Ann. Inst. Statist. Math.
Vol. 52, No. 4, 722–736 (2000)

JOINT BAYESIAN ANALYSIS OF FACTOR SCORES AND STRUCTURAL PARAMETERS IN THE FACTOR ANALYSIS MODEL*

SIK-YUM LEE[1] AND JIAN-QING SHI[2]

[1] *Department of Statistics, Chinese University of Hong Kong, China*
[2] *Department of Statistics, University of Warwick, U.K.*

(Received June 25, 1998; revised May 14, 1999)

Abstract. A Bayesian approach is developed to assess the factor analysis model. Joint Bayesian estimates of the factor scores and the structural parameters in the covariance structure are obtained simultaneously. The basic idea is to treat the latent factor scores as missing data and augment them with the observed data in generating a sequence of random observations from the posterior distributions by the Gibbs sampler. Then, the Bayesian estimates are taken as the sample means of these random observations. Expressions for implementing the algorithm are derived and some statistical properties of the estimates are presented. Some aspects of the algorithm are illustrated by a real example and the performance of the Bayesian procedure is studied using simulation.

Key words and phrases: Posterior distributions, conjugate prior, hyper-parameters, factor scores, Gibbs sampler, posterior mean, posterior covariance matrix, simulation study.

1. Introduction

Factor analysis is an important statistical method that has wide practical applications. Historically, it was developed by psychometricians (e.g., Thurstone (1944)) originally for the analysis of individuals' scores on mental tests; however, the model has recently been applied to a much wider range of situations; for example, analyzing sets of economic quantities, sets of tests of attitudes and behaviors, and sets of physical measurements. The definition of the model is given by the following equation:

$$\boldsymbol{y} = \boldsymbol{\Lambda}\boldsymbol{\zeta} + \boldsymbol{\delta}, \tag{1.1}$$

where $\boldsymbol{y}$ is a $p \times 1$ observed random vector, $\boldsymbol{\Lambda}$ is a $p \times r$ factor loading matrix, $\boldsymbol{\zeta}$ is a $r \times 1$ vector of factor scores and $\boldsymbol{\delta}$ is a $p \times 1$ random vector of error measurements. In this model, the random vector $\boldsymbol{y}$ is expressed as a linear combination of a relatively small number of latent factors in $\boldsymbol{\zeta}$ and a residual vector $\boldsymbol{\delta}$. Hence, the original interest was to find $\boldsymbol{\Lambda}$ and $\boldsymbol{\zeta}$.

Suppose the distribution of $\boldsymbol{\delta}$ is $N[\mathbf{0}, \boldsymbol{\Psi}]$, where $\boldsymbol{\Psi}$ is a diagonal matrix. In the classical analysis of the model, the latent vector of factor scores $\boldsymbol{\zeta}$ was either treated as a random vector or as a vector of incidental parameters that varies from observation to observation. In the latter case, it has been shown by Anderson and Rubin (1956) that the likelihood function based on a random sample $\{\boldsymbol{y}_i, i = 1, \ldots, n\}$ does not have

* The work described in the paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. CUHK4026/97H).

a maximum and the joint maximum likelihood estimates of the factor scores and the unknown parameters in $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ do not exist. Hence, it became a common practice to treat the factor scores as random. Under the assumption that $\boldsymbol{\zeta}$ is distributed as $N[\mathbf{0}, \mathbf{\Phi}]$ for some positive definite covariance matrix $\mathbf{\Phi}$, the model was then analyzed within the covariance structure analysis (also known as structural equation modeling) framework, where the structural parameter vector $\boldsymbol{\theta}$ that contains the unknown parameters in $\mathbf{\Lambda}$, $\mathbf{\Phi}$ and $\mathbf{\Psi}$ is estimated via standard approaches such as maximum likelihood (ML) and generalized least squares (Browne (1974); Jöreskog (1978); Lee and Jennrich (1979); and Bentler (1983)). Under this framework, the estimation of factor scores is not involved and the main application of the factor analysis model is to find a plausible structure for the covariance matrix of $\boldsymbol{y}$ that can explain the underlying theory of the practical situation. To obtain a solution of the basic problem in estimating the factor scores, most existing methods (e.g., Lawley and Maxwell (1971); Bartholomew (1981)) have to assume the unrealistic assumption that the structural parameter vector $\boldsymbol{\theta}$ is known. In practice, since $\boldsymbol{\theta}$ is unknown, it is replaced by its estimate and the sampling errors are ignored.

This article develops a Bayesian approach in estimating jointly the parameter vector $\boldsymbol{\theta}$ and the factor scores in a confirmatory factor analysis model. According to our knowledge, existing published work relating to Bayesian analysis of the factor analysis model is rather limited. Based on the restrictive exploratory factor analysis model, Martin and McDonald (1975) proposed a Bayesian procedure to handle Heywood cases. Bartholomew (1981) gave Bayesian estimates of the factor scores in a general confirmatory factor analysis model; however, elements in $\boldsymbol{\theta}$ were not estimated. Based on the same confirmatory model, Lee (1981) used a hierarchical Bayesian approach to estimate $\boldsymbol{\theta}$, but factor scores were not estimated. The proposed Bayesian procedure in this article is more general than the above cited work in one or more aspects. We will consider the general confirmatory model, and joint Bayesian estimates of $\boldsymbol{\theta}$ and the factor scores will be produced simultaneously. Hence, direct factor scores estimates that do not express in terms of $\boldsymbol{\theta}$ are produced. It will be shown that the joint Bayesian estimates are generally better than the classical ML estimates, and they are fairly robust to the prior information on the values of the hyper-parameters in the prior distributions.

The idea of our procedure is to utilize the distributional properties of the fundamental latent factor scores by treating them as missing data. This missing data set will be augmented with the observed data to generate a sequence of random observations of $\boldsymbol{\theta}$ and factor scores from the appropriate posterior distributions via the Gibbs sampler. Then, based on the simulated sample, the analysis of the model can be carried out easily by means of some standard data analysis methods.

The structure of the paper is as follows. The main Bayesian theory for the confirmatory factor analysis model is developed in Section 2. Expressions for implementation of the Gibbs sampler in generating the random observations from the appropriate posterior distributions are also presented in this section. Here, some related statistical properties will be discussed as well. To illustrate some aspects of the algorithm and the performances of the Bayesian procedure, results from analysis of a real example and a simulation study are reported in Section 3. Finally, Section 4 presents some concluding comments. Some technical derivations are presented in the Appendix.

2. Bayesian analysis of the factor analysis model

Consider a random sample $\{\boldsymbol{y}_i,\ i = 1, \ldots, n\}$ that satisfies the factor analysis model as given in (1.1). Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ be the observed data matrix, $\boldsymbol{Z} = (\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n)$ be

the matrix of latent factor scores, and $\boldsymbol{\theta}$ be the structural parameter vector that contains the unknown elements of $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ in the covariance structure $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ of the model. It is assumed that the model $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is identified; and in order to achieve this, some appropriate elements in $\boldsymbol{\Lambda}$ may set equal to fixed known values. From a Bayesian point of view, this is equivalent to assigning the fixed values to these parameters with probability one, and in the analysis, they are not estimated. In our analysis, we will use the following essential idea behind the EM algorithm (Dempster *et al.* (1977)): treat the latent factor scores in $\boldsymbol{Z}$ as hypothetical missing data, and augment the observed data set $\boldsymbol{Y}$ with $\boldsymbol{Z}$ to develop the Bayesian procedure. In the posterior analysis, a sufficiently large sample of $(\boldsymbol{\theta}, \boldsymbol{Z})$ from the joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{Z}$ given $\boldsymbol{Y}$ is generated by the Gibbs sampler algorithm (Geman and Geman (1984)) as follows: At the $(j+1)$-th iteration with a current values of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}^{(j)}$,

(c1) generate $\boldsymbol{Z}^{(j+1)}$ from $p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta}^{(j)})$,

(c2) generate $\boldsymbol{\theta}^{(j+1)}$ from $p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{Z}^{(j)})$; update j, return to (c1) and continue,

where $p(\cdot \mid \cdot)$ indicates the conditional density function. It has been shown (Geman and Geman (1984); Geyer (1992)) that under mild conditions and for sufficiently large j, say J, the joint distribution of $(\boldsymbol{\theta}^{(J)}, \boldsymbol{Z}^{(J)})$ converges at an exponential rate to the desired posterior distribution $[\boldsymbol{\theta}, \boldsymbol{Z} \mid \boldsymbol{Y}]$. Hence, $[\boldsymbol{\theta}, \boldsymbol{Z} \mid \boldsymbol{Y}]$ can be approximated by the empirical distribution of $\{(\boldsymbol{\theta}^{(t)}, \boldsymbol{Z}^{(t)}) : t = J+1, \ldots, J+T\}$ where T is chosen to give sufficient precision to the empirical distribution. To obtain a more nearly independent sample, observations may be collected in cycles with indices $t = J+s, J+2s, \ldots, J+Ts$ for some spacing s (see, Gelfand and Smith (1990)). However, in most practical applications a small s will suffice for many statistical analyses such as getting estimates of the parameters and standard errors, see Zeger and Karim (1991), Albert and Chib (1993). To implement the algorithm, conditional distributions of $\boldsymbol{\theta}$ given $(\boldsymbol{Y}, \boldsymbol{Z})$, and $\boldsymbol{Z}$ given $(\boldsymbol{Y}, \boldsymbol{\theta})$ are required. These distributions are discussed as follows.

2.1 *Conditional distribution of $\boldsymbol{Z}$ given $(\boldsymbol{Y}, \boldsymbol{\theta})$*

The derivation of $p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta})$ is based on the definition of the model and the distributional properties of the random vectors $\boldsymbol{y}_i$ and $\boldsymbol{\zeta}_i$. It is noted that for $i = 1, \ldots, n$, $\boldsymbol{\zeta}_i$ are mutually independent; and $\boldsymbol{y}_i$ are also mutually independent given $(\boldsymbol{\zeta}_i, \boldsymbol{\theta})$. Hence, we have

$$p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{\zeta}_i \mid \boldsymbol{y}_i, \boldsymbol{\theta}) \propto \prod_{i=1}^{n} p(\boldsymbol{\zeta}_i \mid \boldsymbol{\theta})\ p(\boldsymbol{y}_i \mid \boldsymbol{\zeta}_i, \boldsymbol{\theta}). \tag{2.1}$$

Moreover, since the conditional distributions of $\boldsymbol{\zeta}_i$ given $\boldsymbol{\theta}$, and $\boldsymbol{y}_i$ given $(\boldsymbol{\zeta}_i, \boldsymbol{\theta})$ are $N(\mathbf{0}, \boldsymbol{\Phi})$ and $N(\boldsymbol{\Lambda}\boldsymbol{\zeta}_i, \boldsymbol{\Psi})$ respectively, it can be shown that the conditional distribution of $\boldsymbol{\zeta}_i$ given $(\boldsymbol{y}_i, \boldsymbol{\theta})$ is given by

$$[\boldsymbol{\zeta}_i \mid \boldsymbol{y}_i, \boldsymbol{\theta}] \overset{D}{=} N[(\boldsymbol{\Phi}^{-1} + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{y}_i, (\boldsymbol{\Phi}^{-1} + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}], \tag{2.2}$$

where $[\cdot \mid \cdot]$ denotes the underlying conditional distribution. Hence, the conditional distribution of $\boldsymbol{Z}$ given $(\boldsymbol{Y}, \boldsymbol{\theta})$ can be obtained from (2.1) and (2.2).

2.2 *Conditional distribution of $\boldsymbol{\theta}$ given $(\boldsymbol{Y}, \boldsymbol{Z})$ with conjugate prior distributions*

The conditional distribution of $\boldsymbol{\theta}$ given $(\boldsymbol{Y}, \boldsymbol{Z})$ is proportional to $p(\boldsymbol{\theta})p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\theta})$. Hence, it is necessary to select the prior probability density function $p(\boldsymbol{\theta})$ that represents the prior information of $\boldsymbol{\theta}$. Based on the factor analysis as defined in (1.1), we first note that with given $\boldsymbol{Z}$ the underlying model becomes a regression model with parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ only. On the other hand, only the parameter matrix $\boldsymbol{\Phi}$ is involved in the distribution of the latent factor $\boldsymbol{\zeta}_i$. Hence, it is reasonable to assume that the prior distributions of $(\boldsymbol{\Lambda}, \boldsymbol{\Psi})$ and $\boldsymbol{\Phi}$ are independent. As a result, in this paper we specify the prior distribution as follows:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Psi}) = p(\boldsymbol{\Lambda}, \boldsymbol{\Psi})\, p(\boldsymbol{\Phi}). \tag{2.3}$$

Moreover, the distribution of $\boldsymbol{Y}$ only depends on $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ when $\boldsymbol{z}$ is given, and the distribution of $\boldsymbol{Z}$ only involves $\boldsymbol{\Phi}$. Consequently, it follows that

$$\begin{aligned} p(\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\Phi} \mid \boldsymbol{Y}, \boldsymbol{Z}) &= p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{Z}) \propto p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{Y} \mid \boldsymbol{\theta}, \boldsymbol{Z})p(\boldsymbol{Z} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &= p(\boldsymbol{Y} \mid \boldsymbol{\theta}, \boldsymbol{Z})p(\boldsymbol{Z} \mid \boldsymbol{\theta})p(\boldsymbol{\Lambda}, \boldsymbol{\Psi})p(\boldsymbol{\Phi}) \\ &= [p(\boldsymbol{\Lambda}, \boldsymbol{\Psi})\, p(\boldsymbol{Y} \mid \boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{Z})] \cdot [p(\boldsymbol{Z} \mid \boldsymbol{\Phi})\, p(\boldsymbol{\Phi})]. \end{aligned} \tag{2.4}$$

Since the first term of the product on the right hand side of (2.4) depends only on $(\boldsymbol{\Lambda}, \boldsymbol{\Psi})$ while the second term depends only on $\boldsymbol{\Phi}$, the marginal conditional densities $p(\boldsymbol{\Lambda}, \boldsymbol{\Psi} \mid \boldsymbol{Y}, \boldsymbol{Z})$ and $p(\boldsymbol{\Phi} \mid \boldsymbol{Y}, \boldsymbol{Z})$ are proportional to $p(\boldsymbol{\Lambda}, \boldsymbol{\Psi})p(\boldsymbol{Y} \mid \boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{Z})$ and $p(\boldsymbol{Z} \mid \boldsymbol{\Phi})p(\boldsymbol{\Phi})$, respectively.

Now, we need to select prior distributions for $(\boldsymbol{\Lambda}, \boldsymbol{\Psi})$ and $\boldsymbol{\Phi}$. Based on the justifications and rationale given in Raiffa and Schlaifer (1961), Lee (1981), Lindley and Smith (1972), Broemeling (1985), and Press and Shigemasu (1989), the following conjugate type prior distributions are considered. Let ψ_{kk} and $\boldsymbol{\Lambda}_k'$ be the k-th diagonal elements of $\boldsymbol{\Psi}$ and the k-th row of $\boldsymbol{\Lambda}$ respectively. For any $k \neq h$, we assume that the prior distribution of ψ_{kk} is independent of ψ_{hh}, and $\boldsymbol{\Lambda}_k$ is independent of $\boldsymbol{\Lambda}_h$; moreover,

$$\begin{aligned} &\psi_{kk}^{-1} \overset{D}{=} \text{Gamma}[\alpha_{0k}, \beta_{0k}], \qquad [\boldsymbol{\Lambda}_k \mid \psi_{kk}] \overset{D}{=} N[\boldsymbol{\Lambda}_{0k}, \psi_{kk}\boldsymbol{H}_{0k}], \quad \text{and} \\ &\boldsymbol{\Phi}^{-1} \overset{D}{=} W[\boldsymbol{R}_0, \rho_0, r], \end{aligned} \tag{2.5}$$

where $W[\cdot, \cdot, \cdot]$ denotes the Wishart distribution, α_{0k}, β_{0k}, $\boldsymbol{\Lambda}_{0k}$, ρ_0 and the positive definite matrices $\boldsymbol{H}_{0k}$ and $\boldsymbol{R}_0$ are hyper-parameters whose values are assumed to be given from the prior information of previous studies or other sources. It will be shown by our simulation study in the next section that the Bayesian solution is fairly robust to the selected values of these hyper-parameters.

Let $\gamma_k = \psi_{kk}^{-1}$, $\boldsymbol{Y}_k'$ be the k-th row of $\boldsymbol{Y}$, $\boldsymbol{\Omega}_k = (\boldsymbol{H}_{0k}^{-1} + \boldsymbol{Z}\boldsymbol{Z}')^{-1}$, $\boldsymbol{\mu}_k = \boldsymbol{\Omega}_k(\boldsymbol{H}_{0k}^{-1}\boldsymbol{\Lambda}_{0k} + \boldsymbol{Z}\boldsymbol{Y}_k)$, and $\beta_k = \beta_{0k} + 2^{-1}(\boldsymbol{Y}_k'\boldsymbol{Y}_k - \boldsymbol{\mu}_k'\boldsymbol{\Omega}_k^{-1}\boldsymbol{\mu}_k + \boldsymbol{\Lambda}_{0k}'\boldsymbol{H}_{0k}^{-1}\boldsymbol{\Lambda}_{0k})$, it can be shown as in the Appendix that for $k = 1, \ldots, p$, the conditional distribution of $(\boldsymbol{\Lambda}_k, \gamma_k)$ given $\boldsymbol{Y}$ and $\boldsymbol{Z}$ is independently distributed as the following Normal Gamma distribution (Broemeling (1985)):

$$[\gamma_k \mid \boldsymbol{Y}, \boldsymbol{Z}] \overset{D}{=} \text{Gamma}[n/2 + \alpha_{0k}, \beta_k], \qquad \text{and} \qquad [\boldsymbol{\Lambda}_k \mid \boldsymbol{Y}, \boldsymbol{Z}, \gamma_k] \overset{D}{=} N[\boldsymbol{\mu}_k, \gamma_k^{-1}\boldsymbol{\Omega}_k]. \tag{2.6}$$

Since $p(\boldsymbol{\Lambda}_k, \gamma_k \mid \boldsymbol{Y}, \boldsymbol{Z}) = p(\gamma_k \mid \boldsymbol{Y}, \boldsymbol{Z})p(\boldsymbol{\Lambda}_k \mid \boldsymbol{Y}, \boldsymbol{Z}, \gamma_k)$, the conditional distribution of $(\boldsymbol{\Lambda}_k, \gamma_k)$ given $(\boldsymbol{Y}, \boldsymbol{Z})$ can be obtained via (2.6).

Moreover, from the prior distribution of $\boldsymbol{\Phi}^{-1}$ given in (2.5) and the distribution of $\boldsymbol{\zeta}_i$ given $\boldsymbol{\Phi}$, which is $N(\mathbf{0}, \boldsymbol{\Phi})$, it can be shown that

$$\begin{aligned} p(\boldsymbol{\Phi} \mid \boldsymbol{Y}, \boldsymbol{Z}) \propto & \left[|\boldsymbol{\Phi}|^{-(\rho_0+r+1)/2} \exp\left\{ -\frac{1}{2} \operatorname{tr}[\boldsymbol{R}_0^{-1}\boldsymbol{\Phi}^{-1}] \right\} \right] \\ & \cdot \left[|\boldsymbol{\Phi}|^{-n/2} \exp\left\{ \left(-\frac{1}{2} \sum_{i=1}^{n} \boldsymbol{\zeta}_i' \boldsymbol{\Phi}^{-1} \boldsymbol{\zeta}_i \right) \right\} \right] \\ = & |\boldsymbol{\Phi}|^{-(n+\rho_0+r+1)/2} \exp\left\{ -\frac{1}{2} \operatorname{tr}[\boldsymbol{\Phi}^{-1}(\boldsymbol{Z}\boldsymbol{Z}' + \boldsymbol{R}_0^{-1})] \right\}. \end{aligned}$$

Since the right hand side of above equation is proportional to the density function of an inverted Wishart distribution (see, Zellner (1971)), we have

$$[\boldsymbol{\Phi} \mid \boldsymbol{Y}, \boldsymbol{Z}] \stackrel{D}{=} IW[(\boldsymbol{Z}\boldsymbol{Z}' + \boldsymbol{R}_0^{-1}), n + \rho_0, r], \tag{2.7}$$

where $IW[\cdot, \cdot, \cdot]$ denotes the inverted Wishart distribution.

From results given in (2.4), (2.6) and (2.7), the derivation of the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{Z})$ is completed. This distribution will be extended to handle the general situation with fixed known elements in $\boldsymbol{\Lambda}$ as follows.

Let $c_{kj} = 0$ if λ_{kj} is a fixed parameter; and $c_{kj} = 1$ if λ_{kj} is an unknown parameter for $k = 1, \ldots, p$, $j = 1, \ldots, r$, and $r_k = c_{k1} + \cdots + c_{kr}$. Moreover, let $\boldsymbol{\Lambda}_k^{*\prime}$ be the 1 by r_k row vector that contains the unknown parameters in $\boldsymbol{\Lambda}_k$; $\boldsymbol{Z}_k^*$ be the r_k by n submatrix of $\boldsymbol{Z}$ such that for $j = 1, \ldots, r$, all the rows corresponding to $c_{kj} = 0$ are deleted; and $\boldsymbol{Y}_k^{*\prime} = (y_{k1}^*, \ldots, y_{kn}^*)$ with

$$y_{ki}^* = y_{ki} - \sum_{j=1}^{r} \lambda_{kj} \zeta_{ji} (1 - c_{kj}).$$

The conjugate prior distributions defined in (2.5) about the loading matrix becomes

$$[\boldsymbol{\Lambda}_k^* \mid \psi_{kk}] \stackrel{D}{=} N[\boldsymbol{\Lambda}_{0k}^*, \psi_{kk} \boldsymbol{H}_{0k}^*], \tag{2.8}$$

for some hyper-parameters $\boldsymbol{\Lambda}_{0k}^*$ and $\boldsymbol{H}_{0k}^*$. Let $\boldsymbol{\Omega}_k^* = (\boldsymbol{H}_{0k}^{*-1} + \boldsymbol{Z}_k^* \boldsymbol{Z}_k^{*\prime})^{-1}$, $\boldsymbol{\mu}_k^* = \boldsymbol{\Omega}_k^* (\boldsymbol{H}_{0k}^{*-1} \boldsymbol{\Lambda}_{0k}^* + \boldsymbol{Z}_k^* \boldsymbol{Y}_k^*)$, and $\beta_k^* = \beta_{0k} + \frac{1}{2}(\boldsymbol{Y}_k^{*\prime} \boldsymbol{Y}_k^* - \boldsymbol{\mu}_k^{*\prime} \boldsymbol{\Omega}_k^{*-1} \boldsymbol{\mu}_k^* + \boldsymbol{\Lambda}_{0k}^{*\prime} \boldsymbol{H}_{0k}^{*-1} \boldsymbol{\Lambda}_{0k}^*)$. Then, for $k = 1, \ldots, p$, it can be shown from exactly the same reasonings as given in the Appendix that the posterior distributions of $(\boldsymbol{\Lambda}_k^*, \gamma_k)$ and $\boldsymbol{\Phi}$ corresponding to the conjugate priors are respectively given by:

$$\begin{aligned} & [\gamma_k \mid \boldsymbol{Y}, \boldsymbol{Z}] \stackrel{D}{=} \text{Gamma}[n/2 + \alpha_{0k}, \beta_k^*], \quad [\boldsymbol{\Lambda}_k^* \mid \boldsymbol{Y}, \boldsymbol{Z}, \gamma_k] \stackrel{D}{=} N[\boldsymbol{\mu}_k^*, \gamma_k^{-1} \boldsymbol{\Omega}_k^*], \quad \text{and} \\ & [\boldsymbol{\Phi} \mid \boldsymbol{Y}, \boldsymbol{Z}] \stackrel{D}{=} IW[(\boldsymbol{Z}\boldsymbol{Z}' + \boldsymbol{R}_0^{-1}), n + \rho_0, r]. \end{aligned} \tag{2.9}$$

Based on similar reasonings, the Bayesian procedure can be extended to cover linear constraints over the factor loadings. In this situation, the corresponding conditional distribution may be slightly complicated than the Normal-Gamma distribution as given in (2.9). Some general algorithms, such as the Metropolis-Hastings (Metropolis *et al.* (1953); and Hastings (1970)) algorithm can be used to generate the required observations.

It is advantageous to use the conjugate prior distributions in the Bayesian analysis if we have reasonable prior information about the hyper-parameters. Moreover, as will

be demonstrated by our example and simulation study, the proposed procedure is fairly robust to their selected values. It should also be noted that the conditional distributions that required by the Gibbs sampler as given in (2.2) and (2.9) are very simple and drawing observations from them requires little computational effort. Although the Bayesian procedure developed in this paper is based on the conjugate prior, analysis with other prior distributions can be considered similarly with the same reasonings.

2.3 *Bayesian estimates and their statistical properties*

Let $\{(\boldsymbol{\theta}^{(t)}, \boldsymbol{Z}^{(t)}), t = 1, \ldots, T\}$ be the random observations of $(\boldsymbol{\theta}, \boldsymbol{Z})$ generated by the Gibbs sampler from the joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{Z}$ given $\boldsymbol{Y}$, $E(\boldsymbol{\theta} \mid \boldsymbol{Y})$ and $\mathrm{Var}(\boldsymbol{\theta} \mid \boldsymbol{Y})$ be the posterior mean vector and the posterior covariance matrix, respectively. The Bayesian estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ can be easily obtained from the simulated random observations as

$$\hat{\boldsymbol{\theta}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\theta}^{(t)}, \tag{2.10}$$

which is the sample estimate of $E(\boldsymbol{\theta} \mid \boldsymbol{Y})$, the posterior mean of $\boldsymbol{\theta}$ given $\boldsymbol{Y}$. Based on the results given in Geyer (1992), $\hat{\boldsymbol{\theta}}$ tends to $E(\boldsymbol{\theta} \mid \boldsymbol{Y})$ in probability as T tends to infinity (independent of n). An estimate of the posterior covariance matrix can be obtained easily via the simulated sample as well:

$$\widehat{\mathrm{Var}}(\boldsymbol{\theta} \mid \boldsymbol{Y}) = \frac{1}{T-1}\sum_{t=1}^{T} (\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}})'. \tag{2.11}$$

This estimate tends to $\mathrm{Var}(\boldsymbol{\theta} \mid \boldsymbol{Y})$ in probability as T tends to infinity (also independent of n). Other statistical inference on $\boldsymbol{\theta}$ can be carried out based on the simulated sample as well.

Assessing the plausibility of a proposed model is always a fundamental issue in the analysis. Based on the posterior predictive assessment as discussed in Rubin (1984), Gelman *et al.* (1996) recently proposed an approach for model-checking in a Bayesian framework. It has been shown that (see Gelman *et al.* (1996) and the references therein) this approach is conceptually and computationally simple, and is very useful in model-checking for a wide varieties of complicated situations. Moreover, the required computation is a hyproduct of the common Bayesian simulation procedures such as the Gibbs sampler or its related algorithms. Therefore, it is natural for us to apply their procedure in achieving a goodness of fit assessment for our proposed model. To save space, we just present the procedure; see Gelman *et al.* (1996) for a more detailed discussion.

As suggested by Gelman *et al.* (1996), we use the following sums of squares of standardized residuals as a discrepancy variable in the procedure:

$$D(\boldsymbol{Y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{y}_i'(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1}\boldsymbol{y}_i.$$

Given $\{\boldsymbol{\theta}^{(t)}, t = 1, \ldots, T\}$, the following steps are performed for each t: (i) Given $\boldsymbol{\theta}^{(t)}$, simulate a replicated data set, $\boldsymbol{Y}^{\mathrm{rep}(t)} = (\boldsymbol{y}_1^{\mathrm{rep}(t)}, \ldots, \boldsymbol{y}_n^{\mathrm{rep}(t)})$, from its posterior predictive distribution $p(\boldsymbol{Y}^{\mathrm{rep}} \mid C, \boldsymbol{\theta}^{(t)})$, where C denotes the proposed factor analysis model with covariance structure $\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$. (ii) Calculate $D(\boldsymbol{Y}, \boldsymbol{\theta}^{(t)})$ and $D(\boldsymbol{Y}^{\mathrm{rep}(t)}, \boldsymbol{\theta}^{(t)})$.

Having obtained $\{(D(\boldsymbol{Y},\boldsymbol{\theta}^{(t)}), D(\boldsymbol{Y}^{\text{rep}(t)},\boldsymbol{\theta}^{(t)})), t = 1,\ldots,T\}$, a scatter-plot of $D(\boldsymbol{Y},\boldsymbol{\theta}^{(t)})$ against $D(\boldsymbol{Y}^{\text{rep}(t)},\boldsymbol{\theta}^{(t)})$ is constructed to make a graphical assessment. A sample p-value of posterior predictive assessment is defined as the proportion of the pairs for which $D(\boldsymbol{Y}^{\text{rep}(t)},\boldsymbol{\theta}^{(t)})$ exceeds $D(\boldsymbol{Y},\boldsymbol{\theta}^{(t)})$. Since the procedure uses the available $\boldsymbol{\theta}^{(t)}$, the computational burden is rather light.

We now consider the posterior analysis about the factor scores. For any given individual $\boldsymbol{y}_i$, let $E(\boldsymbol{\zeta}_i \mid \boldsymbol{y}_i)$ and $\text{Var}(\boldsymbol{\zeta}_i \mid \boldsymbol{y}_i)$ be the posterior mean and the posterior covariance matrix, and $\boldsymbol{\zeta}_{io}$ be the true factor scores of $\boldsymbol{y}_i$. Similar to (2.10), a Bayesian estimates $\hat{\boldsymbol{\zeta}}_i$ of $\boldsymbol{\zeta}_{io}$ can be obtained based on the simulated sample from the posterior distribution as

$$\hat{\boldsymbol{\zeta}}_i = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\zeta}_i^{(t)}, \qquad i = 1,\ldots,n, \tag{2.12}$$

where $\boldsymbol{\zeta}_i^{(t)}$ is the i-th column of $\boldsymbol{Z}^{(t)}$. This gives a direct Bayesian estimate that does not express in terms of the structural parameters or their estimates. Based on similar reasonings as in Geyer (1992), it can be shown that $\hat{\boldsymbol{\zeta}}_i$ is a consistent estimate of $E(\boldsymbol{\zeta}_i \mid \boldsymbol{y}_i)$. A consistent estimate of $\text{Var}(\boldsymbol{\zeta}_i \mid \boldsymbol{y}_i)$ can be similarly obtained as in (2.11) with $\boldsymbol{\theta}$ replaced by $\boldsymbol{\zeta}_i$. In practice, standard error estimates of elements in $\hat{\boldsymbol{\zeta}}_i$ can be obtained via $\text{Var}(\hat{\boldsymbol{\zeta}}_i \mid \boldsymbol{y}_i)$. It should be noted that both $E(\boldsymbol{\zeta}_i \mid \boldsymbol{y}_i)$ and $\text{Var}(\boldsymbol{\zeta}_i \mid \boldsymbol{y}_i)$ are difficult to assess using the existing theory of factor analysis, see Bartholomew (1981).

3. Example and simulation studies

In this section, we present a real example and a simulation study to illustrate the performance of the proposed Bayesian approach.

3.1 *Numerical example: The language data*

Fuller ((1987), p. 154) provided a real data set from a study about the writing skill of nonnative speakers of English. One hundred faculty members were asked to read and give scores to two essays using a five point scale for eleven items. The information on each item in the data set is the sum of scores on that item for the two essays. A part of the data set that involved six items was analyzed by Fuller (1987). To illustrate various aspects of our proposed algorithm, this part of the data set was reanalyzed based on the assumption that the random observations are coming from a multivariate normal population with a factor analysis model. As suggested in Fuller (1987), the following structure of $\boldsymbol{\Lambda}$ is considered in the analysis:

$$\boldsymbol{\Lambda}' = \begin{bmatrix} \lambda_{11} & \lambda_{21} & 0 & 0 & 1 & 0 \\ \lambda_{12} & \lambda_{22} & \lambda_{32} & \lambda_{42} & 0 & 1 \end{bmatrix},$$

where elements with '0' and '1' are treated as fixed known values. Hence, the structural parameter vector $\boldsymbol{\theta}$ contains the unknown elements of $\boldsymbol{\Lambda}$, the upper triangular elements of $\boldsymbol{\Phi}$ and diagonal elements of $\boldsymbol{\Psi}$. The total number of unknown structural parameters is 15.

Bayesian estimates with conjugate prior distributions are first obtained via our proposed method. In practice, values of the hyper-parameters should be selected based on the historical prior information of that particular research. In this example, we do not have any historical information, so for illustration, ML estimates that obtained from

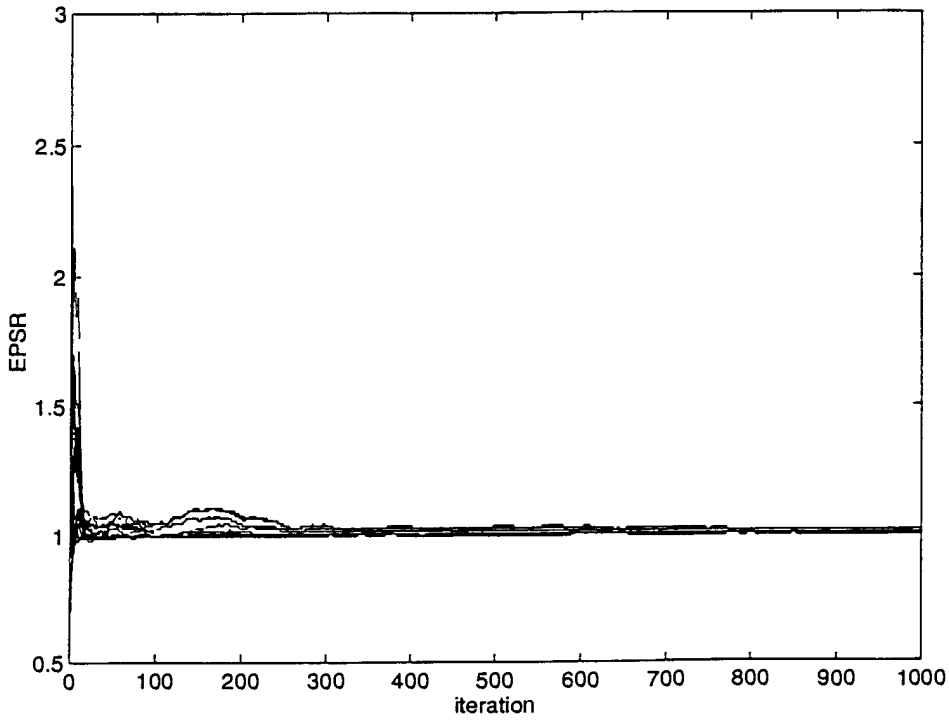

Fig. 1. Values of EPSR for all 15 parameters for the Language Data (BAY I).

the first forty observations are used to provide values for elements in $\mathbf{\Lambda}_{0k}^*$, $\boldsymbol{R}_0$, α_{0k} and β_{0k}, $k = 1, \ldots, p$. Let $\tilde{\mathbf{\Lambda}}_k^*$, $\tilde{\psi}_{kk}$ and $\tilde{\mathbf{\Phi}}$ be the corresponding ML estimates. We select $\mathbf{\Lambda}_{0k}^* = \tilde{\mathbf{\Lambda}}_k^*$; and since $E(\psi_{kk}) = \beta_{0k}/(\alpha_{0k} - 1)$ and $E(\mathbf{\Phi}) = \boldsymbol{R}_0^{-1}/(\rho_0 - r - 1)$, we take $\alpha_{0k} = 3$, $\rho_0 = r+4$, $\beta_{0k} = (\alpha_{0k}-1)\tilde{\psi}_{kk}$ and $\boldsymbol{R}_0^{-1} = (\rho_0 - r - 1)\tilde{\mathbf{\Phi}}$. Finally, for convenience, we take $\boldsymbol{H}_{0k} = \boldsymbol{I}_2$, a 2×2 identity matrix. The Bayesian estimates were obtained with the remaining sixty observations. The convergence of the Gibbs sampler is monitored by the following method as described in Gelman (1996). Based on different starting values of the structural parameters and factor scores, three parallel sequences of observations are generated and the 'estimated potential scale reduction (EPSR)' values corresponding to the 15 parameters are calculated sequentially as the runs proceed. As suggested by Gelman (1996), convergence of these sequences has been achieved if the EPSR values are all less than 1.2. Figure 1 presents the plots of the EPSR values against the iteration numbers. We observe from this figure that the sequences converged rapidly. The values of EPSR for all parameters are less than 1.2 after 100 iterations and less than 1.1 after about 200 iterations.

After the sequences have been converged, a total of $T = 4000$ observations are collected with $s = 1$. Then, the Bayesian estimates and their standard errors estimates are obtained via (2.10) and (2.11), respectively. The sample posterior predictive p-value is equal to 0.476, indicating the proposed model fits the sample data. Moreover, ML estimates of the structural parameters that based on the given data set $\boldsymbol{Y}$ are also obtained from LISREL VIII (Jöreskog and Sörbom (1994)). The Bayesian estimates (BAY), the ML estimates and estimates of the corresponding standard errors are presented in Table 1. We observe that the ML estimate of ψ_{55}, the unique variance corresponding to the fifth item, is equal to 0.06. Hence, this estimate is very close to an improper

Table 1. Bayesian estimates (BAY) and ML estimates and their standard errors for the Language Data.

	Estimates		Standard Errors			Estimates		Standard Errors	
Parameters	ML	BAY	ML	BAY	Parameters	ML	BAY	ML	BAY
λ_{11}	0.77	1.04	0.18	0.18	ψ_{11}	0.97	0.79	0.21	0.18
λ_{12}	0.28	0.06	0.18	0.20	ψ_{22}	1.03	1.02	0.17	0.16
λ_{21}	0.53	0.65	0.14	0.14	ψ_{33}	0.97	1.03	0.20	0.21
λ_{22}	0.50	0.42	0.16	0.17	ψ_{44}	0.97	1.01	0.21	0.21
λ_{32}	1.19	1.25	0.14	0.19	ψ_{55}	0.06	0.40	0.26	0.13
λ_{42}	1.27	1.34	0.15	0.19	ψ_{66}	0.89	0.93	0.17	0.17
ϕ_{11}	2.35	1.92	0.43	0.35					
ϕ_{12}	1.15	1.09	0.25	0.24					
ϕ_{22}	1.41	1.31	0.32	0.31					

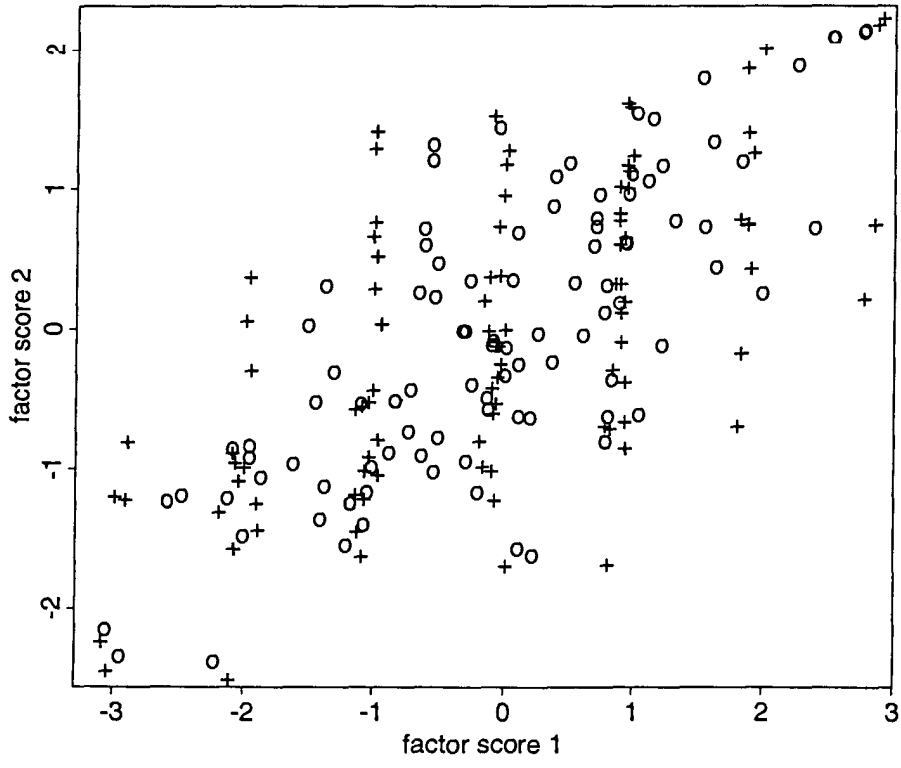

Fig. 2. Plots of the estimates of the factor scores for the Language Data.

Heywood case. As suggested by Lee (1980), Heywood cases in the ML estimation can be avoided by imposing an inequality constraints on ψ_{kk} with a penalty function. In our Bayesian approach, the conjugate prior distribution of ψ_{kk}^{-1} specified ψ_{kk} in a region of positive values and hence has a similar effect as adding a penalty function. As a result, no Heywood cases are found in the Bayesian solution because of the penalty function

induced by the prior distribution on ψ_{kk}^{-1}. This phenomenon agrees with the results in Martin and McDonald (1975).

The Gibbs sampler also generates a sequence of factor scores, $\{\boldsymbol{Z}^{(t)},\ t = 1, \ldots, T\}$, from the joint posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{Z})$ given $\boldsymbol{Y}$. For each individual $\boldsymbol{y}_i$ the corresponding Bayesian estimate of factor score, $\hat{\boldsymbol{\zeta}}_i$, is obtained from this sequence of observation via (2.12). For completeness, scores estimates were also obtained by the plots of the factor score estimates are presented in Fig. 2. More detailed comparisons with the existing methods will be given in the simulation study.

Bayesian estimates of this example have also been obtained with $s = 50$ and 100. We found that the estimates obtained with different s are very closed to each other. Using $s = 1$, Bayesian estimates have also been obtained with $T = 2000$ and 8000. We found that these estimates are very close to the previous one with $T = 4000$ (difference only at the third decimal place). Hence, it seems that the choice of s and T is not important in our analysis.

3.2 *A simulation study*

Results of a simulation study will be presented to give some ideas on the accuracy of the Bayesian estimates and to compare their performances with some estimates obtained by the existing methods. Several confirmatory factor analysis models with various factor loading structures and different numbers of variables and factors have been considered. Results obtained from each of these different cases are quite similar, hence to save space only detailed results obtained from the model with the following specifications are presented:

$$\boldsymbol{\Lambda}' = \begin{bmatrix} 0.8 & 0 & \lambda_{31} & \lambda_{41} & \lambda_{51} & 0 & 0 & 0 \\ 0 & 0.8 & 0 & 0 & 0 & \lambda_{62} & \lambda_{72} & \lambda_{82} \end{bmatrix}, \qquad \boldsymbol{\Phi} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{bmatrix}$$

and $\boldsymbol{\Psi} = \text{diag}\{\psi_{11}, \ldots, \psi_{88}\}$. To identify the model, only ϕ_{11}, ϕ_{12}, ϕ_{22}, the diagonal elements of $\boldsymbol{\Psi}$ and the λ_{kj} in $\boldsymbol{\Lambda}$ are treated as unknown parameters, while others with the given preassigned values are considered as fixed known parameters. Hence, there are a total of 17 free unknown structural parameters in this model.

For each k, the true population value, $\psi_{kk,0}$, is obtained based on a simulated observation ψ_{kk}^{-1} from Gamma$[10, 9 \times .36]$, while the true value of $\lambda_{kj,0}$ is then obtained based on a simulated value from $N(0.8, \psi_{kk,0})$. The true value of $\boldsymbol{\Phi}_0$ is obtained based on a generated observation of $\boldsymbol{\Phi}^{-1}$ from the Wishart distribution $W[\boldsymbol{R}^{*-1}/17, 20, 2]$ with $\boldsymbol{R}^*(1,1) = \boldsymbol{R}^*(2,2) = 1.0$ and $\boldsymbol{R}^*(1,2) = 0.6$. To create data sets for the simulation studies, random samples $\{\boldsymbol{\zeta}_i\}$ and $\{\boldsymbol{\delta}_i\}$ of size n were generated respectively from the corresponding multivariate normal distributions with mean vector zero and true population covariance matrices $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Psi}_0$ as given from above. Then a random sample $\{\boldsymbol{y}_i,\ i = 1, \ldots, n\}$ was obtained with the true population matrix $\boldsymbol{\Lambda}_0$ according to equation (1.1). Three sample sizes $n = 100$, 300 and 500 were selected.

Based on our proposed procedure with the Gibbs sampler, the following Bayesian estimates were obtained with different types of prior distributions; (i) BAY I: estimates based on conjugate priors with hyper-parameters obtained via the following procedure: First, a data set $\{\boldsymbol{y}_i\}$ with $n = 300$ was simulated, then the ML estimates of the 17 unknown structural parameters were obtained from the simulated data set. Finally, we took $\alpha_{0k} = 10$ and $\boldsymbol{H}_{0k} = \boldsymbol{I}_2$ for all k, $\rho_0 = 20$ and the values of other hyper-parameters such that the means of the prior distributions were equal to the corresponding ML estimates just obtained. The conjugate prior distributions obtained via this procedure represent distributions that give good prior information for the Bayesian analysis.

Table 2. RMS between the structural parameters estimates and the true values.

	$n = 100$			$n = 300$			$n = 500$		
	BAY		ML	BAY		ML	BAY		ML
	I	II		I	II		I	II	
λ_{31}	.091	.096	.159	.057	.057	.081	.048	.048	.075
λ_{41}	.095	.105	.116	.059	.062	.057	.053	.056	.050
λ_{51}	.110	.122	.131	.077	.081	.077	.057	.060	.059
λ_{62}	.091	.102	.109	.063	.067	.073	.052	.052	.053
λ_{72}	.087	.091	.101	.061	.061	.065	.051	.051	.051
λ_{82}	.107	.113	.119	.063	.067	.066	.053	.056	.054
ϕ_{11}	.196	.215	.275	.145	.159	.158	.113	.126	.114
ϕ_{12}	.141	.145	.160	.091	.090	.092	.069	.068	.069
ϕ_{22}	.206	.253	.261	.126	.136	.135	.100	.108	.100
ψ_{11}	.069	.095	.110	.043	.058	.046	.034	.045	.033
ψ_{22}	.062	.089	.066	.045	.049	.053	.030	.033	.031
ψ_{33}	.060	.071	.107	.044	.046	.055	.035	.038	.047
ψ_{44}	.060	.054	.077	.037	.035	.045	.028	.027	.030
ψ_{55}	.067	.075	.100	.048	.051	.075	.039	.039	.047
ψ_{66}	.080	.085	.085	.065	.070	.075	.052	.057	.055
ψ_{77}	.080	.080	.086	.050	.047	.048	.041	.040	.039
ψ_{88}	.060	.062	.079	.042	.040	.046	.034	.033	.036
$\sum$ RMS	1.662	1.853	2.141	1.116	1.176	1.247	.889	.937	.943

(ii) BAY II: estimates based on conjugate prior distributions with the following hyper-parameters: $\alpha_{0k} = 10$, $\lambda^*_{0kj} = 0.4$, $\beta_{0k} = 9 \times 0.18$ and $\boldsymbol{H}_{0k} = \boldsymbol{I}_2$ for all k, $\rho_0 = 20$, $\boldsymbol{R}_0(1,1) = \boldsymbol{R}_0(2,2) = 1/17$ and $\boldsymbol{R}_0(1,2) = 0$. It is noted that the prior values of λ^*_{0kj}, β_{0k} and $\boldsymbol{R}_0$ are quite different from those in BAY I. After the algorithm has been converged, we took $s = 1$ and $T = 1000$ in obtaining the Bayesian estimates. In analyzing the structural parameters, for comparison sake, ML estimates were also obtained from the LISREL VIII (Jöreskog and Sörbom, (1994)) program. The root mean squares (RMS) between the estimates and the corresponding true values based on 100 replications are reported in Table 2. To give a rough idea of the overall accuracy, the sums of the RMS across the estimates are also presented in the last row. From this table, we observe that (i) BAY I estimates that obtained with reasonably good prior information are accurate and obviously better than the ML estimates. (ii) BAY II estimates with the rough prior information are not as good as BAY I estimates, but the differences are rather minor. This result indicates that Bayesian approach with conjugate prior is robust to the choices of hyper-parameters. Apparently, when $n = 100$, and 300, the BAY II estimates with some inaccurate prior information are still better than the ML estimates. As expected, when $n = 500$, the differences between the two estimates are minor. (iii) As expected, increasing the sample size improves the accuracy of the estimates.

From previous sections, we have shown that the Bayesian estimates of the factor scores possess more established properties than the estimates obtained by the existing methods. In the study of the empirical performance of the factor scores estimates, in addition to our Bayesian estimates, those using the existing regression method (REG) (see, Lawleg and Maxwell (1971)) were also obtained. For each $n = 100$, 300 and 500, and for each replication, the RMS between the n estimates obtained by various methods and

Table 3. RMS between the factor scores estimates and the true values ($n = 100$).

		Methods		
Replications		Bay I	Bay II	REG
20th	$\hat{\zeta}(1)$	.279(.343)	.282(.323)	.333
	$\hat{\zeta}(2)$	.436(.413)	.423(.394)	.416
40th	$\hat{\zeta}(1)$	.238(.243)	.243(.237)	.270
	$\hat{\zeta}(2)$	.269(.303)	.271(.295)	.276
60th	$\hat{\zeta}(1)$	.299(.292)	.302(.284)	.298
	$\hat{\zeta}(2)$	.273(.308)	.304(.317)	.298
80th	$\hat{\zeta}(1)$	.313(.250)	.311(.234)	.388
	$\hat{\zeta}(2)$	.299(.279)	.309(.285)	.298
100th	$\hat{\zeta}(1)$	.307(.324)	.307(.306)	.301
	$\hat{\zeta}(2)$	.345(.388)	.362(.364)	.354
$\sum$ RMS		3.058	3.114	3.232

Note: Averages of the standard error estimates across the n factor score estimates are in parentheses.

the true factor scores $\{\boldsymbol{\zeta}_i\}$ were computed. Unlike the estimation of the structural parameters that can be based on the information from the whole sample $\{\boldsymbol{y}_i, i = 1, \ldots, n\}$, in estimating the true factor score $\boldsymbol{\zeta}_{i0}$, the available information is the single particular observed individual $\boldsymbol{y}_i$ with only p measurements. With this limited information, it is expected that $\hat{\boldsymbol{\zeta}}_i$ may not be very close to $\boldsymbol{\zeta}_{i0}$ and its standard error may be quite large. Moreover, the accuracy of the estimates is not improved with larger n. For each replication and each n, we have a RMS between the estimates and the true factor scores. It will take too much journal space to present the RMS for all the 100 replications and for all $n = 100$, 300, and 500. Hence, to save space, only the RMS corresponding to the 20th, 40th, ..., 100th replications with $n = 100$ are reported in Table 3, together with the averages of the standard error estimates across the corresponding n factor score estimates. The sums of the RMS across the estimates are also presented in the last row. The other RMS values and averages of standard error estimates have similar behaviours and order of magnitudes. From these results, it can be seen that the Bayesian estimates based on conjugate prior distributions are slightly better than the REG estimates. Moreover, it seems that the Bayesian estimates with conjugate priors are fairly robust to the choice of the hyper-parameter values.

4. Discussion

Factor analysis involves a model that contains both observed and latent variables. In the past years, since the joint ML estimates of the structural parameters and the factor scores do not exist (Anderson and Rubin (1956)), the traditional methods in estimating the latent factor scores are not perfect because they depend on the unrealistic assumption that the structural parameters are known. In practice, the unknown structural parameters are replaced by the estimates, ignoring the corresponding sampling errors and the statistical properties of the estimates. In computing the standard errors of the traditional factor scores estimates, some general Bootstrapping approaches may be considered. But such approaches cannot be applied in a naive way and the computational burden involved may be heavy. In this paper, a Bayesian procedure has been developed to provide joint estimates of the structural parameters and the factor scores

which do not express in terms of the structural parameters. In previous sections, it has been shown that the proposed procedure is efficient, simple to implement and produces reasonably accurate estimates.

It is well-known that the Gibbs sampler is a powerful algorithm that can be applied to solve complicated problems with ill-conditions and/or a large numbers of parameters, see Gilks *et al.* (1996), Meng and van Dyk (1997), and the references therein. In this article, the basic idea in computing the Bayesian estimates is to treat the factor scores as missing data. Then, the efficient algorithm that based on the Gibbs sampler can be applied by augmenting the observed data with the latent factor scores. In this way, the distributional properties of the latent factor scores are incorporated in the analysis. From (2.2) and (2.9), it can be seen that the conditional distributions required by the Gibbs sampler are very simple, hence the effort in simulating observations from these distributions is light. Based on our experience from the empirical studies, the convergence of the Gibbs sampler is quite rapid and the computational burden is not heavy. Moreover, the rate of convergence is improved with the increase of sample sizes. For example, in the simulation study, the average numbers of iterations required to achieve convergence for $n = 100$, 300 and 500 are roughly 200, 150 and 100, respectively. The computing time to create 100 replications in our simulation study with $n = 100$ is roughly ten minutes using a HP9000 series 735/125 workstation.

Acknowledgements

The authors are greatly indebted to two reviewers for valuable comments. The assistance of Michael K. H. Leung and Esther L. S. Tam is also acknowledged.

Appendix: Derivation of $p(\boldsymbol{\Lambda}_k, \gamma_k \mid \boldsymbol{Y}, \boldsymbol{Z})$

We now consider the derivation of $p(\boldsymbol{\Lambda}, \boldsymbol{\Psi} \mid \boldsymbol{Y}, \boldsymbol{Z})$ that proportional to $p(\boldsymbol{\Lambda}, \boldsymbol{\Psi})p(\boldsymbol{Y} \mid \boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{Z})$. Let $\gamma_k = \psi_{kk}^{-1}$; for convenience and without lost of generality, we will work with γ_k. From (2.5), the conjugate prior densities of γ_k, and $\boldsymbol{\Lambda}_k$ given γ_k are proportional to $\gamma^{\alpha_{0k}-1}\exp(-\beta_{0k}\gamma_k)$ and $\gamma_k^{r/2}\exp\{-\frac{1}{2}(\boldsymbol{\Lambda}_k - \boldsymbol{\Lambda}_{0k})'\boldsymbol{H}_{0k}^{-1}(\boldsymbol{\Lambda}_k - \boldsymbol{\Lambda}_{0k})\gamma_k\}$, respectively. Also, from (1.1), it can be seen that the likelihood of $\boldsymbol{Y}$ is given by

$$p(\boldsymbol{Y} \mid \boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{Z}) \propto |\boldsymbol{\Psi}|^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\Lambda}\boldsymbol{\zeta}_i)'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}_i - \boldsymbol{\Lambda}\boldsymbol{\zeta}_i)\right\}.$$

Let $\boldsymbol{Y}_k'$ be the k-th row of $\boldsymbol{Y}$, y_{ki} be the i-th component of $\boldsymbol{Y}_k'$, $\boldsymbol{A} = (\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{Y}_k$, and $a = \boldsymbol{Y}_k'\boldsymbol{Y}_k - \boldsymbol{Y}_k'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{Y}_k = \boldsymbol{Y}_k'\boldsymbol{Y}_k - \boldsymbol{A}'(\boldsymbol{Z}\boldsymbol{Z}')\boldsymbol{A}$, the exponential term in $p(\boldsymbol{Y} \mid \boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{Z})$ can be expressed as

$$\begin{aligned}
&-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\Lambda}\boldsymbol{\zeta}_i)'\boldsymbol{\Psi}^{-1}(\boldsymbol{y}_i - \boldsymbol{\Lambda}\boldsymbol{\zeta}_i)\\
&\quad= \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{p}\psi_{kk}^{-1}(y_{ki} - \boldsymbol{\Lambda}_k'\boldsymbol{\zeta}_i)^2\\
&\quad= -\frac{1}{2}\sum_{k=1}^{p}\left\{\psi_{kk}^{-1}\left[\sum_{i=1}^{n}y_{ki}^2 - 2\boldsymbol{\Lambda}_k'\sum_{i=1}^{n}y_{ki}\,\boldsymbol{\zeta}_i + \mathrm{tr}\left(\boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k'\sum_{i=1}^{n}\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i'\right)\right]\right\}\\
&\quad= -\frac{1}{2}\sum_{k=1}^{p}\left\{\psi_{kk}^{-1}\left[\boldsymbol{Y}_k'\boldsymbol{Y}_k - 2\boldsymbol{\Lambda}_k'\boldsymbol{Z}\boldsymbol{Y}_k + \boldsymbol{\Lambda}_k'(\boldsymbol{Z}\boldsymbol{Z}')\boldsymbol{\Lambda}_k\right]\right\}
\end{aligned}$$

$$= -\frac{1}{2}\sum_{k=1}^{p}\Big\{\psi_{kk}^{-1}\left[\boldsymbol{Y}_k'\boldsymbol{Y}_k - \boldsymbol{Y}_k'\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{Y}_k\right] + \psi_{kk}^{-1}\left[\boldsymbol{\Lambda}_k - (\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{Y}_k\right]'(\boldsymbol{Z}\boldsymbol{Z}')\left[\boldsymbol{\Lambda}_k - (\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\boldsymbol{Y}_k\right]\Big\}$$

$$= -\frac{1}{2}\sum_{k=1}^{p}\{\psi_{kk}^{-1}[a + (\boldsymbol{\Lambda}_k - \boldsymbol{A})'(\boldsymbol{Z}\boldsymbol{Z}')(\boldsymbol{\Lambda}_k - \boldsymbol{A})]\}.$$

Therefore, it follows from the likelihood of $\boldsymbol{Y}$ and the conjugate densities of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ that

$$p(\boldsymbol{\Lambda}, \boldsymbol{\Psi}^{-1} \mid \boldsymbol{Y}, \boldsymbol{Z}) \propto \prod_{k=1}^{p}\Big[\gamma_k^{n/2+r/2+\alpha_{0k}-1}\exp\Big\{-\frac{1}{2}\gamma_k[(\boldsymbol{\Lambda}_k - \boldsymbol{A})'(\boldsymbol{Z}\boldsymbol{Z}')(\boldsymbol{\Lambda}_k - \boldsymbol{A}) + (\boldsymbol{\Lambda}_k - \boldsymbol{\Lambda}_{0k})'\boldsymbol{H}_{0k}^{-1}(\boldsymbol{\Lambda}_k - \boldsymbol{\Lambda}_{0k})'] - \gamma_k(\beta_{0k} + a/2)\Big\}\Big]$$

$$= \prod_{k=1}^{p} p(\boldsymbol{\Lambda}_k, \gamma_k \mid \boldsymbol{Y}, \boldsymbol{Z}).$$

From the above equation, it can be seen that the conditional distributions of $(\boldsymbol{\Lambda}_k, \gamma_k)$ given $(\boldsymbol{Y}, \boldsymbol{Z})$ are mutually independent for $k = 1, \ldots, p$. Hence, it suffices to derive $p(\boldsymbol{\Lambda}_k, \gamma_k \mid \boldsymbol{Y}, \boldsymbol{Z})$.

Let $\boldsymbol{\Omega}_k = (\boldsymbol{H}_{0k}^{-1} + \boldsymbol{Z}\boldsymbol{Z}')^{-1}$ and $\boldsymbol{\mu}_k = \boldsymbol{\Omega}_k(\boldsymbol{H}_{0k}^{-1}\boldsymbol{\Lambda}_{0k} + \boldsymbol{Z}\boldsymbol{Y}_k)$, it follows that

$$(\boldsymbol{\Lambda}_k - \boldsymbol{A})'(\boldsymbol{Z}\boldsymbol{Z}')(\boldsymbol{\Lambda}_k - \boldsymbol{A}) + (\boldsymbol{\Lambda}_k - \boldsymbol{\Lambda}_{0k})'\boldsymbol{H}_{0k}^{-1}(\boldsymbol{\Lambda}_k - \boldsymbol{\Lambda}_{0k})'$$
$$= (\boldsymbol{\Lambda}_k - \boldsymbol{\mu}_k)'\boldsymbol{\Omega}_k^{-1}(\boldsymbol{\Lambda}_k - \boldsymbol{\mu}_k) - \boldsymbol{\mu}_k'\boldsymbol{\Omega}_k^{-1}\boldsymbol{\mu}_k + \boldsymbol{A}'\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{A} + \boldsymbol{\Lambda}_{0k}'\boldsymbol{H}_{0k}^{-1}\boldsymbol{\Lambda}_{0k}.$$

Hence,

$$p(\boldsymbol{\Lambda}_k, \gamma_k \mid \boldsymbol{Y}, \boldsymbol{Z}) = p(\gamma_k \mid \boldsymbol{Y}, \boldsymbol{Z})p(\boldsymbol{\Lambda}_k \mid \boldsymbol{Y}, \boldsymbol{Z}, \gamma_k)$$
$$\propto [\gamma_k^{n/2+\alpha_{0k}-1}\exp\{-\beta_k\gamma_k\}] \cdot \Big[\gamma_k^{r/2}\exp\Big\{-\frac{1}{2}(\boldsymbol{\Lambda}_k - \boldsymbol{\mu}_k)'\boldsymbol{\Omega}_k^{-1}(\boldsymbol{\Lambda}_k - \boldsymbol{\mu}_k)\gamma_k\Big\}\Big]$$

where $\beta_k = \beta_{0k} + 2^{-1}(\boldsymbol{Y}_k'\boldsymbol{Y}_k - \boldsymbol{\mu}_k'\boldsymbol{\Omega}_k^{-1}\boldsymbol{\mu}_k + \boldsymbol{\Lambda}_{0k}'\boldsymbol{H}_{0k}^{-1}\boldsymbol{\Lambda}_{0k})$. Thus, the posterior distribution of $(\boldsymbol{\Lambda}_k, \gamma_k)$ given $\boldsymbol{Y}$ and $\boldsymbol{Z}$ is the following Normal-Gamma distribution (Broemeling (1985)):

$$[\gamma_k \mid \boldsymbol{Y}, \boldsymbol{Z}] \overset{D}{=} \text{Gamma}[n/2 + \alpha_{0k},\ \beta_k], \quad \text{and} \quad [\boldsymbol{\Lambda}_k \mid \boldsymbol{Y}, \boldsymbol{Z}, \gamma_k] \overset{D}{=} N[\boldsymbol{\mu}_k, \gamma_k^{-1}\boldsymbol{\Omega}_k].$$

REFERENCES

Albert, J. H. and Chib, S. (1993). Bayesian analysis of Binary and polychotomous response data, *J. Amer. Statist. Assoc.*, **88**, 669–679.

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis, *Proc. Third Berkeley Symp. on Math. Statist. Prob.*, Vol. 5 110–150, University of California Press, Berkeley.

Bartholomew, D. J. (1981). Posterior analysis of the factor model, *British J. Math. Statist. Psych.*, **34**, 93–99.

Bentler, P. M. (1983). Some contributions to efficient statistics for structural models: Specification and estimation of moment structures, *Psychometrika*, **48**, 493–517.

Broemeling, L. D. (1985). *Bayesian Analysis of Linear Models*, Marcel Dekker INC, New York.

Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures, *South African Statist. J.*, **8**, 1–24.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. Roy. Statist. Soc. Ser. B*, **39**, 1–38.

Fuller, W. A. (1987). *Measurement Error Models*, Wiley, New York.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *J. Amer. Statist. Assoc.*, **85**, 398–409.

Gelman, A. (1996). Inference and monitoring convergence, *Markov Chain Monte Carlo in Practice* (eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter), 131–144, Chapman & Hall, London.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies, *Statist. Sinica*, **6**, 733–759.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion), *Statist. Sci*, **7**, 473–511.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). Introducing Markov chain Monte Carlo, *Markov Chain Monte Carlo in Practice* (eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter), 1–19, Chapman & Hall, London.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications, *Biometrika*, **57**, 97–109.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices, *Psychometrika*, **43**, 443–477.

Jöreskog, K. G. and Sörbom, D. (1994). LISREL VIII, *A Guide to the Program and Applications*, SPSS, Inc, Chicago, IL.

Lawley, D. N. and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*, 2nd ed., American Elsevier, New York.

Lee, S. Y. (1980). Estimation of covariance structure models with parameters subject to functional restraints, *Psychometrika*, **45**, 309–324.

Lee, S. Y. (1981). A Bayesian approach to confirmatory factor analysis, *Psychometrika*, **46**, 153–160.

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

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion), *J. Roy. Statist. Soc. Ser. B*, **34**, 1–42.

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

Meng, X.-L. and van Dyk, D. (1997). The EM algorithm — an old folk-song sung to a fast new time, *J. Roy. Statist. Soc. Ser. B*, **59**, 511–567.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machine, *Journal of Chemical Physics*, **21**, 1087–1091.

Press, S. J. and Shigemasu, K. (1989). Bayesian inference in factor analysis, *Contributions to probability and statistics: Essays in honor of Ingram Olkin* (eds. L. J. Gleser, M. D. Perlman, S. J. Press, and A. R. Sampson), 271–287, Springer Verlag, New York.

Raiffa, H. and Schlaifer, L. R. (1961). *Applied Statistical Decision Theory*, 2nd ed., Graduate School of Business Adiministration, Harvard, Boston.

Rubin, D. B. (1984). Bayesian justifiable and relevant frequency calculations for the applied statistician, *Ann. Statist.*, **12**, 1151–1172.

Thurstone, L. L. (1944). A multiple group method of factoring the correlation matrix, *Psychometrika*, **10**, 73–78.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; A Gibbs sampling approach, *J. Amer. Statist. Assoc.*, **86**, 79–86.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*, Wiley, New York.