Bayesian Analysis (2017) **12**, Number 4, pp. 1039–1067

Bayesian Variable Selection Regression of Multivariate Responses for Group Data

B. Liquet[*,†], K. Mengersen[‡], A. N. Pettitt[§], and M. Sutton[¶],

Abstract. We propose two multivariate extensions of the Bayesian group lasso for variable selection and estimation for data with high dimensional predictors and multi-dimensional response variables. The methods utilize spike and slab priors to yield solutions which are sparse at either a group level or both a group and individual feature level. The incorporation of group structure in a predictor matrix is a key factor in obtaining better estimators and identifying associations between multiple responses and predictors. The approach is suited to many biological studies where the response is multivariate and each predictor is embedded in some biological grouping structure such as gene pathways. Our Bayesian models are connected with penalized regression, and we prove both oracle and asymptotic distribution properties under an orthogonal design. We derive efficient Gibbs sampling algorithms for our models and provide the implementation in a comprehensive R package called `MBSGS` available on the Comprehensive R Archive Network (CRAN). The performance of the proposed approaches is compared to state-of-the-art variable selection strategies on simulated data sets. The proposed methodology is illustrated on a genetic dataset in order to identify markers grouping across chromosomes that explain the joint variability of gene expression in multiple tissues.

Keywords: Bayesian variable selection, multivariate regression, sparsity, spike and slab.

1 Introduction

In this article, we consider the challenging task of developing a fully Bayesian sparse regression analysis for the situation when the numbers of predictors is larger than observations for a multivariate response and covariates grouped by blocks with the sparsity for blocks and within blocks. It is well established that the incorporation of prior knowledge on the structure existing in the data for potential grouping of the covariates is key to more accurate prediction and improved interpretability. In genomics, genes within the same pathway have similar functions and act together in regulating a biological

[*] Laboratoire de Mathematiques et de leurs Applications, Université de Pau et des Pays de l'Adour, UMR CNRS 5142, Pau, France

[†] ARC Centre of Excellence for Mathematical and Statistical Frontiers and School of Mathematical Sciences at Queensland University of Technology, Brisbane, Australia, benoit.liquet@qut-edu.au

[‡] ARC Centre of Excellence for Mathematical and Statistical Frontiers and School of Mathematical Sciences at Queensland University of Technology, Brisbane, Australia, k.mengersen@qut-edu.au

[§] ARC Centre of Excellence for Mathematical and Statistical Frontiers and School of Mathematical Sciences at Queensland University of Technology, Brisbane, Australia, a.pettitt@qut-edu.au

[¶] ARC Centre of Excellence for Mathematical and Statistical Frontiers and School of Mathematical Sciences at Queensland University of Technology, Brisbane, Australia, m5.sutton@hdr.qut-edu.au

DOI: 10.1214/17-BA1081

system. These genes can add up to have a larger effect and therefore can be detected as a group (i.e., at a pathway or gene set level). Incorporation of this grouping structure is becoming increasingly common due to the success of geneset enrichment analysis approaches (Subramanian et al., 2005). For instance, the incorporation of group structure in regression analysis has been found to be effective for biomarker identification (Yuan and Lin, 2006; Meier et al., 2008; Puig et al., 2009; Simon and Tibshirani, 2012). Penalized regression methods are a popular approach for incorporating group structure and performing variable selection. Among these methods, Yuan and Lin (2006) proposed the group lasso by placing an L_2 penalty on the size of the regression coefficients. This method has drawn attention due to its ability to simultaneously perform group variable selection and estimate regression coefficients. The method was later extended by Meier et al. (2008) to logistic regression and modified by Puig et al. (2009) and Simon (2013) to consider non-orthonormal predictor matrices. Although the group lasso penalty can improve the quality of the variable selection, it requires a strong group-sparsity (Huang and Zhang, 2010), and cannot yield sparsity within a group. Ma et al. (2007) proposed a supervised group lasso which selects both significant gene clusters and significant genes within clusters for logistic binary classification and Cox survival analysis. Simon (2013) proposed a sparse group lasso penalty by combining an L_1 penalty with a group lasso to yield sparsity at both the group and individual feature level. Zhou (2010) applied this approach to genomic feature identification. Garcia et al. (2014) developed a sparse group-subgroup lasso to accommodate selecting important groups, subgroups and individual predictors. In a regression context, with a multivariate response variable, Li et al. (2015) have recently proposed a multivariate sparse group lasso. A review of group variable selection methods is presented by Huang et al. (2012). Recently, Liquet et al. (2016b) proposed a sparse group partial least squares approach for dealing with structured data in a genomic context.

In a Bayesian framework Xu and Ghosh (2015) proposed a Bayesian group lasso using spike and slab priors for group variable selection. Rockova and Lesaffre (2014) have recently developed rapid computational procedures based on the expectation maximization (EM) algorithm for a hierarchical model incorporating grouping information. Recently, Stingo et al. (2011) proposed a partial least squares approach for pathway and gene selection using variable selection priors and Markov chain Monte Carlo (MCMC) for computation. However, these Bayesian procedures only deal with a univariate response variable.

In some cases, the outcome can be complex and may consist of several correlated measures of continuous variables (e.g., metabolic syndrome). Figure 1 illustrates the most general situation of p predictors (typically OMICs measures) belonging to G groups being analyzed in relation to q correlated outcomes based on n observations.

The matrix $\mathbb{X}$ can be divided into G sub-matrices (groups) $\mathbb{X}_g : n \times m_g$ where m_g is the number of covariates in group g. For example, in gene expression data this sub-matrix may represent gene pathways or be factor level indicators for categorical data. The aim is to select only a few groups of $\mathbb{X}$ which are related to the multivariate response $\mathbb{Y}$. Further, sometimes we would like sparsity with respect to which groups are selected and which coefficients are nonzero within each group. For exam-

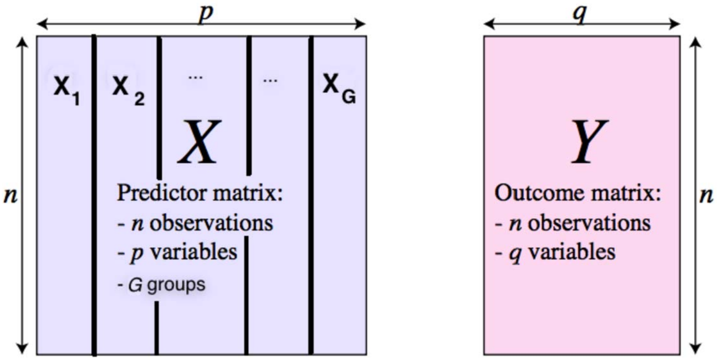

Figure 1: General representation of data for modeling multivariate outcomes given a group structure of the covariates.

ple, we might be interested in identifying important genes within selected gene pathways.

The multivariate response and predictor combination can be modelled using the multivariate Gaussian linear model:

$$\mathbb{Y} - \sum_{g=1}^{G} \mathbb{X}_g \mathbb{B}_g \sim MN_{n\times q}\left(\mathbf{0}_{p\times q}, \mathbb{I}_n, \Sigma\right), \tag{1}$$

where $\mathbb{Y}$ is a $n\times q$ matrix of responses, and $\mathbb{B}_g$ is a $m_g \times q$ matrix of regression coefficients associated to the sub-matrix predictors $\mathbb{X}_g$. $MN(\cdot,\cdot,\cdot)$ indicates the normal matrix-variate as defined in Dawid (1981) where Σ $(q \times q)$ controls the responses' residual correlation (the variance-covariance matrix of $\mathbb{Y} - \sum_{g=1}^{G} \mathbb{X}_g \mathbb{B}_g$) and the observations are treated as independent (e.g., no familial structure is assumed in the data, with I_n denoting the $n\times n$ identity matrix). Setting $q = 1$, the Gaussian linear model simplifies to

$$Y - \sum_{g=1}^{G} \mathbb{X}_g \beta_g \sim N_n\left(0, \sigma^2 \mathbb{I}_n\right), \tag{2}$$

where Y is a $n \times 1$ vector, β_g is a $p_g \times 1$ vector of regression coefficients associated with the group g, and σ^2 corresponds to the variance of the error term. We denote with $N_n(\cdot,\cdot)$ the $n-$variate normal distribution.

One way to analyze the multivariate Gaussian model (1) is to consider a collection of q regression problems in $\mathbb{R}^p$. Using this framework, we could fit q different univariate regression problems using any of the previous variable selection methods. However, in many applications the response matrix $\mathbb{Y}$ contains variables that are strongly correlated. As such, one would expect that the underlying covariates would be related. One approach for this type of problem would be to posit that there is some underlying subset of the coefficients in $\mathbb{X}$ which is related to all components of the response $\mathbb{Y}$.

A frequentist way to tackle this problem is offered by Friedman et al. (2010) who suggest using a group lasso penalty to select variables which are related to all components of the response $\mathbb{Y}$. Extensions of this approach have been proposed to provide simultaneous estimation of the precision matrix (inverse of Σ) and of the regression coefficients (see e.g., Rothman et al. (2010), Lee and Liu (2012), Cai et al. (2013)). However, these methods do not take into account the group structure of the predictors.

In the context of multivariate multiple regression, different approaches have been developed to account for the biological group structure within the predictor matrix (see e.g., Wen (2014), Zhu et al. (2014)). For imaging genomics data, Greenlaw et al. (2016) have recently proposed a Bayesian hierarchical modeling formulation where the posterior mode corresponds to the estimator proposed by Wang et al. (2012). However, their approach is limited to the case of $\Sigma = \sigma^2 \mathbb{I}_q$ corresponding to the independent phenotypes.

Following the ideas of Xu and Ghosh (2015), we develop more general Bayesian hierarchical models for variable selection with a group structure in the context of correlated multivariate response variables. For the univariate response model (2), a Bayesian group lasso model with spike and slab priors has been developed by Xu and Ghosh (2015) providing variable selection at the group level. The authors have also proposed a hierarchical spike and slab prior structure to select variables both at the group level and within each group. The posterior mode of their models was shown to provide shrinkage similar to the group lasso and sparse group lasso. Using this formulation, all groups were shrunk equally. Inspired by the adaptive group lasso (Wang and Leng, 2008; Nardi and Rinaldo, 2008) and the Bayesian adaptive lasso (Leng et al., 2014) we generalize their approach to allow for different amounts of shrinkage for different groups and coefficients.

For the multivariate response model (1), we define in Section 2 a Multivariate Bayesian Group Lasso with Spike and Slab prior (hereafter referred to as MBGL-SS) which enables only variable selection at the group level. Our Bayesian model is connected with penalized regression. We highlight properties of the posterior median estimator such as an Oracle property for group variable selection. We also derive asymptotic distributions under orthogonal designs. Then, we derive efficient Gibbs sampling algorithms for our group Bayesian lasso models with spike and slab priors. In order to select variables at both the group level and the individual level, we define a Multivariate Bayesian Sparse Group Selection with Spike and Slab priors (hereafter referred to as MBSGS-SS) in Section 3. Section 4 presents simulation studies to evaluate the performance of our approaches in terms of variable selection and prediction performance compared to frequentist approaches such as lasso, group lasso, and sparse group lasso. We illustrate the method in Section 5, with a challenging genetic data set where SNPs are used to predict the gene expression data in four tissue types. The final section concludes with a brief discussion.

2 Multivariate group lasso with spike and slab prior (MBGL-SS)

In this section, we consider the problem of Bayesian shrinkage estimation with group variables. The group Bayesian lasso was first proposed for a univariate response, by

Kyung et al. (2010) who showed that a scale mixture of Normals with Gamma hyper priors for β_g enables shrinkage of coefficients at the group level. While this method was shown to have shrinkage properties similar to the group lasso, Xu and Ghosh (2015) have stressed that the posterior mean and median do not produce exact zero estimates. To obtain sparsity at the group level, they proposed a hierarchical Bayesian group lasso model with an independent spike and slab type prior. We exploit and extend their approach for multivariate responses and propose the following hierarchical multivariate Bayesian group lasso model with an independent spike and slab prior for each group variable $\mathbb{B}_g$:

$$\mathbb{Y}|\mathbb{X}, \mathbb{B}, \Sigma \sim MN_{n\times q}(\mathbb{X}\mathbb{B}, \Sigma, \mathbb{I}_n), \tag{3}$$

$$Vec(\mathbb{B}_g^T|\Sigma, \tau_g, \pi_0) \overset{ind}{\sim} (1-\pi_0)N_{m_g q}(0, \mathbb{I}_{m_g} \otimes \tau_g^2\Sigma) + \pi_0\delta_0(Vec(\mathbb{B}_g^T)), \quad g = 1, \dots, G, \tag{4}$$

$$\tau_g^2 \overset{ind}{\sim} \text{Gamma}\left(\frac{m_g q + 1}{2}, \frac{\lambda_g^2}{2}\right), \quad g = 1, \dots, G, \tag{5}$$

$$\Sigma \sim \text{IW}(d, Q), \tag{6}$$

$$\pi_0 \sim \text{Beta}(a, b), \tag{7}$$

where $\delta_0(Vec(\mathbb{B}_g^T))$ denotes a point mass at $\mathbf{0} \in \mathbb{R}^{m_g q}$, $\mathbb{B}_g$ is the $m_g \times q$ regression coefficient matrix for the group g and denote β_{ij}^g ($i = 1, \dots, m_g$ and $j = 1, \dots, q$) as the elements of this matrix. The prior density of Σ is assumed to follow an inverse Wishart distribution (denoted IW) where d and Q are respectively the degrees of freedom and a positive finite scale matrix such that $\text{E}(\Sigma) = Q/(d-2)$. The matrix Q is defined as $Q = k\mathbb{I}_q$ with the hyper parameter k being comparable in size with the likely error variance of $\mathbb{Y}$ given $\mathbb{X}$ and $d = 3$ representing the smallest integer value ensuring the existence of $\text{E}(\Sigma)$.

Fixing π at $\frac{1}{2}$ is often recommended since it assigns equal probabilities to all submodels in the regression. Instead of fixing π_0 at $\frac{1}{2}$, we use a conjugate beta prior on π_0, $\pi_0 \sim \text{Beta}(a, b)$ to incorporate potential prior knowledge on the sparsity of the model. This choice has been adopted by Scheipl et al. (2012) and Xu and Ghosh (2015). By setting $a = b = 1$, it gives a uniform prior for π_0 with mean 0.5 but allows for spread in the prior. However, one can choose an informative prior such as $\pi_0 \sim \text{Beta}(20, 40)$ (see Scheipl et al. (2012)) to encourage a sparser model for high dimensional data.

The value of λ_g controls the amount of shrinkage for the gth group of coefficients. This parameter needs to be carefully tuned to provide the correct amount of shrinkage for the estimation. A large value of λ_g will result in parameters that are extremely biased towards zero, whilst small values of λ_g will lead to poor variable selection properties.

In Xu and Ghosh (2015), the shrinkage for each group is controlled by a single parameter. Consequently, larger groups of variables will be less affected by the shrinkage and more likely to be selected. In the original group lasso, Yuan and Lin (2006) propose weighting the shrinkage parameters by the size of their group to reduce the effect of different group sizes. We propose a global shrinkage parameterization of λ_g by setting $\lambda_g = \sqrt{m_g}\lambda$ where λ is a global shrinkage parameter, and m_g is the size of the group.

We take an empirical Bayes approach to estimate the value of λ from the data using marginal maximum likelihood (Park and Casella, 2008; Xu and Ghosh, 2015). Since the marginal likelihood function and marginal posterior for λ are intractable, a Monte Carlo EM algorithm is used. The kth EM update for λ is:

$$\lambda^{(k)} = \sqrt{\frac{G + qp}{\sum_{g=1}^{G} m_g \mathrm{E}_{\lambda^{(k-1)}}\left[\tau_g^2 | \mathbb{Y}\right]}},$$

in which the posterior expectation of τ_g^2 is replaced by the Monte Carlo sample average of τ_g^2 generated in the Gibbs sample based on $\lambda^{(k-1)}$. We name this choice of λ the *"global shrinkage parameter"*.

Inspired by the adaptive group lasso (Wang and Leng, 2008; Nardi and Rinaldo, 2008) and the Bayesian adaptive lasso (Leng et al., 2014) we propose an *"adaptive shrinkage parameter"* λ_g for each group. The adaptive shrinkage parameter can be estimated using a Monte Carlo EM algorithm where the kth update for λ_g is:

$$\lambda_g^{(k)} = \sqrt{\frac{1 + qm_g}{m_g \mathrm{E}_{\lambda^{(k-1)}}\left[\tau_g^2 | \mathbb{Y}\right]}}.$$

2.1 Connection to penalized regression and alternate reformulation of the model

To place our method in a context with the existing Bayesian group lasso and the penalized multivariate regression, we observe the marginal prior for $\mathbb{B}_g$. Integrating out the term τ_g^2 in (4) using prior (5), the marginal prior distribution is a mixture of a point mass at $\mathbf{0} \in \mathbb{R}^{m_g q}$ and a $m_g q$-dimensional K-distribution:

$$Vec(\mathbb{B}_g^T) \mid \Sigma, \pi_0 \sim (1-\pi_0)\mathrm{MK}\left(\frac{m_g q - 1}{2},\ \frac{\lambda_g^2}{2},\ \mathbf{0},\ \mathbb{I}_{m_g} \otimes \Sigma\right) + \pi_0 \delta_o Vec(\mathbb{B}_g^T), \quad (8)$$

where $\mathrm{MK}(\alpha, \beta, \boldsymbol{\mu}, \boldsymbol{\Gamma})$ denotes the Multivariate K-distribution as defined by Eltoft et al. (2006) with parameter set $\{\alpha, \beta, \boldsymbol{\mu}, \boldsymbol{\Gamma}\}$. In general the multivariate K-distribution does not have a closed form; however, for the parameters specified in (8) the density function is given by

$$\mathrm{MK}\left(\frac{m_g q - 1}{2},\ \frac{\lambda_g^2}{2},\ \mathbf{0},\ \mathbb{I}_{m_g} \otimes \Sigma\right) \propto \left(\frac{\lambda_g}{|\Sigma|}\right)^{m_g q} \exp(-\lambda_g \|Vec(\mathbb{B}_g^T)\|_{\mathbb{I}_{m_g} \otimes \Sigma}), \quad (9)$$

where $\|\boldsymbol{z}\|_\Gamma = \sqrt{\boldsymbol{z^T \Gamma^{-1} z}}$.

For a single response, the correlation matrix Σ will be a scalar denoted by σ and the MK distribution from (9) will reduce to the m_g-dimensional Multi-Laplace distribution,

$$\text{M-Laplace}\left(\mathbf{0}, \frac{\sigma}{\lambda_g}\right) \propto \left(\frac{\lambda_g}{\sigma}\right)^{m_g} \exp\left(-\frac{\lambda_g}{\sigma}\|Vec(\mathbb{B}_g^T)\|_2\right). \quad (10)$$

From (8) and (10), we can observe that the marginal prior for $Vect(\mathbb{B}_g)$ with a single response variable reduces to a point mass at $\mathbf{0} \in \mathbb{R}^{m_g}$, and a term that matches the Bayesian group lasso with shrinkage parameter λ_g (Raman et al., 2009; Kyung et al., 2010; Leng et al., 2014).

In the multivariate setting, the distribution for the slab part (9) can be interpreted as a generalization of the regular Bayesian group lasso that accounts for the correlations between the response variables. We note that there are many possible ways to extend the Laplace distribution for multivariate random variables. A general class of multivariate priors has been considered for group-sparse modeling by Babacan et al. (2014). While they do not consider correlated multivariate responses, they note that the multivariate Laplace distribution is part of a rich family of heavy tailed distributions. Importantly, it has been shown that for spike and slab priors using a heavy tailed distribution for the slab part results in optimal estimation risk with the posterior median estimator (Johnstone and Silverman, 2004).

To see the connection between our method and penalized regression we re-parameterize the regression coefficients: $Vec(\mathbb{B}^T) = \gamma_g \mathbf{b}_g$ where γ_g is an indicator taking a value 0 or 1 and $\mathbf{b}_g = (b_g^{(1,1)}, b_g^{(1,2)}, \ldots, b_g^{(m_g,q)})^T$. Guided by the marginal prior distribution (8) we place an MK distribution on $\mathbf{b}_g$ using the parameters from (9) and a Bernoulli prior on γ_g,

$$\mathbf{b}_g \mid \Sigma \sim \mathrm{MK}\left(\tfrac{m_g q-1}{2},\ \tfrac{\lambda_g^2}{2},\ \mathbf{0},\ \mathbb{I}_{m_g} \otimes \Sigma\right), \tag{11}$$

$$\gamma_g \mid \pi_0 \sim \mathrm{Ber}(1-\pi_0), \tag{12}$$

for $g = 1, 2, \ldots, G$.

Using the formulation $Vec(\mathbb{B}^T) = \gamma_g \mathbf{b}_g$, the marginal prior for $Vec(\mathbb{B}_g^T)$ will match the marginal prior (8). The negative log likelihood of the model and the prior defined by the above formulation is:

$$-\frac{1}{2}\|Vec(\mathbb{Y}^T) - Vec(\mathbb{B}^T\mathbb{X}^T)\|^2_{\mathbb{I}_n \otimes \Sigma} + \sum_{g=1}^{G} \lambda_g \|\mathbf{b}_g\|_{\mathbb{I}_{m_g} \otimes \Sigma} + \log\left(\frac{1-\pi_0}{\pi_0}\right) \sum_{g=1}^{G} \gamma_g + const.$$

In the case where the matrix Σ is set to $\sigma^2 \mathbb{I}_q$ we have $\|\mathbf{b}_g\|_{\mathbb{I}_{m_g} \otimes I_q \sigma^2} = \sigma^{-1}\|\mathbf{b}_g\|_2$. In this setting the likelihood becomes

$$-\frac{1}{2\sigma^2}\|\mathbb{Y} - \mathbb{X}\mathbb{B}\|_F^2 + \frac{1}{\sigma}\sum_{g=1}^{G} \lambda_g \|\mathbf{b}_g\|_2 + \log\left(\frac{1-\pi_0}{\pi_0}\right) \sum_{g=1}^{G} \gamma_g + const.$$

Thus the posterior mode of the regression model is equivalent to a penalized regression problem where groups are penalized with an ℓ_2 norm (see Li et al. (2015)) and the number of nonzero groups is penalized in an ℓ_0-like penalty. Setting $q = 1$ we obtain an expression similar to the likelihood found by Xu and Ghosh (2015). Once again our expression differs because we introduced a group spesific λ_g to allow for different shrinkage across groups.

2.2 Median thresholding estimator

A key point of this section is to highlight the benefits of using the posterior median estimator in spike and slab type models for both selection and estimation at the same time. We generalize to a multivariate response variable the thresholding results of the posterior median estimator proposed by Xu and Ghosh (2015), who have also generalized the thresholding results of Johnstone and Silverman (2004). We first show that the posterior median estimator enables one to perform group variable selection by obtaining a zero coefficient for some groups. Then, we express the posterior median as a soft thresholding estimator. Finally, we show that the median thresholding estimator is consistent in model selection and has optimal asymptotic estimation rate.

Posterior median estimator

Consider one group:

$$Z_{m\times 1} \sim f(z-\mu),$$

$$\mu \sim \pi_0 \delta_0(\mu) + (1-\pi_0)\gamma(\mu),$$

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

$$c = \frac{\int f(-v)\gamma(v)dv}{f(0)} \leq \frac{\int f(0)\gamma(v)dv}{f(0)} = 1,$$

Xu and Ghosh (2015) stated the following theorem:

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

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

In the case of a block orthogonal design matrix $\mathbb{X}$ (i.e., $\mathbb{X}_i^T\mathbb{X}_j = 0$ for $i \neq j$), we have for $1 \leq g \leq G$

$$Vec(\widehat{\mathbb{B}}_g^T) = Vec\left(((\mathbb{X}_g^T\mathbb{X}_g)^{-1}\mathbb{X}_g^T\mathbb{Y})^T\right) \sim N_{m_g q}\left(Vec(\mathbb{B}_g^T), (\mathbb{X}_g^T\mathbb{X}_g)^{-1} \otimes \Sigma\right).$$

By Theorem 1, assuming $\pi_0 > \frac{c}{1+c}$, then there exists $t(\pi_0) > 0$, such that the marginal posterior median of β_{ij}^g under the prior (4) satisfies

$$\mathrm{Med}(\beta_{ij}^g|\widehat{\mathbb{B}}_g) = 0 \text{ for any } 1 \leq i \leq m_g \text{ and } 1 \leq j \leq q,$$

when $||Vec(\widehat{\mathbb{B}}_g^T)||_2 < t$. As noted by Xu and Ghosh (2015), the marginal posterior median estimator of the gth group of regression coefficients is zero when the norm of the corresponding block least square estimator is less than a certain threshold.

Posterior median as a soft thresholding estimator

We assume now that the design matrix $\mathbb{X}$ is orthogonal, i.e., $\mathbb{X}^T\mathbb{X} = n\mathbb{I}_p$ and consider the model defined by (3) and (4) with fixed τ_g^2 $(1 \leq g \leq G)$. Under this model the posterior distribution of $\mathbb{B}_g$ is a spike and slab,

$$Vec(\mathbb{B}_g^T)|\mathbb{X},\mathbb{Y} \sim (1-l_g)N_{m_gq}\left((1-D_g)Vec(\mathbb{B}_{LS,g}^T), \frac{1-D_g}{n}\mathbb{I}_{m_g} \otimes \Sigma\right) + l_g\delta_0(Vec(\mathbb{B}_g^T)),$$

where $\mathbb{B}_{LS,g}$ is the least squares estimator of $\mathbb{B}_g$, $D_g = \frac{1}{1+n\tau_g^2}$, and

$$\begin{aligned} l_g &= p(\mathbb{B}_g = 0|rest) \\ &= \frac{\pi_0}{\pi_0 + (1-\pi_0)(\tau_g^2)^{-\frac{m_g(q-1)}{2}}(1+n\tau_g^2)^{-\frac{m_g}{2}} \exp\left\{(1-D_g)nTr[\Sigma^{-1}\mathbb{B}_{LS,g}^T\mathbb{B}_{LS,g}]\right\}}. \end{aligned}$$

Thus the marginal posterior distribution of β_{ij}^g $(1 \leq i \leq m_g$ and $1 \leq j \leq q)$ conditional on the observed data is also a spike and slab distribution,

$$\beta_{ij}^g|\mathbb{X},\mathbb{Y} \sim (1-l_g)N\left((1-D_g)\hat{\beta}_{LS,ij}^g, \frac{1-D_g}{n}\Sigma_{jj}\right) + l_g\delta_0(\beta_{ij}^g),$$

where Σ_{jj} is the j-th diagonal element of Σ. The resulting median is a soft thresholding estimator defined by

$$\hat{\beta}_{ij}^{Med,g} = \text{Med}(\beta_{ij}^g|\mathbb{X},\mathbb{Y}) = \text{sgn}\left(\hat{\beta}_{LS,ij}^g\right)\left((1-D_g)|\hat{\beta}_{LS,ij}^g| - \frac{\sqrt{\Sigma_{jj}}}{\sqrt{n}}Q_g\sqrt{1-D_g}\right)_+,$$

where z_+ denotes the positive part of z and $Q_g = \phi^{-1}(\frac{1}{2(1-\min(\frac{1}{2},l_g))})$. For a univariate response $(q = 1)$ the matrix Σ reduces to the scalar σ^2, and our result matches the previous work of Xu and Ghosh (2015). In the multivariate frequentist setting, Li et al. (2015) have proposed an iterative algorithm which utilizes a similar soft thresholding function to incorporate group structure in estimating the regression estimates.

Oracle property

Let $\mathbb{B}^0, \mathbb{B}_g^0, \beta_{ij}^{0,g}$ denote the true values of $\mathbb{B}, \mathbb{B}_g, \beta_{ij}^g$, respectively. Define the index vector of the true model as $\mathcal{A} = (I(||Vec(\mathbb{B}_g)||_2 \neq 0), g = 1,\dots,G)$, and the index vector of the model selected by a certain thresholding estimator $\hat{\mathbb{B}}_g$ as $\mathcal{A}_n = (I(||Vec(\hat{\mathbb{B}}_g)||_2 \neq 0), g = 1,\dots,G)$. Model selection consistency is attained if and only if $\lim_n P(\mathcal{A}_{n\to\infty} = \mathcal{A})$.

Under an orthogonal design, the median thresholding estimator has the oracle property.

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

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

and asymptotic normality,

$$\sqrt{n}\left(Vec(\hat{\mathbb{B}}_{\mathcal{A}}^{Med}) - Vec(\mathbb{B}_{\mathcal{A}}^{0})\right) \xrightarrow{d} N(\mathbf{0}, \Sigma \otimes \mathbb{I}).$$

The proof follows the same steps as the proof of Theorem 4 in Xu and Ghosh (2015). For asymptotic normality, the result comes from the fact that $\sqrt{n}(\hat{\beta}_{ij}^{Med,g} - \hat{\beta}_{ij}^{LS,g}) \xrightarrow{p} 0$, and $\sqrt{n}(Vec(\hat{\mathbb{B}}^{LS}) - Vec(\mathbb{B}^{0})) \xrightarrow{d} N(\mathbf{0}, \Sigma \otimes \mathbb{I})$.

2.3 Gibbs sampler

An efficient block Gibbs sampler (Hobert and Geyer, 1998) is used for simulating from the posterior distribution. The full posterior distribution of all the unknown parameters conditional on the data is

$$p(\mathbb{B}, \boldsymbol{\tau}^2, \Sigma, \pi_0 | \mathbb{Y}, \mathbb{X}) \propto p(\mathbb{Y}|\mathbb{B}, \boldsymbol{\tau}^2, \Sigma, \pi_0) \times p(\mathbb{B}|\boldsymbol{\tau}^2, \Sigma, \pi_0) \times p(\boldsymbol{\tau}^2) \times p(\Sigma) \times p(\pi_0), \quad (13)$$

where

$$\begin{aligned}
p(\mathbb{Y}|\mathbb{B}, \boldsymbol{\tau}^2, \Sigma, \pi_0) &\propto |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2} Tr\left[(\mathbb{Y} - \mathbb{X}\mathbb{B})\Sigma^{-1}(\mathbb{Y} - \mathbb{X}\mathbb{B})^T\right]\right\}, \\
p(\mathbb{B}|\boldsymbol{\tau}^2, \Sigma, \pi_0) &= \prod_{g=1}^{G} p(\mathbb{B}_g|\tau_g^2, \Sigma, \pi_0), \\
p(\mathbb{B}_g|\tau_g^2, \Sigma, \pi_0) &\propto (1-\pi_0)(2\pi)^{-\frac{qm_g}{2}} (\tau_g^2)^{-\frac{qm_g}{2}} |\Sigma|^{-\frac{m_g}{2}} \exp \\
&\quad - \left\{\frac{1}{2\tau_g^2} Tr\left[\mathbb{B}_g \Sigma^{-1} \mathbb{B}_g^T\right]\right\} I[\mathbb{B}_g \neq 0] + \pi_0 \delta_0(Vec(\mathbb{B}_g^T)), \\
p(\tau_1, \ldots, \tau_g) &\propto \prod_{g=1}^{G} (\lambda_g^2)^{\frac{qm_g+1}{2}} (\tau_g^2)^{\frac{qm_g+1}{2}-1} \exp\left(-\frac{\lambda_g^2}{2}\tau_g^2\right), \\
p(\pi_0) &\propto \pi_0^{a-1}(1-\pi_0)^{b-1}, \\
p(\Sigma) &\propto |\Sigma|^{-\frac{d+q+1}{2}} \exp\left\{-\frac{1}{2} Tr(Q\Sigma^{-1})\right\}.
\end{aligned}$$

Conditional posterior distribution

Let $\mathbb{B}_{(g)}$ denote the $\mathbb{B}$ matrix without the gth group, and $\mathbb{X}_{(g)}$ denote the covariate matrix corresponding to $\mathbb{B}_{(g)}$, that is,

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

where $\mathbb{X}_g$ is the design matrix corresponding to $\mathbb{B}_g$.

- **The conditional posterior distribution of** $\mathbb{B}_g$

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

$$Vec(\mathbb{B}_g^T)|\text{rest} \sim (1-l_g)N_{m_g q}\left(Vec(\mathbb{M}_g^T), \Sigma_g \otimes \Sigma\right) + l_g\delta_0(Vec(\mathbb{B}_g^T)), \quad g = 1, \dots, G, \quad (14)$$

where

$$l_g = p(\mathbb{B}_g = 0|rest) = \frac{\pi_0}{\pi_0 + (1-\pi_0)(\tau_g^2)^{-\frac{qm_g}{2}}|\Sigma_g|^{\frac{q}{2}}\exp\left\{\frac{1}{2}Tr[\Sigma^{-1}\mathbb{M}_g^T\Sigma_g^{-1}\mathbb{M}_g]\right\}}.$$

- **The conditional posterior distribution of $\alpha_g^2 = \frac{1}{\tau_g^2}$**

$$\alpha_g^2|\text{rest} \sim \begin{cases} \text{Inverse Gamma}\left(\text{shape} = \frac{m_g q+1}{2}, \text{scale} = \frac{\lambda_g^2}{2}\right), & \text{if } \mathbb{B}_g = 0, \\ \text{Inverse Gaussian}\left(\frac{\lambda_g}{(Tr[\mathbb{B}_g\Sigma^{-1}\mathbb{B}_g^T])^{-1/2}}, \lambda_g^2\right), & \text{if } \mathbb{B}_g \neq 0, \end{cases}$$

where the inverse Gaussian distribution is defined in Folks and Chhikara (1978) and the inverse Gamma distribution in Gelman et al. (2014).

- **The conditional posterior distribution of Σ**

$$\Sigma|\text{rest} \sim \text{IW}\left(d + n + \sum_{g=1}^{G} m_g Z_g, (\mathbb{Y} - \mathbb{X}\mathbb{B})^T(\mathbb{Y} - \mathbb{X}\mathbb{B}) + \mathbb{B}^T\mathbb{D}_{\boldsymbol{\tau}}\mathbb{B} + Q\right),$$

where

$$Z_g = \begin{cases} 1 & \text{if } \mathbb{B}_g \neq 0, \\ 0 & \text{if } \mathbb{B}_g = 0 \end{cases} \quad \text{and} \quad \mathbb{D}_{\boldsymbol{\tau}} = diag(\frac{1}{\tau_1^2}\mathbb{I}_{m_1}, \dots, \frac{1}{\tau_{m_G}^2}\mathbb{I}_{m_G}).$$

- **The conditional posterior distribution of π_0**

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

Remark. From the Gibbs sampler, different strategies are used to select models and predictors. The highest posterior probability model (denoted here after HPPM) is estimated by recording at each simulation (iteration of the Gibbs sampler) the generated model. Then, the generated models are tabulated to find the model that has the highest frequency. Thus HPPM defines the selected relevant groups. An alternative choice is the median estimator which is found by taking the element wise median of the samples of $\mathbb{B}_g$ generated by the Gibbs sampler.

3 Multivariate sparse group selection with spike and slab prior (MBSGS-SS)

The MBGL-SS is tailored for problems that only require group level sparsity. However, sometimes we would like to combine both sparsity of groups and within each group. For example, if the predictor matrix contains genes, we might be interested in identifying particularly important genes in pathways of interest. For a univariate response, Xu and Ghosh (2015) defined a Bayesian sparse group lasso which offers shrinkage effects at both the group level and also within a group. However, the authors stressed that the model does not produce a sparse model since the posterior mean/median estimators are never exactly set to zero. To overcome this drawback, the authors defined a hierarchical Bayesian sparse group selection with spike and slab prior using spike and slab type priors (named BSGS-SS) for both group variable selection and individual variable selection. We exploit the same idea for a multivariate response variable.

Model specification

First, we reparametrize the coefficient matrices to tackle the two kinds of sparsity separately:

$$\mathbb{B}_g = \boldsymbol{V}_g^{\frac{1}{2}} \tilde{\mathbb{B}}_g, \text{ where } \boldsymbol{V}_g^{\frac{1}{2}} = \mathrm{diag}\{\tau_{g_1}, \dots, \tau_{gm_g}\}, \ \tau_{gj} \geq 0, \ g = 1, \dots, G; \ j = 1, \dots, m_g, \tag{15}$$

where $\tilde{\mathbb{B}}_g$, when nonzero, follows the distribution $Vec(\tilde{\mathbb{B}}_g^T) \sim N_{m_g q}(\mathbf{0}, \mathbb{I}_{m_g} \otimes \Sigma)$. Thus the diagonal element of $\boldsymbol{V}_g^{\frac{1}{2}}$ control the magnitude of the elements of $\mathbb{B}_g$. To select variables at the group level, we assume the multivariate spike and slab prior for each $Vec(\tilde{\mathbb{B}}_g^T)$:

$$Vec(\tilde{\mathbb{B}}_g^T | \Sigma, \tau_g, \pi_0) \overset{ind}{\sim} (1-\pi_0) N_{m_g q}(0, \mathbb{I}_{m_g} \otimes \Sigma) + \pi_0 \delta_0(Vec(\tilde{\mathbb{B}}_g^T)), \quad g = 1, \dots, G. \tag{16}$$

We denote the j-th row of $\mathbb{B}_g$ by $\mathbb{B}_g^j$ and the j-th row of $\tilde{\mathbb{B}}_g$ by $\tilde{\mathbb{B}}_g^j$. Note that when $\tau_{gj} = 0$, the row $\mathbb{B}_g^j$ is set to zero, even when the corresponding row $\tilde{\mathbb{B}}_g^j$ is nonzero. In order to choose variables within each relevant group, we assume the following spike and slab prior for each τ_{gj}:

$$\tau_{gj} \overset{ind}{\sim} (1-\pi_1) N^+(0, s^2) + \pi_1 \delta_0(\tau_{gj}), \quad g = 1, \dots, G; \ j = 1, \dots, m_g, \tag{17}$$

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

Prior specification

- We assume an Inverse Wishart prior for $\Sigma \sim \mathrm{IW}(d, Q)$
- We assume conjugate beta hyper-priors for π_0 and π_1:

$$\pi_0 \sim \mathrm{Beta}(a_1, a_2), \quad \pi_1 \sim \mathrm{Beta}(c_1, c_2). \tag{18}$$

- We use a conjugate inverse gamma prior for $s^2 \sim$ Inverse Gamma$(1, t)$, and estimate t with the Monte Carlo EM algorithm. For the k-th EM update,

$$t^{(k)} = \frac{1}{\mathrm{E}_{t^{(k-1)}}\left[\frac{1}{s^2} | \mathbb{Y}\right]},$$

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

3.1 Gibbs sampler

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

Joint posterior

$$
\begin{aligned}
&p(\tilde{\mathbb{B}}, \boldsymbol{\tau}^2, \Sigma, \pi_0, \pi_1, s^2 | \mathbb{Y}, \mathbb{X}) \\
&\quad \propto |\Sigma|^{-n/2} \exp\left\{ -\frac{1}{2} Tr \left[\left(\mathbb{Y} - \sum_{g=1}^{G} \mathbb{X}_g \boldsymbol{V}_g^{\frac{1}{2}} \tilde{\mathbb{B}}_g \right) \Sigma^{-1} \left(\mathbb{Y} - \sum_{g=1}^{G} \mathbb{X}_g \boldsymbol{V}_g^{\frac{1}{2}} \tilde{\mathbb{B}}_g \right)^T \right] \right\} \\
&\quad \times \prod_{g=1}^{G} (1-\pi_0)(2\pi)^{-\frac{q m_g}{2}} |\Sigma|^{-\frac{m_g}{2}} \exp\left\{ -\frac{1}{2} Tr \left[\tilde{\mathbb{B}}_g \Sigma^{-1} \tilde{\mathbb{B}}_g^T \right] \right\} I[\tilde{\mathbb{B}}_g \neq 0] + \pi_0 \delta_0(Vec(\tilde{\mathbb{B}}_g^T)) \\
&\quad \times \prod_{g=1}^{G} \prod_{j=1}^{m_g} \left[(1-\pi_1) 2 (2\pi s^2)^{-\frac{1}{2}} \exp\left\{ -\frac{\tau_{gj}^2}{2s^2} \right\} I[\tau_{gj} > 0] + \pi_1 \delta_0(\tau_{gj}) \right] \\
&\quad \times |\Sigma|^{-\frac{d+q+1}{2}} \exp\left\{ -\frac{1}{2} Tr(Q\Sigma^{-1}) \right\} \\
&\quad \times \pi_0^{a_1 - 1} (1-\pi_0)^{a_2 - 1} \\
&\quad \times \pi_1^{c_1 - 1} (1-\pi_1)^{c_2 - 1} \\
&\quad \times t (s^2)^{-2} \exp\left\{ -\frac{t}{s^2} \right\}.
\end{aligned}
$$

Let $\mathbb{B}_{(gj)}$ denote the $\mathbb{B}$ without the jth row vector in the gth group, and $\mathbb{X}_{(gj)}$ denote the covariate matrix corresponding to $\mathbb{B}_{(gj)}$, that is,

$$
\mathbb{X}_{(gj)} = (x_{1,1}, \ldots, x_{1,m_1}, \ldots, x_{g,1} \ldots, x_{g,j-1}, x_{g,j+1}, \ldots, x_{g,m_g}, \ldots, x_{G,m_G}).
$$

The posterior distribution of $\tilde{\mathbb{B}}_g$

Let $\mathbb{M}_g = \Sigma_g \boldsymbol{V}_g^{\frac{1}{2}} \mathbb{X}_g^T (\mathbb{Y} - \mathbb{X}_{(g)} \boldsymbol{V}_{(g)}^{\frac{1}{2}} \tilde{\mathbb{B}}_{(g)})$, $\Sigma_g = (\mathbb{I}_{m_g} + \boldsymbol{V}_g^{\frac{1}{2}} \mathbb{X}_g^T \mathbb{X}_g \boldsymbol{V}_g^{\frac{1}{2}})^{-1}$, then the conditional posterior distribution of $\tilde{\mathbb{B}}_g$ is a spike and slab distribution

$$
Vec(\tilde{\mathbb{B}}_g^T)|\text{rest} \sim (1-l_g) N_{m_g q} \left(Vec(\mathbb{M}_g^T), \Sigma_g \otimes \Sigma \right) + l_g \delta_0(Vec(\tilde{\mathbb{B}}_g^T)), \quad g = 1, \ldots, G, \tag{19}
$$

where

$$
l_g = p(\tilde{\mathbb{B}}_g = 0 | rest) = \frac{\pi_0}{\pi_0 + (1-\pi_0)|\Sigma_g|^{\frac{q}{2}} \exp\left\{ \frac{1}{2} Tr[\Sigma^{-1} \mathbb{M}_g^T \Sigma_g^{-1} \mathbb{M}_g] \right\}}. \tag{20}
$$

The conditional posterior distribution of τ_{gj}

The conditional posterior distribution of τ_{gj} is a spike and slab distribution:

$$\tau_{gj}|\text{rest} \sim (1-q_{gj})N^{+}(u_{gj}, v_{gj}^2) + q_{gj}\delta_0(\tau_{gj}),\ g=1,\ldots,G;\ j=1,\ldots,m_G, \tag{21}$$

where $u_{gj} = Tr[\Sigma^{-1}(\mathbb{Y}^T - \mathbb{B}_{(gj)}^T\mathbb{X}_{(gj)}^T)\mathbb{X}_{gj}\tilde{\mathbb{B}}_{gj}]/\{Tr[\Sigma^{-1}\tilde{\mathbb{B}}_{gj}^T\mathbb{X}_{gj}^T\mathbb{X}_{gj}\tilde{\mathbb{B}}_{gj}] + \frac{1}{s^2}\}$, $v_{gj}^2 = (Tr[\Sigma^{-1}\tilde{\mathbb{B}}_{gj}^T\mathbb{X}_{gj}^T\mathbb{X}_{gj}\tilde{\mathbb{B}}_{gj}] + \frac{1}{s^2})^{-1}$

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

The conditional posterior distribution of Σ

The conditional posterior distribution of Σ is an inverse Wishart distribution:

$$\Sigma|\text{rest} \sim \text{IW}\left(d+n+\sum_{g=1}^{G} m_g Z_g, (\mathbb{Y}-\mathbb{X}\mathbb{B})^T(\mathbb{Y}-\mathbb{X}\mathbb{B}) + \tilde{\mathbb{B}}^T\tilde{\mathbb{B}} + Q\right),$$

where

$$Z_g = \begin{cases} 1 & \text{if } \tilde{\mathbb{B}}_g \neq 0, \\ 0 & \text{if } \tilde{\mathbb{B}}_g = 0. \end{cases}$$

The conditional posterior distribution of π_0 and π_1

$$\pi_0|\text{rest} \sim \text{Beta}\left(a_1 + G - \sum_{g=1}^{G} Z_g, a_2 + \sum_{g=1}^{G} Z_g\right),$$
$$\pi_1|\text{rest} \sim \text{Beta}\left(\#(\tau_{gj}=0) + c_1, \#(\tau_{gj}\neq 0) + c_2\right).$$

The conditional posterior distribution of s^2

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

Remark. The MBGL-SS and MBSGS-SS methods have been designed to tackle the situation where a subset of predictors in $\mathbb{X}$ are related to all components of the response $\mathbb{Y}$. Meaning, that our models have not been designed to allow for sparseness within a regressor across the response variables. However, the median estimator does do this, as it were, for free. This nice feature of the median estimator is highlighted in both the simulation study and the case study application.

4 Simulation studies

To investigate the properties of our approach, our first simulation study was conducted in the univariate setting to show the behavior of the BGL-SS (Bayesian Group Lasso with Spike and Slab prior) proposed by Xu and Ghosh (2015) compared to the proposed extension including the group size effect related to the shrinkage part of our model. We compared different approaches (such as lasso, group lasso, sparse group lasso) in terms of prediction and variable selection accuracy performance. Then, a second simulation study was performed with a multivariate response to demonstrate the good prediction and variable selection accuracy performance of MBGL-SS and MBSGS-SS when compared with BGL-SS, BSGS-SS (Bayesian Sparse Group selection with spike and slab priors defined in Xu and Ghosh (2015)) and two lasso methods (denoted mlasso and MRCE) for a multivariate response. The mlasso method has been implemented in the `glmnet` R package and assumes that there is some underlying subset of the coefficients in $\mathbb{X}$ which are related to all components of the response $\mathbb{Y}$. The multivariate lasso (mlasso) problem is stated as:

$$\hat{\mathbb{B}}_{mlasso} = \underset{\mathbb{B}\in\mathbb{R}^{p\times q}}{\operatorname{argmin}} \left\{ \|\mathbb{Y} - \mathbb{X}\mathbb{B}\|_F^2 + \lambda \sum_{j=1}^{q} \|\mathbb{B}^{jT}\|_2 \right\},$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $\mathbb{B}^{jT}$ denotes the jth row of $\mathbb{B}$. The MRCE method proposed by Rothman et al. (2010) has been implemented in the `MRCE` R package (Rothman, 2017). The multivariate regression with covariance estimation (MRCE) method producing a sparse estimator of $\mathbb{B}$ which depends on the inverse of the covariance matrix $\Omega = \Sigma^{-1}$ is stated as:

$$(\hat{\mathbb{B}}, \hat{\Omega}) = \underset{\mathbb{B}\in\mathbb{R}^{p\times q}, \Omega\in\mathbb{R}^{q\times q}}{\operatorname{argmin}} \left\{ L(\mathbb{B}, \Omega) + \lambda_1 \sum_{j\neq j'} |w_{j'j}| + \lambda_2 \sum_{j=1}^{q}\sum_{k=1}^{q} |\beta_{jk}| \right\}, \qquad (22)$$

where $L(\mathbb{B}, \Omega)$ is the negative log-likelihood function and $\Omega = [w_{j'j}]$.

Note that BGL-SS and BSGS-SS are designed for univariate response but are also used in the multivariate response simulation (viewed as q univariate regressions) and the results are pooled to obtain prediction performance. The different Bayesian methods used have been implemented in our R package `MBSGS` (Liquet and Sutton, 2017) available on CRAN.

The posterior mean and posterior median are both used as our Bayes estimators and we compare their variable selection and prediction performance.

For our Bayesian methods, we generate data from the full posterior distribution with a Gibbs Sampler, running 20000 iterations in which the first 10000 are burn-ins. For MBGL-SS and BGL-SS, we set $a = b = 1$ for the hyperparameters relating to π_0. For MBSGS-SS and BSGS-SS, hyperparameters relating to the Beta distributions of π_0 and π_1 are chosen to be $a_1 = a_2 = c_1 = c_2 = 1$. As suggested by Brown et al. (1998), a weak prior information requires a small value for the hyperparameter d. We set $d = 3$ which is just a convenient small value, the smallest integer value for which the expectation of

Σ exists. Instead of an arbitrary setting of the hyperparameter k for the expectation of the error variance ($\mathrm{E}(\Sigma) = kI_q$), we propose a practical way to fix it in the spirit of an Empirical Bayes approach. As adopted by Petretto et al. (2010) and Liquet et al. (2016a), we perform q univariate regressions which enable us to derive an estimate of the error variances. We fix k to be the average of the q residual error variance from the univariate models. In the case of $p > n$, q univariate forward regressions are performed to derive an estimate of the error variance.

We use the `glmnet` R package (Friedman et al., 2010) to perform the lasso method for univariate and multivariate responses. The `SGL` R package (Simon et al., 2013) is used to perform the group and sparse group lasso methods for the univariate setting. The `MRCE` R package (Rothman, 2017) is used to perform the multivariate regression with covariance estimation. The tuning parameters for the frequentist methods are calibrated using 5-fold cross-validation.

4.1 Univariate setting

In this simulation setting, we investigate both the effect of correlation between predictors and the group size effect.

True models

The data have been generated from the following univariate model:

$$y = \mathbb{X}\beta + \epsilon \quad \text{where} \quad \epsilon \sim N_n(0, \sigma\mathbb{I}_n),$$

where each row of the predictor matrix $\mathbb{X}$ is generated from a multivariate Normal distribution with zero mean and covariance matrix $\Sigma_{\mathbb{X}} = (1-\rho)I_p + \rho\mathbf{1}_p\mathbf{1}_p^T$ where the correlation ρ is given according the simulation setting, $\mathbf{1}_m$ is the m-length vector of ones. We consider the following two simulation settings:

- Model 1. We simulated data sets with $n = 120$ observations and $p = 20$ covariates divided into 4 groups with 5 covariates each. We randomly sampled 80 observations to train the model and used the remaining 40 for comparing performance prediction. Let $\beta^T = ((0.3, -1, 0, 0.5, 0.01), \mathbf{0}_5, \mathbf{0.8}_5, \mathbf{0}_5)$, where the notation $\mathbf{x}_l$ denotes a vector of length l with x values. We varied the pairwise correlation $\rho \in \{0,\ 0.5,\ 0.75\}$ between covariates. We specify $\sigma = 3$.

- Model 2. We simulated a data set with $n = 120$ observations and $p = 130$ covariates divided into 5 groups with respectively 5, 5, 20, 50 and 50 covariates. We randomly sampled 80 observations to train the model and used the remaining 40 for comparing performance prediction. Let $\beta^T = ((0.3, -1, 0, 0.5, 0.01), \mathbf{0}_5, \mathbf{0.8}_5, \mathbf{0}_{50}, \mathbf{0}_{50})$. We vary the pairwise correlation $\rho \in \{0,\ 0.5,\ 0.75\}$ between covariates. We specify $\sigma = 3$. This model enables us to investigate both the effect of correlation between predictors and the group size effect. For this model we also investigate the behavior of our methods when the sample size increases (200 and 300 observations for training and 40 observations for comparing prediction performance).

For the simulated data for model 1, Table 1 of the Supplementary Material (Liquet et al., 2017) presents the model selection accuracy over 50 replications for the different methods designed for univariate response variables. The models are compared with true and false positive rates and with Matthews correlation coefficient.

For both BGL-SS and BSGS-SS, the median thresholding model (MTM) outperforms all other methods including the highest posterior probability model (HPPM) whatever the values of the pairwise correlation. Lasso, group lasso (gLasso) and sparse group lasso (sgLasso) tend to select more variables than the spike and slab methods. A similar pattern for the simulations was noted in Xu and Ghosh (2015). Our extension of the BGL-SS model incorporating the group size effect gives similar results to the traditional one if we use the global shrinkage parameterization of λ_g while the adaptive shrinkage parameterization tends to select more variables.

The results regarding the prediction performance (mean square error of prediction) are presented in Table 2 of the Supplementary Material. The medians of the mean squared prediction error are compared for the 12 methods. The bootstrapped standard errors of the medians are given in the parentheses. BSGS-SS and BGL-SS methods gave similar results and outperform the frequentist lasso approaches which are adversely impacted by the correlation between predictors. Note that in this case the posterior mean estimator and posterior median estimator have similar performances.

Tables 3 and 4 of the Supplementary Material present performance results for Model 2, where the number of predictors in each group varies. This structure of the data clearly affects the BGL-SS model proposed by Xu and Ghosh (2015) especially when the predictors are correlated and with small sample size. Our modifications of the model (including the group size effect to the shrinkage parameter) combined with the *"adaptive shrinkage parameter"* give better results, especially with the Median Thresholding Model. We can note that the BSGS-SS model is not affected by this structure of the data and out performs all the other methods. Only for high correlation and small sample size are the frequentist approaches competitive in terms of variable selection compared with the BGL-SS model with *"adaptive shrinkage parameter"*. However, regardless of the pairwise correlations between the predictors, the frequentist methods have worse prediction performances.

4.2 Multivariate setting

True models

The data have been generated from the following multivariate model:

$$\mathbb{Y} = \mathbb{X}\mathbb{B} + \mathbb{E} \quad \text{where} \quad \mathbb{E} \sim MN_{n\times q}(0, \Sigma, \mathbb{I}_n), \quad \mathbb{B} = [\mathbb{B}^1, \mathbb{B}^2, \mathbb{B}^3] \text{ and } \; q = 3.$$

For all models, we assumed strong levels of correlation between the first and second outcomes, and weaker levels for the other pairwise correlations. Specifically, we defined

$$\Sigma = \begin{pmatrix} 1 & 0.95 & 0.5 \\ 0.95 & 1 & 0.3 \\ 0.5 & 0.3 & 1 \end{pmatrix}. \tag{23}$$

We considered the following nine simulations setting:

- Model 1. We simulated data sets with $n = 100$ observations and $p = 20$ covariates divided into four groups with five covariates each. We randomly sampled 60 observations to train the model and used the remaining 40 for comparing performance prediction. Let

$$\mathbb{B}^T = \begin{pmatrix} 0.3 & -1 & 0 & 0.5 & 0.01 & \mathbf{0}_5 & \mathbf{0.8}_5 & \mathbf{0}_5 \\ 0.2 & -1.1 & 0 & 0.6 & 0.02 & \mathbf{0}_5 & \mathbf{0.7}_5 & \mathbf{0}_5 \\ 0.1 & -1.2 & 0 & 0.7 & 0.03 & \mathbf{0}_5 & \mathbf{0.6}_5 & \mathbf{0}_5 \end{pmatrix}. \tag{24}$$

  The pairwise correlation between covariates is set equal to 0.5.

- Model 2. The simulation setting is the same as for the previous model except for the true $\mathbb{B}$. Let

$$\mathbb{B}^T = \begin{pmatrix} 0.3 & -1 & 0 & 0.5 & 0.01 & \mathbf{0}_5 & \mathbf{0.8}_5 & \mathbf{0}_5 \\ 0 & 0 & 0 & 0 & 0.0 & \mathbf{0}_5 & \mathbf{0.7}_5 & \mathbf{0}_5 \\ 0.1 & -1.2 & 0 & 0.7 & 0.03 & \mathbf{0}_5 & \mathbf{0.6}_5 & \mathbf{0}_5 \end{pmatrix}. \tag{25}$$

  In this simulation, some relevant predictors (1,2, 3 and 4) are not associated for the second response variables.

- Model 3. We consider the situation where $n < p$. We simulated data set with $n = 60$ and $p = 80$ covariates divided into 16 groups with 5 covariates each. We use 40 observations to train the model and used the remaining 20 for comparing performance prediction. Let

$$\mathbb{B}^T = \begin{pmatrix} 0.5 & 1 & 1.5 & 2 & 2.5 & \mathbf{0}_5 & 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & \mathbf{0}_5 & \dots & \mathbf{0}_5 \\ 0.25 & 0.5 & 0.75 & 1 & 1.25 & \mathbf{0}_5 & 0.05 & 0.1 & 0.15 & 0.2 & 0.25 & \mathbf{0}_5 & \dots & \mathbf{0}_5 \\ 0.2 & 0.4 & 0.6 & 0.8 & 1 & \mathbf{0}_5 & \frac{1}{30} & \frac{2}{30} & \frac{3}{30} & \frac{4}{30} & \frac{5}{30} & \mathbf{0}_5 & \dots & \mathbf{0}_5 \end{pmatrix}. \tag{26}$$

  We define the jth predictor in group g as $X_{gj} = z_g + z_{gj}$, where z_g and z_{gj} are independent standard normal variates, $g = 1, \dots, 16$; $j = 1, 2, \dots, 5$. Thus predictors within a group are correlated with pairwise correlation $\frac{1}{2}$ while the predictors in different groups are independent.

- Model 4. The simulation setting is the same as for the previous model except for the true $\mathbb{B}$. Let

$$\mathbb{B}^T = \begin{pmatrix} 0.5 & 1 & 1.5 & 2 & 2.5 & \mathbf{0}_5 & 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & \mathbf{0}_5 & \dots & \mathbf{0}_5 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{0}_5 & 0.05 & 0.1 & 0.15 & 0.2 & 0.25 & \mathbf{0}_5 & \dots & \mathbf{0}_5 \\ 0.2 & 0.4 & 0.6 & 0.8 & 1 & \mathbf{0}_5 & 0 & 0 & 0 & 0 & 0 & \mathbf{0}_5 & \dots & \mathbf{0}_5 \end{pmatrix}. \tag{27}$$

  As in model 2, some relevant covariates are non-zero for two responses and zero for the other one.

- Model 5. We simulated data with $n = 100$ and $p = 40$ covariates divided into 4 groups with 10 covariates each. We randomly sampled 60 observations to train the model and used the remaining 40 for comparing performance prediction. Let
$$\mathbb{B}^T = \begin{pmatrix} \mathbf{0}_{10} & \mathbf{2}_{10} & \mathbf{0}_{10}, & \mathbf{2}_{10} \\ \mathbf{0}_{10} & \mathbf{1}_{10} & \mathbf{0}_{10}, & \mathbf{1}_{10} \\ \mathbf{0}_{10} & \mathbf{0.5}_{10} & \mathbf{0}_{10}, & \mathbf{0.5}_{10} \end{pmatrix}. \tag{28}$$
Predictors have been simulated in the same way as in model 2.

- Model 6. The simulation setting is the same as for the previous model except for the true $\mathbb{B}$. Let
$$\mathbb{B}^T = \begin{pmatrix} \mathbf{0}_{10} & (\mathbf{2}_5, \mathbf{0}_5) & \mathbf{0}_{10}, & (\mathbf{2}_5, \mathbf{0}_5) \\ \mathbf{0}_{10} & (\mathbf{1}_5, \mathbf{0}_5) & \mathbf{0}_{10}, & (\mathbf{1}_5, \mathbf{0}_5) \\ \mathbf{0}_{10} & (\mathbf{0.5}_5, \mathbf{0}_5) & \mathbf{0}_{10}, & (\mathbf{0.5}_5, \mathbf{0}_5) \end{pmatrix}. \tag{29}$$

- Model 7. We simulated data with $n = 240$ and $p = 500$ covariates divided into 50 groups with 10 covariates each. We randomly sampled 200 observations to train the model and used the remaining 40 for comparing performance prediction. Let
$$\mathbb{B}^T = \begin{pmatrix} \mathbf{0}_{10} & \mathbf{2}_{10} & \mathbf{0}_{10}, & \mathbf{2}_{10} & \mathbf{0}_{10} & \dots & \mathbf{0}_{10} \\ \mathbf{0}_{10} & \mathbf{1}_{10} & \mathbf{0}_{10}, & \mathbf{1}_{10} & \mathbf{0}_{10} & \dots & \mathbf{0}_{10} \\ \mathbf{0}_{10} & \mathbf{0.5}_{10} & \mathbf{0}_{10}, & \mathbf{0.5}_{10} & \mathbf{0}_{10} & \dots & \mathbf{0}_{10} \end{pmatrix}. \tag{30}$$
Predictors have been simulated in the same way as in model 2.

- Model 8. We simulated data with $n = 240$ and $p = 1000$ covariates divided into 50 groups with 20 covariates each. We randomly sampled 200 observations to train the model and used the remaining 40 for comparing performance prediction. Let
$$\mathbb{B}^T = \begin{pmatrix} \mathbf{0}_{20} & \mathbf{2}_{20} & \mathbf{0}_{20}, & \mathbf{2}_{20} & \mathbf{0}_{20} & \dots & \mathbf{0}_{20} \\ \mathbf{0}_{20} & \mathbf{1}_{20} & \mathbf{0}_{20}, & \mathbf{1}_{20} & \mathbf{0}_{20} & \dots & \mathbf{0}_{20} \\ \mathbf{0}_{20} & \mathbf{0.5}_{20} & \mathbf{0}_{20}, & \mathbf{0.5}_{20} & \mathbf{0}_{20} & \dots & \mathbf{0}_{20} \end{pmatrix}. \tag{31}$$
Predictors have been simulated in the same way as in model 2.

- Model 9. We simulated data with $n = 120$ observations and $p = 130$ covariates divided into 5 groups with respectively 5, 5, 20, 50 and 50 covariates. We randomly sampled 80 observations to train the model and used the remaining 40 for comparing performance prediction. Let
$$\mathbb{B}^T = \begin{pmatrix} 0.3 & -1 & 0 & 0.5 & 0.01 & \mathbf{0}_5 & \mathbf{0.8}_{20} & \mathbf{0}_{50} & \mathbf{0}_{50} \\ 0.2 & -1.1 & 0 & 0.6 & 0.02 & \mathbf{0}_5 & \mathbf{0.7}_{20} & \mathbf{0}_{50} & \mathbf{0}_{50} \\ 0.1 & -1.2 & 0 & 0.7 & 0.03 & \mathbf{0}_5 & \mathbf{0.6}_{20} & \mathbf{0}_{50} & \mathbf{0}_{50} \end{pmatrix}. \tag{32}$$
We vary the pairwise correlation $\rho \in \{0,\ 0.5,\ 0.75\}$ between covariates. This model enables us to investigate both the effect of correlation between predictors and the group size effect. For this model we can also investigate the behavior of our methods when the sample size increases (200 and 300 observations for training and 40 observations for comparing performance prediction).

Note that models 1, 2, 6 and 9 have sparsity at the group level and also sparsity within nonzero groups while models 3, 7 and 8 have only sparsity at the group level. Models 2, 4 and 5 present the case where some relevant covariates are not related to all the responses variables.

For the first 8 models we used the *"global shrinkage parameter"* version of our MBGL-SS which has better performance in the univariate setting when all the groups have the same size. For model 9, we performed both the *"adaptive"* and *"global"* parameterizations of λ_g for our MBGL-SS model.

Numerical results

Table 5 of the Supplementary Material presents the model selection accuracy over 50 replications for methods designed for multivariate response variables. The models are compared with true and false positive rates and with Matthews correlation coefficient. For the MBGL-SS and the MBSGS-SS, both the median thresholding model (MTM) and the highest posterior probability model (HPPM) are compared. Both median thresholding model (MTM) and the highest posterior probability model (HPPM) outperform lasso methods for multivariate responses (implemented in `glmnet` and `MRCE` R packages) which does not take into account the information of the data (group) structure. As expected, the MTM model which is more parsimonious has a lower false positive rate than the HPPM model. However, the HPPM model gives better result for the true positive rate when a multivariate sparse group selection model is applied.

Models 2 and 4, correspond to the scenario where some relevant covariates are not associated with all response variables, MBGL-SS has a higher false positive rate since the method produces an estimator which gives non-zero estimates for all coefficients within a selected group regardless of the response variables. However, MBSGS-SS is not impacted by this situation because the method produces an estimator which can give zero estimates for some coefficients within a selected group. This result is highlighted in the application section.

For a large number of predictors (Models 7 and 8) compared to the number of observations, MBGL-SS has poor performance while MBSGS-SS attains very good results. In these simulations, MBGL-SS gives good results for larger sample sizes ($n = 500$ for Model 7 and $n = 900$ for Model 8). Note that MRCE methods failed dramatically in these situations and give the worst performances of all simulated models. For the current version of the MRCE optimization problem (22), the diagonal elements of the optimization variable corresponding to the error precision matrix are not penalized. Consequently, when $p > n$ a global minimizer of the penalized likelihood optimization can fail to exist. For this reason, it is not recommended to use the current MRCE approach when $p > n$.

Table 6 of the Supplementary Material presents the median mean squared prediction error for all simulation settings based on 50 replications.

The bootstrapped standard errors of the medians are given in the parentheses. BGL-SS and BSGS-SS have been performed on each response variable and we have derived and reported the median mean squared error of prediction corresponding to the multivariate response. As expected the multivariate models MBGL-SS and MBSGS-SS outperform the univariate model applied to each response variable. The MSBGS-SS also outperforms the lasso models for all simulation settings. Finally, we remark that the MSBGS-SS always performs better than the MBGL-SS and shows very good behavior when there is strong within-group sparsity (Model 6).

Results from Model 9 (corresponding to a model with different group size effect) for different pairwise correlations between predictors and different sample sizes are presented in Tables 7 and 8 of the Supplementary Material. From these tables we observe that:

- When the predictors are independent, MBGL-SS performs well both for variable selection and prediction. As expected the results improve when the sample size increases. These approaches always outperform the lasso models in this independent setting. However, the MBGL-SS methods (both the "adaptive" and "global" parameterizations of λ_g) are impacted by moderate and high correlations between predictors especially for small sample size ($n = 80$). For a larger sample size ($n = 200$), the "global shrinkage parameter" mostly gives better results for the prediction performance, but the large standard error of the median of the mean squared error of prediction indicates that some of the runs over the 50 replications gave poor results. For larger sample sizes, the models are competitive with the other approaches. We can note that the "global shrinkage parameter" always gives better results than the "adaptive shrinkage parameter" which only gives good results for the large sample size ($n = 900$, not shown in these tables).

- MBSGS-SS models outperform all the other approaches except in the case of independent predictors and small sample size. For this setting only, the method failed to select the signal of the model which is also the case for lasso models.

- As previously observed the median thresholding model (MTM) which is more parsimonious has slightly lower false positive rate than the HPPM model.

5 Application to real data

In this section, we present the results of the application of our approaches to investigate genetic regulation. To discover the genetic causes of variation in the expression (i.e. transcription) of genes, gene expression data are treated as a quantitative phenotype while genotype data (SNPs) are used as predictors, a type of analysis known as expression Quantitative Trait Loci (eQTL). Here, we use a dataset which comes from a larger study (Heinig et al. (2010)) from which we selected the Hopx genes, as in Petretto et al. (2010). This dataset has been also analyzed by Liquet et al. (2016a) who used a Bayesian model to identify a parsimonious set of predictors that explains the joint variability of gene expression in four tissues (adrenal gland, fat, heart, and kidney). However, their model does not take into account the group structure of the predictors.

	Correlation				Summary statistics	
	ADR	Fat	Heart	Kidney	Mean	Variance
ADR	1.00	0.46	0.44	0.70	4.72	0.07
Fat		1.00	0.24	0.42	8.23	0.09
Heart			1.00	0.44	8.79	1.61
Kidney				1.00	6.65	0.07

Chromosome	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Group size	74	67	63	60	39	45	52	43	31	51	21	26	33	22	15	27	18	30	34	19

Table 1: Top: pairwise correlations among the four tissues (ADR: adrenal gland, fat, heart and kidney). Bottom: repartition of the SNPs along the chromosomes.

The dataset consists of 770 SNPs in 29 inbred rats as a predictor matrix ($n = 29$, $p = 770$), and the 29 measured expression levels in the 4 tissues as the outcome ($q = 4$). A full description of the dataset is provided in Petretto et al. (2010) and available from the R package `R2GUESS` (Liquet et al. (2016a)). Table 1 presents the means, variances, and pairwise correlation structure among the four tissues, noting similar means and variances except for heart, which is larger, and moderate positive correlation. The table also shows the partitioning of the SNPs among the 20 chromosomes of the rats. Thus, the chromosome information defines the group structure of the predictor matrix.

We ran our MBGL-SS and MBSGS-SS models using this group structure and the multivariate lasso which does not take into account the group structure. The multivariate lasso selects 69 SNPs which come from the 20 chromosomes. The MBGL-SS selects only the two first groups corresponding to the SNPs from chromosomes 1 and 2. However, the simulation study showed that MBGL-SS methods are impacted by moderate and high correlations between predictors, especially for small sample sizes. Therefore, we focus our analysis on the MBSGS-SS model. Using the thresholding median estimator, the method selects 32 SNPs which belong to only 8 groups/chromosomes. Table 2 presents the posterior median estimators of the selected SNPs (meaning that the median estimator produced an estimate exactly equal to 0 for all others SNPs). Note that some SNPs in the selected chromosomes are not associated (median estimator exactly equal to 0) with all the four tissue types. Although the model does not allow for sparseness within the SNP across tissue types, that is, setting some regression parameters to 0, the median estimator does do this, as it were, for free. We note this in Table 2 for chromosomes 14, 15 and 19 in particular.

The SNP `D14Mit3` (from chromosome 10), which has been previously identified by Liquet et al. (2016a) as the most associated with the four levels of expression, is also selected by our method with the highest estimate (0.334) for the heart tissue. The four estimates for SNP `D14Mit3` for the four tissue types are substantially larger than estimates for any other SNPs. We can consider the statistical significance, estimate (posterior mean, median) compared with the posterior standard deviation for the SNP regression parameters. We note that the posterior standard deviation of the regression parameter for each selected non-zero median estimate was in the range 0.11 to 0.64. In particular, the SNP `D14Mit3` estimate was 0.334 with posterior standard deviation 0.639, a "Z-value" of about 0.5. All other SNPs with non-zero estimates have "Z-values" close to 0.0. Here the study size was small, $n = 29$, explaining to some extent the lack of power of the analysis but nevertheless when the SNP `D14Mit3` estimates are compared with the other SNPs' estimates they are substantially larger. In terms of choosing important chromosomes in addition to chromosome 10, on which SNP `D14Mit3` lies, chromosomes 2 and 7 have a larger number of non-zero SNP estimates than other chromosomes. The importance of chromosomes was investigated using an estimate of the probability of inclusion (EPI) presented in Table 3. The EPI is defined as an empirical version of l_g defined in (20) and shows the importance of chromosomes 2, 3, 4, 7, 10, and 14, all with EPI equal to 1.0.

6 Concluding remarks

In this paper, we have proposed Bayesian methods for group-sparse modeling in the context of a multivariate correlated response variable. Our models are based on spike and slab type priors which facilitate variable selection. In the case of the group variable selection, we have shown the importance of including the group size information in the shrinkage part of our model. We have shown that the posterior median estimator could both select and estimate the regression coefficients. Simulation results showed excellent performance of the posterior median estimator for variable selection and prediction error. This estimator obtains similar results as the highest probability model in terms of true and false positive rates. Moreover, this estimator can produce sparseness within the regression coefficient across the response variables even though our models have not been specifically designed for this purpose.

Simulation results also suggest that the multivariate Bayesian group lasso with spike and slab prior is strongly influenced by a combination of different group size structures and high correlation between predictors. The multivariate Bayesian sparse group selection with spike and slab prior does not suffer in this situation. This approach seems to be the most powerful method for variable selection and prediction performance in the presence of group structure data except in the extreme case of independent predictors and small sample size. All numerical results from this article can be reproduced using our R package `MBSGS` (Liquet and Sutton, 2017) available on CRAN.

We have illustrated our methods on a challenging dataset involving gene expression data ($q = 4, n = 29$) and SNP explanatory variables ($p = 770$) with the group structure ($G = 20$) defined by chromosomes. Our approach effectively found a significant SNP and chromosome while suggesting five other chromosomes could possibly be of interest. We noted the small sample size, ($n = 29$), indicating a lack of power for this study.

Our current development is restricted to non-overlapping groups. To use the present approaches with overlapping groups, such as groups of genes (like in Gene Ontology), an extension in the spirit of Stingo et al. (2011) who uses two sets of binary indicators for group and individual level selection would be required.

Chromosome	SNP Name	ADR	Fat	Heart	Kidney
2	D2Rat147	0.00553	0.00238	-	0.00329
2	D2Rat222	0.00442	0.00116	-	0.00305
2	D2CebrP476s2	0.00123	-	-	-
2	D2Rat69	0.00715	0.01748	0.00730	0.00620
2	D4Ucsf2	0.00054	-	-	-
2	D7Cebr205s3	0.00246	-	0.00950	0.00461
3	D7Cebr14C16s2	0.00209	0.00326	-	0.00049
4	D7Rat112	0.00035	0.00001	-	-
4	D7Rat19	0.01113	0.01800	0.03680	0.01828
4	Cyp11b2	0.00075	0.00374	-	0.00394
7	D10Ntr32	0.00123	-	0.01112	0.00143
7	D10Rat31	0.00031	0.00573	0.00442	0.00316
7	D10Cebr39s2	0.00280	0.00490	0.00821	0.00586
7	Es13	0.00539	-	0.00924	0.00419
7	D10Rat226	0.00415	0.00006	0.00987	0.00372
7	D14Rat36	0.00036	-	0.03076	-
7	D14Cebrp312s2	0.00004	-	0.05427	-
10	**D14Mit3**	**0.04963**	**0.05415**	**0.33434**	**0.07491**
10	D15Rat21	0.00937	0.00569	0.03140	0.01704
10	D19Utr1	0.00149	0.00297	0.00251	0.00487
10	Ednra	0.00026	-	-	-
10	D2Mit16	-	0.00077	-	-
10	D2Rat70	-	0.00190	-	-
10	D3Cebr204s4	-	0.00042	-	-
14	D4Rat49	-	0.00102	0.00092	0.00401
14	D7Mit6	-	0.00002	-	-
14	D10Rat102	-	0.00112	-	-
14	D4Rat252	-	-	-0.00184	-
14	Myc	-	-	0.00669	-
15	D10Mit3	-	-	0.00104	-
19	D14Rat8	-	-	0.00058	-
19	D14Rat52	-	-	0.00361	-

Table 2: Posterior median estimators of the selected SNPs by MBSGS-SS model on the real dataset.

In terms of computation, these methods are very practical. The current version of our package runs, for example, an MBSG-SS model in around two minutes and an MBGL-SS model in around one minute for the simulated Model 1 (Section 4.2) for a sample size ($n = 900$) with 20000 iterations including 10000 for the burnin. Further improvements of the code, such as parallelization over the group structure, are in progress to speed

Chromosome	1	2	3	4	5	6	7	8	9	10
EPI	0.00	1.00	1.00	1.00	0.72	0.00	1.00	0.00	0.00	1.00
Chromosome	11	12	13	14	15	16	17	18	19	20
EPI	0.83	0.12	0.46	1.00	0.93	0.88	0.79	0.59	0.89	0.40

Table 3: Empirical estimation of the Probability of Inclusion of each chromosome (EPI).

up the computational time for tackling Big Data sets which are common for genomics studies where predictors are embedded in a grouping structure such as gene pathways. In this context of genetics studies, some further extensions of our model are under investigation such as integrating different group penalties given a biological prior of the pathways or different distribution priors for each group.

Supplementary Material

Supplementary Material of the "Bayesian Variable Selection Regression of Multivariate Responses for Group Data" (DOI: 10.1214/17-BA1081SUPP; .pdf).

References

Babacan, S. D., Nakajima, S., and Do, M. N. (2014). "Bayesian Group-Sparse Modeling and Variational Inference." *IEEE Transactions on Signal Processing*, 62(11): 2906–2921. MR3225154. 1045

Brown, P. J., Vannucci, M., and Fearn, T. (1998). "Multivariate Bayesian Variable Selection and Prediction." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60(3): 627–641. 1053

Cai, T. T., Li, H., Liu, W., and Xie, J. (2013). "Covariate-adjusted precision matrix estimation with an application in genetical genomics." *Biometrika*, 100(1): 139. MR3034329. 1042

Dawid, A. P. (1981). "Some matrix-variate distribution theory: notational considerations and a Bayesian application." *Biometrika*, 68: 265–274. 1041

Eltoft, T., Kim, T., and Lee, T.-W. (2006). "Multivariate Scale Mixture of Gaussians Modeling." In *Independent Component Analysis and Blind Signal Separation*, 799–806. Springer, Berlin, Heidelberg. 1044

Folks, J. and Chhikara, R. (1978). "The inverse Gaussian distribution and its statistical application–a review." *Journal of the Royal Statistical Society. Series B*, 263–289. 1049

Friedman, J., Hastie, T., and Tibshirani, R. (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, 33(1): 1–22. MR1082147. 1042, 1054

Garcia, T. P., Muller, S., Carroll, R. J., and Walzem, R. L. (2014). "Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data." *Bioinformatics*, 30(6): 831–837. 1040

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA. 1049

Greenlaw, K., Szefer, E., Graham, J., Lesperance, M., and Nathoo, F. S. (2016). "A Bayesian Group Sparse Multi-Task Regression Model for Imaging Genetics." *ArXiv:1605.02234*. 1042

Heinig, M., Petretto, E., Wallace, C., Bottolo, L., Rotival, M., Lu, H., Li, Y., Sarwar, R., Langley, S., Bauerfeind, A., Hummel, O., Lee, Y., Paskas, S., Rintisch, C., Saar, K., Cooper, J., Buchan, R., Gray, E., Cyster, J., Erdmann, J., Hengstenberg, C., Maouche, S., Ouwehand, W., Rice, C., Samani, N., Schunkert, H., Goodall, A., Schulz, H., Roider, H., Vingron, M., Blankenberg, S., Munzel, T., Zeller, T., Szymczak, S., Ziegler, A., Tiret, L., Smyth, D., Pravenec, M., Aitman, T., Cambien, F., Clayton, D., Todd, J., Hubner, N., and Cook, S. (2010). "A Trans-Acting Locus Regulates an Anti-Viral Expression Network and Type 1 Diabetes Risk." *Nature*, 467(7314): 460–464. 1059

Hobert, J. P. and Geyer, C. J. (1998). "Geometric Ergodicity of Gibbs and Block Gibbs Samplers for a Hierarchical Random Effects Model." *Journal of Multivariate Analysis*, 67(2): 414–430. 1048

Huang, J., Breheny, P., and Ma, S. (2012). "A Selective Review of Group Selection in High-Dimensional Models." *Statistical Science*, 27(4): 481–499. 1040

Huang, J. and Zhang, T. (2010). "The Benefit of Group Sparsity." *Annals of Statistics*, 38(4): 1978–2004. 1040

Johnstone, I. M. and Silverman, B. W. (2004). "Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences." *Annals of Statistics*, 32(4): 1594–1649. 1045, 1046

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). "Penalized regression, standard errors, and Bayesian lassos." *Bayesian Analysis*, 5(2): 369–411. 1043, 1045

Lee, W. and Liu, Y. (2012). "Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood." *Journal of Multivariate Analysis*, 111: 241–255. 1042

Leng, C., Tran, M.-N., and Nott, D. (2014). "Bayesian adaptive Lasso." *Annals of the Institue of Statistical Mathematics*, 66(2): 221–244. 1042, 1044, 1045

Li, Y., Nan, B., and Zhu, J. (2015). "Multivariate Sparse Group Lasso for the Multivariate Multiple Linear Regression with an Arbitrary Group Structure." *Biometrics*, 71(2): 354–363. 1040, 1045, 1047

Liquet, B., Bottolo, L., Campanella, G., Richardson, S., and Chadeau–Hyam, M. (2016a). "R2GUESS: A Graphics Processing Unit-Based R Package for Bayesian Variable Selection Regression of Multivariate Responses." *Journal of Statistical Software*, 69(1): 1–32. URL https://www.jstatsoft.org/index.php/jss/article/view/v069i02. 1054, 1059, 1060

Liquet, B., Lafaye de Micheaux, P., Hejblum, B., and Thiebaut, R. (2016b). "Group and Sparse Group Partial Least Square Approaches Applied in Genomics Context." *Bioinformatics*, 3(1): 35–42. 1040

Liquet, B., Mengersen, K., Pettitt, A. N., and Sutton, M. (2017). "Supplementary Material of the "Bayesian Variable Selection Regression of Multivariate Responses for Group Data"." *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/17-BA1081SUPP. 1055

Liquet, B. and Sutton, M. (2017). *MBSGS: Multivariate Bayesian Sparse Group Selection with Spike and Slab*. R package version 1.1.0. URL http://CRAN.R-project.org/package=MBSGS. 1053, 1061

Ma, S., Song, X., and Huang, J. (2007). "Supervised Group Lasso with Applications to Microarray Data Analysis." *BMC Bioinformatics*, 8(1): 60. 1040

Meier, L., Svd, G., and Buhlmann, P. (2008). "The group lasso for logistic regression." *Journal of the Royal Statistical Society Series B*, 70(Part 1): 53–71. 1040

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. 1042, 1044

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–686. 1044

Petretto, E., Bottolo, L., Langley, S. R., Heinig, M., McDermott-Roe, C., Sarwar, R., Pravenec, M., Hübner, N., Aitman, T. J., Cook, S. A., and Richardson, S. (2010). "New Insights into the Genetic Control of Gene Expression Using a Bayesian Multi-Tissue Approach." *PLOS Computational Biology*, 6(4): e1000737. 1054, 1059, 1060

Puig, A., Wiesel, A., and Hero, A. (2009). "A multidimensional shrinkage-thresholding operator." In *Statistical Signal Processing, 2009. SSP '09. IEEE/SP 15th Workshop on*, 113–116. IEEE. 1040

Raman, S., Fuchs, T. J., Wild, P. J., Dahl, E., and Roth, V. (2009). "The Bayesian group-Lasso for Analyzing Contingency Tables." In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, 881–888. New York, NY, USA: ACM. 1045

Rockova, V. and Lesaffre, E. (2014). "Incorporating Grouping Information in Bayesian Variable Selection with Applications in Genomics." *Bayesian Analysis*, 9(1): 221–258. 1040

Rothman, A. J. (2017). *MRCE: Multivariate Regression with Covariance Estimation*. R package version 2.1. URL https://CRAN.R-project.org/package=MRCE. 1053, 1054

Rothman, A. J., Levina, E., and Zhu, J. (2010). "Sparse Multivariate Regression With Covariance Estimation." *Journal of Computational and Graphical Statistics*, 19(4): 947–962. 1042, 1053

Scheipl, F., Fahrmeir, L., and Kneib, T. (2012). "Spike-and-Slab Priors for Function Selection in Structured Additive Regression Models." *Journal of the American Statistical Association*, 107(500): 1518–1532. 1043

Simon, N. (2013). "A sparse group lasso." *Journal of Computational and Graphical Statistics*, 22(2): 231–245. 1040

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). *SGL: Fit a GLM (or cox model) with a combination of lasso and group lasso regularization*. R package version 1.1. URL `http://CRAN.R-project.org/package=SGL`. 1054

Simon, N. and Tibshirani, R. (2012). "Standarization and the group lasso penalty." *Statistica Sinica*, 22: 983–1001. 1040

Stingo, F. C., Chen, Y. A., Tadesse, M. G., and Vannucci, M. (2011). "Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes." *The Annals of Applied Statistics*, 5(3): 1978–2002. 1040, 1061

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E., and Mesirov, J. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proceedings of the National Academy of Sciences of the United States of America*, 102(43): 15545–50. 1040

Wang, H. and Leng, C. (2008). "A note on adaptive group lasso." *Computational Statistics & Data Analysis*, 52(12): 5277–5286. 1042, 1044

Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S. L., Saykin, A. J., Shen, L., and the Alzheimer's Disease Neuroimaging Initiative, F. (2012). "Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort." *Bioinformatics*, 28(2): 229–237. 1042

Wen, X. (2014). "Bayesian Model Selection in Complex Linear Systems, as Illustrated in Genetic Association Studies." *Biometrics*, 70(1): 73–83. 1042

Xu, X. and Ghosh, M. (2015). "Bayesian Variable Selection and Estimation for Group Lasso." *Bayesian Analysis*, 10(4): 909–936. 1040, 1042, 1043, 1044, 1045, 1046, 1047, 1048, 1050, 1053, 1055, 1067

Yuan, M. and Lin, Y. (2006). "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society Series B*, 68(Part 1): 49–67. 1040, 1043

Zhou, H. (2010). "Association screening of common and rare genetic variants by penalized regression." *Bioinformatics*, 26(19): 2375–2382. 1040

Zhu, H., Khondker, Z., Lu, Z., and Ibrahim, J. (2014). "Bayesian Generalized Low Rank Regression Models for Neuroimaging Phenotypes and Genetic Markers." *Journal of the American Statistical Association*, 109(507): 977–990. 1042

Acknowledgments

The authors wish to thank Xiaofan Xu and Malay Ghosh for sharing the code implemented in Xu and Ghosh (2015). The authors will also to thank Dr. Marc Chadeau–Hyam for his valuable inputs and comments on the paper and the model. The authors would also like to acknowledge the ARC Centre of Excellence for Mathematical and Statistical Frontiers for supporting this research. Finally, we thank the High Performance Computing and Research Support (HPC) group from Queensland University Technology.