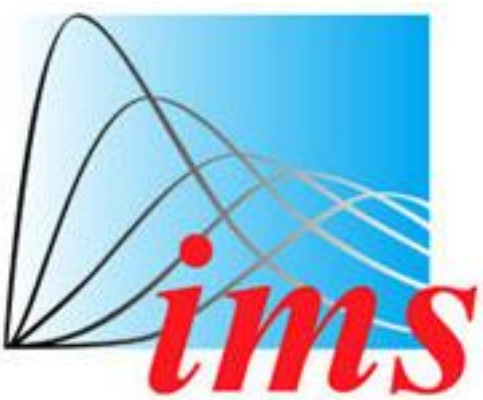




*The Annals of Applied Statistics*
2019, Vol. 13, No. 4, 2611–2636
https://doi.org/10.1214/19-AOAS1271


# BAYESIAN INDICATOR VARIABLE SELECTION TO INCORPORATE HIERARCHICAL OVERLAPPING GROUP STRUCTURE IN MULTI-OMICS APPLICATIONS

BY LI ZHU[1,*], ZHIGUANG HUO[†], TIANZHOU MA[1,‡], STEFFI OESTERREICH[*] AND GEORGE C. TSENG[1,*]

*University of Pittsburgh[*], University of Florida[†] and University of Maryland[‡]*

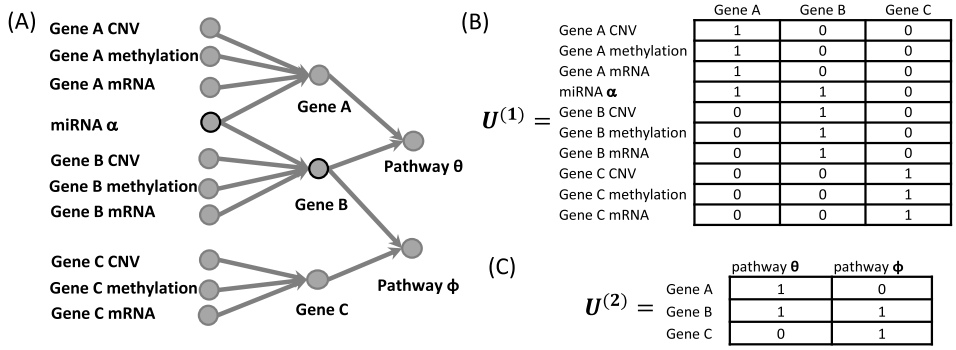
Variable selection is a pervasive problem in modern high-dimensional data analysis where the number of features often exceeds the sample size (a.k.a. small-n-large-p problem). Incorporation of group structure knowledge to improve variable selection has been widely studied. Here, we consider prior knowledge of a hierarchical overlapping group structure to improve variable selection in regression setting. In genomics applications, for instance, a biological pathway contains tens to hundreds of genes and a gene can be mapped to multiple experimentally measured features (such as its mRNA expression, copy number variation and methylation levels of possibly multiple sites). In addition to the hierarchical structure, the groups at the same level may overlap (e.g., two pathways can share common genes). Incorporating such hierarchical overlapping groups in traditional penalized regression setting remains a difficult optimization problem. Alternatively, we propose a Bayesian indicator model that can elegantly serve the purpose. We evaluate the model in simulations and two breast cancer examples, and demonstrate its superior performance over existing models. The result not only enhances prediction accuracy but also improves variable selection and model interpretation that lead to deeper biological insight of the disease.

**1. Introduction.** Variable selection is a pervasive problem in statistical applications, intended to search for the best model 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. For high-dimensional regression problems, it is commonly believed that only a small set of features have a nontrivial effect on the outcome, while most other features have little or no effect. In the literature, the penalized regression method—lasso (Tibshirani (1996)) uses an $l_1$ norm penalty to achieve variable selection, however it tends to randomly select one out of a set of highly correlated variables while ignoring the others. Zou and Hastie (Zou and Hastie (2005)) proposed the elastic net method with a combination of $l_1$ and $l_2$ norm penalties to overcome this problem. When prior information of grouped variables is available and variable selection by groups is desired, Yuan and Lin (2006) proposed the group lasso penalty

Received November 2018; revised May 2019.
[1]Supported in part by NIH Grants R01CA190766, R01MH111601 and R21LM012752.
*Key words and phrases.* Bayesian variable selection, hierarchical overlapping group structure, overlapping groups, spike and slab.

so that variables inside the same group are selected or dropped together. In order to further allow sparsity within selected groups, Simon et al. (2013) proposed the sparse group lasso with both an $l_1$ norm penalty and a group lasso penalty. In the counterpart of Bayesian framework, variable selection can be viewed as identifying nonzero variables (or elimination of variables very close to zero) in the posterior distribution. Tibshirani (1996) pointed out that the lasso estimator is equivalent to the posterior median of a Gaussian model using the 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 the group lasso and the elastic net. Mitchell and Beauchamp (1988) proposed another popular type of prior called the "spike and slab" prior, which is a mixture of a point mass at zero (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, Hernández-Lobato and Dupont (2013) generalized the spike-and-slab prior for group feature selection and implemented the expectation propagation algorithm. Xu and Ghosh (2015) and Zhang et al. (2014a) extended the spike-and-slab prior to achieve sparsity both at the group level and within groups. Under mild conditions, the posterior median estimator for a normal mean sample with the spike-and-slab prior is a soft-thresholding estimator with desired selection consistency and asymptotic normality properties (Johnstone and Silverman (2004), Xu and Ghosh (2015)). Chen et al. (2016) developed a similar Bayesian model by introducing separate binary selection indicators for each group and each feature inside each group, which can also lead to sparsity at the group level and within groups.

All aforementioned methods allow only nonoverlapping and single layer group structures. In this paper, we consider a motivating example that requires incorporation of a hierarchical overlapping group structure. Suppose SNP array, methylation array, miRNA array and RNA-seq are performed on $n$ tumor 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 has brought new statistical challenges (see Richardson, Tseng and Sun (2016) for review). Denote $p$ as the total umber of variables in the union of all CNV, methylation sites, miRNA and mRNA expression features. The input data $X = \{x_{ij}\}$ is a $n \times p$ matrix, where $n$ is the number of samples. Figure 1(A) shows an example of hierarchical overlapping group structure with two layers of groups. In the first layer of groups, four features belong to the gene A group: mRNA, CNV and methylation probe of gene A, and miRNA $\alpha$ that targets gene A (knowledge known a priori from miRNA target database). Similarly, group gene B contains three multi-omics features, and miRNA $\alpha$ also targets this gene. The group structure of gene A and B, at the first layer, is an example of overlapping group structure as they are both targeted by miRNA $\alpha$. In the second layer, pathway $\theta$ contains these two genes, and another pathway $\phi$ contains




| | Gene A | Gene B | Gene C |
|---|---|---|---|
| Gene A CNV | 1 | 0 | 0 |
| Gene A methylation | 1 | 0 | 0 |
| Gene A mRNA | 1 | 0 | 0 |
| miRNA α | 1 | 1 | 0 |
| Gene B CNV | 0 | 1 | 0 |
| Gene B methylation | 0 | 1 | 0 |
| Gene B mRNA | 0 | 1 | 0 |
| Gene C CNV | 0 | 0 | 1 |
| Gene C methylation | 0 | 0 | 1 |
| Gene C mRNA | 0 | 0 | 1 |

| | pathway θ | pathway ϕ |
|---|---|---|
| Gene A | 1 | 0 |
| Gene B | 1 | 1 |
| Gene C | 0 | 1 |

FIG. 1. (*A*) *An example of a hierarchical overlapping group structure in a multi-omics dataset. Multi-omics features* (*mRNA expression*, *copy number variation* (*CNV*) *and DNA methylation*) *are mapped to genes* (*level*-1 *groups*), *and genes are grouped into pathways* (*level*-2 *groups*). *Some multi-omics features may belong to multiple gene groups. For example*, *miRNA* α *regulates both gene A and gene B. A gene may also belong to multiple pathways due to its multiple functions*, *such as gene B.* (*B*) $U^{(1)}$ *membership matrix denotes if a multi-omics feature belongs to a certain gene.* (*C*) $U^{(2)}$ *membership matrix denotes if a gene belongs to a certain pathway.*

gene B and C. As a result, pathways $\theta$ and $\phi$ represent overlapping groups at the second layer as they both contain gene B. To formally represent such structure, we introduce two membership matrices for this example in Figure 1(B) and 1(C). $U^{(1)}$ is a matrix with a row dimension equal to the number of multi-omics features (i.e., $p = 10$), and a column dimension equal to the number of genes (i.e., $m_1 = 3$, $m_1$ is the total gene number). $U^{(1)}_{jk} = 1$ denotes multi-omics feature $j$ $(1 \le j \le p)$ belonging to gene $k$ $(1 \le k \le m_1)$, otherwise $U^{(1)}_{jk} = 0$. Furthermore, we also introduce $U^{(2)}$ matrix with a row dimension equal to the number of genes (i.e., $m_1 = 3$), and a column dimension equal to the number of pathways (i.e., $m_2 = 2$, $m_2$ is the number of pathways). Again, $U^{(2)}_{kl} = 1$ denotes that gene $k$ $(1 \le k \le m_1)$ belongs to pathway $l$ $(1 \le l \le m_2)$, otherwise $U^{(2)}_{kl} = 0$. In this paper, we consider a multi-omics linear regression setting $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$, where dependent variable $Y = \{y_i\}_{1 \le i \le n}$ is the clinical outcome, $X = \{x_{ij}\}_{1 \le i \le n, 1 \le j \le p}$ represents measurements of multi-omics features, and feature number $p$ is usually much larger than sample size $n$. Since $p >> n$, variable selection that properly incorporates prior biological knowledge is crucial. In our situation, the group structure of "multi-omics feature $\Rightarrow$ gene $\Rightarrow$ pathway" demonstrates a hierarchical overlapping group structure that brings challenges for variable selection in the regression setting.

A similar but simplified version of this structure was studied by Zhao, Rocha and Yu (2009) and Jenatton et al. (2011), in which, features are structured to form a tree, but the groups defined by nodes at the same depth are not allowed to overlap. They designed a specific group penalty, so that a child node group will only be selected when its parent node is selected. For general overlapping group structure,

Jacob, Obozinski and Vert (2009) proposed the concept of latent feature decomposition, which led to the solution support as the union of groups. Similarly, in the Bayesian framework, Zhang et al. (2014b) decomposed the marginal regression coefficient of a feature shared by multiple groups to be the sum of partial effects contributed by each group. With the hierarchical overlapping group structure, the target function of penalized regression approaches generally becomes intractable to optimize. A Bayesian hierarchical model provides a natural alternative for incorporating the hierarchical overlapping group 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-omics features, genes or pathways are selected. For overlapping groups, we adopt from Zhang et al. (2014a) the additive effect assumption for each overlapping group. We will show that incorporation of the hierarchical overlapping group structure enhances prediction accuracy and improves both feature selection and model interpretation.

The paper is organized as follows. In Section 2, we review the indicator variable selection model, and propose a Bayesian indicator variable selection model with single-layer and hierarchical (multi-layer) overlapping group structures. We describe the detailed MCMC algorithms for each model and extend the models to binary and survival outcomes. In Section 3, we illustrate the capabilities and limitations of existing methods compared to our proposed model. In Section 4, four simulations are demonstrated to compare the performance of the proposed model and other existing methods. We further apply the model to data from two real examples in breast cancer, using multi-omics features to predict estrogen receptor (ER) status and histological 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 linear regression setting, in which $Y = (Y_1, \ldots, Y_n)^T$ denotes the outcomes for $n$ samples, and $X$ denotes an $n \times p$ covariate matrix for $p$ variables. Assume data are centered and thus the intercept can be omitted. Under linear regression assumptions, $Y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$, 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. Denoting the binary indicator as $\gamma_j$, the indicator variable selection model is

$$Y_i = \sum_{j=1}^{p} \beta_j \gamma_j x_{ij} + \epsilon_i, \quad \epsilon_i \sim N\big(0, \sigma^2\big),$$

$$\beta = (\beta_1, \ldots, \beta_p)^T \sim N(0, D_0), \gamma_j \sim \text{Bern}(\pi).$$

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:

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

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 a model with interactions. However, if the prior is too vague, mixing can be poor, as the sampled values of $\beta_j$ may only rarely be in the region with high posterior support (O'Hara and Sillanpää (2009)). Other alternatives have been proposed (George and McCulloch (1993)), but most of them require additional parameter tuning.

2.2. *SOG*: *Bayesian indicator 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 hierarchical 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 are mapped to $m_1$ genes. We define a $p \times m_1$ matrix $U^{(1)}$ to denote the group membership of level-0 variables, with $U_{j,k}^{(1)} = 1$ denoting that level-0 variable $j$ belongs to level-1 group $k$, and $U_{j,k}^{(1)} = 0$ otherwise. We propose the following model:

$$\begin{aligned}
&Y_i \sim N\left(\sum_{j=1}^{p}\sum_{k=1}^{m_1} x_{ij}\beta_{jk}, \sigma^2\right),\\
&\big(\beta_{jk}|U_{jk}^{(1)} = 1\big) = \gamma_k^{(1)}\gamma_{jk}^{(0)} b_{jk},\\
&\big(\beta_{jk}|U_{jk}^{(1)} = 0\big) \sim \delta_0(\cdot),\\
&\gamma_k^{(1)} \sim \text{Bern}\big(\pi^{(1)}\big), \gamma_{jk}^{(0)} \sim \text{Bern}\big(\pi_k^{(0)}/R_j\big), b_{jk} \sim N\big(0, s^2\big), p\big(\sigma^2\big) \propto 1/\sigma^2,
\end{aligned}$$

where $R_j = \sum_{k=1}^{m_1} U_{jk}^{(1)}$ is the number of level-1 groups which includes level-0 variable $j$. The reason for scaling by $R_j$ in the prior of $\gamma_{jk}^{(0)}$ is to ensure the same selection probability and variance of the marginal effect $\beta_j = \sum_{k=1}^{m_1} U_{jk}^{(1)}\beta_{jk}$ in the prior distribution. The justification is outlined below in Remark (2). In SOG, $\gamma_k^{(1)}$ can be interpreted as the selection indicator for level-1 group $k$; if $\gamma_k^{(1)} = 1$, $\gamma_{jk}^{(0)}$ can be interpreted as the selection indicator for level-0 variable $j$ belonging to the level-1 group $k$; $\beta_{jk} \neq 0$ if and only if $\gamma_k^{(1)} = 1$ and $\gamma_{jk}^{(0)} = 1$. A singleton will be treated as a group with itself as its only member.

Markov chain Monte Carlo (MCMC) is used for model fitting. When groups do not overlap, all the full conditional distributions are available for Gibbs sampling; otherwise, Metropolis–Hastings is used to update $\pi_k^{(0)}$. See Section 1 of the Supplementary Material (Zhu et al. (2019)).

*Remarks*.

(1) $U^{(1)}$ is a sparse matrix, most of whose entries are 0's and a few are 1's. $\sum_{k=1}^{m_1} U_{jk}^{(1)}$ is the number of level-1 groups that level-0 variable $j$ belongs to. If $\sum_{k=1}^{m_1} U_{jk}^{(1)} > 1$, level-0 variable $j$ belongs to multiple groups. $\sum_{j=1}^{p} U_{jk}^{(1)}$ is the number of level-0 variables that belong to level-1 group $k$. If $\sum_{j=1}^{p} U_{jk}^{(1)} U_{jk'}^{(1)} \geq 1$, level-1 groups $k$ and $k'$ overlap. $\beta$ is also a $p \times m_1$ sparse matrix, with $\beta_{jk} \neq 0$ only when $U_{jk}^{(1)} = 1$.

(2) Assuming $\pi_k^{(0)} = \pi^{(0)}$ for all $1 \leq k \leq m_1$, the prior probability $\Pr(\beta_j \neq 0) = 1 - \prod_{k=1}^{m_1} \Pr(\beta_{jk} \neq 0)^{U_{jk}^{(1)}} = 1 - (1 - \pi^{(1)}\pi^{(0)}/R_j)^{R_j} \approx 1 - (1 - \pi^{(1)}\frac{\pi^{(0)}}{R_j} R_j)$ (if ignoring higher order terms) $= \pi^{(1)}\pi^{(0)}$, which is free of $R_j$. Meanwhile, the variance of $\beta_j$ in the prior distribution is $\mathrm{Var}(\beta_j) = E(\beta_j^2) = E(\sum_{k=1}^{m_1} U_{jk}^{(1)} \beta_{jk})^2 = R_j(\pi^{(1)}\frac{\pi^{(0)}}{R_j} s^2) = \pi^{(1)}\pi^{(0)} s^2$, which is also free of $R_j$.

(3) In the case of features belonging to multiple groups, such as feature 1 shared by group 1 and 2, we assume $\beta_1 = \beta_{11} + \beta_{12}$. Here, partial effects ($\beta_{11}$, $\beta_{12}$) are not separately identifiable in the classical frequentist sense, since different parameter values can correspond to the same likelihood through the equal sum. This may seem to violate another definition of identifiability in the Bayesian framework, which we will refer to as "unidentifiable by likelihood" (Gelfand and Sahu (1999)). However, with an informative prior, or if the separate parameters share information from other parameters (e.g., $\beta_{11}$ shares information with other parameters in group 1 and $\beta_{12}$ shares information with other parameters in group 2 in our case), identifiability is not an issue, although slow convergence or unstable MCMC can be a problem (Eberly and Carlin (2000)). Nevertheless, the marginal parameter $\beta_1$ is our main interest of inference and is always identifiable by likelihood no matter in a frequentist or a Bayesian framework.

(4) For binary indicators $\gamma_k^{(1)}$ and $\gamma_{jk}^{(0)}$, there are three situations potentially not identifiable by likelihood (suppose two features belong to group $k$): (1) $\gamma_k^{(1)} = 0$, and $\gamma_{1k}^{(0)} = \gamma_{2k}^{(0)} = 0$; (2) $\gamma_k^{(1)} = 1$, and $\gamma_{1k}^{(0)} = \gamma_{2k}^{(0)} = 0$; and (3) $\gamma_k^{(1)} = 0$, and $\gamma_{1k}^{(0)} = 1$ or $\gamma_{2k}^{(0)} = 1$. Chen et al. (2016) used a conditional prior to avoid situation (3), so that whenever $\gamma_k^{(1)}$ is zero, $\gamma_{1k}^{(0)}$ and $\gamma_{2k}^{(0)}$ have to be zero. This conditional prior can be adopted into our model easily, but it still cannot distinguish situation (1) and (2) by avoiding a "false group" with all zero features in situation (2). Stingo et al. (2011) imposed three additional constraints for interpretability and

identifiability. When $\gamma_k^{(1)} = 0$, they forced $\gamma_{1k}^{(0)} = \gamma_{1k}^{(0)} = 0$; and if $\gamma_k^{(1)} = 1$, they eliminated the possibility of having $\gamma_{jk}^{(0)} = 0$ for all $j = 1, \ldots, p$. However, this constrained prior makes the Gibbs sampling infeasible. Thus, they have to adopt the Metropolis–Hastings algorithm, which can be inefficient when multi-layers of groups are introduced and feature dimension becomes large. Therefore, we decided not to add constraints in our prior, at a price that individual indicators $\gamma_k^{(1)}$ and $\gamma_{jk}^{(0)}$ may not be interpretable occasionally. Instead, they are used to impose group level and variable level sparsity. Variable selection eventually is determined at level-0 variable level by $\eta_{jk} = \gamma_k^{(1)}\gamma_{jk}^{(0)}$. Higher level group selection will be defined through group impact score (i.e., pathway impact score, PIS; see Section 5.1) to provide interpretation of selection at different layers of groups.

(5) $s^2$ controls the magnitude of the effect size. Here, for simplicity, we assume all $b_{jk}$ are from the same distribution with common $s^2$. However, when dealing with multi-omics data in all our applications, we let $s^2$ be platform specific. In other words, methylation, CNV and gene expression can have different levels of variability.

(6) We assign hyper-priors: $\pi^{(1)} \sim \text{Beta}(a_1, b_1)$, $\pi_k^{(0)} \sim \text{Beta}(a_0, b_0)$, and $s^2 \sim \text{Inverse} - \text{Gamma}(a_s, b_s)$. If prior information is not available, we set $a_1 = b_1 = a_0 = b_0 = 1$, and $p(s^2) \propto 1/s^2$ (i.e., $a_s = b_s \approx 0$) as a noninformative prior. When the group size varies, borrowing information across groups will stabilize the estimate of $\pi_k^{(0)}$ for groups with small size. We consider two possible ways of information sharing: one is to assume that genes can be categorized into clusters, each with cluster-specific sparsity prior (Lock and Dunson (2017)), and the other is to use a common informative prior to stabilize the estimates. Since the former option is similar to the design of level-2 group sparsity, which will be proposed later, in this situation, we choose the second option and propose an empirical Bayes approach to estimate $a_0$ and $b_0$: (1) We first apply lasso regression (Tibshirani (1996)), ignoring any group structure; (2) Then, group specific sparsity $\hat{\pi}_k^{(0)}$ is estimated by counting the number of nonzero coefficients inside each group $k$; (3) Finally, by moment matching, hyper-parameters are estimated as $\hat{a}_0 = (\frac{1-\hat{E}}{\hat{V}} - \frac{1}{\hat{E}})\hat{E}^2$ and $\hat{b}_0 = (\frac{1-\hat{E}}{\hat{V}} - \frac{1}{\hat{E}})(1-\hat{E})\hat{E}$, where $\hat{E}$ and $\hat{V}$ are the sample expectation and variance of $\hat{\pi}_k^{(0)}$ $(k = 1, \ldots, m_1)$. A simulation was conducted to evaluate the performance of borrowing information using the proposed empirical Bayes approach (simulation V in Section 3 of the Supplementary Material, Zhu et al. (2019)). When a large number of groups with a reasonable number of variables inside each group exist, borrowing information can better estimate $\pi_k^{(0)}$. When the number of groups or the number of variables in each group is small, this approach may produce inaccurate estimate of $\pi_k^{(0)}$, because $a_0$ and $b_0$ cannot be accurately inferred. Due to the pros and cons of borrowing information to help estimate group specific sparsity, we allow users to choose the new empirical Bayes approach or the original noninformative approach in our R package. Users can decide which approach to use by

evaluating performance in cross-validation. For all simulations and applications in this paper, we will apply the original noninformative prior.

2.3. *MOG*: *Bayesian indicator variable selection with multi-layer hierarchical overlapping groups*. In the presence of multi-layer (say $s$ layers) hierarchical overlapping groups, we define $U^{(1)}, \ldots, U^{(s)}$, each with dimension $p \times m_1, m_1 \times m_2, \ldots, m_{s-1} \times m_s$ respectively, to specify the group structures. The multi-level omics data example in the Introduction (Figure 1(A)) corresponds to a structure with $s = 2$. Below, we use $s = 2$ to illustrate the motivating example, but the model can be extended to $s > 2$. The proposed model for two-layer overlapping groups is

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

where $D_k = \sum_{l=1}^{m_2} U_{kl}^{(2)}$ is the number of level-2 groups which share level-1 group $k$. Similar to $R_j$ in SOG, $D_k$ and $R_j$ here are used to ensure the same selection probability and variance for the marginal effect $\beta_j$ in the prior distribution. In MOG, $\gamma_l^{(2)}$ 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 $\gamma_{jkl}^{(0)} = 1$.

When prior information is not available, we assign noninformative hyperpriors similar to SOG: $\pi^{(2)} \sim \text{Beta}(1, 1)$, $\pi_l^{(1)} \sim \text{Beta}(1, 1)$, $\pi_{kl}^{(0)} \sim \text{Beta}(1, 1)$, and $p(s^2) \propto 1/s^2$. MCMC sampling is described in Section 1 of the Supplementry Material (Zhu et al. (2019)).

*Remarks*.

(1) 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.

(2) Asymptotic properties of SOG and MOG under orthogonal design are provided in the Supplementry Materials. Briefly, we show that the posterior median estimator of $\beta_{jkl}$ is a soft-thresholding estimator with selection consistency and

asymptotic normality, when the design matrix is orthogonal, $p$ is fixed and $n \to \infty$. Although the assumptions generally do not hold in multi-omics applications, the properties provide some insights to the proposed method.

2.4. *Extension to binary and survival outcomes*. For a binary outcome, we adopt the data augmentation from Albert and Chib (1993) introducing latent variable $Z_i$ $(i = 1, \ldots, n)$ to replace $Y_i$ in the regression

$$Y_i = \begin{cases} 1, & \text{if } Z_i \geq 0, \\ 0, & \text{otherwise,} \end{cases} \qquad Z_i = \beta_0 + x_i^T \beta + \varepsilon_i, \varepsilon_i \sim N(0, 1),$$

where $\beta_0$ is the intercept, for which a noninformative prior $N(0, 100)$ is given.

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

$$\begin{cases} \log(t_i) = Z_i, & \text{if } \delta_i = 1, \\ \log(t_i) < Z_i, & \text{if } \delta_i = 0, \end{cases} \qquad Z_i = \beta_0 + x_i^T \beta + \varepsilon_i, \varepsilon_i \sim N\big(0, \sigma^2\big).$$

## 3. Related methods.

3.1. *Capabilities and limitations of existing methods*. Many methods have been proposed for variable selection with or without group structures. Here, we illustrate the major capabilities and limitations of several related methods comparing to SOG/MOG. Table 1 tabulates the key features and comparison of all methods.

- Penalized regression.
  - Lasso (Tibshirani (1996)): One of the most popular variable selection methods using an $l_1$ penalty at individual variable level without any group structure.
  - Group lasso (GL) (Yuan and Lin (2006)): Group selection is performed using an $l_2$ penalty, and each group is either entirely selected or entirely dropped.
  - Sparse group lasso (SGL) (Simon et al. (2013)): The penalty term combines an $l_2$ norm penalty on group level and an $l_1$ penalty on individual variable level to achieve both group selection and sparsity inside a selected group. However, it is only applicable to single-layer group structure.
  - Tree structured group lasso (TGL) (Zhao, Rocha and Yu (2009), Liu et al. (2009)): It is designed for hierarchical tree structured variables, and it can lead to a sparsity pattern in which a group defined by child node is always selected after its parent node. However, groups defined by nodes at the same depth are not allowed to overlap, and thus TGL is not applicable to overlapping group structure.

TABLE 1
*Compare MOG/SOG to some existing methods*

| Method | Feature selection | Exact zero in feature selection | Group selection | Exact zero in group selection | Varying sparsity inside groups | Overlapping groups | Multi-layer groups | Reference |
|---|---|---|---|---|---|---|---|---|
| MOG | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| SOG | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | χ | |
| BSGS | ✓ | ✓ | ✓ | ✓ | χ | ◇ | χ | Chen et al. (2016) |
| BSGS-SS | ✓ | ✓ | ✓ | ✓ | χ | ◇ | χ | Xu and Ghosh (2015) |
| HSVS | ✓ | χ | ✓ | ✓ | ✓ | ◇ | χ | Zhang et al. (2014a) |
| TGL | ✓ | ✓ | ✓ | ✓ | – | χ | ✓ | Zhao, Rocha and Yu (2009) |
| SGL | ✓ | ✓ | ✓ | ✓ | – | χ | χ | Simon et al. (2013) |
| GL | χ | χ | ✓ | ✓ | – | χ | χ | Yuan and Lin (2006) |
| Lasso | ✓ | ✓ | χ | χ | – | χ | χ | Tibshirani (1996) |

✓ indicates it can be achieved; χ indicates it cannot be achieved; ◇ indicates it cannot be achieved by the original method, but it can be achieved by an extended version in this paper; – indicates it is not applicable.

- Bayesian methods.
  - BSGS (Bayesian sparse group lasso, Chen et al. (2016)): This Bayesian hierarchical model is similar to SOG except that it predetermines some hyper-parameters, assumes common group sparsity, and assumes conditional priors on binary indicators (see Section 2.2 Remark (4)). It does not allow a multi-layer hierarchical group structure.
  - BSGS-SS (Bayesian sparse group selection with spike-and-slab, Xu and Ghosh (2015)): Compared to SOG or BSGS, this Bayesian hierarchical model constructs binary indicators differently. But it still assumes common group sparsity and does not allow a multi-layer hierarchical group structure.
  - HSVS (Hierarchical structured variable selection, Zhang et al. (2014a)): This is another Bayesian indicator variable selection model similar to SOG. The method applies Laplace prior and does not generate exact zero estimates in MCMC. Sparsity is achieved by truncation at an arbitrary threshold. It does not allow a multi-layer hierarchical group structure.

3.2. *Implementation and evaluation to compare with other existing models*. We compared our model to three existing Bayesian models BSGS (Chen et al. (2016)), BSGS-SS (Xu and Ghosh (2015)) and HSVS (Zhang et al. (2014a)), all of which can perform variable selection at the group level and within groups. Since BSGS requires all hyper-parameters to be predetermined, we set them to the software default if available. The choice of $\tau^2$ in BSGS, which serves the same purpose as $s^2$ in SOG, is a sensitive hyper-parameter without a default value, and the details will be discussed in each simulation. When overlapping groups existed, the same decomposition assumption was applied to all Bayesian methods. When dealing with binary outcomes, we applied the same data augmentation in Section 2.4 to BSGS-SS and HSVS. The built-in function for binary outcome in BSGS package reported a fatal error, so we excluded it from our comparison. We also compared our model to lasso (Tibshirani (1996)), group lasso (GL) (Yuan and Lin (2006)), sparse group lasso (SGL) (Simon et al. (2013)) and tree structured group lasso (TGL) (Zhao, Rocha and Yu (2009)). Since TGL reduces to SGL when only a single-layer of groups exist, and it does not allow groups of the same level to overlap, we only evaluated its performance in simulation IV in which tree structured variables were simulated. The mixing weight $\alpha$ in SGL was set to be 0.95 by software default, thus more similar to lasso. The performance was evaluated by accuracy of both variable selection and prediction. In all the simulations and applications, data were randomly split into five folds, with four folds as training sets and one fold as the testing set.

In terms of variable selection performance, when the true $\beta$ is known in simulation, the performance of variable selection relies on a tuning parameter or a cutoff. To eliminate the influence of arbitrary cutoffs in different methods, we derived sensitivity and specificity of variable selection under different cutoffs and

calculated the area under the receiver operating curves (AUC) for a fair comparison. MOG (SOG), BSGS and BSGS-SS can obtain exact zero estimates inside groups in each MCMC iteration, so level-0 variables were sorted according to the posterior mean of the selection probability, which was calculated as $\hat{\Pr}(\beta_j \neq 0|Y, X) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(\beta_j^{(b)} \neq 0)$, where $\beta_j^{(b)}$ is the $b$th iteration of totally $B$ converged MCMC samples. HSVS uses Laplace prior within groups and cannot obtain exact zeros inside a group if the group is selected, even though the estimates are shrunk towards zero. As a result, we sorted the features based on $\max(p_{\text{pos}}, 1 - p_{\text{pos}})$, where $p_{\text{pos}}$ is the posterior mean of $P(\beta_j > 0|Y)$. For lasso, GL, SGL and TGL, we applied multiple tuning parameters that detected different numbers of variables and formed the basis of ROC curve. Default tuning parameter sequences were used in lasso, GL and SGL, whereas TGL calculated the max tuning parameter automatically, and we selected a sequence of ratios as 0.01 to 1, with an increment of 0.01. For MOG (SOG), BSGS and BSGS-SS, which produces coefficients as exact zeroes, we also controlled Bayesian false discovery rate (BFDR, Newton et al. (2004)) at the nominal level of 10% to compare their true FDRs (the number of false positives/the number of claimed positives) and false omission rates (FOR, the number of false negatives/number of claimed negatives).

To evaluate model prediction accuracy, the coefficient estimates need to be calculated. All Bayesian methods (MOG, SOG, BSGS, BSGS-SS and HSVS) used the posterior median estimator, whereas the penalized regression methods (Lasso, GL, SGL and TGL) used tuning parameters selected by tenfold cross-validation. For continuous outcomes, we compared prediction mean squared error (MSE) in the testing set, that is, $\text{MSE} = \frac{1}{n_{te}}\sum_{i=1}^{n_{te}}(x_{\text{test},i}^T\hat{\beta} - y_{\text{test},i})^2$, where $n_{te}$ is the sample size in the testing set and $y_{\text{test},i}$ is the $i$th observation in the testing set. If the outcomes were binary, we sorted the samples in the testing set based on the predicted probability and calculated the prediction AUC.

*R* was used to implement all methods, except that TGL was implemented in Matlab. Gibbs sampler of all Bayesian models used 3000 MCMC iterations (2000 as burn-in) in simulations, and 20,000 iterations (10,000 as burn-in) in applications. BSGS, by default, uses Monte Carlo standard error (MCSE) for convergence diagnosis, and it only updates one group at each iteration. To make comparison fair but also save time, we applied 30,000 iterations (20,000 as burn-in) simulations with 10 groups in simulation I and II, and 200,000 iterations (100,000 as burn-in) in simulation III with 100 groups. In the end, we only included simulations which achieved MCSE below 0.1. When groups overlaped, SOG/MOG used the Metropolis–Hastings algorithm keeping 5000 iterations from stationary distribution, which was monitored by Gekewe diagnosis (Geweke et al. (1991)). We applied R packages MBSGS, glmnet, grplasso, SGL and Matlab package SLEP (Liu et al. (2009)), for BSGS-SS, lasso, GL, SGL and TGL, respectively. R packages/functions for BSGS and HSVS was provided by the original authors. We provided data and programming code in github (https://github.com/lizhu06/MOG_code_data) to reproduce all results in this paper.

## 4. Simulations.

4.1. *Simulation I*: *Single-layer nonoverlapping groups*. We first simulated data with single-layer nonoverlapping 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 $k$ ($k = 1, \dots, m_1$), we drew $z_k^{(1)}$ independently from $N(0, 1)$, and sampled $x_{ij} = (z_k^{(1)} + e_{ij})/\sqrt{2}$, where $e_{ij} \sim N(0, 1)$, $1 \le i \le n$, and $1 \le j \le p$. The total number of effective $\beta_{jk}$'s with corresponding $U_{jk}^{(1)} = 1$ was 200. We set 50 out of those 200 $\beta$'s to be nonzero, generated from $N(0, 5)$. Other $\beta$'s were set to be 0. We set the sparsity to vary among level-1 groups: group 1 had all 20 $\beta$'s as nonzero; group 2 and 3 had 10 out of 20 $\beta$'s as nonzero; group 4 and 5 had 5 out of 20 $\beta$'s as nonzero. All other groups had all $\beta$'s as zero. The outcomes were generated as $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$.

We repeated the simulation 100 times and compared the variable selection and prediction performance of SOG with BSGS, BSGS-SS, HSVS, lasso, GL and SGL with the evaluation criteria described in Section 3.2. We applied two values of $\tau^2$ in BSGS, one being the truth $\tau^2 = 5$, and the other being $\tau^2 = 1$ to evaluate the sensitivity.

From Table 2, we can see that SOG had the best variable selection performance and prediction accuracy, with the highest AUC and the smallest MSE. For BSGS, even with the large amount of MCMC iterations, the number of valid simulations (MCSE $< 0.1$) left were limited, with 54 and 11 simulations left for $\tau^2 = 5$ and 1 respectively. Among the valid simulations, BSGS with the correct setting ($\tau^2 = 5$) had the similar feature selection AUC with SOG, but MSE was slightly larger. This is probably because BSGS estimates $\beta$ as the average of its non-zero MCMC samples, which may be biased as it ignores the zeros; in addition, it assumes the same sparsity inside each group. BSGS-SS also had similar feature selection AUC and slightly higher prediction MSE, possibly also due to the same assumption of equal within group sparsity. HSVS had larger MSE and smaller AUC, which was likely because the Laplace prior failed to provide exact zero estimates. Lasso and GL both had poor performance as expected, since lasso does not consider group structure and GL does not consider sparsity within selected groups. SGL improved

TABLE 2

*Variable selection and prediction performance comparisons for simulation I, II, III and IV (mean(SE)). AUC is calculated based on posterior selection probability for Bayesian models and different tuning parameters for penalized regression models. True FDR (number of false positives/ number of claimed positives) and FOR (number of false negatives/ number of claimed negatives) are calculated given simulation truth under nominal level of BFDR* 0.1 *for Bayesian models which can produce exact zero estimates; MSE is calculated using posterior median for Bayesian models and using best tuning parameter from cross-validation for penalized regression models*

| | | Feature selection | | | |
|---|---|---|---|---|---|
| | | Cutoff-free | Control nominal BFDR = 0.1 | | Prediction |
| | Model | AUC | True FDR | True FOR | MSE |
| Simulation I | SOG | 0.99 (0.00) | 0.08 (0.00) | 0.03 (0.00) | 3.15 (0.13) |
| single-layer | BSGS ($\tau^2 = 5$) | 0.97 (0.00) | 0.06 (0.01) | 0.06 (0.00) | 7.39 (0.98) |
| non-overlapping | BSGS ($\tau^2 = 1$) | 0.95 (0.01) | 0.15 (0.04) | 0.07 (0.00) | 7.92 (0.8) |
| | BSGS-SS | 0.97 (0.00) | 0.45 (0.02) | 0.01 (0.00) | 7.07 (2.50) |
| | HSVS | 0.96 (0.00) | – | – | 6.68 (0.31) |
| | Lasso | 0.78 (0.00) | – | – | 28.3 (1.44) |
| | GL | 0.51 (0.00) | – | – | 193.75 (11.31) |
| | SGL | 0.74 (0.00) | – | – | 41.64 (1.82) |
| Simulation II | SOG | 0.98 (0.00) | 0.11 (0.01) | 0.04 (0.01) | 5.27 (0.93) |
| single-layer | BSGS ($\tau^2 = 5$) | 0.96 (0.01) | 0.07 (0.02) | 0.06 (0.00) | 10.06 (2.11) |
| overlapping | BSGS ($\tau^2 = 1$) | 0.87 (0.02) | 0.26 (0.06) | 0.07 (0.00) | 23.66 (3.10) |
| | BSGS-SS | 0.97 (0.00) | 0.44 (0.01) | 0.01 (0.00) | 5.57 (1.04) |
| | HSVS | 0.97 (0.00) | – | – | 5.93 (0.29) |
| Simulation III | MOG | 0.99 (0.00) | 0.03 (0.00) | 0.04 (0.00) | 0.75 (0.03) |
| U = 0.2 | SOG | 0.97 (0.02) | 0.02 (0.02) | 0.11 (0.05) | 3.92 (1.29) |
| two-layer | BSGS | 0.86 (0.00) | 0.03 (0.01) | 0.25 (0.00) | 29.64 (1.05) |
| overlapping | BSGS-SS | 0.92 (0.00) | 0.02 (0.00) | 0.22 (0.01) | 8.91 (0.33) |
| | HSVS | 0.82 (0.01) | – | – | 11.85 (0.46) |
| | Lasso | 0.74 (0.00) | – | – | 8.96 (0.25) |
| | GL | 0.75 (0.00) | – | – | 5.64 (0.17) |
| | SGL | 0.74 (0.00) | – | – | 8.52 (0.24) |
| Simulation III | MOG | 1.00 (0.00) | 0.1 (0.00) | 0.00 (0.00) | 0.54 (0.01) |
| U = 0.5 | SOG | 1.00 (0.00) | 0.1 (0.01) | 0.00 (0.00) | 2.09 (0.09) |
| two-layer | BSGS | 0.94 (0.01) | 0.08 (0.02) | 0.04 (0.00) | 17.01 (1.36) |
| overlapping | BSGS-SS | 0.96 (0.00) | 0.07 (0.01) | 0.11 (0.01) | 28.99 (1.86) |
| | HSVS | 0.98 (0.01) | – | – | 4.57 (0.81) |
| | Lasso | 0.77 (0.00) | – | – | 42.15 (1.30) |
| | GL | 0.81 (0.00) | – | – | 20.51 (0.69) |
| | SGL | 0.75 (0.00) | – | – | 43.10 (1.24) |

TABLE 2
(*Continued*)

| | | Feature selection | | | Prediction |
|---|---|---|---|---|---|
| | | Cutoff-free | Control nominal BFDR = 0.1 | | |
| | Model | AUC | True FDR | True FOR | MSE |
| Simulation IV | MOG | 1.00 (0.00) | 0.03 (0.00) | 0.04 (0.00) | 0.76 (0.028) |
| U = 0.2 | TGL | 0.86 (0.04) | – | – | 5.47 (1.37) |
| two-layer | Lasso | 0.74 (0.00) | – | – | 9.21 (0.19) |
| non-overlapping | GL | 0.77 (0.00) | – | – | 6.00 (0.20) |
| | SGL | 0.74 (0.00) | – | – | 8.34 (0.18) |
| Simulation IV | MOG | 1.00 (0.00) | 0.10 (0.00) | 0.00 (0.00) | 0.55 (0.015) |
| U = 0.5 | TGL | 0.88 (0.04) | – | – | 18.47 (6.05) |
| two-layer | Lasso | 0.77 (0.00) | – | – | 42.26 (1.05) |
| non-overlapping | GL | 0.80 (0.00) | – | – | 22.14 (0.90) |
| | SGL | 0.76 (0.00) | – | – | 42.69 (0.97) |

feature selection AUC over GL as expected, but it implicitly assumes equal proportion of true nonzero $\beta$'s in each group. For SOG, BSGS, and BSGS-SS, the posterior distribution of feature selection can allow for the control of BFDR. Under nominal level of BFDR at 10%, the actual FDR given the simulation truth are shown in Table 2. BSGS-SS was anti-conservative with 45% true FDR, while SOG and BSGS ($\tau^2 = 5$) properly controlled true FDR at 8% and 6%, respectively. In addition to smaller true FDR, SOG had only slightly higher FOR than BSGS-SS and lower than BSGS, showing its better feature selection performance.

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, BSGS-SS and HSVS. The setting was 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 Figure 2(A)). In other words, we set level-0 variable 1 to belong to both level-1 group 1 and 2, and level-0 variable 41 to belong to both group 3 and 4. To maintain the within group correlation of 0.5, for variables shared by more than one group, such as $x_{i,1}$, we first generated "pseudo" variables such as $x_{i,11}$ $(k = 1)$ and $x_{i,12}$ $(k = 2)$ as described in Section 4.1, and then set $x_{i,1}$ as the average of $x_{i,11}$ and $x_{i,12}$. $\beta_{jk}$'s and outcome $Y_i$ were generated the same way as in Section 4.1. Table 2 shows the evaluation results using 100 simulated data sets.

From the results, SOG continued to have the best variable selection and prediction performance. In fact, the results were very similar to simulation I. Even though we introduced partial coefficients due to overlapping groups (e.g., $\beta_{11}$ and $\beta_{12}$) which were unidentifiable by likelihood, we were still able to estimate the marginal effects (e.g., $\beta_1 = \beta_{11} + \beta_{12}$), which were identifiable.

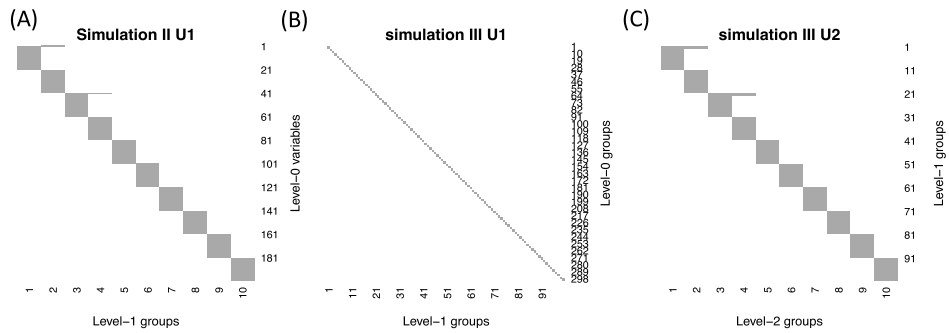



FIG. 2. *$U^{(1)}$ matrix in simulation II, $U^{(1)}$ and $U^{(2)}$ in simulation III. Grey denotes* 1, *which means variable/group in the row belongs to the group in the column*; *white denotes* 0.

4.3. *Simulation III*: *Two-layer overlapping groups*. In this simulation, we simulated 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 Figure 2(B) and 2(C). $U^{(1)}$ had a block diagonal structure, that is, every three features belonged to one level-1 group; $U^{(2)}$ had a block diagonal structure in the most parts except $U^{(2)}_{1,1} = U^{(2)}_{1,2} = 1$ and $U^{(2)}_{21,3} = U^{(2)}_{21,4} = 1$, that is, level-1 group 1 belonged to level-2 group 1 and 2; level-1 group 21 belonged to level-2 group 3 and 4.

In this setting, we only had overlapping level-2 groups while level-1 groups were disjoint. As a result, we could still compare MOG to SOG, BSGS, 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 the 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 $x_{ij} = z^{(1)}_k + z^{(2)}_l + e_{ij}$, where $e_{ij} \sim N(0, 0.5)$. In this way, $\mathrm{Var}(X_{ij}) = 1$. For variables belonging to the same level-1 group, the correlation was 0.5; for variables belonging to the same level-2 group but different level-1 groups, the correlation was 0.2. Variables shared by more than one group were generated the same way as in simulation II. Outcomes were also generated as $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$.

We set 5 out of 10 level-2 groups to contain relevant features. Inside these 5 predictive level-2 groups, we set 4 out of 10 level-1 groups to have strong signals, in which all three features had coefficients $\beta \sim \mathrm{Unif}(2U, 3U)$ ($U$ will vary); we set the other 2 of 10 level-1 groups to have medium signals, in which all three features had coefficients $\beta \sim \mathrm{Unif}(U, 2U)$; the remaining 4 level-1 groups had all 3 features with coefficients as zero. We set $U$ to be 0.2 and 0.5. In this setting, we did not have the true $\tau^2$ to set in BSGS. Instead, we tested $\tau^2 = 1, 2, \ldots, 5$, performing threefold cross-validation in the training set, and then selected the one with the smallest MSE.

Table 2 shows the comparison results for 100 replicates. MOG had the best performance in both variable selection and prediction, especially when $U$ was small. When $U = 0.2$, SOG had better performance than other models and MOG further improved SOG, demonstrating the benefit of incorporating level-2 grouping structure. When the signal was weak, BSGS had a severe convergence issue, even with 200,000 MCMC iterations, which also impaired its feature selection and prediction performance. BSGS-SS had smaller FDR but higher FOR than SOG, because it assumes the same sparsity inside groups. Inside the groups with weak signals, it missed some features weakly predictive of the outcome. At $U = 0.5$, all four Bayesian models obtained similar good performance in feature selection, but for prediction MSE, MOG still outperformed other Bayesian models. Lasso, GL and SGL had poor selection and prediction performance even when $U$ was large. GL performed better than Lasso and SGL, because sparsity was not designed inside level-1 groups in this simulation.

4.4. *Simulation IV*: *Two-layer nonoverlapping groups*. This simulation was designed to compare the performance of MOG and TGL since TGL does not allow groups at the same level to overlap as in simulation III. The only difference from the setting of simulation III was that level-2 groups did not overlap, so it was a straightforward extension of the tree structure which included multiple trees. The implementation of TGL is described in Section 3.2, and results are shown in Table 2.

Compared to lasso, GL and SGL, TGL had better variable selection and prediction performance as expected. However, MOG still outperformed TGL in the tree structure setting, regarding both variable selection and prediction. The improved performance is possibly because penalized regression methods are optimization-based and cannot incorporate complex structure and information flow as efficiently and naturally as Bayesian hierarchical models.

## 5. Applications.

5.1. *Predict ER+ versus ER− breast cancer.* We applied MOG to n = 727 (560 ER+ and 167 ER−) breast cancer patients retrieved from The Cancer Genome Atlas (TCGA). Each sample had mRNA expression, methylation and copy number variations (CNV) available. This application 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 constructed one summary methylation value for each gene by averaging the $\beta$ values within 50 kb of the gene starting position. $\beta$ values were later transformed to $M$ values ($M = \log(\beta/(1-\beta))$ to better fit model assumptions. After this filtering step, 14,976 features were left, of which 5125 were from mRNA expression data, 4816 were from CNV data, and 5035 were from methylation data.

Since BSGS-SS and HSVS are computationally intensive, we want to further filter genes and pathways. We first tested each mRNA expression feature for equal expression levels in ER+ and ER− groups. Then, we only kept the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways containing 40–50 genes and having 80% of the genes with mRNA expression significantly different in two groups ($t$-test $p < 0.05$), and filtered out features mapped to the genes that were not included in those selected pathways. 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. Among 824 features, 276 were from gene expression data, 274 were from CNV data and the remaining 274 were from methylation data.

For another more realistic setting, we relaxed our filtering criteria. We kept the KEGG pathways containing 20–50 genes and filtered out the features mapped to the genes which were not included in the selected pathways. In this way, $p =$ 11,785 multi-level omics features (1316 from mRNA expression data, 1292 from CNV data and 1302 from methylation data) belonging to $m_1 = 1316$ genes (level-1 groups) in $m_2 = 123$ pathways (level-2 groups) were left for analysis. This setting was used to compare the performance of MOG, SOG, lasso, GL and SGL.

Obviously, the "ER signaling pathway" should predict the ER status well. It was included in both 8 and 123 pathways selected and could 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 used tenfold cross-validation in the training set to select tuning parameters. Performance was evaluated using fivefold cross-validation by keeping the original case/control ratio in all folds. Each time, four folds of the ER+ and ER− samples were left for training, and one fold was left for testing. To avoid local optimal trapping and save time, when applying MOG and SOG, we used estimates from lasso as initial values. It took BSGS-SS and HSVS 1.4 and 19.7 hours to complete eight pathways example respectively, much longer than that of MOG (0.1 hours). These two models, especially HSVS, became inapplicable in larger data set such as those with 123 pathways.

To prioritize variable and group selection, we defined a feature impact score ($\mathrm{FIS}_j$) in MOG as the posterior average of the selection probability of feature $j$, that is, $\mathrm{FIS}_j = \mathrm{AVE}(\sum_{k=1}^{m_1}\sum_{l=1}^{m_2}\gamma_l^{(2)}\gamma_{kl}^{(1)}\gamma_{jkl}^{(0)}U_{jk}^{(1)}U_{kl}^{(2)})$, where $\mathrm{AVE}(\cdot)$ was the average over all MCMC iterations. The pathway impact score ($\mathrm{PIS}_l$) was then defined as the average of the selection probability of all level-0 variables included in pathway $l$, that is, $\mathrm{PIS}_l = \mathrm{AVE}(\sum_{j=1}^{p}\sum_{k=1}^{m_1}\gamma_l^{(2)}\gamma_{kl}^{(1)}\gamma_{jkl}^{(0)}U_{jk}^{(1)}U_{kl}^{(2)})$. In SOG, FIS and PIS were defined similarly, $\mathrm{FIS}_j = \mathrm{AVE}(\sum_{k=1}^{m_1}\gamma_k^{(1)}\gamma_{jk}^{(0)}U_{jk}^{(1)})$ and $\mathrm{PIS}_l = \mathrm{AVE}(\sum_{j=1}^{p}\sum_{k=1}^{m_1}\gamma_k^{(1)}\gamma_{jk}^{(0)}U_{jk}^{(1)}U_{kl}^{(2)})$. Setting $\gamma_k^{(1)} = 1$, denoting $\gamma_{jk}^{(0)} = 1$ if $\beta_{jk} \neq 0$, and denoting $\gamma_{jk}^{(0)} = 0$ otherwise, the definitions of FIS and PIS for BSGS-SS and HSVS are the same as SOG. We ranked the pathways and variables

TABLE 3

*Top pathways/features and prediction results in breast cancer ER+/− application. Results are from fivefold cross-validation*

A. 8 pathways

| Bayesian model | Top pathway by PIS | PIS | Top 3 selected features by FIS | $AUC_1$[a] (SD) | $AUC_2$[b] (SD) |
|---|---|---|---|---|---|
| MOG | ER signaling | 0.109 | ESR1-mRNA, ESR1-methyl, ESR1-CNV | 0.943 (0.008) | 0.949 (0.010) |
| SOG | ER signaling | 0.053 | ESR1-mRNA, ESR1-methyl, NME3-mRNA | 0.945 (0.009) | 0.948 (0.010) |
| BSGS-SS | ER signaling | 0.020 | ESR1-mRNA, NME3-mRNA, ADCY9-mRNA | 0.947 (0.009) | 0.948 (0.013) |
| HSVS | ER signaling | 0.027 | ESR1-mRNA, ESR1-CNV, ESR1-methyl | 0.942 (0.012) | 0.944 (0.012) |
| Penalized regression | Top pathway by Fisher's exact test | Fisher's exact test p-val | – | AUC (SD) | |
| Lasso | Calcium signaling | 0.179 | – | 0.945 (0.008) | |
| GL | ER signaling | 0.999 | – | 0.943 (0.011) | |
| SGL | ER signaling | 0.152 | – | 0.764 (0.108) | |
| TGL | AMPK signaling | 0.204 | | 0.946 (0.010) | |

B. 123 pathways

| Bayesian model[c] | Top pathway by PIS | PIS | Top 3 selected features by FIS | $AUC_1$ (SD) | $AUC_2$(SD) |
|---|---|---|---|---|---|
| MOG | ER signaling | 0.044 | ESR1-mRNA, ESR1-methyl, ESR1-CNV | 0.940 (0.013) | 0.944 (0.011) |
| SOG | Prolactin signaling | 0.031 | ESR1-mRNA, MARCKS-mRNA, ESR1-methyl | 0.943 (0.009) | 0.944 (0.011) |
| Penalized regression | Top pathway by Fisher's exact test | Fisher's exact test p-val | – | AUC (SD) | |
| Lasso | Dilated cardiomyopathy | 0.0004 | – | 0.946 (0.006) | |
| GL | RNA transport | 0.528 | – | 0.942 (0.011) | |
| SGL | Prolactin signaling | 0.021 | – | 0.681 (0.111) | |
| TGL | Adrenergic signaling | 0.007 | – | 0.944 (0.009) | |

[a]Plug-in $\hat{\beta}$ (posterior median).

[b]Model averaging.

[c]Computation not affordable with 123 pathways in BSGS-SS and HSVS models.

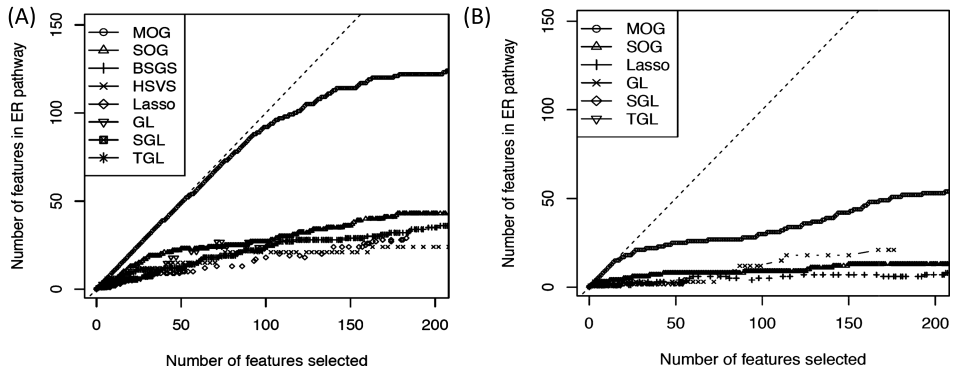



FIG. 3. *The number of top selected features versus the number of selected features belonging to the ER signaling pathway in breast cancer ER+/− application using* (*A*) 8 *pathways and* (*B*) 123 *pathways*.

based on their impact scores averaged over fivefold cross validations in Table 3. Top 20 selected multi-omics features by MOG are also listed in Table S1. Penalized regression models including lasso, GL and SGL, cannot readily prioritize variables and pathways. Instead, we performed pathway enrichment analysis applying Fisher's exact test to features selected at least once in fivefold cross-validation to prioritize the top pathways.

It is well known that the mRNA expression of ESR1 is predictive of ER status, defined by the immunohistochemistry (IHC) assay of estrogen receptor (ER). In both settings with 8 and 123 pathways, MOG detected the ER signaling pathway as the top selected pathway with the highest PIS, and ESR1-mRNA, ESR1-methyl and ESR1-CNV were among the top selected features. To obtain a better sense of the feature selection, we plotted the number of selected features ranked by FIS (x-axis) versus the number of selected features belonging to the ER signaling pathway (y-axis) in Figure 3. For lasso, GL, SGL and TGL, for which FIS was not available, we used the feature selection results with the first fold data left out, as leaving different folds out gave similar results. Most of the top features selected by MOG, belonged to the ER signaling pathway (e.g., 92 out of top 100 in Figure 3(A)). Nonetheless other models had much fewer features in ER signaling (e.g., SOG had 27 out of top 100 in Figure 3(A)).

To compare the prediction performance, we calculated ER prediction AUC for samples in the testing set. For Bayesian models, we performed two predictions: (1) plugging posterior median estimates of $\beta$ into $\hat{\Pr}(Y_i = 0) = \Phi(X\hat{\beta}^{\text{Med}})$ to obtain $\text{AUC}_1$; (2) using model averaging by calculating posterior mean of $\Phi(X\hat{\beta})$ to generate $\text{AUC}_2$. For lasso, GL, and SGL, we selected tuning parameter from tenfold cross-validation and plugged in $\hat{\beta}$. Having strong predictive genes such as ESR1, all models generated high AUCs in the testing set as expected. Comparing the two AUCs, $\text{AUC}_2$ was slightly higher than $\text{AUC}_1$ in general for the Bayesian models,

consistent with the common belief that averaging over all models from MCMC provides better predictive ability than using a single plug-in estimate. MOG using model averaging predictor generated the highest prediction AUC although the differences were not statistically significant given the almost perfect prediction for all models.

5.2. *Predict invasive lobular carcinoma* (*ILC*) *versus invasive ductal carcinoma* (*IDC*). We next applied MOG to predict histological subtypes (ILC/IDC) for 669 patients (496 IDCs, 173 ILCs) in the same TCGA data set. Invasive lobular carcinoma (ILC) constituting 10% of all invasive breast cancer cases, is the second most frequently diagnosed subtype, following invasive ductal carcinoma (IDC, 80%) (Ciriello et al. (2015)). We chose the same 123 KEEG pathways to compare the performance of MOG, SOG, lasso, GL and SGL. Variable selection and prediction performances are summarized in Table 4, and top 20 multi-omics features selected by MOG are listed in Table S2. Similar to ER status, there exists a well-known strong predictor CDH1 mRNA expression, as the loss of CDH1 is the hallmark of ILC (Ciriello et al. (2015)). Thus all models had good prediction AUCs. Since ILC is a less-studied subtype in breast cancer research, there is no annotated pathway specifically for this histologic subtype. The pathways identified by MOG provides proof-of-principle, as the top identified pathway termed "Endometrial Cancer" not only includes E-cadherin (CDH1), but also contains PI3K and Akt, two kinases that are activated as a result of loss of CDH1 (Ciriello et al. (2015), Teo et al. (2018)). And finally, there are a couple of genes such as APC, TCF7/TCF7L (Ravindranath and Cadigan (2016)) and LEF1 (Santiago et al. (2017)) that all belong to the Wnt signaling pathway, highlighting a unique role for this pathway as we (Sikora et al. (2016)) and others (Turashvili et al. (2007), van Hengel et al. (1999)) have previously shown. Another top pathway identified is related to "Amoebiasis", and it includes many genes known to play diverse roles in movement and motility of cells, such as serpins, laminins and extracellular movement, which we hypothesize is likely related to the different behavior of ILC cells, as a result of loss CDH1, and decreased cell-cell attachment, a phenotype that we have recently described in great detail (Tasdemir et al. (2018)).

**6. Conclusion and discussion.** In modern small-n-large-p applications, effective variable selection has become an increasingly important component in statistical methodologies. Models 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 hierarchical overlapping group structure that is commonly seen in the "multi-level omics features $\Rightarrow$ genes $\Rightarrow$ pathways" scenario in genomic applications. Our proposed Bayesian indicator variable selection model has several innovations and advantages for the targeted problem. First, the Bayesian hierarchical model and indicator variable selection model allow for natural incorporation of hierarchical group structure with

TABLE 4

*Top pathways/features and prediction results in breast cancer ILC/IDC application. Results are from fivefold cross-validation*

| Bayesian model[c] | Top pathway from PIS | PIS | Top 3 selected features by FIS | $AUC_1$[a] (SD) | $AUC_2$[b] (SD) |
|---|---|---|---|---|---|
| MOG | Endometrial cancer | 0.064 | CDH1-mRNA, LAMA3-mRNA, CDH1-methyl | 0.941 (0.008) | 0.949 (0.014) |
| SOG | Viral myocarditis | 0.044 | CDH1-mRNA, MAP3K1-mRNA, SHROOM1-mRNA | 0.911 (0.010) | 0.950 (0.012) |
| Penalized regression | Top pathway by Fisher's exact test | Fisher's exact test p-val | – | AUC (SD) | |
| Lasso | Thyroid hormone synthesis | 0.008 | – | 0.956 (0.009) | |
| GL | Notch signaling | 0.593 | – | 0.953 (0.011) | |
| SGL | Endometrial cancer | 0.017 | – | 0.901 (0.011) | |
| TGL | AMPK signaling | 0.005 | – | 0.955 (0.010) | |

[a]Plug-in $\hat{\beta}$ (posterior median).

[b]Model averaging.

[c]Computation not affordable with 123 pathways in BSGS-SS and HSVS models.

fast MCMC sampling. Second, we explicitly model group-specific proportions of nonzero $\beta$ values (i.e., $\pi_k^{(0)}$) for different sparsity levels in selected groups. Thirdly, our Bayesian approach easily allows for a latent decomposition assumption to incorporate overlapping groups. Fourth, the proposed model can be extended to more than two layers of overlapping group structure. The result gives clear interpretation of which features, genes and pathways contribute to the prediction. Finally, the posterior distribution from MCMC samples provides easy post hoc inferences, such as characterization of variability and BFDR control of feature selection. Using four simulation settings and two breast cancer examples, we demonstrated superior performance of the proposed method for hierarchical overlapping group (MOG) structure in terms of variable selection, prediction accuracy and model interpretation.

Our proposed model has several limitations to be improved in the future. First, as noted in the paper, the MCMC mixing rate in the indicator model can be unstable, leading to slow convergence. Although our current simulation and application can be implemented adequately, we expect worse performance when $p$ increases or the data signal becomes weaker. A modification to spike-and-slab prior with a small-variance Gaussian spike might alleviate the computing difficulty. Second, in SOG/MOG, feature sparsity varies by gene groups. To better allow for heterogeneity among multi-omics platforms, a more sophisticated sparsity modeling may be needed to allow for different levels of sparsity in different platforms. Specifically, taking MOG as an example, we can design feature sparsity prior through a probit function: $\gamma_{jkl}^{(0)} \sim \text{Bern}(\Phi(\mu_{kl}^{(0)}/R_j + \mu_m))$, where $\Phi(\cdot)$ is the CDF of the standard normal distribution, $\mu_{kl}^{(0)}$ is the feature selection strength of gene $k$ in pathway $l$, and $\mu_m$ is the feature selection strength of multi-omics platform $m$. Since this implementation may bring computational challenges and significantly slow down computing time, we did not implement it in this paper to allow for practical omics applications, but we consider this as a future extension. As large data sets with complex prior information structure continue to accumulate in data science, we expect to encounter the hierarchical overlapping group structure more often in the future, and the proposed method can better incorporate prior information to improve statistical learning performance.

An efficient R package "MOG" calling C++ using RcppEigen (Bates and Eddelbuettel (2013)) is available at github (https://github.com/lizhu06/MOG). Processed data and R code to reproduce result is available at github (https://github.com/lizhu06/MOG_exp). The computing time for MOG to predict ER+/ER− with 123 pathways is 2.33 hours, and computing time to predict ILC/IDC is 2.26 hours with 16 CPU cores, 1.4 GHz and 128 GB RAM.

**Acknowledgments.** We would like to thank Dr. Lin Zhang for providing the code of the HSVS method, Dr. Kuo-Jung Lee and Dr. Ray-Bing Chen for sending us the most updated version of BSGS package. We would also like to thank all

reviewers for many insightful comments that led to significant improvement of our work.

## SUPPLEMENTARY MATERIAL

**Supplement to "Bayesian indicator variable selection to incorporate hierarchical overlapping group structure in multi-omics applications"** (DOI: 10.1214/19-AOAS1271SUPP; .pdf). We provide MCMC sampling details, proof of asymptotic properties of SOG and MOG, an additional simulation borrowing information across groups in SOG and supporting tables for applications.

## REFERENCES

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

BATES, D. and EDDELBUETTEL, D. (2013). Fast and elegant numerical linear algebra using the RcppEigen package. *J. Stat. Softw.* **52** 1–24.

CHEN, R.-B., CHU, C.-H., YUAN, S. and WU, Y. N. (2016). Bayesian sparse group selection. *J. Comput. Graph. Statist.* **25** 665–683. MR3533632

CIRIELLO, G., GATZA, M. L., BECK, A. H., WILKERSON, M. D., RHIE, S. K., PASTORE, A., ZHANG, H., MCLELLAN, M., YAU, C. et al. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163** 506–519.

EBERLY, L. E. and CARLIN, B. P. (2000). Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Stat. Med.* **19** 2279–2294.

GELFAND, A. E. and SAHU, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *J. Amer. Statist. Assoc.* **94** 247–253. MR1689229

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GEWEKE, J. et al. (1991). Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN.

HERNÁNDEZ-LOBATO, D., HERNÁNDEZ-LOBATO, J. M. and DUPONT, P. (2013). Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation. *J. Mach. Learn. Res.* **14** 1891–1945. MR3104499

JACOB, L., OBOZINSKI, G. and VERT, J.-P. (2009). Group lasso with overlap and graph lasso. In *Proceedings of the* 26*th Annual International Conference on Machine Learning* 433–440. ACM, New York.

JENATTON, R., MAIRAL, J., OBOZINSKI, G. and BACH, F. (2011). Proximal methods for hierarchical sparse coding. *J. Mach. Learn. Res.* **12** 2297–2334. MR2825428

JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. MR2089135

KUO, L. and MALLICK, B. (1998). Variable selection for regression models. *Sankhya, Ser. B* **60** 65–81. MR1717076

KYUNG, M., GILL, J., GHOSH, M. and CASELLA, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal.* **5** 369–411. MR2719657

LIU, J., JI, S., YE, J. et al. (2009). SLEP: Sparse learning with efficient projections. The Biodesign Institute, Arizona State University, Tempe, AZ.

LOCK, E. F. and DUNSON, D. B. (2017). Bayesian genome- and epigenome-wide association studies with gene level dependence. *Biometrics* **73** 1018–1028. MR3713135

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. MR0997578

NEWTON, M. A., NOUEIRY, A., SARKAR, D. and AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–176.

O'HARA, R. B. and SILLANPÄÄ, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Anal.* **4** 85–117. MR2486240

PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. MR2524001

RAVINDRANATH, A. and CADIGAN, K. (2016). The role of the C-clamp in wnt-related colorectal cancers. *Cancers* **8** 74.

RICHARDSON, S., TSENG, G. C. and SUN, W. (2016). Statistical methods in integrative genomics. *Annu. Rev. Stat. Appl.* **3** 181–209.

SANTIAGO, L., DANIELS, G., WANG, D., DENG, F.-M. and LEE, P. (2017). Wnt signaling pathway protein LEF1 in cancer, as a biomarker for prognosis and a target for treatment. *American Journal of Cancer Research* **7** 1389.

SIKORA, M. J., JACOBSEN, B. M., LEVINE, K., CHEN, J., DAVIDSON, N. E., LEE, A. V., ALEXANDER, C. M. and OESTERREICH, S. (2016). WNT4 mediates estrogen receptor signaling and endocrine resistance in invasive lobular carcinoma cell lines. *Breast Cancer Res.* **18** 92.

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245. MR3173712

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. *Ann. Appl. Stat.* **5** 1978–2002. MR2884929

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357

TASDEMIR, N., BOSSART, E. A., LI, Z., ZHU, L., SIKORA, M. J., LEVINE, K. M., JACOBSEN, B. M., TSENG, G. C., DAVIDSON, N. E. et al. (2018). Comprehensive phenotypic characterization of human invasive lobular carcinoma cell lines in 2D and 3D cultures. *Cancer Res.* 6209–6222.

TEO, K., GÓMEZ-CUADRADO, L., TENHAGEN, M., BYRON, A., RÄTZE, M., VAN AMERSFOORT, M., RENES, J., STRENGMAN, E., MANDOLI, A. et al. (2018). E-cadherin loss induces targetable autocrine activation of growth factor signalling in lobular breast cancer. *Sci. Rep.* **8** 15454.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TURASHVILI, G., BOUCHAL, J., BAUMFORTH, K., WEI, W., DZIECHCIARKOVA, M., EHRMANN, J., KLEIN, J., FRIDMAN, E., SKARDA, J. et al. (2007). Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC Cancer* **7** 55.

VAN HENGEL, J., VANHOENACKER, P., STAES, K. and VAN ROY, F. (1999). Nuclear localization of the p120ctn armadillo-like catenin is counteracted by a nuclear export signal and by E-cadherin expression. *Proc. Natl. Acad. Sci. USA* **96** 7980–7985.

XU, X. and GHOSH, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Anal.* **10** 909–936. MR3432244

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574

ZHANG, L., BALADANDAYUTHAPANI, V., MALLICK, B. K., MANYAM, G. C., THOMPSON, P. A., BONDY, M. L. and DO, K.-A. (2014a). Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 595–620. MR3258055

ZHANG, L., MORRIS, J. S., ZHANG, J., ORLOWSKI, R. Z. and BALADANDAYUTHAPANI, V. (2014b). Bayesian joint selection of genes and pathways: Applications in multiple myeloma genomics. *Cancer Inform.* **13** 113.

ZHAO, P., ROCHA, G. and YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **37** 3468–3497. MR2549566

ZHU, L., HUO, Z., MA, T., OESTERREICH, S. and TSENG, G. C. (2019). Supplement to "Bayesian indicator variable selection to incorporate hierarchical overlapping group structure in multi-omics applications." DOI:10.1214/19-AOAS1271SUPP.

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

L. ZHU
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF PITTSBURGH
PITTSBURGH, PENNSYLVANIA 15261
USA
E-MAIL: liz86@pitt.edu

Z. HUO
DEPARTMENT OF BIOSTATISTICS
COLLEGE OF PUBLIC HEALTH AND HEALTH PROFESSIONS
COLLEGE OF MEDICINE
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA 32611
USA
E-MAIL: zhuo@ufl.edu

T. MA
DEPARTMENT OF EPIDEMIOLOGY AND BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
UNIVERSITY OF MARYLAND
COLLEGE PARK, MARYLAND 20742
USA
E-MAIL: tim28@pitt.edu

S. OESTERREICH
DEPARTMENT OF PHARMACOLOGY AND CHEMICAL BIOLOGY
UNIVERSITY OF PITTSBURGH
PITTSBURGH, PENNSYLVANIA 15261
USA
E-MAIL: oesterreichs@upmc.edu

G. C. TSENG
DEPARTMENTS OF BIOSTATISTICS, HUMAN GENETICS, AND COMPUTATIONAL BIOLOGY
UNIVERSITY OF PITTSBURGH
PITTSBURGH, PENNSYLVANIA 15261
USA
E-MAIL: ctseng@pitt.edu