Statistics and Its Interface Volume 6 (2013) 175–186

Doubly regularized Cox regression for high-dimensional survival data with group structures

Tong Tong Wu*,† and Sijian Wang

The goal of this research is to integrate group structures to the Cox proportional hazards model with ultra high-dimensional predictors. By doubly regularizing the partial likelihood based on the Cox model with convex penalties, this method is able to perform group selection and within-group selection simultaneously. Compared with methods ignoring the structure information, our method yields better variable selection and more accurate prediction. The convexity of our regularized objective function makes the method numerically stable especially when the number of predictors far exceeds the number of the observations. A fast coordinate descent algorithm is exploited to avoid matrix operations and speed up the computation. Numerical experiments on simulated data demonstrate the good performance of our doubly regularized method. We analyze the TCGA ovarian cancer data with this new method.

AMS 2000 subject classifications: 62J07, 62N01, 65K10.
Keywords and phrases: Coordinate descent, Genetic pathways, Group structures, Lasso, Survival analysis.

1. INTRODUCTION

Ovarian cancer is the fifth leading cause of cancer deaths among women in the United States. In the year 2010, 21,880 new cases and 13,850 deaths were estimated in the U.S. [27]. The majority of ovarian cancers are diagnosed late as it produces no symptoms until the tumor has widespread. At late stages, ovarian cancer is difficult to treat and often fatal. Fortunately, the advancement of recent biotechnology offers the promise of precise and objective diagnostics by measuring the genomic features for each patient. If gene activities can be related to survival of ovarian cancer, more accurate and specific cancer diagnosis will be possible, which will lead to targeted treatment and improved survival rate.

This paper is motivated by the analysis of the ovarian cancer gene expression data from The Cancer Genome Atlas (TCGA) project. We aim to extend the current regularized method for high-dimensional survival data to incorporate group structures of predictors. Specifically, we would like to incorporate the pathway information of genes, which is available in the TCGA data. A pathway is a group of genes that are involved in the same biological process or have similar biological functions. Those genes are co-regulated and their expression levels are expected to be highly correlated. The pathways structures are believed to be biologically important to understand the complicated process of cancer occurrence and development [21]. As Dr. Bert Vogelstein stated in the 101st meeting of the American Association for Cancer Research on April 19, 2010, "virtually all of the 300 or so known driver genes are part of a few core pathways, ... more driver genes will be discovered ... but most of them will be infrequently mutated ... and most of the cancer genes will be members of these same core pathways". From the statistical point of view, since the genes may perform as groups rather than individuals, the statistical accuracy and efficiency may be improved by intergrading the pathway information into the analysis [36]. By analyzing the TCGA ovarian cancer data, we aim to (1) identify important core pathways and to identify important genes within those pathways related to ovarian cancer survival; (2) build a predictive model for the survival of future patients based on the identified genetic signatures.

The Cox proportional hazards model [5] is considered in the current paper to study the dependence of survival time T and predictors of high-dimensionality. For high-dimensional problems, there are two existing approaches for dimension reduction: feature selection and feature extraction [23]. Feature extraction methods, which project the original feature spaces into lower dimensional spaces, may lack good scientific interpretation and may provide no clue what to target for therapy. Feature selection methods, instead, choose a possible best subset from the original features and retain the scientific interpretation. Here, the feature selection approach will be taken to generate simple and stable models that are interpretable.

The variable selection problem has been studied extensively for the Cox proportional hazards model, such as lasso [18, 28, 31], adaptive lasso [39, 43], and SCAD [11]. These methods can automatically remove unimportant variables

*Corresponding author.
†Wu's research is supported in part by NSF Grant CCF-0926181.

by shrinking some regression coefficients to be exactly zero. [9] and [41] extended the sure screening procedure of [12] to Cox's proportional hazards model. The idea is to screen out variables with small marginal associations with survival outcomes. However, when the predictors are grouped, e.g. genes belong to the same pathway, these methods fail to integrate the grouping information and still treat variables individually. The variable selection is therefore performed based on the strength of the individual variables rather than the strength of the groups.

Recently the variable selection problem with grouped predictors has been considered by several authors. [38] and [40] introduced the group lasso and CAP methods that penalize the L_2-norm (Euclidean norm) and L_∞-norm of the coefficients within each group in linear regression, respectively. [26] applied the group lasso penalty to the Cox proportional hazards model. Based on the boosting technique, [25] and [34], respectively, developed a group additive regression model and a nonparametric pathway-based regression model to identify groups of genomic features that are related to several clinical phenotypes including the survival outcome. All these group variable selection methods have a common limitation: they select variables in an "all-in-or-all-out" fashion. In other words, when one variable in a group is selected, all other variables in the same group are also selected. Thus, these methods only conduct "group selection" but no "within-group selection", i.e, they do not select important variables within the identified groups. The reality, however, may be that, for example, some genes in a pathway may not be related to the phenotype although the pathway as a whole is involved in the biological process.

In order to achieve sparsity within groups, [20] imposed a bridge L_γ-norm penalty on coefficients within each group in linear regression. [42] and [33] proposed a hierarchical lasso penalty for group variable selection in linear regression and Cox regression, respectively. When the groups are not overlapped, the hierarchical lasso penalty is equivalent to the bridge L_γ-norm penalty with $\gamma = 0.5$. One possible drawback of these methods is that the objective functions are no longer convex, which may cause numerical problems in practice.

In this article, we introduce a doubly regularized Cox regression (DrCox) with convex penalties to achieve both group selection and within-group selection for high-dimensional survival data. The convexity of the doubly penalized objective function is more numerically stable compared to the method of [33], especially when the number of predictors far exceeds the sample size. In order to tackle the high-dimensionality of the data and nondifferentiability of the objective function, a fast computational method based on cyclic coordinate descent [13, 36] is employed to efficiently implement this new method.

The remainder of the paper is organized as follows. Section 2 formulates the doubly regularized Cox regression based on the partial likelihood. Cyclic coordinate descent algorithms are derived for parameter estimation. Sections 3 and 4 report our numerical tests of the doubly regularized Cox regression on simulated and real data. The paper is concluded with a short summary.

2. DOUBLY REGULARIZED COX REGRESSION

For the ease of representation, we first describe the general framework for variable selection via regularized partial likelihood of the Cox model. We then formulate the doubly regularized Cox regression and derive the parameter estimates via cyclic coordinate descent algorithms.

2.1 Variable selection via regularized partial likelihood

Suppose that the p variables occur in K groups. Assume the kth group has p_k variables, and we denote the variables in the kth group by $\boldsymbol{X}_{(k)} = (X_{k1}, \ldots, X_{kp_k})^T$. The corresponding regression coefficients for the kth group are $\boldsymbol{\beta}_{(k)} = (\beta_{k1}, \ldots, \beta_{kp_k})^T$. We first assume that these K groups do not overlap, i.e., each variable belongs to only one group. In Section 2.4, we consider the overlap case, i.e., variables are allowed to belong to multiple groups.

For a sample of n subjects, let T_i and C_i denote the survival time and the censoring time for subject $i = 1, \ldots, n$. The observed survival time is defined by $Y_i = \min\{T_i, C_i\}$ and the censoring indicator is $\delta_i = \boldsymbol{I}(T_i \le C_i)$. We denote $\boldsymbol{X}_{i,(k)} = (X_{i,k1}, \ldots, X_{i,kp_k})^T$ to be the p_k variables in the kth group for the ith subject and $\boldsymbol{X}_i = (\boldsymbol{X}_{i,(1)}^T, \ldots, \boldsymbol{X}_{i,(K)}^T)^T$ to be the total p variables for the ith subject. The survival time T_i and the censoring time C_i are assumed to be conditionally independent given $\boldsymbol{X}_i$. Furthermore, the censoring mechanism is assumed to be noninformative. The observed data can be represented by the triplets $\{(Y_i, \delta_i, \boldsymbol{X}_i),\ i = 1, \ldots, n\}$.

We consider the following Cox proportional hazards model [5]

$$h(t|X) = h_0(t) \exp\Big(\sum_{k=1}^{K} \sum_{j=1}^{p_k} \beta_{kj} X_{kj}\Big).$$

If the failure times are continuous, it is reasonable to assume that there are no ties in the observed times. The partial likelihood is

$$L_n(\boldsymbol{\beta}) = \prod_{i \in D} \frac{\exp\Big(\sum_{k=1}^{K} \boldsymbol{X}_{i,(k)}^T \boldsymbol{\beta}_{(k)}\Big)}{\sum_{l \in R_i} \exp\Big(\sum_{k=1}^{K} \boldsymbol{X}_{l,(k)} \boldsymbol{\beta}_{(k)}\Big)},$$

where D is the set of indices of observed failures, and R_i is the set of indices of the subjects who are at risk at time Y_i.

Let the log-partial likelihood be $\ell_n(\boldsymbol{\beta}) = \log\{L_n(\boldsymbol{\beta})\}/n$. Variable selection can then be realized by minimizing the penalized negative log-partial likelihood function

$$-\ell_n(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}),$$

where $P_\lambda(\boldsymbol{\beta})$ is a penalty function on the coefficients $\boldsymbol{\beta}$. Possible examples of $P_\lambda(\boldsymbol{\beta})$ include, but are not limited to, the lasso penalty [31]

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{k=1}^{K} \sum_{j=1}^{p_k} |\beta_{kj}|,$$

group lasso penalty [38]

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{k=1}^{K} \sqrt{\sum_{j=1}^{p_k} \beta_{kj}^2},$$

and the hierarchical lasso penalty [33]

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{k=1}^{K} \sqrt{\sum_{j=1}^{p_k} |\beta_{kj}|}.$$

The lasso penalty $\sum_{k=1}^{K}\sum_{j=1}^{p_k}|\beta_{kj}|$ is singular at the individual level, i.e., when individual parameter $\beta_{kj} = 0$. Hence, it is able to perform selection for individual variables as some estimates $\hat{\beta}_{kj}$ are exactly zero [10, 30]. The group lasso penalty $\sum_{k=1}^{K}\sqrt{\beta_{k1}^2 + \cdots + \beta_{kp_k}^2}$ performs group selection [38] as it is singular at the group level, i.e. when $\boldsymbol{\beta}_{(k)} = 0$. The group lasso penalty is more effective than the lasso penalty in removing unimportant groups. It is easy to see that the group lasso penalty $\|\boldsymbol{\beta}_{(k)}\|_2$ is no longer singular at any point if any $\beta_{kj} \neq 0$. Therefore, the group lasso penalty $\|\boldsymbol{\beta}_{(k)}\|_2$ selects a group of variables in an "all-in-or-all-out" fashion. In other words, once one variable in a group is selected, the whole group will be selected.

The hierarchical lasso penalty is singular at both the group level and the individual level, and therefore it achieves the desired sparsity at both the group level and the within-group level. However, the hierarchical lasso penalty is not convex, which may cause numerical instability, especially when $p \gg n$.

2.2 Doubly regularized Cox regression

To achieve the goal of both group and within-group variable selection and to overcome the non-convexity drawback, we penalize the log-partial likelihood by a mixture of the lasso penalty and group lasso penalty

$$(1)\quad g(\boldsymbol{\beta}) = -\ell_n(\boldsymbol{\beta}) + \lambda_1 \sum_{k=1}^{K}\sum_{j=1}^{p_k}|\beta_{kj}| + \lambda_2 \sum_{k=1}^{K}\sqrt{\sum_{j=1}^{p_k}\beta_{kj}^2}$$
$$= -\ell_n(\boldsymbol{\beta}) + \lambda_1 \sum_{k=1}^{K}\|\boldsymbol{\beta}_{(k)}\|_1 + \lambda_2 \sum_{k=1}^{K}\|\boldsymbol{\beta}_{(k)}\|_2,$$

where $\|\boldsymbol{\beta}_{(k)}\|_1 = \sum_{j=1}^{p_k}|\beta_{kj}|$, $\|\boldsymbol{\beta}_{(k)}\|_2 = \sqrt{\sum_{j=1}^{p_k}\beta_{kj}^2}$, and λ_1 and λ_2 are two nonnegative tuning constants, which control the strength of variables selection. The larger are the tuning constants, the less variables are retained in the model. The tuning constants can be determined using k-fold cross validation [18], generalized cross validation [11, 39], information based criterion like AIC or BIC [17], or independent validation set [19, 44].

Given the singularities of lasso and group lasso penalties discussed in the previous section, their combination will achieve the goal of both group and within-group selections. It is also straightforward to verify that the objective function (1) is convex and strictly convex when $\lambda_2 > 0$. Note that the double penalties in this paper are different than the penalties in elastic net [44], which uses a mixture of lasso penalty and ridge penalty (summation of squared Euclidean norms).

This type of regularization methods with a mixture of penalties were originally independently developed by [14, 36] and [24] in linear regression. [37] also use the double penalties in multicategory classification. These papers show that the mixtures of penalties perform superior to lasso penalty in linear regression and multi-class classification problem. The new doubly regularized Cox regression method inherits these good properties from its predecessors for censored data.

2.3 Cyclic coordinate descent algorithm

To tackle the high-dimensionality of the data, we exploit a cyclic coordinate descent algorithm, which has been shown to be computationally efficient [13, 15, 36]. The idea is to optimize the objective function (1) by updating parameters one by one. Variable selection is achieved as only the important parameters that can "escape" from the pressure of penalties will get updated. The avoidance of matrix operations explains its fast computing speed and numeric stability, especially for large systems. At the same time, coordinate descent is able to handle the nondifferentiability of the objective function.

In the nonoverlap case, where each variable belongs to only one group, estimation of parameters and selection of important variables can be conducted via the minimization of (1). The negative partial likelihood $-\ell_n(\boldsymbol{\beta})$ is convex and twice continuously differentiable, which allows us to implement Newton's method. Although the lasso penalty is nondifferentiable, i.e., there is no derivative at the origin, fortunately, its directional derivatives along the forward and backward directions are available. For the group lasso

penalty $\|\boldsymbol{\beta}_{(k)}\|_2$, there are two situations to consider. When $\|\boldsymbol{\beta}_{(k)}\|_2 = 0$ at the current value, $\|\boldsymbol{\beta}_{(k)}\|_2 = |\beta_{kj}|$ is a function of β_{kj} if we consider the component β_{kj}. Thus, minimization with respect to β_{kj} reduces to the standard update with a lasso penalty with tuning constant $\lambda_1 + \lambda_2$. If $\|\boldsymbol{\beta}_{(k)}\|_2 > 0$ at the current value, then $\|\boldsymbol{\beta}_{(k)}\|_2$ is twice differentiable with

$$\frac{\partial}{\partial \beta_{kj}} \|\boldsymbol{\beta}_{(k)}\|_2 = \frac{\beta_{kj}}{\|\boldsymbol{\beta}_{(k)}\|_2},$$
$$\frac{\partial^2}{\partial \beta_{kj}^2} \|\boldsymbol{\beta}_{(k)}\|_2 = \frac{1}{\|\boldsymbol{\beta}_{(k)}\|_2}\Big(1 - \frac{\beta_{kj}^2}{\|\boldsymbol{\beta}_{(k)}\|_2^2}\Big).$$

If e_{kj} is the coordinate direction along which β_{kj} varies, then the forward and backward directional derivatives of β_{kj} are

$$\begin{aligned} & d_{e_{kj}} g(\boldsymbol{\beta}) \\ &= \lim_{t \downarrow 0} \frac{g(\boldsymbol{\beta} + t e_{kj}) - g(\boldsymbol{\beta})}{t} \\ &= -\frac{\partial}{\partial \beta_{kj}} \ell_n(\boldsymbol{\beta}) \\ &\quad + \begin{cases} (\lambda_1 + \lambda_2)(-1)^{I(\beta_{kj} < 0)} & \text{if } \|\boldsymbol{\beta}_{(k)}\|_2 = 0 \\ \lambda_1 (-1)^{I(\beta_{kj} < 0)} + \lambda_2 \frac{\beta_{kj}}{\|\boldsymbol{\beta}_{(k)}\|_2} & \text{if } \|\boldsymbol{\beta}_{(k)}\|_2 > 0, \end{cases} \end{aligned}$$

and

$$\begin{aligned} & d_{-e_{kj}} g(\boldsymbol{\beta}) \\ &= \lim_{t \downarrow 0} \frac{g(\boldsymbol{\beta} - t e_{kj}) - g(\boldsymbol{\beta})}{t} \\ &= \frac{\partial}{\partial \beta_{kj}} \ell_n(\boldsymbol{\beta}) \\ &\quad + \begin{cases} (\lambda_1 + \lambda_2)(-1)^{I(\beta_{kj} > 0)} & \text{if } \|\boldsymbol{\beta}_{(k)}\|_2 = 0 \\ \lambda_1 (-1)^{I(\beta_{kj} > 0)} - \lambda_2 \frac{\beta_{kj}}{\|\boldsymbol{\beta}_{(k)}\|_2} & \text{if } \|\boldsymbol{\beta}_{(k)}\|_2 > 0, \end{cases} \end{aligned}$$

where $I(\cdot)$ is an indicator function equal to 1 if the condition in the parentheses is satisfied and 0 otherwise, and

$$\begin{aligned} & \frac{\partial}{\partial \beta_{kj}} \ell_n(\boldsymbol{\beta}) \\ &= \sum_{i \in D} \left\{ x_{i,kj} - \frac{\sum_{l \in R_i} \exp\Big(\sum_{k=1}^K \boldsymbol{X}_{l,(k)}^T \boldsymbol{\beta}_{(k)}\Big) x_{l,kj}}{\sum_{l \in R_i} \exp\Big(\sum_{k=1}^K \boldsymbol{X}_{l,(k)}^T \boldsymbol{\beta}_{(k)}\Big)} \right\}. \end{aligned}$$

After obtaining the directional derivatives, we need to decide which parameters to be updated and the direction for updating. If both of the directional derivatives $d_{e_{kj}} g(\boldsymbol{\beta})$ and $d_{-e_{kj}} g(\boldsymbol{\beta})$ are nonnegative, then the update for β_{kj} is skipped. If either directional derivative is negative, then we solve for the minimum along the corresponding direction. It is impossible for both directional derivatives to be negative due to the convexity of $g(\boldsymbol{\beta})$. After identifying the direction to update the parameter, one can use Newton's method to solve for the minimum. The update at iteration $m+1$ is given by

$$\begin{aligned} \beta_{kj}^{m+1} = \beta_{kj}^m & + \frac{-\frac{\partial}{\partial \beta_{kj}} \ell_n(\boldsymbol{\beta}^m) + \lambda_1 (-1)^{I(\beta_{kj}^m < 0)}}{\frac{\partial^2}{\partial \beta_{kj}^2} \ell_n(\boldsymbol{\beta}^m)} \\ & + \frac{\lambda_2 \{(-1)^{I(\beta_{kj}^m < 0)} I_1(\beta_{(k)}^m) + \frac{\beta_{kj}^m}{\|\boldsymbol{\beta}_{(k)}^m\|_2} I_2(\beta_{(k)}^m)\}}{\frac{\partial^2}{\partial \beta_{kj}^2} \ell_n(\boldsymbol{\beta}^m)} \end{aligned}$$

where $\boldsymbol{\beta}^m$ is the estimate at iteration m, $I_1(\cdot) = I(\|\cdot\|_2 = 0)$, and $I_2(\cdot) = I(\|\cdot\|_2 > 0)$.

All parameters are recommended to be initiated at the origin. For problems with sparse solutions, most updates are skipped and many parameters never leave from their initial values of 0. This is another reason why cyclic coordinate descent is fast, additionally to the fact of no matrix operations involved.

The convergence of coordinate descent to a local optimizer is a straightforward extension of the results in [29] and [32] for the nonoverlap case. For the overlap case discussed in the next section, the convergence results still hold since in our setting one predictor is associated with only one coefficient, which is updated separately from the rest of coefficients.

2.4 Doubly penalized Cox regression for overlap cases

The group structure we have considered in previous sections does not have overlaps, i.e. each variable belongs to only one group. In practice, however, a variable can belong to several groups. For example, one gene can be shared by many different pathways. In this section, we extend the proposed method for problems with overlaps.

To allow for overlapping, we modify the notation and rewrite the objective function (1). The p variables are denoted by $X_1, \ldots, X_p$ and their corresponding regression coefficients are $\beta_1, \ldots, \beta_p$. We let $V_k \subseteq \{1, 2, \ldots, p\}$ be the set of indices of variables in the kth group. We consider the following objective function designed for the overlap case:

$$(2) \qquad g(\boldsymbol{\beta}) = -\ell_n(\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^p |\beta_p| + \lambda_2 \sum_{k=1}^K \sqrt{\sum_{j \in V_k} \beta_j^2}.$$

Note that predictor X_j is only associated with one coefficient β_j. It is easy to see that the objective function (2) reduces to the objective function (1) when there is no overlap among the p variables.

The cyclic coordinate descent algorithm can be easily modified for the objective function (2). If we consider the coordinate direction e_j for β_j, the forward and backward directional derivatives of β_j are

$$
\begin{aligned}
&d_{e_j} g(\boldsymbol{\beta}) \\
&\quad = \lim_{t \downarrow 0} \frac{g(\boldsymbol{\beta} + t e_j) - g(\boldsymbol{\beta})}{t} \\
&\quad = -d_{e_j} \ell_n(\boldsymbol{\beta}) + \lambda_1 (-1)^{I(\beta_j < 0)} \\
&\qquad + \lambda_2 \sum_{k \in G_j} \{(-1)^{I(\beta_j < 0)} I_1(\boldsymbol{\beta}_{(k)}) \\
&\qquad + \frac{\beta_j}{\|\boldsymbol{\beta}_{(k)}\|_2} I_2(\boldsymbol{\beta}_{(k)})\}
\end{aligned}
$$

and

$$
\begin{aligned}
&d_{-e_j} g(\boldsymbol{\beta}) \\
&\quad = \lim_{t \downarrow 0} \frac{g(\boldsymbol{\beta} - t e_j) - g(\boldsymbol{\beta})}{t} \\
&\quad = -d_{-e_j} \ell_n(\boldsymbol{\beta}) + \lambda_1 (-1)^{I(\beta_j > 0)} \\
&\qquad + \lambda_2 \sum_{k \in G_j} \{(-1)^{I(\beta_j > 0)} I_1(\boldsymbol{\beta}_{(k)}) \\
&\qquad - \frac{\beta_j}{\|\boldsymbol{\beta}_{(k)}\|_2} I_2(\boldsymbol{\beta}_{(k)})\}.
\end{aligned}
$$

where $G_j \subseteq \{1, 2, \ldots, K\}$ are the indices of groups that X_j belongs to.

After determining the direction for updating, the coefficient can be updated by

$$
\begin{aligned}
&\beta_j^{m+1} \\
&\quad = \beta_j^m + \frac{-\frac{\partial}{\partial \beta_j} \ell_n(\boldsymbol{\beta}^m) + \lambda_1 (-1)^{I(\beta_j^m < 0)}}{\frac{\partial^2}{\partial \beta_j^2} \ell_n(\boldsymbol{\beta}^m)} \\
&\qquad + \frac{\lambda_2 \sum_{k \in G_j} \{(-1)^{I(\beta_j^m < 0)} I_1(\boldsymbol{\beta}_{(k)}^m) + \frac{\beta_j^m}{\|\boldsymbol{\beta}_{(k)}^m\|_2} I_2(\boldsymbol{\beta}_{(k)}^m)\}}{\frac{\partial^2}{\partial \beta_j^2} \ell_n(\boldsymbol{\beta}^m)}.
\end{aligned}
$$

3. SIMULATION STUDIES

3.1 Comparison to competing methods in nonoverlap and overlap cases

In this section, we apply the doubly regularized Cox regression (DrCox) in four simulation settings and compare it to Cox regression with lasso penalty (lasso-Cox), Cox regression with group lasso penalty (glasso-Cox) and Cox regression with hierarchical lasso penalty (hlasso-Cox). The first three settings are overdetermined, i.e., $p < n$; the fourth setting is underdetermined, i.e., $n < p$. The first two examples are nonoverlap cases and the other two examples are overlap cases. In all settings, the survival time is generated from an exponential distribution with $h(t|\boldsymbol{x}) = \exp(\boldsymbol{x}^T \boldsymbol{\beta})$. The censoring time is generated from a uniform distribution $U(0, c)$, where c is chosen to achieve a 35% censoring rate. Detailed settings of training data are provided below. In all the simulation examples, we set the ranges of the two tuning parameters wide enough to cover the cases where zero predictor to several hundred predictors could be selected.

Example 1. We generate $n = 50, 100$ training samples, $p = 24$ variables, and $K = 3$ groups with 8 variables in each group. The three groups are independent. In groups 1 and 2, variables are generated from $N(0, 1)$ with $\text{cov}(x_{1i}, x_{1j}) = 0.5^{|i-j|}$. In group 3, variables are generated from independent $N(0, 1)$. The corresponding coefficients are:

$$
\boldsymbol{\beta} = (\underbrace{1.5, -0.8, 0, 0, 0, 1.2, 0, 0}_{8}, \underbrace{0}_{8}, \underbrace{0}_{8})'.
$$

Example 2. We generate $n = 100, 200$ training samples, $p = 40$ variables, and $K = 8$ groups of different group sizes. The group ID's for the covariate vector are

$$
\begin{aligned}
&\underbrace{1, \ldots, 1}_{6}, \underbrace{2, \ldots, 2}_{4}, \underbrace{3, \ldots, 3}_{6}, \underbrace{4, \ldots, 4}_{5}, \underbrace{5, \ldots, 5}_{4}, \\
&\underbrace{6, \ldots, 6}_{5}, \underbrace{7, \ldots, 7}_{4}, \underbrace{8, \ldots, 8}_{6}.
\end{aligned}
$$

We first generate $W_1, \ldots, W_{40}$ independently from a standard normal distribution. Then we generate $Z_1, \ldots, Z_8$ from a standard normal distribution with $\text{cov}(Z_k, Z_{k'}) = 0.5^{|k-k'|}$. Then variables are generated by $X_{kj} = (W_{kj} + Z_k)/\sqrt{2}$. The corresponding coefficients are

$$
\begin{aligned}
\boldsymbol{\beta} = (&\underbrace{1.2, -0.8, 1.6, 0, 0, 0}_{6}, \underbrace{1, -0.9, -1.1, -1.3}_{4}, \\
&\underbrace{0}_{6}, \underbrace{0}_{5}, \underbrace{0}_{4}, \underbrace{1.5, 0, 0, 0, 0}_{5}, \underbrace{0}_{4}, \underbrace{0}_{6})'.
\end{aligned}
$$

In this example, there are three important groups, i.e., at least one variable within the group has nonzero coefficient. The three groups represent three situations: all variables within the group are important (group 2), some variables within the group are important (group 1), and very few variables within the group are important (group 6).

Example 3. We generate $n = 100, 200$ training samples, $p = 48$ variables, and $K = 8$ groups with some groups being overlapped. The group ID's are

$$
\begin{aligned}
&\underbrace{1, \ldots, 1}_{8} \; \underbrace{\overbrace{2, 2, 2, 2, 2}^{5}}_{\underbrace{3, 3, 3, 3, 3}_{5}} \; \underbrace{4, \ldots, 4}_{8} \; \underbrace{5, \ldots, 5}_{8} \\
&\underbrace{\overbrace{6, 6, 6, 6, 6}^{5}}_{\underbrace{7, 7, 7, 7, 7}_{5}} \; \underbrace{8, \ldots, 8}_{8}.
\end{aligned}
$$

Notice that groups 2 and 3 have two overlapped variables, and so do groups 6 and 7. The first 24 variables $X_1, \ldots, X_{24}$ are generated from a standard normal distribution with $\text{cov}(X_j, X_{j'}) = 0.5^{|j-j'|}$, and the rest $X_{25}, \ldots, X_{48}$ are also generated from a standard normal distribution with

$\mathrm{cov}(X_j, X_{j'}) = 0.5^{|j-j'|}$. $X_1, \ldots, X_{24}$ and $X_{25}, \ldots, X_{48}$ are independent. The corresponding coefficients are

$$\underbrace{0}_{8}\ \frac{\overbrace{1.3, 0, 1.5, 0, -1}^{5}}{\underbrace{0, -1, 0, -2, -1.1}_{5}}\ \underbrace{0}_{8}\ \underbrace{0}_{8}\ \frac{\overbrace{1.4, 0, 0.8, 0, 1}^{5}}{\underbrace{0, 1, 0, 1.6, 0}_{5}}\ \underbrace{0}_{8}.$$

Example 4. We generate $n = 100, 200$ training samples, $p = 148$ variables, and $K = 24$ groups of different group sizes. Some groups are overlapped. The 24 groups can be divided into four blocks, each having the same group assignment. The group ID's for variables in the first block are

$$\underbrace{1, \ldots, 1}_{8}\ \frac{\overbrace{2, 2, 2, 2, 2}^{5}}{\underbrace{3, 3, 3, 3, 3}_{5}}\ \underbrace{4, \ldots, 4}_{8}\ \underbrace{5, \ldots, 5}_{5}\ \underbrace{6, \ldots, 6}_{8}.$$

The four blocks are independent with each other. Within the four blocks, variables are generated from N(0,1) with $\mathrm{cov}(X_j, X_{j'}) = 0.5^{|j-j'|}$, respectively. The corresponding coefficients are

$$\boldsymbol{\beta} = \Big(\underbrace{0}_{8}, \frac{\overbrace{1.3, 0, 1.5, 0, -1}^{5}}{\underbrace{0, -1, 0, -2, -1.1}_{5}}, \underbrace{0}_{8}, \underbrace{1.2, 1.8, 0, 0, 0}_{5}, \underbrace{0}_{8}, \underbrace{0}_{8}, \frac{\overbrace{1.4, 0, 0.8, 0, 1.4}^{5}}{\underbrace{0, 1.4, 0, 1.6, 0}_{5}}, \underbrace{0}_{8}, \underbrace{-0.9, -1.1, 0, 0, 0}_{5}, \underbrace{0}_{8}, \underbrace{0}_{74}\Big).$$

For each of the settings above, we generate an independent validation set with 500 samples to select the optimal tuning constants that maximize the partial likelihood by grid search. The estimate at the optimal λ's will be tested on an independent testing sample with 500 observations. Since the quality of the parameter estimates are naturally of interest, we always re-estimate the active parameters to offset the shrinkage for all three methods [8, 35, 36].

Table 1 reports the simulation results of 100 replicates. The predictors are divided into three categories: $X_{\mathcal{A}}$, which denotes important predictors within important groups; $X_{\mathcal{B}}$, which denotes unimportant predictors within important groups; and $X_{\mathcal{C}}$, which denotes unimportant predictors within unimportant groups. For example, in simulation 1, $X_{\mathcal{A}} = \{X_1, X_2, X_6\}$, $X_{\mathcal{B}} = \{X_3, X_4, X_5, X_7, X_8\}$, and the rest belong to $X_{\mathcal{C}}$. The variable selection performance is reported in columns 4–6. In each block (example), the first line is the true numbers of predictors in each of three categories. For each method, the average numbers of selected predictors in each of three categories over 100 replicates are reported with the corresponding standard errors in the parenthesis. We expect the numbers in column $X_{\mathcal{A}}$ to be as large as possible (of course it cannot exceed the truth) and the numbers in columns $X_{\mathcal{B}}$ and $X_{\mathcal{C}}$ to be as small as possible.

The variable selection performances of four methods are compared from three different aspects: (1) removing unimportant groups (i.e. $X_{\mathcal{C}}$); (2) removing unimportant variables in important groups (i.e. $X_{\mathcal{B}}$); and (3) selecting important variables in important groups (i.e. $X_{\mathcal{A}}$). First, glasso-Cox, hlasso-Cox, and DrCox perform better than lasso-Cox in all four examples in removing unimportant groups. This is because the three methods utilize the group structure which leads to a more efficient removal of unimportant groups than lasso-Cox. Glasso-Cox, hlasso-Cox and DrCox perform comparably in Example 1. Glasso-Cox removes all unimportant groups in the overdetermined settings (Examples 1–3), but it picks up more variables in unimportant groups than hlasso-Cox and DrCox in the underdetermined setting (Example 4). Second, lasso-Cox is the best in terms of removing unimportant variables in important groups in all four examples due to its flexibility of selection based on individual predictors. Glasso-Cox fails to remove the unimportant variables in important groups because of its "all-in-or-all-out" property. DrCox has better performance in removing unimportant variables in important groups than hlasso-Cox in all four examples. Third, glasso-Cox, hlasso-Cox, and DrCox have similar performance in all four examples in terms of selecting important variables $X_{\mathcal{A}}$. As sample size n increases, better selection results are obtained. In summary, our DrCox method has a good balance between group selection and individual predictor selection. It is not only as effective as glasso-Cox and hlasso-Cox in removing unimportant groups but also as comparable as the lasso-Cox in removing unimportant variables in important groups.

To measure the prediction accuracy, we follow [31] and calculate the model error (ME) at the optimal λ_1 and λ_2

$$\mathrm{ME} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \Sigma (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

where Σ is the covariance matrix of the predictors. The model error of the three methods is reported in Column 7 of Table 1. In the first three examples ($n > p$), the DrCox and hlasso-Cox have comparable model errors, which are much smaller than the model errors of lasso-Cox and glasso-Cox. In the fourth example ($p > n$), the DrCox method has the smallest model error among all four methods. As sample size n increases, better predictions are obtained. The last

Table 1. Simulation results for Examples 1–4 over 100 random replications. Columns 2 to 3 list the sample size n and methods. Columns 4 to 6 report the average number of selected variables in $X_{\mathcal{A}}$, $X_{\mathcal{B}}$, and $X_{\mathcal{C}}$ with standard errors appearing in parentheses. Column 7 reports the average model errors (ME) and the corresponding standard errors. The last column is the average training time in seconds under the optimal tuning constants

	n	Method	$X_{\mathcal{A}}$	$X_{\mathcal{B}}$	$X_{\mathcal{C}}$	ME	Time
Example 1		truth	3	5	16		
	50	lasso-Cox	2.53 (0.05)	0.71 (0.09)	0.62 (0.09)	0.518 (0.030)	0.013
		glasso-Cox	3 (0)	5 (0)	0 (0)	0.760 (0.056)	0.039
		hlasso-Cox	3 (0)	3.26 (0.10)	0.04 (0.03)	0.276 (0.028)	5.710
		DrCox	2.87 (0.03)	1.86 (0.13)	0.1 (0.05)	0.351 (0.029)	0.026
	100	lasso-Cox	2.95 (0.02)	1.14 (0.08)	0.75 (0.1)	0.167 (0.156)	0.083
		glasso-Cox	3 (0)	5 (0)	0 (0)	0.225 (0.019)	0.163
		hlasso-Cox	3 (0)	3.32 (0.10)	0.06 (0.06)	0.127 (0.008)	18.563
		DrCox	2.99 (0.01)	1.71 (0.12)	0.12 (0.04)	0.122 (0.013)	0.215
Example 2		truth	8	7	25		
	100	lasso-Cox	7.04 (0.11)	1.13 (0.12)	2.33 (0.24)	1.121 (0.088)	0.131
		glasso-Cox	8 (0)	7 (0)	0 (0)	0.961 (0.092)	0.269
		hlasso-Cox	7.99 (0.01)	3.38 (0.13)	0.64 (0.05)	0.523 (0.028)	7.931
		DrCox	7.98 (0.01)	3.02 (0.2)	0.28 (0.07)	0.537 (0.048)	0.213
	200	lasso-Cox	8 (0)	1.27 (0.11)	1.95 (0.17)	0.225 (0.019)	0.753
		glasso-Cox	8 (0)	7 (0)	0 (0)	0.279 (0.024)	1.127
		hlasso-Cox	8 (0)	3.80 (0.12)	0.65 (0.05)	0.166 (0.008)	39.421
		DrCox	8 (0)	1.91 (0.15)	0.17 (0.06)	0.165 (0.016)	1.207
Example 3		truth	9	7	32		
	100	lasso-Cox	8.81 (0.04)	3.07 (0.13)	2.25 (0.21)	1.392 (0.104)	0.145
		glasso-Cox	8.95 (0.04)	6.95 (0.04)	0 (0)	2.166 (0.213)	0.371
		hlasso-Cox	9 (0)	4.02 (0.12)	0.63 (0.09)	1.117 (0.063)	9.794
		DrCox	8.96 (0.03)	3.64 (0.16)	0.59 (0.12)	1.147(0.106)	0.195
	200	lasso-Cox	9 (0)	3.24 (0.13)	1.08 (0.12)	0.330 (0.034)	0.582
		glasso-Cox	9 (0)	7 (0)	0 (0)	0.435 (0.044)	1.63
		hlasso-Cox	9 (0)	3.97 (0.12)	0.54 (0.05)	0.344 (0.025)	78.903
		DrCox	9 (0)	3.82 (0.15)	0.46 (0.08)	0.316 (0.034)	0.745
Example 4		truth	13	13	122		
	100	lasso-Cox	11.61 (0.15)	2.86 (0.14)	4.62 (0.31)	4.565 (0.385)	0.375
		glasso-Cox	11.1 (0.28)	11.07 (0.29)	1.62 (0.32)	9.472 (0.721)	0.876
		hlasso-Cox	12.79 (0.06)	7.14 (0.17)	0.40 (0.12)	3.68 (0.22)	13.607
		DrCox	12.59 (0.09)	4.76 (0.22)	3.22 (0.29)	2.772 (0.250)	0.466
	200	lasso-Cox	12.98 (0.01)	3.4 (0.13)	3.1 (0.27)	0.753 (0.079)	0.467
		glasso-Cox	13 (0)	13 (0)	0.97 (0.26)	1.154 (0.100)	1.118
		hlasso-Cox	13 (0)	9.67 (0.16)	1.47 (0.25)	0.661 (0.031)	150.551
		DrCox	13 (0)	4.38 (0.22)	1.23 (0.15)	0.615 (0.064)	1.16

column of Table 1 reports the computing time in seconds on a personal computer at the optimal value of λ_1 and λ_2. Note that Lasso-Cox, glasso-Cox, and DrCox are implemented in Fortran 90, while hlasso-Cox is implemented in R.

To better understand the effects of tuning constants on our DrCox method, the average negative log-partial likelihood $-\ell_n(\boldsymbol{\beta})$ of the independent validation set and the average number of nonzero predictors based on 100 replicates are plotted over the grid of λ_1 and λ_2 in Figure 1. The optimal $(\lambda_1, \lambda_2) = (0.11, 0.18)$ is achieved at the minimum of average $-\ell_n(\boldsymbol{\beta})$, which leads to an average total number of 5.01 nonzero predictors. The occurrence of a unique minimum is a consequence of the convexity of the objective function.

3.2 Large groups vs. large numbers of groups

We focus on ultra-high-dimensional settings in this section. Two situations are considered: one has a large number of predictors in one group (Examples 5–6), and the other has a large number of groups (Example 7).

Example 5. We generate $n = 100$ training samples, $p = 1,000$ variables, and $K = 3$ groups with 8 variables in the first two groups. The three groups are independent. In groups 1 and 2, variables are generated from $N(0,1)$ with $\text{cov}(x_{1i}, x_{1j}) = 0.5^{|i-j|}$. In group 3, variables are generated from independent $N(0,1)$. The corresponding coefficients

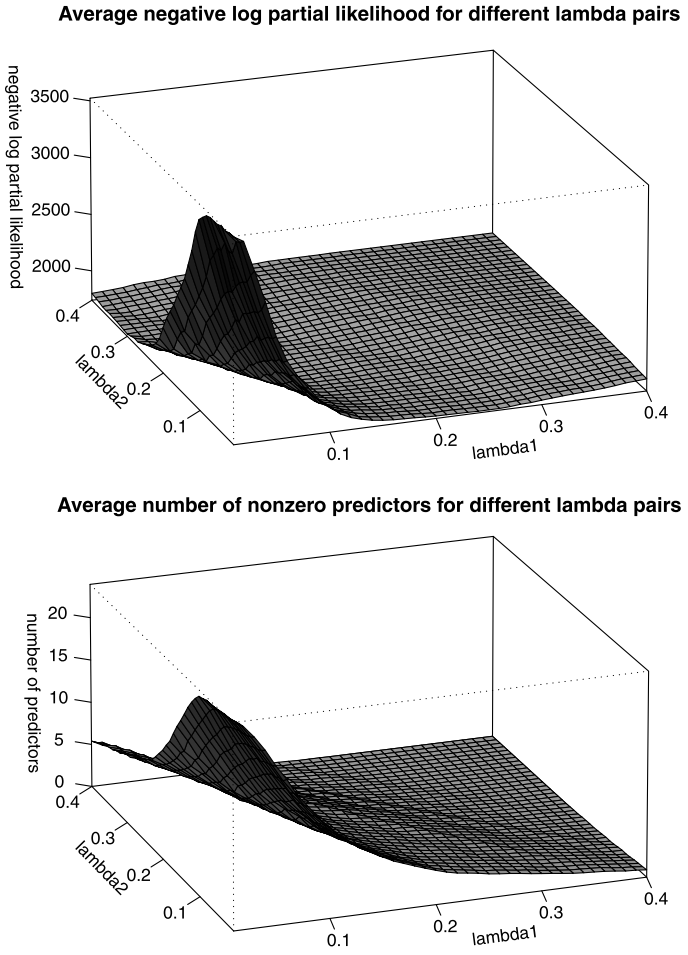

Figure 1. Upper panel: 3D-plot of average $-\ell_n(\boldsymbol{\beta})$ versus λ_1 and λ_2 for DrCox in Example 1. Lower panel: 3D-plot of average number of nonzero predictors versus λ_1 and λ_2 for DrCox in Example 1.

are:

$$\boldsymbol{\beta} = (\underbrace{1.5, -0.8, 0, 0, 0, 1.2, 0, 0}_{8}, \underbrace{0}_{8}, \underbrace{0}_{984})'.$$

Example 6. We use the same setting as Example 5, except $p = 5,000$. The corresponding coefficients are:

$$\boldsymbol{\beta} = (\underbrace{1.5, -0.8, 0, 0, 0, 1.2, 0, 0}_{8}, \underbrace{0}_{8}, \underbrace{0}_{4984})'.$$

Example 7. $n = 100$, $p = 1,000$ and $K = 100$. Every 10 predictors form one group. The corresponding coefficients are:

$$\boldsymbol{\beta} = (\underbrace{1.5, -0.8, 0, 0, 0, 1.2, 0, 0, 0, 0}_{10}, \underbrace{0}_{10}, \ldots, \underbrace{0}_{10})'.$$

The results are reported in Table 2. For large p or K, the lasso method works much worse than the other methods for small sample sizes ($n = 100$) in selecting the true predictors $X_{\mathcal{A}}$. The group lasso methods select all the predictors in important groups. Our method selects almost all important predictors, while eliminating unimportant predictors $X_{\mathcal{B}}$ and $X_{\mathcal{C}}$. The hierarchical lasso method does not work in ultra-high-dimensional settings due to the computational speed and numerical instability.

3.3 Misspecification of groups

Unlike elastic net [44], in which no grouping information is required, our method needs to specify the groups. It is interesting to investigate the effects of misspecification of groups. We consider the following two examples. In the first example, the overlap groups are collapsed and the number of groups are wrong. In the second example, the overlapping predictors are put in one group.

Example 8. Suppose the data are generated in the same way as Example 3 with $n = 100$ and $p = 48$. However, the grouping information is misspecified. Instead of $K = 8$, the number of groups is misspecified as 6. The overlapping groups of 3 and 4 are considered as one group with 8 predictors, and so are groups 6 and 7. The misspecified group ID's are

$$\underbrace{1,\ldots,1}_{8}\ \underbrace{\begin{matrix}2,2,2,2,2\\ 2,2,2,2,2\end{matrix}}_{8}\ \underbrace{3,\ldots,3}_{8}\ \underbrace{4,\ldots,4}_{8}$$
$$\underbrace{\begin{matrix}5,5,5,5,5\\ 5,5,5,5,5\end{matrix}}_{8}\ \underbrace{6,\ldots,6}_{8}.$$

Example 9. The setting is the same as Example 8, except the grouping information is misspecified as:

$$\underbrace{1,\ldots,1}_{8}\ \begin{matrix}\overbrace{2,2,2,2,2}^{5}\\ \underbrace{3,3,3}_{3}\end{matrix}\ \underbrace{4,\ldots,4}_{8}\ \underbrace{5,\ldots,5}_{8}$$
$$\begin{matrix}\overbrace{6,6,6}^{3}\\ \underbrace{7,7,7,7,7}_{5}\end{matrix}\ \underbrace{8,\ldots,8}_{8}.$$

The results are reported in Table 3. It is easy to see that the lasso method is not affected by the mis-grouping, as expected. The group lasso method selects all the misspecified predictors and is affected most. The DrCox method and hierarchical method are quite robust and not too sensitive to the misspecification.

4. TCGA OVARIAN CANCER DATA ANALYSIS

As mentioned in the beginning, this research was motivated by the ovarian cancer study from The Can-

Table 2. Simulation results for Examples 5–7 over 100 random replications for $n = 100$. Column 2 lists the number of predictors p and the number of groups K. Columns 4 to 6 report the average number of selected variables in $X_{\mathcal{A}}$, $X_{\mathcal{B}}$, and $X_{\mathcal{C}}$ with standard errors appearing in parentheses. Column 7 reports the average model errors (ME) and the corresponding standard errors. The last column is the average training time in seconds under the optimal tuning constants

	(p, K)	Method	$X_{\mathcal{A}}$	$X_{\mathcal{B}}$	$X_{\mathcal{C}}$	ME	Time
Example 5	$(1000, 3)$	truth	3	5	992		
		lasso-Cox	2.02 (0.03)	0.14 (0.04)	0.18 (0.06)	0.596 (0.024)	0.331
		glasso-Cox	3 (0)	5 (0)	0 (0)	0.247 (0.023)	0.994
		hlasso-Cox	NA				
		DrCox	2.99 (0.01)	2.51 (0.15)	0 (0)	0.155 (0.016)	0.734
Example 6	$(5000, 3)$	truth	3	5	4992		
		lasso-Cox	1.93 (0.03)	0.12 (0.04)	0.32 (0.07)	0.750 (0.038)	2.225
		glasso-Cox	3 (0)	5 (0)	0 (0)	0.384 (0.071)	6.78
		hlasso-Cox	NA				
		DrCox	2.92 (0.03)	2.81 (0.15)	0.2 (0.1)	0.236 (0.039)	5.245
Example 7	$(1000, 100)$	truth	3	7	990		
		lasso-Cox	2.06 (0.03)	0.15 (0.04)	0.35 (0.1)	0.595 (0.024)	0.334
		glasso-Cox	3 (0)	7 (0)	0 (0)	0.337 (0.029)	0.977
		hlasso-Cox	NA				
		DrCox	2.98 (0.01)	2.88 (0.18)	0.06 (0.03)	0.169 (0.016)	0.707

Table 3. Simulation results for Examples 8–9 over 100 random replications for $n = 100$. Columns 3 to 5 report the average number of selected variables in $X_{\mathcal{A}}$, $X_{\mathcal{B}}$, and $X_{\mathcal{C}}$ with standard errors appearing in parentheses. Column 6 reports the average model errors (ME) and the corresponding standard errors. The last column is the average training time in seconds under the optimal tuning constants

	Method	$X_{\mathcal{A}}$	$X_{\mathcal{B}}$	$X_{\mathcal{C}}$	ME	Time
Example 8	truth	9	7	32		
	lasso-Cox	8.77 (0.05)	3.28 (0.13)	2.71 (0.22)	1.645 (0.135)	0.041
	glasso-Cox	9 (0)	7 (0)	0 (0)	2.011 (0.200)	0.097
	hlasso-Cox	9 (0)	4.05 (0.18)	0.75 (0.06)	1.516 (0.147)	10.214
	DrCox	8.95 (0.02)	4.26 (0.16)	0.28 (0.07)	1.284 (0.127)	0.069
Example 9	truth	9	7	32		
	lasso-Cox	8.77 (0.05)	3.28 (0.13)	2.71 (0.22)	1.645 (0.135)	0.041
	glasso-Cox	8.86 (0.07)	6.9 (0.05)	0.99 (0.27)	2.399 (0.233)	0.094
	hlasso-Cox	9 (0)	3.59 (0.13)	0.58 (0.05)	1.434 (0.122)	15.675
	DrCox	8.94 (0.02)	4.07 (0.15)	0.76 (0.13)	1.314 (0.128)	0.061

cer Genome Atlas (TCGA) project. The gene expression data of ovarian cancer are publicly available at http://cancergenome.nih.gov. This study is expected to produce high-quality, large-sample, and well-curated data. Up to the date of September 11, 2010, 503 samples have been collected. The 134 independent samples described in the paper of [3] are used as a test set.

Our analysis is based on 1,863 genes from 15 core pathways suggested by [21], which are 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.

We apply our doubly penalized Cox regression method to the TCGA data. The whole dataset is randomly split into one training set and one validation set with equal size. The training set was used for model fitting, and the validation set is used for tuning constants selection. Under the optimal tuning constants, 4 pathways and 36 genes are selected from the pool of 1,863 genes in the 15 core pathways. The selected pathways are: Apoptosis (5 out of 163 genes are selected), cell cycle (11 out of 238 genes are selected), MAPK signaling pathway (18 out of 481 genes are selected) and Wnt signaling pathway (11 out of 259 genes are selected). The detailed gene list is available upon request.

The identified pathways are biologically meaningful and consistent with the existing scientific findings. Apoptosis pathway is well-known to be related to the development of cancer and its activation is a key mechanism by which cytotoxic drugs kill tumor cells [6]. [7] reported the prognostic impact of apoptosis pathway in ovarian cancer. The MAPK

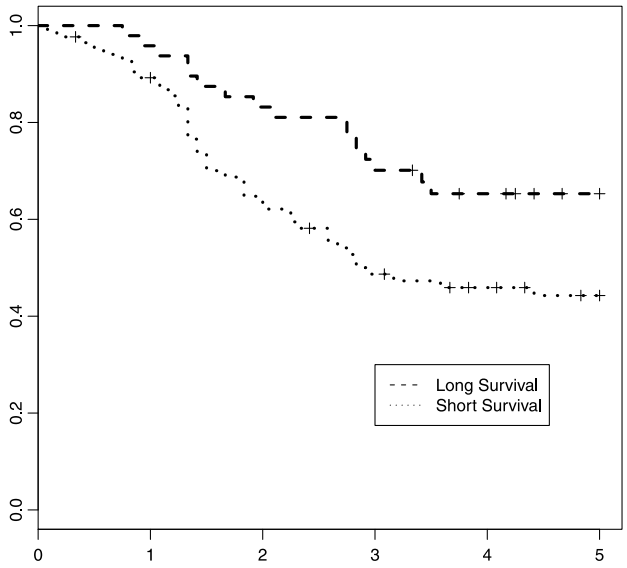

Figure 2. Ovarian cancer survival curves (Kaplan-Meier) for the high and low risk groups of the 134 independent testing samples. The p-value of the log rank test is 0.01.

signal transduction cascade is dysregulated in a majority of human tumors [2]. It is expected to play an important role in molecular diagnostics and molecular therapeutics for low-grade ovarian cancer [1]. [22] identified upregulated genes involved in the MAPK signaling pathway in ovarian cancer tissues. Wnt signaling pathway is best known for its role in tumorigenesis. [16] demonstrated the difference in Wnt signaling pathway between normal ovarian and cancer cell lines and between benign tissue and ovarian cancer. They also pointed out that those differences implicate that Wnt signaling leads to ovarian cancer development despite the fact that gene mutations are uncommon.

In order to predict the survival for subjects in the independent ovarian cancer dataset using the model with the selected 4 pathways, we first calculate the Breslow estimate [4] of the cumulative baseline hazard. The risk score $X\hat{\boldsymbol{\beta}}$ is computed to find the 50% survival probability at three years for the subjects in the TCGA training set, which is used as the threshold for the high and low risk groups. The risk scores for the subjects in the dataset of [3] are then computed using $\hat{\boldsymbol{\beta}}$ obtained from the TCGA training set and subjects are assigned into the high and low risk groups by comparing with the threshold. Out of the 134 subjects in [3], 48 are in the high risk group, and 86 are in the low risk group. The Kaplan-Meier curves (Fig. 2) of these two groups are well separated with a p-value of the log-rank test equals to 0.01.

As a comparison, we also apply lasso, group lasso and hierarchical lasso methods to the TCGA ovarian cancer data. The hierarchical lasso fails to fit the model using the existing R code because of the high dimensionality. The lasso method only identifies one gene and failed to predict on the independent test set. One possible explanation is that the signal-to-noise ratio in the TCGA ovarian cancer data might be relatively low. The lasso method selects genes based on their individual strength, and hence failed to identify important genes whose individual effects are weak. The group lasso method picks up no genes. We think this may be caused by the "all-in-all-out" selection nature of the group lasso method. The 15 pathways in TCGA ovarian cancer data have too many overlaps. When the group lasso method picks up one gene, it has to pick up all the other genes in that pathway and this may result in picking up too many unimportant genes and reduce the prediction ability. The doubly penalized method takes a good balance of individual selection and group selection, therefore it produces good selection results.

5. CONCLUSION

In this paper, we impose convex penalties for both group selection and within group selection on the Cox regression model for high-dimensional survival data. This doubly regularized method not only keeps the advantage of group lasso in effectively removing unimportant groups, but also maintains the flexibility of selecting important variables within the identified groups. To tackle the high-dimensionality and nondifferentiability problems in optimization, we develop efficient coordinate descent algorithms for nonoverlap and overlap cases, respectively. This new method has been demonstrated to perform well in several simulation settings. We analyze the motivating TCGA ovarian cancer data using the new method to predict the patients' survivals. The gene-pathway signature is tested on an independent ovarian cancer dataset and well separates the high and low risk groups.

Our method does not distinguish the contribution of one predictor from overlapping groups. Specifically, one predictor X_j is associated with only one coefficient β_j. Therefore, there is no identifiability problem for predictors in overlapping groups. Another advantage of this setting is that the existing convergence results [29, 32] can be directly extended to our method. Of course, one can always argue that it is interesting to determine the contribution of one predictor from different groups. However, identifiability and uniqueness of estimation might be problems in that case.

One reviewer raised two interesting questions about grouping. The first question is how to define a "group". Some people might define "groups" by correlation and consider predictors correlated with each other from one group. Our definition in this paper is based on both correlation and coefficients, as well as their location. For example, in Example 1, although the last 8 predictors are independent, they have the same 0 coefficients so we define them to belong to one group. Similarly in Example 3, the first 24 predictors are generated from one normal distribution, but groups 1 and 4 are located in different places and the predictors within each group share the same 0 coefficients. By our definition,

they belong to different groups. However, we are not worried about the definition of groups in real data applications as the definition is usually clear and has no ambiguity in real life. Another interesting question raised by the reviewer is the effects of misspecification of groups. Different than elastic net [44], in which no grouping information is required, our method needs the prior information of grouping. The grouping information also needs to be accurate. The two simulation examples in Section 3.3 show that the lasso method is not affected by mis-grouping, as expected, and the group lasso method selects all the misspecified predictors. The effects of misspecification on DrCox are very moderate.

ACKNOWLEDGEMENTS

The authors thank the Editor, Associate Editor, and two referees for their constructive comments that helped substantially to improve the article. T. T. Wus research is supported in part by NSF grant CCF-0926181.

Received 11 November 2011

REFERENCES

[1] Bast, R. and Mills, G. (2010). Personalizing therapy for ovarian cancer: Brcaness and beyond. *Journal of Clinical Oncology* **28** 3545–3548.

[2] Basu, S., Harfouche, R., Soni, S., Chimote, G., Mashelkar, R. A. and Sengupta, S. (2009). Nanoparticle-mediated targeting of mapk signaling predisposes tumor to chemotherapy. *PNAS* **106** 7957–7961.

[3] Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., Joshi, M. B., Harpole, D., Lancaster, J. M., Berchuck, A., Olson, J. A., Marks, J. R., Dressman, H. K., West, M., and Nevins, J. R. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439** 353–357.

[4] Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics* **30** 89–99.

[5] Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. R. Statist. Soc. B* **34** 187–220. MR0341758

[6] Debatin, K. M. (2004). Apoptosis pathways in cancer and cancer therapy. *Cancer Immunol Immunother* **53** 153–159.

[7] Duiker, E. W., van der Zee, A. G., de Graeff, P., Ek, W. B.-v., Hollema, H., de Bock, G. H., de Jong, S. and de Vries, E. G. (2010). The extrinsic apoptosis pathway and its prognostic impact in ovarian cancer. *Gynecologic Oncology* **116** 549–555.

[8] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32** 407–499. MR2060166

[9] Fan, J., Feng, Y. and Wu, Y. (2010). High-dimensional variable selection for cox's proportional hazards model. *Institute of Mathematical Statistics Collections* **30** 70–86.

[10] Fan, J. and Li, R. (2001). Variable selection via non concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360. MR1946581

[11] Fan, J. and Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *The Annals of Statistics* **6** 74–99. MR1892656

[12] Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B* **70** 849–911. MR2530322

[13] Friedman, J., Hastie, T., Hoefling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1** 302–332. MR2415737

[14] Friedman, J., Hastie, T. and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. http://arxiv.org/abs/1001.0736.

[15] Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *J Comp and Graph Stat* **7** 397–416. MR1646710

[16] Gatcliffe, T., Monk, B., Planutis, K. and Holcombe, R. (2008). Wnt signaling in ovarian tumorigenesis. *Int. J. Gynecol. Cancer* **18** 954–962.

[17] Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer Verlag. MR1876599

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

[19] Guo, J. (2010). Simultaneous variable selection and class fusion for high-dimensional linear discriminant analysis. *Biostatistics* **11**, 4, 599.

[20] Huang, J., Ma, S., Xie, H. and Zhang, C. (2009). A group bridge approach for variable selection. *Biometrika* **96** 339–355. MR2507147

[21] 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., Iacobuzio-Donahue, C., Eshleman, J. R., Kern, S. E., Hruban, R. H., Karchin, R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V. E. and Kinzler, K. W. (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321** 1801–1806.

[22] Lancaster, W. (2004). Analysis of genes differentially expressed in a human ovarian cancer model. http://www.worldcat.org/title/analysis-of-genes-differentially-expressed-in-a-human-ovarian-cancer-model/oclc/74274688.

[23] Li, H., Zhang, K. and Jiang, T. (2005). Robust and accurate cancer classification with gene expression profiling. In: *Proc. of the 2005 IEEE Computational Systems Bioinformatics Conference*. IEEE Computer Society, Washington, DC, USA, 310–321. http://dl.acm.org/citation.cfm?id=1084011.1084135.

[24] Li, Y., Nan, B., Wang, S. and Zhu, J. (2012). Group and within group variable selection via convex penalties.

[25] Luan, Y. and Li, H. (2008). Group additive regression models for genomic data analysis. *Biostatistics* **9** 100–113.

[26] Ma, S., Song, X. and Huang, J. (2007). Supervised group lasso with applications to microarray data analysis. *BMC Bioinformatics* **8** 60–76.

[27] NCI. (2010). National cancer insitute: Ovarian cancer. Website. http://www.cancer.gov/cancertopics/types/ovarian.

[28] Park, M. Y. and Hastie, T. (2006). Penalized logistic regression for detecting gene interactions. Tech. Rep. 2006-15, Department of Statistics, Stanford University.

[29] Ruszczyński, A. (2006). *Nonlinear Optimization*. Princeton University Press, Princeton, NJ. MR2199043

[30] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc., Series B* **58** 267–288. MR1379242

[31] Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statist. Med.* **16** 385–395.

[32] Tseng, P. (2001). Convergence of block coordinate descent method for nondifferentiable maximization. *J. Optim. Theory Appl.* **109** 473–492. MR1835069

[33] Wang, S., Nan, B., Zhou, N. and Zhu, J. (2009). Hierarchically penalized cox regression for censored data with grouped variables and its oracle property. *Biometrika* **96** 307–322. MR2507145

[34] Wei, Z. and Li, H. (2007). Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics* **8** 265–284.

[35] Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E. and Lange, K. (2009). Genomewide association analysis by lasso penalized logistic regression. *Bioinformatics* **25** 714–721.
[36] Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.* **2** 224–244. MR2415601
[37] Wu, T. T. and Lange, K. (2010). Multicategory vertex discriminant analysis for high-dimensional data. *Ann. Appl. Statist.* **4**, 4, 1698–1721. MR2829933
[38] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B* **68** 49–67. MR2212574
[39] Zhang, H. and Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika* **94**, 3, 691–703. MR2410017
[40] Zhao, P., Rocha, G. and Yu, B. (2009). Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics* **37** 3468–349. MR2549566
[41] Zhao, S. and Li, Y. (2009). Survival analysis with ultra-high dimensional covariates: identifying predictive genes for cancer disease. *Harvard University Biostatistics Working Paper Series*, 111.
[42] Zhou, N. and Zhu, J. (2007). Group variable selection via a hierarchical lasso and its oracle property. Tech. rep., Dept. Statistics, Univ. of Michigan.
[43] Zou, H. (2008). A note on path-based variable selection in the penalized proportional hazards model. *Biometrika* **95** 241–247. MR2409726
[44] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* **67** 301–320. MR2137327

Tong Tong Wu
2234B SPH Building
Department of Epidemiology & Biostatistics
University of Maryland, College Park 20742
USA
E-mail address: ttwu@umd.edu

Sijian Wang
1212 MSC
Departments of Biostatistics & Medical Informatics and Statistics
University of Wisconsin, Madison 53792
USA
E-mail address: swang@biostat.wisc.edu