Biometrika (2009), **96**, 2, *pp*. 307–322
doi: 10.1093/biomet/asp016

Printed in Great Britain

Hierarchically penalized Cox regression with grouped variables

BY S. WANG AND B. NAN

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
sijiwang@umich.edu bnan@umich.edu

N. ZHOU AND J. ZHU

Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
nfzhou@umich.edu jizhu@umich.edu

SUMMARY

In many biological and other scientific applications, predictors are often naturally grouped. For example, in biological applications, assayed genes or proteins are grouped by biological roles or biological pathways. When studying the dependence of survival outcome on these grouped predictors, it is desirable to select variables at both the group level and the within-group level. In this article, we develop a new method to address the group variable selection problem in the Cox proportional hazards model. Our method not only effectively removes unimportant groups, but also maintains the flexibility of selecting variables within the identified groups. We also show that the new method offers the potential for achieving the asymptotic oracle property.

Some key words: Cox model; Group variable selection; Lasso; Microarray; Oracle property; Regularization.

1. INTRODUCTION

A problem of interest in censored survival data analysis is to study the dependence of survival time T on a p-dimensional vector of predictors X. The proportional hazards model (Cox, 1972) is one of the most popular models in the literature and has been widely studied. The hazard function of a subject given predictors X is specified by

$$h(t \mid X) = h_0(t) \exp(\beta' X), \tag{1}$$

where $h_0(t)$ is a completely unspecified baseline hazard function and $\beta = (\beta_1, \ldots, \beta_p)'$ is an unknown vector of regression coefficients.

In practice, predictors are rarely all important, i.e. some components of β are zeros. Effective variable selection often leads to parsimonious models with better prediction accuracy and easier interpretation.

In this article, we investigate the variable selection problem in the Cox model when predictors can be naturally grouped. We are interested in selecting important groups as well as important individual variables within identified groups. We propose a hierarchically penalized Cox regression model for simultaneous variable selection at both the group and within-group levels. One motivation arises from genomic research. Genomic data can often be naturally divided into small sets based on biological knowledge. For example, when analyzing microarray gene expression data, one can group genes into functionally similar sets as in The Gene Ontology Consortium (2000),

or into known biological pathways such as the Kyoto encyclopedia of genes & genomes pathways (Kanehisa & Goto, 2002). Making use of the group information can help to identify both pathways and genes within the pathways related to the phenotypes, and hence improves understanding of biological processes.

Many variable selection methods based on penalized partial likelihood have been proposed for the Cox proportional hazards model, such as the lasso (Tibshirani, 1996, 1997) and the smoothly clipped absolute deviation (Fan & Li, 2001, 2002). By shrinking some of the regression coefficients to be exactly zero, these methods automatically remove unimportant variables. However, when the predictors are grouped, one drawback of these methods is that they treat variables individually and hence tend to perform variable selection based on the strength of the individual variables rather than the strength of the group.

The variable selection problem with grouped predictors has recently been tackled by several authors. Antoniadis & Fan (2001) and Cai (2001) discussed blockwise thresholding, P. Zhao, G. Rocha and B. Yu, in an unpublished technical report for the University of California at Berkley, proposed a method that penalizes the L_∞-norm of the coefficients within each group. Yuan & Lin (2006) extended the lasso to penalize the L_2-norm of the coefficients within each group for linear regression models, and Wang & Leng proposed an adaptive grouped lasso in their 2006 unpublished technical report from the National University of Singapore. Based on the boosting technique, Luan & Li (2008) and Wei & Li (2007), respectively, proposed 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 of these group variable selection methods have a common limitation: they select variables in an all-in-all-out fashion, i.e. when one variable in a group is selected, all other variables in the same group are also selected. Thus, these methods do not do variable selection within an identified group. 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 this article, we propose a new method based on the lasso criterion to address the group variable selection problem in the Cox model, which we call the hierarchically penalized Cox regression method. Our method not only effectively removes unimportant groups, but also maintains the flexibility of selecting variables within identified groups. Furthermore, we show that our method achieves the asymptotic oracle property of Fan & Li (2001, 2002), i.e. the method performs as well as if the correct underlying model were provided in advance.

2. VARIABLE SELECTION VIA PENALIZED PARTIAL LIKELIHOOD

We assume that the p variables in X can be divided into K groups. Let the kth group have p_k variables, denoted by $X_{(k)} = (X_{k1}, \ldots, X_{kp_k})'$, and let $\beta_{(k)} = (\beta_{k1}, \ldots, \beta_{kp_k})'$ represent the corresponding regression coefficients. We first assume that the K groups do not overlap, i.e. each variable belongs to only one group. In § 6, we allow variables to belong to multiple groups.

Suppose that a random sample of n subjects is observed. Let T_i and C_i be the survival time and the censoring time for subject i, let $Y_i = \min\{T_i, C_i\}$ be the observed time and let $\delta_i = I(T_i \leqslant C_i)$ be the censoring indicator. We use $X_{i,(k)} = (X_{i,k1}, \ldots, X_{i,kp_k})'$ to denote the p_k variables in the kth group for the ith subject, and $X_i = (X'_{i,(1)}, \ldots, X'_{i,(K)})'$ to denote the total p variables for the ith subject. We assume that T_i and C_i are conditionally independent given X_i and that the censoring mechanism is noninformative. The observed data then can be represented by the triplets $\{(Y_i, \delta_i, X_i),\ i = 1, \ldots, n\}$.

The proportional hazards model (1) can be written as

$$h(t \mid X) = h_0(t) \exp\left(\sum_{k=1}^{K} \sum_{j=1}^{p_k} \beta_{kj} X_{kj} \right) = h_0(t) \exp\{\beta'_{(1)} X_{(1)} + \cdots + \beta'_{(K)} X_{(K)}\}.$$

We consider continuous failure times and for simplicity assume that there are no ties in the observed times. Then the partial likelihood is

$$L_n(\beta) = \prod_{i \in D} \frac{\exp\left(\sum_{k=1}^{K} \beta'_{(k)} X_{i,(k)} \right)}{\sum_{l \in R_i} \exp\left(\sum_{k=1}^{K} \beta'_{(k)} X_{l,(k)} \right)},$$

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 $\ell_n(\beta)$ denote $\log\{L_n(\beta)\}$. Variable selection can be realized by maximizing the penalized log partial likelihood function

$$\frac{1}{n} \ell_n(\beta) - \sum_{k=1}^{K} \sum_{j=1}^{p_k} p_{\lambda_n}(\beta_{kj}),$$

where $p_{\lambda_n}(\beta_{kj})$ is a penalty function. Tibshirani (1997) proposed use of the lasso penalty

$$p_{\lambda_n}(\beta_{kj}) = \lambda_n |\beta_{kj}|, \tag{2}$$

and Fan & Li (2002) proposed use of the smoothly clipped absolute deviation penalty

$$p'_{\lambda_n}(|\beta_{kj}|) = I(|\beta_{kj}| \leqslant \lambda_n) + \frac{(a\lambda_n - |\beta_{kj}|)_+}{(a-1)\lambda_n} I(|\beta_{kj}| > \lambda_n) \quad (a > 2) \tag{3}$$

in the Cox regression. Fan & Li (2001) suggested using $a = 3{\cdot}7$, a value also used later in our numerical examples. Both (2) and (3) treat variables individually without taking into account the group structure.

3. Hierarchically penalized Cox regression

To make use of the group structure among the predictors, we reparameterize β_{kj} as

$$\beta_{kj} = \gamma_k \theta_{kj} \quad (k = 1, \ldots, K;\ j = 1, \ldots, p_k), \tag{4}$$

where $\gamma_k \geqslant 0$ for identifiability. This factorization reflects the information that all β_{kj} $(j = 1, \ldots, p_k)$ belong to the kth group by treating each β_{kj} hierarchically. Parameter γ_k controls all β_{kj} $(j = 1, \ldots, p_k)$ as a group at the first level of the hierarchy; the θ_{kj}s reflect differences within the kth group at the second level of the hierarchy. Let $\theta_{(k)}$ denote $(\theta_{k1}, \ldots, \theta_{kp_k})'$ for $k = 1, \ldots, K$, which gives $\beta_{(k)} = \gamma_k \theta_{(k)}$. The partial likelihood function can be written as

$$L_n(\gamma, \theta) = \prod_{i \in D} \frac{\exp\left(\sum_{k=1}^{K} \gamma_k \theta'_{(k)} X_{i,(k)} \right)}{\sum_{l \in R_i} \exp\left(\sum_{k=1}^{K} \gamma_k \theta'_{(k)} X_{l,(k)} \right)}.$$

Let $\ell_n(\gamma, \theta)$ denote $\log\{L_n(\gamma, \theta)\}$. For the purpose of variable selection, we consider the penalized log partial likelihood

$$\max_{\gamma_k, \theta_{kj}} \frac{1}{n}\ell_n(\gamma, \theta) - \lambda_\gamma \sum_{k=1}^{K} \gamma_k - \lambda_\theta \sum_{k=1}^{K}\sum_{j=1}^{p_k} |\theta_{kj}|\,, \tag{5}$$

subject to $\gamma_k \geqslant 0\,(k = 1, \ldots, K)$, where $\lambda_\gamma \geqslant 0$ and $\lambda_\theta \geqslant 0$ are two tuning parameters. Parameter λ_γ controls the estimators at the group level and can effectively remove unimportant groups: if γ_k is shrunk to zero, all β_{kj} in the kth group will be equal to zero. Parameter λ_θ controls the estimators at the variable-specific level: if γ_k is not equal to zero, some θ_{kj} and hence the corresponding β_{kj} $(j = 1, \ldots, p_k)$ can still be shrunk to zero. It is clearly seen that for fixed β and given values of λ_γ and λ_θ, the maximizer of (5), where $\ell_n(\gamma, \theta)$ is constant, is unique.

Since there are two tuning parameters in (5), it may be complicated to choose their values in practice. However, it turns out that λ_γ and λ_θ can be simplified into one tuning parameter. Specifically, with $\lambda = \lambda_\gamma \lambda_\theta$, we can show that (5) is equivalent to

$$\max_{\gamma_k, \theta_{kj}} \frac{1}{n}\ell_n(\gamma, \theta) - \sum_{k=1}^{K} \gamma_k - \lambda \sum_{k=1}^{K}\sum_{j=1}^{p_k} |\theta_{kj}|\,, \tag{6}$$

subject to $\gamma_k \geqslant 0\,(k = 1, \ldots, K)$. The meaning of equivalence is illustrated in:

LEMMA 1. *Let* $(\hat\gamma^*, \hat\theta^*)$ *be a local maximizer of* (5). *Then there exists a local maximizer* $(\hat\gamma^\dagger, \hat\theta^\dagger)$ *of* (6) *such that* $\hat\gamma_k^*\hat\theta_{kj}^* = \hat\gamma_k^\dagger\hat\theta_{kj}^\dagger$. *Similarly, if* $(\hat\gamma^\dagger, \hat\theta^\dagger)$ *is a local maximizer of* (6), *then there exists a local maximizer* $(\hat\gamma^*, \hat\theta^*)$ *of* (5) *such that* $\hat\gamma_k^*\hat\theta_{kj}^* = \hat\gamma_k^\dagger\hat\theta_{kj}^\dagger$.

Lemma 1 indicates that we only need to tune one parameter $\lambda = \lambda_\gamma\lambda_\theta$ in (6). Furthermore, criterion (6) for the hierarchically penalized Cox regression using γ_k and θ_{kj} can be written in an equivalent form using the original regression coefficients β_{kj}.

LEMMA 2. *If* $(\hat\gamma, \hat\theta)$ *is a local maximizer of* (6), *then* $\hat\beta$, *where* $\hat\beta_{kj} = \hat\gamma_k\hat\theta_{kj}$, *is a local maximizer of*

$$\max_{\beta_{kj}} \frac{1}{n}\ell_n(\beta) - 2\lambda^{1/2} \sum_{k=1}^{K} \left(\sum_{j=1}^{p_k} |\beta_{kj}|\right)^{1/2}. \tag{7}$$

On the other hand, if $\hat\beta$ *is a local maximizer of* (7), *then* $(\hat\gamma, \hat\theta)$ *is a local maximizer of* (6), *where* $\hat\gamma_k = (\lambda\sum_{j=1}^{p_k} |\hat\beta_{kj}|)^{1/2}$ *and* $\hat\theta_{kj} = \hat\beta_{kj}/\hat\gamma_k$ *if* $\hat\gamma_k \neq 0$ *and zero otherwise.*

The proofs of Lemmas 1 and 2 are given in the Appendix. As we will see in §§ 4 and 5, the numerical computation is based on (6) while the proof of asymptotic properties is based on (7). The penalty in (7) reduces to a bridge penalty (Frank & Friedman, 1993) when $p_k = 1$ for all k.

4. COMPUTATIONAL ALGORITHM

To estimate γ_k and θ_{kj} in (6), we can use an iterative approach, i.e. we first fix γ_k and estimate θ_{kj}, then fix θ_{kj} and estimate γ_k, and we iterate between these steps until convergence is achieved. Specifically, the algorithm is

Step 0. Centre and normalize X_{kj}, and obtain an initial value $\gamma_k^{(0)}$ for each γ_k; for example, $\gamma_k^{(0)} = 1$. Let $m = 1$.

Step 1. At the mth iteration, let $\tilde{X}_{i,kj} = \gamma_k^{(m-1)} X_{i,kj}$ $(k = 1, \ldots, K,\ j = 1, \ldots, p_k)$ and estimate $\theta_{kj}^{(m)}$ by

$$\theta_{kj}^{(m)} = \arg\max_{\theta_{kj}} \frac{1}{n} \log \left\{ \prod_{i \in D} \frac{\exp\left(\sum_{k=1}^{K} \sum_{j=1}^{p_k} \theta_{kj} \tilde{X}_{i,kj}\right)}{\sum_{l \in R_i} \exp\left(\sum_{k=1}^{K} \sum_{j=1}^{p_k} \theta_{kj} \tilde{X}_{l,kj}\right)} \right\} - \lambda \sum_{k=1}^{K} \sum_{j=1}^{p_k} |\theta_{kj}|.$$

Step 2. Let $\tilde{X}_{i,k} = \sum_{j=1}^{p_k} \theta_{kj}^{(m)} X_{i,kj}$ $(k = 1, \ldots, K)$ and estimate $\gamma_k^{(m)}$ by

$$\gamma_k^{(m)} = \arg\max_{\gamma_k \geqslant 0} \frac{1}{n} \log \left\{ \prod_{i \in D} \frac{\exp\left(\sum_{k=1}^{K} \gamma_k \tilde{X}_{i,k}\right)}{\sum_{l \in R_i} \exp\left(\sum_{k=1}^{K} \gamma_k \tilde{X}_{l,k}\right)} \right\} - \sum_{k=1}^{K} \gamma_k.$$

Step 3. Repeat Steps 1 and 2 until $\gamma_k^{(m)}$ and $\theta_{kj}^{(m)}$ converge. Let $\beta_{kj} = \gamma_k^{(m)} \theta_{kj}^{(m)}$ be the final solution.

Since at each step, the value of objective function (6) is nondecreasing, the algorithm always converges. Step 1 is a lasso-type problem, and we can use one of the algorithms proposed in Fan & Li (2002), Gui & Li (2005), Zhang & Lu (2007) or Park & Hastie (2007) to efficiently solve for θ_{kj}. Step 2 is a nonnegative garrote algorithm, and we can use one of the algorithms in Fan & Li (2002) or Yuan & Lin (2007) to efficiently solve for γ_k.

5. Asymptotic theory

5·1. *Preparation and general results*

To state the results in a fairly general setting, we consider the penalized log partial likelihood function with a general penalty function. Let

$$Q_n(\beta) = \frac{1}{n} \ell_n(\beta) - \sum_{k=1}^{K} p_{\lambda_n}^{(k)}(|\beta_{(k)}|), \tag{8}$$

where $p_{\lambda_n}^{(k)}(|\beta_{(k)}|) = p_{\lambda_n}^{(k)}(|\beta_{k1}|, \ldots, |\beta_{kp_k}|)$ is a general p_k-variate penalty function for parameters in the kth group, and satisfies the following two conditions:

$$p_{\lambda_n}^{(k)}(|\beta_{(k)}|) \geqslant 0 \quad (\beta_{(k)} \in R^{p_k}), \quad p_{\lambda_n}^{(k)}(0) = 0; \tag{9}$$

$$p_{\lambda_n}^{(k)}(|\beta_{(k)}|) \geqslant p_{\lambda_n}^{(k)}(|\beta_{(k)}^*|) \quad (|\beta_{kj}| \geqslant |\beta_{kj}^*|;\ j = 1, \ldots, p_k). \tag{10}$$

The penalty functions $p_{\lambda_n}^{(k)}(\cdot)$ $(k = 1, \ldots, K)$ in (8) are not necessarily the same for all groups. We also allow $p_{\lambda_n}^{(k)}(\cdot)$ to depend on the tuning parameter λ_n which varies with n.

Similar to Andersen & Gill (1982), we consider a finite time interval $[0, \tau]$ with $\tau < \infty$. The regularity conditions (A)–(D) of Andersen & Gill (1982) are also assumed here.

We write the true parameter vector as $\beta^0 = (\beta_{\mathcal{A}}^{0\prime}, \beta_{\mathcal{B}}^{0\prime}, \beta_{\mathcal{C}}^{0\prime})'$, where $\mathcal{A} = \{(k, j) : \beta_{kj}^0 \neq 0\}$, $\mathcal{B} = \{(k, j) : \beta_{kj}^0 = 0,\ \beta_{(k)}^0 \neq 0\}$, and $\mathcal{C} = \{(k, j) : \beta_{(k)}^0 = 0\}$. Here $\mathcal{A}$ contains the indices of nonzero coefficients, $\mathcal{B}$ contains the indices of zero coefficients that belong to nonzero groups, and $\mathcal{C}$ contains the indices of zero coefficients that belong to zero groups. Thus $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are disjoint and partition the set of all indices of coefficients. We write $\mathcal{D} = \mathcal{B} \cup \mathcal{C}$, which contains the indices of all zero coefficients.

Let $I(\beta^0)$ be the Fisher information matrix based on the log partial likelihood function for all $\beta = \beta^0$, and $I_{\mathcal{A}}(\beta_{\mathcal{A}}^0)$ the Fisher information matrix for $\beta_{\mathcal{A}} = \beta_{\mathcal{A}}^0$ knowing that $\beta_{\mathcal{D}}^0 = 0$. We also

define

$$a_n = \max_{(k,j)} \left\{ \frac{\partial p^{(k)}_{\lambda_n}(|\beta^0_{k1}|, \ldots, |\beta^0_{kp_k}|)}{\partial |\beta_{kj}|} : \beta^0_{kj} \neq 0 \right\},$$

$$b_n = \max_{(k,j)} \left\{ \left| \frac{\partial^2 p^{(k)}_{\lambda_n}(|\beta^0_{k1}|, \ldots, |\beta^0_{kp_k}|)}{\partial |\beta_{kj}|^2} \right| : \beta^0_{kj} \neq 0 \right\}.$$

Let $\|a\|$ denote the l_2-norm of a. The following two theorems state main asymptotic results for the penalized partial likelihood with a general penalty function.

THEOREM 1. *Under regularity conditions* (A)–(D) *of* Andersen & Gill (1982), *if* $a_n = O_p(n^{-1/2})$ *and* $b_n \to 0$, *then there exists a local maximizer* $\hat{\beta}_n$ *of* $Q_n(\beta)$ *in* (8) *such that* $\|\hat{\beta}_n - \beta^0\| = O_p(n^{-1/2})$.

THEOREM 2. *Let* $\hat{\beta}_n = (\hat{\beta}'_{n,\mathcal{A}}, \hat{\beta}'_{n,\mathcal{B}}, \hat{\beta}'_{n,\mathcal{C}})'$ *be a root-n consistent local maximizer of* $Q_n(\beta)$. *Assume that regularity conditions* (A)–(D) *of* Andersen & Gill (1982) *hold.*

(a) For $(k, j) \in \mathcal{D}$, *i.e.* $\beta^0_{kj} = 0$, *if* $n^{1/2}\partial p^{(k)}_{\lambda_n}(|\hat{\beta}_{n,k1}|, \ldots, |\hat{\beta}_{n,kp_k}|)/\partial |\beta_{kj}| \to \infty$ *as* $n \to \infty$, *then we have* $\hat{\beta}_{n,kj} = 0$ *with probability approaching* 1.

(b) For $(k, j) \in \mathcal{A}$, *i.e.* $\beta^0_{kj} \neq 0$, *if* $b_n \to 0$ *and* $n^{1/2}\partial p^{(k)}_{\lambda_n}(|\beta^0_{k1}|, \ldots, |\beta^0_{kp_k}|)/\partial |\beta_{kj}| \to 0$, *then under (a) we have that* $n^{1/2}(\hat{\beta}_{n,\mathcal{A}} - \beta^0_{\mathcal{A}})$ *converges in distribution to a zero-mean normal random variable with covariance matrix* $I_{\mathcal{A}}(\beta^0_{\mathcal{A}})^{-1}$.

Theorem 1 indicates that by choosing proper penalty functions $p^{(k)}_{\lambda_n}$ and a proper λ_n, there exists a root-n consistent penalized partial likelihood estimator. Theorem 2 implies that by choosing proper penalty functions $p^{(k)}_{\lambda_n}$ and a proper λ_n, the corresponding penalized partial likelihood estimator possesses the sparse property that $\hat{\beta}_{n,\mathcal{D}} = 0$ with probability tending to 1. Furthermore, the estimators for the nonzero coefficients, $\hat{\beta}_{n,\mathcal{A}}$, have the same asymptotic distribution as they would have if the zero coefficients were known in advance. Asymptotically, therefore, the penalized partial likelihood estimator can perform as well as if the true underlying model were provided in advance, i.e. it possesses the oracle property of Fan & Li (2001, 2002). Proofs of these theorems follow the spirit of Fan & Li (2001, 2002), but are nontrivial extensions due to the group structure of the penalty.

5·2. *Asymptotic results and further improvement*

We will show the asymptotic results for the hierarchically penalized Cox regression based on criterion (7). If we write $2\lambda^{1/2}$ in (7) as λ_n, then based on Theorems 1 and 2 we have

COROLLARY 1. *If* $\lambda_n = O_p(n^{-1/2})$, *then there exists a root-n consistent local maximizer* $\hat{\beta}_n = (\hat{\beta}'_{n,\mathcal{A}}, \hat{\beta}'_{n,\mathcal{B}}, \hat{\beta}'_{n,\mathcal{C}})'$ *for the hierarchically penalized Cox regression* (7); *if further* $n^{3/4}\lambda_n \to \infty$ *as* $n \to \infty$, *then* $\hat{\beta}_{n,\mathcal{C}} = 0$ *with probability tending to* 1.

This implies that the hierarchically penalized Cox regression can effectively remove unimportant groups. If $\mathcal{B}$ is nonempty, then the sparse property may not hold for the above root-n consistent estimator, i.e. there is no guarantee that $\hat{\beta}_{n,\mathcal{B}} = 0$ with probability approaching 1. Thus although the hierarchically penalized Cox regression can effectively remove unimportant groups, it cannot effectively remove unimportant variables within the important groups.

To tackle this limitation, we apply the adaptive idea used in Breiman (1995), Shen & Ye (2002), Wang et al. (2007), Zhang & Lu (2007), Zhao & Yu (2006), Zou (2006, 2008), and others, which is to penalize different coefficients differently. We consider

$$\max_{\beta_{kj}} \; Q_n^W(\beta) = \frac{1}{n}\ell_n(\beta) - \lambda_n \sum_{k=1}^{K} \left(\sum_{j=1}^{p_k} w_{n,kj}|\beta_{kj}| \right)^{1/2}, \tag{11}$$

where the $w_{n,kj}$ are prespecified nonnegative weights. The intuition is that if the effect of a variable is strong, we would like the corresponding weight to be small, hence the corresponding coefficient is lightly penalized. If the effect of a variable is not strong, we would like the corresponding weight to be large, hence the corresponding coefficient is heavily penalized. The next theorem shows that, by controlling weights properly, the adaptive hierarchically penalized Cox regression (11) possesses the oracle property as stated in Theorem 2.

THEOREM 3. *Define*

$$w_{n,\max}^{\mathcal{A}} = \max\{w_{n,kj} : (k,j) \in \mathcal{A}\}, \quad w_{n,\min}^{\mathcal{A}} = \min\{w_{n,kj} : (k,j) \in \mathcal{A}\};$$
$$w_{n,\max}^{\mathcal{D}} = \max\{w_{n,kj} : (k,j) \in \mathcal{D}\}, \quad w_{n,\min}^{\mathcal{D}} = \min\{w_{n,kj} : (k,j) \in \mathcal{D}\}.$$

Under regularity conditions (A)–(D) *of* Andersen & Gill (1982), *if* $n^{1/2}\lambda_n w_{n,\max}^{\mathcal{A}} w_{n,\min}^{\mathcal{A}-1/2} \to 0$, $\lambda_n w_{n,\max}^{\mathcal{A}2} w_{n,\min}^{\mathcal{A}-3/2} \to 0$, *and* $n^{1/2}\lambda_n w_{n,\min}^{\mathcal{D}}(w_{n,\max}^{\mathcal{D}} + w_{n,\max}^{\mathcal{A}})^{-1/2} \to \infty$ *as* $n \to \infty$, *then there exists a root-n consistent local maximizer* $\hat{\beta}_n$ *of* $Q_n^W(\beta)$ *such that* $\hat{\beta}_{n,\mathcal{D}} = 0$ *with probability tending to* 1 *and* $n^{1/2}(\hat{\beta}_{n,\mathcal{A}} - \beta_{\mathcal{A}}^0) \to N(0, I_{\mathcal{A}}(\beta_{\mathcal{A}}^0)^{-1})$ *in distribution.*

The remaining question, how we specify λ_n and the weights $w_{n,kj}$ so that the conditions in Theorem 3 are satisfied, is answered by the following corollary.

COROLLARY 2. *Let* $\tilde{\beta}_n$ *be a* n^{α}*-consistent estimator, i.e.* $\|\tilde{\beta}_n - \beta^0\| = O_p(n^{-\alpha})$ *with* $0 < \alpha \leqslant 1/2$. *If we choose* $\lambda_n = n^{-1/2}/\log(n)$ *and* $w_{n,kj} = 1/|\tilde{\beta}_{n,kj}|^r$, *where* $r > 0$, *then there exists a root-n consistent local maximizer* $\hat{\beta}_n$ *of* $Q_n^W(\beta)$ *such that* $\hat{\beta}_{n,\mathcal{D}} = 0$ *with probability tending to* 1 *and* $n^{1/2}(\hat{\beta}_{n,\mathcal{A}} - \beta_{\mathcal{A}}^0) \to N\{0, I_{\mathcal{A}}(\beta_{\mathcal{A}}^0)^{-1}\}$ *in distribution as* $n \to \infty$.

In practice, we can choose $\tilde{\beta}_n = \arg\max_{\beta} \ell_n(\beta)$, the estimator from the unpenalized partial likelihood when $p < n$, or $\tilde{\beta}_n = \arg\max_{\beta}\{\ell_n(\beta) - \lambda_n^* \sum_{k=1}^{K}\sum_{j=1}^{p_k} \beta_{kj}^2\}$, the ridge regression estimator with λ_n^* properly tuned when $p > n$.

6. HIERARCHICALLY PENALIZED COX REGRESSION IN THE OVERLAP CASE

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.

With slightly different notation, we reparameterize each β_j as

$$\beta_j = \theta_j \sum_{k \in G_j} \gamma_k, \quad \gamma_k \geqslant 0 \quad (k = 1, \ldots, K; j = 1, \ldots, p), \tag{12}$$

where G_j is the index set of groups to which variable X_j belongs. This is a natural generalization of the decomposition (4) for the non-overlap case. We still treat each β_j hierarchically. Parameter γ_k at the first level controls the contribution of β_j to group k, while θ_j at the second level of the

hierarchy reflects the specific effect of variable X_j. One can see that when each variable belongs to only one group, the factorization (12) reduces to (4).

Under this factorization, the corresponding partial likelihood function can be written as

$$L_n^{OL}(\gamma, \theta) = \prod_{i \in D} \frac{\exp\{\sum_{j=1}^{p} (\sum_{k \in G_j} \gamma_k)\theta_j X_{ij}\}}{\sum_{l \in R_i} \exp\{\sum_{j=1}^{p} (\sum_{k \in G_j} \gamma_k)\theta_j X_{lj}\}}.$$

Let $\ell_n^{OL}(\gamma, \theta) = \log\{L_n^{OL}(\gamma, \theta)\}$. For variable selection, we mimic (6) and consider

$$\max_{\gamma_k \geqslant 0, \theta_j} \frac{1}{n} \ell_n^{OL}(\gamma, \theta) - \sum_{k=1}^{K} \gamma_k - \lambda \sum_{j=1}^{p} |\theta_j|. \tag{13}$$

The iterative algorithm given in § 4 can then be used to estimate γ_k and θ_j in (13).

7. NUMERICAL STUDIES

7·1. *Simulations*

In this subsection, we use simulation to demonstrate hierarchically penalized Cox regression, and compare it to unpenalized Cox regression as well as the lasso approach. We also assess the adaptive hierarchically penalized Cox regression in (11), where weights w_{kj} are specified in Corollary 2 with $r = 1$.

Four examples are considered. In the first three, $p < n$; in the fourth, $p > n$. In all cases, the survival time is generated from the exponential distribution with $h(t \mid x) = \exp(\beta' x)$, and the censoring time is generated from the uniform distribution $U(0, c)$, where c is chosen to yield a 30% censoring rate. Detailed settings are given below.

Example 1. There are $n = 50$ subjects, $p = 24$ variables and $K = 3$ non-overlapping groups with eight variables in each group. In groups 1 and 2, variables are generated from $N(0, 1)$ with $\mathrm{cov}(X_{1i}, X_{1j}) = 0{\cdot}5^{|i-j|}$. In group 3, variables are generated from $N(0, 1)$ independently of one another. Variables between groups are independent. The corresponding coefficients are

$$\beta = \Big(\underbrace{1{\cdot}5, -0{\cdot}8, 0, 0, 0, 1{\cdot}2, 0, 0}_{8}, \underbrace{0}_{8}, \underbrace{0}_{8} \Big)'.$$

Example 2. There are $n = 100$ subjects, $p = 40$ variables and $K = 8$ non-overlapping groups with different group sizes. Group labels for the covariate vector are

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

We first generate $W_1, \ldots, W_{40}$ independently from the standard normal distribution and $Z_1, \ldots, Z_8$ from the standard normal distribution with $\mathrm{cov}(Z_k, Z_{k'}) = 0{\cdot}5^{|k-k'|}$, then obtain covariates by $X_{kj} = 2^{-1/2}(W_{kj} + Z_k)$. The corresponding coefficients are

$$\beta = \Big(\underbrace{1{\cdot}2, -0{\cdot}8, 1{\cdot}6, 0, 0, 0}_{6}, \underbrace{1, -0{\cdot}9, -1{\cdot}1, -1{\cdot}3}_{4}, \underbrace{0}_{6}, \underbrace{0}_{5}, \underbrace{0}_{4}, \underbrace{1{\cdot}5, 0, 0, 0, 0}_{5}, \underbrace{0}_{4}, \underbrace{0}_{6} \Big)'.$$

In this example, we have three important groups, groups 1, 2 and 6, each with at least one important variable. These groups represent three situations: many variables within the group are important, all variables within the group are important, and very few variables within the group are important.

Example 3. There are $n = 100$ subjects, $p = 48$ variables and $K = 8$ groups where some groups overlap. Group labels for variables are

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

Groups 2 and 3 have two overlapped variables, as do groups 6 and 7. We generate $X_1, \ldots, X_{24}$ from the standard normal distribution with $\mathrm{cov}(X_j, X_{j'}) = 0{\cdot}5^{|j-j'|}$, and $X_{25}, \ldots, X_{48}$ from the standard normal distribution with $\mathrm{cov}(X_j, X_{j'}) = 0{\cdot}5$. Variables $X_1, \ldots, X_{24}$ are independent of $X_{25}, \ldots, X_{48}$. The corresponding coefficients are

$$\beta = \underbrace{0}_{8}\ \overbrace{\underbrace{1{\cdot}3, 0, 1{\cdot}5, 0, -1}_{0, -1, 0, 0, 0}}^{5}\ \underbrace{0}_{8}\ \underbrace{0}_{8}\ \overbrace{\underbrace{1{\cdot}4, 0, 0{\cdot}8, 0, 1{\cdot}0}_{0, 1{\cdot}0, 0, 1{\cdot}6, 0}}^{5}\ \underbrace{0}_{8}.$$

Example 4. There are $n = 100$ subjects, $p = 148$ variables and $K = 24$ groups with different group sizes and some groups overlap. These 24 groups are divided into four independent blocks, and each block has the same structure for group assignments, but with different group labels. For example, group labels for variables in the first block are

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

In blocks 1 and 3, variables are generated from $N(0, 1)$ with $\mathrm{cov}(X_j, X_{j'}) = 0{\cdot}5^{|j-j'|}$, respectively. In blocks 2 and 4, variables are generated from $N(0, 1)$ with $\mathrm{cov}(X_j, X_{j'}) = 0{\cdot}5$, respectively. The corresponding coefficients are

$$\beta = \left(\underbrace{0}_{8}, \overbrace{\underbrace{1{\cdot}3, 0, 1{\cdot}5, 0, -1}_{0, -1, 0, 0, 0}}^{5}\ \underbrace{0}_{8}, \underbrace{1{\cdot}2, -1{\cdot}8, 0, 0, 0}_{5}, \underbrace{0}_{8}, \right.$$
$$\left. \underbrace{0}_{8}, \overbrace{\underbrace{1{\cdot}4, 0, 0{\cdot}8, 0, 1{\cdot}4}_{0, 1{\cdot}4, 0, 1{\cdot}6, 0}}^{5}\ \underbrace{0}_{8}, \underbrace{-0{\cdot}9, 1{\cdot}1, 0, 0, 0}_{5}, \underbrace{0}_{8}, \underbrace{0}_{74} \right).$$

For each of the settings above, we generate two independent data sets with the same sample size: a training set to fit the model, and a validation set to select the tuning parameter λ that maximizes the partial likelihood. Using the selected λ, we follow Tibshirani (1997) and use the model error $\mathrm{ME} = (\hat{\beta}_n - \beta)' \Sigma (\hat{\beta}_n - \beta)$ to measure the prediction accuracy, where Σ is the covariance matrix of predictors. We repeat the simulation 100 times and compute the average model errors and their corresponding standard errors. We also record two other measures: the frequencies of variables being selected and the relative biases of the coefficient estimates for

Table 1. *Variable selection frequency and relative bias for different methods*

	Frequency for $X_{\mathcal{A}}$ (%) (min, med, max)	Frequency for $X_{\mathcal{B}}$ (%) (min, med, max)	Frequency for $X_{\mathcal{C}}$ (%) (min, med, max)	Relative bias for $X_{\mathcal{A}}$ (%) (min, med, max)
Example 1				
MPLE	–	–	–	(111.5, 111.7, 114.9)
Lasso	(94, 100, 100)	(30, 37, 47)	(27, 37, 47)	(25·7, 28·3, 47·1)
HPCox	(100, 100, 100)	(67, 74, 77)	(0, 3, 5)	(1·9, 2·8, 9·3)
ALasso	(87, 100, 100)	(13, 16, 20)	(10, 15, 23)	(16·5, 20·2, 37·8)
SCAD	(82, 99, 99)	(3, 8, 9)	(4, 7, 12)	(1·1, 3·3, 6·9)
AHPCox	(94, 100, 100)	(21, 28, 38)	(0, 1, 2)	(3·2, 4·5, 12·2)
Example 2				
MPLE	–	–	–	(62·2, 65·6, 70·0)
Lasso	(100, 100, 100)	(32, 43, 51)	(35, 43, 55)	(21·6, 26·2, 38·6)
HPCox	(100, 100, 100)	(52, 62, 68)	(0, 1, 6)	(10·8, 13·0, 18·0)
ALasso	(97, 100, 100)	(9, 13, 17)	(9, 15, 22)	(12·2, 17·1, 27·9)
SCAD	(98, 99, 100)	(1, 3, 4)	(1, 5, 8)	(2·9, 7·5, 8·3)
AHPCox	(100, 100, 100)	(9, 13, 26)	(0, 1, 3)	(5·8, 8·0, 14·2)
Example 3				
MPLE	–	–	–	(97·8, 104·3, 107·2)
Lasso	(100, 100, 100)	(26, 34, 46)	(23, 37, 50)	(26·4, 28·4, 34·5)
HPCox	(100, 100, 100)	(0, 56, 80)	(0, 1, 3)	(7·6, 15·0, 16·6)
ALasso	(96, 99, 100)	(7, 16, 19)	(8, 14, 28)	(14·6, 19·4, 27·9)
SCAD	(95, 100, 100)	(1, 4, 7)	(0, 5, 12)	(1·8, 5·2, 7·2)
AHPCox	(99, 100, 100)	(0, 16, 38)	(0, 1, 2)	(3·8, 7·3, 13·5)
Example 4				
Ridge	–	–	–	(14·1, 42·2, 70·1)
Lasso	(75, 100, 100)	(12, 21, 37)	(8, 19, 35)	(42·4, 50·6, 75·7)
HPCox	(90, 99, 99)	(0, 57, 77)	(0, 1, 4)	(11·9, 22·0, 37·6)
ALasso	(77, 99, 100)	(1, 11, 36)	(2, 10, 18)	(37·1, 45·9, 78·3)
SCAD	(75, 95, 99)	(0, 1, 3)	(0, 1, 4)	(0·01, 1·9, 17·3)
AHPCox	(98, 100, 100)	(0, 20, 64)	(0, 0, 5)	(15·7, 23·1, 47·4)

$X_{\mathcal{A}}$, the collection of all important variables; $X_{\mathcal{B}}$, the collection of all unimportant variables within important groups; $X_{\mathcal{C}}$, the collection of all unimportant variables in unimportant groups; (min, med, max), minimum, median and maximum selection frequencies for variables in $X_{\mathcal{A}}$, $X_{\mathcal{B}}$ and $X_{\mathcal{C}}$, respectively, or relative biases for variables in $X_{\mathcal{A}}$; MPLE, the maximum partial likelihood estimation; Lasso, the lasso estimation; HPCox, the hierarchically penalized Cox regression method; ALasso, the adaptive lasso regression method; SCAD, the smoothly clipped absolute deviation estimation; AHPCox, the adaptive hierarchically penalized Cox regression method; Ridge, the ridge regression method.

important variables. The absolute biases of the coefficient estimates for unimportant variables are very small for all methods except the partial likelihood estimation, thus not reported. The relative biases of the coefficient estimates for important variables are calculated as $|\hat{\beta}_{n,kj} - \beta^0_{kj}|/|\beta^0_{kj}|$. The results are summarized in Table 1. The weights for both the adaptive lasso and the adaptive hierarchically penalized Cox regression methods are obtained from the usual unpenalized Cox regression model in Examples 1–3 and from the ridge regression method in Example 4.

The hierarchically penalized Cox regression and lasso methods perform similarly in identifying important variables, but the former is more effective in removing unimportant groups. The hierarchically penalized Cox regression methods tend to select more unimportant variables than the lasso method in important groups, but can be improved using the adaptive method. Results of the adaptive hierarchically penalized Cox regression method are also compared with

Table 2. *Model error for different methods*

Example	MPLE/Ridge	Lasso	HPCox	ALasso	SCAD	AHPCox
1	23·89 (7.19)	0·51 (0·02)	0·31 (0·02)	0·35 (0·03)	0·40 (0·05)	0·25 (0·02)
2	9·12 (0.61)	0·88 (0·04)	0·38 (0·02)	0·52 (0·03)	0·42 (0·05)	0·29 (0·02)
3	45·50 (7.98)	1·47 (0·07)	0·56 (0·04)	1·12 (0·19)	0·58 (0·08)	0·42 (0·04)
4	208·24 (9.00)	4·74 (0·17)	1·50 (0·11)	4·34 (0·24)	2·05 (0·32)	1·40 (0·08)

The numbers in the parentheses are corresponding standard errors. The abbreviations are the same as in Table 1.

two other methods that have the oracle property: the adaptive lasso and the smoothly clipped absolute deviation method. We can see that the adaptive lasso performs similarly to the adaptive hierarchically penalized Cox regression method in terms of frequencies of selecting important variables and excluding unimportant variables in important groups, but less effective in removing unimportant groups. The smoothly clipped absolute deviation is the most effective method in removing unimportant variables in important groups. It performs slightly worse in selecting important variables and removing unimportant groups than the adaptive hierarchically penalized Cox regression method. In terms of coefficient estimation for important variables, the adaptive hierarchically penalized Cox regression has similar biases to the hierarchically penalized Cox regression and both are better than the adaptive lasso, while the smoothly clipped absolute deviation method is the least biased and the lasso is the most biased among all the regularization methods considered.

In terms of prediction, from Table 2 we see that the adaptive hierarchically penalized Cox regression method is most accurate. The hierarchically penalized Cox regression method is the second best and much better than the lasso in all examples. That the smoothly clipped absolute deviation method performs less well comparing to the proposed methods is probably due to its more variable parameter estimation: its median variance of coefficient estimates for $X_{\mathcal{A}}$ is from about 50% to almost 200% larger than that of the proposed methods in Examples 1–3. All the considered shrinkage methods perform better than the unpenalized or the ridge Cox regression method. This illustrates that regularization should be considered for estimation and prediction when covariate effects are sparse.

7·2. *Real data example*

We analyze a breast cancer microarray gene expression dataset to demonstrate the application of the hierarchically penalized Cox regression method in identifying pathways that are related to breast cancer survival. The dataset was presented in Miller et al. (2005), where the gene expression levels were profiled on 251 frozen primary breast cancer tissues resected in Uppsala County, Sweden, from January 1, 1987 to December 31, 1989 using Affymetrix Chip HG-133A. Among these patients, 236 had follow-up information in terms of time and event of disease-specific survival. In our analysis, we use the same pathways as those in Wei & Li (2007) and Luan & Li (2008), which were obtained by merging the Affymetrix data with the cancer-related pathways provided by SuperArray on the web site http://superarray.com/. There are 245 genes in 33 cancer-related sub-pathways. See Table 2 in Luan & Li (2008) for the list of all pathways. Some genes belong to multiple pathways. Our goal is to identify the pathways that are related to survival time in breast cancer patients.

The final model is selected based on the generalized crossvalidation criterion (Fan & Li, 2002; Zhang et al., 2006; Zhang & Lu, 2007), and three pathways are identified. The three pathways are: regulation of cell cycle, in which 48 out of 75 genes are selected; cell growth and maintenance, in

which 40 out of 62 genes are selected and small GTPase-medicated signal transduction, in which 6 out of 11 genes are selected. The first two pathways were also identified by Miller et al. (2005), Wei & Li (2007) and Luan & Li (2008). In total, 82 genes are selected, among which 12 genes are shared by two pathways. The detailed gene list is available upon request.

We then use the model with the selected three pathways to predict the survival for subjects in an independent breast cancer dataset that was reported by Sotiriou et al. (2006). The dataset contains gene expressions obtained by the same microarray platform for 94 patients from the John Radcliffe Hospital in Oxford, U.K. First, we estimate the cumulative baseline hazard function using the Breslow estimator (Breslow, 1974). Then we compute the risk score $X'\hat{\beta}_n$ that yields a 50% survival probability at five years for subjects in Miller's dataset, which is chosen to be the threshold for the high and low risk groups. Finally, we compute the risk scores for subjects in Sotiriou's dataset using $\hat{\beta}_n$ obtained from Miller's dataset, and assign each subject in Sotiriou's dataset into the high- or low-risk group based on the comparison to the threshold. Among the total of 94 subjects, 26 are in the high-risk group, and 68 are in the low-risk group. Kaplan–Meier curves of these two groups are well separated with a p-value of the log-rank test less than 0·001.

ACKNOWLEDGEMENT

We thank Professor Hongzhe Li for providing the two breast cancer datasets. Bin Nan also thanks the Isaac Newton Institute for Mathematical Sciences for hosting his visits while this work was under revision. The work was partially supported by several U.S. National Science Foundation grants.

APPENDIX

Proofs

Proof of Lemma 1. In the proofs we use $|a|$ to denote the l_1-norm of a. Let $Q^*(\lambda_\gamma, \lambda_\theta, \gamma, \theta)$ denote the criterion that we would like to maximize in equation (5), let $Q^\dagger(\lambda, \gamma, \theta)$ denote the corresponding criterion in equation (6), and let $(\hat{\gamma}^*, \hat{\theta}^*)$ denote a local maximizer of $Q^*(\lambda_\gamma, \lambda_\theta, \gamma, \theta)$. We will prove that $(\hat{\gamma}^\dagger = \lambda_\gamma \hat{\gamma}^*, \hat{\theta}^\dagger = \hat{\theta}^*/\lambda_\gamma)$ is a local maximizer of $Q^\dagger(\lambda, \gamma, \theta)$.

We immediately have $Q^*(\lambda_\gamma, \lambda_\theta, \gamma, \theta) = Q^\dagger(\lambda, \lambda_\gamma \gamma, \theta/\lambda_\gamma)$. Since $(\hat{\gamma}^*, \hat{\theta}^*)$ is a local maximizer of $Q^*(\lambda_\gamma, \lambda_\theta, \gamma, \theta)$, there exists $\delta > 0$ such that if (γ', θ') satisfies $|\gamma' - \hat{\gamma}^*| + |\theta' - \hat{\theta}^*| < \delta$, then $Q^*(\lambda_\gamma, \lambda_\theta, \gamma', \theta') \leqslant Q^*(\lambda_\gamma, \lambda_\theta, \hat{\gamma}^*, \hat{\theta}^*)$.

Choose δ' such that $\delta'/\min(\lambda_\gamma, 1/\lambda_\gamma) \leqslant \delta$. Then for any (γ'', θ'') satisfying $|\gamma'' - \hat{\gamma}^\dagger| + |\theta'' - \hat{\theta}^\dagger| < \delta'$, we have

$$\left|\frac{\gamma''}{\lambda_\gamma} - \hat{\gamma}^*\right| + |\lambda_\gamma \theta'' - \hat{\theta}^*| \leqslant \frac{\lambda_\gamma \left|\frac{\gamma''}{\lambda_\gamma} - \hat{\gamma}^*\right| + \frac{1}{\lambda_\gamma}\left|\lambda_\gamma \theta'' - \hat{\theta}^*\right|}{\min\left(\lambda_\gamma, \frac{1}{\lambda_\gamma}\right)} < \frac{\delta'}{\min\left(\lambda_\gamma, \frac{1}{\lambda_\gamma}\right)} \leqslant \delta.$$

Hence

$$Q^\dagger(\lambda, \hat{\gamma}'', \hat{\theta}'') = Q^*(\lambda_\gamma, \lambda_\theta, \hat{\gamma}''/\lambda_\gamma, \lambda_\gamma \hat{\theta}'') \leqslant Q^*(\lambda_\gamma, \lambda_\theta, \hat{\gamma}^*, \hat{\theta}^*) = Q^\dagger(\lambda, \hat{\gamma}^\dagger, \hat{\theta}^\dagger).$$

Therefore, $(\hat{\gamma}^\dagger = \lambda_\gamma \hat{\gamma}^*, \hat{\theta}^\dagger = \hat{\theta}^*/\lambda_\gamma)$ is a local maximizer of $Q^\dagger(\lambda, \gamma, \theta)$.

Similarly, we can prove that for any local maximizer $(\hat{\gamma}^\dagger, \hat{\theta}^\dagger)$ of $Q^\dagger(\lambda, \gamma, \theta)$, there is a corresponding local maximizer $(\hat{\gamma}^*, \hat{\theta}^*)$ of $Q^*(\lambda_\gamma, \lambda_\theta, \gamma, \theta)$ such that $\hat{\gamma}_k^* \hat{\theta}_{kj}^* = \hat{\gamma}_k^\dagger \hat{\theta}_{kj}^\dagger$. □

Proof of Lemma 2. Suppose $(\hat{\gamma}, \hat{\theta})$ is a local maximizer of (6). Let $\hat{\beta}$ satisfy $\hat{\beta}_{kj} = \hat{\gamma}_k \hat{\theta}_{kj}$. It is trivial that $\hat{\gamma}_k = 0$ if and only if $\hat{\theta}_{(k)} = 0$. Hence if $\hat{\gamma}_k \neq 0$, then $|\hat{\beta}_{(k)}| \neq 0$, and we further have $\hat{\gamma}_k = (\lambda|\hat{\beta}_{(k)}|)^{1/2}$ and $\hat{\theta}_{(k)} = \hat{\beta}_{(k)}/\hat{\gamma}_k$, which can be shown in the following.

Let β be fixed at $\hat{\beta}$. Then $Q^\dagger(\lambda, \gamma, \theta)$ only depends on the penalty. For some k with $|\hat{\beta}_{(k)}| \neq 0$, the corresponding penalty term is $-\gamma_k - \lambda \sum_{j=1}^{p_k} |\hat{\beta}_{kj}|/\gamma_k$, which is maximized at $\hat{\gamma}_k = (\lambda|\hat{\beta}_{(k)}|)^{1/2}$.

Let $Q(\lambda, \beta)$ be the corresponding criterion in equation (7). We first show that $\hat{\beta}$ is a local maximizer of $Q(\lambda, \beta)$, i.e. there exists a $\delta' > 0$ such that if $|\triangle\beta| < \delta'$, then $Q(\lambda, \hat{\beta} + \triangle\beta) \leqslant Q(\lambda, \hat{\beta})$. Denote $\triangle\beta = \triangle\beta^{(1)} + \triangle\beta^{(2)}$, where $\triangle\beta^{(1)}_{(k)} = 0$ if $|\hat{\beta}_{(k)}| = 0$ and $\triangle\beta^{(2)}_{(k)} = 0$ if $|\hat{\beta}_{(k)}| \neq 0$. We thus have $|\triangle\beta| = |\triangle\beta^{(1)}| + |\triangle\beta^{(2)}|$.

We first show $Q(\lambda, \hat{\beta} + \triangle\beta^{(1)}) \leqslant Q(\lambda, \hat{\beta})$ for some δ'. By the argument given at the beginning of the proof, we have $\hat{\gamma}_k = (\lambda|\hat{\beta}_{(k)}|)^{1/2}$ and $\hat{\theta}_{(k)} = \hat{\beta}_{(k)}/\hat{\gamma}_k$ if $|\hat{\gamma}_k| \neq 0$, and $\hat{\theta}_{(k)} = 0$ if $|\hat{\gamma}_k| = 0$. Furthermore, let $\hat{\gamma}'_k = (\lambda|\hat{\beta}_{(k)} + \triangle\beta^{(1)}_{(k)}|)^{1/2}$ and $\hat{\theta}'_{(k)} = (\hat{\beta}_{(k)} + \triangle\beta^{(1)}_{(k)})/\hat{\gamma}'_k$ if $|\hat{\gamma}_k| \neq 0$, and let $\hat{\gamma}'_k = 0$ and $\hat{\theta}'_{(k)} = 0$ if $|\hat{\gamma}_k| = 0$. Then we have $Q^\dagger(\lambda, \hat{\gamma}', \hat{\theta}') = Q(\lambda, \hat{\beta} + \triangle\beta^{(1)})$ and $Q^\dagger(\lambda, \hat{\gamma}, \hat{\theta}) = Q(\lambda, \hat{\beta})$. Hence we only need to show $Q^\dagger(\lambda, \hat{\gamma}', \hat{\theta}') \leqslant Q^\dagger(\lambda, \hat{\gamma}, \hat{\theta})$. As $(\hat{\gamma}, \hat{\theta})$ is a local maximizer of $Q^\dagger(\lambda, \gamma, \theta)$, there exists a δ such that for any (γ', θ') satisfying $|\gamma' - \hat{\gamma}| + |\theta' - \hat{\theta}| < \delta$, we have $Q^\dagger(\lambda, \gamma', \theta') \leqslant Q^\dagger(\lambda, \hat{\gamma}, \hat{\theta})$. Straightforward calculation shows that, for $a = \min\{|\hat{\beta}_{(k)}| : |\hat{\beta}_{(k)}| \neq 0,\ k = 1, \ldots, K\}$, $b = \max\{|\hat{\beta}_{(k)}| : |\hat{\beta}_{(k)}| \neq 0,\ k = 1, \ldots, K\}$, and $\delta' < a/2$, we have

$$|\hat{\gamma}'_k - \hat{\gamma}_k| \leqslant \frac{\lambda\big|\triangle\beta^{(1)}_{(k)}\big|}{(2\lambda a)^{1/2}}, \quad \big|\hat{\theta}'_{(k)} - \hat{\theta}_{(k)}\big| \leqslant \big|\triangle\beta^{(1)}_{(k)}\big|\left\{\frac{1}{(\lambda a/2)^{1/2}} + \frac{b}{a(\lambda a)^{1/2}}\right\}.$$

Therefore, we are able to choose a δ' satisfying $\delta' < a/2$ such that $|\hat{\gamma}' - \hat{\gamma}| + |\hat{\theta}' - \hat{\theta}| < \delta$ when $|\triangle\beta^{(1)}| < \delta'$. Hence we have $Q^\dagger(\lambda, \hat{\gamma}', \hat{\theta}') \leqslant Q^\dagger(\lambda, \hat{\gamma}, \hat{\theta})$ due to the local maximality. Hence $Q(\lambda, \hat{\beta} + \triangle\beta^{(1)}) \leqslant Q(\lambda, \hat{\beta})$.

Next we show $Q(\lambda, \hat{\beta} + \triangle\beta^{(1)} + \triangle\beta^{(2)}) \leqslant Q(\lambda, \hat{\beta} + \triangle\beta^{(1)})$. This is trivial when $\triangle\beta^{(2)} = 0$. If $\triangle\beta^{(2)} \neq 0$, then $\triangle\beta^{(1)} = 0$ and we have

$$Q(\lambda, \hat{\beta} + \triangle\beta^{(1)} + \triangle\beta^{(2)}) - Q(\lambda, \hat{\beta} + \triangle\beta^{(1)}) = (\triangle\beta^{(2)})' n^{-1} \frac{\partial \ell_n(\beta^*)}{\partial\beta} - 2\sum_{k=1}^{K} \big(\lambda\big|\triangle\beta^{(2)}_{(k)}\big|\big)^{1/2},$$

where β^* is a vector between $\hat{\beta} + \triangle\beta^{(1)} + \triangle\beta^{(2)}$ and $\hat{\beta} + \triangle\beta^{(1)}$. Since $|\triangle\beta^{(2)}| < \delta'$, for a small enough δ', the second term in the above equality dominates the first term, hence we have $Q(\lambda, \hat{\beta} + \triangle\beta^{(1)} + \triangle\beta^{(2)}) \leqslant Q(\lambda, \hat{\beta} + \triangle\beta^{(1)})$. Thus we have shown that there exists a $\delta' > 0$ such that if $|\triangle\beta| < \delta'$, then $Q(\lambda, \hat{\beta} + \triangle\beta) \leqslant Q(\lambda, \hat{\beta})$, which implies that $\hat{\beta}$ is a local maximizer of $Q(\lambda, \beta)$.

Similarly, we can prove that if $\hat{\beta}$ is a local maximizer of $Q(\lambda, \beta)$, then $(\hat{\gamma}, \hat{\theta})$ is a local maximizer of $Q^\dagger(\lambda, \gamma, \theta)$, and where $\hat{\gamma}_k = (\lambda|\hat{\beta}_{(k)}|)^{1/2}$ and $\hat{\theta}_{(k)} = \hat{\beta}_{(k)}/\hat{\gamma}_k$ if $|\hat{\beta}_{(k)}| \neq 0$, and $\hat{\gamma}_k = 0$ and $\hat{\theta}_{(k)} = 0$ if $|\hat{\beta}_{(k)}| = 0$. □

Proof of Theorem 1. Let s be the number of nonzero groups. Without loss of generality, we assume that $\beta^0_{(k)} \neq 0$ $(k = 1, \ldots, s)$ and $\beta^0_{(k)} = 0$ $(k = s + 1, \ldots, K)$.

Let s_k be the number of nonzero coefficients in group k $(k = 1, \ldots, s)$. Again, without loss of generality, we assume that $\beta^0_{kj} \neq 0$ $(k = 1, \ldots, s;\ j = 1, \ldots, s_k)$ and $\beta^0_{kj} = 0$ $(k = 1, \ldots, s;\ j = s_k + 1, \ldots, p_k)$.

To prove the consistency, it is sufficient to show that for any given $\epsilon > 0$, there exists a constant C such that

$$\mathrm{pr}\left\{\sup_{\|u\|=C} Q_n\big(\beta^0 + n^{-1/2}u\big) < Q_n(\beta^0)\right\} \geqslant 1 - \epsilon. \tag{A1}$$

This implies that with a probability of at least $1 - \epsilon$, there exists a local maximum in the ball $\{\beta^0 + n^{-1/2}u : \|u\| \leqslant C\}$. Hence, there exists a local maximizer $\hat{\beta}_n$ such that $\|\hat{\beta}_n - \beta^0\| = O_p(n^{-1/2})$. Since p_{λ_n} satisfies conditions (9) and (10), we have

$$\begin{aligned} D_n(u) &= Q_n\big(\beta^0 + n^{-1/2}u\big) - Q_n(\beta^0) \\ &\leqslant n^{-1}\left\{\ell_n\big(\beta^0 + n^{-1/2}u\big) - \ell_n(\beta^0)\right\} \\ &\quad - \sum_{k=1}^{s}\left\{p^{(k)}_{\lambda_n}\big(\big|\beta^0_{k1} + n^{-1/2}u_{k1}\big|, \ldots, \big|\beta^0_{kp_k} + n^{-1/2}u_{ks_k}\big|, 0\big) - p^{(k)}_{\lambda_n}\big(\big|\beta^0_{k1}\big|, \ldots, \big|\beta^0_{ks_k}\big|, 0\big)\right\} \\ &= A - B. \end{aligned}$$

By Taylor expansion of the partial likelihood function, we have

$$\begin{aligned} A &= n^{-1/2}\left\{n^{-1/2}\frac{\partial \ell_n(\beta^0)}{\partial \beta}\right\}' n^{-1/2}u - \frac{1}{2}n^{-1}u'\left\{-n^{-1}\frac{\partial^2 \ell_n(\beta^0)}{\partial \beta^2}\right\}u + n^{-1}o_p(n^{-1}\|u\|^2) \\ &\leqslant n^{-1}O_p(1)|u| - \frac{1}{2}n^{-1}u'\{I(\beta^0)+o_p(1)\}u + n^{-1}o_p(n^{-1}\|u\|^2) \\ &\leqslant p^{1/2}n^{-1}\|u\|O_p(1) - \frac{1}{2}n^{-1}u'I(\beta^0)u + o_p(n^{-1}\|u\|^2) \ = A_1 + A_2 + A_3. \end{aligned}$$

By Taylor expansion of the penalty function, we have

$$\begin{aligned} B &= \sum_{k=1}^{s}\left\{\sum_{j=1}^{s_k}\frac{\partial p_{\lambda_n}^{(k)}(|\beta_{k1}^0|,\ldots,|\beta_{kp_k}^0|)}{\partial|\beta_{kj}|}\operatorname{sgn}(\beta_{kj}^0)n^{-1/2}u_{kj}\right. \\ &\quad \left. + \frac{1}{2}\sum_{i=1}^{s_k}\sum_{j=1}^{s_k}\frac{\partial^2 p_{\lambda_n}^{(k)}(|\beta_{k1}^0|,\ldots,|\beta_{kp_k}^0|)}{\partial|\beta_{ki}|\partial|\beta_{kj}|}n^{-1}u_{ki}u_{kj}\right\} + o_p\{n^{-1}(u_{k1}^2+\cdots+u_{ks_k}^2)\} \\ &\leqslant q_1^{1/2}n^{-1/2}a_n\|u\| + \frac{1}{2}n^{-1}b_n\|u\|^2 + o_p(n^{-1}\|u\|^2) \\ &= q_1^{1/2}\|u\|O_p(n^{-1}) + o_p(n^{-1}\|u\|^2) = B_1 + B_2, \end{aligned}$$

where $q_1 = \sum_{k=1}^{s} p_k$. We can see that, by choosing a sufficiently large C, A_2 dominates A_1, A_3, B_1, B_2 uniformly in $\|u\| = C$. Hence, (A1) holds. □

Proof of Theorem 2. First we prove the sparsity: $pr(\hat{\beta}_{n,kj} = 0) \to 1$ as $n \to \infty$ if $\beta_{kj}^0 = 0$. Using Taylor expansion, we have

$$\begin{aligned} \frac{\partial Q_n(\hat{\beta}_n)}{\partial \beta_{kj}} &= n^{-1}\frac{\partial \ell_n(\beta^0)}{\partial \beta_{kj}} + \sum_{k',j'} n^{-1}\frac{\partial^2 \ell_n(\beta^*)}{\partial \beta_{k'j'}\partial \beta_{kj}}(\hat{\beta}_{n,k'j'} - \beta_{k'j'}^0) \\ &\quad - \frac{\partial p_{\lambda_n}^{(k)}(|\hat{\beta}_{n,k1}|,\ldots,|\hat{\beta}_{n,kp_k}|)}{\partial|\beta_{kj}|}\operatorname{sgn}(\hat{\beta}_{n,kj}), \end{aligned} \tag{A2}$$

where β^* lies between $\hat{\beta}_n$ and β^0. By Theorem 3.2 in Andersen & Gill (1982) and the fact that $\hat{\beta}_n$ is a root-n consistent estimator, the first two terms on the right-hand side of (A2) are both $O_p(n^{-1/2})$. Hence we have

$$\frac{\partial Q_n(\hat{\beta}_n)}{\partial \beta_{kj}} = n^{-1/2}\left\{O_p(1) - n^{1/2}\frac{\partial p_{\lambda_n}^{(k)}(|\hat{\beta}_{n,k1}|,\ldots,|\hat{\beta}_{n,kp_k}|)}{\partial|\beta_{kj}|}\operatorname{sgn}(\hat{\beta}_{n,kj})\right\}.$$

If $n^{1/2}\partial p_{\lambda_n}^{(k)}(|\hat{\beta}_{n,k1}|,\ldots,|\hat{\beta}_{n,kp_k}|)/\partial|\beta_{kj}| \to \infty$ with probability tending to 1 as $n \to \infty$, then for an arbitrary $\epsilon > 0$ and $k = s+1,\ldots,K$, when n is large we have

$$\frac{\partial Q_n(\hat{\beta}_n)}{\partial \beta_{kj}} < 0,\ 0 < \hat{\beta}_{n,kj} < \epsilon, \qquad \frac{\partial Q_n(\hat{\beta}_n)}{\partial \beta_{kj}} > 0,\ -\epsilon < \hat{\beta}_{n,kj} < 0.$$

Therefore, $\operatorname{pr}(\hat{\beta}_{n,kj} = 0) \to 1$ as $n \to \infty$.

Second, we prove the asymptotic normality. Following Theorem 1 and the sparsity property that we just have shown, there exists a root-n consistent estimator $\hat{\beta}_n = (\hat{\beta}_{n,\mathcal{A}}, 0)'$ that satisfies the equation

$$\frac{\partial Q_n(\hat{\beta}_n)}{\partial \beta_{kj}} = 0,\ (k,j) \in \mathcal{A}.$$

Hence we have

$$
\begin{aligned}
0 &= n^{-1}\frac{\partial \ell_n(\hat{\beta}_{n,\mathcal{A}},0)}{\partial \beta_{\mathcal{A}}} - \sum_{k=1}^{s}\left\{\frac{\partial p_{\lambda_n}^{(k)}(|\hat{\beta}_{n,k1}|,\ldots,|\hat{\beta}_{n,ks_k}|,0)}{\partial|\beta_{\mathcal{A}}|}\operatorname{sgn}(\hat{\beta}_{n,\mathcal{A}})\right\}\\
&= n^{-1}\frac{\partial \ell_n(\beta^0_{\mathcal{A}},0)}{\partial \beta_{\mathcal{A}}} + n^{-1}\frac{\partial^2 \ell_n(\beta^*_{\mathcal{A}},0)}{\partial \beta^2_{\mathcal{A}}}(\hat{\beta}_{n,\mathcal{A}}-\beta^0_{\mathcal{A}})\\
&\quad - \sum_{k=1}^{s}\left\{\frac{\partial p_{\lambda_n}^{(k)}(|\beta^0_{k1}|,\ldots,|\beta^0_{ks_k}|,0)}{\partial|\beta_{\mathcal{A}}|}\operatorname{sgn}(\beta^0_{\mathcal{A}}) + \frac{\partial^2 p_{\lambda_n}^{(k)}(|\beta^0_{k1}|,\ldots,|\beta^0_{ks_k}|,0)}{\partial|\beta_{\mathcal{A}}|^2}(\hat{\beta}_{n,\mathcal{A}}-\beta^0_{\mathcal{A}})\right\}\\
&\quad + o_p(n^{-1/2}),
\end{aligned}
$$

where $\beta^*_{\mathcal{A}}$ lies between $\hat{\beta}_{n,\mathcal{A}}$ and $\beta^0_{\mathcal{A}}$. If $b_n \to 0$ and $n^{1/2}\partial p_{\lambda_n}^{(k)}(|\beta^0_{k1}|,\ldots,|\beta^0_{kp_k}|)/\partial|\beta_{kj}| \to 0$ as $n\to\infty$, it follows by Theorem 3.2 in Andersen & Gill (1982) and Slutsky's lemma that $n^{1/2}(\hat{\beta}_{\mathcal{A}}-\beta^0_{\mathcal{A}})$ converges in distribution to a normal random variable with mean zero and variance $I_1(\beta^0_{\mathcal{A}})^{-1}$. □

Proof of Corollary 1. It is straightforward to check that the corresponding conditions in Theorems 1 and 2(a) hold for the penalty function $p_{\lambda_n}^{(k)}(|\beta_{(k)}|) = \lambda_n(|\beta_{k1}|+\cdots+|\beta_{kp_k}|)^{1/2}$. The details are omitted. □

Proof of Theorem 3. We only need to check that the conditions in Theorems 1 and 2 hold for the penalty function $p_{\lambda_n}^{(k)}(|\beta_{(k)}|) = \lambda_n(w_{n,k1}|\beta_{k1}|+\cdots+w_{n,kp_k}|\beta_{kp_k}|)^{1/2}$.

First, we prove the root-n consistency. For $\beta_{kj}\in\mathcal{A}$, i.e. $\beta^0_{kj}\neq 0$, we have

$$
\begin{aligned}
a_n &= \max_{(k,j)\in\mathcal{A}}\frac{\partial p_{\lambda_n}(|\beta^0_{k1}|,\ldots,|\beta^0_{kp_k}|)}{\partial|\beta_{kj}|} = \max_{(k,j)\in\mathcal{A}}\frac{1}{2}\lambda_n w_{n,kj}\left(w_{n,k1}|\beta^0_{k1}|+\cdots+w_{n,kp_k}|\beta^0_{kp_k}|\right)^{-1/2}\\
&\leqslant \frac{1}{2}\lambda_n w^{\mathcal{A}}_{n,\max} w^{\mathcal{A}\,-1/2}_{n,\min} M^{-1/2},
\end{aligned}
$$

$$
\begin{aligned}
b_n &= \max_{(k,j)\in\mathcal{A}}\left|\frac{\partial^2 p_{\lambda_n}(|\beta^0_{k1}|,\ldots,|\beta^0_{kp_k}|)}{\partial|\beta_{kj}|^2}\right| = \frac{1}{4}\lambda_n w^2_{n,kj}\left(w_{n,k1}|\beta^0_{k1}|+\cdots+w_{n,kp_k}|\beta^0_{kp_k}|\right)^{-3/2}\\
&\leqslant \frac{1}{4}\lambda_n w^{\mathcal{A}\,2}_{n,\max} w^{\mathcal{A}\,-3/2}_{n,\min} M^{-3/2},
\end{aligned}
$$

where $M = \min_k(|\beta^0_{k1}|+\cdots+|\beta^0_{ks_k}|)$ is a finite constant. Then the consistency follows from Theorem 1.

Second, we prove the sparsity. Assume $\hat{\beta}_{n,kj}$ is any root-n consistent local maximizer of $Q^W_n(\beta)$, then we can find a constant M^*, such that $|\hat{\beta}_{n,kj}|\leqslant M^*$ for all (k,j) with probability tending to 1. Then for $(k,j)\in\mathcal{D}$, i.e. $\beta^0_{kj}=0$, we have

$$
\begin{aligned}
n^{1/2}\frac{\partial p_{\lambda_n}(|\hat{\beta}_{n,k1}|,\ldots,|\hat{\beta}_{n,kp_k}|)}{\partial|\beta_{kj}|} &= \frac{n^{1/2}\lambda_n w_{n,kj}}{2(w_{n,k1}|\hat{\beta}_{k1}|+\cdots+w_{n,kp_k}|\hat{\beta}_{kp_k}|)^{1/2}}\\
&\geqslant \frac{n^{1/2}\lambda_n w^{\mathcal{D}}_{n,\min}}{2M^{*1/2}\left(w^{\mathcal{A}}_{n,\max}+w^{\mathcal{D}}_{n,\max}\right)^{1/2}}.
\end{aligned}
$$

Therefore, when $n^{1/2}\lambda_n w^{\mathcal{D}}_{n,\min}/(w^{\mathcal{A}}_{n,\max}+w^{\mathcal{D}}_{n,\max})^{1/2}\to\infty$, by Theorem 2(a) we have $\operatorname{pr}(\hat{\beta}_{n,\mathcal{D}}=0)\to 1$.

Finally, the asymptotic normality follows from Theorem 2(b) in a manner similar to the proof of consistency. □

Proof of Corollary 2. We only need to verify that $w_{n,kj}=|\tilde{\beta}_{n,kj}|^r$ satisfy the conditions in Theorem 3. Let $A=\max\{\beta^0_{kj}\}$ and $B=\min\{\beta^0_{kj}:\beta^0_{kj}\neq 0\}$. Then by the consistency of $\tilde{\beta}_n$, it is easy to show that $w^{\mathcal{A}}_{n,\max}\to B^{-r}$ and $w^{\mathcal{A}}_{n,\min}\to A^{-r}$. Thus, when taking $\lambda_n = n^{-1/2}/\log(n)$, we have $n^{1/2}\lambda_n w^{\mathcal{A}}_{n,\max}/w^{\mathcal{A}\,1/2}_{n,\min}\to 0$ and $\lambda_n w^{\mathcal{A}\,2}_{n,\max}/w^{\mathcal{A}\,3/2}_{n,\min}\to 0$.

For each (k, j) with $\beta_{kj}^0 = 0$, we have $\tilde{\beta}_{n,kj} = O_p(n^{-\alpha})$ with $0 < \alpha \leqslant 1/2$. Therefore, $w_{n,\,\min}^{\mathcal{D}} / (w_{n,\,\max}^{\mathcal{D}} + w_{n,\,\max}^{\mathcal{A}})^{1/2} = O_p(n^{\alpha/2})$. When taking $\lambda_n = n^{-1/2}/\log(n)$, we have $n^{1/2}\lambda_n w_{n,\,\min}^{\mathcal{D}} / (w_{n,\,\max}^{\mathcal{D}} + w_{n,\,\max}^{\mathcal{A}})^{1/2} \to \infty$. □

REFERENCES

ANDERSEN, P. K. & GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100–20.

ANTONIADIS, A. & AND FAN, J. (2001). Regularization of wavelet approximations (with discussions). *J. Am. Statist. Assoc.* **96**, 939–67.

BREIMAN, L. (1995). Better subset regression using the non-negative garrote. *Technometrics* **37**, 373–84.

BRESLOW, N. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89–99.

CAI, T. (2001). Discussion of "Regularization of wavelet approximations," by Antoniadis & Fan. *J. Am. Statist. Assoc.* **96**, 960–2.

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

FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

FAN, J. & LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30**, 74–99.

FRANK, I. E. & FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35**, 109–48.

GUI, J. & 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**, 3001–8.

KANEHISA, M. & GOTO, S. (2002). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30.

LUAN, Y. & LI, H. (2008). Group additive regression models for genomic data analysis. *Biostatistics* **9**, 100–13.

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. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Nat. Acad. Sci.* **102**, 13550–5.

PARK, M. Y. & HASTIE, T. (2007). L_1-regularization path algorithm for generalized linear models. *J. R. Statist. Soc.* B **69**, 659–77.

SHEN, X. & YE, J. (2002). Adaptive model selection. *J. Am. Statist. Assoc.* **97**, 210–21.

SOTIRIOU, C., WIRAPATI, P., LOI, S., HARRIS, A., FOX, S., SMEDS, J., NORDGREN, H., FARMER, P., PRAZ, V., HAIBE-KAINS, B., DESMEDT, C., LARSIMONT, D., CARDOSO, F., PETERSE, H., NUYTEN, D., BUYSE, M., VAN DE VIJVER, M. J., BERGH, J., PICCART, M. & DELORENZI, M. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Nat. Cancer Inst.* **98**, 262–72.

THE GENE ONTOLOGY CONSORTIUM (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 259.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

TIBSHIRANI, R. (1997). The lasso method for variable selection in the Cox model. *Statist. Med.* **16**, 385–95.

WANG, H., LI, G. & TSAI, C. L. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **69**, 63–78.

WEI, Z. & LI, H. (2007). Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics* **8**, 265–84.

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

YUAN, M. & LIN, Y. (2007). On the nonnegative garrote estimator. *J. R. Statist. Soc.* B **69**, 143–61.

ZHANG, H. H., LIU, Y., WU, Y. & ZHU, J. (2006). Variable selection for multicategory SVM via sup-norm regularization. *Electron. J. Statist.* **2**, 149–67.

ZHANG, H. H. & LU, W. (2007). Adaptive-lasso for Cox's proportional hazard model. *Biometrika* **94**, 691–703.

ZHAO, P. & YU, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–63.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.

ZOU, H. (2008). A note on path-based variable selection in the penalized proportional hazards model. *Biometrika* **95**, 241–7.

[*Received December* 2007. *Revised October* 2008]