Bayesian Group Bridge for Bi-level Variable Selection

Himel Mallick^{a,b,*}, Nengjun Yi^c

^a*Department of Biostatistics, Harvard T. H. Chan School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA*

^b*Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA*

^c*Section on Statistical Genetics, Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, 1665 University Boulevard, Birmingham, AL 35294, USA*

Abstract

A Bayesian bi-level variable selection method (BAGB: Bayesian Analysis of Group Bridge) is developed for regularized regression and classification. This new development is motivated by grouped data, where generic variables can be divided into multiple groups, with variables in the same group being mechanistically related or statistically correlated. As an alternative to frequentist group variable selection methods, BAGB incorporates structural information among predictors through a group-wise shrinkage prior. Posterior computation proceeds via an efficient MCMC algorithm. In addition to the usual ease-of-interpretation of hierarchical linear models, the Bayesian formulation produces valid standard errors, a feature that is notably absent in the frequentist framework. Empirical evidence of the attractiveness of the method is illustrated by extensive Monte Carlo simulations and real data analysis. Finally, several extensions of this new approach are presented, providing a unified framework for bi-level variable selection in general models with flexible penalties.

Keywords: Bayesian Regularization, Bayesian Variable Selection, Bi-level Variable Selection, Group Bridge, Group Variable Selection, MCMC

*Corresponding author

Email address: hmallick@hsph.harvard.edu (Himel Mallick)

1. Introduction

In many scientific applications, covariates are naturally grouped. One such example arises in association studies, where genetic variants can be divided into multiple groups, with variants in the same group being biologically related or statistically correlated. When the features to be associated possess an inherent grouping structure, traditional variable selection methods are known to produce nonsensical results (Breheny and Huang, 2009; Breheny, 2015), a phenomenon that becomes particularly apparent when categorical variables with multiple levels are present (e.g. in an ANOVA model) or nonparametric components are approximated by a linear combination of some basis functions (e.g. in an additive model) (Yuan and Lin, 2006; Qian et al., 2016). Any standard variable selection method thus offers at best a partial solution to the problem, and the longstanding interest in sparse group selection in many fields including genetics and ecology has led to a fair amount of recent work extending these individual variable selection approaches to grouped predictors.

Specifically, several methods in the context of linear models have been proposed; virtually all enforce within-group sparsity and identify predictive features at both group and individual levels (hereafter referred to as ‘bi-level variable selection’). The only technique that cannot do bi-level variable selection is the group LASSO (Yuan and Lin, 2006), which performs an ‘all-in-all-out’ group selection using an L_2 penalization. Huang et al. (2009) noted that the resulting estimates are not guaranteed to be consistent and proposed group bridge to achieve ‘oracle group selection’ consistency using a generalized L_1 regularization. Simon et al. (2013) similarly proposed sparse group LASSO (Tibshirani, 1996) by using a convex combination of LASSO and group LASSO penalties on the coefficients. While several other methods have been proposed (Breheny, 2015; Geng et al., 2015), these three have remained the most popular, being thoroughly studied and routinely used in various benchmarks (Breheny and Huang, 2009; Huang et al., 2012).

Among the pros and cons of these individual frequentist approaches, they all

lack a systematic approach to inference and are not flexible in simultaneously tuning multiple shrinkage parameters (Leng et al., 2014). Huang et al. (2009) proposed an inferential procedure based on a ridge regression approximation of the covariance matrix. However, this approximation leads to standard error
35 estimates of 0 when the corresponding estimated coefficient is 0. An alternative approach to obtaining standard error estimates is the bootstrap, which is known to have serious numerical difficulties when the true parameter values are either exactly 0 or close to 0; in such cases, bootstrap sampling introduces a bias that does not vanish asymptotically (Kyung et al., 2010; Knight and Fu, 2000). This
40 means that investigators typically must use the resulting estimate without a quantification of its uncertainty, and they must further compromise the desired precision of the tuning parameters.

We address both these issues by providing a flexible, fully Bayesian approach to perform simultaneous variable selection and group identification. It
45 provides coherent inference through the associated posterior and estimates the tuning parameters as an automatic byproduct of the Markov chain Monte Carlo (MCMC) procedure. We have implemented this method as BAGB (Bayesian Analysis of Group Bridge); it estimates the model parameters by using a structured shrinkage prior which enforces group-wise sparsity on the coefficients. In
50 this study, we also introduce a new data-augmentation strategy based on a novel scale mixture of multivariate uniform (SMU) hierarchical representation. The SMU distribution allows us to implement a computationally efficient Gibbs sampling algorithm, which suggests interesting generalizations in a unified modeling framework. Taken together, the major contributions of this paper are as
55 follows: (i) introduction of a group sparsity-promoting prior in Bayesian regularization, (ii) the SMU representation (defined in Proposition 2), which is central to the proposed methodology, (iii) a simple Gibbs sampler with tractable full conditional posterior distributions, (iv) data-adaptive estimation of the tuning parameters, and (v) extended characterization of Bayesian sparse group selec-
60 tion for modeling a variety of outcomes.

2. The Model

2.1. Background and Notation

Consider the normal linear regression model

$$\mathbf{y} = \mu \mathbf{1}_n + X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (1)$$

where $\mathbf{y}$ is the $n \times 1$ vector of responses, X is the $n \times p$ matrix of regressors, $\boldsymbol{\beta}$ is the $p \times 1$ vector of coefficients to be estimated, and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of independent and identically distributed normal errors with mean 0 and variance σ^2 . We assume that there is an intrinsic grouping structure among the predictors, i.e. $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K)'$ (K is the number of groups consisting of m_k elements each with $\sum_{k=1}^K m_k = p$), where $\boldsymbol{\beta}_k$ is the group-specific coefficient vector ($k = 1, \dots, K$). Without loss of generality, we assume that $\mathbf{y}$ and X are centered so that μ can be omitted from the model. The group bridge estimator (Huang et al., 2009) results from the following regularization problem

$$(\mathbf{y} - X\boldsymbol{\beta})'(\mathbf{y} - X\boldsymbol{\beta}) + \lambda \sum_{k=1}^K w_k \|\boldsymbol{\beta}_k\|_1^\alpha, \quad (2)$$

where $\|\boldsymbol{\beta}_k\|_1$ is the L_1 norm of $\boldsymbol{\beta}_k$, $\lambda > 0$ is the tuning parameter which controls the degree of penalization, $\alpha > 0$ is the concavity parameter which controls the shape of the penalty function, w_k 's ($w_k > 0$, $k = 1, \dots, K$) are known weights depending on some preliminary estimates, and for any $\mathbf{z} = (z_1, \dots, z_q)' \in \mathbb{R}^q$ (the q -dimensional Euclidean space), $\|\mathbf{z}\|_s$ is defined as $\|\mathbf{z}\|_s = (|z_1|^s + \dots + |z_q|^s)^{\frac{1}{s}}$, $s > 0$. Contrary to the original formulation (2), we consider a multi-parameter version that gives rise to the following regularization problem

$$(\mathbf{y} - X\boldsymbol{\beta})'(\mathbf{y} - X\boldsymbol{\beta}) + \sum_{k=1}^K \lambda_k \|\boldsymbol{\beta}_k\|_1^\alpha. \quad (3)$$

Introducing multiple parameters provides a natural way to pool the information among variables within a group, while also accommodating differential shrinkage through group-specific tuning parameters. Specifically, this allows us to propose

a Bayesian approach that incorporates the structural information among predictors by means of a suitable prior distribution as follows (hereafter referred to as ‘group bridge prior’)

$$\pi(\boldsymbol{\beta}|\alpha, \lambda_1, \dots, \lambda_K) \propto \prod_{k=1}^K \exp(-\lambda_k \|\boldsymbol{\beta}_k\|_1^\alpha).$$

As opposed to minimizing (3), BAGB solves the group bridge problem by using a Gibbs sampler that involves constructing a Markov chain having the joint posterior for $\boldsymbol{\beta}$ as its stationary distribution. Unlike most existing sparsity-inducing priors, the BAGB joint posterior distribution is multimodal. Therefore, both classical and Bayesian approaches to group bridge estimation must encounter practical difficulties related to multimodality. However, as noted in Polson et al. (2014), multimodality is one of the strongest arguments for pursuing a Bayesian approach, as we are able to handle model uncertainty in a systematic manner by fully exploring the posterior distribution, unlike a frequentist approach which forces one to summarize by means of a single point estimate. As it will be clear from the simulation studies and real data analyses, the proposed Gibbs sampler seems to be very effective at exploring the whole parameter space, unlike classical group bridge estimator, which suffers from a number of computational drawbacks due to non-convexity that limits its applicability in high dimensional data analysis (Breheny, 2015).

2.2. Prior Formulation

With a slight abuse of notation, we drop the dependence on the group index k , and consider a q -dimensional group bridge prior as follows

$$\pi(\boldsymbol{\beta}) = C(\lambda, \alpha) \exp(-\lambda \|\boldsymbol{\beta}\|_1^\alpha),$$

where $C(\lambda, \alpha)$ is the normalizing constant. The normalizing constant $C(\lambda, \alpha)$ can be explicitly calculated as $C(\lambda, \alpha) = \frac{\lambda^{\frac{q}{\alpha}} \Gamma(q+1)}{2^q \Gamma(\frac{q}{\alpha} + 1)}$ (detailed in **Appendix A**). Hence, the normalized group bridge prior is

$$\pi(\boldsymbol{\beta}) = \frac{\lambda^{\frac{q}{\alpha}} \Gamma(q+1)}{2^q \Gamma(\frac{q}{\alpha} + 1)} \exp(-\lambda \|\boldsymbol{\beta}\|_1^\alpha). \quad (4)$$

This proves that the group bridge prior is a proper prior and in any Bayesian
80 analysis with the group bridge prior, the propriety of the posterior will be re-
tained. The posterior mode estimate under a normal likelihood and the group
bridge prior (4) matches exactly with the solution of (3). The group bridge prior
can be construed as a multivariate extension of the Bayesian bridge prior (Pol-
son et al., 2014); while Polson et al. (2014) considers independent and identical
85 generalized Gaussian priors on the coefficients, we assign independent multivari-
ate group bridge priors on each group of coefficients to take into account the
structural blocking information among predictors and also to mimic the prop-
erty of the group bridge penalty. Although the above prior formulation may
appear conceptually modest, there are considerable challenges to capitalizing
90 on the BAGB prior distribution, underlined primarily by the fact that a scale
mixture of multivariate normal (SMN) representation is not explicitly available,
due to the lack of a closed-form expression of the mixing density function. The
BAGB posterior inference is further impeded by the multimodality of the re-
sulting posterior and presence of multiple shrinkage parameters. To overcome
95 these challenges, we appeal to the SMU representation; with the SMU repre-
sentation, we are able to formulate a hierarchical model, which alleviates the
problem of estimating many parameters. The resulting Gibbs sampler seems to
be very effective at exploring the multimodal parameter space, ensuring enriched
posterior inference.

100 2.3. Properties

Proposition 1: *Let, $\beta \sim \pi(\beta)$, the group bridge prior defined in (4). Then, β has the following stochastic representation*

$$\beta \stackrel{D}{=} RU,$$

where $U = \frac{\beta}{\|\beta\|_1}$ is a random vector uniformly distributed on the unit L_1 norm
sphere in $\mathbb{R}^q$, and $R = \|\beta\|_1$ is an absolutely continuous non-negative random
variable independent of U with density function

$$g(r) = \frac{qr^{q-1}\lambda^{\frac{q}{\alpha}}e^{-\lambda r^\alpha}}{\Gamma(\frac{q}{\alpha} + 1)}, \quad r > 0,$$

which corresponds to a generalized Gamma distribution.

Proof: See **Appendix B**.

Based on the above proposition, we provide a straightforward algorithm on
105 how to sample directly from the group bridge prior (4) (see **Appendix C**),
allowing us to investigate the shape of the group bridge prior for varying α
(Figures 1-2). When $0 < \alpha < 1$, the density function is strongly peaked at zero,
suggesting that priors in this range are suitable for variable selection by shrink-
ing small coefficients to zero while minimally shrinking the large coefficients.
110 Specifically, when $\alpha \leq 1$, the density preserves its strong peak at zero (Figure 1);
when $\alpha > 1$, the sharpness of the density diminishes with increasing α (Figure
2), suggesting that priors in this range may not be appropriate if variable selec-
tion is desired. Based on this, we consider $0 < \alpha < 1$ throughout this article,
and develop our Gibbs sampler based on a novel mixture representation of the
115 group bridge prior, which we describe next.

2.4. Scale Mixture of Multivariate Uniform Representation

Proposition 2: *The group bridge prior (4) can be written as a scale mix-
ture of multivariate uniform (SMU) distribution, the mixing density being a
particular Gamma distribution, i.e. $\beta|u \sim \text{Multivariate Uniform}(A)$, where
120 $A = \{\beta \in \mathbb{R}^q : \|\beta\|_1^\alpha < u\}$, $u > 0$ and $u \sim \text{Gamma}(\frac{q}{\alpha} + 1, \lambda)$.*

Proof: Detailed in **Appendix D**.

(a) $\alpha = 0.05$

(b) $\alpha = 0.2$

(c) $\alpha = 0.5$

(d) $\alpha = 0.8$

Figure 1: *Density plots of four bivariate group bridge priors corresponding to $\alpha = \{0.05, 0.2, 0.5, 0.8\}$ and $\lambda = 1$, based on 10,000 samples drawn from the group bridge prior distribution.*

(a) $\alpha = 1$

(b) $\alpha = 2$

(c) $\alpha = 5$

(d) $\alpha = 10$

Figure 2: *Density plots of four bivariate group bridge priors corresponding to $\alpha = \{1, 2, 5, 10\}$ and $\lambda = 1$, based on 10,000 samples drawn from the group bridge prior distribution.*

2.5. Hierarchical Representation

Based on Proposition 2, we formulate our hierarchical representation as follows

$$\begin{aligned}
\mathbf{y}|X, \boldsymbol{\beta}, \sigma^2 &\sim N_n(X\boldsymbol{\beta}, \sigma^2 I_n), \\
\boldsymbol{\beta}_k|u_k, \alpha &\sim \text{Multivariate Uniform}(\Omega_k) \text{ independently for } k = 1, \dots, K, \\
\text{where } \Omega_k &= \{\boldsymbol{\beta}_k \in \mathbb{R}^{m_k} : \|\boldsymbol{\beta}_k\|_1^\alpha < u_k\}, \\
u_1, \dots, u_K|\lambda_1, \dots, \lambda_K, \alpha &\sim \prod_{k=1}^K \text{Gamma}(\frac{m_k}{\alpha} + 1, \lambda_k), \\
\sigma^2 &\sim \pi(\sigma^2),
\end{aligned} \tag{5}$$

where a non-informative scale-invariant marginal prior is assumed on σ^2 , i.e.
 $\pi(\sigma^2) \propto 1/\sigma^2$.

2.6. Full Conditional Distributions

Introduction of $\mathbf{u} = (u_1, u_2, \dots, u_K)'$ enables us to derive the full conditional distributions as follows

$$\boldsymbol{\beta}|\mathbf{y}, X, \mathbf{u}, \sigma^2, \lambda_1, \dots, \lambda_K, \alpha \sim N_p(\hat{\boldsymbol{\beta}}_{\text{OLS}}, \sigma^2(X'X)^{-1}) \prod_{k=1}^K I\{\|\boldsymbol{\beta}_k\|_1^\alpha < u_k\}, \tag{6}$$

$$\mathbf{u}|\mathbf{y}, X, \boldsymbol{\beta}, \lambda_1, \dots, \lambda_K, \alpha \sim \prod_{k=1}^K \text{Exponential}(\lambda_k) I\{u_k > \|\boldsymbol{\beta}_k\|_1^\alpha\}, \tag{7}$$

$$\sigma^2|\mathbf{y}, X, \boldsymbol{\beta} \sim \text{Inverse Gamma} \left(\frac{n-1}{2}, \frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})'(\mathbf{y} - X\boldsymbol{\beta}) \right), \tag{8}$$

where $I(\cdot)$ denotes an indicator function. The proof involves only simple algebra, and is therefore omitted.

2.7. Sampling Coefficients and the Latent Variables

(6), (7) and (8) lead us to a Gibbs sampler that starts at initial guesses for $\boldsymbol{\beta}$ and σ^2 and iterates the following steps:

1. Generate u_k from the left-truncated exponential distribution $\text{Exponential}(\lambda_k) I\{u_k > \|\boldsymbol{\beta}_k\|_1^\alpha\}$ using inversion method, which can be done as follows:

- 135 (a) Generate $u_k^* \sim \text{Exponential}(\lambda_k)$,
 (b) $u_k = u_k^* + \|\beta_k\|_1^\alpha$, $k = 1, \dots, K$.
2. Generate β from a truncated multivariate normal distribution proportional to the posterior distribution of β (described in **Appendix E**).
 3. Generate σ^2 from the Inverse Gamma distribution proportional to the posterior distribution of σ^2 described in (8).

140 2.8. Sampling Hyperparameters

To update the tuning parameters $(\lambda_1, \dots, \lambda_K)$, we work directly with the group bridge prior, marginalizing out the latent variables u_k 's. From (5), we observe that the posterior for λ_k , given β , is conditionally independent of $\mathbf{y}$. Therefore, if λ_k has a $\text{Gamma}(a, b)$ prior, we can update the tuning parameters by generating samples from their joint conditional posterior distribution

$$\lambda_1, \dots, \lambda_K | \mathbf{y}, X, \beta, \alpha \propto \prod_{k=1}^K \lambda_k^{a + \frac{m_k}{\alpha} - 1} \exp\{-\lambda_k(b + \|\beta_k\|_1^\alpha)\}.$$

The concavity parameter α is usually prefixed beforehand. However, it can also be estimated by assigning a suitable prior $\pi(\alpha)$. Since $0 < \alpha < 1$, a natural choice for prior on α is the Beta distribution. Assuming a Beta (c, d) prior on α , the posterior distribution of α is given by

$$\pi(\alpha | \mathbf{y}, X, \beta, \lambda_1, \dots, \lambda_K) \propto \alpha^{c-1} (1-\alpha)^{d-1} \prod_{k=1}^K \frac{\lambda_k^{m_k/\alpha}}{\Gamma(\frac{m_k}{\alpha} + 1)} \exp\{-\sum_{k=1}^K \lambda_k \|\beta_k\|_1^\alpha\}. \quad (9)$$

To sample from (9), one can use a random walk Metropolis-Hastings algorithm with the prior Beta (c, d) as the transition distribution and acceptance probability

$$w = \min\left\{1, \frac{\pi(\alpha' | \mathbf{y}, X, \beta, \lambda_1, \dots, \lambda_K)}{\pi(\alpha | \mathbf{y}, X, \beta, \lambda_1, \dots, \lambda_K)}\right\},$$

where α is the present state and α' is a candidate draw from the prior distribution Beta (c, d) . In our experience, a Beta $(10, 10)$ prior leads to better rate of convergence and mixing as compared to the uniform prior, which is therefore recommended as the default prior for real data analyses.

Table 1: *Description of variables in the birth weight dataset.*

Group	Variables	Description
Group 1	<i>age1, age2, age3</i>	Orthogonal polynomials of first, second, and third degree representing mother's age in years
Group 2	<i>lwt1, lwt2, lwt3</i>	Orthogonal polynomials of first, second, and third degree representing mother's weight in pounds at last menstrual period
Group 3	<i>white, black</i>	Indicator functions for mother's race ("other" is the reference group)
Group 4	<i>smoke</i>	Smoking status during pregnancy
Group 5	<i>ptl1, ptl2m</i>	Indicator functions for 1 and ≥ 2 previous premature labors ("no previous premature labors" is the reference category)
Group 6	<i>ht</i>	History of hypertension
Group 7	<i>ui</i>	Presence of uterine irritability
Group 8	<i>ftv1, ftv2, ftv3m</i>	Indicator functions for 1, 2, and ≥ 3 physician visits during the first trimester ("no visits" is the reference category)

In this section, we evaluate the statistical properties of BAGB and compare its performance with three competing approaches viz. group LASSO (GL), sparse group LASSO (SGL), and group bridge (GB). We make use of the default parameters available in the **R** software (<https://www.r-project.org/>) implementations of these methods (**grpreg** and **SGL**), none of which assume invariance under within-group orthonormality of the design matrix. All these individual frequentist methods provide an estimate of the degrees of freedom (DF) as their standard outputs, the only exception being SGL, in which, DF is estimated as the number of non-zero coefficients in the model. Regarding the selection of the tuning parameters, Huang et al. (2009) indicated that tuning based on BIC (Schwarz, 1978) in general leads to better variable selection performance as compared to other model selection criteria such as AIC (Akaike, 1974), C_p (Mallows, 1973), and GCV (Golub et al., 1979). Therefore, the associated tuning parameters are selected by BIC for all the frequentist methods under consideration, which is defined as

$$\text{BIC} = \log(\|y - X\hat{\beta}\|_2^2/n) + \log(n) * \text{DF}/n.$$

For BAGB, we set the hyperparameters corresponding to λ_j 's as $a = 1$ and $b = 0.1$, which is relatively flat and results in high posterior probability near the MLE (Kyung et al., 2010). BAGB estimates are posterior means using the post-burn-in samples of the Gibbs sampler. To decide on the burn-in number, we calculate the potential scale reduction factor $\hat{R}$ (Gelman and Rubin, 1992) using the **coda** package in **R**. Once $\hat{R} < 1.1$ for all parameters of interest, we continue to draw the desired number of iterations to obtain samples from the joint posterior distribution. The convergence of the MCMC algorithm is also investigated by trace and ACF plots of the generated samples. In particular, the Markov chain is run for a total of 30,000 iterations, of which 15,000 are discarded as burn-in (in this and the simulation study as well) and the rest are used as draws from the posterior distribution for making inference. This choice

Figure 3: *Beta densities for different hyperparameter values.*

of running parameters appear to work satisfactorily based on the convergence diagnostics. The response is centered and the predictors are normalized to have zero means and unit variances before applying any group regularization method. For variable selection with BAGB, we make use of the Scaled Neighbourhood Criterion (SNC) (with threshold $\tau = 0.5$) (Li and Lin, 2010), which is defined as $SNC = P\{(|\beta| > \sqrt{\text{Var}(\beta|\text{Data})}|\text{Data})\}$.

3.1. The Birth Weight Example

First, we illustrate our proposal by re-examining the birth weight dataset from Hosmer and Lemeshow (1989). We make use of a reparametrized data frame available from the **R** package **grpreg**. The original dataset, collected at Baystate Medical Center, Springfield, Massachusetts, USA in 1986, consists of the birth weights of 189 babies in kilograms (the response variable of interest) and 8 predictors concerning the mother. Among the eight predictors, two are continuous (mother's age in years and mother's weight in pounds at the last

menstrual period) and six are categorical (mother’s race, smoking status during pregnancy, number of previous premature labors, history of hypertension, presence of uterine irritability, and number of physician visits). Since a preliminary
175 analysis suggested that nonlinear effects of both mother’s age and weight might exist, both these effects are modeled by using third-order polynomials (Yuan and Lin, 2006). The reparametrized data frame includes $p = 16$ predictors corresponding to $K = 8$ groups consisting of 3, 3, 2, 1, 2, 1, 1, and 3 variables respectively (Table 1). As in the original analysis (Yuan and Lin, 2006), we
180 randomly select three-quarters of the observations (151 cases) for model fitting and reserve the rest of the data as the test set.

We first investigate the possible susceptibility of the results to the Beta prior by considering several values of the concavity parameter that covers a range of non-informative settings (Figure 1). This sensitivity analysis reveals that
185 BAGB significantly outperforms the other three methods in terms of prediction accuracy regardless of the values of the hyperparameters (Table 2). We repeat the random selection of training and test sets many times and obtain similar results as in Table 2. This analysis further reveals that the posterior estimate of α can be sensitive to the choice of the prior distribution. Following similar
190 arguments in Chakraborty and Guo (2011), we recommend that the mean of the Beta distribution is kept above 0.3 and the variance is kept relatively small.

Next, we compare the prediction accuracy of both group bridge estimators for varying α , which clearly suggests preference for the BAGB model across the values of α (Table 3). We observe that GB solutions do not always co-
195 incide with posterior mode estimates (Figure 2). This can be contributed to BAGB’s ability to account for uncertainty associated with the hyperparameters, which is ignored altogether by frequentist GB. Another intrinsic feature of this diagram is the noticeable multimodality in the joint distribution of the BAGB posterior. Clearly, summarizing β by means of a single point estimate
200 may not be reasonable, which unfortunately is the only available estimate in the frequentist framework. We observe that, for this particular dataset, BAGB performs better for fixed α , possibly because not much variance is explained by

Table 2: *Summary of birth weight data analysis. ‘MSE’ reports average of mean squared errors on the test data. Variables selected by each method as well as an estimate of the concavity parameter (whenever applicable) is also provided.*

Method	Selected Variables	MSE	Estimated α
BAGB - Beta (1, 9)	<i>white, ui</i>	710151.1	0.04
BAGB - Beta (2, 8)	<i>white, ptl1, ui</i>	701676.6	0.04
BAGB - Beta (3, 7)	<i>white, ui</i>	696373.8	0.06
BAGB - Beta (4, 6)	<i>white, ui</i>	697454.7	0.4
BAGB - Beta (5, 5)	<i>white, ptl1, ui</i>	693814.3	0.07
BAGB - Beta (6, 4)	<i>white, ht, ui</i>	664035.1	0.3
BAGB - Beta (7, 3)	<i>white, ptl1, ui</i>	683937.7	0.1
BAGB - Beta (8, 2)	<i>white, ptl1, ui</i>	686022.2	0.1
BAGB - Beta (9, 1)	<i>white, ui</i>	681118.1	0.1
BAGB - Beta (10, 10)	white, ht, ui	662615.2	0.3
BAGB - Uniform	<i>white, ui</i>	699259.5	0.06
GB	<i>lwt1, lwt3, white, ptl1, ht, ui</i>	758162.2	Prefixed at 0.5
GL	<i>lwt1, lwt2, lwt3, white, black, smoke, ptl1, ptl2m, ht, ui</i>	744161.4	Not applicable
SGL	<i>white, ui</i>	868976.3	Not applicable

Table 3: *A comparison of GB and BAGB with respect to MSE’s for varying α based on 38 test set observations.*

Method	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.9$
BAGB	705374.0	672173.1	654960.2	640146.8	625812.3
GB	839626.6	838866.3	758162.2	764094.2	728786.2

Figure 4: *Marginal posterior densities (based on 15,000 posterior samples) of the 16 predictors in the birth weight data. Solid red line: penalized likelihood (group bridge) solution with λ chosen by BIC. Dashed green line: BAGB posterior means. Dotted purple line: BAGB posterior modes. α is prefixed as 0.5.*

introducing the priors, resulting in slightly worse performance with estimated α (Tables 2–3). The mixing of the MCMC chain is reasonably good, confirmed in
205 the trace and ACF plots and convergence diagnostics (data not shown). Thus, in addition to mimicking the bi-level variable selection property of frequentist group variable selection methods, BAGB unravels the full posterior summary of the coefficients, with better uncertainty assessment of the tuning parameters.

3.2. Simulation Experiments

210 Next, we conduct extensive Monte Carlo simulation experiments to benchmark our method as compared to other state-of-the-art group variable selection methods. For a fair comparison, we prefix $\alpha = 0.5$ for both group bridge methods. To evaluate the prediction performance, following Zou (2006), we calculate the median of relative prediction errors (MRPE) based on 100 replications. The
215 relative prediction error (RPE) is defined as $\text{RPE} = E(\hat{y} - X\beta^*)/\sigma^2$, where β^* is the true coefficient vector and $\hat{y}$ is the predicted response. After partitioning each synthetic data into a training set and a test set, models are fitted on the training set and RPE’s are calculated on the test set. For any variable selection procedure, we measure its performance by Sensitivity (Sen) and Specificity
220 (Spe), as defined in Chen and Wang (2013). We also report the percentage of occasions on which correct groups are selected (averaged over 100 replications by various estimators). A group is selected if any of the variables within that group is selected.

We simulate data from the true model $\mathbf{y} = X\beta + \epsilon$, $\epsilon \sim N(\mathbf{0}, \sigma^2 I)$. Follow-
225 ing Huang et al. (2009), we consider two scenarios. For the generating models in Example 1, the number of groups is moderately large, the group sizes are equal and relatively large, and within each group, the coefficients are either all nonzero or all zero. In Example 2, the group sizes vary and there are coefficients equal to zero in a nonzero group. For a relatively complete evaluation, we vary
230 the signal-to-noise ratio (SNR) from low to moderate to high for varying sample sizes $\{n_T, n_P\}$, where SNR is defined as $\text{SNR} = \sqrt{\frac{\text{Var}(X\beta)}{\sigma^2}}$, n_T denotes the size of the training set, and n_P denotes the size of the testing set. Most previous

simulation experiments (including those reported in Huang et al. (2009)) have only considered high SNR values that do not accurately reflect the actual characteristics of real data (Levina and Zhu, 2008). Therefore, with the simulation study attempted in this paper, we are able to investigate the possible sensitivity of the methods in realistic synthetic data, which is an important contribution to the benchmarking of existing and future methods. In particular, we consider four different levels of SNR from low (1) to moderate (3, 5) to high (10) and three sample sizes $\{n_T, n_P\} = \{200, 200\}$ and $\{400, 400\}$. σ^2 is chosen such that the desired level of SNR is achieved.

Simulation 1 (“All-In-All-Out”) In this simulation example, we assume that there are $K = 5$ groups and each group consists of eight covariates, resulting in $p = 40$ predictors. The design matrix X consists of 5 sub-matrices, i.e. $X = (X_1, X_2, X_3, X_4, X_5)$, where $X_j = (x_{8(j-1)+1}, \dots, x_{8(j-1)+8})'$ for any $j \in \{1, \dots, 5\}$. The covariates $x_1, \dots, x_{40}$ are generated as follows:

1. First simulate $R_1, \dots, R_{40}$ independently from the standard normal distribution.
2. Simulate Z_j 's from the multivariate normal distribution with mean 0, variance 1, and pairwise correlations between x_{j_1} and x_{j_2} equal to $0.5^{|j_1 - j_2|}$ $\forall j_1 \neq j_2$, for $j_1, j_2 = 1, \dots, 5$.
3. Obtain x_j 's as $x_j = (Z_{g_j} + R_j)/\sqrt{2}$, $j = 1, \dots, 40$, where g_j is the smallest integer greater than $(j-1)/8$ and the x_j 's with the same value of g_j belong to the same group.

The coefficients corresponding to the first two groups are given as $(\beta_1, \dots, \beta_8)' = (0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4)'$, $(\beta_9, \dots, \beta_{16})' = (2, \dots, 2)'$. We set the rest of the coefficients to zero. Thus, the coefficients in each group are either all nonzero or all zero, leading to a “All-In-All-Out” scenario.

Simulation 2 (“Not-All-In-All-Out”) In this example, we assume that the coefficients in a group can be either all zero, all nonzero or partly zero, lead-

Table 4: *Results from Simulation 1. MRPE & Correct Selection (percentage of occasions on which correct groups are selected, averaged over 100 replications) of various estimators are reported. The numbers in parentheses are the corresponding standard errors of the RPE's based on 500 bootstrap resampling.*

$n_T = n_P = 200, \text{SNR} = 1$				
Method	Sen	Spe	MRPE	% Correct Selection
BAGB	0.175	0.921	0.009 _(0.019)	11
GB	0.241	0.881	0.012 _(0.013)	7
GL	0.535	0.663	0.049 _(0.011)	27
SGL	0.076	0.965	0.068 _(0.014)	13
$n_T = n_P = 200, \text{SNR} = 3$				
BAGB	0.352	0.903	0.011 _(0.020)	41
GB	0.352	0.937	0.036 _(0.014)	21
GL	0.880	0.493	0.088 _(0.015)	77
SGL	0.376	0.837	0.076 _(0.022)	72
$n_T = n_P = 200, \text{SNR} = 5$				
BAGB	0.478	0.892	0.026 _(0.021)	64
GB	0.451	0.948	0.049 _(0.017)	35
GL	0.955	0.443	0.112 _(0.018)	91
SGL	0.538	0.768	0.056 _(0.022)	94
$n_T = n_P = 200, \text{SNR} = 10$				
BAGB	0.682	0.889	0.043 _(0.020)	88
GB	0.584	0.958	0.081 _(0.018)	56
GL	1.000	0.393	0.128 _(0.016)	100
SGL	0.698	0.743	0.051 _(0.017)	100
$n_T = n_P = 400, \text{SNR} = 1$				
BAGB	0.237	0.932	0.016 _(0.008)	20
GB	0.292	0.934	0.023 _(0.009)	13
GL	0.760	0.553	0.047 _(0.009)	53
SGL	0.172	0.915	0.067 _(0.014)	31
$n_T = n_P = 400, \text{SNR} = 3$				
BAGB	0.505	0.912	0.023 _(0.009)	67
GB	0.482	0.962	0.034 _(0.011)	32
GL	0.980	0.417	0.064 _(0.010)	96
SGL	0.584	0.770	0.039 _(0.008)	96
$n_T = n_P = 400, \text{SNR} = 5$				
BAGB	0.669	0.910	0.029 _(0.009)	92
GB	0.586	0.982	0.053 _(0.008)	52
GL	0.995	0.387	0.072 _(0.010)	99
SGL	0.698	0.741	0.045 _(0.008)	100
$n_T = n_P = 400, \text{SNR} = 10$				
BAGB	0.848	0.910	0.036 _(0.010)	100
GB	0.795	0.998	0.053 _(0.009)	93
GL	1.000	0.320	0.085 _(0.011)	100
SGL	0.841	0.732	0.046 _(0.008)	100

Table 5: *Results from Simulation 2. MRPE & Correct Selection (percentage of occasions on which correct groups are selected, averaged over 100 replications) of various estimators are reported. The numbers in parentheses are the corresponding standard errors of the RPE's based on 500 bootstrap resampling.*

$n_T = n_P = 200, \text{SNR} = 1$				
Method	Sen	Spe	MRPE	% Correct Selection
BAGB	0.916	0.732	0.137 _(0.015)	93
GB	0.720	0.925	0.225 _(0.021)	29
GL	0.982	0.608	0.226 _(0.018)	88
SGL	0.507	0.920	0.363 _(0.033)	27
$n_T = n_P = 200, \text{SNR} = 3$				
BAGB	0.988	0.674	0.146 _(0.015)	100
GB	0.978	0.851	0.185 _(0.014)	99
GL	1.000	0.454	0.212 _(0.021)	100
SGL	0.878	0.903	0.568 _(0.016)	100
$n_T = n_P = 200, \text{SNR} = 5$				
BAGB	0.994	0.653	0.153 _(0.014)	100
GB	0.993	0.839	0.192 _(0.014)	100
GL	1.000	0.379	0.229 _(0.022)	100
SGL	0.903	0.938	0.870 _(0.026)	100
$n_T = n_P = 200, \text{SNR} = 10$				
BAGB	0.999	0.630	0.175 _(0.018)	100
GB	0.999	0.830	0.190 _(0.021)	100
GL	1.000	0.282	0.260 _(0.029)	100
SGL	0.915	0.962	1.641 _(0.049)	100
$n_T = n_P = 400, \text{SNR} = 1$				
BAGB	0.977	0.704	0.078 _(0.010)	100
GB	0.928	0.895	0.115 _(0.015)	77
GL	1.000	0.527	0.114 _(0.011)	100
SGL	0.858	0.870	0.192 _(0.014)	100
$n_T = n_P = 400, \text{SNR} = 3$				
BAGB	0.996	0.658	0.082 _(0.011)	100
GB	0.994	0.850	0.094 _(0.012)	100
GL	1.000	0.408	0.124 _(0.010)	100
SGL	0.919	0.963	0.451 _(0.013)	100
$n_T = n_P = 400, \text{SNR} = 5$				
BAGB	0.999	0.639	0.091 _(0.011)	100
GB	0.999	0.841	0.096 _(0.014)	100
GL	1.000	0.291	0.128 _(0.011)	100
SGL	0.929	0.986	0.705 _(0.016)	100
$n_T = n_P = 400, \text{SNR} = 10$				
BAGB	1.000	0.618	0.109 _(0.009)	100
GB	1.000	0.837	0.109 _(0.011)	100
GL	1.000	0.299	0.148 _(0.011)	100
SGL	0.934	0.994	1.360 _(0.033)	100

ing to a “Not-All-In-All-Out” scenario. We also assume that the group sizes differ across groups. There are $K = 6$ groups made up of three groups each of size 10 and three groups each of size 4, resulting in $p = 42$ predictors. The design matrix X consists of 6 sub-matrices, i.e. $X = (X_1, X_2, X_3, X_4, X_5, X_6)$, where $X_j = (x_{10(j-1)+1}, \dots, x_{10(j-1)+10})'$ for any $j \in \{1, 2, 3\}$ and $X_j = (x_{4(j-4)+31}, \dots, x_{4(j-4)+34})'$ for any $j \in \{4, 5, 6\}$. The covariates $x_1, \dots, x_{42}$ are generated as follows:

1. First simulate $R_1, \dots, R_{42}$ independently from the standard normal distribution.
2. Simulate Z_j 's from the multivariate normal distribution with mean 0, variance 1, and pairwise correlations between x_{j_1} and x_{j_2} equal to $0.5^{|j_1-j_2|}$ $\forall j_1 \neq j_2$, for $j_1, j_2 = 1, \dots, 6$.
3. For $j = 1, \dots, 30$, let g_j be the largest integer less than $j/10 + 1$, and for $j = 31, \dots, 42$, let g_j be the largest integer less than $(j - 30)/4 + 1$. Then we obtain x_j 's as $x_j = (Z_{g_j} + R_j)/\sqrt{2}$, $j = 1, \dots, 42$.

The coefficients are given as follows

$$(\beta_1, \dots, \beta_{10})' = (0.5, -2, 0.5, 2, -1, 1, 2, -1.5, 2, -2)',$$

$$(\beta_{11}, \dots, \beta_{20})' = (-1.5, 2, 1, -2, 1.5, 0, \dots, 0)',$$

$$(\beta_{21}, \dots, \beta_{30})' = (0, \dots, 0)', \quad (\beta_{31}, \dots, \beta_{34})' = (2, -2, 1, 1.5)',$$

$$(\beta_{35}, \dots, \beta_{38})' = (-1.5, 1.5, 0, 0)', \quad (\beta_{39}, \dots, \beta_{42})' = (0, \dots, 0)'.$$

The operating characteristics measuring the variable selection and prediction performance of various methods are summarized in Tables 3-4. First, in low SNR (high noise) settings, BAGB vastly outperforms existing methods, regardless of the sample size and sparsity structure of the data generating model. Even for moderate to high SNRs, it consistently produces the smallest prediction error across all simulation examples. This is due to the ability of BAGB to account for uncertainty in the tuning parameters in a systematic fashion that leads to the better prediction performance of the method. Second, none of the methods uniformly dominate the others in variable selection performance.

In the presence of within-group sparsity (Simulation 2), BAGB has higher or similar sensitivity as compared to both GB and SGL (lower specificity), which indicates that as far as bi-level variable selection is concerned, BAGB is not as ‘aggressive’ as the frequentist methods in excluding predictors. This holds true even in the absence of within-group sparsity (Simulation 1), when the corresponding SNR is moderate or high, although it maintains good specificity. For low SNR scenarios in Simulation 1, it has slightly worse sensitivity than frequentist GB estimator, although it maintains better specificity. Third, consistent with previous studies (Huang et al., 2009), group LASSO has the lowest specificity in all the simulation examples, regardless of the sample size, which is expected as it tends to select all variables in an unimportant group. However, it has the highest sensitivity across all examples. Fourth, in terms of percentage of correct groups selected, BAGB always perform better than frequentist GB estimator in Simulation 1, although outperformed by both GL and SGL, especially in low and moderate SNR scenarios. All the methods have similar performance in selecting correct groups in Simulation 2. Lastly, variable selection performance of BAGB depends on the threshold parameter τ . In practice, using a higher threshold value τ would result in a higher sensitivity but a lower specificity. Therefore, depending on the importance of sensitivity and specificity of a particular study, the researcher may choose the appropriate threshold.

Additional Simulations with Small Samples Sizes and Categorical

Predictors To investigate the impact of small sample sizes and categorical covariates on the performance of the proposed method, as suggested by an anonymous reviewer, we conduct additional simulation experiments. In particular, we repeat the above-mentioned simulation studies (Simulation 1 and Simulation 2) with $\{n_T, n_P\} = \{100, 100\}$ and $\text{SNR} = \{1, 3, 5\}$. In addition, we carry out another simulation study with categorical covariates (Simulation 3), where we generate 15 latent variables $Z_1, Z_2, \dots, Z_{15}$ from a multivariate normal distribution with mean 0, variance 1, and pairwise correlations between Z_i and Z_j equal to $0.5^{|i-j|} \forall i \neq j$ (Park and Yoon, 2011). Each Z_i is trichotomized as

Table 6: *Results from Simulation 3. MRPE and % Correct Selection (percentage of occasions on which correct groups are selected, averaged over 100 replications) of various estimators are reported. The numbers in parentheses are the corresponding standard errors of the RPE's based on 500 bootstrap resampling.*

Simulation 1, $n_T = n_P = 100$, SNR = 1				
Method	Sen	Spe	MRPE	% Correct Selection
BAGB	0.206	0.870	0.088 _(0.023)	20
GB	0.304	0.754	0.114 _(0.013)	29
GL	0.350	0.793	0.111 _(0.022)	16
SGL	0.064	0.970	0.103 _(0.028)	12
Simulation 1, $n_T = n_P = 100$, SNR = 3				
BAGB	0.322	0.854	0.107 _(0.022)	36
GB	0.238	0.905	0.107 _(0.018)	13
GL	0.745	0.557	0.175 _(0.028)	54
SGL	0.192	0.908	0.164 _(0.027)	44
Simulation 1, $n_T = n_P = 100$, SNR = 5				
BAGB	0.414	0.843	0.120 _(0.021)	54
GB	0.311	0.921	0.120 _(0.015)	23
GL	0.865	0.483	0.203 _(0.027)	73
SGL	0.359	0.845	0.120 _(0.023)	77
Simulation 2, $n_T = n_P = 100$, SNR = 1				
BAGB	0.821	0.653	0.384 _(0.024)	72
GB	0.446	0.934	0.470 _(0.022)	7
GL	0.856	0.688	0.467 _(0.032)	41
SGL	0.220	0.957	0.651 _(0.034)	16
Simulation 2, $n_T = n_P = 100$, SNR = 3				
BAGB	0.961	0.586	0.498 _(0.034)	96
GB	0.835	0.850	0.583 _(0.042)	60
GL	0.993	0.479	0.579 _(0.060)	94
SGL	0.602	0.898	1.059 _(0.079)	70
Simulation 2, $n_T = n_P = 100$, SNR = 5				
BAGB	0.980	0.565	0.540 _(0.032)	100
GB	0.953	0.808	0.554 _(0.039)	91
GL	1.000	0.345	0.598 _(0.051)	100
SGL	0.720	0.887	1.323 _(0.084)	85
Simulation 3, $n_T = n_P = 100$				
BAGB	0.980	0.812	0.060 _(0.019)	99
GB	0.910	0.946	0.060 _(0.008)	92
GL	0.993	0.776	0.080 _(0.013)	18
SGL	0.780	0.780	0.120 _(0.019)	16
Simulation 3, $n_T = n_P = 200$				
BAGB	0.900	0.768	0.110 _(0.01)	89
GB	0.768	0.924	0.110 _(0.01)	63
GL	0.970	0.766	0.110 _(0.01)	13
SGL	0.548	0.945	0.230 _(0.02)	5

0, 1, or 2 given it is smaller than $\Phi^{-1}(\frac{1}{3})$, larger than $\Phi^{-1}(\frac{2}{3})$, or in between. Then y is generated from

$$y = 1.8I(Z_1 = 1) - 1.2I(Z_1 = 0) + I(Z_3 = 1) + 0.5I(Z_3 = 0) + I(Z_5 = 1) + I(Z_5 = 0) + \epsilon,$$

where $I(\cdot)$ is the indicator function and the noise ϵ is normally distributed with variance σ^2 chosen so that the SNR is approximately 1.8. For each run, 100 observations are collected with two sample sizes: $\{n_T, n_P\} = \{100, 100\}$ and
310 $\{n_T, n_P\} = \{200, 200\}$. The results (presented in Table 6) indicates that BAGB performs as well as or better than frequentist methods in prediction accuracy. Here also, no method dominates the others with respect to variable selection performance. In terms of percentage of occasions on which correct groups are selected, BAGB exhibits superior performance in both Simulations 2 and 3,
315 although outperformed by both GL and SGL in Simulation 1. In Table 7, we also report the computing time of our method for each of the simulation studies. We note that the runtime of our algorithm is not directly comparable to that of the frequentist methods, as it depends on many factors unrelated to the theoretical properties of the sampler that can affect the run time significantly
320 (e.g. programming language, compiler options, whether lower level language is used or not, libraries one links to, and system hardware, among others). Beyond that, how one implements each Gibbs steps will impact the time it takes to run. While the method is more computationally intensive than calculating a solution path for a classical penalized regression approach like group bridge,
325 this is the price to be paid for exploring uncertainty in parameter estimates and the computational demands of the method are certainly less than Bayesian approaches which search the model space directly. The results presented here employ the default options and libraries used by the **R** software. However, the computation time of BAGB can be significantly reduced either by using
330 efficient programming of low-level languages like C++ or implementing GPU programming, which we intend to do in future.

Table 7: Average computing time (in minutes) of BAGB based on 100 replications. The numbers in parentheses are the corresponding standard errors.

Simulation 1 ($p = 40$)					
$n_T = n_P = 100$		$n_T = n_P = 200$		$n_T = n_P = 400$	
SNR=1	5.22(0.21)	SNR=1	5.21(0.22)	SNR=1	4.42(0.17)
SNR=3	5.19(0.21)	SNR=3	5.20(0.09)	SNR=3	4.30(0.21)
SNR=5	4.30(0.21)	SNR=5	5.10(0.19)	SNR=5	5.14(0.22)
		SNR=10	4.25(0.21)	SNR=10	5.12(0.18)
Simulation 2 ($p=42$)					
$n_T = n_P = 100$		$n_T = n_P = 200$		$n_T = n_P = 400$	
SNR=1	5.40(0.10)	SNR=1	5.54(0.20)	SNR=1	4.71(0.13)
SNR=3	5.53(0.22)	SNR=3	5.51(0.22)	SNR=3	4.50(0.09)
SNR=5	5.43(0.20)	SNR=5	5.52(0.22)	SNR=5	5.52(0.12)
		SNR=10	4.60(0.16)	SNR=10	5.40(0.17)
Simulation 3 ($p = 30$)					
$n_T = n_P = 100$		$n_T = n_P = 200$			
5.50(0.19)		5.49(0.22)			

4. Extensions

4.1. Extension to General Models

In this section, we briefly discuss how BAGB can be extended to several other models (e.g. GLM, Cox's model, etc.) beyond the linear regression. Our extension is based on the least squares approximation (LSA) by Wang and Leng (2007). Following Wang and Leng (2007), for a general model, the conditional distribution of $\mathbf{y}$ is given by

$$\mathbf{y}|\boldsymbol{\beta} \sim \exp \left\{ -\frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})' \hat{\Sigma}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \right\},$$

where $\tilde{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ and $\hat{\Sigma}^{-1} = \delta^2 L(\boldsymbol{\beta}) / \delta \boldsymbol{\beta}^2$. Thus, we can easily extend our method to several other models by approximating the corresponding likelihood by normal likelihood. Combining the SMU representation of the group bridge prior and the LSA approximation of the general likelihood, the hierarchical

Figure 5: For the birth weight data, posterior mean Bayesian group bridge estimates and corresponding 95% equal-tailed credible intervals based on 15,000 Gibbs samples using a logistic regression model. Overlaid are the corresponding GB, GL, and SGL estimates.

representation of BAGB for general models can be written as

$$\begin{aligned}
\mathbf{y}|\boldsymbol{\beta} &\sim \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})' \hat{\boldsymbol{\Sigma}}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\right\}, \\
\boldsymbol{\beta}_k|u_k, \alpha &\sim \text{Multivariate Uniform}(\Omega_k), \text{ independently for } k = 1(1)K, \\
\text{where } \Omega_k &= \{\boldsymbol{\beta}_k \in \mathbb{R}^{m_k} : \|\boldsymbol{\beta}_k\|_1^\alpha < u_k\}, \\
u_1, \dots, u_K|\lambda_1, \dots, \lambda_K, \alpha &\sim \prod_{k=1}^K \text{Gamma}\left(\frac{m_k}{\alpha} + 1, \lambda_k\right), \\
\lambda_1, \dots, \lambda_K &\sim \prod_{k=1}^K \text{Gamma}(a, b), \\
\alpha &\sim \text{Beta}(c, d).
\end{aligned}$$

As before, an efficient Gibbs sampler can be easily carried out based on the full conditionals, which are easily derived as before (detailed in **Appendix F**).

4.2. Application to the Birthweight Data - Logistic Regression Illustration

As a second example, we re-visit the birth weight data. We devote this example to evaluate the performance of BAGB in general models. The binary response variable of interest is whether the maternal birth weight is less than 2.5 kg ($Y = 1$) or not ($Y = 0$). As before, we randomly select 151 observations for

Table 8: *Summary of the birth weight data analysis results using logistic regression. ‘AUC’ reports the area under the curve (AUC) estimates on the test data. ‘Test Error’ reports the misclassification errors on the test data.*

	BAGB	GB	GL	SGL
Test Error	0.21	0.24	0.26	0.21
AUC	0.78	0.67	0.71	0.64

model fitting and use the rest as the test set to validate the test errors. Table 5 summarizes the comparison of test errors, which reveals that BAGB performs the best and SGL closely follows next. Figure 5 describes the 95% equal-tailed credible intervals for the regression parameters for this data set corresponding to $\alpha = 0.5$. We emphasize that, among the methods depicted here, only BAGB can provide valid standard errors, based on a geometrically ergodic Markov chain. It is interesting to note that all the estimates are inside the credible intervals, which indicates that the resulting conclusion will be similar regardless of which method is used. Hence, the analyses show strong support for the use of the proposed method.

4.3. Composite Group Bridge Penalty

Seetharaman (2013) proposed the composite group bridge estimator as an extension of the original group bridge penalty that results from the following regularization problem

$$(\mathbf{y} - X\boldsymbol{\beta})'(\mathbf{y} - X\boldsymbol{\beta}) + \sum_{k=1}^K \lambda_k \left\{ \sum_{j=1}^{m_k} |\beta_{kj}|^\gamma \right\}^\alpha,$$

where λ_k 's are the tuning parameters ($\lambda_k > 0, k = 1, \dots, K$), and α and γ are the concavity parameters ($0 < \alpha < 1, 0 < \gamma < 1$). As before, the hierarchical representation of the composite group bridge estimator can be formulated as

follows

$$\begin{aligned}
\mathbf{y}|X, \boldsymbol{\beta}, \sigma^2 &\sim N_n(X\boldsymbol{\beta}, \sigma^2 I_n), \\
\boldsymbol{\beta}_k|u_k, \alpha, \gamma &\sim \text{Multivariate Uniform}(\Omega_k), \\
\text{where } \Omega_k &= \{\boldsymbol{\beta}_k \in \mathbb{R}^{m_k} : \sum_{j=1}^{m_k} |\beta_{k_j}|^\gamma < u_k^{\frac{1}{\alpha}}\} \text{ independently for } k = 1, \dots, K, \\
u_1, \dots, u_K|\lambda_1, \dots, \lambda_K, \alpha, \gamma &\sim \prod_{k=1}^K \text{Gamma}(\frac{m_k}{\alpha} + 1, \lambda_k), \\
\sigma^2 &\sim \pi(\sigma^2), \quad \pi(\sigma^2) \propto 1/\sigma^2, \\
\lambda_1, \dots, \lambda_K &\sim \prod_{k=1}^K \text{Gamma}(a, b), \quad \alpha \sim \text{Beta}(c, d), \quad \gamma \sim \text{Beta}(c^*, d^*).
\end{aligned}$$

The full conditional distributions can be derived as before, which are provided in **Appendix G**.

5. Discussion

355 Despite recent methodological advances, there is a surprising dearth of published methods devoted to the Bayesian sparse group selection problem. Here, we describe a Bayesian method, BAGB, for bi-level variable selection in regularized regression and classification. Assuming a group-wise shrinkage prior on the coefficients, BAGB estimates the model parameters using a computation-
360 ally efficient MCMC algorithm. Unlike established frequentist methods that summarize inference using a single point estimate, BAGB yields uncertainty estimates of the coefficients and enables coherent inference through the posterior distribution. Further, by allowing data-adaptive tuning of multiple shrinkage parameters, it provides a novel and natural treatment of the uncertainty associated with the hyperparameters. Realistic benchmark results reveal that BAGB
365 outperforms a variety of published methods in both estimation and prediction. Finally, we have presented a unified framework which is applicable to models above and beyond the linear regression.

Analysis using SMU mixture representation provided a new data-augmentation
370 strategy for fully Bayesian inference with general likelihoods. This mixture formulation is in sharp contrast to previous studies that relied primarily on SMN

hierarchical representations of the associated priors (Kyung et al., 2010). In addition to facilitating efficient computation, this new multivariate representation provided interesting generalizations including several non-trivial variants of the proposed method. This is an important contribution to the characterization of existing and future analogous MCMC algorithms for other sparse Bayesian models, which to the best of our knowledge has not been explored in depth before. In our experience, the resulting Gibbs sampler is efficient with good rates of convergence, opening up new avenues for the potential application of BAGB in various scientific disciplines.

BAGB is similar in essence to Bayesian Group LASSO (Kyung et al., 2010), which considers independent multivariate Laplace priors on the group-specific coefficients. However, there are several differences between these two methods. Similar to its frequentist cousin, Bayesian group LASSO cannot do bi-level variable selection, and therefore, it is expected to perform poorly for problems that are actually sparse at both group and individual levels (Hernández-Lobato et al., 2013). In addition, Bayesian group LASSO assumes invariance under within-group orthonormal transformations (Kyung et al., 2010), which is difficult to justify without making untestable assumptions on the design matrix. Due to these fundamental differences, a comparison with Bayesian group LASSO is not attempted in this paper.

We anticipate several computational and methodological improvements that may further refine BAGB’s performance. While BAGB uses the heuristic SNC criterion (Li and Lin, 2010) for posterior inference, it can be highly sensitive to the choice of the threshold (Li and Pati, 2017). Alternative methods may improve on this, such as the posterior 2-means (Chakraborty and Guo, 2011; Li and Pati, 2017). Although we have developed our method for working with non-overlapping groups, we recognize that multiple features can belong to multiple overlapping groups in many practical problems, for which the current version of BAGB may not be applicable although we suggest it as a future work. Another path for future investigation includes a computationally efficient adaptation of our approach to fast algorithms for rapid estimation and speedy inference. We

thus hope to refine both the scalability and applicability of BAGB in future studies.

405 Acknowledgements

We thank the Associate Editor and the two anonymous reviewers for their helpful comments. This work was supported in part by the research computing resources acquired and managed by University of Alabama at Birmingham IT Research Computing. Any opinions, findings, and conclusions or recommendations
410 tions expressed in this material are those of the authors and do not necessarily reflect the views of the University of Alabama at Birmingham. Himel Mallick was supported in part by the research grants U01 NS041588 and NSF 1158862. Nengjun Yi was supported in part by the research grant NIH 5R01GM069430-08.

415 Appendix A: Calculating the Normalizing Constant

Here we derive the normalizing constant for the group bridge prior.

Proof: The group bridge prior is given by $\pi(\boldsymbol{\beta}) = C(\lambda, \alpha) \exp\{-\lambda \|\boldsymbol{\beta}\|_1^\alpha\}$, where $\boldsymbol{\beta} \in \mathbb{R}^q$, $\lambda > 0$, $\alpha > 0$, and $\mathbb{R}^q$ denotes the q -dimensional Euclidean space. Consider the following transformation

$$\begin{cases} z_i = \frac{\beta_i}{\|\boldsymbol{\beta}\|_1}, & i = 1, \dots, (q-1), \\ r = \|\boldsymbol{\beta}\|_1. \end{cases}$$

The Jacobian of the above transformation is $|J| = r^{q-1}$.

Therefore, the joint pdf of $(r, z_1, z_2, \dots, z_{q-1})$ is given by

$$f(r, z_1, z_2, \dots, z_{q-1}) = C(\lambda, \alpha) \exp\{-\lambda r^\alpha\} r^{q-1},$$

where $-1 < z_i < 1$, $i = 1, \dots, (q-1)$ and $\sum_{i=1}^{q-1} |z_i| < 1$.

Note that,

$$\int_{-1}^1 \cdots \int_{-1}^1 \sum_{i=1}^{q-1} |z_i| < 1 \cdots \int_{-1}^1 dz_1 \cdots dz_{q-1} = \frac{2^q}{\Gamma(q)}.$$

Therefore, we have,

$$1 = C(\lambda, \alpha) \frac{2^q}{\Gamma(q)} \int_0^\infty \exp\{-\lambda r^\alpha\} r^{q-1} dr \implies C(\lambda, \alpha) = \frac{\lambda^{\frac{q}{\alpha}} \Gamma(q+1)}{2^q \Gamma(\frac{q}{\alpha} + 1)}.$$

This completes the proof.

Appendix B: Stochastic Representation

Here we study various properties of the group bridge prior in light of its connection to the multivariate L_1 norm *spherically symmetric distributions* (also known as *simplicially contoured distributions*) (Fang et al., 1989). To this end, we introduce the following notations and lemma.

Definition 1: A random vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)'$ with location parameter $\boldsymbol{\mu}$ and scale parameter $\boldsymbol{\Sigma}$ follows an L_1 norm spherically symmetric distribution if its density function f is of the form $f(\boldsymbol{\beta}) = \psi(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{\beta} - \boldsymbol{\mu})\|_1)$ for some $\psi : \mathbb{R}^+ \rightarrow \mathbb{R}^+$.

Definition 2: A random vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)'$ follows a multivariate uniform distribution in an arbitrary region $A \subset \mathbb{R}^q$ if its density function f is of the form

$$f(\boldsymbol{\beta}) = \begin{cases} \frac{1}{\Omega(A)}, & \text{if } \boldsymbol{\beta} \in A, \\ 0 & \text{otherwise,} \end{cases}$$

where $\Omega(A) = \text{Volume of } A$.

Lemma: If f is of the form $f(\boldsymbol{\beta}) = \psi(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{\beta} - \boldsymbol{\mu})\|_1)$ for some $\psi : \mathbb{R}^+ \rightarrow \mathbb{R}^+$, it can also be expressed as

$$f(\boldsymbol{\beta}) = |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \frac{\Gamma(q)}{2^q} \frac{g(r(\boldsymbol{\beta}))}{r(\boldsymbol{\beta})^{q-1}}, \quad r(\boldsymbol{\beta}) > 0,$$

where $r(\boldsymbol{\beta}) = \|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{\beta} - \boldsymbol{\mu})\|_1$, $g(\cdot)$ is the density of $r(\boldsymbol{\beta})$.

Proof: Without loss of generality, we assume that the random vector $\boldsymbol{\beta}^*$ follows an L_1 norm spherically symmetric distribution with $\boldsymbol{\Sigma} = \mathbf{I}_q$ and $\boldsymbol{\mu} = \mathbf{0}$, where

$\mathbf{I}_q$ is the $q \times q$ identity matrix and $\mathbf{0}$ is the q -dimensional vector with all entries equal to 0. Following the proof of Lemma 1.4 in Fang et al. (1989), we have the following

$$\int_{\mathbb{R}^q} \zeta(\|\boldsymbol{\beta}^*\|_1) d\boldsymbol{\beta}^* = \frac{2^q}{\Gamma(q)} \int_0^\infty \zeta(u) u^{q-1} du,$$

for any non-negative measurable function ζ . Therefore, for any non-negative measurable function η , we define $\zeta = \eta \cdot \psi$ and get the following

$$E(\eta(\|\boldsymbol{\beta}^*\|_1)) = \frac{2^q}{\Gamma(q)} \int_0^\infty \eta(u) \psi(u) u^{q-1} du.$$

Therefore, the density of $r(\boldsymbol{\beta}^*)$ is

$$g(r) = \frac{2^q}{\Gamma(q)} \psi(r) r^{q-1} \implies \psi(r(\boldsymbol{\beta}^*)) = \frac{\Gamma(q)}{2^q} \frac{g(r(\boldsymbol{\beta}^*))}{r(\boldsymbol{\beta}^*)^{q-1}}.$$

Now, consider the transformation $\boldsymbol{\beta} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\beta}^*$. We get,

$$f(\boldsymbol{\beta}) = |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \frac{\Gamma(q)}{2^q} \frac{g(r(\boldsymbol{\beta}))}{r(\boldsymbol{\beta})^{q-1}}.$$

This completes the proof.

420 **Proof of Proposition 1:** By applying Lemma 1, the proof is an immediate consequence of the fact that the group bridge prior belongs to the family of multivariate L_1 norm spherically symmetric distributions, as defined above.

Appendix C: Directly Sampling from the Group Bridge Prior

Based on Proposition 1, we observe that when $\alpha = 1$, $\boldsymbol{\beta}$ follows a joint distribution of q i.i.d Laplace distributions, which can be decomposed as $\boldsymbol{\beta} = R\mathbf{U}$,
425 where $\mathbf{U} = \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|}$ follows a multivariate ‘ L_1 norm sphere’ uniform distribution and R follows a particular gamma distribution. Therefore, any sample from the joint distribution of i.i.d Laplace distributions, divided by its L_1 norm can be regarded as a sample generated from the multivariate ‘ L_1 norm sphere’ uniform
430 distribution (a multivariate uniform distribution defined in an L_1 norm sphere). Therefore, given λ and α , the algorithm for generating observations from the q -dimensional group bridge prior proceeds as follows:

Step I: Generate an observation r^* from $\text{Gamma}(\frac{q}{\alpha}, \lambda)$ and calculate $r = r^{*1/\alpha}$.

Step II: Generate $\mathbf{z} = (z_1, \dots, z_q)'$ from the joint distribution of q i.i.d Laplace distributions with scale parameter 0 and location parameter 1 (these parameters
435 can be arbitrarily chosen).

Step III: Calculate $\mathbf{u} = \frac{\mathbf{z}}{\|\mathbf{z}\|_1}$. $\mathbf{u}$ can be regarded as a sample generated from the multivariate ' L_1 norm sphere' uniform distribution.

Step IV: $\mathbf{y} = \mathbf{u} * r$ is a sample generated from the desired group bridge prior.

440 **Appendix D: Scale Mixture of Multivariate Uniform Representation**

Proof of Proposition 2: Consider the L_1 norm open ball $A = \{\boldsymbol{\beta} \in \mathbb{R}^q : \|\boldsymbol{\beta}\|_1 < u^{\frac{1}{\alpha}}, u > 0, \alpha > 0\}$. The volume of A is given by $V_q(A) = \frac{2^q u^{\frac{q}{\alpha}}}{\Gamma(\frac{q}{\alpha} + 1)}$. We have,

$$\begin{aligned} \pi(\boldsymbol{\beta}) &= \frac{\lambda^{\frac{q}{\alpha}} \Gamma(q+1)}{2^q \Gamma(\frac{q}{\alpha} + 1)} \exp\{-\lambda \|\boldsymbol{\beta}\|_1^\alpha\} \\ &= \frac{\lambda^{\frac{q}{\alpha}+1} \Gamma(q+1)}{2^q \Gamma(\frac{q}{\alpha} + 1)} \int_{\|\boldsymbol{\beta}\|_1 < u^{\frac{1}{\alpha}}} \exp\{-\lambda u\} du \\ &= \int_{\|\boldsymbol{\beta}\|_1 < u^{\frac{1}{\alpha}}} \frac{1}{V_q(A_1)} \frac{\lambda^{\frac{q}{\alpha}+1}}{\Gamma(\frac{q}{\alpha} + 1)} u^{\frac{q}{\alpha}} \exp\{-\lambda u\} du = \int \pi(\boldsymbol{\beta}|u) \pi(u) du. \end{aligned}$$

This completes the proof.

Appendix E: A Gibbs Sampler for Sampling from the Multivariate Normal Distribution Truncated On An L_1 Norm Open Ball

I. Problem Statement

Let $\Omega_n(\rho)$ denote the following L_1 norm open ball in $\mathbb{R}^n$:

$$\Omega_n(\rho) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_1 < \rho\}, \quad (10)$$

445 where $\rho > 0$ is the radius of $\Omega_n(\rho)$ and $\|\cdot\|_1$ denotes the L_1 norm. With a slight abuse of notation, let us denote an n -dimensional truncated multivariate normal (TMVN) distribution defined on the open ball $\Omega_n(R)$ with mean vector μ and covariance matrix Σ as $N_{\Omega_n}(\mu, \Sigma)$. The following describes an efficient Gibbs sampler to sample from this distribution.

II. Gibbs Sampler for the TMVN Distribution

The full conditional distributions of the above TMVN distribution are truncated univariate normals. Therefore, the generation of $\tilde{\mathbf{x}}^{(t)}$ distributed according to a $N_{\Omega_n}(\mu, \Sigma)$ distribution proceeds by successively drawing the components $\tilde{x}_i^{(t)}, i = 1, \dots, n$ from their conditional distributions, i.e. at the step $(t + 1)$:

$$\tilde{x}_i^{(t+1)} | \tilde{\mathbf{x}}_{-i}^{(t)} \sim \text{TUVN}(\mu_i^{*(t)}, \sigma_i^{*2}, -L_i^{(t)}, L_i^{(t)}), \quad i = 1, \dots, n,$$

where $\tilde{\mathbf{x}}_{-i}^{(t)} = (\tilde{x}_1^{(t)}, \dots, \tilde{x}_{i-1}^{(t)}, \tilde{x}_{i+1}^{(t)}, \dots, \tilde{x}_n^{(t)})'$ and TUVN denotes a two-sided truncated univariate normal distribution truncated in $(-L_i^{(t)}, L_i^{(t)})$, $L_i^{(t)} > 0$, $i = 1, 2, \dots, n$ with mean $\mu_i^{*(t)}$ and variance σ_i^{*2} , where the expressions μ_i^*, σ_i^{*2} , and L_i are given as follows

$$\begin{aligned} \mu_i^{*(t)} &= \mu_i + s_i^T \Sigma_{-i}^{-1} (\tilde{\mathbf{x}}_{-i}^{(t)} - \boldsymbol{\mu}_{-i}), \\ \sigma_i^{*2} &= \sigma_{i,i}^2 - s_i^T \Sigma_{-i}^{-1} s_i, \\ L_i^{(t)} &= \rho - \sum_{k \neq i} |\tilde{x}_k^{(t)}|, \end{aligned} \tag{11}$$

where $\boldsymbol{\mu}_{-i}^{(t)} = (\mu_1^{(t)}, \dots, \mu_{(i-1)}^{(t)}, \mu_{(i+1)}^{(t)}, \dots, \mu_n^{(t)})'$, μ_i is the i^{th} element of μ , s_i is the $(n-1) \times 1$ vector obtained from the i^{th} column of Σ by removing the i^{th} row and Σ_{-i}^{-1} denotes the $(n-1) \times (n-1)$ matrix obtained from Σ^{-1} by removing the i^{th} row and i^{th} column. The sampling according to the above TUVN distributions
455 can be easily done by implementing an efficient algorithm proposed by Li and Ghosh (2015).

III. Sampling Coefficients in the Bayesian Group Bridge Regression

It is easy to note that the truncation region in BAGB posterior is group-wise
460 independent. Therefore, at each step, we sample from the group-wise conditional distribution (using the algorithm described above), conditional on other parameters, i.e. $\beta_k | \beta_{-k}, \mathbf{u}, \boldsymbol{\lambda}, \alpha, \sigma^2, y, X \sim N_{\Omega_k}(\mu_k, \Sigma_k)$, where $N_{\Omega_k}(\mu_k, \Sigma_k)$ is the m_k -dimensional TMVN distribution defined on the open ball $\Omega_k(u)$ with

mean vector μ_k and covariance matrix Σ_k , which are easily obtained by par-
 465 tioning the corresponding full variance-covariance matrix, and $\Omega_k(u_k) = \{\beta_k \in \mathbb{R}^{m_k} : \|\beta_k\|_1 < u_k^{1/\alpha}\}$, and $\beta_{-k} = (\beta_1, \dots, \beta_{k-1}, \beta_{k+1}, \dots, \beta_K)$, $k = 1, \dots, K$.

Appendix F: The BAGB Posterior Distributions for General Models

The full conditional distributions for the general models are given as

$$\begin{aligned} \beta|\mathbf{y}, X, \mathbf{u}, \lambda_1, \dots, \lambda_K, \alpha &\sim N_p(\tilde{\beta}, \hat{\Sigma}) \prod_{k=1}^K I\{\|\beta_k\|_1^\alpha < u_k\}, \\ \mathbf{u}|\mathbf{y}, X, \beta, \lambda_1, \dots, \lambda_K, \alpha &\sim \prod_{k=1}^K \text{Exponential}(\lambda_k) I\{u_k > \|\beta_k\|_1^\alpha\}, \\ \lambda_1, \dots, \lambda_K|\mathbf{y}, X, \beta, \alpha &\propto \prod_{k=1}^K \lambda_k^{a + \frac{m_k}{\alpha} - 1} \exp\{-\lambda_k(b + \|\beta_k\|_1^\alpha)\}. \\ \alpha|\mathbf{y}, X, \beta, \lambda_1, \dots, \lambda_K &\propto \alpha^{c-1}(1-\alpha)^{d-1} \prod_{k=1}^K \frac{\lambda_k^{m_k/\alpha}}{\Gamma(\frac{m_k}{\alpha} + 1)} \exp\{-\sum_{k=1}^K \lambda_k \|\beta_k\|_1^\alpha\}. \end{aligned}$$

Appendix G: The BAGB Posterior Distributions for Composite Group Bridge

The full conditional distributions for the composite group bridge (Seetharaman, 2013) can be derived as

$$\begin{aligned} \beta|\mathbf{y}, X, \mathbf{u}, \sigma^2, \lambda_1, \dots, \lambda_K, \alpha, \gamma &\sim N_p(\hat{\beta}_{\text{OLS}}, \sigma^2(X'X)^{-1}) \prod_{k=1}^K I\{\sum_{j=1}^{m_k} |\beta_{k_j}|^\gamma < u_k^{\frac{1}{\alpha}}\}, \\ \mathbf{u}|\mathbf{y}, X, \beta, \lambda_1, \dots, \lambda_K, \alpha, \gamma &\sim \prod_{k=1}^K \text{Exponential}(\lambda_k) I\{u_k > (\sum_{j=1}^{m_k} |\beta_{k_j}|^\gamma)^\alpha\}, \\ \sigma^2|\mathbf{y}, X, \beta &\sim \text{Inverse Gamma}(\frac{n-1}{2}, \frac{1}{2}(\mathbf{y} - X\beta)'(\mathbf{y} - X\beta)), \\ \lambda_1, \dots, \lambda_K|\mathbf{y}, X, \beta, \alpha, \gamma &\propto \prod_{k=1}^K \lambda_k^{a + \frac{m_k}{\alpha} - 1} \exp\{-\lambda_k(b + \{\sum_{j=1}^{m_k} |\beta_{k_j}|^\gamma\}^\alpha)\}. \\ \alpha|\mathbf{y}, X, \beta, \lambda_1, \dots, \lambda_K, \gamma &\propto \frac{\alpha^{c-1}(1-\alpha)^{d-1}}{\{\Gamma(\frac{1}{\gamma} + 1)\}^p} \prod_{k=1}^K \frac{\lambda_k^{m_k/\alpha} \Gamma(\frac{m_k}{\gamma} + 1)}{\Gamma(\frac{m_k}{\alpha} + 1)} \exp\{-\sum_{k=1}^K \lambda_k (\sum_{j=1}^{m_k} |\beta_{k_j}|^\gamma)^\alpha\}, \end{aligned}$$

$$\gamma|\mathbf{y}, X, \boldsymbol{\beta}, \lambda_1, \dots, \lambda_K, \alpha \propto \frac{\gamma^{c^*-1}(1-\gamma)^{d^*-1}}{\{\Gamma(\frac{1}{\gamma} + 1)\}^p} \prod_{k=1}^K \frac{\lambda_k^{m_k/\alpha} \Gamma(\frac{m_k}{\gamma} + 1)}{\Gamma(\frac{m_k}{\alpha} + 1)} \exp\{-\sum_{k=1}^K \lambda_k (\sum_{j=1}^{m_k} |\beta_{k_j}|^\gamma)^\alpha\}.$$

470 When $\gamma = 1$, the above reduces to the BAGB estimator.

References

- H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.
- P. Breheny and J. Huang. Penalized methods for bi-level variable selection. *Statistics and its interface*, 2(3):369, 2009.
- Patrick Breheny. The group exponential lasso for bilevel variable selection. *Biometrics*, 71(3):731–740 2015.
- Sounak Chakraborty and Ruixin Guo. A bayesian hybrid huberized support vector machine and its applications in high-dimensional medical data. *Computational Statistics & Data Analysis*, 55(3):1342–1356, 2011.
- Qixuan Chen and Sijian Wang. Variable selection for multiplyimputed data with application to dioxin exposure study. *Statistics in medicine*, 32(21):3646–3659
- K. T. Fang, S. Kotz, and K. W. Ng. *Symmetric Multivariate and Related Distributions*. Chapman and Hall, London, 1989.
- 485 Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.
- Zhigeng Geng, Sijian Wang, Menggang Yu, Patrick O Monahan, Victoria Champion, and Grace Wahba. Group variable selection via convex logexpsum penalty with application to a breast cancer survivor study. *Biometrics*, 71(1):
- 490 53–62 2015.
- Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2): 215–223, 1979.

- Daniel Hernández-Lobato, José Miguel Hernández-Lobato, and Pierre Dupont.
 495 Generalized spike-and-slab priors for bayesian group feature selection using
 expectation propagation. *Journal of Machine Learning Research*, 14(1):1891–
 1945, 2013.
- D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. John Wiley &
 Sons, 1989.
- 500 J. Huang, S. Ma, H. Xie, and C. Zhang. A group bridge approach for variable
 selection. *Biometrika*, 96(2):339–355, 2009.
- J. Huang, P. Breheny, and S. Ma. A selective review of group selection in
 high-dimensional models. *Statistical Science*, 27 (4), 2012.
- K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of Statis-*
 505 *tics*, pages 1356–1378, 2000.
- M. Kyung, J. Gill, M. Ghosh, and G. Casella. Penalized regression, standard
 errors, and bayesian lassos. *Bayesian Analysis*, 5:369–412, 2010.
- C. Leng, M. Tran, and D. Nott. Bayesian adaptive lasso. *Annals of the Institute*
of Mathematical Statistics, 66(2):221–244, 2014.
- 510 E. Levina and J. Zhu. Discussion of “sure independence screening for ultrahigh
 dimensional feature space”. *Journal of the Royal Statistical Society: Series*
B (Statistical Methodology), 70(5):897–898, 2008.
- Hanning Li and Debdeep Pati. Variable selection using shrinkage priors. *Com-*
putational Statistics & Data Analysis, 107:107–119, 3 2017.
- 515 Q. Li and N. Lin. The bayesian elastic net. *Bayesian Analysis*, 5(1):151–70,
 2010.
- Yifang Li and Sujit K Ghosh. Efficient sampling methods for truncated mul-
 tivariate normal and student-t distributions subject to linear inequality con-
 straints. *Journal of Statistical Theory and Practice*, 9(4):712–732

- 520 Colin L Mallows. Some comments on c p. *Technometrics*, 15(4):661–675, 1973.
- C. Park and Y. J. Yoon. Bridge regression: adaptivity and group selection. *Journal of Statistical Planning and Inference*, 141:3506–3519, 2011.
- N. G. Polson, J. G. Scott, and J. Windle. The bayesian bridge. *Journal of the Royal Statistical Society, Series B (Methodological)*, 76(4):713–733, 2014.
- 525 Wei Qian, Yi Yang, and Hui Zou. Tweedie’s compound poisson model with grouped elastic net. *Journal of Computational and Graphical Statistics*, 25(2):606–625
- G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.
- 530 Indu Seetharaman. Consistent bi-level variable selection via composite group bridge penalized regression. Master’s thesis, Kansas State University, 2013.
- N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1996.
- 535 H. Wang and C. Leng. Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048, 2007.
- M. Yuan and N. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 68:49–67, 2006.
- 540 H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.