Sparse Bayesian infinite factor models

Anirban Bhattacharya
David B. Dunson

Department of Statistical Science,
Duke University, Durham, NC 27708
email: ab179@stat.duke.edu dunson@stat.duke.edu

Abstract

We focus on sparse modeling 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 toward zero as the column index increases. We use our prior on a parameter expanded loadings matrix to avoid the order dependence typical in factor analysis models and develop a highly 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 loadings 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 after chemotherapy from gene expression data.

KEYWORDS: Adaptive; Factor analysis; High-dimensional; 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 $\Lambda\Lambda' + \Sigma$, where Λ is a $p \times k$ factor loadings matrix with $k << p$ and Σ is a $p \times p$ diagonal matrix with non-negative diagonal entries. A popular approach to ensure identifiability of the loadings elements is to constrain the loadings matrix to be lower triangular with positive diagonal entries [Geweke and Zhou, 1996]. Factor models have been traditionally applied in behavioral 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, small n applications such as gene expression studies.

A Bayesian specification of the factor model [Arminger and Muthén, 1998, Song and 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. Also, posterior inference tends to be sensitive to certain hyperparameters, with elicitation difficult. To address these issues, Ghosh and Dunson [2009] use parameter expansion [Liu and Wu, 1999, Gelman, 2006] to induce a heavy-tailed default prior distribution on the loadings 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 works in this direction [Polasek, 1997] involve computation of the marginal likelihoods under models with different numbers of factors. Lopes and West [2004] proposed a reversible jump Markov chain Monte Carlo algorithm to allow for uncertainty in the number of factors. Lee and Song [2002] instead developed a path sampling approach. A more recent method infers the number of factors by zeroing a subset of the loadings elements using Bayesian variable selection priors [Lucas et al., 2006, Carvalho et al., 2008, Schnatter and Lopes, 2009]. 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 toward 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 free of order dependence. We provide results on the support of our prior and explore the induced sparsity in the loadings with respect to the column index, obtaining a bound on approximation error resulting from truncating the loadings matrix at finitely many columns. The truncation bound justifies using a finite approximation. We propose a straightforward Gibbs sampler for posterior computation that scales well in large p problems. The structure admits a jointly conjugate conditional posterior for the loadings, thereby allowing block updating. An adaptive Gibbs approach is proposed to allow selection of the truncation level and number of important factors. A fast algorithm is proposed for calculating an approximate maximum *a posteriori* estimate of the covariance matrix.

The proposed prior is relevant in a broad class of applications outside the factor analysis setting. Recently, there has been a significant amount of work on latent feature models using the Indian buffet process [Griffiths and Ghahramani, 2006, Thibaux and Jordan, 2007]. The Indian buffet process defines a probability distribution on infinite binary matrices, with weighted versions finding applications in factor analysis [Knowles and Ghahramani, 2007, Meeds et al., 2007, Rai and Daumé, 2009]. Our approach can be used as a simpler and more computationally efficient alternative to the Indian buffet process to define priors on latent feature matrices.

2 Bayesian factor models

2.1 Model and prior specification

The generic form of a latent factor model is given by

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

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 $\Omega = \Lambda\Lambda' + \Sigma$. 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.

Note that the above decomposition of Ω is not unique and actually there are infinitely many possible, since $\Lambda_1 = \Lambda P$ also satisfies the above condition for any semi-orthogonal matrix P ($PP' = I$). One specific choice of P is given by $P = [I_k \ : \ 0_{k\times r}]$ which simply appends r zero columns to the right of Λ. Clearly, the loadings matrix Λ is not identifiable without further restrictions on its structure. Following Geweke and Zhou [1996], a common approach to ensure identifiability is to constrain Λ to a block lower triangular form, with the diagonal elements strictly positive. This constraint induces order dependence among the responses [Aguilar and West, 2000, West, 2003, Lopes and West, 2004], with the choice of the first k response variables being an important modeling decision [Carvalho et al., 2008]. One of the key points we want to emphasize is that the uniqueness of the decomposition is not necessary in many applications, including inference on the covariance matrix and prediction. Our approach avoids the need for a unique decomposition, which frees us to define priors with much better properties computationally while simplifying the theory.

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

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

Using the Cauchy-Schwartz inequality, it is straightforward to show that all the entries of $\Lambda\Lambda'$ are finite if and only if the condition in (2) is satisfied. Denote Θ_Σ to be the set of $p \times p$ diagonal matrices with non-negative entries and Θ to be all $p \times p$ positive semi-definite matrices. Consider the function $g\colon \Theta_\Lambda \times \Theta_\Sigma \to \Theta$ corresponding to $g(\Lambda, \Sigma) = \Lambda\Lambda' + \Sigma$.

Lemma 2.1 *For any* $(\Lambda, \Sigma) \in \Theta_\Lambda \times \Theta_\Sigma$, $g(\Lambda, \Sigma) \in \Theta$.

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 2.2 states a regularity property of g, which is later used to prove sup-norm support of the proposed prior.

Lemma 2.2 *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 given by

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

for $p\times\infty$ matrices $\Lambda = (\lambda_{jh}), \Lambda_0 = (\lambda^0_{jh})$, and $d_\infty(A, B) = \max_{1\le r,s\le p} |a_{rs} - b_{rs}|$ 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 loadings matrix without imposing any restriction on the loadings 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,

$$\lambda_{jh} \mid \phi_{jh}, \tau_h \sim N(0, \phi_{jh}^{-1}\tau_h^{-1}), \ \phi_{jh} \sim \text{Ga}(3/2, 3/2), \ \tau_h = \prod_{l=1}^{h} \delta_l,$$

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

where $\delta_l, \ l = 1, \ldots, \infty$, are independent, τ_h is a global shrinkage parameter for the hth column and ϕ_{jh}'s are local shrinkage parameters for the elements

in the hth column. Note that the τ_h's are stochastically increasing under the restriction $a_2 > 1$, which favors more shrinkage as the column index increases. If we only use the global shrinkage parameter, the prior has a tendency to over-shrink the non-zero 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. Properties of the proposed prior are described next.

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

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 $\Omega = \Lambda\Lambda' + \Sigma$ be the induced covariance matrix. We can approximate Ω by $\Omega_T = \Lambda_T\Lambda_T' + \Sigma$ where Λ_T denote the matrix obtained by setting the columns of Λ from $T+1$ onwards to zero or equivalently discarding those higher indexed columns. The following theorem states that the prior probability of Ω_T being arbitrarily close to Ω in an appropriate sense converges exponentially fast to 1 as T tends to ∞.

Theorem 2.4 *If* $a_2 > 2$, *then for any* $\epsilon > 0$,

$$pr\big\{d_\infty\big(\Omega, \Omega_T\big) > \epsilon\big\} < \frac{6pb}{\epsilon(1-a)}a^T \text{ for } T > \frac{\log\{6pb/\epsilon(1-a)\}}{\log(1/a)},$$

where $b = E\big(\delta_1^{-1}\big)$ *and* $a = E\big(\delta_2^{-1}\big)$.

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 $a_2 > 1 + b_2$.

Next we investigate the support of our prior. Let Π denote the the induced prior on Θ, then $\Pi = (\Pi_\Lambda \otimes \Pi_\Sigma) \circ g^{-1}$ 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 2.5 shows that the proposed prior has sup-norm support on Θ,

Proposition 2.5 *If Ω_0 is any $p \times p$ covariance matrix and $B_\epsilon^\infty(\Omega_0)$ is an ϵ-neighborhood of Ω_0 under the sup-norm, then $\Pi\{B_\epsilon^\infty(\Omega_0)\} > 0$ for any $\epsilon > 0$.*

Proposition 2.5 shows that our proposed prior has large support, so places positive probability in arbitrary small neighborhoods around any covariance matrix. We use Proposition 2.5 to show weak consistency of the posterior distribution of Ω in Theorem 2.6. Denote $K(\Omega_0, \Omega)$ to be the Kullback-Leibler divergence between $N_p(0, \Omega_0)$ and $N_p(0, \Omega)$ given by

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

Theorem 2.6 *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 the posterior distribution of Ω is weakly consistent.

The weak consistency of the posterior follows from the famous Schwartz [1965] theorem, since any Kullback-Leibler neighborhood of the true density has positive probability using Proposition 2.5.

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 $\Omega \overset{\mathcal{D}}{=} \Omega_\pi$, where $\Omega_\pi = (w_{\pi_r \pi_s})$ with π any permutation of $\{1, \ldots, p\}$ and $\Omega = (w_{rs})$. We have $w_{rs} = \sum_{h=1}^\infty \lambda_{rh}\lambda_{sh} = \lambda_r'\lambda_s$, where $\lambda_j = (\lambda_{j1}, \lambda_{j2}, \ldots)'$. Note that $\lambda_{rh} \mid \tau_h \sim t_3(0, \tau_h^{-1})$ and λ_{rh}'s are conditionally independent given $\tau = (\tau_1, \tau_2, \ldots)'$. Since the marginal prior on λ_r is the same for every r, $w_{rs} \overset{\mathcal{D}}{=} w_{r's'}$ for any $(r, s) \neq (r', s')$ such that $r \neq s, r' \neq s'$. The permutation invariance follows from the fact that $w_{rr} \overset{\mathcal{D}}{=} w_{r'r'}$ for any $1 \leq r, r' \leq p$.

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

$$E(w_{rs}) = \sum_{h=1}^\infty E\Big\{ E(\lambda_{rh}\lambda_{sh} \mid \tau_h) \Big\} = 0$$

and

$$\begin{aligned} E\big(w_{rs}^2\big) &= E\big\{tr(\lambda_r'\lambda_s\lambda_s'\lambda_r)\big\} = tr\Big\{E\big(\lambda_r\lambda_r'\lambda_s\lambda_s'\big)\Big\} \\ &= tr\Big[E\big\{E(\lambda_r\lambda_r' \mid \tau)E(\lambda_s\lambda_s' \mid \tau)\big\}\Big] = 9\sum_{h=1}^{\infty} E(\tau_h^{-2}). \end{aligned}$$

Thus $E\big(w_{rs}^2\big)$ is finite if $d = E(\delta_1^{-2})$ is finite and $c = E(\delta_2^{-2}) < 1$ and in that case $E\big(w_{rs}^2\big) = 9d/(1-c)$. 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 $k^* << p$ columns. An adaptive strategy for inference on the truncation level k^* is described in section 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' , then the λ_j's have independent conditionally conjugate posteriors given by

$$\pi(\lambda_j \mid -) \sim N_{k^*}\Big(\big(D_j^{-1} + \sigma_j^{-2}\eta'\eta\big)^{-1}\eta'\sigma_j^{-2}y^{(j)}, \big(D_j^{-1} + \sigma_j^{-2}\eta'\eta\big)^{-1}\Big),$$

where $\eta' = [\eta_1, \ldots, \eta_n]$, $D_j^{-1} = \text{diag}(\phi_{j1}\tau_1, \ldots, \phi_{jk^*}\tau_{k^*})$ and $y^{(j)} = (y_{1j}, \ldots, y_{nj})'$ for $j = 1, \ldots, p$.

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

$$\pi(\sigma_j^{-2} \mid -) \sim \text{Ga}\Big(a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2}\sum_{i=1}^{n}(y_{ij} - \lambda_j'\eta_i)^2\Big),$$

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

$$\pi(\eta_i \mid -) \sim N_{k^*}\Big(\big(I_{k^*} + \Lambda_{k^*}'\Sigma^{-1}\Lambda_{k^*}\big)^{-1}\Lambda_{k^*}'\Sigma^{-1}y_i, \big(I_{k^*} + \Lambda_{k^*}'\Sigma^{-1}\Lambda_{k^*}\big)^{-1}\Big),$$

Step 4. Sample ϕ_{jh} from

$$\pi(\phi_{jh} \mid -) \sim \text{Ga}\bigg(\frac{\nu+1}{2}, \frac{\nu+\tau_h\lambda_{jh}^2}{2}\bigg),$$

with $\nu = 3$.

Step 5. Sample δ_1 from

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

and for $h \geq 2$, sample δ_h from

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

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

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 to the dimension p of the outcomes. Theorem 2.4 provides theoretical justification for using a loadings matrix with $T << p$ non-zero columns. However, we need a computational strategy for choosing an appropriate level of truncation k^*. Ideally, 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 section 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 neighborhood 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 $\tilde{k} \geq k^*$. However, in order to be assured that $\tilde{k} \geq k^*$, 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 and 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 $u_t \leq p(t)$, we monitor the columns in the loadings having all elements within some pre-specified small neighborhood 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 non-redundant columns.

The most commonly used 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 and 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 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. A distinct 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 and Rosenthal [2007]. In addition, we save computation

by discarding the unimportant factors. We have observed in a number of simulation examples that our adaptive sampler gets close to the true number of factors within a few hundred iterations, irrespective of the starting number of 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, $\Omega^{(t)} = \Lambda^{(t)}_{\tilde{k}^{(t)}} \Lambda^{(t)'}_{\tilde{k}^{(t)}} + \Sigma^{(t)}$ represent draws from the approximated marginal posterior distribution of Ω given $y_i, i = 1, \ldots, n$, 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 [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 considered a number of simulation examples to illustrate our approach and compare with relevant methods. We simulated $y_i, i = 1, \ldots, 200$, from a p dimensional normal distribution with zero mean and covariance matrix $\Omega = \Lambda\Lambda' + \Sigma$ 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 $\text{Ga}(1, 0.25)$ 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 allocated the location of the zeros in each column and simulated the non-zero elements independently from a normal distribution with mean 0 and variance 9.

We chose three (p, k) combinations with moderate to large p, namely $(100, 5)$, $(500, 10)$ and $(1000, 15)$. For each pair we considered 50 simulation replicates. We ran the adaptive Gibbs sampler for 25000 iterations with a burn-in of 5000, and collected every 5th sample to thin the chain. We used 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) were chosen to be 1 and 0.3 respectively. We placed $\text{Ga}(2, 1)$ priors on a_1 and a_2. We chose α_0 and

α_1 in the adaptation probability $p(t)$ as -1 and -5×10^{-4} respectively. We monitored the columns in the loadings having all elements less than 10^{-4} in magnitude and proceeded by adapting the number of factors as in section 3.2. For the stochastic EM algorithm, we chose a burn-in of 100 and monitored the estimated covariance matrix every 10 iterations. We stopped the chain when the sup-norm distance between the estimated covariance matrix at the current iteration was within a small tolerance level compared to the estimate 10 iterations previously.

The average of the estimated number of factors across the replicates was 6.82, 10 and 14.40 corresponding to $k = 5, 10$ and 15 with empirical 95% interval for the number of factors given by $[5, 8]$, $[9, 11]$ and $[13, 16]$ respectively.

The estimated covariance matrix in each case was close to the true value, with small mean square error, average and maximum absolute bias. We compared estimation of the covariance matrix to a recent method by Bickel and 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. The average, best and worst case performances are tabulated. Based on Table 1, the proposed shrinkage approach does significantly better than the Bickel and 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. It is well known that ordinary least squares fails in the large p, small n paradigm. Modifications using penalized techniques have been studied extensively. The lasso [Tibshirani, 1996] and the elastic net [Zou and 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].

Let $y_i = (z_i, x_i')'$, $i = 1, \ldots, n$, where x_i's are $(p-1)$ dimensional predictors and z_i's are the response. For ease of illustration, we assume the z_i's 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_i's 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 given by

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

For the simulation examples described in section 4.1, let $z_i = y_{i1}$ and $x_i = (y_{i2}, \ldots, y_{ip})'$. 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_i's 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 similarly to lasso and elastic net, but has the advantage of quantifying predictive uncertainty.

The joint Gaussian model implies that $E(z_i \mid x_i) = x_i'\beta$, with $\beta = \Omega_{xx}^{-1}\, \Omega_{zx}$, where the Ω matrix is partitioned as

$$\Omega = \begin{pmatrix} \Omega_{zz} & \Omega_{zx} \\ \Omega_{xz} & \Omega_{xx} \end{pmatrix}$$

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 $\Omega_{xx}^{(t)} = \Lambda_x^{(t)} \Lambda_x^{(t)'} + \Sigma_{xx}^{(t)}$, 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 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.

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. The lasso and the elastic net perform variable selection by zeroing a subset of the coefficients. 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 $\pi_j \leq \tilde{h}$, with $\tilde{h}$ chosen to minimize the mean square prediction error. We compare the proportion of false positives and power to lasso and elastic net, summaries are in Table 4. As before, we report the average, worst and best case performances. From Table 4, it is evident that our approach does similar to lasso and elastic net in identifying the signal, but outperforms them comprehensively in terms of protecting against false positives. A visual summary of the above facts are provided in fig 1. For each p, the true $p-1$ regression coefficients and the estimated ones (using our thresholding approach and elastic net) are plotted against their index for one randomly chosen replicate.

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 let $\phi_{jh} \sim \text{Ga}(\nu/2, \nu/2)$, with $\nu = 3.5, 4$ 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 the elements of β were 55%, 53% and 48% for the three cases respectively, suggesting excellent MCMC efficiency.

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] analyzed biopsy samples from 240 patients with untreated diffuse large-B-cell lymphoma and identified 17 genes predictive of survival after chemotherapy. Segal [2006] re-analyzed 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. Segal [2006] used the same training-validation split.

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. Segal [2006] evaluated a number of competing approaches and advocated using residual finesse to lessen the computational burden associated with penalized methods in proportional hazards model. Gui and Li [2004], Ma and Huang [2007] are among others who performed variable selection with this data using penalized methods, Gui and Li [2004] used L_1-penalized Cox regression while Ma and Huang [2007] proposed a penalized additive risk survival model. For all of these methods, 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 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_i's to create a p dimensional vector $y_i = (z_i, x_i')'$, where $p = 7400$ and $z_i = \log(1 + T_i)$. We can now model the y_i's jointly as in section 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 section 3, only 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 5,000 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 initially focus on identifying an important subset of the features. We thresholded the posterior mean of the regression coefficients as described in section 4.2 to perform 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-center B-cell signature, MHC class II signature and Lymph-node signature, while no genes in the proliferation signature group were selected. The top features mentioned in Gui and Li [2004] 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 and Li [2004] 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 0.958. There are several groups of highly correlated predictors in the selected set of 17.

Segal [2006] obtained 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 1.3069 and 0.8928 while the same for lasso trained with the uncensored observations in the training sample were 1.2828 and 0.8982. The proportion of times the predicted survival times for the censored observations exceeded the censoring time was 0.54.

Prediction of survival times based only on the gene expressions obtained at the beginning of the study seems to be too optimistic, since a number of other factors can affect mortality given the length of the median follow up time. However, the feature selection seems to be more feasible, and there is some agreement across the different studies. Three of the signature groups mentioned in Rosenwald et al. [2002] seem to be important in characterizing mortality, as all the selected features belonged to those groups. We have seen in a number of simulation examples that our thresholding approach protects

well against false positives, which leads us to believe that the expression of the selected features play an important role in survival probability of patients having large-B-cell lymphoma. 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 ten genes selected on all occasions.

6 Discussion

This paper proposes a sparse shrinkage prior on factor loadings for modeling large covariance matrices. The proposed prior allows introduction of infinitely many factors, with sparsity manifest through an increasing degree of shrinkage with the column index. The emphasis is on prediction and variable selection in large p, small n settings using a low-rank approximation of the joint covariance matrix of the predictors and the response. One does not require identifiability of the loadings in such settings, a fact which has been exploited throughout to define the prior on a parameter expanded loadings matrix. A method based on thresholding the posterior mean of the regression coefficients is proposed for variable selection, which seems promising for selecting correlated batches of predictors. The challenging issue of inference on the number of factors is addressed easily using the shrinkage approach. The shrinkage prior has attractive theoretical properties and allows block updating of the loadings, resulting in efficient posterior computation and mixing. To make the approach robust with respect to the initial number of factors, an adaptive Gibbs sampler is used to tune the number of factors as the sampler progresses. The adaptive approach avoids the need for using a very conservative initial choice, thereby improving the computational efficiency.

7 Acknowledgement

This research was partially supported by grant number R01 ES017240-01 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH). The authors would like to thank Mark Segal for sharing the DLBCL data.

Appendix

Proof of Lemma 2.1

It is enough to show that, for any $\Lambda \in \Theta_\Lambda$, $\Lambda\Lambda'$ is positive semi-definite. For any vector $v \in \mathcal{R}^p$, $v'\Lambda\Lambda'v$ is finite since all elements of $\Lambda\Lambda'$ are finite. The proof is completed by observing that $v'\Lambda\Lambda'v = ||\Lambda'v||^2 \geq 0$ where $||\cdot||$ denotes the Euclidian norm.

Proof of Lemma 2.2

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

$$\begin{aligned} \big|w_{rs} - w^0_{rs}\big| &\leq \sum_{h=1}^\infty \big|\lambda^0_{rh}\psi_{sh}\big| + \sum_{h=1}^\infty \big|\lambda^0_{sh}\psi_{rh}\big| + \sum_{h=1}^\infty \big|\psi_{rh}\psi_{sh}\big| + \epsilon \\ &\leq (2M_0 + 1)\epsilon + \epsilon^2, \end{aligned}$$

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

Proof of Proposition 2.3

Clearly $\Pi_\Sigma\big(\Theta_\Sigma\big) = 1$, so it is enough to show $\Pi_\Lambda\big(\Theta_\Lambda\big) = 1$. The ϕ_{jh}'s are independent of the δ_h's. 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 Cauchy-Schwartz inequality,

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

and thus

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

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

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

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

Proof of Theorem 2.4

Write $\Lambda\Lambda' = \Lambda_T\Lambda_T' + \Delta_T$. Clearly $d_\infty(\Omega, \Omega_T) = \max_{1\leq r,s\leq p} \big|a_{rs}^T\big|$, where a_{rs}^T is the rsth entry of Δ_T given by $a_{rs}^T = \sum_{h=T+1}^{\infty} \lambda_{rh}\lambda_{sh}$. An application of the Cauchy-Schwartz inequality as in the previous proof gives

$$\Big|\sum_{h=T+1}^{\infty} \lambda_{rh}\lambda_{sh}\Big| \leq \max_{1\leq j\leq p}\Big(\sum_{h=T+1}^{\infty} \lambda_{jh}^2\Big), \quad \text{which implies } d_\infty(\Omega, \Omega_T) = \max_{1\leq j\leq p} a_{jj}^T$$

Now, for a fixed $\epsilon > 0$,

$$\begin{aligned}
\mathrm{pr}\big\{d_\infty(\Omega, \Omega_T) \leq \epsilon\big\} &= E\big\{\mathrm{pr}(a_{11}^T \leq \epsilon, \ldots, a_{pp}^T \leq \epsilon \mid \delta)\big\} \\
&= E\big\{\mathrm{pr}(a_{11}^T \leq \epsilon \mid \delta)^p\big\} > \big[E\{\mathrm{pr}(a_{11}^T \leq \epsilon \mid \delta)\}\big]^p \\
&= \big[1 - E\{\mathrm{pr}(a_{11}^T > \epsilon \mid \delta)\}\big]^p \geq \left[1 - E\left\{\frac{E(a_{11}^T \mid \delta)}{\epsilon}\right\}\right]^p \\
&= \left\{1 - \frac{E(a_{11}^T)}{\epsilon}\right\}^p
\end{aligned}$$

where the equality in the second line follows from the fact that a_{ii}^T are conditionally i.i.d. given δ and the subsequent two inequalities use Jensen's and Chebyshev's inequality respectively. Now

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

where $b = E\big(\delta_1^{-1}\big), a = E\big(\delta_2^{-1}\big) < 1$ 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 $0 < x \leq 1.5$ to get

$$\mathrm{pr}\big\{d_\infty(\Omega, \Omega_T) \leq \epsilon\big\} \geq \exp\left\{\frac{-6pb}{\epsilon(1-a)}a^T\right\} \text{ if } T > \frac{\log(2b/\epsilon(1-a))}{\log(1/a)}$$

Hence

$$\mathrm{pr}\big\{d_\infty(\Omega, \Omega_T) > \epsilon\big\} \leq 1 - \exp\left\{\frac{-6pb}{\epsilon(1-a)}a^T\right\} \leq \frac{6pb}{\epsilon(1-a)}a^T$$

for

$$\frac{6pb}{(1-a)\epsilon}a^T < 1 \text{ or } T > \frac{\log(6pb/\epsilon(1-a))}{\log(1/a)}.$$

Proof of Proposition 2.5

Let Λ_* be a $p\times k$ matrix ($k \le p$) and $\Sigma_0 \in \Theta_\Sigma$ such that $\Omega_0 = \Lambda_*\Lambda_*' + \Sigma_0$. 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 proof of Lemma 2.2. By Lemma 2.2, $g\big\{B_{\epsilon_1}(\Lambda_0, \Sigma_0)\big\} \subset B_\epsilon^\infty(\Omega_0)$, and thus $B_{\epsilon_1}(\Lambda_0, \Sigma_0) \subset g^{-1}\big\{B_\epsilon^\infty(\Omega_0)\big\}$. Now $\Pi\big\{B_\epsilon^\infty(\Omega_0)\big\} = (\Pi_\Lambda \otimes \Pi_\Sigma) \circ g^{-1}\big\{B_\epsilon^\infty(\Omega_0)\big\} \ge \Pi_\Lambda \otimes \Pi_\Sigma(B_{\epsilon_1}(\Lambda_0, \Sigma_0))$. Clearly, $\Pi_\Sigma\{\Sigma : d_\infty(\Sigma, \Sigma_0) < \epsilon_1\} > 0$, so it is enough to show $\Pi_\Lambda\{\Lambda : d_2(\Lambda, \Lambda_0) < \epsilon_1\} > 0$. We have,

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

by the following Lemma.
Lemma: Fix $1 \le j \le p$. For any $\epsilon > 0$, $\mathrm{pr}\big\{\sum_{h=1}^{\infty}(\lambda_{jh} - \lambda_{jh}^0)^2 < \epsilon_1^2/p \mid \delta\big\} > 0$ almost surely.

Proof: Note that $\lambda_{jh}^0 = 0$ for $h > k$. Thus for any $T \ge k$,

$$\begin{aligned}
\mathrm{pr}\big\{\sum_{h=1}^{\infty}(\lambda_{jh} - \lambda_{jh}^0)^2 < \epsilon \mid \delta\big\} &\ge \mathrm{pr}\big\{\sum_{h=1}^{T}(\lambda_{jh} - \lambda_{jh}^0)^2 < \epsilon/2,\ \sum_{h=T+1}^{\infty}\lambda_{jh}^2 < \epsilon/2 \mid \delta\big\} \\
&= \mathrm{pr}\big\{\sum_{h=1}^{T}(\lambda_{jh} - \lambda_{jh}^0)^2 < \epsilon/2 \mid \delta\big\}\,\mathrm{pr}\big\{\sum_{h=T+1}^{\infty}\lambda_{jh}^2 < \epsilon/2 \mid \delta\big\}
\end{aligned}$$

By Theorem 2.4, $\mathrm{pr}(\sum_{h=T+1}^{\infty}\lambda_{jh}^2 < \epsilon/2) \to 1$ as $T \to \infty$, hence we can find $T_0 > k$ such that $\mathrm{pr}(\sum_{h=T_0+1}^{\infty}\lambda_{jh}^2 < \epsilon/2) > 0$ and thus $\mathrm{pr}(\sum_{h=T_0+1}^{\infty}\lambda_{jh}^2 \mid \delta) > 0$ almost surely. The proof is completed by observing that $\mathrm{pr}\big\{\sum_{h=1}^{T}(\lambda_{jh} - \lambda_{jh}^0)^2 < \epsilon/2 \mid \delta\big\} > 0$ almost surely for any $T < \infty$.

Proof of Theorem 2.6

Fix $\epsilon > 0$, $\Omega_0 \in \Theta$. Observe that

$$\log\frac{N(y; 0, \Omega_0)}{N(y; 0, \Omega)} = \frac{1}{2}\log\frac{\det\Omega_0}{\det\Omega} - \frac{1}{2}y'(\Omega_0^{-1} - \Omega^{-1})y,$$

which implies,

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

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 $\mathcal{R}$, 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 $\lambda_1 \leq \ldots \leq \lambda_p$ are the eigenvalues of $\Omega^{-1}\Omega_0$. Since Ω and Ω_0 are both positive definite,

$$0 \leq \lambda_1 \leq \frac{x'\Omega_0 x}{x'\Omega x} \leq \lambda_p,$$

where x is any p dimensional vector with $x'x = 1$. For any $x \in \mathcal{R}^p$ with $x'x = 1$,

$$\Big|\frac{x'\Omega_0 x}{x'\Omega x} - 1\Big| \quad = \quad \frac{|x'\Omega_0 x - x'\Omega x|}{x'\Omega x}$$

Now

$$|x'\Omega_0 x - x'\Omega x| \quad \leq \quad \sum_{i=1}^p \sum_{j=1}^p |w_{ij} - w_{ij}^0||x_i x_j| \leq d_\infty(\Omega_0, \Omega)(\sum_{i=1}^p |x_i|)^2 \leq p d_\infty(\Omega_0, \Omega),$$

and

$$x'\Omega x = x'\Omega_0 x + (x'\Omega_0 x - x'\Omega x) \geq \lambda_{\min}(\Omega_0) + (x'\Omega_0 x - x'\Omega x),$$

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

$$\Big|\frac{x'\Omega_0 x}{x'\Omega x} - 1\Big| \quad = \quad \frac{|x'\Omega_0 x - x'\Omega x|}{x'\Omega x} \leq \frac{p\,\epsilon_3}{\lambda_{\min}(\Omega_0)/2} < \epsilon/p,$$

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

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

which proves Theorem 2.6.

Figure 1. Estimation of regression coefficients in the simulation study. MGPS (thrs) denotes our thresholding approach for variable selection.

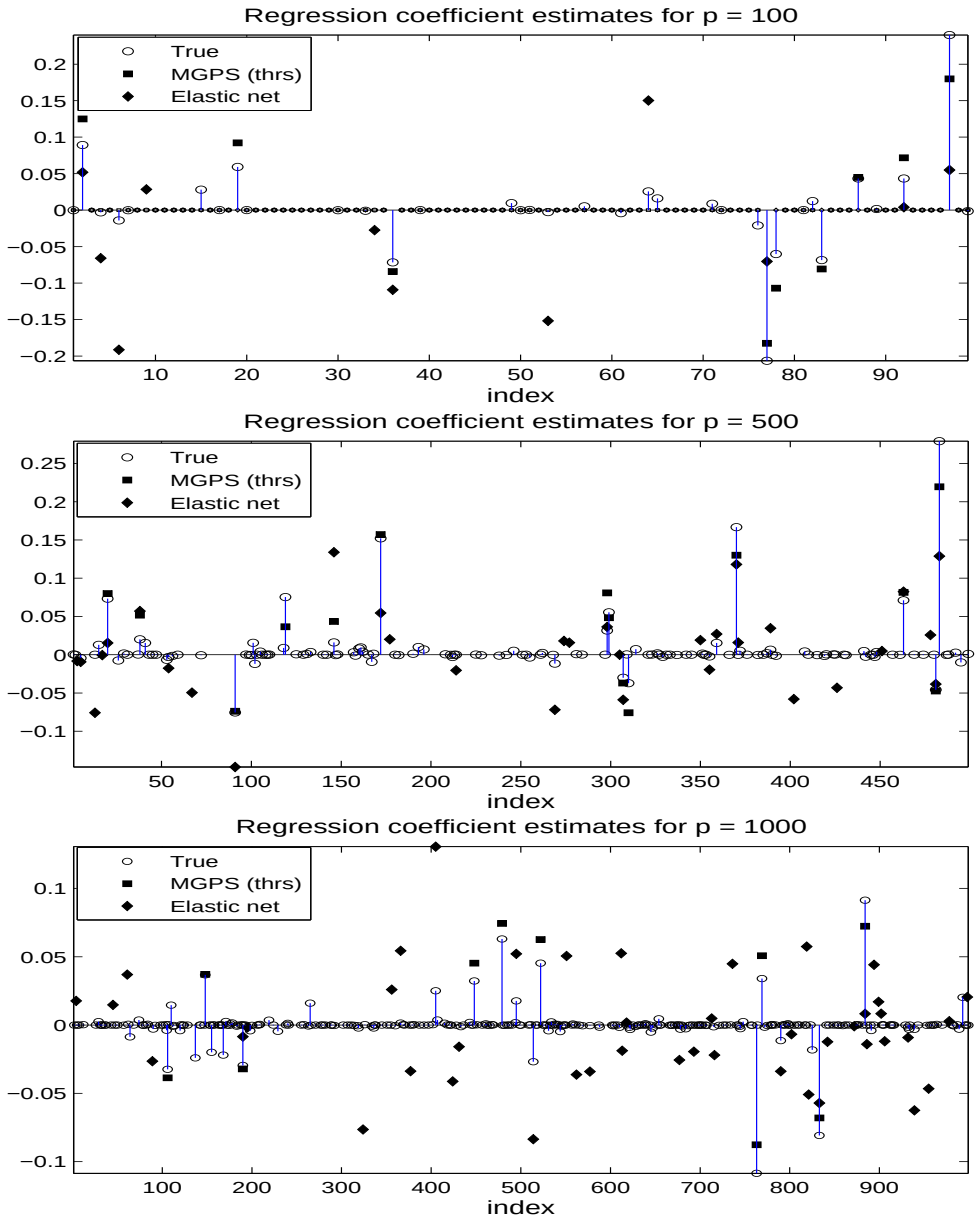

Table 1. Simulation study performance in estimating covariance matrix

true (p, k)	(100, 5)			(500, 10)			(1000, 15)		
method	MGPS	Band	MAP	MGPS	Band	MAP	MGPS	Band	MAP
mse									
mean	0.0022	0.0132	0.0015	0.0010	0.0041	0.0011	0.0009	0.0034	0.0014
min	0.0010	0.0099	0.0006	0.0002	0.0037	0.0005	0.0002	0.0019	0.0005
max	0.0035	0.0162	0.0033	0.0025	0.0046	0.0029	0.0035	0.0052	0.0029
aab									
mean	0.0185	0.0310	0.0097	0.0058	0.0059	0.0033	0.0041	0.0048	0.0030
min	0.0127	0.0249	0.0055	0.0040	0.0055	0.0021	0.0024	0.0043	0.0019
max	0.0246	0.0493	0.0147	0.0092	0.0086	0.0053	0.0062	0.0054	0.0046
mab									
mean	0.5086	1.1130	0.4474	0.9542	1.1777	0.9773	1.1453	1.1463	1.0793
min	0.3878	0.9979	0.2465	0.5027	1.0480	0.6443	0.5255	1.1098	0.7465
max	0.7409	1.3113	1.0534	1.5192	1.3111	1.6243	2.4219	2.3981	2.2136

Table 2. Predictive performance in the simulation study

true (p, k)	(100, 5)			(500, 10)			(1000, 15)		
method	MGPS	Lasso	EN	MGPS	Lasso	EN	MGPS	Lasso	EN
mspe									
mean	0.6285	0.5475	0.5470	0.4075	0.3755	0.3775	0.9492	0.8725	0.8831
min	0.3177	0.3312	0.3294	0.1793	0.2195	0.2180	0.5672	0.5538	0.5588
max	0.8947	0.7908	0.7789	0.8572	0.5694	0.5648	1.4837	1.4389	1.4392
aape									
mean	0.6220	0.5866	0.5859	0.5125	0.4880	0.4887	0.7961	0.7472	0.7493
min	0.4682	0.4679	0.4684	0.3339	0.3758	0.3746	0.5975	0.5892	0.5924
max	0.8453	0.7300	0.7240	0.8043	0.5833	0.5878	0.9892	0.9848	0.9869
mape									
mean	2.1903	2.0676	2.0692	1.7058	1.6634	1.6784	2.5424	2.4807	2.4833
min	1.3647	1.4326	1.3990	1.2104	1.1704	1.1775	1.8341	1.8325	1.8043
max	3.1475	2.9162	2.8931	2.9505	2.6968	2.6227	3.2713	3.0703	3.0744

Table 3. Performance in estimation of the regression coefficients in the simulation study

true (p, k)	(100, 5)			(500, 10)			(1000, 15)		
method	MGPS	Lasso	EN	MGPS	Lasso	EN	MGPS	Lasso	EN
mse	0.0011	0.0012	0.0013	0.0001	0.0003	0.0004	0.0000	0.0001	0.0001
aab	0.0101	0.0124	0.0130	0.0017	0.0039	0.0041	0.0009	0.0018	0.0019
mab	0.1761	0.2073	0.2113	0.1725	0.2533	0.2445	0.1026	0.1090	0.1226

Table 4. Variable selection performance in the simulation study

true (p, k)	(100, 5)			(500, 10)			(1000, 15)		
method	MGPS	Lasso	EN	MGPS	Lasso	EN	MGPS	Lasso	EN
false positives									
mean	0	0.0861	0.0694	0	0.0359	0.0325	0	0.0269	0.0246
min	0	0	0	0	0.0020	0	0	0.0070	0.0070
max	0	0.2626	0.2525	0	0.1363	0.1423	0	0.0851	0.0971
power									
mean	0.7217	0.7596	0.7765	0.7546	0.7610	0.7685	0.7107	0.7195	0.7151
min	0.6768	0.7172	0.7374	0.7295	0.7475	0.7575	0.7047	0.7077	0.7067
max	0.8081	0.7980	0.8283	0.7916	0.7896	0.7916	0.7347	0.7287	0.7247

Table 5. Feature selection in the DLBCL 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

References

O. Aguilar and M. West. Bayesian dynamic factor models and variance matrix discounting for portfolio allocation. *Journal of Business and Economic Statistics*, 18:338–357, 2000.

T. Ando. Bayesian factor analysis with fat-tailed factors and its exact marginal likelihood. *Journal of Multivariate Analysis*, 100:1717–1726, 2009.

G. Arminger and B.O. Muthén. A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika*, 63:271–300, 1998.

P.J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36:199–227, 2008.

C. Carvalho, J. Lucas, Q. Wang, J. Nevins, and M. West. High-dimensional sparse factor modelling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association*, 103:1438–1456, 2008.

G. Celeux, D. Chauveau, and J. Diebolt. Stochastic versions of the EM algorithm: An experimental study in the mixture case. *Journal of Statistical Computation and Simulation*, 55:287–314, 1996.

A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 3:515–534, 2006.

J. Geweke and G. Zhou. Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies*, 9:557–587, 1996.

J. Ghosh and D.B. Dunson. Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18:306–320, 2009.

Tom Griffiths and Zoubin Ghahramani. Infinite latent feature models and the Indian buffet process. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 475–482. MIT Press, Cambridge, MA, 2006.

J. Gui and H. Li. Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *http://repositories.cdlib.org/cbmb/L1Cox/*, 2004.

D. Knowles and Z. Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. In *7th International Conference on Independent Component Analysis and Signal Separation*. 2007.

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

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

H.F. Lopes and M. West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14:41–67, 2004.

Joseph E. Lucas, C. Carvalho, Q. Wang, A. Bild, J.R. Nevins, and M. West. Sparse statistical modelling in gene expression genomics. In K.A. Do, P. Müller, and M. Vannucci, editors, *Bayesian Inference for Gene Expression and Proteomics*, pages 155–176. Cambridge University Press, 2006.

S. Ma and J. Huang. Additive risk survival model with microarray data. *BMC Bioinformatics*, 8:192, 2007.

Edward Meeds, Zoubin Ghahramani, Radford M. Neal, and Sam T. Roweis. Modeling dyadic data with binary latent factors. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 977–984. MIT Press, Cambridge, MA, 2007.

W. Polasek. Factor analysis and outliers: A Bayesian approach. In *Discussion Paper, University of Basel*. 1997.

Piyush Rai and Hal Daumé. The infinite hierarchical factor regression model. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1321–1328. 2009.

G.O. Roberts and J.S. Rosenthal. Coupling and ergodicity of adaptive MCMC. *J. Appl. Prob.*, 44:458–475, 2007.

A. Rosenwald, G. Wright, W. C. Chan, J. M. Connors, E. Campo, R. I. Fisher, R. D. Gascoyne, H. K. Mueller-Hermelink, E. B. Smeland, et al. The use of molecular profiling to predict survival after chemotheropy for diffuse large-B-cell lymphoma. *The New England Journal of Medicine*, 346:1937–1947, 2002.

S.F. Schnatter and H.F. Lopes. Parsimonious bayesian factor analysis when the number of factors is unknown. Technical report, The University of Chicago Booth School of Business, 2009.

L. Schwartz. On Bayes procedures. *Z. Wahrsch. Verw. Gebiete*, 4:1026, 1965.

M.R. Segal. Microarray gene expression data with linked survival phenotypes: diffuse large-B-cell lymphoma revisited. *Biostatistics*, 7:268–285, 2006.

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

R. Thibaux and M. I. Jordan. Hierarchical Beta processes and the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*. 2007.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58:267–288, 1996.

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

H. Zou and T. Hastie. Regularization and variable selection via the Elastic Net. *J. Royal. Statist. Soc B.*, 67:301–320, 2005.