Bayesian low rank and sparse covariance matrix decomposition

Lin Zhang, Abhra Sarkar, Bani K. Mallick
Department of Statistics, Texas A&M University

Abstract

We consider the problem of estimating high-dimensional covariance matrices of a particular structure, which is a summation of low rank and sparse matrices. This covariance structure has a wide range of applications including factor analysis and random effects models. We propose a Bayesian method of estimating the covariance matrices by representing the covariance model in the form of a factor model with unknown number of latent factors. We introduce binary indicators for factor selection and rank estimation for the low rank component combined with a Bayesian lasso method for the sparse component estimation. Simulation studies show that our method can recover the rank as well as the sparsity of the two components respectively. We further extend our method to a graphical factor model where the graphical model of the residuals as well as selecting the number of factors is of interest. We employ a hyper-inverse Wishart prior for modeling decomposable graphs of the residuals, and a Bayesian graphical lasso selection method for unrestricted graphs. We show through simulations that the extended models can recover both the number of latent factors and the graphical model of the residuals successfully when the sample size is sufficient relative to the dimension.

Keywords: Bayesian lasso; covariance matrix; graphical factor model; hyper-inverse Wishart
1 INTRODUCTION

Estimation of covariance matrices is a fundamental issue in multivariate analysis with many statistical applications including modeling genetic data, brain imaging data, climate data, and many other fields. Suppose \( y_1, \ldots, y_n \) are \( q \)-dimensional random vectors which independently follow a multivariate Gaussian distribution \( \mathcal{N}_q(\mu, \Sigma) \). It is well known that the sample covariance \( \hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})' \) is not a stable estimator of the population covariance matrix, \( \Sigma \), when the dimension of the covariance matrix is large relative to the sample size.

A number of approaches have been proposed for a stable estimation of the high-dimensional covariance matrix efficiently from a frequentist perspective. They are mainly based on developing regularized estimators through banding or tapering of the sample covariance matrix (Bickel and Levina, 2008; Wu and Pourahmadi, 2010). Alternative regularization methods have been developed by banding the cholesky or inverse cholesky factors (Wu and Pourahmadi, 2003; Huang et al., 2006; Levina et al., 2008); thresholding the sample covariance matrix (Bickel and Levina, 2008A; Cai and Liu, 2011) or regularizing the principal component analysis (Johnstone and Lu, 2009). Leonard and Hsu (1992), Chiu et al. (1996), and Deng and Tsui (2010) modeled the matrix logarithm of the covariance matrix. Ledoit and Wolf (2004) constructed a shrinkage estimator which is a linear combination of the sample covariance matrix and a pre-chosen matrix. Alternative parsimonious modeling methods by identifying zero off-diagonal elements in the covariance matrix or its inverse have been proposed by Yuan and Lin (2007), Friedman et al. (2008), Levina et al. (2008), Bien and Tibshirani (2011), among others. In a Bayesian framework, Wong et al. (2003) used a selection prior for off-diagonal elements of the partial correlation matrix to identify zeros in the inverse covariance matrix. Talluri et al. (2009) and Wang (2012) used the Bayesian graphical lasso priors for sparse inverse covariance matrix estimation. Furthermore, the hyper-inverse Wishart (HIW) prior was employed for covariance selection given a decomposable Gaussian graphical model (Lauritzen, 1996; Giudici and Green, 1999; Armstrong et al., 2009), which was extended for nondecomposable graphical models by Giudici and Green (1999), Roverato (2002), Brooks et al. (2003), and Atay-Kayis and Massam (2005).
Luo (2011) introduced a different covariance structure for high-dimensional datasets, which can be decomposed into the summation of a low-rank and sparse matrix as

\[ \Sigma = L + S, \]  

(1)

where \( L \) is a low rank symmetric component and \( S \) is a sparse symmetric positive definite component. This decomposition of covariance matrices for dimension reduction has a wide range of applications including factor analysis, random effects and conditional covariance models, where the low rank component \( L \) indicates that the variation of the random vector can mostly be explained by a small number of common factors or principal components, and the sparse part \( S \) displays the variance/covariance between the variables conditional on these latent factors. The covariance matrices in the models are not sparse in themselves, and hence banding, thresholding, or other parsimonious modeling methods are not appropriate to apply directly to these models. However, parsimonious modeling could be used for the covariance of the residuals after removing latent factors. In contrast to the popular principal component or low rank approximation methods, this decomposition method has the ability to provide a full rank covariance estimator. Luo (2011) proposed a frequentist approach LOREC, which regularizes \( \hat{\Sigma} \) by the Frobenius norm and uses a composite penalty on the trace norm of \( L \) and the \( l_1 \) norm of \( S \) to achieve the low-rank and sparse component estimation, respectively. This method provides a point estimate of the covariance matrix ignoring uncertainties from different sources. In this paper, we propose a Bayesian approach to estimate the covariance matrix with the decomposition structure in equation (1). We also extend the method for graphical factor analysis.

We represent the \( q \times q \) low-rank matrix \( L \) utilizing a singular value decomposition (SVD) as follows

\[ L = M D_\tau M^T, \]  

(2)

where \( M \in \mathbb{R}^{q \times r^*} \), the diagonal matrix \( D_\tau = \text{diag}(\tau_1^2, ..., \tau_{r^*}^2) \in \mathbb{R}^{r^* \times r^*} \) consists of singular values of \( L \), and \( r^* \) denotes the true rank of \( L \). This decomposition ensures positive definiteness of \( \Sigma \) and determines the rank of \( L \), which is given by the dimension of \( D_\tau \). When \( L \) is low-rank, \( r^* << q \). This representation of the covariance matrix has the same structure as in a factor analytic model, where \( M \) could be viewed as the latent factor loadings matrix, the singular values \( \tau_k^2 \)'s as the variances of the latent factors. In our Bayesian method, we jointly estimate \( L \) and \( S \) through
the factor analytic model, which has an unknown number of factors and a sparse covariance matrix of the residuals. To estimate the rank of \( L \), i.e. the number of true factors, we introduce a binary indicator for each factor separating the factor selection and the singular value estimation, and we assign conjugate type priors such that most of the parameters have closed-form conditional posterior distributions, which facilitates the Markov chain Monte Carlo (MCMC) computation and also allows an automatic choice of the regularization parameters.

In statistical applications such as gene expression and financial data analysis, we often are interested in the graphical model of the variables, which is flagged by the zero pattern in the off-diagonal of \( S \) inverse. In these cases, the sparsity in \( C = S^{-1} \) instead of \( S \) is desired for graphical model inference. For this purpose, we extend our method to a graphical factor model so that it achieves sparsity in estimating \( C \). We employ a conjugate HIW prior on \( S \) in the graphical factor model when \( S \) is restricted to a decomposable graph. We further extend the model for unrestricted graphs by using a Bayesian graphical lasso selection prior. We show through simulations that the extended model can recover both the number of latent factors and the graphical model of the residuals successfully when the sample size is sufficient relative to the dimension.

The rest of the paper is organized as follows. In Section 2 we describe our proposed Bayesian model for low rank and sparse covariance matrix estimation. In Section 3 we report results from simulation studies to assess the operating characteristics of our method. A real data analysis of gene expression data is included in Section 4. In Section 5 we extend our method to develop a graphical factor model to handle data where selection of the latent factors as well as inference of the graphical model among variables are both of interest, and show the application of the graphical factor model to a gene expression dataset. We provide a discussion and conclusion in Section 6.
2 BAYESIAN LOW RANK AND SPARSE COVARIANCE MODEL

2.1 Likelihood Model

Consider a $q \times n$ data matrix $y$, with each column vector $y_i$ for $i = 1, \ldots, n$ following an identical and independent (iid) Gaussian distribution

$$y_i \sim N_q(0, \Sigma). \quad (3)$$

We assume that the covariance matrix $\Sigma$ could be decomposed as a sum of a low rank component $L$ and a sparse component $S$, with $L$ to be represented in the form of SVD as in formula (2). Hence we have

$$\Sigma = MD_\tau M^T + S. \quad (4)$$

Note that the representation of the covariance in formula (4) can be viewed from the standpoint of a latent factor analytic model:

$$y_i = Mf_i + \epsilon_i, \quad (5)$$

where $f_i = (f_{1i}, \ldots, f_{ri})^T$ is the value of an $r^*$-dimensional random vector of latent factors in the $i^{th}$ replicate, $M$ is the $q \times r^*$ latent factor loadings matrix, the diagonal elements in $D_\tau$ are the variances of $f_i$, and $S$ is the covariance matrix of the $\epsilon_i$.

Factor analytic models have been extensively studied for summarizing the variance and covariance patterns in multivariate data. With advances in computational tools, Bayesian methods for factor analysis have been rapidly developed. See, for example, Geweke and Zhou (1996), Aguilar and West (2000), and Rowe (2003) among others. Lopes and West (2004) explored the inference on the number of latent factors in a factor model with a reversible jump MCMC method. Other recent Bayesian factor analysis methods incorporated different modeling structures through the columns of the factor loadings matrix (Lopes and Carvalho, 2007; Carvalho et al., 2008). Hoff
(2007) proposed a model based singular value decomposition method introducing the unknown rank scenario. Using the above decomposition, Fan et al. (2008), Hoff and Niu (2012) proposed a regression type estimator assuming \( f_i \) is observable. However, most of the methods assume that the covariance matrix of the residuals \( S \) is diagonal. That is, all the association among the observed variables are exclusively contributed to the latent factors.

In our proposed method we assume that \( S \) is a sparse covariance matrix. That is, we allow \( S \) to be nondiagonal so that the variables could be correlated conditional on the latent factors. Grzebyk, Wild and Chouanière (2004) gave a sufficient condition for the identification of a multi-factor model with correlated residuals as in (4). Since we do not know the rank of \( L \), i.e. the number of latent factors \( r^* \) in the model, we need to estimate the number of factors and the variances of the factors jointly. To separate these two tasks, we introduce an extra binary indicator matrix \( Z \). Our proposed Bayesian model is

\[
\Sigma = M(ZD_\tau)M^T + S, \tag{6}
\]

where \( D_\tau \) is a diagonal matrix with positive diagonal elements \( \tau_k^2 \) for \( k = 1, \ldots, r \) for some \( r > r^* \), and \( Z \) is a diagonal matrix with binary entries \( z_k \in \{0,1\} \) for \( k = 1, \ldots, r \) along the diagonal. While \( \tau_k^2 \) gives the variance of the \( k^{th} \) latent factor, the indicator \( z_k \) determines if the latent factor is included in the model. In this way, we separate the recovery of the rank and the estimation of the singular values of \( L \). The rank of \( L \), \( r^* \), is only determined by the number of 1's in the diagonal of \( Z \). Now the estimation of \( r^* \) is equivalent to selecting the true number of latent factors in a factor analytic model. In our method, we choose a relatively large integer \( r \leq q \) which is supposed to be much larger than \( r^* \), and expect that the diagonal entries of \( Z \) are sparse. If the estimates of \( Z \) diagonals are not sparse, we increase the value of \( r \).

We can rewrite the likelihood model in (3) and (6) as the regression-type representation of a latent factor analysis model:

\[
\begin{align*}
\mathbf{y}_i &= M\mathbf{Z}\mathbf{f}_i + \mathbf{\epsilon}_i, \quad \text{for } i = 1, \ldots, n \\
\mathbf{f}_{ki} &\sim \mathcal{N}(0, \tau_k^2), \quad \text{for } k = 1, \ldots, r \\
\mathbf{\epsilon}_i &\sim \mathcal{N}_q(0, S),
\end{align*}
\tag{7}
\]
If $z_k = 1$, the $k^{th}$ factor, $k = 1, \ldots, r$, is a true factor of the variables $Y$ with variance $\tau_k^2$; otherwise, the $k^{th}$ factor is not included in the factor model. By rewriting the model in the form of a linear regression problem, we can assign conjugate priors to $M$, $Z$ and $D_r$, which leads to closed-form full conditional distributions of the parameters and facilitates the posterior sampling using a Gibbs algorithm.

### 2.2 Prior Specification

#### 2.2.1 Low Rank Component

To complete model specification, we need to assign priors to the set of parameters $\{M, Z, D_r, S\}$ in the hierarchical likelihood model (7), where $M$, $Z$, $D_r$ give the low rank component $L$, and $S$ is the sparse component. Let $m_{jk}$ be the element of $M$ in the $j^{th}$ row and $k^{th}$ column, and $M_k = (m_{1k}, \ldots, m_{qk})^T$ be the $k^{th}$ column vector of $M$, which could be viewed as the loading vector of the factor $k$ on the variables. We assume that $M_k$ has a Gaussian prior:

$$M_k \sim N_q \left(0, \frac{1}{q} I_q \right), \quad k = 1, \ldots, r,$$

where $I_q$ is a $q \times q$ identity matrix. Note that for large values of $q$, the columns of $M$ are approximately orthogonal. Furthermore, we notice that $M$ and $D$ are unidentifiable. By assigning the prior variance of $m_{ik}$ to be $1/q$, we reduce the variability of $m_{ik}$, shift the variability to the single element $\tau_k^2$ in $D_r$, and obtain a relatively stable estimate of $L$.

The binary diagonal matrix $Z$ is modeled as

$$z_k \sim \text{Bernoulli}(p_k), \quad k = 1, \ldots, r,$$

where $p_k$ is the prior probability of $z_k = 1$. The values of $p_k$ determine the strength of the penalization that is assigned to the rank of $L$ as $r^* = \text{rank}(L)$ is equivalent to the number of $z_k$’s that equal 1. Since $L$ is assumed to be low-rank, most of the prior probabilities are expected to be small or zero. We model these probabilities with the following hyper-prior distribution:

$$p_k \sim (1 - \pi) \mathcal{I}\{p_k = 0\} + \pi \text{Beta}(a_p, b_p),$$
where $\mathcal{I}(\cdot)$ is the indicator function. The hyper-prior of $p_k$ is a Beta distribution mixed with a point mass at 0 with probability $\pi$, where $\pi$ is drawn from a Beta prior distribution

$$
\pi \sim \text{Beta}(a_\pi, b_\pi).
$$

The sparseness of the diagonals of $Z$ is explicitly imposed through the hyperparameters $(a_p, b_p)$ and $(a_\pi, b_\pi)$. When $a_p/(a_\pi + b_\pi) \ll 1$, $p_k$ has a high prior probability to be zero; when $a_p/(a_\pi + b_\pi) \ll 1$, $p_k$ is still likely to be close to zero, if it is not exactly zero. To impose high penalization on the rank of $L$, we choose $(a_\pi, b_\pi) = (1/q, 1 - 1/q)$ and $(a_\pi, b_\pi) = (1, r)$.

Each diagonal entry $\tau_k^2$ in $D_\tau$ corresponds to the variance of the $k^{th}$ factor. We can assign a conjugate Inverse-Gamma prior

$$
\tau_k^2 \sim \text{IG}(a_\tau, b_\tau), \quad k = 1, \ldots, r,
$$

which leads to a closed form of the posterior conditional distribution.

### 2.2.2 Sparse Component

In order to achieve adaptive shrinkage of the sparse component $S$, we use a Bayesian lasso prior for the sparse $S$ estimation. In the graphical lasso method (Yuan and Lin 2007; Friedman et al. 2008), an $l_1$ penalty term is assigned to $S$, which, in a Bayesian framework, is equivalent to a prior distribution as follows

$$
S|\lambda \propto \prod_{j < j'} \exp\{-\lambda |S_{jj'}|\} \prod_{j = 1}^q \exp\{-\frac{\lambda}{2} |S_{jj}|\} \mathcal{I}\{S \in S^+\},
$$

where $S^+$ is the space of all positive definite $S$, and $\lambda > 0$ is the regularization parameter. The prior is a joint density of exponential distributions on the diagonal elements $S_{jj}$, $j = 1, \ldots, q$, and Laplace distributions on the off-diagonal elements $S_{jj'}$, $j < j'$, with the constraint $S \in S^+$. However, the Bayesian lasso method shrinks but does not set the off-diagonal elements to exact zeros, which is desired for the sparse $S$ estimation. To this end, we modify the method by placing a point mass at 0 in the Laplace prior for the off-diagonal elements. The prior on $S$ is then as follows:

$$
S|\lambda, \rho \sim \text{con}^{-1}_{\lambda, \rho} \prod_{j < j'} \left[(1 - \rho_{jj'})\mathcal{I}\{S_{jj'} = 0\} + \rho_{jj'}\text{Laplace}(\lambda)\right] \cdot \prod_{j = 1}^q \text{Exp}\left(\frac{\lambda}{2}\right) \cdot \mathcal{I}\{S \in S^+\},
$$

8
where Laplace($\lambda$) is a Laplace distribution with $\lambda$ to be the rate parameter, and Exp($\frac{1}{2}$) is an exponential distribution with rate $\frac{1}{2}$. Here, $\text{con}_{\lambda, \rho} = \int_{S \in \mathcal{S}^+} \prod_{j < j'} \left[ (1 - \rho_{jj'})I\{S_{jj'} = 0\} + \rho_{jj'} \text{Laplace}(\lambda) \right] \prod_{j=1}^{q} \text{Exp}(\frac{1}{2}) dS$ is the normalizing constant, which depends on $\lambda$ and $\rho = \{\rho_{jj'}, j < j'\}$, and is analytically intractable.

In this construction, the hyper-parameter $\lambda$ shrinks the covariance elements toward zero, while $\rho_{jj'}$ controls the probability that the $(j, j')$ element will be enforced to be a zero. In the Bayesian framework, we can assign hyperpriors to $\lambda$ and $\rho_{jj'}$ and implement the regularization parameters in the MCMC algorithm for posterior inference. However, since the normalizing term $\text{con}_{\lambda, \rho}$ cannot be evaluated analytically, standard prior distributions on the hyper-parameters will lead to intractable posterior distributions. Following Wang (2012), we consider an extension of the conjugate priors on $\lambda$ and $\rho$

$$(\lambda, \rho) \sim \text{con}_{\lambda, \rho} \Gamma(a_\lambda, b_\lambda) \prod_{j < j'} \text{Beta}(a_\rho, b_\rho).$$

The prior distribution is proper as $\text{con}_{\lambda, \rho} < \int_{S \in \mathbb{R}(q+1)/2} \prod_{j < j'} \left[ (1 - \rho_{jj'})I\{S_{jj'} = 0\} + \rho_{jj'} \text{Laplace}(\lambda) \right] \prod_{j=1}^{q} \text{Exp}(\frac{1}{2}) dS = 1$. Given the joint hyperprior, the full conditional distributions for $\lambda$ and $\rho_{jj'}$ are independent Gamma and Beta distributions, respectively. In our experiments, we specify $(a_\lambda, b_\lambda) = (1, 1)$ for a diffuse prior for $\lambda$, and $(a_\rho, b_\rho) = (0.5, 0.5)$ for a noninformative prior of $\rho_{jj'}$. 
The complete hierarchical model can be summarized as

\[
\begin{align*}
    y_i & \sim N_q(MZf_i, S), \quad i = 1, \ldots, n \\
    f_{ki} & \sim N(0, \tau_k^2), \quad k = 1, \ldots, r \\
    M_k & \sim \frac{1}{q}N_q(0, \frac{1}{q}I_q), \\
    z_k & \sim \text{Bernoulli}(p_k), \\
    p_k & \sim (1 - \pi)I\{p_k = 0\} + \pi\text{Beta}(a_p, b_p), \\
    \pi & \sim \text{Beta}(a_\pi, b_\pi), \\
    \tau_k^2 & \sim \text{IG}(a_\tau, b_\tau),
\end{align*}
\]

(8)

\[
\begin{align*}
    S|\lambda, \rho & \sim \text{con}^{-1}_{\lambda, \rho} \prod_{j<j'} \left[ (1 - \rho_{jj'})I\{S_{jj'} = 0\} + \rho_{jj'}\text{Laplace}(\lambda) \right] \\
    & \quad \cdot \prod_{j=1}^q \text{Exp}(\lambda) \cdot I\{S \in S^+\}, \\
    (\lambda, \rho) & \sim \text{con}_{\lambda, \rho}\text{Gamma}(a_\lambda, b_\lambda) \prod_{j<j'} \text{Beta}(a_\rho, b_\rho),
\end{align*}
\]

(10)

where \( j \) denotes the variable, \( i \) denotes the sample, \( k \) denotes the latent factor, \( j = 1, \ldots, q \), \( i = 1, \ldots, n \), and \( k = 1, \ldots, r \).

2.3 Posterior Inference

In this section, we present the full conditional posterior distributions and a framework to carry out MCMC calculations. Note that the full conditional distributions of most parameters are available in a closed form, allowing for a straightforward Gibbs sampling algorithm, except for the off-diagonal elements in the sparse matrix \( S \). In this case, we employ an independent Metropolis Hastings (MH) sampler within the Gibbs sampler to generate posterior samples of the off-diagonal elements in \( S \).

1. Sampling the factor loadings matrix \( M \):

Let \( M_k = (m_{1k}, \ldots, m_{qk})^T \) be the \( k^{th} \) column vector of \( M \), and \( M_{(-k)} \) be the matrix of \( M \) excluding the \( k^{th} \) column. The full conditional distribution of \( M_k \) for \( k = 1, \ldots, r \) is

\[
M_k|y, M_{(-k)}, f, S \sim N_q(\mu_k^M, \Sigma_k^M),
\]
where \( \Sigma^M_k = \left\{ S^{-1}(\sum_{i=1}^{n} f_{ki}^2) + qI_q \right\}^{-1} \), and \( \mu^M_k = \Sigma^M_k S^{-1}(y - M(-k)f_{(-k)} \cdot f_k^T) \). Hence we can draw samples of each column of \( M \) from a multivariate Gaussian distribution.

2. **Sampling the random factors \( f \):**

Let \( f \) be the \( r \times n \) matrix with \( f_{ki} \) to be the value of \( k^{th} \) factor in \( i^{th} \) replicate. Then \( f_k = (f_{k1}, \ldots, f_{kn}) \) is the \( k^{th} \) row vector of \( f \), and \( f_{(-k)} \) denotes the matrix of \( f \) excluding the \( k^{th} \) row. Note \( f_k \) could be viewed as the unobserved values of the random factor \( k \) in the \( n \) replicates. The full conditional distribution of the transpose of \( f_k \) for \( k = 1, \ldots, r \) is

\[
f_k^T | y, M, f_{(-k)}, S, z_k, \tau^2_k \sim (1 - z_k)N_q(0, \tau^2_k I_n) + z_k N_n(\mu^f_k, \sigma^f_k I_n),
\]

where \( \sigma^f_k = \left( M_k^T S^{-1} M_k + \tau^2_k \right)^{-1} \), and \( \mu^f_k = \sigma^f_k \left\{ M_k^T S^{-1}(y - M(-k)f_{(-k)} \cdot f_k^T) \right\}^T \). Hence we can draw samples of each row of \( f \) from a mixture of two multivariate Gaussian distributions.

3. **Sampling the binary diagonal matrix \( Z \):**

The full conditional of each diagonal element of \( Z \), \( z_k \), is a Bernoulli distribution:

\[
z_k | y, M, f, S, \tau^2_k, p_k \sim Bernoulli(p^*_k),
\]

where \( p^*_k = 1 - (1 - p_k) / \left\{ 1 - p_k + p_k \cdot \left( \frac{\sigma^f_k}{\tau^2_k} \right)^2 \cdot \exp \left( -\frac{1}{2} \frac{(\mu^f_k)^T \mu^f_k}{\sigma^f_k} \right) \right\} \).

4. **Sampling the probabilities \( p_k \) and \( \pi \):**

The full conditional of \( p_k \) for \( k = 1, \ldots, r \) is

\[
p_k | z_k = 1 \sim Beta(a_p + 1, b_p),
\]

\[
p_k | z_k = 0 \sim \left( 1 - \pi^* \right) I\{ p_k = 0 \} + \pi^* Beta(a_p, b_p + 1),
\]

where \( \pi^* = \frac{a_b}{a_p + b_p - a_p} \).

The full conditional of \( \pi \) is

\[
\pi \sim Beta \left( a_\pi + \sum_k I\{ p_k = 0 \}, b_\pi + \sum_k I\{ p_k \neq 0 \} \right).
\]

5. **Sampling the positive diagonal matrix \( \tau^2 \):**

The full conditional of each diagonal element of \( \tau^2_k \), for \( k = 1, \ldots, r \) is

\[
\tau^2_k | z_k, \pi \sim (1 - z_k)IG(a_\tau, b_\tau) + z_k IG \left( a_\tau + \frac{n}{2}, b_\tau + \frac{f_k f_k^T}{2} \right).
\]
6. Sampling the diagonal elements $S_{jj}$:

The full posterior conditional density of the diagonal element $S_{jj}$, for $j = 1, \ldots, q$, is

$$p(S_{jj} | \cdot) \propto \det(S)^{-\frac{d}{2}} \exp \left\{ -\frac{\text{tr}(\Lambda S^{-1})}{2} - \frac{\lambda S_{jj}}{2} \right\} \mathcal{I}\{S \in S^+\},$$

where $\Lambda = (y - Mf)(y - Mf)^T$. Without loss of generality, suppose that $j = q$. By the properties of matrix inverse and matrix determinant, we have

$$p(S_{jj} | \cdot) \propto (S_{jj} - c)^{-\frac{d}{2}} \exp\left\{ -\frac{d}{2}(S_{jj} - c)^{-1} - \frac{\lambda}{2} S_{jj} \right\} \mathcal{I}\{S_{jj} > c\},$$

where $c = S_{j,-j}S_{j,-j}^{-1}S_{j,j}$, and $d = [S_{j,-j}S_{j,-j}^{-1} - 1] : \Lambda : [S_{j,-j}S_{j,-j}^{-1} - 1]^T$. Note that the indicator function $\mathcal{I}\{S_{jj} > c\}$ ensures the positive definiteness of $S$ conditional on the other elements in $S$. The derivation of the full conditional distributions for the elements of $S$ is detailed in Appendix. The above distribution is in a closed form. We now transform $S_{jj}$ to $\nu = S_{jj} - c$, then the conditional density of $\nu$ is

$$p(\nu | \cdot) \propto \nu^{-\frac{d}{2}} \exp\left\{ -(d/\nu + \lambda \nu)/2 \right\} \mathcal{I}\{\nu > 0\},$$

which is a generalized inverse Gaussian (GIG) distribution with parameters $(1 - n/2, d, \lambda)$. Therefore, at each MCMC iteration, we can draw a sample of $\nu$ from the GIG distribution, and obtain $S_{jj} = \nu + c$.

7. Sampling the off-diagonal elements $S_{jj'}$:

The full posterior conditional density of the off-diagonal element $S_{jj'}$, $j < j'$ is

$$p(S_{jj'} | \cdot) \propto \det(S)^{-\frac{d}{2}} \exp \left\{ -\frac{\text{tr}(\Lambda S^{-1})}{2} \right\} \left[ (1 - \rho_{jj'})\mathcal{I}\{S_{jj'} = 0\} + \rho_{jj'} \frac{\lambda}{2} \exp(-\lambda |S_{jj'}|) \right] \mathcal{I}\{S \in S^+\}.$$

Without loss of generality, suppose that $j = q - 1$ and $j' = q$. By the properties of matrix inverse and matrix determinant, we have

$$p(S_{jj'} | \cdot) \propto \left\{ 1 - \frac{(S_{jj'} - B_{12})^2}{(S_{jj} - B_{11})(S_{jj'} - B_{22})} \right\}^{-\frac{d}{2}} \cdot \exp \left\{ -\frac{1}{2} \frac{(S_{jj'} - B_{22})D_{11} + (S_{jj} - B_{11})D_{22} - 2(S_{jj'} - B_{12})D_{12}}{(S_{jj} - B_{11})(S_{jj'} - B_{22}) - (S_{jj'} - B_{12})^2} \right\} \cdot \left[ (1 - \rho_{jj'})\mathcal{I}\{S_{jj'} = 0\} + \rho_{jj'} \frac{\lambda}{2} \exp(-\lambda |S_{jj'}|) \right] \cdot \mathcal{I}\{(S_{jj'} - B_{12})^2 < (S_{jj} - B_{11})(S_{jj'} - B_{22})\},$$

where $B = S_{jj',-(jj')}S_{(jj'),jj'}^{-1}S_{-(jj'),jj'}^{-1}$, and $D = [S_{jj',-(jj')}S_{(jj'),jj'}^{-1} - I_2] : \Lambda : [S_{jj',-(jj')}S_{(jj'),jj'}^{-1} - I_2]^T$. We transform $S_{jj'}$ to $\nu = S_{jj'} - B_{12}$ and let $a = S_{jj} - B_{11}, b = S_{jj'} - B_{22}$, then the
conditional density of \( \nu \) is

\[
p(\nu | \cdot) \propto (1 - \rho) g(-B_{12}) \mathcal{I}\{ \nu = -B_{12}, \nu^2 < ab \} + \frac{\rho \lambda}{2} g(\nu),
\]

where \( g(\nu) = (1 - \frac{\nu^2}{ab})^{-n/2} \exp \left\{ -\frac{bD_{11} + aD_{22} - 2D_{12}\nu}{2(ab - \nu^2)} - \lambda |\nu + B_{12}| \right\} \mathcal{I}\{ \nu^2 < ab \} \). Note that the positive definite constraint on \( S \) is ensured by the indicator \( \mathcal{I}\{ \nu^2 < ab \} \). That is, the posterior distribution of \( \nu \) is only over the region \((-\sqrt{ab}, \sqrt{ab})\).

The continuous part of the conditional distribution of \( \nu \), \( g(\nu) \), cannot be sampled directly. Furthermore, \( g(\nu) \) is nonconcave and therefore the sampler may be trapped locally if we use the random-walk MH algorithm within the Gibbs sampling. Since \( g(\nu) \) only has density over \((-\sqrt{ab}, \sqrt{ab})\) and is zero elsewhere, we construct a piecewise uniform proposal distribution approximating \( g(\nu) \).

We choose \( \kappa - 1 \) equally spaced grids between \((-\sqrt{ab}, \sqrt{ab})\), \(-\sqrt{ab} = \nu_0 < \nu_1 < \cdots < \nu_\kappa = \sqrt{ab} \), which divide the domain of \( \nu \) into \( \kappa \) intervals of width \( 2\sqrt{ab}/\kappa \). The piecewise uniform is as follows:

\[
g_a(\nu) = \begin{cases} 
g(\frac{\nu_0 + \nu_1}{2}) & \text{if } \nu_0 < \nu \leq \nu_1 \\
g(\frac{\nu_1 + \nu_2}{2}) & \text{if } \nu_1 < \nu \leq \nu_2 \\
& \cdots \\
g(\frac{\nu_\kappa-1 + \nu_\kappa}{2}) & \text{if } \nu_{\kappa-1} < \nu < \nu_\kappa \end{cases}
\]

The independent MH proposal for generating \( \nu = S^c_{j,j'} - B_{12} \) is given by

\[
q(\nu | \cdot) \propto (1 - \rho) g(-B_{12}) \mathcal{I}\{ \nu = -B_{12}, \nu^2 < ab \} + \frac{\rho \lambda}{2} g_a(\nu).
\]

Samples of \( \nu \) could be generated using an inverse-CDF method from the proposal distribution and the proposal \( \nu^* \) is accepted with the probability

\[
\alpha = \min \left\{ 1, \frac{p(\nu^*)}{q(\nu^*)} / \frac{p(\nu^c)}{q(\nu^c)} \right\},
\]

where \( \nu^c = S^c_{j,j'} - B_{12} \) denotes the current state of \( \nu \). The piecewise uniform proposal distribution avoids the local-trap problem and can be sampled easily using an inverse-CDF method. Furthermore, \( q(\nu | \cdot) \) approximates the distribution \( p(\nu | \cdot) \) more accurately with the increases of the number of grids. Based on our simulations, 100 grids are enough for a fast convergence of \( S^c_{j,j'} \).
8. **Sampling the shrinkage parameter** $\lambda$:

The full conditional of $\lambda$ is

$$\lambda | S \sim \text{Gamma}(a_\lambda + m, b_\lambda + \sum_{j < j'} |S_{jj'}| + \frac{1}{2} \sum_{j=1}^{q} S_{jj}),$$

where $m$ equals to size of the set $\{(j, j') : j \leq j', S_{jj'} \neq 0\}$.

9. **Sampling the selection parameters** $\rho_{jj'}$:

The full conditional of $\rho_{jj'}$ for $j < j'$ is

$$\rho_{jj'}|S \sim \text{Beta}(a_\rho + I\{S_{jj'} \neq 0\}, b_\rho + I\{S_{jj'} = 0\}).$$

**3 SIMULATION STUDIES**

We conducted a detailed simulation study to evaluate the operating characteristics of our method. We considered three covariance models to generate the data:

- **Model 1**: $\Sigma = MDM^T + I$, where $M \in \mathbb{R}^{q \times 3}$ with orthonormal columns, and $D = \text{diag}(8, 8, 8) \cdot (q/n)$. This covariance model comes from a factor model with independent residuals.

- **Model 2**: $\Sigma = 0.311^T + S$, where $S$ is block diagonal with each square block matrix $B$ of dimension 5, and $B = 0.711^T + 0.3I$. This covariance matrix simulates a random effect model, with the covariance of the residuals to be block diagonal.

- **Model 3**: $\Sigma = MDM^T + S$, where $M$ and $D$ are the same as in model 1, and $S$ is a block diagonal matrix as in model 2. This covariance model comes from a factor model with a block diagonal covariance matrix of the residuals.

For each model, 50 observations were generated from the multivariate Gaussian distribution $\mathcal{N}_q(0, \Sigma)$ with varying dimensions $q = 50, 100, \text{and } 200$. We compared our proposed Bayesian model for low-rank and sparse covariance decomposition with the frequentist LOREC method (Luo, 2011).
in estimating the covariance matrices as well as recovering the rank of \( L \) and sparsity of \( S \). The estimates of the parameters using the Bayesian method were based on the posterior samples of 10,000 iterations after 5000 burn-in iterations. The number of latent factors was estimated by the mode of its posterior distribution, and the support of the sparse \( S \) was selected by a false discovery rate (FDR)-based model selection algorithm (Bonato et al. 2010) with an FDR of 0.20. The tuning parameters for the LOREC estimators are picked by 5-fold cross validation using the Bregman divergence loss as in Luo (2011).

Table 1 compares the performance of covariance estimation using our Bayesian method, Luo’s LOREC method, and the sample covariance over 20 replicates measured by the \( l_1 \) norm and the Frobenius norm. The two matrix norms are defined as follows: Let \( X = (X_{ij}) \) be any matrix; \(|X|_1 = \sum_i \sum_j |X_{ij}| \) gives the \( l_1 \) norm, and \(|X|_F = \sqrt{\sum_i \sum_j X_{ij}^2} \) gives the Frobenius norm.

The Bayesian method overall is comparable to the LOREC estimator and is better than the sample covariance estimator. While the LOREC estimator performs better for the random effect model, our Bayesian estimator has lower loss for the factor model with independent residuals. Both of them are better than the sample covariance in all models.

Table 2 summarizes the inference results in the recovery of rank of \( L \) and sparsity of \( S \). The table shows that the Bayesian estimator can recover the true rank of the low rank components with high frequencies for all the three models, with the successful recovery rates close to the LOREC estimator. Furthermore, our method has much lower false positive rates in support recovery of \( S \) when \( S \) is non-diagonal, at the price of a little higher false negative rates. The above results indicate that our method can recover both the rank and the sparsity of the two components with high frequencies.

4 COVARIANCE ESTIMATION FOR GENE EXPRESSION DATASET

In this section, we applied the Bayesian low rank and sparse decomposition model to estimating the covariance of a gene expression dataset from Stranger et al. (2005). The dataset is composed of
60 unrelated individuals of Northern and Western European ancestry from Utah (CEU) of which gene expression levels were measured throughout the genome. We considered 100 genes in the dataset which are most variable among all the genes available in the gene expression profile. Thus we had \( n = 60 \) and \( q = 100 \) in our dataset.

We estimated the covariance matrix using our Bayesian method and compared it to the LOREC estimator and the sample covariance. Figure 1 shows a heatmap showing the absolute intensities of three covariance matrix estimates. Compared to the sample covariance matrix, the LOREC estimator regulates the sample covariance estimate by shrinking all the off-diagonal elements uniformly, whereas the Bayesian decomposition model shrinks more of the elements with strong signals on the top left corner while keeps the abundant elements with moderate signals at the same time.

The Bayesian method and the LOREC estimator also decompose the covariance matrix into low rank and sparse components in the gene expression data. The LOREC estimator identifies a rank 1 component, and our Bayesian method identifies a low rank component of rank 2. The singular vector of a rank 1 component is equivalent (up to a multiplying constant) to the loading in a single factor model, and therefore we obtain the single loading vector from the rank 1 component of the LOREC estimator. We also obtain the loadings matrix corresponding to the two random factors identified by our Bayesian decomposition model. Figure 2 shows the scatter plots of the two loadings by the Bayesian model versus the single loading by the frequentist LOREC method, one of which corresponds to a correlation close to 1. It suggests that the Bayesian model identifies one latent factor with a similar loading as the LOREC estimator.

Figure 3 displays the sparse support of the residual covariance component obtained by the Bayesian decomposition method as well as the LOREC estimator. The LOREC estimator identifies nonzero correlations predominantly on the left corner of the sparse component, whereas the Bayesian decomposition model detects nonzero correlations overspreading the sparse matrix. This difference in the support of the sparse component agrees with the patterns in the covariance estimators plotted in Figure 1.
In many applications such as gene expression data and financial data analysis, we are often interested in the inference of the graphical Markov model as they represent the conditional dependence associations among the set of observed variables. In these applications, it is generally accepted that the variation of a variable is directly regulated by a small subset of other variables, and many approaches including HIW and graphical lasso have been proposed for sparse estimation of the inverse covariance or concentration matrix as mentioned at the beginning of the paper. However, the associations among the observed variables could be, in many cases, partially contributed to some unobserved latent factors. For example, in gene expression data analysis, the genes included in the data might be commonly regulated by certain unknown genes or environmental factors lying upstream of the signaling pathway, whose levels cannot be measured. The existence of such latent factors could mask the true conditional independence structures among the observed variables. For this purpose, we are interested to identify the latent factors and infer the sparse graphical Markov model among the observed variables after removing the effects of these latent factors.

As shown in the likelihood model (7), we represented the covariance model as a latent factor model, where we assume that the covariance matrix $S$ of the residuals $\epsilon_i$ is sparse, and use a Bayesian lasso selection prior to achieve sparsity in $S$. In this section, we extend the latent factor model given by (7) by assuming that the concentration matrix of the residuals, $C = S^{-1}$, is sparse, whose nonzero pattern in the off-diagonal elements gives the conditional dependence arising out of a graphical model of the residuals. The extended latent factor model with sparse $C$ is a sparse graphical factor model with the number of factors unknown.

Giudici (2001) introduced the concept of a graphical factor model, which generalizes factor analytic models by allowing the concentration matrix of the residuals to have non-zero off-diagonal elements. He used an HIW prior (Dawid and Lauritzen 1993) for inference on the concentration matrices restricted to decomposable graphical models, and assigned a uniform prior on all decomposable graphs. We make the following contributions in our proposed model. Firstly, we recover the number of factors as well as the graphical model in a graphical factor model framework. Sec-
ondly, we propose a novel prior on the decomposable graphs for the HIW method, which induces adaptive sparsity in the inferred graphical model. Finally, we extend the method for unrestricted graphs by using a Bayesian graphical lasso selection method. This framework allows for additional flexibility both in the aspect of the analysis of latent factors as well as for graphical models.

5.1 The Graphical Factor Model with Unknown Number of Factors

Before introducing the graphical factor model, we first describe the notations in a graphical model. Let $Y$ be a $q$-dimensional vector of random variables. A conditional independence graph is a pair of $G = (V, E)$ with the vertex set $V = \{1, ..., q\}$ and the edge set $E \subseteq V \times V$. Nodes $j$ and $j'$ are adjacent or connected in $G$ if $(j, j') \in E$, whereas $j$ and $j'$ are conditionally independent if $(j, j') \notin E$. A graph $G$ with $E = V \times V$ is called a complete graph. Complete subgraphs $P \subset V$ are called cliques; the joint subset of two cliques is called a separator denoted by $Q$. If a graph $G$ could be partitioned into a sequence of subgraphs $(P_1, Q_2, P_2, ..., P_K)$ such that $V = \bigcup_k P_k$ and $Q_k = P_{k-1} \cap P_k$ are complete for all $k = 1, ..., K$, $G$ is called a decomposable graph (Lauritzen 1996). For a covariance matrix $S$ of the variables $Y$, let $C = S^{-1}$ be the concentration matrix. Nodes $j$ and $j'$ are conditionally independent, given other nodes, if and only if $C_{jj'} = 0$. Thus, the graph $G$ is given by the configuration of nonzero off-diagonal elements of $C$: $E = \{(j, j') : C_{jj'} \neq 0\}$.

In a classical factor model as in formula (5), there is a pre-specified number of factors $r^*$ and the residuals are assumed to be independent with the covariance matrix $S = \text{diag}(s_{11}, \ldots, s_{qq})$. We relax the two assumptions in our model as follows: (i) the number of underlying factors, $r^*$, is unknown, and is thought to be much smaller than some pre-specified integer $r$. Since the number of common factors is usually small, we pick a moderate to larger value of $r$ and expect only a small fraction of the factors are selected. (ii) The inverse covariance/concentration matrix of the residuals, $C = S^{-1}$, is a sparse matrix allowing to be nondiagonal. That is, we allow nonzero off-diagonal elements in the concentration matrix $C$ so that the unobserved variables could be dependent on each other conditional on the latent factors.

A sufficient condition for identification of a graphical factor model with a single factor and multiple
factors is given in Stanghellini (1997) and Guidici (2001) respectively. However, from a Bayesian viewpoint, identification is of less theoretical concern but more important for posterior computation as discussed in Guidici (2001). When the graphical factor model is unidentifiable (with more than one solutions), the likelihood would be flat, and the posterior distribution of parameters would be multimodal except for extremely informative priors.

Our objective is to select the true number of factors out of \( r \) candidate factors, and to recover the sparse graphical model of the residuals represented by the nonzero pattern in \( C \) as well. For factor selection, as described in Section 2, we introduce binary indicators \( Z \) to determine if each candidate factor is included or excluded from the factor model, and utilize the same modeling method as in the Bayesian covariance decomposition model for factor selection and loadings estimation in the graphical factor mode. That is, we assign the same priors for \( M \), \( Z \), and \( D_\tau \) as in formulas (9). Hence the graphical factor model has the same likelihood as the Bayesian covariance decomposition model in formulas (8), and the same hierarchy for factor estimation as in formulas (9). The difference primarily lies in the assumption and the modeling method for \( S \), the covariance matrix of the residuals. In the following section, we will mainly present the methods of modeling \( S \), whose inverse matrix is assumed to be sparse.

5.2 Bayesian Hierarchical Model for Decomposable Graphs

In this section, we focus on the modeling method when the graphical models of the residuals are restricted to be decomposable. The Bayesian model for unrestricted graphs will be discussed in the next section.

When the graphical model of the residuals is restricted to be decomposable, we allow the covariance matrix of the residuals \( S \) to follow a mixture of HIW priors over decomposable graphs as

\[
S \sim \text{HIW}(G, \delta, \Phi), \\
G \sim p(G),
\]

where \( \delta \in \mathbb{R}^+ \) is a fixed degree-of-freedom parameter, \( \Phi \) is a symmetric positive-definite scale matrix, and \( p(G) \) is the mixing prior over decomposable graphs. The HIW distribution was introduced
by Dawid and Lauritzen (1993) with the probability density function (pdf) given by

$$p(S|G, \delta, \Phi) = \frac{\prod_{k=1}^{K} p(S_{P_k} | \delta, \Phi_{P_k})}{\prod_{k=1}^{K} p(S_{Q_k} | \delta, \Phi_{Q_k})},$$

where $P_k$ and $Q_k$ are the cliques and separators of the graph $G$ respectively. The term $p(S_{P_k} | \delta, \Phi_{P_k})$ denotes the inverse Wishart (IW) density of $S_{P_k} \sim \text{IW}(\delta, \Phi_{P_k})$ with the pdf

$$p(S_{P_k} | \delta, \Phi_{P_k}) \propto |S_{P_k}|^{-(\delta/2+|P_k|)} \exp \left\{ -\frac{1}{2} \text{tr}(\Phi_{P_k}^{-1} S_{P_k}^{-1}) \right\}.$$

The HIW distribution is a conjugate prior distribution for the covariance matrix $S$. Specifically, if $q$-dimensional random vectors $x_1, \ldots, x_n$ follow an iid multivariate normal distribution $\mathcal{N}_p(0, S)$, and $S$ follows $\text{HIW}(G, \delta, \Phi)$, the posterior of $S$ is $S|x, G \sim \text{HIW}(G, \delta + n, \Phi + x^T x)$. The closed form of the posterior distribution for $S$ plays a key part in the posterior inference based on an MCMC algorithm. In our model, we consider $\delta = 3$ as reflecting the lack of prior information on $S$, and specify $\Phi = I_q$.

Let $e_{ij}$ be a binary indicator denoting whether the $(i, j)$ is included or excluded from the graphical model. One option for the mixing prior $p(G)$ is to assign an independent prior probability of an edge, $p(e_{ij})$, to each pair of nodes $(i, j)$, so that

$$\pi(G) = \prod_{(i,j) \in E} p(e_{ij} = 1) \cdot \prod_{(i,j) \notin E} p(e_{ij} = 0).$$

Such prior uses a uniform probability controlling the sparsity of the graph $G$, and requires prior knowledge of the sparsity of the graph to choose the probability parameter. In this paper, we propose a new prior on $G$ which induces adaptive sparsity in the graphical models as:

$$G \propto \exp(-|G|^\xi),$$

where $\xi$ is a positive value penalizing on the size of the graph $G$. Varying $\xi$ penalizes a graph size with different strength. A large value of $\xi$ (> 3) constrains the graph to be extremely sparse, while a value of $\xi$ near zero approximates a uniform prior on all graphs. In the Bayesian method, we assign a uniform prior between 0 and a large value (e.g. 5) on $\xi$, and estimate $\xi$ using an MCMC algorithm. Such choice of prior lets the data choose the intensity of the penalization on the graph size and leads to adaptive sparsity in the inferred graph $G$. 

20
5.3 Bayesian Hierarchical Model for Unrestricted Graphs

In this section, we assume the graphical model of the residuals is unrestricted. The HIW prior was extended for nondecomposable graphical models by Giudici and Green (1999), Roverato (2002), Brooks et al. (2003), and Atay-Kayis and Massam (2005). However, sampling nondecomposable graphs from HIW distributions induces extensive computational burden. In Section 2.2.2, we employ a Bayesian graphical lasso selection prior to achieve sparse estimation of $S$. We now apply the Bayesian graphical lasso selection prior on $C$, the inverse of $S$. As mentioned previously, the graphical lasso prior on a matrix $C$ is equivalent to a joint density of exponential distributions on the diagonal elements $C_{jj}$, $j = 1, \ldots, q$, and Laplace distributions on the off-diagonal elements $C_{jj'}$, $j < j'$. The Bayesian graphical lasso method does not set the off-diagonal elements in $C$ to exact zeros. Hence similarly, we add a point mass at 0 in the Laplace priors. The priors are detailed as follows:

$$C | \lambda^C, \rho^C \sim \text{con}^{-1}_{\lambda^C, \rho^C} \prod_{j<j'} \left[ (1 - \rho^C_{jj'}) \mathcal{I}(C_{jj'} = 0) + \rho^C_{jj'} \text{Laplace}(\lambda^C) \right] \cdot \prod_{j=1}^{q} \text{Exp}\left(\frac{\lambda^C}{2}\right) \cdot \mathcal{I}(C \in C^+),$$

where $C^+$ is the space of all positive-definite concentration matrices, and $\text{con}_{\lambda^C, \rho^C} = \int_{C \in C^+} \prod_{j<j'} \left[ (1 - \rho^C_{jj'}) \mathcal{I}(C_{jj'} = 0) + \rho^C_{jj'} \text{Laplace}(\lambda^C) \right] \prod_{j=1}^{q} \text{Exp}\left(\frac{\lambda^C}{2}\right) dC$ is the normalizing constant, which depends on $\lambda^C$ and $\rho^C = \{\rho^C_{jj'}, j < j'\}$.

Again, we choose an extension of the conjugate priors on $\lambda^C$ and $\rho^C$ to induce computationally tractable full conditional distributions for posterior inference:

$$(\lambda^C, \rho^C) \sim \text{con}_{\lambda^C, \rho^C} \text{Gamma}(a_{\lambda}, b_{\lambda}) \prod_{j<j'} \text{Beta}(a_{\rho}, b_{\rho}).$$

As mentioned previously, the hyper-parameter $\lambda^C$ shrinks the entries in $C$ toward zero, while $\rho^C_{jj'}$ controls the probability that the $(j, j')$ element will be enforced to be a zero. Based on our experiments, we find that noninformative priors on $\rho^C_{jj'}$ would lead to significant inaccuracy in estimating $S$, and influence the factor selection and loadings estimation. In this case, we specify $(a_{\lambda}, b_{\lambda}) = (1, 1)$ for a diffuse prior for $\lambda^C$, and $(a_{\rho}, b_{\rho}) = (1, q)$ for a sparse prior of $\rho^C_{jj'}$. 

21
5.4 Posterior Inference

We derive the full conditionals for all the parameters and perform the posterior inference using a Gibbs sampling algorithm. Note that the prior specification for the parameters \( \{M, z_k, \tau_k\} \) in the graphical factor analysis models parallels the hierarchical model in formulas (9), so the full conditionals of these parameters for a Gibbs algorithm are similar to those in Section 2.3. In this section, we just present the sampling algorithm of the parameter set \( \{S, G, \xi\} \) for a decomposable graph of the residuals, and \( \{C, \lambda^C, \rho^C\} \) for an unrestricted graph of the residuals, in sequence.

5.4.1 Sampling \( \{S, G, \xi\} \) for decomposable graphical models

1. Sampling of \( S \): The full conditional distribution of \( S \) is

\[
S \mid G, y, M, f \sim \text{HIW}(G, \delta + n, \Phi + \Lambda),
\]

where \( \Lambda = (y - Mf)(y - Mf)^T \). Hence we can generate posterior samples of \( S \) directly from the HIW distribution conditional on other parameters.

2. Sampling of \( G \): The conditional distribution of \( G \) is

\[
G \mid y, M, f, \xi \propto h(G, \delta, \Phi) \cdot \exp\{-|G|^\xi\}.
\]

The term \( h(G, \delta, \Phi) \) is the normalizing constant for the HIW \((G, \delta, \Phi)\) distribution given by

\[
h(G, \delta, \Phi) = \prod_{k=1}^{K} \frac{\Phi^{P_k} |\delta + |P_k|-1|^{\frac{P_k}{2}}}{\Gamma|P_k| \left( \frac{\delta + |P_k|-1}{2} \right)^{P_k}} \cdot \prod_{k=2}^{K} \frac{\Phi^{Q_k} |\delta + |Q_k|-1|^{\frac{Q_k}{2}}}{\Gamma|Q_k| \left( \frac{\delta + |Q_k|-1}{2} \right)^{Q_k}}.
\]

where \( \Gamma_p(x) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma(x + (1 - j)/2) \) is the multivariate gamma function. Note that the conditional distribution of \( G \) is marginalized over \( S \) and hence only dependent on \( M, f \) and \( \xi \). This marginalized posterior conditional of \( G \) leads to a collapsed Gibbs algorithm in sampling \( G \), accelerating the graphical model search task. To sample the graph \( G \) from the conditional distribution, we use a random walk MH algorithm within the Gibbs sampling.

Let \( \{e_{jj'} : j < j'\} \) be the set of edge indicators where \( e_{jj'} = 1 \) if \((j, j') \in E\) and \( e_{jj'} = 0 \) otherwise. In an iteration with the current state of graph \( G^c = (V, E^c) \), we choose a pair
(j, j') at random and change the state of the edge, i.e. \( e_{jj'}^p = 1 - e_{jj'}^c \). If the proposed state \( G^p = (V, E^p) \) is decomposable, the proposal is accepted as a new state with the probability \( \alpha(G^c, G^p) = \min \left\{ 1, \frac{p(G^c | y, M^c, f^c, \xi^c)}{p(G^p | y, M^c, f^c, \xi^c)} \right\} \), where \( p(\cdot) \) denotes the posterior conditional distribution of \( G \). If the proposed state \( G^p = (V, E^p) \) is not decomposable, then choose another pair \((j, j')\) until the proposal graph is decomposable.

3. Sampling of \( \xi \): The conditional distribution of \( \xi \) is

\[
\xi | G \propto \exp \left\{ -|G| \xi \right\} I\{\xi \in (0, 5)\}.
\]

We use a random-walk MH algorithm to generate posterior samples of \( \xi \). Given the current state \( \xi^c \), generate a proposal \( \log(\xi^p) \) from a normal distribution \( \mathcal{N}(\log(\xi^c), \sigma^2_\xi) \), with the standard deviation \( \sigma_\xi \) chosen properly. We generate the MCMC samples of \( \xi \) in the log scale to ensure positivity. The proposal \( \xi^p \) is then accepted with the probability

\[
\alpha(\xi^c, \xi^p) = \min \left\{ 1, \frac{p(\xi^p | G^c)}{p(\xi^c | G^c)} \right\},
\]

where \( p(\cdot) \) denotes the full conditional distribution of \( \xi \).

5.4.2 Sampling \( \{C, \lambda^C, \rho^C\} \) for unrestricted graphical models

For convenience, let \( \Lambda = (y - Mf)(y - Mf)^T \).

1. Sampling of the diagonal elements of \( C, C_{jj} \), for \( j = 1, \ldots, q \): The full conditional density of \( C_{jj} \) is

\[
p(C_{jj} | \cdot) \propto (\text{det} C)^{n/2} \exp \left\{ -\frac{1}{2} \Lambda_{jj} C_{jj} - \frac{\lambda^C}{2} C_{jj} \right\} I\{C \in \mathbb{C}^+\}.
\]

Without loss of generality, suppose that \( j = q \). Let \( C = R'R \) be the Cholesky decomposition of \( C \) where the matrix \( R = (R_{jj'}) \) is upper triangular. Then

\[
p(C_{jj} | \cdot) \propto (C_{jj} - c)^{n/2} \exp \left\{ -\left( \frac{\Lambda_{jj}}{2} + \frac{\lambda^C}{2} \right) C_{jj} \right\} I\{C_{jj} > c\},
\]
where $c = \sum_{j=1}^{q-1} R_{j,q}^2$ does not depend on $C_{jj}$, and the indicator function $\mathcal{I}\{C_{jj} > c\}$ ensures $C$ to be positive definite. Let $\nu = (C_{jj} - c)$, then the conditional distribution of $\nu$ is

$$p(\nu \cdot) \propto \nu^{n/2} \exp \left\{ - \frac{\Lambda_{jj}}{2} + \frac{\lambda^C}{2} \nu \right\} \mathcal{I}\{\nu > 0\},$$

which follows Gamma($\frac{a}{2} + 1, \frac{\Lambda_{jj} + \lambda^C}{2}$). Hence, we can draw samples of $\nu$ from the Gamma distribution first, and obtain $C_{jj} = \nu + c$.

2. Sampling of the off-diagonal elements of $C$, $C_{jj'}$, for $j < j'$: The complete conditional density of $C_{jj'}$ is

$$p(C_{jj'} \cdot) \propto (\det C)^{n/2} \exp(-\Lambda_{jj',jj'})p(C_{jj'}|C_j^C, \lambda^C)\mathcal{I}\{C \in C^+\}.$$

Without loss of generality, suppose that $j = q - 1$ and $j' = q$. Then using Lemma 2 of Wong et al. (2003),

$$p(C_{jj'} \cdot) \propto \left\{ 1 - \frac{(C_{jj'} - a)^2}{cb^2} \right\}^{n/2} \exp(-\Lambda_{jj',jj'})p(C_{jj'}|\rho_{jj'}, \lambda^C)\mathcal{I}\{|C_{jj'} - a| < b\sqrt{c}\},$$

$$\propto \mathcal{I}\{|C_{jj'} - a| < b\sqrt{c}\} \cdot \left[ (1 - \frac{\rho_{jj'}}{c^2}) \left\{ 1 - \frac{(C_{jj'} - a)^2}{cb^2} \right\}^{n/2} \mathcal{I}\{C_{jj'} = 0\} + \rho_{jj'} \left\{ 1 - \frac{(C_{jj'} - a)^2}{cb^2} \right\}^{n/2} \exp(-\Lambda_{jj',jj'} - \lambda^C|C_{jj'}|)\mathcal{I}\{C_{jj'} \neq 0\} \right],$$

where $a = \sum_{j=1}^{q-2} R_{j,q-1} R_{j,q}$, $b = R_{q-1,q-1}$, and $c = R_{q-1,q}^2 + R_{q,q}^2$ do not depend on $C_{jj'}$. Now transform $C_{jj'}$ to $\nu = (C_{jj'} - a)/(b\sqrt{c})$, and let $\kappa = -a/(b\sqrt{c})$. The full conditional density of $\nu$ is

$$p(\nu \cdot) \propto (1 - \rho_{jj'}^C)(1 - \kappa^2)^{n/2}\mathcal{I}\{\nu = \kappa, |\nu| < 1\} + \frac{\rho_{jj'}^C \lambda^C}{2} g(\nu),$$

where $g(\nu) = (1 - \nu^2)^{n/2} \exp\{-\Lambda_{jj'}(\nu b\sqrt{c} + a) - \lambda^C|\nu b\sqrt{c} + a|\}\mathcal{I}\{|\nu| < 1\}$. The positive definite constraint on $C$ is ensured by the indicator function $\mathcal{I}\{|\nu| < 1\}$. The continuous part of the conditional distribution of $\nu$, $g(\nu)$, cannot be sampled directly. Since $g(\nu)$ only has density over $(-1, 1)$ and is zero elsewhere, we can use an independent MH algorithm as the sampling algorithm for $S_{jj'}$ in Section 2.3. The details of the independent MH algorithm are explained in Section 2.3. Briefly, we choose $\kappa - 1$ equally spaced grids between $(-1, 1)$, $-1 = \nu_0 < \nu_1 < \cdots < \nu_\kappa = 1$, which divide the domain of $\nu$ into $\kappa$ intervals of width $2/\kappa$,
and construct a piecewise uniform distribution $g_a(\nu)$ approximating $g(\nu)$. The independent MH proposal for generating $\nu$ is then given by

$$q(\nu|\cdot) \propto (1 - \rho_{jj'}^C)(1 - \kappa^2)^{n/2}\mathcal{I}\{\nu = \kappa, |\nu| < 1\} + \frac{\rho_{jj'}^C\lambda^C}{2}g_a(\nu).$$

Samples of $\nu$ could be generated using an inverse-CDF method from the proposal distribution, and the proposal $\nu^*$ is accepted with the probability

$$\alpha = \min \left\{ 1, \frac{p(\nu^*)}{p(\nu^c)} \right\},$$

where $\nu^c = (C_{jj'}^c - a)/(b\sqrt{c})$ denotes the current state of $\nu$. Samples of $C_{jj'}$ are obtained as $C_{jj'} = \nu b \sqrt{c} + a$.

3. Sampling the shrinkage parameter $\lambda^C$:

The full conditional of $\lambda$ is

$$\lambda^C|C \sim \text{Gamma}(a_\lambda + m, b_\lambda + \sum_{j < j'} |C_{jj'}| + \frac{1}{2} \sum_{j=1}^q C_{jj}),$$

where $m$ equals to size of the set $\{(j, j') : j \leq j', C_{jj'} \neq 0\}$.

4. Sampling the selection parameters $\rho_{jj'}^C$:

The full conditional of $\rho_{jj'}^C$ for $j < j'$ is

$$\rho_{jj'}^C|C \sim \text{Beta}(a_\rho + \mathcal{I}\{C_{jj'} \neq 0\}, b_\rho + \mathcal{I}\{C_{jj'} = 0\}).$$

### 5.5 Simulation

In this section, we examined the graphical factor models using simulated datasets. We considered three graphical factor models to generate datasets.

- **Model 4:** An initial simulation study considered a one-factor model for a 30-dimensional random vector. The factor loading vector $M$ was randomly generated with $\|M\| = 1$, the variance of the factor was set to be $\tau^2 = 4$, and the covariance matrix of the residuals $S$ corresponded to an AR(1) model with $S_{jj'} = 0.7^{|j-j'|}$.  

- Model 5: The second simulation considered a two-factor model for a 30-dimensional random vector. The $q \times 2$ factors loadings matrix $M = (M_1, M_2)$ with $|\{j : M_{1j} \neq 0\}| = q/2$, $\{j : M_{2j} \neq 0\} = \{1, \ldots, q\}\setminus\{j' : M_{1j'} \neq 0\}$, and $|M_1| = |M_2| = 1$. The variances of the factors were $\tau^2 = (4, 4)$, and the covariance matrix of the residuals $S$ was a block diagonal matrix with each square block matrix $B$ of dimension 5, and $B = 0.711^T + 0.3I$.

- Model 6: The third simulation considered a one-factor model for a 30-dimensional random vector. The factor loading vector $M$ was randomly generated with $\|M\| = 1$, the variance of the factor was set to be $\tau^2 = 4$, and the covariance matrix of the residuals $S$ corresponds to an nondecomposable graphical model depicted in Figure 4.

We generated datasets with varying sample sizes $n = 100$, and 300 for each model. The proposed HIW based graphical factor model for decomposable graphs was used to recover the number of factors and the graph in model 4 and 5, and the lasso based graphical factor model for unrestricted graphs was used for model 6. The estimates were based on the posterior samples of 10,000 iterations after 5,000 burn-in iterations. The number of latent factors was estimated by the mode of the posterior distribution, and the graphical model was based on the FDR-based model selection method with an overall FDR of 0.20.

Table 3 summarizes the results in estimating the number of factors and the graphical model of the residuals over 20 replications. The results indicate that our Bayesian graphical factor models can recover the number of true latent factors most of the time. Besides, it can recover the graphical model of the residuals with both low rates of false positives and false negatives when the sample size is sufficient relative to the dimension.

5.6 Real Data Analysis

We also applied the graphical factor models to a microarray gene expression dataset from Liu et al. (2011), which was obtained from 176 primary breast cancer patients. We focused on 15 mRNA transcripts whose coding genes are known to lie in the estrogen receptor (ER) pathway. The estrogen pathway regulates a variety of genes and plays key roles in the development or progression
of breast cancer.

We analyzed the data with the two graphical factor models. The HIW based graphical factor model identifies one latent factors, and the lasso based model does not detect any latent factors. The adjacency matrices corresponding to the inferred graphical models by the HIW and lasso based methods are plotted in Figure 5 (a) and (b) respectively, depicting the conditional dependency relationship among the variables. Some of the genes have multiple sets of oligonucleotide sequences on the microarray, and hence have multiple appearances including MYBL1 (MYBL1a, MYBL1b, MYBL1c, MYBL1d) and XBP (XBP1a, XBP1b). The plots of the adjacency matrices show that the HIW based method derived a graph sparser than the graph inferred by the graphical lasso selection method, with the direct regulatory relationships identified in (a) mostly included in (b) as well. The difference in the sparsity level of the graphs and the inferred number of latent factors could be due to the restriction of decomposition in the HIW method.

We conducted a further simulation study based on the results of the above real data analysis. We treated the means of the posterior distributions of the parameters by the lasso based method as the true values, generated simulated data with the same dimension and sample size, i.e. $q = 15$ and $n = 176$, and then reapplied the two graphical factor models to the simulated datasets. We conducted 20 runs of the simulation based on the real data analysis. The adjacency matrices in Figure 5 (c) and (d) show the edges that are selected for over half of the runs by the HIW based and the lasso based graphical factor model respectively. We note that the simulation results are consistent with the real data analysis in that the graphical model identified by the HIW based model is much sparser than the lasso based model, and the inferred conditional associations in the real analysis are mostly re-identified in the simulations.

### 6 DISCUSSION

In this paper, we propose a Bayesian method for estimating covariance matrices of a particular structure, which is a summation of a low rank and a sparse component. Different from the frequentist LOREC method, which is based on the sample covariance estimate only, our Bayesian method
of covariance decomposition is likelihood-based. Hence, it takes the variability of the variables into consideration in case of $p >> n$ and shrinks the covariance elements of varying intensities differently, as indicated in the real data analysis. Furthermore, in stead of the point estimates of the parameters as in the frequentist LOREC, the Bayesian covariance decomposition method yields posterior distributions of the parameters providing a degree of uncertainties in model inference. We model the low rank and sparse component in the form of a latent factor model with correlated residuals. The representation of the decomposable covariance facilitates a Bayesian inference by using conjugate priors on all the parameters except for the off-diagonal elements in the sparse component as well as providing estimates of factor loadings in addition to the estimate of the low rank and sparse component. Simulations indicate that such representation favors the covariance estimation for a latent factor model but does not perform as good as the frequentist LOREC method for a random effect model.

We further extend our method to a graphical factor model, in which we perform inference on both the number of latent factors and the sparse graphical model of the residuals. Simulation studies show that the methods can successfully recover the number of factors as well as the graphical model when the sample size is sufficient relative to the dimension. However, simulations (not presented here) also indicate that in the case of $p >> n$, the methods tend to choose over-sparse graphical models of the residuals. This is reasonable: when the sample size is small, the estimate of the covariance of the residuals $S$ would be inaccurate resulting in significant change in estimating the $S$ inverse. In the high-dimensional low-sample-size condition, the Bayesian methods would choose the most sparse estimate of $S$ inverse that fits the data.

In the graphical factor analysis, we use the HIW method to model a decomposable graph of the residuals and the Bayesian graphical lasso method for an unrestricted graph. One way to determine which method fits the data better in application is to calculate the marginal likelihoods of the inferred models based on the MCMC outputs as discussed in Chib (1995) and Chib and Jeliazkov (2001). Briefly, as the marginal likelihood of a model $M$ is the normalizing constant of the posterior density, we can obtain the logarithm of the marginal likelihood as

$$\log m(y|M) = \log f(y|M, \theta^*) + \log \pi(\theta^*|M) - \log \pi(\theta^*|y, M),$$
where $\theta^*$ is some selected point, such as the posterior mode, of the parameter vector $\theta$ for model $M$. Suppose that the parameter vector could be split into $B$ blocks $\theta^* = (\theta^*_1, \ldots, \theta^*_B)$. Then, the posterior density at the $\theta^*$ could be written as

$$
\pi(\theta^* | y, M) = \pi(\theta^*_1 | y, M)\pi(\theta^*_2 | y, M, \theta^*_1) \cdots \pi(\theta^*_B | y, M, \theta^*_1, \ldots, \theta^*_{B-1}),
$$

where the first term is the marginal ordinate that can be estimated by $\theta_1$’s posterior distribution from the initial Gibbs samples, and the remaining terms are reduced conditional ordinates. Each reduced conditional ordinate, $\pi(\theta^*_t | y, M, \theta^*_1, \ldots, \theta^*_{t-1})$, can be estimated by $\theta_t$’s posterior distribution from a reduced Gibbs MCMC run in which $\theta_1, \ldots, \theta_{t-1}$ are fixed at $\{\theta^*_1, \ldots, \theta^*_{t-1}\}$ and sampling is over $\{\theta_t, \ldots, \theta_B\}$. Since the full conditional distributions of the parameters are explicitly given for the HIW and lasso based graphical factor models, it is straightforward to estimate the reduced conditional ordinates by fixing certain parameter blocks and running additional reduced Gibbs iterations without new programming required. Hence, we can subsequently estimate the posterior ordinate $\pi(\theta^* | y, M)$ and then the marginal likelihood of the inferred models.

**APPENDIX: DERIVATION OF THE CONDITIONAL DENSITIES FOR THE ELEMENTS IN $S$**

**A.1 The Conditional Densities for the Diagonal Elements in $S$**

For convenience, let $\Lambda = (y - Mf)(y - Mf)^T$.

For each diagonal element $S_{jj}$ in $S$, the full conditional density is

$$
p(S_{jj} | \cdot) \propto \det(S)^{-n/2} \exp\{-\frac{1}{2} tr(S^{-1}\Lambda) - \frac{\lambda}{2} S_{jj}\} \mathcal{I}\{S \in S^+\}.
$$

Without loss of generality, suppose that $j = p$. Let $S = \begin{bmatrix} S_{-j,-j} & S_{-j,j} \\ S_{j,-j} & S_{jj} \end{bmatrix}$. With the property of
matrices, we have
\[
\det(S) = \det(S_{-j,-j}) \cdot \det(S_{jj} - S_{j,-j}S_{-j,-j}^{-1}S_{j,j}^{-1}),
\]
\[\propto (S_{jj} - c), \text{ where } c = S_{j,-j}S_{-j,-j}^{-1}S_{j,j}^{-1}.
\]

\[
S^{-1} = \begin{bmatrix}
S_{-j} + S_{-j,-j}^{-1}S_{-j,j}(S_{jj} - c)^{-1}S_{j,-j}S_{-j,-j}^{-1} & -S_{-j,-j}^{-1}S_{j,j}(S_{jj} - c)^{-1} \\
-(S_{jj} - c)^{-1}S_{j,-j}S_{-j,j}^{-1} & (S_{jj} - c)^{-1}
\end{bmatrix},
\]
\[
= \begin{bmatrix}
S_{-j,-j}^{-1} & 0 \\
0 & 0
\end{bmatrix} + \begin{bmatrix}
S_{-j,-j}^{-1}S_{-j,j}^{-1} \\
-1
\end{bmatrix}(S_{jj} - c)^{-1} \begin{bmatrix}
S_{j,-j}S_{-j,j}^{-1} & -1
\end{bmatrix}.
\]

\[
tr(S^{-1} \Lambda) = tr\left(\begin{bmatrix}
S_{-j,-j}^{-1} & 0 \\
0 & 0
\end{bmatrix} \Lambda\right) + tr\left(\begin{bmatrix}
S_{-j,-j}^{-1}S_{-j,j}^{-1} \\
-1
\end{bmatrix}(S_{jj} - c)^{-1} \begin{bmatrix}
S_{j,-j}S_{-j,j}^{-1} & -1
\end{bmatrix} \Lambda\right),
\]
\[
= tr\left(\begin{bmatrix}
S_{-j,-j}^{-1} & 0 \\
0 & 0
\end{bmatrix} \Lambda\right) + d(S_{jj} - c)^{-1},
\]
where \(d = \begin{bmatrix}
S_{j,-j}S_{-j,j}^{-1} & -1
\end{bmatrix} \Lambda \begin{bmatrix}
S_{-j,-j}^{-1}S_{-j,j}^{-1} \\
-1
\end{bmatrix}.
\]

Hence, we have
\[
p(S_{jj}|\cdot) \propto (S_{jj} - c)^{-n/2} \exp\left\{-\frac{d}{2}(S_{jj} - c)^{-1} - \frac{\lambda}{2}S_{jj}\right\} I\{S_{jj} > c\}.
\]

### A.2 The Conditional Densities for the Off-diagonal Elements in S

The conditional density of the off-diagonal element of S, \(S_{jj'}\), is
\[
p(S_{jj'}|\cdot) \propto \det(S)^{-n/2} \exp\left\{-\frac{1}{2}tr(S^{-1} \Lambda)\right\}\left[1 - \rho_{jj'}I\{S_{jj'} = 0\} + \rho_{jj'}\frac{\lambda}{2} \exp(-\lambda|S_{jj'}|)\right] I\{S \in S^+\}.
\]

Without loss of generality, suppose that \(j = p - 1\) and \(j' = p\). Let
\[
S = \begin{bmatrix}
S_{-(jj')} & S_{-(jj'),jj'} \\
S_{jj',-(jj')} & S_{jj',jj'}
\end{bmatrix},
\]
where
\[
S_{jj',jj'} = \begin{bmatrix}
S_{jj} & S_{jj'} \\
S_{jj'} & S_{jj'}
\end{bmatrix}.
\]

30
With the property of matrices, we have

$$\det(S) \propto \det(S_{jj',jj'} - B),$$

where

$$B = S_{jj',-(jj')}S_{-(jj'),jj'}^{-1},$$

$$S^{-1} = \begin{bmatrix} S_{-(jj')}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} S_{-(jj')}^{-1}S_{-(jj'),jj'}^{-1} \\ 0 \\ 0 \end{bmatrix} (S_{jj'} - B)^{-1} \begin{bmatrix} S_{jj',-(jj')}^{-1}S_{-(jj')}^{-1} & -I_2 \\ S_{jj',-(jj')}^{-1}S_{-(jj')}^{-1} & -I_2 \end{bmatrix},$$

$$\text{tr}(S^{-1} \Lambda) = \text{tr} \left( \begin{bmatrix} S_{jj'}^{-1} & 0 \\ 0 & 0 \end{bmatrix} S \right) + \text{tr} \left( (S_{jj'} - B)^{-1} D \right),$$

where

$$D = \begin{bmatrix} S_{jj',-(jj')}^{-1}S_{-(jj')}^{-1} & -I_2 \end{bmatrix} \Lambda \begin{bmatrix} S_{-(jj')}^{-1}S_{-(jj'),jj'}^{-1} \\ -I_2 \end{bmatrix}.$$

Hence, we have

$$p(S_{jj'}|\cdot) \propto \det(S_{jj',jj'} - B)^{-n/2} \exp \left\{ -\frac{1}{2} \text{tr} ((S_{jj'} - B)^{-1} D) \right\} \times p(S_{ij}|\lambda, \rho_{jj'} \Lambda \{S \in S^+\},$$

$$\times \left\{ 1 - \frac{(S_{jj'} - B_{12})^2}{(S_{jj} - B_{11})(S_{jj',jj'} - B_{22})} \right\}^{-\frac{n}{2}} \cdot \exp \left\{ -\frac{1}{2} \frac{(S_{jj'} - B_{22})D_{11} + (S_{jj} - B_{11})D_{22} - 2(S_{jj'} - B_{12})D_{12}}{(S_{jj} - B_{11})(S_{jj',jj'} - B_{22}) - (S_{jj'} - B_{12})^2} \right\} \cdot \left\{ 1 - \rho_{jj'} \Lambda \{S_{jj'} = 0\} + \rho_{jj'} \frac{\lambda}{2} \exp(-\lambda |S_{jj'}|) \right\} \cdot \Lambda \{ (S_{jj'} - B_{12})^2 < (S_{jj} - B_{11})(S_{jj',jj'} - B_{22}) \}.$$

**REFERENCES**

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

Armstrong, H., Carter, C., Wong, K., and Kohn, R. (2009). Bayesian covariance matrix estimation using a mixture of decomposable graphical models. *Statistical Computation, 19 (3)*, 303-316.

Atay-Kayis, A. and Massam, H. (2005). The marginal likelihood for decomposable and non-decomposable graphical Gaussian models. *Biometrika, 92*, 317-335.

Baladandayuthapani, V., Ji, Y., Talluri, R., Neito-Barajas, L. E., and Morris J. (2010). Bayesian random segmentation models to identify shared copy number aberrations for array CGH data. *Journal of the American Statistical Association, 105*, 390-400.
Bickel, O. and Levina (2008). Covariance regularization by thresholding. *The Annals of Statistics, 36*, 2577-2604.

Bickel, O. and Levina (2008A). Regularized estimation of large covariance matrices. *The Annals of Statistics, 36*, 199-227.

Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika, 98*, 807-820.

Bonato, V., Baladandayuthapani, V., Broom, B. M., Sulman, E. P., Aldape, K. D., and Do, K. (2010). Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics, 27*, 359-367.

Brooks, S., Giudici, P., and Roberts, G. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65*, 39.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association, 106*, 672-684.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J., Wang, Q., and West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association, 103*, 1438-1456.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association, 90*, 1313-1321.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association, 96*, 270-281.

Chiu, T. Y. M., Leonard, T., and Tsui, K. W. (1996). The matrix-logarithmic covariance model. *Journal of the American Statistical Association, 91*, 198-210.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics, 21 (3)*, 1272-1317.

Deng X., and Tsui, K. W. (2010). Penalized covariance matrix estimation using a matrix-logarithm transformation. *Journal of Computational and Graphical Statistics, accepted.*

Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics, 147*, 186-197.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics, 9 (3)*, 432-441.

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

Giudici P. (2001). Bayesian inference for graphical factor analysis models. *Psychometrika, 66 (4)*, 577-592.

Giudici, P. and Green, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika, 86*, 785-801.

Grzybyk, M., Wild, P. and Chouanière, D. (2004). On identification of multi-factor models with correlated residuals. *Biometrika, 91 (1)*, 141-151.
Hoff, P. (2007). Model averaging and dimension selection for the singular value decomposition. *Journal of the American Statistical Association, 102*, 674-685.

Hoff, P. and Niu, X. (2012). A covariance regression model. *Statistica Sinica, 22*, 729-753.

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika, 93*, 85-98.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of American Statistics Association, 104*, 682-693.

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press, Oxford.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis, 88 (2)*, 365-411.

Leonard, T. and Hsu, J. S. J. (1992). Bayesian inference for a covariance matrix. *The Annals of Statistics, 20*, 1669-1696.

Levina, E., Rothman, A. J. and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics, 2 (1)*, 245-263.

Liu, R. Z., Graham, K., Glubrecht, D. D., Germain, D. R. et al. (2011). Association of FABP5 expression with poor survival in triple-negative breast cancer: implication for retinoic acid therapy. *The American Journal of Pathology, 178 (3)*, 997-1008.

Lopes, H. F. and Carvalho, C. M. (2007). Factor stochastic volatility with time varying loadings and Markov switching regimes. *Journal of Statistical Planning and Inference, 137*, 3082-3091.

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

Luo, X. (2011). High dimensional low rank and sparse covariance matrix estimation via convex minimization. [arXiv:1111.1133](http://arxiv.org/abs/1111.1133).

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika, 86*, 677-690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika, 87*, 425-435.

Roverato, A. (2002). Hyper-inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics, 29*, 391-411.

Rowe, D. B. (2003). *Multivariate Bayesian Statistics*. Chapman & Hall, London.

Stranger, B. E., Forrest, M. S., Clark, A. G., Minichiello, M. J., Deutsch, S., et al. (2005). Genome-wide associations of gene expression variation in humans. *PLoS Genetics, 1 (6)*, e78.

Stanghellini, E. (1997). Identification of a single-factor model using graphical Gaussian rules. *Biometrika, 84 (1)*, 241-244.

Talluri, R., Baladandayuthapani, V. and Mallick, B. K. (2009). Bayesian sparse graphical models and their mixtures using lasso selection priors. *Journal of American Statistical Association, Submitted*. Department of Statistics, Texas A&M University, College Station, Texas.
Wang, H. (2012). The Bayesian Graphical Lasso and Efficient Posterior Computation. Working. Department of Statistics, University of South Carolina, Columbia, South Carolina.

Wong, F., Carter, C., and Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika*, 90, 809-830.

Wu, W.B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90, 831-844.

Wu, W.B. and Pourahmadi, M. (2010). Banding sample auto covariance matrices of stationary processes. *Statistica Sinica*, 19, 1755.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94 (1), 19-35.
Table 1: Simulation results for Bayesian low rank and sparse matrix decomposition for model 1, 2 and 3. The mean $L_1$ and Frobenius norms of the difference matrices between the estimated and true covariance matrices over 20 replications are presented in the table with the standard deviations in parentheses. See Section 3 for details about the models.

| q  | model 1 |           |           |           |           |           |
|----|---------|-----------|-----------|-----------|-----------|-----------|
|    |         | $L_1$ norm | Frobenius | $L_1$ norm | Frobenius | $L_1$ norm | Frobenius |
| 50 | Bayesian | 11.90 (1.82) | 8.46 (0.82) | 12.89 (1.63) | 11.62 (1.46) | 13.40 (2.02) | 9.78 (0.70) |
|    | LOREC   | 13.64 (1.76) | 9.15 (0.68) | 11.84 (1.77) | 9.12 (0.98) | 13.55 (1.73) | 9.75 (0.70) |
|    | Sample  | 15.18 (2.12) | 11.63 (0.78) | 13.98 (2.95) | 11.06 (1.27) | 13.69 (1.99) | 10.75 (0.86) |
| 100| Bayesian | 15.28 (2.45) | 10.00 (0.78) | 25.48 (2.74) | 16.13 (1.17) | 20.89 (3.81) | 15.93 (1.00) |
|    | LOREC   | 16.54 (2.19) | 10.37 (0.96) | 23.67 (2.98) | 14.18 (1.68) | 21.97 (3.18) | 15.40 (0.96) |
|    | Sample  | 20.74 (1.90) | 17.36 (0.52) | 26.18 (4.64) | 20.08 (1.80) | 25.39 (5.00) | 19.51 (1.43) |
| 200| Bayesian | 29.90 (3.98) | 18.16 (1.15) | 49.83 (6.28) | 32.71 (9.94) | 35.58 (4.10) | 25.23 (1.21) |
|    | LOREC   | 32.79 (4.06) | 19.31 (1.77) | 48.96 (7.51) | 28.02 (6.09) | 39.66 (3.37) | 23.68 (1.04) |
|    | Sample  | 42.58 (2.56) | 35.42 (0.97) | 54.49 (4.96) | 37.60 (1.66) | 45.16 (5.25) | 35.76 (1.75) |
Table 2: Simulation results for Bayesian low rank and sparse matrix decomposition for model 1, 2 and 3. The top panel shows the mean successful rates of rank recovery and the means of the estimated ranks over 20 replications with the standard deviations in parentheses. The bottom panel shows the mean false positives and false negatives in the support recovery of the sparse component $S$. FP: number of false positive discoveries; FN: number of false negative discoveries. See Section 3 for details about the models.

| q   | Bayesian | mean(se) | LOREC | mean(se) | Bayesian | mean(se) | LOREC | mean(se) | Bayesian | mean(se) | LOREC | mean(se) |
|-----|----------|----------|--------|----------|----------|----------|--------|----------|----------|----------|--------|----------|
| 50  | 95       | 2.95 (0.22) | 100    | 3.00 (0.00) | 90       | 0.90 (0.31) | 40     | 2.05 (1.00) | 80       | 2.80 (0.41) | 40     | 2.35 (0.93) |
| 100 | 90       | 2.90 (0.31) | 80     | 1.20 (0.41) | 95       | 2.95 (0.22) | 100    | 1.00 (0.00) | 85       | 2.85 (0.37) | 50     | 2.35 (0.93) |
| 200 | 100      | 3.00 (0.00) | 90     | 0.92 (0.28) | 90       | 2.90 (0.31) | 80     | 2.70 (0.66) | 80       | 2.90 (0.31) | 50     | 2.35 (0.93) |

| q   | Bayesian | FN (FP) | LOREC | FN (FP) | Bayesian | FN (FP) | LOREC | FN (FP) |
|-----|----------|---------|--------|---------|----------|---------|--------|---------|
| 50  | 0 (0)    | 7.75 (4.06) | 0 (0)  | 6.00 (15.75) | 0 (0)    | 40.95 (11.38) | 0 (0)  | 188.25 (50.49) | 0 (0)    | 13.45 (7.96) | 0 (0)  | 518.25 (93.35) |
| 100 | 0 (0)    | 13.70 (6.33) | 0 (0)  | 15.90 (4.28) | 0 (0)    | 68.90 (15.48) | 0 (0)  | 508.25 (97.80) | 0 (0)    | 33.40 (6.06) | 0 (0)  | 1011.60 (304.40) |
| 200 | 0 (0)    | 2.00 (1.73) | 0 (0)  | 0.80 (1.47) | 0 (0)    | 34.60 (41.13) | 0 (0)  | 1175.60 (265.00) | 0 (0)    | 6.4 (4.03) | 0 (0)  | 1848.30 (495.00) |
Table 3: Simulation results for Bayesian graphical factor analysis models for model 4, 5 and 6. The top panel shows the mean successful rates of number of factor recovery and the means of the estimated numbers of latent factors over 20 replications with the standard deviations in parentheses. The bottom panel shows the mean false positives and false negatives in the edge selection for the graphical models of the residuals. FP: number of false positive edges; FN: number of false negative edges. See Section 5.5 for details about the models.

| Recovery of Number of Factors | model 4                      | model 5                      | model 6                      |
|-------------------------------|------------------------------|------------------------------|------------------------------|
| n                             | % (1 factor) | mean (se) | % (2 factor) | mean (se) | % (1 factor) | mean (se) |
|-------------------------------|--------------|-----------|--------------|-----------|--------------|-----------|
| 100                           | 95           | 1.05 (0.22) | 70           | 2.40 (0.68) | 75           | 1.25 (0.44) |
| 300                           | 100          | 1.00 (0.00) | 95           | 1.95 (0.22) | 100          | 1.00 (0.00) |

| Graph of Residuals            | model 4                      | model 5                      | model 6                      |
|-------------------------------|------------------------------|------------------------------|------------------------------|
| n                             | FN (se) | FP (se) | FN (se) | FP (se) | FN (se) | FP (se) |
|-------------------------------|--------|--------|--------|--------|--------|--------|
| 100                           | 0.05 (0.22) | 7.15 (0.49) | 17.35 (3.69) | 6.00 (1.68) | 0.45 (1.05) | 15.80 (2.75) |
| 300                           | 0.00 (0.00) | 6.90 (0.31) | 3.30 (1.59) | 9.00 (1.86) | 0.00 (0.00) | 17.25 (1.80) |
Figure 1: Heatmaps of the absolute value of covariance estimates by (a) sample covariance (b) Bayesian decomposition method (c) LOREC estimator.
Figure 2: Scatter plots of the single factor loadings identified by LOREC versus the two factors loadings identified by the Bayesian decomposition model. The correlation between the two factor loadings vectors on the left is 0.98, and the correlation between the two factors on the right in 0.12.
Figure 3: Matrix plot indicating the sparse support of the residual covariance component by (a) Bayesian decomposition method (b) LOREC estimator.
Figure 4: Graphical structure of the residuals in model 6 in the simulations in Section 5.5
Figure 5: Top panel: The adjacency matrix of the genes involved in ER pathway depicting the graphical model of the residuals inferred by (a) the HIW based graphical factor model (b) the lasso based graphical factor model. Bottom panel: The adjacency matrix depicting the edges that are selected for over half of the simulation runs by (c) the HIW based graphical factor model and (d) the lasso based graphical factor model, respectively. The simulated datasets were generated from a graphical factor model where the parameters were the posterior mean estimates of the real data analysis by the lasso based graphical factor model.