Bayesian Analysis (2015) **10**, Number 4, pp. 909–936

Bayesian Variable Selection and Estimation for Group Lasso

Xiaofan Xu* and Malay Ghosh†

Abstract. The paper revisits the Bayesian group lasso and uses spike and slab priors for group variable selection. In the process, the connection of our model with penalized regression is demonstrated, and the role of posterior median for thresholding is pointed out. We show that the posterior median estimator has the oracle property for group variable selection and estimation under orthogonal designs, while the group lasso has suboptimal asymptotic estimation rate when variable selection consistency is achieved. Next we consider bi-level selection problem and propose the Bayesian sparse group selection again with spike and slab priors to select variables both at the group level and also within a group. We demonstrate via simulation that the posterior median estimator of our spike and slab models has excellent performance for both variable selection and estimation.

Keywords: group variable selection, spike and slab prior, Gibbs sampling, median thresholding.

1 Introduction

Group structures of predictors arise naturally in many statistical applications:

- In a regression model, a multi-level categorical predictor is usually represented by a group of dummy variables.
- In an additive model, a continuous predictor may be represented by a group of basis functions to incorporate nonlinear relationship.
- Grouping structure of variables may be introduced into a model to make use of some domain specific prior knowledge. Genes in the same biological pathway, for example, form a natural group.

For a thorough review of the application of group variable selection methods in statistical problems, one may refer to Huang et al. (2012), in which semiparametric regression models, varying coefficients models, seemingly unrelated regressions and analysis of genomic data are discussed.

It is usually desirable to use the prior information on the grouping structure to select variables group-wise. Depending on the application, selecting individual variables in a group may or may not be relevant. We will discuss variable selection methods which

*Department of Statistics, University of Florida, xiaofanxufl@gmail.com
†Department of Statistics, University of Florida, ghoshm@stat.ufl.edu

 DOI: 10.1214/14-BA929

only conduct variable selection at the group level, as well as bi-level selection methods that select variables both at the group level and within group level.

Specifically, we consider a linear regression problem with G factors (groups):

$$\boldsymbol{Y}_{n\times 1} = \sum_{g=1}^{G} \boldsymbol{X}_g \boldsymbol{\beta}_g + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon}_{n\times 1} \sim \boldsymbol{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, $\boldsymbol{\beta}_g$ is a coefficients vector of length m_g, and $\boldsymbol{X}_g$ is an $n \times m_g$ covariate matrix corresponding to the factor $\boldsymbol{\beta}_g$, $g = 1, 2, \ldots, G$. Let p be the total number of predictors, so $p = \sum_{g=1}^{G} m_g$. In the following article, we will use factor and group interchangeably to denote a group of predictors that are formed naturally.

Penalized regression methods have been very popular for the power to select relevant variables and estimate regression coefficients simultaneously. Among them the lasso (Tibshirani, 1996), which puts an upper bound on the L_1-norm of the regression coefficients, draws much attention for its ability to both select and estimate. A distinctive feature of the lasso is that it can produce exact 0 estimates, resulting in automatic model selection with suitably chosen penalty parameter. Least Angle Regression (LARS) makes the lasso even more attractive because the full lasso solution path can be computed with the cost of only one least squares estimation by a modified LARS algorithm (Efron et al., 2004).

With multi-factor analysis of variance problems in mind, Yuan and Lin (2006) proposed the group lasso which generalizes the lasso in order to select grouped variables (factors) for accurate prediction in regression. The group lasso estimator is obtained by solving

$$\min_{\boldsymbol{\beta}} \left\| \boldsymbol{Y} - \sum_{g=1}^{G} \boldsymbol{X}_g \boldsymbol{\beta}_g \right\|_2^2 + \lambda \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_2. \tag{2}$$

We note that the lasso is a special case of the group lasso when all the groups have size 1, i.e., $m_1 = m_2 = \cdots = m_G = 1$.

One major issue with the lasso-type estimates is that it is difficult to give satisfactory standard errors since the limit distribution of the lasso estimator is very complicated (Knight and Fu, 2000; Chatterjee and Lahiri, 2011). But the Bayesian formulation of the lasso can produce reliable standard errors without any extra efforts. Tibshirani (1996) suggested that the lasso estimator is equivalent to the posterior mode with independent double exponential prior for each regression coefficient. Motivated by the fact that the double exponential distribution can be expressed as a scale mixture of normal distributions, Park and Casella (2008) developed a fully Bayesian hierarchical model and an efficient Gibbs sampler for the lasso problem. Kyung et al. (2010) later extended this model and proposed a fairly general fully Bayesian formulation which could accommodate various lasso variations, including the group lasso, the fused lasso (Tibshirani et al., 2004) and the elastic net (Zou and Hastie, 2005) (see also Raman et al. 2009).

Zero inflated mixture priors, an important subclass of spike and slab priors (Mitchell and Beauchamp, 1988), have been utilized towards a Bayesian approach for variable selection. George and McCulloch (1997) used zero inflated normal mixture priors in the hierarchical formulation for variable selection in a linear regression model. To select random effects, Chen and Dunson (2003) allowed some random effects to effectively drop out of the model by choosing mixture priors with point mass at zero for the random effects variances in a linear mixed effects model. Zhao and Sarkar (2012) developed new multiple intervals for selected parameters under the Bayesian lasso model with zero inflated mixture priors.

Point mass mixture priors are also studied by Johnstone and Silverman (2004) for estimation of possibly sparse sequences of Gaussian observations, with an emphasis on utilizing the posterior median, which is proven to be a soft thresholding estimator like the lasso but with data adaptive thresholds. Heavy tailed distributions like double exponential for the continuous part of the mixture are advocated for the purpose of achieving optimal estimation risk. Posterior concentration of such priors on sparse sequences is studied by Castillo and Van Der Vaart (2012).

Following Johnstone and Silverman (2004), Yuan and Lin (2005) combined the power of point mass mixture priors and double exponential distributions in variable selection and estimation, and showed that the resulting empirical Bayes estimator is closely related to the lasso estimator. Lykou and Ntzoufras (2013) proposed a similar mixture prior and focused on specifying the shrinkage parameter λ based on Bayes factors. Zhang et al. (2014) generalized this prior for group variable selection and proposed the hierarchical structured variable selection (HSVS) method for simultaneous selection of grouped variables and variables within a group. They also extended the HSVS method to account for within group serial correlations by using Bayesian fused lasso technique for within group selection. These authors used an FDR-based variable selection technique at the group level and posterior credible intervals for selection of within group variables. The paper considered an interesting application to molecular inversion probe studies in breast cancer.

In this paper, instead of taking a traditional Bayesian approach to group lasso problem (Kyung et al., 2010; Raman et al., 2009), we will develop a Bayesian group lasso model with spike and slab priors (hereafter referred to as BGL-SS) for problems that only require variable selection at the group level. Our procedure consists of a multivariate point mass mixture prior similar to Zhang et al. (2014) and produces exact 0 estimates at the group level to facilitate group variable selection. Marginal posterior median is proven to be a soft thresholding estimator, and can automatically select variables. Simulation results suggest that while prediction accuracy is comparable to the group lasso, median thresholding results in substantial reduction of false positive rate in comparison to the latter.

Another important problem we focus in this paper is the bi-level selection. Simon et al. (2012) proposed sparse group lasso to produce exact 0 coefficients at the group level and also within a group. The sparse group lasso estimator of $\boldsymbol{\beta}$ is given by

$$\min_{\boldsymbol{\beta}} \left(\left\| \boldsymbol{Y} - \sum_{g=1}^{G} \boldsymbol{X}_g \boldsymbol{\beta}_g \right\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_2 \right). \tag{3}$$

With the prior of the form

$$\pi(\boldsymbol{\beta}) \propto \exp\left\{-\lambda_1\|\boldsymbol{\beta}\|_1 - \lambda_2 \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_2\right\}, \tag{4}$$

the posterior mode for problem (1) is equivalent to the sparse group lasso estimator. We will show that (4) can also be expressed as a scale mixture of normals and therefore we can easily provide a full Bayesian implementation of sparse group lasso (BSGL). Next, to improve the BSGL model, which undershrinks the coefficients and cannot automatically select variables, we utilize a hierarchical spike and slab prior structure to select variables both at the group level and within each group. We will refer to this as Bayesian sparse group selection with spike and slab priors (BSGS-SS). We will demonstrate the significant improvement in variable selection and prediction power via simulation examples.

Although our BSGS-SS method is similar to the HSVS method of Zhang et al. (2014), which also focuses on selection of both group variables and variables within selected groups, it differs from the latter in the following sense. To select variables within a group, the HSVS method assumes independent double exponential priors on the regression coefficients and conducts selection via posterior credible intervals. They need to decide the significance level and deal with the complex issue of multiplicity adjustment. Our priors, with another spike and slab distribution at the individual level, can automatically select and estimate variables with posterior median thresholding. So our posterior median estimator can be a good default estimator and has great variable selection and prediction performance.

We stress that a key point of this paper is to advocate the use of posterior median estimator in spike and slab type models as an alternative sparse estimator to the (sparse) group lasso estimator since the former can also select and estimate at the same time. Under an orthogonal design, we will show that they are both soft thresholding estimators, and the median thresholding estimator is consistent in model selection and has optimal asymptotic estimation rate, while the group lasso has to sacrifice estimation rate to achieve selection consistency. The selected model by median thresholding has far lower false positive rate than the model chosen by lasso methods in all our simulation examples. It has even slightly better model selection accuracy than the model with largest posterior probability, which is often a gold standard for stochastic model selection (George and McCulloch, 1997; Geweke, 1994). Also the prediction performance of posterior median estimator is better than the corresponding lasso methods and is marginally better than that of posterior mean. This is not surprising since the latter is a Bayesian model averaging estimator and is widely believed to have optimal prediction performance (Clyde, 1999; Hoeting et al., 1999; Brown et al., 2002).

Griffin and Brown (2012) also addressed the variable selection problem. Their goal was not only to examine whether or not just some of the regression coefficients are zeros, but also whether there exists some clustering or grouping of random effects. They met their target by considering normal-gamma priors. In a later paper, Griffin and Brown (2013) used the same priors, but primarily with the objective of robustifying as well as combining ridge priors with g-priors (Zellner, 1986).

In Section 2, we assume independent multivariate zero inflated mixture prior for each factor in our fully Bayesian formulation of the group lasso (BGL-SS), and derive a Gibbs sampler to compute the posterior mean and median as our estimators of the coefficients. We introduce posterior median thresholding in this section and prove a frequentist oracle property of our procedure for orthogonal designs. Bi-level selection methods are developed in Section 3. In Section 3.1, we will introduce a fully Bayesian hierarchical model for the sparse group lasso and an efficient Gibbs sampler. We further improve this model in Section 3.2 with spike and slab type priors and propose the BSGS-SS model in order to automatically select variables and improve prediction performance. Simulation results are given in Section 4 in which our BGL-SS and BSGS-SS methods show significant improvement in variable selection as compared to the frequentist group lasso and traditional Bayesian group lasso methods. We conclude with a brief discussion in Section 5.

2 Bayesian Group Lasso with Spike and Slab Prior (BGL-SS)

2.1 Model Formulation

We consider the regression problem with grouped variables in (1). Kyung et al. (2010) demonstrated that the prior

$$\pi\left(\boldsymbol{\beta}_g\right) \propto \exp\left\{-\frac{\lambda}{\sigma}\|\boldsymbol{\beta}_g\|_2\right\}, \tag{5}$$

a multivariate generalization of the double exponential prior, can also be expressed as a scale mixture of normals with Gamma hyperpriors. Specifically, with

$$\boldsymbol{\beta}_g|\tau_g^2,\sigma^2 \overset{ind}{\sim} \boldsymbol{N}_{m_g}\left(\mathbf{0},\tau_g^2\sigma^2\boldsymbol{I}_{m_g}\right), \tau_g^2 \overset{ind}{\sim} \text{Gamma}\left(\frac{m_g+1}{2},\frac{\lambda^2}{2}\right), \tag{6}$$

the marginal distribution of $\boldsymbol{\beta}_g$ is of the form (5). This Bayesian formulation encourages shrinkage of coefficients at the group level and provides comparable prediction performance with the group lasso. However, this approach, based on estimation of $\boldsymbol{\beta}_g (g = 1,\ldots,G)$ by posterior means or medians, does not produce exact 0 estimates. To introduce sparsity at the group level and facilitate group variable selection, we assume a multivariate zero inflated mixture prior for each $\boldsymbol{\beta}_g$. We propose the following hierarchical Bayesian group lasso model with an independent spike and slab type prior for each factor $\boldsymbol{\beta}_g$:

$$\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2 \sim \boldsymbol{N}_n(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n), \tag{7}$$

$$\boldsymbol{\beta}_g|\sigma^2,\tau_g^2 \overset{ind}{\sim} (1-\pi_0)\boldsymbol{N}_{m_g}(\mathbf{0},\sigma^2\tau_g^2\boldsymbol{I}_{m_g}) + \pi_0\delta_0(\boldsymbol{\beta}_g), \quad g = 1,2,\ldots,G, \tag{8}$$

$$\tau_g^2 \overset{ind}{\sim} \text{Gamma}\left(\frac{m_g+1}{2},\frac{\lambda^2}{2}\right), \quad g = 1,2,\ldots,G, \tag{9}$$

$$\sigma^2 \sim \text{Inverse Gamma}\left(\alpha,\gamma\right), \quad \sigma^2 > 0, \tag{10}$$

where $\delta_0(\boldsymbol{\beta}_g)$ denotes a point mass at $\mathbf{0} \in \mathbb{R}^{m_g}$, $\boldsymbol{\beta}_g = (\beta_{g1}, \ldots, \beta_{gm_g})^T$. In this paper, a limiting improper prior is used for σ^2, $\pi(\sigma^2) = 1/\sigma^2$.

Fixing π_0 at $\frac{1}{2}$ is a popular choice since it assigns equal prior probabilities to all submodels and represents no prior information on the true model. Instead of fixing π_0, we place a conjugate beta prior on it, $\pi_0 \sim \text{Beta}(a, b)$. We prefer $a = b = 1$ since it gives a prior mean $\frac{1}{2}$ and also allows a prior spread. Under sparsity, for example, in gene selection problems, one may need $\pi_0 \equiv \pi_{0n}$ where $\pi_{0n} \to 1$ as $n \to \infty$.

The value of λ should be carefully tuned. A very large value of λ will overshrink the coefficients and thus yields severely biased estimates; $\lambda \to 0$ will lead to a very diffuse distribution for the slab part and the null model will always be preferred no matter what data we have because of the Lindley paradox (Lindley, 1957). A conjugate gamma prior can be placed on the penalty parameter, $\lambda^2 \sim \text{Gamma}(r, \delta)$. Instead, we will take an empirical Bayes approach and estimate λ from data using marginal maximum likelihood. Since marginal likelihood function for λ is intractable, a Monte Carlo EM algorithm (Casella, 2001; Park and Casella, 2008) can be used to estimate λ. The kth EM update for λ is

$$\lambda^{(k)} = \sqrt{\frac{p+G}{\sum_{g=1}^{G} E_{\lambda^{(k-1)}}\left[\tau_g^2 \mid \boldsymbol{Y}\right]}},$$

in which the posterior expectation of τ_g^2 will be replaced by the sample average of τ_g^2 generated in the Gibbs sampler based on $\lambda^{(k-1)}$.

It should be noted that (8) is essentially a special case of the prior used in Zhang et al. (2014) which conducts shrinkage at both the group level and also the individual level by using independent exponential hyperpriors to induce lasso shrinkage for individual variables. However, focusing on group level selection only, BGL-SS instead uses group lasso prior on the slab part and is tailored for problems that only require group level sparsity.

2.2 Marginal Prior for $\boldsymbol{\beta}_g$ and Connection with Penalized Regression

Integrating out τ_g^2 in (8) and (9), the marginal prior for $\boldsymbol{\beta}_g$ is a mixture of point mass at $\mathbf{0} \in \mathbb{R}^{m_g}$ and a Multi-Laplace distribution:

$$\boldsymbol{\beta}_g|\sigma^2 \sim (1-\pi_0)\,\text{M-Laplace}\left(\mathbf{0}, \frac{\sigma}{\lambda}\right) + \pi_0\delta_0\left(\boldsymbol{\beta}_g\right), \tag{11}$$

where the density function for an m_g-dimensional Multi-Laplace distribution is

$$\text{M-Laplace}\left(\boldsymbol{x}|0, c^{-1}\right) \propto c^{m_g} \exp\left(-c\|x\|_2\right). \tag{12}$$

We can observe from (11) that the marginal prior for $\boldsymbol{\beta}_g$ has two shrinkage effects: one is the point mass at $\mathbf{0}$ which leads to exact 0 coefficients; the other, same as the one considered in the Bayesian group lasso (Kyung et al., 2010; Raman et al., 2009), results in shrinkage at the group level. Combining these two components together facilitates variable selection at the group level and shrinks coefficients in the selected groups at

the same time. For the special case when the dimension of $\boldsymbol{\beta}_g$ is 1, i.e., $m_g = 1$, (11) reduces to a one-dimensional mixture distribution with a point mass at 0 and a double exponential distribution. This has been thoroughly studied by Johnstone and Silverman (2004) and Castillo and Van Der Vaart (2012) for estimation of sparse normal means, and by Yuan and Lin (2005) and Lykou and Ntzoufras (2013) for Bayesian variable selection. Importantly, it was shown that a heavy-tailed distribution for the slab part, such as a double-exponential distribution or a Cauchy-like distribution, is advantageous since that it results in optimal estimation risk with posterior median estimator and optimal posterior contraction rate for sparse means. We will generalize the thresholding result of Johnstone and Silverman (2004) on the posterior median to our multivariate spike and slab type prior (8).

To see the connection between our model and the penalized regression problem, we reparametrize the regression coefficients: $\boldsymbol{\beta}_g = \gamma_g \boldsymbol{b}_g$, where γ_g is an indicator that only takes value 0 or 1, and $\boldsymbol{b}_g = (b_{g1}, b_{g2}, \ldots, b_{gm_g})^T$. We then place a Multi-Laplace prior on $\boldsymbol{b}_g$ and a Bernoulli prior on γ_g,

$$\boldsymbol{b}_g|\sigma \overset{ind}{\sim} \text{M-Laplace}\left(\mathbf{0}, \frac{\sigma}{\lambda}\right), \quad g = 1, 2, \ldots, G, \tag{13}$$

$$\gamma_g \overset{ind}{\sim} \text{Bernoulli}\left(1 - \pi_0\right), \quad g = 1, 2, \ldots, G. \tag{14}$$

Note that with this configuration, the marginal prior distribution of $\boldsymbol{\beta}_g$ is still (11) and this model can only be identified up to $\boldsymbol{\beta}_g = \gamma_g \boldsymbol{b}_g$. The negative log-likelihood under the model (1) and the above prior is

$$-\log L\left(\boldsymbol{b}, \boldsymbol{\gamma}|\boldsymbol{Y}\right) = \frac{1}{2\sigma^2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{\sigma}\sum_{g=1}^{G}\|\boldsymbol{b}_g\|_2 + \log\left(\frac{1-\pi_0}{\pi_0}\right)\sum_{g=1}^{G}\gamma_g + const.$$

Thus the posterior mode of the regression model (1) under this new parametrization is equivalent to the solution of a penalized regression problem with an L_2-penalty on each group of coefficients and an L_0-like penalty, penalizing the number of nonzero groups in the predictors. Solving this penalization regression problem is extremely hard for problems with a moderate to large number of groups of covariates because of the combinatorial optimization problem induced by the L_0-like norm. We would also like to point out that for the special case when all the groups have size 1, if we replace the Laplace prior with Normal prior, it becomes the so-called Bernoulli–Gaussian model or Binary Mask model, and has been applied to variable selection (Kuo and Mallick, 1998) and signal process problems (Zhou et al., 2009; Soussen et al., 2011).

2.3 Posterior Median as an Adaptive Thresholding Estimator

Regarding the wavelet-based nonparametric problem, Abramovich et al. (1998) demonstrated that the traditional Bayes rule with respect to L_2-loss function is a shrinkage rule while the posterior median, which is a Bayes estimator corresponding to L_1-loss, is a thresholding estimator with spike and slab priors. Johnstone and Silverman (2004)

showed that under spike and slab priors for normal means problem, the posterior median is a random thresholding estimator with a couple of desirable properties under fairly general conditions. In this section, we will generalize the thresholding results of Johnstone and Silverman (2004) to multivariate spike and slab priors, with (8) as a special case. First, we focus on only one group:

$$\boldsymbol{Z}_{m\times 1} \sim f\left(\boldsymbol{z}-\boldsymbol{\mu}\right), \tag{15}$$

$$\boldsymbol{\mu} \sim \pi_0\delta_0\left(\boldsymbol{\mu}\right)+\left(1-\pi_0\right)\gamma\left(\boldsymbol{\mu}\right), \tag{16}$$

where $\boldsymbol{Z}$ is an m-dimensional random variable, and $\gamma(\cdot)$ and $f(\cdot)$ are both density functions for m-dimensional random vectors. $f(\boldsymbol{t})$ is maximized at $\boldsymbol{t}=\boldsymbol{0}$. Let $\mathrm{Med}(\mu_i|\boldsymbol{z})$ denote the marginal posterior median of μ_i given data. We define

$$c=\frac{\int f\left(-\boldsymbol{v}\right)\gamma\left(\boldsymbol{v}\right)d\boldsymbol{v}}{f\left(\boldsymbol{0}\right)}\leq\frac{\int f\left(\boldsymbol{0}\right)\gamma\left(\boldsymbol{v}\right)d\boldsymbol{v}}{f\left(\boldsymbol{0}\right)}=1,$$

Then we have the following theorem:

Theorem 1. *Suppose* $\pi_0>\frac{c}{1+c}$, *then there exists a threshold* $t(\pi_0)>0$, *such that when* $\|\boldsymbol{z}\|_2<t$,

$$Med\left(\mu_i|\boldsymbol{z}\right)=0,\ \textit{for any}\ 1\leq i\leq m.$$

Proof. The posterior odds of $\boldsymbol{\mu}\neq\boldsymbol{0}$ given $\boldsymbol{Z}=\boldsymbol{0}$ is

$$\begin{aligned}Odds\left(\boldsymbol{\mu}\neq\boldsymbol{0}|\boldsymbol{Z}=\boldsymbol{0}\right)&=\frac{1-\pi_0}{\pi_0}\frac{\int f\left(\boldsymbol{0}-\boldsymbol{v}\right)\gamma\left(\boldsymbol{v}\right)d\boldsymbol{v}}{f\left(\boldsymbol{0}\right)}\\&=\frac{1-\pi_0}{\pi_0}c\\&<1.\end{aligned}$$

Note that $Odds(\boldsymbol{\mu}\neq\boldsymbol{0}|\boldsymbol{Z}=\boldsymbol{z})$ is a continuous function of $\boldsymbol{z}$. Hence, there exists $t(\pi_0)>0$, such that when $\|\boldsymbol{z}\|_2<t$, $Odds(\boldsymbol{\mu}\neq\boldsymbol{0}|\boldsymbol{Z}=\boldsymbol{z})<1$. Therefore, when $\|\boldsymbol{z}\|_2<t$, for any $i(1\leq i\leq m)$, $P(\mu_i=0|\boldsymbol{Z}=\boldsymbol{z})\geq P(\boldsymbol{\mu}=\boldsymbol{0}|\boldsymbol{Z}=\boldsymbol{z})>\frac{1}{2}$, and we conclude that $\mathrm{Med}(\mu_i|\boldsymbol{z})=0$. □

Suppose now the design matrix X in (7) is block orthogonal, i.e., $\boldsymbol{X}_i^T\boldsymbol{X}_j=\boldsymbol{0}$ for $i\neq j$. Then for $1\leq g\leq G$ we have

$$\hat{\boldsymbol{\beta}}_g=\left(\boldsymbol{X}_g^T\boldsymbol{X}_g\right)^{-1}\boldsymbol{X}_g^T\boldsymbol{Y}\sim N_{m_g}\left(\boldsymbol{\beta}_g,\sigma^2\left(\boldsymbol{X}_g^T\boldsymbol{X}_g\right)^{-1}\right).$$

By Theorem 1, suppose $\pi_0>\frac{c}{1+c}$, then there exists $t(\pi_0)>0$, such that the marginal posterior median of β_{gj} under the prior (8) satisfies

$$\mathrm{Med}\left(\beta_{gj}|\hat{\boldsymbol{\beta}}_g\right)=0\quad\text{for any}\ 1\leq j\leq m_g$$

when $\|\hat{\boldsymbol{\beta}}_g\|_2<t$. Thus the marginal posterior median estimator of the gth group of regression coefficients is zero when the norm of the corresponding block least square estimator is less than certain threshold.

To illustrate the random thresholding property of posterior median estimator, we further assume that the design matrix $\boldsymbol{X}$ is orthogonal, i.e., $\boldsymbol{X}^T\boldsymbol{X} = n\boldsymbol{I}_p$ for the rest of this subsection and consider the model defined by (7) and (8) with fixed $\tau_{g,n}^2 (1 \le g \le G)$. Note that we use the subscript n here to emphasize that τ_g^2 depends on n for developing the asymptotic theory. Under this model, the posterior distribution of $\boldsymbol{\beta}_g$ conditional on the data is still a multivariate spike and slab distribution,

$$\boldsymbol{\beta}_g|\boldsymbol{Y},\boldsymbol{X} \sim l_{g,n}\delta_0\left(\boldsymbol{\beta}_g\right) + \left(1-l_{g,n}\right)\boldsymbol{N}_{m_g}\left(\left(1-B_{g,n}\right)\hat{\boldsymbol{\beta}}_g^{LS}, \frac{\sigma^2}{n}\left(1-B_{g,n}\right)\boldsymbol{I}\right),$$

where $\hat{\boldsymbol{\beta}}_g^{LS}$ is the least squares estimator of $\boldsymbol{\beta}_g$, $B_{g,n} = \frac{1}{1+n\tau_{g,n}^2}$, and

$$l_{g,n} = P\left(\boldsymbol{\beta}_g = \boldsymbol{0}|\boldsymbol{Y},\boldsymbol{X}\right) = \frac{\pi_0}{\pi_0 + \left(1-\pi_0\right)\left(1+n\tau_{g,n}^2\right)^{-m_g/2}\exp\left\{\frac{\left(1-B_{g,n}\right)}{2\sigma^2}n\|\hat{\boldsymbol{\beta}}_g^{LS}\|_2^2\right\}}.$$

Thus the marginal posterior distribution for $\beta_{gj} (1 \le j \le m_g)$ conditional on the observed data is also a spike and slab distribution,

$$\beta_{gj}|\boldsymbol{Y},\boldsymbol{X} \sim l_{g,n}\delta_0\left(\beta_{gj}\right) + \left(1-l_{g,n}\right)N\left(\left(1-B_{g,n}\right)\hat{\beta}_{gj}^{LS}, \frac{\sigma^2}{n}\left(1-B_{g,n}\right)\right).$$

The resulting median, a soft thresholding estimator, is given by

$$\hat{\beta}_{gj}^{Med} \triangleq \mathrm{Med}\left(\beta_{gj}|\boldsymbol{Y},\boldsymbol{X}\right) = \mathrm{sgn}\left(\hat{\beta}_{gj}^{LS}\right)\left(\left(1-B_{g,n}\right)|\hat{\beta}_{gj}^{LS}| - \frac{\sigma}{\sqrt{n}}Q_{g,n}\sqrt{1-B_{g,n}}\right)_+, \tag{17}$$

where z_+ denotes the positive part of z, and $Q_{g,n} = \Phi^{-1}(\frac{1}{2(1-\min(\frac{1}{2},l_{g,n}))})$. This is similar to the group lasso estimator (Yuan and Lin, 2006) which can also be expressed as a soft thresholding estimator under an orthogonal design:

$$\hat{\beta}_{gj}^{GL} = \left(1 - \frac{\lambda_n}{n\|\hat{\boldsymbol{\beta}}_g^{LS}\|_2}\right)_+ \hat{\beta}_{gj}^{LS} = \mathrm{sgn}\left(\hat{\beta}_{gj}^{LS}\right)\left(|\hat{\beta}_{gj}^{LS}| - \frac{\lambda_n}{n}\cdot\frac{|\hat{\beta}_{gj}^{LS}|}{\|\hat{\boldsymbol{\beta}}_g^{LS}\|_2}\right)_+.$$

It should be noted that the L_2-norm of the shrinkage vector for the gth group is λ_n/n, which is a fixed amount and does not relate to the relative importance of each factor. It is expected that such a penalty could be excessive and adversely affect the estimation efficiency and model selection consistency (Wang and Leng, 2008). We will demonstrate this point for an orthogonal design.

Remark 1. One interesting observation from (17) is the interaction of the spike part and the slab part in the posterior inference. The spike part leads to a soft thresholding estimator that can select variables automatically and the thresholds depend on π_0, while the hyperparameter in the slab part, $\tau_{g,n}^2$ (or λ if the gamma hyperprior is assumed) decides the shrinkage factor $B_{g,n}$.

Let $\boldsymbol{\beta}^0, \boldsymbol{\beta}_g^0, \beta_{gj}^0$ denote the true values of $\boldsymbol{\beta}, \boldsymbol{\beta}_g, \beta_{gj}$, respectively. Define the index vector of the true model as $\mathcal{A} = (I(\|\boldsymbol{\beta}_g\|_2 \ne 0), g = 1,2,\ldots,G)$, and the index vector of the model selected by certain thresholding estimator $\hat{\boldsymbol{\beta}}_g$ as $\mathcal{A}_n = (I(\|\hat{\boldsymbol{\beta}}_g\|_2 \ne 0), g = 1,2,\ldots,G)$. Model selection consistency is attained if and only if $\lim_n P(\mathcal{A}_n = \mathcal{A}) = 1$.

Lemma 2. *If* $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$, *then* $\limsup_n P(\mathcal{A}_n^{GL} = \mathcal{A}) < 1$.

Proof. Note that for any g such that $\|\boldsymbol{\beta}_g\|_2 = 0$,

$$P\left(\|\hat{\boldsymbol{\beta}}_g^{GL}\|_2 = 0\right) = P\left(\|\hat{\boldsymbol{\beta}}_g^{LS}\|_2 \leq \frac{\lambda_n}{n}\right) = P\left(\|\sqrt{n}\hat{\boldsymbol{\beta}}_g^{LS}\|_2 \leq \frac{\lambda_n}{\sqrt{n}}\right),$$

where $\sqrt{n}\hat{\boldsymbol{\beta}}_g^{LS} \xrightarrow{d} \boldsymbol{Z}$, $\boldsymbol{Z} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I})$, and $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$. Thus by Fatou's Lemma,

$$\limsup_n P\left(\mathcal{A}_n^{GL} = \mathcal{A}\right) \leq \limsup_n P\left(\|\hat{\boldsymbol{\beta}}_g^{GL}\|_2 = 0\right) \leq P\left(\|\boldsymbol{Z}\|_2 \leq \lambda_0\right) < 1. \qquad \square$$

We can observe from above lemma that in order for the group lasso to consistently select variables, we must have $\lambda_n/\sqrt{n} \to \infty$. But this condition does not give optimal estimation rate, as demonstrated by the following lemma.

Lemma 3. *If* $\lambda_n/\sqrt{n} \to \infty$, *then*

$$\frac{n}{\lambda_n}\left(\hat{\boldsymbol{\beta}}^{GL} - \boldsymbol{\beta}^0\right) \xrightarrow{p} \boldsymbol{C},$$

where $\boldsymbol{C} = (\boldsymbol{\beta}_g^0/\|\boldsymbol{\beta}_g^0\|_2, g = 1, \ldots, G)^T$, *is a vector of constants depending on the true model.*

Proof. For any $g (1 \leq g \leq G)$,

$$\begin{aligned}
&\frac{n}{\lambda_n}\left(\hat{\boldsymbol{\beta}}_g^{GL} - \boldsymbol{\beta}_g^0\right) \\
&= \frac{\sqrt{n}}{\lambda_n}\sqrt{n}\left(\hat{\boldsymbol{\beta}}_g^{LS} - \boldsymbol{\beta}_g^0\right) - \frac{n}{\lambda_n}\left(1 - \frac{\lambda_n}{n\|\hat{\boldsymbol{\beta}}_g^{LS}\|_2}\right) I\left(n\|\hat{\boldsymbol{\beta}}_g^{LS}\|_2 < \lambda_n\right) - \frac{1}{\|\hat{\boldsymbol{\beta}}_g^{LS}\|_2}\hat{\boldsymbol{\beta}}_g^{LS} \\
&\xrightarrow{p} -\frac{1}{\|\boldsymbol{\beta}_g^0\|_2}\boldsymbol{\beta}_g^0
\end{aligned}$$

by noting that $\sqrt{n}(\hat{\boldsymbol{\beta}}_g^{LS} - \boldsymbol{\beta}_g^0) = O_p(1)$, $\frac{\sqrt{n}}{\lambda_n} \to 0$, $I(n\|\hat{\boldsymbol{\beta}}_g^{LS}\|_2 < \lambda_n) \xrightarrow{p} 0$ and applying Slutsky's theorem. $\square$

Thus the convergence rate of the group lasso estimator is n/λ_n, which is slower than $\sqrt{n}$. Adaptive group lasso (Wang and Leng, 2008; Nardi and Rinaldo, 2008) was proposed to overcome this limitation. By using different regularization parameter that depends on the least square estimators for different factors, the adaptive group lasso enjoys oracle property. We will show that the median thresholding estimator also has the oracle property under an orthogonal design.

Theorem 4. *Assume orthogonal design matrix, i.e.,* $\boldsymbol{X}^T\boldsymbol{X} = n\boldsymbol{I}_p$. *Suppose* $\sqrt{n}\tau_{g,n}^2 \to \infty$ *and* $\log(\tau_{g,n}^2)/n \to 0$ *as* $n \to \infty$, *for* $g = 1, \ldots, G$, *then the median thresholding estimator has oracle property, that is, variable selection consistency,*

$$\lim_{n\to\infty} P\left(\mathcal{A}_n^{Med} = \mathcal{A}\right) = 1$$

and asymptotic normality,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{Med}-\boldsymbol{\beta}_{\mathcal{A}}^{0}\right)\xrightarrow{d}\boldsymbol{N}\left(\mathbf{0},\sigma^{2}\boldsymbol{I}\right).$$

Proof. First we observe that $\lim_{n\to\infty}\sqrt{n}B_{g,n}=0$ since $\sqrt{n}\tau_{g,n}^{2}\to\infty$ as $n\to\infty$, $g=1,\ldots,G$.

For g such that $\|\boldsymbol{\beta}_{g}^{0}\|_{2}=0$, since $\sqrt{n}\hat{\boldsymbol{\beta}}_{g}^{LS}=O_{p}(1)$ and $n\tau_{g,n}^{2}\to\infty$, $l_{g,n}\xrightarrow{p}1$ as $n\to\infty$. The probability of correctly classifying this factor is

$$\begin{aligned}P\left(\|\hat{\boldsymbol{\beta}}_{g}^{Med}\|_{2}=0\right)&=P\left(\left(1-B_{g,n}\right)|\hat{\beta}_{gj}^{LS}|\leq\frac{\sigma}{\sqrt{n}}Q_{g,n}\sqrt{1-B_{g,n}},j=1,...,m_{g}\right)\\&=\prod_{j=1}^{m_{g}}P\left(T_{g,n}^{j}\leq1\right)\\&\to1\text{ as }n\to\infty\end{aligned}$$

where $T_{g,n}^{j}\triangleq\sigma\sqrt{1-B_{g,n}}\cdot\sqrt{n}|\hat{\beta}_{gj}^{LS}|/Q_{g,n}\xrightarrow{p}0$ for all $1\leq j\leq m_{g}$ by Slutsky's theorem.

For g such that $\|\boldsymbol{\beta}_{g}^{0}\|_{2}\neq0$, since $\hat{\boldsymbol{\beta}}_{g}^{LS}\xrightarrow{p}\boldsymbol{\beta}_{g}^{0}$ and $\log(\tau_{g,n}^{2})/n\to0$, $l_{g,n}\xrightarrow{p}0$ as $n\to\infty$. The probability of correctly identifying this factor is

$$\begin{aligned}P\left(\|\hat{\boldsymbol{\beta}}_{g}^{Med}\|_{2}\neq0\right)&=P\left(\left(1-B_{g,n}\right)|\hat{\beta}_{gj}^{LS}|>\frac{\sigma}{\sqrt{n}}Q_{g,n}\sqrt{1-B_{g,n}},j=1,...,m_{g}\right)\\&=\prod_{j=1}^{m_{g}}P\left(1/T_{g,n}^{j}<1\right)\\&\to1\text{ as }n\to\infty\end{aligned}$$

where $1/T_{g,n}^{j}\xrightarrow{p}0$ for all $1\leq j\leq m_{g}$ by Slutsky's theorem. Thus we have proved variable selection consistency. For asymptotic normality, we only need to show that $\sqrt{n}(\hat{\beta}_{gj}^{Med}-\hat{\beta}_{gj}^{LS})\xrightarrow{p}0$, and then the result follows from the fact that $\sqrt{n}(\hat{\beta}_{gj}^{LS}-\beta_{gj}^{0})\xrightarrow{d}N(0,\sigma^{2})$. Note that $\sqrt{n}B_{g,n}\to0$, $\hat{\boldsymbol{\beta}}_{g}^{LS}\xrightarrow{p}\boldsymbol{\beta}_{g}^{0}$, $l_{g,n}\to0$ and $\sqrt{n}I(T_{g,n}^{j}\leq1)\xrightarrow{p}0$. Then

$$\begin{aligned}&\left|\sqrt{n}\left(\hat{\beta}_{gj}^{Med}-\hat{\beta}_{gj}^{LS}\right)\right|\\&\quad=\left(\sqrt{n}B_{g,n}|\hat{\beta}_{gj}^{LS}|-\sqrt{1-B_{g,n}}Q_{g,n}\right)I\left(T_{g,n}^{j}>1\right)+\sqrt{n}|\hat{\beta}_{gj}^{LS}|I\left(T_{g,n}^{j}\leq1\right)\\&\quad\xrightarrow{p}0\end{aligned}$$

by Slutsky's theorem. Therefore, we conclude $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{Med}-\boldsymbol{\beta}_{\mathcal{A}}^{0})\xrightarrow{d}\boldsymbol{N}(\mathbf{0},\sigma^{2}I)$. □

2.4 Gibbs Sampler

The full posterior distribution of all the unknown parameters conditional on data is

$$\begin{aligned}
&p(\boldsymbol{\beta}, \boldsymbol{\tau}^2, \sigma^2, \pi_0 | \boldsymbol{Y}, \boldsymbol{X}) \\
&\quad \propto (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{Y} - X\boldsymbol{\beta})^T(\boldsymbol{Y} - X\boldsymbol{\beta})\right\} \\
&\quad \times \prod_{g=1}^{G}\left[(1-\pi_0)(2\pi\sigma^2\tau_g^2)^{-\frac{m_g}{2}} \exp\left\{-\frac{\boldsymbol{\beta}_g^T\boldsymbol{\beta}_g}{2\sigma^2\tau_g^2}\right\} I\left[\boldsymbol{\beta}_g \neq \mathbf{0}\right] + \pi_0\delta_0(\boldsymbol{\beta}_g)\right] \\
&\quad \times \prod_{g=1}^{G}\left(\lambda^2\right)^{\frac{m_g+1}{2}}\left(\tau_g^2\right)^{\frac{m_g+1}{2}-1}\exp\left(-\frac{\lambda^2}{2}\tau_g^2\right) \\
&\quad \times \pi_0^{a-1}\left(1-\pi_0\right)^{b-1} \\
&\quad \times \left(\sigma^2\right)^{-\alpha-1}\exp\left\{-\frac{\gamma}{\sigma^2}\right\}.
\end{aligned}$$

We utilize an efficient block Gibbs sampler (Hobert and Geyer, 1998) to simulate from the posterior distribution above. To estimate the highest posterior probability model, we record the model selected at each simulation and tabulate them to find the model that appears most often. Let $\boldsymbol{\beta}_{(g)}$ denote the $\boldsymbol{\beta}$ vector without the gth group, that is,

$$\boldsymbol{\beta}_{(g)} = \left(\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{g-1}^T, \boldsymbol{\beta}_{g+1}^T, \ldots, \boldsymbol{\beta}_G^T\right)^T.$$

Let $\boldsymbol{X}_{(g)}$ denote the covariate matrix corresponding to $\boldsymbol{\beta}_{(g)}$, that is,

$$\boldsymbol{X}_{(g)} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{g-1}, \boldsymbol{X}_{g+1}, \ldots, \boldsymbol{X}_G),$$

where $\boldsymbol{X}_g$ is the design matrix corresponding to $\boldsymbol{\beta}_g$.

The Gibbs Sampler we used to generate from the posterior distribution is given below

- Let $\boldsymbol{\mu}_g = \boldsymbol{\Sigma}_g\boldsymbol{X}_g^T(\boldsymbol{Y} - \boldsymbol{X}_{(g)}\boldsymbol{\beta}_{(g)}), \boldsymbol{\Sigma}_g = (\boldsymbol{X}_g^T\boldsymbol{X}_g + \frac{1}{\tau_g^2}\boldsymbol{I}_{m_g})^{-1}$, then the conditional posterior distribution of $\boldsymbol{\beta}_g$ is a spike and slab distribution,

$$\boldsymbol{\beta}_g|\text{rest} \sim (1-l_g)\boldsymbol{N}(\boldsymbol{\mu}_g, \sigma^2\boldsymbol{\Sigma}_g) + l_g\delta_0(\boldsymbol{\beta}_g), \quad g = 1, \ldots, G,$$

where

$$\begin{aligned}
l_g &= p(\boldsymbol{\beta}_g = 0|\text{rest}) \\
&= \frac{\pi_0}{\pi_0 + (1-\pi_0)(\tau_g^2)^{-\frac{m_g}{2}}|\boldsymbol{\Sigma}_g|^{\frac{1}{2}}\exp\left\{\frac{1}{2\sigma^2}\|\boldsymbol{\Sigma}_g^{\frac{1}{2}}X_g^T(\boldsymbol{Y} - \boldsymbol{X}_{(g)}\boldsymbol{\beta}_{(g)})\|_2^2\right\}}.
\end{aligned}$$

Remark 2. $\boldsymbol{Y} - \boldsymbol{X}_{(g)}\boldsymbol{\beta}_{(g)}$ is the residual vector when we exclude the gth factor $\boldsymbol{\beta}_g$ in our regression model. Each element of $X_g^T(\boldsymbol{Y} - \boldsymbol{X}_{(g)}\boldsymbol{\beta}_{(g)})$ is proportional to the correlation between the each covariate in the gth group and this residual vector.

- Let $\alpha_g^2 = \frac{1}{\tau_g^2},\ g = 1, 2, \ldots, G$. Then

$$\alpha_g^2|\text{rest} \sim \begin{cases} \text{Inverse Gamma}\left(\text{shape} = \frac{m_g+1}{2}, \text{scale} = \frac{\lambda^2}{2}\right), & \text{if } \boldsymbol{\beta}_g = 0, \\ \text{Inverse Gaussian}\left(\frac{\lambda\sigma}{\|\beta_g\|_2}, \lambda^2\right), & \text{if } \boldsymbol{\beta}_g \neq 0. \end{cases}$$

- $$\sigma^2|rest \sim \text{Inverse Gamma}\Bigg(\frac{n}{2} + \frac{1}{2}\sum_{g=1}^{G} m_g Z_g + \alpha, \\ \frac{1}{2}[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^T \boldsymbol{D}_\tau^{-1}\boldsymbol{\beta}] + \gamma\Bigg)$$

  where $Z_g = \begin{cases} 1, & \text{if } \boldsymbol{\beta}_g \neq \mathbf{0}, \\ 0, & \text{if } \boldsymbol{\beta}_g = \mathbf{0}, \end{cases}$ $\boldsymbol{D}_\tau = \text{diag}\{\tau_1^2, \tau_2^2, \ldots, \tau_G^2\}$.

- $$\pi_0|\text{rest} \sim \text{Beta}\left(a + \sum_{g=1}^{G} Z_g, b + \sum_{g=1}^{G} m_g - \sum_{g=1}^{G} Z_g\right).$$

3 Bi-level Selection

We have introduced BGL-SS for group level variable selection in the last section but it is not always suitable for the problem. In many applications, it may be desirable to select variables at both the group level and the individual level. In a genetic association study (Huang et al., 2012), for example, genetic variations in the same gene form a natural group. But one genetic variation related to the disease does not necessarily mean that all the other variations in the same gene are also associated with the disease. We propose methods for selecting variables simultaneously at both levels in this section.

3.1 Bayesian Sparse Group Lasso (BSGL)

Model Formulation

With a combination of L_1- and L_2-penalty, the sparse group lasso (Simon et al., 2012) has the desirable property of both group-wise sparsity and within group sparsity. Assuming the following independent multivariate priors on each group of regression coefficients in (1),

$$\pi(\boldsymbol{\beta}_g) \propto \exp\left\{-\frac{\lambda_1}{2\sigma^2}\|\boldsymbol{\beta}_g\|_1 - \frac{\lambda_2}{2\sigma^2}\|\boldsymbol{\beta}_g\|_2\right\}, \quad g = 1, 2, \ldots, G, \tag{18}$$

then the sparse group lasso estimator in (3) is equivalent to the MAP solution under this prior.

To find a Bayesian representation of the sparse group lasso where all posterior conditionals are of standard form and thus greatly simplify computation, we follow the approach of Park and Casella (2008) and Kyung et al. (2010), and express the prior as a two level hierarchical structure including independent $\mathbf{0}$ mean Gaussian priors on $\boldsymbol{\beta}_g$'s with parameters $\boldsymbol{\tau}_g, \gamma_g$ and hyperpriors on $\boldsymbol{\tau}_g, \gamma_g$.

To enable shrinkage both at the group level and within a group, we propose the following Bayesian hierarchical model which we refer to as Bayesian sparse group lasso (BSGL).

$$\boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2 \sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n\right), \tag{19}$$

$$\boldsymbol{\beta}_g|\boldsymbol{\tau}_g, \gamma_g, \sigma^2 \sim N\left(\mathbf{0}, \sigma^2 \boldsymbol{V}_g\right), \quad g = 1, \ldots, G, \tag{20}$$

where $\boldsymbol{V}_g = \text{diag}\{(\frac{1}{\tau_{gj}^2} + \frac{1}{\gamma_g^2})^{-1},\ j = 1, 2, \ldots, m_g\}$. Then we place the following multivariate prior on $\boldsymbol{\tau}_g, \gamma_g$

$$\pi\left(\tau_{g1}^2, \ldots, \tau_{gm_g}^2, \gamma_g^2\right) = c_g\left(\lambda_1^2, \lambda_2^2\right) \prod_{j=1}^{m_g} \left[\left(\tau_{gj}^2\right)^{-\frac{1}{2}} \left(\frac{1}{\tau_{gj}^2} + \frac{1}{\gamma_g^2}\right)^{-\frac{1}{2}}\right] \left(\gamma_g^2\right)^{-\frac{1}{2}} \times \exp\left\{-\frac{\lambda_1^2}{2} \sum_{j=1}^{m_g} \tau_{gj}^2 - \frac{\lambda_2^2}{2} \gamma_g^2\right\}. \tag{21}$$

Although this prior has a complicated form and an unknown normalizing constant depending on λ_1 and λ_2, all the resulting full conditionals in the Gibbs sampler are standard distributions and thus are easy and fast to sample from. The propriety of the prior given in (21) is proved in the appendix.

With above hierarchical priors, the marginal prior on $\boldsymbol{\beta}_g$ is

$$\pi\left(\boldsymbol{\beta}_g|\sigma^2\right) \propto \exp\left\{-\frac{\lambda_1}{\sigma}\|\boldsymbol{\beta}_g\|_1 - \frac{\lambda_2}{\sigma}\|\boldsymbol{\beta}_g\|_2\right\},$$

which is a prior of the form (18) with our two level hierarchical prior specification.

Hyperparameter Specification

The specification of hyperparameters λ_1^2, λ_2^2 is very important because it expresses our prior belief of sparsity and the amount of shrinkage. We place a hyper-prior on them instead of imposing fixed values. Define $C(\lambda_1^2, \lambda_2^2) = \prod_{g=1}^{G} c_g(\lambda_1^2, \lambda_2^2)$. The following prior is assigned to λ_1^2 and λ_2^2,

$$p\left(\lambda_1^2, \lambda_2^2\right) \propto C^{-1}\left(\lambda_1^2, \lambda_2^2\right) \left(\lambda_1^2\right)^p \left(\lambda_2^2\right)^{G/2} \exp\left\{-d_1\lambda_1^2 - d_2\lambda_2^2\right\},$$

where $d_1 > 0, d_2 > 0$. It is easy to show that this prior is proper. To make it a moderately diffuse prior, we specify small values for d_1 and d_2, $d_1 = d_2 = 10^{-1}$.

3.2 Bayesian Sparse Group Selection with Spike and Slab Prior (BSGS-SS)

Although the Bayesian sparse group lasso has shrinkage effects at both the group level and also within a group, it does not produce sparse model since the posterior mean/median estimators are never exact 0. To achieve sparsity at both levels for variable selection purpose, and to improve out-of-sample prediction performance, we propose the Bayesian Sparse Group Selection with Spike and Slab prior (BSGS-SS), which utilizes spike and slab type priors for both group variable selection and individual variable selection. The difficulty of this problem lies in how to introduce both types of sparsity with spike and slab priors.

Model Specification

We reparametrize the coefficients vectors to tackle the two kinds of sparsity separately:

$$\boldsymbol{\beta}_g = \boldsymbol{V}_g^{\frac{1}{2}} \boldsymbol{b}_g, \text{ where } \boldsymbol{V}_g^{\frac{1}{2}} = \operatorname{diag}\left\{\tau_{g1}, \ldots, \tau_{gm_g}\right\}, \ \tau_{gj} \geq 0, \ g = 1, \ldots, G; \ j = 1, \ldots, m_g, \tag{22}$$

where $\boldsymbol{b}_g$, when nonzero, has a $\mathbf{0}$ mean multivariate normal distribution with identity matrix as its covariance matrix. Thus the diagonal elements of $\boldsymbol{V}_g^{\frac{1}{2}}$ control the magnitude of elements of $\boldsymbol{\beta}_g$. To select variables at the group level, we assume the following multivariate spike and slab prior for each $\boldsymbol{b}_g$:

$$\boldsymbol{b}_g \overset{ind}{\sim} (1 - \pi_0) \boldsymbol{N}_{m_g}\left(\mathbf{0}, \boldsymbol{I}_{m_g}\right) + \pi_0 \delta_0\left(\boldsymbol{b}_g\right), \quad g = 1, \ldots, G. \tag{23}$$

Note that when $\tau_{gj} = 0$, β_{gj} is essentially dropped out of the model even when $b_{gj} \neq 0$. So in order to choose variables within each relevant group, we assume the following spike and slab prior for each τ_{gj}:

$$\tau_{gj} \overset{ind}{\sim} (1 - \pi_1) N^+\left(0, s^2\right) + \pi_1 \delta_0\left(\tau_{gj}\right), \quad g = 1, \ldots, G; \ j = 1, \ldots, m_g, \tag{24}$$

where $N^+(0, s^2)$ denotes a normal $N(0, s^2)$ distribution truncated below at 0. Note that this truncated normal distribution has mean $\sqrt{\frac{2}{\pi}} s$ and variance s^2.

Remark 3. If $m_g = 1$, $\beta_g = \tau_g b_g$ is a scalar, and still has a spike and slab distribution. The prior probability of $\beta_g = 0$ is $1 - (1 - \pi_0)(1 - \pi_1)$, which is larger than both π_0 and π_1, but smaller than $\pi_0 + \pi_1$. As a comparison, the sparse group lasso penalty for the gth group of coefficients becomes $(\lambda_1 + \lambda_2)\|\beta_g\|_1$ when $m_g = 1$. Thus the penalty parameter is the sum of the individual level penalty parameter λ_1, and the group level penalty parameter λ_2.

Remark 4. Alternatively, we could enforce both types of sparsity by generalizing the binary masking model of Kuo and Mallick (1998). We can reparameterize the regression coefficients as $\beta_{gj} = \gamma_g^{(1)} \gamma_{gj}^{(2)} b_{gj}$, where $\gamma_g^{(1)}$ is a binary indicator of whether the gth

group of coefficients are all 0, and $\gamma_{gj}^{(2)}$ indicates whether $\beta_{gj} = 0$. The following priors are assumed:

$$\begin{aligned}
\gamma_g^{(1)} &\sim \text{Bernoulli}\,(\pi_0)\,, \quad g = 1, \ldots, G, \\
\gamma_{gj}^{(2)} &\sim \text{Bernoulli}\,(\pi_1)\,, \quad g = 1, \ldots, G;\ j = 1, \ldots, m_g, \\
b_{gj} &\sim N\left(0, s^2\right), \quad g = 1, \ldots, G;\ j = 1, \ldots, m_g.
\end{aligned}$$

We expect that the above alternative formulation to have comparable performance with the BSGS-SS model that we proposed. Stingo et al. (2011) also uses two sets of binary indicators for group and individual level selection for a more specific group selection problem, in which groups may be overlapping and certain dependence structure among variables exists.

Instead of specifying fixed values for hyperparameters, typical non-informative priors are used. We assume an inverse gamma prior for the error variance σ^2, where shape and scale parameters are chosen to be relatively small:

$$\sigma^2 \sim \text{Inverse Gamma}\,(\alpha, \gamma)\,, \quad \alpha = 0.1,\ \gamma = 0.1. \tag{25}$$

To decide the values of hyperparameters π_0, π_1, we assume conjugate beta hyper-priors:

$$\pi_0 \sim \text{Beta}\,(a_1, a_2)\,, \quad \pi_1 \sim \text{Beta}\,(c_1, c_2)\,. \tag{26}$$

For s^2, we place a conjugate inverse gamma prior on it,

$$s^2 \sim \text{Inverse Gamma}\,(1, t)\,,$$

and estimate t with the Monte Carlo EM algorithm (Casella, 2001; Park and Casella, 2008). For the kth EM update,

$$t^{(k)} = \frac{1}{E_{t^{(k-1)}}\left[\frac{1}{s^2} \mid \boldsymbol{Y}\right]},$$

where the posterior expectation of $\frac{1}{s^2}$ is estimated from the Gibbs samples based on $t^{(k-1)}$.

Therefore, with the above model specification, the joint posterior of $\boldsymbol{b}, \tau^2, \sigma^2, \pi_0, \pi_1$ conditional on observed data is

$$\begin{aligned}
&p\left(\boldsymbol{b}, \boldsymbol{\tau}^2, \sigma^2, \pi_0, \pi_1, s^2 \mid \boldsymbol{Y}, \boldsymbol{X}\right) \\
&\quad \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left\|\boldsymbol{Y} - \sum_{g=1}^{G} \boldsymbol{X}_g \boldsymbol{V}_g^{\frac{1}{2}} \boldsymbol{b}_g\right\|_2^2\right\} \\
&\quad \times \prod_{g=1}^{G}\left[(1-\pi_0)\,(2\pi)^{-\frac{m_g}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{b}_g^T \boldsymbol{b}_g\right\} I\left[\boldsymbol{b}_g \neq 0\right] + \pi_0 \delta_0\left(\boldsymbol{b}_g\right)\right] \\
&\quad \times \prod_{g=1}^{G}\prod_{j=1}^{m_g}\left[(1-\pi_1) \cdot 2\left(2\pi s^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{\tau_{gj}^2}{2s^2}\right\} I\left[\tau_{gj} > 0\right] + \pi_1 \delta_0\left(\tau_{gj}\right)\right]
\end{aligned}$$

$$\times \left(\sigma^2\right)^{-\alpha-1} \exp\left\{-\frac{\gamma}{\sigma^2}\right\}$$
$$\times \pi_0^{a_1-1} \left(1-\pi_0\right)^{a_1-1}$$
$$\times \pi_1^{c_1-1} \left(1-\pi_1\right)^{c_1-1}$$
$$\times t \left(s^2\right)^{-2} \exp\left\{-\frac{t}{s^2}\right\}.$$

Gibbs Sampler

Similar to Subsection 2.4, we define the coefficients vector without the jth element in the gth group as

$$\boldsymbol{\beta}_{(gj)} = \left(\beta_{11}, \ldots, \beta_{1m_1}, \ldots, \beta_{g1}, \ldots, \beta_{g,j-1}, \beta_{g,j+1}, \ldots, \beta_{gm_g}, \ldots, \beta_{Gm_G}\right)^T,$$

and the covariates matrix corresponding to $\boldsymbol{\beta}_{(gj)}$ as

$$\boldsymbol{X}_{(gj)} = \left(\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1m_1}, \ldots, \boldsymbol{x}_{g1}, \ldots, \boldsymbol{x}_{g,j-1}, \boldsymbol{x}_{g,j+1}, \ldots, \boldsymbol{x}_{gm_g}, \ldots, \boldsymbol{x}_{Gm_G}\right).$$

- The posterior distribution of $\boldsymbol{b}_g$ conditioning on everything else is still a multivariate spike and slab distribution,

$$\boldsymbol{b}_g \mid \text{rest} \sim l_g \delta_0\left(\boldsymbol{b}_g\right) + \left(1-l_g\right) \boldsymbol{N}_{m_g}\left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right),$$

where l_g, the posterior probability of $\boldsymbol{b}_g$ equal to $\mathbf{0}$ given the remaining parameters, is

$$\begin{aligned} l_g &= P\left(\boldsymbol{b}_g = \mathbf{0} | \text{rest}\right) \\ &= \frac{\pi_0}{\pi_0 + \left(1-\pi_0\right) \mid \boldsymbol{\Sigma} \mid^{\frac{1}{2}} \exp\left\{\frac{1}{2\sigma^4} \parallel \boldsymbol{\Sigma}_g^{\frac{1}{2}} \boldsymbol{V}_g^{\frac{1}{2}} \boldsymbol{X}_g^T \left(\boldsymbol{Y} - \boldsymbol{X}_{(g)} V_{(g)}^{\frac{1}{2}} \boldsymbol{b}_{(g)}^{\frac{1}{2}}\right) \parallel_2^2\right\}}, \end{aligned}$$

$\boldsymbol{\mu}_g = \frac{1}{\sigma^2} \boldsymbol{\Sigma}_g^{\frac{1}{2}} \boldsymbol{V}_g^{\frac{1}{2}} \boldsymbol{X}_g^T (\boldsymbol{Y} - \boldsymbol{X}_{(g)} V_{(g)}^{\frac{1}{2}} \boldsymbol{b}_{(g)}^{\frac{1}{2}})$, and $\boldsymbol{\Sigma}_g = (\boldsymbol{I}_{m_g} + \frac{1}{\sigma^2} \boldsymbol{V}_g^{\frac{1}{2}} \boldsymbol{X}_g^T \boldsymbol{X}_g \boldsymbol{V}_g^{\frac{1}{2}})^{-1}$.

- The conditional posterior of τ_{gj} is a spike and slab distribution, with the slab a positive part normal distribution:

$$\tau_{gj} \mid \text{rest} \sim q_{gj} \delta_0\left(\tau_{gj}\right) + \left(1-q_{gj}\right) \boldsymbol{N}^+\left(u_{gj}, v_{gj}^2\right), \; g = 1, 2, \ldots, G; \; j = 1, 2, \ldots, m_G,$$

where $u_{gj} = \frac{1}{\sigma^2} v_{gj}^2 (\boldsymbol{Y} - \boldsymbol{X}_{(gj)} \boldsymbol{\beta}_{(gj)})^T \boldsymbol{X}_{gj} b_{gj}$, $v_{gj}^2 = (\frac{1}{s^2} + \frac{1}{\sigma^2} \boldsymbol{X}_{gj}^T \boldsymbol{X}_{gj} b_{gj}^2)^{-1}$ and

$$q_{gj} = p\left(\tau_{gj} = 0 \mid \text{rest}\right) = \frac{\pi_1}{\pi_1 + 2\left(1-\pi_1\right)\left(s^2\right)^{-\frac{1}{2}} \left(v_{gj}^2\right)^{\frac{1}{2}} \exp\left\{\frac{u_{gj}^2}{2v_{gj}^2}\right\} \left[\Phi\left(\frac{u_{gj}}{v_{gj}}\right)\right]}.$$

- $$\sigma^2 | \text{rest} \sim \text{Inverse Gamma}\left(\frac{n}{2} + \alpha, \frac{1}{2} \parallel \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \parallel_2^2 + \gamma\right).$$

- With conjugate Beta priors, the posteriors of π_0 and π_1 conditional on everything else continue to be Beta distributions:

$$\pi_0 \mid \text{rest} \sim \text{Beta}\left(\#\left(\boldsymbol{b}_g = 0\right) + a_1, \#\left(\boldsymbol{b}_g \neq 0\right) + a_2\right),$$
$$\pi_1 \mid \text{rest} \sim \text{Beta}\left(\#\left(\tau_{gj} = 0\right) + c_1, \#\left(\tau_{gj} \neq 0\right) + c_2\right).$$

- With conjugate inverse gamma prior, the conditional posterior of s^2 is still an inverse gamma distribution:

$$s^2 \mid \text{rest} \sim \text{Inverse Gamma}\left(1 + \frac{1}{2}\#\left(\tau_{gj} = 0\right), t + \frac{1}{2}\sum_{g,j}\tau_{gj}^2\right).$$

4 Simulation

We simulate data from the following true model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \text{ where } \epsilon_i \overset{iid}{\sim} N(0, \sigma^2),\ i = 1, 2, \ldots, n.$$

For the following examples, we compare the variable selection accuracy and prediction performance of BGL-SS, BSGL, BSGS-SS with 4 other models: linear regression, the Group Lasso (GL), the Sparse Group Lasso (SGL) and the Bayesian Group Lasso (BGL), when applicable. Five examples are considered in our simulations. The third one is from the original lasso paper (Tibshirani, 1996).

- Example 1. We simulate a data set with 100 observations and 20 covariates, which are divided into 4 groups with 5 covariates each. We randomly sample 60 observations to train the model and use the remaining 40 to compare the prediction performance of proposed model with other lasso variations. Let

$$\boldsymbol{\beta} = \left((0.3, -1, 0, 0.5, 0.01), \mathbf{0}, (0.8, 0.8, 0.8, 0.8, 0.8), \mathbf{0}\right),$$

  where $\mathbf{0}$ is the 0 vector of length 5. The pairwise correlation between covariates x_i and x_j is 0.5 for $i \neq j$. We specify $\sigma = 3$.

- Example 2. This example is a large p small n problem with $n = 60$ and $p = 80$. 40 observations are randomly sampled to train the model and the remaining 20 are used to compare the prediction performance. 80 predictors are grouped into 16 groups of 5 covariates each. We define the jth predictor in group g as $X_{gj} = z_g + z_{gj}$, where z_g and z_{gj} are independent standard normal variates, $g = 1, \ldots, 16;\ j = 1, 2, \ldots, 5$. Thus predictors within a group are correlated with pairwise correlation $\frac{1}{2}$ while the predictors in different groups are independent. Let

$$\boldsymbol{\beta} = \left((1, 2, 3, 4, 5), \mathbf{0}, (0.1, 0.2, 0.3, 0.4, 0.5), \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}\right)$$

  where $\mathbf{0}$ is the 0 vector of length 5. We use $\sigma = 2$.

- Example 3. In this example, we simulate a data set with $n = 100$ and $p = 40$. 60 observations are used to train the model and the remaining 40 are used for testing the predictions. Let $\boldsymbol{\beta} = (\mathbf{0}, \mathbf{2}, \mathbf{0}, \mathbf{2})$, where $\mathbf{0}$ and $\mathbf{2}$ are both of length 10, with all elements 0 or 2, respectively. We simulate predictors in the same way as in Example 2 except for necessary dimension changes. The error standard deviation σ is 2.

- Example 4. This example is the same as Example 3 except the true coefficients

$$\boldsymbol{\beta} = (\mathbf{0}, (2, 2, 2, 2, 2, 0, 0, 0, 0, 0), \mathbf{0}, (2, 2, 2, 2, 2, 0, 0, 0, 0, 0)),$$

  where $\mathbf{0}$ is a 0 vector of length 10. So this example, like Example 1, has sparsity at the group level and also sparsity within nonzero groups.

- Example 5. This example is taken from Yuan and Lin (2006). $Z_1, Z_2, \ldots, Z_{20}$ and W were independently generated from the standard normal distribution, and we define $X_i = \frac{(Z_i + W)}{\sqrt{2}}$. The first 10 covariates are each expanded to a third order polynomial thus we have 10 factors consisting of third order polynomial terms. The last 10 covariates are each trichotomized as $0, 1, 2$ if it is smaller than $\phi^{-1}(1/3)$, larger than $\phi^{-1}(2/3)$, or between them. The simulation model is

$$Y = \left(X_3 + X_3^2 + X_3^3\right) + \left(\frac{2}{3}X_6 - X_6^2 + \frac{1}{3}X_6^3\right) + 2I\left(X_{11} = 0\right) + I\left(X_{11} = 1\right) + \epsilon$$

  where $\epsilon \sim N(0, 2^2)$. We simulate 200 samples and use 100 for training and the rest 100 for testing. We have 20 factors with 50 covariates in total.

SPArse Modeling Software (SPAM) is the most stable program we have found for fitting group lasso and sparse group lasso (see Mairal et al., 2010; Jenatton et al., 2011), and we use 5-fold cross-validation to choose optimal λs. For BGL-SS, we have conjugate prior on π_0, so we only need to specify suitable hyperparameters, which we choose $a = 1, b = 1$. For BSGS-SS, we have beta priors on both π_0 and π_1, and we set $a_1 = a_2 = c_1 = c_2 = 1$. For Bayesian models, we generate from the full posterior distribution with a Gibbs Sampler running 10000 iterations in which the first 5000 are burn-ins. Posterior mean and posterior median are both used as our Bayes estimators and we will compare their variable selection and prediction performance. To summarize the prediction errors, we calculate the median mean squared error in 50 simulations.

In Table 1, we summarize the model selection accuracy of different methods. For both BGL-SS and BSGS-SS, the median thresholding model (MTM) and the highest posterior probability model (HPPM) are compared by true and false positive rate. We also list the group lasso and sparse group lasso results for comparison. Median thresholding model, which is more parsimonious, outperforms all other methods including the corresponding highest posterior probability model. The group lasso and the sparse group lasso with penalty parameters chosen by cross validation tend to select much more variables than our spike and slab methods. Leng et al. (2004) showed that when the tuning parameter is selected by minimizing the prediction error, the lasso procedure is inconsistent in

	BGL-SS		BSGS-SS		GL	SGL
	MTM	HPPM	MTM	HPPM		
Example 1						
TPR	0.96	0.98	0.79	0.89	0.97	0.90
FPR	0.23	0.48	0.09	0.19	0.65	0.53
Example 2						
TPR	0.90	0.91	0.82	0.92	0.98	0.87
FPR	0.06	0.12	0.02	0.02	0.39	0.16
Example 3						
TPR	1.00	1.00	1.00	1.00	1.00	1.00
FPR	0.00	0.00	0.02	0.03	0.44	0.26
Example 4						
TPR	1.00	1.00	1.00	1.00	1.00	1.00
FPR	0.34	0.34	0.22	0.34	0.79	0.32
Example 5						
TPR	0.97	0.99	0.91	0.94	0.99	0.94
FPR	0.14	0.54	0.02	0.02	0.40	0.30

Table 1: Mean True/False Positive Rate for six methods in five simulation examples, based on 50 simulations.

variable selection in general. It is suspected (Wang and Leng, 2008) that the group lasso may suffer the same variable selection inconsistency which may explain why the group lasso and the sparse group lasso tends to select more variables and have higher false positive rate in our simulation. On the other hand, model selected by median thresholding has very low false positive rate and even outperforms the gold standard of Bayesian variable selection – the highest posterior probability model.

Table 2 summarizes the median mean squared prediction error for all 5 simulated examples using 9 methods to fit the simulated data, based on 50 replications. The bootstrapped standard errors of the medians are given in the parentheses. A couple of observations can be made from Table 2:

- BGL-SS is comparable with the group lasso in prediction except in Example 2, and BSGS-SS outperforms the sparse group lasso in all examples;
- Posterior mean estimator and posterior median estimator have very close prediction error;
- BGL and BSGL does not predict as well as their frequentist counterpart, GL and SGL;
- When there is no obvious sparsity within relevant groups, BGL-SS usually performs favorably or sometimes better than BSGS-SS; but when there is significant sparsity within relevant groups (Example 4), BSGS-SS is very good at identifying within group sparsity and thus further improves the prediction performance from BGL-SS;

	Example 1	Example 2	Example 3	Example 4	Example 5
BGL-SS with mean	9.69(0.35)	6.79(0.39)	6.45(0.29)	6.41(0.34)	5.24(0.17)
BGL-SS with median	9.76(0.40)	6.60(0.43)	6.46(0.25)	6.40(0.32)	5.08(0.18)
BSGS-SS with mean	10.07(0.38)	5.51(0.21)	6.83(0.42)	5.37(0.15)	4.83(0.16)
BSGS-SS with median	10.37(0.34)	5.59(0.32)	6.51(0.38)	5.38(0.12)	4.92(0.15)
Group Lasso	9.82(0.51)	5.99(0.33)	5.91(0.38)	6.98(0.46)	5.30(0.16)
Sparse Group Lasso	10.48(0.55)	5.75(0.45)	6.88(0.34)	5.90(0.28)	5.22(0.23)
Bayesian Group Lasso	10.53(0.34)	8.24(0.51)	7.89(0.24)	7.48(0.41)	6.46(0.23)
Bayesian Sparse Group lasso	10.08(0.47)	10.55(0.56)	10.21(0.37)	8.65(0.41)	6.03(0.16)
Linear Regression	11.19(0.42)	–	12.71(0.96)	12.68(1.03)	8.71(0.54)

Table 2: Median mean squared error for nine methods in five simulation examples, based on 50 replications.

	Fix π_0			Hyperprior		
	0.20	0.50	0.80	$a = b = 0.50$	$a = b = 1.00$	$a = b = 1.50$
MTM	0.10	0.05	0.10	0.15	0.10	0.10
HPPM	0.30	0.05	0.05	0.55	0.55	0.30
GL	0.55	0.55	0.55	0.55	0.55	0.55

Table 3: Sensitivity analysis for BGL-SS using Example 1.

- The fact that BGL-SS does not predict well in Example 2 suggests that a flat prior with mean $\frac{1}{2}$ on π_0 does not work well for high-dimensional problems in which most groups of predictors are 0. We note that it still works much better than the group lasso in terms of variable selection even with this flat prior.

Now we demonstrate the sensitivity of BGL-SS for model selection to the specification of π_0. We fix π_0 at 0.2, 0.5, 0.8 and assume Beta$(0.5, 0.5)$, Beta$(1, 1)$, Beta$(1.5, 1.5)$ priors. Table 3 shows that the misclassification error, the percentage of misclassified variables, of the median thresholding model and the highest probability model with different specification of π_0. For comparison we append the result of the group lasso, with penalty parameter chosen by cross-validation, in the last row. For all choices of π_0, the median thresholding model is very stable and misclassifies at most three variables, while the highest probability model is very sensitive to the choice of π_0. We also note that although the misclassification error of the group lasso is much higher, its prediction error is comparable to the BGL-SS in this example as we have seen in Table 2.

Posterior mean and median estimators of our spike and slab models are compared in Table 4. Two variations of Example 1, with $\sigma = 1$ or $\sigma = 3$, respectively, are both fitted by BGL-SS and BSGS-SS model. For both cases, the posterior median estimators both produce 0 estimates and correctly identify the two most important factors. When the signal-to-noise ratio is high, the posterior mean estimates shrink coefficients of redundant variables to very small values. But when there is too much noise, posterior mean does not have enough shrinkage effects to help us with variable selection. Regarding within group sparsity, β_3 was shrunk to 0 by BSGS-SS at the cost of shrinking β_5, which has a very small true value, 0.01.

		$\sigma = 3$				$\sigma = 1$			
		BGL-SS		BSGS-SS		BGL-SS		BSGS-SS	
	True	Median	Mean	Median	Mean	Median	Mean	Median	Mean
β_1	0.3	0.09	0.127	0	0.115	0.291	0.289	0.288	0.282
β_2	−1	−0.611	−0.633	−0.656	−0.666	−1.103	−1.056	−1.274	−1.27
β_3	0	−0.056	−0.086	0	−0.03	−0.065	−0.065	0	−0.022
β_4	0.5	0.619	0.637	0.861	0.812	0.695	0.675	0.789	0.788
β_5	0.01	0	0.011	0	0.014	−0.01	−0.008	0	0.003
β_6	0	0	0.158	0	0.111	0	0.006	0	0.011
β_7	0	0	−0.144	0	−0.071	0	−0.005	0	−0.008
β_8	0	0	0.11	0	0.067	0	0.004	0	0.007
β_9	0	0	−0.183	0	−0.071	0	−0.007	0	−0.014
β_{10}	0	0	0.165	0	0.105	0	0.006	0	0.014
β_{11}	0.8	1.534	1.522	1.555	1.555	1.237	1.232	1.23	1.231
β_{12}	0.8	0.271	0.279	0.053	0.187	0.696	0.693	0.689	0.685
β_{13}	0.8	0.877	0.876	0.728	0.709	0.948	0.952	0.906	0.905
β_{14}	0.8	0.73	0.737	0.66	0.666	0.956	0.942	1.055	1.053
β_{15}	0.8	0.744	0.741	0.527	0.532	0.928	0.926	0.919	0.917
β_{16}	0	0	−0.128	0	−0.059	0	−0.002	0	−0.003
β_{17}	0	0	0.111	0	0.078	0	0.002	0	0.006
β_{18}	0	0	0.177	0	0.131	0	0.003	0	0.006
β_{19}	0	0	−0.023	0	−0.003	0	−0.001	0	−0.003
β_{20}	0	0	−0.056	0	−0.015	0	−0.002	0	−0.003

Table 4: Posterior mean and posterior median estimators under spike and slab models using Example 1 with two different error variances.

5 Discussion

The primary goal of the group lasso is to both select groups of variables and estimate corresponding coefficients. Previous Bayesian approaches via multivariate scale mixture of normals do have shrinkage effects at the group level but do not yield sparse estimators.

Spike and slab type priors facilitate variable selection by putting a point mass at 0, or in the case of group variable selection, a multivariate point mass at $\mathbf{0}_{m\times 1}$ for an m-dimensional coefficients group. Since the posterior mean estimator still does not produce sparse estimators, two variable selection criterion were proposed. Highest posterior probability model (Geweke, 1994; Kuo and Mallick, 1998; George and McCulloch, 1997) is a very popular one since via Gibbs sampling simulations we could easily obtain the model and an estimate of its corresponding posterior probability. Alternatively, one can use FDR based variable selection which selects variables with marginal inclusion probability larger than certain threshold and we could choose the threshold to control the overall average Bayesian FDR rate (Bonato et al., 2011; Zhang et al., 2014). Median probability model is advocated by Barbieri and Berger (2004) due to its optimal prediction performance. We note that this is the special case of FDR based methods with thresholds set to $\frac{1}{2}$. Our median thresholding model is more parsimonious than the median probability model because the median of a variable with a spike and slab

distribution is 0 if and only if the probability for it to be either larger or smaller than 0 are both less than $\frac{1}{2}$.

Posterior median estimator is distinctive in the Bayesian methods since it can both select and estimate automatically like the lasso estimator. We demonstrate in this paper that it can achieve superior variable selection accuracy and good prediction performance at the same time. It tends to select fewer variables than group lasso methods but achieves similar or sometimes better prediction error. Compared to the highest probability model, the median thresholding model is at least as good as and sometimes better than it in terms of true and false positive rate.

Appendix A: Propriety of (21)

Prior (21) is proper since

$$\begin{aligned}
&\prod_{j=1}^{m_g}\left[\left(\tau_{gj}^2\right)^{-\frac{1}{2}}\left(\frac{1}{\tau_{gj}^2}+\frac{1}{\gamma_g^2}\right)^{-\frac{1}{2}}\right]\left(\gamma_g^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{\lambda_1^2}{2}\sum_{j=1}^{m_g}\tau_{gj}^2-\frac{\lambda_2^2}{2}\gamma_g^2\right\}\\
&=\prod_{j=1}^{m_g}\left[\left(1+\frac{\tau_{gj}^2}{\gamma_g^2}\right)^{-\frac{1}{2}}\right]\left(\gamma_g^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{\lambda_1^2}{2}\sum_{j=1}^{m_g}\tau_{gj}^2-\frac{\lambda_2^2}{2}\gamma_g^2\right\}\\
&\le\left(\gamma_g^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{\lambda_2^2}{2}\gamma_g^2\right\}\exp\left\{-\frac{\lambda_1^2}{2}\sum_{j=1}^{m_g}\tau_{gj}^2\right\}.
\end{aligned}$$

Appendix B: Marginal Prior for The Bayesian Sparse Group Lasso

With (20)(21), the marginal prior on $\boldsymbol{\beta}_g$ is:

$$\begin{aligned}
\pi\left(\boldsymbol{\beta}_g|\sigma^2\right)\propto&\int_0^\infty\cdots\int_0^\infty\prod_{j=1}^{m_g}\left(\frac{1}{\tau_{gj}^2}+\frac{1}{\gamma_g^2}\right)^{\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{m_g}\left(\frac{1}{\gamma_g^2}+\frac{1}{\tau_{gj}^2}\right)\beta_{gj}^2\right\}\\
&\times\left(\prod_{j=1}^{m_g}\left(\frac{1}{\tau_{gj}^2}+\frac{1}{\gamma_g^2}\right)^{-\frac{1}{2}}\left(\tau_{gj}^2\right)^{-\frac{1}{2}}\right)\left(\gamma_g^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{\lambda_1^2}{2}\sum_{j=1}^{m_g}\tau_{gj}^2-\frac{\lambda_2^2}{2}\gamma_g^2\right\}\\
&\times\left(\prod_{j=1}^{m_g}d\tau_{gj}^2\right)d\gamma_g^2\\
\propto&\prod_{j=1}^{m_g}\int_0^\infty\left(\tau_{gj}^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\frac{\beta_{gj}^2}{\sigma^2}\frac{1}{\tau_{gj}^2}+\lambda_1^2\tau_{gj}^2\right)\right\}d\tau_{gj}^2\\
&\times\int_0^\infty\left(\gamma_g^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\frac{\|\boldsymbol{\beta}_g\|_2^2}{\sigma^2}\frac{1}{\gamma_g^2}+\lambda_2^2\gamma_g^2\right)\right\}d\gamma_g^2
\end{aligned}$$

$$\propto \exp\left\{-\frac{\lambda_1}{\sigma}\|\boldsymbol{\beta}_g\|_1 - \frac{\lambda_2}{\sigma}\|\boldsymbol{\beta}_g\|_2\right\}.$$

Appendix C: Gibbs Sampler for BSGL

The joint posterior probability density function of $\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}, \sigma^2$ given $\boldsymbol{Y}, \boldsymbol{X}$ is

$$\begin{aligned}
&\pi\left(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}\right) \\
&\quad\propto \left(\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)\right\} \\
&\quad\times \prod_{g=1}^{G}\left[\left(\sigma^2\right)^{-\frac{m_g}{2}} \prod_{j=1}^{m_g}\left(\frac{1}{\tau_{gj}^2}+\frac{1}{\gamma_g^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{m_g}\left(\frac{1}{\gamma_g^2}+\frac{1}{\tau_{gj}^2}\right)\beta_{gj}^2\right\}\right] \\
&\quad\times \prod_{g=1}^{G}\left[\left(\prod_{j=1}^{m_g}\left(\frac{1}{\tau_{gj}^2}+\frac{1}{\gamma_g^2}\right)^{-\frac{1}{2}}(\tau_{gj}^2)^{-\frac{1}{2}}\right)(\gamma_g^2)^{-\frac{1}{2}} \exp\left\{-\frac{\lambda_1^2}{2}\sum_{j=1}^{m_g}\tau_{gj}^2 - \frac{\lambda_2^2}{2}\gamma_g^2\right\}\right]\frac{1}{\sigma^2} \\
&\quad\propto \left(\sigma^2\right)^{-\frac{n+m}{2}-1} \prod_{g=1}^{G}\left[\prod_{j=1}^{m_g}(\tau_{gj})^{-\frac{1}{2}}(\gamma_g^2)^{-\frac{1}{2}}\right] \exp\left\{-\frac{\lambda_1^2}{2}\sum_{g=1}^{G}\sum_{j=1}^{m_g}\tau_{gj}^2 - \frac{\lambda_2^2}{2}\sum_{g=1}^{G}\gamma_g^2\right\} \\
&\quad\times \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right) - \frac{1}{2\sigma^2}\sum_{g=1}^{G}\sum_{j=1}^{m_g}\left(\frac{1}{\gamma_g^2}+\frac{1}{\tau_{gj}^2}\right)\beta_{gj}^2\right\}.
\end{aligned}$$

Then we can generate from the posterior distribution using the following full conditional posteriors,

$$\begin{aligned}
\sigma^2|\text{rest} &\sim \text{Inverse Gamma}\left(\frac{m+n}{2}, \frac{1}{2}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right) + \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{V}^{-1}\boldsymbol{\beta}\right), \\
\gamma_g^2|\text{rest} &\overset{ind}{\sim} \text{Inverse Gaussian}\left(\frac{\sigma\lambda_2}{\|\boldsymbol{\beta}_g\|_2^2}, \lambda_2^2\right), \quad g = 1, \ldots, G, \\
\tau_{gj}^2|\text{rest} &\overset{ind}{\sim} \text{Inverse Gaussian}\left(\frac{\sigma\lambda_1}{|\beta_{gj}|}, \lambda_1^2\right), \quad g = 1, \ldots, G;\ j = 1, \ldots, m_g, \\
\boldsymbol{\beta}|\text{rest} &\sim N\left((\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{V}^{-1})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{V}^{-1})^{-1}\right), \\
\lambda_1^2|\text{rest} &\sim \text{Gamma}\left(p+1, \frac{\|\boldsymbol{\tau}\|_2^2}{2} + d_1\right), \\
\lambda_2^2|\text{rest} &\sim \text{Gamma}\left(\frac{G}{2}+1, \frac{\|\boldsymbol{\gamma}\|_2^2}{2} + d_2\right)
\end{aligned}$$

where

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_1 & 0 & \ldots & 0 \\ 0 & \boldsymbol{V}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \boldsymbol{V}_G \end{pmatrix}.$$

References

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998). "Wavelet thresholding via a Bayesian approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4): 725–749. MR1649547. doi: http://dx.doi.org/10.1111/1467-9868.00151. 915

Barbieri, M. M. and Berger, J. O. (2004). "Optimal predictive model selection." *The Annals of Statistics*, 32(3): 870–897. MR2065192. doi: http://dx.doi.org/10.1214/009053604000000238. 930

Bonato, V., Baladandayuthapani, V., Broom, B. M., Sulman, E. P., Aldape, K. D., and Do, K.-A. (2011). "Bayesian ensemble methods for survival prediction in gene expression data." *Bioinformatics*, 27(3): 359–367. doi: http://dx.doi.org/10.1093/bioinformatics/btq660. 930

Brown, P. J., Vannucci, M., and Fearn, T. (2002). "Bayes model averaging with selection of regressors." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3): 519–536. MR1924304. doi: http://dx.doi.org/10.1111/1467-9868.00348. 912

Casella, G. (2001). "Empirical Bayes Gibbs sampling." *Biostatistics (Oxford, England)*, 2(4): 485–500. doi: http://dx.doi.org/10.1093/biostatistics/2.4.485. 914, 924

Castillo, I. and Van Der Vaart, A. (2012). "Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences." *The Annals of Statistics*, 40(4): 2069–2101. MR3059077. doi: http://dx.doi.org/10.1214/12-AOS1029. 911, 915

Chatterjee, A. and Lahiri, S. (2011). "Bootstrapping Lasso Estimators." *Journal of the American Statistical Association*, 106(494): 608–625. MR2847974. doi: http://dx.doi.org/10.1198/jasa.2011.tm10159. 910

Chen, Z. and Dunson, D. (2003). "Random effects selection in linear mixed models." *Biometrics*, 59(4): 762–769. MR2025100. doi: http://dx.doi.org/10.1111/j.0006-341X.2003.00089.x. 911

Clyde, M. A. (1999). "Bayesian model averaging and model search strategies." In: *Bayesian statistics, 6 (Alcoceber, 1998)*, 157–185. New York: Oxford Univ. Press. MR1723497. 912

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). "Least angle regression." *The Annals of Statistics*, 32(2): 407–499. MR2060166. doi: http://dx.doi.org/10.1214/009053604000000067. 910

George, E. and McCulloch, R. (1997). "Approaches for Bayesian variable selection." *Statistica Sinica*, 7: 339–374. 911, 912, 930

Geweke, J. F. (1994). "Variable selection and model comparison in regression." Working Paper 539, Federal Reserve Bank of Minneapolis. 912, 930

Griffin, J. E. and Brown, P. J. (2012). "Structuring shrinkage: some correlated priors for regression." *Biometrika*, 99(2): 481–487. MR2931267. doi: http://dx.doi.org/10.1093/biomet/asr082. 912

— (2013). "Some priors for sparse regression modelling." *Bayesian Analysis*, 8(3): 691–702. MR3102230. doi: http://dx.doi.org/10.1214/13-BA827. 912

Hobert, J. P. and Geyer, C. J. (1998). "Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model." *Journal of Multivariate Analysis*, 67(2): 414–430. MR1659196. doi: http://dx.doi.org/10.1006/jmva.1998.1778. 920

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). "Bayesian model averaging: a tutorial." *Statistical Science*, 14(4): 382–401. MR1765176. doi: http://dx.doi.org/10.1214/ss/1009212519. 912

Huang, J., Breheny, P., and Ma, S. (2012). "A Selective Review of Group Selection in High-Dimensional Models." *Statistical Science*, 27(4): 481–499. MR3025130. doi: http://dx.doi.org/10.1214/12-STS392. 909, 921

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. (2011). "Proximal methods for hierarchical sparse coding." *Journal of Machine Learning Research*, 12: 2297–2334. MR2825428. 927

Johnstone, I. M. and Silverman, B. W. (2004). "Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences." *The Annals of Statistics*, 32(4): 1594–1649. MR2089135. doi: http://dx.doi.org/10.1214/009053604000000030. 911, 915, 916

Knight, K. and Fu, W. (2000). "Asymptotics for lasso-type estimators." *Annals of Statistics*, 1356–1378. MR1805787. doi: http://dx.doi.org/10.1214/aos/1015957397. 910

Kuo, L. and Mallick, B. (1998). "Variable selection for regression models." In: *Bayesian Analysis, Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, 60(1): 65–81. MR1717076. 915, 923, 930

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). "Penalized regression, standard errors, and Bayesian lassos." *Bayesian Analysis*, 5(2): 369–411. MR2719657. doi: http://dx.doi.org/10.1214/10-BA607. 910, 911, 913, 914, 922

Leng, C., Lin, Y., and Wahba, G. (2004). "A note on the LASSO and related procedures in model selection." *Statistica Sinica*. Technical report. 927

Lindley, D. V. (1957). "A statistical paradox." *Biometrika*, 44(1/2): 187–192. MR0896257. doi: http://dx.doi.org/10.1093/biomet/44.1-2.187. 914

Lykou, A. and Ntzoufras, I. (2013). "On Bayesian lasso variable selection and the specification of the shrinkage parameter." *Statistics and Computing*, 23(3): 361–390. MR3041441. doi: http://dx.doi.org/10.1007/s11222-012-9316-x. 911, 915

Mairal, J., Jenatton, R., Obozinski, G., and Bach, F. (2010). "Network flow algorithms for structured sparsity." *arXiv:1008.5209 [cs, stat]*. 927

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian variable selection in linear regression." *Journal of the American Statistical Association*, 83(404): 1023–1032. MR0997578. doi: http://dx.doi.org/10.1080/01621459.1988.10478694. 911

Nardi, Y. and Rinaldo, A. (2008). "On the asymptotic properties of the group lasso estimator for linear models." *Electronic Journal of Statistics*, 2: 605–633. Zentralblatt MATH identifier: 06165707. MR2426104. doi: http://dx.doi.org/10.1214/08-EJS200. 918

Park, T. and Casella, G. (2008). "The Bayesian lasso." *Journal of the American Statistical Association*, 103(482): 681–686. MR2524001. doi: http://dx.doi.org/10.1198/016214508000000337. 910, 914, 922, 924

Raman, S., Fuchs, T. J., Wild, P. J., Dahl, E., and Roth, V. (2009). "The Bayesian group-lasso for analyzing contingency tables." In: *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, 881–888. New York, NY, USA: ACM. 910, 911, 914

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2012). "A sparse-group lasso." *Journal of Computational and Graphical Statistics*. MR3173712. doi: http://dx.doi.org/10.1080/10618600.2012.681250. 911, 921

Soussen, C., Idier, J., Brie, D., and Duan, J. (2011). "From Bernoulli–Gaussian deconvolution to sparse signal restoration." *IEEE Transactions on Signal Processing*, 59(10): 4572–4584. MR2882966. doi: http://dx.doi.org/10.1109/TSP.2011.2160633. 915

Stingo, F. C., Chen, Y. A., Tadesse, M. G., and Vannucci, M. (2011). "Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes." *The Annals of Applied Statistics*, 5(3): 1978–2002. Zentralblatt MATH identifier: 1228.62150. MR2884929. doi: http://dx.doi.org/10.1214/11-AOAS463. 924

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288. MR1379242. 910, 926

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2004). "Sparsity and smoothness via the fused lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1): 91–108. MR2136641. doi: http://dx.doi.org/10.1111/j.1467-9868.2005.00490.x. 910

Wang, H. and Leng, C. (2008). "A note on adaptive group lasso." *Computational Statistics & Data Analysis*, 52(12): 5277–5286. MR2526593. doi: http://dx.doi.org/10.1016/j.csda.2008.05.006. 917, 918, 928

Yuan, M. and Lin, Y. (2005). "Efficient empirical Bayes variable selection and estimation in linear models." *Journal of the American Statistical Association*, 100(472): 1215–1225. MR2236436. doi: http://dx.doi.org/10.1198/016214505000000367. 911, 915

— (2006). "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68(1): pp. 49–67. MR2212574. doi: http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x. 910, 917, 927

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." In: *Bayesian inference and decision techniques*, volume 6 of *Studies in Bayesian Econometrics and Statistics*, 233–243. Amsterdam: North-Holland. MR0881437. 912

Zhang, L., Baladandayuthapani, V., Mallick, B. K., Manyam, G. C., Thompson, P. A., Bondy, M. L., and Do, K.-A. (2014). "Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. MR3258055. 911, 912, 914, 930

Zhao, Z. and Sarkar, S. (2012). "On credible intervals for selected parameters under the zero-inflated mixture prior in high dimensional inference." *Unpublished manuscript*. 911

Zhou, M., Chen, H., Ren, L., Sapiro, G., Carin, L., and Paisley, J. W. (2009). "Non-parametric Bayesian dictionary learning for sparse image representations." In *Advances in Neural Information Processing Systems*, 2295–2303. 915

Zou, H. and Hastie, T. (2005). "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301–320. MR2137327. doi: http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x. 910

Acknowledgments

Ghosh's research was partially supported by NSF Grants DMS-1007417 and SES-1026165. The authors want to thank the reviewers for their helpful comments which substantially improved the paper.