Bayesian Analysis (2020) **15**, Number 3, pp. 909–935

Consistent Group Selection with Bayesian High Dimensional Modeling

Xinming Yang* and Naveen N. Narisetty†

Abstract. In many applications with high dimensional covariates, the covariates are naturally structured into different groups which can be used to perform efficient statistical inference. We propose a Bayesian hierarchical model with a spike and slab prior specification to perform group selection in high dimensional linear regression models. While several penalization methods and more recently, some Bayesian approaches are proposed for group selection, theoretical properties of Bayesian approaches have not been studied extensively. In this paper, we provide novel theoretical results for group selection consistency under spike and slab priors which demonstrate that the proposed Bayesian approach has advantages compared to penalization approaches. Our theoretical results accommodate flexible conditions on the design matrix and can be applied to commonly used statistical models such as nonparametric additive models for which very limited theoretical results are available for the Bayesian methods. A shotgun stochastic search algorithm is adopted for the implementation of our proposed approach. We illustrate through simulation studies that the proposed method has better performance for group selection compared to a variety of existing methods.

Keywords: group selection, spike and slab priors, Bayesian variable selection, shotgun stochastic search.

1 Introduction

Variable selection is a crucial statistical tool especially in high dimensional data settings as it provides interpretability of the learned model and also often helps to improve prediction power by removing irrelevant predictors. Several variable selection methods from frequentist and Bayesian viewpoints have been proposed in the literature. Some of the popular choices of penalty based methods include least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996), bridge penalization (Frank and Friedman, 1993), smoothly clipped absolute deviation (SCAD) estimator (Fan and Li, 2001), and minimax concave penalty (MCP) estimator (Zhang, 2010). Several Bayesian variable selection methods have been developed in the literature with a variety of prior structures (George and McCulloch, 1993, 1997; Ishwaran and Rao, 2005; Johnson and Rossell, 2012; Narisetty and He, 2014; Ročková and George, 2014, 2018; Spitzner, 2019).

In many applications, the predictors under consideration naturally exhibit a grouping structure due to their inherent similarities. For example, in gene expression data, genes can be classified into different groups based on the phenotypic traits they control; in stock market data, stocks from the same sector form a group. Group structure

*Department of Statistics, University of Illinois at Urbana-Champaign, xyang104@illinois.edu
†Department of Statistics, University of Illinois at Urbana-Champaign, naveen@illinois.edu

 https://doi.org/10.1214/19-BA1178

also arises naturally in many statistical models. In multiple factor analysis of variance (ANOVA) models, dummy variables encoding the same factor form a group; in polynomial regressions or nonparametric additive models, the basis functions involving the same predictor become a group. Variable selection is thus translated to group selection in these problems as we desire to perform selection at the group level. Under these scenarios, traditional variable selection methods which do not incorporate the group information would not be efficient (Breheny and Huang, 2009). Incorporating the grouping structure naturally available in the predictors or implied by the statistical model helps in performing more precise variable selection. Therefore, it is important to develop variable selection methods accounting for group structure. While there are several existing frequentist methods and studies of their theoretical properties available for group selection (Yuan and Lin, 2006; Bach, 2008; Wang and Leng, 2008; Nardi and Rinaldo, 2008; Wang et al., 2008), the Bayesian approaches and their theoretical properties have been much less explored. Given the recent insights about the advantages of Bayesian approaches for variable selection and their model selection consistency properties under weaker conditions compared to penalization approaches (Johnson and Rossell, 2012; Narisetty and He, 2014), it is important to investigate them for group selection.

Let us consider the linear regression model with G groups:

$$Y = \sum_{g=1}^{G} X_g \beta_g + \epsilon, \tag{1.1}$$

where Y is an $n \times 1$ vector, $\epsilon \sim N_n(0, \sigma^2 I)$, and X_g and β_g are respectively the $n \times m_g$ design matrix and the $m_g \times 1$ coefficient vector corresponding to the g'th group. The group Lasso estimator (Yuan and Lin, 2006) was proposed to perform group selection and is defined as the minimizer of the following objective function:

$$GL(\beta) := \frac{1}{2} \left\| Y - \sum_{g=1}^{G} X_g \beta_g \right\|_2^2 + \lambda \sum_{g=1}^{G} \sqrt{m_g} \|\beta_g\|_2.$$

This is a natural extension of the Lasso (Tibshirani, 1996) by applying the L_1 penalty to the L_2 norms of the group coefficients. The theoretical properties of the group Lasso estimator were studied in Bach (2008) and Nardi and Rinaldo (2008). Group selection consistency is only guaranteed under restrictive correlation conditions which extend the irrepresentable condition (Zhao and Yu, 2006) of Lasso. To improve that, Wang and Leng (2008) and Nardi and Rinaldo (2008) studied an adaptive group Lasso method that generalizes the adaptive Lasso (Zou, 2006) in the group setting. With adaptive weights of penalty on each group, the adaptive group Lasso achieves group selection consistency under more flexible conditions. Non-convex group penalized methods like group SCAD and group MCP can be formulated similarly by applying the corresponding penalty to the L_2 norms of the group coefficients. Various penalized methods have also been proposed to perform sparse group selection. We mention group bridge (Huang et al., 2009), sparse-group Lasso (Simon et al., 2013), and group exponential Lasso (Breheny, 2015), among others. An extensive review of penalized group selection approaches is provided in Huang et al. (2012).

To formulate the group selection problem in the Bayesian framework, multivariate-Laplacian priors over each group of coefficients were considered by Raman et al. (2009) and Kyung et al. (2010). This approach is often referred to as Bayesian group Lasso. The maximum a posteriori (MAP) estimator of Bayesian group Lasso is equivalent to the group Lasso estimator, however, Bayesian group Lasso provides the entire posterior distribution on the parameter space as opposed to a single point estimator provided by group Lasso. The posterior distribution of Bayesian group Lasso has the flexibility to incorporate prior knowledge and can also be used in a more general way (such as posterior mean, uncertainty estimation) as opposed to one summary statistic in the form of the MAP estimator. However, model selection consistency results for the Bayesian group Lasso procedure are not extensively studied. Hernández-Lobato et al. (2013) introduced generalized spike-and-slab priors for group selection problems and used expectation propagation to perform approximate inference. Xu and Ghosh (2015) considered adopting the prior of Bayesian group Lasso as the slab prior and introduced Bayesian group Lasso with spike and slab priors (BGL-SS). Under the orthogonal design, the posterior median estimator of BGL-SS was shown to have the oracle property (Fan and Li, 2001). However, there are no theoretical results established under general designs beyond the orthogonal design.

In this paper, we propose spike and slab priors to perform model selection at the group level in the Bayesian framework. Spike and slab priors would be imposed on the group coefficients rather than individual coefficients. Meanwhile, a binary latent variable would be explicitly introduced for every group to indicate whether the corresponding group is active or not. We establish strong selection consistency (Johnson and Rossell, 2012) of our method under a general setup allowing both the number of groups and the size of the true model to go to infinity. With the proposed method, we consider the application of it in some special cases of group selection problems including nonparametric additive models and seemingly unrelated regressions and demonstrate that strong selection consistency continues to hold for these models. To the best of our knowledge, Shang and Li (2014) provide the only existing result for model selection consistency of Bayesian nonparametric additive model in high dimensions. However, they established strong selection consistency of their method when the sparsity level is correctly specified through the model hyperparameter, which is usually not available in reality.

We place a point mass at zero as the spike prior and a multivariate normal with diagonal covariance matrix as the slab prior. Though this prior is similar to generalized spike-and-slab priors and BGL-SS, our method differs from them in two aspects. First, following the same idea of Narisetty and He (2014), we suggest that the slab prior should be sample size dependent to achieve appropriate shrinkage. With this specification, our proposed method is shown to have strong selection consistency under more general designs. Second, we perform group selection based on the inference on the latent binary variables rather than the group coefficients. We propose a shotgun stochastic search algorithm similar to the one by Hans et al. (2007) to search for the MAP estimator in the model space rather than generating samples from the posterior. Our algorithm includes a deterministic searching layer on top of the algorithm of Hans et al. (2007) to make it explore a larger model space and to converge in fewer iterations.

In the theoretical analysis, we consider a quite flexible setup to accommodate both realistic design situations as well as commonly used statistical models such as the nonparametric additive models. Our method is shown to be able to asymptotically select all the groups containing any active individual predictors. Our theoretical results are stronger than any of the existing Bayesian group selection consistency results (Shang and Li, 2014; Xu and Ghosh, 2015) in terms of the weakness of assumptions and the generality of the results. Compared with traditional variable selection methods ignoring the group structure, our proposed group selection method achieves consistency under weaker conditions on the signal strength as it makes use of the group information. When there exists a large within-group correlation, it is likely that traditional variable selection methods will select only some of the individual predictors in the group and ignore the others. In contrast, our theoretical results show that our proposed method will not suffer from a large within-group correlation and performs well under this scenario.

The rest of the paper is organized as follows. We introduce the proposed method and theoretical results in Section 2. The application of our proposed method in some special cases is discussed in Section 3. Algorithms for the implementation of the method are discussed in Section 4. Simulation and real data studies are conducted in Section 5 and Section 6, respectively. Finally, we draw the conclusion in Section 7.

2 Bayesian Group Selection

Consider the regression model (1.1) and let $m = (m_1, \ldots, m_G)$ denote the numbers of individual variables within each group and $p = \sum m_g$ is the total number of individual variables. For the purpose of simplicity, we assume the design matrix of every group X_g, $g = 1, \ldots, G$, to be full-ranked. Furthermore, we center Y and normalize every column of the design matrix $X = (X_1, \ldots, X_G)$. We introduce a latent binary random vector $Z = (Z_1, \ldots, Z_G)$ with entries equal to either 1 (active) or 0 (not) to indicate selection of groups.

2.1 Model

Our Bayesian hierarchical model is given by:

$$\begin{aligned} Y \mid (X, \beta, \sigma^2) &\sim N(X\beta, \sigma^2 I_n), \\ \beta_g \mid (Z_g, \sigma^2) &\sim (1 - Z_g)\, \delta_0(\beta_g) + Z_g\, N(0_{m_g}, \sigma^2 \tau^2 I_{m_g}), \\ Z_g &\sim \text{Bernoulli}(q), \\ \sigma^2 &\sim \text{Inv-Gamma}(\gamma_1, \gamma_2), \end{aligned} \tag{2.1}$$

where $g = 1, \ldots, G$ and τ^2, q are the prior parameters which depend on n. In the model, we use $\delta_0(\beta_g)$, a point mass at zero, as the spike prior and $N(0_{m_g}, \sigma^2 \tau^2 I_{m_g})$ as the slab prior. We treat τ^2 and q as tuning parameters and let them go to infinity and zero, respectively, along with n to achieve appropriate sparsity at the group level. We shall perform group selection based on the MAP estimator of the posterior distribution of Z.

We work with Model (2.1) for its simplicity and convenience in studying the theoretical properties. However, several alternative priors and further hierarchies on the hyper-

parameters of the Model (2.1) can be considered in practice. For instance, a Beta prior on q and a Gamma prior on τ^2 can be specified. For the slab prior, one can alternatively adopt an objective Zellner's g-prior (Zellner, 1986), that is, $N\big(0_{m_g}, \sigma^2\tau^2(X_g'X_g)^{-1}\big)$. This formulation is actually equivalent to first transform X_g, the design matrix of the g'th group, to $X_g(X_g'X_g)^{-1/2}$, and then place the original slab prior described in (2.1) on the new group coefficients corresponding to $X_g(X_g'X_g)^{-1/2}$. By this procedure, the design matrix for each group is orthonormalized, so which slab prior to choose is a matter of whether orthonormalization within the group is needed. The distinction between the two different choices of the slab prior will be discussed further later.

In our implementation, we need the sample size for selecting our hyperparameters of the Model (2.1), which is not completely a subjective Bayesian approach as the prior may not be interpreted entirely as a representation of prior knowledge. The subjective aspect of our prior comes from the form of the prior specification that induces group sparsity while the specific values for the hyperparameters are not chosen in a subjective manner. Motivated by the difficulty in high dimensions for having prior belief about the distribution of a high dimensional parameter in its entirety, several authors (George and Foster, 2000; Park and Casella, 2008; Scott and Berger, 2010; Xu and Ghosh, 2015; Ročková and George, 2014, 2016; Gan et al., 2018) take this approach for selecting hyperparameters while incorporating prior belief through the prior structures. These approaches as well as our proposal can be considered under the broad umbrella of objective priors that include a wide range of priors that could be data-dependent implicitly or explicitly such as Jeffreys' prior (Jeffreys, 1961), reference prior (Bernardo, 1979), and Zellner's g-prior (Zellner, 1986).

Our approach of specifying hyperparameters that depend on the sample size is quite similar to empirical Bayes strategies, which specify hyperparameters based on the data and can therefore also be sample size dependent. Empirical Bayes strategies use more specific information from the data compared to our approach as they often employ a likelihood maximization strategy. For instance, George and Foster (2000) estimated both the hyperparameter in their g-prior and prior inclusion probability based on a marginal or conditional maximum likelihood criterion; Scott and Berger (2010) specified the prior inclusion probability by marginal maximum likelihood; Park and Casella (2008), Kyung et al. (2010), and Xu and Ghosh (2015) specified the parameters of their hyperpriors within a Monte Carlo expectation-maximization (EM) algorithm. In all these approaches, hyperparameters are chosen by maximizing certain form of likelihood that requires access to the full data while our specification of the hyperparameters depends only on the sample size.

These objective approaches are in contrast with fully subjective Bayesian approaches such as Spitzner (2019) where prior distributions are entirely characterized based on prior knowledge before the data are collected. If a specific prior knowledge on the prior distribution is available that allows fixing the prior hyperparameters at specific values, this can certainly be done in our framework and the approach would be fully Bayesian. Another alternative is to place a prior on the hyperparameters but selection of its hyperparameters is also crucial since the procedures can be sensitive to those parameters in high dimensions.

We now introduce some of the notations to be used in this paper.

Model indices: Let k, a binary vector of length G with entries equal to either 1 or 0, be the index of the active groups present in a regression model. For instance, the regression model $Y = X_2\beta_2 + X_3\beta_3 + \epsilon$ will be indexed by the binary vector $(0, 1, 1, 0, \cdots, 0)_{G\times 1}$. Let t denote the index of the true model which has 1's for all the groups containing active individual variables. For example, if we have 6 variables divided into three groups with $m_1 = m_2 = m_3 = 2$ and the active variables are the first 3 variables which are in groups 1 and 2, then $t = (1, 1, 0)$.

Model symbols: For a regression model indexed by k, denote $X(k)$, r_k, $P(k)$, $\beta(k)$, and $m(k)$ as the corresponding design matrix, rank of $X(k)$, projection matrix corresponding to $X(k)$, regression coefficients, and the vector of group sizes, respectively. For a vector, the operator $|\cdot|$ outputs the L_1 norm of the vector. For example, $|k| = \sum k_g$ is the number of active groups in the model indexed by k; $|m(k)| = \sum [m(k)]_g$ is the number of active individual variables in the model indexed by k.

2.2 Posterior Distribution of Z

We now proceed to derive the marginal posterior distribution of Z as group selection will be performed based on the corresponding MAP estimator. As will be seen later, this marginalization would be helpful not only in our theoretical analysis but also for computation.

The posterior probability of the model indexed by k can be computed by summing and integrating out β from the joint posterior distribution of β and Z which is given by

$$P(Z = k \mid Y, \sigma^2) \propto Q_k \left(\frac{q}{1-q}\right)^{|k|} \exp\left\{-\frac{1}{2\sigma^2}\tilde{R}_k\right\}, \tag{2.2}$$

where

$$Q_k = \left|D_k + X(k)'X(k)\right|^{-1/2}\left|D_k\right|^{1/2},$$
$$\tilde{R}_k = Y'\left\{I - X(k)\left[D_k + X(k)'X(k)\right]^{-1}X(k)'\right\}Y, \text{ and } D_k = \tau^{-2}I_{|m(k)|}.$$

In the posterior probability (2.2), $\tilde{R}_k$ can be interpreted as regularized residual sum of squares; $[q/(1-q)]^{|k|}$ deals with the number of active groups in the model indexed by k; Q_k represents the variation within the design matrix $X(k)$. To see the last point, rewrite Q_k as $|I + \tau^2 X(k)X(k)'|^{-1/2}$ by Sylvester's determinant theorem. Recall from principal components analysis that $|X(k)'X(k)|$ is the product of the variances of the principal components of $X(k)$. Including more groups in the model would introduce more variation within the design matrix $X(k)$. Similar expression of the posterior probability (2.2) without a group structure is obtained by imposing Bayesian shrinking and diffusing priors (Narisetty and He, 2014) on the regression coefficients.

Remark. Looking for the model that maximizes the posterior probability (2.2) becomes the familiar problem of the trade-off between the residual sum of squares and the model

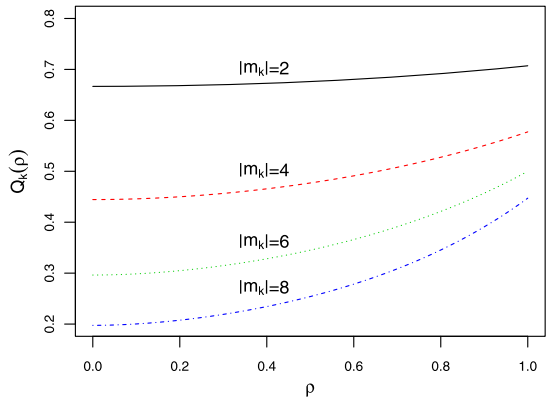

Figure 1: $Q_k(\rho)$ versus ρ given $n = 50$ and $\tau^2 = 0.01$.

size as seen in many traditional model selection criteria like the Akaike information criterion (AIC) and the Bayesian information criterion (BIC).

We now illustrate the philosophy of whether the orthonormalization within the group is needed or not. If no orthonormalization is performed, the within-group correlation would be favored in the sense that given the regularized residual sum of squares and model size, the posterior probability (2.2) would be larger with a larger within-group correlation. To see this, suppose k is the index of an one-group model, that is, $|k| = 1$ and the correlation between every pair of two individual variables is equal to ρ:

$$\frac{X(k)'X(k)}{n} \to (1-\rho)I_{|m(k)|} + \rho J_{|m(k)|},$$

where $J_{|m(k)|}$ denotes the all-ones matrix. We have

$$\begin{aligned} Q_k(\rho) &= \left|I_{|m(k)|} + \tau^2 X(k)X'(k)\right|^{-1/2} \\ &\to \left|\left[n\tau^2(1-\rho)+1\right]I_{|m(k)|} + n\tau^2\rho J_{|m(k)|}\right|^{-1/2} \\ &= \left\{\left(1+\frac{|m(k)|\rho}{1-\rho+1/n\tau^2}\right)\left[1+n\tau^2(1-\rho)\right]^{|m(k)|}\right\}^{-1/2}, \end{aligned}$$

which increases and therefore so does the posterior probability (2.2) when ρ goes up as can be seen in Figure 1. Alternatively, if orthonormalization within the group is done, or equivalently, a g prior is adopted as the slab prior, this feature of favoring the within-group correlation would be discarded. We believe that the within-group correlation should be favored, so we stick to our original slab prior in Model (2.1).

2.3 Theoretical Results

We shall now provide the strong selection consistency of our Bayesian hierarchical model (2.1) in the sense that the posterior probability (2.2) of the true model goes to one as the

sample size goes to infinity. We consider a general design allowing both the number of groups G and the size of the true model $|t|$ to go to infinity and assume σ^2 to be fixed for simplicity. We first introduce the following notations to be used in our theoretical results:

Operations of model indices: For a regression model indexed by k, we use $k^c = 1_G - k$ as the index of its complementary model. For two models indexed by k and w respectively, the set operations $k \cup w$ and $k \cap w$ index the models corresponding to the union and intersection of the covariates indexed by k and w, respectively. In addition, $k \subset w$ (or $w \supset k$) denotes that the model indexed by k is a submodel of the model indexed by w, and $k \subsetneq w$ (or $w \supsetneq k$) denotes strict inclusion. The submodel of the model indexed by k that only includes (or excludes) its i'th active group by $k_{(i)}$ (or $k_{(-i)}$) for $i = 1, \ldots, |k|$.

Rates: For sequences a_n and b_n, $a_n \sim b_n$ means $a_n/b_n \to c$ for some constant $c > 0$; $a_n \preceq b_n$ (or $b_n \succeq a_n$) means $a_n = O(b_n)$; $a_n \prec b_n$ (or $b_n \succ a_n$) means $a_n = o(b_n)$.

Next, we describe the conditions used to achieve strong selection consistency.

Define

$$\Delta_1 = \inf_{\{i=1,\ldots,|t|\}} \left\| \left[I - P(t_{(-i)}) \right] X(t_{(i)}) \beta(t_{(i)}) \right\|_2^2,$$

where $P(t_{(-i)})$ is the projection matrix corresponding to the model indexed by $t_{(-i)}$ which is the submodel of the true model that only excludes the i'th active group, and

$$\lambda_{\min} = \inf_{\{k:\, k \supset t,\, r_k \le (K+1) r_t\}} \phi_{\min} \left(\frac{X(k)' X(k)}{n} \right),$$
$$\bar{\lambda} = \inf_{\{k:\, |k \cap t^c| > 0\}} \bar{\phi} \left(\frac{X(k)' [I - P(k \cap t)] X(k)}{n} \right),$$

where $\phi_{\min}$ outputs the minimum nonzero eigenvalue of the input matrix and $\bar{\phi}$ outputs the geometric mean of the nonzero eigenvalues of the input matrix.

For a fixed K, define

$$\Delta_2 = \inf_{\{k:\, r_k \le K r_t,\, k \not\supset t\}} \left\| \left[I - P(k) \right] X(t) \beta(t) \right\|_2^2.$$

We assume the following regularity conditions:

(A.1) On the number of groups: $G \to \infty$ and $G = e^{o(n)}$.

(A.2) On the marginal prior probability: $q \sim G^{-1}$.

(A.3) On the variance of the slab prior: $n\tau^2 \sim G^{2+\eta} (\bar{\lambda})^{-\eta}$ and $n\tau^2 \succ \lambda_{\min}^{-1}$, for some $\eta > 0$.

(A.4) Sensitivity condition: $\Delta_1 > (1+\epsilon_1)\sigma^2 r_t \left[(4+\eta) \log G - \eta \log \bar{\lambda} \right]$, for some $\epsilon_1 > 0$.

(A.5) Specificity condition: $\Delta_2 > (1+\epsilon_2)\sigma^2 r_t \left[(4+\eta) \log G - \eta \log \bar{\lambda} \right]$, for some $K > \max\{8/\eta + 1,\, \eta/(\eta - 1)\}$ and $\epsilon_2 > 0$.

Here, conditions (A.1)–(A.3) are primarily related to the rates of the prior parameters and conditions (A.4) and (A.5) are concerned with the identifiability of the true model.

Condition (A.1) allows the number of groups G to grow near-exponentially along with the sample size n. This is the strongest result available in the group selection literature. Under a similar near-exponential rate for the number of groups, Wei and Huang (2010) established the selection consistency of adaptive group Lasso and Shang and Li (2014) achieved the selection consistency for nonparametric additive models.

Condition (A.4) deals with the identifiability of the active groups and condition (A.5) is about preventing the selection of the inactive groups. Both the conditions essentially require that for any false model of moderate size that misses some active group, it cannot fit the mean response $X(t)\beta(t)$ well enough and its residual sum of squares would be lower bounded at a certain rate. It is worth noting that our conditions depend on the number of groups G instead of the number of variables p which can be much larger than G. Otherwise, the conditions would be more restrictive if they have p in place of G. This illustrates an advantage of incorporating the group information. Under the orthogonal design matrix, that is, $X'X = nI$, conditions (A.4) and (A.5) could be further simplified as

$$\inf_{\{i=1,\ldots,|t|\}} \|\beta(t_{(i)})\|_2 > \left(cr_t \frac{\log G}{n}\right)^{1/2},$$

for some $c > 0$. Therefore, the infimum group signal strength is allowed to shrink to zero with the sample size at a fast rate.

When the correlations between the covariates are high, conditions (A.4) and (A.5) can still be satisfied as long as the active coefficients are strong enough. To see this, consider the simple case of having one active group with design matrix X_1 and one inactive group with design matrix X_2. Here, we assume X_1 and X_2 are orthonormalized and the correlation matrix between the two groups $X_2'X_1 = \rho J$ for simplicity. Thus,

$$\begin{aligned}\|(I - P_2)X_1\beta_1\|_2^2 &= \|X_1\beta_1 - X_2X_2'X_1\beta_1\|_2^2 \\ &= \|\beta_1\|_2^2 - \|X_2'X_1\beta_1\|_2^2 \\ &= \|\beta_1\|_2^2 - \rho^2 \|J\beta_1\|_2^2.\end{aligned}$$

The difference between $\|\beta_1\|_2^2$ and $\rho^2\|J\beta_1\|_2^2$ could be lower bounded even when ρ is large as long as the active coefficients β_1 are large enough. In contrast, Bach (2008) derived the sufficient and necessary conditions for selection consistency of group Lasso as an extension of the irrepresentable condition (Zhao and Yu, 2006) of Lasso. For the selection consistency of group Lasso, the strength of the signals within the active groups does not play a role in Conditions (4) and (5) of Bach (2008). Due to this, once the correlation structure fails to satisfy Conditions (4) and (5) of Bach (2008), no matter how large the active signals are, selection consistency would not be achieved by group Lasso, which is quite restrictive and can be easily violated.

Theorem 2.1. *Under conditions (A.1)–(A.5), our proposed Bayesian hierarchical model (2.1) has strong model selection consistency property. That is,*

$$P(Z = t \mid Y) \xrightarrow{P} 1 \quad \text{as } n \to \infty.$$

A proof of Theorem 2.1 is provided in the Supplementary Material (Yang and Narisetty, 2019).

3 Applications to Specific Statistical Models

So far, we have introduced our Bayesian hierarchical model under general regression models with group structures present. We now discuss how our proposed method can be applied to some special statistical models which can be formulated as group selection problems. Group selection methods have a lot of applications in statistical problems as well as in real data analysis, as discussed in Huang et al. (2012). Here, we consider the application of our proposed method in nonparametric additive models and seemingly unrelated regressions and illustrate that our established strong selection consistency continues to hold under these special cases.

3.1 Nonparametric Additive Models

One natural application of group selection methods is the nonparametric additive model:

$$Y = \sum_{j=1}^{p} f_j(X_j) + \epsilon,$$

where f_j's are some unknown smooth univariate functions and $f_j(X_j)$'s are usually referred to as nonparametric components. A class of polynomial spline functions is used to approximate the unknown functions f_j's. Every continuous function can be approximated arbitrarily well by polynomial splines using a sufficient number of knots with a fixed order. The class of polynomial spline functions evaluated for the same individual feature are grouped naturally and thus the selection of the nonparametric components becomes a group selection problem.

Suppose only the first α nonparametric components are active so that the true model is indexed by $t = (\underbrace{1, \ldots, 1}_{\alpha}, \underbrace{0, \ldots, 0}_{p-\alpha})$. Use $\mathcal{S}(\mathcal{K})$ to denote the function space of polynomial splines of order l with simple knots $\mathcal{K} = \{\xi_1, \ldots, \xi_N\}$ where $a < \xi_1 < \cdots < \xi_N < b$. The dimension of $\mathcal{S}(\mathcal{K})$ would be $d = N + l$. Denote a set of basis functions of $\mathcal{S}(\mathcal{K})$ as $\{\phi_1(x), \ldots, \phi_d(x)\}$ and every X_j is expanded as a group of d predictors, $\{\phi_1(X_j), \ldots, \phi_d(X_j)\}$, to approximate the corresponding $f_j(X_j)$.

To ensure that the nonparametric components can be approximated well enough, we make the following assumptions:

(B.1) $\sum_{i=1}^{n} f_j(X_{ij}) = 0$, $j = 1, \ldots, \alpha$.

(B.2) Every X_j is bounded with $X_j \in [a_j, b_j]$, $j = 1, \ldots, \alpha$.

(B.3) $f_j \in L_\infty^l[a_j, b_j]$, $j = 1, \ldots, \alpha$, where $L_\infty^l[a_j, b_j] = \{f : D^{l-1} f \in AC[a_j, b_j], D^l f \in L_\infty[a_j, b_j]\}$ with D to be the differential operator and AC standing for absolute continuity.

(B.4) $N \succeq n^{1/(2l)}$ and the knots are equally spaced.

(B.5) $\alpha \preceq \log p$.

Here, assumption (B.1) deals with the model identifiability and (B.2)–(B.4) are standard assumptions to ensure the approximation power of polynomial splines. We assume the absolute continuity for the smoothness of f_j's in (B.3). In contrast, Huang et al. (2010) assumed f_j's are Hölder continuous for nonparametric additive models under a random design.

Now, we provide the strong selection consistency of our method under nonparametric additive models. We still consider a general design allowing both the number of groups p and the size of the true model α to go to infinity.

Theorem 3.1. *Under assumptions (B.1)–(B.5) and conditions (A.1)–(A.5), our proposed Bayesian hierarchical model (2.1) has strong model selection consistency when the data-generating model is a nonparametric additive model:*

$$P(Z = t \mid Y) \xrightarrow{P} 1 \quad \text{as } n \to \infty.$$

The proof of Theorem 3.1 follows the lines of Theorem 2.1 except that there is an additional approximation error in the linear model which needs to be controlled by the increasing approximation power of polynomial splines. A proof of Theorem 3.1 is provided in the Supplementary Material (Yang and Narisetty, 2019).

Bach (2008) applied group Lasso to nonparametric additive models and established model selection consistency under conditions originating from the irrepresentable condition (Zhao and Yu, 2006). Ravikumar et al. (2009) proposed sparse additive models (SpAM) for nonparametric additive models and assumed conditions similar to the irrepresentable condition for model selection consistency. Therefore, similar to the earlier discussion, our results hold under weaker conditions on the covariates. Shang and Li (2014) proposed a Bayesian nonparametric size-control model which involves a size-control prior that restricts the scope of the target models. Their method has strong selection consistency only when the hyperparameter associated with the size-control prior is correctly specified. This can be violated easily as the sparsity information is usually not available.

3.2 Seemingly Unrelated Regressions

We will now discuss the generality of our results with the seemingly unrelated regressions (SUR) model (Zellner, 1962; Lounici et al., 2011; Obozinski et al., 2011; Huang et al., 2012) as an example. The SUR model consists of a set of regression equations:

$$\boldsymbol{y}^{(i)} = \boldsymbol{X}^{(i)}\boldsymbol{\beta}^{(i)} + \boldsymbol{\epsilon}^{(i)}, \quad i = 1, \ldots, T,$$

where $\boldsymbol{y}^{(i)}$, $\boldsymbol{X}^{(i)}$, $\boldsymbol{\beta}^{(i)}$, and $\boldsymbol{\epsilon}^{(i)}$ are respectively the $n_i \times 1$ response vector, the $n_i \times p$ design matrix, the $p \times 1$ coefficient vector, and the $n_i \times 1$ error vector corresponding to the i'th regression. Equivalently, the equations can be written as

$$\begin{pmatrix} \boldsymbol{y}^{(1)} \\ \boldsymbol{y}^{(2)} \\ \vdots \\ \boldsymbol{y}^{(T)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^{(1)} & 0 & \cdots & 0 \\ 0 & \boldsymbol{X}^{(2)} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{X}^{(T)} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^{(1)} \\ \boldsymbol{\beta}^{(2)} \\ \vdots \\ \boldsymbol{\beta}^{(T)} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}^{(1)} \\ \boldsymbol{\epsilon}^{(2)} \\ \vdots \\ \boldsymbol{\epsilon}^{(T)} \end{pmatrix}. \tag{3.1}$$

Denote the whole design matrix and coefficient vector in (3.1) as $\boldsymbol{X}$ and $\boldsymbol{\beta}$, respectively. The seemingly unrelated regressions become related when the j'th predictors, $j = 1, \ldots, p$, of the regression models describe similar features. Thus, we can reasonably assume that the j'th predictors in the regression equations are more likely to be either included or excluded together, that is, the active set of each individual regression would be the same. Under this formulation, the coefficient vector $\boldsymbol{\beta}$ of length Tp is divided into p groups and the columns with the same remainder after their indices being divided by p stay in the same group. Now due to this group structure, it can be formulated as a group selection problem.

When our proposed method is applied to the seemingly unrelated regressions, we can retain group selection consistency without any further effort and the conditions can be directly translated and further simplified. Denote the design matrix, projection matrix, and coefficient vector of the model indexed by k within the i'th regression as $\boldsymbol{X}^{(i)}(k)$, $\boldsymbol{P}^{(i)}(k)$, and $\boldsymbol{\beta}^{(i)}(k)$, respectively. Then, in the sensitivity condition (A.4),

$$\begin{aligned}\Delta_1 &= \inf_{\{j=1,\ldots,|t|\}} \left\| \left[I - \boldsymbol{P}(t_{(-j)}) \right] \boldsymbol{X}(t_{(j)}) \boldsymbol{\beta}(t_{(j)}) \right\|_2^2 \\ &= \inf_{\{j=1,\ldots,|t|\}} \sum_{i=1}^{T} \left\| \left[I - \boldsymbol{P}^{(i)}(t_{(-j)}) \right] \boldsymbol{X}^{(i)}(t_{(j)}) \boldsymbol{\beta}^{(i)}(t_{(j)}) \right\|_2^2 .\end{aligned}$$

This implies that an active feature would be selected even when its coefficients are small in some individual regressions as long as the overall signal across the T regressions is strong enough. Similarly, the specificity condition (A.5) would tell that the inactive features would not be selected even when the correlations between the active features and the inactive features are large within some individual regressions as long as they are not large across all regressions.

4 Computation

We shall use the MAP estimator of $P(Z \mid Y)$ to perform group selection, so the implementation of the proposed method can be viewed as an optimization problem of finding the model with the maximum posterior probability among all possible models, which is a computationally challenging problem. The discreteness of the parameter space of Z facilitates the use of shotgun stochastic search algorithm (hereafter referred to as SSS) (Hans et al., 2007). Alternatively, Markov chain Monte Carlo (MCMC) methods like the Gibbs sampler can be adopted and inference can be performed with the samples drawn from $P(Z \mid Y)$. To deal with the unknown σ^2, we let the inverse Gamma prior be flat with both shape and scale parameters equal to 0.01 and integrate it out from the posterior distribution (2.2).

4.1 Shotgun Stochastic Search Algorithm

Shotgun stochastic search algorithm attempts to find models having high posterior probabilities by systematically searching the high posterior density regions of the model

space as opposed to MCMC methods which attempt to approximate the posterior distribution on the whole model space. Following the same notations and definitions as in Hans et al. (2007), we let Γ denote the set that will contain the models with large posterior probabilities and $\text{nbd}(k) = \gamma^+(k) \cup \gamma^-(k) \cup \gamma^\circ(k)$ denote the neighborhood of k where $\gamma^+(k)$, $\gamma^-(k)$, and $\gamma^\circ(k)$ are "addition" neighbors, "deletion" neighbors and "replacement" neighbors, respectively. More specifically,

$$\begin{aligned}\gamma^+(k) &= \{w : |w| = |k| + 1 \text{ and } w \supset k\},\\ \gamma^-(k) &= \{w : |w| = |k| - 1 \text{ and } w \subset k\},\\ \gamma^\circ(k) &= \{w : |w| = |k| \text{ and } |w \cap k| = |k| - 1\}.\end{aligned}$$

Given a starting model indexed by $k^{(0)}$, set $\Gamma^{(0)} = \{k^{(0)}\}$ and choose a constant B as the maximum number of models we would like to keep in Γ and a constant C for stopping criteria. Then iterate for $t = 0, \ldots, T$:

Step 1. Update $\Gamma^{(t+1)}$ to be $\Gamma^{(t)} \cup \text{nbd}(k^{(t)})$ and keep the top B models in $\Gamma^{(t+1)}$ having the largest $P(Z \mid Y)$.

Step 2. For every newly enrolled model indexed by w in $\Gamma^{(t+1)}$ that is not in $\Gamma^{(t)}$, update $\Gamma^{(t+1)}$ to be $\Gamma^{(t+1)} \cup \text{nbd}(w)$ and keep the top B models in $\Gamma^{(t+1)}$ having the largest $P(Z \mid Y)$.

Step 3. Sample k^+, k^-, k° from $\gamma^+(k^{(t)})$, $\gamma^-(k^{(t)})$, and $\gamma^\circ(k^{(t)})$ respectively with probabilities proportional to the posterior probabilities normalized within each set of neighbors.

Step 4. Sample $k^{(t+1)}$ from $\{k^+, k^-, k^\circ\}$ with probabilities proportional to the normalized posterior probability.

Step 5. Stop if $\Gamma^{(t+1)}$ remains unchanged during the past C iterations.

The top model in the Γ after running SSS would be our MAP estimator of the posterior distribution $P(Z \mid Y)$.

Here, step 2 is the only difference from the original SSS algorithm (Hans et al., 2007). We add this deterministic step to ensure that the model space around the local maximums of $P(Z \mid Y)$ would be explored more exhaustively at each iteration. Step 2 does not involve $k^{(t)}$, so it would not interfere with the stochastic steps. Compared with the original SSS algorithm, our modified algorithm can converge in fewer iterations and explore a broader region of the model space as the neighborhood of the current top models can be fully explored at each iteration. The dominant part of the computational complexity at each iteration comes from the inversion of $D_k + X(k)'X(k)$. The inverse can be efficiently computed using the Woodbury matrix identity which gives the computational complexity to be $O(n|m(k)|^2)$, where $|m(k)|$ is the number of active individual variables in the model indexed by k. A stochastic approximation Monte Carlo (SAMC) algorithm (Liang et al., 2007, 2013) can also be used for posterior computation but it is a posterior sampling algorithm that generates samples from $P(Z \mid Y)$ as opposed to the SSS algorithm that aims to find the MAP model.

4.2 Gibbs Sampling

Alternatively, we can use a standard Gibbs sampling method to generate samples from $P(Z \mid Y)$. As Z is a binary vector, the full conditional distributions are all Bernoulli distributions:

$$P(Z_g = 1 \mid Z_{-g} = k_{-g}, Y) = \frac{P(Z_g = 1, Z_{-g} = k_{-g} \mid Y)}{P(Z_g = 1, Z_{-g} = k_{-g} \mid Y) + P(Z_g = 0, Z_{-g} = k_{-g} \mid Y)},$$

where $Z_{-g} = (Z_1, \ldots, Z_{g-1}, Z_{g+1}, \ldots, Z_G)$.

If the posterior samples of the coefficients β are desired, they could be generated along with Z and σ^2 from the joint posterior distribution $P(Z, \beta, \sigma^2 \mid Y)$ by Gibbs sampling. The full conditional distributions would still be standard distribution due to the use of conjugate priors. The crucial point is that β and Z need to be blocked together to make Gibbs sampling work. Otherwise, the Markov chain would not be irreducible because of the use of a point mass as the spike prior. A Gibbs sampler to draw samples from $P(Z, \beta, \sigma^2 \mid Y)$ is provided in the Supplementary Material (Yang and Narisetty, 2019). We prefer to sample $Z \mid Y$ directly due to the fact that the analytical integration over β would be beneficial as it results in fast mixing and thus speeds up convergence (George and McCulloch, 1997).

Remark on bi-level selection. When within-group selection of variables is also important, many bi-level selection methods such as the hierarchical structured variable selection (HSVS) method (Zhang et al., 2014) and Bayesian sparse group selection with spike and slab priors (BSGS-SS) (Xu and Ghosh, 2015) have been proposed. It is possible to extend our proposed model (2.1) to induce within-group sparsity by introducing another set of binary variables $S_g = (S_{g1}, \ldots, S_{gm_g})$ for every group of coefficients $\beta_g = (\beta_{g1}, \ldots, \beta_{gm_g})$ to indicate individual level selection. By allowing some of the S_{gi}'s to be inactive when Z_g is active would then result in within-group sparsity. However, the introduction of S_g's would increase the computational burden because the dimension of the model space would be 2^p which can be much larger than 2^G. Therefore, we focus on group selection in this paper and the computational aspects of the model for bi-level selection are deferred for future research.

5 Simulation Results

We will refer to our proposed method as Group Spike and Diffusing prior (GSD) and the estimates computed from SSS and Gibbs sampling algorithms as GSD-SSS and GSD-Gibbs, respectively. To test the performance of our method, we compare it with existing methods including adaptive group Lasso (agLasso), group Lasso (gLasso), group SCAD (gSCAD), group MCP (gMCP), and BGL-SS under different settings. Additionally, for our last simulation setting where there is bi-level sparsity, we also implement the sparse-group Lasso (SGL) method as a bi-level selection approach for comparison.

For good model selection performance, choice of the hyperparameters τ^2 and q are important as the conditions for our theoretical results have demonstrated. In Figure 2, we plot the number of active groups in the MAP estimator under different choices of τ^2

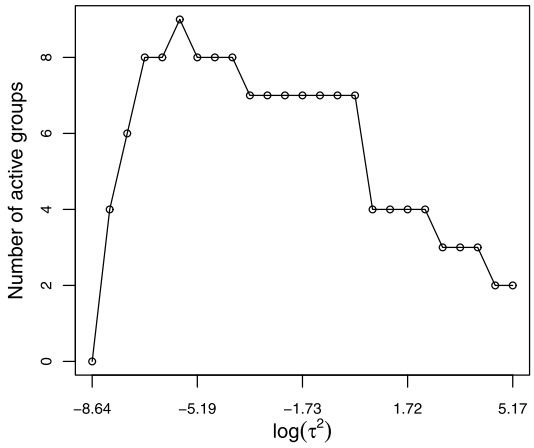

Figure 2: Sparsity level of the MAP estimator under different choices of τ^2.

and a fixed q based on a simulation dataset generated as described in Case 2. As can be seen, the sparsity level does not change rapidly so a very fine tuning of τ^2 may not be needed. In the following simulation studies, we take $q = 1/G$ and tune τ^2 for five different values equal to $G^{2.5}/(10^i n)$ with $i = 0, 1, 2, 3$, and 4. We choose the optimal τ^2 by the mean squared prediction error produced by 5-fold cross-validation using ridge regression. From our simulation studies, this coarse grid of tuning parameters is already able to yield quite promising results. Note that this implies that our hyperparameters are data-dependent and the resultant procedure is similar to an empirical-Bayesian approach which is a common practice in the high dimensional Bayesian literature due to sparse prior knowledge about the hyperparameters. In our model formulation, both q and τ^2 could be tuned if desired. However, the shrinkage effects of q and τ^2 are related to each other and several authors in prior research have observed that it is sufficient to tune one of them by setting the other at a reasonable value (Narisetty and He, 2014; Ročková and George, 2014, 2016; Gan et al., 2018). For this reason, in our empirical studies, we fix q and tune only τ^2. Another possible alternative would be to tune q by fixing τ^2 which we do not pursue in the paper. The choice of $1/G$ for q is motivated by our condition (A.2). From our empirical work, this procedure is already able to yield quite promising results. Otherwise, one may also consider tuning q but it does not seem necessary in most practical contexts.

When implementing our method using SSS, we set B, T and C as 10, 100 and 10, respectively, where B is the number of top models we record, T is the number of iterations, and C is related to the stopping criterion. In the Gibbs sampler, we take a burn-in period of 1000 iterations followed by 1000 iterations to compute the posterior.

We consider 6 simulation designs under the linear regression model (1.1), each with the same sample size of 100. Depending on the context, the number of groups would be either 50 or 100. Group size is taken to be one of 4, 5, 6, 7, or 8 randomly with equal probability on all the designs. The X covariates are generated independently

from a multivariate normal distribution with zero mean, unit variance, and different correlation matrices on different designs. Coefficients of inactive covariates are set to be 0. The errors at each observation are i.i.d. standard normal. In each design, we consider two subcases of weak and strong signal strength with active coefficients to be 0.3 and 1, respectively, unless otherwise specified. Here are the different cases considered:

Case 1 (Baseline design). The number of groups $G = 50$ and the first 3 groups are active. The correlations at both the group and within-group levels are equally 0.5.

Case 2 (Dense model design). The first 7 groups are active with the number of groups and correlation structure same as in Case 1.

Case 3 (High dimensional design). The number of groups $G = 100$ with the first 3 groups to be active. The correlation structure is the same as in Case 1.

Case 4 (Confounding group design). There are $G = 50$ groups with the first 3 groups to be active. The correlations within the first 3 groups and the fourth group (confounding group) are 0.3 and 0.8, respectively. The correlation between a variable from the first 3 groups and a variable from the fourth group is 0.5. All the other entries of the correlation matrix are equal to 0.1.

Case 5 (Bi-level sparsity design). The number of groups $G = 50$ with the same correlation structure as in Case 1. The first 5 groups are active but there are 2 inactive covariates within each of the 5 groups. Active coefficients are equal to 1 under the strong signal strength setting and 0.5 under the weak signal strength setting.

Case 6 (Bi-level high sparsity design). The number of groups $G = 50$ with the same correlation structure as in Case 1. The first 5 groups are active with 1 active covariate in the first 3 groups and 2 active covariates in the other 2 active groups. Active coefficients are equal to 1 under the strong signal strength setting and 0.8 under the weak signal strength setting.

Each simulation design is repeated for 500 times and the results are summarized in Tables 1–6 where four measures are reported: $Z = t$ is the proportion that the selected model is the true model; $Z \supset t$ is the proportion that the selected model contains the true models; the area under the curve (AUC) is the average area under the receiver operating characteristic (ROC) curve; $t \in \text{Path}$ is the proportion that the true model is a candidate model in the solution path.

Comparing different implementations of our method using SSS and Gibbs sampling, overall the two algorithms give similar results but SSS has a slightly better performance in most cases. The computational times for GSD-SSS and GSD-Gibbs using a MacBook Pro with 2.9 GHz Intel Core i5 processor, 8.00 GB memory, and macOS Sierra are reported in Table 7 which show that GSD-SSS is much faster than GSD-Gibbs.

Comparing our method with the competitors, we have the following observations:

- Generally speaking, when signal strength is weak, our method can outperform the competitors except for agLasso in terms of the measure $Z = t$. As signal strength gets stronger, the consistency conditions of our method are easier to satisfy so that

	$Z = t$	$Z \supset t$	AUC	$t \in$ Path
	Active coefficients = 1			
agLasso	0.948	1.000	1.000	1.000
gLasso	0.014	1.000	0.981	0.416
gMCP	0.552	0.560	0.869	0.624
gSCAD	0.468	0.760	0.873	0.534
BGL-SS	0.846	0.986	0.997	0.986
GSD-Gibbs	0.978	1.000	1.000	1.000
GSD-SSS	0.984	1.000	1.000	1.000
	Active coefficients = 0.3			
agLasso	0.514	0.974	0.998	0.910
gLasso	0.004	0.996	0.972	0.288
gMCP	0.008	0.080	0.726	0.050
gSCAD	0.000	0.398	0.778	0.046
BGL-SS	0.436	0.778	0.953	0.606
GSD-Gibbs	0.342	0.872	0.977	0.738
GSD-SSS	0.348	0.880	0.978	0.746

Table 1: Performance of group selection methods in Case 1 (Baseline design).

	$Z = t$	$Z \supset t$	AUC	$t \in$ Path
	Active coefficients = 1			
agLasso	0.930	0.930	0.994	0.930
gLasso	0.002	1.000	0.915	0.008
gMCP	0.000	0.000	0.623	0.000
gSCAD	0.000	0.024	0.666	0.000
BGL-SS	0.008	0.900	0.977	0.548
GSD-Gibbs	0.998	1.000	0.996	1.000
GSD-SSS	0.998	1.000	0.999	1.000
	Active coefficients = 0.3			
agLasso	0.384	0.548	0.949	0.542
gLasso	0.000	0.996	0.910	0.004
gMCP	0.000	0.000	0.615	0.000
gSCAD	0.000	0.016	0.653	0.000
BGL-SS	0.024	0.584	0.929	0.134
GSD-Gibbs	0.360	0.474	0.924	0.448
GSD-SSS	0.422	0.526	0.948	0.512

Table 2: Performance of group selection methods in Case 2 (Dense model design).

our method has much better performance and can outperform all the competitors including agLasso.

- Cases 2 and 3 are modifications of Case 1 in terms of sparsity and dimensionality, respectively. In either case, our method works the best in all the measures under strong signal design. When signal strength is weak, agLasso is the most competitive method and our method is still comparable.

	$Z = t$	$Z \supset t$	AUC	$t \in$ Path
	Active coefficients = 1			
agLasso	0.944	1.000	1.000	1.000
gLasso	0.002	1.000	0.986	0.288
gMCP	0.398	0.402	0.813	0.478
gSCAD	0.352	0.588	0.865	0.418
BGL-SS	0.254	0.944	0.989	0.944
GSD-Gibbs	1.000	1.000	1.000	1.000
GSD-SSS	1.000	1.000	1.000	1.000
	Active coefficients = 0.3			
agLasso	0.502	0.932	0.992	0.876
gLasso	0.000	0.998	0.980	0.160
gMCP	0.002	0.028	0.674	0.022
gSCAD	0.000	0.216	0.748	0.024
BGL-SS	0.300	0.396	0.888	0.360
GSD-Gibbs	0.686	0.774	0.960	0.722
GSD-SSS	0.696	0.778	0.962	0.736

Table 3: Performance of group selection methods in Case 3 (High dimensional design).

	$Z = t$	$Z \supset t$	AUC	$t \in$ Path
	Active coefficients = 1			
agLasso	0.504	1.000	0.993	0.660
gLasso	0.000	1.000	0.979	0.000
gMCP	0.712	0.726	0.875	0.772
gSCAD	0.660	0.728	0.839	0.692
BGL-SS	0.932	1.000	1.000	1.000
GSD-Gibbs	0.996	1.000	0.999	1.000
GSD-SSS	0.996	1.000	1.000	1.000
	Active coefficients = 0.3			
agLasso	0.188	0.960	0.985	0.362
gLasso	0.000	1.000	0.977	0.000
gMCP	0.070	0.408	0.817	0.374
gSCAD	0.004	0.566	0.812	0.542
BGL-SS	0.650	0.838	0.953	0.790
GSD-Gibbs	0.698	0.894	0.961	0.874
GSD-SSS	0.700	0.896	0.959	0.872

Table 4: Performance of group selection methods in Case 4 (Confounding group design).

- Case 4 is a more challenging scenario with the presence of the fourth group as a confounding group. The correlations between the fourth group and the active groups are even larger than those within the active groups. Thus, the fourth group alone could explain a large proportion of the variation of the response variable. In this case, gLasso has a bad performance because it includes the confounding group along with the active groups at most times. Without a good initial estimator,

	$Z = t$	$Z \supset t$	AUC	$t \in$ Path
	Active coefficients = 1			
agLasso	0.720	0.982	0.998	0.978
gLasso	0.008	1.000	0.942	0.046
gMCP	0.202	0.216	0.808	0.260
gSCAD	0.092	0.406	0.805	0.152
BGL-SS	0.482	0.966	0.994	0.914
GSD-Gibbs	0.944	1.000	0.991	1.000
GSD-SSS	0.954	1.000	0.994	1.000
SGL	0.032	1.000	0.934	0.320
	Active coefficients = 0.5			
agLasso	0.460	0.888	0.989	0.780
gLasso	0.000	0.980	0.936	0.038
gMCP	0.034	0.104	0.773	0.088
gSCAD	0.006	0.276	0.778	0.038
BGL-SS	0.398	0.862	0.981	0.700
GSD-Gibbs	0.602	0.766	0.950	0.758
GSD-SSS	0.600	0.772	0.958	0.764
SGL	0.004	0.998	0.926	0.108

Table 5: Performance of group selection methods in Case 5 (Bi-level sparsity design).

	$Z = t$	$Z \supset t$	AUC	$t \in$ Path
	Active coefficients = 1			
agLasso	0.206	0.866	0.983	0.632
gLasso	0.008	0.922	0.927	0.036
gMCP	0.310	0.836	0.976	0.714
gSCAD	0.024	0.930	0.936	0.376
BGL-SS	0.588	0.922	0.991	0.876
GSD-Gibbs	0.798	0.858	0.932	0.844
GSD-SSS	0.798	0.856	0.934	0.846
SGL	0.006	1.000	0.975	0.328
	Active coefficients = 0.8			
agLasso	0.114	0.726	0.971	0.436
gLasso	0.000	0.874	0.918	0.026
gMCP	0.084	0.560	0.947	0.372
gSCAD	0.002	0.804	0.919	0.176
BGL-SS	0.300	0.706	0.955	0.570
GSD-Gibbs	0.472	0.496	0.882	0.502
GSD-SSS	0.470	0.498	0.887	0.502
SGL	0.006	1.000	0.970	0.188

Table 6: Performance of group selection methods in Case 6 (Bi-level high sparsity design).

	GSD-SSS	GSD-Gibbs
Case 1	0.346	11.303
Case 2	0.996	25.069
Case 3	0.476	18.904
Case 4	0.277	13.618
Case 5	0.599	19.253
Case 6	0.385	11.069

Table 7: Average computational times (in min) for GSD-SSS and GSD-Gibbs under different simulation designs based on 10 replications.

agLasso also works poorly. On the contrary, our proposed method can still perform well demonstrating that it is more flexible with different correlation structures.

- Cases 5 and 6 have sparsity at both the group and within-group levels. For every active group in Case 5, 25% to 50% of the predictors are inactive. The within-group sparsity is even higher in Case 6 with only 1 or 2 active covariates within each active group. From Tables 5–6, we see that the penalized group selection methods suffer from the within-group sparsity and the bi-level selection method SGL tends to select more groups than the true model. In contrast, our method is able to accommodate this bi-level sparsity situation as our consistency conditions only rely on the overall signal strength of the whole group.

- Overall, our simulation studies indicate that our proposed method can perform well under a variety of configurations with different dimensionalities, model complexities, sparsity levels, and correlation structures.

6 Application to Gene Expression Data

We use the real dataset from Keller et al. (2018b) to evaluate the performance of our proposed Bayesian group selection method. The dataset contains 21771 expression levels of the islet gene and diabetes-related phenotypes of 378 mice to study the expression quantitative trait loci for pancreatic islet function. The data are available on Dryad Digital Repository (Keller et al., 2018a).

The nonparametric additive model is applied to study the relationship between the gene expression data and two phenotype variables, homeostatic model assessment (HOMA) of insulin resistance (IR) and pancreatic islet function (B). The two phenotype responses are highly correlated with a correlation of 0.819. We adopt the seemingly unrelated regressions (SUR) model to fit the two phenotype responses together to select a common set of covariates relevant for both the phenotypes.

As this is a high dimensional problem with 21771 predictors, we first screen out 500 genes for each of the responses by quantile-adaptive nonlinear variable screening (He et al., 2013) and take the 258 common genes plus the gender predictor to fit a SUR model. He et al. (2013) is a more flexible generalization of the sure independent screening approach of Fan and Lv (2008) and incorporates the information at multiple

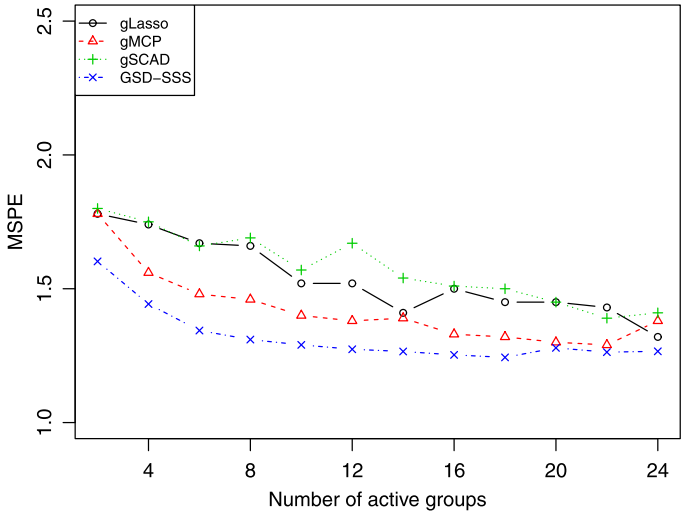

Figure 3: MSPE of SUR model versus number of active groups.

quantile levels. We approximate the nonparametric additive components using cubic splines with 10 knots. Thus, we have 259 groups in total and the final model used is given by

$$\boldsymbol{Y} = \sum_{g=1}^{259} \boldsymbol{X_g}\boldsymbol{\beta_g} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y}$ is the concatenation of the two phenotype responses HOMA-IR and HOMA-B and

$$\boldsymbol{X_g} = \begin{pmatrix} X_g & 0 \\ 0 & X_g \end{pmatrix},$$

with X_g being the g'th predictor evaluated at the basis functions of cubic splines.

To perform group selection for the above model, we use GSD-SSS along with gLasso, gMCP, and gSCAD. For evaluating performance of the models selected by different methods, we randomly split the data into a training set with 80 percent of the observations and a test set with the remaining 20 percent. We only use the training set for model fitting as well as tuning parameter selection and the test set is used to calculate the prediction error for the purpose of performance evaluation. This process is replicated in parallel on a cluster machine with 24 cores resulting in 24 replications.

In Figure 3, we plot the average mean squared prediction errors (MSPE) at different model sizes by GSD-SSS along with gLasso, gMCP, and gSCAD. For the SSS algorithm, we set B to be 50 and to accelerate computation, we abandon the deterministic step 2 of the algorithm and set C for stopping criterion to be 5. We follow the same tuning parameter selection procedure as in Section 5 and the only difference is that the tuning is finer with more values considered for the prior variance parameter τ^2. More specifically, we choose τ^2 from the set of values $\{G^{2.5}/(10^i n) : i = 0, \cdots, b\}$, where $b = 4$ for

Gene ID	Proportion	Gene ID	Proportion
ENSMUSG00000020102	0.792	ENSMUSG00000021708	0.250
ENSMUSG00000027188	0.750	ENSMUSG00000020953	0.250
ENSMUSG00000024563	0.625	ENSMUSG00000005566	0.250
ENSMUSG00000059187	0.625	ENSMUSG00000017615	0.250
ENSMUSG00000029168	0.500	ENSMUSG00000061032	0.208
ENSMUSG00000020653	0.458	ENSMUSG00000003355	0.167
ENSMUSG00000030659	0.458	ENSMUSG00000003660	0.167
ENSMUSG00000017264	0.417	ENSMUSG00000038150	0.125
ENSMUSG00000049823	0.417	ENSMUSG00000059921	0.125
ENSMUSG00000023110	0.417	ENSMUSG00000038007	0.125
ENSMUSG00000027642	0.375	ENSMUSG00000053091	0.125
ENSMUSG00000040972	0.292		

Table 8: Selected Genes under the SUR model along with the proportion of times they were selected in 24 replications.

simulation studies and $b = 10$ for real data analysis. We perform the ridge regression with 5-fold cross-validation on the training set to get coefficients estimates for GSD-SSS and use the test set to calculate MSPE. We can discover that GSD-SSS has the best performance in terms of the average MSPE across different model sizes. This indicates that compared to the competitors, our method selects more powerful groups for prediction at different sparsity levels. In Table 8, we report the genes selected at least twice by GSD-SSS in the 24 replications along with the corresponding proportion of times they were selected.

7 Conclusion

In this paper, we propose a Bayesian hierarchical model with a spike and slab prior specification to perform group selection in high dimensional linear regression models. The group selection consistency of our method is established under mild conditions and we show that this consistency result can be retained for important statistical models including nonparametric additive models and seemingly unrelated regressions. Shotgun stochastic search and Gibbs sampling algorithms can be used for the implementation of our proposed approach. We notice that our proposed shotgun stochastic search algorithm exhibits more computational efficiency due to its fast computation and also has better empirical performance compared to a standard Gibbs sampling algorithm. Our simulation and real data studies indicate that the proposed method has better performance for group selection compared to a variety of state-of-the-art competing methods.

The focus of our paper is to provide a general framework for group selection problems and can certainly be generalized to special cases such as the multivariate response model similar to Greenlaw et al. (2017); Liquet et al. (2017) by considering special covariance matrix structures for the errors. Other generalizations such as auto-regressive models can also be potentially studied within our framework by considering a general structure for the error covariance matrix.

Supplementary Material

Consistent Group Selection with Bayesian High Dimensional Modeling: Supplementary Material (DOI: 10.1214/19-BA1178SUPP; .pdf). Proofs of Theorem 2.1 and Theorem 3.1 and a Gibbs sampler for drawing samples from $P(Z, \beta, \sigma^2 \mid Y)$ are provided in the Supplementary Material.

References

Bach, F. R. (2008). "Consistency of the group Lasso and multiple kernel learning." *Journal of Machine Learning Research*, 9(Jun): 1179–1225. MR2417268. 910, 917, 919

Bernardo, J. M. (1979). "Reference posterior distributions for Bayesian inference." *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2): 113–128. MR0547240. 913

Breheny, P. (2015). "The group exponential Lasso for bi-level variable selection." *Biometrics*, 71(3): 731–740. MR3402609. doi: https://doi.org/10.1111/biom.12300. 910

Breheny, P. and Huang, J. (2009). "Penalized methods for bi-level variable selection." *Statistics and Its Interface*, 2(3): 369. MR2540094. doi: https://doi.org/10.4310/SII.2009.v2.n3.a10. 910

Fan, J. and Li, R. (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American Statistical Association*, 96(456): 1348–1360. MR1946581. doi: https://doi.org/10.1198/016214501753382273. 909, 911

Fan, J. and Lv, J. (2008). "Sure independence screening for ultrahigh dimensional feature space." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5): 849–911. MR2530322. doi: https://doi.org/10.1111/j.1467-9868.2008.00674.x. 928

Frank, L. E. and Friedman, J. H. (1993). "A statistical view of some chemometrics regression tools." *Technometrics*, 35(2): 109–135. 909

Gan, L., Narisetty, N. N., and Liang, F. (2018). "Bayesian regularization for graphical models with unequal shrinkage." *Journal of the American Statistical Association*, 1–14. MR4011774. doi: https://doi.org/10.1080/01621459.2018.1482755. 913, 923

George, E. I. and Foster, D. P. (2000). "Calibration and empirical Bayes variable selection." *Biometrika*, 87(4): 731–747. MR1813972. doi: https://doi.org/10.1093/biomet/87.4.731. 913

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 909

George, E. I. and McCulloch, R. E. (1997). "Approaches for Bayesian variable selection." *Statistica Sinica*, 339–373. 909, 922

Greenlaw, K., Szefer, E., Graham, J., Lesperance, M., Nathoo, F. S., and Initiative, A. D. N. (2017). "A Bayesian group sparse multi-task regression model for imaging genetics." *Bioinformatics*, 33(16): 2513–2522. 930

Hans, C., Dobra, A., and West, M. (2007). "Shotgun stochastic search for "large p" regression." *Journal of the American Statistical Association*, 102(478): 507–516. MR2370849. doi: https://doi.org/10.1198/016214507000000121. 911, 920, 921

He, X., Wang, L., and Hong, H. G. (2013). "Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data." *The Annals of Statistics*, 41(1): 342–369. MR3059421. doi: https://doi.org/10.1214/13-AOS1087. 928

Hernández-Lobato, D., Hernández-Lobato, J. M., and Dupont, P. (2013). "Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation." *The Journal of Machine Learning Research*, 14(1): 1891–1945. MR3104499. 911

Huang, J., Breheny, P., and Ma, S. (2012). "A selective review of group selection in high-dimensional models." *Statistical Science*, 27(4). MR3025130. doi: https://doi.org/10.1214/12-STS392. 910, 918, 919

Huang, J., Horowitz, J. L., and Wei, F. (2010). "Variable selection in nonparametric additive models." *The Annals of Statistics*, 38(4): 2282. MR2676890. doi: https://doi.org/10.1214/09-AOS781. 919

Huang, J., Ma, S., Xie, H., and Zhang, C.-H. (2009). "A group bridge approach for variable selection." *Biometrika*, 96(2): 339–355. MR2507147. doi: https://doi.org/10.1093/biomet/asp020. 910

Ishwaran, H. and Rao, J. S. (2005). "Spike and slab variable selection: frequentist and Bayesian strategies." *The Annals of Statistics*, 730–773. MR2163158. doi: https://doi.org/10.1214/009053604000001147. 909

Jeffreys, H. (1961). *Theory of Probability (3rd edition)*. Oxford University Press. MR0187257. 913

Johnson, V. E. and Rossell, D. (2012). "Bayesian model selection in high-dimensional settings." *Journal of the American Statistical Association*, 107(498): 649–660. MR2980074. doi: https://doi.org/10.1080/01621459.2012.682536. 909, 910, 911

Keller, M., Gatti, D., Schueler, K., Rabaglia, M., Stapleton, D., Simecek, P., Vincent, M., Allen, S., Broman, A., Bacher, R., Kendziorski, C., Broman, K., Yandell, B., Churchill, G., and Attie, A. (2018a). "Data from: Genetic drivers of pancreatic islet function." URL https://doi.org/10.5061/dryad.pj105. 928

Keller, M. P., Gatti, D. M., Schueler, K. L., Rabaglia, M. E., Stapleton, D. S., Simecek, P., Vincent, M., Allen, S., Broman, A. T., Bacher, R., Kendziorski, C., Broman, K. W., Yandell, B. S., Churchill, G. A., and Attie, A. D. (2018b). "Genetic Drivers of Pancreatic Islet Function." *Genetics*, 209(1): 335–356. URL http://www.genetics.org/content/209/1/335. 928

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: https://doi.org/10.1214/10-BA607. 911, 913

Liang, F., Liu, C., and Carroll, R. J. (2007). "Stochastic approximation in Monte Carlo computation." *Journal of the American Statistical Association*, 102(477): 305–320. MR2345544. doi: https://doi.org/10.1198/016214506000001202. 921

Liang, F., Song, Q., and Yu, K. (2013). "Bayesian subset modeling for high-dimensional generalized linear models." *Journal of the American Statistical Association*, 108(502): 589–606. MR3174644. doi: https://doi.org/10.1080/01621459.2012.761942. 921

Liquet, B., Mengersen, K., Pettitt, A., and Sutton, M. (2017). "Bayesian variable selection regression of multivariate responses for group data." *Bayesian Analysis*, 12(4): 1039–1067. MR3724978. doi: https://doi.org/10.1214/17-BA1081. 930

Lounici, K., Pontil, M., Van De Geer, S., and Tsybakov, A. B. (2011). "Oracle inequalities and optimal inference under group sparsity." *The Annals of Statistics*, 39(4): 2164–2204. MR2893865. doi: https://doi.org/10.1214/11-AOS896. 919

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. MR2426104. doi: https://doi.org/10.1214/08-EJS200. 910

Narisetty, N. N. and He, X. (2014). "Bayesian variable selection with shrinking and diffusing priors." *The Annals of Statistics*, 42(2): 789–817. MR3210987. doi: https://doi.org/10.1214/14-AOS1207. 909, 910, 911, 914, 923

Obozinski, G., Wainwright, M. J., and Jordan, M. I. (2011). "Support union recovery in high-dimensional multivariate regression." *The Annals of Statistics*, 39(1): 1–47. MR2797839. doi: https://doi.org/10.1214/09-AOS776. 919

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–686. MR2524001. doi: https://doi.org/10.1198/016214508000000337. 913

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*, 881–888. ACM. 911

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). "Sparse additive models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5): 1009–1030. MR2750255. doi: https://doi.org/10.1111/j.1467-9868.2009.00718.x. 919

Ročková, V. and George, E. I. (2014). "EMVS: The EM approach to Bayesian variable selection." *Journal of the American Statistical Association*, 109(506): 828–846. MR3223753. doi: https://doi.org/10.1080/01621459.2013.869223. 909, 913, 923

Ročková, V. and George, E. I. (2016). "Fast Bayesian factor analysis via automatic rotations to sparsity." *Journal of the American Statistical Association*, 111(516): 1608–

1622. MR3601721. doi: https://doi.org/10.1080/01621459.2015.1100620. 913, 923

Ročková, V. and George, E. I. (2018). "The spike-and-slab Lasso." *Journal of the American Statistical Association*, 113(521): 431–444. MR3805383. 909

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 38(5): 2587–2619. MR2722450. doi: https://doi.org/10.1214/10-AOS792. 913

Shang, Z. and Li, P. (2014). "High-dimensional Bayesian inference in nonparametric additive models." *Electronic Journal of Statistics*, 8(2): 2804–2847. MR3299123. doi: https://doi.org/10.1214/14-EJS963. 911, 912, 917, 919

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). "A sparse-group Lasso." *Journal of Computational and Graphical Statistics*, 22(2): 231–245. MR3173712. doi: https://doi.org/10.1080/10618600.2012.681250. 910

Spitzner, D. J. (2019). "Subjective Bayesian testing using calibrated prior probabilities." *Brazilian Journal of Probability and Statistics*, 33(4): 861–893. MR3996320. doi: https://doi.org/10.1214/18-BJPS424. 909, 913

Tibshirani, R. (1996). "Regression shrinkage and selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 267–288. MR1379242. 909, 910

Wang, H. and Leng, C. (2008). "A note on adaptive group Lasso." *Computational Statistics & Data Analysis*, 52(12): 5277–5286. MR2526593. doi: https://doi.org/10.1016/j.csda.2008.05.006. 910

Wang, L., Li, H., and Huang, J. Z. (2008). "Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements." *Journal of the American Statistical Association*, 103(484): 1556–1569. MR2504204. doi: https://doi.org/10.1198/016214508000000788. 910

Wei, F. and Huang, J. (2010). "Consistent group selection in high-dimensional linear regression." *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 16(4): 1369. MR2759183. doi: https://doi.org/10.3150/10-BEJ252. 917

Xu, X. and Ghosh, M. (2015). "Bayesian variable selection and estimation for group Lasso." *Bayesian Analysis*, 10(4): 909–936. MR3432244. doi: https://doi.org/10.1214/14-BA929. 911, 912, 913, 922

Yang, X. and Narisetty, N. N. (2019). "Consistent Group Selection with Bayesian High Dimensional Modeling: Supplementary Material." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1178SUPP. 917, 919, 922

Yuan, M. and Lin, Y. (2006). "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1): 49–67. MR2212574. doi: https://doi.org/10.1111/j.1467-9868.2005.00532.x. 910

Zellner, A. (1962). "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias." *Journal of the American Statistical Association*, 57(298): 348–368. MR0139235. 919

Zellner, A. (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions." *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti*, 233–243. 913

Zhang, C.-H. (2010). "Nearly unbiased variable selection under minimax concave penalty." *The Annals of Statistics*, 38(2): 894–942. MR2604701. doi: https://doi.org/10.1214/09-AOS729. 909

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)*, 63(4): 595–620. MR3258055. doi: https://doi.org/10.1111/rssc.12053. 922

Zhao, P. and Yu, B. (2006). "On model selection consistency of Lasso." *Journal of Machine Learning Research*, 7(Nov): 2541–2563. MR2274449. 910, 917, 919

Zou, H. (2006). "The adaptive Lasso and its oracle properties." *Journal of the American Statistical Association*, 101(476): 1418–1429. MR2279469. doi: https://doi.org/10.1198/016214506000000735. 910

Acknowledgments

We thank the Editor-in-Chief, an Editor, an Associate Editor, and a Reviewer for their encouraging and insightful comments on a previous version of the paper. The research is partially supported by the NSF Award DMS-1811768.