Statistics and Computing (2021) 31:77
https://doi.org/10.1007/s11222-021-10052-4

# Penalized Cox's proportional hazards model for high-dimensional survival data with grouped predictors

**Xuan Dang**[1] · **Shuai Huang**[2] · **Xiaoning Qian**[1]

Received: 8 July 2020 / Accepted: 13 September 2021 / Published online: 30 September 2021


**Abstract**
The rapid development of next-generation sequencing technologies has made it possible to measure the expression profiles of thousands of genes simultaneously. Often, there exist group structures among genes manifesting biological pathways and functional relationships. Analyzing such high-dimensional and structural datasets can be computationally expensive and results in the complicated models that are hard to interpret. To address this, variable selection such as penalized methods are often taken. Here, we focus on the Cox's proportional hazards model to deal with censoring data. Most of the existing penalized methods for Cox's model are the group lasso methods that show deficiencies, including the over-shrinkage problem. In addition, the contemporary algorithms either exhibit the loss of efficiency or require the group-wise orthonormality assumption. Hence, efficient algorithms for general design matrices are needed to enable practical applications. In this paper, we investigate and comprehensively evaluate three group penalized methods for Cox's model: the group lasso and two nonconvex penalization methods—group SCAD and group MCP—that have several advantages over the group lasso. These methods are able to perform group selection in both non-overlapping and overlapping cases. We have developed the fast and stable algorithms and a new package *grpCox* to fit these models without the initial orthonormalization step. The runtime of *grpCox* is improved significantly over the existing packages, such as *grpsurv* (for the non-overlapping case), *grpregOverlap* (overlapping), and *SGL*. In addition, *grpCox* is better than *grpsurv* and comparable with *SGL* in terms of variable selection performances. Comprehensive studies on both simulation and real-world cancer datasets demonstrate the statistical properties of our *grpCox* implementations with the group lasso, SCAD, and MCP regularization terms.

**Keywords** Penalized method · High-dimensional · Survival analysis · Group-wise descent · Majorization-minimization (MM) approach

## 1 Introduction

The Cox's proportional hazards model (Cox 1972) is commonly used to study the relationship between survival time and a set of covariates in high-dimensional space as potential predictors for survival time. To tackle the curse of dimensionality and construct robust and interpretable models that generalize well, variable selection approaches, including penalization-based methods, are often taken.

Variable selection for the Cox's proportional hazards model has been extensively studied, including implementations based on lasso (Tibshirani 1996; Gui and Li 2005; Park and Hastie 2006), adaptive lasso (Zhang and Lu 2007; Zou 2008), the smoothly clipped absolute deviation (SCAD) (Fan and Li 2002), to name a few. These methods can automatically select the important covariates by shrinking the coefficients of unimportant covariates to be exactly zero. However, these methods fail to produce good results when there exist group structures in covariates. A common group structure example is where each categorical covariate is expressed through a set of dummy variables. Group structures can also be introduced by integration of prior knowledge that is scientifically meaningful. For example, in gene expression analysis, genes belonging to the same biological pathway

✉ Xuan Dang
xuandt89@tamu.edu

Shuai Huang
shuaih@uw.edu

Xiaoning Qian
xqian@ece.tamu.edu

[1] Texas A&M University, College Station, TX 77840, USA

[2] University of Washington, Seattle, WA 98195, USA

have similar functions and act together in regulating a biological system. These genes can be considered as a group.

Group selection in various statistical modeling problems has been considered in literature. Yuan and Lin (2006) introduced the group lasso for linear regression with the $l_2-$norm of the coefficients for a group of covariates in the penalty function. Meir et al. (2008) extended it to logistic regression. Zhao et al. (2009) used a general composite absolute penalty, which treats the group lasso as a special case. Wang et al. (2007) introduced group SCAD to linear regression. The group minimax concave penalty (MCP) was presented in Huang et al. (2012). Breheny and Huang (2015) introduced nonconvex penalties for linear and logistic regression. These works require the group-wise orthonormal condition to implement their algorithms. The solutions of the group lasso with non-orthonormal matrices for linear regression, logistic regression and SVM classifiers have been developed in literature (Puig et al. 2011; Simon et al. 2013; Yang and Zou 2015).

There are, however, few extensions to the Cox's model. Ma et al. (2007) applied the supervised group lasso to select both significant gene clusters and significant genes within these clusters for both logistic binary classification and Cox's survival model, for which the lasso and group lasso methods were implemented separately. In the first step, it identified important genes within each group based on the lasso formulation. In the second step, it selected important groups using the group lasso formulation. Simon et al. (2013) introduced the sparse group lasso method combining the lasso with group lasso formulations to yield sparsity at both the group and individual levels for the Cox's proportional hazards model. Wu and Wang (2013) introduced the doubly regularized Cox regression that can deal with a mixture of individual sparsity and group sparsity with the extension to an overlapping case. Very recently, Belhechmi et al. (2020) presented a statistical approach that can handle sparse group lasso cases with superior variable selection performance.

In these existing penalized Cox's model with group structures, only the group lasso formulation has been considered because the group lasso penalty is convex for relatively straightforward optimization solutions. However, the group lasso penalty has deficiencies. Namely, large penalties are imposed on large coefficients, which leads to over-shrinking of large coefficients. As a result, the estimates of model coefficients are biased. To avoid over-shrinkage, the group lasso implementations often tend to reduce the penalty levels, which in turn results in selecting many variables. With the "oracle" property in SCAD and MCP penalty, the estimations having the same limiting distribution as the true model, both the group SCAD and group MCP formulations have been studied (Wang et al. 2008; Huang et al. 2012; Breheny and Huang 2015). However, to the best of our knowledge, there is no effort to apply either the group SCAD or group MCP formulation in the Cox's model.

In this paper, we investigate and comprehensively evaluate the group lasso, the group SCAD, and the group MCP penalized Cox's models. More critically, these three group penalty formulations with different mathematical structures, we would like to derive scalable and efficient optimization algorithms and open-access packages for more general group penalized Cox's models.

The existing group lasso based Cox's model implementations have used different algorithms to solve the corresponding optimization problem. Ma et al. (2007) used a blockwise coordinate descent algorithm (Kim et al. 2006) to solve the group lasso problem. Wu and Wang (2013) used the cyclic coordinate descent algorithm and Simon et al. (2013) used Nesterov's method. More recently, a group-wise descent algorithm was implemented in the R package *grpreg*, whose *grpsurv* function for the group penalized Cox's model as an extension of the methods presented in Breheny and Huang (2015). We will focus on developing and evaluating the group-wise descent algorithm for three group penalized Cox's models for its simplicity, speed, and stability. We have tried the cyclic coordinate descent algorithm, and found it inferior in both timing and accuracy to the group-wise descent algorithm. Specifically, while the group-wise algorithm can produce exact solutions for a single group in one step, the cyclic coordinate descent algorithm requires multiple iterations to converge to the same solution that leads to a loss of efficiency. Although Nesterov's method is a more general optimization method than the group-wise descent algorithm, it appears to be empirically slower than the group-wise descent algorithm for the specific problem of optimizing the group penalized Cox's models as shown in our running time comparison. The existing group-wise descent algorithm implemented in *grpreg* requires the group-wise orthonormal condition. Specifically, it needs to do an initial orthonormalization step, which leads to a different problem that is not equivalent to the original group lasso formulation (Simon and Tibshiran 2011; Huang et al. 2012). In particular, the new problem is to apply the $l_2-$penalty on the linear predictors instead of the original coefficients. Moreover, even though we can do orthonormalization for each group to make the observed data satisfy the group-wise orthonormal condition, the group-wise orthonormal condition can be easily violated when removing a fraction of the data or perturbing the dataset in bootstrap or sub-sampling as pointed out in Yang and Zou (2015). Therefore, it is more favorable to solve the design matrices without the group-wise orthonormal condition. Our aim is to use the group-wise descent algorithm to handle the general design matrices of the three group penalized Cox's models. To achieve it, we adopt the majorization-minimization approach (Lange et al. 2000; Hunter and Lange 2004) to derive the majorizing (surrogate) function of the

objective function with closed-form expressions for a single group in gradient computation. We demonstrate that this algorithm is fast and efficient, and provide an open-access R package *grpCox*. Both simulation studies and real-world case studies provide comprehensive evaluation of our developed optimization algorithm for the three group penalized Cox's models.

The remainder of the article is organized as follows. Section 2 formulates the non-overlapping group penalized Cox's proportional hazards model. We introduce the majorization-minimization approach and group-wise descent algorithm for solving the group penalized Cox's model. Section 3 presents the extension with overlapping group penalty. Simulation results are reported in Sect. 4. The illustrations of our methods with real-world survival datasets are presented in Sect. 5. Section 6 concludes the article with discussion.

## 2 Non-overlapping groups

In this section, we present the Cox's model with non-overlapping groups of covariates as potential survival predictors, i.e. each potential predictor belongs to one and only one group. We first describe the general framework for group selection via the penalized partial likelihood of the Cox's model. We then derive the group-wise descent algorithms combining with the majorization-minimization approach for model inference.

### 2.1 Model formulation

Consider the standard survival data set of $N$ subjects represented by the triplets $\{(Y_i, X^{(i)}, \delta_i)\}_{i=1}^N$, where $Y_i$ denotes the survival time, $X^{(i)}$ a $P-$dimensional covariate vector, and $\delta_i$ the censoring indicator. With $T_i$ and $C_i$ denoting the survival time and the censoring time for subject $i$, the survival time $Y_i$ is defined by $Y_i = \min\{T_i, C_i\}$ and the censoring indicator is defined as $\delta_i = \mathbf{I}_{T_i \le C_i}$. Suppose that $P$ covariates belong to $J$ non-overlapping groups $I_j$'s such that $\{1, 2, \ldots, P\} = \cup_{j=1}^J I_j$ where the number of covariates in group $I_j$ is $p_j$ and $I_j \cap I_{j'} = \emptyset$ for $j \neq j'$. The $P-$dimensonal covariate vector for subjet $i$ is $X^{(i)} = (X_1^{(i)}, \ldots, X_J^{(i)})$, where $X_j^{(i)}$ is a $p_j-$dimensional covariate vector of the $j^{th}$ group for subject $i$. The corresponding coefficients of the covariates in the $j^{th}$ group are $\beta_j$. The standard Cox's proportional hazards model of the hazard for patient $i$ at time $t$ can be written as Cox (1972):

$$h(t|X^{(i)}) = h_0(t)\exp\big(X^{(i)}\mathbf{fi}\big) = h_0(t)\exp\big(\sum_{j=1}^J X_j^{(i)}\mathbf{fi}_j\big), \quad (1)$$

where $h_0(t)$ is the baseline hazard function.

Assume there is no ties in the observed times, and the censoring is non-informative. Let $t_1 < t_2 < \cdots < t_D$ be the distinct observed times where $D$ is the number of unique observed failures. $R_i$ is the set of indices of the subjects who are at risk at time $t_i$. The partial likelihood function is given by

$$L(\beta) = \prod_{i=1}^D \frac{\exp\big(\sum_{j=1}^J X_j^{(i)}\mathbf{fi}_j\big)}{\sum_{l\in R_i}\exp\big(\sum_{j=1}^J X_j^{(l)}\mathbf{fi}_j\big)}, \quad (2)$$

Penalization is one of the important variable selection methods, which can be applied to the Cox's model for better understanding survival predictors when $P$ is large by minimizing the penalized partial likelihood function

$$\mathcal{L}(\beta) = -\frac{1}{N}\log\big(L(\beta)\big) + P_{\lambda,\gamma}(\beta) = \ell(\beta) + P_{\lambda,\gamma}(\beta), \quad (3)$$

where

$$\ell(\beta) = -\frac{1}{N}\sum_{i=1}^D\Big[(\sum_{j=1}^J X_j^{(i)}\mathbf{fi}_j) - \log\Big(\sum_{l\in R_i}\exp(\sum_{j=1}^J X_j^{(l)}\mathbf{fi}_j)\Big)\Big],$$

and the penalty term $P_{\lambda,\gamma}(\beta)$ can take different forms.

- Group lasso (Yuan and Lin 2006): $P_\lambda(\beta) = \lambda\sum_j\sqrt{p_j}\,\|\beta_j\| = \sum_j \lambda_j\|\beta_j\|$, where $\lambda_j = \lambda\sqrt{p_j}$, $j = 1, \ldots, J$.
- Group smoothly clipped absolute deviation (SCAD) (Wang et al. 2007): $P_{\lambda,\gamma}(\beta) = \sum_j \mathrm{S}_{\lambda,\gamma}\big(\|\beta_j\|\big)$ with

$$\mathrm{S}_{\lambda,\gamma}\big(\|\beta_j\|\big) = \begin{cases} \lambda_j\|\beta_j\|, & \text{if } \|\beta_j\| \le \lambda_j, \\ \frac{\gamma\lambda_j\|\beta_j\| - 0.5(\|\beta_j\|^2+\lambda_j^2)}{\gamma-1}, & \text{if } \lambda_j < \|\beta_j\| \le \gamma\lambda_j, \\ \frac{\lambda_j^2(\gamma^2-1)}{2(\gamma-1)}, & \text{if } \|\beta_j\| > \gamma\lambda_j. \end{cases} \quad (4)$$

- Group minimax concave penalty (MCP) (Huang et al. 2012): $P_{\lambda,\gamma}(\beta) = \sum_j \mathrm{M}_{\lambda,\gamma}\big(\|\beta_j\|\big)$ with

$$\mathrm{M}_{\lambda,\gamma}\big(\|\beta_j\|\big) = \begin{cases} \lambda_j\|\beta_j\| - \frac{\|\beta_j\|^2}{2\gamma} & \text{if } \|\beta_j\| \le \gamma\lambda_j, \\ \frac{1}{2}\gamma\lambda_j^2 & \text{if } \|\beta_j\| > \gamma\lambda_j. \end{cases} \quad (5)$$

Here $\|\cdot\|$ denotes the Euclidean vector norm. We scale by a factor of $\frac{1}{N}$ for convenience.

Given the survival data, the Cox's model inference is to learn $\beta$ that minimizes the penalized partial likelihood function. Specifically,

$$\beta_{opt} = \underset{\beta}{\operatorname{argmin}}\Big[\ell(\beta) + P_{\lambda,\gamma}(\beta)\Big] = \underset{\beta}{\operatorname{argmin}}\,\mathcal{L}(\beta). \quad (6)$$

### 2.2 Majorization-minimization (MM) approach

The negative log partial likelihood $\ell(\beta)$ is convex and twice continuously differentiable. We adopt the majorization-minimization (MM) approach Lange et al. (2000), Hunter and Lange (2004) that involves majorizing the negative log partial likelihood $\ell(\beta)$. We derive the upper bound of $\ell(\beta)$ as the majorizing/surrogate objective function through its Hessian matrix.

Denote $\eta = X\beta$, then $\eta$ is a $N-$dimensional vector whose $i^{th}$ element is $\eta_i = X^{(i)}\beta$. We have

$$\ell(\eta) = -\frac{1}{N}\sum_{i=1}^{D}\left[\eta_i - \log\Big(\sum_{l\in R_i}\exp(\eta_l)\Big)\right],$$

We can calculate the first- and second-order derivatives of $\ell(\beta)$; in particular, via the chain rule: $\ell^{'}(\beta) = X\ell^{'}(\eta)$ and $\ell^{''}(\beta) = X^T\ell^{''}(\eta)X$. Let $U = \ell^{'}(\eta)$ and $H = \ell^{''}(\eta)$ denote the corresponding gradient vector and Hessian matrix, respectively. We can write

$$U_d = \frac{\partial \ell}{\partial \eta_d} = -\frac{1}{N}\left[I_d - \sum_{i\in C_d}\frac{\exp(\eta_d)}{\sum_{l\in R_i}\exp(\eta_l)}\right],$$

where $C_d$ is the set of subjects $i$'s with $t_d \geq t_i$.
For the Hessian matrix $H$:

- If $d \neq k$, then

$$H_{d,k} = -\frac{1}{N}\left[\sum_{i\in C_d}\frac{\exp(\eta_d)}{\sum_{l\in R_i}\exp(\eta_l)}\right]\left[\sum_{i\in C_k}\frac{\exp(\eta_k)}{\sum_{l\in R_i}\exp(\eta_l)}\right],$$

  where $C_k$ is the set of subjects $i$'s with $t_k \geq t_i$.
- If $d = k$, e.g. the diagonal element,

$$H_{d,d} = \frac{1}{N}\sum_{i\in C_d}\left[\frac{\exp(\eta_d)}{\sum_{l\in R_i}\exp(\eta_l)} - \frac{\exp(\eta_d)\sum_{i\in C_d}\exp(\eta_d)}{\big(\sum_{l\in R_i}\exp(\eta_l)\big)^2}\right],$$

Let $w_d = \frac{1}{\sqrt{N}}\left[\sum_{i\in C_d}\frac{\exp(\eta_d)}{\sum_{l\in R_i}\exp(\eta_l)}\right]$, then $-H_{d,k} = w_d w_k$, and $H_{d,d} = \frac{1}{\sqrt{N}}w_d - w_d^2$.
Let $z^* = (z_1^*, z_2^*, \ldots, z_P^*)$ be a $P-$dimensional vector, and $B$ be a $P \times P$ matrix defined by $B = sX^TX$ where $s = \max_d\big(\frac{1}{\sqrt{N}}w_d\big)$. We have

$$(z^*)^T(B - \ell^{''}(\beta))z^* = (Xz^*)^T(s\mathbf{I}_N - \ell^{''}(\eta))(Xz^*),$$

where $\mathbf{I}_N$ is a $N \times N$ identity matrix. Let $Xz^* = z = (z_1, z_2, \ldots, z_N)$ be a $N-$dimensional vector, then

$$(z^*)^T(B - \ell^{''}(\beta))z^* = z^T(s\mathbf{I}_N - \ell^{''}(\eta))z$$

$$\begin{aligned}
&= \sum_{d=1}^{N} z_d\Big(z_d\big(s - H_{d,d}\big) + \sum_{k\neq d}^{N} z_k\big(-H_{d,k}\big)\Big)\\
&= \sum_{d=1}^{N}\big(s - H_{d,d}\big)z_d^2 + \sum_{d=1}^{N} z_d\sum_{k\neq d}^{N} z_k\big(-H_{d,k}\big)\\
&= \sum_{d=1}^{N}\big(s - H_{d,d}\big)z_d^2 + \sum_{d=1}^{N} z_d\sum_{k\neq d}^{N} z_k\big(w_d w_k\big)\\
&= \sum_{d=1}^{N}\big(s - H_{d,d}\big)z_d^2 + \sum_{d=1}^{N}(w_d z_d)\sum_{k\neq d}^{N}(w_k z_k)\\
&= \sum_{d=1}^{N}\big(s - H_{d,d} - w_d^2\big)z_d^2 + \big(\sum_{d=1}^{N} w_d z_d\big)^2\\
&\geq \sum_{d=1}^{N}\big(s - H_{d,d} - w_d^2\big)z_d^2\\
&\geq \sum_{d=1}^{N}\big(s - \frac{1}{\sqrt{N}}w_d\big)z_d^2 \geq 0
\end{aligned}$$

Therefore, $(B - \ell^{''}(\beta))$ is nonnegative definite. It is worth nothing that without loss of generality, we may standardize the covariates first, as the estimated coefficients of the covariates can always be transformed back to the original scales for the sake of interpretation. We have $B = sX^TX \approx s\big(s^{'}N\mathbf{I}_P\big) = \tau\mathbf{I}_P$, where $\tau = s^{'}N\max_d\big(\frac{1}{\sqrt{N}}w_d\big)$ and $\mathbf{I}_P$ is a $P \times P$ identity matrix. Here $s^{'} = \frac{N}{P}$, if $N \geq P$, and $\frac{P}{N}$, if $N < P$.

Let $\beta^*$ be the current solution of $\beta$, we can define the majorizing (surrogate) function of the negative log partial likelihood $\ell(\beta)$ as

$$\mathcal{M}(\beta|\beta^*) = \ell(\beta^*) + \ell^{'}(\beta^*)^T(\beta - \beta^*) + \frac{\tau}{2}(\beta - \beta^*)^T(\beta - \beta^*),$$

We further write the majorizing function of the objective function for the group penalized Cox's model in (6) as

$$\begin{aligned}
\mathcal{Q}(\beta|\beta^*) &= \ell(\beta^*) + \ell^{'}(\beta^*)^T(\beta - \beta^*)\\
&\quad + \frac{\tau}{2}(\beta - \beta^*)^T(\beta - \beta^*) + P_{\lambda,\gamma}(\beta).
\end{aligned} \tag{7}$$

### 2.3 Group-wise descent algorithm

Now the estimator based on the majorizing function is defined as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ \mathcal{Q}\big(\beta|\beta^*\big). \tag{8}$$

The asymptotic properties of this estimator have been investigated with the corresponding theorem and proofs given in "Appendix" 1. Here we focus on the optimization algorithm.

To solve the minimization problem, we use the group-wise descent algorithm. This algorithm is essentially the same as the algorithm in Yuan and Lin (2006) though we solve for general design matrices of the Cox's model. The idea behind it is that the algorithm optimizes the objective function with respect to a single group at a time, iteratively cycling through all groups until convergence conditions are satisfied. The overall structure of the group-wise descent algorithm is shown in Algorithm 1. In this algorithm, $\beta^*$ refers to the current value of the Cox's model coefficients while $\hat{\beta}_j$, $\hat{\beta}$ are the updated values. This algorithm is suitable for fitting group lasso, group SCAD, and group MCP models since all three have closed-form expressions for a single-group update $\hat{\beta}_j$. These three group models have different mathematical formulations, so the closed-form expressions of a single-group updates for three models are different. The following parts present the derivations of $\hat{\beta}_j$ for three models. We prove that the algorithm possesses the descent property. Furthermore, we employ techniques to speed up the implementations of the corresponding algorithms considerably. Let us begin with the group lasso.

**Algorithm 1** Group-wise descent algorithm for the group penalized Cox's model.

Initialize $\beta^*$
**repeat**
  **for** $j = 1, 2, \ldots, J$ **do**
    Update $\hat{\beta}_j$ according to (10) for group lasso, (11) for group SCAD, or (12) for group MCP
  **end**
  Update $\beta^* = \hat{\beta}$
**until** Convergence of $\beta^*$ ;

#### 2.3.1 Group lasso

The majorizing function (7) for the group lasso Cox's model can be written as

$$\mathcal{Q}(\beta|\beta^*) = \ell(\beta^*) + \ell^{'}(\beta^*)^T(\beta - \beta^*) + \frac{\tau}{2}(\beta - \beta^*)^T(\beta - \beta^*) + \sum_j \lambda_j \|\beta_j\|,$$

Let $\mathcal{Q}_j^{'}(\beta)$ be the partial derivative of $\mathcal{Q}(\beta)$ with respect to the group $j$. We have

$$\mathcal{Q}_j^{'}(\beta) = \ell_j^{'}(\beta^*) + \tau(\beta_j - \beta_j^*) + \begin{cases} \lambda_j \frac{\beta_j}{\|\beta_j\|}, & \text{if } \beta_j \neq \mathbf{0} \\ \lambda_j \|\mathbf{v}\|, & \text{if } \beta_j = \mathbf{0}. \end{cases} \quad (9)$$

where $\mathbf{v}$ is any vector satisfying $\|\mathbf{v}\| \leq 1$. Denote $\hat{\beta}_j$ is the solution to (9). It has the following closed-form expression

$$\hat{\beta}_j = \left(1 - \frac{\lambda_j}{\tau\|\mathbf{r}\|}\right)_+ \mathbf{r}, \quad (10)$$

where $\mathbf{r} = \beta_j^* - \frac{\ell_j^{'}(\beta^*)}{\tau}$ and $(x)_+ = \max\{x, 0\}$.

#### 2.3.2 Group SCAD

The majorizing function (7) for the group SCAD Cox's model can be written as

$$\mathcal{Q}(\beta) = \ell(\beta^*) + \ell^{'}(\beta^*)^T(\beta - \beta^*) + \frac{\tau}{2}(\beta - \beta^*)^T(\beta - \beta^*) + \sum_j \mathrm{S}_{\lambda,\beta}\big(\|\beta_j\|\big),$$

The optimal solution is characterized by the partial derivative equation.

– If $\|\beta_j\| \leq \lambda_j$, then

$$\begin{cases} \ell_j^{'}(\beta^*) + \tau(\beta_j - \beta_j^*) + \lambda_j \frac{\beta_j}{\|\beta_j\|} = 0, & \text{if } \beta_j \neq \mathbf{0} \\ \ell_j^{'}(\beta^*) - \tau\beta_j^* + \lambda_j\|\mathbf{v}\| = 0, & \text{if } \beta_j = \mathbf{0}. \end{cases}$$

– If $\lambda_j < \|\beta_j\| \leq \gamma\lambda_j$, then

$$\ell_j^{'}(\beta^*) + \tau(\beta_j - \beta_j^*) + \frac{\gamma\lambda_j \frac{\beta_j}{\|\beta_j\|} - \beta_j}{\gamma - 1} = 0$$

– If $\|\beta_j\| > \gamma\lambda_j$, then

$$\ell_j^{'}(\beta^*) + \tau(\beta_j - \beta_j^*) = 0$$

where $\mathbf{v}$ is any vector satisfying $\|\mathbf{v}\| \leq 1$. By solving these equations, we find the final solutions

$$\hat{\beta}_j = \begin{cases} \left(1 - \frac{\lambda_j}{\tau\|\mathbf{r}\|}\right)_+ \mathbf{r}, \text{ if } \|\mathbf{r}\| \leq (\lambda_j + \frac{\lambda_j}{\tau}), \\ \frac{\tau(\gamma-1)}{\tau(\gamma-1)-1}\left(1 - \frac{\gamma\lambda_j}{\tau(\gamma-1)\|\mathbf{r}\|}\right)\mathbf{r}, \text{ if } \begin{cases} \tau(\gamma-1) - 1 > 0, \\ (\lambda_j + \frac{\lambda_j}{\tau}) < \|\mathbf{r}\| \leq \gamma\lambda_j, \end{cases} \\ \mathbf{r}, \text{ if } \|\mathbf{r}\| > \gamma\lambda_j. \end{cases} \quad (11)$$

where $\mathbf{r} = \beta_j^* - \frac{\ell_j^{'}(\beta^*)}{\tau}$ and $(x)_+ = \max\{x, 0\}$.

#### 2.3.3 Group MCP

The majorizing function (7) for the group MCP Cox's model can be written as

$$\mathcal{Q}(\beta) = \ell(\beta^*) + \ell'(\beta^*)^T(\beta - \beta^*) + \frac{\tau}{2}(\beta - \beta^*)^T(\beta - \beta^*) + \sum_j \mathrm{M}_{\lambda,\beta}(\|\beta_j\|),$$

The optimal solution is characterized by the partial derivative equation.

- If $\|\beta_j\| \le \gamma\lambda_j$, then

$$\begin{cases} \ell_j'(\beta^*) + \tau(\beta_j - \beta_j^*) + \lambda_j \frac{\beta_j}{\|\beta_j\|} - \frac{1}{\gamma}\beta_j = 0, & \text{if } \beta_j \neq \mathbf{0} \\ \ell_j'(\beta^*) - \tau\beta_j^* + \lambda_j\|\mathbf{v}\| = 0, & \text{if } \beta_j = \mathbf{0}. \end{cases}$$

- If $\|\beta_j\| > \gamma\lambda_j$, then

$$\ell_j'(\beta^*) + \tau(\beta_j - \beta_j^*) = 0.$$

where $\mathbf{v}$ is any vector satisfying $\|\mathbf{v}\| \le 1$. By solving these equations, we find the final solutions

$$\hat{\beta}_j = \begin{cases} \frac{\tau\gamma}{\tau\gamma - 1}\left(1 - \frac{\lambda_j}{\tau\|\mathbf{r}\|}\right)_+ \mathbf{r}, & \text{if } \|\mathbf{r}\| \le \gamma\lambda_j, \tau\gamma - 1 > 0 \\ \mathbf{r}, & \text{if } \|\mathbf{r}\| > \gamma\lambda_j. \end{cases} \quad (12)$$

where $\mathbf{r} = \beta_j^* - \frac{\ell_j'(\beta^*)}{\tau}$ and $(x)_+ = \max\{x, 0\}$.

### 2.4 The descent property of group-wise descent algorithm

The surrogate function $\mathcal{Q}$ have two properties

$$\mathcal{Q}(\beta_j^*|\beta^*) = \mathcal{L}(\beta_j^*),$$
$$\mathcal{Q}(\beta_j|\beta^*) \ge \mathcal{L}(\beta_j) \text{ for all } \beta_j.$$

From that we can prove the descent property of the group-wise descent algorithm. The descent property is stated as follows. At every iteration of the proposed group-wise descent algorithms, let $\beta^*$ and $\hat{\beta}$ denote the current value and the updated value of the coefficient estimator, respectively. Then the value of the objective function $\mathcal{L}(\beta)$ decreases, i.e., $\mathcal{L}(\hat{\beta}) \le \mathcal{L}(\beta^*)$.

***Proof*** From the second property of the surrogate function $\mathcal{Q}$ we have $\mathcal{L}(\hat{\beta}_j) \le \mathcal{Q}(\hat{\beta}_j)$. In addition, according to (8) we have $\mathcal{Q}(\hat{\beta}_j) \le \mathcal{Q}(\beta_j^*)$. Therefore, $\mathcal{L}(\hat{\beta}_j) \le \mathcal{Q}(\beta_j^*) = \mathcal{L}(\beta_j^*)$, which justifies the descent property of the group-wise descent algorithm. In other words, the objective function decreases after updating all groups in a cycle. □

**Lemma 1** *The objective function $\mathcal{Q}(\beta_j)$ is strictly convex with respect to $\beta_j$ for the group lasso with $\tau > 0$, for the group SCAD with $\tau(\gamma - 1) > 1$, and for the group MCP with $\tau\gamma > 1$.*

***Proof*** Although $\mathcal{Q}(\beta_j)$ is not differentiable, it does possess twice directional derivatives everywhere. Let $\nabla_d^2\mathcal{Q}(\beta_j)$ be the second order directional derivatives along the direction $d$, and denote $\epsilon^* = \min_{\beta_j, d} \nabla_d^2\mathcal{Q}(\beta_j)$. Then, we have

- $\epsilon^* = \tau$ for group lasso
- $\epsilon^* = \tau - \frac{1}{\gamma - 1}$ for group SCAD
- $\epsilon^* = \tau - \frac{1}{\gamma}$ for group MCP.

These are positive under the conditions specified in the lemma. In other words, $\nabla_d^2\mathcal{Q}(\beta_j)$ for all $\beta_j$ and $d$, which means that the function $\mathcal{Q}(\beta_j)$ is strictly convex. □

***Remark*** The objective function for the group lasso penalty is convex, thus the descent property of the algorithm implies the unique solution. However, the objective functions for the group SCAD and group MCP penalty are sums of convex and nonconvex components, thus it is possible that the algorithms converge to a local minimum.

### 2.5 Active set updates

To improve the computational speed, we have constructed an active set $A = \{\hat{\beta}_j \neq \mathbf{0}\}$ that takes advantage of the sparsity of $\beta$. As shown in Algorithm 1, we only need to update the nonzero coefficients $\hat{\beta}_j$ in $A$ after a complete cycle has run through all the groups, i.e., when $\beta^* = 0$, $\hat{\beta}_j$ will stay zero if $\| - \frac{\ell_j'(\mathbf{0})}{\tau}\| \le \frac{\lambda_j}{\tau}$ or $\|\ell_j'(\mathbf{0})\| \le \lambda_j$; otherwise, $\hat{\beta}_j$ will be updated and stored in the active set if $\|\ell_j'(\mathbf{0})\| > \lambda_j$. Therefore, the number of updates is reduced significantly and the rate of convergence of the algorithm is improved. The algorithm will stop if another complete cycle does not change this set. Note that the active set $A$ can only become larger after each update, so the algorithm will always stop after a finite number of updates. More details of its convergence property can be found in Meir et al. (2008).

### 2.6 Pathwise solution

The above procedure is just for one fixed value of $\lambda$. However, in general, it is of interest to be able to compute the optimal solution for a range of $\lambda$ values. Thus, we aim to compute the regularization path (denoted as $\hat{\beta}(\lambda)$) where $\lambda \in [0, \infty]$. It can be shown that $\hat{\beta}(\lambda)$ turns out to be a piecewise linear, continuous function of $\lambda$ Mairal and Yu (2012). In other words, we only need to compute the solutions on the change points in this path, denoted $\lambda_{max} \ge \lambda_1 \ge \cdots \ge \lambda_{min} \ge 0$.

We can start with $\lambda_{max}$ that is any value sufficiently large for which the entire coefficients $\beta^* = 0$. Notice that when $\beta^* = 0$, $\hat{\beta}_j$ will stay zero if $\| - \ell_j^{'}(\mathbf{0})\| \leq \lambda_j = \lambda\sqrt{p_j}$. Hence, we can set

$$\lambda_{max} = \max_j \left( \frac{\| - \ell_j^{'}(\mathbf{0})\|}{\sqrt{p_j}} \right).$$

Following the suggestions made by Simon (2012), we can ignore solutions that are close to 0 and set $\lambda_{min} = \epsilon\lambda_{max}$, then, compute the solutions over $m + 1$ values defined as $\lambda_i = \lambda_{max}\left(\frac{\lambda_{min}}{\lambda_{max}}\right)^{\frac{i}{m}}$, for $i = 0, 1, \ldots, m$. We set $\epsilon = 0.05$, if $N < P$, and 0.001, if $N \geq P$. In doing this, the algorithm usually converges well because we could use the preceding solution (i.e., for $\lambda_i$) as the initial values to obtain the solution for $\lambda_{i-1}$. It is worth noting that when $N < P$ and $\lambda$ is small, the log likelihood estimates can be $\infty$. Therefore, when implementing our *grpCox* package, we terminates the regularization path if it occurs.

### 2.7 Selection of the tuning parameters

With a path of solutions, we need to select an optimal one. The natural choice is by cross validation. However, the partial likelihood of the Cox's model is not as well defined as the Gaussian log likelihood or any exponential family on the left out samples using the traditional cross-validation, which leads to poor results. To tackle it, we have used the cross-validation method as described in Verweij and Houwelingen (1993) proposed for the Cox's model, in which data are split into $k$ parts, use $k - 1$ parts to train the model, and then, validate the learned model on the whole data set. The cross-validated log-partial likelihood for a given part $i$ and $\lambda$ is $\widehat{\mathrm{CV}}_i(\lambda) = \mathcal{L}(\hat{\beta}_{-i}) - \mathcal{L}_{-i}(\hat{\beta}_{-i})$, which can be used as the goodness-of-fit estimate of the solution. Here, $\hat{\beta}_{-i}$ and $\mathcal{L}_{-i}$ are the optimal coefficients and its corresponding log-partial likelihood for data excluding part $i$. The total goodness-of-fit, $\widehat{\mathrm{CV}}(\lambda)$, is the sum of all $\widehat{\mathrm{CV}}_i(\lambda)$. We find the optimal $\hat{\lambda}_{cvl}$ that maximizes $\widehat{\mathrm{CV}}(\lambda)$.

This method alone produces high true positive rates (TPR) but often also with high false positive rates (FPR) for group lasso. We have implemented another approach proposed in Ternes et al. (2016) to reduce FPR without significant reduction of TPR. Let $p_\lambda$ be the number of non-zero coefficients in the model for a given $\lambda$, the optimal $\lambda$ maximizes

$$\widehat{\mathrm{CV}}(\lambda) - \frac{\widehat{\mathrm{CV}}(\hat{\lambda}_{cvl}) - \widehat{\mathrm{CV}}(\lambda_{max})}{p_{\hat{\lambda}_{cvl}}} * p_\lambda, \text{ for } \lambda \in \left[\hat{\lambda}_{cvl}, \lambda_{max}\right].$$

Intuitively, it reduces the sparsity of the model $p_\lambda$ without decreasing much the goodness-of-fit of the model $\widehat{\mathrm{CV}}(.)$. The simulation studies for the second approach are presented in "Appendix" 2.

## 3 Overlapping groups

We have considered the non-overlapping group structure in the previous sections. In practice, however, a predictor can belong to several groups. For example, one gene can be shared by many different pathways. In this section, we extend the proposed methods for problems with overlapping groups. Note that the sparse group selection, which yields group-wise and within-group sparsity, can be considered as a special case of an overlapping group. Specifically, in this case, many groups would be of size 1.

Let us modify the notations and rewrite the penalty functions. Let $\mathcal{G} = \{g_1, \ldots, g_{|\mathcal{G}|}\}$ denote a set of groups as a partition of $\{1, \ldots, P\}$, $\boldsymbol{\beta}_g \in \mathcal{R}^{|g|}$ a subvector of $\boldsymbol{\beta}$, and $p_g$ the number of covariates in each group $g$. The objective function becomes

$$\mathcal{L}(\beta) = -\frac{1}{N}\log\big(L(\beta)\big) + \Omega_{\lambda,\gamma}(\beta), \tag{13}$$

where

- Overlapping group lasso: $\Omega_\lambda(\beta) = \lambda \sum_{g\in\mathcal{G}} \sqrt{p_g}\|\beta_g\|$ $= \sum_{g\in\mathcal{G}} \lambda_g\|\beta_g\|$ with $\lambda_g = \lambda\sqrt{p_g}$.
- Overlapping group smoothly clipped absolute deviation (SCAD): $\Omega_{\lambda,\gamma}(\beta) = \sum_{g\in\mathcal{G}} \mathrm{S}_{\lambda,\gamma}\big(\|\beta_g\|\big)$ with

$$\mathrm{S}_{\lambda,\gamma}\big(\|\beta_g\|\big) = \begin{cases} \lambda_g\|\beta_g\|, & \text{if } \|\beta_g\| \leq \lambda_g, \\ \frac{\gamma\lambda_g\|\beta_g\| - 0.5(\|\beta_g\|^2 + \lambda_g^2)}{\gamma - 1}, & \text{if } \lambda_g < \|\beta_g\| \leq \gamma\lambda_g, \\ \frac{\lambda_g^2(\gamma^2 - 1)}{2(\gamma - 1)}, & \text{if } \|\beta_g\| > \gamma\lambda_g. \end{cases}$$

- Overlapping group minimax concave penalty (MCP): $\Omega_{\lambda,\gamma}(\beta) = \sum_{g\in\mathcal{G}} \mathrm{M}_{\lambda,\gamma}\big(\|\beta_g\|\big)$ with

$$\mathrm{M}_{\lambda,\gamma}\big(\|\beta_g\|\big) = \begin{cases} \lambda_g\|\beta_g\| - \frac{\|\beta_g\|^2}{2\gamma} & \text{if } \|\beta_g\| \leq \gamma\lambda_g, \\ \frac{1}{2}\gamma\lambda_g^2 & \text{if } \|\beta_g\| > \gamma\lambda_g. \end{cases}$$

where $\|\cdot\|$ is the Euclidean vector norm.

Also, it is worth clarifying about how the overlapping group works. For example, consider $P = 3$ and $G = 2$, two groups sharing one covariate, and only the first group affecting the survival outcome. When the second group is not selected, all of its coefficients are shrunk to zeros. On the other hand, as the first group is selected, all of its coefficients are nonzeros. One approach, presented in Jacob et al. (2009), Obozinski et al. (2011), considered *unions* of groups: the shared covariates are selected in the final model. Another approach, presented in Jenatton et al. (2011), considered *intersections* of groups: the shared covariates are not selected. In our paper, we consider the *union* approach.

The main difficulty in solving (13) is from the non-separable $\{\beta_g\}_{g\in\mathcal{G}}$ in the non-smooth penalty $\Omega_{\lambda,\gamma}(\beta)$. The

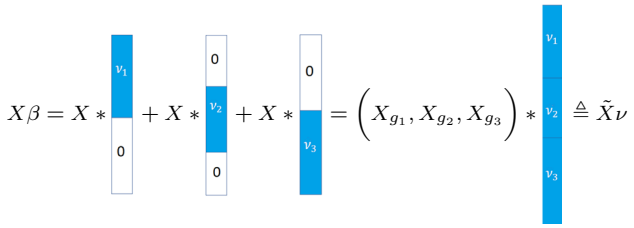

**Fig. 1** The coefficient decomposition of overlapping groups

overlapping character makes the computation of the subgradient with respect to $\boldsymbol{\beta}_g$ in the group-wise descent algorithm challenging. To tackle this problem, we have adopted the latent group approach Jacob et al. (2009), Obozinski et al. (2011) that replicates a variable in whatever group it appears; then fits the non-overlapping group models. Note that "latent" here does not imply the case that the group structure is unobservable - we consider the cases where the group structure is known in advance, which is called *predefined* group structure. Rather, "latent" implies the set of latent variables, which are formed as linear combinations of predefined groups. Next, we discuss with more details.

Let $\nu_g \in \mathcal{R}^P$ be a vector that is zero everywhere except in those positions corresponding to the elements of group $g$, and let $\mathcal{V}_g \subseteq \mathcal{R}^P$ be the subspace of these possible vectors $\nu_g$. Hence, $\boldsymbol{\beta} = \sum_{g=1}^{|\mathcal{G}|} \nu_g$. Figure 1 illustrates the idea how to transform $X\beta = \tilde{X}\nu$, where $\nu$ is the latent variable, and $\tilde{X}$ is the replicated variable matrix.

We can reformulate the objective function (13) in the latent variable space as

$$\mathcal{L}(\nu) = -\frac{1}{N}\log\big(L(\nu)\big) + \Omega_{\lambda,\gamma}(\nu), \tag{14}$$

Three penalty formulations can be similarly defined:

- Overlapping group lasso: $\Omega_\lambda(\nu) = \lambda \sum_{g\in\mathcal{G}} \sqrt{p_g}\|\nu_g\| = \sum_{g\in\mathcal{G}} \lambda_g \|\nu_g\|$ with $\lambda_g = \lambda\sqrt{p_g}$.
- Overlapping group smoothly clipped absolute deviation (SCAD): $\Omega_{\lambda,\gamma}(\nu) = \sum_{g\in\mathcal{G}} \mathrm{S}_{\lambda,\gamma}\big(\|\nu_g\|\big)$ with

$$\mathrm{S}_{\lambda,\gamma}\big(\|\nu_g\|\big) = \begin{cases} \lambda_g\|\nu_g\|, & \text{if } \|\nu_g\| \le \lambda_g, \\ \frac{\gamma\lambda_g\|\nu_g\| - 0.5(\|\nu_g\|^2+\lambda_g^2)}{\gamma-1}, & \text{if } \lambda_g < \|\nu_g\| \le \gamma\lambda_g, \\ \frac{\lambda_g^2(\gamma^2-1)}{2(\gamma-1)}, & \text{if } \|\nu_g\| > \gamma\lambda_g. \end{cases}$$

- Overlapping group minimax concave penalty (MCP): $\Omega_{\lambda,\gamma}(\nu) = \sum_{g\in\mathcal{G}} \mathrm{M}_{\lambda,\gamma}\big(\|\nu_g\|\big)$ with

$$\mathrm{M}_{\lambda,\gamma}\big(\|\nu_g\|\big) = \begin{cases} \lambda_g\|\nu_g\| - \frac{\|\nu_g\|^2}{2\gamma} & \text{if } \|\nu_g\| \le \gamma\lambda_g, \\ \frac{1}{2}\gamma\lambda_g^2 & \text{if } \|\nu_g\| > \gamma\lambda_g. \end{cases}$$

where $\|\cdot\|$ is the Euclidean vector norm.

Here, $L(\nu)$ is analogous to $L(\beta)$, but it is worth noting that $L(\beta)$ is computed in the original $\beta$ space using the design matrix $X$ while $L(\nu)$ is computed in the latent $\nu$ space using the replicated variable matrix $\tilde{X}$. In the latent (expanded and non-overlapping) space of dimension $\sum_{g\in\mathcal{G}}|g|$, the formulation has the same structure as the non-overlapping group formulations discussed previously. This allows us to apply the same solution procedure presented in the previous sections.

## 4 Simulation studies

In this section, we first show the efficiency of our proposed algorithms and package *grpCox* (Dang 2020) by comparing the running time to fit the entire path of solutions with other publicly available R packages. We also compare these packages in term of variable selection. Then, we illustrate the similarities and differences between three group regularization methods: group lasso, group SCAD, and group MCP in both the non-overlapping group and overlapping group settings. Finally, we compare the performance of three methods in terms of variable selection and model accuracy in both the non-overlapping group and overlapping group cases.

### 4.1 Setup

We generate data with $N$ observations and $P$ covariates from the following model:

$$Y^{true} = \exp\big(X\beta\big),$$

where $Y^{true}$ is the true survival time. The censoring time $C$ is generated from a exponential distribution with the mean $U\exp\big(X\beta\big)$, where $U$ is randomly generated from a uniform distribution $U(0, c)$. The recorded survival time is $Y = \min\{Y^{true}, C\}$. The observation is censored if $C < Y^{true}$. We choose different $c$ to achieve different censoring rates. The original covariates $X$ are generated from a multivariate normal distribution with a zero mean vector and the correlation matrix $\mathbf{C}$ as an autoregressive matrix where $\mathbf{C}_{ij} = \rho^{|i-j|}$ and $0 \le \rho \le 1$. The reason to use an autoregressive correlation matrix is that we could flexibly tune the correlation between covariates by setting $\rho$ values: $\rho = 0$ means no correlation between covariates, while $\rho = 1$ means that the covariates are perfectly correlated as duplicates of each other. In all the simulations, we fix $\gamma = 3.7$ for the group SCAD formulation as suggested in Fan and Li (2001), and $\gamma = 3$ for the group MCP formulation as suggested in Zhang (2010).

We evaluate the variable selection performance of these methods by presenting the model sizes, true positive rate (TPR), and false positive rate (FPR). These measures are

defined as

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \text{ and } \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}},$$

where TP, FP, FN, TN are the number of true positives, false positives, false negatives and false negatives, respectively. For all simulations, we create a path of 50 $\lambda$ values, apply 10-fold cross-validation described above to select the optimal $\lambda$ for variable selection.

We evaluate model accuracy by root mean square error (RMSE) that is given by

$$\text{RMSE} = \sqrt{\frac{1}{P}\sum_{p=1}^{P}(\beta_p - \hat{\beta}_p)^2}.$$

Recall that $P$ is the number of covariates.

## 4.2 Time and quality comparison with other packages

In this section, we compare the running time of our R package, *grpCox*, in which we implement our methods, with other publicly available R packages for fitting models. We also compare them in term of variable selection using TPR and FPR measurement.

### 4.2.1 Non-overlapping groups

We consider two other R packages *SGL* Simon et al. (2013) and *grpsurv*, which is a part of the *grpreg* package Breheny and Huang (2015). Note that *SGL* package is not for the overlapping group case.

We consider three high-dimensional settings $(N, P) = \{(50, 1000), (100, 3000), (150, 4500)\}$. In this set of experiments, $\beta$ is sparse including 100 nonzero elements and $(P - 100)$ zero elements. Each group includes 10 covariates, and the corresponding numbers of groups $J$ are set to 100, 300, 450. No censoring, and $\rho = 0.5$. We set $\alpha = 0$ for the group lasso penalty when implementing the *SGL* package. We compute the 50 $\lambda$ value solution paths of the group penalized Cox models for 100 independent data sets, and report the average running time. The 10-fold cross-validation is used for model selection. The results are shown in Table 1.

The running time results show that *grpCox* is faster than *grpsurv*, and both of them run much faster than *SGL*. Among different methods, group lasso is the fastest that followed by the group SCAD and group MCP. It can be explained that the upper bound for group lasso is sufficiently tight and convex, which leads to faster convergence.

From Table 1, it can be seen that the TPR values of *grpCox* are much higher than *grpsurv* while the FPR values of *grpCox* are a bit higher than *grpsurv*. In other words, *grpCox* gives better results than *grpsurv* in term of variable selection. In addition, *grpCox* is comparable with *SGL* in term of variable selection. It can be explained that both *grpCox* and *SGL* can handle general design matrices while *grpsurv* does an initial orthonormalization step, which can be easily violated when applying cross-validation to select models. Even worse, it may cause the significant differences in TPR and FPR for group SCAD and group MCP from group lasso in *grpsurv* results.

In addition, we would like to show how these methods scale with $N$ and $P$. We run simulations with $\rho = 0$, 20% censoring rate fixed and different setups for the number of subjects $N$ and the number of covariates $P$. For each $(N, P)$ pair, we solve for a path of 50 $\lambda$ values. Figure 2 shows the corresponding runtime for fixed $P$ as $N$ changes, and for fixed $N$ as $P$ changes. We can see that all three methods are scalable to both $N$ and $P$ and handle large $N$ and large $P$ well. The presented setups are with the maximum $N$ at 50000 and the maximum $P$ at 450000.

### 4.2.2 Overlapping groups

We consider one available R package *grpregOverlap* Zeng and Breheny (2016). Here, we show the running time of three high-dimensional overlapping settings with $N = 50$ samples for each. 20% censoring, and $\rho = 0.5$. Firstly, the equal group case includes $P = 802$ covariates with 100 groups of 10 covariates with two of them overlapping between two successive groups, and there are 81 nonzero covariates. Secondly, the unequal group case includes $P = 835$ covariates with 30 groups of 8 covariates with two of them overlapping between two successive groups, 30 groups of 11 covariates with three of them overlapping between two successive groups, and 40 groups of 15 covariates with five of them overlapping between two successive groups. There are 98 nonzero covariates. Lastly, the sparse case includes $P = 1000$ covariates with 100 groups of 10 covariates. There are 10 sparse groups. We also include the running time of the *SGL* package with $\alpha = 0.5$ for the sparse group case. Note that *grpregOverlap* does not include the model selection for Cox's model, so we choose not to report the TPRs and FPRs for all packages. The running time results are summarized in Table 2. It can be seen that for group lasso, *grpCox* is faster than *grpregOverlap* that followed by *SGL*. For group SCAD and group MCP, *grpCox* is faster than *grpregOverlap* in the sparse group setting, but a bit slower in the equal and unequal settings.

### 4.2.3 $N \geq P$ problems

We show that *grpCox* also can deal with large datasets by considering the running time results for three combinations of $(N, P) = \{(100, 50), (300, 100), (6000, 1000)\}$. The corresponding numbers of equal groups $J$ are set to 10, 10,

**Table 1** Comparison of *grpCox* with publicly available packages in the non-overlapping settings

| Package | Method | $\{N = 50, P = 1000\}$ | | | $\{N = 100, P = 3000\}$ | | | $\{N = 150, P = 4500\}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | time | TPR | FPR | time | TPR | FPR | time | TPR | FPR |
| grpCox | Group lasso | 0.05 | 0.50 | 0.10 | 0.15 | 0.97 | 0.15 | 0.26 | 1 | 0.15 |
| | Group SCAD | 0.30 | 0.54 | 0.10 | 0.33 | 0.99 | 0.06 | 0.52 | 1 | 0.13 |
| | Group MCP | 0.28 | 0.47 | 0.08 | 0.31 | 0.99 | 0.04 | 0.50 | 1 | 0.12 |
| grpsurv | Group lasso | 0.08 | 0.10 | 0.05 | 0.28 | 0.59 | 0.06 | 0.52 | 0.98 | 0.08 |
| | Group SCAD | 0.18 | 0.09 | 0.03 | 0.72 | 0.46 | 0.04 | 1.31 | 0.86 | 0.04 |
| | Group MCP | 0.14 | 0.01 | 0.01 | 0.48 | 0.16 | 0.01 | 0.85 | 0.56 | 0.02 |
| SGL | Group lasso | 7.55 | 0.33 | 0.06 | 38.73 | 1 | 0.10 | 87.84 | 1 | 0.10 |

The mean time, average TPRs, and average FPRs, over 100 independent data sets and a 50 $\lambda$ values path, are reported. The time is in seconds

**Table 2** Comparison of *grpCox* with publicly available packages in the overlapping and $N > P$ settings

| | Package | Method | Equal group | Unequal group | Sparse group |
|---|---|---|---|---|---|
| Overlapping | | $\mathbf{N < P}$ | $\{N = 50, P = 802\}$ | $\{N = 50, P = 835\}$ | $\{N = 50, P = 1000\}$ |
| | grpCox | Group lasso | 0.17 | 0.17 | 1.46 |
| | | Group SCAD | 0.34 | 0.30 | 1.63 |
| | | Group MCP | 0.33 | 0.31 | 1.65 |
| | grpregOverlap | Group lasso | 0.28 | 0.29 | 2.57 |
| | | Group SCAD | 0.27 | 0.26 | 2.52 |
| | | Group MCP | 0.26 | 0.26 | 2.52 |
| | SGL | Group lasso | - | - | 8.14 |
| Non-overlapping | | $\mathbf{N \geq P}$ | $\{N = 100, P = 50\}$ | $\{N = 300, P = 100\}$ | $\{N = 6000, P = 1000\}$ |
| | grpCox | Group lasso | 0.03 | 0.06 | 5.97 |
| | | Group SCAD | 0.03 | 0.06 | 5.75 |
| | | Group MCP | 0.02 | 0.06 | 5.53 |
| | grpsurv | Group lasso | 0.05 | 0.20 | 35.59 |
| | | Group SCAD | 0.03 | 0.19 | 16.42 |
| | | Group MCP | 0.02 | 0.11 | 15.71 |
| | SGL | Group lasso | 2.06 | 9.84 | - |

The mean time, over 100 independent data sets and a 50 $\lambda$ values path, is reported in seconds

100. In this set of experiments, $\beta$ is sparse with $P/10$ elements are nonzero. We set $\alpha = 0$ for the group lasso penalty when implementing the *SGL* package. We compute the 50 $\lambda$ value solution paths of the group penalized Cox models for 100 independent data sets, and report the average running time. However, we could run the *SGL* package with reasonable running time on small data sets only. The results are shown in Table 2. The results are consistent with high-dimensional cases: *grpCox* is faster than *grpsurv*, and both of them run much faster than *SGL*. However, group SCAD and group MCP are a bit faster than group lasso especially in the $(N, P) = (6000, 1000)$ case of *grpsurv* implementation that are presumably because their solution paths tend to be more sparse Breheny and Huang (2015).

*Note*: Group SCAD and MCP models depend on an additional parameter $\gamma$. In particular, small changes of $\gamma$ can lead the implementations terminate at different $\lambda$ values along the regularization path, which results in big running time changes. Here we used the fix $\gamma$ values suggested in Fan and Li (2001) for group SCAD and Zhang (2010) for group MCP that gave good results in term of variable selection and model accuracy (more details presented the following parts.) How to determine the optimal $\gamma$ value, however, definitely needs further investigation.

### 4.3 Comparison of three group penalized Cox's models

In this section, we illustrate the similarities and differences between three group regularization methods: group lasso, group SCAD, and group MCP in both the non-overlapping and overlapping group settings using simulated data.

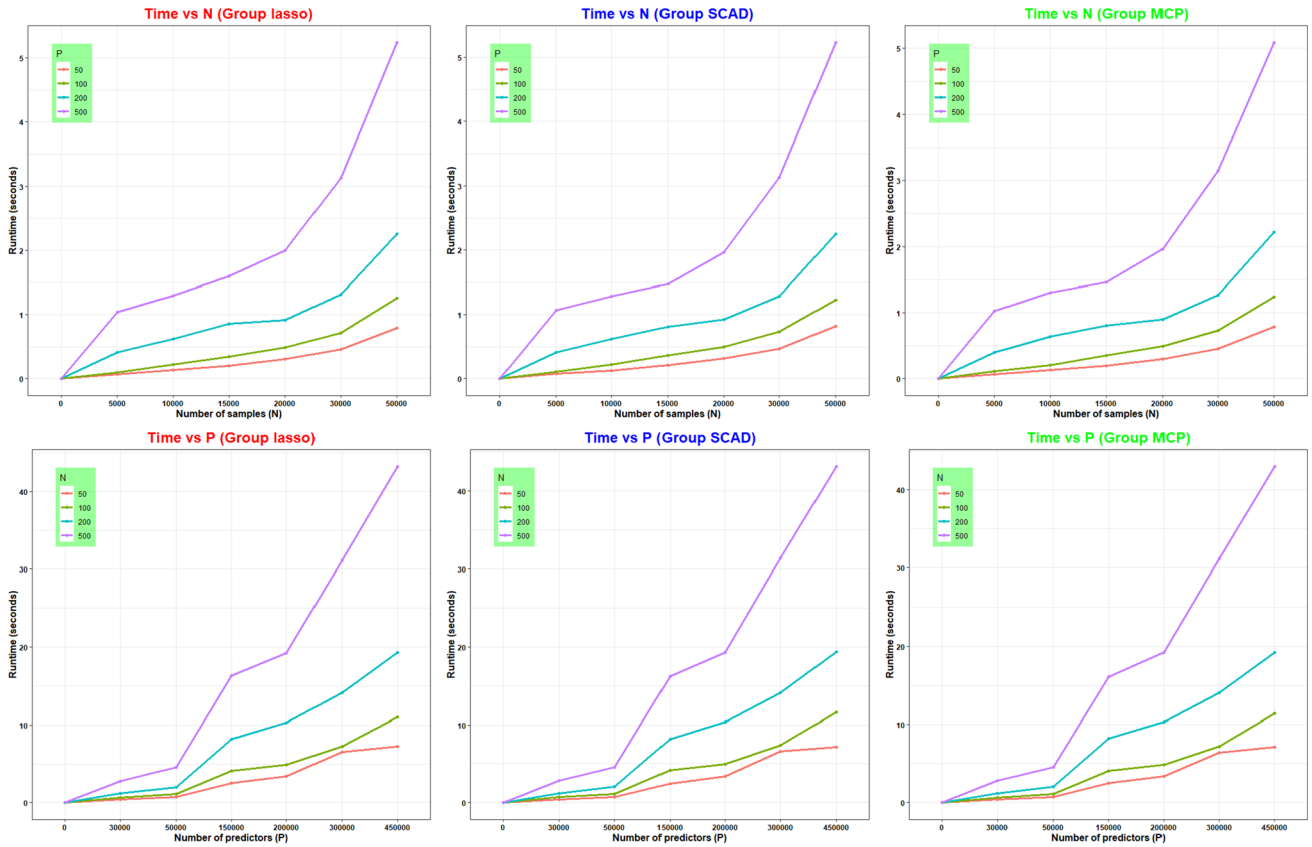


**Fig. 2** Plots of average runtime over 100 trials for 50 λ-value paths. The runtime is in seconds

#### 4.3.1 Non-overlapping groups

We consider a simple example with five primary covariates that are generated from a multivariate normal distribution with the zero mean vector and the correlation matrix $\mathbf{C}$ with $\mathbf{C}_{ij} = \rho^{|i-j|}$ and $\rho = 0.5$. The true survival time is generated as follows:

$$Y^{true} = \exp(X_1 + X_1^2 + X_1^3 - 0.7X_5 - 0.95X_5^2 - 0.8X_5^3).$$

In other words, this model includes nine covariates that can be divided into three groups: the first group is $\{X_1, X_1^2, X_1^3\}$, the second group $\{X_2, X_3, X_4\}$, and the third group $\{X_5, X_5^2, X_5^3\}$. Note that the first and third groups have nonzero coefficients while the second group has zero coefficients. The sample size $N$ is 50, and the censoring rate is 20%. We create a path of 50 values of λ. The resulting solution paths are shown in Fig. 3.

It is easy to see that the group selection selects a group of covariates in an "all-in-or-all-out" fashion. In other words, once one covariate of a group is selected, the whole group will be selected. In addition, the group SCAD and group MCP methods eliminate some of the bias towards zero among the true nonzero groups. In particular, when $\log(\lambda)$ is between -1.17 and -1.88, they produce the estimated model including only the nonzero covariates (the "oracle" model).

#### 4.3.2 Overlapping groups

We also consider a simple example with six covariates that are generated from a multivariate normal distribution with the zero mean vector and the correlation matrix $\mathbf{C}$ with $\mathbf{C}_{ij} = \rho^{|i-j|}$ and $\rho = 0.5$. There are five groups defined as $g_1 = \{X_1, X_2, X_3\}$, $g_2 = \{X_1, X_4\}$, $g_3 = \{X_2, X_4, X_5\}$, $g_4 = \{X_3, X_5\}$, $g_5 = \{X_6\}$. The true survival time is generated as follows:

$$Y^{true} = \exp(0.8X_1 + X_2 + 2X_3 + X_5).$$

The sample size $N$ is 100, and the censoring rate is 20%. We create a path of 50 λ values. The resulting solution paths are shown in Fig. 3. The results are consistent with the results of the non-overlapping group cases. The group SCAD and group MCP methods again reduce the bias towards zero among the true nonzero groups. In particular, when $\log(\lambda)$ is between -1.5 and -2.76, they produce the estimated model including only the nonzero covariates.

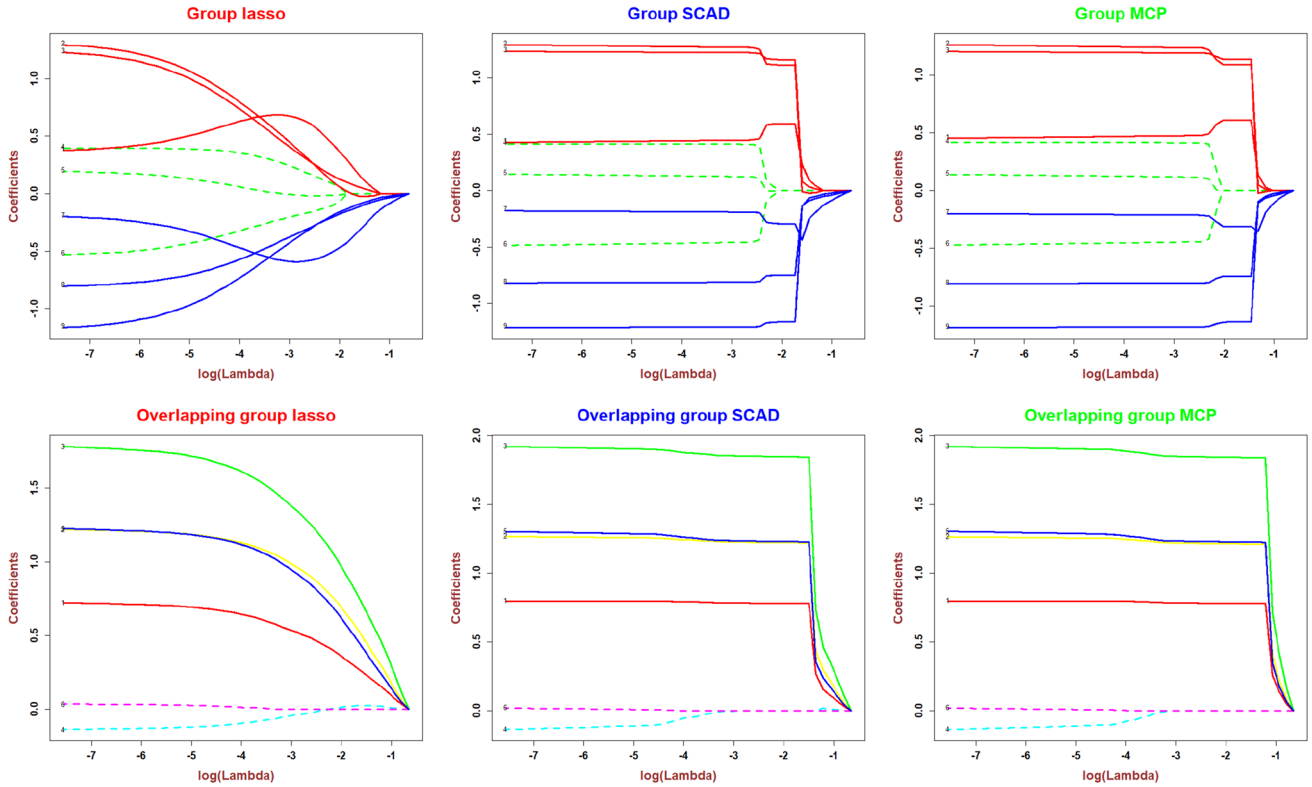


**Fig. 3** Solution paths for the group lasso, group SCAD, and group MCP models. The solid lines are for signal variables while the dashed lines are for noise variables

## 4.4 Comparison of three group penalized Cox's models with non-overlapping groups

In this section, we compare the performance of three group regularization methods in terms of variable selection and model accuracy using simulated data. In here, the model size is given in terms of the number of groups. Clearly, the true model size is the number of nonzero groups. The group size is the number of covariates of each group.

### 4.4.1 Effect of the coefficient magnitude

We focus on high dimensional cases, therefore, we generate $N = 100$ observations with $P = 400$ covariates that include 100 groups, each with 4 elements. There are five nonzero groups whose coefficient magnitudes are $\pm\beta$ where $\beta$ is a scalar, and ninety-five other groups are zero groups. We vary $|\beta|$ between 0.25 and 1.5. We also investigate the effects of the censoring setting by considering two scenarios: no censoring and right censoring with 20% censoring rate.

The results in terms of estimation accuracy and model sizes are shown in Fig. 4. The results show that when the coefficient values are small, all three methods have the same RMSE values. However, group SCAD and group MCP methods perform better with decreasing RMSE values, while the group lasso method performs increasingly poorly. Moreover, group SCAD and group MCP methods always select smaller models and approach the true model size while the group lasso method often selects too many covariates. Comparing group SCAD and group MCP, the two are nearly identical in terms of estimation accuracy. However, the group MCP method selects smaller models than the group SCAD method does.

The TPR and FPR results are summarized in Table 3. They illustrate that when the coefficients are small, group lasso does variable selection better than group SCAD and group MCP. However, group MCP begins doing better variable selection than group SCAD that produces better variable selection than group lasso.

### 4.4.2 Effect of the group size

We use the same setting as it was described previously, but the group sizes are different. We consider two different cases. In the first case, the group size is 10, and the number of groups is 40. The first two groups are nonzero groups; other groups are zero groups. These results are shown in Fig. 5 and Table 4. In the second case, the group size is 20, and the number of groups is 20. Only the first group was nonzero group; other

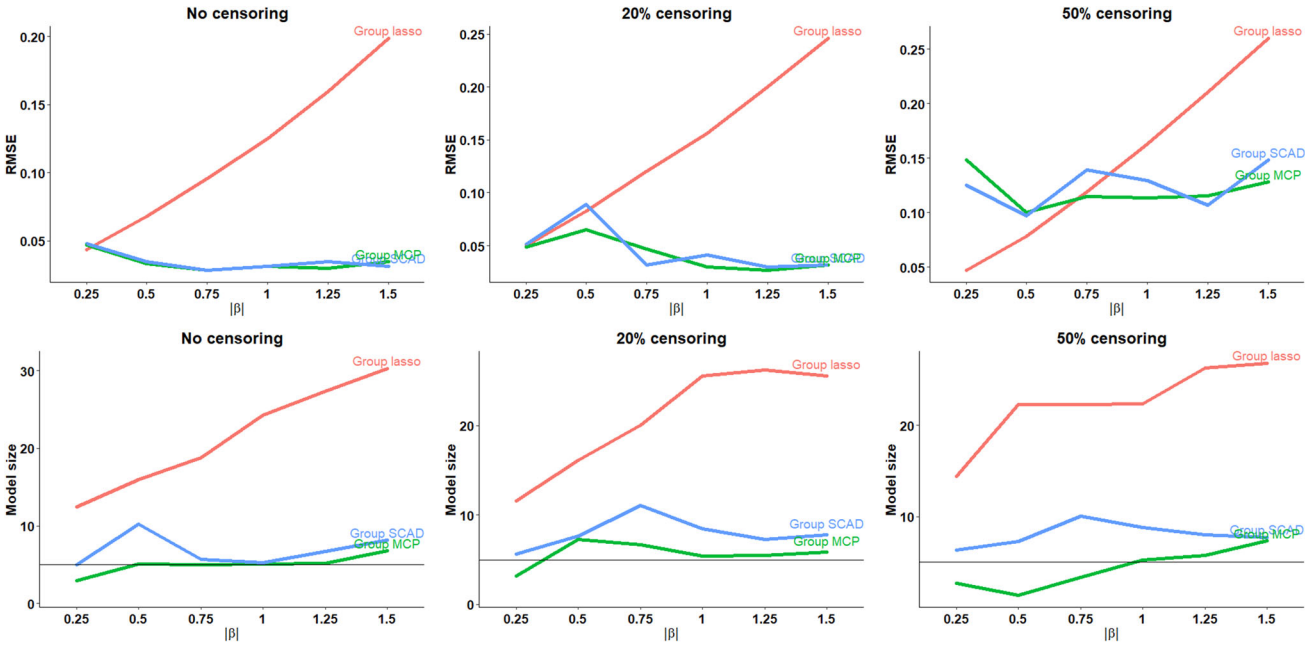


**Fig. 4** The impact of the coefficient magnitude and censoring rate on group regularization methods when the group size is 4. The *black line* is the true model size (*5*)

groups were zero groups. The results are shown in Fig. 5 and Table 4.

Figure 5 shows the same pattern as in Fig. 4. However, when the group size increases, the RMSE values decrease. Comparing Tables 3 and 4 , it can be seen that when the group size increases, group lasso performs worse with much higher FPR values. The group SCAD gives higher TPR values, but a little bit higher FPR values when the coefficient magnitude increases. The group MCP gives better performance when the group size increases.

#### 4.4.3 Effect of censoring

We investigate the performance of three methods with respect to the censoring rate. We use the same setting, in which the group size is 4 with the higher censoring rate 50%. The results are summarized in Fig. 4 and Table 3. From Fig. 4 and Table 3, on one hand, it can be seen that there is no big difference in terms of RMSE, model size, and variable selection (TPR and FPR) between no censoring and 20% censoring. On the other hand, 50% censoring affects slightly on group lasso and group SCAD, but strongly on group MCP especially when the coefficients are small. It may be explained by the fact that the presence of censoring reduces the available sample size, which leads to inconsistent estimation.

#### 4.4.4 Effect of covariate correlation

In all the above simulations, we set the population correlation $\rho = 0$. In other words, covariates are generated independently from the standard normal distribution. In this section, we still set the group size to be 4, no censoring, but the values of $\rho$ at 0.2, 0.5 and 0.9. The results are shown in Fig. 6 and Table 5. It can be seen that when the population correlation is mild, e.g. not larger than 0.5, all the three models work fine. In particular, the group MCP formulation performs the best while the group lasso performs the worst in terms of TPR and FPR values. The model with the group MCP penalty also leads to smaller models that approach the true model sizes compared to much bigger models from the group lasso model. When the population correlation is high at 0.9, all three models have bigger RMSE and smaller TPR values. The group MCP and group SCAD formulations still derive models with similar size as in the mild population correlation cases. The group lasso formulation becomes more conservative, which leads to smaller selected models whose sizes are close to the true size.

**Table 3** Average true positive rate (TPR) and false positive rate (FPR) values of three group regularization methods over 100 replications for different coefficient magnitude values and different censoring scenarios when the group size is 4

| | $\lvert\beta\rvert$ | Group lasso | | Group SCAD | | Group MCP | |
|---|---|---|---|---|---|---|---|
| | | TPR | FPR | TPR | FPR | TPR | FPR |
| No censoring | 0.25 | 0.95 | 0.09 | 0.73 | 0.02 | 0.71 | 0.00 |
| | 0.50 | 1 | 0.12 | 0.91 | 0.04 | 1 | 0.01 |
| | 0.75 | 1 | 0.15 | 1 | 0.02 | 1 | 0.01 |
| | 1.00 | 1 | 0.21 | 1 | 0.01 | 1 | 0.01 |
| | 1.25 | 1 | 0.24 | 1 | 0.03 | 1 | 0.01 |
| | 1.50 | 1 | 0.27 | 1 | 0.04 | 1 | 0.03 |
| 20% censoring | 0.25 | 0.54 | 0.09 | 0.50 | 0.04 | 0.50 | 0.01 |
| | 0.50 | 1 | 0.13 | 0.84 | 0.04 | 1 | 0.05 |
| | 0.75 | 1 | 0.17 | 1 | 0.07 | 1 | 0.03 |
| | 1.00 | 1 | 0.22 | 1 | 0.05 | 1 | 0.01 |
| | 1.25 | 1 | 0.23 | 1 | 0.03 | 1 | 0.02 |
| | 1.50 | 1 | 0.22 | 1 | 0.04 | 1 | 0.02 |
| 50% censoring | 0.25 | 0.75 | 0.12 | 0.65 | 0.04 | 0.33 | 0.01 |
| | 0.50 | 1 | 0.19 | 0.91 | 0.04 | 0.33 | 0.00 |
| | 0.75 | 1 | 0.19 | 0.93 | 0.07 | 0.66 | 0.00 |
| | 1.00 | 1 | 0.19 | 1 | 0.05 | 0.92 | 0.02 |
| | 1.25 | 1 | 0.23 | 1 | 0.04 | 0.98 | 0.02 |
| | 1.50 | 1 | 0.24 | 0.96 | 0.04 | 0.97 | 0.04 |

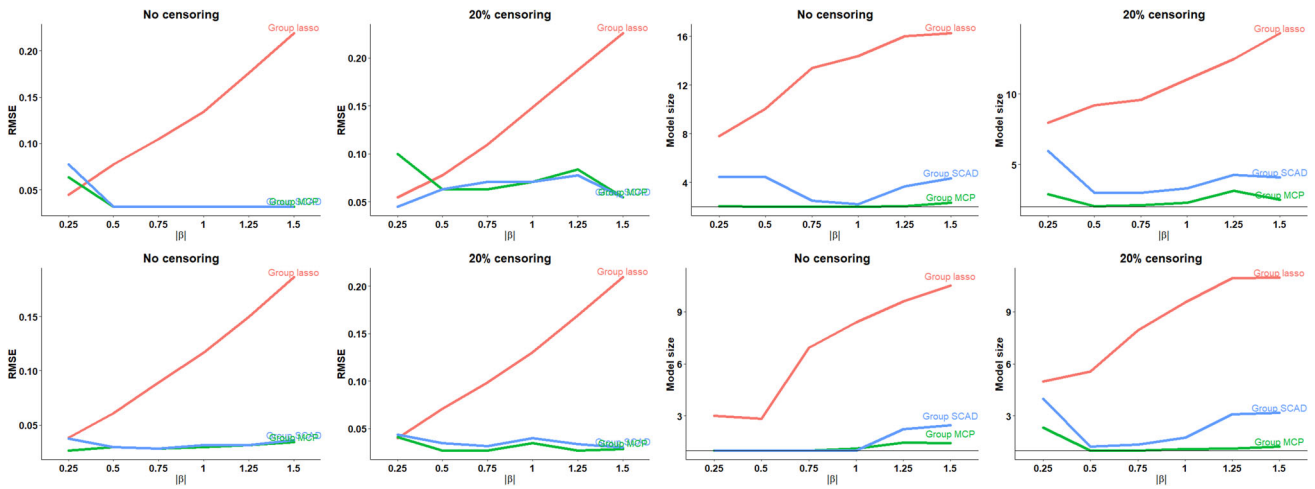


**Fig. 5** The impact of increasing coefficient magnitude on group regularization methods when the group size is 10 (first row) and 20 (second row). The *black line* on the right is the true model size

## 4.5 Comparison of three group penalized Cox's models with overlapping groups

In this section, we compare the performance of three group regularization methods in terms of variable selection and model accuracy using simulated data. In here, the model size is the number of nonzero covariates.

.

### 4.5.1 Equal group size

We generate $N = 50$ observations with $P = 162$ covariates $X_1, \ldots, X_{162}$. There are 20 groups of 10 covariates with two of them overlapping between two successive groups: $\{1, \ldots, 9, 10\}$, $\{9, \ldots, 17, 18\}, \ldots, \{153, \ldots, 162\}$. The nonzero covariates are $X_{25}, X_{26}, \ldots, X_{42}$.

**Table 4** Average true positive rate (TPR), and false positive rate (FPR) values of three group regularization methods over 100 replications for different coefficient magnitude values and different censoring scenarios

| Group size | | $\|\beta\|$ | Group lasso | | Group SCAD | | Group MCP | |
|---|---|---|---|---|---|---|---|---|
| | | | TPR | FPR | TPR | FPR | TPR | FPR |
| 10 | No censoring | 0.25 | 1 | 0.15 | 0.78 | 0.08 | 0.51 | 0.03 |
| | | 0.50 | 1 | 0.21 | 1 | 0.06 | 1 | 0.00 |
| | | 0.75 | 1 | 0.30 | 1 | 0.01 | 1 | 0.00 |
| | | 1.00 | 1 | 0.33 | 1 | 0.01 | 1 | 0.01 |
| | | 1.25 | 1 | 0.37 | 1 | 0.04 | 1 | 0.00 |
| | | 1.50 | 1 | 0.38 | 1 | 0.06 | 1 | 0.01 |
| | 20% censoring | 0.25 | 1 | 0.16 | 1 | 0.10 | 0.97 | 0.03 |
| | | 0.50 | 1 | 0.19 | 1 | 0.03 | 1 | 0.00 |
| | | 0.75 | 1 | 0.20 | 1 | 0.03 | 1 | 0.00 |
| | | 1.00 | 1 | 0.24 | 1 | 0.03 | 1 | 0.01 |
| | | 1.25 | 1 | 0.28 | 1 | 0.06 | 1 | 0.03 |
| | | 1.50 | 1 | 0.32 | 1 | 0.05 | 1 | 0.01 |
| 20 | No censoring | 0.25 | 1 | 0.10 | 1 | 0.00 | 1 | 0.00 |
| | | 0.50 | 1 | 0.10 | 1 | 0.00 | 1 | 0.00 |
| | | 0.75 | 1 | 0.31 | 1 | 0.00 | 1 | 0.00 |
| | | 1.00 | 1 | 0.39 | 1 | 0.00 | 1 | 0.01 |
| | | 1.25 | 1 | 0.45 | 1 | 0.06 | 1 | 0.02 |
| | | 1.50 | 1 | 0.50 | 1 | 0.08 | 1 | 0.02 |
| | 20% censoring | 0.25 | 1 | 0.21 | 1 | 0.16 | 1 | 0.07 |
| | | 0.50 | 1 | 0.24 | 1 | 0.01 | 1 | 0.00 |
| | | 0.75 | 1 | 0.37 | 1 | 0.02 | 1 | 0.00 |
| | | 1.00 | 1 | 0.45 | 1 | 0.04 | 1 | 0.01 |
| | | 1.25 | 1 | 0.52 | 1 | 0.11 | 1 | 0.01 |
| | | 1.50 | 1 | 0.52 | 1 | 0.12 | 1 | 0.01 |

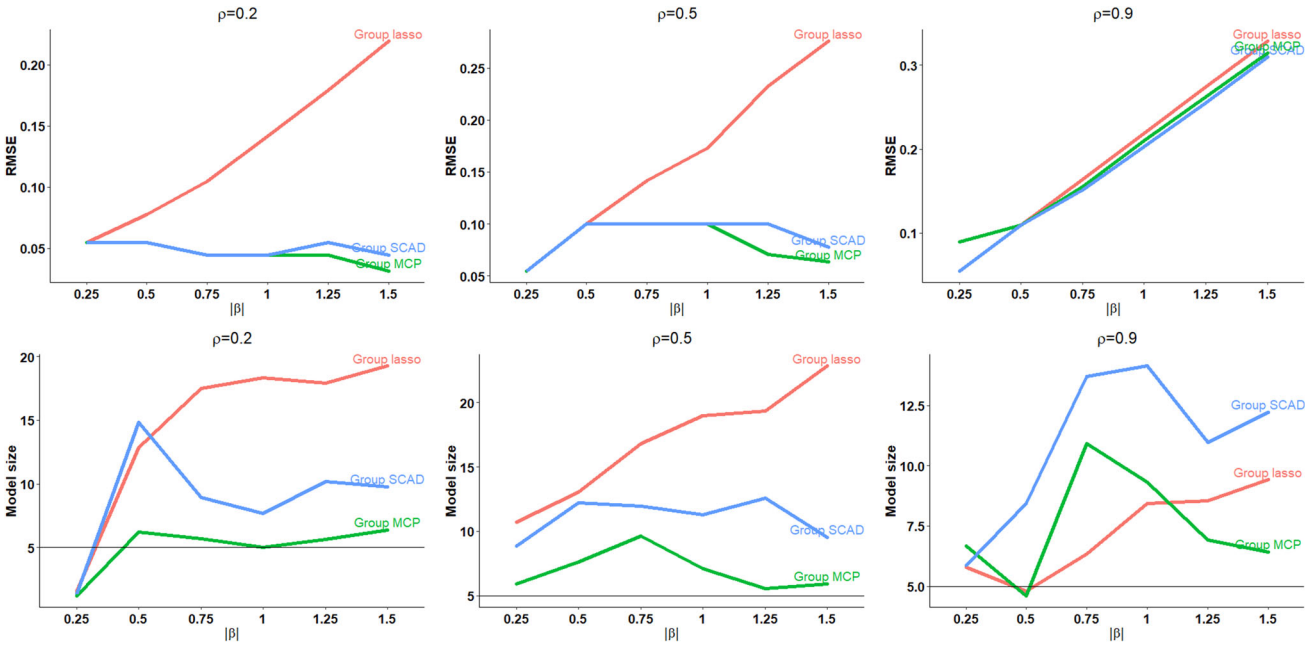


**Fig. 6** The impact of increasing the coefficient magnitude and the population correlation on group regularization methods when the size is 4. The *black line* is the true model size (*5*)

**Table 5** Average true positive rate (TPR), and false positive rate (FPR) values over 100 replications for three group regularization models with different coefficient magnitude and population correlation values when the group size is 4

| $\rho$ | $\lvert\beta\rvert$ | Group lasso | | Group SCAD | | Group MCP | |
|---|---|---|---|---|---|---|---|
| | | TPR | FPR | TPR | FPR | TPR | FPR |
| 0 | 0.25 | 1 | 0.10 | 1 | 0.00 | 1 | 0.00 |
| | 0.50 | 1 | 0.10 | 1 | 0.00 | 1 | 0.00 |
| | 0.75 | 1 | 0.31 | 1 | 0.00 | 1 | 0.00 |
| | 1.00 | 1 | 0.39 | 1 | 0.00 | 1 | 0.01 |
| | 1.25 | 1 | 0.45 | 1 | 0.06 | 1 | 0.02 |
| | 1.50 | 1 | 0.50 | 1 | 0.08 | 1 | 0.02 |
| 0.2 | 0.25 | 0.07 | 0.01 | 0.04 | 0.01 | 0.02 | 0.01 |
| | 0.50 | 1 | 0.01 | 0.99 | 0.11 | 0.99 | 0.02 |
| | 0.75 | 1 | 0.14 | 1 | 0.05 | 1 | 0.02 |
| | 1.00 | 1 | 0.15 | 1 | 0.04 | 1 | 0.01 |
| | 1.25 | 1 | 0.14 | 1 | 0.06 | 1 | 0.02 |
| | 1.50 | 1 | 0.16 | 1 | 0.06 | 1 | 0.02 |
| 0.5 | 0.47 | 1 | 0.09 | 0.52 | 0.07 | 0.47 | 0.04 |
| | 0.50 | 0.97 | 0.10 | 0.80 | 0.09 | 0.78 | 0.05 |
| | 0.75 | 1 | 0.13 | 1 | 0.08 | 1 | 0.06 |
| | 1.00 | 1 | 0.16 | 1 | 0.08 | 1 | 0.03 |
| | 1.25 | 1 | 0.16 | 1 | 0.09 | 1 | 0.02 |
| | 1.50 | 1 | 0.20 | 1 | 0.06 | 1 | 0.02 |
| 0.9 | 0.25 | 0.25 | 0.05 | 0.25 | 0.05 | 0.26 | 0.06 |
| | 0.50 | 0.50 | 0.03 | 0.49 | 0.07 | 0.27 | 0.04 |
| | 0.75 | 0.51 | 0.04 | 0.50 | 0.12 | 0.44 | 0.09 |
| | 1.00 | 0.53 | 0.07 | 0.50 | 0.13 | 0.38 | 0.08 |
| | 1.25 | 0.53 | 0.07 | 0.50 | 0.09 | 0.38 | 0.06 |
| | 1.50 | 0.51 | 0.08 | 0.51 | 0.11 | 0.27 | 0.06 |

#### 4.5.2 Effect of the number of overlapping covariates among groups

We continue considering the setting with the equal group size but set the varying number of overlapping covariates between two successive groups to 3, 4, 5, 6, 7, and 8. The results are shown in Table 6. It shows clearly that group SCAD and group MCP select smaller models with smaller RMSE values than group lasso does. In terms of variable selection performances, group SCAD and group MCP produce better results than group lasso. Overall the change of overlap covariates among groups does not affect performances by group SCAD and group MCP. On the other hand, it has strong effect upon group lasso.

#### 4.5.3 Unequal group size

We generate $N = 50$ observations with $P = 185$ covariates $X_1, \ldots, X_{185}$. There are 11 groups: 5 groups with 8 covariates per group, 10 groups with 11 covariates per group, and 6 groups with 15 covariates per group. There are two covariates overlapping between two successive groups. The nonzero covariates are $X_1, X_2, \ldots, X_{14}$.

#### 4.5.4 Sparse group example

As we mentioned above, the sparse group selection is a special case of the overlapping group. Here, we provide one example. We generate $N = 50$ observations with $P = 60$ covariates $X_1, \ldots, X_{60}$. Each covariate is treated as a group whose size is 1. In addition, there are 15 groups whose size was 4. The nonzero covariates include $X_1, X_2, X_9, X_{10}, X_{11}, X_{12}, X_{21}$. In other words, out of fifteen 4-covariate groups, there are two groups that have sparse group effects.

#### 4.5.5 Results

For all three settings above, we consider the population correlation $\rho = 0.5$ with 20% right censoring. We create a path of 50 $\lambda$ values and use 10-fold cross-validation to select the final model. The results of 100 replications are summarized in Table 7. The results in terms of TPR, FPR, model size, and RMSE values are consistent with the results of the non-overlapping group cases presented above: group SCAD and group MCP give better results in term of variable selection and model accuracy.

### 4.6 Misspecification of group structures

As described above, our methods need pre-defined group structures. We would like to investigate the effects of erroneous specification of groups. We consider an example with $N = 100$, $P = 80$, and the "correct" underlying group structure:

$$\underbrace{1,\ldots,10}_{group1}\ \underbrace{11,\ldots,20}_{group2}\ \underbrace{21,\ldots,26}_{group3}\ \underbrace{25,\ldots,30}_{group4}\ \underbrace{31,\ldots,40}_{group5}$$
$$\underbrace{41,\ldots,50}_{group6}\ \underbrace{51,\ldots,57}_{group7}\ \underbrace{55,\ldots,60}_{group8}\ \underbrace{61,\ldots,70}_{group9}\ \underbrace{71,\ldots,80}_{group10},$$

in which there are non-overlapping groups and overlapping groups. Notice that groups 3 and 4 have two overlapped covariates, and groups 7 and 8 have three overlapped covariates. We set the population correlation $\rho = 0.5$ with 50% censoring rate. The corresponding coefficients are

$$\underbrace{0,\ldots,0}_{group1-2}\ \underbrace{1.5,0,1.5,0,-2,0}_{group3}\ \underbrace{-2,0,0,-2,-1,-2}_{group4}\ \underbrace{0,\ldots,0}_{group5-6}$$
$$\underbrace{1.4,0,1,0,1.8,0,0}_{group7}\ \underbrace{1.8,0,0,1,1.6,1.2}_{group8}\ \underbrace{0,\ldots,0}_{group9-10}.$$

**Table 6** Results for overlapping group settings with different overlapping covariates between two successive groups over 100 replications

| No. of overlapping covariates | | TPR | FPR | Model size | RMSE |
|---|---|---|---|---|---|
| | Truth | | | 18 | |
| 2 | Group lasso | 1 | 0.36 | 70.58 | 0.42 |
| | Group SCAD | 1 | 0 | 18 | 0.36 |
| | Group MCP | 1 | 0 | 18 | 0.36 |
| | Truth | | | 17 | |
| 3 | Group lasso | 1 | 0.53 | 83.47 | 0.43 |
| | Group SCAD | 1 | 0 | 17 | 0.30 |
| | Group MCP | 1 | 0 | 17 | 0.30 |
| | Truth | | | 16 | |
| 4 | Group lasso | 1 | 0.50 | 70.68 | 0.47 |
| | Group SCAD | 1 | 0 | 16 | 0.29 |
| | Group MCP | 1 | 0.07 | 23.22 | 0.31 |
| | truth | | | 15 | |
| 5 | Group lasso | 1 | 0.62 | 71.35 | 0.46 |
| | Group SCAD | 1 | 0.36 | 47.4 | 0.23 |
| | Group MCP | 1 | 0.03 | 17.9 | 0.20 |
| | truth | | | 14 | |
| 6 | Group lasso | 1 | 0.63 | 59.36 | 0.58 |
| | Group SCAD | 1 | 0.06 | 19.2 | 0.26 |
| | Group MCP | 1 | 0.03 | 16.2 | 0.23 |
| | Truth | | | 13 | |
| 7 | Group lasso | 0.96 | 0.86 | 58.76 | 0.58 |
| | Group SCAD | 0.96 | 0.10 | 18.24 | 0.46 |
| | Group MCP | 0.90 | 0.08 | 15.98 | 0.40 |
| | Truth | | | 12 | |
| 8 | Group lasso | 1 | 0.81 | 41.24 | 0.61 |
| | Group SCAD | 1 | 0.35 | 24.80 | 0.35 |
| | Group MCP | 1 | 0.30 | 21.72 | 0.29 |

**Table 7** Results for overlapping group settings over 100 replications

| | | TPR | FPR | Model size | RMSE |
|---|---|---|---|---|---|
| Equal group | Truth | | | 18 | |
| | Group lasso | 1 | 0.36 | 70.58 | 0.42 |
| | Group SCAD | 1 | 0 | 18 | 0.36 |
| | Group MCP | 1 | 0 | 18 | 0.36 |
| Unequal group | Truth | | | 14 | |
| | Group lasso | 1 | 0.43 | 87.52 | 0.37 |
| | Group SCAD | 1 | 0.01 | 16.05 | 0.25 |
| | Group MCP | 1 | 0 | 14.55 | 0.25 |
| Sparse group | Truth | | | 7 | |
| | Group lasso | 1 | 0.27 | 21.61 | 0.29 |
| | Group SCAD | 1 | 0.05 | 9.56 | 0.24 |
| | Group MCP | 1 | 0.02 | 7.83 | 0.24 |

Then we consider two examples with the misspecified groups for inference. In the first example, the number of groups are incorrect because the overlapping groups are collapsed:

$$\underbrace{1,\ldots,10}_{group1}\ \underbrace{11,\ldots,20}_{group2}\ \underbrace{21,\ldots,30}_{group3}\ \underbrace{31,\ldots,40}_{group4}\ \underbrace{41,\ldots,50}_{group5}$$
$$\underbrace{51,\ldots,60}_{group6}\ \underbrace{61,\ldots,70}_{group7}\ \underbrace{71,\ldots,80}_{group8}.$$

In the second example, there are no overlapping covariates because the overlapping covariates are put into one group.

$$\underbrace{1,\ldots,10}_{group1}\ \underbrace{11,\ldots,20}_{group2}\ \underbrace{21,\ldots,26}_{group3}\ \underbrace{27,\ldots,30}_{group4}\ \underbrace{31,\ldots,40}_{group5}$$
$$\underbrace{41,\ldots,50}_{group6}\ \underbrace{51,\ldots,57}_{group7}\ \underbrace{58,\ldots,60}_{group8}\ \underbrace{61,\ldots,70}_{group9}\ \underbrace{71,\ldots,80}_{group10}.$$

The results are shown in Table 8. It can be seen that our methods are quite robust and not affected by the group structure misspecification.

We consider additional settings with a large number of overlapping covariates and the number of zero groups being more than the number of non-zero groups in "Appendix" 3.

## 5 Real-world case studies

An important motivation for developing our methods is to perform gene selection for biomarker discovery from gene expression data using the prior knowledge about group structures. We apply our methods to analyze both ovarian cancer and breast cancer data as detailed below. The grouping of genes into predefined gene sets is based on the curated database, MSigDB (MSi 2021).

### 5.1 Data

The ovarian cancer data are downloaded from The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov). It includes gene expression data for 12,043 genes in 593 samples. We first map gene probes to gene symbols and remove the duplicated genes. We use the 15 KEGG subsets of canonical pathways suggested in Jones et al. (2008). The subsets include apoptosis, cell adhesion molecules, cell cycle, base excision repair, nucleotide excision repair, mismatch repair, non-homologous end joining, Hedgehog signaling pathway, mTOR signaling pathway, Jak-STAT signaling pathway, Notch signaling pathway, Phosphatidylinositol signaling system, MAPK signaling pathway, TGF-beta signaling pathway, and Wnt signaling pathway. These gene sets include 1,347 genes in total. After removing the samples without survival information, 580 samples remain.

We use the breast cancer dataset compiled by Van de Vijer et al. (2002), which includes gene expression data for 21,463 genes in 295 breast cancer samples. Out of 295 samples there are 216 censoring samples. We first map gene probes to gene symbols and remove the duplicated genes, with the final expression data consisting of 9,950 genes. We use the gene sets from Subramanian et al. (2005) containing 427 gene sets. We restrict the analysis to the 2,663 genes that are in at least one gene set.

### 5.2 Methods

We apply our methods (group lasso, group SCAD, and group MCP) with 5-fold cross validation.

In addition, we run univariate test to select genes and pathways for evaluation. For gene-level analysis, where each gene is tested one at a time, we use the *RegParallel* function of the *RegParallel* package (Blighe and Lasky-Su 2021) with the embedded *coxph* function of the *survival* package Therneau (2021) to compute the adjusted p-values for multiple comparisons with multiple FDR and FWER methods (7 methods in total (Holm 1979; Hochberg 1988; Hommel 1988; Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001)). For pathway-level analysis, where each pathway is tested one at a time, we first convert the gene-level expression data matrix into pathway-level variables using the *GSVA* package (Hänzelmann et al. 2013), then apply the *coxph* function and compute the adjusted p-values. The significance threshold 0.05 is used to select the genes or pathways.

### 5.3 Results and discussion

#### 5.3.1 Analysis of ovarian cancer data

In univariate test, there is no gene or pathway selected using the significance level 0.05, which shows that it is often subjective relying on (adjusted) p-values for biomarker identification depending on univariate tests. This again motivates why we would like to develop our penalized survival model with different group regularization terms to consider candidate covariates together. For comparison purpose, we consider 54 genes selected based on the raw p-values at the significance 0.05 and top four pathways with the smallest p-values. Its results and the results using our methods (group lasso, group SCAD, and group MCP) are summarized in Table 9. More details about genes and pathways selected by univariate test and our methods are provided in Tables 3, 4, 5 in the Supporting Information.

First, comparing different models using *grpCox*, the results are consistent with the simulation results when group lasso selects a relatively larger model than group SCAD and group MCP do.

**Table 8** Results for misspecified group structures over 100 replications

| | | TPR | FPR | Model size | RMSE |
|---|---|---|---|---|---|
| | truth | | | 12 | |
| Correct | Group lasso | 1 | 0.70 | 59.8 | 0.45 |
| specification | Group SCAD | 1 | 0.34 | 35.5 | 0.18 |
| | Group MCP | 1 | 0.17 | 24.1 | 0.16 |
| | truth | | | 12 | |
| First | Group lasso | 1 | 0.70 | 59.8 | 0.45 |
| misspecification | Group SCAD | 1 | 0.28 | 31.3 | 0.17 |
| | Group MCP | 1 | 0.13 | 21.4 | 0.15 |
| | truth | | | 12 | |
| Second | Group lasso | 1 | 0.71 | 61.8 | 0.46 |
| misspecification | Group SCAD | 1 | 0.30 | 32.8 | 0.18 |
| | Group MCP | 1 | 0.17 | 24.2 | 0.16 |

**Table 9** Pathways and genes selected by different methods for ovarian cancer data

| Methods | | Selected pathways | No. of unique genes No. of genes | No. of selected unique genes |
|---|---|---|---|---|
| | Group lasso | KEGG_NON_HOMOLOGOUS_END_JOINING, KEGG_HEDGEHOG_SIGNALING_PATHWAY, KEGG_TGF_BETA_SIGNALING_PATHWAY, KEGG_WNT_SIGNALING_PATHWAY | 252/304 | 208 |
| grpCox | Group SCAD | KEGG_NON_HOMOLOGOUS_END_JOINING, KEGG_TGF_BETA_SIGNALING_PATHWAY, KEGG_WNT_SIGNALING_PATHWAY | 232/248 | 194 |
| | Group MCP | KEGG_NON_HOMOLOGOUS_END_JOINING, KEGG_TGF_BETA_SIGNALING_PATHWAY, KEGG_WNT_SIGNALING_PATHWAY | 232/248 | 194 |
| Univariate test | Gene-level | - | 1098/1347 | 54 |
| | Pathway-level | KEGG_BASE_EXCISION_REPAIR, KEGG_TGF_BETA_SIGNALING_PATHWAY, KEGG_WNT_SIGNALING_PATHWAY, KEGG_MISMATCH_REPAIR | 271/293 | 271 |

**Table 10** Pathways and genes selected by different methods for breast cancer data. Note that 293 selected pathways using univariate tests are listed in the Supporting Information

| Methods | | No. of selected pathways | No. of unique genes No. of genes | No. of selected unique genes |
|---|---|---|---|---|
| | Group lasso | GCM_ATM, GCM_BCL2L1, GNF2_CDH11, GNF2_CEBPA, GCM_PPP1CC, GNF2_PTX3, GNF2_TPT1, GNF2_GLTSCR2, GNF2_CYP2B6 | 289/361 | 151 |
| grpCox | Group SCAD | GCM_ATM, GNF2_CDH11, GNF2_TPT1 | 90/90 | 43 |
| | Group MCP | GNF2_CDH11 | 25/25 | 20 |
| Univariate test | Gene-level | - | 2663/42526 | 5 |
| | Pathway-level | 293 | 2197/35517 | 2197 |

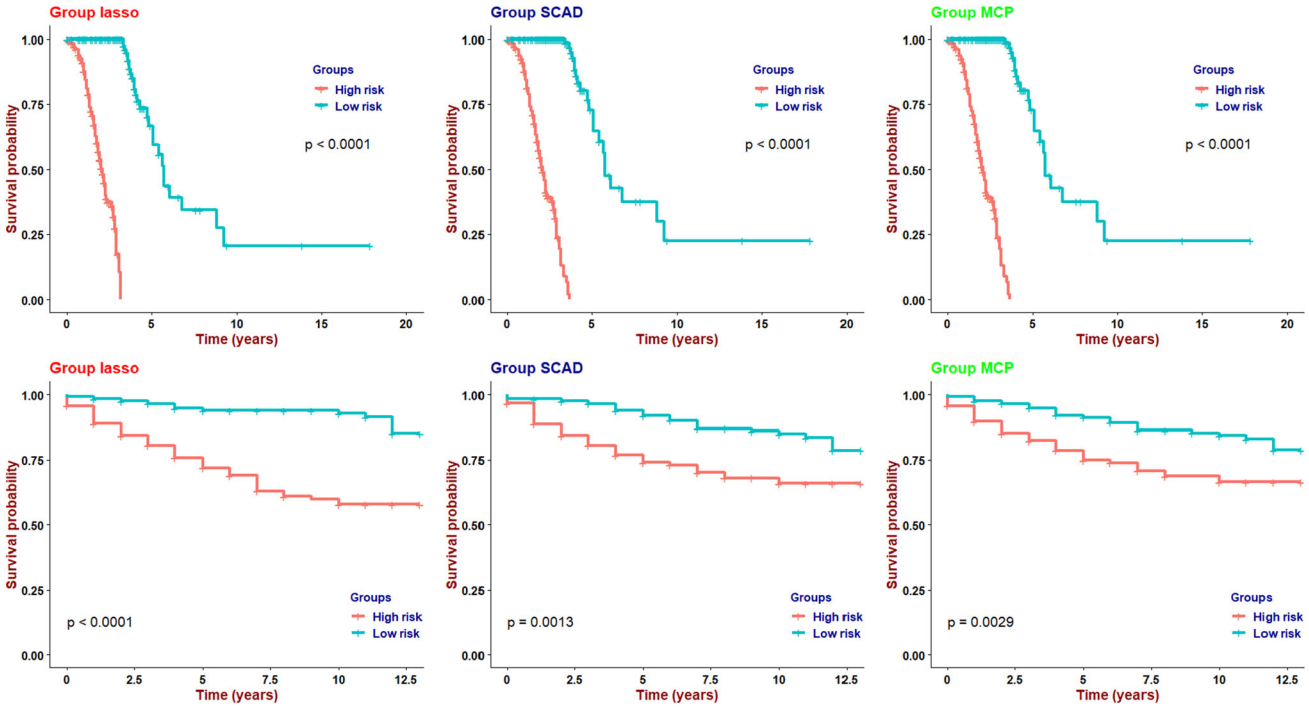


**Fig. 7** Survival curves for the high and low risk groups of the independent testing samples of ovarian cancer (first row) and breast cancer (second row)

Second, we compare the results of univariate tests and our *grpCox* package using the group lasso penalty since the results selected by group lasso include all the selections using group SCAD and MCP. At gene-level, among 15 overlapping selected genes, there are 6 genes have been reported in the literature as ovarian cancer biomarkers. Among non-overlapping genes identified by our *grpCox*, there are 38 genes showing biologically meaning. In contrast, among 54 genes by univariate tests, there are only additional 6 genes showing biological relevance.

At pathway-level, all selected pathways using our methods are biologically meaningful. The identified pathways appear to be biologically meaningful in ovarian cancer. Non homologous end joining (NHEJ) pathway is known to repair double strand breaks. Defective NHEJ has been found in up to 50% of ovarian cancers (McCormick et al. 2017; Gee et al. 2018). Overexpression or pathway activation by gene mutations among genes of the Hedgehog signaling in ovarian tumorigenesis play the crucial role in the development and progression of ovarian cancer (Szkandera et al. 2013; Otsuka et al. 1981). Wnt signaling pathway is well-known to play a role in tumorigenesis. Gatcliffe et al. (2008) demonstrated the difference in Wnt signaling pathway between normal ovarian and cancer cell lines. They also pointed out that those differences implicate that Wnt signaling leads to ovarian cancer development despite the fact that gene mutations are uncommon. TGF-$\beta$ signaling pathway behaves as both a tumor suppressor in ovarian physiology as well as acting as a tumor promoter that controls proliferation in ovarian cancer (Alsina-Sanchis et al. 2016, 2017). Two other pathways selected by univariate test are also biologically meaningful. It is clear again that considering genes together can help understand underlying cellular processes. However, the results in Table 9 show that when univariate test selects pathway, it selects all genes in this pathway, which is less flexible compared to the group penalized survival models. In fact, *grpCox* naturally takes care of the gene-pathway relationships in the model formulations and results in simultaneous selection of relevant genes and pathways. In other words, *grpCox* jointly considers potential effects, which may lead to better biomarker identification results.

### 5.3.2 Analysis of breast cancer data

Similarly, in univariate test results, very few genes, either one or five genes depending on the adopted multiple testing adjustment method, are selected. Five genes are selected with the significance level 0.05 based on the FDR and Benjamini-Hochberg correction. There are 293 pathways out of 427 pathways are selected. Its results and the results using our methods (group lasso, group SCAD, and group MCP) are summarized in Table 10. More details about genes and pathways selected by univariate tests and our methods are provided in Tables 6, 7, 8, 9 in the Supporting Information.

Similarly, the results of different models using *grpCox* are consistent with the simulation results when group lasso

selects a relatively larger model than group SCAD and group MCP do.

Next, we compare the results of univariate tests and *grpCox* package. At gene-level, among three overlapping selected genes, *TBCB* gene has been reported as breast cancer biomarker. Among non-overlapping genes using *grpCox*, there are 33 genes showing biological relevance. Among non-overlapping genes by univariate test, the other selected genes have not been reported to be relevant to breast cancer specifically.

At pathway-level, there are three overlapping pathways in which GCM_ATM and GCM_PPP1CC pathways all being biologically relevant. For example, the gene *ATM* in the GCM_ATM pathway associated with increased breast cancer risk (Ahmed and Rahman 2006; Goldgar et al. 2011). In addition, the gene *CDH11* in the GNF2_CDH11 pathway has been found to be overexpressed in breast cancer (Sarrio et al. 2008; Li et al. 2014; Assefnia et al. 2014). The collagen genes *COL1A2, COL3A1, COL6A1* are correlated significantly during breast cancer development and progression (Sengupta et al. 2003; Loss et al. 2010; Brisson et al. 2015; Xiong et al. 2014; Bertucci et al. 2002; Lin et al. 2018).

All non-overlapping pathways using *grpCox* are biologically meaningful. Among 290 non-overlapping pathways using univariate test, consider top 6 pathways with smallest adjusted p-values, there are five among them showing biological relevance. However, univariate test at pathway-level again shows less flexible when selecting relevant genes than *grpCox*.

#### 5.3.3 Validation of results

The results that are selected by our methods are further analyzed.

For the ovarian cancer data, we use the independent dataset described in Etemadmoghadam et al. (2009) as a test set. This dataset contains 285 samples and 53,433 genes. After removing the samples without survival information, there are 276 samples in total. We first compute the estimated coefficients $\hat{\beta}$, and the risk scores $X\hat{\beta}$. Their median value is used as the threshold for the high and low risk groups. The samples are assigned into the high and low risk groups by comparing with the threshold. The survival curves of these two groups are shown in Fig. 7. These two curves of all methods are well separated with a $p-$value of the log-rank test is smaller than 0.0001.

For the breast cancer data, we use the independent dataset described in Miller et al. (2005) as a test set. This dataset contains 251 samples and 24,712 genes. After removing the samples without survival information and selecting genes appearing in the selected genes in Table 10, there are 236 samples with 181 censoring samples. We first compute the estimated coefficients $\hat{\beta}$, and the risk scores $X\hat{\beta}$. Their median value is used as the threshold for the high and low risk groups. The samples are assigned into the high and low risk groups by comparing with the threshold. The survival curves of these two groups are shown in Fig. 7. It shows that the $p-$values of the log-rank tests for three models are much smaller than 0.01: the $p-$value of the group lasso is the smallest, followed by the group SCAD and the group MCP. In other words, the selected genes sets of group SCAD and group MCP are much smaller than the selected genes set of group lasso, and still classify the patients in independent breast cancer dataset into high risk and low risk groups well.

## 6 Discussion

The high-dimensional problems for survival data, in which $P$ exceeds $N$, are increasingly common thanks to our advancing data collection and storage capability. Introducing the additional structures into these problems especially group structures, is natural for incorporating prior knowledge to achieve robust and interpretable survival models. This paper has presented three group selection methods for high-dimensional data with censoring in the framework of the Cox's proportional hazards model. The proposed methods have been demonstrated in solving problems of both non-overlapping group and overlapping group cases. Since the sparse group lasso that can yield both individual and group sparsity is a special case of overlapping group lasso, our methods can effectively select important groups as well as identifying the important covariates within the selected groups.

The group-wise descent algorithms combining with the MM approach have been developed to solve the corresponding optimization problems. Thanks to the MM approach, the proposed algorithms have a proven descent property. Several computational tricks have been implemented to speed up the group-wise descent algorithms, including the screening, active set, and warm-start approaches. An open-access implementation can be found in our R package *grpCox*. Our experiments have demonstrated that *grpCox* is faster than *grpsurv*, *grpregOverlap* and much faster than *SGL*. In addition, *grpCox* is better than *grpsurv* and comparable with *SGL* in term of variable selection.

We have studied the group lasso, group SCAD, and group MCP Cox's models. These methods perform well in several simulation settings. The group lasso enjoys its convexity but it tends to select a model that is more complicated than the underlying model. It leads to relatively high false positive group selection rates. On the other hand, the nonconvex penalties, including group SCAD and group MCP, show the promising grouped variable selection results with oracle properties. We have analyzed the TCGA ovarian cancer data and breast cancer data using available pathway infor-

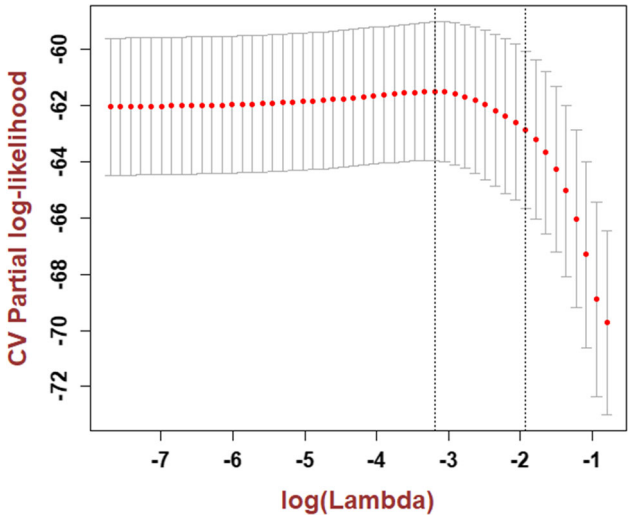


**Fig. 8** Plot of the cross-validation log-partial likelihood against the log of λ values along the regularization path

mation to construct gene groups. The selected genes have been tested on independent ovarian cancer and breast cancer datasets. The results show that the high and low risk groups are well separated. In other words, group SCAD and group MCP methods are powerful alternatives to the group lasso Cox's model for grouped variable selection. It is worth mentioning that we have used fixed $\gamma$ parameters ($\gamma = 3.7$ for group SCAD as suggested in Fan and Li (2001) and $\gamma = 3$ for group MCP as suggested in Zhang (2010)) in this paper. However, by adjusting $\gamma$ values, group MCP can resemble the group lasso with $\gamma = \infty$ and group SCAD as well. Clearly, the choice of $\gamma$ has a big impact. Therefore, how to determine the values of additional tuning parameters of group SCAD and group MCP for the Cox's model is an important research question for further investigation.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11222-021-10052-4.

**Acknowledgements** We are grateful to three anonymous reviewers for their helpful comments and constructive feedback, which help significantly improve the preliminary version of this paper. We thank Texas A&M High Performance Research Computing for providing computational resources to perform experiments in this work. This work was supported in part by the National Science Foundation (NSF)–Division of Communication & Computing Foundations (CCF) awards #1553281, #1718513, #1715027, NSF–Division of Information & Intelligent Systems (IIS) award #1812641, and the JDRF award #2-SRA-2018-513-S-B.

## Appendix 1

We have studied the statistical properties of the estimators: consistency and convergence rate as follows.

The partial likelihood

$$\ell_n(\beta) = -\frac{1}{n}\sum_{i=1}^{D}\left[(\sum_{j=1}^{J} X_j^{(i)}\mathbf{fi}_j) - \log\Big(\sum_{l\in R_i}\exp(\sum_{j=1}^{J} X_j^{(l)}\mathbf{fi}_j)\Big)\right],$$

where the penalty term $P_{\lambda,\gamma}(\beta)$ can be denoted as $P_{\lambda_n}(\beta)$ since $\gamma$ for group SCAD and group MCP are fixed. Here, $\ell_n(\beta)$, $\lambda_n$ denote the partial likelihood and tuning parameter changing with the sample size $n$, respectively.

Let the true parameter be $\beta_0 = \left(\beta_{01}^T, \beta_{02}^T\right)^T$ where $\beta_{01}$ consists of all nonzero groups and $\beta_{02}$ consists of all remaining zero groups. The objective function is

$$\mathcal{Q}_n(\beta, \lambda_n) = \ell_n(\beta_0) + \ell_n^{'}(\beta_0)^T(\beta - \beta_0) + \frac{\tau}{2}(\beta - \beta_0)^T(\beta - \beta_0) + P_{\lambda_n}(\beta).$$

Correspondingly, the minimizer of $\mathcal{Q}_n(\beta, \lambda_n)$ is $\beta_n = \left(\beta_{n1}^T, \beta_{n2}^T\right)^T$ where $\beta_n = \underset{\beta}{\operatorname{argmin}}\ \mathcal{Q}_n(\beta, \lambda_n)$.

Define $a_n = \max\{P_{\lambda_n}^{'}(\|\beta_{j0}\|) : \|\beta_{j0}\| \neq 0\}$ and $b_n = \max\{P_{\lambda_n}^{''}(\|\beta_{j0}\|) : \|\beta_{j0}\| \neq 0\}$.

**Theorem 1** *(Consistency and convergence rate) If $P_{\lambda_n}(\|\beta\|)$ simultaneously satisfies two conditions: $a_n = O_p(n^{-1/2})$ and $b_n \to 0$, then $\beta_n$ is a root-n consistent estimator for $\beta_0$ with rate $n^{-1/2}$, i.e. $\|\beta_n - \beta_0\| = O_p(n^{-1/2})$.*

***Proof*** According to Theorem 3.2 in Andersen and Gill (1982) two results hold

$$-\ell^{'}(\beta_0) \xrightarrow{p} n^{-1/2}\mathcal{N}(0, \Sigma)$$
$$\ell''(\beta^*) \xrightarrow{p} n\Sigma \text{ for any random } \beta^* \xrightarrow{p} \beta_0$$

Then, $\ell''(\beta^*) = n(\Sigma + O_p(1))$,

where $\Sigma$ is the positive definite Fisher information matrix. Consider a constant ball, $B(C) = \{\beta_0 + \alpha_n\mathbf{u} : \|\mathbf{u}\| \leq C\}$ and its boundary $\partial B(C)$ where $C > 0$ and $\alpha_n = n^{-1/2} + a_n$. Therefore, $O_p(\alpha_n) = O_p(a_n) = O_p(n^{-1/2})$. To prove $\|\beta_n - \beta_0\| = O_p(n^{-1/2})$, it is sufficient to prove that for any $\epsilon > 0$, there exists a large constant $C$ such that

$$\mathrm{P}\left(\sup_{\beta\in\partial B(C)} Q_n(\beta, \lambda_n) < Q(\beta_0, \lambda_n)\right) \geq 1 - \epsilon. \tag{15}$$

This implies that with probability at least $1-\epsilon$ (or goes to 1), $Q_n(\beta, \lambda_n)$ has a local minimum in the ball $B(C)$ for a given $\lambda_n$.

Denote $D_n(\mathbf{u}) = Q_n(\beta, \lambda_n) - Q(\beta_0, \lambda_n)$, we have

$$D_n(\mathbf{u}) = \ell^{'}(\beta_0)^T(\beta - \beta_0) + \frac{\tau}{2}(\beta - \beta_0)^T(\beta - \beta_0) + P_{\lambda_n}(\beta) - P_{\lambda_n}(\beta_0) = D_1 + D_2.$$

**Table 11** Results for group lasso using different cross-validation methods to select hyperparameters over 100 replications

| Censoring rate | $\rho$ | First CV method | | | Second CV method | | |
|---|---|---|---|---|---|---|---|
| | | Model size | TPR | FPR | Model size | TPR | FPR |
| No censoring | 0 | 95.8 | 1 | 0.19 | 30 | 1 | 0.02 |
| | 0.2 | 79.5 | 1 | 0.15 | 20 | 1 | 0 |
| | 0.5 | 119.6 | 1 | 0.26 | 30 | 1 | 0.02 |
| 20% censoring | 0 | 102 | 1 | 0.21 | 33.2 | 1 | 0.03 |
| | 0.2 | 94.1 | 1 | 0.19 | 25.1 | 1 | 0.01 |
| | 0.5 | 122.6 | 1 | 0.27 | 32.2 | 1 | 0.03 |

**Table 12** Results for group SCAD using different cross-validation methods to select hyperparameters over 100 replications

| Censoring rate | $\rho$ | First CV method | | | Second CV method | | |
|---|---|---|---|---|---|---|---|
| | | Model size | TPR | FPR | Model size | TPR | FPR |
| No censoring | 0 | 20 | 1 | 0 | 20 | 1 | 0 |
| | 0.2 | 40 | 1 | 0.05 | 39 | 1 | 0.04 |
| | 0.5 | 58.1 | 1 | 0.10 | 23.1 | 1 | 0.01 |
| 20% censoring | 0 | 80 | 1 | 0.15 | 30.9 | 1 | 0.02 |
| | 0.2 | 40.4 | 1 | 0.05 | 29.7 | 1 | 0.02 |
| | 0.5 | 83.7 | 1 | 0.17 | 27.6 | 0.91 | 0.02 |

**Table 13** Results for group MCP using different cross-validation methods to select hyperparameters over 100 replications

| Censoring rate | $\rho$ | First CV method | | | Second CV method | | |
|---|---|---|---|---|---|---|---|
| | | Model size | TPR | FPR | Model size | TPR | FPR |
| No censoring | 0 | 20 | 1 | 0 | 20 | 1 | 0 |
| | 0.2 | 20 | 1 | 0 | 20 | 1 | 0 |
| | 0.5 | 20.4 | 1 | 0.00 | 19.5 | 0.98 | 0 |
| 20% censoring | 0 | 29.5 | 1 | 0.02 | 20 | 1 | 0 |
| | 0.2 | 32 | 1 | 0.03 | 20 | 1 | 0 |
| | 0.5 | 36.9 | 1 | 0.04 | 16.2 | 0.65 | 0.01 |

Consider that

$$\begin{aligned} D_1 &= \ell'(\beta_0)^T(\beta-\beta_0) + \frac{\tau}{2}(\beta-\beta_0)^T(\beta-\beta_0) \\ &= O_p(n^{-1/2})\alpha_n\mathbf{u} + \frac{\tau}{2}\alpha_n^2\mathbf{u}^T\mathbf{u} \\ &= O_p(C\alpha_n^2) + O_p(C^2\alpha_n^2). \end{aligned}$$

Consider $D_2$ using Taylor expansion, we have

$$\begin{aligned} D_2 &= P_{\lambda_n}(\beta) - P_{\lambda_n}(\beta_0) \\ &= \sum_j P'_{\lambda_n}(\|\beta_{j0}\|)(\|\beta_{j0}+\alpha_n\mathbf{u}_j\| - \|\beta_{j0}\|) + \frac{1}{2}(\|\beta_{j0}+ \\ &\alpha_n\mathbf{u}_j\| - \|\beta_{j0}\|)^T\left(P''_{\lambda_n}(\|\beta_{j0}\|)(\|\beta_{j0}+\alpha_n\mathbf{u}_j\| - \|\beta_{j0}\|)\right) \\ &\le \sum_j a_n\alpha_n\|\mathbf{u}_j\| + b_n\alpha_n^2\|\mathbf{u}_j\|^2 \\ &\le \sum_j \alpha_n^2 C + b_n\alpha_n^2C^2 = J(\alpha_n^2C + b_n\alpha_n^2C^2). \end{aligned}$$

Because $b_n \to 0$, $D_2 \to O_p(C\alpha_n^2)$. By choosing a sufficiently large $C$, $D_1$ dominates $D_2$. Thus, inequality (15) holds.

## Appendix 2

We present the simulation studies of the second cross-validation approach described in Section 2.7 to select the tuning parameters λ and evaluate its variable selection performance.

In Fig. 8, each dot represents the logarithm of the λ values along the solution path, and the error bars provide the confidence intervals for the cross-validation log-partial-likelihood. The left vertical bar indicates the maximum cross-validation partial-log-likelihood using the first method Verweij and Houwelingen (1993) while the right one shows the maximum cross-validation log-partial-likelihood using the second method Ternes et al. (2016).

We continue considering $N = 100$ observations and $P = 400$ covariates with 40 groups, each with 10 elements. There

**Table 14** Results for misspecified group structures over 100 replications

| | | TPR | FPR | Model size | RMSE |
|---|---|---|---|---|---|
| | truth | | | 4 | |
| Correct | Group lasso | 1 | 0.70 | 40 | 0.24 |
| specification | Group SCAD | 1 | 0.52 | 31 | 0.14 |
| | Group MCP | 1 | 0.35 | 22.2 | 0.12 |
| | truth | | | 4 | |
| First | Group lasso | 1 | 0.71 | 40.5 | 0.26 |
| misspecification | Group SCAD | 1 | 0.53 | 31.2 | 0.14 |
| | Group MCP | 1 | 0.50 | 29.3 | 0.15 |
| | truth | | | 4 | |
| Second | Group lasso | 1 | 0.71 | 40.3 | 0.26 |
| misspecification | Group SCAD | 1 | 0.50 | 29 | 0.13 |
| | Group MCP | 1 | 0.40 | 25 | 0.13 |
| | truth | | | 4 | |
| Third | Group lasso | 1 | 0.70 | 40.2 | 0.26 |
| misspecification | Group SCAD | 1 | 0.50 | 29.5 | 0.13 |
| | Group MCP | 1 | 0.41 | 25.6 | 0.13 |
| | truth | | | 4 | |
| Fourth | Group lasso | 1 | 0.75 | 42.2 | 0.25 |
| misspecification | Group SCAD | 1 | 0.42 | 26 | 0.12 |
| | Group MCP | 1 | 0.35 | 21.9 | 0.12 |

are two non-zero groups. The coefficient magnitude $|\beta| = 0.5$, the values of the population correlation $\rho$ are 0, 0.2 and 0.5, the censoring rates are 0% and 20%. The results are summarized in Tables 11, 12, and 13 . It can be seen that using the second cross-validation method always results in smaller models than using the first cross-validation method. For group lasso, it produces better variable selection results with much smaller FPR values. For group SCAD and MCP, it often gives better results, but sometimes suppresses too much, e.g., in group MCP case with 20% censoring, $\rho = 0.5$. Therefore, the second cross-validation method should be used with caution.

## Appendix 3

We present additional settings based on the reviewer's suggestions: settings with a large number of overlapping covariates and the number of zero groups being more than the number of non-zero groups. More specifically, we have performed an additional experiment using the simulated data with $N = 100$, $P = 55$, in which there are 10 groups of size 10 and 50% covariates overlap between two successive groups. The "correct" underlying group structure is given by

$$\underbrace{1,\ldots,10}_{group1}\ \underbrace{6,\ldots,15}_{group2}\ \underbrace{11,\ldots,20}_{group3}\ \underbrace{16,\ldots,25}_{group4}\ \underbrace{21,\ldots,30}_{group5}$$

$$\underbrace{26,\ldots,35}_{group6}\ \underbrace{31,\ldots,40}_{group7}\ \underbrace{36,\ldots,45}_{group8}\ \underbrace{41,\ldots,50}_{group9}\ \underbrace{46,\ldots,55}_{group10}.$$

We set the population correlation $\rho = 0.5$ with 30% censoring rate. The corresponding coefficients are

$$\underbrace{0,\ldots,0}_{group1-2}\ \underbrace{0,0,0,0,0,1.5,0,0,-2,0}_{group3}\ \underbrace{1.5,0,0,-2,0,0,0,0,0,0}_{group4}$$
$$\underbrace{0,\ldots,0}_{group5-6}\ \underbrace{0,0,0,0,0,1.4,0,0,0,1.8}_{group7}$$
$$\underbrace{1.4,0,0,0,1.8,0,0,0,0,0}_{group8}\ \underbrace{0,\ldots,0}_{group9-10}.$$

Then we consider four setups with the misspecified group structures for inference. In the first setup, the number of groups are incorrect because the overlapping groups are collapsed as follows:

$$\underbrace{1,\ldots,10}_{group1}\ \underbrace{6,\ldots,15}_{group2}\ \underbrace{11,\ldots,25}_{group3}\ \underbrace{21,\ldots,30}_{group4}\ \underbrace{26,\ldots,35}_{group5}$$
$$\underbrace{31,\ldots,45}_{group6}\ \underbrace{41,\ldots,50}_{group7}\ \underbrace{46,\ldots,55}_{group8}.$$

In the second setup, the misspecified group structure deviates from the ground truth more significantly will all the overlapping covariates put into one group:

$$\underbrace{1, 3, 5, 7, 9, 11, 13, 15}_{group1}\ \underbrace{2, 4, \ldots, 12, 14, 16, 17, 18, 19, 20, 21, 22}_{group2}$$
$$\underbrace{16, \ldots, 25}_{group3}\ \underbrace{21, \ldots, 30}_{group4}\ \underbrace{26, \ldots, 35}_{group5}\ \underbrace{31, \ldots, 45}_{group6}\ \underbrace{41, \ldots, 50}_{group7}$$
$$\underbrace{46, \ldots, 55}_{group8}.$$

Similar as the first setup, the third and fourth setups are defined as follows:

$$\underbrace{1, \ldots, 20}_{group1}\ \underbrace{16, \ldots, 25}_{group2}\ \underbrace{21, \ldots, 30}_{group3}\ \underbrace{26, \ldots, 35}_{group4}\ \underbrace{31, \ldots, 45}_{group5}$$
$$\underbrace{41, \ldots, 50}_{group6}\ \underbrace{46, \ldots, 55}_{group7}$$

and

$$\underbrace{1, \ldots, 10}_{group1}\ \underbrace{6, \ldots, 20}_{group2}\ \underbrace{16, \ldots, 25}_{group3}\ \underbrace{21, \ldots, 30}_{group4}\ \underbrace{26, \ldots, 40}_{group5}$$
$$\underbrace{36, \ldots, 45}_{group6}\ \underbrace{41, \ldots, 50}_{group7}\ \underbrace{46, \ldots, 55}_{group8}$$

The results shown in Table 14 confirm our expectation: the setup with the collapsed groups including several non-zero (active) groups produces worse results than the cases with the collapsed groups with none or only one non-zero group. More clearly, the first setup in the table including two collapsed groups (group3 and group5), where each of them consists of two non-zero groups, has the worst variable selection performance. Both the second and third misspecification setups including only one group (group5) that is collapsed from two non-zero groups have almost the same performance, better than the first misspecification setup. The fourth mispecification setup with no misspecified group collapsed from two non-zero groups has the best performance. We hypothesize that the probability of variables being incorrectly selected increases due to the ignorance of the overlapping property of active elements in the collapsed groups and the larger group sizes of these collapsed groups. In other words, FPR increases and then corresponding RMSE increases.

## References


Ahmed, M., Rahman, N.: Atm and breast cancer susceptibility. Oncogene **25**(43), 5906–11 (2006)

Alsina-Sanchis, E., Figueras, A., Lahiguera Vidal, A., Casanovas, O., Graupera, M., Villanueva, A., Viñals, F.: The tgf pathway stimulates ovarian cancer cell proliferation by increasing igf1r levels. Int. J. Cancer **139**(8), 1894–903 (2016)

Alsina-Sanchis, E., Figueras, A., Gil-Martín, M., Pardo, B., Piulats, J.M., Martí, L., Ponce, J., Matias-Guiu, X., Vidal, A., Villanueva, A., Viñals, F.: Tgf controls ovarian cancer cell proliferation. Int. J. Mol. Sci. **18**(8) (2017)

Andersen, P.K., Gill, R.D.: Cox's regression model for counting processes: a large sample study. Ann. Stat. **10**(4), 1100–1120 (1982)

Assefnia, S., Dakshanamurthy, S., Guidry-Auvil, J.M., Hampel, C., Anastasiadis, P.Z., Kallakury, B., Uren, A., Foley, D.W., Brown, M.L., Shapiro, L., Brenner, M., Haigh, D., Byers, S.: Cadherin-11 in poor prognosis malignancies and rheumatoid arthritis: common target, common therapies. Oncotarget **5**(6), 1458–74 (2014)

Belhechmi, S., De Bin, R., Rotolo, F., Michiels, S.: Accounting for grouped predictor variables or pathways in high-dimensional penalized Cox regression models. BMC Bioinf. **21**(277) (2020)

Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B **57**, 289–300 (1995)

Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. Ann. Stat.B **29**, 1165–1188 (2001)

Bertucci, F., Nasser, V., Granjeaud, S., Eisinger, F., Adelaïde, J., Tagett, R., Loriod, B., Giaconia, A., Benziane, A., Devilard, E., Jacquemier, J., Viens, P., Nguyen, C., Birnbaum, D., Houlgatte, R.: Gene expression profiles of poor-prognosis primary breast cancer correlate with survival. Hum. Mol. Genet. **11**(8), 863–72 (2002)

Blighe, K., Lasky-Su, J.: Regparallel: Standard regression functions in r enabled for parallel processing over large data-frames (2021)

Breheny, P., Huang, J.: Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Stat. Comput. **25**, 173–187 (2015)

Brisson, B.K., Mauldin, E.A., Lei, W., Vogel, L.K., Power, A.M., Lo, A., Dopkin, D., Khanna, C., Wells, R.G., Pure, E.: Estimation of mean sojourn time in breast cancer screening using a Markov chain model of entry to and exit from preclinical detectable phase. Am. J. Pathol. **185**(5), 1471–86 (2015)

Cox, D.R.: Regression models and life-tables. J. R. Stat. Soc. B **34**(1), 187–220 (1972)

Dang, X.: grpCox: Penalized Cox model for high-dimensional data with grouped predictors. (2020) https://CRAN.R-project.org/package=grpCox, R package version 1.0-1

Etemadmoghadam, D., deFazio, A., Beroukhim, R., Mermel, C.: Integrated genome-wide dna copy number and expression analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas. Clin. Cancer Res. **15**(4), 1417–27 (2009)

Fan, J., Li, R.: Variable selection for cox's proportional hazards model and frailty model. Ann. Stat. **6**, 74–99 (2002)

Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)

Gatcliffe, T.A., Monk, B.J., Planutis, K., Holcombe, R.F.: Wnt signaling in ovarian tumorigenesis. Int. J. Gynecol. Cancer **18**(5), 954–62 (2008)

Gee, M.E., Faraahi, Z., McCormick, A., Edmondson, R.: Dna damage repair in ovarian cancer: unlocking the heterogeneity. J. Ovarian Res. **11**(50),(2018)

Goldgar, D.E., Healey, S., Dowty, J.G., Da-Silva, L., Chen, X., Spurdle, A.B., Terry, M.B., Daly, M.J., Buys, S.M., Southey, M.C., Andrulis, I., John, E.M., Khanna, K.K., Hopper, J.L., Oefner, P.J., Lakhani, S., Chenevix-Trench, G.: Rare variants in the atm gene and risk of breast cancer. Breast Cancer Res. **13**(4) (2011)

Gui, J., Li, H.: Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Biofinformatics **21**(13), 3001–3008 (2005)

Hänzelmann, S., Castelo, R., Guinney, J.: GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinf. **14**(7) (2013)

Hochberg, Y.: A sharper bonferroni procedure for multiple tests of significance. Biometrika **75**, 800–80 (1988)

Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Stat. **6**, 65–70 (1979)
Hommel, G.: A stagewise rejective multiple test procedure based on a modified bonferroni test. Biometrika **75**, 383–386 (1988)
Huang, J., Breheny, P., Ma, S.: A selective review of group selection in high-dimensional models. Stat. Sci. **27**(4), 481–499 (2012)
Hunter, D., Lange, K.: A tutorial on mm algorithms. Am. Stat. **58**(1), 30–37 (2004)
Jacob, L., Obozinski, G., Vert, J.: Group lasso with overlap and graph lasso. In International Conference on Machine Learning, Montreal, Canada, Proceedings of the 26th annual international conference on machine learning, pp. 433–440, (2009)
Jenatton, R., Mairal, G., Obozinski, G., Bach, F.: Proximal methods for hierarchical sparse coding. J. Mach. Learn. Res. **12**, 2297–2334 (2011)
Jones, S., Zhang, X., Parsons, D.W., Lin, J.C., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A., Hong, S.M., Fu, B., Lin, M.T., Calhoun, E.S., Kamiyama, M., Walter, K., Nikolskaya, T., Nikolsky, Y., Hartigan, J., Smith, D.R., Hidalgo, M., Leach, S.D., Klein, A.P., Jaffee, E.M., Goggins, M., Maitra, A., IacobuzioDonahue, C., Eshleman, J.R., Kern, S.E., Hruban, R.H., Karchin, R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V.E., Kinzler, K.W.: Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science **321**, 1801–1806 (2008)
Kim, Y., Kim, J., Kim, Y.: Blockwise sparse regression. Stat. Sin. **16**, 375–390 (2006)
Lange, K., Hunter, D., Yang, I.: Optimization transfer using surrogate objective functions (with discussion). J. Comput. Graph. Stat. **9**(1), 1–20 (2000)
Li, Y., Chao, F., Huang, B.: Hoxc8 promotes breast tumorigenesis by transcriptionally facilitating cadherin-11 expression. Oncotarget **5**(9), 2596–607 (2014)
Lin, Z., Zhu, G., Tang, D., Bu, J., Zou, J.: High expression of col6a1 correlates with poor prognosis in patients with breast cancer. Int. J. Clin. Exp. Med. **11**(11), 12157–12164 (2018)
Loss, L.A., Sadanandam, A., Durinck, S., Nautiyal, S., Flaucher, D., Carlton, V.E., Moorhead, M., Lu, Y., Gray, J.W., Faham, M., Spellman, P., Parvin, B.: Prediction of epigenetically regulated genes in breast cancer cell lines. BMC Bioinf. **11**(305) (2010)
Ma, S., Song, X., Huang, J.: Supervised group lasso with applications to microarray data analysis. BMC Bioinf. **8**, 60–76 (2007)
Mairal, J., Yu, B.: Complexity analysis of the lasso regularization path (2012)
McCormick, A., Donoghue, P., Dixon, M., O'Sullivan, R., O'Donnell, R., Murray, J., Kaufmann, A., Curtin, N., Edmondson, R.: Ovarian cancers harbour defects in non-homologous end joining resulting in resistance to rucaparib. Clin. Cancer Res. **23**(8), 2050–2060 (2017)
Meir, L., Van de Geer, S., Buhlmann, P.: The group lasso for logistic regression. J. R. Stat. Soc. Ser. B (Methodol.) **70**(1), 53–71 (2008)
Miller, L.D., Smeds, J., George, J., Vega, V.B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E.T., Bergh, J.: An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc. Natl. Acad. Sci. U.S.A. **102**(38), 13550–13555 (2005)
Molecular signatures database v7.4. (2021) https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp
Obozinski, G., Jacob, L., Vert, J.: Group lasso with overlaps: the latent group lasso approach. arXiv (2011)
Otsuka, A., de Paolis, A., Tocchini-Valentini, G.P.: Ribonuclease "xlai," an activity from xenopus laevis oocytes that excises intervening sequences from yeast transfer ribonucleic acid precursors. Mol. Cell. Biol. **1**(3), 269–280 (1981)
Park, M.Y., Hastie, T.: Penalized logistic regression for detecting gene interactions. Tech report, Stanford University, United States, Tech. Rep (2006)
Puig, A., Wiesel, A., Fleury, G., Hero, A.: Multidimensional shrinkage-thresholding operator and group lasso penalties. IEEE Signal Process. Lett. **18**, 363–366 (2011)
Sarrio, D., Rodriguez-Pinilla, S.M., Hardisson, D., Cano, A., Moreno-Bueno, G., Palacios, J.: Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype. Cancer Res. **68**(4), 989–997 (2008)
Sengupta, P.K., Smith, E.M., Kim, K., Murnane, M.J., Smith, B.D.: Dna hypermethylation near the transcription start site of collagen alpha2(i) gene occurs in both cancer cell lines and primary colorectal cancers. Can. Res. **63**, 1789–1797 (2003)
Simon, N.: Regularization paths for coxś proportional hazards model via coordinate descent. J. Stat. Softw. **39**(5), 53–66 (2012)
Simon, N., Tibshiran, R.: Standardization and the group lasso penalty. Stat. Sin. **22**, 983–1001 (2011)
Simon, N., Friedman, J., Hastie, T., Tibshirani, R.: A sparse-group lasso. J. Comput. Graph. Stat. **22**(2), 231–245 (2013)
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U.S.A. **102**(43), 15545–50 (2005)
Szkandera, J., Kiesslich, T., Haybaeck, J., Gerger, A., Pichler, M.: Hedgehog signaling pathway in ovarian cancer. Int. J. Mol. Sci. **14**(1), 1179–1196 (2013)
Ternes, N., Rotolo, F., Michiels, S.: Empirical extensions of the lasso penalty to reduce the false discovery rate in high-dimensional Cox regression models. Stat. Med. **35**(15), 2561–73 (2016)
Therneau, T.M.: A package for survival analysis in R. https://CRAN.R-project.org/package=survival, R package version 3.2-11 (2021)
Tibshirani, R.: The lasso method for variable selection in the cox model. Stat. Med. **16**(4), 385–395 (1996)
Van de Vijer, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D., Schreiber, G.J., Peterse, J.L., CW, R., Marton, M.J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T.W., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H., Bernards, R.: A gene-expression signature as a predictor of survival in breast cancer. New Engl. J. Med. **347**(25), 1999–2009 (2002)
Verweij, P.J., Houwelingen, H.C.: Cross-validation in survival analysis. Stat. Med. **12**(24), 385–395 (1993)
Wang, L., Chen, G., Li, H.: Group scad regression analysis for microarray time course gene expression data. Bioinformatics **23**(12), 1486–1494 (2007)
Wang, L., Li, H., Huang, J.: Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. J. Am. Stat. Assoc. **103**(484), 1556–1569 (2008)
Wu, T., Wang, S.: Doubly regularized cox regression for high-dimensional survival data with group structures. Stat. Interface **6**, 175–186 (2013)
Xiong, G., Deng, L., Zhu, J., Xu, R.: Prolyl-4-hydroxylase subunit 2 promotes breast cancer progression and metastasis by regulating collagen deposition. BMC Cancer **14**(1) (2014)
Yang, Y., Zou, H.: A fast unified algorithm for solving group-lasso penalize learning problems. Stat. Comput. **25**, 1129–1141 (2015)
Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B (Methodol.) **68**(1), 49–67 (2006)
Zeng, Y., Breheny, P.: Overlapping group logistic regression with applications to genetic pathway selection. Cancer Inf. **15**, 179–187 (2016)
Zhang, H., Lu, W.: Adaptive lasso for cox's proportional hazards model. Biometrika **94**(3), 691–703 (2007)

Zhang, C.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**(2), 894–942 (2010)
Zhao, P., Rocha, G., Yu, B.: The composite absolute penalties family for grouped and hierarchical variable selection. Ann. Stat. **37**(6A), 3468–3497 (2009)
Zou, H.: A note on path-based variable selection in the penalized proportional hazards model. Biometrika **95**(1), 241–247 (2008)