Journal of Data Science **9**(2011), 243-259

Bayesian Information Criterion and Selection of the Number of Factors in Factor Analysis Models

Kei Hirose[1], Shuichi Kawano[2],
Sadanori Konishi[3] and Masanori Ichikawa[4]
[1]*Kyushu University,* [2]*University of Tokyo,*
[3]*Chuo University and* [4]*Tokyo University of Foreign Studies*

Abstract: In maximum likelihood exploratory factor analysis, the estimates of unique variances can often turn out to be zero or negative, which makes no sense from a statistical point of view. In order to overcome this difficulty, we employ a Bayesian approach by specifying a prior distribution for the variances of unique factors. The factor analysis model is estimated by EM algorithm, for which we provide the expectation and maximization steps within a general framework of EM algorithms. Crucial issues in Bayesian factor analysis model are the choice of adjusted parameters including the number of factors and also the hyper-parameters for the prior distribution. The choice of these parameters can be viewed as a model selection and evaluation problem. We derive a model selection criterion for evaluating a Bayesian factor analysis model. Monte Carlo simulations are conducted to investigate the effectiveness of the proposed procedure. A real data example is also given to illustrate our procedure. We observe that our modeling procedure prevents the occurrence of improper solutions and also chooses the appropriate number of factors objectively.

Key words: EM algorithm, factor analysis, model selection criterion, number of factors, prior distribution.

1. Introduction

Factor analysis provides a useful tool to draw information from data by exploring the covariance structure among observed variables in terms of a smaller number of unobserved variables. Successful applications have been reported in various fields of research including the social and behavioral sciences.

The factor analysis model is usually estimated by maximum likelihood methods under the assumption that the observations are normally distributed. In practice, however, the maximum likelihood estimates of unique variances can often turn out to be zero or negative. Such estimates are known as improper solutions, and many authors have studied these inappropriate estimates both from a theoretical point of view and also by means of numerical examples (see, e.g., Jöreskog, 1967; van Driel, 1978; Sato, 1987; Kano and Ihara, 1994; Kano, 1998; Krijnen *et al.*, 1998). Various causes of improper solutions in structural equation models, including confirmatory factor analysis model, have been also explored (see, e.g., Anderson and Gerbing, 1984; Boomsma, 1985; Gerbing and Anderson, 1987; Chen *et al.*, 2001; Flora and Curran, 2004).

In order to prevent the occurrence of improper solutions in factor analysis model, we take a Bayesian approach by specifying a prior distribution for the variances of unique factors. In Bayesian factor analysis, the choice of a prior distribution is a fundamental issue. Martin and McDonald (1975) used a prior distribution for the elements of unique variances. Press (1982) used a natural conjugate prior distribution for factor loadings and unique variances. Akaike (1987) introduced a prior distribution using the information extracted from the knowledge of the likelihood function. Recently, a Bayesian approach based on the Markov chain Monte Carlo (MCMC) algorithms has received an amount of attention in the Bayesian factor analysis (see, e.g., Lee and Song, 2002; Lopes and West, 2004). Basically, a conjugate prior distribution is used for MCMC-based algorithm and estimates can be obtained via an algorithm based on the Gibbs sampler.

Another important point in Bayesian factor analysis model is the choice of adjusted parameters including the number of factors and hyper-parameters in the prior distribution. Regarding selection of the number of factors, the AIC (Akaike, 1973) and BIC (Schwarz, 1978) have been widely used, and some other selection procedures have also been developed by several researchers (see, e.g., Bozdogan, 1987; Press and Shigemasu, 1999). However, these procedures cannot provide suitable values of hyper-parameters included in the prior distribution. A selection procedure via the MCMC algorithms has also been widely used in the Bayesian factor analysis (see, e.g., Lee and Song, 2002; Lopes and West, 2004; Dunson, 2006; Fokouè, 2009). Although the MCMC-based selection method is certainly attractive, we take a different approach that selects both the number of factors and the values of hyper-parameters in the prior distribution since the MCMC-based procedure sometimes requires much computational load.

In this paper, we introduce a proper prior distribution for the variances of unique factors by extending a basic idea given in Akaike (1987). The Bayesian factor analysis model is estimated by EM algorithm, for which we provide the expectation and maximization steps within a general framework of EM algorithms. We treat a selection of parameters, which include the number of factors and the hyper-parameters for the prior distribution, as a model selection and evaluation problem, and derive a model selection criterion from a Bayesian viewpoint for evaluating a Bayesian factor analysis model. The proposed modeling procedure enables us to choose the number of factors and the values of hyper-parameters in the prior distribution simultaneously.

The remainder of this paper is organized as follows: Section 2 describes the maximum likelihood factor analysis and its related problems. In Section 3, we introduce a prior distribution according to the basic idea given by Akaike (1987), and provide a model estimation procedure using EM algorithm. Section 4 derives a model selection criterion for evaluating a Bayesian factor analysis model. Section 5 presents numerical results for both artificial and real datasets. Some concluding remarks are given in Section 6.

2. Maximum Likelihood Factor Analysis and Its Related Problems

Let $\boldsymbol{X}= (X_1, \cdots, X_p)'$ be a p-dimensional observable random vector with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. The factor analysis model is

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{F} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\Lambda} = (\lambda_{ij})$ is a $p \times k$ matrix of factor loadings, and $\boldsymbol{F} = (F_1, \cdots, F_k)'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_p)'$ are unobservable random vectors. The elements of $\boldsymbol{F}$ and $\boldsymbol{\varepsilon}$ are called common factors and unique factors, respectively. It is assumed that $E(\boldsymbol{F}) = \mathbf{0}$, $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $E(\boldsymbol{F}\boldsymbol{F}') = \mathbf{I}_k$, $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \boldsymbol{\Psi}$ and $E(\boldsymbol{F}\boldsymbol{\varepsilon}') = \mathbf{0}$, where $\mathbf{I}_k$ is the identity matrix of order k and $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix with i-th diagonal element ψ_i which is called unique variance. Under these assumptions, the variance-covariance matrix of $\boldsymbol{X}$ can be expressed as

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}.$$

The i-th diagonal element of $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ is called communality, which measures the percent of variance in x_i explained by all the factors. It is well-known that factor loadings have a rotational indeterminacy since both $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}\mathbf{T}$ generate the same covariance matrix $\boldsymbol{\Sigma}$, where $\mathbf{T}$ is an arbitrary orthogonal matrix.

Assume that the common factors $\boldsymbol{F}$ and the unique factors $\boldsymbol{\varepsilon}$ are, respectively, distributed according to multivariate normal distributions

$$\boldsymbol{F} \sim N_k(\mathbf{0}, \mathbf{I}_k) \quad \text{and} \quad \boldsymbol{\varepsilon} \sim N_p(\mathbf{0}, \boldsymbol{\Psi}).$$

Suppose that we have a random sample of N observations $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ from the p-dimensional normal population $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$. Then the log-likelihood function is given by

$$\log f(\boldsymbol{X}_N|\boldsymbol{\Lambda}, \boldsymbol{\Psi}) = -\frac{N}{2}\left\{p\log(2\pi) + \log|\boldsymbol{\Sigma}| + \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S})\right\}, \tag{1}$$

where $\mathbf{X}_N = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)'$, $f(\mathbf{X}_N|\mathbf{\Lambda}, \mathbf{\Psi})$ is the likelihood function and $\mathbf{S} = (s_{ij})$ is the sample variance-covariance matrix

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})',$$

with $\bar{\boldsymbol{x}}$ being the sample mean vector. For convenience, let us consider the discrepancy function given by

$$q(\mathbf{\Lambda}, \mathbf{\Psi}) = \log|\mathbf{\Sigma}| + \mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{S}) - \log|\mathbf{S}| - p. \tag{2}$$

The maximum likelihood estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are given as the solutions of $\partial q(\mathbf{\Lambda}, \mathbf{\Psi})/\partial\mathbf{\Lambda} = \mathbf{0}$ and $\partial q(\mathbf{\Lambda}, \mathbf{\Psi})/\partial\mathbf{\Psi} = \mathbf{0}$. Since the solutions cannot be expressed in a closed form, some iterative procedures are required to obtain the maximum likelihood estimates $\hat{\mathbf{\Lambda}}_{\mathrm{ML}}$ and $\hat{\mathbf{\Psi}}_{\mathrm{ML}}$. In maximum likelihood factor analysis, numerical algorithms have been proposed by several authors (see, e.g., Jöreskog, 1967; Jennrich and Robinson, 1969; Clarke, 1970).

In practice, the maximum likelihood estimates of unique variances can often turn out to be zero or negative, which have been called improper solutions. van Driel (1978) categorized the causes of improper solutions into the following three types:

(i) sampling fluctuation,
(ii) there exist no appropriate factor analysis models for extraction of beneficial information from the data,
(iii) indefiniteness of the model.

van Driel (1978) distinguished among the causes of improper solutions by using the standard errors of diagonal elements of $\hat{\mathbf{\Psi}}_{\mathrm{ML}}$, whereas the theoretical method for distinguishing the causes of improper solutions remains to be established.

What should be noted is how to prevent the occurrence of improper solutions. In order to handle this problem, several attempts have been made for parameter estimation. For example, the parameters are estimated (a) under the condition that $\psi_i \geq 0.005$ for $i = 1, \cdots, p$ (see Jöreskog, 1967), (b) after eliminating variables for which the estimates are improper and (c) by utilizing a Bayesian procedure. Some problems still remain, however, in approaches (a) and (b). In approach (a), the variances of unique factors are provided subjectively, whereas those should be estimated. The approach (b) often yields inappropriate estimates even when variables that cause the improper solutions are eliminated. We, therefore, focus our attention on the approach (c) that estimates the parameters included in the factor analysis model with the help of Bayesian procedure.

3. Bayesian Factor Analysis Model

Akaike (1987) introduced a prior distribution using the information extracted from the knowledge of the likelihood function. In this section we first give a brief review of his prior distribution and its related problems, and introduce a proper prior distribution according to the basic idea given by Akaike (1987). We then develop an estimation procedure for Bayesian factor analysis model along with the technique of EM algorithm.

3.1 Prior Distributions

First, we give a brief review of a prior distribution proposed by Akaike (1987). Akaike (1987) showed that for a given $\mathbf{\Psi}$, the minimum value of the discrepancy function $q(\mathbf{\Lambda}, \mathbf{\Psi})$ in (2) with respect to $\mathbf{\Lambda}$ is given by

$$q_k(\mathbf{\Psi}) = \sum_{i=k+1}^{p} (\theta_i - \log \theta_i) + (p - k), \tag{3}$$

where $\theta_1 > \cdots > \theta_p$ are the eigenvalues of $\mathbf{\Psi}^{-1/2}\mathbf{S}\mathbf{\Psi}^{-1/2}$. Note that the number of factors k in (3) will be selected by using a model selection criterion given in Section 4.

It can be seen from Equation (3) that $q_k(\mathbf{\Psi})$ is minimized when the values of $\theta_{k+1}, \cdots, \theta_p$ are chosen as close to one as possible since a function $x - \log x$ $(x > 0)$ has a minimum at $x = 1$, and that the values of $\theta_1, \cdots, \theta_k$ do not directly affect the function (3). Therefore, there is a possibility that the larger eigenvalues of $\hat{\mathbf{\Psi}}_{\mathrm{ML}}^{-1/2}\mathbf{S}\hat{\mathbf{\Psi}}_{\mathrm{ML}}^{-1/2}$ turn out to be extremely large. This implies that some diagonal elements of $\hat{\mathbf{\Psi}}_{\mathrm{ML}}$ can become zero. In order to prevent the occurrence of improper solutions, the parameter estimation should be done under the restriction that the values of $\theta_1, \cdots, \theta_k$ are not too large.

Akaike (1987) thus added a penalty term $\rho \sum_{i=1}^{k} \theta_i$ with $\rho > 0$ to the discrepancy function $q(\mathbf{\Lambda}, \mathbf{\Psi})$ given in (2) and minimized the following function with respect to $\mathbf{\Lambda}$ and $\mathbf{\Psi}$:

$$q^*(\mathbf{\Lambda}, \mathbf{\Psi}) = q(\mathbf{\Lambda}, \mathbf{\Psi}) + \rho \sum_{i=1}^{k} \theta_i.$$

The additional term prevents the occurrence of improper solutions because it does not allow the values of $\theta_1, \cdots, \theta_k$ to be infinite. Under the constraint that $\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}$ is a diagonal matrix which removes the rotational indeterminacy, $\sum_{i=1}^{k} \theta_i$ is equal to $\mathrm{tr}(\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda} + \mathbf{I}_k)$ (see, e.g., Lawley and Maxwell, 1971), which leads to a prior distribution proposed by Akaike (1987) in the following:

$$K \exp\left\{-\frac{N\rho}{2}\,\mathrm{tr}(\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1/2})\right\}, \tag{4}$$

where K denotes the normalizing constant and ρ can be considered as a hyper-parameter. Akaike (1987) considered this distribution as a standard spherical prior distribution of the factor loadings and did not adopt the prior for $\boldsymbol{\Psi}$.

This prior distribution has an advantage that it prevents the occurrence of improper solutions if a value of the hyper-parameter ρ is suitably chosen. We have, however, no prior convictions about factor loadings in exploratory factor analysis, because $\boldsymbol{\Lambda}$ has a rotational indeterminacy. Hence it is natural to define a prior distribution for the diagonal elements of $\boldsymbol{\Psi}$ rather than $\boldsymbol{\Lambda}$.

We, therefore, propose adding a penalty term $\rho\sum_{i=1}^{p}\theta_i$ to $q(\boldsymbol{\Lambda},\boldsymbol{\Psi})$ in (2) and then minimize the function given by

$$\begin{aligned} q^{**}(\boldsymbol{\Lambda},\boldsymbol{\Psi}) &= q(\boldsymbol{\Lambda},\boldsymbol{\Psi}) + \rho\sum_{i=1}^{p}\theta_i \\ &= q(\boldsymbol{\Lambda},\boldsymbol{\Psi}) + \rho\,\mathrm{tr}(\boldsymbol{\Psi}^{-1/2}\mathbf{S}\boldsymbol{\Psi}^{-1/2}). \end{aligned} \tag{5}$$

It is reasonable to add the term $\rho\sum_{i=1}^{p}\theta_i$ instead of $\rho\sum_{i=1}^{k}\theta_i$ to $q(\boldsymbol{\Lambda},\boldsymbol{\Psi})$ in (2) since the values of $\theta_{k+1},\cdots,\theta_p$ are close to one and could be ignored relative to the very large values of $\theta_1,\cdots,\theta_k$. From Equation (5), the prior distribution for $\boldsymbol{\Psi}$ is thus given by

$$\pi(\boldsymbol{\Psi}|\rho) = K\prod_{i=1}^{p}\exp\left\{-\frac{N\rho s_{ii}}{2}\psi_i^{-1}\right\}. \tag{6}$$

The inverses of the diagonal elements of $\boldsymbol{\Psi}$ have exponential distributions that yield the normalizing constant $K = \prod_{i=1}^{p} N\rho s_{ii}/2$. In contrast, the normalizing constant for the prior distribution in (4) is infinite. Note that it is difficult to derive a model selection criterion, which will be described in the Section 4, with the prior distribution given in (4) since the model selection criterion depends on K.

Our proposed prior distribution is closely related to that of Martin and McDonald (1975) given by

$$K\prod_{i=1}^{p}\exp\left\{-\frac{N\alpha_i}{2}\psi_i^{-1}\right\}, \tag{7}$$

where $\alpha_1,\cdots,\alpha_p$ are hyper-parameters of the prior distribution. However, it is difficult to specify these hyper-parameters when p is large. They recommended restricting these hyper-parameters by requiring that $\alpha_i = s_{ii}\alpha$, in which case their prior distribution coincides with the proposed prior distribution in (6). From these descriptions, our proposed prior distribution seems to be a minor modification of Martin and McDonald (1975). It is noted, however, that the proposed distribution in (6) has the theoretical justification for preventing the occurrence of improper solutions because the prior distribution is introduced according to the theoretical scheme given by Akaike (1987), whereas the prior distribution in (7) is heuristically provided. In addition, Martin and McDonald (1975) subjectively selected a hyper-parameter α which controls the trade-off between the log-likelihood and the penalty term, while a model selection criterion presented in Section 4 enables us to choose the hyper-parameter objectively.

3.2 Estimation

For the prior distribution $\pi(\boldsymbol{\Psi}|\rho)$ defined by (6), the posterior distribution is given by

$$\begin{aligned}\pi(\boldsymbol{\Lambda}, \boldsymbol{\Psi}|\mathbf{X}_N) &= \frac{f(\mathbf{X}_N|\boldsymbol{\Lambda}, \boldsymbol{\Psi})\pi(\boldsymbol{\Psi}|\rho)}{\displaystyle\int\int f(\mathbf{X}_N|\boldsymbol{\Lambda}, \boldsymbol{\Psi})\pi(\boldsymbol{\Psi}|\rho)d\boldsymbol{\Lambda}d\boldsymbol{\Psi}} \\ &\propto f(\mathbf{X}_N|\boldsymbol{\Lambda}, \boldsymbol{\Psi})\pi(\boldsymbol{\Psi}|\rho).\end{aligned}$$

We estimate the parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ by a posterior mode. Taking the logarithm of the equation gives the penalized log-likelihood function

$$l_\rho(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) = \log f(\mathbf{X}_N|\boldsymbol{\Lambda}, \boldsymbol{\Psi}) - \frac{N\rho}{2}\mathrm{tr}(\boldsymbol{\Psi}^{-1/2}\mathbf{S}\boldsymbol{\Psi}^{-1/2}), \tag{8}$$

where the hyper-parameter ρ can be considered as a regularization parameter. We estimate the parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ in the Bayesian factor analysis model by maximizing the penalized log-likelihood function given in (8).

One of the beneficial methods to obtain the maximum penalized likelihood estimates is an EM algorithm. Rubin and Thayer (1982) suggested using an EM algorithm in maximum likelihood factor analysis. The advantage of the EM algorithms is that even if the likelihood function is not concave with respect to the parameters, the algorithm leads to a (local) maximization of the function. Bentler and Tanaka (1983) pointed out the problems in the EM algorithm for factor analysis, whereas Rubin and Thayer (1983) addressed the problem of Bentler and Tanaka's (1983) discussion.

We employ an EM algorithm to obtain the maximum penalized likelihood estimates. We provide the expectation and maximization steps for the Bayesian factor analysis model within a general framework of EM algorithms. We regard the common factors as missing variables, and maximize the complete-data log-likelihood using a posterior distribution for the missing variables. The iterative procedure is given by

$$\hat{\mathbf{\Lambda}} = (\mathbf{S}\mathbf{\Sigma}^{-1}\mathbf{\Lambda})(\mathbf{B} + \mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{S}\mathbf{\Sigma}^{-1}\mathbf{\Lambda})^{-1}, \tag{9}$$

$$\hat{\mathbf{\Psi}} = \mathrm{Diag}\left[\mathbf{S} - 2\mathbf{S}\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Lambda}}\mathbf{B}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Lambda}}\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{S}\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\hat{\mathbf{\Lambda}}' + \rho\mathbf{S}\right], \tag{10}$$

where $\mathbf{B} = \mathbf{I}_k - \mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{\Lambda}$. For detailed procedure for estimation of factor analysis models via EM algorithms, we refer to Rubin and Thayer (1982) and Tipping and Bishop (1999).

In order to eliminate the rotational indeterminacy from $\mathbf{\Lambda}$, we impose restrictions that $\lambda_{ij} = 0$ $(i > j)$ (see, for example, Anderson and Rubin; 1956).

4. Model Selection Criterion

In the Bayesian factor analysis model, we still have crucial issues to be solved: the choice of a hyper-parameter ρ for the prior distribution and the number of factors k. In this section we derive a model selection criterion for evaluating a Bayesian factor analysis model.

The generalized Bayesian information criterion (GBIC), proposed by Konishi *et al.* (2004), enables us to choose adjusted parameters including the hyper-parameter ρ and the number of factors k simultaneously by extending the Bayesian information criterion (BIC) proposed by Schwarz (1978). The basic idea of BIC is to select a model from a set of candidate models by maximizing the posterior probability. The BIC only deals with models estimated by the maximum likelihood method, whereas the model selection criterion GBIC can be applied to models estimated by the maximum penalized likelihood method. For model selection criteria we refer to Konishi and Kitagawa (2008) and references given therein.

Suppose that $\boldsymbol{\theta}$ is a parameter vector given by

$$\boldsymbol{\theta} = (\boldsymbol{\lambda}'_{.1}, \boldsymbol{\lambda}'_{.2}, \cdots, \boldsymbol{\lambda}'_{.k}, \mathrm{Diag}(\mathbf{\Psi})')',$$

where $\boldsymbol{\lambda}_{.i} = (\lambda_{i,i}, \lambda_{i+1,i}, \cdots, \lambda_{p,i})'$. We used the definition of $\boldsymbol{\lambda}_{.i}$ which consists of only the lower elements of $\mathbf{\Lambda}$ because it eliminates the rotational indeterminacy as described in the previous section. Let $f(\mathbf{X}_N|\hat{\boldsymbol{\theta}})$ be the estimated model by maximum penalized likelihood methods. Then we have a statistical model

$$f(\mathbf{X}_N|\hat{\boldsymbol{\theta}}) = (2\pi)^{-\frac{Np}{2}}|\hat{\mathbf{\Sigma}}|^{-\frac{N}{2}} \exp\left\{-\frac{N}{2}\mathrm{tr}\left(\hat{\mathbf{\Sigma}}^{-1}\mathbf{S}\right)\right\},$$

where $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}$. The model selection criterion GBIC for Bayesian factor analysis is given by

$$\begin{aligned}\mathrm{GBIC} = & -p^* \log(2\pi) + p^* \log N + \log |J_\rho(\hat{\boldsymbol{\theta}})| + N \left\{ p \log(2\pi) + \log |\hat{\mathbf{\Sigma}}| + \mathrm{tr}(\hat{\mathbf{\Sigma}}^{-1}\mathbf{S}) \right\} \\ & -2 \sum_{i=1}^{p} \log \left(\frac{N\rho s_{ii}}{2} \right) + N\rho \sum_{i=1}^{p} (s_{ii} \hat{\psi}_i^{-1}), \qquad (11)\end{aligned}$$

where p^* is the number of parameters given by $p(k+1) - k(k-1)/2$ and $J_\rho(\hat{\boldsymbol{\theta}})$ is a second order differential of the penalized log-likelihood function given by

$$J_\rho(\hat{\boldsymbol{\theta}}) = -\frac{1}{N} \left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \left\{ \log f(\mathbf{X}_N | \boldsymbol{\theta}) + \log \pi(\mathbf{\Psi}|\rho) \right\} \bigg|_{\hat{\boldsymbol{\theta}}} \right].$$

We choose optimum values of the hyper-parameter ρ and the number of factors k which simultaneously minimize the value of the model selection criterion in (11). The derivation of the GBIC is given by Hirose *et al.* (2008).

Other traditional model selection criteria include AIC (Akaike, 1973) and BIC (Schwarz, 1978). It should be noted that the AIC and BIC often select a model which causes improper solutions because these model selection criteria only evaluate models estimated by the maximum likelihood method. These model selection criteria are given by

$$\begin{aligned}\mathrm{AIC} &= -2 \log f(\mathbf{X}_N | \hat{\mathbf{\Lambda}}_{\mathrm{ML}}, \hat{\mathbf{\Psi}}_{\mathrm{ML}}) + 2p^*, \\ \mathrm{BIC} &= -2 \log f(\mathbf{X}_N | \hat{\mathbf{\Lambda}}_{\mathrm{ML}}, \hat{\mathbf{\Psi}}_{\mathrm{ML}}) + p^* \log N.\end{aligned}$$

5. Numerical Examples

Monte Carlo simulations and a real data example are used to examine the efficiency of the proposed procedure. We investigate how well the Bayesian modeling strategy with GBIC performs well in the sense that it prevents the occurrence of improper solutions and can select the true number of factors.

5.1 Numerical Comparison

Monte Carlo simulations were conducted to investigate the performance of our proposed procedure in various covariance structures and samples. In this simulation study we focus on the choice of the number of factors and compare the performance of GBIC with that of AIC and BIC.

We consider various datasets which are likely to produce improper solutions due to sampling fluctuations. van Driel (1978) showed that improper solutions sometimes occur when one of the diagonal elements of Ψ is close to zero. In addition, improper solutions often arise when the number of samples is small or communalities are large. Taking these natures of improper solutions into account, we considered four models, which are given in Table 1, and three variants for the number of observations, $N = 30$, $N = 50$ and $N = 100$.

Table 1: Four models for simulated datasets

	parameters (a1)			parameters (a2)			parameters (b1)				parameters (b2)			
i	Λ		Ψ	Λ		Ψ	Λ			Ψ	Λ			Ψ
1	0.95	0.0	0.10	0.9	0.0	0.19	0.95	0.0	0.0	0.10	0.9	0.0	0.0	0.19
2	0.00	0.7	0.51	0.0	0.9	0.19	0.00	0.7	0.0	0.51	0.0	0.9	0.0	0.19
3	0.70	0.0	0.51	0.8	0.0	0.36	0.00	0.0	0.7	0.51	0.0	0.0	0.8	0.36
4	0.70	0.0	0.51	0.6	0.0	0.64	0.70	0.0	0.0	0.51	0.8	0.0	0.0	0.36
5	0.00	0.7	0.51	0.0	0.8	0.36	0.70	0.0	0.0	0.51	0.7	0.0	0.0	0.51
6	0.00	0.7	0.51	0.0	0.8	0.36	0.00	0.7	0.0	0.51	0.0	0.8	0.0	0.36
7	0.00	0.7	0.51	0.0	0.7	0.51	0.00	0.7	0.0	0.51	0.0	0.7	0.0	0.51
8	0.00	0.7	0.51	0.0	0.7	0.51	0.00	0.0	0.7	0.51	0.0	0.0	0.8	0.36
9	0.00	0.7	0.51	0.0	0.6	0.64	0.00	0.0	0.7	0.51	0.0	0.0	0.7	0.51
10	0.00	0.7	0.51	0.0	0.6	0.64	0.00	0.0	0.7	0.51	0.0	0.0	0.6	0.64

The models (a1) and (b1) are constructed based on the simulations of *close to zero* data in van Driel (1978) since the value of ψ_1 given by these models is 0.10, which is small compared with the other diagonal elements of $\boldsymbol{\Psi}$. The difference between models (a1) and (b1) is that we considered 2 factor model for model (a1) whereas 3 factor model is used for model (b1). We also used models given by (a2) and (b2).

When each dataset was generated 1000 times, we often obtained improper solutions. The frequencies of improper solutions were

$N = 30$:	(a1): 420 times,	(a2): 416 times,	(b1): 668 times,	(b2): 540 times,
$N = 50$:	(a1): 266 times,	(a2): 237 times,	(b1): 407 times,	(b2): 266 times,
$N = 100$:	(a1): 230 times,	(a2): 89 times,	(b1): 243 times,	(b2): 78 times.

We chose the adjusted parameters including a hyper-parameter of prior distribution and the number of factors using the model selection criterion GBIC given by (11). The minimum GBIC was selected for varying values of k and ρ. We also selected the number of factors using AIC and BIC, which only deal with the models estimated by the maximum likelihood method, to compare the performance of AIC and BIC with that of GBIC. The maximum likelihood estimates were obtained under the condition that $\psi_i \geq 0.005$ for $i = 1, \cdots, p$ (see Jöreskog, 1967).

Table 2 shows that how many times the model selection criteria selected each number of factors out of 1000 datasets. For example, the AIC selected the one factor model 9 times out of 1000 datasets in model (a1) when $N = 30$.

Table 2: Comparisons of model selection criteria for simulated datasets generated by (a1), (a2), (b1) and (b2). The bold text in the left column represents the true number of factors.

		$N = 30$			$N = 50$			$N = 100$		
	k	AIC	BIC	GBIC	AIC	BIC	GBIC	AIC	BIC	GBIC
(a1):	1	9	139	27	0	16	0	0	0	0
	2	712	853	964	776	982	989	776	1000	986
	3	228	8	9	188	2	11	195	0	14
	4	40	0	0	35	0	0	24	0	0
	5	11	0	0	1	0	0	5	0	0
(a2):	1	4	146	36	0	19	0	0	0	0
	2	714	844	958	766	981	993	784	1000	996
	3	225	10	6	204	0	7	188	0	4
	4	48	0	0	28	0	0	24	0	0
	5	9	0	0	2	0	0	0	0	0
(b1):	1	2	170	5	0	8	0	0	0	0
	2	77	395	265	7	248	28	0	7	0
	3	700	433	730	786	741	964	793	993	990
	4	184	1	0	184	3	6	196	0	10
	5	37	1	0	23	0	2	11	0	0
(b2):	1	0	17	0	0	0	0	0	0	0
	2	11	148	84	0	7	0	0	0	0
	3	712	830	916	782	991	996	796	998	995
	4	231	5	0	201	2	4	193	2	5
	5	46	0	0	17	0	0	11	0	0

When $N = 30$, the AIC often selected the large number of factors, whereas the BIC sometimes chose the small number of factors, which means the performance of the AIC and BIC are poor. However, the GBIC performed well compared with AIC and BIC since the GBIC most often selected the true number of factors for all models. The BIC selected 1 factor 170 times and the AIC selected 5 factors 37 times for the model (b1). However, the GBIC rarely selected 1 factor or 5 factors for this model. As a result, even if the GBIC does not select the true number of factors, the outcome may not be significantly worse such as the result of AIC and BIC.

When $N = 50$, the performance of the AIC is still poor. The performance of BIC is not poor, but the GBIC brings better result compared with the BIC for model (b1).

When $N = 100$, the performance of the GBIC and BIC is quite well, but the AIC still often selects the large number of factors. The performance of the BIC is slightly better than that of GBIC. It seems that the modeling procedure with

BIC is preferable to our proposed procedure based on the GBIC when N is large. However, the maximum likelihood procedure with BIC is not able to prevent the occurrence of improper solutions whereas our proposed procedure with GBIC can prevent the occurrence of improper solutions.

5.2 Job Application Dataset

We illustrate our modeling procedure through a job application dataset in Kendall (1980). This dataset contains 48 applicants for a certain job, who have been scored on $p = 15$ variables regarding their acceptability. The variables are

(1) Form of letter application, (2) Appearance, (3) Academic ability,
(4) Likeability, (5) Self confidence, (6) Lucidity,
(7) Honesty, (8) Salesmanship, (9) Experience,
(10) Drive, (11) Ambition, (12) Grasp,
(13) Potential, (14) Keenness to join, (15) Suitability.

We compared the performance of AIC, BIC with that of GBIC. The AIC and BIC selected 7 factor model and 4 factor model, respectively, each of which resulted in improper solutions since we obtained improper solutions when $k \geq 4$. The model selection criterion GBIC also selected 4 factor model. Hereafter we focus on the 4 factor model.

Before we illustrate our procedure, we show how the choice of a hyper-parameter is an important point. The maximum likelihood estimate of ψ_{14} was -0.000, which is apparently inappropriate. To overcome this problem, we employed our modeling method. We obtained a maximum penalized likelihood estimate of ψ_{14}, when $\rho = 0.00001$, 0.01 and 1, which is given by 0.005, 0.130 and 1.608, respectively. When $\rho = 0.00001$ the estimate of ψ_{14} was too close to zero. This shows that we were not able to prevent the occurrence of improper solutions. In comparison, the estimate of ψ_{14} was too large when $\rho = 1$. However, when $\rho = 0.01$, we obtained an appropriate estimate of ψ_{14} compared with that obtained when $\rho = 0.00001$ and $\rho = 1$.

It is important to identify the cause of the improper solutions. The maximum likelihood estimates of $\mathbf{\Psi}$ and the standard deviation $\hat{\sigma}_{\psi_i}$ of $N^{1/2}\psi_i/s_{ii}$ (see (5.50) in Lawley and Maxwell, 1971) for $k = 2$ to 4 are shown in Table 3. For $k = 4$, the maximum likelihood estimates $\hat{\psi}_1$, $\hat{\psi}_3$, $\hat{\psi}_7$, $\hat{\psi}_{13}$, $\hat{\psi}_{14}$ were less than the corresponding estimates for $k = 3$. These results for the estimates of unique variances suggest that we have identified some new common factors. Moreover, van Driel (1978) found that the value of the standard deviation $\hat{\sigma}_{\psi_i}$ may be large if the cause of improper solutions is the indefiniteness, whereas it is not especially large for each i when $k = 4$. These results indicate that the improper solutions are probably due to sampling fluctuations rather than indefiniteness of the model.

Table 3: Maximum likelihood estimates of unique variances and the standard deviations of $N^{1/2}\psi_i/s_{ii}$ for $k = 2$ to 4 in the job application data.

	$k=2$		$k=3$		$k=4$	
i	$\hat{\psi}_i$	$\hat{\sigma}_{\psi_i}$	$\hat{\psi}_i$	$\hat{\sigma}_{\psi_i}$	$\hat{\psi}_i$	$\hat{\sigma}_{\psi_i}$
1	0.546	0.834	0.535	0.816	0.444	0.757
2	0.717	0.779	0.701	0.777	0.688	0.777
3	0.951	0.449	0.945	0.473	0.523	0.779
4	0.741	0.772	0.000	0.717	0.199	0.566
5	0.139	0.340	0.109	0.291	0.112	0.249
6	0.191	0.382	0.196	0.381	0.194	0.350
7	0.795	0.757	0.445	0.771	0.341	0.800
8	0.171	0.346	0.144	0.304	0.133	0.263
9	0.366	0.766	0.360	0.747	0.360	0.756
10	0.247	0.460	0.238	0.447	0.225	0.403
11	0.178	0.361	0.157	0.325	0.140	0.271
12	0.192	0.377	0.204	0.390	0.153	0.285
13	0.208	0.404	0.183	0.357	0.089	0.195
14	0.600	0.779	0.420	0.671	−0.000	0.001
15	0.190	0.553	0.188	0.534	0.250	0.569

The estimates of Λ and Ψ obtained by using the proposed method are given in Table 4. The estimates of factor loadings Λ are rotated by varimax method (Kaiser, 1958). It can be seen from Table 4 that the proposed procedure prevents the occurrence of improper solutions and we can obtain the interpretable common factors in the following: *Career and Adequacy*, *Motivation and Ability*, *Academic Capability* and *Character*. For this reason, the proposed procedure performs well in that case.

6. Concluding Remarks

In maximum likelihood factor analysis, there arise situations in which the estimates of unique variances go to zero or become negative. To prevent the occurrence of such improper solutions, we used a Bayesian approach by specifying a proper prior distribution for unique variances. The proposed prior distribution is based on the prior distribution given by Akaike (1987). In practice, an optimal choice of the number of factors is also of importance for exploring the covariance structure. We derived the model selection and evaluation criterion GBIC from a Bayesian point of view, and used it to choose adjusted parameters that include the hyper-parameter for the proposed prior distribution and the number of factors. Monte Carlo simulations and a real data example were used to investigate the efficiency of the proposed procedure. We observed that our modeling strategy with GBIC prevents the occurrence of improper solutions and also selects the

Table 4: The estimates of factor loading Λ and unique variances Ψ obtained by the proposed method in the job application data.

i	factor 1	factor 2	factor 3	factor 4	unique variances
1	0.717	0.130	−0.107	0.118	0.453
2	0.154	0.449	0.131	0.255	0.703
3	0.116	0.072	0.735	−0.018	0.451
4	0.242	0.226	−0.053	0.848	0.178
5	−0.092	0.915	−0.083	0.149	0.135
6	0.120	0.837	0.063	0.303	0.200
7	−0.211	0.248	−0.019	0.740	0.356
8	0.238	0.893	−0.072	0.084	0.144
9	0.777	0.090	0.180	−0.051	0.363
10	0.386	0.767	−0.052	0.174	0.240
11	0.180	0.899	−0.056	0.107	0.154
12	0.267	0.790	0.180	0.348	0.161
13	0.343	0.730	0.261	0.428	0.109
14	0.366	0.430	−0.509	0.549	0.130
15	0.781	0.362	0.103	0.059	0.254

suitable number of factors simultaneously.

As a future research topic, it is interesting to construct a modeling procedure for preventing the occurrence of improper solutions in structural equation models including confirmatory factor analysis.

Acknowledgement

The authors would like to thank the anonymous reviewer for the helpful comments and suggestions.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory* (Petrov and, B. N. and Csaki, F., eds.), Akademiai Kiado, 267-281. (Reproduced in *Breakthroughs in Statistics* **1**, S. Kotz and N. L. Johnson eds., Foundations and Basic Theory, Springer–Verlag, (1992) 610-624.)

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika* **52**, 317-332.

Anderson, J. C. and Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and good ness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika* **49**, 155-173.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken.

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **5**, (pp. 111-150). University of California Press, Berkeley.

Bentler, P. M. and Tanaka, J. S. (1983). Problems with EM algorithms for ML factor analysis. *Psychometrika* **48**, 247-251.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in lisrel maximum likelihood estimation. *Psychometrika* **50**, 229-242.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52**, 345-370.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J. and Kirby, J. B. (2001). Improper solutions in structural equation models, causes, consequences, and strategies. *Sociological Methods Research* **29**, 468-508.

Clarke, M. R. B. (1970). A rapidly convergent method for maximum-likelihood factor analysis. *British Journal of Mathematical and Statistical Psychology* **23**, 43-52.

Dunson, D. B. (2006). Efficient Bayesian model averaging in factor analysis. *ISDS Discussion Paper*, Duke University.

Flora, D. B. and Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods* **9**, 466-491.

Fokouè, E. (2009). Bayesian computation of the intrinsic structure of factor analytic models. *Journal of Data Science* **7**, 285-311.

Gerbing, D. W. and Anderson, J. C. (1987). Improper solutions in the analysis of covariance structures: Their interpretability and a comparison of alternate respecifications. *Psychometrika* **52**, 99-111.

Hirose, K., Kawano, S., Konishi, S. and Ichikawa, M. (2008). Bayesian factor analysis and model selection. Preprint, MHF2008-2 , Kyushu University.

Jennrich, R. I. and Robinson, S. M. (1969). A Newton-Raphson algorithm for maximum likelihood factor analysis. *Psychometrika* **34**, 111-123.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika* **32**, 443-482.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**, 187-200.

Kano, Y. (1998). Improper solutions in exploratory factor analysis: Causes and treatments. In A. Rizzi, M. Vichi and H. Bock (Eds.), *Advances in Data Sciences and Classification.* Springer-Verlag, Berlin.

Kano, Y. and Ihara, M. (1994). Identification of inconsistent variates in factor analysis. *Psychometrika* **59**, 5-20.

Kendall, M. G. (1980). *Multivariate Analysis*, 2nd.ed. Charles Griffin, London.

Konishi, S., Ando, T. and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika* **91**, 27-43.

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling.* Springer, New York.

Krijnen, W. P., Dijkstra, T. K. and Gill, R. D. (1998). Conditions for factor (in)determinacy in factor analysis. *Psychometrika* **63**, 359-367.

Lawley, D. N. and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*, 2nd ed. Butterworths, London.

Lee, S. Y. and Song, X. Y. (2002). Bayesian selection on the number of factors in a factor analysis model. *Behaviormetrika* **29**, 23-40.

Lopes, H. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14**, 41-67.

Martin, J. K. and McDonald, R. P. (1975). Bayesian estimation in unrestricted factor analysis: A treatment for Heywood cases. *Psychometrika* **40**, 505-517.

Press, S. J. (1982). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference.* Krieger, New York.

Press, S. J. and Shigemasu, K. (1999). A note on choosing the number of factors. *Communications in Statistics-Theory and Methods* **28**, 1653-1670.

Rubin, D. B. and Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika* **47**, 69-76.

Rubin, D. B. and Thayer, D. T. (1983). More on EM for ML factor analysis. *Psychometrika* **48**, 253-257.

Sato, M. (1987). Pragmatic treatment of improper solutions in factor analysis. *Annals of the Institute of Statistical Mathematics* **39**, 443-455.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461-464.

Tierny, L. and Kanade, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82-86.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B* **61**, 611-622.

van Driel, O. P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika* **43**, 225-243.

Receied October 18, 2009; accepted January 18, 2010.

Kei Hirose
Graduate School of Mathematics
Kyushu University
744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan
k-hirose@math.kyushu-u.ac.jp

Shuichi Kawano
Human Genome Center, Institute of Medical Science
University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
skawano@ims.u-tokyo.ac.jp

Sadanori Konishi
Department of Mathematics
Chuo University
1-3-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
konishi@math.chuo-u.ac.jp

Masanori Ichikawa
Tokyo University of Foreign Studies
3-11-1 Asahi-cho, Fuchu-shi, Tokyo 183-8534, Japan
Ichikawa.M@tufs.ac.jp