Acta Mathematicae Applicatae Sinica, English Series
Vol. 33, No. 2 (2017) 485–496
DOI: 10.1007/s10255-017-0676-z
http://www.ApplMath.com.cn & www.SpringerLink.com

Acta Mathematicae Applicatae Sinica, English Series

Bayesian Lasso with Neighborhood Regression Method for Gaussian Graphical Model

Fan-qun LI[1,2,†], **Xin-sheng ZHANG**[1]

[1]Department of Statistics, School of Management, Fudan University, Shanghai 200433, China
[2]Institute of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu 233000, China ([†]E-mail: 11110690005@fudan.edu.cn)

Abstract In this paper, we consider the problem of estimating a high dimensional precision matrix of Gaussian graphical model. Taking advantage of the connection between multivariate linear regression and entries of the precision matrix, we propose Bayesian Lasso together with neighborhood regression estimate for Gaussian graphical model. This method can obtain parameter estimation and model selection simultaneously. Moreover, the proposed method can provide symmetric confidence intervals of all entries of the precision matrix.

Keywords gaussian graphical model; regression; precision matrix; Bayesian Lasso; Frobenius loss
2000 MR Subject Classification 62F15; 62H12

1 Introduction

Gaussian graphical models can provide effective tools for discovering conditional independence relationships among variables (Lauritzen[12]; Whittaker[21]). Consider the p-dimensional multivariate Gaussian distribution random variable $X = (X_1, \cdots, X_p)^T \sim N(\mu, \Sigma)$. Assume that matrix Σ is non-singular, the conditional independence structure of the distribution can be represented by a graphical model $\mathfrak{G} = (V, E)$, where $V = \{1, \cdots, p\}$ is the set of nodes and E the set of edges in $(V \times V)$. According to the connection of precision matrix $\Omega = \Sigma^{-1} := (\omega_{ij})_{p\times p}$ and partial correlation coefficient matrix, an edge (i, j) is not contained in the edge set E if and only if X_i is conditionally independent of X_j given the remaining variables. Whether edge (i, j) is not contained in the set E can be represented by whether the corresponding entry of the precision matrix Ω is zero.

The Lasso of Tibshirani[18] is usually used to estimate the regression coefficient $\beta = (\beta_1, \cdots, \beta_p)^T$ in the linear model

$$Y = \mathbf{X}\beta + \varepsilon, \tag{1.1}$$

where Y is the $n \times 1$ vector of response, $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_p)$ is the $n \times p$ matrix of predictors. The Lasso solution $\widehat{\beta}$ minimizes the residual sum of squares with restraining the ℓ_1-norm of the coefficient vector β,

$$\widehat{\beta}_L = \arg\min_{\beta} ||Y - \mathbf{X}\beta||_2^2 + \lambda||\beta||_1, \qquad \lambda \geq 0,$$

where $||\cdot||_2$ denotes the vector ℓ_2-norm, and $||\cdot||_1$ denotes the vector ℓ_1-norm.

Recently, the graphical Lasso is not only used for graphical model selection (Meinshausen and Buhlmann[13]), but also widely used for simultaneous graphical model determination and parameter estimation (Yuan and Lin[24]; Banerjee et al.[4]; Fridman et al.[9]). In Gaussian

Manuscript received September 24, 2013. Revised June 1, 2015.
Supported by the National Natural Science Foundation of China (No. 11571080).

random field, the graphical Lasso is a form to maximize the penalized log-likelihood

$$\log\det(\Omega) - \mathrm{tr}(n^{-1}S\Omega) - \lambda||\Omega||_1, \qquad \text{for } \lambda \geq 0,$$

where $\frac{1}{n}S$ is the empirical covariance matrix, $||\Omega||_1$ denotes the sum of the absolute values of the elements of the positive definite matrix Ω, λ is tuning parameter. Friedman et al.[9] developed a simple algorithm for the Lasso using a coordinate descent procedure. Particularly Meinshausen and Buhlmann[13] proposed neighborhood selection (NS) method to estimate the conditional independence restrictions separately for each node in the graph, NS can equivalently construct graphs for Gaussian models. However, the shortage of the method is that NS can not estimate the precision matrix. Yuan[25] replaced the Lasso selection by a Dantzig type modification, where first the ratios between the off-diagonal elements ω_{ij} and the corresponding diagonal element ω_{ii} were estimated, and then the diagonal element ω_{ii} were obtained given the estimated ratios ω_{ij}/ω_{ii}. At last the off-diagonal elements ω_{ij} were obtained since the ratios ω_{ij}/ω_{ii} and ω_{ii} were estimated. However, the estimations were obtained in two-step procedure. Motivated by these view points, we propose a Bayesian method for graphical model which aims to estimate precision matrix and construct the graph simultaneously.

The Lasso has a Bayesian interpretation. Tibshirani[18] suggested that Lasso estimates can be interpreted as posterior mode estimate when the regression parameters have independent and identical Laplace priors. Motivated by this connection, Figueiredo[8], Bae and Mallick[3], and Yuan and Lin[24] proposed Laplace-like priors in Bayesian Lasso method for linear regression. Park and Casella[14] proposed explicit treatment of Bayesian Lasso regression, and provided a full Bayesian analysis using a conditional Laplace prior specification of the form

$$\pi(\beta|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\{-\lambda|\beta_j|/\sqrt{\sigma^2}\}, \tag{1.2}$$

and used the noninformative scale-invariant marginal prior $\pi(\sigma^2) = 1/\sigma^2$ on σ^2. They extended the Bayesian Lasso regression model to hierarchical model with prior $\beta|\sigma^2, \tau_1^2, \cdots, \tau_p^2 \sim \mathcal{N}_p(0_p, D_\tau)$, where D_τ=diag $(\tau_1^2, \cdots, \tau_p^2)$, and obtained point estimates of the regression coefficients by placing prior distributions on the hyper-parameters $\tau_1^2, \cdots, \tau_p^2$ and σ^2. Hans[6] introduced new aspects of broad Bayesian treatment of Lasso regression, the author provided a direct characterization of the coefficients posterior distribution and new Gibbs sampler for Bayesian Lasso regression.

Bayesian approaches to statistical inference often apply the conjugate prior distribution for parameters. Jones et al.[10], Scott and Carvalho[17], Wang and West[19] relied on the conjugate hyper-inverse Wishart (HIW) prior for covariance matrix Σ. Equivalently, Roverato[15], Atay-Kayis and Massam [2] applied the G-Wishart prior for the precision matrix. Wang and Carvalho[23] pointed out that sampling from (HIW) for non-decomposable graphs is not trivial and depends on computationally extensive Monte Carlo methods. Wong et al.[20] placed point mass priors at elements of the partial correlation matrix and non-informative priors for the non-zero elements.

The graphical Lasso also has a Bayesian interpretation. With the following prior $P(\Omega|\lambda) \propto \prod_{i<j} \exp\{-\lambda|\omega_{ij}\}\prod_{=1}^{p} \exp\big\{-\frac{\lambda}{2}|\omega_{ii}\big\}\, I(\Omega \in M^+)$, where M^+ denotes the set of all definite positive symmetric matrices, for any fixed values $\lambda \geq 0$, the posterior mode of Ω is the graphical Lasso estimate with tuning parameter $\rho = \lambda/n$. Wang[22] presented a Bayesian graphical Lasso procedure for simultaneous parameter estimation and structural learning. The author proposed hierarchical models and priors, under which the posterior distribution of the elements $\omega_{ij}, (i \leq j)$ of precision matrix Ω are introduced. Full conditional distributions of ω_{ij} can be approximated by a Gaussian density respectively.

In this paper, we view a Gaussian graphical model as a linear regression model, and propose Bayesian Lasso method for neighborhood selection. At the same time, we use a Bayesian method to estimate the variances of the errors, so we can estimate parameters of precision matrix and construct the graph simultaneously.

The rest of the paper is as follows. In Section 2, we propose Bayesian Lasso together with neighborhood regression estimate denoted by BLNRE for Gaussian graphical model, and introduce how to choose the penalty parameter λ. We also provide the method to shrink the estimation of the entries of the precision matrix for model selection. In Section 3, We present simulation results and comparison with the results that come from neighborhood selection (NS) (Meinshausen and Buhlmann[13]) and from Graphical Lasso (GLasso) (Fridman et al.,[9]). At last, the flow cytometry data set from Sachs et al.[16] is analyzed in Section 4.

2 Methodology

2.1 Regression and Precision Matrix of Gaussian Graphical Model

Let $X = (X_1, \cdots, X_p)^T$ be a p-dimensional random variable following multivariate normal distribution with mean μ and covariance matrix Σ. We shall denote $\Sigma_{\backslash i,\backslash j}$ the submatrix of Σ with its i-th row and its j-th column removed. $\Sigma_{i,\backslash j}$ or $\Sigma_{\backslash i,j}$ denotes the i-th row of Σ with the j-th column removed or the j-th column of Σ with the i-th row removed. It is well known that if we partition X, μ, Σ as follows (Anderson[1]):

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

then the conditional distribution of $X^{(1)}$ given $X^{(2)} = x^{(2)}$ is

$$X^{(1)}|x^{(2)} \sim N(\mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)} - \mu^{(2)}), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

Specially, let $X^{(1)} = X_i$, $X^{(2)} = X_{\backslash i}$, then the conditional distribution of X_i given the remaining variables $X_{\backslash i}$ follows from normal distribution,

$$X_i|X_{\backslash i} = x_{\backslash i} \sim N(\mu_i + \Sigma_{i,\backslash i}\Sigma_{\backslash i,\backslash i}^{-1}(x_{\backslash i} - \mu_{\backslash i}), \sigma_{ii} - \Sigma_{i,\backslash i}\Sigma_{\backslash i,\backslash i}^{-1}\Sigma_{\backslash i,i}).$$

Equivalently, we have the following regression equation

$$X_i = \alpha_i + X_{\backslash i}^T\beta_{(i)} + \varepsilon_i, \tag{2.1}$$

where $\alpha_i = \mu_i - \Sigma_{i,\backslash i}\Sigma_{\backslash i,\backslash i}^{-1}\mu_{\backslash i}$, $\beta_{(i)} = \Sigma_{\backslash i,\backslash i}^{-1}\Sigma_{\backslash i,i}$ is a $p-1$ dimensional vector and $\varepsilon_i \sim N(0, \sigma_{ii} - \Sigma_{i,\backslash i}\Sigma_{\backslash i,\backslash i}^{-1}\Sigma_{\backslash i,i})$ is independent of $X_{\backslash i}$, μ_i is the i-th element of the mean μ, σ_{ii} is the (i,i)-th element of the covariance matrix Σ.

Because $\Omega = \Sigma^{-1}$, then the i-th column of Ω can be written as

$$\omega_{ii} = (\sigma_{ii} - \Sigma_{i,\backslash i}\Sigma_{\backslash i,\backslash i}^{-1}\Sigma_{\backslash i,i})^{-1},$$
$$\Omega_{\backslash i,i} = -(\sigma_{ii} - \Sigma_{i,\backslash i}\Sigma_{\backslash i,\backslash i}^{-1}\Sigma_{\backslash i,i})^{-1}\Sigma_{\backslash i,\backslash i}^{-1}\Sigma_{\backslash i,i},$$

combining with (2.1), we have

$$\omega_{ii} = (Var(\varepsilon_i))^{-1}, \tag{2.2}$$
$$\Omega_{\backslash i,i} = -(Var(\varepsilon_i))^{-1}\beta_{(i)}. \tag{2.3}$$

Therefore, an estimate of Ω can be potentially obtained by regressing X_i over $X_{\backslash i}$. Furthermore, (2.3) can be written as $\beta_{(i)} = -\Omega_{\backslash i,i}/\omega_{ii}$. This denotes that a zero entry on the i-th column of

the precision matrix Ω implies a zero entry in the corresponding regression coefficient $\beta_{(i)}$ and vice versa.

2.2 Neighborhood Selection and Parameters Estimation with Bayesian Lasso

Meinshausen and Buhlmann[13] proposed Lasso estimate for $\beta_{(i)}$ in (2.1) which is given by

$$\widehat{\beta}_{(i),\lambda} = \arg\min_{\beta} \left(||\mathbf{X}_i - \mathbf{X}_{\backslash i}\beta||_2^2 + \lambda||\beta||_1 \right), \tag{2.4}$$

where $\mathbf{X}$ is $n \times p$ matrix containing n independent observations of X, $\mathbf{X}_i$ is the i-th column of $\mathbf{X}$, $\mathbf{X}_{\backslash i}$ is the sub-matrix of $\mathbf{X}$ with i-th column removed, β is a $(p-1)$-dimensional vector.

Using the Bayesian interpretation of the Lasso, $\widehat{\beta}_{(i),\lambda}$ in (2.4) can be viewed as the mode of the posterior distribution of β, $\widehat{\beta}^B_{(i)} = \arg\max_\beta p(\beta|\mathbf{X}_{\backslash i}, \sigma^2, \tau)$, when independent, double-exponential prior distributions are placed on the $(p-1)$-dimensional regression coefficients $\beta_{(i)}$.

In this paper, we suggest the following hierarchical representation of the full model similar as Park and Casella[14] (we omit the sub-label i of $\tau_{1,i}, \cdots, \tau_{p-1,i}$ for convenience):

$$\begin{aligned}
&X_i|X_{\backslash i}, \beta_{(i)}, \sigma_i^2 \sim \mathcal{N}(X_{\backslash i}^T\beta_{(i)}, \sigma_i^2),\\
&\beta_{(i)}|\sigma_i^2, \tau_1^2, \cdots, \tau_{p-1}^2 \sim \mathcal{N}(0_{p-1}, \sigma_i^2 D_\tau),\\
&D_\tau = \mathrm{diag}\,(\tau_1^2, \cdots, \tau_{p-1}^2),\\
&\sigma_i^2, \tau_1^2, \cdots, \tau_{p-1}^2|\lambda^2 \sim \pi(\sigma_i^2)d\sigma_i^2 \prod_{j=1}^{p-1} \frac{\lambda^2}{2} \exp(-\lambda^2\tau_j^2/2)d\tau_j^2,
\end{aligned}$$

where $\sigma_i^2, \tau_1^2, \cdots, \tau_{p-1}^2 > 0$.

Then the full conditional distributions of $\beta_{(i)}$, σ_i^2, $\tau_1^2, \cdots, \tau_{p-1}^2$ are as follows:

$$\begin{aligned}
&\beta_{(i)}|\mathbf{X}, \sigma_i^2, \tau_1^2, \cdots, \tau_{p-1}^2 \sim \mathcal{N}_{p-1}(A^{-1}\mathbf{X}_{\backslash i}\mathbf{X}_i, \sigma_i^2 A^{-1}),\\
&\sigma_i^2|\mathbf{X}, \beta_{(i)}, \tau_1^2, \cdots, \tau_{p-1}^2 \sim \mathrm{IG}(\frac{\mathrm{n+p-1}}{2}, (\mathbf{X}_\mathrm{i} - \mathbf{X}_{\backslash \mathrm{i}}\beta_{(\mathrm{i})})^\mathrm{T}(\mathbf{X}_\mathrm{i} - \mathbf{X}_{\backslash \mathrm{i}}\beta_{(\mathrm{i})}))/2 + \beta_\mathrm{i}^\mathrm{T}\mathrm{D}_\tau^{-1}\beta_{(\mathrm{i})}/2),\\
&\tau_j^{-2}|\mathbf{X}, \beta_{(i)}, \sigma_i^2 \sim \mathrm{IN}(\mu', \lambda'), \mathrm{j} = 1, \cdots, \mathrm{p}-1,
\end{aligned}$$

where $\mathrm{IG}(a, b)$ denotes inverse-gamma with shape parameter a and scale parameter b, and $\mathrm{IN}(\mu', \lambda')$ denotes inverse-Gaussian distribution with parameters μ', λ', $A = \mathbf{X}^T\mathbf{X} + D_\tau^{-1}$, and here $\mu' = \sqrt{\lambda^2\sigma^2/\beta_j^2}$, $\lambda' = \lambda^2$. These full conditional distribution form the basis for an efficient Gibbs sampler. For $i = 1, \cdots, p$, we obtain estimation of $\widehat{\beta}_{(i)}$ and $\widehat{\sigma}_i^2$ by Gibbs sampling. Combining (2.2) and (2.3), we have $\widehat{\omega}_{ii} = \widehat{\sigma}_i^{-2}$, $\widehat{\Omega}_{\backslash i,i} = -(\widehat{\sigma}_i^2)^{-1}\widehat{\beta}_{(i)}$. Then we obtain the estimation of Ω denoted by $\widehat{\Omega} = (\widehat{\omega}_{ij})$.

2.3 The Choice of the Penalty Parameter λ

The graphical Lasso requires the selection of the penalty parameter λ. Typically one can estimate this parameter by cross-validation (Fridman et al.[9]), or by BIC criterion (Yan[25]). For the Bayesian framework, each iteration of the algorithm involves running the Gibbs sampler using a λ. The full Bayesian method for choosing λ is applied in the context of Bayesian Lasso regression models (Park and Casella[14]; Kyung et al.[11]). Alternatively, one can use the empirical Bayesian estimate to provide a point estimate of λ (Park and Casella[14]). In

this paper we consider the conjugate gamma prior on λ^2 according to the prior of λ^2 given in Yuan[25]. $\lambda^2 \sim \mathrm{Ga}(\mathrm{r}, \mathrm{s})$, and the full conditional distribution for λ^2 is

$$\lambda^2|\tau_1^2, \cdots, \tau_{p-1}^2 \sim \mathrm{Ga}\Big(\mathrm{r} + \mathrm{p} - 1, \mathrm{s} + \sum_{\mathrm{j}=1}^{\mathrm{p}-1} \tau_{\mathrm{j}}^2/2\Big),$$

where r, s are the hyper-parameters. With the full conditional distribution, we can obtain λ via Gibbs sampler in each iteration.

2.4 Construction of Graphical Model and Symmetrization

Graphical model selection is an important problem for a sparse graphical model. The classical graphical Lasso procedure is able to produce possible $\beta_{(i)}^k = 0$ in the problem (2.1). However, we place continuous prior distribution on $\beta_{(i)}$, hence has zero posterior probability on the event $\beta_{(i)}^k = 0$. So we can not obtain exact zeros for true-zero entries of precision matrix Ω through Gibbs sampling. In order to get graphical model selection, we shall apply a threshold to obtain a sparse precision matrix. Carvalho et al.[5] proposed a method for classification under absolutely continuous priors. If we choose to study sparsity in the case where θ is a vector of normal means:$(Y|\theta) \sim \mathbf{N}(\theta, \sigma^2 I)$, then authors proposed discrete mixture models accounting for the presence of sparsity

$$\theta_i \sim (1-p)\delta_0 + pg(\theta_i)),$$

where p is including probability $(P_r(\theta_i \neq 0))$ and $g(\theta_i)$ is prior distribution for θ_i when $\theta_i \neq 0$. Under a discrete mixture model, the posterior mean of θ_i is

$$E(\theta_i|Y) = \omega_i \cdot E_g(\theta_i|Y, \theta_i \neq 0),$$

where w_i is posterior including probability (i.e. $P_r(\theta_i \neq 0|Y)$). Then a possible threshold is to call $\theta_i \neq 0$ if the prior distribution $g(\theta_i)$ yields $w_i > 0.5$, and to call $\theta = 0$ otherwise. In our situation (2.1), under discrete and continuous mixture prior, the Bayesian posterior mean estimator of $\beta_{(i)}^k$ is

$$\widehat{\beta}_{(i)}^k = \omega_k \cdot E_g(\beta_{(i)}^k|\mathbf{X}, \beta_{(i)}^k \neq 0), \qquad k = 1, 2, \cdots, i-1, i+1, \cdots, p,$$

where g is the continuous prior distribution for non-zero $\beta_{(i)}^k$. Consider the graphical Bayesian Lasso prior (1.2) which can also shrink $\beta_{(i)}^k$ towards to zero, its posterior mean estimator $\widetilde{\beta}_{(i)}^k$ of $\beta_{(i)}^k$ can be written as

$$\widetilde{\beta}_{(i)}^k = \widetilde{\omega}_k \cdot E_g(\beta_{(i)}^k|\mathbf{X}, \beta_{(i)}^k \neq 0), \qquad k = 1, 2, \cdots, i-1, i+1, \cdots, p,$$

where $\widetilde{w}_k$ is the amount of shrinkage applied by the Bayesian graphical Lasso prior on $E_g(\beta_{(i)}^k|\mathbf{X})$. Then similar as Carvalho et al.[5], we claim the event $\beta_{(i)}^k \neq 0$ if and only if $\widetilde{\omega}_k > 0.5$, and $\beta_{(i)}^k = 0$ otherwise.

As for the choice of prior distribution $g(\cdot)$ of $\beta_{(i)}$, we consider conjugate multivariate normal distribution, $\beta_{(i)} \sim \mathbf{N}(0, \tau^2 I)$, where τ is unknown. Then $E_g(\beta_{(i)}|\mathbf{X}) = H^{-1}K$, where $H = I/\tau^2 + \mathbf{X}_{\setminus i}^T \mathbf{X}_{\setminus i}/\sigma_i^2$, $K = \mathbf{X}_{\setminus i}^T \mathbf{X}_i$. In our simulation, we replace σ_i^2 by residual covariance $\widehat{\sigma}_i^2$ of regression model (2.1), and let $\tau = 1$. This just is the ridge regression coefficients of $\beta_{(i)}$ in (2.1).

After constructing the graphical model, we can obtain the sparse estimate of the precision matrix Ω based on the estimate $\widehat{\Omega}$ in Section 2.2, and the sparse estimation denoted by $\widehat{\Omega}_1$.

However, we do not impose the symmetry condition on Ω in Section 2.2 and as a result the estimate $\widehat{\Omega}_1$ is not symmetric. The symmetric estimate of Ω denoted by $\widetilde{\Omega}$ can be obtained by symmetrizing $\widehat{\Omega}_1 = (\widehat{\omega}^1_{ij})$ as follow,

$$\widetilde{\Omega} = (\widetilde{\omega}_{ij}), \widetilde{\omega}_{ij} = \widetilde{\omega}_{ji} = \widehat{\omega}^1_{ij} I\{|\widehat{\omega}^1_{ij}| \leq |\widehat{\omega}^1_{ji}|\} + \widehat{\omega}^1_{ji} I\{|\widehat{\omega}^1_{ij}| > |\widehat{\omega}^1_{ji}|\}. \tag{2.5}$$

In the other words, we take the smaller magnitude of $\widehat{\omega}_{ij}$ and $\widehat{\omega}_{ji}$ for $\widetilde{\omega}_{ij}$ and $\widetilde{\omega}_{ji}$. In our simulation, we also consider the performance of the initial Bayesian Lasso estimator obtained by symmetrized $\widehat{\Omega}$.

3 Simulation

Example 1. The simulation experiments are concerned with performances of Bayesian Lasso together with neighborhood regression estimate (BLNRE), Neighborhood selection (NS) and Graphical Lasso (GLasso) in structure learning and in terms of estimation of precision matrix. We consider three different models in this example.

• Model 1. An AR(2) model with $\omega_{ii} = 1,\ \omega_{i-1,i} = \omega_{i,i-1} = 0.5$ and $\omega_{i-2,i} = \omega_{i,i-2} = 0.25$ and $\omega_{ij} = 0$ otherwise.

• Model 2. A star model with every node connected to the first node, with $\omega_{ii} = 1,\ \omega_{1i} = \omega_{i1} = 0.3$ and $\omega_{ij} = 0$ otherwise.

• Model 3. A circle model with $\omega_{1,1} = \omega_{10,10} = 1,\ \omega_{i,i} = 1.36$, for $i = 2, 3, \cdots, 9,\ \omega_{i,i-1} = \omega_{i-1,i} = -0.6$ and $\omega_{1,10} = \omega_{10,1} = 0.6$ and $\omega_{ij} = 0$ otherwise.

Fig.1 gives the image plots of the true graph of the three models. In the images plots, the lightest colour corresponds to elements of the matrix that have minimal absolute value and the darkest colour correspond to elements of matrix that have maximal absolute value. Shades of grey corresponded to interpolated values between the minimum and the maximum.

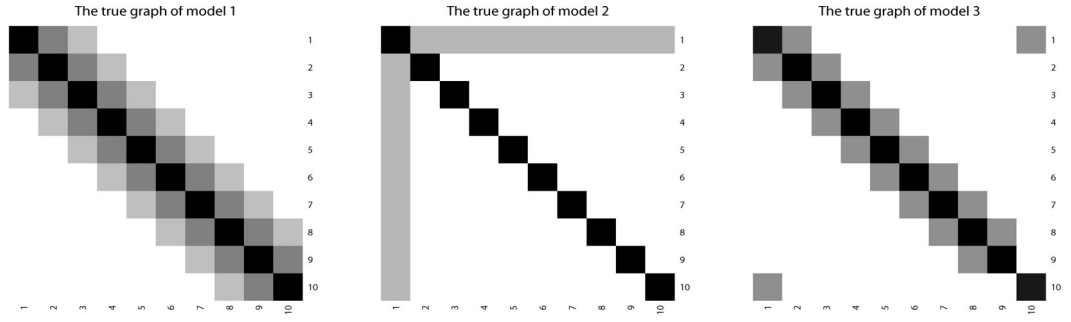

Fig. 1. Image Plots of the True Graphs of Model 1-3 Respectively

For each model, we generate 500 groups data with each group having $n = 100$ observations, and consider the case $p = 10$. In order to assess the effect of the two hyper-parameters r and s to the simulation, we let $s = 0.1, 0.3, \cdots, 3.9$ and $r = 1, 1.5, 2, 2.5$, the simulation results are reported in Fig.2. Fig.2 shows that among the range of r and s, their values don't affect the simulation results.

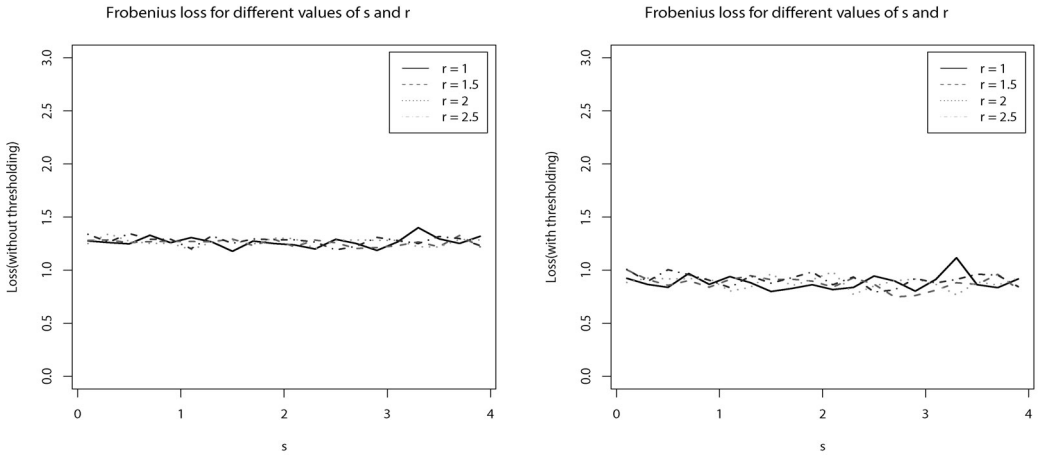

Fig. 2. Frobenius loss with different values of r and s

Then for each generated sample, we fit Bayesian regression Lasso with hyper-parameters $r = 2$, $s = 1.5$ for the prior distribution of λ^2. The Bayesian regression Lasso estimations are obtained via iterations of the Gibbs sampler. To assess the performance of the structure learning, we calculate the percentages of true non-zeros estimated as non-zeros (TP) and the percentages of true zeros estimated as zeros (TN) based on the estimated $\widehat{\omega}_{ij}$, and compare these results with the results obtained by NS (Meinshausen and Buhlmann[13]) and GLasso (Fridman et al.[9]). We use AIC criterion for GLasso to select the tuning parameter λ . The mean and the standard error of TN and TP are reported in Table 1. In this paper, the result of NS is obtained as follows:

Step 1. Using the lasso estimate of $\beta_{(i)}$ which is given by (2.4), we obtain the neighborhood set $\widehat{ne}_i$ of the i-th node via lars algorithm (Efron[7]). In the procedure ,we use AIC criterion for the selection of tuning parameter λ.

Step 2. For $i = 1, 2, \cdots, p$, we obtain a initial graph of the model which is not symmetric. Using "*or*" rule (Meinshausen and Buhlmann[13]) i.e. letting there is not an edge between i and j if $i \notin \widehat{ne}_j$ or $j \notin \widehat{ne}_i$, we obtain a symmetric estimate of the graph denoted by NS.

In addition, our method also uses the "or" rule for symmetric model selection , because the symmetrization strategy (2.5) (in Section 2.4) implies the "or" rule.

As shown in Table 1, the TP is 1 for all models and for all methods, BLNRE method get large TN for all models compared to GLasso, but less than NS. This is tolerable, after all the goal of NS is to identify the underlying graphical model.

Table 1. Summary of Performance of Structure Learning Measured by TN and TP

Model		BLNRE	NS	GLasso
Model 1	TN	0.652(0.043)	0.803(0.041)	0.571(0.076)
	TP	1.000(0.000)	1.000(0.000)	1.000 (0.000)
Model 2	TN	0.774(0.040)	0.825(0.035)	0.734(0.058)
	TP	1.00(0.000)	1.00(0.000)	1.000(0.000)
Model 3	TN	0.703(0.053)	0.780(0.048)	0.682(0.071)
	TP	1.000(0.000)	1.000(0.00)	1.000(0.00)

In order to compare the performances of the estimates of these three models, we consider Frobenius loss: Loss=tr $((\widehat{\Omega} - \Omega)(\widehat{\Omega} - \Omega))$, and the mean and the standard error of the loss

are reported in Table 2. The goal of the neighborhood selection is to identify the underlying graphical model, so we use a simple two-step procedure to obtain the estimate of the precision matrix. Firstly, we obtain the graph NS via neighborhood selection. Secondly, we use the Iterative Proportional Scaling algorithm (IPS) (Lauritzen[12]) for the maximum likelihood estimate based on the selected model. The results denoted by NS+IPS are shown in Table 2. We present the results of BLNRE in two cases, the loss $tr((\widehat{\Omega}-\Omega)(\widehat{\Omega}-\Omega))$ for the without threshold estimate (BLNRE(initial)) and the loss $tr((\widetilde{\Omega}-\Omega)(\widetilde{\Omega}-\Omega))$ for the threshold estimate (BLNRE) in Table 2, respectively. Table 2 shows that without threshold, the loss of BLNRE is substantially large than NS+IPS, while the loss of BLNRE is less than NS+IPS with threshold. But the loss of the two cases of BLNRE is less than GLasso.

Table 2. Summary of Performance of Estimation Measured by the Frobenius Loss

Model	BLNRE(initial)	BLNRE	NS+IPS	GLasso
Model 1	1.243(0.354)	0.934(0.598)	1.209(0.636)	2.291(0.201)
Model 2	1.406(0.287)	0.851(0.603)	0.968(0.529)	1.533(0.176)
Model 3	2.179(0.524)	1.407(0.556)	1.675(0.808)	3.407(0.468)

Fig.3 gives the box plots of all entries of the precision matrix of Models 1–3. Fig.4 gives the 95% credible intervals of all entries of the precision matrix of Models 1–3. These figures are plotted based on the estimate $\widehat{\Omega}$ stated in Section 2.2. Figs. 3, 4 show that ω_{ij} and ω_{ji} have almost the same box plots and confidence intervals. Moreover, the confidence intervals of the true zero entries all contain zero point.

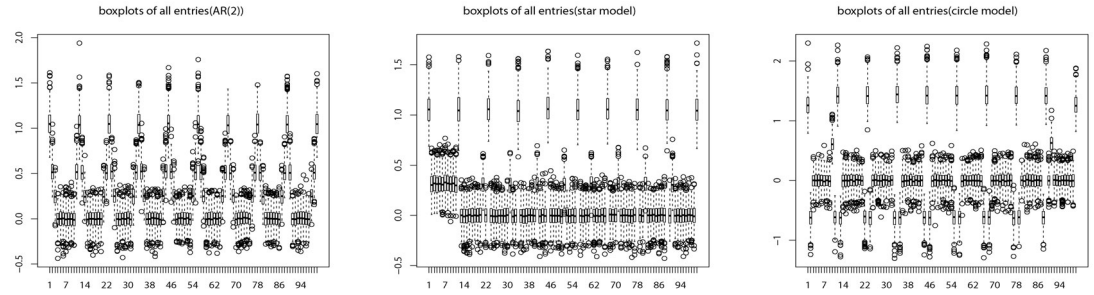

Fig. 3. The box plots of all entries of the estimated precision matrices of Models 1–3.

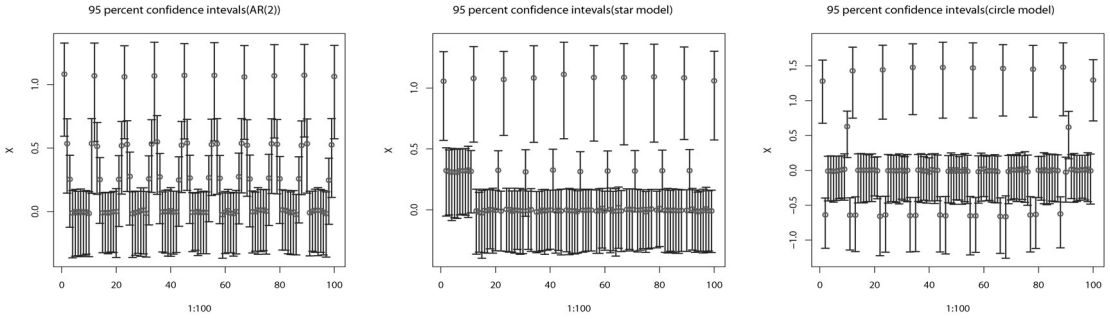

Fig. 4. The 95 percent confidence intervals of all entries of the estimated precision matrices of Models 1–3

Then let p=10,20,30, we plot the image plots of the precision matrices of Model 1 with fixed sample size (n=100). The image plots are presented in Fig. 5.

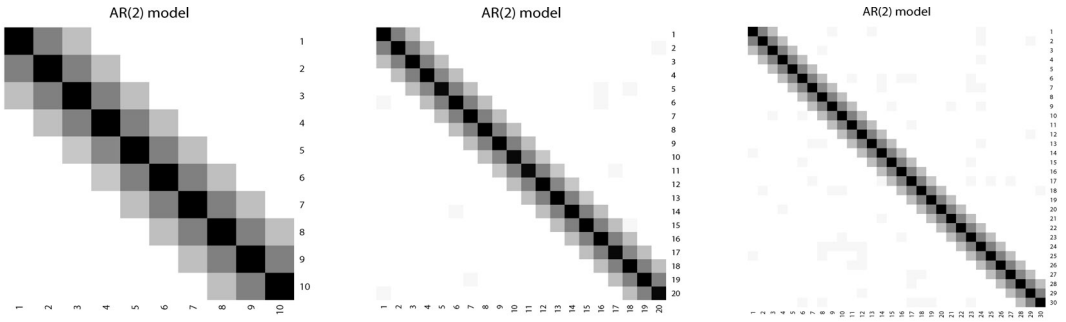

Fig. 5. Image plots of the estimated precision matrices of Model 1 for p=10, 20, 30 respectively.

Example 2. This simulation experiments are concerned with performances of BLNRE estimator for several different p. In this simulation we considered two model: Model 3 and the following model.

• Model 4. An AR(3) model with $\omega_{ij} = 0.5^{|i-j|},\ |i-j| \leq 3,\ i \neq j$ and $\omega_{ii} = 1$, and $\omega_{ij} = 0$ otherwise.

For each model, we let $p = 10, 20, \cdots, 50, 60$ and the sample size n satisfies $p/n \approx 0.1$. The BLNRE estimates are based on 200 repeating simulation. We calculated the normalized Frobenius loss $\frac{1}{p}tr((\widehat{\Omega} - \Omega)(\widehat{\Omega} - \Omega))$ and $\frac{1}{p}tr((\widetilde{\Omega} - \Omega)(\widetilde{\Omega} - \Omega))$ for every model, i.e., without threshold and with threshold, respectively. the results are reported in Fig.6. As shown in Fig.6, our method has stability when p and n increase simultaneously.

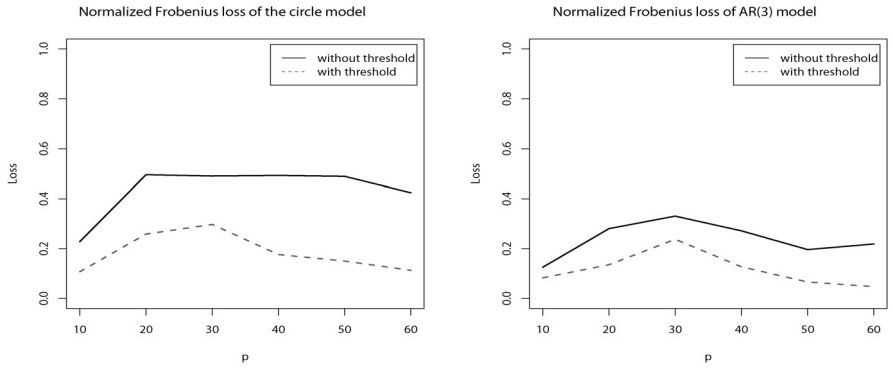

Fig 6. Normalized Frobenius loss of Models 3–4 for different p, the left one is for Model 3, and the right one is for Model 4.

4 Real Data Analysis

In this example, we analyze the flow cytometry dataset from Sachs et al.[16]. The data consists of flow cytometry measurements of $p = 11$ phosphorylated proteins and phospholipids, and

$n = 7466$ cells. The 11 proteins are Raf, Erk, P38, Jnk, AKT, Mek, PKA, PKC, Plc, PIP2, PIP3 respectively. Sachs et al.[16] constructed Bayesian network with the data, they fitted a directed acyclic graph (DAG) shown in Fig.7. Friedman et al.[9] proposed the graphical lasso method (GL) to analyze the data set, they fitted a set of undirected graphs for different values of the penalty parameter ρ. Their results show that given a range of penalty parameters, the graphical lasso has agreement with the DAG of about 50% for both edges and non-edges. Wang[22] applied the BGL to these data, and reported the 95% confidence intervals for the zero off-diagonal elements estimated at two different values of penalty parameter λ. We propose BLNRE method to the data set to the estimation of precision matrix and structure learning of the 11 proteins. We let the hyper-parameters r=1 and s=0.5 for the prior distribution of λ. The image of the estimated precision matrix is shown in Fig.8, and the corresponding constructed graph is shown in Fig.9. As shown in Fig.9, the graph has about 68% of the edges and 85% of the non-edges agreement with the undirected graph implied by the DAG.

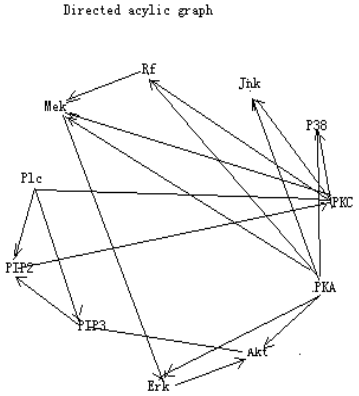

Fig. 7. Directed acylic graph from the flow cytometry data.

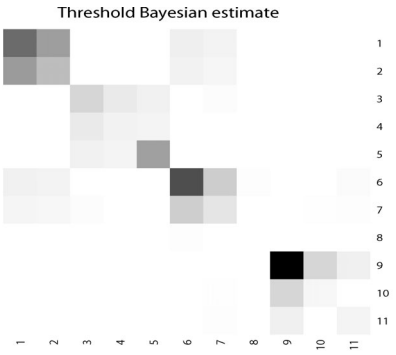

Fig. 8. Image plot of the estimated precision matrix.

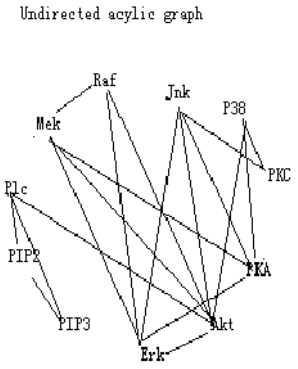

Fig. 9. Fitted undirected graph.

5 Discussion

High dimensional graphical model selection and estimation is becoming more and more common in various scientific and technological fields. In this paper, based on the connection of multivariate linear regression and entries of precision matrix, we propose Bayesian method to estimate the entries of precision matrix and recover the model structure simultaneously. Numerical studies show that BLNRE method has better performance for precision matrix estimation under Frobenius norm loss criterion. And we find that BLNRE estimation is stable for different p according to the normalized Frobenius norm loss.

Acknowledgements. We would like to thank all Associate Editors and gratefully acknowledge the helpful comments of the reviewers that substantially improved the paper.

References

[1] Anderson, T.W. An introduction to multivariate statistical analysis. Wiley-Interscience, London, 2003
[2] Atay-Kayis, A, Massam, H. The marginal likelihood for decomposable and non-decomposable graphical Gaussian models. *Biometrika*, 92: 317–335 (2005)
[3] Bae, K., Mallick, B.K. Gene selection using a two-Level hierarchical Bayesian model. *Bioinformatics*, 20: 3423–3430 (2004)
[4] Banerjee, O., El Ghaoui, L, d'Aspremont, A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9: 485–516 (2008)
[5] Carvalho, C.M., Polson, N.G., Scott, J.G. The horseshoe estimator for sparse signals. *Biometrika*, 97: 465–480 (2010)
[6] Hans, C. Bayesian lasso regression. *Biometrika*, 96: 835–845 (2009)
[7] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. Least angle regression. *Ann. Statist*, 32: 409–499 (2004)
[8] Figueiredo, M.A.T. Adaptive sparseness for supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 25: 1150–1159 (2003)
[9] Friedman, J., Hastie, T., Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9: 432–441 (2008)
[10] Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., West, M. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20: 388–400 (2005)
[11] Kyung, M., Je, G., Malay, G., George, C. Penalized Regression, Standard Errors, and Bayesian Lassos. *Bayesian Analysis*, 2: 369–412 (2010)
[12] Lauritzen, S.L. Graphical models. Clarendon Press, Oxford, 1996
[13] Meinshausen, N., Buhlmann, P. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34: 1436–1462 (2006)
[14] Park, T., Casella, G. The Bayesian Lasso. *Journal of the American Statistical Association*, 103: 681–686 (2008)
[15] Roverato, A. 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 (2002)

[16] Sachs, K., Perez, O., Peer, D., Lauffenburger, D.A., Nolan, G.P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308: 523–529 (2005)
[17] Scott, J.G., Carvalho, C.M. Feature-inclusion stochastic search for Gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17: 790–808 (2008)
[18] Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Ser. B*, 58: 267–288 (1996)
[19] Wang, H., West, M. Bayesian analysis of matrix normal graphical models. *Biometrika*, 96: 821–834 (2009)
[20] Wong, F., Carter, C.K., Kohn, R. Efficient estimation of covariance selection models. *Biometrika*, 90: 809–830 (2003)
[21] Whittaker, J. Graphical Models in Applied Multivariate Statistics. John Wiley and Sons, Chichester, 1990
[22] Wang, H. The Bayesian graphical Lasso and efficient posterior computation. *Bayesian Analysis*, 7: 771–790 (2012)
[23] Wang, H., Carvalho, C.M. Simulation of hyper-inverse wishart distributions for non-decomposable graphs. *Electronic Journal of Statistics*, 4: 1470–1475 (2010)
[24] Yuan, M., Lin, Y. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94: 19–35 (2007)
[25] Yuan, M. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11: 2261–2286 (2010)