^2 \right], \\
\alpha_2 &= E^*_{\boldsymbol{U}} E^*_{\boldsymbol{Y}} \left[\operatorname{tr}\left\{ (\boldsymbol{W} - \boldsymbol{\Sigma}_*)\boldsymbol{S}^{-1} \right\}^2 \right], \\
\alpha_3 &= E^*_{\boldsymbol{Y}} \left[\operatorname{tr}\left\{ (\boldsymbol{S} - \boldsymbol{\Sigma}_*)\boldsymbol{S}^{-1}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\Sigma}_*)\boldsymbol{S}^{-1} \right\} \right], \\
\alpha_4 &= E^*_{\boldsymbol{Y}} \left[\operatorname{tr}\left\{ (\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\Sigma}_*)(\boldsymbol{\Sigma}_*^{-1} + \boldsymbol{S}^{-1})(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\Sigma}_*)(\boldsymbol{\Sigma}_*^{-1} - \boldsymbol{S}^{-1}) \right\} \right], \\
\alpha_5 &= E^*_{\boldsymbol{U}} E^*_{\boldsymbol{Y}} \left[\operatorname{tr}\left\{ (\boldsymbol{W} - \boldsymbol{\Sigma}_*)(\boldsymbol{\Sigma}_*^{-1} + \boldsymbol{S}^{-1})(\boldsymbol{W} - \boldsymbol{\Sigma}_*)(\boldsymbol{\Sigma}_*^{-1} - \boldsymbol{S}^{-1}) \right\} \right].
\end{aligned}$$

It follows from $\sqrt{n}(\boldsymbol{S} - \boldsymbol{\Sigma}_*) = O_p(1)$ that $\alpha_1 = E^*_{\boldsymbol{Y}}[\operatorname{tr}\{(\boldsymbol{S} - \boldsymbol{\Sigma}_*)\boldsymbol{\Sigma}_*^{-1}\}^2] + o(n^{-1})$ and $\alpha_2 = E^*_{\boldsymbol{U}}[\operatorname{tr}\{(\boldsymbol{W} - \boldsymbol{\Sigma}_*)\boldsymbol{\Sigma}_*^{-1}\}^2] + o(n^{-1})$. The two equations further lead to $\alpha_1 - \alpha_2 = o(n^{-1})$. Similarly, it follows from $\sqrt{n}(\boldsymbol{S}^{-1} - \boldsymbol{\Sigma}_*^{-1}) = O_p(1)$ and $\sqrt{n}(\boldsymbol{W} - \boldsymbol{\Sigma}_*) = O_p(1)$ that $\alpha_5 = o(n^{-1})$. These directly imply that $\beta_1 = 2\alpha_1 + o(n^{-1})$, $\beta_2 = \alpha_1 + \alpha_3 + o(n^{-1})$ and $\beta_3 = \alpha_4 + o(n^{-1})$. Substituting them into (A.1) yields

$$B = \alpha_3 + \frac{1}{2}\alpha_4 + o(n^{-1}). \tag{A.3}$$

By applying (A.2) to α_3 and α_4 using $\sqrt{n}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}) = O_p(1)$, we obtain

$$\alpha_3 = \alpha_{31} - \alpha_{32}, \quad \alpha_4 = -\alpha_{41} - \alpha_{42} + \alpha_{43} + o(n^{-1}), \tag{A.4}$$

where

$$\begin{aligned}
\alpha_{31} &= E^*_{\boldsymbol{Y}}\left[\mathrm{tr}\left\{(\boldsymbol{S} - \boldsymbol{\Sigma}_*)\boldsymbol{S}^{-1}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0})\boldsymbol{S}^{-1}\right\}\right], \\
\alpha_{32} &= E^*_{\boldsymbol{Y}}\left[\mathrm{tr}\left\{(\boldsymbol{S} - \boldsymbol{\Sigma}_*)\boldsymbol{S}^{-1}(\boldsymbol{\Sigma}_* - \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0})\boldsymbol{S}^{-1}\right\}\right], \\
\alpha_{41} &= E^*_{\boldsymbol{Y}}\left[\mathrm{tr}\left\{(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0})(\boldsymbol{\Sigma}_*^{-1} + \boldsymbol{S}^{-1})(\boldsymbol{\Sigma}_* - \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0})(\boldsymbol{\Sigma}_*^{-1} - \boldsymbol{S}^{-1})\right\}\right], \\
\alpha_{42} &= E^*_{\boldsymbol{Y}}\left[\mathrm{tr}\left\{(\boldsymbol{\Sigma}_* - \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0})(\boldsymbol{\Sigma}_*^{-1} + \boldsymbol{S}^{-1})(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0})(\boldsymbol{\Sigma}_*^{-1} - \boldsymbol{S}^{-1})\right\}\right], \\
\alpha_{43} &= E^*_{\boldsymbol{Y}}\left[\mathrm{tr}\left\{(\boldsymbol{\Sigma}_* - \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0})(\boldsymbol{\Sigma}_*^{-1} + \boldsymbol{S}^{-1})(\boldsymbol{\Sigma}_* - \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0})(\boldsymbol{\Sigma}_*^{-1} - \boldsymbol{S}^{-1})\right\}\right].
\end{aligned} \tag{A.5}$$

Combining (A.3), (A.4) and (A.5) yields

$$B = \alpha_{31} - \alpha_{32} - \frac{1}{2}(\alpha_{41} + \alpha_{42} - \alpha_{43}) + o(n^{-1}). \tag{A.6}$$

Next, we give some stochastic expansions that will be used to obtain the leading terms of the quantities in (A.5). Let

$$\boldsymbol{z} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\varepsilon}_i, \quad \boldsymbol{V} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i' - \boldsymbol{I}_p),$$

where $\boldsymbol{\varepsilon}_i = \boldsymbol{\Sigma}_*^{-1/2}(\boldsymbol{y}_i - \boldsymbol{\mu}_*)$. Then

$$\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}_*^{-1/2} = \boldsymbol{I}_p + \frac{1}{\sqrt{n}}\boldsymbol{V} - \frac{1}{n}(\boldsymbol{z}\boldsymbol{z}' - \boldsymbol{I}_p) + O_p(n^{-3/2}). \tag{A.7}$$

Inverting (A.7) leads to

$$\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{S}^{-1}\boldsymbol{\Sigma}_*^{1/2} = \boldsymbol{I}_p - \frac{1}{\sqrt{n}}\boldsymbol{V} + \frac{1}{n}\left\{\boldsymbol{V}^2 + (\boldsymbol{z}\boldsymbol{z}' - \boldsymbol{I}_p)\right\} + O_p(n^{-3/2}). \tag{A.8}$$

Since $\hat{\boldsymbol{\theta}}$ is defined by (2.2), there exists

$$\frac{\partial}{\partial\boldsymbol{\theta}}\hat{d}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}, \boldsymbol{S})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \boldsymbol{\Delta}_{\hat{\boldsymbol{\theta}}}'(\boldsymbol{\Sigma}_*^{1/2} \otimes \boldsymbol{\Sigma}_*^{1/2})\mathrm{vec}(\boldsymbol{S}^{-1}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}\boldsymbol{S}^{-1} - \boldsymbol{S}^{-1}) = \boldsymbol{0}_q,$$

where $\boldsymbol{\Delta}_{\boldsymbol{\theta}}$ is given by (2.7). Substituting (A.8) into the above equation yields

$$\boldsymbol{\Delta}_{\hat{\boldsymbol{\theta}}}'\mathrm{vec}(\boldsymbol{\Omega}_{\hat{\boldsymbol{\theta}}}) = \frac{1}{\sqrt{n}}\boldsymbol{\Delta}_{\hat{\boldsymbol{\theta}}}'\boldsymbol{\Gamma}_{\hat{\boldsymbol{\theta}}}\mathrm{vec}(\boldsymbol{V}) + O_p(n^{-1}), \tag{A.9}$$

where $\boldsymbol{\Omega}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Gamma}_{\boldsymbol{\theta}}$ are given by (2.6). Notice that $\partial\mathrm{vec}(\boldsymbol{\Omega}_{\boldsymbol{\theta}})'\boldsymbol{\Delta}_{\boldsymbol{\theta}}/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \boldsymbol{H}_{\boldsymbol{\theta}_0}$, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = O_p(1)$ and both sides of (A.9) are functions of $\hat{\boldsymbol{\theta}}$. Applying the Taylor expansion

to (A.9) at $\boldsymbol{\theta}_0$ and comparing the $O_p(n^{-1/2})$ term on both sides of the resulting equation, we obtain

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \boldsymbol{H}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}'\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0}\mathrm{vec}(\boldsymbol{V}) + O_p(n^{-1/2}).$$

The above equation further leads to

$$\mathrm{tr}\{(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0})\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{A}\boldsymbol{\Sigma}_*^{-1/2}\} = \frac{1}{\sqrt{n}}\mathrm{vec}(\boldsymbol{A})'\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}\boldsymbol{H}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}'\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0}\mathrm{vec}(\boldsymbol{V}) + O_p(n^{-1}), \quad \text{(A.10)}$$

where $\boldsymbol{A}$ is any $p \times p$ constant matrix.

Applying (A.7), (A.8) and (A.10) to $\boldsymbol{S}$, $\boldsymbol{S}^{-1}$ and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}$ in (A.5), respectively, we obtain

$$\begin{aligned}
\alpha_{31} &= \frac{1}{n}E_{\boldsymbol{Y}}^*\left[\mathrm{vec}(\boldsymbol{V})'\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}\boldsymbol{H}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}'\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0}\mathrm{vec}(\boldsymbol{V})\right] + o(n^{-1}),\\
\alpha_{32} &= -\frac{1}{\sqrt{n}}E_{\boldsymbol{Y}}^*\left[\mathrm{tr}(\boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\theta}_0})\right] + \frac{1}{n}E_{\boldsymbol{Y}}^*\left[\mathrm{tr}\{(\boldsymbol{z}\boldsymbol{z}' - \boldsymbol{I}_p)\boldsymbol{\Omega}_{\boldsymbol{\theta}_0}\}\right]\\
&\quad + \frac{2}{n}E_{\boldsymbol{Y}}^*\left[\mathrm{vec}(\boldsymbol{V})'(\boldsymbol{I}_p \otimes \boldsymbol{\Omega}_{\boldsymbol{\theta}_0})\mathrm{vec}(\boldsymbol{V})\right] + o(n^{-1}),\\
\alpha_{41} &= -\frac{2}{n}E_{\boldsymbol{Y}}^*\left[\mathrm{vec}(\boldsymbol{V})'(\boldsymbol{I}_p \otimes \boldsymbol{\Omega}_{\boldsymbol{\theta}_0})\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}\boldsymbol{H}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}'\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0}\mathrm{vec}(\boldsymbol{V})\right] + o(n^{-1}),\\
\alpha_{42} &= -\frac{2}{n}E_{\boldsymbol{Y}}^*\left[\mathrm{vec}(\boldsymbol{V})'(\boldsymbol{\Omega}_{\boldsymbol{\theta}_0} \otimes \boldsymbol{I}_p)\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}\boldsymbol{H}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}'\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0}\mathrm{vec}(\boldsymbol{V})\right] + o(n^{-1}),\\
\alpha_{43} &= \frac{2}{\sqrt{n}}E_{\boldsymbol{Y}}^*\left[\mathrm{tr}(\boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\theta}_0}^2)\right] - \frac{2}{n}E_{\boldsymbol{Y}}^*\left[\mathrm{tr}\left\{(\boldsymbol{z}\boldsymbol{z}' - \boldsymbol{I}_p)\boldsymbol{\Omega}_{\boldsymbol{\theta}_0}^2\right\}\right]\\
&\quad - \frac{1}{n}E_{\boldsymbol{Y}}^*\left[\mathrm{vec}(\boldsymbol{V})'\left\{(\boldsymbol{\Omega}_{\boldsymbol{\theta}_0} \otimes \boldsymbol{\Omega}_{\boldsymbol{\theta}_0}) + 2(\boldsymbol{\Omega}_{\boldsymbol{\theta}_0}^2 \otimes \boldsymbol{I}_p)\right\}\mathrm{vec}(\boldsymbol{V})\right] + o(n^{-1}).
\end{aligned} \quad \text{(A.11)}$$

Notice that $E_{\boldsymbol{Y}}^*[\boldsymbol{V}] = \boldsymbol{O}_{p,p}$ and $E_{\boldsymbol{Y}}^*[\boldsymbol{z}\boldsymbol{z}' - \boldsymbol{I}_p] = \boldsymbol{O}_{p,p}$. Equation (2.5) implies

$$\begin{aligned}
E_{\boldsymbol{Y}}^*[\mathrm{vec}(\boldsymbol{V})'\boldsymbol{A}\mathrm{vec}(\boldsymbol{V})] &= E_{\boldsymbol{Y}}^*[\mathrm{vec}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' - \boldsymbol{I}_p)'\boldsymbol{A}\mathrm{vec}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' - \boldsymbol{I}_p)]\\
&= \kappa(\boldsymbol{A}) + \mathrm{tr}\{(\boldsymbol{I}_{p^2} + \boldsymbol{K}_p)\boldsymbol{A}\},
\end{aligned} \quad \text{(A.12)}$$

where $\boldsymbol{A}$ is any $p^2 \times p^2$ matrix. Combining (A.11) and (A.12) leads to

$$\begin{aligned}
\alpha_{31} &= \frac{1}{n}\left[\kappa(\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}\boldsymbol{H}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}'\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0}) + \mathrm{tr}\left\{(\boldsymbol{I}_{p^2} + \boldsymbol{K}_p)\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}\boldsymbol{H}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}'\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0}\right\}\right] + o(n^{-1}),\\
\alpha_{32} &= \frac{2}{n}\left\{\kappa(\boldsymbol{\Omega}_{\boldsymbol{\theta}_0} \otimes \boldsymbol{I}_p) + (p+1)\mathrm{tr}\boldsymbol{\Omega}_{\boldsymbol{\theta}_0}\right\} + o(n^{-1}),\\
\alpha_{41} &= -\frac{2}{n}\Big[\kappa((\boldsymbol{I}_p \otimes \boldsymbol{\Omega}_{\boldsymbol{\theta}_0})\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}\boldsymbol{H}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}'\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0})\\
&\quad + \mathrm{tr}\left\{(\boldsymbol{I}_{p^2} + \boldsymbol{K}_p)(\boldsymbol{I}_p \otimes \boldsymbol{\Omega}_{\boldsymbol{\theta}_0})\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}\boldsymbol{H}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}'\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0}\right\}\Big] + o(n^{-1}),\\
\alpha_{42} &= -\frac{2}{n}\Big[\kappa((\boldsymbol{\Omega}_{\boldsymbol{\theta}_0} \otimes \boldsymbol{I}_p)\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}\boldsymbol{H}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}'\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0})\\
&\quad + \mathrm{tr}\left\{(\boldsymbol{I}_{p^2} + \boldsymbol{K}_p)(\boldsymbol{\Omega}_{\boldsymbol{\theta}_0} \otimes \boldsymbol{I}_p)\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}\boldsymbol{H}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\theta}_0}'\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0}\right\}\Big] + o(n^{-1}),\\
\alpha_{43} &= -\frac{1}{n}\left\{\kappa(\boldsymbol{\Omega}_{\boldsymbol{\theta}_0} \otimes \boldsymbol{\Omega}_{\boldsymbol{\theta}_0}) + 2\kappa(\boldsymbol{\Omega}_{\boldsymbol{\theta}_0}^2 \otimes \boldsymbol{I}_p) + (\mathrm{tr}\boldsymbol{\Omega}_{\boldsymbol{\theta}_0})^2 + (2p+3)\mathrm{tr}\boldsymbol{\Omega}_{\boldsymbol{\theta}_0}^2\right\} + o(n^{-1}).
\end{aligned} \quad \text{(A.13)}$$

Notice that $(\boldsymbol{\Omega}_{\boldsymbol{\theta}_0} \otimes \boldsymbol{I}_p) + (\boldsymbol{I}_p \otimes \boldsymbol{\Omega}_{\boldsymbol{\theta}_0}) - (\boldsymbol{I}_p \otimes \boldsymbol{I}_p) = \boldsymbol{\Gamma}_{\boldsymbol{\theta}_0}$. Moreover, using (see Magnus & Neudecker, 1999, p. 47) $\boldsymbol{K}_p(\boldsymbol{\Sigma}_*^{-1/2} \otimes \boldsymbol{\Sigma}_*^{-1/2}) = (\boldsymbol{\Sigma}_*^{-1/2} \otimes \boldsymbol{\Sigma}_*^{-1/2})\boldsymbol{K}_p$ and $\boldsymbol{K}_p \mathrm{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}) = \mathrm{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0})$, we obtain $\mathrm{tr}(\boldsymbol{\Pi}_{\boldsymbol{\theta}_0}\boldsymbol{K}_p) = \mathrm{tr}\boldsymbol{\Pi}_{\boldsymbol{\theta}_0}$, where $\boldsymbol{\Pi}_{\boldsymbol{\theta}_0}$ is given by (2.8). These equations imply that

$$\begin{aligned} \alpha_{31} - \frac{1}{2}(\alpha_{41} + \alpha_{42}) &= \frac{1}{n}\left[\kappa(\boldsymbol{\Pi}_{\boldsymbol{\theta}_0}) + \mathrm{tr}\left\{(\boldsymbol{I}_{p^2} + \boldsymbol{K}_p)\boldsymbol{\Pi}_{\boldsymbol{\theta}_0}\right\}\right] + o(n^{-1}) \\ &= \frac{1}{n}\left\{\kappa(\boldsymbol{\Pi}_{\boldsymbol{\theta}_0}) + 2\mathrm{tr}\boldsymbol{\Pi}_{\boldsymbol{\theta}_0}\right\} + o(n^{-1}). \end{aligned} \tag{A.14}$$

Substituting (A.14) and α_{32} and α_{43} in (A.13) into (A.6) yields the equation (2.9) in Theorem 1.

Acknowledgment

Hirokazu Yanagihara's research was supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), #17700274, 2005–2007.

References

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

[2] Anderson, T. W. (1969). Statistical inference for covariance matrices with linear structure. In *Multivariate Analysis, II* (*Proc. Second Internat. Sympos., Dayton, Ohio*), pp. 55–66 Academic Press, New York.

[3] Balakrishnan, N., & Ambagaspitiya, R. S. (1994). On skew Laplace distribution. *Technical Report, Department of Mathematics & Statistics, McMaster University*, Hamilton, Ontario, Canada.

[4] Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons, Inc., New York.

[5] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

[6] Browne, M. W. & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behav. Res.*, **24**, 445–455.

[7] Cudeck, R. & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behav. Res.*, **18**, 147–167.

[8] Fujikoshi, Y. & Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika*, **84**, 707–716.

[9] De Gooijer, J. G. (1995). Cross-validation criteria for covariance structures. *Comm. Statist. Simulation Comput.*, **24**, 111–147.

[10] Kano, Y. & Miura, A. (2002). *Graphical Multivariate Analysis by AMOS, EQS and CALIS* (expanded edition). Gendai-Sugaku-Sha, Kyoto, Japan. (in Japanese).

[11] Magnus, J. R. & Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics* (revised edition). John Wiley & Sons, New York.

[12] Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

[13] Mallows, C. L. (1995). More comments on C_p. *Technometrics*, **37**, 362–372.

[14] Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, **57**, 519–530.

[15] Siotani, M., Hayakawa, T. & Fujikoshi, Y. (1985). *Modern Multivariate Statistical Analysis: A Graduate Course and Handbook.* American Sciences Press, Columbus, Ohio.

[16] Swain, A. J. (1975). A class of factor analysis estimation procedures with common asymptotic sampling properties. *Psychometrika*, **40**, 315–335.

[17] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–25.

[18] Yanagihara, H. (2005). Selection of covariance structure models in nonnormal data by using information criterion: an application to data from the survey of the Japanese national character. *Proc. Inst. Statist. Math.*, **53**, 133–157 (in Japanese).

[19] Yanagihara, H. (2007a). A family of estimators for multivariate kurtosis in a nonnormal linear regression model. *J. Multivariate Anal.*, **98**, 1–29.

[20] Yanagihara, H. (2007b). Conditions for robustness to nonnormality on test statistics in a GMANOVA model. *J. Japan Statist. Soc.*, **37**, 135–155.

[21] Yanagihara, H. & Yuan, K.-H. (2005). Four improved statistics for contrasting means by correcting skewness and kurtosis. *British J. Math. Statist. Psych.*, **58**, 209–237.

[22] Yanagihara, H., Yuan, K.-H., Fujisawa, H. & Hayashi, K. (2007). A class of cross-validatory model selection criteria. *TR No.* 07-01, *Statistical Research Group*, *Hiroshima University*, Hiroshima, Japan.

[23] Yuan, K.-H. & Bentler, P. M. (2000). Inferences on correlation coefficients in some classes of nonnormal distributions. *J. Multivariate Anal.*, **72**, 230–248.

[24] Yuan, K.-H. & Chan, W. (2005). On nonequivalence of several procedures of structural equation modeling. *Psychometrika*, **70**, 791–798.

TABLE 1. Biases, frequencies and MSEs of information criteria ($\boldsymbol{\Sigma}_* = \boldsymbol{I}_6 + \mathbf{1}_6\mathbf{1}_6'$)

Distri-			C_p		CC_p		MC_p		$MC_{p,\mathrm{N}}$	
bution	Model	R_p	Bias	(Freq.)	Bias	(Freq.)	Bias	(Freq.)	Bias	(Freq.)
	1	0.96	0.25	(0.00)	0.26	(0.01)	0.07	(0.00)	0.05	(0.00)
	2	0.64	0.18	(90.12)	0.19	(88.00)	0.06	(72.94)	0.05	(75.71)
1	3	0.98	0.17	(0.00)	0.19	(0.00)	0.03	(0.00)	0.01	(0.00)
	4	0.65	0.10	(9.50)	0.12	(11.37)	0.02	(21.49)	0.00	(20.01)
	5	0.86	0.02	(0.38)	0.05	(0.62)	0.05	(5.58)	0.02	(4.28)
	MSE		0.21		0.21		0.22		0.43	
	1	1.23	0.44	(0.01)	0.43	(0.08)	0.15	(0.02)	0.23	(0.00)
	2	0.92	0.38	(62.37)	0.36	(81.37)	0.14	(60.68)	0.22	(43.02)
2	3	1.25	0.44	(0.00)	0.35	(0.00)	0.14	(0.00)	0.28	(0.00)
	4	0.94	0.38	(36.75)	0.27	(18.31)	0.13	(34.12)	0.28	(49.47)
	5	1.22	0.38	(0.87)	0.22	(0.24)	0.22	(5.18)	0.38	(7.51)
	MSE		0.32		0.31		0.33		0.61	
	1	0.86	0.18	(0.00)	0.19	(0.00)	0.04	(0.00)	-0.01	(0.00)
	2	0.52	0.11	(98.09)	0.13	(92.14)	0.04	(80.82)	-0.01	(91.26)
3	3	0.87	0.06	(0.00)	0.13	(0.00)	-0.00	(0.00)	-0.10	(0.00)
	4	0.53	-0.02	(1.69)	0.06	(6.52)	-0.01	(12.80)	-0.11	(5.35)
	5	0.71	-0.13	(0.22)	0.01	(1.34)	0.01	(6.38)	-0.13	(3.39)
	MSE		0.17		0.17		0.18		0.36	
	1	1.25	0.47	(0.01)	0.46	(0.07)	0.17	(0.01)	0.25	(0.00)
	2	0.95	0.40	(61.15)	0.39	(81.75)	0.16	(59.49)	0.24	(41.49)
4	3	1.28	0.47	(0.00)	0.37	(0.00)	0.17	(0.00)	0.31	(0.00)
	4	0.96	0.41	(38.12)	0.30	(17.99)	0.16	(35.23)	0.31	(51.38)
	5	1.25	0.41	(0.72)	0.26	(0.20)	0.26	(5.27)	0.41	(7.13)
	MSE		0.33		0.32		0.34		0.62	
	1	1.50	0.65	(0.02)	0.62	(0.09)	0.27	(0.06)	0.42	(0.00)
	2	1.21	0.58	(41.41)	0.56	(77.66)	0.26	(53.41)	0.40	(25.01)
5	3	1.53	0.72	(0.00)	0.52	(0.00)	0.29	(0.01)	0.56	(0.00)
	4	1.22	0.66	(57.52)	0.45	(22.13)	0.28	(41.27)	0.56	(66.50)
	5	1.57	0.73	(1.05)	0.44	(0.11)	0.44	(5.25)	0.73	(8.49)
	MSE		0.44		0.41		0.44		0.79	
	1	2.48	1.55	(0.02)	1.51	(0.26)	1.05	(0.14)	1.30	(0.00)
	2	2.21	1.49	(24.85)	1.45	(73.81)	1.03	(42.84)	1.29	(13.49)
6	3	2.61	1.80	(0.02)	1.49	(0.04)	1.22	(0.04)	1.64	(0.00)
	4	2.30	1.71	(72.70)	1.41	(25.84)	1.19	(51.07)	1.61	(72.56)
	5	3.13	2.29	(2.41)	1.84	(0.05)	1.84	(5.91)	2.29	(13.95)
	MSE		0.82		0.60		0.73		1.52	

The model with the smallest risk among all the candidate models is marked in bold, which is also the best model (overspecified with fewest number of parameters).

TABLE 2. Biases, frequencies and MSEs of information criteria ($\mathbf{\Sigma}_* = \mathrm{diag}(1,1,2,2,3,3)$)

Distribution	Model	R_p	C_p Bias	C_p (Freq.)	CC_p Bias	CC_p (Freq.)	MC_p Bias	MC_p (Freq.)	$MC_{p,\mathrm{N}}$ Bias	$MC_{p,\mathrm{N}}$ (Freq.)
1	1	1.11	0.29	(1.47)	0.30	(1.37)	0.06	(0.47)	0.05	(0.53)
	2	1.12	0.27	(0.22)	0.28	(0.28)	0.05	(0.09)	0.04	(0.11)
	3	0.64	0.12	(83.47)	0.13	(81.44)	0.03	(69.07)	0.01	(71.93)
	4	0.65	0.10	(14.29)	0.12	(15.96)	0.03	(22.64)	0.01	(21.39)
	5	0.86	0.02	(0.55)	0.06	(0.95)	0.06	(7.73)	0.02	(6.04)
	MSE		0.22		0.22		0.24		0.43	
2	1	1.38	0.50	(2.92)	0.46	(9.87)	0.15	(4.48)	0.24	(1.26)
	2	1.38	0.48	(0.45)	0.45	(1.98)	0.14	(1.19)	0.23	(0.24)
	3	0.95	0.42	(81.48)	0.30	(75.46)	0.15	(66.04)	0.32	(70.68)
	4	0.95	0.40	(14.50)	0.28	(12.50)	0.15	(22.76)	0.31	(21.54)
	5	1.22	0.38	(0.65)	0.22	(0.19)	0.22	(5.52)	0.38	(6.29)
	MSE		0.35		0.37		0.37		0.61	
3	1	1.01	0.21	(0.56)	0.23	(0.15)	0.04	(0.02)	-0.03	(0.14)
	2	1.01	0.19	(0.10)	0.22	(0.05)	0.03	(0.01)	-0.04	(0.04)
	3	0.53	-0.00	(84.21)	0.08	(79.43)	0.00	(68.43)	-0.11	(71.97)
	4	0.53	-0.02	(14.56)	0.06	(17.91)	-0.00	(22.17)	-0.11	(21.51)
	5	0.71	-0.13	(0.57)	0.01	(2.47)	0.01	(9.37)	-0.13	(6.34)
	MSE		0.17		0.18		0.19		0.36	
4	1	1.40	0.51	(2.94)	0.48	(9.95)	0.17	(4.34)	0.26	(1.27)
	2	1.40	0.50	(0.53)	0.47	(2.01)	0.16	(1.22)	0.25	(0.29)
	3	0.97	0.44	(81.04)	0.33	(74.44)	0.18	(66.28)	0.34	(70.60)
	4	0.98	0.43	(14.78)	0.31	(13.39)	0.18	(22.50)	0.33	(21.40)
	5	1.25	0.41	(0.70)	0.26	(0.21)	0.26	(5.67)	0.41	(6.43)
	MSE		0.36		0.38		0.38		0.62	
5	1	1.64	0.71	(2.99)	0.65	(16.23)	0.27	(7.13)	0.44	(1.35)
	2	1.63	0.69	(0.53)	0.63	(3.17)	0.26	(2.20)	0.43	(0.27)
	3	1.25	0.72	(81.10)	0.51	(67.97)	0.33	(64.14)	0.62	(70.31)
	4	1.25	0.71	(14.68)	0.49	(12.56)	0.33	(21.72)	0.61	(21.47)
	5	1.58	0.74	(0.70)	0.45	(0.07)	0.45	(4.80)	0.74	(6.60)
	MSE		0.48		0.50		0.50		0.79	
6	1	2.55	1.55	(2.86)	1.48	(23.28)	1.00	(8.93)	1.27	(1.25)
	2	2.54	1.53	(0.51)	1.45	(4.71)	0.99	(3.23)	1.26	(0.35)
	3	2.60	2.07	(80.97)	1.72	(60.99)	1.51	(62.80)	1.97	(69.97)
	4	2.60	2.05	(14.80)	1.70	(10.98)	1.51	(21.37)	1.96	(21.38)
	5	3.16	2.32	(0.86)	1.87	(0.03)	1.87	(3.67)	2.32	(7.05)
	MSE		1.06		0.88		0.99		1.60	

The model with the smallest risk among all the candidate models is marked in bold, which is also the best model (overspecified with fewest number of parameters) except for distribution condition 6.