NAME: SANNA TOURAY

PANTHER ID: 002800278

COURSE: MONTE CARLO

TITLE: FINAL PROJECT REPORT

Bayesian Variable Selection and Estimation for Group Lasso

Xiaofan Xu[1] and Malay Ghosh[2]

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 basic 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.

[1] Department of Statistics, University of Florida, xiaofanxufl@gmail.com

[2] Department of Statistics, University of Florida, ghoshm@stat.ufl.edu

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 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):

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

where, $\boldsymbol{\epsilon}_{n\times 1} \sim \boldsymbol{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, β_g is a coefficients vector of length $\boldsymbol{m_g}$, and $\mathbf{X}_g$ is an $n \times m_g$ covariate matrix corresponding to the factor β_g, $g = 1, 2..., G$. 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., $m_1 = m_2 = \cdots = m_G = 1$.

1.1 Motivation for Bayesian Approaches

• In many applications(e.g. genomics, finance),

predictors often form grouped naturally(e.g gene pathway in biological data, economic sector).

• Bayesian methods provide credible intervals, better probabilistic interpretation.

• Challenges with traditional lasso: lacks standard errors, variable selection issues

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

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

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 β_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 $\beta_g (g = 1,\ldots,G)$ 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 β_g. We propose the following hierarchical Bayesian group lasso model with an independent spike and slab type prior for each factor β_g:

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

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

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

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

where $\delta_0(\beta_g)$ denotes a point mass at $\mathbf{0} \in R^{mg}$, $\beta_g = (\beta_{g1},\ldots,\beta_{gmg})^T$. 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 $n \to \infty$.

2.2 Marginal Prior for β_g and Connection with Penalized Regression

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

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

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

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

We can observe from (11) that the marginal prior for β_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 β_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: $\beta_g = \gamma_g b_g$, where γ_g is an indicator that only takes value 0 or 1, and $b_g = (b_{g1}, b_{g2}, \ldots, b_{gmg})^T$. We then place a Multi-Laplace prior on b_g and a Bernoulli prior on γ_g,

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

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

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

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

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 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 $\beta_{(g)}$ denote the β 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 $X_{(g)}$ denote the covariate matrix corresponding to $\beta_{(g)}$, that is,

$$X_{(g)} = (X_1, \ldots, X_g - 1, X_{g+1}, \ldots, X_G),$$

where X_g is the design matrix corresponding to β_g.

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

- Let $\mu_g = \boldsymbol{\Sigma}_g X_g{}^T(Y - X_{(g)}\boldsymbol{\beta}_{(g)})$, $\boldsymbol{\Sigma}_g = (\boldsymbol{X}_g^T\boldsymbol{X}_g + \frac{1}{\tau_g^2}\boldsymbol{I}_{m_g})^{-1}$, then the conditional posterior distribution of β_g is a spike and slab distribution, $\beta_g|\text{rest} \sim (1 - l_g)\text{N}(\mu_g, \sigma^2\boldsymbol{\Sigma}_g) + l_g\delta_0(\beta_g)$, $g = 1, \ldots, G$, where $l_g = p(\beta_g = 0|\text{rest})$

$$= \frac{\pi_0}{\pi_0 + (1-\pi_0)(\tau_g^2)^{-\frac{m_g}{2}}|\boldsymbol{\Sigma}_g|^{\frac{1}{2}}\exp\left\{\frac{1}{2\sigma^2}\|\boldsymbol{\Sigma}_g^{\frac{1}{2}}X_g^T(\boldsymbol{Y} - \boldsymbol{X}_{(g)}\boldsymbol{\beta}_{(g)})\|_2^2\right\}}.$$

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

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 β_g's with parameters τ_g, γ_g and hyperpriors on τ_g, γ_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 $V_g = \mathrm{diag}\{(\frac{1}{\tau_{gj}^2} + \frac{1}{\gamma_g^2})^{-1},\ j = 1, 2, \ldots, m_g\}$. Then we place the following multivariate prior on τ_g, γ_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 β_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^2_1,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^2_1 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:

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

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

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

Note that when $\tau_{gj} = 0$, β_{gj} is essentially dropped out of the model even when $b_{gj} = 0.6$ 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 $(\lambda_1 + \lambda_2)$kβ_gk1 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 $\gamma_{gj}^{(s2)}$ indicates whether $\beta_{gj} = 0$. The following priors are assumed:

$$\gamma_g^{(1)} \sim \text{Bernoulli}(\pi_0), \qquad g = 1,...,G,$$

$$\gamma_{gj}^{(2)} \sim \text{Bernoulli}(\pi_1), \qquad g = 1,...,G; j = 1,...,m_g,$$

$$b_{gj} \sim N\left(0, s^2\right), \quad g = 1,\ldots,G;\ j = 1,\ldots,m_g.$$

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.

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).

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.

- 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.

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 criterions 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 .

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