Bioinformatics, 34(6), 2018, 901–910
doi: 10.1093/bioinformatics/btx684
Advance Access Publication Date: 25 October 2017
Original Paper

OXFORD

Genome analysis

Group spike-and-slab lasso generalized linear models for disease prediction and associated genes detection by incorporating pathway information

Zaixiang Tang[1,2,3,4], Yueping Shen[1,2], Yan Li[4], Xinyan Zhang[4], Jia Wen[5], Chen'ao Qian[6], Wenzhuo Zhuang[7], Xinghua Shi[5] and Nengjun Yi[4,*]

[1]Department of Biostatistics, School of Public Health, [2]Jiangsu Key Laboratory of Preventive and Translational Medicine for Geriatric Diseases, [3]Center for Genetic Epidemiology and Genomics, Medical College of Soochow University, Suzhou 215123, China, [4]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294, USA, [5]Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA, [6]Department of Bioinformatics and [7]Department of Cell Biology, School of Biology & Basic Medical Science, Soochow University, Suzhou 215123, China

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on June 15, 2017; revised on October 5, 2017; editorial decision on October 21, 2017; accepted on October 24, 2017

Abstract

Motivation: Large-scale molecular data have been increasingly used as an important resource for prognostic prediction of diseases and detection of associated genes. However, standard approaches for omics data analysis ignore the group structure among genes encoded in functional relationships or pathway information.

Results: We propose new Bayesian hierarchical generalized linear models, called group spike-and-slab lasso GLMs, for predicting disease outcomes and detecting associated genes by incorporating large-scale molecular data and group structures. The proposed model employs a mixture double-exponential prior for coefficients that induces self-adaptive shrinkage amount on different coefficients. The group information is incorporated into the model by setting group-specific parameters. We have developed a fast and stable deterministic algorithm to fit the proposed hierarchal GLMs, which can perform variable selection within groups. We assess the performance of the proposed method on several simulated scenarios, by varying the overlap among groups, group size, number of non-null groups, and the correlation within group. Compared with existing methods, the proposed method provides not only more accurate estimates of the parameters but also better prediction. We further demonstrate the application of the proposed procedure on three cancer datasets by utilizing pathway structures of genes. Our results show that the proposed method generates powerful models for predicting disease outcomes and detecting associated genes.

Availability and implementation: The methods have been implemented in a freely available R package BhGLM (http://www.ssg.uab.edu/bhglm/).

Contact: nyi@uab.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Large-scale omics data provide extraordinary opportunities for detecting new biomarkers and building accurate prognostic and predictive models. However, such data also introduce statistical and computational challenges. One particular challenge is that these data are typically high dimensional, which makes penalized sparse models a good fit for analyzing such data. Since the original proposal of the lasso by Tibshirani (1996), penalized regressions for variable selection in high-dimensional settings have attracted considerable attention in modern statistical research (Hastie *et al.*, 2009, 2015; Tibshirani, 1996). These methods put L_1-penalty on the coefficients and shrink many coefficients exactly to zero, thus performing variable selection. Zhang (2007, 2010) introduced a minimax concave penalty (MCP) for high-dimensional regressions, which is motivated by and similar to the smoothly clipped absolute deviation (SCAD) penalty function (Fan and Li, 2001). Both MCP and SCAD aim to eliminate unimportant predictors from the model while leaving important predictors un-penalized. MCP and SCAD are thus asymptotically oracle-efficient (Zhang, 2007, 2010). These penalization approaches have been widely applied for disease prediction and prognosis using large-scale molecular data (Barillot *et al.*, 2012; Rapaport *et al.*, 2007; Sohn *et al.*, 2013; Yuan *et al.*, 2014; Zhang *et al.*, 2013; Zhao *et al.*, 2015).

Furthermore, researchers have noticed that molecular predictors possess natural group structures which can be used in improving analysis. For example, when analyzing gene expression data, one can group genes into functionally similar sets as in Gene Ontology (GO) terms (Gene Ontology, 2015) or into known biological pathways as in the Kyoto encyclopedia of genes and genomes (KEGG) pathways (Kanehisa *et al.*, 2016). In genetic association studies, some single nucleotide polymorphisms (SNPs) may fall within the intragenic and regulatory regions of a given gene, where the genetic region can be considered as a group structure. In microbiome research, microbes are classified at multiple taxonomy levels (i.e. phylum, class, order, family, genus and species), within each taxonomy level having many subdivisions, where these taxonomy levels can be considered as a group structure. These examples show complex group structures among predictors, in various forms including hierarchical and overlapping groups, which have been often ignored in genetic modeling. Nonetheless, it is desirable to incorporate such biological grouping information into modeling, since it may improve both the interpretability and prediction accuracy of the models.

Several methods have been recently proposed to utilize the grouping information of high-dimensional data. A popular method is the group Lasso (Yuan and Lin, 2006), which performs group level selection, including or excluding an entire group of variables. Meier *et al.* (2008) extended the group lasso to logistic regression models and presented an efficient algorithm for high-dimensional problems. Zhao *et al.* (2009) proposed a general composite absolute penalty for group selection, which includes the group lasso as a special case. However, the group lasso does not achieve sparsity within each group, which may introduce a suboptimal model. To overcome this deficiency, Friedman *et al.* (2010a,b) proposed the sparse group lasso (SGL) to achieve sparsity at both group and predictor levels. Simon *et al.* (2013) recently proposed a generalized gradient descent algorithm for SGL, and considered applications of this method for linear, logistic and Cox regressions. Several other methods have also been developed for bi-level selection, such as group bridge (Huang *et al.*, 2009), composite MCP (cMCP) (Breheny and Huang, 2009), group exponential Lasso (Breheny, 2015), group variable selection via convex log-exp-sum penalty (Chen *et al.*, 2014a,b) and doubly sparse approach for group variable selection (Kwon *et al.*, 2016).

Overlapping is a common phenomenon in biological pathway strusctures (i.e. a gene can belong to more one pathway). To deal with overlapping structures, a direct solution is to duplicate overlapping predictors into different groups so that predictors belonging to more than one group can enter the model separately (Jacob *et al.*, 2009; Silver *et al.*, 2012), which has been used to identify pathways associated with a trait of interest (Silver *et al.*, 2013). Several other methods were also proposed for handling overlapping group structure (Chen *et al.*, 2014a,b; Obozinski *et al.*, 2011; Yuan *et al.*, 2013). Ogutu and Piepho (2014) reviewed and compared these regularization methods in genomic prediction. Huang *et al.* (2012) gave a selective review of group selection methods, described several applications of these methods in non-parametric additive models, semiparametric regression, seemingly unrelated regressions, genomic data analysis and genome-wide association studies, and highlighted some issues for further study.

The aforementioned approaches are non-Bayesian approaches. Recently, Ročková and George (2016a,b) proposed a new Bayesian approach, called the spike-and-slab lasso, for high-dimensional normal linear models using the spike-and-slab mixture double-exponential prior distribution. The spike-and-slab prior is the fundamental basis for most Bayesian variable selection approaches and has proved remarkably successful (Chipman, 1996; Chipman *et al.*, 2001; George and McCulloch, 1993, 1997; Ročková and George, 2014, 2016a). The mixture priors have been applied to predictive modeling and variable selection in large-scale genomic studies (de los Campos *et al.*, 2010; Ishwaran and Rao, 2005; Lu *et al.*, 2015; Partovi Nia and Ghannad-Rezaie, 2016; Shankar *et al.*, 2015; Shelton *et al.*, 2015; Yi *et al.*, 2003; Zhou *et al.*, 2013). We have recently incorporated this prior with GLMs and Cox models, and developed the spike-and-slab lasso GLMs and Cox models for prediction and gene detection, respectively (Tang *et al.*, 2017a,b).

In this article, we propose a novel group spike-and-slab lasso GLMs (gsslasso GLMs) for predicting disease outcomes and detecting associated genes by incorporating biological group structures into the spike-and-slab lasso framework. We propose an efficient algorithm to fit the group spike-and-slab lasso GLMs by integrating Expectation-Maximization (EM) steps into the extremely fast cyclic coordinate descent algorithm. We assess the performance of the proposed method via extensive simulations and compare with several commonly used methods. We apply the proposed procedure to three cancer datasets with binary outcomes and thousands of molecular features with pathways information. Our results show that the proposed method not only generates powerful prognostic models for predicting disease outcome but also excels at detecting associated genes.

2 Materials and methods

2.1 The group spike-and-slab lasso GLMs

We consider generalized linear models (GLMs) with a large number of structured predictors. For individual i, we denote the observed value of a continuous or discrete response by y_i, and the jth predictor by x_{ij}. The predictor variables include numerous molecular predictors (e.g. gene expression) and some relevant covariates. Assume that the molecular predictors can be organized into G groups (e.g. biological pathways), and the predictors within one group are biologically related. In certain applications, some variables may belong

to more than one group: e.g. genes can belong to more than one biological pathway. Following the idea of overlap group lasso (Hastie *et al.*, 2015; Jacob *et al.*, 2009; Silver *et al.*, 2012, 2013), we expand the vector of predictors by replicating a variable in whatever group it appears.

In GLMs, the mean of the response variable is related to the linear predictor $X_i\beta$ via a link function h (Gelman *et al.*, 2014; McCullagh and Nelder, 1989):

$$h[E(y_i \mid X_i)] = \beta_0 + \sum_{j=1}^{J} x_{ij}\beta_j = X_i\beta, \tag{1}$$

where β_0 is the intercept, β_j is the coefficient of the jth predictor, X_i contains all variables, and β is a vector of the intercept and all the coefficients. The data distribution is expressed as

$$p(y \mid X\beta, \phi) = \prod_{i=1}^{n} p(y_i \mid X_i\beta, \phi), \tag{2}$$

where ϕ is a dispersion parameter, and the distribution $p(y_i \mid X_i\beta, \phi)$ can take various forms, including Normal, Binomial, and Poisson distributions. Some GLMs (e.g. the binomial and Poisson distributions) do not require a dispersion parameter; that is, ϕ is fixed at 1.

For high dimensional and/or correlated data, the model is often unreliably fitted using the classical maximum likelihood procedure. The problem can be solved by using Bayesian hierarchical modeling or penalization approaches (Gelman *et al.*, 2014; Gelman and Hill, 2007; Hastie *et al.*, 2015). We employ a Bayesian hierarchical modeling approach, which allows us to obtain reliable estimation and more importantly provides an efficient way to incorporate group information. Our hierarchical GLMs specify the spike-and-slab mixture double-exponential (de) prior on the coefficients:

$$\begin{aligned}\beta_j \mid \gamma_j &\sim \mathrm{de}(0,\ (1-\gamma_j)s_0 + \gamma_j s_1)\\ &= \frac{1}{(1-\gamma_j)s_0 + \gamma_j s_1}\exp\left(-\frac{|\beta_j|}{(1-\gamma_j)s_0 + \gamma_j s_1}\right),\end{aligned} \tag{3}$$

where γ_j is the indicator variable: $\gamma_j = 1$ or 0, and the scale parameters, s_0 and s_1 ($s_1 > s_0 > 0$), are small and relatively large (e.g. $s_0 = 0.05$, $s_1 = 1$), inducing strong or weak shrinkage on β_j respectively. Thus the prior is a mixture of the shrinkage prior de(0, s_0) and the weakly informative prior de(0, s_1), which are spike and slab components, respectively.

We specify the distributions of indicator variables by incorporating the group structure. For predictors in group g, the indicator variables are assumed to follow the binomial distribution with the group-specific probability θ_g:

$$\gamma_j \mid \theta_g \sim \mathrm{Bin}\left(\gamma_j \mid 1, \theta_g\right) = \theta_g^{\gamma_j}\left(1-\theta_g\right)^{1-\gamma_j}. \tag{4}$$

The parameter θ_g will be estimated to be large if group g has important predictors, encouraging other predictors in the group more likely to be important. Therefore, the group-specific probability parameters, θ_g, play a role on incorporating the biological similarity of genes within a same pathway into the hierarchical model. For simplicity and convenience, we use the uniform prior for θ_g: $\theta_g \sim U(0, 1)$. Hereafter, the hierarchical GLMs with the group spike-and-slab mixture double-exponential priors are referred to as the group spike-and-slab lasso GLMs.

2.2 Algorithm

We develop a fast deterministic algorithm to fit the group spike-and-slab lasso GLMs. Our algorithm, called the EM coordinate descent algorithm, incorporates EM steps into the cyclic coordinate descent procedure for fitting the penalized lasso GLMs. We derive the EM coordinate descent algorithm based on the log joint posterior density of the parameters $\vartheta = (\beta, \phi, \gamma, \theta)$:

$$\begin{aligned}\log p(\vartheta \mid y, X) &\propto \textstyle\sum_{i=1}^{n} \log p(y_i \mid X_i\beta, \phi) + \sum_{j=1}^{J} \log p(\beta_j \mid \gamma_j)\\ &\quad + \textstyle\sum_{j=1}^{J} \log p(\gamma_j \mid \theta_g) + \sum_{g=1}^{G} p(\theta_g)\\ &\propto \textstyle\sum_{i=1}^{n} \log p(y_i \mid X_i\beta, \phi) - \sum_{j=1}^{J} [(1-\gamma_j)s_0 + \gamma_j s_1]^{-1}|\beta_j|\\ &\quad + \textstyle\sum_{j=1}^{J} \log [\theta_g^{\gamma_j}(1-\theta_g)^{1-\gamma_j}]\end{aligned} \tag{5}$$

The EM coordinate decent algorithm treats the indicator variables as 'missing values' and estimates the parameters (β, ϕ, θ) by averaging the missing values over their posterior distributions, where $\theta = (\theta_1, \ldots, \theta_G)$. For the E-step, we calculate the expectation of the log joint posterior density with respect to the conditional posterior distributions of the missing data. For predictors in group g, the conditional posterior expectation of the indicator variable γ_j can be derived as

$$\begin{aligned}p_j^g &= p(\gamma_j = 1 \mid \beta_j, \theta_g)\\ &= \frac{p(\beta_j \mid \gamma_j = 1, s_1)p(\gamma_j = 1 \mid \theta_g)}{p(\beta_j \mid \gamma_j = 0, s_0)p(\gamma_j = 0 \mid \theta_g) + p(\beta_j \mid \gamma_j = 1, s_1)p(\gamma_j = 1 \mid \theta_g)},\end{aligned} \tag{6}$$

where $p(\gamma_j = 1 \mid \theta_g) = \theta_g$, $p(\beta_j \mid \gamma_j = 1, s_1) = \mathrm{de}(\beta_j \mid 0, s_1)$, and $p(\beta_j \mid \gamma_j = 0, s_0) = \mathrm{de}(\beta_j \mid 0, s_0)$. Therefore, the conditional posterior expectation of $[(1-\gamma_j)s_0 + \gamma_j s_1]^{-1}$ can be obtained by

$$\lambda_j = E([(1-\gamma_j)s_0 + \gamma_j s_1]^{-1} \mid \beta_j) = \frac{1-p_j^g}{s_0} + \frac{p_j^g}{s_1}. \tag{7}$$

For the M-step, we update (β, ϕ, θ) by maximizing the posterior expectation of the log joint posterior density with γ_j and $[(1-\gamma_j)s_0 + \gamma_j s_1]^{-1}$ replaced by their conditional posterior expectations p_j^g and λ_j. From the log joint posterior density, we observe that (β, ϕ), and θ can be updated separately, because the parameters (β, ϕ) are only involved in the first two terms of the log joint posterior density and the probability parameters θ_g are only involved in the third term. Therefore, the parameters (β, ϕ) are updated by maximizing the expression:

$$Q_1(\beta, \phi) = \sum_{i=1}^{n} \log p(y_i \mid X_i\beta, \phi) - \sum_{j=1}^{J} \lambda_j|\beta_j|. \tag{8}$$

The term $\sum_{j=1}^{J} \lambda_j|\beta_j|$ serves as the L_1 lasso penalty with λ_j as the penalty factors, and thus the coefficients can be updated by maximizing $Q_1(\beta, \phi)$ using the cyclic coordinate decent algorithm (Friedman *et al.*, 2010a,b; Hastie *et al.*, 2015). Therefore, the coefficients can be estimated to be zero. The probability parameters $\{\theta_g\}$ are updated by maximizing the expression:

$$Q_2(\theta) = \sum_{j=1}^{J} \left[p_j^g \log \theta_g + \left(1-p_j^g\right) \log\left(1-\theta_g\right)\right]. \tag{9}$$

We can easily obtain: $\theta_g = \frac{1}{J_g}\sum_{j\in g} p_j^g$, where J_g is the number of predictors belonging to group g.

In summary, the EM coordinate decent algorithm proceeds as follows:

i. Choose a starting value for β^0, ϕ^0 and θ_g^0. For example, we can initialize $\beta^0 = 0$, $\phi^0 = 1$ and $\theta_g^0 = 0.5$.
ii. For t = 1, 2, 3 . . .,

E-step: Update γ_j and $[(1-\gamma_j)s_0 + \gamma_j s_1]^{-1}$ by their conditional posterior expectations.

M-step:

i. Update (β, ϕ) using the cyclic coordinate decent algorithm;
ii. Update $\theta_1, \ldots, \theta_G$.

We assess convergence by the criterion: $|d^{(t)} - d^{(t-1)}|/(0.1+|d^{(t)}|) < \varepsilon$, where $d^{(t)} = -2\sum_{i=1}^{n} \log p(y_i \mid X_i\beta^{(t)}, \phi^{(t)})$ is the estimate of deviance at the tth iteration, and ε is a small value (say 10^{-5}).

2.3 Evaluating the predictive performance of a fitted model

There are several measures to evaluate the quality of a fitted GLM (Steyerberg, 2009), including: (i) Deviance: $d = -2\sum_{i=1}^{n} \log p(y_i \mid X_i\widehat{\beta}, \widehat{\phi})$; (ii) Mean squared error (MSE): $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$, where $\widehat{y}_i = h^{-1}(X_i\widehat{\beta})$; For logistic models, we can use two additional measures: (iii) area under the ROC curve (AUC) and (iv) misclassification: $\frac{1}{n}\sum_{i=1}^{n} I(|y_i - \widehat{y}_i| > 0.5)$, where $I(|y_i - \widehat{y}_i| > 0.5) = 1$ if $|y_i - \widehat{y}_i| > 0.5$, and $I(|y_i - \widehat{y}_i| > 0.5) = 0$ if $|y_i - \widehat{y}_i| \leq 0.5$.

To assess the predictive performance of the model, we use the pre-validation method, a variant of cross-validation (Hastie *et al.*, 2015; Tibshirani and Efron, 2002), that randomly partitions the data to K subsets of roughly the same size and uses $(K-1)$ subsets to fit a model. Denote the estimate of coefficients from the data excluding the kth subset by $\widehat{\beta}^{(-k)}$. We calculate the linear predictor $\widehat{\eta}_{(k)} = X_{(k)}\widehat{\beta}^{(-k)}$ for all individuals in the k-th subset of the data, called the cross-validated or pre-validated predicted index. Cycling through K parts, we obtain the cross-validated linear predictor $\widehat{\eta}_i$ for all individuals. We then use $\{y_i, \widehat{\eta}_i\}$ to compute the measures described above. The cross-validated linear predictor for each individual is derived independently of the observed response of the individual, and hence the 'pre-validated' dataset $\{y_i, \widehat{\eta}_i\}$ can essentially be treated as a 'new dataset'. Therefore, the pre-validation procedure provides valid assessment of the predictive performance of the model (Hastie *et al.*, 2015; Tibshirani and Efron, 2002). To get more stable results, we can run K-fold cross-validation multiple times and average the measures over the replicates.

We also perform the pre-validated linear predictor analysis (Hastie *et al.*, 2015; Tibshirani and Efron, 2002), i.e. using the pre-validated linear predictor $\widehat{\eta}_i$ as a continuous covariate to fit the model: $E(y_i \mid \widehat{\eta}_i) = h^{-1}(\mu + \widehat{\eta}_i b)$. We then look at the P-value for testing the hypothesis $b = 0$ or the measures to evaluate the predictive performance. We can transform the continuous pre-validated linear predictor $\widehat{\eta}_i$ into a categorical factor $c_i = (c_{i1}, \ldots, c_{i6})$ based on the quantiles of $\widehat{\eta}_i$, e.g. 5, 25, 50, 75 and 95% quantiles, and then fit the model: $E(y_i \mid c_i) = h^{-1}(\mu + \sum_{k=2}^{6} c_{ik} b_k)$. This allows us to compare statistical significance and prediction between different categories.

2.4 Selecting optimal scale values

The performance of the group spike-and-slab lasso approach can depend on the scale parameters (s_0, s_1). Rather than restricting attention to a single model, we fix the slab scale s_1 (e.g. $s_1 = 1$) which provides no or weak shrinkage, and consider a sequence of L decreasing values $\{s_0^l\}$: $0 < s_0^1 < s_0^2 < \cdots < s_0^L < s_1$, for the spike scale s_0 (Ročková and George, 2014, 2016a; Tang *et al.*, 2017a,b). Increasing the spike scale s_0 tends to include more non-zero coefficients in the model. We can use one of the measures described in the last section to choose a model. This procedure is similar to the lasso implemented in the widely used R package glmnet, which quickly fits the lasso model over a grid of values of λ covering its entire range, giving a sequence of models for users to choose from (Friedman *et al.*, 2010a,b; Hastie *et al.*, 2015).

2.5. Implementation

We have created an R function bmlasso for setting up and fitting the spike-and-slab lasso GLMs and several other R functions for simulating predictor data and several outcomes, for summarizing the fitted models and for evaluating the predictive performance. We have incorporated these functions into the freely available R package BhGLM (http://www.ssg.uab.edu/bhglm/). A clear instruction is also included in the help file of the package.

3 Simulation study

We used simulations to assess the proposed approach, and compare with the lasso implemented in the R package glmnet and several penalization methods that can incorporate group information, including SGL in the R package SGL, overlap group lasso (grlasso), overlap group MCP (grMCP), overlap group SCAD (grSCAD) and overlap group cMCP in the R package grpregOverlap (Zeng and Breheny, 2016). Although the proposed method can be applied to any GLMs, we focused on the hierarchical logistic model because we analyzed binary outcomes in our real datasets (see the next section). In each situation, we simulated two datasets, and used the first one as the training data to fit the models and the second one as the test data to evaluate the predictive values. Our simulation method was similar to Tang *et al.* (2017a,b), but accounted for additional complexities of varied group structures. We considered five simulation scenarios with different complexities, including non-overlap or overlap groups, group sizes, number of non-null groups, and correlation coefficients (r) (Table 1). In simulation Scenarios 2–5, overlap structures were considered. To handle the overlap structures, we duplicated overlapping predictors into groups that predictors belong to (Jacob *et al.*, 2009; Silver *et al.*, 2012).

For each dataset, we generated n (= 500) observations, each with a binary response y_i and a vector of m (= 1000) continuous predictors $X_i = (x_{i1}, \ldots, x_{im})$. The 1000 predictors were organized into 20 groups. The vector X_i was randomly sampled from multivariate normal distribution $N_{1000}(0, \Sigma)$, where the covariance matrix Σ was set to account for varied grouped correlation and overlapped structures under different simulation scenarios. The predictors within a group were simulated to be correlated and those

Table 1. The preset non-zero coefficients and their values of the different simulation scenarios

Simulation scenarios	Non-zero coefficients and effect size							
1. Non-overlap group								
	β_5	β_{20}	β_{40}	β_{210}	β_{220}	β_{240}	β_{975}	β_{995}
2. Overlap group								
	β_5	β_{20}	β_{40}	β_{210}	β_{220}	β_{240}	β_{975}	β_{995}
3. varying group size								
=4/20/50	β_1	β_2	β_3	β_4	β_{501}	β_{502}	β_{503}	β_{504}
4. varying number of non-null groups								
=8	β_5	β_{55}	β_{305}	β_{355}	β_{505}	β_{555}	β_{905}	β_{955}
=3	β_5	β_{15}	β_{25}	β_{355}	β_{365}	β_{375}	β_{905}	β_{915}
=1	β_5	β_{10}	β_{15}	β_{20}	β_{25}	β_{30}	β_{35}	β_{40}
5. varying correlation within group								
$r = 0.0/0.3/0.5/0.7$	β_5	β_{20}	β_{40}	β_{210}	β_{220}	β_{240}	β_{975}	β_{995}
Effect size								
	0.8	−0.7	1.0	−0.9	−0.8	0.9	−1.0	0.7

predictors in different groups were independent. Except for Scenario 5, the correlation coefficient r was set to be 0.6. To simulate the binary response, we first generated Gaussian response z_i from univariate normal distribution $N(\eta_i, 1.6^2)$, where $\eta_i = \beta_0 + \sum_{j=1}^{m} x_{ij}\beta_j$, and then transformed the continuous response to a binary data by setting individuals with the 30% largest continuous response Z as 'affected' ($y_i = 1$) and the other individuals as 'unaffected' ($y_i = 0$). For all the scenarios, we set eight coefficients to be non-zero and the others to be zero (see Table 1).

For each simulation setting, we replicated the simulation 100 times and summarized the results over the replicates. We reported the results on the predictive measures including deviance, MSE, AUC, misclassification in the test data and the accuracy of parameter estimates. We used deviance to choose an optimal model. For the penalization methods, we used 10-fold cross-validation to select an optimal penalty value, which determines an optimal model, and reported the results based on the optimal model. For the proposed spike-and-slab lasso GLMs approach, we fixed slab scale as $s_1 = 1$, and ran a grid value of spike scales: $s_0 = \{0.01 \times k; k = 1, \ldots, 7\}$ to select an optimal model.

3.1. Scenario 1: non-overlap group

In this scenario, there was no overlap among groups. Eight non-zero predictors were placed in three groups. The group sizes and overlap structure are presented below:

Group ID:	1	2	3	...	19	20
Predictors and groupsizes:	x_1–x_{50}	x_{51}–x_{100}	x_{101}–x_{150}		x_{901}–x_{950}	x_{951}–x_{1000}

Table 2 summarizes the deviance, MSE, AUC, and misclassification in the test data. With the deviance as a general measure, we can see that the group spike-and-slab lasso GLMs performed similarly to cMCP, but better than the other methods. The group spike-and-slab lasso GLMs had AUC value similar to that of cMCP, higher than those from the other methods. With the measures, MSE and misclassification, we also observed similar results that the group spike-and-slab lasso GLMs matched the performance of cMCP and outperformed other methods.

Supplementary Figure S1 shows the estimates of coefficients from the group spike-and-slab lasso GLMs and the other methods over 100 replicates. It can be seen that the group spike-and-slab lasso GLMs and cMCP produced estimates close to the simulated values for all the coefficients. This is expected, because the spike-and-slab prior can induce weak shrinkage on larger coefficients and strong shrinkage on zero coefficients. In contrast, other methods gave a strong shrinkage amount on all the coefficients and resulted in the solutions that non-zero coefficients were shrunk and underestimated compared with true values.

3.2. Scenario 2: overlap grouping

In this scenario, we considered overlapped grouping structure, that is, some of the predictors can belong to more than one group. The group sizes and overlap structure are presented below:

Group ID:	1	2	3	...	19	20
Predictors and group size:	x_1–x_{50}	x_{46}–x_{100}	x_{96}–x_{150}		x_{896}–x_{950}	x_{951}–x_{1000}

Table 2. Estimates of four measures over 100 replicates under simulation Scenarios 1 and 2

	Deviance	MSE	AUC	Misclassification
Scenario 1				
gsslasso[a]	470.76(28.04)	0.15(0.01)	0.85(0.02)	0.23(0.02)
lasso	521.40(22.43)	0.17(0.01)	0.82(0.02)	0.26(0.02)
grMCP	575.49(20.32)	0.19(0.01)	0.76(0.02)	0.30(0.03)
grSCAD	583.25(20.43)	0.19(0.01)	0.76(0.02)	0.30(0.02)
cMCP	468.69(31.13)	0.15(0.01)	0.85(0.02)	0.23(0.02)
SGL	538.15(27.01)	0.25(0.02)	0.77(0.02)	0.26(0.02)
Scenario 2				
gsslasso[a]	439.46(33.91)	0.14(0.01)	0.87(0.02)	0.21(0.02)
lasso	487.42(26.65)	0.16(0.01)	0.84(0.02)	0.24(0.02)
grMCP	549.41(20.94)	0.18(0.01)	0.80(0.02)	0.27(0.03)
grSCAD	526.93(22.25)	0.17(0.01)	0.83(0.02)	0.25(0.02)
cMCP	454.56(40.80)	0.15(0.01)	0.87(0.03)	0.22(0.03)

Note: Values in the parentheses are SDs. 'gsslasso' represents the proposed group spike-and-slab lasso GLMs. The slab scales, s_1, are 1 in the analyses.

[a]The optimal $s_0 = 0.05$ for gsslasso method under both Scenarios 1 and 2.

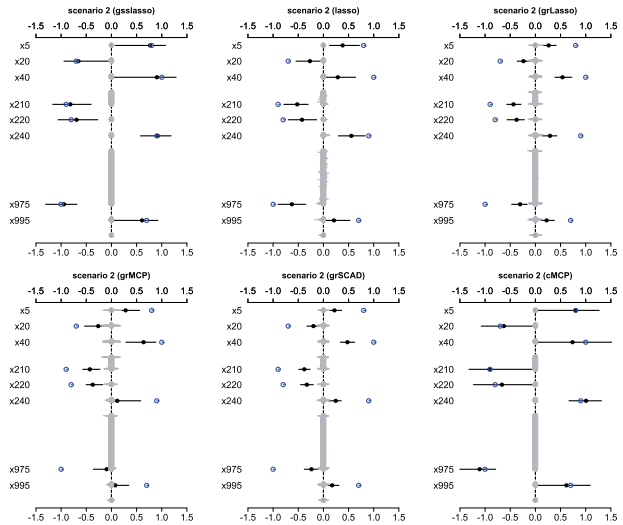

Fig. 1. The parameter estimation averaged over 100 replicates for the group spike-and-slab lasso GLMs (gsslasso), the lasso, grlasso, grMCP, grSCAD and cMCP methods for Scenario 2. Cycles denote the simulated non-zero values. Black points and lines represent the estimated values and the interval estimates of coefficients, respectively

For all the scenarios with overlaps, SGL method was not used for comparison since it cannot handle overlap situation directly. Table 2 summarizes the results from different methods in these scenarios. We can see that for all four measures for the group spike-and-slab lasso GLMs showed better performance than all the other methods. Figure 1 shows the estimates of coefficients from the group spike-and-slab lasso GLMs and the other methods over 100 replicates. It can be seen that the group spike-and-slab lasso GLMs slightly outperformed cMCP and significantly outperformed the other methods. This result suggests that, with complex overlap among groups, the proposed method could still perform well.

Similar to the lasso, the group spike-and-slab lasso GLMs is a path-following strategy for fast dynamic posterior exploration. To fully investigate the impact of the spike scale s_0 on the parameter estimation, we varied the spike scale s_0 over the grid of values: $\{0.001,\ 0.005 \times k; k = 1, \ldots,\ 39\}$, leading to 40 models.

We estimated the coefficients averaged over 100 replicates, to show the characteristics of the group spike-and-slab lasso GLMs. Supplementary Figure S2 presents the solution path for Scenario 2 by the proposed method and the lasso. The solution path of the proposed method is essentially different from that of the lasso model. For the lasso solution, non-zero coefficients can be over-shrunk. However, a spike-and-slab mixture prior has self-adaptive and flexible characteristics, and can help the larger coefficients escape the gravitational pull of the spike.

3.3. Scenario 3: varying group sizes

In this scenario, we assumed that non-zero predictors, $\{x_1, x_2, x_3, x_4\}$ and $\{x_{501}, x_{502}, x_{503}, x_{504}\}$, belong to two groups. To investigate the group size effect on modeling, based on Scenario 2, we simulated the group size and overlap structures as below:

(i) only four non-zero predictors included in a group:

Group ID:	1	2	3	...	11	12	...	19	20
Predictors and group size:	x_1 | x_4	x_5 | x_{100}	x_{96} | x_{150}		x_{501} | x_{504}	x_{505} | x_{600}		x_{896} | x_{950}	x_{951} | x_{1000}

(ii) 20 predictors included in a group:

Group ID:	1	2	3	...	11	12	...	19	20
Predictors and group size	x_1 | x_{20}	x_{21} | x_{100}	x_{96} | x_{150}		x_{501} | x_{520}	x_{521} | x_{600}		x_{896} | x_{950}	x_{951} | x_{1000}

(iii) 50 predictors included in a group:

Group ID:	1	2	3	...	11	12	...	19	20
Predictors and group size	x_1 | x_{50}	x_{46} | x_{100}	x_{96} | x_{150}		x_{501} | x_{550}	x_{546} | x_{600}		x_{896} | x_{950}	x_{951} | x_{1000}

Supplementary Table S1 summarizes the four measures over 100 replicates for all the methods. We can see that the group spike-and-slab lasso GLMs performed slightly better than cMCP, and significantly better than the other methods, especially when the group size increased to 50 predictors. Supplementary Figures S3 and S4 present the estimates of coefficients for all the methods. We can find that the proposed method and cMCP method always produced accurate estimations under varying group sizes. However, the other methods generated shrinkage on the non-zero coefficients and underestimated these larger coefficients.

3.4. Scenario 4: varying the number of non-null group

We varied the number of non-zero groups, to show its effect on modeling:

i. There are eight non-null groups including non-zero coefficients: $\{x_5\}$, $\{x_{55}\}$, $\{x_{305}\}$, $\{x_{355}\}$, $\{x_{505}\}$, $\{x_{555}\}$, $\{x_{905}\}$, and $\{x_{955}\}$;
ii. There are three non-null groups including non-zero coefficients: $\{x_5, x_{15}, x_{25}\}$, $\{x_{355}, x_{365}, x_{375}\}$, and $\{x_{905}, x_{915}\}$;
iii. There is 1 non-*null* group including non-zero coefficients: $\{x_5, x_{10}, x_{15}, x_{20}, x_{25}, x_{30}, x_{35}, x_{40}\}$. Other groups were the same as Scenario 2. The effect sizes of these non-zero coefficients are summarized in Table 1.

Supplementary Table S2 summarizes the four measures over 100 replicates for all the methods. The performance of the group spike-and-slab lasso GLMs and cMCP were similar, and both better than lasso, grlasso, grMCP and grSCAD. When the number of non-null groups was reduced, these methods tended to perform similar with lasso. Supplementary Figures S5 and S6 show the estimates of coefficients for all the methods. Similar to the conclusion of Scenario 3, the proposed method and cMCP method always generated stable and accurate estimations under varying number of non-null groups. Altough for grlasso, grMCP, and grSCAD methods, the increased number of non-null groups usually introduced stronger shrinkage amount and much more noise.

3.5. Scenario 5: varying the correlation within group

In Scenario 2, the correlation coefficient within group was set as 0.6. In this scenario, we set different correlation coefficients within a group: $r = 0.0$, 0.3, 0.5, and 0.7. Others were the same as in Scenario 2. Supplementary Table S3 summarizes the four measures over 100 replicates. We observe that the performances of the group spike-and-slab lasso GLMs and cMCP were consistently better than the other methods, and when the correlation was high, the proposed method was slightly better than cMCP. The comparisons for the estimates of coefficients were similar to those of other scenarios.

4. Applications to real data

We applied our method to analyze three real datasets, sarcoma, ovarian cancer and breast cancer downloaded from The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) (updated at March 2016). To construct the pathways, we used genome annotation tools to map genes to pathways. We first mapped gene symbols to Entrez ids with *R/bioconductor package AnnotationDbi and mygene*, and then mapped all the genes to KEGG pathways using the R/bioconductor package clusterProfiler (Yu *et al.*, 2012). The details of the three datasets are described below.

4.1 TCGA sarcoma dataset (mRNA-sequencing data)

The first dataset includes clinical information and RNAseq expression on sarcoma extracted from TCGA. First, we combined the clinical files with new tumor event records. Clinical data is available for 261 patients. Secondly, the expression data included 259 patients for 20 502 genes with gene names after removing duplicated patients from raw data with 265 samples. We filtered the genes with expressions less or equal to 10 as they showed almost no expression. Furthermore, genes with >50% of zero expression values in the samples were removed. We calculated the variance of expression for each gene, and kept the genes with variance of >20% quantile. We then merged the standardized expression data with new tumor event outcome, and obtained 218 patients with 13 837 genes expression profiles for final analysis. We mapped these genes to 249 pathways including 4919 genes.

4.2 TCGA ovarian cancer dataset (microarray mRNA data)

The second dataset included microarray mRNA expression data and relevant clinical outcome for ovarian cancer from TCGA. Microarray mRNA expression data (Agilent Technologies platform) included 551 tumors with 17 785 genes profiled after removing duplications. We selected the top 80% genes filtered by variance for further analysis. We merged the individuals with gene expression data and those with new tumor events. After removing individuals with missing response, 362 tumors with 14 227 genes were included in our analysis. We mapped these genes to 276 pathways including 4764 genes.

4.3 TCGA breast cancer dataset (mRNA-sequencing data)

The raw data contains 1220 patients and 20 530 genes. After removing the duplication and unknown gene names, there are 1097 patients with 20 502 used for further quality control. Similar with the steps for sarcoma dataset, we filtered the genes with expressions less or equal to 10. Then, genes with >50% of zero expression values in the samples were removed. We calculated the variance of expression for each gene, and kept the genes with variance of >20% quantile. Furthermore, we cleaned the clinical data with clear records about new tumor event, and included 962 patients in our analysis. We then merged the clinical data and expression data, and obtained 960 patients with 14 068 genes. We mapped these genes to 266 pathways including 5026 genes.

The detailed information of genes shared by different pathways are listed in Supplementary Materials S1–S3. Supplementary Table S4 and Supplementary Figure S7 show the distribution of number of genes shared by different number of pathway.

On these three cancer datasets, we aimed to build a logistic model for predicting new tumor event by integrating gene expression data and pathway information. Prior to fitting the models, we standardized all the predictors. For hierarchical (and also penalized) models, it is important to use a roughly common scale for all predictors. For both datasets, we fixed the slab scale s_1 to 1, and varied the spike scale s_0 over the grid of values: $\{0.001, k \times 0.01; k = 1, \ldots, 10\}$, leading to 11 models. We performed 10-fold cross-validation for 10 times, and selected an optimal spike scale s_0 based on the minimum deviance. We firstly performed analysis using the genes included in the pathways. Second, we set the rest genes not included in the pathways as a new group, and then performed analysis again. For comparison, we also analyzed the data using the five existing methods described in the simulation studies.

Table 3 summarizes the measures of performance on these three datasets. For all the three datasets, the proposed group spike-and-slab lasso GLMs generally performed better than the other methods based on the four measures. Supplementary Figure S8 shows the genes detected by the proposed method by using the genes included in the pathways. The results of pathway enrichment analyses were summarized in Supplementary Materials S4–S6. For sarcoma, the detected genes are mainly associated ATP associated genes, similar to several previous studies (Buondonno *et al.*, 2016; Slotkin *et al.*, 2015). For ovarian cancer, the detected genes spread over a wide range of pathways. For breast cancer, there were two pathway involved, according to the detected five genes. Many of these detected genes for the three datasets have not been validated yet in literatures. In addition, we noticed that the genes detected by the existing methods are different. Most of detected genes are not overlapped among these methods, which may be due to the model assumption and the complexity of real data.

We further estimated the pre-validated linear predictor, $\eta_i = X_i\widehat{\beta}$, for each patient, and then grouped the patients on the basis of the pre-validated linear predictor into categorical factor according to 5th, 25th, 50th, 75th and 95th percentiles, denoted by $c_i = (c_{i1}, \ldots, c_{i6})$. We fitted the univariate model $E(y_i \mid \widehat{\eta}_i) = h^{-1}(\mu + \widehat{\eta}_i b)$ and the multivariate model $E(y_i \mid c_i) = h^{-1}(\mu + \sum_{k=2}^{6} c_{ik} b_k)$ by using the pre-validated linear predictor and the categorical factors, respectively. The results are summarized in Supplementary Table S5 for both datasets. Here, we only used the genes included in the pathways and excluded the genes not included in the pathways. As expected, the two models for the datasets showed significant results, indicating that the resulting prediction models were informative.

5. Discussion

The group structure of variables arises naturally in many real statistical modeling problems. Such group structure can be incorporated into a model to take advantage of prior knowledge that is theoretically meaningful and intrinsically encoded in the underlying data. If the group structure is present yet ignored by using models taking into account of solely individual predictors, such models may be inefficient or even inappropriate, leading to low accuracy of genomic prediction. In the article, we have developed a novel hierarchical modeling approach to integrate the variable group information for gene detection and prognostic prediction. The proposed group spike-and-slab lasso GLMs are capable of analyzing large-scale data using various GLMs with group structure, although we focus on modeling molecular profiling data and binary outcome in this study.

The key to our group spike-and-slab lasso GLMs is the introduction of a new prior distribution, i.e. the mixture spike-and-slab double-exponential prior, on the coefficients of each group. The mixture spike-and-slab prior improves the accuracy of coefficient estimation and prognostic prediction by adaptively inducing different amounts of shrinkage for different predictors and thus achieving nice effect of removing irrelevant predictors while supporting the larger coefficients. Similar to other Bayesian approaches, most spike-and-slab variable selection approaches proposed previously use the mixture normal priors on coefficients and employ Markov Chain Monte Carlo (MCMC) algorithms to fit the model (Lu *et al.*, 2015; Partovi Nia and Ghannad-Rezaie, 2016; Shankar *et al.*, 2015; Shelton *et al.*, 2015). However, these MCMC methods are computationally intensive for analyzing large-scale and high-dimensional genetic data. Instead, we develop an efficient EM coordinate descent algorithm to fit the proposed model, which incorporates EM steps into the fast cyclic coordinate descent algorithm. The E-steps involve calculating the posterior expectations of the indicator variable γ_j and the scale S_j for each coefficient of each group, and the M-steps employ a fast algorithm, i.e. the cyclic coordinate descent algorithm (Friedman *et al.*, 2010a,b; Hastie *et al.*, 2015; Simon *et al.*, 2011), to update the coefficients group by group. The resulted EM coordinate descent algorithm converges rapidly, and is capable of identifying important predictors and building promising predictive models from a large number of candidates organized into various groups.

The group spike-and-slab lasso proposed in this study maintains the advantages of two popular methods for high-dimensional data analysis (Ročková and George, 2016a), i.e. Bayesian variable selection (Chipman, 1996; Chipman *et al.*, 2001; George and McCulloch, 1993, 1997; Ročková and George, 2014) and the

Table 3. The measures of optimal group spike-and-slab lasso (gsslasso) and the five penaliztion models for TCGA sarcoma, ovarian cancer and breast cancer dataset by 10 times 10-fold cross validation

No. of pathway/gene	Methods	Deviance	AUC	MSE	Misclassification	No. non-zero gene
TCGA sarcoma ($n = 218$)						
249/4919[a]	gsslasso[b]	257.72 (3.08)	0.69 (0.01)	0.20 (0.00)	0.30 (0.02)	14
	lasso	265.60 (8.76)	0.66 (0.03)	0.21 (0.01)	0.32 (0.01)	52
	grlasso	246.31 (6.23)	0.73 (0.02)	0.19 (0.01)	0.29 (0.02)	13
	grMCP	260.13 (4.82)	0.67 (0.02)	0.21 (0.01)	0.33 (0.01)	31
	grSCAD	248.86 (5.54)	0.72 (0.02)	0.20 (0.01)	0.29 (0.02)	422
	cMCP	267.21 (4.83)	0.62 (0.02)	0.21 (0.00)	0.32 (0.01)	15
TCGA ovarian cancer ($n = 362$)						
276/4764	gsslasso[b]	434.17 (4.01)	0.67 (0.02)	0.20 (0.00)	0.29 (0.01)	116
	lasso	442.62 (3.69)	0.64 (0.01)	0.21 (0.00)	0.31 (0.02)	48
	grlasso	465.03 (3.04)	0.55 (0.03)	0.22 (0.00)	0.33 (0.01)	10
	grMCP	461.39 (0.49)	0.51 (0.02)	0.22 (0.00)	0.33 (0.00)	19
	grSCAD	460.78 (0.50)	0.53 (0.01)	0.22 (0.00)	0.33 (0.00)	919
	cMCP	450.71 (7.38)	0.62 (0.02)	0.23 (0.01)	0.32 (0.02)	25
TCGA Breast cancer ($n = 960$)						
266/5026	gsslasso[b]	976.83 (27.82)	0.63 (0.01)	0.16 (0.00)	0.21 (0.00)	5
	lasso	957.89 (5.33)	0.62 (0.01)	0.16 (0.00)	0.21 (0.00)	8
	grlasso	973.87 (1.38)	0.50 (0.00)	0.16 (0.00)	0.20 (0.00)	11
	grMCP	972.95 (4.95)	0.57 (0.01)	0.16 (0.00)	0.21 (0.00)	11
	grSCAD	975.60 (3.61)	0.55 (0.01)	0.16 (0.00)	0.21 (0.00)	11
	cMCP	975.23 (21.55)	0.60 (0.02)	0.16 (0.00)	0.21 (0.00)	2

Note: Values in the parentheses are SDs.

[a]in TCGA sarcoma data, we mapped 4919 genes into 249 pathways. The rest genes were put together as an additional group. The results using all genes are provided as Supplementary Table S6. The same is true for ovarian and breast cancer datasets.

[b]The optimal s_0 for gsslasso are $s_0 = 0.03$, $s_0 = 0.07$ and $s_0 = 0.0005$ for the three datasets, respectively.

penalized lasso (Hastie *et al.*, 2015; Tibshirani, 1996, 1997), and bridges these two methods into one unifying framework. Similar to the lasso, the proposed method can shrink many coefficients exactly to zero, thus automatically achieving sparsity within group for feature selection, and the output is characterized by the solution path. Without the slab component, the output of the proposed model would be equivalent or similar to the lasso solution path. Instead, the solution path of the model with the spike-and-slab prior is different from the lasso. The large coefficients are usually included in the proposed model with weak or none shrinkage, while the lasso may undesirably shrink large coefficients.

The proposed group spike-and-slab lasso GLMs well suit to handle both overlapping group structure and non-overlapping group situations. The proposed group spike-and-slab lasso approach always outperforms other methods on the simulated datasets covering different scenarios with overlapping or non-overlapping group structures. Not surprisingly, the performance of the proposed method depends on the scale parameter of the double-exponential prior. We evaluated the performance of the proposed model on the different combinations of prior scales (s_0, s_1). Our results showed that slab scale s_1 had little influence on the deviance, while the spike scale s_0 strongly affected model performance (Tang *et al.*, 2017a,b). A slab scale s_1 value introducing weak shrinkage amount would be helpful to include relevant variables into the model. Hence, we suggest a path-following strategy for fast dynamic posterior exploration of the proposed models, which is similar to the approach of Ročková and George (2014, 2016a). Usually, a path-following strategy can be implemented by first fixing the slab scale s_1 (e.g. $s_1 = 1$), running a grid of values of spike scale s_0 from a reasonable range, e.g. (0, 0.1), and then selecting an optimal according to cross-validation. The fast speed of the proposed algorithm makes it feasible to consider several or 10 of reasonable values for selecting an optimal s_0. To evaluate this strategy, we compared the performance of the proposed method and several other methods under different simulation settings. The prediction performance of the proposed method is always slightly better than cMCP method, and significantly better than all other methods, especially when the variables within group are highly correlated as often observed in genetic data. In addition, for most of real data, like the sarcoma and ovarian cancer datasets, the sample size and effect size might be small. To evaluate the performance of the proposed method under these conditions, we further performed simulation study based on Scenario 2 with only half sample size ($n = 250$) and half effect size for eight non-zero predictors. The results are summarized in Supplementary Table S7. It could be found that the proposed could still perform slightly better than other methods.

Due to the complexity of group structures in real datasets, as expected, the prediction accuracy of the proposed method incorporating pathway information was improved on the TCGA datasets. In addition to pathways used in this study, more sophisticated grouping strategies could potentially enhance prediction accuracy and further improve the models (Ogutu and Piepho, 2014). Such complex grouping strategies could result from a single or mixture of various sources reflecting the underlying biological structure including haplotype blocks or genetic regions covering nearby genetic markers, subnetworks, communities, clusters or modules of many types of biological networks (e.g. regulatory networks, signaling networks, protein–protein interaction networks, metabolic networks). Therefore, we expect that the proposed models will be further improved by more strategically designed grouping strategy that captures the complicated grouping structure in real genetic data.

The method we present here has attractive features which point to several further extensions. For example, the proposed group spike-and-slab lasso GLMs can be extended to Cox proportional hazards model for censored survival data, truncated regressions for extreme phenotyping designs, ordered logistic or probit regressions

for ordinal response (e.g. disease severity), and conditional logistic regression for matched case-control studies. These models can also be extended to incorporate multiple level group structure, such as multiple taxonomy levels in microbiome data (i.e. phylum, class, order, family, genus and species), which is difficult for SGL or many other methods. Additionally, incorporating multiple level group structures, like three-level group structure, i.e. SNP-gene-pathway, might also be an interesting topic. Besides the spike-and-slab mixture double-exponential prior used in the proposed models, we should investigate theoretical and empirical properties of other priors. An important example is Cauchy distribution, a special case of Student-*t* distribution. Cauchy distribution has a spike at zero and includes heavier tails, and thus can be included as an appropriate prior to handle high-dimensional data.

Acknowledgements

We thank the reviewers and the associate editor for their constructive suggestions and comments that have improved the manuscript. We acknowledge the contributions of the TCGA Research Network.

Funding

This work was supported in part by research grants from USA National Institutes of Health (R03-DE024198, R03-DE025646) and National Science Foundation (IIS-1502172), grants from China Scholarship Council, the National Natural Science Foundation of China (81573253, 81773541, and 81673448), and funds from the Priority Academic Program Development of Jiangsu Higher Education Institutions at Soochow University, Natural Science Foundation of Jiangsu Province China (BK 20161218).

Conflict of Interest: none declared.

References

Barillot,E. *et al.* (2012) *Computational Systems Biology of Cancer.* Chapman & Hall/CRC Mathematical and Computational Biology CRC Press, Boca Raton, FL, USA.

Breheny,P. (2015) The group exponential lasso for bi-level variable selection. *Biometrics*, **71**, 731–740.

Breheny,P. and Huang,J. (2009) Penalized methods for bi-level variable selection. *Stat. Interf.*, **2**, 369–380.

Buondonno,I. *et al.* (2016) Mitochondria-targeted doxorubicin: a new therapeutic strategy against doxorubicin-resistant osteosarcoma. *Mol. Cancer Ther.*, **15**, 2640–2652.

Chen,C. *et al.* (2014a) O(1) Algorithms for Overlapping Group Sparsity. *2014 22nd International Conference on Pattern Recognition*. Institute of Electrical and Electronics Engineers Inc., pp. 1645–1650.

Chen,Y. *et al.* (2014b) Variable selection in linear models. *Wiley Interdiscip. Rev. Comput. Stat.*, **6**, 1–9.

Chipman,H. (1996) Bayesian variable selection with related predictions. *Can. J. Stat.*, **24**, 17–36.

Chipman,H. *et al.* (2001) The Practical Implementation of Bayesian Model Selection. *Lecture Notes-Monograph Series.*, **38**, 65–134.

de los Campos,G. *et al.* (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat. Rev. Genet.*, **11**, 880–886.

Fan,J. and Li,R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, **96**, 1348–1360.

Friedman,J. *et al.* (2010a) A note on the group lasso and a sparse group lasso, *Technical report*, Department of Statistics, Stanford University.

Friedman,J. *et al.* (2010b) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

Gelman,A. *et al.* (2014) *Bayesian Data Analysis*. New York: Chapman & Hall/CRC Press.

Gelman,A. and Hill,J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Gene Ontology,C. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.

George,E.I. and McCulloch,R.E. (1993) Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.*, **88**, 881–889.

George,E.I. and McCulloch,R.E. (1997) Approaches for Bayesian variable selection. *Stat. Sin.*, **7**, 339–373.

Hastie,T. *et al.* (2009) *The Elements of Statistical Learning*. New York, NY: Springer-Verlag.

Hastie,T. *et al.* (2015) *Statistical Learning with Sparsity - the Lasso and Generalization*. CRC Press, New York.

Huang,J. *et al.* (2012) A Selective review of group selection in high-dimensional models. *Stat. Sci.*, **27**, 481–499.

Huang,J. *et al.* (2009) A group bridge approach for variable selection. *Biometrika*, **96**, 339–355.

Ishwaran,H. and Rao,J.S. (2005) Spike and slab gene selection for multigroup microarray data. *J. Am. Stat. Assoc.*, **100**, 764–780.

Jacob,L. *et al.* (2009) Group lasso with overlap and graph lasso. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, Montreal, Quebec, Canada, pp. 433–440.

Kanehisa,M. *et al.* (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*, **44**, D457–D462.

Kwon,S. *et al.* (2016) A doubly sparse approach for group variable selection. *Ann. Inst. Stat. Math.*, **69**, 1–29.

Lu,Z.H. *et al.* (2015) Multiple SNP set analysis for genome-wide association studies through Bayesian latent variable selection. *Genet. Epidemiol.*, **39**, 664–677.

McCullagh,P. and Nelder,J.A. (1989) *Generalized Linear Models*. London: Chapman and Hall.

Meier,L. *et al.* (2008) The group lasso for logistic regression. *J. Royal Stat. Soc. Ser. B*, **70**, 53–71.

Obozinski,G. *et al.* (2011) Group lasso with overlaps: the latent group lasso approach. *Research report*, arXiv preprint arXiv:1110.0413, October 2011.

Ogutu,J.O. and Piepho,H.P. (2014) Regularized group regression methods for genomic prediction: Bridge, MCP, SCAD, group bridge, group lasso, sparse group lasso, group MCP and group SCAD. *BMC Proc.*, **8**, S7.

Partovi Nia,V. and Ghannad-Rezaie,M. (2016) Agglomerative joint clustering of metabolic data with spike at zero: A Bayesian perspective. *Biom. J.*, **58**, 387–396.

Rapaport,F. *et al.* (2007) Classification of microarray data using gene networks. *BMC Bioinformatics*, **8**, 1–15.

Ročková,V. and George,E.I. (2014) EMVS: the EM approach to Bayesian variable selection. *J. Am. Stat. Assoc.*, **109**, 828–846.

Ročková,V. and George,E.I. (2016a) The Spike-and-Slab LASSO, *J. Am. Stat. Assoc.*, doi: 10.1080/01621459.2016.1260469.

Ročková,V. and George,E.I. (2016b) Bayesian penalty mixing: the case of a non-separable penalty. In: Frigessi, A. *et al.* (eds.) *Statistical Analysis for High-Dimensional Data: The Abel Symposium 2014*. New York: Springer International Publishing, Cham, pp. 233–254.

Shankar,J. *et al.* (2015) A systematic evaluation of high-dimensional, ensemble-based regression for exploring large model spaces in microbiome analyses. *BMC Bioinformatics*, **16**, 31.

Shelton,J.A. *et al.* (2015) Nonlinear spike-and-slab sparse coding for interpretable image encoding. *PLoS One*, **10**, e0124088.

Silver,M. *et al.* (2013) Pathways-driven sparse regression identifies pathways and genes associated with high-density lipoprotein cholesterol in two Asian cohorts. *PLoS Genet.*, **9**, e1003939.

Silver,M. *et al.* (2012) Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. *Stat. Appl. Genet. Mol. Biol.*, **11**, Article 7.

Simon,N. *et al.* (2011) Regularization paths for cox's proportional hazards model via coordinate descent. *J. Stat. Softw.*, **39**, 1–13.

Simon,N. *et al.* (2013) A sparse-group Lasso. *J. Comput. Graph. Stat.*, **22**, 231–245.

Slotkin,E.K. *et al.* (2015) MLN0128, an ATP-competitive mTOR kinase inhibitor with potent in vitro and in vivo antitumor activity, as potential therapy for bone and soft-tissue sarcoma. *Mol. Cancer Ther.*, **14**, 395–406.

Sohn,I. *et al.* (2013) Predictive modeling using a somatic mutational profile in ovarian high grade serous carcinoma. *PLoS One*, **8**, e54089.

Steyerberg,E.W. (2009) *Clinical Prediction Models: A Practical Approch to Development, Validation, and Updates*. New York: Springer.

Tang,Z. *et al*. (2017a) The spike-and-slab lasso cox model for survival prediction and associated genes detection. *Bioinformatics*, **33**, 2799–2807.

Tang,Z. *et al*. (2017b) The spike-and-slab lasso generalized linear models for prediction and associated genes detection. *Genetics*, **205**, 77–88.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. Ser. B*, **58**, 267–288.

Tibshirani,R. (1997) The lasso method for variable selection in the Cox model. *Stat Med*, **16**, 385–395.

Tibshirani,R.J. and Efron,B. (2002) Pre-validation and inference in microarrays. *Stat. Appl. Genet. Mol. Biol*., **1**, 1–18.

Yi,N. *et al*. (2003) Stochastic search variable selection for mapping multiple quantitative trait loci. *Genetics*, **165**, 867–883.

Yu,G. *et al*. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics*, **16**, 284–287.

Yuan,L. *et al*. (2013) Efficient methods for overlapping group lasso. *IEEE Trans. Pattern Anal. Mach. Intell*., **35**, 2104–2116.

Yuan,M. and Lin,Y. (2006) Model selection and estimation in regression with grouped variables. *J. Royal Stat. Soc. Ser. B*, **68**, 49–67.

Yuan,Y. *et al*. (2014) Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat. Biotechnol*., **32**, 644–652.

Zeng,Y. and Breheny,P. (2016) Overlapping group logistic regression with applications to genetic pathway selection. *Cancer Informatics*, **15**, 179–187.

Zhang,C.H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat*., **38**, 894–942.

Zhang,C. (2007) Penalized linear unbiased selection. *Technical Report*, Department of Statistics and Bioinformatics, Rutgers University, 2007-2003.

Zhang,W. *et al*. (2013) Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. *PLoS Comput. Biol*., **9**, e1002975.

Zhao,P. *et al*. (2009) The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Stat*., **37**, 3468–3497.

Zhao,Q. *et al*. (2015) Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. *Brief Bioinform*., **16**, 291–303.

Zhou,X. *et al*. (2013) Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet*., **9**, e1003264.