Bayesian Analysis (2014) **9**, Number 2, pp. 449–474

Bayesian Regularization via Graph Laplacian

Fei Liu *, Sounak Chakraborty †, Fan Li ‡, Yan Liu §, and Aurelie C. Lozano ¶

Abstract. Regularization plays a critical role in modern statistical research, especially in high-dimensional variable selection problems. Existing Bayesian methods usually assume independence between variables a priori. In this article, we propose a novel Bayesian approach, which explicitly models the dependence structure through a graph Laplacian matrix. We also generalize the graph Laplacian to allow both positively and negatively correlated variables. A prior distribution for the graph Laplacian is then proposed, which allows conjugacy and thereby greatly simplifies the computation. We show that the proposed Bayesian model leads to proper posterior distribution. Connection is made between our method and some existing regularization methods, such as Elastic Net, Lasso, Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) and Ridge regression. An efficient Markov Chain Monte Carlo method based on parameter augmentation is developed for posterior computation. Finally, we demonstrate the method through several simulation studies and an application on a real data set involving key performance indicators of electronics companies.

Keywords: Bayesian analysis, Elastic Net, Grouping, Lasso, OSCAR, Regularization, Ridge regression, Variable selection

1 Introduction

Regularization plays a critical role in modern statistical research, especially in high dimensional variable selection problems. For example, the ridge regression (Hoerl and Kennard 1970) utilizes the L_2 norm of regression coefficients as the regularization term. Though marked improvement in prediction has been observed over the ordinary least squares, the ridge regression does not lead to sparse estimates. Unlike the ridge regression, the least absolute shrinkage and selection operator (Lasso), proposed in Tibshirani (1996), is based on the regularization of the L_1 norm of regression coefficients. Lasso results in simultaneous shrinkage and variable selection, as many of the coefficients will be estimated exactly as zero. For problems where the explanatory variables are possibly highly correlated, a variety of penalty terms have been proposed to incorporate the grouping structures of variables. See, for example, the Elastic Net (EN)

*Liu is Assistant Professor, Queens College, The City University of New York, Flushing, NY Fei.Liu@qc.cuny.edu

†Chakraborty is Associate Professor, Department of Statistics, University of Missouri, Columbia, MO chakrabortys@missouri.edu

‡Li is Assistant Professor, Department of Statistical Science, Duke University, Durham, NC fl35@stat.duke.edu

§Liu is Assistant Professor, Department of Computer Science, University of Southern California, Los Angles, CA yanliu.cs@usc.edu

¶Lozano is Research Staff Member, IBM Watson Research Center, Yorktown Heights, NY aclozano@us.ibm.com

 DOI:10.1214/14-BA860

penalty in Zou and Hastie (2005) and the grouped Lasso penalty in Yuan and Lin (2006). Recent developments include the octagonal shrinkage and clustering (OSCAR) penalty in Bondell and Reich (2008), the correlation-based penalty in Tutz and Ulbricht (2009). More recently the graph-constrained (Grace) penalty and the adaptive graph-constrained (aGrace) penalty of Li and Li (2010) is successfully used to model biological graphs and networks in genomic data sets.

In the Bayesian framework, regularization problems are formulated through shrinkage priors. For example, the Bayesian Lasso has been discussed in Tipping (2001), Park and Casella (2008), and Hans (2009). Recent advances in Bayesian Elastic Net can be found in Kyung et al. (2010), Bornn et al. (2010), Li and Lin (2010), and Hans (2011). For binary data, Chakraborty and Guo (2011) proposed a Bayesian hybrid Huberized support vector machine with an elastic-net prior for variable selection for microarray data. There have also been enormous developments of prior distributions under the Bayesian variable selection framework (George and McCulloch 1993; Smith and Kohn 1996; George and McCulloch 1997; Kuo and Mallick 1998). More recently, Li and Zhang (2010) and Vannucci and Stingo (2011) developed Bayesian variable selection for structured variables and showed their enormous potential in genomics and biological pathway selections. All these methods assume either independence *a priori* between variables or a completely known dependence structure. Explicitly modeling the dependence structure between variables is challenging, especially in high dimensional problems.

Real world applications, however, have suggested a substantial need for modeling the dependence structure. Our case study example in Section 5.2 provides one such example from the area of business analytics. The primary interest is to predict the future revenue of a company using predictor variables of both financial performance metrics (such as Revenue growth, earnings per shares (EPS), productivity (Revenue/Employee), ROA (Return on Asset), Market Cap Growth, etc.) and lower-level operational metrics (such as Revenue per R&D Spend, Business Weeks Investing 4 Future Index, etc.). It can be seen that these predictor variables are correlated. Moreover, in many cases, the dependence could be either positively correlated or negatively correlated. For example, revenue growth and productivity are positively correlated while innovation index is negatively correlated with revenue per R&D Spend. How we can incorporate the dependence to make better predictions and to automatically infer the underlying dependency relations poses great challenges, which is the main focus of this paper.

Modeling the dependence structure between variables is also desirable from the methodological point of view. In most regularization problems, the interest lies in identifying a subset from a large number of predictor variables, that have some legitimate predictive power on the response. Modeling the dependence structure enables borrowing information across variables, thus leading to better predictive power (Storey and Tibshirani 2003; Kim and Xing 2009; Li and Li 2010). Additionally, it also overcomes the difficulty of collinearity in the presence of highly correlated predictors by imposing identifiability constraints through the dependence structure.

In this article, we propose a Bayesian regularization approach to model the dependence

structure in normal linear models. Our method explicitly characterizes the dependence structure between variables through a graph Laplacian matrix in the spectral graph theory (Ng et al. 2002; Li and Li 2010). Graph Laplacian matrices are the main tools for spectral clustering, whose primary interest is to detect similarity between data points. Traditional graph Laplacian matrices, however, only allow for positive partial correlations among nodes or variables. We generalize the graph Laplacian to allow for negative partial correlations. A prior distribution is then proposed for the generalized graph Laplacian. We show that the proposed prior leads to proper posterior distributions. It can be shown that the proposed Bayesian model is in favor of sparsity and clustering *a priori*. We also establish the connections between the proposed method and the existing regularization methods such as Lasso, EN, and OSCAR. For computation, we develop an efficient Markov chain Monte Carlo (MCMC) algorithm, based on data augmentation.

The rest of this article is organized as follows. Section 2 provides a brief review of the related literature. Section 3 introduces the proposed Bayesian model and the theoretical support. Section 4 describes the computational algorithms. Section 5 gives numerical results, both in simulations and in a business analytics context. Section 6 concludes with a discussion. Details of proofs and computations are relegated to the Appendices.

2 Related literature

Consider the normal linear regression:

$$\boldsymbol{Y} \quad = \quad \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,, \boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)\,, \tag{1}$$

where $\boldsymbol{Y}$ is the $n \times 1$ vector of the dependent variables, $\boldsymbol{X}$ is the $n \times p$ design matrix with the (i, j)th element x_{ij} $(i = 1, \ldots, n; j = 1, \ldots, p)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is the vector of the regression coefficients, and $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.. An estimate for $\boldsymbol{\beta}$ is sparse if most of the β_j are set to zero. Sparsity is critical especially when one wants to identify the signal variables from a massive number of predictors. Grouping variables arises as one wants to detect clusters of the signal variables (nonzero β's). This is of particular interest when there are highly correlated predictors in the data, yet the correlation structure is unknown. In an extreme case where there is only one group, all the signal variables are expected to take exactly the same value, whereas when there are no groups, the signal variables are expected to take different values. The grouping idea is most relevant to the applications of detecting similarly behaving variables or features in a data set. Nevertheless, as suggested in Zou and Hastie (2005), grouping variables can lead to improvements of the prediction accuracy. The main objective of this article is to develop a Bayesian regularized method, which simultaneously encourages sparsity and grouping.

2.1 Laplace matrix of graphs

The *Laplace matrices* of graphs or the *graph Laplacians* are the main tools for spectral clustering algorithms, whose focuses are to find good clusters in the machine learning and pattern recognition literature. von Luxburg (2007) provided an excellent review on the graph Laplacians. Let $x_1, \ldots, x_n$ be a set of n data points. Define $s_{ij} \geq 0$ as the measure of similarity between any pair of points x_i and x_j. A *similarity graph* $G = (V, E)$ can be used to represent the data, where each vertex v_i represents the data point x_i, and two vertices v_i and v_j are connected by an edge weighted by s_{ij} if the similarity measure s_{ij} is positive. For an undirected graph $G = (V, E)$ with vertex set $V = (v_1, \ldots, v_n)$, its *weighted adjacency matrix* is defined as $\boldsymbol{W} = (w_{ij})_{i,j=1,\ldots,n}$ with $w_{ij} = w_{ji} \geq 0$, where $w_{ij} = 0$ implies that the vertices v_i and v_j are disconnected. The degree of a vertex $v_i \in V$ is then defined as $d_i = \sum_{j=1}^{n} w_{ij}$. The *Laplace matrix* of the graph G is defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, where $D = \mathrm{diag}(d_1, \ldots, d_n)$.

The spectral clustering algorithms have been shown to be very effective in detecting clusters (see, e.g., Ng et al. 2002; von Luxburg 2007, and references therein). For the regression problem in (1), however, they are not immediately applicable due to the following limitations. First, they directly cluster the observations $x_1, \ldots, x_n$ rather than the unobserved quantities such as the regression coefficients. Second, they assume that information about the weighted adjacency matrix is readily available, which is not the case under the scenario of our consideration. Finally, the restriction that $w_{ij} \geq 0$ implies that they assume positive partial correlations among all pairs of the variables *a priori*, which may be unrealistic for real world applications. The restriction that $w_{ij} \geq 0$ is due to that fact that w_{ij} is the number of edges from vertex i to vertex j. In Section 3, we will extend the graph Laplacians to overcome these difficulties.

2.2 Regularized approaches

Classical regularization methods minimize the residual sum of squares subject to an imposed penalty term. A variety of penalty terms have been proposed in the literature, among which most relevant to this paper are Ridge, Lasso, EN, and OSCAR. The Ridge regression circumvents the issue of predictor collinearity by a penalty term, which is defined as the L_2 norm of the regression coefficients. To achieve a sparse solution, the Lasso utilizes the L_1 norm as the penalty term. In light of the need to include or exclude together strongly correlated predictors and to retain the sparsity at the same time, the EN utilizes a combination of the L_1 and L_2 norms as the penalty term. Recently, motivated by the need to determine the predictive clusters of the selected variables, the OSCAR penalty combines the L_1 and L_∞ norms. The OSCAR penalty has an octagonal shape, which results in the exact grouping property in that the coefficients of the same group are exactly equal. We summarize these methods in Table 1.

Table 1: Penalty terms in Ridge, Lasso, EN and OSCAR.

Method	Tuning Parameters	Penalty
Ridge	λ	$\lambda \sum_{j=1}^{p} \beta_j^2$
Lasso	λ	$\lambda \sum_{j=1}^{p} \mid \beta_j \mid$
EN	λ_1, λ_2	$\lambda_1 \sum_{j=1}^{p} \mid \beta_j \mid + \lambda_2 \sum_{j=1}^{p} \beta_j^2$
OSCAR	λ, c	$\lambda \sum_{j=1}^{p} \mid \beta_j \mid + c\lambda \sum_{j<k} \max\{ \mid \beta_j \mid , \mid \beta_k \mid \}$

2.3 Bayesian regularized approaches

Bayesian shrinkage priors corresponding to the Lasso, grouped Lasso, EN and Fused Lasso have been proposed (Park and Casella 2008; Hans 2009; Kyung et al. 2010; Li and Lin 2010). Under the shrinkage prior formulations, the penalties correspond to the special choices of priors and are expressed as the "scale-mixture" of normal and gamma distributions. We summarize the mixture representation of the Bayesian shrinkage priors in Table 2. Here, $\boldsymbol{D}_\tau = \text{diag}(\tau_1^2, \ldots, \tau_p^2)$ in the Bayesian Lasso; $\boldsymbol{D}_\tau^* = \text{diag}\{(\frac{1}{\tau_j^2} + \lambda_2)^{-1}\}$ in the Bayesian EN; m_k is the prespecified number of variables in the kth group in the Bayesian Group Lasso, $p = \sum_{k=1}^{K} m_k$; and in the Bayesian Fused Lasso, the diagonal elements for Σ_β are $\frac{1}{\tau_i^2} + \frac{1}{\omega_{i-1}^2} + \frac{1}{\omega_i^2}$, $i = 1, \ldots, p$, and the off diagonals are $-\frac{1}{\omega_i^2}$, $i = 1, \ldots, p$.

Table 2: Scale Mixture Representations of the Bayesian Shrinkage Priors.

Methods	$\pi(\boldsymbol{\beta} \mid \sigma^2, \tau_1, \ldots, \tau_p)$	$\pi(\tau_1, \ldots, \tau_p)$
Lasso	$\text{N}_p(0, \sigma^2 \boldsymbol{D}_\tau)$	$\prod_{j=1}^{p} \frac{\lambda^2}{2} \exp\left(-\lambda^2 \tau_j^2/2\right)$
EN	$\text{N}_p(0, \sigma^2 \boldsymbol{D}_\tau^*)$	$\prod \frac{\lambda_1^2}{2} \exp\left(\lambda_1^2 \tau_j^2/2\right)$
Group Lasso (with K groups)	$\text{N}_{m_k}(0, \sigma^2 \tau_k^2 I_{m_k})$	$\prod_{k=1}^{K} \text{Gamma}((m_k+1)/2, \lambda^2/2)$
Fused Lasso	$\text{N}_p(0, \sigma^2 \Sigma_\beta)$	$\prod_{j=1}^{p} \frac{\lambda_1^2}{2} \exp\left(-\lambda_1^2 \tau_j^2/2\right)$ and $\prod_{j=1}^{p-1} \frac{\lambda_2^2}{2} \exp\left(-\lambda_2^2 \omega_j^2/2\right)$

Among these priors, only the Group Lasso prior and the Fused Lasso prior take into account the dependence structure among variables, which is assumed completely known. This, however, may not be the case in many applications. Motivated by this gap, here, we propose a Bayesian regularization method for correlated variables with an unknown dependence structure.

3 The proposed method

3.1 Formulation

To avoid the computational cost of inverting a covariance matrix, we directly model the dependence structure through the precision matrix. Conditioning on σ^2, we assign the prior distribution for $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta} \,|\, \sigma^2 \sim \mathrm{N}_p(\mathbf{0}, \frac{\sigma^2}{r}\boldsymbol{\Lambda}^{-1}) \,,$$

where $\boldsymbol{\Lambda}$ is the precision matrix, taking the form,

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1+\lambda_{11}+\sum_{j\neq 1}|\lambda_{1j}| & \lambda_{12} & \dots & \lambda_{1p} \\ \lambda_{21} & 1+\lambda_{22}+\sum_{j\neq 2}|\lambda_{2j}| & \dots & \lambda_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1} & \dots & \dots & 1+\lambda_{pp}+\sum_{j\neq p}|\lambda_{pj}| \end{pmatrix}, \quad (2)$$

with $\lambda_{ij} = \lambda_{ji}$, $\lambda_{ii} > 0$ and the hyperparameter $r \geq 0$. As noted in Park and Casella (2008), conditioning on σ^2 is important to guarantee the unimodality of the full posterior distribution. This is critical for the fast convergence of the Gibbs sampler. Letting $\boldsymbol{\lambda}$ be the collection of all elements in $\boldsymbol{\Lambda}$, we propose the following prior distribution for $\boldsymbol{\lambda}$,

$$\pi(\boldsymbol{\lambda}) \propto C_{a,b}|\boldsymbol{\Lambda}|^{-1/2}\prod_{i=1}^{p}\lambda_{ii}^{-3/2}\exp\left(-\frac{a^2}{2\lambda_{ii}}\right)1(\lambda_{ii}>0)\prod_{j<i}|\lambda_{ij}|^{-3/2}\exp\left(-\frac{b^2}{2|\lambda_{ij}|}\right) \quad (3)$$

where $C_{a,b}$ is the normalizing constant and a, b and r are hyperparameters. It is to be noted that the off diagonal elements of Λ can take both positive and negative values. The prior for σ^2 is specified as $\pi(\sigma^2) \propto 1/\sigma^2$.

We call the prior distribution defined in equations (2) and (3) the *Graph Laplacian prior* (GL-prior) because $\boldsymbol{\Lambda}$ can be considered as an extended version of the graph Laplacian matrix. In fact, the connections between the GL-prior and the graph Laplacian matrix can be immediately seen by defining the degree of a vertex i as $d_i = r + \lambda_{ii} + \sum_{j=1}^{p}|\lambda_{ij}|$ and the "weighted adjacency matrix" $\boldsymbol{S} = (\lambda_{ij})_{i,j=1,\dots,n}$ in (2). There are, however, the following differences in between. First, the graph Laplacian matrix is assumed known in the spectral clustering algorithms, whereas the matrix here is completely unknown and needs to be learned from the data. Second, the off-diagonal elements in $\boldsymbol{\Lambda}$ can take both positive and negative values in the GL-prior formulation whereas the off-diagonal elements are all negative in the original graph Laplacians. The advantage of this generalization is that it allows for both positive and negative partial correlations between two coefficients (the original graph Laplacian matrix only allows for positive partial correlations). Finally, we remark in Proposition 1 (with the proof given in Appendix 6) that being diagonally dominant, $\boldsymbol{\Lambda}$ is positive definite and invertible, and thus is a valid form for the precision matrix.

Proposition 1. *The precision matrix defined in* (2) *is symmetric and positive semidefinite.*

The specification in (2) and (3) are the key components in our method. (2) explicitly models the dependence structure through the precision matrix, while the prior distribution in (3) makes it possible to draw inference upon such dependence structure via information provided by the data. Note that there is a factor of $|\boldsymbol{\Lambda}|^{-1/2}$ in the prior distribution of $\boldsymbol{\lambda}$ in (3). This suggests that λ_{ij}s are not independent. The advantage of including this factor in the prior is that it cancels out with $|\boldsymbol{\Lambda}|^{1/2}$ in the likelihood, leading to a closed form marginal prior distribution for $\boldsymbol{\beta}$ after integrating $\boldsymbol{\lambda}$ out. This, described in the next Proposition, is our key result in establishing the connections between our method and the classical regularization methods.

Proposition 2. *Let* $c_{ij} = sign(\lambda_{ij})$. *Conditioning on* $\boldsymbol{c} = \{c_{ij}, j < i\}$ *and* σ^2, $\pi(\boldsymbol{\beta} \,|\, \boldsymbol{c}, \sigma^2)$, *the prior distribution of* $\boldsymbol{\beta}$, *can be written as*

$$\pi(\boldsymbol{\beta} \,|\, \boldsymbol{c}, r, a, b, \sigma^2) \propto (2\pi\sigma^2)^{-p/2} \exp\left\{ -\frac{1}{2\sigma^2} \left(r \sum_i \beta_i^2 + ra\sigma \sum_i |\beta_i| + rb\sigma \sum_{j<i} |\beta_i + c_{ij}\beta_j| \right) \right\}. \tag{4}$$

Note that in (4), r, $ra\sigma$ and $rb\sigma$ correspond to the tuning parameters of the L_2, L_1, and a piecewise L_1 regularization, respectively. We further establish the connection with the OSCAR penalty term as follows. Let $c_{ij} = -\text{sign}(\beta_i\beta_j)$ and write $\max\{|\beta_i|, |\beta_j|\} = (|\beta_i|+|\beta_j|)/2 + (||\beta_i|-|\beta_j||)/2 = (|\beta_i|+|\beta_j|)/2 + (|\beta_i + c_{ij}\beta_i|)/2$, which implies that the OSCAR penalty term is a combination of $\sum_i |\beta_i|$ and $\sum_{j<i} |\beta_i + c_{ij}\beta_j|$. The marginal prior in (4) clearly indicates that our method is in favor of both sparsity and grouping, where a and b reflect the degree of sparseness and that of grouping, respectively. We further illustrate this fact with the 3-d density plot of $\pi(\boldsymbol{\beta} \,|\, \sigma^2)$ and 2-d contour plot of $-\log(\pi(\beta))$ in a 2-dimensional case in Figure 1. Here, we set $a = b = r = \sigma^2 = 1$. The contour plot is octagonal in shape, with the eight vertices being joined by an arc instead of by a straight line. Similar to the Lasso, the four vertices that lie on the horizontal or vertical axis are critical for selection of variables. The remaining four vertices, on the other hand, play the important role of grouping the highly correlated variables. Similar to the EN, the edges in the 2-d contour plots have some curvature, which is due to the quadratic term in the prior (4). From the connections with the OSCAR penalty, the last term in (4) effectively penalizes the pairwise L_∞ norm of the coefficients. As a result, our method is in favor of coefficient groups with modest or small effect sizes. This makes our method suitable for problems with modest or small coefficients such as genetic association studies.

3.2 Theoretical results

In this section, we establish some theoretical results for the proposed method. Due to the length of details, we relegate the proofs to the appendix.

We first show that the prior distribution for $\boldsymbol{\lambda}$ is proper in the next Proposition.

Proposition 3. *The prior distribution defined in* (3) *is proper.*

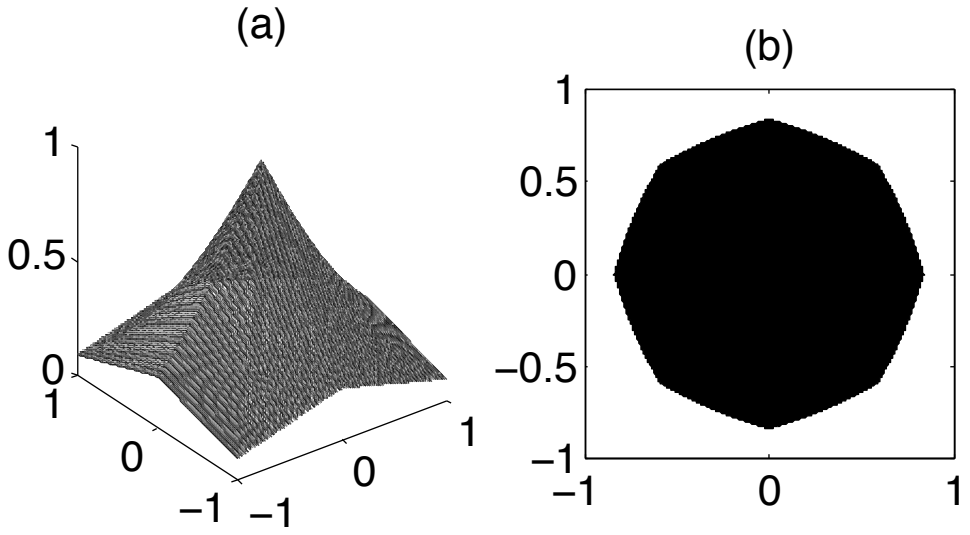

Figure 1: (a) the 3-d plot of the (unnormalized) density function $\pi(\boldsymbol{\beta} \,|\, \sigma^2)$ and (b) the 2-d contour plot of $-\log(\pi(\boldsymbol{\beta} \,|\, \sigma^2))$. The parameters are set to $r = 1, a = 1, b = 1, \sigma^2 = 1$.

Despite a proper distribution on $\boldsymbol{\lambda}$, we specify an improper prior distribution on σ^2. Thus, we still need to show that the proposed model leads to a proper posterior distribution. In the proposition to follow, we show that the posterior distribution is proper under a mild condition.

Proposition 4. *The joint posterior distribution for* $\sigma^2, \boldsymbol{\beta}, \boldsymbol{\lambda}$ *is proper if* $\boldsymbol{y}^{'}\boldsymbol{y} \neq 0$.

This proposition assures the validity of Bayesian inference. The condition $\boldsymbol{y}^{'}\boldsymbol{y} \neq 0$ is equivalent to $\boldsymbol{y} \neq \mathbf{0}$. In fact, as can be seen from the proof, the posterior distribution is dominated by the ridge regression. In addition, it suggests that the proposed method can be applied to "small n, large p" problems.

4 Posterior Computation

Let $\mathbf{D} = (\boldsymbol{X}, \boldsymbol{y})$. The likelihood function from model (1) is

$$\mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \sigma^2; \mathbf{D}) = (2\pi\sigma^2)^{-n/2} \exp\left\{-(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{'}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/2\sigma^2\right\}.$$

Combined with the priors $\pi(\sigma^2) \propto 1/\sigma^2$, $\pi(\boldsymbol{\beta} \,|\, \boldsymbol{\lambda}, \sigma^2)$ and $\pi(\boldsymbol{\lambda})$, we obtain the joint posterior distribution as

$$\begin{aligned}\pi(\sigma^2, \boldsymbol{\beta}, \boldsymbol{\lambda} \,|\, \mathbf{D}) \quad \propto \\ \sigma^{-(n+p+2)} \left\{ \prod_i \lambda_{ii}^{-3/2} \prod_{j<i} |\lambda_{ij}|^{-3/2} \right\} \exp\left\{ -(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/2\sigma^2 \right\} \\ \times \exp\left\{ -\frac{r}{2\sigma^2} \boldsymbol{\beta}' \boldsymbol{\Lambda} \boldsymbol{\beta} - \frac{a^2}{2} \sum_i \lambda_{ii}^{-1} - \frac{b^2}{2} \sum_{j<i} |\lambda_{ij}|^{-1} \right\} . \end{aligned} \tag{5}$$

The full conditional posterior distributions for σ^2 and $\boldsymbol{\beta}$ have closed forms. Specifically, we have $\left(\boldsymbol{\beta} \,|\, \sigma^2, \boldsymbol{\lambda}, \mathbf{D}\right) \sim \mathrm{N}_p(\boldsymbol{\mu}_{\boldsymbol{\beta}},\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ with

$$\boldsymbol{\mu}_{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + r\boldsymbol{\Lambda})^{-1} \boldsymbol{X}' \widetilde{\boldsymbol{y}}\,, \quad \text{and} \quad \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \sigma^2 (\boldsymbol{X}'\boldsymbol{X} + r\boldsymbol{\Lambda})^{-1}\,.$$

After integrating out $\boldsymbol{\beta}$ in (5), we have

$$\pi(\sigma^2 \,|\, \boldsymbol{\lambda}, \mathbf{D}) \propto (\sigma^2)^{-n/2-1} \exp\{ -\boldsymbol{y}' \left(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + r\boldsymbol{\Lambda})^{-1} \boldsymbol{X}' \right) \boldsymbol{y}/2\sigma^2 \}\,,$$

which implies that $\sigma^2 \,|\, \boldsymbol{\lambda}, \mathbf{D} \sim \text{Inv-Gamma}\left(n/2, \boldsymbol{y}' \left(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + r\boldsymbol{\Lambda})^{-1} \boldsymbol{X}' \right) \boldsymbol{y}/2 \right)$, an inverse gamma distribution with the shape parameter and rate parameter as specified.

Finally, the full conditional posterior distribution for $\boldsymbol{\lambda}$ can be obtained as

$$\pi(\boldsymbol{\lambda} \,|\, \boldsymbol{\beta}, \sigma^2, \mathbf{D}) \propto \prod_i \lambda_{ii}^{-3/2} \prod_{j<i} |\lambda_{ij}|^{-3/2} \exp\{ -\frac{r}{2\sigma^2} \boldsymbol{\beta}' \boldsymbol{\Lambda} \boldsymbol{\beta} - \frac{a^2}{2} \sum_i \lambda_{ii}^{-1} - \frac{b^2}{2} \sum_{i<j} |\lambda_{ij}|^{-1} \}\,. \tag{6}$$

This distribution does not have closed form. In the section to follow, we will develop an efficient MCMC method for posterior inference based on parameter augmentation.

4.1 Markov Chain Monte Carlo method

For posterior computation, we use MCMC methodology (Gelfand and Smith 1990). We propose an efficient Gibbs sampler based on parameter augmentation.

Let $\boldsymbol{\theta}_1 = (\sigma^2, \boldsymbol{\beta})$. Within each iteration of the Gibbs sampler, we update $\boldsymbol{\theta}_1$ as a block. The posterior distribution of $\boldsymbol{\theta}_1$, conditioning on $\boldsymbol{\lambda}$, can be written as the product of two terms: $\pi(\boldsymbol{\theta}_1 \,|\, \boldsymbol{\lambda}, \mathbf{D}) = \pi(\sigma^2 \,|\, \boldsymbol{\lambda}, \mathbf{D}) \pi(\boldsymbol{\beta} \,|\, \sigma^2, \boldsymbol{\lambda}, \mathbf{D})$. All these distributions are in closed forms and are very straightforward to be sampled from, as discussed earlier.

To draw from $\pi(\boldsymbol{\lambda} \,|\, \boldsymbol{\theta}_1, \mathbf{D})$, we first augment the parameter space. Let $\eta_{ij} = |\lambda_{ij}|$ and $c_{ij} = \text{sign}(\lambda_{ij})$. For notational simplicity, we set $\boldsymbol{\eta} = \{\eta_{ij}, i = 1, \ldots, p; j = 1, \ldots, p\}$ and $\boldsymbol{c} = \{c_{ij}, j < i\}$. Conditioning on $\boldsymbol{\theta}_1$, the joint posterior distribution for $\boldsymbol{c}, \boldsymbol{\eta}$ is independent of μ and can be written as $\pi(\boldsymbol{c}, \boldsymbol{\eta} \,|\, \boldsymbol{\theta}_1, \mathbf{D}) = \pi(\boldsymbol{c} \,|\, \boldsymbol{\beta}, \sigma^2, \mathbf{D}) \pi(\boldsymbol{\eta} \,|\, \boldsymbol{c}, \boldsymbol{\beta}, \sigma^2, \mathbf{D})$. For the

first term $\pi(\boldsymbol{c}\,|\,\boldsymbol{\beta},\sigma^2,\mathbf{D})$, note that c_{ij} is a discrete random variable and it can take only one of the two values, $+1$ or -1. Furthermore, from (3), c_{ij} are mutually independent conditioning on $\boldsymbol{\beta}$ and σ^2. Let p_{ij} be the probability that $c_{ij}=1$ conditioning on $\boldsymbol{\beta}$ and σ^2 ($c_{ij}=-1$ with probability $1-p_{ij}$), we have

$$p_{ij}=[1+\exp\{-rb(|\beta_i-\beta_j|-|\beta_i+\beta_j|)/2\sigma\}]^{-1}\ , j<i\,.$$

For $\pi(\boldsymbol{\eta}\,|\,\boldsymbol{c},\boldsymbol{\beta},\sigma^2,\mathbf{D})$, plugging $\boldsymbol{\beta}^{'}\boldsymbol{\Lambda}\boldsymbol{\beta}=\sum_i(1+\lambda_{ii})\beta_i^2+\sum_i\sum_{j<i}|\lambda_{ij}|(\beta_i+c_{ij}\beta_j)^2$ into (6) and eliminating the irrelevant terms, we have

$$\pi(\boldsymbol{\eta}\,|\,\boldsymbol{c},\boldsymbol{\beta},\sigma^2,\mathbf{D})\propto$$
$$\prod_i\eta_{ii}^{-3/2}\prod_{j<i}\eta_{ij}^{-3/2}\exp\{-\sum_i\left(\frac{r\beta_i^2\eta_{ii}}{2\sigma^2}+\frac{a^2}{2\eta_{ii}}\right)-\sum_i\sum_{j<i}\frac{r(\beta_i+c_{ij}\beta_j)^2\eta_{ij}}{2\sigma^2}-\frac{b^2}{2}\sum_{j<i}\eta_{ij}^{-1}\}\,.$$

This implies that, conditioning on $\boldsymbol{c}$, $\boldsymbol{\beta}$ and σ^2, η_{ij} are mutually independent. In addition, for η_{ii}, we have $\pi(\eta_{ii}\,|\,\boldsymbol{c},\boldsymbol{\beta},\sigma^2,\mathbf{D})\propto\eta_{ii}^{-3/2}\exp(-a^2\eta_{ii}^{-1}/2-r\beta_i^2\eta_{ii}/2\sigma^2)$. Completing the square, we have

$$\pi(\eta_{ii}\,|\,\boldsymbol{c},\boldsymbol{\beta},\sigma^2,\mathbf{D})\propto\frac{1}{\eta_{ii}^{3/2}}\exp\left[-\frac{r\beta_i^2\left(\eta_{ii}-a\sigma|\sqrt{r}\beta_i|^{-1}\right)^2}{2\sigma^2\eta_{ii}}\right]\,.$$

Therefore $(\eta_{ii}\,|\,\boldsymbol{\beta},\sigma^2,\mathbf{D})\sim$ Inv-Gaussian$(a\sigma|\sqrt{r}\beta_i|^{-1},a^2)$, where Inv-Gaussian$(\mu,\lambda)$ represents an Inverse Gaussian distribution, whose density function is defined as

$$f(x;\mu,\lambda)=\left[\frac{\lambda}{2\pi x}\right]^{1/2}\exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2x}\right\}\ , x>0,\mu>0,\lambda>0\,.$$

Similarly, for η_{ij}, we have

$$\pi(\eta_{ij}\,|\,\boldsymbol{c},\boldsymbol{\beta},\sigma^2,\mathbf{D})\propto\eta_{ij}^{-3/2}\exp\left\{-\frac{r|\beta_i+c_{ij}\beta_j|^2}{2\sigma^2\eta_{ij}}\left(\eta_{ij}-b\sigma|\sqrt{r}(\beta_i+c_{ij}\beta_j)|^{-1}\right)^2\right\}\,.$$

Thus, $(\eta_{ij}\,|\,\boldsymbol{c},\boldsymbol{\beta},\sigma^2,\mathbf{D})\sim$ Inv-Gaussian$(b\sigma|\sqrt{r}(\beta_i+c_{ij}\beta_j)|^{-1},b^2)$.

The Gibbs sampler iterates through the following steps:

(i) Update σ^2 by sampling from $\pi(\sigma^2\,|\,\boldsymbol{\lambda},\mathbf{D})$, which is an inverse Gamma distribution.

(ii) Update $\boldsymbol{\beta}$ by sampling from $\pi(\boldsymbol{\beta}\,|\,\sigma^2,\boldsymbol{\lambda},\mathbf{D})$, which is a multivariate normal distribution.

(iii) Update $\boldsymbol{c}$ by sampling from $\pi(\boldsymbol{c}\,|\,\boldsymbol{\beta},\sigma^2,\mathbf{D})$, where $c_{ij}=1$ with probability p_{ij} and $c_{ij}=-1$ with probability $1-p_{ij}$.

(iv) Update $\boldsymbol{\eta}$ by sampling from $\pi(\boldsymbol{\eta}\,|\,\boldsymbol{c},\boldsymbol{\beta},\sigma^2,\mathbf{D})$, the product of independent inverse Gaussian distributions.

(v) Set $\lambda_{ii}=\eta_{ii}$ and $\lambda_{ij}=c_{ij}\eta_{ij}$, for $j\le i$ and $i=1,\ldots,p$.

At the end of the MCMC simulation, we obtain a sample of draws from the posterior distribution in (5). After the burn-in period, we obtain N MCMC draws

$$\left\{\sigma^{2(h)}, \boldsymbol{\beta}^{(h)}, \boldsymbol{\lambda}^{(h)}; h = 1, \ldots, N\right\}.$$

This leads to our final samples. The results to follow are based on the MCMC samples from the posterior.

4.2 Selection of hyperparameters

For hyperparameters r, a, and b, we assign the following prior

$$\pi(r, a, b) \propto C_{a,b}^{-1} r^{h_r - 1} \exp(-g_r r) \exp(-g_a a) \exp(-g_b b).$$

To allow a relatively flat prior, we recommend small values for g_a, h_b, g_b (we set these values to 0.01 in our numerical experiments). Conditioning on $\boldsymbol{c}$ and $\boldsymbol{\beta}$, we have

$$\begin{aligned}
[r \,|\, a, b, \boldsymbol{c}, \boldsymbol{\beta}] &\sim \text{Gamma}\left(\frac{p}{2} + h_r, \frac{\sum_i \beta^2}{2\sigma^2} + \frac{a\sum_i |\beta_i|}{2\sigma} + \frac{b\sum_{i<j} |\beta_i + c_{ij}\beta_j|}{2\sigma}\right), \\
[a \,|\, r, b, \boldsymbol{c}, \boldsymbol{\beta}] &\sim \text{Exp}\left(g_a + \frac{r\sum_i |\beta_i|}{2\sigma}\right), \\
[b \,|\, r, a, \boldsymbol{c}, \boldsymbol{\beta}] &\sim \text{Exp}\left(g_b + \frac{r\sum_{i<j} |\beta_i + c_{ij}\beta_j|}{2\sigma}\right).
\end{aligned}$$

At each iteration, we update these hyperparameters by drawing samples from their full conditional distributions.

5 Numerical Results

5.1 Simulations

In this section, we compare the performance of our method with that of Lasso, EN, OSCAR, Bayesian Lasso (BLasso) and Bayesian Elastic Net (BEN). In specific, we are interested in the predictive performance when the predictor variables are not independent, with an underlying correlation structure. We simulate data from the regression model $y = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$ with $\epsilon \sim \text{N}(0, \sigma^2)$. We consider five different scenarios, which are very similar to those in Zou and Hastie (2005) and Bondell and Reich (2008). For each scenario, we generate training data and testing data. We first apply each method to the training data to obtain the estimates of the coefficients. We then obtain the estimates for the testing data using the coefficients estimates and calculate the test mean squared error (MSE), calculated as $\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^2/n$. For the Bayesian methods, we use the posterior means as the estimates of the regression coefficients. To select the tuning parameters for Lasso, EN, and OSCAR, we perform five-fold cross validations. The combination of tuning parameters that produce the lowest cross validation scores

is then used in the final Lasso, EN and OSCAR models. We repeat this procedure 100 times and report the 10th, 50th, and 90th percentile of the test MSEs.

The five simulation scenarios are:

(i) Under Scenario 1, each training data set has sample size $n.train = 20$ and each testing data set has sample size $n.test = 200$. The variance for the error term is $\sigma = 3$. The correlation between the ith column and the jth column of the design matrix is $\text{corr}(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0.7^{|i-j|}$. The true regression coefficients are $\boldsymbol{\beta} = (3, 2, 1.5, 0, 0, 0, 0, 0)$.

(ii) Scenario 2 is the same as above, except $\boldsymbol{\beta} = (3, 0, 0, 1.5, 0, 0, 0, 2)$.

(iii) Scenario 3 is the same as above, except $\boldsymbol{\beta} = (\underbrace{0.85, \ldots, 0.85}_{8})$.

(iv) Under Scenario 4, each training data set has sample size $n.train = 100$, and each testing data set has sample size $n.test = 400$. $\sigma = 15$, $\text{corr}(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0.5$, and

$$\boldsymbol{\beta} = (\underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10}) .$$

(v) Under Scenario 5, each training data set has sample size $n.train = 100$, and each testing data set has sample size $n.test = 400$. $\sigma = 15$, and $\boldsymbol{\beta} = (\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{25})$. Letting $\epsilon_i^x \sim \text{N}(0, 0.16)$, the predictors are generated from

$$\begin{cases} \boldsymbol{x}_i = Z_1 + \epsilon_i^x, & Z_1 \sim \text{N}(0,1), \quad i = 1, \ldots, 5 \\ \boldsymbol{x}_i = Z_2 + \epsilon_i^x, & Z_1 \sim \text{N}(0,1), \quad i = 6, \ldots, 10 \\ \boldsymbol{x}_i = Z_3 + \epsilon_i^x, & Z_1 \sim \text{N}(0,1), \quad i = 11, \ldots, 15 \\ \boldsymbol{x}_i \sim \text{N}(0,1), & i = 16, \ldots, 40. \end{cases}$$

In Table 3, we show the 10th, 50th, and 90th percentile of the test MSEs for our method, and that of Lasso, EN, OSCAR, BLasso and BEN. As we can see from the table, our method is highly competitive under all five scenarios. It has the best performance under Scenarios 2-5, and the second to the best performance under Scenario 1. In addition, we can also see that EN and BEN generally perform better than Lasso, OSCAR and BLasso. When the correlations between predictors are moderate such as under scenarios 1-4, our method performs similarly to EN and BEN. When the predictors is highly correlated (Scenario 5), a significant improvement over EN and BEN has been observed. The last column in Table 3 shows the 10th, 50th, and 90th percentiles of the running time (in seconds) for our method. The method is very efficient in terms of computation. Under Scenarios 1-3, the dimension is relatively low and the median computational time is about 6 to 7 seconds for 2000 MCMC iterations. Even when the dimension increases to $p = 40$ as under scenarios 4-5, the median computation time is only about 30-40 seconds to generate 2000 MCMC iterations. Overall, our method improves the predictive performance over the existing methods and it is computationally efficient.

In Table 4, we report the operating characteristics measuring our variable selection performance. The strategy for variable selection is to use the scaled neighborhood criterion (SNC) (Li and Lin 2010). A variable is included if the posterior probability $P\left\{|\beta_j| > \sqrt{var(\beta_j|D)} \,\middle|\, D\right\}$ exceeds a certain threshold, ψ. Following Li and Lin (2010), we set $\psi = 0.5$. The operating characteristics reported in Table 4 are true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), and the negative predictive value (NPV). In the ideal case, only the truly relevant variables should be selected by a method, all the four measures should be 1 (or 100 %). From Table 4, we can see that our method outperforms all other competing methods under all five simulation scenarios.

Study	Pct	Method	Lasso	EN	OSCAR	BLasso	BEN	Time
1	10	10.05	10.01	9.43	10.88	10.55	10.10	5.85
	50	11.62	13.00	11.27	15.11	12.55	11.76	6.73
	90	14.76	18.20	15.46	24.12	16.39	15.21	19.37
2	10	9.17	9.62	9.61	11.00	9.73	9.49	6.03
	50	11.92	13.19	12.76	14.55	12.47	12.00	6.77
	90	14.73	19.35	17.12	20.73	15.51	14.54	17.96
3	10	9.22	10.60	9.46	10.45	9.68	9.17	5.93
	50	10.40	13.19	10.97	13.97	11.24	10.72	6.28
	90	13.42	20.09	15.37	21.87	15.16	13.54	18.19
4	10	231.72	255.73	236.13	309.00	265.53	243.07	31.12
	50	262.16	288.80	267.20	372.09	300.51	269.72	37.86
	90	295.91	340.63	299.95	449.06	339.24	301.05	52.23
5	10	239.77	253.38	248.76	251.93	275.75	241.76	57.23
	50	333.52	509.67	460.04	374.54	362.66	353.19	80.06
	90	748.83	1629.92	1512.43	1100.50	1337.27	767.93	90.74

Table 3: Test MSEs of the simulation studies. The results are based on 100 replicated data sets. The tuning parameters in Lasso and EN are chosen according to 5-fold cross validation. Pct stands for percentile. Time represent the time to compute in seconds

5.2 Real Data Analysis

In this section, we present some results of applying our method on a real world data set involving key performance indicators (KPI's) of electronics companies. The problem of monitoring and analyzing performance indicators of corporations is important in business investment decision making, and has received considerable attention (Kaplan and Norton 1992, 1996). This particular data set was obtained from Standard and Poor's Compustat database, available at *http://www.compustat.com*.

The data set consists of values of various performance indicators for electronics companies that are in the industry group of "semiconductors and semiconductor equipments." Specifically, quarterly data over the duration of three years were pulled, for companies

Simulation Study	Methods	TPR %	TNR %	PPV %	NPV
1	Our Method	100.0	100.0	100.0	100.0
	Lasso	100.0	80.0	75.0	100.0
	EN	100.0	60.0	60.0	100.0
	OSCAR	100.0	93.8	96.3	99.9
	BLasso	100.0	85.0	88.8	98.3
	BEN	100.0	87.3	92.6	96.4
2	Our Method	100.0	98.0	100.0	99.8
	Lasso	100.0	84.0	78.0	96.9
	EN	87.5	65.0	69.0	98.0
	OSCAR	99.8	95.6	97.0	97.8
	BLasso	98.9	83.8	85.8	98.0
	BEN	94.3	84.2	94.4	99.1
3	Our Method	100.0	NA	100.0	NA
	Lasso	100.0	NA	100.0	NA
	EN	91.5	NA	100.0	NA
	OSCAR	97.4	NA	100.0	NA
	BLasso	98.3	NA	100.0	NA
	BEN	95.0	NA	100.0	NA
4	Our Method	100.0	92.5	93.0	100
	Lasso	97.0	75.0	79.1	96.3
	EN	100.0	45.0	64.5	100.0
	OSCAR	100.0	35.0	61.0	100.0
	BLasso	98.6	82.0	84.5	98.3
	BEN	99.1	71.0	77.3	98.6
5	Our Method	98.6	91.0	94.8	97.5
	Lasso	95.1	68.0	83.2	89.3
	EN	97.8	45.0	74.7	92.4
	OSCAR	93.9	92.4	95.3	90.0
	BLasso	96.3	82.4	90.1	93.0
	BEN	96.9	84.6	91.2	94.2

Table 4: Simulation results based on 100 replications. Overall median of operating characteristics

having at least 25 million dollars in annual revenue. The performance indicators in the data set include financial performance metrics such as Revenue growth, EBIT (Earnings before Interest and Tax) margin, productivity (Revenue/Employee), ROA (Return on Asset), Market Cap Growth, Earnings per Share (EPS), PE (Price Earning) Ratio, and Beta. The data also include lower level (operational) metrics such as Revenue per R&D Spend, Business Week's Investing 4 Future Index, Capital Expenditure/Revenue, Current Ratio, Working Capital/Revenue, COGS/Revenue (Cost of Goods Sold), SG&A (Selling, General & Administrative Expense) Revenue, Operating Cash Flow/Revenue, Inventory Cost/Revenue, Inventory Turnover, Cash conversion cycle in days, and Net Working Capital Ratio.

For many of these metrics, we consider both "absolute" values and the "CAGR" values or the "Compound Annual Growth Rate", which measures the annual rate of growth of the KPI in question. We note that some normalization and outlier filtering were

performed in generating these data: for example, the values in each column were normalized by subtracting the sample mean and dividing it by the standard deviation, and outliers that fall outside of 3 standard deviations of the mean in each column were treated as "missing" values and replaced with the median of that column.

There are many interesting questions we can ask from the data. One that particularly interests us is what performance indicators will affect the future values of revenue growth, and how they interact with each other. In order to answer this question, we run our method with lagged variables of 2 time stamps of all variables as predictor variables, leading to a total number of 60 predictors in the model. The response variable in the model is the revenue growth. We use the posterior mean to estimate the coefficients. The estimate is not sparse in the absolute sense, but our solution is close to a sparse one. In fact, as suggested in Figure 2, the largest 22 coefficients contribute more than 90% of the L_2 norms of the coefficient estimates. We thus threshold the rest of the coefficients to zero and reported the selected variables in Appendix B. It is interesting to note that the revenue growth not only depends highly on financial performance metrics, such as EBIT margin, PE ratio, Earnings per Share (EPS) and Beta, but is also related to lower-level operational metrics, such as Innovation Index, Inventory Turnover, and Cash conversion cycle in days.

Based on the MCMC draws of $\boldsymbol{\Lambda}$, we can estimate the dependence structure. Let $\widehat{\boldsymbol{\Lambda}}$ be the posterior mean of $\boldsymbol{\Lambda}$, we may estimate the covariance matrix of $\boldsymbol{\beta}$ by $(\boldsymbol{X}^{'}\boldsymbol{X}+\widehat{\boldsymbol{\Lambda}})^{-1}$ and use it to quantify the dependence. In Figure 4, we compare the estimated correlation matrix and sample correlation matrix of $\boldsymbol{X}$. Since our primary goal is to infer the dependence structure, we take the absolute values of each entry. As indicated from the heatmap, the dependence structure given by our method is significantly more concise than the sample correlation matrix. In a practical application such as this one, the number of variables of interest tends to be sizable. For practical use, therefore, it is critical that the presented information is concise for reasonable interpretability. We should point out that the estimate of $\boldsymbol{\Lambda}$ is not sparse. The sparsity is more likely due to the shrinkage effect on $\boldsymbol{\beta}$. In Figure 3 we report the posterior estimate (off-diagonals) of the correlation matrix.

Figure 5 shows the resulting graph by removing the edges with values less than or equal to 0.1, where the thickness of the edges corresponds to the strengths of the correlation. From the graph, we can see that there usually exist strong correlations between lagged variables of the same feature (i.e. lag-1 and lag-2), such as PE ratio, EBIT margin, and Inventory turnover. Interestingly, Figure 5 suggests that the lag-2 variable of EBIT (earnings before interest and taxes) Margin has strong correlations with the lag-1 variable of CAGR Inventory Turnover. It might be useful to group them into one group for future modeling. The validity of such observations would need to be verified by further investigation, but it is at least suggestive of the potential value that our method could provide in applications to corporate level business decision making.

Finally, for comparison purposes, we apply the g-prior approach for Bayesian variable selection of Liang et al. (2008). With 60 predictors in this data, it is impossible to enumerate all the 2^{60} models in the model space as in Liang et al. (2008). Instead, we

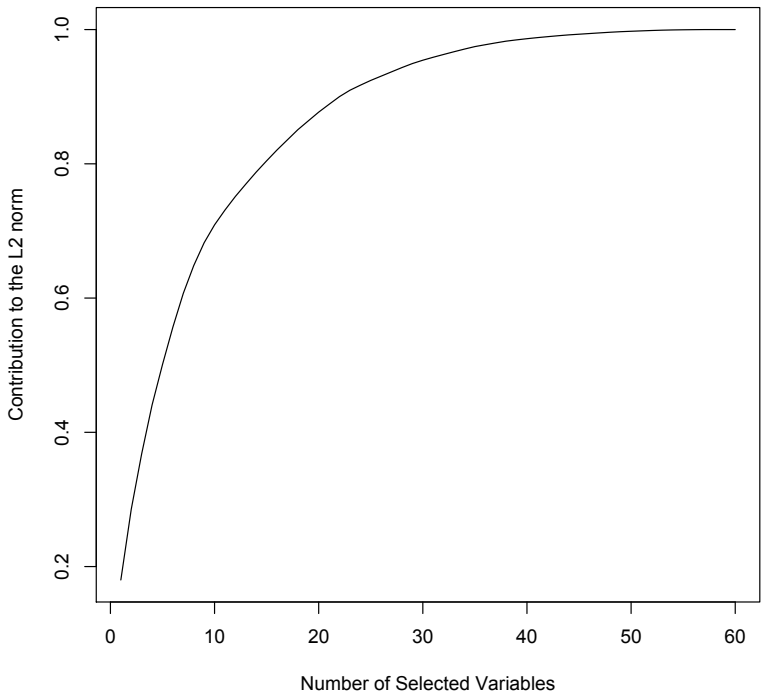

Figure 2: The percentage of the contribution to the L_2 norm of $\boldsymbol{\beta}$ as a function of the number of selected variables. The largest 22 coefficients contribute more than 90%.

use the Gibbs sampler algorithm available for moderately large dimensional problems suggested in Garcia-Donato and Martinez-Beneito (2013). This can be done by using the R function *GibbsBvs()* in the R package *BayesVarSel* (http://cran.r-project.org/web/packages/BayesVarSel/BayesVarSel.pdf) on this data set. Among the 22 non-zero coefficients that are selected by our method, 15 of them are also selected by the g-prior approach. We also notice that our approach results in a sparser model than the g-prior approach (the g-prior approach selects 35 non-zero coefficients). This is likely due to the extra L_1 norm penalty and the pairwise L_1 norm penalty introduced by the dependence structure of the coefficients. We list the variables that are selected by the g-prior approach in the Appendix.

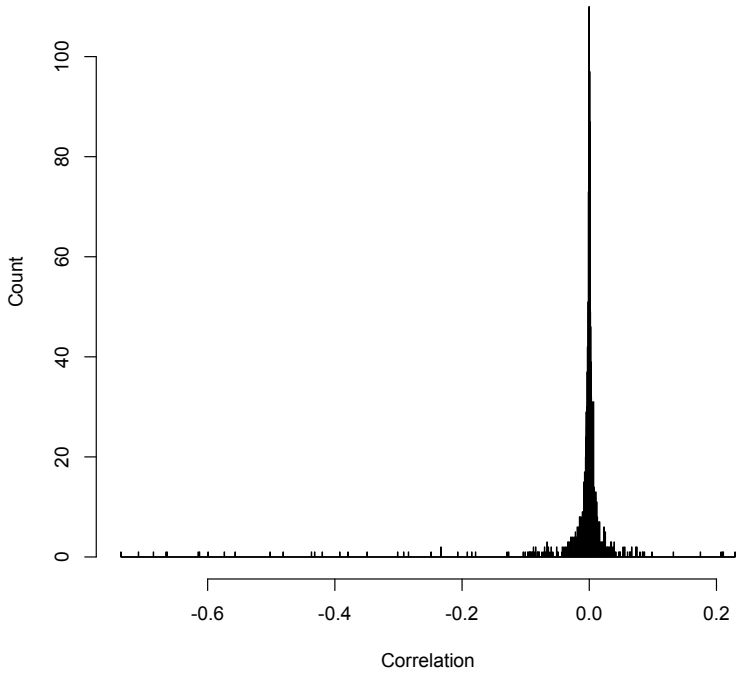

Figure 3: The histogram of the off-diagonal entries in the estimated correlation matrix of $\boldsymbol{\beta}$.

6 Discussion

We have introduced a new Bayesian method for regularized regression, which provides inference on the inter-relationship between variables by explicitly modeling through a graph Laplacian matrix. Our formulation has a strong motivation from the dependence structure as defined by the undirected graphs in spectral clustering. The prior distribution proposed for the graph Laplacian matrix allows us to learn the dependence structure from the data. This can be critical in real applications where the dependence structure is unknown. In the event when prior knowledge is available, some coefficients should be set equal. One may use the average of the corresponding columns as a predictor variable in the model.

We have established the connection between our method and the classical regularized regression methods, which suggests that our method is in favor of sparseness and variable clustering. For posterior computation, an efficient Gibbs sampler has been developed

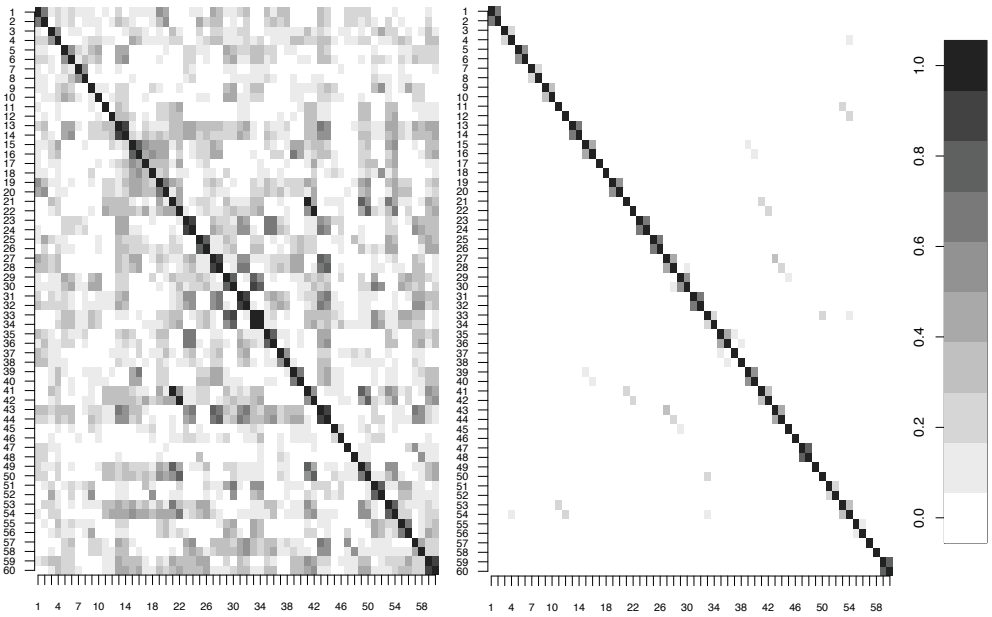

Figure 4: Left: the sample correlation matrix of the design matrix; Right: the estimated correlation matrix of $\boldsymbol{\beta}$.

based on parameter augmentation. The proposed Gibbs sampler is found to be very convenient and highly efficient. We have applied the proposed method with success to simulation studies and a real data analysis.

The proposed method provides a general framework to incorporate dependence structure among predictors in Bayesian regularized regression. For example, extension to the generalized linear models (GLM) in McCullagh and Nelder (1989) is straightforward within the Bayesian hierarchical modeling framework. The major challenge in such an extension is the posterior computation, due to the complexity of the likelihood function. The quadratic approximation in Gelman et al. (2004) may be used to speed up the computation.

For fast posterior computation, we develop an MCMC algorithm based on parameter augmentation. The parameter augmentation leads to closed forms for the conditional posterior distributions, which greatly simplifies the computation. From our experience, the algorithm works well for moderately large dimension problems (e.g., about 100 predictors). While sampling $\boldsymbol{\beta}$, the algorithm involves inverting the $p \times p$ matrix $(\boldsymbol{X}^{'}\boldsymbol{X} + r\boldsymbol{\Lambda})$ in each iteration. For very high dimensional problems with more than thousands of predictors, this computation becomes infeasible. For such problems, we propose a scalable Expectation Maximization (EM) based algorithm for statistical inference based on the Maximum a posteriori (MAP) estimates. The algorithm leverages an efficient

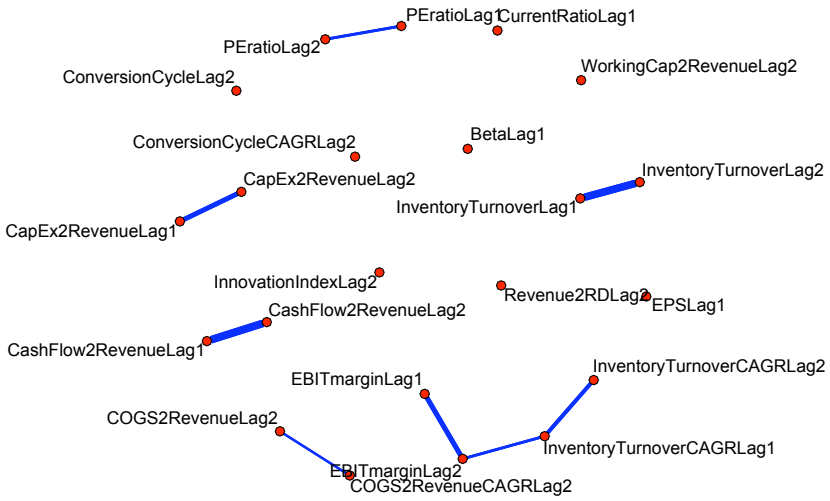

Figure 5: The graph corresponding to the estimated correlation matrix of $\boldsymbol{\beta}$ for the selected variables. The thickness of the edges reflects the strength of the correlation.

augmented Lagrangian approach for the maximization step, and is very efficient for high dimensional problems such as microarray data. Due to the length of technical details, we address this approach in a separate paper.

An alternative approach that may be used to tackle the high dimensional problem is to first reduce the dimensionality via the "spike and slab" prior (Mitchell and Beauchamp 1988; George and McCulloch 1993; Chipman 1996; George and McCulloch 1997; Clyde et al. 1998; Kuo and Mallick 1998; Ishwaran and Rao 2005), and then model the dependence for the selected coefficients as discussed in this paper. Since the dependence structure relies on which coefficients are selected, the technical challenge is to develop a coherent modeling framework for all possible models. This will be a focus of our future work.

References

Bondell, H. and Reich, B. (2008). "Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR." *Biometrics*, 64: 115–123. 450, 459

Bornn, L., Gottardo, R., and Doucet, A. (2010). "Grouping Priors and the Bayesian Elastic Net." Technical Report 254, The University of British Columbia, Department of Statistics. 450

Chakraborty, S. and Guo, R. (2011). "Bayesian Hybrid Huberized SVM and its Applications in High Dimensional Medical Data." *Computational Statistics and Data Analysis*, 55(3): 1342 – 1356. 450

Chipman, H. (1996). "Bayesian variable selection with related predictors." *Canadian Journal of Statistics*, 24: 17–36. 467

Clyde, M., Parmigiani, G., and Vidakovic, B. (1998). "Multiple shrinkage and subset slection in wavelets." *Biometrika*, 85: 391–401. 467

Garcia-Donato, G. and Martinez-Beneito, M. (2013). "On sampling strategies in Bayesian variable selection problems with large model spaces." *Journal of the American Statistical Association*, 108: 340–352. 464, 473

Gelfand, A. E. and Smith, A. F. M. (1990). "Sampling-based Approaches for Calculating Marginal Densities." *Journal of the American Statistial Association*, 85: 398–409. 457

Gelman, A., Carlin, J., Stern, H., and Rubin, R. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC. 466

George, E. I. and McCulloch, R. E. (1993). "Variable Selection Via Gibbs Sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 450, 467

— (1997). "Approaches for Bayesian Variable Selection." *Statistica Sinica*, 7: 339–373. 450, 467

Hans, C. (2009). "Bayesian lasso regression." *Biometrika*, 96(4): 835–845. 450, 453

— (2011). "Elastic Net Regression Modeling With the Orthant Normal Prior." *Journal of the American Statistical Association*, 106: 1383–1393. 450

Hoerl, A. E. and Kennard, R. W. (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics*, 12(1): 55–67. 449

Ishwaran, H. and Rao, J. S. (2005). "Spike and slab variable selection: frequentist and Bayesian strategies." *Annals of Statistics*, 33(2): 730–773. 467

Kaplan, R. and Norton, D. (1992). "The Balanced Scorecard - Measures that Drive Performance." In *Harvard Business Review*, 71–79. 461

— (1996). *The Balanced Scorecard: Translating Strategy into Action*. Boston, MA: Harvard Business School Press. 461

Kim, S. and Xing, E. P. (2009). "Statistical Estimation of Correlated Genome Associations to a Quantitative Trait Network." *Public Library of Science Genetics*, 5(8). 450

Kuo, L. and Mallick, B. (1998). "Variable selection for regression models." *Sankhya Series B*, 60: 65–81. 450, 467

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). "Penalized Regression, Standard Errors, and Bayesian Lassos." *Bayesian Analysis*, 5(2): 369–412. 450, 453

Li, C. and Li, H. (2010). "Variable Selection and Regression Analysis for Graph-structured Covariates with an Application to Genomics." *Annals of Applied Statistics*, 4(3): 1498–1516. 450, 451

Li, F. and Zhang, N. (2010). "Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics." *Journal of the American Statistical Association*, 105(491): 1202–1214. 450

Li, Q. and Lin, N. (2010). "The Bayesian Elastic Net." *Bayesian Analysis*, 5(1): 151–170. 450, 453, 461

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g Priors for Bayesian Variable Selection." *Journal of the American Statistical Association*, 103(481): 410–423. 463, 473

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall / CRC. 466

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian variable selection in linear regression." *Journal of American Statistical Association*, 83: 1023–1036. 467

Ng, A., Jordan, M., and Weiss, Y. (2002). "On spectral clustering: analysis and an algorithm." In Dietterich, T., Becker, S., and Ghahramani, Z. (eds.), *Advances in Neural Information Processing Systems 14*. MIT Press. 451, 452

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–686. 450, 453, 454

Smith, M. and Kohn, R. (1996). "Nonparametric Regression Using Bayesian Variable Selection." *Journal of Econometrics*, 75: 317–343. 450

Storey, J. D. and Tibshirani, R. (2003). "Statistical significance for genomewide studies." *Proceedings of the National Academy of Sciences. USA*, 100: 9440 – 9445. 450

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society, Series B*, 58(1): 267–288. 449

Tipping, M. E. (2001). "Sparse Bayesian Learning and the Relevance Vector Machine." *Journal of Machine Learning Research*, 1: 211–244. 450

Tutz, G. and Ulbricht, J. (2009). "Penalized regression with correlation-based penalty." *Statistical Computing*, 19: 239–253. 450

Vannucci, M. and Stingo, F. (2011). "Bayesian Models for Variable Selection that Incorporate Biological Information (with discussion)." In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (eds.), *Bayesian Statistics*, volume 9, 659–678. Oxford University Press. 450

von Luxburg, U. (2007). "A Tutorial on Spectral Clustering." *Statistics and Computing*, 17(4): 395–416. 452

Yuan, M. and Lin, Y. (2006). "Model Selection and Estimation in Regression with Grouped Variables." *Journal of the Royal Statistical Society, Series B*, 68(1): 49–67. 450

Zou, H. and Hastie, T. (2005). "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society, Series B*, 67(2): 301–320. 450, 451, 459

Appendix A: Proofs

- Proof of Proposition 1.

  Proof Letting $\boldsymbol{x} = (x_1, \ldots, x_p)^{'} \in \mathrm{R}^p$, $\boldsymbol{x} \neq \mathbf{0}$, we have,

$$\begin{aligned}
\boldsymbol{x}^{'}\boldsymbol{\Lambda}\boldsymbol{x} &= \sum_i (1+\lambda_{ii})x_i^2 + \sum_i \sum_{j\neq i} |\lambda_{ij}| x_i^2 + \sum_i \sum_{j\neq i} \lambda_{ij} x_i x_j \\
&= \sum_i (1+\lambda_{ii})x_i^2 + \sum_i \sum_{j<i} |\lambda_{ij}| (x_i^2 + x_j^2) + \sum_i \sum_{j<i} 2\lambda_{ij} x_i x_j \\
&= \sum_i (1+\lambda_{ii})x_i^2 + \sum_i \sum_{j<i} |\lambda_{ij}| (x_i^2 + x_j^2 + 2c_{ij} x_i x_j) \\
&= \sum_i (1+\lambda_{ii})x_i^2 + \sum_i \sum_{j<i} |\lambda_{ij}| (x_i + c_{ij} x_j)^2 > 0\,,
\end{aligned}$$

  where $c_{ij} = \mathrm{sign}(\lambda_{ij})$. By definition, $\boldsymbol{\Lambda}$ is positive semidefinite.

- Proof of Proposition 2.

  Proof Conditioning on $\boldsymbol{\Lambda}$, σ^2, r, a and b, the prior distribution for $\boldsymbol{\beta}$ can be written as

$$\pi(\boldsymbol{\beta} \,|\, \boldsymbol{\lambda}, \sigma^2) \propto \frac{(\sigma^2)^{p/2}}{|\boldsymbol{\Lambda}|^{1/2}} \exp\left\{ -\frac{r}{2\sigma^2} \left(\sum_{i=1}^{p} (1+\lambda_{ii})\beta_i^2 + \sum_{j<i} |\lambda_{ij}| (\beta_i + c_{ij}\beta_j)^2 \right) \right\}.$$

where $c_{ij} = \text{sign}(\lambda_{ij})$. Integrating out $\boldsymbol{\lambda}$ with respect to the prior distributions in (3), we have,

$$\begin{aligned}
&\pi(\boldsymbol{\beta} \,|\, \boldsymbol{c}, \sigma^2, r, a, b) = \int \pi(\boldsymbol{\beta} \,|\, \boldsymbol{\lambda}, \sigma^2, r, a, b)\pi(\boldsymbol{\lambda} \,|\, a, b) d\boldsymbol{\lambda} \\
&\propto \exp\left(-\frac{r}{2\sigma^2}\sum_{i=1}^{p}\beta_i^2\right)\prod_{i=1}^{p}\int_0^{\infty}\exp\left(-\frac{r\lambda_{ii}}{2\sigma^2}\beta_i^2\right)\lambda_{ii}^{-\frac{3}{2}}\exp\left(-\frac{a^2}{2\lambda_{ii}}\right)\mathrm{d}\lambda_{ii} \\
&\times\prod_{i=1}^{p}\prod_{j<i}\int_{-\infty}^{\infty}\exp\left\{-\frac{r|\lambda_{ij}|}{2\sigma^2}(\beta_i + c_{ij}\beta_j)^2\right\}|\lambda_{ij}|^{-3/2}\exp\left(-\frac{b^2}{2|\lambda_{ij}|}\right)\mathrm{d}\lambda_{ij}\,.
\end{aligned}$$

Applying the following fact to the above equation,

$$\frac{1}{a}\exp(-a|z|) = \int_0^{\infty}(2\pi)^{-1/2}t^{-3/2}\exp\left(-\frac{z^2 t}{2}\right)\exp\left(-\frac{a^2}{2t}\right)\mathrm{d}t$$

we have,

$$\pi(\boldsymbol{\beta} \,|\, \boldsymbol{c}, \sigma^2) \propto (\sigma^2)^{-p/2}\exp\left\{-\frac{r}{2\sigma^2}\left(\sum_i \beta_i^2 + a\sigma\sum_i|\beta_i| + b\sigma\sum_{j<i}|\beta_i + c_{ij}\beta_j|\right)\right\},$$

completing the proof.

- Proof of Proposition 3

Proof We first show that $|\boldsymbol{\Lambda}| \geq 1$. Let $\boldsymbol{A} = \boldsymbol{\Lambda} - \boldsymbol{I}_n$. It is easy to see that $\boldsymbol{A}$ is symmetric and positive semidefinite. Let $\boldsymbol{D} = \text{diag}(a_1, \ldots, a_p)$ where $0 \leq a_1 \leq a_2 \leq \cdots \leq a_p$ are the p eigenvalues of $\boldsymbol{A}$. There exists an orthonormal matrix $\boldsymbol{T}$ such that $\boldsymbol{A} = \boldsymbol{T}\boldsymbol{D}\boldsymbol{T}'$. Therefore, we have

$$\boldsymbol{\Lambda} = \boldsymbol{A} + \boldsymbol{I}_n = \boldsymbol{T}\boldsymbol{D}\boldsymbol{T}' + \boldsymbol{T}\boldsymbol{T}' = \boldsymbol{T}(\boldsymbol{D} + \boldsymbol{I}_n)\boldsymbol{T}',$$

and, consequently, $|\boldsymbol{\Lambda}| = \prod_i (a_i + 1) \geq 1$. Next, we show that $\pi(\boldsymbol{\lambda})$ is proper. Note that

$$\begin{aligned}
&\int |\boldsymbol{\Lambda}|^{-1/2}\prod_{i=1}^{p}2\lambda_{ii}^{-3/2}\exp\left(-\frac{a^2}{2\lambda_{ii}}\right)1(\lambda_{ii} > 0)\prod_{j<i}|\lambda_{ij}|^{-3/2}\exp\left(-\frac{b^2}{2|\lambda_{ij}|}\right)\mathrm{d}\boldsymbol{\lambda} \\
&\leq \int\prod_{i=1}^{p}\frac{a^2}{2\lambda_{ii}^{3/2}}\exp\left(-\frac{a^2}{2\lambda_{ii}}\right)1(\lambda_{ii} > 0)\prod_{j<i}|\lambda_{ij}|^{-3/2}\exp\left(-\frac{b^2}{2|\lambda_{ij}|}\right)\mathrm{d}\boldsymbol{\lambda} \\
&= \left(\prod_{i=1}^{p}\int_0^{\infty}\lambda_{ii}^{-3/2}\exp\left(-\frac{a^2}{2\lambda_{ii}}\right)\mathrm{d}\lambda_{ii}\right)\left(\prod_{i=1}^{p}\prod_{j<i}\int_{-\infty}^{\infty}|\lambda_{ij}|^{-3/2}\exp\left(-\frac{b^2}{2|\lambda_{ij}|}\right)\mathrm{d}\lambda_{ij}\right).
\end{aligned}$$

Note that in the above, the integrands are kernels of the inverse gamma densities. Therefore, the integral is finite, completing the proof.

- Proof of Proposition 4.

  Proof We only need to show that the marginal likelihood is finite. Let $\pi(\boldsymbol{c})$ be the prior probability for each configuration, $\pi(\boldsymbol{c}) \propto 1$. Given the tuning parameters r, a, b, the marginal likelihood of $\boldsymbol{y}$, $m(\boldsymbol{y})$, can be written as

$$
\begin{aligned}
m(\boldsymbol{y}) &= \sum_{\boldsymbol{c}} \pi(\boldsymbol{c}) \int p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) \pi(\boldsymbol{\beta} \,|\, \boldsymbol{c}, \sigma^2) \pi(\sigma^2) \mathrm{d}\boldsymbol{\beta} \mathrm{d}\sigma^2 \\
&\propto \sum_{\boldsymbol{c}} \pi(\boldsymbol{c}) \int (\sigma^2)^{-n/2-1} \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{'} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right) \times (\frac{\sigma^2}{r})^{-p/2} \\
&\times \exp\left(-\frac{1}{2\sigma^2} (r\boldsymbol{\beta}^{'}\boldsymbol{\beta} + ra\sigma \sum_i |\beta_i| + rb\sigma \sum_{j<i} |\beta_j + c_{ij}\beta_i|) \right) \mathrm{d}\boldsymbol{\beta} \mathrm{d}\sigma^2 \\
&\leq \sum_{\boldsymbol{c}} \pi(\boldsymbol{c}) \int (\sigma^2)^{-\frac{(n+p)}{2}-1} r^{p/2} \exp\left(-\frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{'} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + r\boldsymbol{\beta}^{'}\boldsymbol{\beta}}{2\sigma^2} \right) \mathrm{d}\boldsymbol{\beta} \mathrm{d}\sigma^2 \\
&= \int (\sigma^2)^{-\frac{(n+p)}{2}-1} r^{p/2} \exp\left(-\frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{'} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + r\boldsymbol{\beta}^{'}\boldsymbol{\beta}}{2\sigma^2} \right) \mathrm{d}\boldsymbol{\beta} \mathrm{d}\sigma^2 \\
&\propto \int \frac{(\sigma^2)^{-n/2-1} r^{p/2}}{|\boldsymbol{X}^{'}\boldsymbol{X} + r\boldsymbol{I}_n|^{1/2}} \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{y}^{'} \left\{ \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X} + r\boldsymbol{I}_n)^{-1}\boldsymbol{X}^{'} \right\} \boldsymbol{y} \right) \mathrm{d}\sigma^2 \,.
\end{aligned}
$$

  Next, we show that $\frac{|\boldsymbol{X}^{'}\boldsymbol{X} + r\boldsymbol{I}_n|^{-1/2}}{r^{-p/2}} \leq 1$ and $y^{'} \left\{ \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X} + r\boldsymbol{I}_n)^{-1}\boldsymbol{X}^{'} \right\} \boldsymbol{y} > 0$. Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}$ be the singular value decomposition (SVD) of $\boldsymbol{X}$, i.e., $\boldsymbol{U}$ is an $n \times n$ orthonormal matrix , $\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_p)$ is a $p \times p$ diagonal matrix, such that $d_j \geq 0$, for $j = 1, \ldots, p$, and $\boldsymbol{V}$ is a $p \times p$ orthonormal matrix. Therefore, we have $\boldsymbol{U}\boldsymbol{U}^{'} = \boldsymbol{I}_n$ and $\boldsymbol{V}\boldsymbol{V}^{'} = \boldsymbol{I}_p$. As a result,

$$
\begin{aligned}
r^{p/2} |\boldsymbol{X}^{'}\boldsymbol{X} + r\boldsymbol{I}_n|^{-1/2} &\leq r^{p/2} |\boldsymbol{V}^{'}\boldsymbol{D}^{'}\boldsymbol{D}\boldsymbol{V} + r\boldsymbol{I}_n|^{-1/2} \\
&\leq r^{p/2} |\boldsymbol{D}^2 + r\boldsymbol{I}_n|^{-1/2} \\
&\leq r^{p/2} r^{-p/2} \\
&= 1 \,.
\end{aligned}
$$

  Additionally, we have

$$
\begin{aligned}
& \boldsymbol{y}^{'} \left\{ \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X} + r\boldsymbol{I}_n)^{-1}\boldsymbol{X}^{'} \right\} \boldsymbol{y} \\
&= \mathbf{n} y^{'} \left\{ \boldsymbol{I}_n - \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{'}(\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^{'} + r\boldsymbol{I}_n)^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{'} \right\} \boldsymbol{y} \\
&= \boldsymbol{y}^{'}\boldsymbol{U} \left\{ \boldsymbol{I}_n - \boldsymbol{D}(\boldsymbol{D}^2 + r\boldsymbol{I}_n)^{-1}\boldsymbol{D} \right\} \boldsymbol{U}\boldsymbol{y}.
\end{aligned}
$$

  Note that $\boldsymbol{D}(\boldsymbol{D}^2 + r\boldsymbol{I}_n)^{-1}\boldsymbol{D}$ is a diagonal matrix, with the ith diagonal element equal to $d_i^2/(d_i^2 + r) < 1$. Therefore, the above is greater than 0 as long as $\boldsymbol{y} \neq \boldsymbol{0}$.

Plugging these inequalities into above, we have

$$\begin{aligned} m(\boldsymbol{y}) \leq & c \int (\sigma^2)^{-n/2-1} \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^{'}\left\{\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-}\boldsymbol{X}^{'}\right\}\boldsymbol{y}\right) \mathrm{d}\sigma^2 \\ & \propto \left\{\boldsymbol{y}^{'}\left\{\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X} + r\boldsymbol{I}_n)^{-1}\boldsymbol{X}^{'}\right\}\boldsymbol{y}\right\}^{-n/2} \\ & < \infty\,, \end{aligned}$$

completing the proof.

Appendix B: Selected variables in KPI data analysis

Using our approach, the selected variables in KPI analysis are shown in Table 5.

BetaLag1 (∗)	CapEx2RevenueLag1 (∗)
CapEx2RevenueLag2	CashFlow2RevenueLag1
CashFlow2RevenueLag2 (∗)	COGS2RevenueLag2 (∗)
COGS2RevenueCAGRLag2 (∗)	ConversionCycleLag2 (∗)
ConversionCycleCAGRLag2 (∗)	CurrentRatioLag1 (∗)
EBITmarginLag1 (∗)	EBITmarginLag2 (∗)
EPSLag1 (∗)	InnovationIndexLag2
InventoryTurnoverLag2 (∗)	InventoryTurnoverLag1
InventoryTurnoverCAGRLag1	InventoryTurnoverCAGRLag2 (∗)
PEratioLag1 (∗)	PEratioLag2 (∗)
Revenue2RDLag2	WorkingCap2RevenueLag2

Table 5: Variables selected by our method. Variables followed by ∗ are also selected by the g-prior approach.

The variables shown in Table 6 are selected by *GibbsBvs()* in the R package *BayesVarSel* of Garcia-Donato and Martinez-Beneito (2013), which implements a Gibbs sampler algorithm for the g-prior approach in Liang et al. (2008).

Acknowledgments

The authors thank the Associate editor and two anonymous referees for their valuable suggestions and comments. We would also like to thank the University of Missouri Bioinformatics Consortium for letting us use their Dell EM64T Lewis cluster system. This work is supported by the National Science Foundation under the award number DMS-1106717.

BetaLag1 (∗)	BetaLag2
CapEx2RevenueLag1 (∗)	CapEx2RevenueCAGRLag1
CashFlow2RevenueLag2 (∗)	CashFlow2RevenueCAGRLag2
COGS2RevenueLag2 (∗)	COGS2RevenueCAGRLag2 (∗)
ConversionCycleLag1	ConversionCycleLag2 (∗)
ConversionCycleCAGRLag1	ConversionCycleCAGRLag2 (∗)
CurrentRatioLag1 (∗)	EBITmarginLag1 (∗)
EBITmarginLag2 (∗)	EPSLag1 (∗)
EPSLag2	FlexibilityLag1
FlexibilityLag2	InnovationIndexLag1
InnovationIndexCAGRLag1	Inventory2RevenueLag2
Inventory2RevenueCAGRLag2	InventoryTurnoverLag2 (∗)
InventoryTurnoverCAGRLag2 (∗)	MarketCapGrowthLag1
MarketCapGrowthLag2	NetworkingCapRatioLag1
NetworkingCapRatioLag2	PEratioLag1 (∗)
PEratioLag2 (∗)	RevenueGrowthLag1
RevenueGrowthLag2	RevPerEmployeeLag2
WorkingCap2RevenueLag1	

Table 6: Variables selected by the g-prior approach. Variables followed by ∗ are also selected by our method.