Tang et al. BMC Bioinformatics (2019) 20:94
https://doi.org/10.1186/s12859-019-2656-1

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Gsslasso Cox: a Bayesian hierarchical model for predicting survival and detecting associated genes by incorporating pathway information

Zaixiang Tang[1,2,5], Shufeng Lei[1,2], Xinyan Zhang[3], Zixuan Yi[4], Boyi Guo[5], Jake Y. Chen[6], Yueping Shen[1,2*] and Nengjun Yi[5*]

Abstract

Background: Group structures among genes encoded in functional relationships or biological pathways are valuable and unique features in large-scale molecular data for survival analysis. However, most of previous approaches for molecular data analysis ignore such group structures. It is desirable to develop powerful analytic methods for incorporating valuable pathway information for predicting disease survival outcomes and detecting associated genes.

Results: We here propose a Bayesian hierarchical Cox survival model, called the group spike-and-slab lasso Cox (gsslasso Cox), for predicting disease survival outcomes and detecting associated genes by incorporating group structures of biological pathways. Our hierarchical model employs a novel prior on the coefficients of genes, i.e., the group spike-and-slab double-exponential distribution, to integrate group structures and to adaptively shrink the effects of genes. We have developed a fast and stable deterministic algorithm to fit the proposed models. We performed extensive simulation studies to assess the model fitting properties and the prognostic performance of the proposed method, and also applied our method to analyze three cancer data sets.

Conclusions: Both the theoretical and empirical studies show that the proposed method can induce weaker shrinkage on predictors in an active pathway, thereby incorporating the biological similarity of genes within a same pathway into the hierarchical modeling. Compared with several existing methods, the proposed method can more accurately estimate gene effects and can better predict survival outcomes. For the three cancer data sets, the results show that the proposed method generates more powerful models for survival prediction and detecting associated genes. The method has been implemented in a freely available R package BhGLM at https://github.com/nyiuab/BhGLM.

Keywords: Cox survival models, Grouped predictors, Hierarchical modeling, Lasso, Pathway, Spike-and-slab prior

Background

Survival prediction from high-dimensional molecular data is an active topic in the fields of genomics and precision medicine, especially for various cancer studies. Large-scale omics data provide extraordinary opportunities for detecting biomarkers and building accurate prognostic and predictive models. However, such high-dimensional data also introduce statistical and computational challenges. Tibshirani [1, 2] has proposed a novel penalized method, lasso, for variable selection in high-dimensional data, which has attracted considerable attention in modern statistical research. Thereafter, several penalized methods were developed, like minimax concave penalty (MCP) method by Zhang [3, 4], smoothly clipped absolute deviation (SCAD) penalty method by Fan and Li [5]. These penalization approaches have been widely

* Correspondence: shenyueping@suda.edu.cn; nyi@uab.edu
[1]Department of Biostatistics, School of Public Health, Medical College of Soochow University, University of Alabama at Birmingham, Suzhou 215123, China
[5]Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294-0022, USA
Full list of author information is available at the end of the article

applied for disease prediction and prognosis using large-scale molecular data [6–11].

Furthermore, the group structures among molecular variables was noticed in analysis. For example, genes can be grouped into known biological pathways or functionally similar sets. Genes within a same biological pathway may be biologically related and statistically correlated. Incorporating such biological grouping information into statistical modeling can improve the interpretability and efficiency of the models. Several penalization methods have been proposed had been proposed to utilize the grouping information, such as group Lasso method [12], sparse group lasso (SGL) [13, 14]. group bridge [15], composite MCP [16], composite absolute penalty method [17], group exponential Lasso [18], group variable selection via convex log-exp-sum penalty method [19], and doubly sparse approach for group variable selection [20]. Some of these methods perform group level selection, including or excluding an entire group of variables. Others can perform bi-level selection, achieving sparsity within each group. Huang et al. [21] and Ogutu et al. [22] reviewed these penalization methods in prediction and highlighted some issues for further study.

Ročková and George [23, 24] recently 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. Based on the Bayesian framework, we have recently incorporated the spike-and-slab mixture double-exponential prior into generalized linear models (GLMs) and Cox survival models, and developed the spike-and-slab lasso GLMs and Cox models for predicting disease outcomes and detecting associated genes [25, 26]. More recently, we have developed the group spike-and-slab lasso GLMs [27] to incorporate biological pathways.

In this article, we aim to develop the group spike-and-slab lasso Cox model (gsslasso Cox) for predicting disease survival outcomes and detecting associated genes by incorporating biological pathway information. An efficient algorithm was proposed to fit the group spike-and-slab lasso Cox model by integrating Expectation-Maximization (EM) steps into the extremely fast cyclic coordinate descent algorithm. The novelty is incorporating group or biological pathway information into the spike-and-slab lasso Cox model for predicting disease survival outcomes and detecting associated genes. The performance of the proposed method was evaluated via extensive simulations and comparing with several commonly used methods. The proposed procedure was also applied to three cancer data sets with thousands of gene expression values and their pathways information. Our results show that the proposed method not only generates powerful prognostic models for survival prediction, but also excels at detecting associated genes.

Methods

The group spike-and-slab lasso Cox models

In Cox survival model, variables $y_i = (t_i,\ d_i)$ for each individual is the survival outcome. The censoring indicator d_i takes 1 if the observed survival time t_i for individual i is uncensored. The d_i takes 0 if it is censored. For individual i, the true survival time is assumed by T_i. Therefore, when $T_i = t_i,\ d_i = 1$, whereas when $T_i > t_i,\ d_i = 0$. The predictor variables include numerous molecular predictors (e.g., gene expression) and some relevant demographic/clinical covariates. Assume that the predictors can be organized into G groups (e.g., biological pathways) based on existing biological knowledge. It should be indicated that the group could overlap each other. For example, one or some genes can belong to two or more biological pathway. Following the idea of overlap group lasso [28–31], we performed a restructure step by replicating a variable in whatever group it appears to expand the vector of predictors.

In Cox proportional hazards model, it usually assumes that the hazard function of survival time T takes the form [32, 33]:

$$h(t|X) = h_0(t)\ \exp(X\beta) \tag{1}$$

where the baseline hazard function $h_0(t)$ is unspecified, X and β are the vectors of explanatory variables and coefficients, respectively, and $X\beta$ is the linear predictor or called the prognostic index.

Fitting classical Cox models is to estimate β by maximizing the partial log-likelihood [34]:

$$pl(\beta) = \sum_{i=1}^{n} d_i\ \log\left(\ \exp(X_i\beta) / \sum_{i' \in R(t_i)}\ \exp(X_{i'}\beta)\right) \tag{2}$$

where $R(t_i)$ is the risk set at time t_i. In the presence of ties, the partial log-likelihood can be approximated by the Breslow or the Efron methods [35, 36]. The standard algorithm for maximizing the partial log-likelihood is the Newton-Raphson algorithm [32, 37].

For high dimensional and/or correlated data, the classical model fitting is often unreliable. The problem can be solved by using Bayesian hierarchical modeling or penalization approaches [31, 38, 39]. We here propose a Bayesian hierarchical modeling approach, which allows us to simultaneously analyze numerous predictors and more importantly provides an efficient way to incorporate group information. Our hierarchical Cox models

employ the spike-and-slab mixture double-exponential (de) prior on the coefficients:

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

where, s_0 and s_1 are the preset scale parameters, which are small and relatively large ($0 < s_0 < s_1$), inducing strong or weak shrinkage on β_j, respectively. γ_j is the indicator variable: γ_j = 1 or 0. Equivalently, this prior can be expressed as (1 - γ_j) de(0, s_0) + γ_j de(0, s_1), a mixture of the shrinkage prior de(0, s_0) and the weakly informative prior de(0, s_1), which are the spike and slab components of the prior distribution, respectively.

We incorporate the group structure by proposing a group-specific Berllouli distribution for the indicator variables. For predictors in group g, the indicator variables are assumed to follow the Berllouli distribution with the group-specific probability θ_g:

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

If group g includes important predictors, the parameter θ_g will be estimated to be relatively large, implying other predictors in the group more likely to be important. Therefore, the group-specific Berllouli prior plays a role on incorporating the biological similarity of genes within a same pathway into the hierarchical model. For the probability parameters, we adopt a beta prior, $\theta_g \sim \mathrm{beta}(a, b)$, setting $a = b = 1$ yielding the uniform hyper prior $\theta_g \sim U(0, 1)$ that will be used in later sections to illustrate our method. Hereafter, the above hierarchical Cox models are referred to as the group spike-and-slab lasso Cox model.

The EM coordinate descent algorithm

We have developed a fast deterministic algorithm, called the EM coordinate descent algorithm to fit the spike-and-slab lasso Cox models by estimating the posterior modes of the parameters [26]. The EM coordinate descent algorithm incorporates EM steps into the cyclic coordinate descent procedure for fitting the penalized lasso Cox models, and has been shown to be fast and efficient for analyzing high-dimensional survival data [26]. We here extend the EM coordinate descent algorithm to fit the group spike-and-slab lasso Cox models. We derive the algorithm based on the log joint posterior density of the parameters $\vartheta = (\beta, \gamma, \theta)$:

$$\log p(\beta, \gamma, \theta|t, d) \propto \log p(t, d|\beta, h_0) + \sum_{j=1}^{J} \log p\left(\beta_j|S_j\right) + \sum_{j=1}^{J} \log p\left(\gamma_j|\theta_g\right) + \sum_{g=1}^{G} \log p(\theta_g) \tag{5}$$

The log-likelihood function, $\log p(t, d|\beta, h_0)$, is proportional to the partial log-likelihood $pl(\beta)$ defined in Eq. (2) or the Breslow or the Efron approximation in the presence of ties [35, 36], if the baseline hazard function h_0 is replaced by the Breslow estimator [37, 40]. Therefore, the log joint posterior density can be expressed as

$$\log p(\beta, \gamma, \theta|t, d) \propto pl(\beta) - \sum_{j=1}^{J} S_j^{-1}\left|\beta_j\right| + \sum_{j=1}^{J}\left(\gamma_j \log\theta_g + \left(1-\gamma_j\right)\log\left(1-\theta_g\right)\right) + \sum_{g=1}^{G}\left((a-1)\log\theta_g + (b-1)\log\left(1-\theta_g\right)\right) \tag{6}$$

where $pl(\beta)$ is the partial likelihood described in (2), and $S_j = (1 - \gamma_j)s_0 + \gamma_j s_1$.

In EM coordinate decent algorithm, the indicator variables γ_j were treated the as 'missing values'. The parameters (β, θ) were estimated by averaging the missing values over their posterior distributions. For the E-step, the expectation of the log joint posterior density was calculated 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

$$p_j^g = p\left(\gamma_j = 1|\beta_j, \theta_g, t, d\right) = \frac{p\left(\beta_j|\gamma_j = 1, s_1\right)p\left(\gamma_j = 1|\theta_g\right)}{p\left(\beta_j|\gamma_j = 0, s_0\right)p\left(\gamma_j = 0|\theta_g\right) + p\left(\beta_j|\gamma_j = 1, s_1\right)p\left(\gamma_j = 1|\theta_g\right)} \tag{7}$$

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

$$E\left(S_j^{-1}|\beta_j\right) = E\left(\frac{1}{\left(1-\gamma_j\right)s_0 + \gamma_j s_1}|\beta_j\right) = \frac{1-p_j^g}{s_0} + \frac{p_j^g}{s_1} \tag{8}$$

From Eqs. (7) and (8), we can see that the estimates of p_j and S_j are larger for larger coefficientsβ_j, leading to different shrinkage for different coefficients.

For the M-step, parameters (β, θ) were updated by maximizing the posterior expectation of the log joint posterior density with γ_j and S_j^{-1} replaced by their conditional posterior expectations. From the log joint

posterior density, we can see that β and θ can be updated separately, because the coefficients β are only involved in $pl(\beta)-\sum_{j=1}^{J} S_j^{-1}|\beta_j|$ and the probability parameter θ is only in $\sum_{j=1}^{J}(\gamma_j \log\theta_g + (1-\gamma_j)\log(1-\theta_g)) + \sum_{g=1}^{G}((a-1)\log\theta_g + (b-1)\log(1-\theta_g))$. Therefore, the coefficients β are updated by maximizing the expression:

$$Q_1(\beta) = pl(\beta)-\sum_{j=1}^{J}\hat{S}_j^{-1}\left|\beta_j\right| \tag{9}$$

where $\hat{S}_j^{-1}$ is the conditional posterior expectation of S_j^{-1} as derived above. Given the scale parameters S_j, the term $\sum_{j=1}^{J}\hat{S}_j^{-1}|\beta_j|$ serves as the L_1 lasso penalty with $\hat{S}_j^{-1}$ as the penalty factors, and thus the coefficients can be updated by maximizing $Q_1(\beta)$ using the cyclic coordinate decent algorithm, which is extremely fast and can estimate some coefficients exactly to zero [31, 41]. The probability parameters $\{\theta_g\}$ are updated by maximizing the expression:

$$Q_2(\theta) = \sum_{j=1}^{J}\left[p_j^g \log\theta_g + \left(1-p_j^g\right)\log\left(1-\theta_g\right)\right] + \sum_{g=1}^{G}\left((a-1)\log\theta_g + (b-1)\log\left(1-\theta_g\right)\right) \tag{10}$$

We can easily obtain:

$$\theta_g = \frac{\sum_{j\in g} p_j^g + a - 1}{J_g + a + b - 2} \tag{11}$$

where J_g is the number of predictors belonging to group g.

Totally, the framework of the proposed EM coordinate decent algorithm was summarized as follows:

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

E-step: Update γ_j and S_j^{-1} by their conditional posterior expectations.

M-step:

a) Update β using the cyclic coordinate decent algorithm;
b) Update $(\theta_1, \quad , \theta_G)$ by Eq. (11).

We assess convergence by the criterion: $|d^{(t)} - d^{(t-1)}| / (0.1 - |d^{(t)}|) < \varepsilon$, where $d^{(t)} = -2pl(\beta^{(t)})$ is the estimate of deviance at the t^{th} iteration, and ε is a small value (say 10^{-5}).

Evaluation of predictive performance

We can use several ways to measure the performance of a fitted group lasso Cox model, including the partial log-likelihood (PL), the concordance index (C-index), the survival curves, and the survival prediction error [37]. The partial log-likelihood function measures the overall quality of a fitted Cox model, and thus is usually used to choose an optimal model [37, 41, 42]. The standard way to evaluate the performance of a model is to fit the model using a data set and then calculate the above measures with independent data. A variant of cross-validation [31, 43], called pre-validation method was used in the present study to evaluate the performance. The data was randomly split to K subsets of roughly the same size. The $(K - 1)$ subsets was used to fit a hierarchical Cox model. The estimate of coefficients denoted as $\hat{\beta}^{(-k)}$ from the data excluding the k-th subset. The prognostic indices $\hat{\eta}_{(k)} = X_{(k)}\hat{\beta}^{(-k)}$, called the cross-validated or pre-validated prognostic index, were calculated for all individuals in the k-th subset of the data. Cross-validated prognostic indices $\hat{\eta}_i$ for all individuals can be calculated by cycling through all the K parts. Then, $(t_i, d_i, \hat{\eta}_i)$ was used to compute the several measures described above. We can see that the cross-validated prognostic value for each patient is derived independently of the observed response of the patient. Therefore, the 'pre-validated' dataset $(t_i, d_i, \hat{\eta}_i)$ can essentially be treated as a 'new dataset'. This procedure provides valid assessment of the predictive performance of the model [31, 43].

Moreover, we also use an alternative way to evaluate the partial log-likelihood, i.e., the so-called cross-validated partial likelihood (CVPL), defined as [37, 41, 42].

$$CVPL = \sum_{k=1}^{K}\left[pl\left(\hat{\beta}_{(-k)}\right) - pl_{(-k)}\left(\hat{\beta}_{(-k)}\right)\right] \tag{12}$$

where $\hat{\beta}_{(-k)}$ is the estimate of β from all the data except the k-th part, $pl(\hat{\beta}_{(-k)})$ is the partial likelihood of all the data points and $pl_{(-k)}(\hat{\beta}_{(-k)})$ is the partial likelihood excluding part k of the data. By subtracting the log-partial likelihood evaluated on the non-left out data from that evaluated on the full data, we can make efficient use of the death times of the left out data in relation to the death times of all the data.

Selecting optimal scale values

The spike-and-slab double-exponential prior requires two preset scale parameters (s_0, s_1). Following the previous studies [24–26], we set the slab scale s_1 to be relatively large (e.g., 1), and consider a sequence of L decreasing values $\{s_0^l\}$: $s_1 > s_0^1 > s_0^2 > \cdots > s_0^L > 0$, for

the spike scale s_0. We then fit L models with scales $\{(s_0^l, s_1); l = 1, \cdots, L\}$ and select an optimal model using the method described above. This procedure is similar to the lasso implemented in the widely-used R package glmnet, which quickly fits the lasso Cox models over a grid of values of λ covering its entire range, giving a sequence of models for users to choose from [31, 41].

Implementation and software package

We have incorporated the method proposed in this study into the function bmlasso() in our R package BhGLM [44]. The package BhGLM also includes several other functions for summarizing and evaluating the predictive performance, like summary.bh, cv.bh predict.bh. The function in the package is very fast, usually taking several minutes for fitting and evaluating a model with thousands of variables. The package BhGLM is freely available from https://github.com/nyiuab/BhGLM.

Simulation study and real data analysis

Simulation studies

We assessed the proposed approach by extensive simulations, and compared with the lasso implemented in the R package glmnet and several penalization methods that can incorporate group information, including sparse group lasso (SGL) in the R package SGL, overlap group lasso (grlasso), overlap group MCP (grMCP), overlap group SCAD (grSCAD), and overlap group composite MCP (cMCP) in the R package grpregOverlap [45]. Our simulation method was similar to our previous work [26, 27]. 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 scenario 2–5, overlap structures were considered. To handle the overlap structures, we duplicated overlapping predictors into groups that predictors belong to [28, 30]. In each scenario, we simulated two data sets, and used

Table 1 The preset non-zero predictors and their assumed effect values of the different simulation scenarios

Simulation scenarios	Group, non-zero predictors and effect size							
1 non-overlap group								
Group	group1			group5			group20	
predictors	$\{x_5$	x_{20}	$x_{40}\}$	$\{x_{210}$	x_{220}	$x_{240}\}$	$\{x_{975}$	$x_{995}\}$
2 overlap group								
Group	group1			group5			group20	
predictors	$\{x_5$	x_{20}	$x_{40}\}$	$\{x_{210}$	x_{220}	$x_{240}\}$	$\{x_{975}$	$x_{995}\}$
3 varying group size (4/20/50)								
Group	group1				group11			
predictors	$\{x_1$	x_2	x_3	$x_4\}$	$\{x_{501}$	x_{502}	x_{503}	$x_{504}\}$
4 varying number of non-null groups (8/3/1)								
Group	group1	group2	group7	group8	group11	group12	group19	group20
predictors	$\{x_5\}$	$\{x_{55}\}$	$\{x_{305}\}$	$\{x_{355}\}$	$\{x_{505}\}$	$\{x_{555}\}$	$\{x_{905}\}$	$\{x_{955}\}$
Group	group1			group8			group20	
predictors	$\{x_5$	x_{15}	$x_{25}\}$	$\{x_{355}$	x_{365}	$x_{375}\}$	$\{x_{905}$	$x_{915}\}$
Group	group1							
predictors	$\{x_5$	x_{10}	x_{15}	x_{20}	x_{25}	x_{30}	x_{35}	$x_{40}\}$
5 varying correlation within group ($r = 0.0/0.5/0.7$)								
Group	group1			group5			group20	
predictors	$\{x_5$	x_{20}	$x_{40}\}$	$\{x_{210}$	x_{220}	$x_{240}\}$	$\{x_{975}$	$x_{995}\}$
Effect size for above simulation scenarios								
	0.8	−0.7	1.0	−0.9	−0.8	0.9	−1.0	0.7
6 varying effect size								
Group	group1			group5			group20	
predictors	$\{x_5$	x_{20}	$x_{40}\}$	$\{x_{210}$	x_{220}	$x_{240}\}$	$\{x_{975}$	$x_{995}\}$
Effect size for scenario 6								
	(−2, 2)	−0.7	1.0	−0.9	−0.8	0.9	−1.0	0.7

Note:{} quotes the predictors within a group.

the first one as the training data to fit the models and the second one as the test data to evaluate the predictive values. We replicated the simulation 100 times and summarized the results over these replicates. In simulation scenario 6, we vary the effect size of the non-zero coefficient β_5, from − 2 to 2. Other simulation setting are the same with scenario 2. The purpose of this simulation is to see the profile of prior scale along with varying effect size.

Each simulated dataset included $n = 500$ observations, with a censored survival response y_i and a vector of $m = 1000$ continuous predictors$X_i = (x_{i1}, \ldots, x_{im})$. We assumed 20 groups. Each group included about 50 predictors. For example, group 1 and 2 included variables $(x_1, \ldots, x_{50})$ and $(x_{51}, \ldots, x_{100})$, respectively. 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. We simulated several scenarios. The predictors were assumed to be correlated each other with in group and those predictors in different groups were assumed to be independent. The correlation coefficient r was generally set to be 0.5.

To simulate the censored survival response, following the method of Simon [41], we generated the "true" survival time T_i for each individual from the exponential distribution: $T_i \sim \text{Expon}(\exp(\sum_{j=1}^{m} x_{ij}\beta_j))$ and the censoring time C_i for each individual from the exponential distribution: $C_i \sim \text{Expon}(\exp(r_i))$, where r_i were randomly sampled from a standard normal distribution. The observed censored survival time t_i was set to be the minimum of the "true" survival and censoring times, $t_i = \min(T_i, C_i)$, and the censoring indicator d_i was set to be 1 if $C_i > T_i$ and 0 otherwise. Our simulation scenarios resulted in different censoring ratios, but generally below 50%. For all the scenarios, we set eight coefficients to be non-zero and the others to be zero.

Scenario 1: Non-overlap group

In this scenario, each group is independent. There was no any overlap among groups. Eight non-zero predictors$\{x_5, x_{20}, x_{40}\}$, $\{x_{210}, x_{220}, x_{240}\}$, $\{x_{975}, x_{995}\}$ were simulated to be included into three groups, group 1, 5, and 20 (Table 1). The group sizes is 50, including 50 predictors, presented as below:

Group ID:	1	2	…	5	…	19	20
Group setting:	x_1 – x_{50}	x_{51} – x_{100}		x_{201} – x_{250}		x_{901} – x_{950}	x_{951} – x_{1000}

Scenario 2: Overlap grouping

In this scenario, overlapped grouping structure was considered. Only the last group is independent. For example, for group 1 and group 2, there were five predictors (x_{46}, x_{47}, x_{48}, x_{49}, x_{50}) belong to two groups. The setting for eight non-zero predictors and their effect sizes are the same with scenario 1. The group sizes is still 50. The overlap structure are presented below:

Group ID:	1	2	3	…	19	20
Group setting:	x_1 – x_{50}	x_{46} – x_{100}	x_{96} – x_{150}		x_{896} – x_{950}	x_{951} – x_{1000}

Scenario 3: Varying group sizes

Group size means the number of predictors included in a group. A big group size means the group included relative more predictors. The group size may affect the model fitting. In this scenario, we assumed two groups, group 1 and 11, including non-zero predictors, $\{x_1, x_2, x_3, x_4\}$ and $\{x_{501}, x_{502}, x_{503}, x_{504}\}$, respectively. Other simulation setting are similar with scenario 2. To investigate the group size effect on model fitting, we simulated different group size as below:

(1). only four non-zero predictors included in group 1 and 11:

Group ID:	1	2	3	…	11	12	…	19	20
Group setting:	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}

(2). 20 predictors included in group 1 and 11:

Group ID:	1	2	3	…	11	12	…	19	20
Group setting:	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}

(3). 50 predictors included in group 1 and 11:

Group ID:	1	2	3	…	11	12	…	19	20
Group setting:	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}

Scenario 4: Varying the number of non-null group

The true non-zero predictors may be included in some groups. Other zero predictors belong to other groups.

These groups included non-zero predictors called non-null group. The number of non-null group may also affect the model fitting. To evaluate the group number effect, we varied the number of non-null groups, as following:

(1). There are 8 non-null groups including non-zero coefficients: $\{x_5\}$, $\{x_{55}\}$, $\{x_{305}\}$, $\{x_{355}\}$, $\{x_{505}\}$, $\{x_{555}\}$, $\{x_{905}\}$, and $\{x_{955}\}$;
(2). There are 3 non-null groups including non-zero coefficients: $\{x_5, x_{15}, x_{25}\}$, $\{x_{355}, x_{365}, x_{375}\}$, and $\{x_{905}, x_{915}\}$;
(3). There is only 1 non-null group including non-zero coefficients: $\{x_5, x_{10}, x_{15}, x_{20}, x_{25}, x_{30}, x_{35}, x_{40}\}$. The overlap settings were the same with scenario 2. The group number and effect sizes of these non-zero coefficients are shown in Table 1.

Scenario 5: Varying the correlation within group

To evaluate the effect of correlation within group, we set different correlation coefficients within a group: $r = 0.0$, 0.5, and 0.7. Other settings were the same with scenario 2.

Scenario 6: Self-adaptive shrinkage on varying the effect size

The significant feature of the proposed spike-and-slab prior is the self-adaptive shrinkage. To show this property, we performed additional simulation study based on Scenario 1. We fixed the prior scale $(s_0, s_1) = (0.02, 1)$ and varied the effect size of the first simulated non-zero predictor (x_5) from (− 2, 2). We recorded the scale parameters for this non-zero predictor (x_5) and nearby zero effect predictor (x_6), and non-zero predictor (x_{20}) with the simulated effect size − 0.7. These three predictors belong to the same group.

Real data analysis

We applied the proposed gsslasso Cox model to analyze three real datasets, ovarian cancer (OV), lung adenocarcinoma (LUAD), and breast cancer. The whole genome expression data were downloaded from The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) (updated at June 2017). We firstly clean the data to get the clear survival information and potential genes involved in further analysis. The details of the three datasets and clean procedure are described below paragraphs. Secondly, several genome annotation tools were used to construct the pathways information. All the genes were mapped to KEGG pathways by using *R/bioconductor* packages: *mygene*, clusterProfiler *and AnnotationDbi* [46]. The R/Bioconductor mygene package was used to convert gene names to gene ENTREZ ID. The clusterProfiler package was used to get pathway/group information for genes, by loading the gene ENTREZ ID. AnnotationDbi was used primarily to create mapping objects that allow easy access from R to underlying annotation databases, like KEGG in the present study. By using these packages, we mapped the genes into pathways, and got group structure information for further analysis. Only the gene included in pathways were used in further analysis. Thirdly, the proposed method and several penalization approaches used in above simulation study were applied to analyze the survival data with thousands of genes and pathway/group information. We performed 10-fold cross-validation with 10 replicates to evaluate the predictive values of the several models. After model fitting, the non-zero parameters were the detected genes.

TCGA ovarian cancer dataset (mRNA sequencing data)

This dataset contains mRNA expression data and relevant clinical outcome for ovarian cancer (OV) from TCGA. The raw dataset includes 304 patients and 20,503 genes after removing the duplication and unknown gene names. The raw clinical data included 586 patients. We cleaned the clinical survival data from several clinical files, and obtained 582 patients with clear survival information. We merged the individuals both with gene expression data and survival information, and obtained 304 patients with 20,503 genes for further analysis. First, we filtered the genes with expressions values less than 10. Then, genes with more than 30% of zero expression values in the dataset were removed. Furthermore, we calculated the coefficient of variance (CV) of expression values for each gene, and kept the genes with CV of larger than 20% quantile. After these steps, 304 patients with 14,265 genes were included in our analysis. The censoring ratio was 39.5%.We mapped these genes to 271 pathways including 4260 genes.

TCGA lung adenocarcinoma dataset (mRNA sequencing data)

The raw expression data contains 578 patients and 20,530 genes. After removing the duplication and unknown gene names, there are 516 patients with 20,501 used for further analysis. The raw clinical data included 521 patients. We cleaned the clinical data with clear survival records, and included 497 patients in our analysis. We then merged the clinical data and expression data, and obtained 491 patients for with 20,501 genes for quality control. Similar with the steps for ovarian cancer dataset, we filtered the genes with expressions values less than 10. Then, we removed genes with more than 30% of zero expression values in the dataset. Furthermore, we calculated the coefficient of variance (CV) of expression values for each gene, and kept the genes with CV of larger than 20% quantile. After these steps, 491 patients with 14,143 genes were included in our analysis. The censoring ratio was

68.4%. We mapped these genes to 274 pathways including 4266 genes.

TCGA breast cancer dataset (mRNA sequencing data)

The raw expression data contains 1220 patients and 20,530 genes. After removing the duplication and unknown gene names, there are 1097 patients with 20,503 used for further analysis. The raw clinical data included 1097 patients. We cleaned the clinical data with clear survival records, and included 1084 patients in our analysis. We then merged the clinical data and expression data, and obtained 1082 patients for with 20,503 genes for quality control. The same steps used here for breast cancer dataset, we filtered the genes with expressions values less than 10, and removed genes with more than 30% of zero expression values in the dataset. Furthermore, we calculated the coefficient of variance (CV) of expression values for each gene, and kept the genes with CV of larger than 20% quantile. After these steps, 1082 patients with 14,077 genes were included in our analysis. The censoring ratio was 86.0%. We mapped these genes to 275 pathways including 4385 genes.

Results

Simulation results

Predictive performance

Tables 2 and 3 summarizes the CVPL (cross-validated partial likelihood) and C-index in the testing data over 100 replicates for Scenarios 1–5. We observed that the group spike-and-slab lasso Cox model performed similarly with cMCP and outperformed other methods, under different simulation scenarios. These results suggested that, with complex group structures, the proposed method could perform well.

Table 2 Estimates of two measures over 100 replicates under simulation scenario 1 and 2

	Methods	CVPL	C-index
Scenario 1	gsslasso	− 1111.541(52.390)	0.848(0.012)
	lasso	− 1140.742(52.108)	0.836(0.013)
	grplasso	− 1198.449(53.664)	0.792(0.017)
	grMCP	− 1280.783(66.870)	0.736(0.039)
	grSCAD	− 1256.297(57.293)	0.752(0.027)
	cMCP	− 1114.934(53.278)	0.847(0.012)
	SGL	− 1167.902(72.121)	0.826(0.016)
Scenario 2	gsslasso	− 1077.398(56.949)	0.868(0.011)
	lasso	−1114.886(56.200)	0.853(0.012)
	grplasso	− 1161.058 (59.318)	0.825 (0.015)
	grMCP	− 1236.072(67.840)	0.775(0.018)
	grSCAD	− 1219.129(66.240)	0.798(0.020)
	cMCP	− 1078.363 (57.004)	0.866 (0.011)

Note: Values in the parentheses are standard deviations. "gsslasso" represents the proposed group spike-and-slab lasso cox. The slab scales, s_1, are 1 in the analyses. The optimal $s_0 = 0.02$ and $s_0 = 0.03$ for gsslasso cox methods under scenario 1 and 2, respectively. For scenarios with overlap structures, SGL method was not used for comparison since it cannot handle overlap situation directly

Accuracy of parameter estimates

To evaluate the accuracy of parameters estimation, we summarized the average numbers of non-zero coefficients and the mean absolute errors (MAE) of coefficient estimates, defined as MAE = $\sum |\hat{\beta}_j - \beta_j| / m$, in Tables 4 and 5 for different scenarios**.** It was found that the dected number of null-zero coefficients were very close preset number 8, and the values of MAE were very small for the proposed method under different scenarios. The performances of the group spike-and-slab lasso Cox and cMCP were consistently better than the other methods for all the five scenarios, and the proposed method was slightly better than cMCP. These results suggested that the proposed method can generate lowest false positive and unbiased estimation.

The estimates of coefficients from the group spike-and-slab lasso Cox and the other methods over 100 replicates are shown in Fig. 1 and Additional file 1: Figure S1, Additional file 2: Figure S2, Additional file 3: Figure S3, Additional file 4: Figure S4, Additional file 5: Figure S5, Additional file 6: Figure S6 and Additional file 7: Figure S7 for different scenarios. It can be seen that the group spike-and-slab lasso Cox method 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 except for cMCP, gave a strong shrinkage amount on all the coefficients and resulted in the solutions that non-zero coefficients were shrunk and underestimated compared to true values. In addition, higher false positives (grey bars) were observed, except for the group spike-and-slab lasso Cox and cMCP methods.

The self-adaptive shrinkage feature

To show the self-adaptive shrinkage feature, we performed simulation 6. Figure 2 shows the adaptive shrinkage amount on non-zero coefficients x_5, along with the varying effect size. It clearly shows that the proposed spike-and-slab lasso Cox model approach has self-adaptive and flexible characteristics, without affecting the nearby zero coefficient (x_6) and non-zero variable (x_{20}) belong to the same group.

Real data analysis results

There were about one third genes were mapped to pathways for the above three real datasets. The rest genes were put together as an additional group. The detailed information of genes shared by different

Table 3 Estimates of two measures over 100 replicates for varying group size and varying number of non-null group under simulation scenario 3,4 and 5, respectively

scenario 3				scenario 4			scenario 5		
Group size	methods	CVPL	C-index	number of non-null group	CVPL	C-index	Correlation coefficients within group	CVPL	C-index
4/4	gsslasso	− 1130.995 (58.229)	0.829 (0.0513)	8/20	− 1090.819 (53.224)	0.875 (0.010)	$r = 0.0$	−1077.130 (57.084)	0.876 (0.009)
	lasso	−1167.319 (57.844)	0.813 (0.015)		− 1113.349 (52.438)	0.870 (0.010)		− 1104.924 (56.431)	0.870 (0.010)
	grlasso	− 1137.892 (57.414)	0.827 (0.014)		− 1266.185 (57.782)	0.746 (0.018)		− 1174.234 (57.919)	0.829 (0.014)
	grMCP	− 1131.451 (57.960)	0.829 (0.013)		− 1334.359 (58.901)	0.616 (0.029)		− 1287.124 (64.897)	0.747 (0.035)
	grSCAD	− 1132.272 (58.315)	0.829 (0.013)		− 1305.299 (58.587)	0.721 (0.025)		− 1258.988 (62.970)	0.795 (0.026)
	cMCP	−1131.483 (58.339)	0.829 (0.013)		− 1094.230 (52.983)	0.875 (0.010)		−1082.770 (57.483)	0.875 (0.010)
4/20	gsslasso	− 1149.792 (56.801)	0.830 (0.013)	3/20	− 1120.043 (62.936)	0.849 (0.013)	$r = 0.5$	− 1087.823 (56.773)	0.865 (0.011)
	lasso	− 1179.653 (56.986)	0.813 (0.014)		−1149.463 (61.507)	0.836(0.015)		− 1119.388 (56.076)	0.852 (0.013)
	grlasso	−1179.498 (55.463)	0.811 (0.013)		− 1213.466 (62.431)	0.784 (0.018)		− 1157.999 (54.642)	0.828 (0.013)
	grMCP	− 1172.856 (56.712)	0.816 (0.013)		− 1318.758 (64.886)	0.685 (0.033)		− 1226.349 (62.257)	0.778 (0.018)
	grSCAD	−1172.884 (56.852)	0.816 (0.013)		− 1278.753 (62.726)	0.756 (0.023)		− 1208.197 (63.032)	0.801 (0.018)
	cMCP	− 1150.915 (56.806)	0.827 (0.013)		− 1122.606 (62.852)	0.848(0.014)		− 1089.138 (56.817)	0.864 (0.011)
4/50	gsslasso	− 1145.155 (56.523)	0.825 (0.013)	1/20	− 1141.219 (60.329)	0.824 (0.014)	$r = 0.7$	−1113.142 (60.749)	0.852 (0.012)
	lasso	− 1176.796 (56.449)	0.810 (0.015)		−1172.768 (57.418)	0.810 (0.014)		−1130.099 (60.286)	0.834 (0.013)
	grlasso	−1208.999 (55.893)	0.782 (0.017)		− 1180.395 (58.095)	0.802 (0.017)		− 1164.874 (59.066)	0.814 (0.013)
	grMCP	− 1272.423 (73.279)	0.782 (0.082)		− 1178.416 (64.849)	0.808 (0.016)		− 1202.094 (62.653)	0.822 (0.013)
	grSCAD	− 1271.286 (58.185)	0.777 (0.018)		−1178.827 (65.082)	0.808 (0.016)		− 1195.481 (63.401)	0.852 (0.012)
	cMCP	− 1148.318 (56.896)	0.824 (0.013)		− 1147.845 (59.271)	0.821 (0.014)		− 1117.158 (61.869)	0.858 (0.013)

Notes: in scenario 3, group size "4/50" denotes that there are four none-zero coefficients embedded in a group with 50 predictors. The group size is 50. This is true for "4/20" and "4/4". The optimal $s_0 = 0.02$ for different group size settings. In scenario 4, "8/20" denotes that there are 8 non-null groups among 20 groups. Each non-null group includes at least one non-zero coefficients. The optimal $s_0 = 0.02$ for the three settings. In scenario 5, the optimal s_0 are 0.02, 0.03, and 0.04 for different correlation coefficients, 0.0, 0.5, and 0.7 within group, respectively. The slab scales, s_1, are 1 in this scenario 3 4, and 5. Values in the parentheses are standard errors. "gsslasso" represents the proposed group spike-and-slab lasso cox

pathways is listed in Additional file 8: S1, S2, and S3, for ovarian cancer, lung cancer and breast cancer, respectively.

Real data analysis is to build a survival model for predicting the overall survival outcome by integrating gene expression data and pathway information. We standardized all the predictors to use a common scale for all predictors, prior to fitting the models, using the function *covariate*() function in BhGLM package. In our prior distribution, there were to preset parameters, (s_0, s_1). In our real data analysis, we fixed the slab scale s_1 to 1, and varied the spike scale s_0 values by: $\{k \times 0.01; k = 1, \ldots, 9\}$, leading to 9 models. The optimal spike scale s_0 was select by the 10-fold

Table 4 Average number of non-zero coefficients and mean absolute error (MAE) of coefficient estimates over 100 simulations for scenario 1 and 2

	Method	Average Number	MAE
Scenario 1	gsslasso	8.61	0.60 (0.24)
	lasso	51.99	3.77 (0.40)
	grlasso	474.80	12.43 (1.64)
	grMCP	62.00	9.30 (2.56)
	grSCAD	108.80	8.41 (1.25)
	cMCP	14.19	0.96 (0.34)
	SGL	39.79	6.25 (1.65)
Scenario 2	gsslasso	9.74	1.29 (0.84)
	lasso	53.70	4.02 (0.46)
	grlasso	502.05	12.11 (2.08)
	grMCP	57.13	8.04 (0.67)
	grSCAD	167.59	8.77 (0.93)
	cMCP	15.14	0.96 (0.33)

*: the optimal $s_0 = 0.02$ and $s_0 = 0.03$ for gsslasso method under scenario 1 and 2, respectively. For scenarios with overlap structures, SGL method was not used for comparison since it cannot handle overlap situation directly

10-time cross-validation according to the CVPL. Using the optimal s_0, we performed further real data analysis. For comparison, we also analyzed the data using the several existing methods as described in the simulation studies.

We performed 10-fold cross-validation with 10 replicates to evaluate the predictive values of the several models. Table 6 summarizes the measures of the prognostic performance on these three data sets, by only using the genes included in pathway. For all the data sets the proposed group spike-and-slab lasso Cox model performed better than the other methods. The above results used only genes mapped in pathways. Additional file 9 shows the measures of the performance on these three data sets, by using the all genes. The genes which were not mapped into any pathway were put together as an additional group. We can see that the prediction performance of the proposed method were still better than the other methods.

The pathway enrichment analyses for these detected genes were summarized in Additional file 10: S4-S6. Additional file 11: S7 presents the genes detected by the proposed and existed methods. Their standardized effects size were also plotted for the three real data sets. There were many common gene among different methods. For ovarian cancer dataset, the genes CYP2R1 and HLA-DOB detected by the proposed gsslasso method, were also detected by both lasso and cMCP methods. For Lung cancer dataset, several genes detected by the proposed gsslasso method, such

Table 5 Average number of non-zero coefficients and mean absolute error (MAE) of coefficient estimates over 100 simulations for scenario 3, 4, and 5

scenario 3: Group size						
	4/4		4/20		4/50	
	Average Number	MAE	Average Number	MAE	Average Number	MAE
gsslasso	9.09	0.58 (0.22)	9.26	0.64 (0.29)	9.27	0.69 (0.54)
lasso	53.75	3.90 (0.41)	54.58	3.94 (0.43)	53.96	3.93 (0.44)
grlasso	270.78	2.78 (1.22)	455.15	7.06 (1.69)	509.75	10.58 (1.71)
grMCP	13.40	0.57 (0.18)	40.00	3.50 (0.54)	84.50	10.09 (2.52)
grSCAD	56.16	0.85 (0.78)	53.85	3.62 (0.74)	100.00	7.86 (1.77)
cMCP	9.42	0.64 (0.29)	14.64	0.98 (0.33)	16.85	1.05 (0.38)
scenario 4: Number of non-null groups						
	8/20		3/20		1/20	
gsslasso	8.85	0.54 (0.19)	9.12	0.56 (0.19)	9.27	0.68 (0.26)
lasso	52.87	3.58 (0.44)	53.51	3.89 (0.43)	52.49	3.90 (0.41)
grlasso	757.1	19.84 (2.51)	610.25	13.91 (2.08)	461.25	7.64 (1.71)
grMCP	83.95	7.68 (0.60)	46.00	7.49 (1.51)	50.00	4.52 (0.67)
grSCAD	410.3	10.80 (0.82)	142.30	8.04 (0.74)	55.85	4.57 (0.74)
cMCP	13.22	0.83 (0.38)	15.74	0.96 (0.35)	14.81	1.03 (0.31)
scenario 5: Correlation coefficients within group						
	$r = 0$		$r = 0.5$		$r = 0.7$	
gsslasso	9.18	0.85 (0.72)	8.90	1.07 (0.99)	8.11	3.54 (0.54)
lasso	59.10	3.42 (0.35)	52.63	4.04 (0.49)	48.82	5.25 (0.50)
grlasso	557.00	10.27 (1.25)	490.50	11.88 (1.75)	465.90	13.61 (2.45)
grMCP	61.71	7.38 (0.89)	57.40	8.14 (1.29)	53.54	9.59 (0.83)
grSCAD	148.68	7.42 (0.67)	170.61	8.78 (1.16)	194.58	11.28 (1.28)
cMCP	16.72	0.98 (0.47)	14.32	0.98 (0.37)	21.51	3.53 (0.40)

Notes: in scenario 3, group size "4/50" denotes that there are four none-zero coefficients embedded in a group with 50 predictors. The group size is 50. This is true for "4/20" and "4/4". The optimal $s_0 = 0.02$ for different group size settings. The slab scales, s_1, are 1 in this scenario. In scenario 4 "8/20" denotes that there are 8 non-null groups among 20 groups. Each non-null group includs at least one non-zero coefficients. The optimal $s_0 = 0.02$ for the three settings. In scenario 5, the optimal s_0 are 0.02, 0.03, and 0.04 for different correlation coefficients, 0.0, 0.5, and 0.7 within group, respectively. The slab scales, s_1, are 1 in this scenario 3, 4 and 5. Values in the parentheses are standard errors. "gsslasso" represents the proposed group spike-and-slab lasso cox

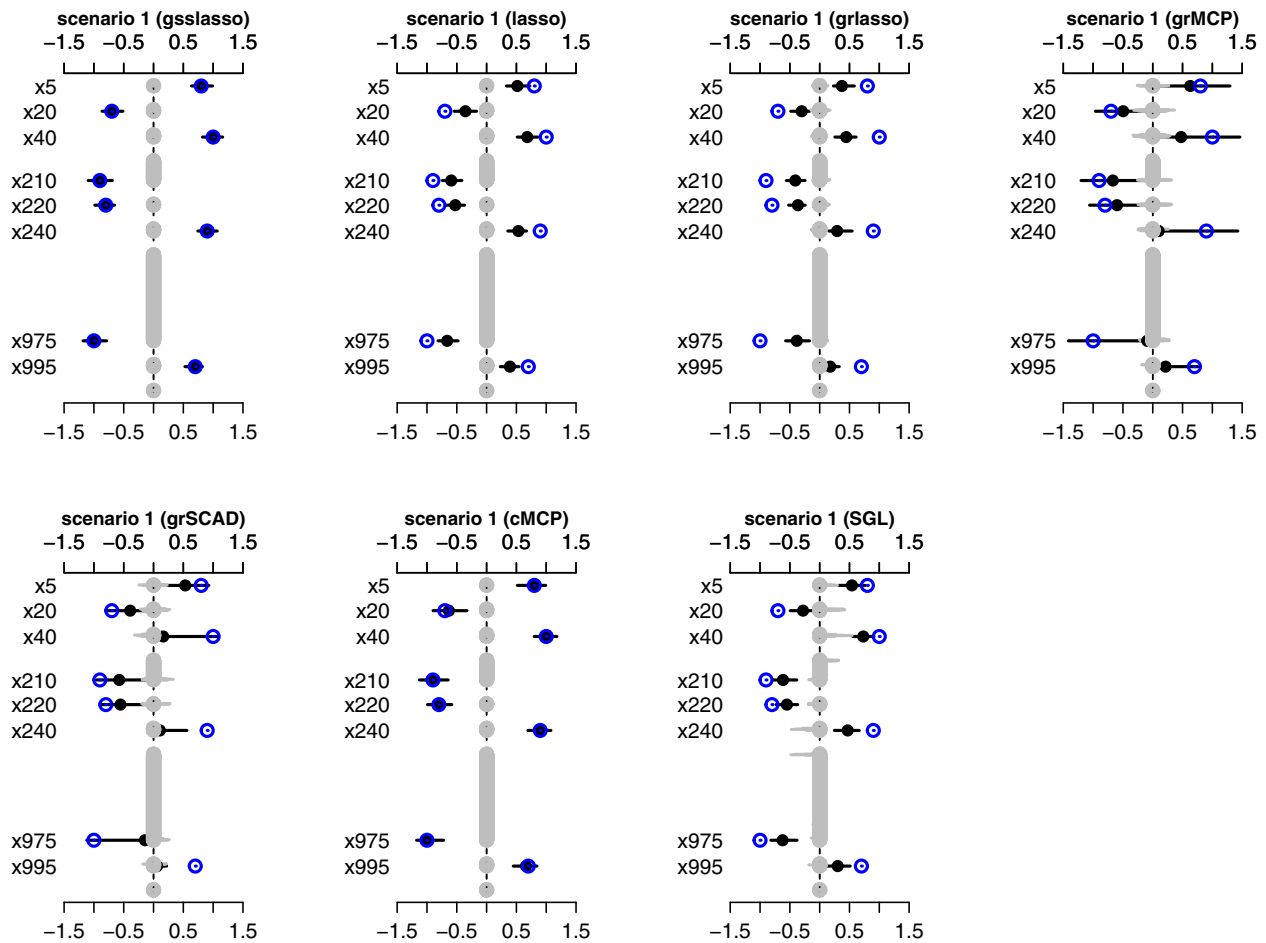

Fig. 1 The parameter estimation averaged over 100 replicates for the group spike-and-slab lasso Cox (gsslasso), the lasso, grlasso, grMCP, grSCAD, SGL and cMCP methods for Scenario 1. Blue cycles are the simulated non-zero values. Black points and lines represent the estimated values and the interval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

as VDAC1, EHHADH, ACAT2, KIT, CCND1, PIK3R1, NRAS, GNPNAT1, and KYNU, were also detected by other method. For, breast cancer dataset, two genes HSPA1A and ABCB5 detected by the proposed gsslasso method were also detected by other method.

We found that most of the genes detected by the proposed method were associated with cancers in previous studies. HABP2, detected in ovarian cancer, was associated with familial nonmedullary thyroid cancer [47]. CYP24A1, the main enzyme responsible for the degradation of active vitamin D, plays an important role in many cancer related cellular processes. The associations between CYP24A1 polymorphisms and cancer susceptibility had been evaluated by many studies [48]. Keratin 8 (KRT8) plays an essential role in the development and metastasis of multiple human cancers. A recent study suggested that in clear cell renal cell carcinoma and gastric cancer, KRT8 upregulation promotes tumor metastasis and associated with poor prognosis [49, 50]. E2F7, detected in lung cancer, involved in several cancer studies, which might act as an independent prognostic factor for breast cancer, and Squamous Cell Carcinoma, and gliomas [51–53]. Most of the genes detected by the proposed method in the three real datasets had been found associated with different cancers. These results may provide some interesting information for further studies.

Discussion

The group structures among various features arise naturally in many biological and medical researches, especially in large-scale omics data. Such grouping information is biologically meaningful and intrinsically encoded in the biological data. Thus it is desirable to incorporate the grouping information into data analysis. Various penalization methods have been designed for such situations [13, 14, 30, 54, 55]. Recently, we have developed a novel hierarchical modeling approach, the group spike-and-slab lasso GLMs, to integrate the variable group information for gene detection and

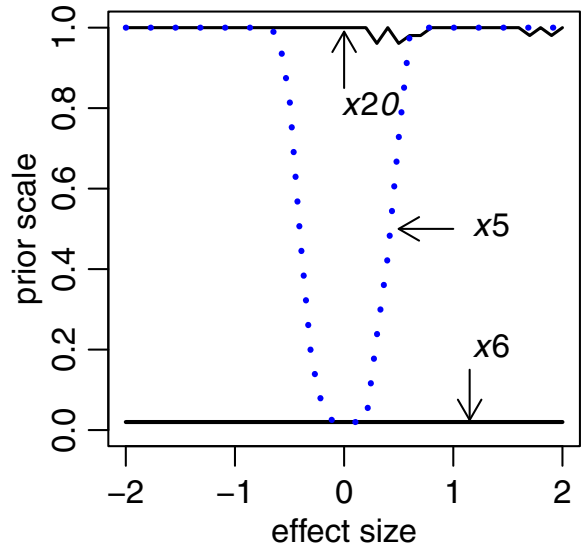

Fig. 2 The adaptive shrinkage amount, along with the varying effect size for x_5. The x_{20} is the none-zero coefficients with the simulated effect size − 0.7, while x_6 is the nearby coefficient with simuated zero effect size

prognostic prediction [27]. In this study, we extended the method to Cox proportional hazards model for analyzing censored survival data.

Similar to the group spike-and-slab lasso GLMs, the key to our group spike-and-slab lasso Cox is the group spike-and-slab double-exponential prior. This prior has significant advantage in variable selection and parameter estimation. It induces weak shrinkage on larger coefficients and strong shrinkage on irrelevant coefficients. In contrast, other methods usually gave a strong shrinkage amount on all the coefficients and resulted in the solutions that non-zero coefficients were shrunk and underestimated. The proposed group spike-and-slab prior allows the model to incorporate the biological similarity of genes within a same pathway into the analysis.

The spike-and-slab prior depends on the spike and slab scale parameters. Our previous study suggested that slab scale s_1 had little influence on model fitting, while the spike scale s_0 strongly affected model performance [25, 26]. A slab scale s_1 value introducing weak shrinkage amount would be helpful to include relevant variables into the model. Therefore, we set $s_1 = 1$ in our analysis. We evaluated the performance of the proposed model on a grid of values of spike scale s_0 from a reasonable range, e.g., (0, 0.1), and then selecting an optimal value using cross-validation. This is a path-following strategy for fast

Table 6 The measures of optimal group spike-and-slab lasso (gsslasso) cox and the lasso cox models for TCGA ovarian cancer, lung adenocarcinoma (LUAD) and breast cancer dataset with pathway genes by 10 times 10-fold cross validation

	Pathway number	Genes included	Methods	CVPL	C-index	Number of non-zero gene
TCGA	271	4260	gsslasso	− 1041.218 (2.118)	0.577 (0.012)	33
ovarian			lasso	− 1042.905 (1.687)	0.533 (0.027)	15
cancer			grlasso	− 1044.110 (12.741)	0.504 (0.014)	24
$N = 304$			grMCP	− 1046.965 (8.604)	0.502 (0.007)	24
			grSCAD	−1042.349 (5.339)	0.503 (0.012)	24
			cMCP	− 1043.373 (2.215)	0.532 (0.019)	13
TCGA	274	4266	gsslasso	− 938.973 (1.675)	0.559 (0.010)	64
LUAD			lasso	− 941.383 (3.720)	0.545 (0.019)	13
$N = 491$			grlasso	− 945.605 (8.137)	0.547 (0.023)	111
			grMCP	− 1092.091 (30.477)	0.512 (0.015)	25
			grSCAD	− 940.358 (1.331)	0.538 (0.021)	123
			cMCP	−942.831 (3.301)	0.530 (0.022)	3
TCGA	275	4385	gsslasso	−996.491 (2.131)	0.640 (0.153)	86
Breast			lasso	−1002.046 (5.356)	0.523 (0.027)	2
cancer			grlasso	− 1001.073 (9.641)	0.590 (0.022)	93
$N = 1082$			grMCP	− 1016.864 (25.290)	0.520 (0.019)	12
			grSCAD	− 1005.299 (2.268)	0.522 (0.007)	24
			cMCP	− 1012.587 (44.339)	0.502 (0.012)	1

Note: Values in the parentheses are standard errors. For group spike-and-slab lasso model, the optimal $s_0 = 0.03$ for three data sets. In TCGA ovarian cancer, we mapped 4260 genes into 271 pathways. The analyses was performed on these genes including in these pathways. The same is true for other two datasets

dynamic posterior exploration of the proposed models, which is similar to the approach of Ročková and George [24, 56]. Additional file 12: Figure S8 a and b show the solution paths under Scenario 2, for the proposed model and the lasso Cox model. Additional file 12: Figure S8 c and d show the profiles of cross-validated palatial log-likelihood by 10-fold cross-validation for the proposed model. These profiles would help to choose optimal tuning parameters. It could be found that, similar to the lasso, the spike-and-slab lasso Cox is a path-following strategy for fast dynamic posterior exploration. However, the solution path is essentially different from that of the lasso model. For the lasso cox model, the number of non-zero coefficient could be a few, even zero if a strong penalty is adopted. However, in the spike-and-slab lasso Cox model, larger coefficients will be always included in the model with weak shrinkage, while irrelevant coefficients are removed (grey path in Additional file 12: Figure S8 a).

Another feature of the proposed spike-and-slab prior is bi-level selection, which is capable of selecting important groups as well as important individual variables within those groups. Several methods perform bi-level selection, including cMCP method [16], SGL [14], and group exponential lasso [18]. The underlying assumption is that the model is sparse at both the group and individual variable levels. The proposed group spike-and-slab lasso Cox model can efficiently perform bi-level selection. In group level, the importance of a group is controlled by the group-specific probability θ_g. Within a group, the spike-and-slab prior allows to perform variable selection by shrinking irrelevant or small effect coefficients exactly to zero, without affecting the prediction performance.

The extensive simulation studies show that the prediction performance of the proposed method is always slightly better than cMCP method, and significantly better than all other methods under different scenarios. In the real data analysis, the prediction accuracy of the proposed method incorporating pathway information was slightly improved compared with the existing methods. This might be mainly due to the complex genetic components involved in the expression data, like haplotype blocks, subnetworks, and interaction among the genes. The present model under the linear assumption may not capture these complexities. More sophisticated strategies could potentially enhance prediction accuracy and further improve the models, by defining more precise biological grouping information.

There are several further extensions of the proposed method. For example, it can also be extended to incorporate multiple level group structure, like three-level group structure, i.e. SNP-gene-pathway. In addition, the proposed model takes the spike-and-slab mixture double-exponential prior. Other priors with a spike at zero and includes heavier tails could be investigated, like Cauchy distribution, a special case of Student-t distribution. The theoretical and empirical properties of other priors are different, which may introduce more interesting results.

Conclusion

Incorporating biological group structure in high-dimensional molecular data analysis can improve the accuracy of disease prediction and power of gene detection. We propose a new hierarchical Cox model, gsslasso Cox, for incorporating biological pathway information for predicting disease survival outcomes and detecting associated genes. We develop a fast and stable deterministic algorithm to fit the proposed models. Extensive simulation studies and real applications show that compared with several existing methods, the proposed approach provides more accurate parameter estimation and survival prediction. The proposed method has been implemented in a freely available R package BhGLM.

Additional files

Additional file 1: Figure S1. The parameter estimation averaged over 100 replicates for the group spike-and-slab lasso Cox (gsslasso), the lasso, grlasso, grMCP, grSCAD, and cMCP methods for Scenario 2. Blue cycles denote the simulated non-zero values. Black points and lines represent the estimated values and the interval estimates of coefficients. (PDF 523 kb)

Additional file 2: Figure S2. The parameter estimation averaged over 100 replicates for the group spike-and-slab lasso Cox (gsslasso), the lasso and grlasso methods for Scenario 3. Blue cycles denote the simulated non-zero values. Black points and lines represent the estimated values and the interval estimates of coefficients. The main title of each plot denotes the varying group size for scenario 3. (PDF 778 kb)

Additional file 3: Figure S3. The parameter estimation averaged over 100 replicates for grMCP, grSCAD, and cMCP methods for Scenario 3. Blue cycles denote the simulated non-zero values. Black points and lines represent the estimated values and the interval estimates of coefficients. The main title of each plot denotes the varying group size for Scenario 3. (PDF 767 kb)

Additional file 4: Figure S4. The parameter estimation averaged over 100 replicates for the group spike-and-slab lasso Cox (gsslasso), the lasso and grlasso methods for Scenario 4. Blue cycles denote the simulated non-zero values. Black points and lines represent the estimated values and the interval estimates of coefficients. The main title of each plot denotes the varying the number of non-null group for Scenario 4. (PDF 791 kb)

Additional file 5: Figure S5. The parameter estimation averaged over 100 replicates for grMCP, grSCAD, and cMCP for Scenario 4. Blue cycles denote the simulated non-zero values. Black points and lines represent the estimated values and the interval estimates of coefficients. The main title of each plot denotes the varying the number of non-null group for Scenario 4. (PDF 783 kb)

Additional file 6: Figure S6. The parameter estimation averaged over 100 replicates for the group spike-and-slab lasso Cox (gsslasso), the lasso and grlasso methods for Scenario 5. Blue cycles denote the simulated non-zero values. Black points and lines represent the estimated values and the interval estimates of coefficients. The main title of each plot denotes the varying the number of non-null group for Scenario 5. (PDF 1054 kb)

Additional file 7: Figure S7. The parameter estimation averaged over 100 replicates for grMCP, grSCAD, and cMCP for Scenario 5. Blue cycles denote the simulated non-zero values. Black points and lines represent the estimated values and the interval estimates of coefficients. The main title of each plot denotes the varying the number of non-null group for Scenario 5. (PDF 778 kb)

Additional file 8: S1, S2, and S3. The detailed information of genes shared by different pathways for ovarian cancer, lung cancer and breast cancer, respectively. (ZIP 213 kb)

Additional file 9: Table S1. The measures of optimal group spike-and-slab lasso (gsslasso) cox and the lasso cox models for TCGA ovarian cancer, lung adenocarcinoma (LUAD) and breast cancer dataset with all genes by 10 times 10-fold cross validation. (DOCX 20 kb)

Additional file 10: S4, S5 and S6. The pathway enrichment analyses for these detected genes for ovarian cancer, lung cancer and breast cancer, respectively. (ZIP 15 kb)

Additional file 11: S7. The detected genes and their standardized effect sizes estimated by the group spike-and-slab lasso Cox model and five existed methods for TCGA real datasets. (PDF 1340 kb)

Additional file 12: Figure S8. The solution path and cross-validated partial loglikelihood profiles of the group spike-and-slab lasso Cox (a, c) and the lasso (b, d) based on the Scenario 2. The colored points on the solution path represent the estimated values of assumed eight non-zero coefficients, and the circles represent true non-zero coefficients. Vertical lines correspond to the optimal models. (PDF 1052 kb)

Abbreviations

cMCP: Composite minimax concave penalty; CVPL: Cross-validated partial likelihood; grlasso: Group lasso; grMCP: Group minimax concave penalty; grSCAD: Group smoothly clipped absolute deviation; gsslasso cox: Group spike-and-slab lasso cox model; lasso: Least absolute shrinkage and selection operator

Acknowledgements

We acknowledge the contributions of the TCGA Research Network.

Funding

This work was supported in part by the National Natural Science Foundation of China (81773541 and 81573253), funds from the Priority Academic Program Development of Jiangsu Higher Education Institutions at Soochow University, grants from China Scholarship Council, Jiangsu Provincial Key Project in Research and Development of Advanced Clinical Technique (BL2018657) to ZT, USA National Institutes of Health (R03-DE024198, R03-DE025646) to NY. The funding body did not played any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

All data for the manuscript will be available without restriction at website: https://cancergenome.nih.gov/. The package BhGLM is freely available from https://github.com/nyiuab/BhGLM.

Authors' contributions

Study conception and design: NY, ZXT. Simulation study and data summary: ZXT, NY. Real data analysis and interpretation: ZXT, YS, SFL, XZ, ZY, BG, JC, NY. Drafting of manuscript: ZXT, YS, SFL, XZ, ZY, BG, JC, NY. All authors read and approved the final version of the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not Applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Department of Biostatistics, School of Public Health, Medical College of Soochow University, University of Alabama at Birmingham, Suzhou 215123, China. [2]Jiangsu Key Laboratory of Preventive and Translational Medicine for Geriatric Diseases, Medical College of Soochow University, Suzhou 215123, China. [3]Department of Biostatistics, Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA 30458, USA. [4]Eastern Virginia Medical School, Norfork, VA 23507, USA. [5]Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294-0022, USA. [6]Informatics Institute, School of Medicine, University of Alabama at Birmingham, Birmingham, AL 35294, USA.

Received: 21 May 2018 Accepted: 28 January 2019
Published online: 27 February 2019

References

1. Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Statistical Soc Series B. 1996;58:267–88.
2. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med. 1997;16(4):385–95.
3. Zhang C. Penalized linear unbiased selection. Rutgers University: Department of Statistics and Bioinformatics; 2007. Technical Report #2007–2003
4. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty; 2010. p. 894–942.
5. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its Oracle properties. J Am Stat Assoc. 2001;96(456):1348–60.
6. Zhang W, Ota T, Shridhar V, Chien J, Wu B, Kuang R. Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. PLoS Comput Biol. 2013;9(3):e1002975.
7. Yuan Y, Van Allen EM, Omberg L, Wagle N, Amin-Mansour A, Sokolov A, Byers LA, Xu Y, Hess KR, Diao L, et al. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat Biotechnol. 2014;32(7):644–52.
8. Sohn I, Sung CO. Predictive modeling using a somatic mutational profile in ovarian high grade serous carcinoma. PLoS One. 2013;8(1):e54089.
9. Rapaport F, Zinovyev A, Dutreix M, Barillot E, Vert J-P. Classification of microarray data using gene networks. BMC Bioinformatics. 2007;8(1):1–15.
10. Barillot E, Calzone L, Hupe P, Vert JP, Zinovyev A. Computational systems biology of Cancer Chapman & Hall/CRC Mathematical & Computational Biology; 2012.
11. Zhao Q, Shi X, Xie Y, Huang J, Shia B, Ma S. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. Brief Bioinform. 2015;16(2):291–303.
12. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B. 2006;68(1):49–67.
13. Friedman J, Hastie T, Tibshirani R. A note on the group lasso and a sparse group lasso. Stanford University: Technical report, Department of Statistics; 2010.
14. Simon N, Friedman J, Hastie T, Tibshirani R. A sparse-group lasso. J Comput Graph Stat. 2013;22(2):231–45.
15. Huang J, Ma S, Xie H, Zhang C-H. A group bridge approach for variable selection. Biometrika. 2009;96(2):339–55.
16. Breheny P, Huang J. Penalized methods for bi-level variable selection. Statistics and its interface. 2009;2(3):369–80.
17. Zhao P, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. Ann Stat. 2009;37(6A):3468–97.
18. Breheny P. The group exponential lasso for bi-level variable selection. Biometrics. 2015;71(3):731–40.
19. Chen Y, Du P, Wang Y. Variable selection in linear models. Wiley Interdisciplinary Reviews: Computational Statistics. 2014;6(1):1–9.
20. Kwon S, Ahn J, Jang W, Lee S, Kim Y. A doubly sparse approach for group variable selection. Ann Inst Stat Math. 2017;69(5):997–1025.
21. Huang J, Breheny P, Ma S. A selective review of group selection in high-dimensional models. Stat Sci. 2012;27(4).
22. Ogutu JO, Piepho HP. Regularized group regression methods for genomic prediction: bridge, MCP, SCAD, group bridge, group lasso, sparse group lasso, group MCP and group SCAD. BMC Proc. 2014;8(Suppl 5):S7.
23. Ročková V, George EI. Bayesian penalty mixing: the case of a non-separable penalty. In: Frigessi A, Bühlmann P, Glad IK, Langaas M, Richardson S, Vannucci M, editors. Statistical analysis for high-dimensional data: the Abel symposium, vol. 2014. Cham: Springer International Publishing; 2016. p. 233–54.

24. Ročková V, George EI: The spike-and-slab lasso. J Am Stat Assoc 2016:Online, DOI: https://doi.org/10.1080/01621459.01622016.01260469.
25. Tang Z, Shen Y, Zhang X, Yi N. The spike-and-slab lasso generalized linear models for prediction and associated genes detection. Genetics. 2017; 205(1):77–88.
26. Tang Z, Shen Y, Zhang X, Yi N. The spike-and-slab lasso Cox model for survival prediction and associated genes detection. Bioinformatics. 2017; 33(18):2799–807.
27. Tang Z, Shen Y, Li Y, Zhang X, Wen J, Qian C, Zhuang W, Shi X, Yi N. Group spike-and-slab lasso generalized linear models for disease prediction and associated genes detection by incorporating pathway information. Bioinformatics. 2018;34(6):901–10.
28. Silver M, Montana G. Alzheimer's disease neuroimaging I: fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. Stat Appl Genet Mol Biol. 2012;11(1):Article 7.
29. Silver M, Chen P, Li R, Cheng CY, Wong TY, Tai ES, Teo YY, Montana G. Pathways-driven sparse regression identifies pathways and genes associated with high-density lipoprotein cholesterol in two Asian cohorts. PLoS Genet. 2013;9(11):e1003939.
30. Jacob L, Obozinski G, Vert J-P. Group lasso with overlap and graph lasso. In: Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Quebec, Canada: 1553431: ACM; 2009. p. 433–40.
31. Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity - the lasso and generalization. New York: CRC Press; 2015.
32. Klein J, Moeschberger M. Survival Analysis. New York: Springer-Verlag; 2003.
33. Ibrahim J, Chen M-H, Debajyoti S. Bayesian survival analysis. New York: Springer-Verlag; 2001.
34. Cox DR. Regression models and life tables. J R Stat Soc. 1972;34:187–220.
35. Breslow NE. Contribution to the discussion of the paper by D. R. Cox. J Royal Stat Soc B. 1972;34:216–7.
36. Efron B. The efficiency of Cox's likelihood function for censored data. J Am Stat Assoc. 1977;72:557–65.
37. van Houwelinggen HG, Putter H. Dynamic prediction in clinical survival analysis. Boca Raton: CRC Press; 2012.
38. Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models. New York: Cambridge University Press; 2007.
39. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. Third ed. New York: Chapman & Hall/CRC Press; 2014.
40. Breslow N. Covariance analysis of censored survival data. Biometrics. 1974;30:89–99.
41. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw. 2011;39(5):1–13.
42. van Houwelingen HC, Bruinsma T, Hart AA, Van't Veer LJ, Wessels LF. Cross-validated Cox regression on microarray gene expression data. Stat Med. 2006;25(18):3201–16.
43. Tibshirani RJ, Efron B. Pre-validation and inference in microarrays. Stat Appl Genet Mol Biol. 2002;1:1–18.
44. Yi N, Tang Z, Zhang X, Guo B. BhGLM: Bayesian hierarchical GLMs and survival models, with applications to genomics and epidemiology. Bioinformatics. 2018. https://doi.org/10.1093/bioinformatics/bty803.
45. Zeng Y, Breheny P. Overlapping group logistic regression with applications to genetic pathway selection. Cancer Informat. 2016;15:179–87.
46. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics : a journal of integrative biology. 2012;16(5):284–7.
47. Gara SK, Jia L, Merino MJ, Agarwal SK, Zhang L, Cam M, Patel D, Kebebew E. Germline HABP2 mutation causing familial nonmedullary thyroid Cancer. N Engl J Med. 2015;373(5):448–55.
48. Zhu M, Qiu S, Zhang X, Wang Y, Souraka TDM, Wen X, Liang C, Tu J. The associations between CYP24A1 polymorphisms and cancer susceptibility: a meta-analysis and trial sequential analysis. Pathology - Research and Practice. 2018;214(1):53-63.
49. Tan HS, Jiang WH, He Y, Wang DS, Wu ZJ, Wu DS, Gao L, Bao Y, Shi JZ, Liu B, et al. KRT8 upregulation promotes tumor metastasis and is predictive of a poor prognosis in clear cell renal cell carcinoma. Oncotarget. 2017;8(44):76189–203.
50. Fang J, Wang H, Liu Y, Ding F, Ni Y, Shao S. High KRT8 expression promotes tumor progression and metastasis of gastric cancer. Cancer Sci. 2017;108(2):178–86.
51. Chu J, Zhu Y, Liu Y, Sun L, Lv X, Wu Y, Hu P, Su F, Gong C, Song E, et al. E2F7 overexpression leads to tamoxifen resistance in breast cancer cells by competing with E2F1 at miR-15a/16 promoter. Oncotarget. 2015;6(31):31944–57.
52. Yin W, Wang B, Ding M, Huo Y, Hu H, Cai R, Zhou T, Gao Z, Wang Z, Chen D. Elevated E2F7 expression predicts poor prognosis in human patients with gliomas. J Clin Neurosci. 2016;33:187–93.
53. Hazar-Rethinam M, de Long LM, Gannon OM, Boros S, Vargas AC, Dzienis M, Mukhopadhyay P, Saenz-Ponce N, Dantzic DDE, Simpson F, et al. RacGAP1 is a novel downstream effector of E2F7-dependent resistance to doxorubicin and is prognostic for overall survival in squamous cell carcinoma. Mol Cancer Ther. 2015;14(8):1939–50.
54. Meier L, van de Geer S, Bühlmann P. The group lasso for logistic regression. J Royal Stat Soc Series B. 2008;70(1):53–71.
55. Zhou N, Zhu J. Group variable selection via a hierarchical lasso and its Oracle property; 2011.
56. Ročková V, George EI. EMVS: the EM approach to Bayesian variable selection. J Am Stat Assoc. 2014;109(504):828–46.