HIROSHIMA MATH. J.
44 (2014), 315–326

Selecting a shrinkage parameter in structural equation modeling with a near singular covariance matrix by the GIC minimization method

Ami KAMADA, Hirokazu YANAGIHARA, Hirofumi WAKAKI and Keisuke FUKUI

(Received February 13, 2012)
(Revised May 7, 2014)

ABSTRACT. In the structural equation modeling, unknown parameters of a covariance matrix are derived by minimizing the discrepancy between a sample covariance matrix and a covariance matrix having a specified structure. When a sample covariance matrix is a near singular matrix, Yuan and Chan (2008) proposed the estimation method to use an adjusted sample covariance matrix instead of the sample covariance matrix in the discrepancy function. The adjusted sample covariance matrix is defined by adding a scalar matrix with a shrinkage parameter to the existing sample covariance matrix. They used a constant value as the shrinkage parameter, which was chosen based solely on the sample size and the number of dimensions of the observation, and not on the data itself. However, selecting the shrinkage parameter from the data may lead to a greater improvement in prediction compared to the use of a constant shrinkage parameter. Hence, we propose an information criterion for selecting the shrinkage parameter, and attempt to select the shrinkage parameter by an information criterion minimization method. The proposed information criterion is based on the discrepancy function measured by the normal theory maximum likelihood. Using the Monte Carlo method, we demonstrate that the proposed criterion works well in the sense that the prediction accuracy of an estimated covariance matrix is improved.

1. Introduction

Structural equation modeling (SEM) has been widely used in many fields, especially in social and behavioral sciences (see e.g., Bollen (1989), and Yuan and Bentler (2007)). In SEM, unknown parameters of a covariance matrix are

The Second author is supported by the Ministry of Education, Science, Sports, and Culture, a Grant-in-Aid for Challenging Exploratory Research, #25540012, 2013–2015.

The Third author is supported by the Ministry of Education, Science, Sports, and Culture, Grant-in-Aid for Scientific Research (C), #24500343, 2012–2014.

2010 *Mathematics Subject Classification*. 62H12, 62F07.

Key words and phrases. bias correction, GIC, model selection, near singular covariance matrix, SEM, shrinkage parameter.

derived by minimizing the discrepancy between a sample covariance matrix and a covariance matrix having a specified structure.

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ be independent random samples from $\boldsymbol{x}$ distributed according to a p-variate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where N is the sample size. We are interested in modeling the population covariance matrix $\boldsymbol{\Sigma}$. Denote the model of interest as $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$. For simplicity, we write $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ as $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. Let $\boldsymbol{S}$ be an unbiased estimator of $\boldsymbol{\Sigma}$, i.e.,

$$\boldsymbol{S} = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})',$$

where $\bar{\boldsymbol{x}}$ is the sample mean of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ defined by $\bar{\boldsymbol{x}} = N^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i$. Then, the candidate model is represented by

$$M : n\boldsymbol{S} \sim W_p(n, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}), \tag{1}$$

where $n = N - 1$. Suppose that $\boldsymbol{\Sigma}_0$ is the true covariance matrix, i.e., $Cov[\boldsymbol{x}] = \boldsymbol{\Sigma}_0$. The true model is represented by

$$M_0 : n\boldsymbol{S} \sim W_p(n, \boldsymbol{\Sigma}_0). \tag{2}$$

If the covariance structure can be correctly specified, then there exists $\boldsymbol{\theta}_0$ such that $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}$. The classical approach to SEM fits the sample covariance matrix $\boldsymbol{S}$ by $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ through minimizing the normal theory maximum likelihood (ML) discrepancy function as

$$F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) = \mathrm{tr}(\boldsymbol{S}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}) - \log|\boldsymbol{S}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}| - p. \tag{3}$$

Then, the ML estimator of $\boldsymbol{\theta}$, which is represented by $\hat{\boldsymbol{\theta}}$, is defined by

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

In general, $\hat{\boldsymbol{\theta}}$ is obtained using a modification of Newton's algorithm (see e.g., Lee and Jennrich (1979)), which requires an iteration process to solve the estimating equation. When $\boldsymbol{S}$ is near singular (not full rank), the iteration process for obtaining $\hat{\boldsymbol{\theta}}$ will be very unstable and may require hundreds of iterations to reach convergence (e.g., Boomsma (1985)). A near singular $\boldsymbol{S}$ often occurs in practical data analysis due to not only small samples but also multicollinearity or missing data even when sample size is quite large (Wothke (1993)). When $\boldsymbol{S}$ is literally singular, it is very likely that the iteration will never converge.

In order to avoid such a problem, Yuan and Chan (2008) proposed a new method in which $\boldsymbol{\theta}$ is estimated by minimizing $F(\boldsymbol{S}_a, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, where $\boldsymbol{S}_a = \boldsymbol{S} + a\boldsymbol{I}_p$, a is a small positive value and $\boldsymbol{I}_p$ is a p-dimensional identity matrix. Here, a is

commonly referred to as the shrinkage parameter. Hence, a new estimator of $\boldsymbol{\theta}$ is defined by

$$\hat{\boldsymbol{\theta}}_a = \arg \min_{\boldsymbol{\theta}} F(\boldsymbol{S}_a, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}).$$

Although $\hat{\boldsymbol{\theta}}_a$ has a constant bias, under LISREL models (see Jöreskog and Sörbom (1996), pp. 1–3), Yuan and Chan (2008) reported that $\hat{\boldsymbol{\theta}}_a$ can be adjusted to a consistent estimator through a simple procedure when the covariance structure is the correct model. The adjusted estimator is defined as

$$\tilde{\boldsymbol{\theta}}_a = \hat{\boldsymbol{\theta}}_a - a\boldsymbol{j},$$

where $\boldsymbol{j}$ is a q-dimensional vector whose elements are ones corresponding to the parameters on the diagonals of the covariance matrix, and otherwise are zero. They also studied for the case that $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is not correctly specified. There exists a unique vector $\boldsymbol{\theta}_*$ such that

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{a*}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_*} + a\boldsymbol{I}_p, \tag{4}$$

where $\boldsymbol{\theta}_{a*}$ is a population parameter minimizing $F(\boldsymbol{\Sigma}_0 + a\boldsymbol{I}_p, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, i.e.,

$$\boldsymbol{\theta}_{a*} = \arg \min_{\boldsymbol{\theta}} F(\boldsymbol{\Sigma}_0 + a\boldsymbol{I}_p, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}). \tag{5}$$

Then, $\hat{\boldsymbol{\theta}}_a$ and $\tilde{\boldsymbol{\theta}}_a$ are consistent for $\boldsymbol{\theta}_{a*}$ and $\boldsymbol{\theta}_*$, respectively. If $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is correctly specified, $\boldsymbol{\theta}_{a*} = \boldsymbol{\theta}_0 + a\boldsymbol{j}$ and $\boldsymbol{\theta}_* = \boldsymbol{\theta}_0$.

The selection of the shrinkage parameter is crucial because if the shrinkage parameter is changed, the estimate will be also changed. In Yuan and Chan (2008), the shrinkage parameter was taken to be a constant, determined by only N and p. This means that the shrinkage parameter was not chosen based on the data. However, it is possible that the prediction could be improved by basing the shrinkage parameter on the data itself. Furthermore, it does not always guarantee that the estimator is proper solution by fixed a. Therefore, we attempt to select the shrinkage parameter based on the predictive Kullback-Leibler (KL) discrepancy (Kullback and Leibler (1951)). The basic idea is to measure the goodness of fit of the model by the risk function assessed by the predictive KL discrepancy. In the present paper, our objective is to select the appropriate value of a by minimizing the risk function. However, we cannot directly use the risk function to select a because the risk function includes unknown parameters. Hence, instead of the risk function itself, we use its estimator.

Akaike's information criterion (AIC) (Akaike (1973)) is an estimator of the risk function assessed by the predictive KL information (for the AIC for SEM, see, e.g., Cudeck and Brown (1983), Akaike (1987), Ichikawa and Konishi (1999), Yanagihara (2005)). The objective of the present study may be

achieved by minimizing the AIC rather than the risk function. In general, the AIC is defined by adding the bias to the risk function, i.e., the number of independent parameters divided by n, to the KL discrepancy function with an estimated parameter, which is referred to as a sample discrepancy function. However, the bias term of the AIC is obtained under the situation that the discrepancy function for estimating $\boldsymbol{\theta}$ is the same as that for evaluating the model fit. In the present paper, the discrepancy function for estimating $\boldsymbol{\theta}$ is

$$F(\boldsymbol{S}_a, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) = F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) + a\,\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}) - \log|\boldsymbol{S}_a| + \log|\boldsymbol{S}|,$$

and that for evaluating the model is $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$. Since the two functions are different, we cannot use the bias term of the ordinary AIC. Therefore, we must revaluate the bias using the same approach as the generalized information criterion (GIC) proposed by Konishi and Kitagawa (1996). Hence, we denote the proposed information criterion as GIC(a). We define GIC(a) by adding an estimator of the revaluated bias to the sample discrepancy function $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a})$. Then, the best a is chosen by minimizing GIC(a).

The remainder of the present paper is organized as follows: In Section 2, we obtain GIC(a) from a stochastic expansion of $\hat{\boldsymbol{\theta}}_a$. In Section 3, we verify the performance of our criteria using the Monte Carlo method. In Section 4, we present conclusions and discussions. The proof of the theorem presented herein is provided in the Appendix.

2. GIC for selecting the shrinkage parameter

In order to select the best a, we consider the risk function between the true model and the candidate model. Let $\mathscr{L}(\boldsymbol{\Sigma})$ be an expected ML discrepancy function defined by

$$\begin{aligned}\mathscr{L}(\boldsymbol{\Sigma}) &= E[F(\boldsymbol{S}, \boldsymbol{\Sigma})] \\ &= \mathrm{tr}(\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1}) - E[\log|\boldsymbol{S}|] + \log|\boldsymbol{\Sigma}| - p.\end{aligned}$$

In this paper, E denotes the expectation under the true model M_0 in (2) with respect to $\boldsymbol{S}$. We measure the discrepancy between the candidate model M in (1) and the true model M_0 in (2) by the predictive KL discrepancy function. Then, we define the risk function assessed by the predictive ML discrepancy in (3) as

$$R = E[\mathscr{L}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a})].$$

We regard the shrinkage parameter a having the smallest R as the principle best model. Obtaining an unbiased estimator of R will allow us to correctly evaluate the discrepancy between the data and the model, which will further

facilitate the selection of the best shrinkage parameter. A rough estimator of R is the sample ML discrepancy function $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})$. However, since $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})$ has a bias, the information criterion can be defined as $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}) + \hat{B}$, where $\hat{B}$ is an estimator of the bias given as

$$B = R - E[F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})]. \tag{6}$$

Henceforth, in order to derive $\hat{B}$, we calculate a limiting value of B.

Let

$$\boldsymbol{\Delta}_{\boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}'} \operatorname{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}), \tag{7}$$

and

$$\boldsymbol{G}_{\boldsymbol{\theta}_{a*}} = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} F(\boldsymbol{\Sigma}_0 + a\boldsymbol{I}_p, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{a*}}, \tag{8}$$

where

$$\begin{aligned} &\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} F(\boldsymbol{\Sigma}_0 + a\boldsymbol{I}_p, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \\ &\quad = 2\boldsymbol{\Delta}_{\boldsymbol{\theta}}'(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\Sigma}_0 + a\boldsymbol{I}_p)\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \otimes \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1})\boldsymbol{\Delta}_{\boldsymbol{\theta}} - \boldsymbol{\Delta}_{\boldsymbol{\theta}}'(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \otimes \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1})\boldsymbol{\Delta}_{\boldsymbol{\theta}} \\ &\qquad - \sum_{i,j}^{q} \operatorname{tr}\{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\Sigma}_0 + a\boldsymbol{I}_p - \boldsymbol{\Sigma}_{\boldsymbol{\theta}})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \ddot{\boldsymbol{\Sigma}}_{\boldsymbol{\theta} ij}\} \boldsymbol{e}_i \boldsymbol{e}_j'. \end{aligned}$$

Here, $\boldsymbol{e}_i$ is a q-dimensional vector, the ith element of which is 1, with all others being 0, and $\ddot{\boldsymbol{\Sigma}}_{\boldsymbol{\theta} ij} = \partial^2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}} / \partial \theta_i \partial \theta_j$. Since $\boldsymbol{\theta}_{a*}$ is the minimizer of $F(\boldsymbol{\Sigma}_0 + a\boldsymbol{I}_p, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, $\boldsymbol{G}_{\boldsymbol{\theta}_{a*}}$ is a nonsingular matrix. Using the above notation, we have the following theorem for the bias.

THEOREM 1. *Suppose that a set of standard regularity conditions, as given in Browne* (1984) *or Yuan and Bentler* (1997)*, is satisfied. Then, the bias of* $E[F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})]$ *is expanded as*

$$B = \frac{2}{n} \operatorname{tr}\{\boldsymbol{\Delta}_{\boldsymbol{\theta}_*} \boldsymbol{G}_{\boldsymbol{\theta}_{a*}}^{-1} \boldsymbol{\Delta}_{\boldsymbol{\theta}_{a*}}' (\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{a*}}^{-1} \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}^{-1} \otimes \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{a*}}^{-1} \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}^{-1})\} + O(n^{-2}). \tag{9}$$

The proof of this theorem, which is derived by modifying the results presented in Yanagihara, Himeno, and Yuan (2010), is given in the Appendix.

By replacing $\boldsymbol{\theta}_{a*}$, $\boldsymbol{\theta}_*$, and $\boldsymbol{\Sigma}_0$ by neglecting $O(n^{-2})$ in (9) with $\hat{\boldsymbol{\theta}}_a$, $\tilde{\boldsymbol{\theta}}_a$, and $\boldsymbol{S}$, respectively, an estimator of B is given by

$$\hat{B} = \frac{2}{n} \operatorname{tr}\{\boldsymbol{\Delta}_{\hat{\boldsymbol{\theta}}_a} \boldsymbol{G}_{\hat{\boldsymbol{\theta}}_a}^{-1} \boldsymbol{\Delta}_{\hat{\boldsymbol{\theta}}_a}' (\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a}^{-1} \boldsymbol{S} \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}^{-1} \otimes \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a}^{-1} \boldsymbol{S} \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}^{-1})\}.$$

Thus, the information criterion for selecting a (GIC(a)) is defined by

$$\mathrm{GIC}(a) = F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}) + \hat{B}.$$

Let A be a set $A = \{a \mid a \geq 0 \text{ and } \tilde{\boldsymbol{\theta}}_a \text{ gives a proper solution}\}$. Then, the best a is chosen by minimizing GIC(a), i.e.,

$$\hat{a} = \arg \min_{a \in A} \mathrm{GIC}(a).$$

When the candidate model is correctly specified, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{a*}} = \boldsymbol{\Sigma}_a$. Then, the bias becomes simple, as in the following corollary.

COROLLARY 1. *If the candidate model is correctly specified, the bias of* $E[F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})]$ *is expanded as*

$$B = \frac{2}{n} q + O(n^{-2}).$$

This corollary indicates that the bias does not depend on a by neglecting the $O(n^{-2})$ term when the candidate model is correctly specified. Hence, the best a is the value that minimizes $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})$ in A.

3. Monte Carlo results

In this section, we compare the risk functions of estimated $\boldsymbol{\Sigma}$ obtained from the following methods.

- Method 1 (new method): We estimate $\boldsymbol{\Sigma}$ by $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_{\hat{a}}}$, where $\hat{a}$ is selected by minimizing GIC(a).
- Method 2 (Yuan and Chan's (YC) method): We estimate $\boldsymbol{\Sigma}$ by $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_{p/N}}$.
- Method 3 (ordinary ML method): We estimate $\boldsymbol{\Sigma}$ by $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}$.

Actually, since $-E[\log|\boldsymbol{S}|] - p$ in the expected ML discrepancy does not depend on the result of a selection of a, we evaluated the following expectations:

$$R_{\mathrm{new}} = E[\mathscr{L}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_{\hat{a}}})] + \alpha, \qquad R_{\mathrm{YC}} = E[\mathscr{L}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_{p/N}})] + \alpha, \qquad R_{\mathrm{ML}} = E[\mathscr{L}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}})] + \alpha,$$

where $\alpha = E[\log|\boldsymbol{S}|] + p$. In the simulation, we used the confirmatory factor model, which is included in the LISREL model, as the true model M_0, i.e., the true covariance matrix is $\boldsymbol{\Sigma}_0 = \boldsymbol{\Lambda}_0 \boldsymbol{\Phi}_0 \boldsymbol{\Lambda}_0' + \boldsymbol{\Psi}_0$, where $\boldsymbol{\Lambda}_0$ is the true factor loading matrix, $\boldsymbol{\Phi}_0$ is the true correlation matrix, and $\boldsymbol{\Psi}_0$ is the true covariance matrix of the measurement errors. In this simulation, we defined $\boldsymbol{\Psi}_0 = \boldsymbol{I}_p - \mathrm{diag}(\boldsymbol{\Lambda}_0 \boldsymbol{\Phi}_0 \boldsymbol{\Lambda}_0')$. As the true model, we used the two models specified by the following parameters:

$$\text{Case 1:}\quad \Lambda_0 = \begin{pmatrix} \boldsymbol{b} & \mathbf{0}_5 \\ \mathbf{0}_5 & \boldsymbol{b} \\ \mathbf{0}_5 & \boldsymbol{b} \end{pmatrix}, \qquad \Phi_0 = \begin{pmatrix} 1.0 & .30 \\ .30 & 1.0 \end{pmatrix},$$

$$\text{Case 2:}\quad \Lambda_0 = \begin{pmatrix} \boldsymbol{b} & \mathbf{0}_5 & \mathbf{0}_5 \\ \mathbf{0}_5 & \boldsymbol{b} & \mathbf{0}_5 \\ \mathbf{0}_5 & \mathbf{0}_5 & \boldsymbol{b} \end{pmatrix}, \qquad \Phi_0 = \begin{pmatrix} 1.0 & .30 & .40 \\ .30 & 1.0 & .30 \\ .40 & .30 & 1.0 \end{pmatrix},$$

where $\boldsymbol{b} = (.70, .70, .75, .80, .80)'$ and $\mathbf{0}_q$ is a q-dimensional vector of zeros. The candidate model used in the simulation was also the confirmatory factor model, i.e., the covariance matrix $\Sigma_\theta = \Lambda\Phi\Lambda' + \Psi$, where $\Psi = \mathrm{diag}(\psi_1, \ldots, \psi_p)$. In the case 1, we used the confirmatory three-factor model as the candidate model. On the other hand, the confirmatory two-factor model was used as the candidate model in the case 2. Hence, λ and Φ in the candidate models were

$$\text{Case 1:}\quad \Lambda = \begin{pmatrix} \lambda_1 & \mathbf{0}_5 & \mathbf{0}_5 \\ \mathbf{0}_5 & \lambda_2 & \mathbf{0}_5 \\ \mathbf{0}_5 & \mathbf{0}_5 & \lambda_3 \end{pmatrix}, \qquad \Phi = \begin{pmatrix} 1.0 & \phi_{12} & \phi_{13} \\ \phi_{12} & 1.0 & \phi_{23} \\ \phi_{13} & \phi_{23} & 1.0 \end{pmatrix},$$

$$\text{Case 2:}\quad \Lambda = \begin{pmatrix} \lambda_1 & \mathbf{0}_5 \\ \mathbf{0}_5 & \lambda_2 \\ \mathbf{0}_5 & \lambda_3 \end{pmatrix}, \qquad \Phi = \begin{pmatrix} 1.0 & \phi_{12} \\ \phi_{12} & 1.0 \end{pmatrix}.$$

It is easy to see that the candidate model in the case 1 is overspecified, and that in the case 2 is underspecified. In order to obtain smaller sample sizes, we chose $N = 30$, 50, and 100. The number of replications is 1000.

In order to calculate R_{new}, R_{YC}, and R_{ML}, we first obtained an estimator of θ for each method using R ver. 2.12.1. We then counted the frequencies when the estimate of θ is the proper solution (i.e., an estimator of Σ is positive define). Next, we recorded the value of $\mathscr{L}(\hat{\Sigma})$ for each method, where $\hat{\Sigma}$ is an estimated Σ for each method. After the replication was finished, we obtained the arithmetic mean of $\mathscr{L}(\hat{\Sigma})$ for each method. If all of the estimators are proper solutions, then the arithmetic mean is regarded as a target risk function.

From Table 1, when $N = 30$ in the case 1, the R_{new} was obtained, but R_{YC} and R_{ML} were not obtained because there were several improper solutions for $a = p/N$ and 0. When $N = 50$ and 100 in the case 1, since there were no improper solutions, we could obtain all risk functions. Then, R_{new} was the smallest. On the other hand, in the case 2, R_{new} and R_{YC} were obtained, but R_{ML} was not obtained. Then, R_{new} was smaller than R_{YC}. Hence, the

Table 1. Frequencies of the proper solutions and the risk functions for each method

Case	N	Frequency			Risk		
		New	YC	ML	New	YC	ML
1	30	1000	996	987	16.8295	—	—
	50	1000	1000	1000	15.9808	15.9858	16.0088
	100	1000	1000	1000	15.5024	15.5044	15.5067
2	30	1000	1000	972	19.2521	19.3887	—
	50	1000	1000	987	16.1748	16.2869	—
	100	1000	1000	990	14.1732	14.2618	—

proposed information criterion works well in the sense that the prediction accuracy of an estimated covariance matrix is improved.

4. Conclusion and discussion

In the present paper, we proposed a GIC for selecting the shrinkage parameter, which is used to obtain the estimator for SEM with a near singular covariance matrix. In order to derive the GIC, we revaluated the bias of the risk function. Then, GIC(a) was obtained by adding the estimator of the revaluated bias to the sample discrepancy function. We have observed that when the candidate model is correctly specified, the bias does not depend on a when the $O(n^{-2})$ term is neglected, i.e., the bias term is equivalent to that of the AIC. This means that the best a is the value that minimizes $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})$ under the condition that $\tilde{\boldsymbol{\theta}}_a$ gives a proper solution. In the Monte Carlo results, an estimate of $\tilde{\boldsymbol{\theta}}_{\hat{a}}$ was always a proper solution, and the risk function of the estimated covariance matrix based on $\tilde{\boldsymbol{\theta}}_{\hat{a}}$ with the selected a was the smallest.

In this paper, we assumed that data has normality. If we do not assume normality to data, a kurtosis will appear in the bias to the risk function. Hence, an estimator of kurtosis will be required to estimate the bias. Unfortunately, Yanagihara (2007) reported that such an estimator gives a poor value unless the sample is huge. When the sample size is large enough, a sample covariance matrix will not become a near singular matrix in most cases. A near singular sample covariance matrix occurs frequently under the small or moderate sample sizes. This is almost the same as a well-known fact that a multicollinearity frequently occur under the small or moderate sample. In practice, we confirmed such results through many simulation experiments. Hence, it is suitable to assume not the large sample case but the small or moderate sample case under a near singular sample covariance matrix. There-

fore, at present, we judge that it is necessary to deal with the case of nonnormal when a sample covariance matrix is a near singular matrix.

Appendix

The derivation of the risk function and the proof of Theorem 1 are presented in this appendix. First, we derive the risk function. In this paper, we measure the discrepancy between the candidate model M in (1) and the true model M_0 in (2) by the following discrepancy function:

$$\int \log \frac{f(\boldsymbol{W}|n, \boldsymbol{\Sigma}_0)}{f(\boldsymbol{W}|n, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})} f(\boldsymbol{W}|n, \boldsymbol{\Sigma}_0) d\boldsymbol{W} = \frac{n}{2}\{\mathscr{L}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}) - \mathscr{L}(\boldsymbol{\Sigma}_0)\}.$$

By omitting the terms that do not depend on a, we have

$$\int F(\boldsymbol{W}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}) f(\boldsymbol{W}|n, \boldsymbol{\Sigma}_0) d\boldsymbol{W} = \mathscr{L}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}).$$

Hence, we define the risk function as R in Section 2.

Next, we prove Theorem 1. The bias of $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})$, defined in (6), can be written as

$$B = E[\mathscr{L}(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}) - F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})] = E[\mathrm{tr}\{\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}^{-1}(\boldsymbol{\Sigma}_0 - \boldsymbol{S})\}]. \tag{A1}$$

Since $\boldsymbol{\Sigma}_0 - \boldsymbol{S} = O_p(n^{-1/2})$ and $E[\boldsymbol{S}] = \boldsymbol{\Sigma}_0$, by applying the Taylor expansion to $\mathrm{tr}\{\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}^{-1}(\boldsymbol{\Sigma}_0 - \boldsymbol{S})\}$ at $\tilde{\boldsymbol{\theta}}_a = \boldsymbol{\theta}_*$, we derive

$$E[\mathrm{tr}\{\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}^{-1}(\boldsymbol{\Sigma}_0 - \boldsymbol{S})\}] = E[\boldsymbol{d}_{\boldsymbol{\theta}_*}(\tilde{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_*)] + O(n^{-2}),$$

where $\boldsymbol{\theta}_*$ is given by (4), and

$$\boldsymbol{d}_{\boldsymbol{\theta}_*} = \frac{\partial}{\partial \boldsymbol{\theta}'} \mathrm{tr}\{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\Sigma}_0 - \boldsymbol{S})\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*}.$$

The remainder term of the above expectation is $O(n^{-2})$ because $\tilde{\boldsymbol{\theta}}_a$ can be expressed as a function of $\boldsymbol{V} = n^{1/2}(\boldsymbol{S} - \boldsymbol{\Sigma}_0)$ which has an asymptotic normality and general cumulants of elements of $\boldsymbol{V}$ may be expanded as a power series in n^{-1} (see e.g., Hall, 1992, p. 46). Indeed, an $n^{-3/2}$ term of the stochastic expansion of $\mathrm{tr}\{\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}^{-1}(\boldsymbol{\Sigma}_0 - \boldsymbol{S})\}$ can be expressed as the third-order polynomial of elements of $\boldsymbol{V}$. Since $\boldsymbol{V}$ has an asymptotic normality, an expectation of the odd-order polynomial of element $\boldsymbol{V}$ becomes $O(n^{-1/2})$. Consequently, the expectation of the $n^{-3/2}$ term of the stochastic expansion becomes not $O(n^{-3/2})$ but $O(n^{-2})$. Let $\boldsymbol{\Gamma}_{\boldsymbol{\theta}} = (\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \otimes \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1})$. From this expression, we obtain

$$\boldsymbol{d}_{\boldsymbol{\theta}_*} = \mathrm{vec}'\{\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}^{-1}(\boldsymbol{S} - \boldsymbol{\Sigma}_0)\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}^{-1}\}\boldsymbol{\Delta}_{\boldsymbol{\theta}_*} = \frac{1}{\sqrt{n}} \mathrm{vec}'(\boldsymbol{V})\boldsymbol{\Gamma}_{\boldsymbol{\theta}_*}\boldsymbol{\Delta}_{\boldsymbol{\theta}_*}. \tag{A2}$$

Since $\hat{\boldsymbol{\theta}}_a$ is the minimizer of $F(\boldsymbol{S}_a, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, $\partial F(\boldsymbol{S}_a, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_a} = \mathbf{0}_q$ is satisfied. Then, under a set of standard regularity conditions, the following equation is derived.

$$\mathbf{0}_q = \boldsymbol{\Delta}'_{\hat{\boldsymbol{\theta}}_a} \operatorname{vec}\{\boldsymbol{\Sigma}^{-1}_{\hat{\boldsymbol{\theta}}_a}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a} - \boldsymbol{\Sigma}_0 - a\boldsymbol{I}_p)\boldsymbol{\Sigma}^{-1}_{\hat{\boldsymbol{\theta}}_a}\} - \frac{1}{\sqrt{n}}\boldsymbol{\Delta}'_{\hat{\boldsymbol{\theta}}_a}\boldsymbol{\Gamma}_{\hat{\boldsymbol{\theta}}_a} \operatorname{vec}(\boldsymbol{V}).$$

Hence, we obtain

$$\boldsymbol{\Delta}'_{\hat{\boldsymbol{\theta}}_a} \operatorname{vec}\{\boldsymbol{\Sigma}^{-1}_{\hat{\boldsymbol{\theta}}_a}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a} - \boldsymbol{\Sigma}_0 - a\boldsymbol{I}_p)\boldsymbol{\Sigma}^{-1}_{\hat{\boldsymbol{\theta}}_a}\} = \frac{1}{\sqrt{n}}\boldsymbol{\Delta}'_{\hat{\boldsymbol{\theta}}_a}\boldsymbol{\Gamma}_{\hat{\boldsymbol{\theta}}_a} \operatorname{vec}(\boldsymbol{V}). \tag{A3}$$

Note that $n^{1/2}(\hat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_{a*}) = O_p(1)$ and that both sides of (A3) are functions of $\hat{\boldsymbol{\theta}}_a$, where $\boldsymbol{\theta}_{a*}$ is given by (5). Applying the Taylor expansion to (A3) at $\hat{\boldsymbol{\theta}}_a = \boldsymbol{\theta}_{a*}$ and comparing the $O_p(n^{-1})$ term on both sides of the resulting equation, we obtain

$$\hat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_{a*} = \frac{1}{\sqrt{n}}\boldsymbol{G}^{-1}_{\boldsymbol{\theta}_{a*}}\boldsymbol{\Delta}'_{\boldsymbol{\theta}_{a*}}\boldsymbol{\Gamma}_{\boldsymbol{\theta}_{a*}} \operatorname{vec}(\boldsymbol{V}) + O_p(n^{-1}),$$

where $\boldsymbol{\Delta}_{\boldsymbol{\theta}}$ and $\boldsymbol{G}_{\boldsymbol{\theta}}$ are given by (7) and (8), respectively. Note that

$$\begin{aligned} E[\operatorname{vec}(\boldsymbol{V}) \operatorname{vec}'(\boldsymbol{V})] &= nE[\operatorname{vec}(\boldsymbol{S} - \boldsymbol{\Sigma}_0) \operatorname{vec}'(\boldsymbol{S} - \boldsymbol{\Sigma}_0)] \\ &= nCov[\operatorname{vec}(\boldsymbol{S})] \\ &= (\boldsymbol{I}_{p^2} + \boldsymbol{K}_p)(\boldsymbol{\Sigma}_0 \otimes \boldsymbol{\Sigma}_0), \end{aligned}$$

where $\boldsymbol{K}_p$ is the commutation matrix (see Magnus and Neudecker (1999), p. 48). Therefore,

$$\begin{aligned} B &= E[\boldsymbol{d}_{\boldsymbol{\theta}_*}(\hat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_{a*})] + O(n^{-2}) \\ &= \frac{1}{n} \operatorname{tr}\{\boldsymbol{\Gamma}_{\boldsymbol{\theta}_*}\boldsymbol{\Delta}_{\boldsymbol{\theta}_*}\boldsymbol{G}^{-1}_{\boldsymbol{\theta}_{a*}}\boldsymbol{\Delta}'_{\boldsymbol{\theta}_{a*}}\boldsymbol{\Gamma}_{\boldsymbol{\theta}_{a*}}(\boldsymbol{I}_{p^2} + \boldsymbol{K}_p)(\boldsymbol{\Sigma}_0 \otimes \boldsymbol{\Sigma}_0)\} + O(n^{-2}). \end{aligned} \tag{A4}$$

Consequently, by using the equations $\boldsymbol{K}_p(\boldsymbol{A} \otimes \boldsymbol{C}) = (\boldsymbol{C} \otimes \boldsymbol{A})\boldsymbol{K}_p$ and $\boldsymbol{K}_p \operatorname{vec}(\boldsymbol{C}) = \operatorname{vec}(\boldsymbol{C}')$ (see Magnus and Neudecker (1999), p. 47), the equation (9) in Theorem 1 is derived.

Acknowledgement

We wish to express our deepest gratitude to Prof. Y. Kano of Osaka University for his valuable advices. We would also like to thank Dr. I. Nagai, Chukyo University, and Mr. S. Imori and Mr. Y. Hashiyama, Hiroshima

University, for their encouragements. Furthermore, the authors wish to thank the referee for helpful suggestions.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In 2nd International Symposium on Information Theory (B. N. Petrov & F. Csaki eds.), Akadémiai Kiadó, Budapest, 267–281.
[2] Akaike, H. (1987). Factor analysis and AIC. Psychometrika, **52**, 317–332.
[3] Bollen, K. A. (1989). Structural Equations with Latent Variables. John Wiley & Sons, Inc., New York.
[4] Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. Psychometrika, **50**, 229–242.
[5] Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. British J. Math. Statist. Psych., **37**, 62–83.
[6] Cudeck, R. and Brown, M. W. (1983). Cross-validation of covariance structures. Multivariate Behav. Res., **18**, 147–167.
[7] Hall, P. (1992). The Bootstrap and Edgeworth Expansion. Springer, New York.
[8] Ichikawa, M. and Konishi, S. (1999). Model evaluation and Information criteria in covariance structure analysis. British J. Math. Statist. Psych., **52**, 285–302.
[9] Jöreskog, K. G. and Sörbom, D. (1996). LISREL 8 User's Reference Guide. Scientific Software International, Chicago.
[10] Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. Biometrika, **83**, 875–890.
[11] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. Ann. Math. Statist., **22**, 79–86.
[12] Lee, S. Y. and Jennrich, R. I. (1979). A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. Psychometrika, **44**, 99–113.
[13] Magnus, J. R. and Neudecker, H. (1999). Matrix Differential Calculus with Applications in Statistics and Econometrics (revised ed.). John Wiley & Sons, Inc., New York.
[14] 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 notional character. Proc. Inst. Statist. Math., **53**, 133–157 (in Japanese).
[15] Yanagihara, H. (2007). A family of estimators for multivariate kurtosis in a nonnormal linear regression model. J. Multivariate Anal., **98**, 1–29.
[16] Yanagihara, H., Himeno, T. and Yuan, K.-H. (2010). GLS discrepancy based information criteria for selecting covariance structure models. Behaviormetrika, **37**, 71–86.
[17] Yuan, K.-H. and Bentler, P. M. (1997). Mean and covariance structure analysis: theoretical and practical improvements. J. Amer. Statist. Assoc., **92**, 767–774.
[18] Yuan, K.-H. and Bentler, P. M. (2007). Structural equation modeling. In Handbook of Statistics **27**: Psychometrics (C. R. Rao & S. Sinharay eds.), Elsevier/North-Holland, Amsterdam, 297–358.
[19] Yuan, K.-H. & Chan, W. (2008). Structural equation modeling with near singular covariance matrices. Comput. Statist. Data Anal., **52**, 4842–4858.
[20] Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In Testing Structural Equation Models (K. A. Bollen & J. S. Long eds.), Sage, Newbury park, CA, 256–293.

Ami Kamada
Biostatistics, Clinical Data Science Department
Takeda Development Center Japan
Pharmaceutical Development Division
Takeda Pharmaceutical Company Limited
1-1 *Doshomachi* 4-*chome, Chuo-ku, Osaka* 540-8645, *Japan*
E-mail: ami.kamada@gmail.com

Hirokazu Yanagihara
Department of Mathematics
Hiroshima University
1-3-1 *Kagamiyama, Higashi-Hiroshima, Hiroshima* 739-8626, *Japan*
E-mail: yanagi@math.sci.hiroshima-u.ac.jp

Hirofumi Wakaki
Department of Mathematics
Hiroshima University
1-3-1 *Kagamiyama, Higashi-Hiroshima, Hiroshima* 739-8626, *Japan*
E-mail: wakaki@math.sci.hiroshima-u.ac.jp

Keisuke Fukui
Department of Mathematics
Hiroshima University
1-3-1 *Kagamiyama, Higashi-Hiroshima, Hiroshima* 739-8626, *Japan*
E-mail: d126313@hiroshima-u.ac.jp