

Bayesian Model Selection in Factor Analytic Models

Joyee Ghosh*and David B. Dunson†

1. Introduction

Factor analytic models are widely used in social science applications to study latent traits, such as intelligence, creativity, stress and depression, that cannot be accurately measured with a single variable. In recent years, there has been a rise in the popularity of factor models due to their flexibility in characterizing multivariate data. 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 widely useful in epidemiologic studies involving complex health outcomes and exposures (Sanchez et al, 2005). Improvements in Bayesian computation permit the routine implementation of latent factor models via Markov chain Monte Carlo (MCMC) algorithms, and a very broad class of models can be fitted easily using the freely-available software package WinBUGS. The literature on methods for fitting and inferences in latent factor models is vast (for recent books, see Loehlin, 2004, and Thompson, 2004).

*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).

In using a factor analytic model for inferences on a covariance structure, it is appealing to formally account for uncertainty in selecting the number of factors. There has been some focus in the frequentist and Bayesian literature on the problem of selection of the number of factors. Press and Shigemasu (1999) propose to choose the number of factors having the highest posterior probability, noting that such an approach improves upon the commonly-used AIC (Akaike, 1987) and BIC (Schwarz, 1978) criteria. For hierarchical models, such as latent factor models, the BIC justification as an approximation to the Bayes factor breaks down (Berger, Ghosh and Mukhopadhyay, 2003), and one may need a different penalty for model complexity (Zhang and Kocka, 2004).

Estimation of posterior probabilities of models having different numbers of latent factors poses major challenges such as (1) how to choose priors for the factor loadings in the list of models corresponding to different numbers of factors; and (2) how to efficiently and accurately estimate posterior model probabilities. Polasek (1997) considered approaches for estimating posterior probabilities based on separate MCMC analyses of models differing only in the number of factors. Although an estimate of the marginal likelihood is not automatically available from the MCMC output, a number of algorithms have been proposed (Chib, 1995; DiCiccio et al., 1997; Gelfand and Dey, 1994; Meng and Wong, 1996).

Lopes and West (2004) proposed a reversible jump MCMC (RJMCMC) algorithm (Green, 1995) to move between models with different numbers of factors, and conducted a thorough comparison with estimators for approximating marginal likelihoods from separate MCMC analyses under each model. In simulation studies, they found that a number of the methods perform poorly relative to RJMCMC and bridge sampling (Meng and Wong, 1996) in terms of proportions of simulations in which the true number of factors is assigned highest posterior probability. A computational challenge in implementing RJMCMC for factor model selection is the difficulty of choosing efficient proposal distributions. Lopes and West (2004) address this problem by constructing proposals using the results of a preliminary MCMC run under

each model. Such an approach is highly computationally demanding, becoming infeasible as the sample size and potential number of factors increases. Motivated by this problem, Carvalho et al. (2008) proposed an evolutionary search algorithm, which provides a useful approach for searching for good factor models in high dimensions.

Lee and Song (2002) developed a method for estimating Bayes factors for selecting the number of factors in a factor analysis model using the idea of path sampling (Gelman and Meng, 1998). They proposed a procedure that is simple to implement and tends to have good performance in terms of accuracy. In order to estimate a single Bayes factor for comparing two competing models, their method requires running of separate MCMC algorithms along a grid corresponding to different values for a path sampling constant. Although such an approach can potentially be implemented in parallel, computational efficiency is nonetheless a concern, particularly if one has many different models under consideration. In addition, if the individual Markov chains exhibit poor mixing, which is a common problem in latent factor models, one may need to run each chain for a very large number of iterations to obtain accurate results.

An additional issue is that it is well known that Bayes factors are sensitive to the choice of prior. In latent factor models, it tends to be difficult to elicit the parameters, since in typical applications there is substantial uncertainty about the true values of the factor loadings and error variances *a priori*. Hence, one would often prefer to choose a vague prior. However, vague priors lead to problems with MCMC convergence, since in the limiting case as the prior variance becomes large one can obtain an improper posterior. In addition, even if this was not an issue, high variance priors tend to systematically favor small models, so that one will tend to select a one factor model if the prior variance is extremely large. In implementing their approach, Lee and Song (2002) used highly informative priors to avoid this problem. However, if their highly informative values are concentrated around the wrong values, as would typically be the case if there is substantial uncertainty *a priori*, then one

would expect poor performance in terms of model selection. For example, one may tend to discard an important factor inappropriately if the priors for the factor loadings have small variance around values far from the truth.

Ghosh and Dunson (2007) proposed an approach for simultaneously addressing the issues of efficient computation and prior specification in latent factor models through the use of a parameter-expansion approach. This approach conveys a dramatic improvement in MCMC efficiency in many cases, while inducing heavy-tailed priors that can be used for robust model selection via the path sampling approach. We have observed good performance of such an approach in a number of simulation studies.

In this chapter, we review the Ghosh and Dunson (2007) approach and propose a new approach for estimating posterior probabilities of models with different numbers of factors. The proposed approach relies on development of parameter expanded Gibbs samplers (Liu and Wu, 1999; Gelman et al., 2007, Ghosh and Dunson, 2007) for all models in our list. Based on the MCMC output we compute estimates for Bayes factors for models differing by one factor. Using a simple identity one can then compute posterior probabilities from these estimates. Clearly this method is computationally less demanding than some of the other methods as we need to run a single MCMC algorithm to estimate a particular Bayes factor. Additionally the use of parameter expansion facilitates good mixing. The method provides reasonably accurate results, based on simulation studies.

Section 2 defines the model when the number of factors is known. Section 3 describes the methodology used for estimating Bayes factors based on MCMC output when the number of factors is unknown. Section 4 presents the results of a simulation study. Section 5 contains an application to rodent organ weight data, and Section 6 discusses the results. The full conditional distributions for the parameter expanded Gibbs Sampler are given in the Appendix.

2. Specification of the Model

We shall first define a factor model when the number of factors is known to be k ,

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

where $\mathbf{\Lambda}$ is a $p \times k$ matrix of factor loadings, $\boldsymbol{\eta}_i = (\eta_{i1}, \dots, \eta_{ik})' \sim N_k(\mathbf{0}, \mathbf{I}_k)$ is a vector of standard normal latent factors, and $\boldsymbol{\epsilon}_i$ is a residual with diagonal covariance matrix $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$. Here the underlying latent factors, $\boldsymbol{\eta}_i$, induce dependence among the components of \mathbf{y}_i , because integrating out the latent factors we can write the marginal distribution of \mathbf{y}_i as $N_p(\mathbf{0}, \boldsymbol{\Omega})$, with $\boldsymbol{\Omega} = \mathbf{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Sigma}$. Thus this model implies that the sharing of common latent factors explains the dependence in the outcomes, and given the latent factors the outcome variables are uncorrelated. For example, the outcomes can be results from various blood tests of an individual and the underlying latent factor may be the health score for that individual. Usually the number of factors is small relative to the number of outcomes ($k \ll p$). This leads 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, especially in applications with moderate to large p .

The above factor model (1) without further constraints is not identifiable under orthogonal rotation. For example if we post-multiply $\mathbf{\Lambda}$ by an orthonormal matrix \mathbf{P} , where \mathbf{P} is such that $\mathbf{P}\mathbf{P}' = \mathbf{I}_k$, we will obtain exactly the same $\boldsymbol{\Omega}$ as in the previous factor model (1). To avoid this issue of non-identifiability, we impose some additional restrictions on the factor loadings matrix $\mathbf{\Lambda}$. We assume that $\mathbf{\Lambda}$ has a full-rank lower triangular structure to ensure identifiability. As the $k(k-1)/2$ elements in the upper triangular part of $\mathbf{\Lambda}$ are restricted to be zero, the number of free parameters in $\mathbf{\Lambda}$ and $\boldsymbol{\Sigma}$ is $q = p(k+1) - k(k-1)/2$. Here, k must be chosen so that $q \leq p(p+1)/2$. We use this restriction as a default, motivated by applications in which the latent factors do not have a pre-specified interpretation but are

included primarily to induce a sparse covariance structure in multivariate data. However, our approach can be easily modified to factor analysis applications in which the different measurements are designed as manifestations of latent traits of interest (e.g., intelligence, stress, etc).

To complete a Bayesian specification of model (1) we would need to specify the prior distributions for the free elements of $\mathbf{\Lambda}$ and Σ . A popular choice is truncated normal priors for the diagonal elements of $\mathbf{\Lambda}$, normal priors for the lower triangular elements, and inverse-gamma priors for $\sigma_1^2, \dots, \sigma_p^2$. 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). However, this prior specification suffers from a few major drawbacks. Firstly specification of the hyperparameters in the prior may be difficult. Prior elicitation is particularly important in this model, because in the limiting case as the prior variance for the normal and inverse-gamma components increases the posterior becomes improper. To address this problem often informative priors are chosen. In the absence of subject matter knowledge sometimes the hyperparameters in the prior are chosen after an initial analysis of the data. Using the data twice in this manner could lead to an underestimation of uncertainty in model selection. Secondly, even if informative priors are used, the Gibbs samplers tend to exhibit extreme slow-mixing.

Ghosh and Dunson (2007) address the above problems by generalizing the idea of Gelman (2006) to induce a new class of robust priors for the factor loadings. They use parameter expansion to induce a class of t or folded-t priors depending on sign constraints on the loadings. In absence of subject matter expertise they recommend using a Cauchy or half-Cauchy prior as a default. They also demonstrate good mixing properties of their parameter expanded Gibbs sampler compared to the traditional Gibbs sampler, using various simulated and real data examples. In this chapter we outline a method to select the number of factors using their prior.

3. Bayesian Uncertainty in the Number of Factors

To allow an unknown number of factors k , we choose a multinomial prior distribution, with $\Pr(k = h) = \kappa_h$, for $h = 1, \dots, m$. We then complete a Bayesian specification through priors on the coefficients within each of the models in the list $k \in \{1, \dots, m\}$. This is accomplished by choosing a prior for the coefficients in the m factor model having the form described in Ghosh and Dunson (2007). We first induce a prior on $\Lambda^{(m)}$ through parameter expansion. For any smaller model $k=h$, the prior for $\Lambda^{(h)}$ is obtained by marginalizing out the columns from $(h+1)$ to m . In this manner, we place a prior on the coefficients in the largest model, while inducing priors on the coefficients in each of the smaller models. We induce a prior on $\Lambda^{(m)}$ by defining the following parameter expanded (PX) factor model:

$$\mathbf{y}_i = \mathbf{\Lambda}^{*(m)} \boldsymbol{\eta}_i^* + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\eta}_i^* \sim N_k(\mathbf{0}, \boldsymbol{\Psi}), \quad \boldsymbol{\epsilon}_i \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}) \quad (2)$$

where $\mathbf{\Lambda}^{*(m)}$ is $p \times m$ working factor loadings matrix having a lower triangular structure without constraints on the elements, $\boldsymbol{\eta}_i^* = (\eta_{i1}^*, \dots, \eta_{im}^*)'$ is a vector of working latent variables, $\boldsymbol{\Psi} = \text{diag}(\psi_1, \dots, \psi_m)$, and $\boldsymbol{\Sigma}$ 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 $\mathbf{y}_i \sim N_p(\mathbf{0}, \mathbf{\Lambda}^{*(m)} \boldsymbol{\Psi} \mathbf{\Lambda}^{*(m)'} + \boldsymbol{\Sigma})$. Clearly, the diagonal elements of $\mathbf{\Lambda}^{*(m)}$ and $\boldsymbol{\Psi}$ are redundant.

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

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

where $\mathcal{S}(x) = -1$ for $x < 0$ and $\mathcal{S}(x) = 1$ for $x \geq 0$. Then, instead of specifying a prior for

$\Lambda^{(m)}$ directly, we induce a prior on $\Lambda^{(m)}$ through a prior for $\Lambda^{*(m)}$, Ψ . In particular, we let

$$\lambda_{jl}^{*(m)} \stackrel{iid}{\sim} \text{N}(0, 1), \quad j = 1, \dots, p, \quad l = 1, \dots, \min(j, m)$$

$$\lambda_{jl}^{(m)} \sim \delta_0, \quad j = 1, \dots, (m-1), \quad l = j+1, \dots, m, \quad \psi_l \stackrel{iid}{\sim} \mathcal{G}(a_l, b_l), \quad l = 1, \dots, m, \quad (4)$$

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 . This prior is conditionally-conjugate, leading to straightforward Gibbs sampling.

Bayesian selection of the number of factors relies on posterior model probabilities:

$$\Pr(k = h | \mathbf{y}) = \frac{\kappa_h \pi(\mathbf{y} | k = h)}{\sum_{l=1}^m \kappa_l \pi(\mathbf{y} | k = l)} \quad (5)$$

where the marginal likelihood under model k , $\pi(\mathbf{y} | k = h)$, is obtained by integrating the likelihood $\prod_i \text{N}_p(\mathbf{y}_i; \mathbf{0}, \Lambda^{(k)} \Lambda^{(k)'} + \Sigma)$ across the prior for the factor loadings $\Lambda^{(k)}$ and residual variances Σ . 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 $BF[h : j] = \{\pi(\mathbf{y} | k = h) / \pi(\mathbf{y} | k = j)\}$ as follows:

$$\Pr(k = h | \mathbf{y}) = \frac{O[h : j] * BF[h : j]}{\sum_{l=1}^m O[l : j] * BF[l : j]} \quad (6)$$

We choose $\kappa_h = 1/m$, which corresponds to a uniform prior for the number of factors k . To obtain an estimate of the Bayes factor $BF[(h-1) : h]$, for comparing models $k=(h-1)$ to $k=h$, we run the PX Gibbs sampler under $k = h$. Let $\{\boldsymbol{\theta}_i^{(h)}, i = 1, \dots, I\}$ denote the I MCMC samples from the PX Gibbs sampler under $k = h$, where $\boldsymbol{\theta}_i^{(h)} = (\Lambda_i^{(h)}, \Sigma_i)$. We can then estimate $BF[(h-1) : h]$ by the following estimator:

$$\widehat{BF}[(h-1) : h] = \frac{1}{I} \sum_{i=1}^I \frac{p(\mathbf{y}|\boldsymbol{\theta}_i^{(h)}, k = h-1)}{p(\mathbf{y}|\boldsymbol{\theta}_i^{(h)}, k = h)} \quad (7)$$

We do this for $h = 2, \dots, m$.

The above estimator is based on the following identity:

$$\begin{aligned} \int \frac{p(\mathbf{y}|\boldsymbol{\theta}^{(h)}, k = h-1)}{p(\mathbf{y}|\boldsymbol{\theta}^{(h)}, k = h)} p(\boldsymbol{\theta}^{(h)}|\mathbf{y}, k = h) d\boldsymbol{\theta}^{(h)} &= \int p(\mathbf{y}|\boldsymbol{\theta}^{(h)}, k = h-1) \frac{p(\boldsymbol{\theta}^{(h)})}{p(\mathbf{y}|k = h)} d\boldsymbol{\theta}^{(h)} \\ &= \frac{p(\mathbf{y}|k = h-1)}{p(\mathbf{y}|k = h)} \end{aligned} \quad (8)$$

We can obtain the Bayes factor for comparing any two models. For example, the Bayes factor for comparing the one factor and the m factor models is obtained as: $BF[1 : m] = BF[1 : 2] * BF[2 : 3] \dots BF[(m-1) : m]$. Using (6), we can estimate the posterior model probabilities in (5). We will refer to this approach as Importance Sampling with Parameter Expansion (IS-PX).

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 . For a detailed description of using path sampling for computing Bayes factors refer to the chapter *Bayesian Model Comparison of Structural Equation Models* in this book. Note that factor models are a special case of structural equation models.

Although their approach is promising in terms of accuracy, it is quite computationally intensive, requiring running of separate MCMC algorithms for each value of t in the grid. In contrast, using IS-PX we avoid the need to run multiple analyses, though it is not clear that this will necessarily improve efficiency, since we may require a long chain to obtain accurate

estimates of the Bayes factors. Given that the true Bayes factors are not available and are analytically intractable, our assessment of the performance of IS-PX will rely on simulations.

In addition to the difficulty of estimating the Bayes factor, an important challenge in Bayes model comparisons is sensitivity to the prior. It is well known that Bayes factors tend to be sensitive to the prior, motivating a rich literature on objective Bayes methods (Berger and Pericchi 1996; Berger and Pericchi 2001). Lee and Song (2002) rely on highly-informative priors 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. Such knowledge is often lacking. This motivated Ghosh and Dunson (2007) to modify the Lee and Song (2002) path sampling approach to allow the use of their default PX-induced priors. They refer to this as path sampling with parameter expansion (PS-PX).

4. Simulation Study

Here we consider two sets of simulation studies and compare our results with those from the PS-PX approach, as reported in Ghosh and Dunson (2007). Let m denote the maximum number of factors in our list. We routinely standardize the data prior to analysis.

4.1 One Factor Model

In the first simulation, $p = 7$, $n = 100$, the true number of factors, $k = 1$ and

$$\begin{aligned}\Lambda &= (0.995, 0.975, 0.949, 0.922, 0.894, 0.866, 0.837)' \\ \text{diag}(\Sigma) &= (0.01, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30).\end{aligned}$$

we take m to be 3, which is also the maximum number of factors resulting in an identifiable model. We repeat this simulation for 100 simulated data sets. To specify the prior for the PX Gibbs sampler, we induce half-Cauchy and Cauchy priors, both with scale parameter 1, for the diagonals and lower triangular elements of $\mathbf{\Lambda}$ respectively. For the residual precisions σ_j^{-2} we take $\mathcal{G}(1, 0.2)$ priors. 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 a burn-in. We note that for IS-PX we need only the samples from the grid $t = 1$ of the PS-PX approach. Here IS-PX chooses the correct model 92/100 times and PS-PX 100/100 times.

4.2 Three Factor Model

For the second simulation study, $p = 10$, $n = 100$, and the true number of factors, $k = 3$. This presents a more difficult scenario as some of the loadings are negative and there is more noise in the data compared to the previous one-factor model.

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

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

For this second simulation example the true model has three factors and the maximum number of factors resulting in an identifiable model is 6. We take $m = 4$ following Ghosh and Dunson (2007). We carry out the simulations exactly as in the previous simulation study. Here IS-PX chooses the correct model 86/100 times compared to 100/100 by PS-PX. In these simulations 10 gridpoints were considered for PS-PX, so given that IS-PX takes

only 1/10 of the run-time its performance is reasonably good. Running the MCMC much longer would improve its performance. A point to be noted here is that in the simulations where the wrong model is chosen, the bigger model with four factors is selected. For these datasets the parameter estimates under both three and four factor models are very similar, with the entries in the fourth column of the loadings matrix being close to zero. Hence even if the bigger model is chosen by IS-PX in some cases, we do not expect much deterioration in parameter estimates.

5. Application to Rodent Organ Weight Data

We illustrate our method for model selection by using it on a real dataset. We have organ weight data from a U.S. National Toxicology Program (NTP) 13 week study of Anthraquinone in female Fischer rats. The goal of such studies is to assess the short term toxicological effects of test agents on a variety of outcomes, including animal and organ body weights. Studies are routinely conducted with 60 animals randomized to approximately six dose groups, including a control. In the Anthraquinone study, doses included 0, 1875, 3750, 7500, 15000 and 30000 ppm.

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.

We can use a factor model here to reduce dimensionality, but it is not clear whether it is appropriate to assume a single factor underlying the different weights or if additional

factors need to be introduced. To address this question using the Anthraquinone data, we repeated the approach described in Section 3 in the same manner as implemented in the simulation examples. Body weights were standardized within each dose group prior to analysis for purposes of studying the correlation structure. Here we ran the algorithm for 100,000 iterations for IS-PX. The maximum possible number of factors was $m = 3$.

The estimated probabilities for the one, two and three factor models under IS-PX are 0.4417, 0.3464 and 0.2120 and using PS-PX are 0.9209, 0.0714 and 0.0077 respectively, as reported in Ghosh and Dunson (2007). The estimated factor loadings for the one and two factor models are presented below:

Table 1: Posterior Summaries Under the One Factor Model for Organ Weight Data

Weight	Parameter	Mean	95% CI
body	λ_1	0.88	[0.67,1.10]
heart	λ_2	0.33	[0.08,0.59]
liver	λ_3	0.52	[0.28,0.77]
lungs	λ_4	0.33	[0.08,0.59]
kidneys	λ_5	0.70	[0.48,0.94]
thymus	λ_6	0.42	[0.17,0.68]

Table 2: Posterior Summaries Under the Two Factor Model for Organ Weight Data

Weight	Parameter	Mean	95% CI	Parameter	Mean	95% CI
body	λ_{11}	0.87	[0.66,1.09]	λ_{12}	0	[0,0]
heart	λ_{21}	0.34	[0.08,0.61]	λ_{22}	0.24	[0.01,0.73]
liver	λ_{31}	0.52	[0.28,0.78]	λ_{32}	-0.15	[-0.63,0.45]
lungs	λ_{41}	0.35	[0.09,0.62]	λ_{42}	0.29	[-0.62,0.88]
kidneys	λ_{51}	0.70	[0.48,0.94]	λ_{52}	-0.06	[-0.44,0.29]
thymus	λ_{61}	0.42	[0.17,0.68]	λ_{62}	-0.14	[-0.66,0.49]

Body weight and kidney weight are the two outcomes having the highest correlation with the latent factor in the one factor analysis. The estimated covariance matrix is similar

under the one and two factor models, and the loadings on the second factor tend to be zero, suggesting that the second factor is not needed. This is also true for the third model. We also examined the likelihoods at the posterior mean of Ω . For the two and three factor models the increase in likelihood seems small compared to the number of additional parameters. In that sense the factor model with one factor seems most appropriate. Hence, the estimates of the posterior probabilities from PS-PX seem more realistic. However, given that we obtained similar estimates of the induced covariance matrix under the one, two and three factor models using the PX approach, it may be that there is little penalty to be paid for the additional complexity involved in fitting the multiple factor models given the regularization implicit in the Bayes approach. However, from the standpoint of ease in interpretation, the more parsimonious one factor model is certainly preferred.

6. Discussion

For analyzing high-dimensional, or even moderate-dimensional, multivariate data, the factor model provides a convenient tool for sparse modeling of the covariance matrix. This kind of data arises in a wide variety of applications ranging from genomics, where the outcomes may be highly correlated measurements on multiple genes, to epidemiology where the goal may be to study the effect of a collection of nutrients, many of which are highly correlated. Usually in such cases, the number of factors to be included in the model is unknown and leads to a challenging model uncertainty problem. A Bayesian solution to this problem proceeds by treating the number of factors as unknown and then assigning prior probabilities to each model in the list. After observing the data, the prior probabilities are updated to obtain posterior probabilities for each model.

In this chapter we have proposed an easy to implement, fully Bayesian approach for estimating the posterior model probabilities via Bayes factors. Model selection based on

fully Bayesian approaches tends to be computationally intensive. Gibbs samplers that exhibit slow-mixing add to the already existing heavy computational burden of model selection. This is why we choose a default heavy-tailed prior for factor loadings (Ghosh and Dunson, 2007), that greatly facilitates mixing. Using a simple identity we show our estimate is unbiased for estimating Bayes factors.

Comparing our method to the path sampling approach (Ghosh and Dunson, 2007; Song and Lee, 2002) based on simulated and real data, we find that it is relatively fast but less accurate. We have found that when it fails to choose the correct model, it usually prefers a model with more factors than the true model, and most of the loadings for the extra factors are close to zero. Since the loadings matrix for the true model can be thought to be nested within a larger factor model, with entries in the last columns equal to zero, we note that choosing a larger model in this case is more desirable than choosing a wrong model with less factors. To estimate the Bayes factor between two models that differ by one factor, our method currently uses the samples only from the larger model. Introducing a single bridge density between the two models may improve results substantially, while adding only negligible computational cost. This still needs further work and seems to be a promising direction for future research.

Acknowledgements

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

References

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

- 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.O., Ghosh, J.K. and Mukhopadhyay, N. (2003). Approximation and consistency of Bayes factors as model dimension grows. *Journal of Statistical Planning and Inference* **112**, 241-258.
- Berger, J., and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109-122.
- Berger, J., and Pericchi, L. (2001). Objective Bayesian methods for model selection: introduction and comparison [with discussion]. *Model Selection*, (P.Lahiri, editor), Institute of Mathematical Statistics Lecture Notes, Monograph Series volume 38, Beachwood Ohio, 135-207.
- 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.
- Chib, S. (1995). Marginal likelihoods from the Gibbs output. *Journal of the American Statistical Association* **90**, 1313-1321.
- DiCiccio, T.J., Kass, R., Raftery, A. and Wasserman, L. (1997). Computing Bayes factors by combining simulations and asymptotic approximations. *Journal of the American Statistical Association* **92**, 903-915.
- Gelfand, A.E. and Dey, D.K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society* **B**, 501-514.
- Gelfand, A.E., Sahu, S.K. and Carlin, B.P. (1995). Efficient parameterisations for normal linear mixed models. *Biometrika* **82**, 479-488.

- 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.
- Gelman, A., van Dyk, D., Huang, Z. and Boscardin, W.J. (2007) Using redundant parameters to fit hierarchical models. *Journal of Computational and Graphical Statistics*, to appear.
- Ghosh, J., and Dunson, D.B. (2007). Default priors and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, revision requested.
- Green, P.J. (1995). Reversible jump Markov chain Monte Carlo and Bayesian model determination. *Biometrika* **82**, 711-732.
- 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, J. and Wu, Y.N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association* **94**, 1264-1274.
- Loehlin, J.C. (2004). Latent Variable Models: An Introduction to Factor, Path and Structural Equation Analysis. *Lawrence Erlbaum Assoc Inc*.
- Lopes, H.F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14**, 41-67.
- Meng, X.L. and Wong, W.H. (1996). Simulating ratios of normalising constants via a simple identity. *Statistica Sinica* **11**, 552-586.

- Polasek, W. (1997). Factor analysis and outliers: a Bayesian approach. Discussion Paper, University of Basel.
- Press, S.J. and Shigemasu, K. (1999). A note on choosing the number of factors. *Communications in Statistics - Theory and Methods* **28**, 1653-1670.
- Rowe, D.B. (1998). Correlated Bayesian factor analysis. Ph.D. Thesis, Department of Statistics, University of California, Riverside, CA.
- 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.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461-464.
- 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.
- Thompson, B. (2004). Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications. APA Books.
- 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.
- Zhang, N.L. and Kocka, T. (2004). Effective dimensions of hierarchical latent class models. *Journal of Artificial Intelligence Research* **21**, 1-17.

Appendix: Full Conditional Distributions for the Gibbs Sampler

Suppose we have a model with k factors, conditional distributions for the PX Gibbs Sampler are presented below:

$$y_{ij} = \mathbf{z}'_{ij} \boldsymbol{\lambda}_j^* + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma_j^2),$$

where $\mathbf{z}_{ij} = (\eta_{i1}^*, \dots, \eta_{ik_j}^*)'$, $\boldsymbol{\lambda}_j^* = (\lambda_{j1}^*, \dots, \lambda_{jk_j}^*)'$ denotes the free elements of row j of $\boldsymbol{\Lambda}^*$, and $k_j = \min(j, k)$ is the number of free elements. Let $\pi(\boldsymbol{\lambda}_j^*) = N_{k_j}(\boldsymbol{\lambda}_{0j}^*, \boldsymbol{\Sigma}_{0j} \boldsymbol{\lambda}_j^*)$ denote the prior for $\boldsymbol{\lambda}_j^*$, the full conditional posterior distributions are as follows:

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

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

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

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