Default Prior Distributions and Efficient Posterior Computation in Bayesian Factor Analysis

Joyee Ghosh[*]and David B. Dunson[†]

Abstract

Factor analytic models are widely used in social sciences. These models have also proven useful for sparse modeling of the covariance structure in multidimensional data. Normal prior distributions for factor loadings and inverse gamma prior distributions for residual variances are a popular choice because of their conditionally conjugate form. However, such prior distributions require elicitation of many hyperparameters and tend to result in poorly behaved Gibbs samplers. In addition, one must choose an informative specification, as high variance prior distributions face problems due to impropriety of the posterior distribution. This article proposes a default, heavy tailed prior distribution specification, which is induced through parameter expansion while facilitating efficient posterior computation. We also develop an approach to allow uncertainty in the number of factors. The methods are illustrated through simulated examples and epidemiology and toxicology applications.

Key words: Bayes factor; Covariance structure; Latent variables; Parameter expansion; Selection of factors; Slow mixing.

[*] Joyee Ghosh is Ph.D. candidate, Department of Statistical Science, Duke University, Durham, NC 27708-0251 (email: joyee@stat.duke.edu);

[†] David B. Dunson is Senior Investigator, Biostatistics Branch, National Institute of Environmental Health Sciences, RTP, NC 27709 and Adjunct Professor, Department of Statistical Science, Duke University, Durham, NC 27708-0251 (email: dunson1@niehs.nih.gov).

1. INTRODUCTION

Factor models have been traditionally used in behavioral sciences, where the study of latent factors such as anxiety and aggression arises naturally. These models also provide a flexible framework for modeling multivariate data by a few unobserved latent factors. In recent years factor models have found their way into many application areas beyond social sciences. For example, latent factor regression models have been used as a dimensionality reduction tool for modeling of sparse covariance structures in genomic applications (West, 2003; Carvalho et al., 2008). In addition, structural equation models and other generalizations of factor analysis are increasingly used in epidemiological studies involving complex health outcomes and exposures (Sanchez et al, 2005). There has been a recent interesting application of factor analysis for reconstruction of gene regulatory networks and unobserved activity profiles of transcription factors (Pournara and Wernisch, 2007).

Improvements in Bayesian computation permit the routine implementation of latent factor models via Markov chain Monte Carlo (MCMC) algorithms. One typical choice of prior distribution for factor models is to use normal and inverse gamma prior distributions for factor loadings and residual variances respectively. These choices are convenient, because they represent conditionally-conjugate forms that lead to straightforward posterior computation by a Gibbs sampler (Arminger, 1998; Rowe, 1998; Song and Lee, 2001).

Although conceptually straightforward, routine implementation of Bayesian factor analysis faces a number of major hurdles. First, it is difficult to elicit the hyperparameters needed in specifying the prior distribution. For example, these hyperparameters control the prior mean, variance and covariance in the factors loadings. In many cases, prior to examination of the data, one may have limited knowledge of plausible values for these parameters, and it can be difficult to convert subject matter expertise into reasonable guesses for the factor loadings and residual variances. Prior elicitation is particularly important in factor analysis, because the posterior distribution is improper in the limiting case as the prior variance for the normal and inverse-gamma components increases. In addition, as for other hierarchical models, use of a diffuse, but proper prior distribution does not solve this problem (Natarajan and McCulloch, 1998). Even for informative prior distributions, Gibbs samplers are commonly very poorly behaved due to high posterior dependence in the parameters leading to extreme slow-mixing.

To simultaneously address the need for default prior distributions and dramatically more efficient and reliable algorithms for posterior computation, this article proposes a parameter expansion approach (Liu and Wu, 1999). As noted by Gelman (2004), parameter expansion provides a useful approach for inducing new families of prior distributions. Gelman (2006) used this idea to propose a class of prior distributions for variance parameters in hierarchical models. Kinney and Dunson (2007) later expanded this class to allow dependent random effects in the context of developing Bayesian methods for random effects selection. Liu, Rubin and Wu (1998) used parameter expansion to accelerate convergence of the EM algorithm, and applied this approach to a factor model. However, to our knowledge, parameter expansion has not yet been used to induce prior distributions and improve computational efficiency in Bayesian factor analysis.

As a robust prior distribution for the factor loadings, we use parameter expansion to induce t or folded-t prior distributions, depending on sign constraints. The Cauchy or half-Cauchy case can be used as a default in cases in which subject matter knowledge is limited. We propose an efficient parameter-expanded Gibbs sampler involving generating draws from standard conditionally-conjugate distributions, followed by a post-processing step to transform back to the inferential parameterization. This algorithm is shown to be dramatically more efficient than standard Gibbs samplers in several examples. In addition, we develop an approach to allow uncertainty in the number of factors, providing an alternative to methods proposed by Lopes and West (2004) and others.

Section 2 defines the model and parameter expansion approach when the number of factors is known. Section 3 presents a comparison of the traditional and the parameter expanded Gibbs sampler, based on the results of a simulation study, when the number of factors is known. Section 4 extends this approach to allow unknown number of factors. Section 5 contains two applications, one to data from a reproductive epidemiology study and the other to data from a toxicology study. Section 6 discusses the results.

2. BAYESIAN FACTOR MODELS

2.1 Model Specification and Standard Prior Distributions

The factor model is defined as follows:

$$\mathbf{y}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma}), \tag{1}$$

where $\boldsymbol{\Lambda}$ is a $p \times k$ matrix of factor loadings, $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{ik})' \sim \mathrm{N}_k(\mathbf{0}, \mathbf{I}_k)$ is a vector of standard normal latent factors, and $\boldsymbol{\epsilon}_i$ is a residual with diagonal covariance matrix $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. The introduction of the latent factors, $\boldsymbol{\eta}_i$, induces dependence, as the marginal distribution of $\mathbf{y}_i$ is $\mathrm{N}_p(\mathbf{0}, \boldsymbol{\Omega})$, with $\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Sigma}$. In practice, the number of factors is small relative to the number of outcomes ($k << p$). Small values of k relative to p lead to sparse models for $\boldsymbol{\Omega}$ containing many fewer than $p(p+1)/2$ parameters. For this reason, factor models provide a convenient and flexible framework for modeling of a covariance matrix, particularly in applications with moderate to large p.

For simplicity in exposition, we leave the intercept out of expression (1). The factor model (1) without further constraints is not identifiable. One can obtain an identical $\boldsymbol{\Omega}$ by multiplying $\boldsymbol{\Lambda}$ by an orthonormal matrix $\mathbf{P}$ defined so that $\mathbf{P}\mathbf{P}' = \mathbf{I}_k$. Following a common convention to ensure identifiability (Geweke and Zhou, 1996), we assume that $\boldsymbol{\Lambda}$ has a full-rank lower triangular structure. The number of free parameters in $\boldsymbol{\Lambda}, \boldsymbol{\Sigma}$ is then $q = p(k+1) - k(k-1)/2$, and k must be chosen so that $q \leq p(p+1)/2$. Although we focus on the case in which the loadings matrix is lower triangular, our methods can be trivially adapted to cases in which structural zeros can be chosen in the loadings matrix based on prior knowledge. For example, the first p_1 measurements may be known to measure the first latent trait but not to measure the other latent traits. This would imply the constraint that $\lambda_{jl} = 0$, for $j = 1, \ldots, p_1$ and $l = 2, \ldots, k$.

To complete a Bayesian specification of model (1), the typical choice specifies truncated normal prior distributions for the diagonal elements of $\boldsymbol{\Lambda}$, normal prior distributions for the lower triangular elements, and inverse-gamma prior distributions for $\sigma_1^2, \ldots, \sigma_p^2$. These choices are convenient, because they represent conditionally-conjugate forms that lead to straightforward posterior computation by a Gibbs sampler (Arminger, 1998; Rowe, 1998; Song and Lee, 2001). Unfortunately, this choice is subject to the problems mentioned in Section 1.

2.2 Inducing Prior Distributions Through Parameter Expansion

In order to induce a heavier tailed prior distribution on the factor loadings to allow specification of a default, proper prior distribution, we propose a parameter expansion (PX) approach. The basic PX idea involves introduction of a *working model* that is over-parameterized. This working model is then related to the *inferential model* through a transformation. Generalizing the approach proposed by Gelman (2006) for prior distribution specification in simple ANOVA models, we define the following PX-factor model:

$$\mathbf{y}_i = \mathbf{\Lambda}^* \boldsymbol{\eta}_i^* + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\eta}_i^* \sim \mathrm{N}_k(\mathbf{0}, \mathbf{\Psi}), \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}_p(\mathbf{0}, \mathbf{\Sigma}) \tag{2}$$

where $\mathbf{\Lambda}^*$ is $p \times k$ working factor loadings matrix having a lower triangular structure without constraints on the elements, $\boldsymbol{\eta}_i^* = (\eta_{i1}^*, \ldots, \eta_{ik}^*)'$ is a vector of working latent variables, $\mathbf{\Psi} = \mathrm{diag}(\psi_1, \ldots, \psi_k)$, and $\mathbf{\Sigma}$ is a diagonal covariance matrix defined as in (1). Note that model (2) is clearly over-parameterized having redundant parameters in the covariance structure. In particular, marginalizing out the latent variables, $\boldsymbol{\eta}_i^*$, we obtain $\mathbf{y}_i \sim \mathrm{N}_p(\mathbf{0}, \mathbf{\Lambda}^* \mathbf{\Psi} \mathbf{\Lambda}^{*\prime} + \mathbf{\Sigma})$. Clearly, the diagonal elements of $\mathbf{\Lambda}^*$ and $\mathbf{\Psi}$ are redundant.

In order to relate the working model parameters in (2) to the inferential model parameters in (1), we use the following transformation:

$$\lambda_{jl} = \mathcal{S}(\lambda_{ll}^*) \lambda_{jl}^* \psi_l^{1/2}, \quad \eta_{il} = \mathcal{S}(\lambda_{ll}^*) \psi_l^{-1/2} \eta_{il}^* \quad \text{for} \quad j = 1, \ldots, p, \quad l = 1, \ldots, k \tag{3}$$

where $\mathcal{S}(x) = -1$ for $x < 0$ and $\mathcal{S}(x) = 1$ for $x \geq 0$. Then, instead of specifying a prior distribution for $\mathbf{\Lambda}$ directly, we induce a prior distribution on $\mathbf{\Lambda}$ through a prior distribution for $\mathbf{\Lambda}^*, \mathbf{\Psi}$. In particular, we let

$$\lambda^*_{jl} \overset{iid}{\sim} \mathrm{N}(0,1), j=1,\ldots,p, l=1,\ldots,\min(j,k), \lambda^*_{jl} \sim \delta_0, j=1,\ldots,(k-1), l=j+1,\ldots,k,$$
$$\psi_l^{-1} \overset{iid}{\sim} \mathcal{G}(a_l,b_l), l=1,\ldots,k, \qquad (4)$$

where δ_0 is a measure concentrated at 0, and $\mathcal{G}(a,b)$ denotes the gamma distribution with mean a/b and variance a/b^2.

The prior distribution is conditionally-conjugate, leading to straightforward Gibbs sampling, as described in Section 2.3. In the special case in which k=1 and $\lambda_{jl} = \lambda$, $\lambda^*_{jl} = \lambda^*$, the induced prior distribution on $\mathbf{\Lambda}$ reduces to the Gelman (2006) half-t prior distribution. In a general case, we obtain a t prior distribution for the off-diagonal elements of $\mathbf{\Lambda}$ and half-t prior distributions for the diagonal elements of $\mathbf{\Lambda}$ upon marginalizing out $\mathbf{\Lambda}^*$ and $\mathbf{\Psi}$. Note that we have induced prior dependence across the elements within a column of $\mathbf{\Lambda}$. In particular, columns having higher ψ_l values will tend to have higher factor loadings, while columns with low ψ_l values tend to have low factor loadings. Such a dependence structure is quite reasonable because factors which tend to have higher precision will have their corresponding factor loadings inflated. On the other hand loadings for factors with low precision will be automatically smaller. We have considered proper prior distributions for the working parameters in this paper. Alternatively improper prior distributions can be used, but then one would need to perform technical calculations to ensure that the chain corresponding to the inferential parameters has the desired posterior distribution as its stationary distribution (Meng and van Dyk, 1999).

2.3 Parameter Expanded Gibbs Sampler

After specifying the prior distribution one can then run an efficient, blocked Gibbs sampler for posterior computation in the PX-factor model. Note that the chains under the overparameterized model actually exhibit very poor mixing, reflecting the lack of identifiability. It is only after transforming back to the inferential model that we obtain improved mixing. We have found this algorithm to be highly efficient, in terms of rates of convergence and mixing, in a wide variety of simulated and real data examples. The conditional distributions are described below.

Model (2) can be written as $y_{ij} = \mathbf{z}_{ij}'\boldsymbol{\lambda}_j^* + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma_j^2)$, where $\mathbf{z}_{ij} = (\eta_{i1}^*, \ldots, \eta_{ik_j}^*)'$, $\boldsymbol{\lambda}_j^* = (\lambda_{j1}^*, \ldots, \lambda_{jk_j}^*)'$ denotes the free elements of row j of $\boldsymbol{\Lambda}^*$, and $k_j = \min(j, k)$ is the number of free elements. Let $\pi(\boldsymbol{\lambda}_j^*) = \mathrm{N}_{k_j}(\boldsymbol{\lambda}_{0j}^*, \boldsymbol{\Sigma}_{0\boldsymbol{\lambda}_j^*})$ denote the prior distribution for $\boldsymbol{\lambda}_j^*$, the full conditional posterior distributions are as follows:

$$\pi\big(\boldsymbol{\lambda}_j^* \,|\, \boldsymbol{\eta}^*, \boldsymbol{\Psi}, \boldsymbol{\Sigma}, \mathbf{y}\big) = \mathrm{N}_{k_j}\bigg(\Big(\boldsymbol{\Sigma}_{0\boldsymbol{\lambda}_j^*}^{-1} + \sigma_j^{-2}\mathbf{Z}_j'\mathbf{Z}_j\Big)^{-1}\Big(\boldsymbol{\Sigma}_{0\boldsymbol{\lambda}_j^*}^{-1}\boldsymbol{\lambda}_{0j}^* + \sigma_j^{-2}\mathbf{Z}_j'\mathbf{Y}_j\Big), \Big(\boldsymbol{\Sigma}_{0\boldsymbol{\lambda}_j^*}^{-1} + \sigma_j^{-2}\mathbf{Z}_j'\mathbf{Z}_j\Big)^{-1}\bigg),$$

where $\mathbf{Z}_j = (\mathbf{z}_{1j}, \ldots, \mathbf{z}_{nj})'$ and $\mathbf{Y}_j = (y_{1j}, \ldots, y_{nj})'$. In addition, we have

$$\begin{aligned}
\pi(\boldsymbol{\eta}_i^* \,|\, \boldsymbol{\Lambda}^*, \boldsymbol{\Sigma}, \boldsymbol{\Psi}, \mathbf{y}) &= \mathrm{N}_k\Big(\big(\boldsymbol{\Psi}^{-1} + \boldsymbol{\Lambda}^{*\prime}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}^*\big)^{-1}\boldsymbol{\Lambda}^{*\prime}\boldsymbol{\Sigma}^{-1}\mathbf{y}_i, \big(\boldsymbol{\Psi}^{-1} + \boldsymbol{\Lambda}^{*\prime}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}^*\big)^{-1}\Big), \\
\pi(\psi_l^{-1} \,|\, \boldsymbol{\eta}^*, \boldsymbol{\Lambda}^*, \boldsymbol{\Sigma}, \mathbf{y}) &= \mathcal{G}\bigg(a_l + \frac{n}{2}, b_l + \frac{1}{2}\sum_{i=1}^n \eta_{il}^{*2}\bigg), \\
\pi(\sigma_j^{-2} \,|\, \boldsymbol{\eta}^*, \boldsymbol{\Lambda}^*, \boldsymbol{\Psi}, \mathbf{y}) &= \mathcal{G}\bigg(c_j + \frac{n}{2}, d_j + \frac{1}{2}\sum_{i=1}^n (y_{ij} - \mathbf{z}_{ij}'\boldsymbol{\lambda}_j^*)^2\bigg),
\end{aligned}$$

where $\mathcal{G}(a_l, b_l)$ is the prior distribution for ψ_l^{-1}, for $l = 1, \ldots, k$, and $\mathcal{G}(c_j, d_j)$ is the prior distribution for σ_j^{-2}, for $j = 1, \ldots, p$.

Hence, the proposed PX Gibbs sampler cycles through simple steps for sampling from normal and gamma full conditional posterior distributions under the working model. After discarding a burn-in and collecting a large number of samples, we then simply apply the transformation in (3) to each of the samples as a post-processing step that produces samples from the posterior distribution under the inferential parameterization. Convergence diagnostics and inferences then rely entirely on the samples after post-processing, with the working model samples discarded.

3. SIMULATION STUDY WHEN THE NUMBER OF FACTORS IS KNOWN

We look at two simulation examples to compare the performance of the traditional and our PX Gibbs sampler. We routinely normalize the data prior to analysis.

3.1 One Factor Model

In our first simulation, we consider one of the examples in Lopes and West (2004). Here $p = 7$, $n = 100$, the number of factors, $k = 1$, $\mathbf{\Lambda} = (0.995, 0.975, 0.949, 0.922, 0.894, 0.866, 0.837)'$ and $\text{diag}(\mathbf{\Sigma}) = (0.01, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30)$. We repeat this simulation for 100 simulated data sets. To specify the prior distributions, for the traditional Gibbs sampler we choose $N_+(0, 1)$ (truncated to be positive) prior distributions for the diagonals and $N(0, 1)$ prior distributions for the lower triangular elements of $\mathbf{\Lambda}$ respectively. For the PX Gibbs sampler we induce half-Cauchy and Cauchy prior distributions, both with scale parameter 1, for the diagonals and lower triangular elements of $\mathbf{\Lambda}$ respectively. In order to induce this prior distribution we take $N(0, 1)$ prior distributions for the free elements of $\mathbf{\Lambda}^*$ and $\mathcal{G}(1/2, 1/2)$ prior distributions for the diagonal elements of $\mathbf{\Psi}$. For the residual precisions σ_j^{-2} we take $\mathcal{G}(1, 0.2)$ prior distributions for both samplers. This choice of hyperparameter values provides a modest degree of shrinkage towards a plausible range of values for the residual precision. For each simulated data set, we run the Gibbs sampler for 25,000 iterations, discarding the first 5,000 iterations as burn-in.

The Effective Sample Size (ESS) gives the size of an independent sample with the same variance as the MCMC sample under consideration (Robert and Casella, 2004), and is hence a good measure of mixing of the chain. We compare the ESS of the variance covariance matrix Ω across the 100 simulations. We find that the PX Gibbs sampler leads to a tremendous gain in ESS for all elements in Ω. This is evident from Figure (1). It is important to note that the target distributions of the MCMC algorithms for the typical prior distribution and the PX-induced prior distribution are not equivalent, so that in some sense the ESS values are not directly comparable. That said, our focus is on choosing a default prior distribution that has good properties in terms of posterior computation, and it is reasonable from this perspective to compare the ESS for the two different approaches. If substantive belief is available allowing one to follow a subjective Bayes approach and choose informative prior distributions, our recommended PX approach can still be used, though one should be careful to choose the hyperparameters based on the prior distributions induced after transforming back to the inferential model (Imai and van Dyk, 2005). We have attempted to choose prior distributions under the two approaches that are roughly comparable in terms of scale, avoiding use of diffuse, but proper prior distributions. Such prior distributions are expected to lead to very poor computational performance and unreliable estimates and inferences, because the limiting case corresponds to an improper posterior distribution. The use of induced heavy-tailed prior distributions after data standardization seems to provide a reasonable objective Bayes solution to the problem, which results in substantially improved computational performance.

3.2 Three Factor Model

For our second simulation, we have $p = 10,\ n = 100$, and the number of factors, $k = 3$. To make the problem more difficult, we set some of the loadings as negative and introduce more noise in the data.

$$\mathbf{\Lambda}' = \begin{pmatrix} 0.89 & 0.00 & 0.25 & 0.00 & 0.80 & 0.00 & 0.50 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.90 & 0.25 & 0.40 & 0.00 & 0.50 & 0.00 & 0.00 & -0.30 & -0.30 \\ 0.00 & 0.00 & 0.85 & 0.80 & 0.00 & 0.75 & 0.75 & 0.00 & 0.80 & 0.80 \end{pmatrix}$$

$$\mathrm{diag}(\mathbf{\Sigma}) = (0.2079, 0.1900, 0.1525, 0.2000, 0.3600, 0.1875, 0.1875, 1.0000, 0.2700, 0.2700).$$

We carried out the simulations exactly as in the previous simulation study. From Figure (2) we find that the gain is not as dramatic as in the previous example. This is not surprising for the following reason.

The idea behind parameter expansion is to introduce auxiliary variables in the original model to form a parameter expanded working model. Thus the working model has an extra set of parameters, and parameters under the original model are obtained by a pre-defined set of reduction functions operating on the extra parameters. The improvement in mixing occurs because of the extra randomness introduced via the auxiliary variables. Gelman (2006) also mentions that parameter expansion diminishes dependence among parameters and hence leads to better mixing.

The off-diagonal elements in the covariance matrix $\mathbf{\Omega}$ represent the correlations between the outcomes as the data are standardized. When some of the outcomes are highly correlated, the corresponding elements in $\mathbf{\Omega}$ are close to one. In such a scenario the likelihood is close to degenerate and the posterior samples under the traditional Gibbs sampler are poorly behaved, exhibiting extreme high autocorrelation for those parameters. Thus in those cases there is a tremendous gain in using the PX approach. On the other hand when the posterior correlation under the traditional Gibbs sampler is low to begin with, there is not much room for improvement using the PX approach, and hence we find a more modest gain.

It should be clarified that the modest improvement in the second example is not due to a fall off in the PX Gibbs performance as dimension increases. Instead, the gain for the first example was attributable to small values for some of the residual variances, leading to very high posterior dependence in the factor loadings. The second example had more moderate residual variances. To verify that the method is scalable to higher dimensions, we repeated the simulation in a case taken from Lopes and West (2004) with $p = 9$, $k = 3$, and with some of the outcomes highly correlated, implying low residual variance. In this case, we again observed a tremendous gain in mixing. We note that the additional computation time needed for the PX approach over the traditional approach is negligible, so that it is reasonable to focus on computational efficiency for the same number of MCMC iterates.

4. BAYESIAN MODEL SELECTION FOR UNKNOWN NUMBER OF FACTORS

4.1 Path Sampling With Parameter Expansion

To allow an unknown number of factors k, we choose a multinomial prior distribution, with $\Pr(k = h) = \kappa_h$, with $\kappa_h = 1/m$, for $h = 1, \ldots, m$. We then complete a Bayesian specification through prior distributions on the coefficients within each of the models in the list $k \in \{1, \ldots, m\}$. This is accomplished by choosing a prior distribution for the coefficients in the m factor model having the form described in Section 2, with the prior distribution for $\Lambda^{(h)}$ for any smaller model $k = h$ obtained by marginalizing out the columns from $(h + 1)$ to m. In this manner, we place a prior distribution on the coefficients in the largest model, while inducing prior distributions on the coefficients in each of the smaller models.

Bayesian selection of the number of factors relies on posterior model probabilities, $\Pr(k = h \,|\, \mathbf{y}) = \{\kappa_h \, \pi(\mathbf{y} \,|\, k = h)\}/\{\sum_{l=1}^{m} \kappa_l \, \pi(\mathbf{y} \,|\, k = l)\}$ where the marginal likelihood under model k, $\pi(\mathbf{y} \,|\, k = h)$, is obtained by integrating the likelihood $\prod_i \mathrm{N}_p(\mathbf{y}_i; \mathbf{0}, \mathbf{\Lambda}^{(k)}\mathbf{\Lambda}^{(k)\prime} + \mathbf{\Sigma})$ across the prior distribution for the factor loadings $\mathbf{\Lambda}^{(k)}$ and residual variances $\mathbf{\Sigma}$. We still need to consider the problem of estimating $\Pr(k = h \,|\, \mathbf{y})$ as the marginal likelihood is not available in closed form. Note that any posterior model probability can be expressed entirely in terms of the prior odds $O[h : j] = \{\kappa_h/\kappa_j\}$ and Bayes factors $\mathrm{BF}[h : j] = \{\pi(\mathbf{y} \,|\, k = h)/\pi(\mathbf{y} \,|\, k = j)\}$ as follows:

$$\Pr(k = h \,|\, \mathbf{y}) = \frac{O[h : j] * \mathrm{BF}[h : j]}{\sum_{l=1}^{m} O[l : j] * \mathrm{BF}[l : j]} \tag{5}$$

Lee and Song (2002) use the path sampling approach of Gelman and Meng (1998) for estimating log Bayes factors. They construct a path using a scalar $t \in [0, 1]$ to link two models M_0 and M_1. They use the same idea as outlined in an example in Gelman and Meng (1998) to construct their path. To compute the required integral they take a fixed set of grid points for t, $t \in [0, 1]$ and then use numerical integration to approximate the integration over t.

Let M_0 and M_1 correspond to models with $(h - 1)$ and h factors respectively. The two models are linked by the path: $M_t : \boldsymbol{y_i} = \mathbf{\Lambda_t}\boldsymbol{\eta_i} + \boldsymbol{\epsilon_i}$, $\mathbf{\Lambda_t} = (\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \ldots, \boldsymbol{\lambda_{(h-1)}}, t\boldsymbol{\lambda_h})$, where $\boldsymbol{\lambda_i}$ is the i^{th} column of the loadings matrix. Thus $t = 0$ and $t = 1$ correspond to M_0 and M_1. Then for a set of fixed and ordered grid-points, $t_{(0)} = 0 < t_{(1)} < \cdots < t_{(S)} < t_{(S+1)} = 1$, we have

$$\widehat{\log}(\mathrm{BF}[h : (h - 1)]) = \frac{1}{2} \sum_{s=o}^{S} (t_{(s+1)} - t_{(s)})(\bar{U}_{(s+1)} + \bar{U}_{(s)}) \tag{6}$$

where $U(\mathbf{\Lambda}, \mathbf{\Sigma}, \boldsymbol{\eta}, \boldsymbol{y}, t) = \sum_{i=1}^{n} (\boldsymbol{y_i} - \mathbf{\Lambda_t}\boldsymbol{\eta_i})'\mathbf{\Sigma}^{-1}(\mathbf{0}^{p\times(h-1)}, \boldsymbol{\lambda_h})\boldsymbol{\eta_i}$ and $\bar{U}_{(s)}$ is the average of $\{U(\mathbf{\Lambda}^{(j)}, \mathbf{\Sigma}^{(j)}, \boldsymbol{\eta}^{(j)}, \boldsymbol{y}, t_{(s)}),\ j = 1, 2, \ldots J\}$ over the J MCMC samples from $p(\mathbf{\Lambda}, \Sigma, \eta | \boldsymbol{y}, t_{(s)})$.

An important challenge in Bayes model comparisons is sensitivity to the prior distribution. It is well known that Bayes factors tend to be sensitive to the prior distribution, motivating a rich literature on objective Bayes methods (Berger and Pericchi, 1996). Lee and Song (2002) rely on highly-informative prior distributions in implementing Bayesian model selection for factor analysis, an approach which is only reliable when substantial prior knowledge is available allowing one to concisely guess a narrow range of plausible values for all of the parameters in the model. Our expectation is that such knowledge is often lacking, motivating our use of default, heavy-tailed prior distributions, a strategy motivated by a desire for Bayesian robustness.

We modify their path sampling approach to allow use of our default PX-induced prior distributions. To do this we run our PX Gibbs sampler for each of the grid points and calculate the parameters under the inferential model simply using (3) and use those to estimate the log Bayes factors as given in (6). We will refer to our approach as path sampling with parameter expansion (PS-PX). Firstly PS-PX eliminates the need to use strongly informative prior distributions. Secondly, model selection based on path sampling is computationally quite intensive. Since PS-PX uses prior distributions with good mixing properties one needs to run the chains for considerably fewer iterations making the procedure much more efficient.

4.2 Simulation Study

Here we consider the same two sets of simulations as in Section 3, but now allowing the number of factors to be unknown. Let m denote the maximum number of factors in our list. For the simulation example considered in Section 3.1, we take m to be 3, which is also the maximum number of factors resulting in an identifiable model. We repeat the simulation for 100 simulated datasets and analyze them using PS-PX, taking 10 equi-spaced grid points in $[0, 1]$ for t. We use the same prior distributions for the parameters within each model as in Section 3.1. The correct model is chosen 100/100 times. We also calculate the BIC for all the models in our list for each dataset based on the maximum likelihood estimates of Λ and Σ. The BIC also chooses the correct model 100/100 times.

For the simulation example in Section 3.2 the true model has three factors and the maximum number of factors resulting in an identifiable model is 6. But here we take $m = 4$. We recommend focusing on lists that do not have large number of factors as sparseness is one of the main goals of factor models. Thus fitting models with as many factors as permitted given identifiability constraints goes against this motivation. We carry out the simulations exactly as in the previous simulation study. Here both PS-PX and the BIC choose the correct model 100/100 times again.

5. APPLICATIONS

5.1 Male Fertility Study

We first illustrate the effect of our parameter expanded Gibbs sampler on mixing when the number of factors is fixed. We have data from a reproductive epidemiology study. Here the underlying latent factor of interest is the latent sperm concentration of subjects. Concentration is defined as sperm count/semen volume. There are three outcome variables: concentration based on i. an automated sperm count system, ii. manual counting (technique 1) and iii. manual counting (technique 2) respectively. We consider the log-transformed concentration as the assumption of normality is more appropriate on the log scale. Here the maximum number of latent factors that results in an identifiable model is one. The model that we consider generalizes the factor analytic model to include covariates at the latent variable level, given as follows:

$$y_{ij} = \alpha_j + \lambda_j \eta_i + \epsilon_{ij},\ \eta_i = \boldsymbol{\beta}' \boldsymbol{x}_i + \delta_i \text{ where } \epsilon_{ij} \sim N(0, {\tau_j}^{-1}),\ \delta_i \sim N(0,1) \tag{7}$$

Following the usual convention we restrict the λ_j's to be positive for sign identifiability. We denote the covariates by $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})'$. There are altogether three study sites.

$$x_{ij} = \begin{cases} 1 & \text{if the sample is from site (j+1)} \\ 0 & \text{otherwise} \end{cases} \quad j = 1, 2. \quad x_{i3} = \begin{cases} 1 & \text{if the time since last ejaculation} \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

The PX model is as follows:

$$y_{ij} = {\alpha_j}^* + {\lambda_j}^* {\eta_i}^* + \epsilon_{ij},\ {\eta_i}^* = \mu^* + \boldsymbol{\beta^*}' \boldsymbol{x}_i + {\delta_i}^* \text{ where } \epsilon_{ij} \sim N(0, {\tau_j}^{-1}),\ \delta_i^* \sim N(0, \psi^{-1}) \tag{8}$$

We have introduced a redundant intercept μ^* in the second level of the model which improves the mixing tremendously. As noted by Gelfand et al. (1995), centering often leads to much better mixing. We relate the PX model parameters in (8) to those of the inferential model in (7) using the transformations

$\alpha_j = \alpha_j^* + \lambda_j^* \mu^*,\ \lambda_j = S(\lambda_j^*) \lambda_j^* \psi^{-1/2},\ \boldsymbol{\beta} = \boldsymbol{\beta^*} \psi^{1/2},\ \eta_i = S(\lambda_j^*) \psi^{1/2} (\eta_i^* - \mu^*),\ \delta_i = \psi^{1/2} \delta_i^*.$

To specify the prior distribution for the traditional Gibbs sampler we choose $\alpha_j \sim N(0,1),\ \lambda_j \sim$

$N_+(0,1)$, $\tau_j \sim \mathcal{G}(1,0.2)$ for $j = 1,2,3$, $\boldsymbol{\beta} \sim N(\mathbf{0}, 10 * I_3)$. For the PX Gibbs sampler we have $\alpha_j^* \sim N(0,1)$, $\lambda_j^* \sim N(0,1)$, $\tau_j \sim \mathcal{G}(1,0.2)$ for $j = 1,2,3$, $\mu^* \sim N(0,1)$, $\boldsymbol{\beta}^* \sim N(\mathbf{0}, 10 * I_3)$, $\psi \sim \mathcal{G}(1/2,1/2)$. We run both the samplers for 25,000 iterations excluding the first 5000 as burn-in, and then compare the performance based on convergence diagnostics such as trace plots and effective sample size (ESS).

It is evident from Figures (3) and (4) that the PX Gibbs sampler dramatically improves mixing. ESS.PX/ESS.Traditional for the upper triangular elements of $\Omega = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Sigma}$, starting from the first row and proceeding from left to right are $162.44, 184.22, 148.47, 143.83, 144.00, 69.98$. We use the Raftery and Lewis Diagnostic to estimate the number of MCMC samples needed for a small Monte Carlo error in estimating 95% credible intervals for the elements of Ω. The required sample size is different for different elements and also varies depending on whether we are trying to estimate the 0.025 or 0.975 quantiles for a particular element of Ω. If we take the maximum sample size over the two quantiles and all the elements then it turns out that the traditional Gibbs Sampler needs to be run for almost an hour whereas the PX Gibbs sampler will take only a little more than a minute to achieve the same accuracy. We also look at tools for inference like posterior means and 95% credible intervals for the different parameters in the model. On the basis of the 95% credible intervals there does not seem to be any significant effect of the covariates like study center and abstinence time on sperm concentration.

5.2 Rodent Organ Weight Study

We next illustrate the approach to model selection using parameter expansion through application to organ weight data from a U.S. National Toxicology Program (NTP) 13 week study of Anthraquinone in female Fischer rats. Studies are routinely conducted with 60 animals randomized to approximately six dose groups. At the end of the study, animals are sacrificed and a necropsy is conducted, with overall body weight obtained along with weights for the heart, liver, lungs, kidneys (combined) and thymus. Although body and organ weights are clearly correlated, a challenge in the analysis of these data is the dimensionality of the covariance matrix. In particular, even assuming a constant covariance across dose groups, it is still necessary to estimate $p(p+1)/2 = 21$ covariance parameters using data from only $n = 60$ animals. Hence, routine analyses rely on univariate approaches applied separately to body weight and the different organ weights.

Here alternatively we can use a factor model to reduce dimensionality. To determine the appropriate number of factors, we implemented the model selection approach described in Section 4 in the same manner as in the simulations. Body weights were normalized within each dose group prior to analysis for purposes of studying the correlation structure. Here the maximum possible number of factors was $m = 3$.

The estimated probabilities for the one, two and three factor models using PS-PX are 0.9209, 0.0714 and 0.0077 respectively. Here the BIC also chooses the one factor model. We also performed some sensitivity analysis by considering i) t prior distributions with 4 d.f. instead of the default Cauchy prior distributions for the factor loadings, ii) $\mathcal{G}(1, 0.4)$ prior distributions instead of $\mathcal{G}(1, 0.2)$ for the precision parameters and iii) both i) and ii) together. The estimated probabilities under i), ii) and iii) were {0.9412, 0.0544, 0.0044}, {0.9373 ,0.0578, 0.0048} and {0.9612, 0.0369, 0.0019} respectively. This suggests that our approach is not very sensitive to the prior distribution. The posterior means of factor loadings under the one factor model corresponding to body, heart, liver, lungs, kidneys and thymus are $0.88, 0.33, 0.52, 0.33, 0.70$ and 0.42 respectively. Body weight and kidney weight are the two outcomes having the highest correlation with the latent factor in the one factor analysis. These results suggest that one can bypass the need to rely on univariate analyses for this data, by using a sparse one factor model instead.

6. DISCUSSION

In analyzing high-dimensional, or even moderate-dimensional, multivariate data, one is faced with the problem of estimating a large number of covariance parameters. The factor model provides a convenient dimensionality-reduction technique. The routine use of normal and gamma prior distributions has proven to be a bottleneck for posterior computation in these models for the very slow convergence exhibited in the Markov chain.

In this article we have proposed a default heavy-tailed prior distribution for factor analytic models, using a parameter expansion approach. The posterior computation can be implemented easily using a Gibbs sampler. This prior distribution leads to a considerable improvement in mixing in the Gibbs chain. Extension of this prior distribution to more general settings with mixed outcomes (Song and Lee, 2007) or to factor regression models and Structural Equation models is straightforward. The computational gain in using the PX Gibbs sampler compared to the traditional one can be tremendous, especially if the outcomes are highly correlated. As evident from the Male Fertility Study data, one may need to run the traditional Gibbs sampler for almost an hour and the PX Gibbs sampler for less than two minutes to achieve the same level of accuracy in estimating 95% credible intervals. Hobert and Marchev (2008) show theoretical support for PX Gibbs samplers. Note that their work is philosophically different than our method, since their goal is to use PX simply to accelerate the MCMC mixing without modifying the prior distribution. We have also outlined a method based on path sampling using our default prior distribution, for computing posterior probabilities of models having different number of factors. Good performance of the method was demonstrated by using simulated data.

Matlab 7.5.0 was used for coding the PX Gibbs sampler and the PS-PX procedure. The package *coda* in R 2.6.1 was used to compute the convergence diagnostics. The data used in the analyses and the code can be downloaded from the JCGS website.

ACKNOWLEDGEMENTS

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences. The authors would like to thank the Editor Dr. David van Dyk, the Associate Editor and the referees for helpful comments.

REFERENCES

Arminger, G. (1998), " A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm," *Psychometrika*, 63, 271-300.

Berger, J., and Pericchi, L. (1996), "The intrinsic Bayes factor for model selection and prediction," *Journal of the American Statistical Association*, 91, 109-122.

Carvalho, C., Lucas, J., Wang, Q., Nevins, J., and West, M. (2008), "High-dimensional sparse factor modelling: Applications in Gene Expression Genomics," *Journal of the American Statistical Association*, to appear.

Gelfand, A.E., Sahu, S.K. and Carlin, B.P. (1995), "Efficient parameterisations for normal linear mixed models." *Biometrika*, 82, 479-488.

Gelman, A. (2004), "Parameterization and Bayesian modeling," *Journal of the American Statistical Association*, 99, 537-545.

Gelman, A. (2006), " Prior distributions for variance parameters in hierarchical models," *Bayesian Analysis* , 3, 515-534.

Gelman, A., and Meng, X.L. (1998), "Simulating normalizing constants: from importance sampling to bridge sampling to path sampling," *Statistical Science*, 13, 163-185.

Geweke, J. and Zhou, G. (1996), "Measuring the pricing error of the arbitrage pricing theory," *Review of Financial Studies*, 9, 557-587.

Hobert, J. P., and Marchev, D. (2008), "A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms," *Annals of Statistics*, to appear.

Imai, K., and van Dyk, D. A. (2005), "A Bayesian analysis of the multinomial probit model using marginal data augmentation," *Journal of Econometrics*, 124, 311-334.

Kinney, S., and Dunson, D.B. (2007), "Fixed and random effects selection in linear and logistic models," *Biometrics*, 63, 690-698.

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

Liu, C., Rubin, D.B., and Wu, Y.N. (1998), "Parameter expansion to accelerate EM: The PX-EM algorithm," *Biometrika*, 85, 755-770.

Liu, J.S., and Wu, Y.N. (1999), "Parameter expansion for data augmentation," *Journal of the American Statistical Association*, 94, 1264-1274.

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

Meng, X. L., and van Dyk, D. A. (1999), "Seeking efficient data augmentation schemes via conditional and marginal augmentation," *Biometrika*, 86, 301-320.

Natarajan, R., and McCulloch, C.E. (1998), "Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference?," *Journal of Computational and Graphical Statistics*, 7, 267-277.

Pournara, I., and Wernisch, L.(2007), "Factor analysis for gene regulatory networks and transcription factor activity profiles," *BMC Bioinformatics*, 8, 61.

Rowe, D.B. (1998), " Correlated Bayesian factor analysis," Ph.D. Thesis, Department of Statistics, University of California, Riverside, CA.

Robert, C. P., and Casella, G. (2004), "Monte Carlo Statistical Methods," Springer-Verlag .

Sanchez, B.N., Budtz-Jorgensen, E., Ryan, L.M., and Hu, H. (2005), "Structural equation models: A review with applications to environmental epidemiology," *Journal of the American Statistical Association*, 100, 1442-1455.

Song, X.Y., and Lee, S.Y. (2001), "Bayesian estimation and test for factor analysis model with continuous and polytomous data in several populations," *British Journal of Mathematical & Statistical Psychology*, 54, 237-263.

Song, X.Y., and Lee, S.Y. (2007), “Bayesian analysis of latent variable models with non-ignorable missing outcomes from exponential family,” *Statistics in Medicine*, 26, 681-693.

West, M. (2003), “Bayesian factor regression models in the large p, small n paradigm,” *Bayesian Statistics*, 7, J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M.West (eds). Oxford University Press.

APPENDIX: EFFECT OF PARAMETER EXPANSION

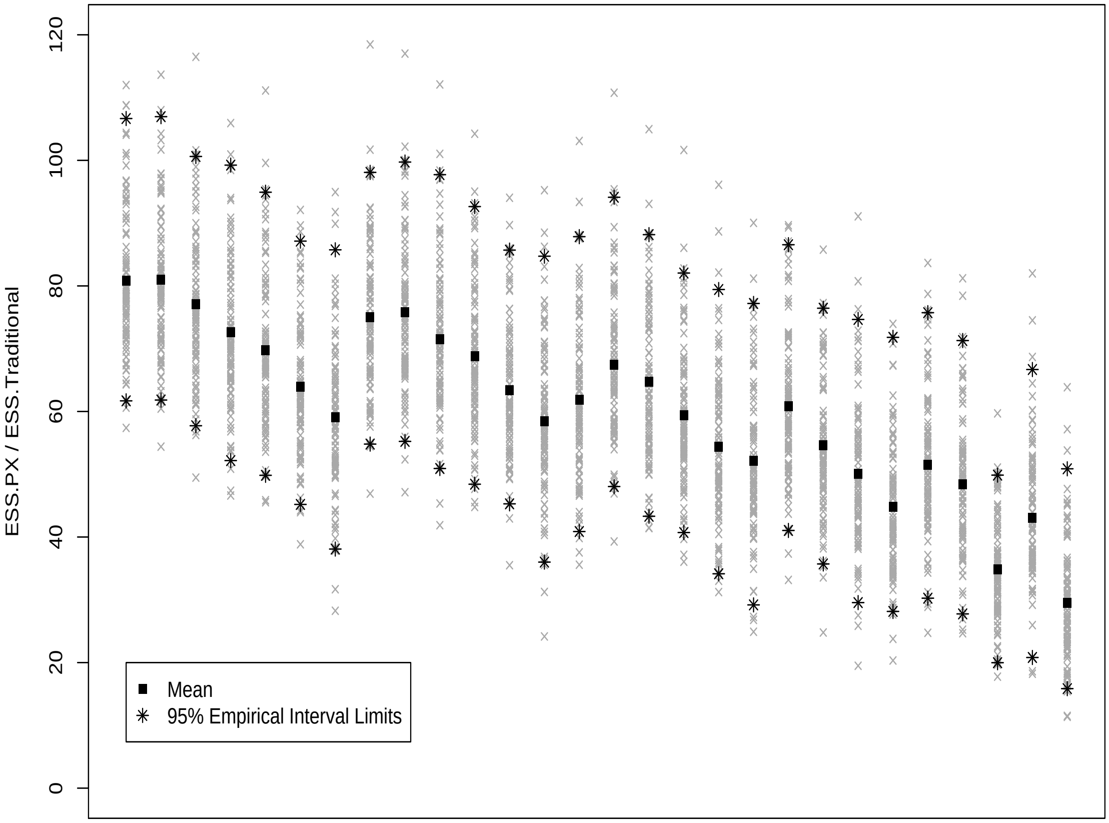

Figure 1: Comparison of the Parameter Expanded and Traditional Gibbs sampler based on the ratio of effective sample size for upper triangular elements of Ω, plotted rowwise from left to right, over 100 simulated datasets for Simulation 3.1

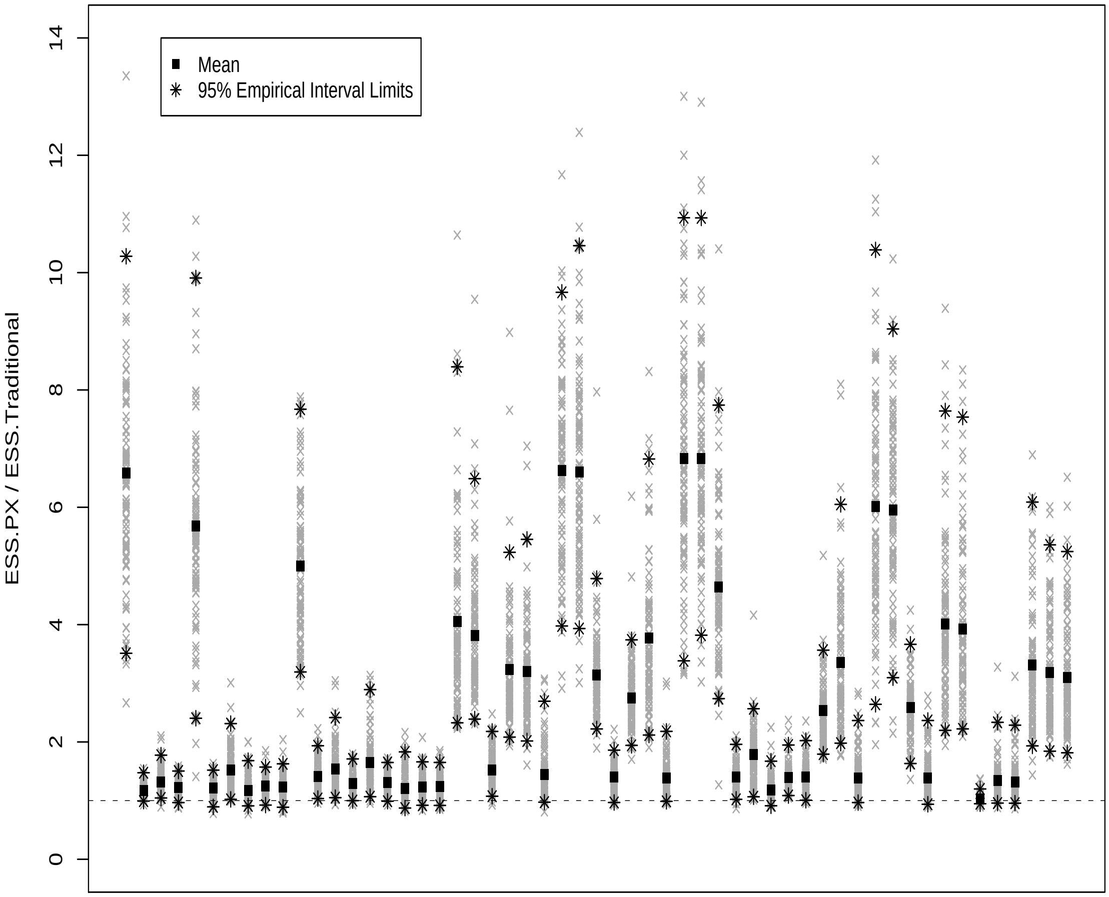

Figure 2: Comparison of the Parameter Expanded and Traditional Gibbs sampler based on the ratio of effective sample size for upper triangular elements of Ω, plotted rowwise from left to right, over 100 simulated datasets for Simulation 3.2

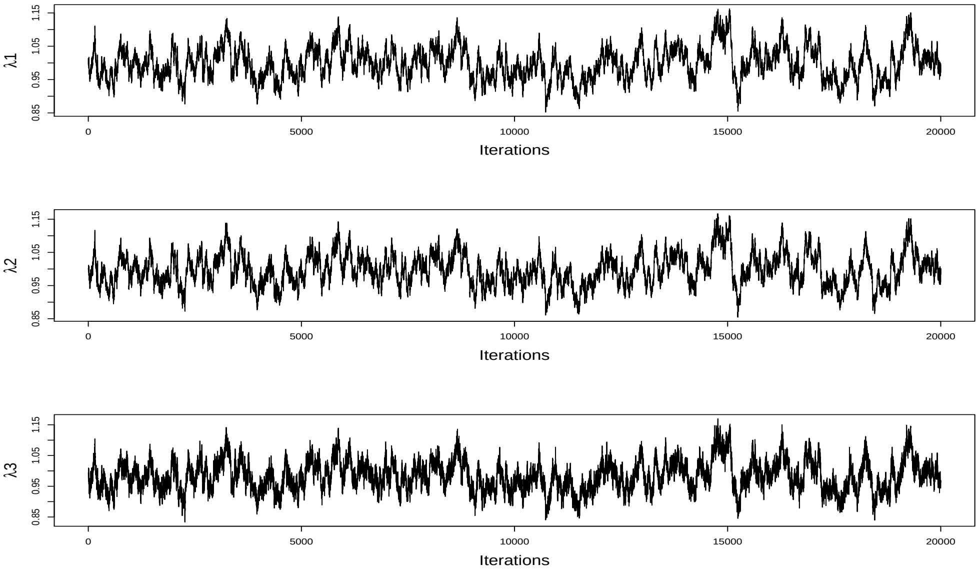

Figure 3: Trace plots of factor loadings exhibiting poor mixing using the Traditional Gibbs sampler in Application 5.1

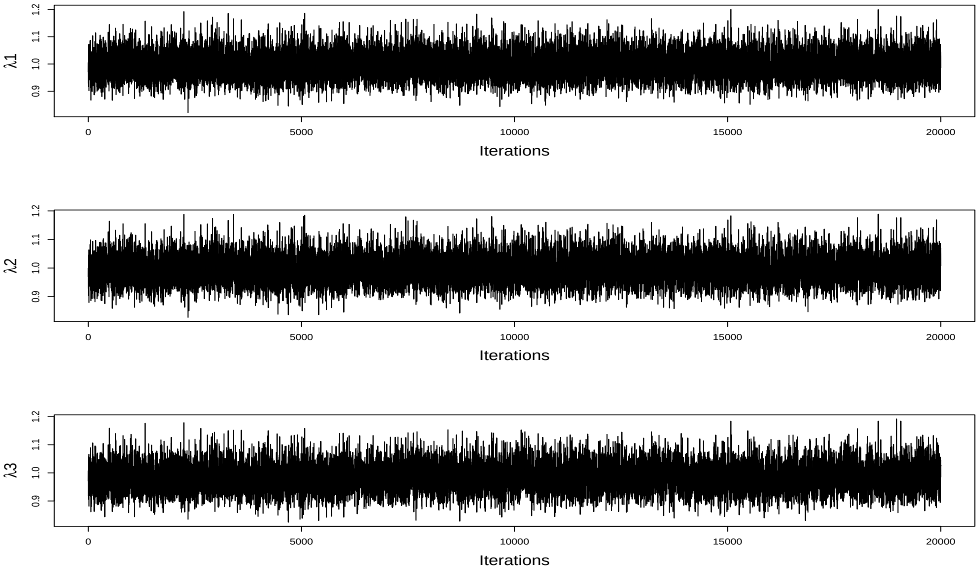

Figure 4: Trace plots of factor loadings exhibiting vastly improved mixing using the Parameter Expanded Gibbs sampler in Application 5.1