Statistica Sinica **24** (2014), 1787-1810
doi:http://dx.doi.org/10.5705/ss.2013.061

GROUP SELECTION IN THE COX MODEL WITH A DIVERGING NUMBER OF COVARIATES

Jian Huang[1], Li Liu[2], Yanyan Liu[2] and Xingqiu Zhao[3]

[1] *University of Iowa,* [2] *Wuhan University and*
[3] *The Hong Kong Polytechnic University*

Abstract: In this article, we propose a variable selection approach in the Cox model when there is a group structure in a diverging number of covariates. Most of the existing variable selection methods are designed for either individual variable selection or group selection, but not for both. The proposed methods are capable of simultaneous group selection and individual variable selection within selected groups. Computational algorithms are developed for the proposed bi-level selection methods, and the properties of the proposed selection methods are established. The proposed group bridge penalized methods are able to correctly select the important groups and variables simultaneously with high probability in sparse models. Simulation studies indicate that the proposed methods work well and two examples are provided to illustrate the applications of the proposed methods to scientific problems.

Key words and phrases: Bi-level selection, coordinate descent algorithm, Cox regression, group bridge penalty, survival data, variable selection consistency.

1. Introduction

Many survival analysis problems focus on estimating the covariate effects on the censored survival outcome. It is increasingly frequent that the number of covariates is large and that it grows as the sample size increases. In these situations, it is desirable to build a valid model selection criteria to identify the important variables and estimate their effects simultaneously. For uncensored data, many variable selection procedures have been proposed, see e.g., the least absolute shrinkage and selection operator (LASSO) (Tibshirani (1996)), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li (2001)), adaptive LASSO (Zou (2006)), the minimum concave penalty (Zhang (2010)), and the seamless-L_0 (SELO) penalty (Dicker, Huang, and Lin (2011)), among others. Some of the variable selection techniques have been extended to censored survival data. For example, Tibshirani (1997) and Zhang and Lu (2007) extended, respectively, the LASSO and the adaptive LASSO variable selection procedures to the Cox model. Fan and Li (2002) extended their nonconcave penalized likelihood approach to

the Cox model. Gui and Li (2005) developed a modified least angle regression (LARS) procedure for the LASSO estimation of Cox model. Cai et al. (2005) studied variable selection for multivariate survival data with a diverging number of covariates. Johnson (2009), Wang et al. (2009) and Du, Ma, and Liang (2010) addressed the problem by combining the LASSO, group LASSO and adaptive LASSO penalties under different models for right-censored survival data.

The aforementioned methods are designed for selecting individual variables. However, grouping structures arise naturally in many statistical modeling problems. For example, as pointed by Ma and Huang (2007), complex diseases such as cancer are often caused by mutations in gene pathways, and it is reasonable to select groups of related genes rather than individual genes. Another example is the additive model with polynomial or nonparametric components, where each component in the additive model may be expressed as a linear combination of a number of basis functions of the original measured variable. In a multi-factor analysis-of-variance (ANOVA) problem, each factor may have several levels that can be expressed through a group of dummy variables. In all these cases, the selection of important measured variables corresponds to the selection of groups of basis functions. Many authors have considered the problem of group selection under various statistical models, see e.g., Yuan and Lin (2007), Antoniadis and Fan (2001), Kim, Kim, and Kim (2006), Meier, van de Geer, and Bühlmann (2008), Zhao, Rocha, and Yu (2009) and references therein. Among others, Huang et al. (2009) considered the problem of simultaneous group and individual variable selection and proposed a group bridge method. To the best of our knowledge, only a few studies have considered group selection methods for survival data. Ma, Song, and Huang (2007) proposed the supervised group lasso for survival data. Most recently, Kim et al. (2012) extended the group LASSO approach of Yuan and Lin (2007) to the Cox model, but did not consider its theoretical properties.

Motivated by the group bridge approach of Huang et al. (2009), we consider Cox regression with a group bridge penalty. The asymptotic properties include the group oracle property, meaning that the proposed estimator consistently identifies the correct group and correct model simultaneously, and is asymptotically normal under some regularity conditions. Furthermore, our oracle selection property allows that the number of regression coefficients $d = d_n$ grows with the number of observations n, provided $d_n^4/n \to 0$.

Tuning parameter selection is an important issue in practice and we compare three selectors: AIC (Akaike (1973)), BIC (Schwarz (1978)), and GCV (Wahba (1990)) through simulation studies. All three can consistently identify the correct group. For variable selection, AIC and GCV selectors perform better than BIC; the classical BIC selector tends to underfit the model. We propose an adjusted

BIC selector by using the idea of Wang et al. (2009) for diverging parameters. Simulation studies show that the adjusted BIC selector performs as well as AIC and GCV selectors in model selection.

The remainder of the paper is organized as follows. In Section 2 we describe the group bridge method for Cox regression and present a coordinate descent algorithm for computation. Asymptotic properties are stated in Section 2, while proofs are given in the Appendix. In Section 3 the proposed selection procedures are evaluated via simulation studies, and in Section 4 we apply the proposed methods to the primary biliary cirrhosis data and breast cancer data. Some concluding remarks are made in Section 5.

2. Model Selection with Group Bridge Penalty

2.1. Cox regression with group bridge penalty

Suppose there are n independent subjects in a large study cohort. For subject i, let $\widetilde{T}_i$ be the failure time and C_i be the potential censoring time, respectively. With right-censoring, one observes the bivariate vector (T_i, δ_i), where $T_i = \min(\widetilde{T}_i, C_i)$ and $\delta_i = I(\widetilde{T}_i \le C_i)$, where $I(\cdot)$ is the indicator function. Let $Z_i(t)$ be a possibly time-dependent d_n-vector of covariates. Assume $\widetilde{T}_i$ and C_i are conditionally independent given $Z_i(\cdot)$, and that the censoring mechanism is noninformative.

Suppose the conditional hazard function of $\widetilde{T}_i$ follows the Cox proportional model

$$h(t|Z_i(s), s \le t) = h_0(t) \exp\{\beta' Z_i(t)\}, \tag{2.1}$$

where $h_0(t)$ is the unspecified baseline hazard function and $\beta = (\beta_1, \ldots, \beta_{d_n})'$ is a vector of unknown regression parameters.

Let $N_i(t) = I(T_i \le t, \delta_i = 1)$ and $Y_i(t) = I(T_i \ge t)$. Assume the process $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_n(t))'$ is left continuous with right-hand limits and satisfies $P(Y_i(t) = 1, 0 \le t \le \tau) > 0$. The negative log partial likelihood function for (2.1) is

$$l_n(\beta) = -\sum_{i=1}^{n} \delta_i \Big[\beta' Z_i(T_i) - \log\Big\{\sum_{j=1}^{n} Y_j(T_i) e^{\beta' Z_j(T_i)}\Big\}\Big].$$

Using the counting process notation, we can rewrite $l_n(\beta)$ as

$$l_n(\beta) = -\sum_{i=1}^{n} \int_0^{T_0} \Big[\beta' Z_i(t) - \log\Big(S^{(0)}(\beta, t)\Big)\Big] dN_i(t),$$

where T_0 is the end time of study and

$$S^{(k)}(\beta, t) = n^{-1} \sum_{i=1}^{n} Y_i(t) Z_i(t)^{\otimes k} \exp\{Z_i'(t)\beta\}, \text{ for } k = 0, 1, 2,$$

with $a^{\otimes k} = 1$, a, and aa' for $k = 0, 1, 2$.

We extend the group bridge approach in Huang et al. (2009) to the Cox proportional hazards model. Suppose the vector of regression coefficients is partitioned into J groups. Let $A_1, \ldots, A_J$ be subsets of $\{1, \ldots, d_n\}$ representing known groupings of the design vectors. For $S \subseteq \{1, \ldots, d_n\}$, let $|S|$ be the cardinality of S and $\beta_S = (\beta_j)_{j \in S}$ the $|S|$-dimensional sub-vector of β containing entries indexed by S. Denote the jth group by $\beta_{A_j} = (\beta_k, k \in A_j)'$. We consider a group bridge penalized partial likelihood (GBPPL) function

$$G_n(\beta) = l_n(\beta) + \lambda_n \sum_{j=1}^{J} c_j \|\beta_{A_j}\|_1^{\gamma}, \tag{2.2}$$

where $\|a\|_q$ is the L_q norm for a vector a, $\lambda_n > 0$ is the penalty level, and the c_j's are constants for the adjustment of the different dimensions of A_j. According to Huang et al. (2009), a simple choice of c_j is $c_j \propto |A_j|^{1-\gamma}$, where $|A|$ is the cardinality of A.

We call the $\widehat{\beta}$ that minimizes (2.2) a group bridge estimator (GBE). Here the groups A_j's are allowed to overlap and their union is allowed to be a proper subset of the whole so that variables not in $\bigcup_{j=1}^{J} A_j$ are not penalized. When $|A_j| = 1, j = 1, \ldots, J$, (2.2) simplifies to the standard bridge criterion. As explained below, when $0 < \gamma < 1$, the group bridge criterion (2.2) can be used for variable selection at the group and individual variable levels simultaneously.

2.2. Asymptotic properties

In this section, we present the asymptotic properties of the group bridge estimator $\widehat{\beta}$. We show that, for $0 < \gamma < 1$, the group bridge estimators correctly select groups of nonzero coefficients with probability converging to one under reasonable conditions. The asymptotic distributions of the estimators of the coefficients in nonzero groups are derived. Proofs are in the Appendix.

Without loss of generality, suppose that

$$\beta_{A_j} \neq 0,\ 1 \leq j \leq J_1, \quad \beta_{A_j} = 0,\ J_1 + 1 \leq j \leq J.$$

Let $B_2 = \bigcup_{j=J_1+1}^{J} A_j$ be the union of the groups with zero coefficients and $B_1 = B_2^c$. Assume without loss of generality that the index is arranged so that $\beta = (\beta'_{B_1}, \beta'_{B_2})'$. Denote by β_0 the true value of β, write β_{0B_1} and β_{0B_2} the true values of β with index belonging to B_1 and B_2, respectively. The true model is fully explained by the first J_1 groups since $\beta_{0B_2} = 0$.

Let $M_i(t) = N_i(t) - \int_0^t h_0(s) \exp\{\beta_0' Z_i(s)\} ds$ be the corresponding martingale for $N_i(t)$. The following conditions are needed:

(C1) $\int_0^{\tau} h_0(t) dt < \infty$;

(C2) There exists a neighborhood $\mathcal{B}$ of the true value β_0 that satisfies the following.

(i) There exist a scalar, vector, and a matrix function $s^{(l)}, l = 0, 1, 2$, defined on $\mathcal{B} \times [0, \tau]$ such that $\sup_{t\in[0,\tau],\beta\in\mathcal{B}} \|S^{(l)}(\beta; t) - s^{(l)}(\beta; t)\|_2 \to 0$ in probability;

(ii) For $l = 0, 1, 2$, the functions $s^{(l)}(\beta, t)$ are bounded and $s^{(0)}(\beta, t)$ is bounded away from zero on $\mathcal{B} \times [0, \tau]$; the family functions $s^{(l)}(\cdot, t)$ are absolutely continuous, for $\beta \in \mathcal{B}$, uniformly in $t \in [0, \tau]$.

(iii) For $e(\beta, t) = s^{(1)}(\beta, t)/s^{(0)}(\beta, t)$, $v(\beta, t) = s^{(2)}(\beta, t)/s^{(0)}(\beta, t) - (e(\beta, t))^{\otimes 2}$, and $I(\beta) = \int_0^\tau v(\beta, t)s^{(0)}(\beta_0, t)h_0(t)dt$, the Fisher information matrix $I(\beta_0)$ is positive definite.

(C3) There exists a matrix $\Gamma = \Gamma(\beta_0)$ such that $\|n^{-1}\sum_{i=1}^n Var(D_i) - \Gamma\| \to 0$, where $D_i = \int_0^\tau [Z_i(t) - e(\beta_0; t)]dM_i(t)$. There exist constants C_1 and C_2 such that $0 < C_1 < \rho_{\min}(\Gamma) \le \rho_{\max}(\Gamma) < C_2 < \infty$ for all n, where $\rho_{\min}(\Gamma)$ and $\rho_{\max}(\Gamma)$ are the minimal and maximal eigenvalues of Γ, respectively.

(C4) With D_{ij} the jth element of D_i, there exists a constant C such that $\sup_{1\le i\le n} \mathrm{E}(D_{ij}^2 D_{il}^2) < C < \infty$ for all $1 \le j, l \le d_n$.

(C5) $d_n^4/n \to 0$.

(C6) If ρ_n and ρ_n^* are the smallest and largest eigenvalues of $I(\beta_0)$. $C_n^* = \max_k \sum_{j=1}^J I(k \in A_j)$ is bounded and

$$\frac{{\lambda_n}^2}{n}\sum_{j=1}^{J_1} {c_j}^2 \|\beta_{0A_j}\|_1^{2\gamma-2}|A_j| \le d_n M_n,\ M_n = O_p(1),$$

where the constants c_j's satisfy $\min_{1\le j\le J} c_j \ge 1$ and $\lambda_n/(n^{\gamma/2}\rho_n^* d_n^{1-\gamma/2}) \to \infty$.

(C7) There exists a constant $r > 0$ such that $\rho_n > r$. For fixed unknown $\{B_1, \beta_{0B_1}, J_1\}$,

$$\lambda_n n^{-1/2} \to \lambda_0,\ \frac{1}{\rho_n} + \rho_n^* + \sum_{j=1}^J {c_j}^2 = O(1),\ \frac{\lambda_n}{n^{\gamma/2} d_n^{1-\gamma/2}} \to \infty \text{ as } n \to \infty.$$

Conditions (C1)–(C5) were required in Cai et al. (2005) for diverging d_n, which guaranteed the local asymptotic quadratic property for the partial likelihood function and hence the asymptotic normality.

Theorem 1. *Under* (C1)–(C6), *we have* $\|\widehat{\beta} - \beta_0\|_2 = O_p(\sqrt{d_n/n})$.

Theorem 2. *Suppose* (C1)–(C7) *hold. If* $\{B_1, \beta_{0B_1}, J_1\}$ *are fixed and unknown, and* $\hat{\beta}_{nB_1}$ *and* $\hat{\beta}_{nB_2}$ *are the estimators of* β_{0B_1} *and* β_{0B_2} *from* $\widehat{\beta}$ *, respectively, then the followings hold.*

(i) $Pr(\widehat{\beta}_{nB_2} = 0) \to 1$.

(ii) $\sqrt{n}(\hat{\beta}_{nB_1} - \beta_{0B_1}) \to_d argmin\{V_1(a) : a \in R^{|B_1|}\}$, *where*

$$V_1(a) = a'W + \frac{1}{2}a'I_{11}(\beta_0)a + \gamma\lambda_0 \sum_{j=1}^{J_1} c_j \|\beta_{0A_j}\|^{\gamma-1} \sum_{k\in A_j\cap B_1} \{a_k sgn(\beta_{0k})I(\beta_{0k} \neq 0) + |a_k|I(\beta_{0k} = 0)\},$$

with W *distributed as* $N(0, I_{11}(\beta_0))$, *and* $I_{11}(\beta_0)$ *the leading* $|B_1| \times |B_1|$ *submatrix of* $I(\beta_0)$ *with* $\beta_{0B_2} = 0$. *In particular, when* $\lambda_0 = 0$, $\sqrt{n}(\hat{\beta}_{nB_1} - \beta_{0B_1}) \to_d I_{11}^{-1}(\beta_0)W \sim N(0, I_{11}^{-1}(\beta_0))$.

Theorem 2 establishes the asymptotic oracle property in group selection. Further, the estimator of coefficients in non-zero groups is $\sqrt{n/d_n}$-consistent and, in general, converges to the argmin of the Gaussian process V_1.

2.3. Computation

Direct minimization of $G_n(\beta)$ is difficult, since the group bridge penalty is not a convex function for $0 < \gamma < 1$. Following Huang et al. (2009), we formulate an equivalent minimization problem that is easier to solve computationally. For $0 < \gamma < 1$, let

$$Q_n(\beta, \theta) = l_n(\beta) + \sum_{j=1}^{J} \theta_j^{1-1/\gamma} {c_j}^{1/\gamma} ||\beta_{A_j}||_1 + \tau \sum_{j=1}^{J} \theta_j,$$

where τ is a penalty parameter. The following is a direct extension of Proposition 1 in Huang et al. (2009), and the proof is omitted.

Proposition 1. *Suppose* $0 < \gamma < 1$. *If* $\lambda_n = \tau^{1-\gamma}\gamma^{-\gamma}(1-\gamma)^{\gamma-1}$, *then* $\widehat{\beta}$ *minimizes* $G_n(\beta)$ *if and only if* $(\widehat{\beta}, \widehat{\theta}_n)$ *minimizes* $Q_n(\beta, \theta)$ *subject to* $\theta > 0$ *for some* $\widehat{\theta}_n > 0$.

Take $\nabla l_n(\beta) = \partial l_n(\beta)/\partial\beta$ and $\nabla^2 l_n(\beta) = \partial^2 l_n(\beta)/\partial\beta^2$ as the gradient vector and Hessian matrix, respectively. Consider the Cholesky decomposition of $\nabla^2 l_n(\beta)$, $\nabla^2 l_n(\beta) = X'X$, and set the pseudo response vector as $Y = (X')^{-1}\{\nabla^2 l_n(\beta)\beta - \nabla l_n(\beta)\}$. Under the conditions in Section 2.2, we see that the log partial likelihood function is smooth with respect to β so that its first two partial derivatives are continuous. By the arguments in Hastie and Tibshirani (1990, pp. 213-214), $l_n(\beta)$ can be approximated by the quadratic form $(1/2)(Y - X\beta)'(Y - X\beta)$. Thus, at each step, we only need to minimize

$$\widetilde{Q}_n(\beta, \theta) = \frac{1}{2}(Y - X\beta)'(Y - X\beta) + \sum_{j=1}^{J} \theta_j^{1-1/\gamma} {c_j}^{1/\gamma} ||\beta_{A_j}||_1 + \tau \sum_{j=1}^{J} \theta_j.$$

The algorithm proceeds as followings.

Step 1. Obtain the initial value of $\beta^{(0)}$ by minimizing the negative log partial likelihood $l_n(\beta)$.

Step 2. At the kth step, compute Y, X based on the current value of $\beta^{(k)}$.

Step 3. Compute

$$\theta_j^{(k)} = c_j \left(\frac{1-\gamma}{\tau\gamma}\right)^{\gamma} \|\beta_{A_j}{}^{(k)}\|_1^{\gamma},\ j = 1, \ldots, J. \tag{2.3}$$

Step 4. Determine the new estimate by coordinate descent as

$$\beta^{(k+1)} = \operatorname{argmin}_{\beta}\Big\{\frac{1}{2}\|Y - X\beta\|_2^2 + \sum_{j=1}^{J}(\theta_j^{(k)})^{1-1/\gamma} c_j{}^{1/\gamma}\|\beta_{A_j}\|_1\Big\}. \tag{2.4}$$

Step 5. Repeat Steps 2-4 until $\|\beta^{(k+1)} - \beta^{(k)}\|/\|\beta^{(k)}\|$ is small.

The algorithm always converges since at each step the non-negative objective function $\widetilde{Q}_n(\beta, \theta)$ decreases. It returns a local minimizer of $\widetilde{Q}_n(\beta, \theta)$ depending on the initial value $\beta^{(0)}$ for a fixed tuning parameter τ, since the group bridge penalty is not convex. To speed up the implementation, we utilize warm starts. We consider a grid of τ values, $\tau_{\max} = \tau_0 > \cdots > \tau_M = 0$, for some large number M. We start with a large value of $\tau = \tau_0$ such that all the solutions are zero, and run the procedure until convergence. Then we decrease τ using the previous solution as a warm start.

The tuning parameter selection procedures is described in Section 2.5. In order to implement it, one must find the penalized partial likelihood path. We employ the efficient adaptive shrinkage method introduced by Zou (2008) to obtain group bridge penalized partial likelihood solution paths.

2.4. Variance estimation

Let $(\widehat{\beta}, \widehat{\theta}) = (\widehat{\beta}(\tau), \widehat{\theta}(\tau))$ be the proposed estimator for a fixed tuning parameter τ. Following Tibshirani (1996) and Fan and Li (2002), the standard errors of $\widehat{\beta}$ can be estimated by using a quadratic approximation. Let

$$\Pi_n(\beta, \theta) = \operatorname{diag}\Big\{\sum_{A_j \ni k} \theta_{0j}^{1-1/\gamma} c_j{}^{1/\gamma} \frac{I(\beta_{0k} \neq 0)}{\beta_{0k}}, k = 1, \ldots, d\Big\},$$
$$\Sigma_n(\beta, \theta) = \Pi_n(\beta, \theta)\beta,$$

where β_{0k} is the kth element of β, and θ_{0j} is the jth element of θ. It can be shown that

$$\Pi_n(\widehat{\beta}, \widehat{\theta}) = \operatorname{diag}\Big\{\gamma\lambda_n \sum_{A_j \ni k} c_j \|\widehat{\beta}_{nA_j}\|^{\gamma-1} \frac{I(\widehat{\beta}_{nk} \neq 0)}{|\widehat{\beta}_{nk}|}, k = 1, \ldots, d\Big\}.$$

Given the initial value of the parameter $(\beta^{(0)}, \theta^{(0)})$, the Newton-Ralphson update is

$$\beta^{(1)} = \beta^{(0)} - [\nabla^2 l_n(\beta^{(0)}) + \Pi_n(\beta^{(0)}, \theta^{(0)})]^{-1}[\nabla^2 l_n(\beta^{(0)}) + \Sigma_n(\beta^{(0)}, \theta^{(0)})].$$

Using the method of Fan and Li (2002), we can approximate the covariance matrix of the group bridge estimator $\widehat{\beta}$ by the sandwich formula

$$\nabla^2 l_n(\widehat{\beta}) + \Pi_n(\widehat{\beta}, \widehat{\theta})]^{-1}\widehat{\mathrm{Cov}(\nabla l_n(\widehat{\beta}))}[\nabla^2 l_n(\widehat{\beta}) + \Pi_n(\widehat{\beta}, \widehat{\theta})]^{-1},$$

where

$$\widehat{\mathrm{Cov}(\nabla^2 l_n(\widehat{\beta}))} = [\nabla^2 l_n(\widehat{\beta}) + \Pi_n(\widehat{\beta}, \widehat{\theta})](\nabla^2 l_n(\widehat{\beta}))^{-1}[\nabla^2 l_n(\widehat{\beta}) + \Pi_n(\widehat{\beta}, \widehat{\theta})].$$

Write $\widehat{\beta} = (\widehat{\beta}'_{C_1}, \widehat{\beta}'_{C_2})'$, where $\widehat{\beta}_{C_1}$ corresponds to the d_1 nonzero components. Correspondingly, we decompose the Hessian matrix as

$$G = \nabla^2 l_n(\widehat{\beta}) = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix},$$

where G_{11} denotes the first $d_1 \times d_1$ submatrix. Take $G_{22.1} = G_{22} - G_{21}G_{11}^{-1}G_{12}$. Similarly, let Π_{11} be the first $d_1 \times d_1$ submatrix of $\Pi = \Pi_n(\widehat{\beta}, \widehat{\theta})$, and let $\widetilde{G}_{11} = G_{11} + \Pi_{11}$. It can be shown that the covariance matrix of $\widehat{\beta}_1$ is estimated by

$$\widehat{\mathrm{Cov}(\widehat{\beta}_1)} = G_{11}^{-1} + (G_{11}^{-1} - \widetilde{G}_{11}^{-1})G_{12}G_{22.1}^{-1}G_{21}(G_{11}^{-1} - \widetilde{G}_{11}^{-1}). \tag{2.5}$$

2.5. Tuning parameter selection

The practical performance of penalized likelihood procedures depends heavily on the choice of a tuning parameter. It is often processed by finding estimators corresponding to a range of tuning parameter values. The preferred estimator is the one corresponding to a tuning parameter value which optimizes some criteria, such as AIC, BIC, or GCV. Let $\widehat{d}(\lambda_n)$ be the set of nonzero coefficients for fixed λ_n. An AIC-type criterion for choosing λ_n is

$$AIC(\lambda_n) = \log\Big\{\frac{l_n(\widehat{\beta})}{n}\Big\} + \frac{2\widehat{d}(\lambda_n)}{n},$$

while a BIC score is defined as

$$BIC(\lambda_n) = \log\Big\{\frac{l_n(\hat{\beta})}{n}\Big\} + \frac{\log(n)\widehat{d}(\lambda_n)}{n}.$$

Since the number of parameters is diverging, we make an adjustment to the BIC criterion by replacing $\log(n)$ in BIC with a positive number $k_n > \log(n)$ to obtain model selection consistency, as suggested by Wang et al. (2009). This adjusted

BIC selector is refereed as BIC_a. We use $k_n = n^{1/(2+\widehat{d}(\lambda_n))}$ in our simulations and demonstrate that this adjusted BIC selector performs well in a variety of settings.

The generalized cross-validation (GCV) function is

$$GCV(\lambda_n) = \frac{l_n(\widehat{\beta})}{n(1 - \widehat{d}(\lambda_n)/n)^2}.$$

The tuning parameter is selected via minimization of AIC(λ_n), GCV(λ_n), or BIC$_a$(λ_n).

3. Simulation Studies

We compared the proposed estimator with the ideal oracle estimator in terms of the model error,

$$ME(\widehat{\beta}) = \mathrm{E}\left[\exp(-\widehat{\beta}'Z) - \exp(-\beta_0'Z)\right]^2.$$

The ideal oracle estimator was calculated assuming the true important covariates are known to be the only covariates in the model and their coefficients are estimated by standard partial likelihood method. All the estimating procedures were carried out using tuning parameter selectors AIC, GCV, and BIC$_a$ with $k_n = n^{1/(2+d(\lambda_n))}$. The relative model error (RME) of the ideal oracle procedure versus the proposed procedure is the ratio of $ME(\widehat{\beta}_c)/ME(\widehat{\beta})$, where $\widehat{\beta}_c$ is the ideal oracle estimator. For each estimate, we recorded the average number of groups selected, the average number of variables selected, and an indicator of whether or not the model produced contains exactly the same groups and variables as the underlying model.

Failure times were generated from (2.1) with $h_0(t) = 1$ in two scenarios. In both, censoring times were Uniform($c/2$, c), where c was chosen to obtain around 20% censoring rate. We fixed $\gamma = 0.5$ and considered $n = 100$, 150, or 200. The covariates were as generated in Examples 1 and 2. Simulation results were based on 400 replications.

Example 1. There were 5 groups and 3 covariates within each group. We generated the covariates $(z_1, \ldots, z_{15})$ as follows. We first simulated $R_1, R_2, \ldots, R_{15}$ as independently standard normal variables, and $Z_1, Z_2, \ldots, Z_5$ from an AR(1) structure model with the initial standard normal distribution and $\mathrm{Cov}(Z_{j_1}, Z_{j_2}) = 0.4^{|j_1 - j_2|}$ for $j_1, j_2 = 1, \ldots, 5$. Then the $(z_1, z_2, \ldots, z_{15})$ were obtained as $z_j = (Z_{g_j} + R_j)/4 \quad (j = 1, \ldots, 15)$, where g_j is the smallest integer greater than $(j - 1)/3$ and the z_j's with the same value of g_j belonging to the same group. The true value of β was taken as $(\beta_1, \beta_2, \beta_3)' = (0.5, 1, 1.5), (\beta_4, \beta_5, \beta_6)' = (1, 1, 1), (\beta_7, \ldots, \beta_{15})' = (0, \ldots, 0)$. Thus, the coefficients within each group are either all nonzero or all zero.

Example 2. We considered a polynomial effect model where the group sizes varied and there were zero coefficients in a nonzero group. There were groups $(A_1, \ldots, A_4)$, 14 covariates with two groups each of size 4 and another two groups each of size 3. With $Z' = (Z_1', Z_2', Z_3', Z_4')$ the group covariate vector, the subvector within group j was $Z_j' = (z_{|A_1|+\cdots+|A_{j-1}|+1}, \ldots, z_{|A_1|+\cdots+|A_j|})$, $|A_j|$ the cardinality of jth group A_j. To generate the covariates $Z' = (z_1, \ldots, z_{14})$, we first generated the covariates (X_1, X_2, X_3, X_4) from an AR(1) structure with the initial standard normal distribution and $\mathrm{Cov}(X_{j_1}, X_{j_2}) = 0.4^{|j_1 - j_2|}$ for $j_1, j_2 = 1, \ldots, 4$. Then the covariate Z was obtained as $Z_j' = (X_j^1, \ldots, X_j^{|A_j|})$, $j = 1, \ldots, 4$. The true value of β was $(\beta_1, \ldots, \beta_4) = (0.5, 1, 0, 0)$, $(\beta_5, \ldots, \beta_8) = (1, 0, 0, 0)$, $(\beta_9, \beta_{10}, \beta_{11}) = (0, 0, 0)$, $(\beta_{12}, \beta_{13}, \beta_{14}) = (0, -1, 0)$.

Tables 1 and 2 summarize the group and variable selection results for Examples 1 and 2 by using the group bridge, adaptive Lasso, and group Lasso penalties. We also report the rates of correctly identifying the true group and variable in Tables 1 and 2. The results indicate that the group bridge penalized partial likelihood consistently reduces model complexity and selects the correct group and variable frequently for all considered tuning parameter selection scores. The three tuning selectors performed well in terms of the percentage of correct group selected. For the number of variables selected, AIC and GCV always selected a smaller set of variables more frequently than BIC_a. In terms of the model error, it seems that BIC_a performed better than AIC and GCV. It can be seen from the tables that the proposed method outperforms the adaptive Lasso method in group selection and the group Lasso method in variable selection.

To test the accuracy of the proposed standard error (SE) formula, we also estimated the standard deviation using (2.5) for each of the 400 simulated data sets. Results based on the three tuning selectors were similar and we only report those of AIC in Tables 3 and 4. It can be seen that the estimated standard deviation corresponding to non-zero entries approximates the sample standard deviation in a reasonable way. The accuracy of SE increases as the sample size increases.

In the following, we consider examples with a large number of covariates.

Example 3. There were ten groups, each with 4 covariates. The covariates were generated as in Example 1 with the true β taken as

$$\begin{aligned}
(\beta_1, \ldots, \beta_4)' &= (1, 0.5, 1, 0.5), (\beta_5, \ldots, \beta_8)' = (0, \ldots, 0), \\
(\beta_9, \ldots, \beta_{12})' &= (-1.5, -1.5, -1.5, -1.5), \\
(\beta_{13}, \ldots, \beta_{16})' &= (1.5, -1, -1.5, 1), \quad (\beta_{17}, \ldots, \beta_{40})' = (0, \ldots, 0).
\end{aligned}$$

Table 1. Simulation results for Example 1.

n	Penalty	Tuning	Group size	Model size	% Corr. sel	% Corr. mod	MRME
100	GBridge	AIC	2.03(0.412)	5.23(1.323)	0.935(0.148)	0.915(0.123)	0.928
		BIC_a	2.19(0.633)	5.71(1.693)	0.926(0.160)	0.911(0.132)	0.917
		GCV	2.02(0.415)	5.21(1.329)	0.935(0.147)	0.915(0.122)	0.928
	ALasso	AIC	2.57(0.685)	5.14(1.064)	0.886(0.137)	0.863(0.063)	0.823
		BIC_a	2.87(0.960)	5.67(1.848)	0.806(0.172)	0.838(0.078)	0.812
		GCV	2.55(0.672)	5.10(1.020)	0.890(0.134)	0.863(0.063)	0.823
	GLasso	AIC	1.66(0.755)	4.98(2.265)	0.752(0.233)	0.752(0.233)	0.390
		BIC_a	2.60(1.128)	7.80(3.384)	0.748(0.204)	0.748(0.204)	0.801
		GCV	1.63(0.720)	4.89(2.160)	0.754(0.236)	0.754(0.236)	0.374
150	GBridge	AIC	2.00(0.296)	5.57(0.981)	0.972(0.111)	0.955(0.098)	0.979
		BIC_a	2.10(0.442)	5.82(1.155)	0.953(0.122)	0.954(0.101)	0.979
		GCV	2.00(0.296)	5.56(0.982)	0.972(0.111)	0.954(0.098)	0.979
	ALasso	AIC	2.28(0.586)	5.17(1.042)	0.944(0.117)	0.902(0.062)	0.884
		BIC_a	2.67(0.863)	5.97(1.470)	0.866(0.173)	0.889(0.081)	0.884
		GCV	2.27(0.582)	5.14(1.032)	0.946(0.116)	0.901(0.062)	0.883
	GLasso	AIC	2.09(0.877)	6.27(2.632)	0.782(0.222)	0.782(0.222)	0.872
		BIC_a	2.97(1.020)	8.91(3.059)	0.754(0.195)	0.754(0.195)	0.907
		GCV	2.00(0.853)	6.00(2.558)	0.784(0.223)	0.784(0.223)	0.865
200	GBridge	AIC	2.02(0.200)	5.83(0.583)	0.989(0.062)	0.981(0.058)	0.998
		BIC_a	2.10(0.351)	6.03(0.680)	0.975(0.079)	0.980(0.057)	0.993
		GCV	2.02(0.200)	5.82(0.585)	0.989(0.062)	0.980(0.058)	0.998
	ALasso	AIC	2.23(0.445)	5.32(0.923)	0.954(0.089)	0.923(0.054)	0.933
		BIC_a	2.72(0.828)	6.23(0.903)	0.856(0.166)	0.903(0.074)	0.923
		GCV	2.23(0.445)	5.32(0.923)	0.954(0.089)	0.923(0.054)	0.933
	GLasso	AIC	2.63(0.939)	7.89(2.817)	0.790(0.185)	0.790(0.185)	0.945
		BIC_a	3.19(0.982)	9.57(2.944)	0.734(0.189)	0.734(0.189)	0.938
		GCV	2.60(0.943)	7.80(2.828)	0.792(0.184)	0.792(0.184)	0.945
True model			2	6	1	1	

GBridge: group bridge penalty; ALasso: adaptive Lasso penalty; GLasso: group lasso; Group size: number of groups selected; Model size: number of variables selected; % Corr. sel.: the portion of occasions on which the model produced contains exactly the same groups as the underlying model; % Corr. mod: the portion of occasions on which the model produced contains exactly the same variables as the underlying model; MRME: median of relative model errors; estimated standard errors in parentheses.

Example 4. There were six groups with three large and three small. The covariates were generated as in Example 1 with the true β taken as

$$\begin{aligned}(\beta_1,\ldots,\beta_{10})' &= (0,1,0,1,0,0,1,0,1,0),(\beta_{11},\ldots,\beta_{20})'\\ &= (-1,-1,-1,-1,0,\ldots,0),\end{aligned}$$

Table 2. Simulation results for Example 2.

n	Penalty	Tuning	Group size	Model size	% Corr. sel	% Corr. mod	MRME
100	GBridge	AIC	3.01(0.187)	5.33(1.044)	0.994(0.046)	0.897(0.075)	0.870
		BIC_a	3.07(0.265)	6.20(1.329)	0.981(0.066)	0.838(0.097)	0.786
		GCV	3.01(0.187)	5.31(1.040)	0.994(0.046)	0.898(0.076)	0.877
	ALasso	AIC	3.64(0.481)	6.36(1.360)	0.840(0.120)	0.809(0.108)	0.763
		BIC_a	3.83(0.381)	7.38(1.509)	0.794(0.095)	0.740(0.113)	0.613
		GCV	3.62(0.487)	6.31(1.334)	0.845(0.122)	0.813(0.106)	0.763
	GLasso	AIC	1.63(1.070)	6.15(3.647)	0.583(0.192)	0.539(0.104)	0.060
		BIC_a	3.35(1.029)	11.95(3.322)	0.758(0.120)	0.376(0.131)	0.229
		GCV	1.39(0.863)	5.33(2.968)	0.553(0.175)	0.558(0.082)	0.056
150	GBridge	AIC	3.01(0.100)	5.27(0.865)	0.998(0.025)	0.909(0.063)	0.901
		BIC_a	3.06(0.238)	6.07(1.171)	0.985(0.059)	0.851(0.084)	0.841
		GCV	3.01(0.100)	5.24(0.847)	0.998(0.025)	0.911(0.062)	0.906
	ALasso	AIC	3.56(0.499)	6.29(1.343)	0.860(0.125)	0.826(0.105)	0.903
		BIC_a	3.76(0.429)	7.25(1.438)	0.810(0.107)	0.759(0.110)	0.794
		GCV	3.53(0.502)	6.25(1.329)	0.868(0.125)	0.829(0.104)	0.903
	GLasso	AIC	2.61(1.399)	9.44(4.654)	0.668(0.151)	0.434(0.141)	0.210
		BIC_a	3.27(1.109)	11.68(3.601)	0.773(0.119)	0.376(0.127)	0.542
		GCV	2.40(1.385)	8.75(4.620)	0.655(0.158)	0.458(0.140)	0.115
200	GBridge	AIC	3.00(0.000)	5.13(0.843)	1.000(0.000)	0.919(0.060)	0.959
		BIC_a	3.04(0.190)	5.84(1.109)	0.991(0.048)	0.868(0.079)	0.908
		GCV	3.00(0.000)	5.11(0.838)	1.000(0.000)	0.921(0.060)	0.960
	ALasso	AIC	3.56(0.499)	6.45(1.242)	0.860(0.125)	0.821(0.092)	0.871
		BIC_a	3.78(0.416)	7.46(1.352)	0.805(0.104)	0.749(0.098)	0.845
		GCV	3.56(0.499)	6.43(1.225)	0.860(0.125)	0.822(0.091)	0.871
	GLasso	AIC	2.69(1.253)	9.82(4.108)	0.708(0.142)	0.443(0.137)	0.435
		BIC_a	3.48(0.847)	12.42(2.602)	0.755(0.094)	0.360(0.118)	0.667
		GCV	2.66(1.257)	9.72(4.127)	0.705(0.144)	0.446(0.137)	0.421
True model			3	4	1	1	

GBridge: group bridge penalty; ALasso: adaptive Lasso penalty; GLasso: group lasso; Group size: number of groups selected; Model size: number of variables selected; % Corr. sel.: the portion of occasions on which the model produced contains exactly the same groups as the underlying model; % Corr. mod: the portion of occasions on which the model produced contains exactly the same variables as the underlying model; MRME: median of relative model errors; estimated standard errors in parentheses.

$$(\beta_{21},\ldots,\beta_{30})' = (0,\ldots,0), (\beta_{31},\ldots,\beta_{34})' = (0,0,-1,1),$$
$$(\beta_{35},\ldots,\beta_{38})' = (-1,1,0,0), \qquad (\beta_{39},\ldots,\beta_{42})' = (0,\ldots,0).$$

The failure and censoring times were generated as in Examples 1 and 2 and the sample sizes $n = 100, 200$, and 300 were considered. The simulation results

Table 3. Variability of the proposed estimators for non-zero entries in simulation study under AIC selection criterion for Example 1. Mean estimated standard deviations across 400 simulated datasets and empirical standard deviation (in parentheses).

n	β_1	β_2	β_3	β_4	β_5	β_6
100	0.412	0.414	0.425	0.423	0.416	0.418
	(0.305)	(0.424)	(0.490)	(0.423)	(0.403)	(0.413)
150	0.332	0.335	0.342	0.336	0.337	0.337
	(0.268)	(0.343)	(0.405)	(0.337)	(0.350)	(0.355)
200	0.286	0.289	0.296	0.290	0.290	0.289
	(0.256)	(0.275)	(0.327)	(0.294)	(0.294)	(0.302)

Table 4. Variability of the proposed estimators for non-zero entries in simulation study under AIC selection criterion for Example 2. Mean estimated standard deviations across 400 simulated datasets and empirical standard deviation (in parentheses).

n	β_1	β_2	β_5	β_{13}
100	0.161	0.172	0.177	0.147
	(0.167)	(0.189)	(0.196)	(0.193)
150	0.127	0.135	0.139	0.120
	(0.133)	(0.133)	(0.151)	(0.149)
200	0.107	0.111	0.118	0.104
	(0.111)	(0.115)	(0.123)	(0.137)

of Examples 3 and 4 are summarized in Tables 5 and 6. They indicate that the group bridge method performs well in selecting both the correct group and variable, and gets more gains than the adaptive Lasso method in group selection and than the group Lasso method in variable selection, the same conclusions as drawn from Examples 1 and 2.

4. Applications

4.1. PBC data analysis

We illustrate the proposed group bridge method by an analysis of a dataset from the Mayo Clinic trial study given in Appendix D of Fleming and Harrington (1991). This study was conducted between 1974 and 1984 containing 312 randomized patients with primary biliary cirrhosis (PBC), a fatal chronic liver disease. Among 312 patients, 152 were assigned to the drug D-penicillanmine, while others were assigned to a control group with placebo drug. Some baseline covariates, such as age, gender, and albumin, were recorded. 140 patients died due to PBC disease during the follow-up. The primary interest was to investigate

Table 5. Simulation results for Example 3.

n	Penalty	Tuning	Group size	Model size	% Corr. sel	% Corr. mod	MRME
100	GBridge	AIC	2.64(0.597)	8.89(2.127)	0.949(0.056)	0.916(0.052)	0.597
		BIC_a	3.75(1.289)	12.48(3.445)	0.907(0.123)	0.925(0.076)	0.673
		GCV	2.58(0.582)	8.68(2.102)	0.945(0.056)	0.912(0.052)	0.576
	ALasso	AIC	5.80(1.576)	11.97(2.372)	0.716(0.157)	0.832(0.062)	0.284
		BIC_a	7.65(1.321)	16.20(3.104)	0.535(0.132)	0.768(0.079)	0.130
		GCV	5.78(1.554)	11.90(2.368)	0.718(0.155)	0.832(0.062)	0.284
	GLasso	AIC	1.33(0.514)	5.32(2.054)	0.823(0.063)	0.823(0.063)	0.171
		BIC_a	2.57(1.628)	10.28(6.514)	0.861(0.125)	0.861(0.125)	0.506
		GCV	1.27(0.468)	5.08(1.873)	0.817(0.059)	0.817(0.059)	0.166
200	GBridge	AIC	2.92(0.302)	10.94(1.394)	0.990(0.030)	0.973(0.035)	0.913
		BIC_a	3.36(0.788)	12.37(1.627)	0.965(0.079)	0.978(0.042)	0.942
		GCV	2.90(0.321)	10.87(1.443)	0.989(0.032)	0.971(0.036)	0.903
	ALasso	AIC	3.55(0.833)	10.22(1.612)	0.945(0.083)	0.927(0.031)	0.767
		BIC_a	5.84(1.650)	14.25(2.587)	0.716(0.165)	0.884(0.056)	0.666
		GCV	3.52(0.797)	10.16(1.613)	0.948(0.080)	0.927(0.031)	0.767
	GLasso	AIC	2.29(0.640)	9.16(2.561)	0.921(0.061)	0.921(0.061)	0.568
		BIC_a	3.41(1.074)	13.64(4.296)	0.917(0.084)	0.917(0.084)	0.878
		GCV	2.22(0.629)	8.88(2.516)	0.918(0.061)	0.918(0.061)	0.555
300	GBridge	AIC	3.00(0.123)	11.73(0.583)	0.999(0.012)	0.993(0.014)	0.985
		BIC_a	3.22(0.542)	12.27(0.946)	0.978(0.054)	0.990(0.023)	0.976
		GCV	2.99(0.100)	11.69(0.636)	0.999(0.010)	0.992(0.016)	0.985
	ALasso	AIC	3.43(0.714)	11.13(1.284)	0.957(0.071)	0.956(0.026)	0.888
		BIC_a	5.69(1.727)	14.64(2.423)	0.731(0.173)	0.904(0.059)	0.826
		GCV	3.40(0.667)	11.05(1.242)	0.960(0.067)	0.955(0.026)	0.888
	GLasso	AIC	2.83(0.667)	11.32(2.670)	0.959(0.055)	0.959(0.055)	0.936
		BIC_a	3.74(0.960)	14.96(3.840)	0.924(0.094)	0.924(0.094)	0.939
		GCV	2.77(0.664)	11.08(2.658)	0.957(0.055)	0.957(0.055)	0.933
True model			3	12	1	1	

GBridge: group bridge penalty; ALasso: adaptive Lasso penalty; GLasso: group lasso; Group size: number of groups selected; Model size: number of variables selected; % Corr. sel.: the portion of occasions on which the model produced contains exactly the same groups as the underlying model; % Corr. mod: the portion of occasions on which the model produced contains exactly the same variables as the underlying model; MRME: median of relative model errors; estimated standard errors in parentheses.

the effectiveness of D-penicillanmine in curing PBC disease. The PBC data have been analyzed by many authors (e.g., Zhang and Lu (2007)).

To compare with the analysis in Zhang and Lu (2007)), we focus on the main effects of the observed 17 risk factors of interest for 276 complete cases in the full model. The 17 covaraites include 10 continuous and 7 categorical variables

Table 6. Simulation results for Example 4.

n	Penalty	Tuning	Group size	Model size	% Corr. sel	% Corr. mod	MRME
100	GBridge	AIC	2.92(0.829)	10.92(2.982)	0.790(0.142)	0.808(0.056)	0.464
		BIC_a	3.68(1.042)	15.03(4.802)	0.835(0.155)	0.788(0.075)	0.374
		GCV	2.91(0.826)	10.85(2.972)	0.788(0.141)	0.808(0.056)	0.463
	ALasso	AIC	4.84(0.927)	11.25(2.400)	0.767(0.147)	0.817(0.067)	0.276
		BIC_a	4.83(0.930)	11.37(2.607)	0.763(0.138)	0.816(0.067)	0.283
		GCV	4.83(0.923)	11.25(2.394)	0.768(0.147)	0.817(0.067)	0.276
	GLasso	AIC	1.22(0.543)	11.84(4.336)	0.533(0.092)	0.657(0.030)	0.158
		BIC_a	1.45(1.158)	13.06(7.588)	0.538(0.094)	0.642(0.078)	0.154
		GCV	1.20(0.471)	11.76(4.103)	0.530(0.080)	0.657(0.030)	0.162
200	GBridge	AIC	2.95(0.704)	12.13(2.136)	0.820(0.117)	0.869(0.048)	0.652
		BIC_a	3.87(0.569)	16.09(2.353)	0.944(0.090)	0.868(0.051)	0.759
		GCV	2.92(0.696)	12.02(2.108)	0.815(0.115)	0.869(0.048)	0.645
	ALasso	AIC	4.29(0.767)	11.43(1.810)	0.888(0.113)	0.925(0.040)	0.674
		BIC_a	4.44(0.727)	12.15(1.777)	0.870(0.119)	0.923(0.041)	0.700
		GCV	4.29(0.767)	11.43(1.810)	0.888(0.113)	0.925(0.040)	0.674
	GLasso	AIC	1.19(0.394)	11.90(3.943)	0.528(0.067)	0.656(0.029)	0.106
		BIC_a	2.64(1.411)	21.12(8.186)	0.727(0.184)	0.614(0.069)	0.554
		GCV	1.18(0.386)	11.80(3.861)	0.527(0.066)	0.656(0.029)	0.105
300	GBridge	AIC	3.30(0.710)	13.12(2.129)	0.882(0.120)	0.900(0.047)	0.780
		BIC_a	3.97(0.342)	16.45(1.829)	0.981(0.054)	0.886(0.041)	0.893
		GCV	3.27(0.713)	13.03(2.082)	0.878(0.120)	0.897(0.048)	0.775
	ALasso	AIC	4.20(0.620)	12.10(1.425)	0.940(0.090)	0.964(0.027)	0.847
		BIC_a	4.56(0.592	13.29(1.282)	0.907(0.099)	0.956(0.029)	0.881
		GCV	4.19(0.615)	12.09(1.422)	0.942(0.090)	0.964(0.027)	0.847
	GLasso	AIC	1.52(0.810)	14.36(5.890)	0.587(0.135)	0.649(0.023)	0.112
		BIC_a	3.71(1.149)	27.20(5.700)	0.868(0.147)	0.586(0.085)	0.714
		GCV	1.45(0.716)	13.96(5.527)	0.575(0.119)	0.650(0.023)	0.109
True model			4	12	1	1	

GBridge: group bridge penalty; ALasso: adaptive Lasso penalty; GLasso: group lasso; Group size: number of groups selected; Model size: number of variables selected; % Corr. sel.: the portion of occasions on which the model produced contains exactly the same groups as the underlying model; % Corr. mod: the portion of occasions on which the model produced contains exactly the same variables as the underlying model; MRME: median of relative model errors; estimated standard errors in parentheses.

as described in Table 7.

The risk factors are naturally clustered into nine categories, measuring such aspects as liver reserve function and demographics. The definitions of the variables are given in Table 5.

We calculate the maximum partial likelihood estimate (MLE) and group

Table 7. PBC data analysis. Dictionary of covariates.

Group	Variable	Type	Definition
Age (G_1)	Z_1	C	Age(years)
Gender(G_2)	Z_2	D	Female gender(0 male and 1 female)
Phynotype(G_3)	Z_3	D	Ascites(0 absence)
	Z_4	D	Hepatomegaly (0 absence and 1 presence)
	Z_5	D	Spiders(0 absence and 1 presence)
	Z_6	D	Edemaoed(0 no edema, 0.5 untreated or successfully treated and 1 unsuccessfully treated)
Liver function damage (G_4)	Z_7	C	Alkaline phosphatase(units/litre)
	Z_8	C	Sgot(liver enzyme in units/ml)
Excretory function of the liver (G_5)	Z_9	C	Serum bilirubin(mg/dl)
	Z_{10}	C	Serum cholesterol(mg/dl)
	Z_{11}	C	Triglyserides(mg/dl)
Liver reserve function(G_6)	Z_{12}	C	Albumin(g/dl)
	Z_{13}	C	Prothrombin time(seconds)
Treatment (G_7)	Z_{14}	D	Penicillamine v.s. placebo (1 control and 2 treatment)
Reflection (G_8)	Z_{15}	D	Stage(histological stage of disease, graded 1,2,3 or 4)
	Z_{16}	C	Urine copper(ug/day)
Haematology (G_9)	Z_{17}	C	Platelets(per cubic ml/1000)

Type: type of variable; Type C: continuous; Type D: Discrete.

bridge estimates under the AIC, BIC_a, and GCV methods. We also show the results of Zhang and Lu (2007) obtained by using LASSO method under the GCV selector. The results are summarized in Table 8, including the estimated coefficients and the corresponding standard errors. The group bridge estimators with AIC and GCV tuning selection criteria are similar and we only present the results from GCV and BIC_a. The LASSO and group Bridge methods suggest that group (G_9, Haematology) should be excluded from the final model. Both group bridge methods suggest deleting Group 4, which includes Alkaline phosphatase and Sgot. In the phynotype group, both methods using AIC and GCV only select spiders and edema. All the methods suggest that the treatment effect is not significant at the 0.05 level and that patients with higher Serum bilirubin level have higher risk in developing PBC.

4.2. Breast cancer data analysis

The breast cancer data set containing the metastasis-free survival times was analyzed. van de Vijver et al. (2002) classified a series of 295 patients with primary breast carcinomas as having a gene-expression signature associated with either a poor or good prognosis. We restricted our study to 144 patients who

Table 8. Estimation results of PBC data with standard errors in parentheses.

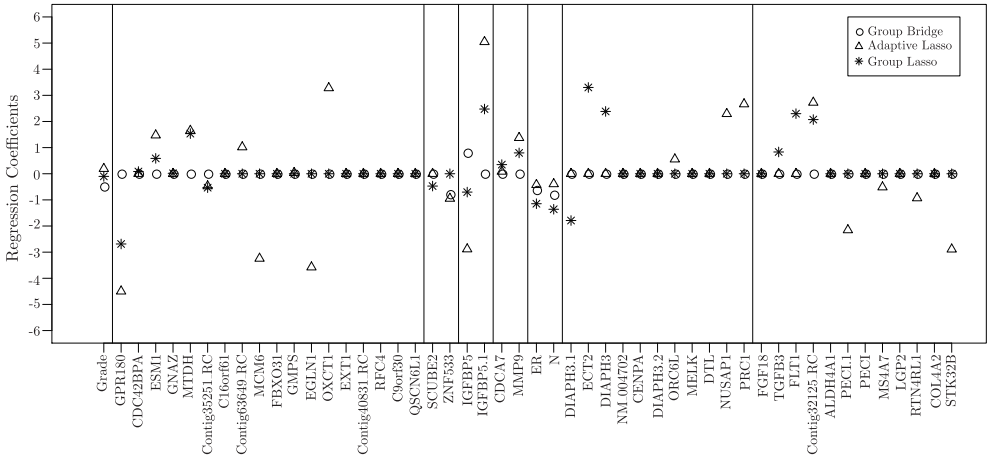
Group	Covariate	MLE	LASSO	Bridge-GCV	Bridge-BIC$_a$
G_1	age	0.029(0.012)	0.033(0.004)	0(−)	0(−)
G_2	Gender	−0.366(0.311)	0(−)	−0.793(0.269)	−0.945(0.244)
G_3	asc	0.088(0.387)	0.107(0.052)	0(−)	0.136(0.394)
	hep	0.026(0.251)	0(−)	0(−)	0.146(0.214)
	spid	0.101(0.244)	0(−)	0.021(0.239)	0.102(0.236)
	oed	1.011(0.394)	0.648(0.177)	0.489(0.402)	0.566(0.412)
G_4	alk	0.000(0.000)	0(−)	0(−)	0(−)
	sgot	0.004(0.002)	0.001(0.000)	0(−)	0(−)
G_5	bil	0.080(0.025)	0.084(0.013)	0.080(0.024)	0.060(0.023)
	chol	0.001(0.000)	0(−)	0(−)	0(−)
	trig	−0.001(0.001)	0(−)	−0.001(0.001)	0(−)
G_6	alb	−0.742(0.308)	−0.548(0.133)	−1.227(0.269)	−1.289(0.281)
	prot	0.233(0.106)	0.125(0.040)	0.128(0.107)	0.124(0.104)
G_7	trt	−0.124(0.215)	0(−)	−0.294(0.199)	−0.237(0.197)
G_8	stage	0.455(0.175)	0.265(0.064)	0.183(0.130)	0(−)
	cop	0.003(0.001)	0.003(0.001)	0.002(0.001)	0(−)
G_9	plat	0.001(0.001)	0(−)	0(−)	0(−)

MLE: maximum partial likelihood; LASSO: Lasso method; Bridge-GCV and Bridge-BIC$_a$: bridge methods by AIC and BIC$_a$ criteria.

had lymph node positive disease. The censoring rate was 66%. The data set can be found in the **R** package *'penalized'*. Five clinical risk factors and 70 gene expression measurements were diameter, d, of the tumor (1 for $>= 2cm$ and 0 for $< 2cm$), number, N, of affected lymph nodes (1 for $1-3$ and 0 for $>= 4$), estrogen receptor, ER, status (1 for positive and 0 for negative), grade of the tumor (1 for well differentiated and 0 otherwise), age of the patient at diagnosis, and the gene expression measurements of 70 prognostic genes.

We first reduced the model dimension to 50 by screening the 25 most unimportant variables. Then we divided the selected 50 variables into eight groups by using dynamic clustering. The proposed group bridge, the adaptive lasso and the group lasso, are used to analyze the data set under AIC, BIC$_a$, and GCV tuning parameters (the estimators of AIC and GCV are the same). The results with AIC and BIC$_a$ methods based on selected variables and groups are shown in Figure 1. Each block in Figure 1 represents a group. Figure 1 shows that the proposed method selects four important groups and five important variables with AIC and six important groups and eleven important variables, while adaptive Lasso and group Lasso methods select more groups and variables.

(a) AIC tuning parameter



(b) BIC_a tuning parameter

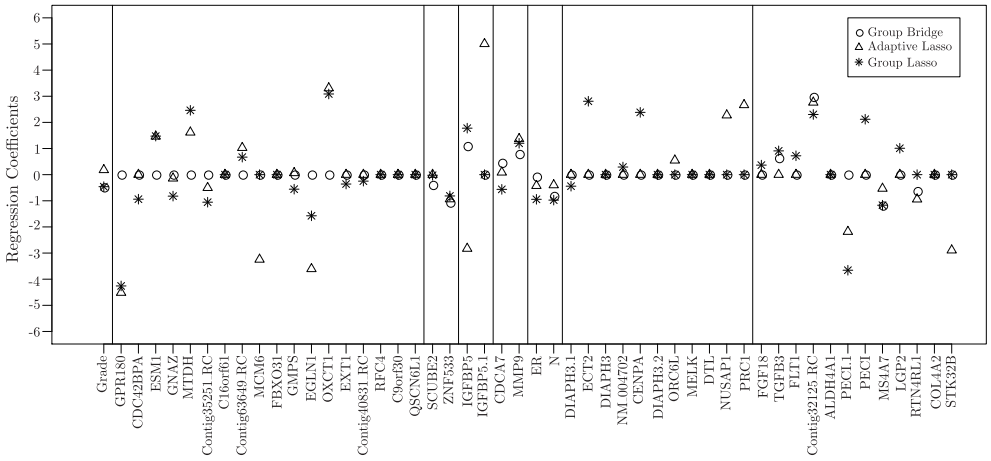


Figure 1. Plots of variable and group selection in the breast cancer data by using the proposed group bridge, the adaptive lasso and the group lasso with AIC and BIC_a.

5. Concluding Remarks

We have not addressed the situation where the number of covariates is larger than the sample size. Further research in this direction is needed for the analysis of high-dimensional survival data.

The classical Cox model assumes that covariates have a log-linear effect on the hazard function, and this can be too rigid in practice. A useful semiparametric generalization of the Cox model is the partially linear Cox model. An efficient estimator for β in this model with fixed number of parameters has been

constructed by Huang (1999). This model is useful when other covariates in addition to high-dimensional genomic data are collected in the study. If X denotes the high-dimensional expression data and W the continuous clinical covariates, we assume that W has much lower dimension than X. The conditional hazard of the failure time is modeled as

$$h(t|X,W) = h_0(t)\exp(\beta' X + \phi_1(W_1) + \cdots \phi_d(W_d)),$$

where h_0 is the unknown baseline hazard function, β is a p-dimensional regression parameter, and $\phi_1, \ldots, \phi_d$ are unknown smooth functions. For example, in the expression profiling study of follicular lymphoma reported in Dave et al. (2004), X may include the expression data, and $W = (W_1, \ldots, W_d)'$ is a vector of clinical covariates. In studies with both clinical variables and high dimensional gene measurements available, the goal is to carry out variable selection with the gene measurements X, while properly adjusting for the effects of W. The log-partial likelihood defined on sieves can be approximated by a sum of independent terms, as in Huang (1999). We expect that this will enable us to prove variable-selection consistency and the group oracle property of the bridge-penalized partial likelihood estimators.

Acknowledgements

The authors are grateful to Professor Jeng-Min Chiou, the associate editor, and the referee for their insightful comments and suggestions that greatly improved the paper. This research was supported in part by the Research Grant Council of Hong Kong (504011). L. Liu's research is supported in part by the Natural Science Foundation of China (11101315, 11171262). Y. Liu's research is supported in part by the Natural Science Foundation of China (11171263, 11371299) and Doctoral Fund of Ministry of Education of China (RFDP20110141110004).

Appendix: Proofs

Proof of Theorem 1. Let $B(C) = \{\beta : \beta = \beta_0 + C\alpha_n u, \|u\|_2 = 1\}$. It is sufficient to prove that, for every $\varepsilon > 0$, there exists $B(C)$ such that $Pr(\inf_{\beta \in B(C)} G_n(\beta) > G_n(\beta_0)) > 1 - \varepsilon$. Since

$$\frac{1}{n}G_n(\beta) - \frac{1}{n}G_n(\beta_0) = \frac{1}{n}[l_n(\beta) - l_n(\beta_0)] + D_n(C),$$

where $D_n(C) = (\lambda_n/n)\Big(\sum_{j=1}^J c_j \|\beta_{A_j}\|_1^\gamma - \sum_{j=1}^J c_j \|\beta_{0A_j}\|_1^\gamma\Big)$ for $\beta \in B(C)$, by Taylor's expansion and results in Cai et al. (2005), we have

$$\frac{1}{n}[l_n(\beta) - l_n(\beta_0)] = \frac{1}{n}(\beta - \beta_0)'\nabla l_n(\beta_0) + \frac{1}{2n}(\beta - \beta_0)'\nabla^2 l_n(\beta^*)(\beta - \beta_0)$$

$$= B_1 + B_2,$$

where β^* is between β and β_0. By the arguments in Cai et al. (2005), $\|\nabla l_n(\beta_0)\|_2 = O_p(\sqrt{nd_n})$ and $n^{-1}\nabla^2 l_n(\beta^*) = I(\beta_0) + o_p(1)$. It follows that B_1 is of order $C\alpha_n^2$ and B_2 is of order $C^2\alpha_n^2$ by the assumption that $I(\beta_0)$ is positive definite. By choosing sufficiently large C, I_2 dominates I_1 uniformly in $\|u\|_2 = 1$.

For the lower bound of $D_n(C)$, it is sufficient to consider the case where $\|\beta_{0A_j}\|_1 \geq \|\beta_{A_j}\|_1$. Since $b^\gamma - a^\gamma \leq 2(b-a)b^{\gamma-1}$ for $0 \leq a \leq b$, it follows that

$$\sum_{j=1}^{J} c_j\|\beta_{0A_j}\|_1^\gamma - \sum_{j=1}^{J} c_j\|\beta_{A_j}\|_1^\gamma \leq 2\sum_{j=1}^{J_1} c_j\|\beta_{0A_j}\|_1^{\gamma-1}(|A_j|\|\beta_{A_j} - \beta_{0A_j}\|_2^2)^{1/2}$$
$$\leq 2\eta_n(\sum_{j=1}^{J}\|\beta_{A_j} - \beta_{0A_j}\|_2^2)^{1/2} \leq 2\eta_n\sqrt{C_n^*}\|\widehat{\beta} - \beta_0\|_2,$$

where $\eta_n^2 = \sum_{j=1}^{J_1} {c_j}^2\|\beta_{0A_j}\|_1^{2\gamma-2}|A_j|$. Therefore,

$$\frac{1}{n}[G_n(\beta_0 + C\alpha_n u) - G_n(\beta_0)] \geq \frac{1}{2}C^2\alpha_n^2 u'[I(\beta) + o_p(1)]u + O_p(C\alpha_n^2) - 2\frac{\lambda_n}{n}\eta_n\sqrt{C_n^*}C\alpha_n. \tag{A.1}$$

Since $-2(\lambda_n/n)\eta_n\sqrt{C_n^*}C\alpha_n$ is of order $C\alpha_n^2$, the first term of the right side of (A.1) dominates the third term uniformly in $\|u\|_2 = 1$, when C is large enough. This completes the proof of Theorem 1.

Proof of Theorem 2 (i). Let $B_2 = \bigcup_{j=J_1+1}^{J} A_j$ and take $\widetilde{\beta}_n = (\widetilde{\beta}_{n1}, \ldots, \widetilde{\beta}_{nd})'$, with $\widetilde{\beta}_{nk} = \widehat{\beta}_{nk}$ if $k \notin B_2$, and 0 otherwise. The KKT condition for (2.4) implies that

$$-\left(\nabla l_n(\hat{\beta})\right)_k = \sum_{j:A_j \ni k} \widehat{\theta}_{nj}^{1-1/\gamma} {c_j}^{1/\gamma}\mathrm{sgn}(\widehat{\beta}_{nk}),\ \forall \widehat{\beta}_{nk} \neq 0, \tag{A.2}$$

where $(a)_k$ is the kth element of vector a.

By (2.3) and the relationship between λ_n and θ_n, we have $\widehat{\theta}_{nj}^{1-1/\gamma} {c_j}^{1/\gamma}\|\widehat{\beta}_{nA_j}\|_1 = \gamma\lambda_n\|\widehat{\beta}_{nA_j}\|_1^\gamma$. Therefore (A.2) implies

$$-\left(\nabla l_n(\hat{\beta})\right)_k = \lambda_n\gamma \sum_{j:A_j \ni k} c_j\|\widehat{\beta}_{nA_j}\|_1^{\gamma-1}\mathrm{sgn}(\widehat{\beta}_{nk}),\quad \forall \widehat{\beta}_{nk} \neq 0.$$

Since $(\widehat{\beta}_{nk} - \widetilde{\beta}_{nk})\mathrm{sgn}(\widehat{\beta}_{nk}) = |\widehat{\beta}_{nk}|I(k \in B_2)$, we have

$$-\left(\nabla l_n(\hat{\beta})\right)'(\widehat{\beta} - \widetilde{\beta}_n) = \sum_{k \in B_2} |\widehat{\beta}_{nk}|\gamma\lambda_n \sum_{j:A_j \ni k} c_j\|\widehat{\beta}_{nA_j}\|_1^{\gamma-1}$$
$$= \lambda_n\gamma\sum_{j=1}^{J} c_j\|\widehat{\beta}_{nA_j}\|_1^{\gamma-1}(\|\widehat{\beta}_{nA_j}\|_1 - \|\widetilde{\beta}_{nA_j}\|_1) = \lambda_n\gamma\sum_{j=J_1+1}^{J} c_j\|\widehat{\beta}_{nA_j}\|_1^\gamma. \tag{A.3}$$

By the definition of $\widehat{\beta}$, we have $G_n(\widehat{\beta}) \geq G_n(\widetilde{\beta}_n)$, or

$$l_n(\widehat{\beta}) + \lambda_n \sum_{j=1}^{J} c_j \|\widehat{\beta}_{nA_j}\|_1^{\gamma} \leq l_n(\widetilde{\beta}_n) + \lambda_n \sum_{j=1}^{J} c_j \|\widetilde{\beta}_{nA_j}\|_1^{\gamma}.$$

Since $\|\widehat{\beta}_{nA_j}\|_1 = 0$ for $j > J_1$, by (A.3) we have

$$\begin{aligned}
&-\frac{1}{n}\left(\nabla l_n(\hat{\beta})\right)'(\widehat{\beta} - \widetilde{\beta}_n) + (1-\gamma)\frac{\lambda_n}{n}\sum_{j=J_1+1}^{J} c_j \|\widehat{\beta}_{nA_j}\|_1^{\gamma} \\
&\quad = \frac{\lambda_n}{n}\sum_{j=1}^{J} c_j \|\widehat{\beta}_{nA_j}\|_1^{\gamma} - \frac{\lambda_n}{n}\sum_{j=1}^{J} c_j \|\widetilde{\beta}_{nA_j}\|_1^{\gamma} \\
&\quad \leq \frac{1}{n}\left[l_n(\widetilde{\beta}_n) - l_n(\widehat{\beta})\right] \\
&\quad = -\frac{1}{n}\left(\nabla l_n(\hat{\beta})\right)'(\widehat{\beta} - \widetilde{\beta}_n) + (\widehat{\beta} - \widetilde{\beta}_n)'[I(\beta_0) + o_p(1)](\widehat{\beta} - \widetilde{\beta}_n) + o_p(\|\widehat{\beta} - \widetilde{\beta}_n\|_2^2).
\end{aligned}$$

It follows that

$$\begin{aligned}
&(1-\gamma)\frac{\lambda_n}{n}\sum_{j=J_1+1}^{J} c_j \|\widehat{\beta}_{nA_j}\|_1^{\gamma} \\
&\quad \leq (\widehat{\beta} - \widetilde{\beta}_n)'[I(\beta_0) + o_p(1)](\widehat{\beta} - \widetilde{\beta}_n) + o_p(\|\widehat{\beta} - \widetilde{\beta}_n\|_2^2) \\
&\quad = (\widehat{\beta} - \widetilde{\beta}_n)'[I(\beta_0)](\widehat{\beta} - \widetilde{\beta}_n) + o_p(\|\widehat{\beta} - \widetilde{\beta}_n\|_2^2). \qquad \text{(A.4)}
\end{aligned}$$

Since the first term of (A.4) dominates the second term for n large enough, with such n, we have

$$(1-\gamma)\frac{\lambda_n}{n}\sum_{j=J_1+1}^{J} c_j \|\widehat{\beta}_{nA_j}\|_1^{\gamma} \leq 2\rho_n^* \|\widehat{\beta} - \widetilde{\beta}_n\|_2^2 = 2\rho_n^* \|\widehat{\beta}_{nB_2}\|_2^2 \leq 2\rho_n^* \|\widehat{\beta} - \beta_0\|_2^2.$$

It follows that

$$(1-\gamma)\lambda_n \sum_{j=J_1+1}^{J} c_j \|\widehat{\beta}_{nA_j}\|_1^{\gamma} \leq 2n\rho_n^* \|\widehat{\beta} - \beta_0\|_2^2 \leq O_p(d_n \rho_n^*).$$

For the lower bound of $\sum_{j=J_1+1}^{J} c_j \|\widehat{\beta}_{nA_j}\|_1^{\gamma}$, since $c_j > 1$ by (C6), we have

$$\sum_{j=J_1+1}^{J} c_j \|\widehat{\beta}_{nA_j}\|_1^{\gamma} \geq \Big(\sum_{j=J_1+1}^{J} \|\widehat{\beta}_{nA_j}\|_1\Big)^{\gamma} \geq \|\widehat{\beta}_{nB_2}\|_1^{\gamma} \geq \|\widehat{\beta}_{nB_2}\|_2^{\gamma}.$$

If $\|\widehat{\beta}_{nB_2}\|_2^{\gamma} > 0$, then

$$(1-\gamma)\lambda_n \leq 2n\rho_n^* \|\widehat{\beta}_{nB_2}\|_2^{2-\gamma} \leq \rho_n^* d_n^{1-\gamma/2} n^{\gamma/2} O_p(1).$$

It follows that $\dfrac{\lambda_n}{n^{\gamma/2}\rho_n^* d_n^{1-\gamma/2}} \le O_p(1)$. Since $\dfrac{\lambda_n}{n^{\gamma/2}\rho_n^* d_n^{1-\gamma/2}} \to \infty$ by (C6),

$$Pr\left(\|\widehat{\beta}_{nB_2}\|_2 > 0\right) \le Pr\Big(\frac{\lambda_n}{n^{\gamma/2}\rho_n^* d_n^{1-\gamma/2}} \le O_p(1)\Big) \to 0.$$

(ii). Since d_1, β_{0B_1} are fixed, $\min_{j\le J_1} \|\beta_{0A_j}\|_1^{1-\gamma} = O(1)$, so that (C7) implies (C6) and

$$\frac{{\lambda_n}^2}{n} \sum_{j=1}^{J_1} {c_j}^2 \|\beta_{0A_j}\|_1^{2\gamma-2} |A_j \bigcap B_1| = O(1).$$

Therefore the proof of Theorem 2 (i) still works with the reduced $\mathcal{X}_1$ and reduced number $d_1 = |B_1|$ of coefficients $\beta_k : k \in B_1$. Thus,

$$\|\widehat{\beta}_{nB_1} - \beta_{0B_1}\|_2^2 = O_p(\frac{1}{n}), \;\; \|\widehat{\beta} - \beta_0\|_2^2 = O_p(\frac{1}{n}).$$

Let $h_n = n^{-1/2}$, and take $V_{1n}(a) = [G_n(\beta_0 + h_n(a', 0')') - G_n(\beta_0)]$, where 0 is a zero vector of dimension $|B_2|$ and $a = (a_1, \ldots, a_{d_1})'$ is a d_1-dimensional constant vector. By part (i) of Theorem 2, with large probability, $\widehat{\beta} - \beta_0 = h_n(\widehat{a}_n', 0')'$, $\widehat{a}_n = \text{argmin}\{V_{1n}(a) : a \in R^{d_1}\}$.

On the other hand, V_{1n} can be rewritten as :

$$\begin{aligned} V_{1n} &= h_n(a', 0')\nabla l_n(\beta_0) + \frac{1}{2} a' I_{11}(\beta_0) a + a' o_p(1) a \\ &\quad + \lambda_n \sum_{j=1}^{J_1} c_j \Big\{ \Big(\sum_{k \in A_j \bigcap B_1} |\beta_{0k} + h_n a_k| \Big)^{\gamma} - \|\beta_{0A_j}\|_1^{\gamma} \Big\} \\ &= T_{1n}(a) + T_{2n}(a). \end{aligned}$$

By Cai et al. (2005), we have $T_{1n} \to_d a'W + \frac{1}{2} a' I_{11}(\beta_0) a$, where $\to_d$ is convergence in distribution. According to Huang et al. (2009), we have

$$T_{2n}(a) \to \gamma\lambda_0 \sum_{j=1}^{J_1} c_j \|\beta_{0A_j}\|_1^{\gamma-1} \sum_{k \in A_j \bigcap B_1} \{a_k \text{sgn}(\beta_{0k}) I(\beta_{0k} \neq 0) + |a_k| I(\beta_{0k} = 0)\}.$$

Therefore, $V_{1n}(a) \to_d V_1(a)$. Since $\widehat{a}_n = O_p(1)$, by the argmin continuous mapping theorem of Kim and Pollard (1990), $\sqrt{n}(\widehat{\beta}_{nB_1} - \beta_{0B_1}) = \widehat{a}_n \to \text{argmin}(V_1(a))$, which completes the proof.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, 267-281.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations (with discussion), *J. Amer. Statist. Assoc.* **96**, 939-967.

Cai, J., Fan, J., Li, R. and Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika* **92**, 303-316.

Dave, S. S., Wright, G. and Tan, B. et al. (2004). Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *The New England J. Medicine* **351**, 2159-2169.

Dicker, L., Huang, B. and Lin, X. (2011). Variable selection and estimation with the seamless-L_0 penalty. *Statist. Sinica*, **32**, 929-962.

Du, P., Ma, S. and Liang, H. (2010). Penalized variable selection procedure for Cox models with semiparametric relative risk. *Ann. Statist.* **38**, 2092-2117.

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

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30**, 74-99.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* Wiley, New York.

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.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models* Chapman and Hall.

Huang, J. (1999). Efficient estimation of the partly linear additive Cox model. *Ann. Statist.* **27**, 1536-1563.

Huang, J., Ma, S., Xie, H. and Zhang, T. (2009). A group bridge approach for variable selection. *Biometrika* **96**, 339-355.

Johnson, B. A. (2009). On lasso for censored data. *Electron. J. Stat.* **3**, 485-506.

Kim, Y., Kim, J. and Kim, Y. (2006). The blockwise sparse regression. *Statist. Sinica* **16**, 375-390.

Kim, J. and D. Pollard (1990). Cube root asymptotics. *Ann. Statist.* **18**, 191-219.

Kim, J., Sohn, I., Jung, S. H, Kim, S. and Park, C. (2012). Analysis of survival data with group Lasso, *Comm. Statist. Simulation Comput.* **41**, 1593-1605.

Ma, S. and Huang, J. (2007). Combining clinical and genomic covariates via Cov-TGDR. *Cancer Informatics* **3**, 371-378.

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

Meier, L., van de Geer, S. and Bühlmann, P. (2008). The group Lasso for logisitc regression. *J. Roy. Statist. Soc. Ser. B* **70**, 53-71.

Schwarz, G. (1978). Estimating the dimension of a model, *Ann. Statist.* **6**, 461-464.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

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

van de Vijver, M. J., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* **347**, 1999-2009.

Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

Wang, S., Nan, B., Zhou, N. and Zhu, J. (2009). Hierarchically penalized Cox regression with grouped variables. *Biometrika* **96**, 307-322.

Yuan, M. and Lin, Y. (2007). On the non-negative garrote estimator. *J. Roy. Statist. Soc. Ser. B* **69**, 143-161.

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

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model, *Biometrika* **94**, 691-703.

Zhao, P., Rocha, G. and Yu, B. (2009). Grouped and hierarchical model selection through composite absolute penalties. *Ann. Statist.* **37**, 3468-3497.

Zou, H. (2006). The adaptive Lasso and its oracle properties, *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H. (2008). A note on path-based variable selection in the penalized proportional hazards model. *Biometrika* **95**, 241-247.

Department of Statistics and Actuarial Science, 241 Schaeffer Hall, University of Iowa, Iowa City, IA 52242-1419, USA.

E-mail: jian-huang@uiowa.edu

School of Mathematics and Statistics, Wuhan University, Wuhan, 430072, Hubei, People's Republic of China.

E-mail: lliu.math@whu.edu.cn

School of Mathematics and Statistics, Wuhan University, Wuhan, 430072, Hubei, People's Republic of China.

E-mail: liuyy@whu.edu.cn

Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.

E-mail: xingqiu.zhao@.polyu.edu.hk

(Received March 2013; accepted January 2014)