SELECTING A SHRINKAGE PARAMETER IN STRUCTURAL EQUATION MODELING WITH A NEAR SINGULAR COVARIANCE MATRIX BY THE GIC MINIMIZATION METHOD

Ami KAMADA

Department of Mathematics, Faculty of Science, Hiroshima University
1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8626, Japan

Abstract

In structural equation modeling (SEM), a covariance parameter is 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 use of an adjusted sample covariance matrix instead of the sample covariance matrix in the discrepancy function for estimating the covariance estimator. The adjusted sample covariance matrix was defined by adding an identity matrix multiplied by 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 attempt to select the shrinkage parameter using an information criterion minimization method. Therefore, we propose an information criterion based on the discrepancy function measured by the normal theory maximum likelihood (ML). Using the Monte Carlo method, we demonstrate that the proposed criterion works well.

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

1. INTRODUCTION

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ be independent random samples from $\boldsymbol{x}$ distributed according to a p-variate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. 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

$$M : n\boldsymbol{S} \sim W_p(n, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}), \tag{1.1}$$

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

$$M_0 : n\boldsymbol{S} \sim W_p(n, \boldsymbol{\Sigma}_0). \tag{1.2}$$

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 structural equation modeling (SEM) fits the sample covariance matrix $\boldsymbol{S}$ by $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ through minimizing the normal theory maximum likelihood (ML) discrepancy function as

$$F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) = \mathrm{tr}(\boldsymbol{S}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}) - \log|\boldsymbol{S}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}| - p. \tag{1.3}$$

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 iteration 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)). 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 $\hat{\boldsymbol{\theta}}_a$ 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), they reported that $\tilde{\boldsymbol{\theta}}_a$ is adjusted to a consistent estimator through a simple procedure when the covariance structure is the correct model. The adjustment is as follows:

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

where $\boldsymbol{j}$ is a q-dimensional vector, the elements of which are 1, corresponding to the parameters on the diagonals of the covariance matrix of the vectors of the measurement errors, and otherwise are zero.

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. Therefore, we attempt to select the shrinkage parameter based on the predictive Kullback-Leibler (KL) discrepancy (Kullback and Leibler (1951)). The basic concept is to measure the goodness of fit of the model by the risk function assessed by 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, 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

Since $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is not always correctly specified, when $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is misspecified, we denote $\boldsymbol{\theta}_{a*}$ as a population parameter minimizing $F(\boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, i.e.,

$$\boldsymbol{\theta}_{a*} = \arg\min_{\boldsymbol{\theta}} F(\boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}),$$

where $\boldsymbol{\Sigma}_a$ is the expectation of $\boldsymbol{S}_a$, i.e., $\boldsymbol{\Sigma}_a = \boldsymbol{\Sigma}_0 + a\boldsymbol{I}_p$. There also exists a unique vector $\boldsymbol{\theta}_*$ such that $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{a*}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_*} + a\boldsymbol{I}_p$ under the LISREL model. Then, $\hat{\boldsymbol{\theta}}_a$ is consistent for $\boldsymbol{\theta}_{a*}$, and $\tilde{\boldsymbol{\theta}}_a$ is consistent for $\boldsymbol{\theta}_*$ (see Yuan and Chan (2008)). If $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is correctly specified, $\boldsymbol{\theta}_{a*} = \boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_* = \boldsymbol{\theta}_0$.

We consider the risk function between the true model and the candidate model. Let $\mathcal{S}$ be a sample covariance matrix that is independent of $\boldsymbol{S}$ and that has the same distribution as $\boldsymbol{S}$. Hence, two matrices $n\boldsymbol{S}$ and $n\mathcal{S}$ are independently and identically distributed according to the Wishart distribution $W_p(n, \boldsymbol{\Sigma}_0)$. The matrix $\mathcal{S}$ is regarded as a future observation or an imaginary new observation. Then, we measure the discrepancy between the candidate model M in (1.1) and the true model M_0 in (1.2) by the following discrepancy function:

$$\int \log \frac{f(\mathcal{S}|n, \boldsymbol{\Sigma}_0)}{f(\mathcal{S}|n, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a})} d\mathcal{S} = \frac{n}{2} E_{\mathcal{S}}\left[F(\mathcal{S}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a}) - F(\mathcal{S}, \boldsymbol{\Sigma}_0)\right],$$

Omitting the terms that do not depend on a, yields

$$\int F(\mathcal{S}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a}) d\mathcal{S} = E_{\mathcal{S}}\left[F(\mathcal{S}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a})\right].$$

Hence, we define the risk function assessed by the predictive ML discrepancy as

$$R = E_{\mathcal{S}} E_{\boldsymbol{S}} \left[F(\mathcal{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}) \right].$$

In the present paper, $E_{\mathcal{S}}$ and $E_{\boldsymbol{S}}$ denote the expectations under the true model M_0 in (1.2) with respect to $\mathcal{S}$ and $\boldsymbol{S}$. We regard the shrinkage parameter a l 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

$$B = R - E_{\boldsymbol{S}}[F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})]. \tag{2.1}$$

The following development contains the technical details for obtaining and estimating B.

Since $\hat{\boldsymbol{\theta}}_a$ is a minimizer of $F(\boldsymbol{S}_a, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ under the conditions of Theorem 1 in Yuan and Chan (2008), then $\hat{\boldsymbol{\theta}}_a$ satisfies

$$\boldsymbol{\Delta}'_{\hat{\boldsymbol{\theta}}_a} \mathrm{vec}(\boldsymbol{\Sigma}^{-1}_{\hat{\boldsymbol{\theta}}_a}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a} - \boldsymbol{S}_a)\boldsymbol{\Sigma}^{-1}_{\hat{\boldsymbol{\theta}}_a}) = \mathbf{0}_q,$$

where $\mathbf{0}_q$ is a q-dimensional vector of zeros, and

$$\boldsymbol{\Delta}_{\boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}'} \mathrm{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}). \tag{2.2}$$

Let

$$\boldsymbol{G}_{\boldsymbol{\theta}_{a*}} = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} F(\boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{a*}},$$

where

$$\begin{aligned} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} F(\boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) = {} & 2\boldsymbol{\Delta}'_{\boldsymbol{\theta}}(\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\theta}} \boldsymbol{\Sigma}_a \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\theta}} \otimes \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\theta}}) \boldsymbol{\Delta}_{\boldsymbol{\theta}} - \boldsymbol{\Delta}'_{\boldsymbol{\theta}}(\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\theta}} \otimes \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\theta}}) \boldsymbol{\Delta}_{\boldsymbol{\theta}} \\ & - \sum_{i,j}^{q} \mathrm{tr}\{\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\theta}}(\boldsymbol{\Sigma}_a - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\theta}} \ddot{\boldsymbol{\Sigma}}_{\boldsymbol{\theta} ij}\} \boldsymbol{e}_i \boldsymbol{e}_j{}'. \end{aligned} \tag{2.3}$$

Here, $\boldsymbol{e}_i$ is a q-dimensional vector, the ith element of which is 1, with all others being 0, and $\ddot{\boldsymbol{\Sigma}}_{\boldsymbol{\theta} ij} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. Since $\boldsymbol{\theta}_{a*}$ is the minimizer of $F(\boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, $\boldsymbol{G}_{\boldsymbol{\theta}_{a*}}$ 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* $E_{\boldsymbol{S}}[F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})]$ *is expanded as*

$$B = \frac{2}{n} \mathrm{tr} \left\{ \boldsymbol{\Delta}_{\boldsymbol{\theta}_*} \boldsymbol{G}^{-1}_{\boldsymbol{\theta}_{a*}} \boldsymbol{\Delta}'_{\boldsymbol{\theta}_{a*}} (\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\theta}_{a*}} \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\theta}_*} \otimes \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\theta}_{a*}} \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\theta}_*}) \right\} + O(n^{-2}). \tag{2.4}$$

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

By replacing $\boldsymbol{\theta}_{a*}$, $\boldsymbol{\theta}_{*}$, and $\boldsymbol{\Sigma}_0$ by neglecting $O(n^{-2})$ in (2.4) with $\hat{\boldsymbol{\theta}}_a$, $\tilde{\boldsymbol{\theta}}_a$, and $\boldsymbol{S}$, respectively, an estimator of B is given by

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

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 $A = \{a \mid a \geqq 0$ and $\tilde{\boldsymbol{\theta}}_a$ 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 become simple, as in the following corollary.

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

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

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 smallest value in A when the model is correctly specified because $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})$ is an increasing function with respect to 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 $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}$.

Namely, the risk functions considered herein are $R_{\mathrm{new}} = E_{\mathcal{S}}E_{\boldsymbol{S}}[F(\mathcal{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_{\hat{a}}})]$, $R_{\mathrm{YC}} = E_{\mathcal{S}}E_{\boldsymbol{S}}[F(\mathcal{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_{p/N}})]$, $R_{\mathrm{ML}} = E_{\mathcal{S}}E_{\boldsymbol{S}}[F(\mathcal{S}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}})]$. The true model M_0 used in the simulation is the confirmatory two-factor model, which is included in the LISREL model, i.e., the true covariance matrix is $\boldsymbol{\Sigma}_0 = \boldsymbol{\Lambda}_0\boldsymbol{\Phi}_0\boldsymbol{\Lambda}'_0 + \boldsymbol{\Theta}_0$, where $\boldsymbol{\Lambda}_0$ is the factor loading matrix, $\boldsymbol{\Phi}_0$ is a 2×2 correlation matrix, and $\boldsymbol{\Theta}_0$ is the covariance with the vectors of the measurement errors. Each of the two factors has five unidimensional indicators. The factor loading and factor correlation matrices in the population are given by

$$\boldsymbol{\Lambda}_0 = \begin{pmatrix} \boldsymbol{b} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{b} \\ \boldsymbol{0} & \boldsymbol{b} \end{pmatrix}, \quad \boldsymbol{\Phi}_0 = \begin{pmatrix} 1.0 & .30 \\ .30 & 1.0 \end{pmatrix},$$

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

	Frequency			Risk		
N	New	YC	ML	New	YC	ML
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

where $\boldsymbol{b} = (.70, .70, .75, .80, .80)'$ and $\mathbf{0} = (0, 0, 0, 0, 0)'$.

The candidate model used in the simulation is the confirmatory three-factor model, which also has five unidimensional indicators, i.e., the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$. The factor loading and factor correlation matrices in the candidate model are given by

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{b}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{b}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{b}_3 \end{pmatrix}, \ \boldsymbol{\Phi} = \begin{pmatrix} 1.0 & \phi_{12} & \phi_{13} \\ \phi_{21} & 1.0 & \phi_{23} \\ \phi_{31} & \phi_{32} & 1.0 \end{pmatrix}.$$

There are 15 factor loadings, three factor correlations, and 15 error variances. In order to obtain smaller sample sizes, we chose $N = 30$, 50, and 100. The number of replications is $1,000$.

In order to calculate R_{new}, R_{YC}, and R_{ML}, we first obtain an estimator of $\boldsymbol{\theta}$ for each method using R Ver. 2.12.1. We then count the frequencies when the estimate of $\boldsymbol{\theta}$ is the proper solution (i.e., an estimator of $\boldsymbol{\Theta}_0$ is the positive define). Then, in order to generate a normal random sample of the sample size N again, we obtain the sample covariance matrix $\mathcal{S}$. Next, we record the value of $F(\mathcal{S}, \hat{\boldsymbol{\Sigma}})$ for each method, where $\hat{\boldsymbol{\Sigma}}$ is an estimated $\boldsymbol{\Sigma}$ for each method. After the replication is finished, we obtain the arithmetic mean of $F(\mathcal{S}, \hat{\boldsymbol{\Sigma}})$ for each method. If all of the estimators are proper solutions, then the arithmetic mean is regarded as a target risk function.

Table 1 shows the frequencies of the proper solutions and the values of the risk functions for each method. From Table 1, when $N = 30$, the R_{new} is obtained, but R_{YC} and R_{ML} are not obtained because there are several improper solutions for $a = p/N$, 0. On the other hand, when $N = 50$ and 100, since there are no improper solutions, we can obtain all risk functions. Then, R_{new} is the smallest. Hence, the proposed information criterion works well.

4. CONCLUSION

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 smallest value among shrinkage parameters making $\tilde{\boldsymbol{\theta}}_a$ a proper solution. In the Monte Carlo results of $\tilde{\boldsymbol{\theta}}_{\hat{a}}$ was proper solutions, and the risk function of the estimated covariance matrix based on $\tilde{\boldsymbol{\theta}}_{\hat{a}}$ with the selected a was the smallest.

APPENDIX

The proof of Theorem 1 is presented in this appendix. The bias of $F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})$, defined in (2.1), can be written as

$$B = E_{\boldsymbol{S}}[E_{\mathcal{S}}[F(\mathcal{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})] - F(\boldsymbol{S}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a})] = E_{\boldsymbol{S}}[\mathrm{tr}\{\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_a}^{-1}(\boldsymbol{\Sigma}_0 - \boldsymbol{S})\}]. \tag{A.1}$$

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

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

where

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

The remainder term is $O(n^{-2})$ because a general moment can be expanded by n^{-1} (see Hall (1992)).

Letting $\boldsymbol{V} = \sqrt{n}(\boldsymbol{S} - \boldsymbol{\Sigma}_0)$, we obtain

$$\begin{aligned}\boldsymbol{d}_{\boldsymbol{\theta}_*} &= \mathrm{vec}'\{\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}^{-1}(\boldsymbol{S} - \boldsymbol{\Sigma}_0)\boldsymbol{\Sigma}_{\boldsymbol{\theta}_*}^{-1}\}\boldsymbol{\Delta}_{\boldsymbol{\theta}_*} \\ &= \frac{1}{\sqrt{n}}\mathrm{vec}'(\boldsymbol{V})\boldsymbol{\Gamma}_{\boldsymbol{\theta}_*}\boldsymbol{\Delta}_{\boldsymbol{\theta}_*}.\end{aligned} \tag{A.2}$$

Then, under a set of standard regularity conditions, it follows from $\partial F(\boldsymbol{S}_a, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_a} = \boldsymbol{0}_q$ that

$$\boldsymbol{0}_q = \boldsymbol{\Delta}'_{\hat{\boldsymbol{\theta}}_a}\mathrm{vec}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a}^{-1}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a} - \boldsymbol{\Sigma}_a)\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a}^{-1}) - \frac{1}{\sqrt{n}}\boldsymbol{\Delta}'_{\hat{\boldsymbol{\theta}}_a}\boldsymbol{\Gamma}_{\hat{\boldsymbol{\theta}}_a}\mathrm{vec}(\boldsymbol{V}),$$

where $\boldsymbol{\Gamma}_{\hat{\boldsymbol{\theta}}_a} = (\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a}^{-1} \otimes \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a}^{-1})$. Hence, we obtain

$$\boldsymbol{\Delta}'_{\hat{\boldsymbol{\theta}}_a}\mathrm{vec}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a}^{-1}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a} - \boldsymbol{\Sigma}_a)\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_a}^{-1}) = \frac{1}{\sqrt{n}}\boldsymbol{\Delta}'_{\hat{\boldsymbol{\theta}}_a}\boldsymbol{\Gamma}_{\hat{\boldsymbol{\theta}}_a}\mathrm{vec}(\boldsymbol{V}). \tag{A.3}$$

Note that $\sqrt{n}(\hat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_{a*}) = O_p(1)$ and that both sides of (A.3) are functions of $\hat{\boldsymbol{\theta}}_a$. Applying the Taylor expansion to (A.3) 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

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

where $\boldsymbol{G}_{\boldsymbol{\theta}}$ is given by (2.3).

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

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

$$\begin{aligned} B &= E_{\boldsymbol{S}}[\boldsymbol{d}_{\boldsymbol{\theta}_*}(\hat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_{a*})] + O(n^{-2}) \\ &= \frac{1}{n}\mathrm{tr}\left\{\boldsymbol{\Gamma}_{\boldsymbol{\theta}_*}\boldsymbol{\Delta}_{\boldsymbol{\theta}_*}\boldsymbol{G}_{\boldsymbol{\theta}_{a*}}^{-1}\boldsymbol{\Delta}'_{\boldsymbol{\theta}_{a*}}\boldsymbol{\Gamma}_{\boldsymbol{\theta}_{a*}}\left(\boldsymbol{I}_{p^2} + \boldsymbol{K}_{pp}\right)\left(\boldsymbol{\Sigma}_0 \otimes \boldsymbol{\Sigma}_0\right)\right\} + O(n^{-2}). \end{aligned} \tag{A.4}$$

Moreover, using (see Magnus and Neudecker 1999, p. 47) $\boldsymbol{K}_p(\boldsymbol{A} \otimes \boldsymbol{C}) = (\boldsymbol{C} \otimes \boldsymbol{A})\boldsymbol{K}_{pp}$, $\boldsymbol{K}_p\mathrm{vec}(\boldsymbol{C}) = \mathrm{vec}(\boldsymbol{C}')$, yields Equation (2.4) in Theorem 1.

ACKNOWLEDGEMENTS

The author is grateful to Prof. Hirofumi Wakaki and Dr. Hirokazu Yanagihara for their valuable comments. The author would also like to thank colleagues, especially Mr. Nagai, Mr. Imori, and Mr. Hashiyama, for providing encouragement and support.

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.), Akademiai Kiado, Budapest, 267–281.

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

[3] Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*, **50**, 229–242.

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

[5] Cudeck, R. & Brown, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, **18**, 147–167.

[6] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

[7] Ichikawa, M. & Konishi, S. (1999). Model evaluation and Information criteria in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology* **52**, 285–302.

[8] Jöreskog, K. G. & Sörbom, D. (1996). *LISREL 8 User's Reference Guide*. Scientific Software International, Chicago.

[9] Konishi, S. & Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika* **83**, 875–890.

[10] Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79–86.

[11] Lee, S. Y. & Jennrich, R. I. (1979). A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. *Psychometrika*, **44**, 99–113.

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

[13] Yanagihara, H. (2005). Selection of covariance structure models in nonnormal data by using information criterion : an application to data from the survey of the Japanege notional character. *Proceedings of the Institute of Statistical Mathematics*, **53**, 133–157. (in Japanese).

[14] Yanagihara, H., Himeno, T. & Yuan, K.-H. (2010), GLS discrepancy based information criteria for selecting covariance structure models. *Behaviormetrika*, **37**, 71–86.

[15] Yuan, K.-H. & Bentler, P. M. (1997). Mean and covariance structure analysis: theoretical and practical improvements. *Journal of American Statistical Association*, **92**, 767–774.

[16] Yuan, K.-H. & Chan, W. (2008). Structural equation modeling with near singular covariance matrices. *Computational Statistics and Data Analysis*, **52**, 4842–4858.

[17] Wothke, W. (1993). Nonpositive definite matrices in structural modeling. *Testing Structural Equation Models* (K. A. Bollen & J. S. Long eds.), Sage, Newbury park, CA, 256–293.