Bayesian indicator variable selection model to incorporate multi-layer overlapping group structure in multi-omics applications

Abstract

Variable selection is a pervasive question in modern high-dimensional data analysis where the number of features often exceeds the sample size (a.k.a. small-n-large-p problem). Incorporation of group structure knowledge to improve variable selection has been widely studied. Here, we consider prior knowledge of a multi-layer overlapping group structure to improve variable selection in regression setting. In genomic applications, for instance, a biological pathway contains tens to hundreds of genes and a gene can contain multiple experimentally measured features (such as its mRNA expression, copy number variation and methylation levels of possibly multiple sites). In addition to the multi-layer structure, the groups may be overlapped (e.g. two pathways can share common genes). Incorporating such hierarchical multi-layer overlapping groups in traditional penalized regression setting produces difficulty in optimization. In this paper, we propose a Bayesian indicator model that can elegantly serve the purpose. We discuss the soft-thresholding property of the posterior median estimator and prove its selection consistency and asymptotic normality under orthogonal design. We apply the model to simulations and two breast cancer examples to demonstrate its superiority over other existing methods. The result not only enhances prediction accuracy but also improves variable selection and model interpretation.

Keywords: Bayesian variable selection; multi-layer group structure; overlapping groups; spike and slab.

1 Introduction

Variable selection is a pervasive question in statistical applications, intended to search for the best models by eliminating unnecessary features. It gains increasing attention particularly in high dimensional data analysis, where the number of features often greatly exceeds the number of samples. In this scenario, it is commonly believed that only a small set of features have large effect on the outcome, while most other features have little or no effect. In the literature, penalized regression methods such as lasso regression (Tibshirani, 1996) have used L1-norm penalty to achieve shrinkage estimation and variable selection. It is well-known that the L1-norm penalty in lasso tends to randomly select one out of a set of highly correlated variables and ignore the others. Zou and Hastie (Zou and Hastie, 2005) proposed an elastic net method with combination of L1 and L2 norm penalty to overcome the problem. When prior information of grouped variables is available and variable selection by groups is desirable, Yuan and Lin (2006) proposed a group lasso penalty to select groups of variables into or out of the model together. In order to further allow sparsity within groups while they are selected, Simon et al. (2013) proposed a sparse group lasso with both L1 penalty and group lasso penalty. In the counterpart of Bayesian framework, variable selection can be viewed as identification of nonzero variables (or elimination of variables very close to zero) in the posterior distribution. Tibshirani (1996) pointed out that lasso estimator is equivalent to posterior median of a Gaussian model using double exponential (Laplace) prior for each variable. Inspired by the hierarchical structure of Laplace prior, Park and Casella (2008) proposed a full Bayesian lasso model and Kyung et al. (2010) further derived the Bayesian version for group lasso and elastic net. Mitchell and Beauchamp (1988) proposed another popular type of "spike and slab" prior with a mixture of a point mass at 0 (or a distribution centered around zero with small variance) and a diffuse uniform or large variance distribution (see also George and McCulloch (1993) and Kuo and Mallick (1998)). Hernández-Lobato et al. (2013) generalized spike-and-slab prior for group feature selection and implemented the expectation propagation algorithm. Xu et al. (2015) and Zhang et al. (2014) extended spike-and-slab to achieve sparsity both at group level and within groups. Under mild conditions, the posterior median estimator for a normal mean sample with spike-and-slab prior is a soft-thresholding estimator with desired selection consistency and asymptotic normality properties (Johnstone and Silverman, 2004; Xu et al., 2015).

All aforementioned methods in penalized regression and Bayesian framework allow only non-overlapping and single layer group structure. In this paper, we consider a motivating example that require incorporation of multi-layer and overlapping group structure. Suppose n tumor tissues are collected. SNP array, methylation array, miRNA array and RNA-seq are performed on these tissues to obtain genome-wide copy number variation (CNV), methylation, miRNA and mRNA expression measurements. Integration of such multi-level omics data has become prevalent in the research of many diseases and brought new statistical challenges (see Richardson et al. (2016) for review). Assume that our total number of variables p is the union of all CNV, methylation sites, miRNA and mRNA expression features. Figure 1 shows an example of multi-layer and overlapping group structure on the p variables. In the first layer of groups, four features belong to the gene A group: mRNA, CNV and methylation probe of gene A, and miRNA α that targets on gene A (knowledge known a priori from miRNA target database). Similarly, gene B group contains three multi-omics features related to itself and miRNA α also targets on this gene. Groups gene A and gene B at the first layer is an example of overlapping group structure as they are both targeted by miRNA α. In the second layer, Pathway θ contains these two genes. Another pathway ϕ contains gene B, C and D. As a result, pathways θ and ϕ represent overlapping grouping at the second layer as they both contain gene B. In summary, the grouping structure of "multi-omics features $\Rightarrow$ gene $\Rightarrow$ pathway" demonstrates a hierarchical (multi-layer) and overlapping group structure that brings challenges for variable selection. Here, we avoid using the term "hierarchical" but replace with "multi-layer" hereafter since all Bayesian models are hierarchical generative models and the term can be confusing. In the literature for overlapping group structure, Jacob et al. (2009) modified the group lasso penalty, which led to a sparse solution with support as a union of groups. Zhang et al. (2014) decomposed the marginal regression coefficients of a feature overlapped in multiple groups to be summation of partial effects. With the new multi-layer and overlapping grouping structure, the target function of penalized regression approaches generally becomes intractable to optimize. Bayesian hierarchical model provides a natural alternative to incorporate the multi-layer hierarchical structure. We propose a multi-layer indicator variable selection model extended from Kuo and Mallick (1998) where three levels of binary indicators illustrate whether the corresponding multi-

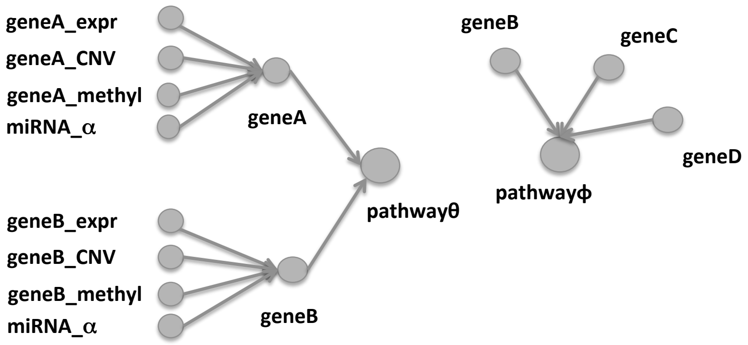


Figure 1: An example of multi-layer overlapping group structure in multi-omics dataset

omics features, genes or pathways are selected. For overlapping groups, we adopt from Zhang et al. (2014) the additive effect assumption for each overlapping group. We will show that incorporation of the multi-layer overlapping group structure enhances prediction accuracy and improve feature selection and model interpretation.

The paper is organized as follows. In Section 2, we review the indicator variable selection model, and propose Bayesian indicator variable selection model with single-layer and multi-layer overlapping group structures. We describe the detailed MCMC algorithms for each model and extend the models for binary and survival outcomes. In Section 3, we prove variable selection consistency and asymptotic normality of parameter estimation from the posterior median estimator under orthogonal design. In Section 4, three simulations are demonstrated to compare the performance of proposed models with other existing methods. We further apply the model to two real examples of using a breast cancer data set to predict estrogen receptor (ER) status and histology subtype (invasive lobular carcinoma (ILC) versus invasive ductal carcinoma (IDC)) in Section 5. Section 6 contains final conclusion and discussion.

2 Methods

2.1 Review of indicator variable selection model

Consider a regression setting, in which $Y = (Y_1, ..., Y_n)^T$ denotes the outcomes for n samples, and X denotes an $n \times p$ covariate matrix for p variables. Under linear regression assumption, $Y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, and $i = 1, \ldots, n$.

Bayesian indicator variable selection model was first proposed in Kuo and Mallick (1998). It embeds binary indicators into regression model to incorporate all 2^p candidate models. Assuming the data are centered and denoting the binary indicator as γ_j, the indicator variable selection model is

$$\begin{aligned} Y_i &= \sum_{j=1}^{p} \beta_j \gamma_j x_{ij} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \\ \beta &= (\beta_1, \ldots, \beta_p)^T \sim N(0, D_0), \quad \gamma_j \sim Bern(\pi). \end{aligned}$$

If $D_0 = s^2 I_{p\times p}$ is a diagonal matrix, where $I_{p\times p}$ is an identity matrix with dimension $p \times p$, and we define $\beta_j^* = \beta_j \gamma_j$, the indicator prior is equivalent to a spike-and-slab prior:

$$\beta_j^* \sim (1 - \pi)\delta_0(\cdot) + \pi N(0, s^2),$$

where $\delta_a(\cdot)$ is a Dirac delta function putting all mass at a.

This method is free of tuning and can be easily extended to more complicated modeling, such as model with interactions. However, if the prior is too vague, mixing can be poor, as the sampled values of β_j may only rarely be in the region with high posterior support (O'Hara et al., 2009). Other alternatives have been proposed (George and McCulloch, 1993), but most of them require additional tuning parameters.

2.2 SOG: Bayesian variable selection with single layer overlapping groups

Motivated by the indicator variable selection model, we propose a Bayesian indicator variable selection model with multi-layer overlapping groups (MOG). We first introduce a simple version with only single layer overlapping groups (SOG).

Under the same linear regression setting in section 2.1, we assume p variables (level-0 variables) belonging to M_1 possibly overlapping groups (level-1 groups). For instance, p experimentally measured features belong to m_1 genes. We define a $p \times m_1$ matrix $U^{(1)}$ to denote the group membership of level-0 variables, with $U^{(1)}_{j,k} = 1$ denoting that level-0 variable j belongs to level-1 group k, and $U^{(1)}_{j,k} = 0$ otherwise. We propose the following model:

$$\begin{gathered} Y_i \sim \left(\sum_{j=1}^{p} \sum_{k=1}^{m_1} x_{ij} \beta_{jk} U^{(1)}_{jk}, \sigma^2 \right), \\ (\beta_{jk} | U^{(1)}_{jk} = 1) = \gamma^{(1)}_k \gamma^{(0)}_{jk} b_{jk}, \quad (\beta_{jk} | U^{(1)}_{jk} = 0) \sim \delta_0(\cdot), \\ \gamma^{(1)}_k \sim Bern(\pi^{(1)}), \quad \gamma^{(0)}_{jk} \sim Bern(\pi^{(0)}_k), \\ b_{jk} \sim N(0, s^2), \quad \sigma^2 \propto 1/\sigma^2, \end{gathered}$$

where $\gamma^{(1)}_k$ can be interpreted as the selection indicator for level-1 group k; if $\gamma^{(1)}_k = 1$, $\gamma^{(0)}_{jk}$ can be interpreted as the selection indicator for level-0 variable j belonging to level-1 group k; $\beta_{jk} \neq 0$ if and only if $\gamma^{(1)}_k = 1$ and $\gamma^{(0)}_{jk} = 1$. Singleton will be treated as a group with itself as its only member.

Remarks:

(1) $U^{(1)}$ is a sparse matrix, most of whose entries are 0's and a few are 1's. $\sum\limits_{k=1}^{m_1} U^{(1)}_{jk}$ is the number of level-1 groups that level-0 variable j belongs to. If $\sum\limits_{k=1}^{m_1} U^{(1)}_{jk} > 1$, level-0 variable j belongs to multiple groups. $\sum\limits_{j=1}^{p} U^{(1)}_{jk}$ is the number of level-0 variables that belong to level-1 group k. If $\sum\limits_{j=1}^{p} U^{(1)}_{jk} U^{(1)}_{jk'} \geq 1$, level-1 groups k and k' overlap.

(2) β is also a $p \times m_1$ sparse matrix, with $\beta_{jk} \neq 0$ only when $U^{(1)}_{jk} = 1$.

(3) If level-0 variable j belongs to multiple groups, each individual coefficient β_{jk} with $U^{(1)}_{jk} = 1$ is not identifiable, but $\beta_j = \sum\limits_{k=1}^{m_1} \beta_{jk}$ is identifiable (Zhang et al., 2014). This latent decomposition is also assumed in Jacob et al. (2009) and Zhang et al. (2014), and it is equivalent to duplicate the features in each group that it belongs to. We apply such duplication technique to other Bayesian methods (such as BSGS-SS and HSVS, which are existing methods we will compare in simulation and real applications) to allow overlapping group structure.

Table 1: Method comparison

Method	Feature selection	Exact zero in feature selection	Group selection	Exact zero in group selection	Varying sparsity inside groups	Reference
MOG (SOG)	✓	✓	✓	✓	✓	
BSGS-SS	✓	✓	✓	✓	✗	Xu (2015)
HSVS	✓	✗	✓	✓	✓	Zhang (2014)
SGL	✓	✓	✓	✓	✗	Simon (2013)
GL	✗	✗	✓	✓	✗	Zou (2005)
Lasso	✓	✓	✗	✗	✗	Tibshirani (1996)

(4) Our model allows group-specific proportions of non-zero β's (i.e. $\pi_k^{(0)}$, $1 \leq k \leq M_1$), which is more realistic in applications. BSGS-SS explicitly assumes the common non-zero proportion in selected groups, and SGL also implicitly assumes common non-zero proportion in each group (see more detailed comparison in table 1).

We assign non-informative hyper-priors: $\pi^{(1)} \sim Beta(1,1)$, $\pi_k^{(0)} \sim Beta(1,1)$, and $s^2 \propto 1/s^2$. With model constructed above, we have the following full conditionals, allowing for Gibbs sampling:

$$Pr(\gamma_k^{(1)} = 1|-) = 1/\left(1 + \frac{(1-\pi^{(1)})}{1-\pi^{(1)}} \exp\left\{\frac{1}{\sigma^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2 \beta_{jk}^2 (U_{jk}^{(1)})^2 - 2\sum_{i=1}^{n} y_{i,k} \sum_{j=1}^{p} x_{ij}\beta_{jk}U_{jk}^{(1)}\right)\right\}\right),$$

$$Pr(\gamma_{jk}^{(0)} = 1|U_{jk}^{(1)} = 1, -) = 1/\left(1 + \frac{(1-\pi_k^{(0)})}{\pi_k^{(0)}} \exp\left\{\frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_{ij}^2 \beta_{jk}^2 (U_{jk}^{(1)})^2 - 2\sum_{i=1}^{n} y_{i,(jk)} x_{ij}\beta_{jk}U_{jk}^{(1)}\right)\right\}\right),$$

$$(b_{jk}|\gamma_k^{(1)}\gamma_{jk}^{(0)} = 0, U_{jk}^{(1)} = 1, -) \sim N(0, s^2),$$

$$(b_{jk}|\gamma_k^{(1)}\gamma_{jk}^{(0)} = 1, U_{jk}^{(1)} = 1, -) \sim N\left(\mu_b = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_{ij}y_{i,(jk)})\sigma_b^2, \sigma_b^2 = \frac{1}{\frac{\sum_{i=1}^{n} x_{ij}^2}{\sigma^2} + \frac{1}{s^2}}\right),$$

$$\sigma^2|- \sim \text{Inverse-Gamma}\left(n/2, 1/2\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\sum_{k=1}^{m_1} x_{ij}\beta_{jk}U_{jk}^{(1)})^2\right),$$

$$s^2|- \sim \text{Inverse-Gamma}\left(\sum_{j=1}^{p}\sum_{k=1}^{m_1} U_{jk}^{(1)}/2, 1/2\sum_{j=1}^{p}\sum_{k=1}^{m_1} b_{jk}^2\right),$$

$$\pi^{(1)}|- \sim Beta\left(\sum_{k=1}^{m_1}\gamma_k^{(1)} + 1, m_1 - \sum_{k=1}^{m_1}\gamma_k^{(1)} + 1\right),$$

$$\pi_k^{(0)}|- \sim Beta\left(\sum_{j=1}^{p}\gamma_{jk}^{(0)}+1, \sum_{j=1}^{P}U_{jk}^{(1)}-\sum_{j=1}^{p}\gamma_{jk}^{(0)}+1\right),$$

where "$-$" stands for all other variables, $y_{i,k} = y_i - \sum_{j=1}^{p}\sum_{k'\neq k} x_{ij}\beta_{jk'}U_{jk'}^{(1)}$, and $y_{i,(jk)} = y_i - \sum_{j'\neq j}\sum_{k=1}^{m_1} x_{ij}\beta_{j'k}U_{j'k}^{(1)} - \sum_{k'\neq k} x_{ij}\beta_{jk'}U_{jk'}^{(1)}$.

2.3 MOG: Bayesian variable selection with multi-layer overlapping groups

In the presence of multi-layer (say s layers) overlapping groups, we define $U^{(1)}$, ..., $U^{(s)}$, each with dimension $p \times m_1$, $m_1 \times m_2$, ..., $m_{s-1} \times m_s$ respectively, to specify the group structures. The multi-level omics data example in the introduction (Fig 1) corresponds to a structure with $s = 2$. In that case, $U_{jk}^{(1)} = 1$ denotes that level-0 variable j (e.g. ESR1 mRNA expression) belongs to level-1 group k (e.g. ESR1 gene); $U_{kl}^{(2)} = 1$ represents that level-1 group k (e.g. ESR1 gene) belongs to level-2 group l (e.g., ER signaling pathway). Below, we use $s = 2$ to illustrate the motivating example but the model can be extended to $s > 2$. The proposed model for 2-layer overlapping groups is

$$\begin{gathered}
Y_i \sim \left(\sum_{j=1}^{p}\sum_{k=1}^{m_1}\sum_{l=1}^{m_2} x_{ij}\beta_{jkl}U_{jk}^{(1)}U_{kl}^{(2)}, \sigma^2\right), \\
(\beta_{jkl}|U_{jk}^{(1)}U_{kl}^{(2)} = 1) = \gamma_l^{(2)}\gamma_{kl}^{(1)}\gamma_{jkl}^{(0)}b_{jkl}, \quad (\beta_{jkl}|U_{jk}^{(1)}U_{kl}^{(2)} = 0) \sim \delta_0(\cdot), \\
\gamma_l^{(2)} \sim Bern(\pi^{(2)}), \quad \gamma_{kl}^{(1)} \sim Bern(\pi_l^{(1)}), \quad \gamma_{jkl}^{(0)} \sim Bern(\pi_{kl}^{(0)}), \\
b_{jkl} \sim N(0, s^2), \quad \sigma^2 \propto 1/\sigma^2,
\end{gathered}$$

where $\gamma_l^{(2)}$ can be interpreted as the selection indicator for level-2 group l; if $\gamma_l^{(2)} = 1$, $\gamma_{kl}^{(1)}$ can be interpreted as the selection indicator for level-1 group k belonging to level-2 group l; if $\gamma_l^{(2)}\gamma_{kl}^{(1)} = 1$, $\gamma_{jkl}^{(0)}$ can be interpreted as the selection indicator for level-0 variable j belonging to level-1 group k and level-2 group l; $\beta_{jkl} \neq 0$ if and only if $\gamma_l^{(2)} = 1$, $\gamma_{jk}^{(1)} = 1$, and $\gamma_{jkl}^{(0)} = 1$.
Remarks: With s layers of overlapping groups, we define $V^{(s_1)(s_2)} = \prod_{t=s_1}^{s_2} U^{(t)}$. $V^{(s_1)(s_2)}$ is a $m_{s_1} \times m_{s_2}$ matrix, with $V_{jq}^{(s_1)(s_2)} \geq 1$ indicating that level-s_1 group j belongs to level-s_2 group q; $V_{jq}^{(s_1)(s_2)} = 0$, otherwise.

We assign hyper-priors: $\pi^{(2)} \sim Beta(1,1)$, $\pi_l^{(1)} \sim Beta(1,1)$, $\pi_{kl}^{(0)} \sim Beta(1,1)$, and $s^2 \propto 1/s^2$. The full conditionals are:

$$Pr(\gamma_l^{(2)} = 1|-) = 1/\left(1 + \frac{(1-\pi^{(2)})}{\pi^{(2)}} \exp\left\{\left(\sum_{i,j,k} x_{ij}^2 \beta_{jkl}^2 (U_{jk}^{(1)} U_{kl}^{(2)})^2 - 2\sum_{i=1}^{n} y_{i,l} \sum_{j,k} x_{ij}\beta_{jkl} U_{jk}^{(1)} U_{kl}^{(2)}\right)/\sigma^2\right\}\right),$$

$$Pr(\gamma_{kl}^{(1)} = 1|U_{kl}^{(2)} = 1, -) = 1/\left(1 + \frac{(1-\pi_l^{(1)})}{\pi_l^{(1)}} \exp\left\{\left(\sum_{i,j} x_{ij}^2 \beta_{jkl}^2 (U_{jk}^{(1)} U_{kl}^{(2)})^2 - 2\sum_{i=1}^{n} y_{i,(kl)} \sum_{j=1}^{p} x_{ij}\beta_{jkl} U_{jk}^{(1)} U_{kl}^{(2)}\right)/\sigma^2\right\}\right),$$

$$Pr(\gamma_{jkl}^{(0)} = 1|U_{jk}^{(1)} U_{kl}^{(2)} = 1, -) = 1/\left(1 + \frac{(1-\pi_{kl}^{(0)})}{\pi_{kl}^{(0)}} \exp\left\{\left(\sum_{i=1}^{n} x_{ij}^2 \beta_{jkl}^2 (U_{jk}^{(1)} U_{kl}^{(2)})^2 - 2\sum_{i=1}^{n} y_{i,jkl} x_{ij}\beta_{jkl} U_{jk}^{(1)} U_{kl}^{(2)}\right)/\sigma^2\right\}\right),$$

$$(b_{jkl}|\gamma_l^{(2)}\gamma_{kl}^{(1)}\gamma_{jkl}^{(0)} = 0, U_{jk}^{(1)} U_{kl}^{(2)} = 1, -) \sim N(0, s^2),$$

$$(b_{jkl}|\gamma_l^{(2)}\gamma_{kl}^{(1)}\gamma_{jkl}^{(0)} = 1, -) \sim N\left(\mu_b = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_{ij} y_{i,jkl})\sigma_b^2, \sigma_b^2 = \frac{1}{\frac{\sum_{i=1}^{n} x_{ij}^2}{\sigma^2} + \frac{1}{s^2}}\right),$$

$$\sigma^2|- \sim \text{IG}\left(n/2, 1/2\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\sum_{k=1}^{m_1}\sum_{l=1}^{m_2} x_{ij}\beta_{jkl} U_{jk}^{(1)} U_{kl}^{(2)})^2\right),$$

$$s^2|- \sim \text{Inverse-Gamma}\left(\sum_{j=1}^{p}\sum_{k=1}^{m_1}\sum_{l=1}^{m_2} U_{jk}^{(1)} U_{kl}^{(2)}/2, 1/2\sum_{j=1}^{p}\sum_{k=1}^{m_1}\sum_{l=1}^{m_2} b_{jkl}^2\right),$$

$$\pi^{(2)}|- \sim Beta\left(\sum_{l=1}^{m_2}\gamma_l^{(2)} + 1, M_1 - \sum_{l=1}^{m_2}\gamma_l^{(2)} + 1\right),$$

$$\pi_l^{(1)}|- \sim Beta\left(\sum_{k=1}^{m_1}\gamma_{kl}^{(1)} + 1, \sum_{k=1}^{m_2} U_{kl}^{(2)} - \sum_{k=1}^{m_1}\gamma_{kl}^{(1)} + 1\right),$$

$$\pi_{kl}^{(0)}|U_{kl}^{(2)} = 1, - \sim Beta\left(\sum_{j=1}^{p}\gamma_{jkl}^{(0)} + 1, \sum_{j=1}^{p} U_{jk}^{(1)} U_{kl}^{(2)} - \sum_{k=1}^{p}\gamma_{jkl}^{(0)} + 1\right)$$

where $y_{i,l} = y_i - \sum_{j=1}^{p}\sum_{k=1}^{m_1}\sum_{l'\neq l} x_{ij}\beta_{jkl'} U_{jk}^{(1)} U_{kl'}^{(2)}$, $y_{i,kl} = y_i - \sum_{j=1}^{p}\sum_{k'=1}^{m_1}\sum_{l'\neq l} x_{ij}\beta_{jk'l'} U_{jk'}^{(1)} U_{k'l'}^{(2)}$-$\sum_{j=1}^{p}\sum_{k'\neq k} x_{ij}\beta_{jk'l} U_{jk'}^{(1)} U_{k'l}^{(2)}$, and $y_{i,jkl} = y_i - \sum_{j=1}^{p}\sum_{k'=1}^{m_1}\sum_{l'\neq l} x_{ij}\beta_{jk'l'} U_{jk'}^{(1)} U_{k'l'}^{(2)} - \sum_{j=1}^{p}\sum_{k'\neq k} x_{ij}\beta_{jk'l} U_{jk'}^{(1)} U_{k'l}^{(2)} - \sum_{j'\neq j} x_{ij}\beta_{j'kl} U_{j'k}^{(1)} U_{kl}^{(2)}$.

2.4 Extension to binary and survival outcome

For binary outcome, we adopt the data augmentation from Albert and Chib (1993) introducing latent variable Z_i $(i = 1, ..., n)$ for each Y_i, and replace Y_i in the regression,

$$Y_i = \begin{cases} 1, & \text{if } Z_i \geq 0 \\ 0, & \text{otherwise} \end{cases}, \quad Z_i = x_i^T\beta + \varepsilon_i, \quad \varepsilon_i \sim N(0,1)$$

For survival outcome, we applied similar data augmentation (Tanner and Wong, 1987) for accelerated failure time (AFT) model, introducing a latent variable Z_i for each time to event t_i and censor indicator δ_i ($\delta_i = 1$ indicating event happened):

$$\begin{cases} \log(t_i) = Z_i, & \text{if } \delta_i = 1 \\ \log(t_i) < Z_i, & \text{if } \delta_i = 0 \end{cases}, \quad Z_i = x_i^T \beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

2.5 Comparison to other existing methods

We compare our method to two existing Bayesian methods BSGS-SS (Xu et al., 2015) and HSVS (Zhang et al., 2014), both of which can perform variable selection at group and within group level. When overlapping groups exist, the duplication approach is applied for all Bayesian methods, under the same assumption. We also compare our method to lasso (Tibshirani, 1996), group lasso (GL) (Zou and Hastie, 2005), and sparse group lasso (SGL) (Simon et al., 2013) under non-overlapping group structure. The mixing weight α in SGL is set to be 0.95 as default, thus more similar to lasso. See a detailed comparison in Table 1. The performance is evaluated by accuracy of both variable selection and prediction. In all the simulations and applications, data are split into 5 folds, with 4 folds as training and one fold as testing set.

In terms of variable selection, when the true β is known in simulation, since all methods requires either a tuning parameter or a cutoff, we derive sensitivity and specificity of variable selection under different cutoffs and calculate area under the receiver operating curves (AUC) for a fair comparison. MOG (SOG) and BSGS-SS can get exact zero estimates inside groups in each MCMC iteration, so features (level-0 variables) can be sorted according to posterior mean of the selection probability, i.e. $P(\beta_{jkl} \neq 0|y)$ (when S=2). HSVS uses Laplace prior within group and cannot obtain exact zero inside a group if the group is selected, even though the estimates are shrunken towards zero. As a result, we sort the features based on $\max(p_{pos}, 1 - p_{pos})$, where p_{pos} is the posterior mean of $P(\beta_{jkl} > 0|y)$ (for S=2). For Lasso, GL, and SGL, we apply multiple tuning parameters (default sequence provided by R packages) that detect different numbers of variables and form the basis of ROC curve. For MOG (SOG) and BSGS-SS, we also apply Bayesian false discovery rate (BFDR, Newton et al. (2004)) controlled at 10% to compare their FDRs and false omission rates (FORs=the number of false negatives/number of negative calls).

Regarding prediction accuracy, all four Bayesian methods (MOG, SOG, BSGS-SS, and HSVS) use posterior median estimates for evaluation. Lasso, GL, and SGL apply 10-fold cross-validation to select tuning parameters. With continuous outcome, we compare prediction mean squared error (MSE) in the testing set, i.e. $MSE = \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} (x_{test,i}^T \hat{\beta} - y_{test,i})^2$, where n_{te} is the sample size in the testing set and $y_{test,i}$ is the i-th observation in the testing set. If the outcome is binary, we sort the samples in the testing set based on the predicted probability and calculate prediction AUC.

R is used to implement all methods. All Bayesian methods use 2000 MCMC iterations with 1000 as burn-in in simulations, and 20,000 MCMC iterations with 10,000 as burn-in in applications. Trace plots are used for convergence diagnostics. We apply R packages MBSGS, glmnet, grplasso, and SGL for BSGS-SS, lasso, GL and SGL, respectively. R function for HSVS is provided by authors.

3 Asymptotic property of posterior median estimator under orthogonal design

Johnstone and Silverman (2004) showed that the posterior median of normal mean model with the spike-and-slab prior is a soft-thresholding estimator. Xu et al. (2015) extended the thresholding results to multi-variate spike-and-slab. Under orthogonal design (i.e. $X^T X = nI$), the posterior median of MOG (SOG) is also a soft-thresholding estimator. We use the same notations as in Xu et al. (2015).

Lemma 1. *Assuming orthogonal design (i.e. $X^T X = nI$) and all levels of groups are disjoint, given $\pi^{(0)}$, $\pi^{(1)}$, $\pi^{(2)}$, s^2, σ^2 are fixed, the posterior median $\hat{\beta}_{jkl}^{Med}$ in MOG is a soft-thresholding estimator.*

Proof: The marginal prior of β_{jkl} in MOG is a “spike-and-slab”:

$$\beta_{jkl} \sim (1 - \pi^{(2)} \pi_l^{(1)} \pi_{kl}^{(0)}) \delta_0(\beta_{jkl}) + \pi^{(2)} \pi_l^{(1)} \pi_{kl}^{(0)} N(0, s^2).$$

Given orthogonal design and $\pi^{(0)}$, $\pi^{(1)}$, $\pi^{(2)}$, s^2, σ^2 are fixed, the posterior distribution of β_{jkl} is:

$$\beta_{jkl}|Y, X \sim (1 - r_{jkl}) \delta_0(\beta_{jkl}) + r_{jkl} N\left((1 - B_n) \hat{\beta}_{jkl}^{LS}, \sigma^2 (1 - B_n)/n\right),$$

where $\hat{\beta}^{LS}_{jkl}$ is the ordinary least square estimator, B_n=$\frac{\sigma^2}{ns^2+\sigma^2}$, and r_{jkl}=$P(\beta_{jkl} \neq 0|Y,X) = \pi^{(2)}\pi_l^{(1)}\pi_{kl}^{(0)} / \left(\pi^{(2)}\pi_l^{(1)}\pi_{kl}^{(0)} + (1-\pi^{(2)}\pi_l^{(1)}\pi_{kl}^{(0)})(1+ns^2/\sigma^2)^{1/2}\exp\{-\frac{(1-B_n)}{2\sigma^2}n(\hat{\beta}^{LS}_{jkl})^2\}\right)$. It can be easily noticed that $P(\beta_{jkl} \neq 0|Y,X) \rightarrow 1$ as $|\hat{\beta}^{LS}_{jkl}| \rightarrow +\infty$, when fix n; $P(\beta_{cgj} \neq 0|Y,X) \rightarrow 0$ as $n \rightarrow \infty$, when $\hat{\beta}^{LS}_{jkl} = 0$. Assuming n is sufficiently large such that $P(\beta_{jkl} \neq 0|Y,X) < 0.5$ when $\hat{\beta}^{LS}_{jkl} = 0$, then there exists a threshold t, such that if $|\hat{\beta}^{LS}_{jkl}| \leq t$, $P(\beta_{jkl} \neq 0|Y,X) < 0.5$, thus $\hat{\beta}^{Med}_{jkl}$=0; and if $|\hat{\beta}^{LS}_{jkl}| > t$, $P(\beta_{jkl} \neq 0|Y,X) > 0.5$, and $\hat{\beta}^{Med}_{jkl} \neq 0$. In MOG, with normal prior and normal likelihood, the threshold t can be solved analytically, and the posterior median is a soft-thresholding estimator given by

$$\hat{\beta}^{Med}_{jkl} = sgn(\hat{\beta}^{LS}_{jkl}) \left((1-B_n)|\hat{\beta}^{LS}_{jkl}| - \frac{\sigma}{\sqrt{n}}\sqrt{1-B_n}\Phi^{-1}\left(\frac{1}{2(1-min(\frac{1}{2}, r_{jkl}))}\right)\right)_+ ,$$

where sgn is the sign function, Φ is the cumulative distribution function (CDF) of standard normal distribution, and $(x)_+$ takes the value of x if $x > 0$, and zero otherwise.

Theorem 1. *Define an index vector of true non-zero features* $\mathcal{A} = (I(\beta^0_{jkl} \neq 0), j = 1,..,p; k = 1,..,m_1; l = 1,...,m_2)$ *and* $\mathcal{A}_n$ *for index vector from posterior median estimator. Under the assumption that* $X^TX = nI$, $\sqrt{n}s^2/\sigma^2 \rightarrow \infty$ *and* $\log(s^2/\sigma^2)/n \rightarrow 0$ *as* $n \rightarrow 0$,

$$\lim_{n\rightarrow\infty} P(\mathcal{A}_n = \mathcal{A}) = 1 \qquad \text{(Selection consistency)}$$

$$\sqrt{n}(\hat{\beta}^{Med} - \beta^0) \rightarrow N(0, \sigma^2 I) \qquad \text{(Asymptotic normality)}$$

Proof: When $\beta^0_{jkl} = 0$, since $\sqrt{n}\hat{\beta}^{LS}_{jkl} = O_p(1)$, and $\sqrt{n}B_n \rightarrow 0$, we get $r_{jkl} \xrightarrow{\text{p}} 0$ and $1/\Phi^{-1}\left(\frac{1}{2(1-min(\frac{1}{2},1-r_{jkl})}\right) \xrightarrow{\text{p}} 0$. Therefore,

$$P(\hat{\beta}^{Med}_{jkl} = 0) = P\left(\frac{\sqrt{1-B_n}}{\sigma\Phi^{-1}\left(\frac{1}{2(1-min(\frac{1}{2},1-r_{jkl})}\right)}\sqrt{n}|\hat{\beta}^{LS}_{jkl}| \leq 1\right) \rightarrow 1.$$

When $\beta^0_{jkl} \neq 0$, $\hat{\beta}^{LS}_{jkl} \xrightarrow{\text{p}} \beta^0_{jkl}$ and $\log(s^2/\sigma^2)/n \rightarrow 0$, thus $r_{jkl} \xrightarrow{\text{p}} 1$ and $\Phi^{-1}\left(\frac{1}{2(1-min(\frac{1}{2},1-r_{jkl})}\right) \xrightarrow{\text{p}} 0$. Therefore,

$$P(\hat{\beta}^{Med}_{jkl} \neq 0) = P\left(\frac{\sigma\Phi^{-1}\left(\frac{1}{2(1-min(\frac{1}{2},1-r_{jkl})}\right)}{\sqrt{1-B_n}\sqrt{n}|\hat{\beta}^{LS}_{jkl}|} < 1\right) \rightarrow 1.$$

This concludes the selection consistency of each coefficient $\hat{\beta}^{Med}_{jkl}$. Under orthogonal design, we get selection consistency for all coefficients.

To prove the asymptotic normality, we consider each coefficient first,

$$|\sqrt{n}(\hat{\beta}^{Med}_{jkl} - \hat{\beta}^{LS}_{jkl})|$$

$$= \left(\sqrt{n}B_n|\hat{\beta}_{jkl}^{LS}| - \sigma\sqrt{1-B_n}\Phi^{-1}\left(\frac{1}{2(1-min(\frac{1}{2}, 1-r_{jkl}))}\right)\right) I(\hat{\beta}_{jkl}^{Med} \neq 0)$$
$$+\sqrt{n}\hat{\beta}_{jkl}^{LS} I(\hat{\beta}_{jkl}^{Med} = 0) \xrightarrow{\text{p}} 0.$$

Since $\sqrt{n}(\hat{\beta}^{LS} - \beta^0) \xrightarrow{\text{d}} N(0, \sigma^2 I)$, by Slutsky theorem, $\sqrt{n}(\hat{\beta}^{Med} - \beta^0) \xrightarrow{\text{d}} N(0, \sigma^2 I)$. ■

If groups overlap, here we assume one level-0 variable is shared by two level-1 groups, each of which belongs to a level-2 group for simplicity, i.e., $\beta_j = \beta_{jkl} + \beta_{jk'l'}$ ($k' \neq k$ and $l' \neq l$). Even though each individual partial effect is not identifiable, the marginal effect β_j is identifiable.

Theorem 2. *Assuming orthogonal design (i.e. $X^T X = nI$), given $\pi^{(0)}$, $\pi^{(1)}$, $\pi^{(2)}$, s^2, σ^2 are known, the posterior median $\hat{\beta}_j^{Med}$ is a soft-thresholding estimator, and $\hat{\beta}^{Med}$ provides variable selection consistency and asymptotic normality*

Proof: Different from the disjoint group setting, the marginal prior for overlapping feature β_j is "one-spike-and-two-slabs":

$$\beta_j \sim \pi_A\pi_B N(0, 2s^2) + (\pi_A(1-\pi_B) + (1-\pi_A)\pi_B) N(0, s^2) + (1-\pi_A)(1-\pi_B)\delta_0(\beta_j),$$

where $\pi_A = \pi^{(2)}\pi_l^{(1)}\pi_{kl}^{(0)}$ and $\pi_B = \pi^{(2)}\pi_{l'}^{(1)}\pi_{k'l'}^{(0)}$. The marginal posterior distribution is a mixture of a point mass at 0 and two normal distributions:

$$\beta_j|Y, X \sim r_A N\left((1-B_n^*)\hat{\beta}_j^{LS}, \sigma^2(1-B_n^*)/n\right) + r_B N\left((1-B_n)\hat{\beta}_j^{LS}, \sigma^2(1-B_n)/n\right) + r_C\delta_0(\beta_j),$$

where $B_n^* = \dfrac{\sigma^2}{2ns^2 + \sigma^2}$, r_A, r_B and r_C are the posterior weights for the three distributions with following forms:

$$r_A = C_{cons}\pi_A\pi_B(1+2n\tau^2)^{-1/2}\exp\left(\frac{(1-B_n^*)}{2\sigma^2}n(\hat{\beta}_j^{LS})^2\right),$$
$$r_B = C_{cons}(\pi_A(1-\pi_B) + (1-\pi_A)\pi_B)(1+n\tau^2)^{-1/2}\exp\left(\frac{(1-B_n)}{2\sigma^2}n(\hat{\beta}_j^{LS})^2\right),$$
$$r_C = C_{cons}(1-\pi_A)(1-\pi_B),$$

where C_{cons} is the normalizing constant.

Since $P(\beta_j = 0|Y, X)$ has the same asymptotic behavior as in non-overlapping group setting, following the same steps, we can prove that with n sufficiently large such that $P(\beta_j = 0|Y, X, \hat{\beta}_j^{LS} = 0) > 0.5$, there exist a threshold t, such that if $|\hat{\beta}_j^{LS}| \leq t$, $\hat{\beta}_j^{Med} = 0$; otherwise

$\hat{\beta}_j^{Med} \neq 0$. However, due to the mixture of three distributions, we lose the analytical solution for the threshold t.

To prove the selection consistency, we need to consider the behaviors of two normal distributions. When $\beta_j^0 = 0$, $r_A \xrightarrow{\text{p}} 0$, $r_B \xrightarrow{\text{p}} 0$, and $r_C \xrightarrow{\text{p}} 1$, hence, $\hat{\beta}_j^{Med} \xrightarrow{\text{p}} 0$. When $\beta_j^0 \neq 0$, $r_C \xrightarrow{\text{p}} 0$, $r_A + r_B \xrightarrow{\text{p}} 1$, $(1 - B_n)\hat{\beta}_j^{LS} \xrightarrow{\text{p}} \beta_j^0 \neq 0$, $(1 - B_n^*)\hat{\beta}_j^{LS} \xrightarrow{\text{p}} \beta_j^0 \neq 0$, $\sigma^2(1 - B_n^*)/n \to 0$, and $\sigma^2(1 - B_n)/n \to 0$, therefore $\hat{\beta}_j^{Med} \xrightarrow{\text{p}} \beta_j^0 \neq 0$.

To prove the asymptotic normality, we define $z_j = \sqrt{n}(\beta_j - \hat{\beta}_j^{LS})$, so that $\hat{z}_j^{Med} = \sqrt{n}(\hat{\beta}_j^{Med} - \hat{\beta}_j^{LS})$. Posterior distribution of z_j is

$$z_j|Y, X \sim r_A N\left(-B_n^*\sqrt{n}\hat{\beta}_j^{LS}, \sigma^2(1 - B_n^*)\right) + r_B N\left(-B_n\sqrt{n}\hat{\beta}_j^{LS}, \sigma^2(1 - B_n)\right) + r_C\delta_0(z_j - \hat{\beta}_j^{LS}),$$

Since $B_n \to 0$, $B_n^* \to 0$, $\sqrt{n}\hat{\beta}_j^{LS} = O_p(1)$, $z_j \xrightarrow{\text{p}} 0$. Following the same steps in the proof of non-overlapping group structure, we get the asymptotic normality of $\hat{\beta}^{Med}$. ■

4 Simulations

4.1 Simulation I: Single-layer non-overlapping groups

We first simulated data with single-layer non-overlapping groups to evaluate the performance of SOG. We set $n = 125$, $p = 200$, $m_1 = 10$, and $U^{(1)}$ with block diagonal structure as below:

$$U^{(1)} = \begin{bmatrix} 1_{20} & 0_{20} & \dots & 0_{20} \\ 0_{20} & 1_{20} & \dots & 0_{20} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{20} & 0_{20} & \dots & 1_{20} \end{bmatrix},$$

where 1_m (0_m) denotes an $m \times 1$ column matrix with all values equal to $1(0)$. In this setting, all 10 level-1 groups are disjoint, each having 20 level-0 variables. To model the within level-1 group correlation to be 0.5, for each level-1 group k (k=1, ..., m_1), we drew $z_k^{(1)}$ independently from $N(0, 1)$, and sampled $x_{ij} = \left(\sum_{k=1}^{m_1} U_{jk}^{(1)} z_k^{(1)} / \sum_{k=1}^{m_1} U_{jk}^{(1)} + e_{ij}\right)/\sqrt{2}$, where $e_{ij} \sim N(0, 1)$, $1 \leq i \leq n$, and $1 \leq j \leq p$. The total number of effective β_{jk}'s with corresponding $U_{jk}^{(1)} = 1$ is 200. We set 50 out of those 200 β's to be non-zero, generated from $N(0, 5)$. Other β's were set to be 0. We set the sparsity varied among level-1 groups: group

1 has all 20 β's as non-zero; group 2 and 3 have 10 out of 20 β's as non-zero; group 4 and 5 have 5 out of 20 β's as non-zero. All other group have all β's as 0.

We repeated the simulation 100 times, each evaluated by 5-fold cross validation. We compared the variable selection and prediction performance of SOG with BSGS-SS, HSVS, lasso, GL, and SGL in Table 2, with the evaluation criteria described in Section 2.5.

From Table 2, we can see that SOG has the best variable selection and prediction accuracy, with the highest AUC and the smallest MSE. BSGS-SS has similar feature selection AUC and higher prediction MSE mainly because it assumes same sparsity proportions inside selected groups. HSVS has larger MSE and smaller AUC. This is probably because the Laplace prior failed to provide exact zero estimates. Lasso and GL both have poor performance as expected, since Lasso does not consider group structure and GL does not consider sparsity within selected groups. SGL improves feature selection AUC over GL as expected, but it implicitly assumes equal proportion of true non-zero β's in each group. For SOG and BSGS-SS, the posterior distribution of feature selection allows control of Bayesian FDR (BFDR). Under nominal BFDR at 10%, the actual FDR given the simulation truth are shown in Table 2. BSGS-SS is anti-conservative with 31% actual FDR, while SOG is properly controlled at 8%. In addition to smaller actual FDR, SOG has similar corresponding FOR, showing its better feature selection performance compared to BSGS-SS.

4.2 Simulation II: Single-layer overlapping groups

We next simulated data with single-layer overlapping groups to evaluate the performance of SOG with BSGS-SS and HSVS. The setting is exactly the same as simulation I in Section 4.1, except now $U^{(1)}_{1,1} = U^{(1)}_{1,2} = 1$ and $U^{(1)}_{41,3} = U^{(1)}_{41,4} = 1$ (see Fig 2). In other words, we set level-0 variable 1 to belong to both level-1 group 1 and 2; level-0 variable 41 to belong to both group 3 and 4. We used the same approach to set the within level-1 group correlation to be 0.5.

We now have 202 β_{jk}'s with corresponding $U^{(1)}_{jk} = 1$, but only 200 of them are identifiable: $\beta_1 = \beta_{1,1} + \beta_{1,2}$ and $\beta_{41} = \beta_{41,3} + \beta_{41,4}$ are identifiable; but each individual term may not. We used the same setting to draw β_{jk}. Two newly added variables $\beta_{1,2}$ and $\beta_{41,4}$ were set to be non-zero and drew from $N(0, 5)$. We applied SOG, BSGS-SS, and HSVS using the same duplication approach. Table 2 shows the evaluation results using 100 simulated data sets.

Table 2: Variable selection and prediction performance comparisons for simulation 1, 2, and 3 (mean(SE))

	Method	Feature selection AUC	MSE	Actual feature FDR	Actual feature FOR
Simulation 1	SOG	0.98 (0.00)	3.81 (0.56)	0.08 (0.00)	0.02 (0.00)
	BSGS-SS	0.98 (0.00)	3.95 (0.80)	0.30 (0.02)	0.01 (0.00)
	HSVS	0.96 (0.00)	6.68 (0.31)	–	–
	Lasso	0.78 (0.00)	27.84 (1.41)	–	–
	GL	0.51 (0.00)	193.75 (11.31)	–	–
	SGL	0.74 (0.00)	41.64 (1.82)	–	–
Simulation 2	SOG	0.99 (0.00)	3.22 (0.13)	0.06 (0.00)	0.03 (0.01)
	BSGS-SS	0.97 (0.00)	5.73 (1.55)	0.45 (0.01)	0.01 (0.00)
	HSVS	0.97 (0.00)	6.39 (0.36)	–	–
Simulation 3 U=0.2	MOG	0.98 (0.00)	4.78 (0.16)	0.02 (0.00)	0.14 (0.00)
	SOG	0.92 (0.00)	7.00 (0.26)	0.06 (0.01)	0.19 (0.00)
	BSGS-SS	0.92 (0.00)	9.13 (0.33)	0.02 (0.00)	0.22 (0.01)
	HSVS	0.84 (0.01)	12.27 (0.48)	–	–
	Lasso	0.74 (0.00)	8.86 (0.25)	–	–
	GL	0.75 (0.00)	5.64 (0.17)	–	–
	SGL	0.74 (0.00)	8.52 (0.24)	–	–
Simulation 3 U=0.5	MOG	1.00 (0.00)	2.54 (0.09)	0.09 (0.00)	0.00 (0.00)
	SOG	0.99 (0.00)	6.37 (1.04)	0.10 (0.00)	0.02 (0.00)
	BSGS-SS	0.96 (0.00)	22.02 (2.08)	0.08 (0.01)	0.10 (0.01)
	HSVS	0.99 (0.01)	4.27 (0.82)	–	–
	Lasso	0.76 (0.00)	42.21 (1.24)	–	–
	GL	0.81 (0.00)	20.51 (0.69)	–	–
	SGL	0.75 (0.00)	43.10 (1.24)	–	–

From the results, SOG continues to have the best variable selection and prediction performance. In fact, the results are very similar to simulation 1, which means overlapping groups do not affect our selection and prediction. Even though we introduced unidentifiable overlapping feature coefficients (e.g. β_{11} and β_{12}), we are still able to estimate the marginal effects (e.g. $\beta_1 = \beta_{11} + \beta_{12}$), which are identifiable.

4.3 Simulation III: Two-layer overlapping groups

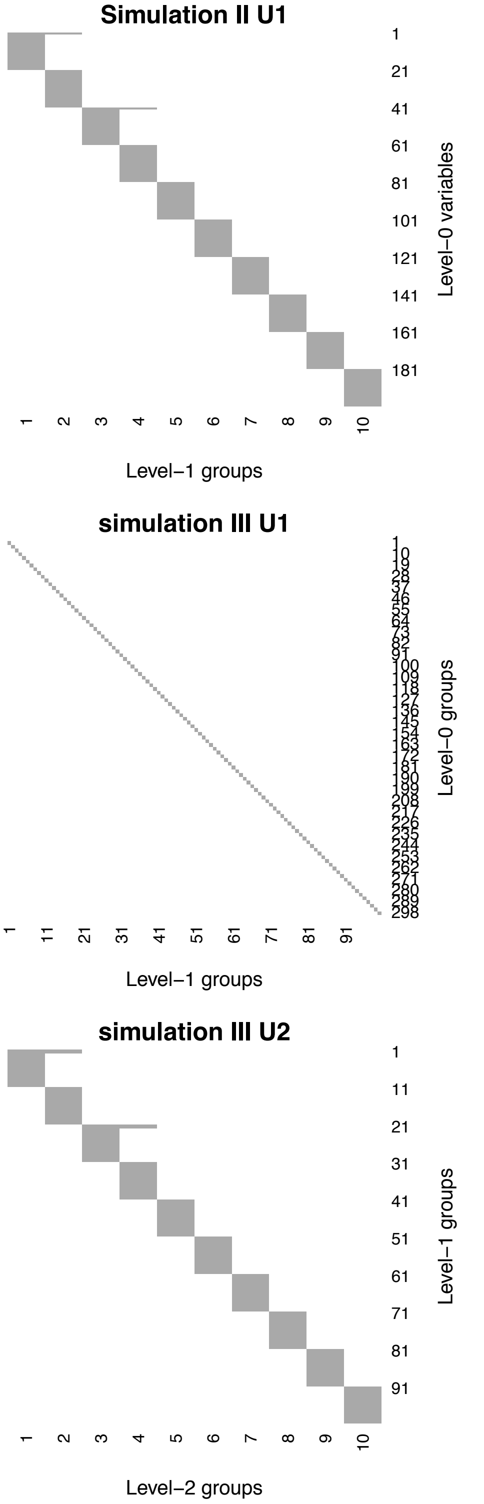
In this simulation, we simulate two-layers of overlapping groups to evaluate the performance of MOG. We set $n = 200$, $p = 300$, $m_1 = 100$, $m_2 = 10$, $U^{(1)}$ and $U^{(2)}$ with structures in Fig 2. $U^{(1)}$ is a block diagonal structured, i.e. every three features belong to one level-1 group, making $p = 300$ features belong to $m_1 = 100$ level-1 groups; $U^{(2)}$ has block diagonal structure in the most part except $U^{(2)}_{1,1} = U^{(2)}_{1,2} = 1$ and $U^{(2)}_{21,3} = U^{(2)}_{21,4} = 1$, i.e. level-1 group 1 belongs to level-2 group 1 and 2; level-1 group 21 belongs to level-2 group 3 and 4.

In this setting, we only have overlapping level-2 groups while level-1 groups are disjoint. As a result, we can still compare MOG to SOG, BSGS-SS, HSVS, GL, and SGL, as they only use level-1 group structure and ignore level-2 group structure. We used a similar approach to model within group correlation. For each level-1 group k, we drew $z^{(1)}_k \sim N(0, 0.3)$; for each level-2 group l, we drew $z^{(2)}_l \sim N(0, 0.2)$; then we set $x_{ij} = \sum\limits_{k=1}^{m_1} \sum\limits_{l=1}^{m_2} U^{(1)}_{jk} U^{(2)}_{kl} (z^{(1)}_k + z^{(2)}_l) / \sum\limits_{k=1}^{m_1} \sum\limits_{l=1}^{m_2} U^{(1)}_{jk} U^{(2)}_{kl} + e_{ij}$, where $e_{ij} \sim N(0, 0.5)$. In this way, $Var(X_{ij}) = 1$. For features belonging to the same level-1 group, the correlation is 0.5; for features belonging to the same level-2 group but different level-1 groups, the correlation is 0.2.

The total number of β_{jk} to be estimated is 306, and among them, 300 are identifiable. We set 5 out of 10 level-2 groups to contain true features. Inside these 5 level-2 groups, we set 4 out of 10 level-1 groups to have strong signals (all three features in each level-1 group have $\beta \sim Unif(2U, 3U)$; U will vary); the other 2 of 10 level-1 groups to have medium signals (all three features in each level-1 group have $\beta \sim Unif(U, 2U)$). The remaining 4 level-1 groups have all 3 features with zero coefficients. We set U to be 0.2 and 0.5.

Table 2 shows the comparison results for 100 simulated data sets. MOG has the best performance in both variable selection ROC and prediction MSE especially when U is small. When U=0.2, SOG has better performance than other methods and MOG further improves

Figure 2: $U^{(1)}$ in simulation II, $U^{(1)}$ and $U^{(2)}$ in simulation III, grey denotes 1, white denotes 0.



SOG. The results shows the benefit of incorporating level-2 grouping structure. BSGS-SS has smaller FDR but higher FOR than SOG. This is because BSGS-SS assumes same sparsity inside groups, with the presence of groups with weak signals, it will miss some weak features. At U=0.5, all four Bayesian methods tend to obtain similar good performance in feature selection, because all of them can perform variable selection both at group and within group level and sparsity for those level-1 groups with signals are not designed (all β's are non-zero). But for prediction MSE, MOG still clearly outperforms other Bayesian methods. Lasso, GL and SGL tend to have poorer selection and prediction performance even when U is large. GL performs better than Lasso and SGL, because sparsity is not designed inside level-1 groups in this simulation.

5 Applications

5.1 Predict ER+ versus ER- breast cancer

We applied MOG to n=727 breast cancer patients retrieved from The Cancer Genome Atlas (TCGA). Each sample has mRNA expression, methylation, and copy number variations (CNV) features available. This application is aimed to predict estrogen receptor (ER) status and identify associated pathways, genes, and multi-level omics features simultaneously. We first filtered out genes with mRNA expression mean and variance below the median and restricted one methylation value for each gene by averaging the M-values within 50kb of the gene starting position. Since BSGS-SS and HSVS are quite computational intensive, we first restricted to 8 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways to compare all seven methods (MOG, SOG, BSGS-SS, HSVS, Lasso, GL, and SGL). Those pathways contain 40-50 genes and more than 80% of genes associated with ER status by simple t-test (p-val=0.05). A total of p = 824 multi-level omics features (level-0 variables) belonging to m_1 = 276 genes (level-1 groups) in m_2 = 8 pathways (level-2 groups) were left for analysis. For a more realistic setting, we then extended to 123 KEGG pathways containing 40-50 genes to compare the performance of MOG, SOG, lasso, GL, and SGL. In this way, p = 11785 multi-level omics features (level-0 variables) belonging to m_1 = 1316 genes (level-1 groups) in m_2 = 123 pathways (level-2 groups) were left for analysis. The “ER signaling pathway” is an obvious pathway that should predict the ER status. It is covered by the both pathway selections and can serve as an internal control.

We applied SOG, BSGS-SS, HSVS, GL, and SGL, by using genes as group structure and ignoring level-2 pathway groups; we also applied lasso ignoring all group structures. Lasso, GL, and SGL use 10-fold cross-validation in training set to select tuning parameters. Performance is evaluated using 5-fold cross-validation. Each time, 4 folds of the data are left for training, and one fold is left for testing.

To prioritize variable and group selection, we define a feature impact score FIS_j in MOG as the posterior average selection probability of feature j, i.e. $FIS_j = AVE(\sum_{k=1}^{M_1}\sum_{l=1}^{M_2}\gamma_l^{(2)}\gamma_{kl}^{(1)}\gamma_{jkl}^{(0)}U_{jk}^{(1)}U_{kl}^{(2)})$, here $AVE(\cdot)$ is the average over all MCMC iteractions. The pathway impact score PIS_l is then defined as the average FIS over all level-0 variables included in pathway l, i.e. $PIS_l = AVE(\sum_{j=1}^{P}\sum_{k=1}^{M_1}\gamma_l^{(2)}\gamma_{kl}^{(1)}\gamma_{jkl}^{(0)}U_{jk}^{(1)}U_{kl}^{(2)})$. In SOG, FIS and PIS are defined similarly, $FIS_j = AVE(\sum_{k=1}^{M_1}\gamma_k^{(1)}\gamma_{jk}^{(0)}U_{jk}^{(1)})$ and $PIS_l = AVE(\sum_{j=1}^{P}\sum_{k=1}^{M_1}\gamma_k^{(1)}\gamma_{jk}^{(0)}U_{jk}^{(1)}U_{kl}^{(2)})$. Using $\gamma_{jk}^{(0)}$ to denote if $\beta_{jk} = 0$, the definitions of FIS and PIS for BSGS-SS and HSVS is the same as SOG. We ranked the pathways and variables based on their impact scores in Table 3. For lasso, GL, and SGL, the methods cannot readily prioritize variables and pathways. We performed Fisher's exact test for pathway enrichment analysis for features selected at least once in 5 fold cross-validation to prioritize the top pathways.

It is well-known that the mRNA expression of ESR1 is predictive for ER status, defined by immunohistochemistry (IHC) assay of estrogen receoptor (ER) . In both data set with 8 and 123 pathways, MOG detected the ER signaling pathway as the top selected pathway with the highest PIS, and ESR1-mRNA and ESR1-methyl are among the top selected features. MOG, SOG, BSGS-SS and HSVS sorted the features based on posterior selection probability averaged over 5 cross-validation. Since averaging over 5 fold cross-validation in lasso, GL, and SGL is infeasible, we sued feature selection result with the first fold data left out, and we found leaving different fold out gave very similar results. To get a better sense of the feature selection, we plotted the number of features selected (x-axis) versus the number of selected features belonging to ER signaling pathway (y-axis) in Fig 3. Most of the top features selected by MOG, belonged to ER signaling pathway (e.g. 98 out of top 100 in Fig

3), nonetheless other methods had much less top features in ER signaling (e.g. SOG has 19 out of top 100 in Fig 3).

We calculated ER prediction AUC for samples in testing set to compare the prediction performance. For Bayesian methods, we performed two predictions: (1) plugging posterior median estimates of β into $Pr(Y_i = 0) = \Phi(X\hat{\beta}^{Med})$ to obtain AUC_1; (2) using model averaging by calculating posterior mean of $\Phi(X\hat{\beta})$ to generate AUC_2. For lasso, GL, and SGL, we selected tuning parameter from 10-fold cross-validation and plugged in $\hat{\beta}$. Having strong predictive genes such as ESR1 gene, all methods have high AUCs in the testing set in both data sets. Comparing two AUCs, AUC_2 tends to be slightly higher than AUC_1 in general for the Bayesian methods, consistent with the common belief that averaging over all models from MCMC provides better predictive ability than using a single plug-in estimate. MOG and SOG using model averaging predictor generate the highest prediction AUC although the difference is not statistically significant given the almost perfect prediction.

5.2 Predict invasive lobular carcinoma (ILC) versus invasive ductal carcinoma (IDC)

We next applied MOG to predict histology subtype (ILC/IDC) for 669 patients in the same TCGA data set. We chose the same 123 KEEG pathways to compare the performance of MOG, SOG, lasso, GL, and SGL. Variable selection and prediction performance are summarized in table 4. Similar to ER status, there exists a well-known strong predictor CDH1 mRNA expression (Ciriello et al., 2015), thus all methods have good prediction $AUCs$. However, MOG has the highest AUC_2 taking model averaging. Since ILC is less-studied subtype in breast cancer research, there is no annotated pathway particularly for this histologic subtype. MOG detects endometrial cancer pathway as the top pathway. In adition to CDH1 mRNA, CNV and methylation, it contains several features of several genes including PIK3CA, PTEN, and TP53, which were shown related to ILC/IDC comparison (Ciriello et al., 2015).

Table 3: Variable selection and prediction results in breast cancer ER+/- application

8 pathways	Top pathway from PIS	PIS	Top 3 selected features	AUC_1[a] (SD)	AUC_2[b](SD)
MOG	ER signaling	0.123	ESR1-mRNA, ESR1-methyl, MMP2-mRNA	0.948 (0.009)	0.948 (0.009)
SOG	ER signaling	0.043	ESR1-mRNA, ESR1-methyl, ESR1-cnv	0.943 (0.014)	0.945 (0.013)
BSGS-SS	ER signaling	0.022	ESR1-mRNA, MMP2-mRNA, NME3-mRNA	0.940 (0.014)	0.944 (0.011)
HSVS	ER signaling	0.034	ESR1-mRNA, ESR1-cnv, ESR1-methyl	0.942 (0.014)	0.945(0.013)
	Top pathway from Fisher's test	Fisher's p-val	–	AUC_1 (SD)	
Lasso	ER signaling	0.12	–	0.943 (0.013)	
GL	Calcium signaling pathway	1	–	0.940 (0.013)	
SGL	Estrogen signaling	0.152	–	0.863 (0.091)	
123 pathways	Top pathway from PIS	PIS	Top 3 selected features	AUC_1 (SD)	AUC_2(SD)
MOG	ER signaling	0.064	ESR1-mRNA, ESR1-methyl, MMP2-mRNA	0.949 (0.011)	0.951 (0.010)
SOG	Prolactin signaling	0.033	ESR1-mRNA, ESR1-cnv, ESR1-methyl	0.946 (0.012)	0.948 (0.013)
Methods	Top pathway from Fisher's test	Fisher's p-val	–	AUC_1 (SD)	
Lasso	Phosphatidylinositol signaling system	0.014	–	0.947 (0.011)	
GL	GABAergic synapse	0.462	–	0.940 (0.012)	
SGL	Prolactin signaling	0.021	–	0.683 (0.113)	

[a] Plug-in $\hat{\beta}$

[b] Model averaging

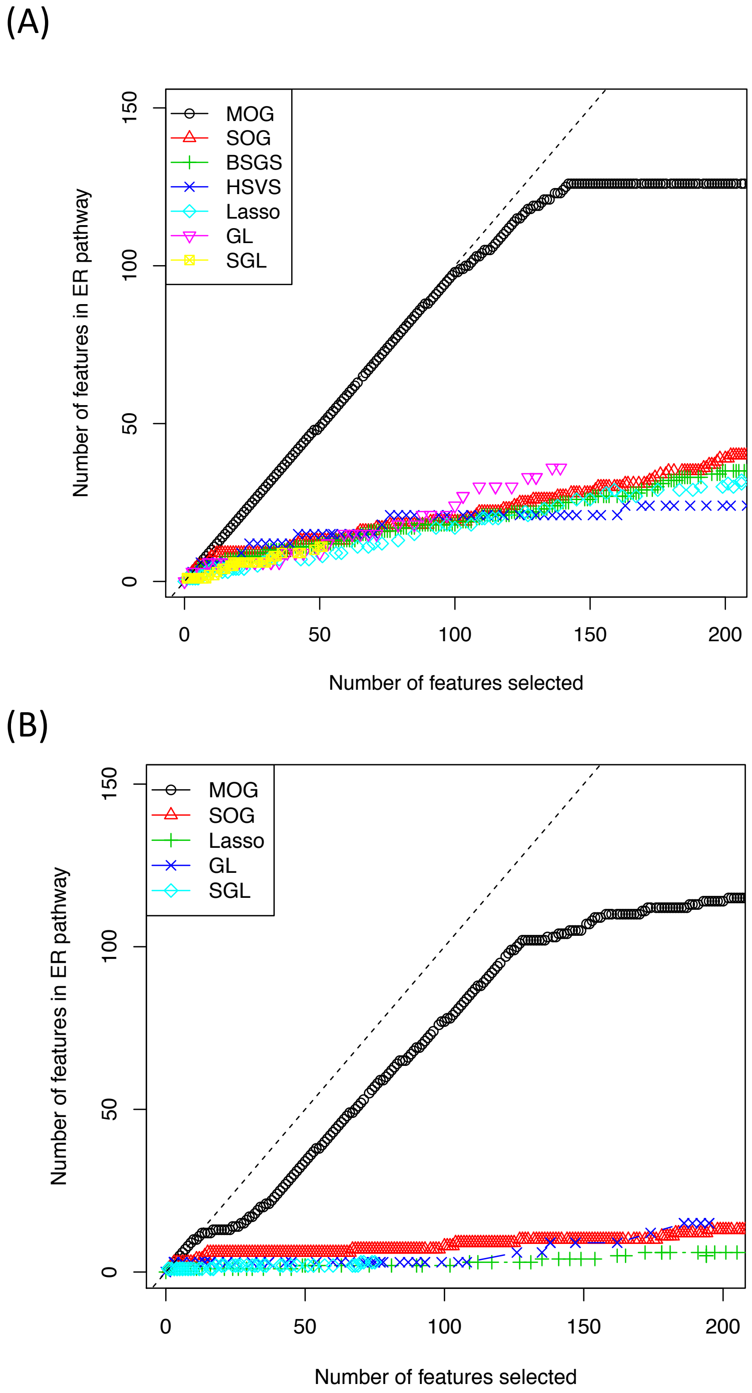


Figure 3: The number of features selected versus the number of selected features belonging to ER signaling pathway in breast cancer ER+/- application using (A) 8 pathways and (B) 123 pathways

Table 4: Variable selection and prediction results in breast cancer ILC/IDC application

Methods	Top pathway from PIS	PIS	Top 3 selected features	$AUC_1{}^a$ (SD)	$AUC_2{}^b$(SD)
MOG	Endometrial cancer	0.103	CDH1-mRNA, CDH1-CNV, CDH1-methyl	0.921 (0.012)	0.963 (0.003)
SOG	Malaria	0.013	CDH1-mRNA, CDH1-CNV,CDH1-methyl	0.913 (0.013)	0.955 (0.003)
Methods	Top pathway from Fisher's test	Fisher's p-val	–	AUC_1 (SD)	
Lasso	RNA degradation	0.010	–	0.959 (0.004)	
GL	Intestinal immune network for IgA production	0.695	–	0.958 (0.002)	
SGL	Endometrial cancer	0.017	–	0.901 (0.014)	

[a] Plug-in $\hat{\beta}$

[b] Model averaging

6 Conclusions

In modern small-n-large-p applications, effective variable selection has become a more and more important component in statistical methodologies. Methods that incorporate prior structural knowledge of variables (e.g. group lasso and fused lasso) can improve variable selection, prediction accuracy and model interpretation. In this paper, we consider a multi-layer overlapping group structure that is commonly seen in the "multi-level omics features ⇒ genes ⇒ pathways" scenario in genomic applications. Our proposed Bayesian indicator variable selection model has several innovations and advantages for the targeted problem. Firstly, Bayesian hierarchical model and indicator variable selection model allow natural incorporation of multi-layer group structure with fast Gibbs sampling computation. Secondly, we explicitly model group-specific proportions of non-zero β values (i.e. $\pi_k^{(0)}$) for different sparsity level in different selected groups. Thirdly, Bayesian approach allows a simple duplication technique for overlapping groups. The duplicated variables assigned into multiple groups have unidentifiable parameters individually but their marginal sum is identifiable. Fourthly, the proposed model is extensible to more than two layers of overlapping group structure. The result gives clear interpretation of which features, genes and pathways contributing to the prediction. Finally, the posterior distribution from MCMC samples provides easy post hoc inferences, such as characterization of variability and Bayesian FDR control of feature selection. Using three simulation settings and two breast cancer examples, we demonstrate superior performance of the proposed method for multi-layer overlapping group (MOG) structure in terms of variable selection, prediction accuracy and model interpretation. We also showed variable selection consistency and asymptotic normality of parameter estimation using posterior median estimate from MOG.

Our proposed method has several shortcomings that will be future directions to improve. As noted in the paper, the MCMC mixing rate in the indicator model can be unstable and converge slowly. Although our currently simulation and application seem to perform adequately, we expect worse performance in this regard when p goes larger or data signal becomes weaker. A modification to spike-and-slab prior with a small-variance Gaussian spike might alleviate the computing difficulty. As large data sets with complex prior information structure continue to grow in data science, we expect to encounter the multi-layer overlapping group structure more and more often in the future and the proposed method will improve performance in its statistical learning.

R package "MOG" is available at github https://github.com/lizhu06.

References

Albert, J. and S. Chib (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American statistical Association 88*(422), 669–679.

Ciriello, G., M. L. Gatza, A. H. Beck, M. D. Wilkerson, S. K. Rhie, A. Pastore, H. Zhang, M. McLellan, C. Yau, C. Kandoth, et al. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell 163*(2), 506–519.

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

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

Jacob, L., G. Obozinski, and J.-P. Vert (2009). Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pp. 433–440. ACM.

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

Kuo, L. and B. Mallick (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, 65–81.

Kyung, M., J. Gill, M. Ghosh, G. Casella, et al. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis 5*(2), 369–411.

Mitchell, T. J. and J. J. Beauchamp (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association 83*(404), 1023–1032.

Newton, M. A., A. Noueiry, D. Sarkar, and P. Ahlquist (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics 5*(2), 155–176.

O'Hara, R. B., M. J. Sillanpää, et al. (2009). A review of bayesian variable selection methods: what, how and which. *Bayesian analysis 4*(1), 85–117.

Park, T. and G. Casella (2008). The bayesian lasso. *Journal of the American Statistical Association 103*(482), 681–686.

Richardson, S., G. C. Tseng, and W. Sun (2016). Statistical methods in integrative genomics. *Annual Review of Statistics and Its Application 3*, 181–209.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics 22*(2), 231–245.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association 82*(398), 528–540.

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

Xu, X., M. Ghosh, et al. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis 10*(4), 909–936.

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

Zhang, L., V. Baladandayuthapani, B. K. Mallick, G. C. Manyam, P. A. Thompson, M. L. Bondy, and K.-A. Do (2014). Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 63*(4), 595–620.

Zhang, L., J. S. Morris, J. Zhang, R. Z. Orlowski, and V. Baladandayuthapani (2014). Bayesian joint selection of genes and pathways: Applications in multiple myeloma genomics. *Cancer informatics 13*(Suppl 2), 113.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.