Biometrika (2011), **98**, 2, *pp.* 291–306
© 2011 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/asr013

Sparse Bayesian infinite factor models

By A. BHATTACHARYA and D. B. DUNSON

Department of Statistical Science, Duke University, Durham, North Carolina 27708-0251, U.S.A.

ab179@stat.duke.edu   dunson@stat.duke.edu

Summary

We focus on sparse modelling of high-dimensional covariance matrices using Bayesian latent factor models. We propose a multiplicative gamma process shrinkage prior on the factor loadings which allows introduction of infinitely many factors, with the loadings increasingly shrunk towards zero as the column index increases. We use our prior on a parameter-expanded loading matrix to avoid the order dependence typical in factor analysis models and develop an efficient Gibbs sampler that scales well as data dimensionality increases. The gain in efficiency is achieved by the joint conjugacy property of the proposed prior, which allows block updating of the loadings matrix. We propose an adaptive Gibbs sampler for automatically truncating the infinite loading matrix through selection of the number of important factors. Theoretical results are provided on the support of the prior and truncation approximation bounds. A fast algorithm is proposed to produce approximate Bayes estimates. Latent factor regression methods are developed for prediction and variable selection in applications with high-dimensional correlated predictors. Operating characteristics are assessed through simulation studies, and the approach is applied to predict survival times from gene expression data.

Some key words: Adaptive Gibbs sampling; Factor analysis; High-dimensional data; Multiplicative gamma process; Parameter expansion; Regularization; Shrinkage.

1. Introduction

Factor models aim to explain the dependence structure among high-dimensional observations through a sparse decomposition of a $p \times p$ covariance matrix Ω as $\Lambda\Lambda^{\mathrm{T}} + \Sigma$, where Λ is a $p \times k$ factor loadings matrix with $k \ll p$ and Σ is a $p \times p$ diagonal matrix with nonnegative diagonal entries. A popular approach to ensure identifiability of the loading elements is to constrain the loading matrix to be lower triangular with positive diagonal entries (Geweke & Zhou, 1996). Factor models have been traditionally applied in behavioural and social sciences, where the latent factors have a natural interpretation as certain unobserved psychological traits. A more recent approach (West, 2003; Carvalho et al., 2008) uses the above sparse characterization as a dimensionality reduction tool in large p and small n applications such as gene expression studies.

A Bayesian specification of the factor model (Arminger & Muthén, 1998; Song & Lee, 2001) commonly uses inverse gamma priors on the residual variances and normal and truncated normal priors on the off-diagonal and diagonal elements of the loadings matrix, respectively. Such choices lead to conditionally conjugate forms of the posterior distribution and enable posterior computation by a straightforward Gibbs sampler. However, it has been observed that these choices lead to poorly behaved Gibbs samplers with slow mixing when some of the outcomes are highly correlated. Posterior inference also tends to be sensitive to certain hyperparameters.

To address these issues, Ghosh & Dunson (2009) use parameter expansion (Liu & Wu, 1999; Gelman, 2006) to induce a heavy-tailed default prior distribution on the loading elements and propose an efficient Gibbs sampler.

Inference on the number of factors in factor analysis models is both conceptually and computationally challenging. Some of the early papers in this direction (discussion paper by Polasek, 1997, University of Basil) involve computation of the marginal likelihoods under models with different numbers of factors. Lopes & West (2004) proposed a reversible jump Markov chain Monte Carlo algorithm to allow for uncertainty in the number of factors. Lee & Song (2002) developed a path sampling approach instead. A more recent method infers the number of factors by zeroing a subset of the loading elements using Bayesian variable selection priors (Lucas et al., 2006; Carvalho et al., 2008); see also the 2009 discussion paper from the University of Chicago Booth School of Business by Schnatter and Lopes. Ando (2009) proposed an approach for calculating the exact marginal likelihood in Bayesian factor analysis with heavy-tailed priors. This method can be used for rapid estimation of the number of factors, but may be sensitive to subjectively chosen priors.

In this article we introduce a multiplicative gamma process shrinkage prior that allows introduction of infinitely many factors, with the loadings increasingly shrunk towards zero as the column index increases. The key to our approach lies in the fact that for purpose of prediction or inference on the covariance matrix, identifiability of the loadings is not necessary. In standard factor models, the identifiability constraints induce undesirable properties, such as a priori order dependence in the off-diagonal entries of the covariance matrix. Our proposed prior is placed on a parameter expanded factor loadings matrix, making the induced prior on the covariance matrix invariant to ordering of the data. The shrinkage prior allows us to adaptively select a truncation of the infinite loadings to one having finite columns, which facilitates the posterior computation while providing an accurate approximation to the infinite factor model.

2. BAYESIAN FACTOR MODELS

2·1. *Model and prior specification*

The generic form of a latent factor model is

$$y_i = \Lambda\eta_i + \epsilon_i, \quad \epsilon_i \sim N_p(0, \Sigma) \quad (i = 1, \ldots, n), \tag{1}$$

where y_i is a p-dimensional continuous response, Λ is a $p \times k$ factor loadings matrix, $\eta_i \sim N_k(0, I_k)$ are latent factors and ϵ_i is an idiosyncratic error with covariance $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. We follow standard practice in normalizing the data prior to analysis and hence do not include an intercept term in (1). Each observation y_i is assumed to have independent components given the factors and dependence among the components is induced by marginalizing over the distribution of the factors, so marginally $y_i \sim N_p(0, \Omega)$ with $\Omega = \Lambda\Lambda^{\mathrm{T}} + \Sigma$. In practical applications involving moderate to large p, the number of factors is typically much smaller than p, inducing a sparse characterization of the unknown covariance matrix Ω.

The above decomposition of Ω is not unique and there are actually infinitely many possibilities, since $\Lambda_1 = \Lambda P$ also satisfies the above condition for any semi-orthogonal matrix P $(PP^{\mathrm{T}} = I)$. The usual lower triangular constraint for identifiability (Geweke & Zhou, 1996) induces order dependence among the responses, with the choice of the first k response variables being an important modelling decision (Carvalho et al., 2008). From a Bayesian perspective, one does not require identifiability of the loading elements for a wide class of applications including covariance matrix estimation, variable selection and prediction. The above fact has been exploited in our approach to define the prior on a parameter-expanded loadings matrix with redundant parameters, resulting in better computational properties while simplifying the theory.

Letting Θ_Λ denote the collection of all matrices Λ with p rows and infinitely many columns such that $\Lambda\Lambda^{\mathrm{T}}$ is a $p \times p$ matrix with all entries finite, we have

$$\Theta_\Lambda = \left\{ \Lambda = (\lambda_{jh}),\ j = 1, \ldots, p,\ h = 1 \ldots, \infty,\ \max_{1 \leqslant j \leqslant p} \sum_{h=1}^{\infty} \lambda_{jh}^2 < \infty \right\}. \tag{2}$$

Using the Cauchy–Schwartz inequality, it is straightforward to show that all the entries of $\Lambda\Lambda^{\mathrm{T}}$ are finite if and only if the condition in (2) is satisfied. Let Θ_Σ denote the set of $p \times p$ diagonal matrices with nonnegative entries and let Θ denote all $p \times p$ positive semi-definite matrices. Consider the function $g\colon \Theta_\Lambda \times \Theta_\Sigma \to \Theta$ corresponding to $g(\Lambda, \Sigma) = \Lambda\Lambda^{\mathrm{T}} + \Sigma$.

Lemma 1. *For any* $(\Lambda, \Sigma) \in \Theta_\Lambda \times \Theta_\Sigma$, *we have* $g(\Lambda, \Sigma) \in \Theta$.

All proofs can be found in the Appendix. The image of $\Theta_\Lambda \times \Theta_\Sigma$ under g is the set $\{\Omega : \Omega = g(\Lambda, \Sigma),\ (\Lambda, \Sigma) \in \Theta_\Lambda \times \Theta_\Sigma\}$. Letting $g^{-1}(\Omega) \subset \Theta_\Lambda \times \Theta_\Sigma$ denote the pre-image of $\Omega \in \Theta$, it is straightforward to show that the set $g^{-1}(\Omega)$ contains at least one element for any $\Omega \in \Theta$, so that the image of $\Theta_\Lambda \times \Theta_\Sigma$ under g is the set Θ. For example, one element corresponds to $(\Lambda, 0_p)$, with $\Lambda = (\Omega^{1/2} : 0_{p\times\infty})$, $\Omega^{1/2}$ a Cholesky decomposition of Ω and 0_p denoting a $p \times p$ matrix of zeros. Thus g is a continuous surjective function. However, g is not bijective, and in general the cardinality of $g^{-1}(\Omega)$ is ∞. Lemma 2 states a regularity property of g, which is later used to prove sup-norm support of the proposed prior.

Lemma 2. *Let* (Λ_0, Σ_0) *be an arbitrary element of* $\Theta_\Lambda \times \Theta_\Sigma$. *For* $\epsilon > 0$, *define the following* ϵ*-ball around* (Λ_0, Σ_0), $B_\epsilon(\Lambda_0, \Sigma_0) = \{(\Lambda, \Sigma) \in \Theta_\Lambda \times \Theta_\Sigma : d_2(\Lambda, \Lambda_0) < \epsilon,\ d_\infty(\Sigma, \Sigma_0) < \epsilon\}$, *where* $d_2(\cdot, \cdot)$ *denotes the* L_2 *distance metric on* Θ_Λ,

$$d_2(\Lambda, \Lambda_0) = \left\{ \sum_{j=1}^{p} \sum_{h=1}^{\infty} (\lambda_{jh} - \lambda_{jh}^0)^2 \right\}^{1/2},$$

for $p \times \infty$ *matrices* $\Lambda = (\lambda_{jh})$, $\Lambda_0 = (\lambda_{jh}^0)$, *and* $d_\infty(A, B) = \max_{1\leqslant r,s\leqslant p} |a_{rs} - b_{rs}|$ *is the sup-norm metric for* $p \times p$ *matrices* $A = (a_{rs})$, $B = (b_{rs})$. *Then, the image* $g\{B_\epsilon(\Lambda_0, \Sigma_0)\}$ *contains values* $\Omega \in \Theta$ *in an* ϵ^* *sized ball in sup norm around* $\Omega_0 = g(\Lambda_0, \Sigma_0)$, *with* ϵ^* *decreasing towards zero monotonically as* ϵ *decreases to zero.*

Observe that d_2 is well defined and finite on Θ_Λ by (2).

We adopt a Bayesian approach and choose independent priors supported on $\Theta_\Lambda \times \Theta_\Sigma$, which in turn induces a prior on $\Omega \in \Theta$ through the operator g. We place the usual inverse gamma priors on the diagonal elements of Σ. To define a prior supported on Θ_Λ, we allow the entries of Λ to decrease in magnitude flexibly as the column index increases. The prior is defined on a parameter-expanded loading matrix without imposing any restriction on the loading elements. The introduction of the redundant parameters simplifies the theory and the induced prior has attractive properties including large support and order-independence. We use a shrinkage-type prior with the degree of shrinkage increasing across the column index as follows,

$$\lambda_{jh} \mid \phi_{jh}, \tau_h \sim N(0, \phi_{jh}^{-1}\tau_h^{-1}), \quad \phi_{jh} \sim \mathrm{Ga}(\nu/2, \nu/2), \ \tau_h = \prod_{l=1}^{h} \delta_l,$$

$$\delta_1 \sim \mathrm{Ga}(a_1, 1), \quad \delta_l \sim \mathrm{Ga}(a_2, 1), \quad l \geqslant 2, \quad \sigma_j^{-2} \sim \mathrm{Ga}(a_\sigma, b_\sigma) \quad (j = 1, \ldots, p), \qquad (3)$$

where δ_l $(l = 1, \ldots, \infty)$, are independent, τ_h is a global shrinkage parameter for the hth column and the ϕ_{jh}s are local shrinkage parameters for the elements in the hth column. The τ_hs are stochastically increasing under the restriction $a_2 > 1$, which favours more shrinkage as the column index increases. If we only use the global shrinkage parameter, the prior has a tendency to over-shrink the nonzero loadings. In gene expression examples involving large p, it is often the case that a relatively small proportion of genes are within each pathway. In such applications, we would like to shrink a subset of the elements strongly towards zero while retaining the sparse signals. We refer to the induced prior on the space of covariance matrices as a multiplicative gamma process shrinkage prior.

2·2. *Properties of the shrinkage prior*

Let $\Pi_\Lambda \otimes \Pi_\Sigma$ denote the prior on (Λ, Σ) defined in (3). We first need to make sure that our prior is well defined so that draws from the above prior are elements of $\Theta_\Lambda \times \Theta_\Sigma$ almost surely.

PROPOSITION 1. *If* $(\Lambda, \Sigma) \sim \Pi_\Lambda \otimes \Pi_\Sigma$, *then* $\Pi_\Lambda \otimes \Pi_\Sigma(\Theta_\Lambda \times \Theta_\Sigma) = 1$.

For computational purposes, we would like to approximate the infinite loadings matrix with a finite matrix having few columns relative to the number of outcomes p. As justification, we obtain theoretical bounds on the truncation approximation error. Let $(\Lambda, \Sigma) \sim \Pi_\Lambda \otimes \Pi_\Sigma$ and $\Omega = \Lambda\Lambda^{\mathrm{T}} + \Sigma$ be the induced covariance matrix. We can approximate Ω by $\Omega_H = \Lambda_H\Lambda_H^{\mathrm{T}} + \Sigma$ where Λ_H denotes the matrix obtained by setting the columns of Λ from $H + 1$ onwards to zero or equivalently discarding those higher indexed columns. The following theorem states that the prior probability of Ω_H being arbitrarily close to Ω in an appropriate sense converges exponentially fast to 1 as H tends to ∞.

THEOREM 1. *If* $a_2 > 2$, *then for any* $\epsilon > 0$,

$$\mathrm{pr}\{d_\infty(\Omega, \Omega_H) > \epsilon\} < \frac{6pb}{\epsilon(1-a)} a^H,$$

for $H > \log\{6pb/\epsilon(1-a)\}/\log(1/a)$, *where* $b = E(\delta_1^{-1})$ *and* $a = E(\delta_2^{-1})$, *with* δ_1 *and* δ_2 *as in (3).*

The proof of the theorem assumes $\nu = 3$ which has been used as a default choice throughout, but the same argument holds for any $\nu > 2$. Although the condition $a_2 > 2$ is sufficient to ensure that $a < 1$, for any $\mathrm{Ga}(a_2, b_2)$ prior on δ_2, the theorem remains valid as long as $E(\delta_2^{-1}) = b_2/(a_2 - 1) < 1$ or $a_2 > 1 + b_2$.

Letting Π denote the induced prior on Θ, $\Pi = (\Pi_\Lambda \otimes \Pi_\Sigma) \circ g^{-1}$ so that for any Borel subset A of Θ, $\Pi(A) = (\Pi_\Lambda \otimes \Pi_\Sigma)\{g^{-1}(A)\}$. Since g is a continuous and hence measurable map, Π is a well-defined probability measure on $(\Theta, \mathcal{A})$, with $\mathcal{A}$ the Borel σ-algebra of subsets of Θ.

PROPOSITION 2. *If* Ω_0 *is any* $p \times p$ *covariance matrix and* $B_\epsilon^\infty(\Omega_0)$ *is an* ϵ*-neighbourhood of* Ω_0 *under the sup-norm, then* $\Pi\{B_\epsilon^\infty(\Omega_0)\} > 0$ *for any* $\epsilon > 0$.

Proposition 2 shows that our proposed prior has large support, so places positive probability in arbitrarily small neighbourhoods around any covariance matrix. We use Proposition 2 to show weak consistency of the posterior distribution of Ω in Theorem 2. Denote $K(\Omega_0, \Omega)$ to be the Kullback–Leibler divergence between $N_p(0, \Omega_0)$ and $N_p(0, \Omega)$,

$$K(\Omega_0, \Omega) = \int \log \frac{N(y; 0, \Omega_0)}{N(y; 0, \Omega)} N(y; 0, \Omega_0)\, dy.$$

THEOREM 2. *Fix $\Omega_0 \in \Theta$. For any $\epsilon > 0$, there exists $\epsilon^* > 0$, such that*

$$\{\Omega : d_\infty(\Omega_0, \Omega) < \epsilon^*\} \subset \{\Omega : K(\Omega_0, \Omega) < \epsilon\},$$

which implies that the posterior distribution of Ω is weakly consistent.

The weak consistency of the posterior follows from the Schwartz (1965) theorem, since any Kullback–Leibler neighbourhood of the true density has positive probability using Proposition 2.

Another attractive property of our prior is that it is free of order dependence, so that the induced prior on Ω is invariant to permutations with Ω having the same distribution as Ω_π, where $\Omega_\pi = (w_{\pi_r \pi_s})$ with π any permutation of $\{1, \ldots, p\}$ and $\Omega = (w_{rs})$. We have $w_{rs} = \sum_{h=1}^{\infty} \lambda_{rh}\lambda_{sh} = \lambda_r^{\mathrm{T}}\lambda_s$, where $\lambda_j = (\lambda_{j1}, \lambda_{j2}, \ldots)^{\mathrm{T}}$. Conditionally on $\tau = (\tau_1, \tau_2, \ldots)^{\mathrm{T}}$, the λ_{rh}'s are independent with $\lambda_{rh} \mid \tau_h \sim t_3(0, \tau_h^{-1})$. Since the marginal prior on λ_r is the same for every r, w_{rs} has the same distribution as $w_{r's'}$ for any $(r, s) \neq (r', s')$ such that $r \neq s$, $r' \neq s'$. The permutation invariance follows from the fact that w_{rr} and $w_{r'r'}$ have the same distribution for any $1 \leqslant r, r' \leqslant p$.

Although the distribution of w_{rs} does not have a simple form, the first two moments of w_{rs} can be obtained as

$$\begin{aligned}
E\left(w_{rs}\right) &= \sum_{h=1}^{\infty} E\left\{E\left(\lambda_{rh}\lambda_{sh} \mid \tau_h\right)\right\} = 0,\\
E\left(w_{rs}^2\right) &= E\left\{\mathrm{tr}(\lambda_r^{\mathrm{T}}\lambda_s\lambda_s^{\mathrm{T}}\lambda_r)\right\} = \mathrm{tr}\left\{E\left(\lambda_r\lambda_r^{\mathrm{T}}\lambda_s\lambda_s^{\mathrm{T}}\right)\right\}\\
&= \mathrm{tr}\left[E\left\{E(\lambda_r\lambda_r^{\mathrm{T}} \mid \tau)E(\lambda_s\lambda_s^{\mathrm{T}} \mid \tau)\right\}\right] = 9\sum_{h=1}^{\infty} E(\tau_h^{-2}).
\end{aligned}$$

Thus $E\left(w_{rs}^2\right)$ is finite if $d = E(\delta_1^{-2})$ is finite and $c = E(\delta_2^{-2}) < 1$ and in that case $E\left(w_{rs}^2\right) = 9d/(1-c)$. One way to ensure the above conditions is to let $a_1 > 2$ and $a_2 > 3$. Hence the induced prior on any of the off-diagonal entries of Ω has mean zero and the parameters a_1, a_2 dictate the existence of higher order moments. We place gamma priors on a_1 and a_2 to learn these key hyperparameters from the data.

3. POSTERIOR COMPUTATION

3·1. *Gibbs sampler with a fixed truncation level*

We propose a straightforward Gibbs sampler for posterior computation after truncating the loadings matrix to have $k^* \ll p$ columns. An adaptive strategy for inference on the truncation level k^* is described in § 3·2. The Gibbs sampler is computationally efficient and mixes rapidly as the shrinkage prior allows block updating of the loadings. The sampler cycles through the following steps.

Step 1. If we denote the jth row of Λ_{k^*} by λ_j^{T} , then the λ_js have independent conditionally conjugate posteriors,

$$\pi(\lambda_j \mid -) \sim N_{k^*}\left\{ \left(D_j^{-1} + \sigma_j^{-2}\eta^{\mathrm{T}}\eta\right)^{-1}\eta^{\mathrm{T}}\sigma_j^{-2}y^{(j)}, \left(D_j^{-1} + \sigma_j^{-2}\eta^{\mathrm{T}}\eta\right)^{-1} \right\},$$

where $\eta = (\eta_1, \ldots, \eta_n)^{\mathrm{T}}$, $D_j^{-1} = \mathrm{diag}(\phi_{j1}\tau_1, \ldots, \phi_{jk^*}\tau_{k^*})$ and $y^{(j)} = (y_{1j}, \ldots, y_{nj})^{\mathrm{T}}$ for $j = 1, \ldots, p$. Given the other parameters, $\pi(\lambda_j \mid -)$ denotes the conditional posterior of λ_j.

Step 2. Sample σ_j^{-2}, $j = 1 \ldots, p$, from conditionally independent posteriors

$$\pi(\sigma_j^{-2} \mid -) \sim \mathrm{Ga}\left\{ a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2}\sum_{i=1}^{n}(y_{ij} - \lambda_j^{\mathrm{T}}\eta_i)^2 \right\}.$$

Step 3. Sample η_i, $i = 1 \ldots, n$, from conditionally independent posteriors

$$\pi(\eta_i \mid -) \sim N_{k^*}\left\{ \left(I_{k^*} + \Lambda_{k^*}^{\mathrm{T}}\Sigma^{-1}\Lambda_{k^*}\right)^{-1}\Lambda_{k^*}^{\mathrm{T}}\Sigma^{-1}y_i, \left(I_{k^*} + \Lambda_{k^*}^{\mathrm{T}}\Sigma^{-1}\Lambda_{k^*}\right)^{-1} \right\}.$$

Step 4. Sample ϕ_{jh} from

$$\pi(\phi_{jh} \mid -) \sim \mathrm{Ga}\left(\frac{\nu + 1}{2}, \frac{\nu + \tau_h\lambda_{jh}^2}{2} \right).$$

Step 5. Sample δ_1 from

$$\pi(\delta_1 \mid -) \sim \mathrm{Ga}\left\{ a_1 + \frac{pk^*}{2}, 1 + \frac{1}{2}\sum_{l=1}^{k^*}\tau_l^{(1)}\sum_{j=1}^{p}\phi_{jl}\lambda_{jl}^2 \right\},$$

and for $h \geqslant 2$, sample δ_h from

$$\pi(\delta_h \mid -) \sim \mathrm{Ga}\left\{ a_2 + \frac{p}{2}(k^* - h + 1), 1 + \frac{1}{2}\sum_{l=h}^{k^*}\tau_l^{(h)}\sum_{j=1}^{p}\phi_{jl}\lambda_{jl}^2 \right\},$$

where $\tau_l^{(h)} = \prod_{t=1, t \neq h}^{l}\delta_t$ for $h = 1, \ldots, k^*$.

Step 6. Update a_1 and a_2 using a Metropolis–Hastings step within the Gibbs sampler.

3·2. *Choosing the number of factors adaptively*

In practical situations, we expect to have relatively few important factors compared with the dimension p of the outcomes. Our proposed model with infinite number of factors obviates the need for pre-specifying the number of factors, while the sparsity favouring prior on the loadings ensures that the effective number of factors would be small when the truth is sparse. However, we need a computational strategy for choosing an appropriate level of truncation k^*. We would like to strike a balance between missing important factors by choosing k^* too small and wasting computation on an overly conservative truncation level. One can think of k^* as the effective number of factors, so that the contribution from adding additional factors is negligible. Starting with a conservative guess $\tilde{k}$ of k^*, the posterior samples of $\Lambda_{\tilde{k}}$ from the Gibbs sampler mentioned in § 3·1 contain information about the effective number of factors. At the tth iteration of the Gibbs sampler, let $m^{(t)}$ denote the number of columns in $\Lambda_{\tilde{k}}$ having all elements in a pre-specified small neighbourhood of zero. Intuitively, $m^{(t)}$ of the factors have a negligible contribution at the tth iteration. Usual shrinkage priors on the loadings exhibit the phenomenon of factor splitting, in which none of the columns have all loadings close to zero even when $\tilde{k}$ is chosen to be greater than the true number of factors. By shrinking increasingly in later columns, we avoid this problem. We define $k^{*(t)} = \tilde{k} - m^{(t)}$ to be the effective number of factors at iteration t.

The above approach has been shown to produce accurate estimates of the true effective number of factors k^* in a number of simulation examples as long as $\tilde{k} \geqslant k^*$. However, in order to be assured that $\tilde{k} \geqslant k^*$, it is typically necessary to choose a very conservative bound in large p applications, which leads to wasted computational effort. Ideally, we would like to discard the redundant factors and continue the sampler with a reduced number of columns in the loadings. With this aim, we modify our sampler described above to an adaptive Gibbs sampler, which tunes the number of factors as the sampler progresses. The adaptations are designed to satisfy the diminishing adaptation condition in Theorem 5 of Roberts & Rosenthal (2007). To be specific, we adapt with probability $p(t) = \exp(\alpha_0 + \alpha_1 t)$ at the tth iteration, with α_0, α_1 chosen so that adaptation occurs around every 10 iterations at the beginning of the chain but decreases in frequency exponentially fast. We generate a sequence u_t of uniform random numbers between 0 and 1. At the tth iteration, if $u_t \leqslant p(t)$, we monitor the columns in the loadings having all elements within some pre-specified small neighbourhood of zero. If the number of such columns drops to zero, we add a column to the loadings and otherwise discard the redundant columns. The other parameters are also modified accordingly. When we add a factor, we sample parameters from the prior distribution to fill in additional columns, and otherwise retain parameters corresponding to the nonredundant columns.

The most common approach for selecting the number of factors relies on fitting the factor model for different choices of k^*, and then using the BIC or another criteria for selection. This approach can be difficult to implement for large p, small n problems in which maximum likelihood estimates often do not exist, and the BIC is not well justified for factor models even for small to moderate p. Lopes & West (2004) compared a number of alternatives, recommending a reversible jump Markov chain Monte Carlo approach that requires a preliminary run for each choice of the number of factors, so it is very computationally intensive. Path sampling faces similar computational hurdles in scaling up to large p. Stochastic search variable selection algorithms have been applied in large p settings, but performance is questionable given the need to update elements of the loadings matrix one at a time, leading to very slow mixing and convergence rates. In the econometrics literature on approximate factor models, there has been recent work (Bai & Ng, 2002; Amengual & Watson, 2007) on consistent estimation of the number of static and dynamic factors as the number of time series and observation times both increase to ∞ at a comparable rate; see also the discussion paper by Onatski (2005), University of Columbia.

A significant advantage of our adaptive method is that a single run provides posterior samples of the parameters as well as information about the number of factors, with convergence of the chain guaranteed by the theory in Roberts & Rosenthal (2007). In addition, we save computation by discarding the unimportant factors. Letting $\tilde{k}^{(t)}$ denote the truncation level at iteration t and $k^{*(t)} = \tilde{k}^{(t)} - m^{(t)}$ denote the effective number of factors, we use the median or mode of $\{k^{*(t)}\}$ after burn-in as an estimate of k^* with credible intervals quantifying uncertainty.

After a reasonable burn-in, $\Omega^{(t)} = \Lambda^{(t)}_{\tilde{k}^{(t)}} \Lambda^{(t)\,\mathrm{T}}_{\tilde{k}^{(t)}} + \Sigma^{(t)}$ represent draws from the approximated marginal posterior distribution of Ω given $y_1, \ldots, y_n$, where $\{\Lambda^{(t)}_{\tilde{k}^{(t)}}, \Sigma^{(t)}\}$ denote posterior samples at the tth iteration. The posterior samples $\Omega^{(t)}$ can be used for inference on Ω. We also propose a fast algorithm for calculating an approximate maximum a posteriori estimate of the covariance matrix. The proposed approach is useful to arrive at a quick working estimate of the covariance matrix. Our proposed stochastic EM approach, (Celeux et al., 1996) approach replaces draws from the conditional posterior distributions of $\Lambda_{\tilde{k}(t)}$, Σ and ϕ in Steps 1, 2 and 4 above by the respective conditional posterior modes.

4. Simulation example

4·1. *Factor selection and covariance matrix estimation*

We consider a number of simulation examples to illustrate our approach and compare with competing methods. We simulated $y_i, i = 1, \ldots, 200$, from a p-dimensional normal distribution with zero-mean and covariance matrix $\Omega = \Lambda\Lambda^{\mathrm{T}} + \Sigma$, where Λ is a $p \times k$ matrix and Σ is a $p \times p$ diagonal matrix. The diagonal elements of Σ^{-1} are drawn independently from a Ga(1,0·25) distribution with mean 4. The number of non-zero elements in each column of Λ are chosen linearly between $2k$ and $k + 1$ in a decreasing fashion. We randomly allocate the location of the zeros in each column and simulate the nonzero elements independently from a normal distribution with mean 0 and variance 9.

We choose three (p, k) combinations with moderate to large p, namely (100, 5), (500, 10) and (1000, 15). For each pair we consider 50 simulation replicates. We run the adaptive Gibbs sampler for 25 000 iterations with a burn-in of 5000, and collect every 5th sample to thin the chain. We use a default choice of $5 \log(p)$ as the starting number of factors. The hyperparameters a_σ and b_σ for σ_j^{-2} in (3) are 1 and 0·3 respectively, while ν is 3. We place Ga(2, 1) priors on a_1 and a_2. We choose α_0 and α_1 in the adaptation probability $p(t)$ as -1 and -5×10^{-4} respectively. We monitor the columns in the loadings having all elements less than 10^{-4} in magnitude and proceed by adapting the number of factors as in § 3·2. For the stochastic EM algorithm, we choose a burn-in of 100 and monitored the estimated covariance matrix every 10 iterations. We stop the chain when the sup-norm distance between the estimated covariance matrix at the current iteration was within a small tolerance level compared with the estimate 10 iterations before.

The average of the estimated number of factors across the replicates is 6·82, 10·00 and 14·40 corresponding to $k = 5$, 10 and 15 with empirical 95% intervals for the number of factors (5, 8), (9, 11) and (13, 16), respectively. The estimated covariance matrix in each case is close to the true value, with small mean square error, average and maximum absolute bias. We compare the estimation of the covariance matrix to a recent method by Bickel & Levina (2008) which bands the sample covariance matrix and proposes a resampling scheme for choosing the optimal banding parameter. The stochastic EM algorithm was also used to arrive at an approximate maximum a posteriori estimate of the covariance matrix. We provide the summaries of the mean square error, average absolute bias and maximum absolute bias for the three methods across the replicates in Table 1. Based on Table 1, the proposed shrinkage approach does significantly better than the Bickel & Levina (2008) method. The stochastic EM algorithm also performs well, especially for smaller values of p.

4·2. *Latent factor regression*

It is common in many application areas to have a massive-dimensional vector of candidate predictors, with many of the predictors being moderately to highly correlated. Modifications using penalized least squares methods have been studied extensively. The lasso (Tibshirani, 1996) and the elastic net (Zou & Hastie, 2005) are two of the most popular such methods. In order to select correlated batches of predictors simultaneously, one can potentially use Bayesian latent factor regression (Lucas et al., 2006; Carvalho et al., 2008).

Table 1. *Comparative performance in covariance matrix estimation in the simulation study. The average, best and worst case performance across* 50 *simulation replicates in terms of mean square error* ($\times 10^2$)*, average absolute bias* ($\times 10^2$) *and maximum absolute bias* ($\times 10^2$) *are tabulated for the different methods*

true (p, k)	(100, 5)			(500, 10)			(1000, 15)		
method	MGPS	Banding	MAP	MGPS	Banding	MAP	MGPS	Banding	MAP
MSE									
mean	0·2	1·3	0·2	0·10	0·4	0·10	0·10	0·3	0·10
min	0·1	0·9	0·1	0·02	0·4	0·05	0·02	0·2	0·05
max	0·3	1·6	0·3	0·20	0·5	0·30	0·4	0·5	0·30
average absolute bias									
mean	1·9	3·1	1·0	0·6	0·6	0·3	0·4	0·5	0·3
min	1·3	2·5	0·6	0·4	0·6	0·2	0·2	0·4	0·2
max	2·5	4·9	1·5	0·9	0·9	0·5	0·6	0·5	0·5
maximum absolute bias									
mean	50·9	111·0	44·8	95·4	117·8	97·7	115·0	115	108·0
min	38·8	99·8	24·7	50·2	105·0	64·4	52·6	111	74·7
max	74·1	131·0	105·0	152·0	131·0	162·0	242·0	240	221·0

MGPS, posterior mean using our proposed multiplicative shrinkage prior; Banding, Banding sample covariance matrix; MAP, approximate maximum a posteriori estimate under our proposed prior; MSE, mean square error.

Let $y_i = (z_i, x_i^{\mathrm{T}})^{\mathrm{T}}$, $i = 1, \ldots, n$, where the x_is are $(p-1)$-dimensional predictors and z_is are the response. For ease of illustration, we assume the z_is to be univariate, though extensions to multivariate cases are straightforward. Also assume that the predictors and response are all continuous. We can use standard data augmentation procedures otherwise. We jointly model the y_is as in (1). Our objective is to predict the response z_{n+1} for a future subject based on the predictors x_{n+1} for that subject and $y_1, \ldots, y_n$. The posterior predictive distribution of $z_{n+1} \mid x_{n+1}, y_1, \ldots, y_n$ is

$$f(z_{n+1} \mid x_{n+1}, y_1, \ldots, y_n) = \int f(z_{n+1} \mid x_{n+1}, \Omega)\, \pi(\Omega \mid y_1, \ldots, y_n)\, \mathrm{d}\Omega.$$

For the simulation examples described in § 4·1, let $z_i = y_{i1}$ and $x_i = (y_{i2}, \ldots, y_{ip})^{\mathrm{T}}$. We randomly selected two locations in the first row of Λ and assigned values 1 and -1 to those locations, with the other elements in the first row set to zero. The remaining rows of the loadings were simulated as mentioned before. We used a randomly chosen training set of size 100 and held out the z_is for the remaining 100 samples. The coverage of 95% predictive intervals averaged across the replicates were 0·95, 0·94 and 0·95, respectively. Table 2 compares the predictive performance with lasso and elastic net. The proposed approach does similar to lasso and elastic net, but has the advantage of quantifying predictive uncertainty.

The joint Gaussian model implies that $E(z_i \mid x_i) = x_i^{\mathrm{T}} \beta$, with $\beta = \Omega_{xx}^{-1}\, \Omega_{zx}$, with the Ω matrix suitably partitioned. The elements of the $(p-1)$-dimensional vector β can be considered as the true regression coefficients of z on x. Letting $\Omega^{(t)}$ denote the posterior samples of Ω, $\beta^{(t)} = \{\Omega_{xx}^{(t)}\}^{-1}\, \Omega_{zx}^{(t)}$ give samples from the posterior distribution of β. Since $\Omega_{xx}^{(t)} = \Lambda_x^{(t)} \Lambda_x^{(t)\,\mathrm{T}} + \Sigma_{xx}^{(t)}$, where $\Lambda_x^{(t)}$ and $\Sigma_{xx}^{(t)}$ are appropriate sub-blocks of $\Lambda^{(t)}$ and $\Sigma^{(t)}$, one can use the Sherman–Morrison–Woodbury formula (Hager, 1989) to invert $\Omega_{xx}^{(t)}$ at each iteration of the Gibbs sampler, which only requires the inverse of a $k^{*(t)} \times k^{*(t)}$ matrix, leading to many-fold speed up in large p settings.

Table 2. *Predictive performance in the simulation study. Average, best and worst case performance across* 50 *simulation replicates are reported for the different methods*

true (p, k)	(100, 5)			(500, 10)			(1000, 15)		
method	MGPS	Lasso	Elastic net	MGPS	Lasso	Elastic net	MGPS	Lasso	Elastic net
mspe									
mean	0·63	0·55	0·55	0·41	0·38	0·38	0·95	0·87	0·88
min	0·32	0·33	0·33	0·18	0·22	0·22	0·57	0·55	0·56
max	0·89	0·79	0·78	0·86	0·57	0·56	1·48	1·44	1·44
aape									
mean	0·62	0·59	0·59	0·51	0·49	0·49	0·80	0·77	0·75
min	0·47	0·47	0·47	0·33	0·38	0·37	0·60	0·59	0·59
max	0·85	0·73	0·72	0·80	0·58	0·59	0·99	0·98	0·99
mape									
mean	2·19	2·07	2·07	1·71	1·66	1·68	2·54	2·48	2·48
min	1·36	1·43	1·40	1·21	1·17	1·18	1·83	1·83	1·80
max	3·15	2·91	2·89	2·95	2·70	2·63	3·27	3·07	3·07

MGPS, our proposed multiplicative shrinkage prior; mspe, mean squared prediction error; aape, average absolute prediction error; mape, maximum absolute prediction error.

Table 3. *Performance in estimating regression coefficients in the simulation study. We report the mean square error* $(\times 10^3)$*, average absolute bias* $(\times 10^3)$ *and maximum absolute bias* $(\times 10^3)$ *averaged across* 50 *simulation replicates for the different methods*

true (p, k)	(100, 5)			(500, 10)			(1000, 15)		
method	MGPS	Lasso	Elastic net	MGPS	Lasso	Elastic net	MGPS	Lasso	Elastic net
MSE	1·1	1·2	1·3	0·1	0·3	0·4	0·0	0·1	0·1
aab	10·1	12·4	13·0	1·7	3·9	4·1	0·9	1·8	1·9
mab	176·1	207·3	211·3	172·5	253·3	244·5	102·6	109·0	122·6

MGPS, our proposed multiplicative shrinkage prior; MSE, mean squared error; aab, average absolute bias; mab, maximum absolute bias.

As shown in Table 3, the estimate of β based on our method was close to the truth in each case, with small mean square error, average and maximum absolute bias. The coverage of 95% credible intervals for the elements of β were 0·96, 0·91 and 0·90 for the three cases, respectively.

The simulation examples were designed to induce correlation in groups of predictors, so that batches of predictors are included in the response model. The sparsity in the loadings ensures that many of the true regression coefficients are exactly equal to zero, with only a few important predictors. We propose a simple algorithm for variable selection in our framework based on thresholding the posterior mean of β. Let $\hat{\beta}_{(1)} < \cdots < \hat{\beta}_{(p-1)}$ denote the ordered values of the posterior means for the $p - 1$ predictors, and let $\pi_j = h$ denote that the jth predictor is the hth smallest in magnitude. Then, our thresholding approach sets $\beta_j = 0$ for all j with $\pi_j \leqslant \tilde{h}$, with $\tilde{h}$ chosen to minimize the mean squared prediction error. Table 4 shows the percentage of false positives and power compared with lasso and elastic net.

The three simulation examples took 2, 14 and 33 seconds per hundred iterations, respectively, in Matlab on a Intel(R) Core(TM) 2 Duo machine. The analyses were repeated with different choices of hyperparameter values. We used $\nu = 3{\cdot}5, 4, 5$ and varied b_σ between 0·1 and 0·5. We also used different multiples of $\log(p)$ between 3 and 10 for the initial number of factors. The results were robust, with the conclusions unchanged. We observed good mixing for the Gibbs sampler using both exploratory and diagnostic tests. The effective sample size averaged across

Table 4. *Variable selection performance in the simulation study. Percentage of false positives and power in detecting the true signal reported across* 50 *simulation replicates* (*average, best and worst case*) *for the different methods*

true (p, k)	(100, 5)			(500, 10)			(1000, 15)		
method	MGPS	Lasso	Elastic net	MGPS	Lasso	Elastic net	MGPS	Lasso	Elastic net
false positives (%)									
mean	0	9	7	0	4·0	3	0	3·0	2·0
min	0	0	0	0	0·2	0	0	0·7	0·7
max	0	26	25	0	14·0	14	0	8·0	10·0
power (%)									
mean	72	76	77	75	76	77	71	72	72
min	68	72	74	73	75	76	70	71	71
max	81	80	83	80	79	79	73	73	72

MGPS, our proposed multiplicative shrinkage prior.

the elements of β were 55, 53 and 48% for the three cases, respectively, suggesting an excellent computational efficiency.

The true loadings were not simulated from our proposed prior in any of the simulation examples. Although our prior on the loadings can concentrate in arbitrarily small neighbourhoods around zero, it does not allow any of the loading elements to be exactly zero. In the simulation study, many of the true loading elements were set equal to zero, and instead of shrinking the nonzero loadings with the column index, they were all drawn from the same N(0, 9) distribution. To assess robustness when the model is not applicable, we ran simulations with correlated factors and/or correlated idiosyncratic error, with the errors drawn from an AR(1) process. The results were robust even in these cases, in particular, we always had similar predictive performance as the elastic net. The adaptive method for factor selection proved to be extremely robust with respect to the choice of the threshold. Although we used 10^{-4} as a default threshold, the conclusions were mostly unchanged even with a threshold as small as 10^{-9}. Also, one can use either of the median or mode of the samples $k^{*(t)}$ as an estimate of the number of factors as they gave the same answer on all occasions. The simulation study clearly highlights the merit of our method in a variety of applications, with much improved performance over competitors in terms of covariance matrix estimation, regression coefficient estimation and variable selection.

5. Diffuse large-B-cell lymphoma application

5·1. *Background*

Lymphoma is a cancer of the white blood cell which occurs when lymphocytes, a type of white blood cell, have abnormal growth. Diffuse large B-cell lymphoma is the most common lymphoma among adults and has a high mortality rate. Rosenwald et al. (2002) analysed biopsy samples from 240 patients with untreated diffuse large B-cell lymphoma and identified 17 genes predictive of survival after chemotherapy. Segal (2006) reanalysed the data using penalized methods. The patients in the study were followed up after collection of biopsy specimens with a median follow-up of 2·8 years. For each patient, a potentially right-censored survival time is available along with 7399 features representing 4128 genes from the Lymphochip cDNA microarray. Rosenwald et al. (2002) divided the patients into a training set of 160 patients and a validation set of 80 patients to gauge predictive performance.

Rosenwald et al. (2002) used hierarchical clustering to identify four signature groups whose expressions were correlated with the survival times. They also identified a subset of 17 genes predictive of overall survival after chemotherapy. Gui & Li (2005), Segal (2006) and

Table 5. *Feature selection in the diffuse large-B-cell lymphoma data*

Unique ID	GenBank ID	Signature	Description
24094	AI476194	lymph	CD63 antigen (melanoma 1 antigen)
17048	AA085368	lymph	CD63 antigen (melanoma 1 antigen)
29636	NM005194	lymph	enhancer binding protein (C/EBP), β
34818	U83461	lymph	solute carrier family 31 (copper transporters), member 2
24394	AA729055	MHC	major histocompatibility complex, class II, DR α

Lymph, lymph-node signature; MHC, major histocompatibility complex; GenBank, National Institute of Health genetic sequence database.

Ma & Huang (2007) analysed this data using penalized methods. In each case, the selected features mostly belonged to one of the four signature groups in Rosenwald et al. (2002), though the individual selected features varied across the methods.

5·2. *Model and results*

Our interest lies in simultaneously identifying an important subset of the features and obtaining a predictive model for the exact survival times. Let T_i denote the survival time for the ith patient and let x_i denote the corresponding 7399 dimensional feature vector. There were 72 patients in the training set whose survival times were right-censored. Possibly due to rounding, there were some survival times equal to zero, so we added one unit to the survival times of all the patients. We took the logarithm of the shifted survival times and appended them to the x_is to create a p-dimensional vector $y_i = (z_i, x_i^{\mathrm{T}})^{\mathrm{T}}$, where $p = 7400$ and $z_i = \log(1 + T_i)$. We model the y_is jointly as in § 4·2 after normalizing them. The joint Gaussian model implies an accelerated failure time model for the survival times, since the conditional mean of the log-shifted survival time z_i given the predictors x_i is linear in x_i. Since the exact survival times are known for the uncensored subjects, the response was normalized with the mean and standard deviations of those subjects only and an intercept for the response was added to the model. A normal prior with zero mean and variance one was placed on the intercept. The posterior computation proceeds exactly as in § 3, but an additional step is needed to impute the shifted log survival times for the censored subjects from a truncated normal distribution, truncated below by the transformed censoring time. We ran the adaptive Gibbs sampler for 25 000 iterations with 5000 burn-in and collected every fifth sample after burn-in to thin the chain. The estimated number of factors was 20, with a 95% credible interval of (19,21).

We thresholded the posterior mean of the regression coefficients as described in § 4·2 to perform a variable selection. The thresholding approach selected 17 features, with all of the features belonging to three of the four signature groups mentioned in Rosenwald et al. (2002). The three signature groups were germinal-centre B-cell signature, major histocompatibility complex class II signature and lymph-node signature, while no genes in the proliferation signature group were selected. The top features mentioned in Gui & Li (2005) and Segal (2006) also come from the same three signature groups. In Table 5, we provide a brief description of the top five genes selected using our approach.

Among the features selected by our approach, the ones with GenBank ID AA729055, AA805575 and X59812 also appear in Gui & Li (2005) and Segal (2006). Although standard penalization methods tend to select one of a correlated group of important predictors, our approach is designed to allow selection of highly correlated predictors into the same model. This is illustrated in Table 5, as the first two predictors have a correlation coefficient of 0·96. There are several groups of highly correlated predictors in the selected set of 17.

Segal (2006) obtained the modest predictive accuracy using a variety of methods, so advocated exercising care before making prognosis based only on the gene expressions. Our analysis also suggested that the gene expression data explain only a small proportion of the variability in the survival times. The 95% predictive intervals for the survival times in the test sample were wide and contained the true survival times for the uncensored observations in all the cases. The mean square prediction error and mean absolute prediction error for the uncensored observations were 1·31 and 0·89 while the same for lasso trained with the uncensored observations in the training sample were 1·28 and 0·90. The proportion of times the predicted survival times for the censored observations exceeded the censoring time was 0·54. We also performed sensitivity analysis by varying ν, initial values of a_1, a_2 and the prior variance of the intercept. The conclusions were unchanged, with the same set of top 10 genes selected on all occasions.

Acknowledgement

This research was partially supported by a grant from the National Institute of Environmental Health Sciences of the National Institutes of Health, U.S.A. The authors would like to thank Mark Segal for sharing the diffuse large B-cell lymphoma data.

Appendix

Proofs

Proof of Lemma 1. It is enough to show that, for any $\Lambda \in \Theta_\Lambda$, $\Lambda\Lambda^{\mathrm{T}}$ is positive semi-definite. For any vector $v \in \Re^p$, $v^{\mathrm{T}}\Lambda\Lambda^{\mathrm{T}}v$ is finite since all elements of $\Lambda\Lambda^{\mathrm{T}}$ are finite. The proof is completed by observing that $v^{\mathrm{T}}\Lambda\Lambda^{\mathrm{T}}v = \|\Lambda^{\mathrm{T}}v\|^2 \geqslant 0$ where $\|\cdot\|$ denotes the Euclidian norm. □

Proof of Lemma 2. Let $\Omega = (\omega_{rs})$, $\Omega_0 = \omega^0_{rs}$, $\lambda_{jh} = \lambda^0_{jh} + \psi_{jh}$, clearly $d_2(\Lambda, \Lambda_0) = \big(\sum_{j=1}^{p}\sum_{h=1}^{\infty}\psi^2_{jh}\big)^{1/2}$. For any $1 \leqslant r, s \leqslant p$,

$$\left|\omega_{rs} - \omega^0_{rs}\right| \leqslant \sum_{h=1}^{\infty}\left|\lambda^0_{rh}\psi_{sh}\right| + \sum_{h=1}^{\infty}\left|\lambda^0_{sh}\psi_{rh}\right| + \sum_{h=1}^{\infty}\left|\psi_{rh}\psi_{sh}\right| + \epsilon \leqslant (2M_0 + 1)\epsilon + \epsilon^2,$$

by the Cauchy–Schwartz inequality, where $M_0 = \big\{\max_{1\leqslant j\leqslant p}\sum_{h=1}^{\infty}(\lambda^0_{jh})^2\big\}^{1/2} < \infty$. Thus $d_\infty(\Omega, \Omega_0) \leqslant \epsilon^*$, with $\epsilon^* = (2M_0 + 1)\epsilon + \epsilon^2$. □

Proof of Proposition 1. Clearly $\Pi_\Sigma\left(\Theta_\Sigma\right) = 1$, so it is enough to show $\Pi_\Lambda\left(\Theta_\Lambda\right) = 1$. The ϕ_{jh}s are independent of the δ_hs. Hence marginalizing over the ϕ_{jh}s yields $\lambda_{jh} \mid \tau_h \sim t_3(0, \tau_h^{-1})$ where $t_\nu(\mu, \sigma^2)$ denotes the t distribution with ν degrees of freedom having location μ and scale σ^2. By the Cauchy–Schwartz inequality,

$$\left(\sum_{h=1}^{\infty}\lambda_{rh}\lambda_{sh}\right)^2 \leqslant \left(\sum_{h=11}^{\infty}\lambda^2_{rh}\right)\left(\sum_{h=1}^{\infty}\lambda^2_{sh}\right) \leqslant \max_{1\leqslant j\leqslant p}\left(\sum_{h=1}^{\infty}\lambda^2_{jh}\right)^2,$$

and thus

$$\left|\sum_{h=1}^{\infty}\lambda_{rh}\lambda_{sh}\right| \leqslant \max_{1\leqslant j\leqslant p}\left(\sum_{h=1}^{\infty}\lambda^2_{jh}\right).$$

Hence all the elements of $\Lambda\Lambda^{\mathrm{T}}$ are bounded in absolute value by M, where $M = \max_{1\leqslant j\leqslant p} M_j$ with $M_j = \sum_{h=1}^{\infty}\lambda^2_{jh}$. Now,

$$E\left(M_j\right) = \sum_{h=1}^{\infty}E\left\{E\left(\lambda^2_{jh} \mid \tau_h\right)\right\} = \sum_{h=1}^{\infty}E\left(\frac{3}{\tau_h}\right) = \sum_{h=1}^{\infty}3ba^{h-1} = \frac{3b}{(1-a)} < \infty,$$

where $b = E\left(\delta_1^{-1}\right)$ and $a = E\left(\delta_2^{-1}\right) < 1$ if $a_2 > 2$. Hence $E(M) \leqslant \sum_{j=1}^{p} E(M_j) < \infty$ and thus M is finite almost surely. It follows that $\Pi_\Lambda \otimes \Pi_\Sigma\left(\Theta_\Lambda \times \Theta_\Sigma\right) = 1$. □

Proof of Theorem 1. Write $\Lambda\Lambda^{\mathrm{T}} = \Lambda_H\Lambda_H^{\mathrm{T}} + \Delta_H$. Clearly $d_\infty(\Omega, \Omega_H) = \max_{1\leqslant r,s\leqslant p} \left|a_{rs}^H\right|$, where a_{rs}^H is the rsth entry of Δ_H so that $a_{rs}^H = \sum_{h=H+1}^{\infty} \lambda_{rh}\lambda_{sh}$. An application of the Cauchy–Schwartz inequality as in the previous proof gives

$$\left|\sum_{h=H+1}^{\infty} \lambda_{rh}\lambda_{sh}\right| \leqslant \max_{1\leqslant j\leqslant p}\left(\sum_{h=H+1}^{\infty} \lambda_{jh}^2\right),$$

which implies $d_\infty(\Omega, \Omega_H) = \max_{1\leqslant j\leqslant p} a_{jj}^H$. Now, for a fixed $\epsilon > 0$,

$$\begin{aligned}
\mathrm{pr}\left\{d_\infty(\Omega, \Omega_H) \leqslant \epsilon\right\} &= E\left\{\mathrm{pr}(a_{11}^H \leqslant \epsilon, \ldots, a_{pp}^H \leqslant \epsilon \mid \delta)\right\}\\
&= E\left\{\mathrm{pr}(a_{11}^H \leqslant \epsilon \mid \delta)^p\right\} > \left[E\left\{\mathrm{pr}(a_{11}^H \leqslant \epsilon \mid \delta)\right\}\right]^p\\
&= \left[1 - E\left\{\mathrm{pr}(a_{11}^H > \epsilon \mid \delta)\right\}\right]^p \geqslant \left[1 - E\left\{\frac{E(a_{11}^H \mid \delta)}{\epsilon}\right\}\right]^p\\
&= \left\{1 - \frac{E(a_{11}^H)}{\epsilon}\right\}^p,
\end{aligned}$$

where the equality in the second line follows from the fact that a_{ii}^H are conditionally independent and identically distributed given δ and the subsequent two inequalities use Jensen's and Chebyshev's inequalities respectively. Now,

$$E(a_{11}^H) = E\{E(a_{11}^H \mid \delta)\} = E\left(\sum_{h=H+1}^{\infty} \frac{3}{\tau_h}\right) = \sum_{h=H+1}^{\infty} E\left(\frac{3}{\tau_h}\right) = \sum_{h=H+1}^{\infty} 3ba^{h-1} = \frac{3b}{(1-a)}a^H,$$

where $b = E\left(\delta_1^{-1}\right)$, $a = E\left(\delta_2^{-1}\right) < 1$ if $a_2 > 2$ and the third equality is a direct consequence of Fubini's theorem. Now use the inequality $(1 - x/2) > \exp(-x)$ if $0 < x \leqslant 1.5$ to get

$$\mathrm{pr}\left\{d_\infty(\Omega, \Omega_H) \leqslant \epsilon\right\} \geqslant \exp\left\{\frac{-6pb}{\epsilon(1-a)}a^H\right\}$$

if $H > \log\{2b/\epsilon(1-a)\}/\log(1/a)$. Hence

$$\mathrm{pr}\left\{d_\infty(\Omega, \Omega_H) > \epsilon\right\} \leqslant 1 - \exp\left\{\frac{-6pb}{\epsilon(1-a)}a^H\right\} \leqslant \frac{6pb}{\epsilon(1-a)}a^H$$

for $6a^H pb/\{(1-a)\epsilon\} < 1$ or $H > \log\{6pb/\epsilon(1-a)\}/\log(1/a)$. □

Proof of Proposition 2. Let Λ_* be a $p \times k$ matrix ($k \leqslant p$) and $\Sigma_0 \in \Theta_\Sigma$ such that $\Omega_0 = \Lambda_*\Lambda_*^{\mathrm{T}} + \Sigma_0$. Set $\Lambda_0 = (\Lambda_* : 0_{p\times\infty})$, then $(\Lambda_0, \Sigma_0) \in \Theta_\Lambda \times \Theta_\Sigma$, with $g(\Lambda_0, \Sigma_0) = \Omega_0$. Fix $\epsilon > 0$, choose $\epsilon_1 > 0$ such that $(2M_0 + 1)\epsilon_1 + \epsilon_1^2 < \epsilon$, with M_0 as in the proof of Lemma 2. By Lemma 2, $g\left\{B_{\epsilon_1}(\Lambda_0, \Sigma_0)\right\} \subset B_\epsilon^\infty(\Omega_0)$, and thus $B_{\epsilon_1}(\Lambda_0, \Sigma_0) \subset g^{-1}\left\{B_\epsilon^\infty(\Omega_0)\right\}$. Now $\Pi\left\{B_\epsilon^\infty(\Omega_0)\right\} = (\Pi_\Lambda \otimes \Pi_\Sigma) \circ g^{-1}\left\{B_\epsilon^\infty(\Omega_0)\right\} \geqslant \Pi_\Lambda \otimes \Pi_\Sigma\{B_{\epsilon_1}(\Lambda_0, \Sigma_0)\}$. Clearly, $\Pi_\Sigma\{\Sigma : d_\infty(\Sigma, \Sigma_0) < \epsilon_1\} > 0$, so it is enough to show Π_Λ

$\{\Lambda : d_2(\Lambda, \Lambda_0) < \epsilon_1\} > 0$. We have,

$$\begin{aligned}
\mathrm{pr}\{d_2(\Lambda, \Lambda_0) < \epsilon_1\} &= \mathrm{pr}\left\{\sum_{j=1}^{p}\sum_{h=1}^{\infty}(\lambda_{jh} - \lambda_{jh}^0)^2 < \epsilon_1^2\right\} \\
&\geqslant \mathrm{pr}\left\{\sum_{h=1}^{\infty}(\lambda_{jh} - \lambda_{jh}^0)^2 < \epsilon_1^2/p,\ j = 1, \ldots, p\right\} \\
&= E_\delta\left[\prod_{j=1}^{p}\mathrm{pr}\left\{\sum_{h=1}^{\infty}(\lambda_{jh} - \lambda_{jh}^0)^2 < \epsilon_1^2/p \mid \delta\right\}\right] > 0
\end{aligned}$$

by the following Lemma. □

LEMMA 3. *Fix $1 \leqslant j \leqslant p$. For any $\epsilon > 0$, pr$\{\sum_{h=1}^{\infty}(\lambda_{jh} - \lambda_{jh}^0)^2 < \epsilon_1^2/p \mid \delta\} > 0$ almost surely.*

Proof of Lemma 3. We have $\lambda_{jh}^0 = 0$ for $h > k$. Thus for any $H \geqslant k$,

$$\begin{aligned}
\mathrm{pr}\left\{\sum_{h=1}^{\infty}(\lambda_{jh} - \lambda_{jh}^0)^2 < \epsilon \mid \delta\right\} &\geqslant \mathrm{pr}\left\{\sum_{h=1}^{H}(\lambda_{jh} - \lambda_{jh}^0)^2 < \epsilon/2,\ \sum_{h=H+1}^{\infty}\lambda_{jh}^2 < \epsilon/2 \mid \delta\right\} \\
&= \mathrm{pr}\left\{\sum_{h=1}^{H}(\lambda_{jh} - \lambda_{jh}^0)^2 < \epsilon/2 \mid \delta\right\}\ \mathrm{pr}\left\{\sum_{h=H+1}^{\infty}\lambda_{jh}^2 < \epsilon/2 \mid \delta\right\}.
\end{aligned}$$

By Theorem 1, pr$(\sum_{h=H+1}^{\infty}\lambda_{jh}^2 < \epsilon/2) \to 1$ as $H \to \infty$, hence we can find $H_0 > k$ such that pr$(\sum_{h=H_0+1}^{\infty}\lambda_{jh}^2 < \epsilon/2) > 0$ and thus pr$(\sum_{h=H_0+1}^{\infty}\lambda_{jh}^2 \mid \delta) > 0$ almost surely. The proof is completed by observing that pr$\{\sum_{h=1}^{H}(\lambda_{jh} - \lambda_{jh}^0)^2 < \epsilon/2 \mid \delta\} > 0$ almost surely for any $H < \infty$. □

Proof of Theorem 2. Fix $\epsilon > 0$, $\Omega_0 \in \Theta$. We have,

$$K(\Omega_0, \Omega) = \frac{1}{2}\log\frac{\det\Omega_0}{\det\Omega} - \frac{1}{2}\mathrm{tr}(\mathrm{I}_p - \Omega^{-1}\Omega_0).$$

Let $u_0 = \det\Omega_0$, find $\epsilon_1 > 0$ such that $|u - u_0| < \epsilon_1$ implies $|\log u - \log u_0| < \epsilon$. Since det$(\cdot)$ is a continuous function from Θ to $\Re$, we can find ϵ_2 such that $d_\infty(\Omega_0, \Omega) < \epsilon_2$ implies $|\det(\Omega_0) - \det(\Omega)| < \epsilon_1$. Now $\mathrm{tr}(\mathrm{I}_p - \Omega^{-1}\Omega_0) = \sum_{i=1}^{p}(1 - \lambda_i)$, where $\lambda_1 \leqslant \ldots \leqslant \lambda_p$ are the eigenvalues of $\Omega^{-1}\Omega_0$. Since Ω and Ω_0 are both positive definite,

$$0 \leqslant \lambda_1 \leqslant \frac{x^{\mathrm{T}}\Omega_0 x}{x^{\mathrm{T}}\Omega x} \leqslant \lambda_p,$$

where x is any p-dimensional vector with $x^{\mathrm{T}}x = 1$. For any $x \in \Re^p$ with $x^{\mathrm{T}}x = 1$,

$$\left|\frac{x^{\mathrm{T}}\Omega_0 x}{x^{\mathrm{T}}\Omega x} - 1\right| = \frac{|x^{\mathrm{T}}\Omega_0 x - x^{\mathrm{T}}\Omega x|}{x^{\mathrm{T}}\Omega x}.$$

Now

$$|x^{\mathrm{T}}\Omega_0 x - x^{\mathrm{T}}\Omega x| \leqslant \sum_{i=1}^{p}\sum_{j=1}^{p}|w_{ij} - w_{ij}^0||x_i x_j| \leqslant d_\infty(\Omega_0, \Omega)\left(\sum_{i=1}^{p}|x_i|\right)^2 \leqslant p\ d_\infty(\Omega_0, \Omega),$$

and

$$x^{\mathrm{T}}\Omega x = x^{\mathrm{T}}\Omega_0 x + (x^{\mathrm{T}}\Omega_0 x - x^{\mathrm{T}}\Omega x) \geqslant \lambda_{\min}(\Omega_0) + (x^{\mathrm{T}}\Omega_0 x - x^{\mathrm{T}}\Omega x),$$

where $\lambda_{\min}(\Omega_0) > 0$ denotes the smallest eigenvalue of Ω_0. Choose $0 < \epsilon_3 < \lambda_{\min}(\Omega_0)/2p$ such that $2p^2\epsilon_3/\lambda_{\min}(\Omega_0) < \epsilon$. We have

$$\left|\frac{x^{\mathrm{T}}\Omega_0 x}{x^{\mathrm{T}}\Omega x} - 1\right| = \frac{|x^{\mathrm{T}}\Omega_0 x - x^{\mathrm{T}}\Omega x|}{x^{\mathrm{T}}\Omega x} \leqslant \frac{p\,\epsilon_3}{\lambda_{\min}(\Omega_0)/2} < \epsilon/p,$$

for all Ω_0 such that $d_\infty(\Omega_0, \Omega) < \epsilon_3$, since $|x^{\mathrm{T}}\Omega_0 x - x^{\mathrm{T}}\Omega x| < \lambda_{\min}(\Omega_0)/2$ and thus $x^{\mathrm{T}}\Omega x > \lambda_{\min}(\Omega_0)/2$. Choose $\epsilon^* = \min\{\epsilon_2, \epsilon_3\}$, then for $d_\infty(\Omega_0, \Omega) < \epsilon^*$, we have,

$$\begin{aligned} K(\Omega_0, \Omega) &\leqslant \frac{1}{2}\left|\log\frac{\det\Omega_0}{\det\Omega}\right| + \frac{1}{2}\left|\mathrm{tr}(\mathrm{I}_p - \Omega^{-1}\Omega_0)\right| \\ &\leqslant \frac{1}{2}\left|\log(\det\Omega_0) - \log(\det\Omega)\right| + \frac{1}{2}\sum_{i=1}^{p}|1-\lambda_i| \\ &\leqslant \frac{\epsilon}{2} + p\max\{|1-\lambda_1|, |1-\lambda_p|\} < \epsilon, \end{aligned}$$

which proves Theorem 2. □

REFERENCES

AMENGUAL, D. & WATSON, M. (2007). Consistent estimation of the number of dynamic factors in a large *N* and *T* panel. *J. Bus. Econ. Statist.* **25**, 91–6.

ANDO, T. (2009). Bayesian factor analysis with fat-tailed factors and its exact marginal likelihood. *J. Mult. Anal.* **100**, 1717–26.

ARMINGER, G. & MUTHÉN, B. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis–Hastings algorithm. *Psychometrika* **63**, 271–300.

BAI, J. & NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70**, 191–221.

BICKEL, P. & LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199–227.

CARVALHO, C., CHANG, J., LUCAS, J., NEVINS, J., WANG, Q. & WEST, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *J. Am. Statist. Assoc.* **103**, 1438–56.

CELEUX, G., CHAUVEAU, D. & DIEBOLT, J. (1996). Stochastic versions of the EM algorithm: an experimental study in the mixture case. *J. Statist. Comp. and Simul.* **55**, 287–314.

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.* **1**, 515–34.

GEWEKE, J. & ZHOU, G. (1996). Measuring the price of the Arbitrage Pricing Theory. *Rev. Finan. Studies* **9**, 557–87.

GHOSH, J. & DUNSON, D. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *J. Comp. Graph. Statist.* **18**, 306–20.

GUI, J. & LI, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* **21**, 3001–8.

HAGER, W. (1989). Updating the inverse of a matrix. *SIAM Rev.* **31**, 221–39.

LEE, S. & SONG, X. (2002). Bayesian selection on the number of factors in a factor analysis model. *Behaviormetrika* **29**, 23–39.

LIU, J. & WU, Y. (1999). Parameter expansion for data augmentation. *J. Am. Statist. Assoc.* **94**, 1264–74.

LOPES, H. & WEST, M. (2004). Bayesian model assessment in factor analysis. *Statist. Sinica* **14**, 41–68.

LUCAS, J., CARVALHO, C., WANG, Q., BILD, A., NEVINS, J. & WEST, M. (2006). Sparse statistical modelling in gene expression genomics. *Bayesian Inference for Gene Expression and Proteomics*, Eds. P. Müller, K. Do, and M. Vannucci, 155–76. Cambridge: Cambridge University Press.

MA, S. & HUANG, J. (2007). Additive risk survival model with microarray data. *BMC Bioinformatics* **8**, 192.

ROBERTS, G. & ROSENTHAL, J. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Prob.* **44**, 458–475.

ROSENWALD, A., WRIGHT, G., CHAN, W. C., CONNORS, J. M., CAMPO, E., FISHER, R. I., GASCOYNE, R. D., MUELLER-HERMELINK, H. K., SMELAND, E. B. et al. (2002). The use of molecular profiling to predict survival after chemotheropy for diffuse large-B-cell lymphoma. *New Engl. J. Med.* **346**, 1937–47.

SCHWARTZ, L. (1965). On Bayes procedures. *Prob. Theory Rel. Fields* **4**, 10–26.

SEGAL, M. (2006). Microarray gene expression data with linked survival phenotypes: diffuse large-B-cell lymphoma revisited. *Biostatistics* **7**, 268–85.

SONG, X. & LEE, S. (2001). Bayesian estimation and test for factor analysis model with continuous and polytomous data in several populations. *Br. J. Math. Statist. Psychol.* **54**, 237–63.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

WEST, M. (2003). Bayesian factor regression models in the "large *p*, small *n*" paradigm. *Bayesian Statist.* **7**, 723–32.

ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B **67**, 301–20.

[*Received November* 2009. *Revised December* 2010]