BAYESIAN FACTOR ANALYSIS AND INFORMATION CRITERION

Hirose, Kei
Graduate School of Mathematics, Kyushu University

Kawano, Shuichi
Graduate School of Mathematics, Kyushu University

Konishi, Sadanori
Faculty of Mathematics, Kyushu University

https://doi.org/10.5109/18995

出版情報：Bulletin of informatics and cybernetics. 40, pp.75-87, 2008-12. Research Association of Statistical Sciences
バージョン：
権利関係：

BAYESIAN FACTOR ANALYSIS AND INFORMATION CRITERION

by

Kei Hirose, **Shuichi Kawano** and **Sadanori Konishi**

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.40

FUKUOKA, JAPAN
2008

Bulletin of Informatics and Cybernetics, Vol. 40, 2008

BAYESIAN FACTOR ANALYSIS AND INFORMATION CRITERION

By

Kei Hirose[*], **Shuichi Kawano**[†] and **Sadanori Konishi**[‡]

Abstract

Factor analysis is one of the most popular methods of multivariate statistical analysis. This technique has been widely used in the social and behavioral sciences to explore the covariance structure among observed variables in terms of a few unobservable variables. In maximum likelihood factor analysis, we often face a problem that the estimates of unique variances turn out to be zero or negative, which is called improper solutions. In order to overcome this difficulty, we employ a Bayesian approach by specifying a prior distribution for model parameters. A crucial issue in Bayesian factor analysis model is the choice of adjusted parameters including hyper-parameters for a prior distribution and also the number of factors. The selection of these parameters can be viewed as a model selection and evaluation problem. We derive an information criterion for evaluating a Bayesian factor analysis model. Our proposed procedure may be used for preventing the occurrence of improper solutions and also for choosing the appropriate number of factors. Monte Carlo simulations are conducted to investigate the efficiency of the proposed procedures.

Key Words and Phrases: Bayesian approach, EM algorithm, Factor analysis, Information criterion.

1. Introduction

Factor analysis provides a useful tool for exploring the covariance structure among a set of observed random variables by constructing a smaller number of random variables called common factors. In maximum likelihood factor analysis, the estimates of unique variances can often turn out to be zero or negative. These estimates, called improper solutions, are clearly inappropriate for an interpretation of the estimated model. There has been a great discussion about these inadmissible 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). In order to prevent the occurrence of improper solutions, we use a Bayesian approach by specifying a prior distribution for parameters. Some authors have proposed prior distributions for a Bayesian factor analysis model. Martin and

[*] Graduate School of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan. k-hirose@math.kyushu-u.ac.jp

[†] Graduate School of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan. s-kawano@math.kyushu-u.ac.jp

[‡] Faculty of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan. konishi@math.kyushu-u.ac.jp

McDonald (1975) used a prior distribution for unique variances. Press (1982) used a natural conjugate prior distribution for factor loadings and unique variances. Akaike (1987) proposed a prior distribution for factor loadings using the knowledge of information extracted from the likelihood function. We use a prior distribution proposed by Akaike (1987), since it is theoretically derived to prevent the occurrence of improper solutions and shown to be efficient numerically. An essential problem in the Bayesian approach is the choice of a hyper-parameter for the prior distribution. However, Akaike (1987) did not discuss a proper choice of a hyper-parameter and selected a hyper-parameter subjectively.

Another crucial aspect in Bayesian factor analysis model is the choice of the number of factors. In maximum likelihood factor analysis, the number of factors can be usually selected by the use of AIC, BIC or the likelihood ratio test. Some other selection procedures for maximum likelihood factor analysis are proposed by Bozdogan (1987), Ninomiya and Yanagihara (2007), and so on. However, even if we can select the number of factors by using these selection methods, the occurrence of improper solutions provokes the problem that it is difficult to interpret the estimated model. The aim of this paper is to introduce procedures for choosing the values of hyper-parameters in the prior distribution and the number of factors simultaneously.

In this paper we use a prior distribution introduced by Akaike (1987) and estimate parameters as a posterior mode, which can be obtained by maximizing the penalized likelihood function with the help of EM algorithm. In order to choose the adjusted parameters that include the hyper-parameters of the prior distribution and the number of factors, we derive a generalized information criterion (Konishi and Kitagawa, 1996) to evaluate models estimated by the maximum penalized likelihood method.

This article is organized as follows: In Section 2, we describe the Bayesian factor analysis model and provide a model estimation procedure with the technique of EM algorithm. Section 3 derives an information criterion for evaluating a Bayesian factor analysis model. Section 4 presents numerical results for artificial datasets. Summary comments are described in Section 5.

2. Bayesian factor analysis model

2.1. Factor analysis model

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 given by

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\Lambda} = (\lambda_{ij})$ is a $p \times k$ $(k < p)$ 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 $\boldsymbol{f}$ and $\boldsymbol{\varepsilon}$ are normally distributed, with $E(\boldsymbol{f}) = \boldsymbol{0}$, $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$, $E(\boldsymbol{\varepsilon}\boldsymbol{f}') = \boldsymbol{O}$, $E(\boldsymbol{f}\boldsymbol{f}') = \boldsymbol{I}_k$ and $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \boldsymbol{\Psi}$, where $\boldsymbol{I}_k$ is the identity matrix of order k, $\boldsymbol{0}$ is a 0-vector, $\boldsymbol{O}$ is a 0-matrix and $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix with i-th diagonal element ψ_i (> 0) which is called unique variance. Consequently, $\boldsymbol{x}$ is normally distributed with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ given by

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}. \tag{2}$$

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.

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

$$\log f(\boldsymbol{X}_N|\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\mu}) = -\frac{1}{2}\left\{Np\log(2\pi) + N\log|\boldsymbol{\Sigma}| + \sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\right\}. \quad (3)$$

The maximum likelihood estimate of mean vector $\boldsymbol{\mu}$ is the sample mean vector $\bar{\boldsymbol{x}}$. By substituting $\boldsymbol{\mu} = \bar{\boldsymbol{x}}$ into (3), the log-likelihood function is given by

$$\log f(\boldsymbol{X}_N|\boldsymbol{\Lambda}, \boldsymbol{\Psi}) = -\frac{N}{2}\left\{p\log(2\pi) + \log|\boldsymbol{\Sigma}| + \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S})\right\}, \quad (4)$$

where $\boldsymbol{S}$ is the sample variance-covariance matrix

$$\boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})'. \quad (5)$$

The unknown parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ are estimated by maximizing the log-likelihood function.

In practical situations, however, estimates of the unique variances can often turn out to be zero or negative, which is called improper solutions. In order to prevent the occurrence of improper solutions, various methods have been considered in the literature (see, e.g., Jöreskog, 1967; Martin and McDonald, 1975; van Driel, 1978; Kano, 1998). In this paper, we employ a Bayesian approach to handle these inadmissible estimates.

2.2. Bayesian approach

Let $\pi(\boldsymbol{\Lambda}, \boldsymbol{\Psi})$ be a prior distribution of the factor loading $\boldsymbol{\Lambda}$ and unique variance $\boldsymbol{\Psi}$. Akaike (1987) proposed a prior distribution

$$\pi(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) = K\exp\left\{-\frac{N\rho}{2}\mathrm{tr}(\boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-\frac{1}{2}})\right\}, \quad (6)$$

which is derived from the information of the likelihood function, where K denotes a normalizing constant and ρ is a hyper-parameter. The posterior distribution is then given by

$$\pi(\boldsymbol{\Lambda}, \boldsymbol{\Psi}|\boldsymbol{X}_N) = \frac{f(\boldsymbol{X}_N|\boldsymbol{\Lambda}, \boldsymbol{\Psi})\pi(\boldsymbol{\Lambda}, \boldsymbol{\Psi})}{\displaystyle\int\int f(\boldsymbol{X}_N|\boldsymbol{\Lambda}, \boldsymbol{\Psi})\pi(\boldsymbol{\Lambda}, \boldsymbol{\Psi})d\boldsymbol{\Lambda}d\boldsymbol{\Psi}}$$

$$\propto f(\boldsymbol{X}_N|\boldsymbol{\Lambda}, \boldsymbol{\Psi})\pi(\boldsymbol{\Lambda}, \boldsymbol{\Psi}). \quad (7)$$

In a Bayesian framework the parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ are estimated through mode of the posterior distribution. It is equivalent to obtain estimates by maximizing the penalized log-likelihood function

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

where the hyper-parameter ρ can be considered as a regularization parameter.

It might be noticed that the penalty term $\mathrm{tr}(\boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-\frac{1}{2}})$ diverges when the diagonal elements of $\boldsymbol{\Psi}$ turn out to be zero. This implies that the penalty term plays a key role in preventing the occurrence of improper solutions.

2.3. Model estimation - EM algorithm

We use an EM algorithm to obtain the penalized maximum likelihood estimates. In maximum likelihood factor analysis, some numerical algorithms have been proposed by earlier authors (see, e.g., Jöreskog, 1967; Jennrich and Robinson, 1969; Clarke, 1970; Lawley and Maxwell, 1971). Rubin and Thayer (1982) suggested using a EM algorithm for a factor analysis model to obtain maximum likelihood estimates. The advantage of the EM algorithm is that even if the likelihood function is not concave with respect to the parameters, the algorithm leads to a maximization of the function. Furthermore, the EM algorithm can easily be adapted to apply to the penalized maximum likelihood procedure. Bentler and Tanaka (1983) pointed out the weaknesses in the EM algorithm for factor analysis, whereas Rubin and Thayer (1983) addressed the problem of Bentler and Tanaka's (1983) discussion.

We give the expectation and maximization steps for estimating the Bayesian factor analysis model within a general framework for EM algorithm. We consider 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

$$\hat{\boldsymbol{\Lambda}} = (\boldsymbol{S}\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})(\boldsymbol{I}_k + \boldsymbol{M}^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{S}\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} + \rho\boldsymbol{M})^{-1},$$
$$\hat{\boldsymbol{\Psi}} = \mathrm{diag}\left[\boldsymbol{S} - 2\boldsymbol{S}\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{M}^{-1}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Lambda}}\boldsymbol{M}^{-1}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Lambda}}\boldsymbol{M}^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{S}\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{M}^{-1}\hat{\boldsymbol{\Lambda}}' + \rho\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'\right],$$

where $\boldsymbol{M} = \boldsymbol{I}_k + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$. The derivation is based on the procedure given by Hirose *et al.* (2008).

It is well known that all factor loadings obtained from the initial loadings by orthogonal transformation have the same ability to reproduce the covariance matrix, and hence the analysis should proceed by imposing some condition on the factor loadings. In maximum likelihood factor analysis, it is convenient to use the constraint that $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ is a diagonal matrix. This constraint is, however, not appropriate for the use of EM algorithm. Then, we use the constraint on factor loadings given by the following equation (see, e.g., Anderson and Rubin, 1956):

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{11} & 0 & 0 & \cdots & 0 & 0 \\ \lambda_{21} & \lambda_{22} & 0 & \cdots & 0 & 0 \\ \lambda_{31} & \lambda_{32} & \lambda_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \lambda_{k-1,1} & \lambda_{k-1,2} & \lambda_{k-1,3} & \cdots & \lambda_{k-1,k-1} & 0 \\ \lambda_{k,1} & \lambda_{k,2} & \lambda_{k,3} & \cdots & \lambda_{k,k-1} & \lambda_{k,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \lambda_{p,1} & \lambda_{p,2} & \lambda_{p,3} & \cdots & \lambda_{p,k-1} & \lambda_{p,k} \end{pmatrix}. \tag{9}$$

3. Model selection criteria

In the Bayesian factor analysis model, we still have crucial problems to be solved: the choices of a hyper-parameter ρ for the prior distribution and the number of factors k. While Akaike (1987) suggested that the AIC can be used to select the number of factors in a Bayesian factor analysis model, he did not discuss the problem of choosing a hyper-parameter and selected a hyper-parameter subjectively.

In this paper, we derive the generalized information criterion (GIC) proposed by Konishi and Kitagawa (1996) for evaluating a Bayesian factor analysis model, to choose the hyper-parameter ρ for the prior distribution and the number of factors k simultaneously. The GIC can be applied to the evaluation of statistical models estimated by various types of estimation procedures that include penalized likelihood procedures.

Suppose that $\boldsymbol{\theta}$ is the parameter vector of factor loadings $\boldsymbol{\Lambda}$ and unique variances $\boldsymbol{\Psi}$ given by

$$\boldsymbol{\theta} = (\boldsymbol{\lambda}'_{\cdot 1}, \boldsymbol{\lambda}'_{\cdot 2}, \cdots, \boldsymbol{\lambda}'_{\cdot k}, \mathrm{diag}(\boldsymbol{\Psi})')', \tag{10}$$

where $\boldsymbol{\lambda}_{\cdot i} = (\lambda_{i,i}, \lambda_{i+1,i}, \cdots, \lambda_{p,i})'$. Let $f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}})$ be the estimated model by penalized maximum likelihood methods. Then we have a statistical model

$$f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}) = (2\pi)^{-\frac{p}{2}} |\hat{\boldsymbol{\Sigma}}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}_n - \bar{\boldsymbol{x}})' \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x}_n - \bar{\boldsymbol{x}}) \right\}, \tag{11}$$

where $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}}$. For the Bayesian factor analysis model, GIC is given as

$$\mathrm{GIC} = -2 \sum_{n=1}^{N} \log f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}) + 2\mathrm{tr}\{R^{-1}Q\}, \tag{12}$$

where R and Q is

$$R = -\frac{1}{N} \sum_{n=1}^{N} \frac{\partial \boldsymbol{\phi}(\boldsymbol{x}_n, \boldsymbol{\theta})'}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -\frac{1}{N} \frac{\partial^2 l_\rho(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}, \tag{13}$$

$$Q = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}_n, \boldsymbol{\theta}) \frac{\partial \log f(\boldsymbol{x}_n|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \tag{14}$$

with

$$\boldsymbol{\phi}(\boldsymbol{x}_n, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \log f(\boldsymbol{x}_n|\boldsymbol{\theta}) - \frac{\rho}{2} \mathrm{tr}(\boldsymbol{\Psi}^{-\frac{1}{2}} \boldsymbol{\Lambda}\boldsymbol{\Lambda}' \boldsymbol{\Psi}^{-\frac{1}{2}}) \right\}. \tag{15}$$

The elements of the matrices R and Q required for calculating the bias correction terms are given by

$$\frac{\partial}{\partial \boldsymbol{\Lambda}} \left\{ \log f(\boldsymbol{x}_n|\boldsymbol{\theta}) - \frac{\rho}{2} \mathrm{tr}(\boldsymbol{\Psi}^{-\frac{1}{2}} \boldsymbol{\Lambda}\boldsymbol{\Lambda}' \boldsymbol{\Psi}^{-\frac{1}{2}}) \right\} = -\boldsymbol{\Omega}_n \boldsymbol{\Lambda} - \rho \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda},$$

$$\frac{\partial}{\partial \boldsymbol{\Psi}} \left\{ \log f(\boldsymbol{x}_n|\boldsymbol{\theta}) - \frac{\rho}{2} \mathrm{tr}(\boldsymbol{\Psi}^{-\frac{1}{2}} \boldsymbol{\Lambda}\boldsymbol{\Lambda}' \boldsymbol{\Psi}^{-\frac{1}{2}}) \right\} = -\frac{1}{2} \mathrm{diag}\boldsymbol{\Omega}_n + \frac{\rho}{2} \boldsymbol{\Psi}^{-2} \odot \mathrm{diag}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}'),$$

$$\frac{\partial^2 l_\rho(\boldsymbol{\Lambda},\boldsymbol{\Psi})}{\partial\lambda_{ab}\partial\lambda_{cd}} = N\Big\{ (\boldsymbol{\Sigma}^{-1})_{a,c}(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{b,d} + (\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{a,d}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{c,b} - (\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1})_{a,c}(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{b,d} - (\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{a,d}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{c,b} - (\boldsymbol{\Sigma}^{-1})_{a,c}(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{b,d} - (\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{a,d}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{c,b} - (\boldsymbol{\Sigma}^{-1})_{a,c}(\boldsymbol{I}_k)_{b,d} + (\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1})_{a,c}(\boldsymbol{I}_k)_{b,d} - \rho(\boldsymbol{\Psi}^{-1})_{a,c}(\boldsymbol{I}_k)_{b,d} \Big\},$$

$$\frac{\partial^2 l_\rho(\boldsymbol{\Lambda},\boldsymbol{\Psi})}{\partial\psi_{i}\partial\lambda_{cd}} = N\Big\{ (\boldsymbol{\Sigma}^{-1})_{c,i}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{i,d} - (\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1})_{c,i}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{i,d} - (\boldsymbol{\Sigma}^{-1})_{c,i}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{i,d} + \rho(\boldsymbol{\Psi}^{-2})_{c,i}(\boldsymbol{\Lambda})_{c,d} \Big\},$$

$$\frac{\partial^2 l_\rho(\boldsymbol{\Lambda},\boldsymbol{\Psi})}{\partial\psi_{i}\partial\psi_{j}} = \frac{N}{2}\Big\{ (\boldsymbol{\Sigma}^{-1})^2_{i,j} - 2(\boldsymbol{\Sigma}^{-1})_{i,j}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1})_{i,j} - 2\rho(\boldsymbol{\Psi}^{-3})_{i,j}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}')_{i,i} \Big\},$$

where $\odot$ is the Hadamard product, $\boldsymbol{\Omega}_n$ is

$$\boldsymbol{\Omega}_n = \boldsymbol{\Sigma}^{-1}\left\{\boldsymbol{\Sigma} - (\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})'\right\}\boldsymbol{\Sigma}^{-1},$$

and $(\boldsymbol{A})_{\alpha,\beta}$ is (α,β)-th element of the matrix $\boldsymbol{A}$. We choose the optimum values of hyper-parameter ρ and the number of factors k which minimize the values of information criterion GIC.

Other traditional model selection criteria include AIC (Akaike, 1973) and BIC (Schwarz, 1978). These model selection criteria only evaluate models estimated by the maximum likelihood method. The AIC and BIC are given by

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

where $\hat{\boldsymbol{\Lambda}}_{\mathrm{ML}}$ and $\hat{\boldsymbol{\Psi}}_{\mathrm{ML}}$ are the maximum likelihood estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$, and p^* is the number of parameters given by $p(k+1) - k(k-1)/2$.

It might be noticed that GIC reduces to TIC (Takeuchi, 1976) when $\rho = 0$. In addition, TIC can reduce to AIC if we assume that a parametric family of distributions include the true model. The relationship among these model selection criteria is detailed in Konishi and Kitagawa (2008). It should be noted that the AIC, TIC and BIC only evaluate models estimated by maximum likelihood methods. This implies that these model selection criteria can often select models which provoke improper solutions.

4. Numerical Examples

In this section two simulation studies are given to examine the performance of our proposed procedure. The purpose of the first simulation study is to confirm that the proposed procedure can prevent the occurrence of improper solutions and select the appropriate number of factors. In a second simulation study we explored Monte Carlo simulations to investigate how well the proposed procedure performs in the sense that the model selection criterion GIC can select an appropriate number of factors.

Table 1: The penalized maximum likelihood estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ when $\rho = 0.00001$, $\rho = 0.01$ and $\rho = 1$

	$\rho = 0.00001$			$\rho = 0.01$			$\rho = 1$		
i	$\hat{\mathbf{\Lambda}}$		$\mathrm{diag}(\hat{\mathbf{\Psi}})$	$\hat{\mathbf{\Lambda}}$		$\mathrm{diag}(\hat{\mathbf{\Psi}})$	$\hat{\mathbf{\Lambda}}$		$\mathrm{diag}(\hat{\mathbf{\Psi}})$
1	.420	.131	.806	.535	.159	.680	.025	.000	.999
2	.156	.492	.733	.177	.498	.712	.080	.001	.987
3	.197	−.174	.931	.236	−.157	.917	−.029	−.000	.998
4	.974	−.120	.037	.720	−.075	.462	−.018	−.000	.999
5	−.116	.373	.847	−.069	.362	.859	.066	.001	.991
6	.029	.761	.420	−.028	.725	.454	.103	.002	.979
7	−.053	.690	.521	−.119	.676	.512	.099	.002	.980

4.1. Simulation results

We considered a factor analysis model with two factors for a seven dimensional observable variables. The data were generated according to $\boldsymbol{x} \sim N(\mathbf{0}, \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi})$, where

$$\mathbf{\Lambda}' = \begin{pmatrix} 0.6 & 0.0 & 0.6 & 0.6 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.6 & 0.0 & 0.0 & 0.6 & 0.6 & 0.6 \end{pmatrix},$$

and $\mathbf{\Psi} = 0.64\,\boldsymbol{I}_7$ with $\boldsymbol{I}_7$ being a 7×7 identity matrix, and the sample sizes were 50. First, maximum likelihood estimates under the condition that $\psi_i \geq 0.005$ $(i = 1, \cdots, p)$ were given by

$$\hat{\mathbf{\Lambda}}'_{\mathrm{ML}} = \begin{pmatrix} 0.414 & 0.154 & 0.194 & 0.990 & -0.116 & 0.032 & -0.049 \\ 0.129 & 0.492 & -0.175 & -0.124 & 0.374 & 0.761 & 0.690 \end{pmatrix},$$

$$\mathrm{diag}(\hat{\mathbf{\Psi}}_{\mathrm{ML}})' = \begin{pmatrix} 0.812 & 0.735 & 0.932 & 0.005 & 0.847 & 0.419 & 0.522 \end{pmatrix}.$$

The estimate of ψ_4 equals 0.005, which is the inappropriate estimate.

Before illustrating our procedure, we show how the choice of a hyper-parameter is an important point. Table 1 summarizes the result of the penalized maximum likelihood estimates when $\rho = 0.00001$, $\rho = 0.01$ and $\rho = 1$. When $\rho = 0.00001$ the estimate of ψ_4 was close to zero, which means that we were not able to prevent the occurrence of improper solutions. The estimate of ψ_4 was not close to zero when $\rho = 1$, whereas the estimates of factor loadings were almost zeros. This shows that the estimates of $\mathbf{\Psi}$ and $\mathbf{\Lambda}$ were not appropriate. However, when $\rho = 0.01$, the proposed method prevented the occurrence of improper solutions and provided appropriate estimates of factor loadings and unique variances.

We chose the adjusted parameters including a hyper-parameter of prior distribution and the number of factors using the information criterion in (12). Figure 1 plots the information criterion against the hyper-parameter ρ. It might be seen from Figure 1 that the GIC is a convex function with respect to ρ at each number of factors on this dataset, while the convexity is not generally guaranteed. Table 2 gives the minimum value of GIC and the value of the hyper-parameter ρ estimated by using GIC when the number of factors is fixed. The value of GIC was a minimum when $k = 2$ and $\rho = 0.0794$, and

Table 2: The values of ρ corresponding to the minimum values of GIC for each number of factors

	$k=1$	$k=2$	$k=3$
ρ	0.0501	0.0794	0.1584
GIC	984.43	978.62	980.18

Table 3: The values of the bias term of AIC and GIC with $\rho = 0$

	$k=1$	$k=2$	$k=3$
AIC	14	20	25
GIC	12.74	20.41	23.22

the estimates of factor loadings $\mathbf{\Lambda}$ and unique variances $\mathbf{\Psi}$ were given by

$$\hat{\mathbf{\Lambda}}' = \begin{pmatrix} 0.495 & 0.150 & 0.219 & 0.559 & -0.038 & -0.048 & -0.125 \\ 0.155 & 0.447 & -0.132 & -0.055 & 0.323 & 0.614 & 0.584 \end{pmatrix},$$
$$\mathrm{diag}(\hat{\mathbf{\Psi}})' = \begin{pmatrix} 0.689 & 0.731 & 0.923 & 0.635 & 0.871 & 0.538 & 0.566 \end{pmatrix}.$$

Unlike the maximum likelihood estimates, the estimate of ψ_4 obtained by the proposed method did not close to zero. Furthermore, the information criterion GIC selected the true number of factors.

We also investigated the relationship between the bias term of GIC and that of AIC. When $\rho = 0$, the bias term in (12) is reduced to AIC if a parametric family of distributions include the true model, hence it can be presumed that the value of the bias term of GIC is close to that of AIC. These values are given in Table 3. The value of the bias term of GIC tends to be smaller than that of AIC when the number of factors is large. This result implies that the GIC might select the larger number of factors than AIC. In the next subsection, we examine the performance of GIC in Bayesian factor analysis from the point of view of selecting the appropriate number of factors.

4.2. Numerical comparison

Monte Carlo simulations were conducted to investigate the performance of the proposed procedure in various covariance structures and sample sizes. We observed that improper solutions often arose when the number of observations was small or communalities were large. Taking this nature of improper solutions into account, we used three models:

Model (a):

$$\mathbf{\Lambda}' = \begin{pmatrix} 0.6 & 0.0 & 0.6 & 0.6 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.6 & 0.0 & 0.0 & 0.6 & 0.6 & 0.6 \end{pmatrix},$$
$$\mathrm{diag}(\mathbf{\Psi})' = \begin{pmatrix} 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 \end{pmatrix},$$

Model (b):

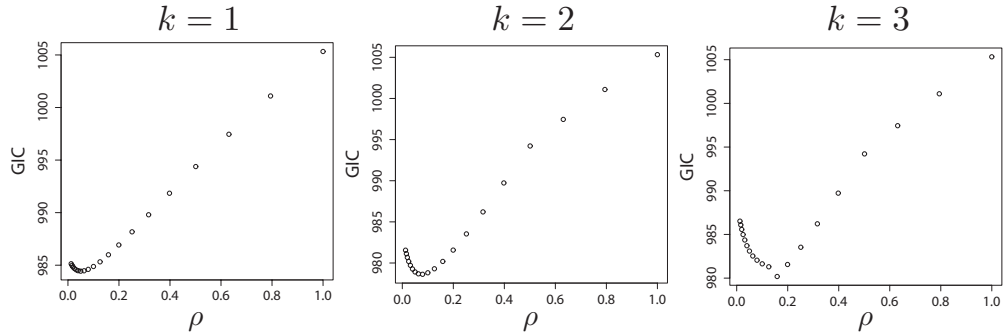

Figure 1: The information criterion GIC for varying values of the hyper-parameter ρ. On the left side $k = 1$, in the center $k = 2$ and on the right side $k = 3$

$$\boldsymbol{\Lambda}' = \begin{pmatrix} 0.6 & 0.0 & 0.0 & 0.6 & 0.6 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.6 & 0.0 & 0.0 & 0.0 & 0.6 & 0.6 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.6 & 0.0 & 0.0 & 0.0 & 0.0 & 0.6 & 0.6 \end{pmatrix},$$

$$\mathrm{diag}(\boldsymbol{\Psi})' = \begin{pmatrix} 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 \end{pmatrix},$$

Model (c):

$$\boldsymbol{\Lambda}' = \begin{pmatrix} 0.6 & 0.0 & 0.0 & 0.0 & 0.6 & 0.6 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.6 & 0.0 & 0.0 & 0.0 & 0.0 & 0.6 & 0.6 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.6 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.6 & 0.6 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.6 & 0.6 \end{pmatrix},$$

$$\mathrm{diag}(\boldsymbol{\Psi})' = \begin{pmatrix} 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 \end{pmatrix},$$

and two variants for the number of observations, $N = 30$ and $N = 50$. Each dataset was generated 100 times and improper solutions were obtained as shown:

$$\begin{array}{lll} N = 30\colon & \text{(a)}\colon 57 \text{ times}, & \text{(b)}\colon 84 \text{ times}, \quad \text{(c)}\colon 94 \text{ times}, \\ N = 50\colon & \text{(a)}\colon 30 \text{ times}, & \text{(b)}\colon 57 \text{ times}, \quad \text{(c)}\colon 70 \text{ times}. \end{array}$$

We chose the adjusted parameters including a hyper-parameter of prior distribution and the number of factors using the information criterion GIC. We also selected the number of factors using AIC and BIC, which only deal with the models estimated by a maximum likelihood method, to compare the performance of AIC and BIC with that of GIC. The maximum likelihood estimates were obtained under the constraint that $\psi_i \geq 0.005$ $(i = 1, \cdots, p)$ (see Jöreskog, 1967).

Table 4 shows that how many times the model selection criterion selected each number of factors out of 100 datasets. For example, AIC selected the one factor model 34 times out of 100 datasets at model (a) when $N = 30$. In most cases, the BIC often selected the smaller number of factors than the true number, whereas the AIC and the GIC made positive results.

Let us explore the results in further detail. In a model (a), the model selection criterion AIC selected the true number of factors than the GIC when $N = 50$, whereas there was few difference between the results selected by AIC and that selected by GIC

Table 4: Comparisons of model selection criteria for simulated datasets.

Model	N	k	AIC	BIC	GIC
(a)	30	1	34	80	16
		2	56	19	55
		3	10	1	29
	50	1	9	59	5
		2	77	41	58
		3	14	0	37
(b)	30	1	22	91	7
		2	44	8	15
		3	29	1	31
		4	5	0	21
		5	0	0	26
	50	1	10	76	5
		2	25	20	9
		3	56	4	46
		4	9	0	21
		5	0	0	19
(c)	30	1	17	95	6
		2	23	4	3
		3	32	1	19
		4	23	0	30
		5	5	0	42
	50	1	4	86	0
		2	14	12	3
		3	28	1	6
		4	49	1	51
		5	5	0	40

when $N = 30$. The result of a model (b) was the same as that of the dataset (a). However, in a model (c), the performance of the GIC was better than that of the AIC. These results can be summarized as follows:

The GIC selected the true number of factors when

1. the sample size was small,
2. the true number of factors was large.

It is observed that the model selection criterion GIC sometimes selected larger number of factors than the true number. In practical situations, we should select the optimal number of factors from not only the information criterion GIC but also various viewpoints.

5. Concluding Remarks

In maximum likelihood factor analysis models, 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 prior distribution proposed by Akaike (1987). In practice, an optimal choice of a hyper-parameter in a prior distribution and the number of factors is also of importance for exploring the covariance structure. We derived an information criterion and used it to choose the hyper-parameter for the prior distribution and the number of factors simultaneously. In order to investigate the effectiveness of the proposed procedure, two simulations studies were conducted. In the first simulation study our proposed procedure prevented the occurrence of improper solutions and chose the optimal number of factors. Monte Carlo simulations in the second simulation study showed that the proposed procedure performed well when N was small and k was large. In practical situations, however, the information criterion tends to choose the large number of factors. A further research of the treatment for improper solutions and the methods of choosing the number of factors should be investigated.

Acknowledgement

The authors would like to thank Professor Masanori Ichikawa of Tokyo University of Foreign Studies for his helpful and constructive comments and suggestions. We are grateful to the anonymous reviewer for careful reading of the manuscript and helpful comments.

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, 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). Berkeley: University of California Press.

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

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

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

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.

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* (pp. 375–382). Springer.

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

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

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

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

Ninomiya, Y. and Yanagihara, H. (2007). On testing for the number of factors in factor analysis via locally conic parametrization. IMPS2007 Invited-Paper Session "Recent developments in multivariate analysis".

Press, S. J. (1982). *Applied Multivariate analysis: using Bayesian and frequentist methods of inference*. Krieger.

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.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Sciences* **153**, 12–18 (in Japanese).

van Driel, O. P. (1978). On various causes of improper solutions in maximum likelihood

factor analysis. *Psychometrika* **43**, 225–243.

Received June 16, 2008
Revised October 6, 2008