Journal of Multivariate Analysis 125 (2014) 50–64

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

SCAD-penalized regression in additive partially linear proportional hazards models with an ultra-high-dimensional linear part

Heng Lian [a,*], Jianbo Li [b], Xingyu Tang [a]

[a] Division of Mathematical Sciences, SPMS, Nanyang Technological University, Singapore, 637371, Singapore

[b] School of Mathematical Sciences, Xuzhou Normal University, Xuzhou, 221116, China

ARTICLE INFO

Article history:
Received 6 June 2013
Available online 19 December 2013

AMS subject classification:
62G20

Keywords:
Akaike information criterion (AIC)
Bayesian information criterion (BIC)
Extended Bayesian information criterion (EBIC)
Cross-validation
Ultra-high dimensional regression
SCAD

ABSTRACT

We consider the problem of simultaneous variable selection and estimation in additive partially linear Cox's proportional hazards models with high-dimensional or ultra-high-dimensional covariates in the linear part. Under the sparse model assumption, we apply the smoothly clipped absolute deviation (SCAD) penalty to select the significant covariates in the linear part and use polynomial splines to estimate the nonparametric additive component functions. The oracle property of the estimator is demonstrated, in the sense that consistency in terms of variable selection can be achieved and that the nonzero coefficients are asymptotically normal with the same asymptotic variance as they would have if the zero coefficients were known a priori. Monte Carlo studies are presented to illustrate the behavior of the estimator using various tuning parameter selectors.

1. Introduction

Parametric proportional hazards models, or Cox models, are probably the oldest and the most popular regression tools for studying the relationships between multiple covariates and censored event times in survival analysis. This class of models assume that the hazard function is related to the covariates by

$$\lambda(t|X) = \lambda_0(t)\exp\{X^T\beta\}, \tag{1}$$

where $\lambda_0(t)$ is the baseline hazard function and X is the covariate vector. The popularity of the proportional hazards models can be at least partially attributed to that the unknown baseline function elegantly disappears from the estimating function when partially likelihood is used, making estimation and inferences much easier.

In this paper, we consider proportional hazards models with the semiparametric risk that has a partially linear structure. More specifically, we assume that the hazard function is given by

$$\lambda(t|W, X) = \lambda_0(t)\exp\{\phi(W) + X^T\beta\}, \tag{2}$$

where now the model contains both the nonparametric component $\phi(W) = \sum_{j=1}^{q}\phi_j(W_j)$ and the parametric component $X^T\beta$, W is q-dimensional and X is p-dimensional. This model combines the flexibility of nonparametric modeling and

* Corresponding author.
E-mail address: henglian@ntu.edu.sg (H. Lian).

0047-259X/$ – see front matter
http://dx.doi.org/10.1016/j.jmva.2013.12.002

parsimony and easy interpretability of parametric modeling. In particular, it avoids the curse of dimensionality of a purely nonparametric model [19,5].

We note that to specify a partially linear model, at least two possible strategies are applied in the literature. One is simply to put discrete covariates in the linear part and continuous ones in the nonparametric part. Another more reasonable approach is to first perform a preliminary analysis with no linear part and then separate the covariates based on the shape of the estimated nonparametric functions. We assume that a partially linear specification is already available one way or another and in this paper do not further consider how the partially linear structure comes by in the first place, although this is an interesting problem in itself.

Variable selection is an important research topic in modern statistics. With many predictors available to include into the model, many of them may not be relevant for prediction and inclusion of these only hurts estimation performance. Recently, there has been considerable interest in investigating the variable selection problem for parametric and nonparametric models. Traditional variable selection methods such as stepwise regression and best subset selection suffer from instability as argued in [2], which is part of the reason why penalization-based method [22,6,38,11,37,30] has gained popularity in recent years. This motivates us to develop a penalization-based approach for simultaneous variable selection and estimation in partially linear Cox models.

Many studies on penalization-based variable selection for semiparametric models, including partially linear models, varying-coefficient models and additive models can be found in the recent literature. For example, Xie and Huang [32] studied variable selection in partially linear models, Wang et al. [25], Wang and Xia [29], Wei et al. [31] and Lian [15] investigated varying-coefficient models, Xue [33], Meier et al. [18], Huang et al. [12] investigated additive models, Li and Liang [14], Liu et al. [16], Wang et al. [28] considered partially linear varying-coefficient models and partially linear additive models for variable selection on the linear part.

For proportional hazards models, penalized variable selection has been considered in [23,7,36,9]. These works do not consider the $p > n$ case and do not consider additive structure. More recently, Bradic et al. [1] has extended it to the case where the number of predictors can be much larger than the sample size, a "large p small n" situation which has attracted much attention in recent years. On the other hand, Du et al. [4] has considered the additive partially linear models in Cox regression with penalized variable selection to identify significant covariates in the linear part. This work however only considered the case where the dimension of the covariates is fixed. This result thus does not apply when the number of covariates diverges. Ma and Du [17] considered weighted least squares estimation procedure for a class of transformation models which includes the accelerated failure time model as a special case. They also considered a high dimensional setting using iterated lasso and Kullback–Leibler geometry for variable selection in the parametric and the nonparametric part respectively, although they did not show asymptotic normality of the parametric part.

In this paper, we consider penalized variable selection for semiparametric Cox models with a diverging number of parameters in the linear part (the number of covariates in the nonparametric part is fixed for technical reasons, also due to that in practice this number is typically small), or even with ultra-high dimensional parameters. We will assume a sparse model. That is, although the number of covariates collected for statistical analysis is very large, only a small subset of covariates is important in predicting the event times. Such an assumption is often reasonable with high-dimensional data. In the ultra-high-dimensional setting we consider here, the full dimensionality might grow exponentially fast with the sample size, with $\log p = O(n^{\delta})$ for some $0 < \delta < 1$, and the true dimensionality involving only important variables diverges as $s = O(n^{\delta})$ for some $\delta < 1/2$. More specific assumptions on the speed of divergence are detailed later in the text. Due to that this small subset of important covariates is unknown to the statistician, its identification is an important issue and the concern is how the large dimension of covariates will affect the estimation of the coefficients. We use the SCAD method to achieve both goals of variable selection and estimation simultaneously. For the nonparametric components, these component functions are approximated and estimated using polynomial splines with the computationally favorable B-spline basis, which allows reasonable approximation of smooth functions with just a small number of basis functions.

The rest of the article is organized as follows. In the next section, we define the SCAD-penalized estimator in the additive partially linear proportional hazards model. We investigate its asymptotic theoretical properties in Section 3, including consistency in both estimation and variable selection, as well as the asymptotic normality of the linear coefficients. Computational aspects are discussed in Section 4 and several criteria for tuning parameter selection are proposed. The finite sample behavior of the SCAD-penalized estimator is illustrated in Section 5 for both $p < n$ and $p > n$. Finally, Appendix A contains all the technical proofs.

2. Penalized estimation with SCAD penalty

Let T^e and T^c be the event time and the censoring time respectively, where the hazard function of T^e is given by (2). Assume that T^e and T^c are independent given the covariates. The true nonparametric functions and parameters will be denoted using a subscript 0. The observable random variables are (T, Δ, W, X) where $T = \min\{T^e, T^c\}$ and $\Delta = I\{T^e \leq T^c\}$ ($I\{\cdot\}$ is the indicator function), $W = (W_1, \ldots, W_q)^T \in R^q$ and $X = (X_1, \ldots, X_p)^T \in R^p$ are the covariate vectors in the nonparametric part and the parametric part respectively. Note that ϕ_{0j} is identifiable only up to a constant and thus we assume $E\Delta\phi_{0j}(W_j) = 0$. We make n i.i.d. observations $(T_i, \Delta_i, W_i, X_i)$.

We use polynomial splines to approximate the nonparametric components. Let $\tau_0 = 0 < \tau_1 < \cdots < \tau_{K'} < 1 = \tau_{K'+1}$ be a partition of $[0, 1]$ into subintervals $[\tau_k, \tau_{k+1}), k = 0, \ldots, K'$ with K' internal knots. A polynomial spline of order r is a function whose restriction to each subinterval is a polynomial of degree $r-1$ and globally $r-2$ times continuously differentiable on $[0, 1]$. The collection of splines with a fixed sequence of knots has a normalized B-spline basis $\{B_1(x), \ldots, B_{\tilde{K}}(x)\}$ with $\tilde{K} = K' + r$. Because of the centering constraint $E\Delta\phi_{0j}(W_j) = 0$, we instead focus on the subspace of spline functions $S_j^0 := \{s : s = \sum_{k=1}^{\tilde{K}} a_{jk}B_k(x), \sum_{i=1}^n \Delta_i s(W_{ij}) = 0\}$ with basis $\{B_{jk}(x) = \sqrt{K}(B_k(x) - \sum_{i=1}^n \Delta_i B_k(W_{ij})/n), k = 1, \ldots, K = \tilde{K} - 1\}$ (the subspace is $K = \tilde{K} - 1$ dimensional due to the empirical version of the constraint). The multiplicative constant $\sqrt{K}$ is incorporated in the basis definition to simplify some expression later in the proofs, as done in [28]. Using spline expansions, we can approximate the nonparametric components by $\phi_{0j}(x) \approx \sum_k a_{jk}B_{jk}(x)$, $1 \le j \le p$.

Using spline expansion introduced above, the problem of estimating ϕ_{0j} is then transformed to the problem of estimating the coefficients $a_j = (a_{j1}, \ldots, a_{jK})^T$. Let $Y_i(t) = 1\{T_i \ge t\}$. Without considering penalized variable selection first, we can estimate (ϕ, β) as the maximizer of the (log-)partial likelihood

$$l(\phi, \beta) = \sum_{i=1}^n \Delta_i \left\{ \phi(W_i) + X_i^T\beta - \log \sum_{k=1}^n Y_k(T_i) \exp[\phi(W_k) + X_k^T\beta] \right\},$$

where $\phi(W_i) = \phi_1(W_{i1}) + \cdots + \phi_q(W_{iq})$, with the constraint $\phi_j = \sum_{k=1}^K a_{jk}B_{jk} \in S_j^0$. Using the notations

$$Z_i = (B_{11}(W_{i1}), \ldots, B_{1\,K}(W_{i1}), \ldots, B_{qK}(W_{iq}))^T$$

and

$$a = (a_1^T, \ldots, a_q^T)^T = (a_{11}, \ldots, a_{qK})^T,$$

the partial likelihood is written equivalently as

$$l(a, \beta) = \sum_{i=1}^n \Delta_i \left\{ Z_i^T a + X_i^T\beta - \log \sum_{k=1}^n Y_k(T_i) \exp[Z_k^T a + X_k^T\beta] \right\}. \tag{3}$$

Finally, we note that when p is large, the maximizer of the partially likelihood might not be well defined due to possible singularity of the information matrix. This problem will be avoided as a by-product of the penalized variable selection.

We use the SCAD penalty to achieve simultaneous variable selection and estimation of β, which is proposed in [6] and defined by its derivative

$$p_\lambda'(|x|) = \lambda \left\{ I(|x| \le \lambda) + \frac{(c\lambda - |x|)_+}{(c-1)\lambda} I(|x| > \lambda) \right\} \quad \text{for some } c > 2, \tag{4}$$

where the notation $(z)_+$ stands for the positive part of z. Fan and Li [6] suggested using $c = 3.7$ for the SCAD penalty function. The SCAD penalty possesses some desirable properties, such as that it results in a sparse model due to the singularity at zero, that it results in an estimator that is continuous in observations, and that it is almost unbiased for large parameter since the derivative of the penalty is zero when x is large. Other penalty functions such as the adaptive Lasso [38,12] or the minimax concave penalty [34] can also be used here and are expected to lead to similar consistency results.

The penalized partial likelihood objective function for estimating both a and β is

$$pl(a, \beta) = l(a, \beta) - n\sum_{j=1}^p p_\lambda(|\beta_j|). \tag{5}$$

Let $(\hat{a}, \hat{\beta})$ be the maximizer of the above penalized partial likelihood, then the SCAD-penalized estimators of β and ϕ_j, $j = 1, \ldots, q$ are $\hat{\beta}$ and $\sum_{k=1}^K \hat{a}_{jk}B_{jk}$, $j = 1, \ldots, q$, respectively.

We have chosen to use B-splines to estimate the nonparametric components. Although other estimation approaches can be applied such as with smoothing splines as in [4], using B-splines has the distinct advantage that it can be implemented with the *glmnet* package in R (see Section 4). This is especially important for our high-dimensional simulations later which require an efficient and robust algorithm. If inferences on β is desired, one can use the sandwich estimator involving the observed information matrix to estimate the variance of $\hat{\beta}$. Although it looks *glmnet* does not return an estimated covariance matrix in its current implementation, this can be implemented easily once a sparse estimate is available. On the other hand, as noted in [10], it is a challenge to demonstrate the consistency of this sandwich estimator.

3. Asymptotic properties of the SCAD-penalized estimator

For theoretical analysis, it is sometimes more convenient to consider the counting process representation of the partial likelihood. Let $N_i(t) = 1\{T_i \le t, \delta = 1\}$ be the right continuous counting process and $Y_i(t) = 1\{T_i \ge t\}$ be the

left-continuous at-risk process for the ith individual. Denote the true risk score by $m_0(W, X) = \phi_0(W) + X^T\beta_0$ where $\phi_0(W) = \phi_{01}(W_1) + \cdots + \phi_{0q}(W_q)$. Let $R = (W, X)$ be all the covariates. Denote by g, h any functions of R (h is possibly vector-valued). Define

$$S_n^{(0)}(g, t) = n^{-1}\sum_{i=1}^{n} Y_i(t)\exp[g(R_i)],$$

$$S_n^{(1)}(g, t)[h] = n^{-1}\sum_{i=1}^{n} Y_i(t)h(R_i)\exp[g(R_i)],$$

$$S_n^{(2)}(g, t)[h] = n^{-1}\sum_{i=1}^{n} Y_i(t)h(R_i)^{\otimes 2}\exp[g(R_i)],$$

$$G_n(g, t)[h] = S_n^{(1)}(g, t)[h]/S_n^{(0)}(g, t),$$

$$V_n(g, t)[h] = S_n^{(2)}(g, t)[h]/S_n^{(0)}(g, t) - G_n(g, t)[h]G_n^T(g, t)[h],$$

where for any vector ξ, $\xi^{\otimes 2}$ simply means $\xi\xi^T$.

Let $S^{(j)}(g, t) = E(S_n^{(j)}(g, t))$, $j = 0, 1, 2$, $G(g, t)[h] = S^{(1)}(g, t)[h]/S^{(0)}(g, t)$, $V(g, t)[h] = S^{(2)}(g, t)[h]/S^{(0)}(g, t) - G(g, t)[h]G^T(g, t)[h]$. We also let $P_{\Delta n}$ be the empirical measure of $(T_i, \Delta_i = 1, R_i)$, that is for any function f of (T, Δ, R), $P_{\Delta n}f = n^{-1}\sum_i \Delta_i f(T_i, \Delta_i, R_i)$, and let P_Δ be the measure of $(T, \Delta = 1, R)$.

The partial likelihood can be rewritten as

$$l(a, \beta) = \sum_{i=1}^{n}\int_0^\tau \{Q_i^T b - \log(S_n^{(0)}(g, t))\}dN_i(t),$$

where $b = (a^T, \beta^T)^T$, $Q_i = (Z_i^T, X_i^T)^T$, and g is a function of R defined by $(a, \beta) : g(R) = \sum_{j=1}^q a_j^T B_j(W_j) + X^T\beta$, $B_j(W_j) = (B_{j1}(W_j), \ldots, B_{jK}(W_j))^T$. Note as usual we only consider events over a finite interval $[0, \tau]$. The score function and the observed information are given by

$$U(a, \beta) = \sum_{i=1}^{n}\int_0^\tau \{Q_i - G_n(g, t)[Q]\}dN_i(t),$$

and

$$I(a, \beta) = \sum_{i=1}^{n}\int_0^\tau V_n(g, t)[Q]dN_i(t),$$

respectively. Define $\mathcal{H}_d$ as the collection of all functions on support $[0, 1]$ whose mth order derivative satisfies the Hölder condition of order r with $d \equiv m+r$. That is, for each $h \in \mathcal{H}_d$, there exists a constant $M_0 \in (0, \infty)$ such that $|h^{(m)}(s) - h^{(m)}(t)| \le M_0|s - t|^r$, for any $s, t \in [0, 1]$.

The following technical conditions are used in the study of asymptotics.

(C1) The covariate vector R has a bounded support (without loss of generality the support is assumed to be $[0, 1]^{(p+q)}$), with the marginal density of each covariate being continuous and bounded away from zero and infinity.

(C2) Only observations with censored event times in a finite interval $[0, \tau]$ are used in the partial likelihood. $P(\Delta = 1|R)$ and $P(T^c > \tau|R)$ are both bounded away from zero with probability one.

(C3) $\phi_{0j}(x) \in \mathcal{H}_d$, $j = 1, 2, \ldots, q$, for some $d > 1/2$. The order of the spline satisfies $r > d + 1/2$.

(C4) Let $\Sigma = \int_0^\tau V(m_0, t)[Q]S^{(0)}(m_0, t)\lambda_0(t)dt$. The eigenvalues of Σ are bounded away from zero and infinity.

Assumptions similar to (C1)–(C4) are contained in [10] or [1]. Boundedness of covariates is assumed for simplicity and might be relaxed to some moment conditions. The term QQ^T appears in the definition of Σ. Under mild assumptions, Huang et al. [12] showed that eigenvalues of EZZ^T are bounded and bounded away from zero. Thus it is expected eigenvalues of EQQ^T are bounded and bounded away from zero if eigenvalues of EXX^T are so and Z and X are not linear dependent. This can in turn be guaranteed with assumptions on the density of (T, W, X) as in assumptions (B5) and (B6) of [10]. Note that Σ is the population information matrix of the model and thus its positive-definiteness is a reasonable assumption.

We first present the results for the semiparametric Cox model with a diverging number of parameters (with $p \ll n$) without using penalty.

Theorem 1 (*Convergence Rates*). *Under conditions* (C1)–(C4), *assume that q is fixed,* $K \to \infty$, $(K+p)^2/n \to 0$, *and let* $(\hat{a}, \hat{\beta})$ *be the maximizer of* $l(a, \beta)$ *in* (3) *and* $\hat{\phi}_j = \sum_k \hat{a}_{jk} B_{jk}$, $\hat{\phi}(w) = \sum_j \hat{\phi}_j(w_j)$. *We have*

$$\|\hat{\phi} - \phi_0\| + \|\hat{\beta} - \beta_0\| = O_p\left(\sqrt{\frac{K+p}{n}} + \frac{1}{K^d}\right).$$

From the rates, it is seen that the optimal choice of K is of order $O(n^{1/(2d+1)})$ as usual. The following theorem shows that in fact the linear coefficients are asymptotically normal and thus converge at the $\sqrt{n}$ rate. For this further notations and assumptions are necessary.

Let $(a^*, h_1^*, \ldots, h_q^*)$ be R^p-valued L_2 functions that minimize $E\Delta\|X - a(T) - h_1(W_1) - \cdots - h_q(W_q)\|^2$. Let $h^*(w) = h_1^*(w_1) + \cdots + h_q^*(w_q)$. Denote $\tilde{\Sigma} = E\Delta(X - a^*(T) - h^*(W))^{\otimes 2}$. By direct calculations, it can be easily verified that $\tilde{\Sigma}$ can also be written as $\int_0^\tau V(m_0, t)[X - h^*(W)]S^{(0)}(m_0, t)\lambda_0(t)dt$ and is thus the information matrix for the linear part after taking into account the effect of the nonparametric part. The p components of h_j^* are denoted by h_{jl}^*, $1 \le l \le p$.

(C5) All h_{jl}^*, $1 \le j \le q$, $1 \le l \le p$, are in $\mathcal{H}_d$. The eigenvalues of $\tilde{\Sigma}$ are bounded and bounded away from zero.

Theorem 2 (*Asymptotic Normality*). *Under the same conditions as assumed in Theorem 1 and in addition* (C5) *holds and* $p = o(n^{1/3})$, *then for any unit p-vector* ν_n, *we have*

$$\sqrt{n}\nu_n^T \tilde{\Sigma}^{1/2}(\hat{\beta} - \beta_0) \to N(0, 1).$$

Thus, informally, we can say $\hat{\beta}$ is asymptotically normal with asymptotic variance $\tilde{\Sigma}^{-1}/n$.

Next we consider penalized variable selection for the semiparametric model. For this we assume a sparse true model in which we can write $\beta_0 = (\beta_0^{(1)}, \beta_0^{(2)} = 0)$, where the dimension of $\beta_0^{(1)}$ is s ($s \le p$). Let $(\hat{a}, \hat{\beta})$ be the maximizer of (5) (note that with abuse of notation, the maximizer for the unpenalized likelihood is denoted with the same notation). Correspondingly, we write $\hat{\beta} = (\hat{\beta}^{(1)}, \hat{\beta}^{(2)})$ where $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ are vectors of lengths s and $p - s$ respectively. The next theorem shows that all the covariates with zero coefficients can be identified with probability approaching 1, provided λ does not converge to zero too fast.

(C6) $\sqrt{\frac{K+p}{n}} + \frac{1}{K^d} \ll \lambda \ll \inf_{1 \le j \le s} |\beta_{0j}|$.

Theorem 3. *Under the same conditions as in Theorem 1, and that* λ *satisfies* (C6), *there exists a local maximizer of the penalized partial likelihood that satisfies*

$$\lim_{n \to \infty} P(\hat{\beta}^{(2)} = 0) = 1.$$

Furthermore, the following theorem demonstrates the oracle property of the SCAD-penalized estimator.

Theorem 4. *Under the same conditions as in Theorem 3, then*

$$\|\hat{\phi} - \phi_0\| + \|\hat{\beta} - \beta_0\| = O_p\left(\sqrt{\frac{K+p}{n}} + \frac{1}{K^d}\right). \tag{6}$$

If furthermore (C5) *holds and* $p^3/n \to 0$ *then*

$$\sqrt{n}\nu_n^T \tilde{\Sigma}_{11}^{1/2}(\hat{\beta}^{(1)} - \beta_0^{(1)}) \to N(0, 1),$$

where $\tilde{\Sigma}_{11}$ *is the* $s \times s$ *principal submatrix of* $\tilde{\Sigma}$ *and* ν_n *is a unit s-vector.*

Now we consider the ultra-high dimensional situation where $p \gg n$ while the number of nonzero coefficients is much smaller and still satisfies the constraint $s = o(n^{1/2})$ (or $s = o(n^{1/3})$ for asymptotic normality to hold). To prove the oracle properties in this ultra-high dimensional case, we first consider an "oracle estimator" where we assume the zero coefficients are known and thus the corresponding covariates are removed from the likelihood. More specifically, the oracle estimator $(\hat{a}^o, \hat{\beta}^o)$ with $\hat{\beta}^o \in R^s$ is defined as the solution to the following maximization problem

$$\arg\max_{a, \beta} \sum_{i=1}^n \int_0^\tau Z_i^T a + X_i^{(1)T}\beta - \log(S_n^{(0)}(g, t))dN_i(t),$$

where $X_i^{(1)} = (X_{i1}, \ldots, X_{is})^T$ and g is a function of R depending on (a, β) defined by $g(R) = Z^T a + X^{(1)T}\beta$.

Based on the results on the unpenalized estimator (Theorems 1 and 2), we know that if $K \to \infty$, $(K+s)^2/n \to 0$, the convergence rate of the oracle estimator is $\sqrt{(K+s)/n} + K^{-d}$, while if in addition $s = o(n^{1/3})$, $\hat{\beta}^o$ is asymptotically normal, under reasonable assumptions. Again with abuse of notation, let $(\hat{a}, \hat{\beta})$ be the SCAD-penalized estimator in the ultra-high-dimensional situation. The following shows that with probability approaching one, the SCAD-penalized estimator is exactly equal to the oracle estimator (by filling in zeros for $\beta^{(2)}$) under some assumptions.

(C7) $\sqrt{(K+s)/n} + K^{-d} + \sqrt{n^{-1}\max\{\log(p), \log(n)\}} \ll \lambda \ll \inf_{1\le j\le s} |\beta_{0j}|$.

(C8) $E[|X_j - S_j^{(1)}(m_0, T)[X]/S^{(0)}(m_0, T)|^m] \le \frac{m!}{2} J^{m-2} D^2$, $m = 2, 3, \ldots$, for some constants J, $D > 0$, for all $1 \le j \le p$, where $S_j^{(1)}(m_0, t)[X]$ is the jth component of $S^{(1)}(m_0, t)[X]$.

Assumptions (C7) and (C8) specify appropriate tuning parameters and implicitly impose a bound for the size of p. In particular, if $\inf_{1\le j\le s} |\beta_{0j}|$ is bounded away from zero so that λ can be chosen to converge to zero arbitrarily slowly, the assumption that $\sqrt{n\log(p)} \ll n\lambda$ actually implies $p \ll \exp\{n\}$. (C8) is similar to condition 3 in [1] which controls the tail probably of appropriate quantities so that Bernstein's inequality can be applied. Also note that the quantity $X_j - S_j^{(1)}(m_0, T)[X]/S^{(0)}(m_0, T)$ indeed has mean zero [20, Lemma 2].

Theorem 5. *Under conditions* (C1)–(C5), (C7) *and* (C8), *and that* $K \to \infty$, $(K+s)^2/n \to 0$, *there exists a local maximizer of the penalized partial likelihood such that*

$$\lim_{n\to\infty} P(\hat{\beta} = (\hat{\beta}^o, 0)) = 1.$$

As discussed above, we immediately have the following corollary.

Corollary 1. *Under conditions assumed in Theorem 5, we have*

(i) $\hat{\beta}^{(2)} = 0$ *with probability approaching* 1.

(ii) $\|\hat{\phi} - \phi_0\| + \|\hat{\beta}^{(1)} - \beta_0^{(1)}\| = O_p(\sqrt{\frac{K+s}{n}} + \frac{1}{K^d})$.

(iii) *If* $s = o(n^{1/3})$, $\sqrt{n} v_n^T \tilde{\Sigma}_{11}^{1/2} (\hat{\beta}^{(1)} - \beta_0^{(1)}) \to N(0, 1)$.

Note that due to assumptions (C7) and (C8) which are needed when $p \gg n$, Theorems 3 and 4 cannot be regarded as a special case of Corollary 1. Thus we will present proofs for both $p \ll n$ and $p \gg n$ based on different techniques.

4. Implementation

As mentioned in Section 2, our approach using B-splines for the nonparametric components allows easy implementation using the existing R package *glmnet* which can deal with the more general elastic-net penalty and includes the lasso penalty as a special case [21]. To utilize this existing implementation, we use the local linear approximation (LLA, [39]) for the penalty function

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(0)}|)(|\beta_j| - |\beta_j^{(0)}|), \quad \text{for } \beta_j \approx \beta_j^{(0)},$$

where $\beta_j^{(0)}$ is an initial estimate of β_j. We can simply set $\beta_j^{(0)} = 0$ for example. For $k = 1, 2, \ldots$, we then repeatedly solve

$$(a^{(k+1)}, \beta^{(k+1)}) = \arg\max \left\{ l(a^{(k)}, \beta^{(k)}) - n \sum_{j=1}^{p} p'_\lambda(|\beta_j^{(k)}|)|\beta_j| \right\} \tag{7}$$

until convergence. It was shown that the LLA algorithm is an instance of the MM (minorize–maximize) algorithm and enjoys the ascent property. For details see [39]. We now only need to note that (7) can be directly solved by the current implementation of *glmnet* which can take into account the factor $p'_\lambda(|\beta_j^{(k)}|)$ in the lasso penalty.

To implement the procedures, we also need to find a data-driven method to choose the regularization parameter λ and number of spline basis K. To ease the computational burden, we use cubic splines with $K = 6$ following [12]. This choice of K is small enough to avoid overfitting in typical problems with sample size not too small, and big enough to flexibly approximate many smooth functions. We also find the results are very similar for K ranging from 5 to 8 in our simulations, and thus we only report the results obtained with $K = 6$ later. To find an appropriate tuning parameter λ, we consider AIC, BIC, EBIC, and 5-fold cross-validation (5CV). The first three criteria are defined by

$$\text{AIC}(\lambda) = -2l(\hat{a}, \hat{\beta}) + 2df,$$
$$\text{BIC}(\lambda) = -2l(\hat{a}, \hat{\beta}) + \log n \cdot df,$$
$$\text{EBIC}(\lambda) = -2l(\hat{a}, \hat{\beta}) + (\log n + \log p) \cdot df,$$

Table 1
Model selection results for different tuning parameter selectors for $n = 100$. # nonzero: number of nonzero coefficients identified. # correct nonzero: number of nonzero coefficients in the true model that are correctly identified as nonzero. The numbers in the subscript are the standard errors based on simulations.

(n, p)	Method	# nonzero	# correct nonzero
(100, 50)	LAS-BIC	$14.64_{2.03}$	$9.84_{0.36}$
	LAS-EBIC	$12.77_{2.19}$	$9.75_{0.45}$
	LAS-AIC	$16.46_{1.36}$	$9.91_{0.28}$
	LAS-5CV	$14.16_{1.68}$	$9.86_{0.34}$
	SCAD-BIC	$10.21_{1.13}$	$9.41_{0.51}$
	SCAD-EBIC	$9.69_{0.91}$	$9.27_{0.50}$
	SCAD-AIC	$12.11_{1.85}$	$9.68_{0.46}$
	SCAD-5CV	$10.32_{1.57}$	$9.38_{0.52}$
(100, 100)	LAS-BIC	$16.82_{4.82}$	$9.59_{0.49}$
	LAS-EBIC	$12.79_{2.64}$	$9.31_{0.70}$
	LAS-AIC	$36.78_{7.01}$	$9.89_{0.31}$
	LAS-5CV	$21.25_{5.36}$	$9.74_{0.44}$
	SCAD-BIC	$11.73_{2.14}$	$9.31_{0.64}$
	SCAD-EBIC	$9.74_{1.43}$	$8.92_{0.73}$
	SCAD-AIC	$22.32_{6.25}$	$9.71_{0.47}$
	SCAD-5CV	$13.58_{3.95}$	$9.44_{0.59}$
(100, 200)	LAS-BIC	$15.01_{4.93}$	$9.21_{0.60}$
	LAS-EBIC	$12.20_{1.93}$	$9.07_{0.63}$
	LAS-AIC	$78.21_{18.43}$	$9.85_{0.38}$
	LAS-5CV	$26.25_{5.58}$	$9.59_{0.51}$
	SCAD-BIC	$12.26_{2.68}$	$9.02_{0.65}$
	SCAD-EBIC	$9.82_{1.47}$	$8.72_{0.69}$
	SCAD-AIC	$54.85_{16.23}$	$9.64_{0.50}$
	SCAD-5CV	$18.78_{7.66}$	$9.28_{0.58}$

respectively, where $\hat{a}$, $\hat{\beta}$ implicitly depends on λ and df is some value representing the "degrees of freedom" of the estimated model, which also depends on λ. Following the existing literature, the degree of freedom here is defined as the number of nonzero entries in $\hat{\beta}$.

Finally, 5CV selects the tuning parameter λ by splitting the data into 5 parts with one part reserved for validation in turn. The estimator obtained from the training part of the data is evaluated by the partial likelihood on the validation part of the data. 5CV is computationally more demanding than the others since the same model fitting procedure needs to be carried out multiple times. Hence in total we have **four** tuning parameter selectors which we will compare in our simulation studies.

5. Simulations

In this section, we conduct some Monte Carlo studies to assess the effectiveness of our proposed method. The performance of the estimators will be assessed by their model selection accuracy as well as estimation accuracy. We generate our data from the Cox model with hazard function given by $(q = 2, s = 10)$

$$\lambda(t|W, X) = \exp\left\{\phi_{01}(W_1) + \phi_{02}(W_2) + \sum_{j=1}^{10} X_j\beta_{0j} - 4\right\},$$

where the two nonparametric components are

$$\phi_{01}(t) = \sin(2\pi t), \qquad \phi_{02}(t) = 4[t(1-t) - 1/6]$$

and the coefficients in the linear part are $\beta = (5.5, 5, 4.5, \ldots, 1)$. The covariates $R = (W, X)$ were generated as follows. We first generate a multivariate $(p+q)$-dimensional Gaussian vector $(V_1, \ldots, V_{p+q})$ with covariance given by $\mathrm{Cov}(V_j, V_{j'}) = (0.2)^{|j-j'|}$. Then the cumulative distribution function of the standard normal distribution is applied to each component to map the components to the range (0, 1), to obtain $W_1, W_2, X_1, \ldots, X_p$. The censoring times are independently generated from an exponential distribution with mean 500. Under our simulation setup, about 25% observations are censored in the generated datasets.

We consider $n = 100, 200$ and $p = 50, 100, 200$. We use cubic splines with $K = 6$ and choose the tuning parameter λ based on various criteria. In each case we repeat the simulation 500 times. For comparison, we also computed the lasso estimator where the penalty function used is $p_\lambda(|\beta_j|) = \lambda|\beta_j|$. We also consider using BIC, EBIC, AIC and 5CV for its tuning parameter selection.

In Tables 1 and 2, we report the model selection results of the lasso estimator (LAS) and the SCAD-penalized estimator for $n = 100$ and $n = 200$ respectively, comparing the results for AIC, BIC, EBIC and 5CV. In the tables, we report the number of

Table 2
Model selection results for different tuning parameter selectors for $n = 200$. # nonzero: number of nonzero coefficients identified. # correct nonzero: number of nonzero coefficients in the true model that are correctly identified as nonzero. The numbers in the subscript are the standard errors based on simulations.

(n, p)	Method	# nonzero	# correct nonzero
(200, 50)	LAS-BIC	$14.75_{2.18}$	$9.92_{0.27}$
	LAS-EBIC	$13.25_{2.11}$	$9.89_{0.31}$
	LAS-AIC	$16.47_{1.49}$	$9.98_{0.14}$
	LAS-5CV	$14.86_{1.82}$	$9.95_{0.21}$
	SCAD-BIC	$10.21_{1.10}$	$9.60_{0.51}$
	SCAD-EBIC	$9.78_{0.81}$	$9.48_{0.54}$
	SCAD-AIC	$12.05_{1.81}$	$9.84_{0.36}$
	SCAD-5CV	$10.79_{1.72}$	$9.65_{0.50}$
(200, 100)	LAS-BIC	$16.27_{4.22}$	$9.70_{0.48}$
	LAS-EBIC	$13.03_{2.04}$	$9.64_{0.50}$
	LAS-AIC	$35.08_{6.74}$	$9.98_{0.14}$
	LAS-5CV	$22.95_{4.97}$	$9.88_{0.32}$
	SCAD-BIC	$10.69_{1.66}$	$9.41_{0.57}$
	SCAD-EBIC	$9.70_{1.05}$	$9.19_{0.59}$
	SCAD-AIC	$20.40_{6.44}$	$9.79_{0.40}$
	SCAD-5CV	$13.67_{3.85}$	$9.52_{0.55}$
(200, 200)	LAS-BIC	$17.41_{4.22}$	$9.47_{0.57}$
	LAS-EBIC	$10.92_{2.69}$	$9.23_{0.58}$
	LAS-AIC	$62.42_{16.47}$	$9.89_{0.31}$
	LAS-5CV	$30.22_{10.16}$	$9.71_{0.47}$
	SCAD-BIC	$11.21_{2.23}$	$9.19_{0.58}$
	SCAD-EBIC	$9.60_{1.16}$	$8.99_{0.61}$
	SCAD-AIC	$39.88_{12.38}$	$9.77_{0.42}$
	SCAD-5CV	$18.94_{8.12}$	$9.47_{0.57}$

nonzero coefficients in the estimated models, as well as the number of nonzero coefficients estimated that are also nonzero in the true model. The following observations can be made based on these tables.

- AIC and 5CV are typically more liberal than BIC and EBIC in the sense that the resulting estimators tend to include more zero components in the estimated model (false positives). This effect is more obvious for larger p.
- For all methods, the false negatives are low in our simulations. That is, most important variables are included in the model.
- Models selected using the SCAD penalty are smaller than the corresponding models selected using the lasso penalty.

Tables 3 and 4 show for each of the ten covariates, the percentage of times it appears in the estimated model. As expected, as the magnitude of the coefficient decreases, it is more likely to be missed by the variable selection procedure.

Fig. 1 shows the estimation errors. Here we define the estimation error to be $\sum_{i=1}^{n}[\exp\{-\sum_{j=1}^{2}\hat{\phi}_j(W_{ij}) - \sum_{j=1}^{p}X_{ij}\hat{\beta}_j\} - \exp\{-\sum_{j=1}^{2}\phi_{0j}(W_{ij}) - \sum_{j=1}^{p}X_{ij}\beta_{0j}\}]^2/n$. This criterion was also used in [1]. To see the meaning of this criterion, note that up to some multiplicative constant $\exp\{-\sum_{j=1}^{2}\phi_{0j}(W_{ij}) - \sum_{j=1}^{p}X_{ij}\beta_{0j}\}$ is the expected survival time given the covariates. In this figure, we also compare different methods with the oracle estimator where the zero coefficients are assumed known and the models are fitted after removing the corresponding covariates without using penalty. From Fig. 1, we see that when $p = 50$, different methods are quite similar to each other in performance. As the dimension increases, it is clear that SCAD-penalized estimators using BIC, EBIC and 5-fold CV emerge as the better performers. The worst performer is probably estimators with tuning parameters selected by AIC. Given that cross-validation requires fitting the model multiple times, BIC and EBIC are recommended.

6. Concluding remarks

In this article, we studied the SCAD-penalized estimator for variable selection and estimation in additive partially linear Cox models with the number of covariates in the linear part possibly larger than sample size. The proposed procedure automatically eliminates the irrelevant components by setting them as zero, and simultaneously estimates the nonzero ones.

Our simulations indicate that BIC/EBIC works better than the other tuning parameter selectors. Various works including Wang et al. [27], Chen and Chen [3], Wang and Xia [29], Wang et al. [26] have demonstrated that BIC/EBIC can consistently identify the nonzero coefficients in both linear models and semiparametric models. We conjecture that similar theoretical results could be demonstrated for the consistency of BIC/EBIC for proportional hazards models, although we are not able to provide a proof currently.

Table 3
Model selection results of different estimators based on 500 replications for $n = 100$. The numbers shown are the percentages among these 500 replications.

(n, p)	Method	Selection frequency									
(100, 50)	LAS-BIC	100	100	100	100	100	100	100	100	100	84
	LAS-EBIC	100	100	100	100	100	100	100	100	99	76
	LAS-AIC	100	100	100	100	100	100	100	100	100	91
	LAS-5CV	100	100	100	100	100	100	100	100	100	86
	SCAD-BIC	100	100	100	100	100	100	100	100	97	44
	SCAD-EBIC	100	100	100	100	100	100	100	100	92	35
	SCAD-AIC	100	100	100	100	100	100	100	100	99	69
	SCAD-5CV	100	100	100	100	100	100	100	100	96	42
(100, 100)	LAS-BIC	100	100	100	100	100	100	100	100	97	62
	LAS-EBIC	100	100	100	100	100	100	100	98	86	47
	LAS-AIC	100	100	100	100	100	100	100	100	99	90
	LAS-5CV	100	100	100	100	100	100	100	100	98	76
	SCAD-BIC	100	100	100	100	100	100	100	99	87	45
	SCAD-EBIC	100	100	100	100	100	100	100	94	72	26
	SCAD-AIC	100	100	100	100	100	100	100	100	96	75
	SCAD-5CV	100	100	100	100	100	100	100	99	92	53
(100, 200)	LAS-BIC	100	100	100	100	100	100	100	100	83	38
	LAS-EBIC	100	100	100	100	100	100	100	99	79	29
	LAS-AIC	100	100	100	100	100	100	100	100	97	88
	LAS-5CV	100	100	100	100	100	100	100	100	93	66
	SCAD-BIC	100	100	100	100	100	100	100	98	76	28
	SCAD-EBIC	100	100	100	100	100	100	100	95	58	19
	SCAD-AIC	100	100	100	100	100	100	100	100	90	74
	SCAD-5CV	100	100	100	100	100	100	100	98	85	45

Table 4
Model selection results of different estimators based on 500 replications for $n = 200$. The numbers shown are the percentages among these 500 replications.

(n, p)	Method	Selection frequency									
(200, 50)	LAS-BIC	100	100	100	100	100	100	100	100	100	92
	LAS-EBIC	100	100	100	100	100	100	100	100	100	89
	LAS-AIC	100	100	100	100	100	100	100	100	100	98
	LAS-5CV	100	100	100	100	100	100	100	100	100	95
	SCAD-BIC	100	100	100	100	100	100	100	100	97	63
	SCAD-EBIC	100	100	100	100	100	100	100	100	96	52
	SCAD-AIC	100	100	100	100	100	100	100	100	100	84
	SCAD-5CV	100	100	100	100	100	100	100	100	98	67
(200, 100)	LAS-BIC	100	100	100	100	100	100	100	100	98	72
	LAS-EBIC	100	100	100	100	100	100	100	100	97	67
	LAS-AIC	100	100	100	100	100	100	100	100	100	98
	LAS-5CV	100	100	100	100	100	100	100	100	99	89
	SCAD-BIC	100	100	100	100	100	100	100	100	93	48
	SCAD-EBIC	100	100	100	100	100	100	100	100	88	31
	SCAD-AIC	100	100	100	100	100	100	100	100	98	81
	SCAD-5CV	100	100	100	100	100	100	100	100	94	58
(200, 200)	LAS-BIC	100	100	100	100	100	100	100	100	95	52
	LAS-EBIC	100	100	100	100	100	100	100	100	91	32
	LAS-AIC	100	100	100	100	100	100	100	100	100	89
	LAS-5CV	100	100	100	100	100	100	100	100	98	73
	SCAD-BIC	100	100	100	100	100	100	100	99	88	32
	SCAD-EBIC	100	100	100	100	100	100	100	98	81	20
	SCAD-AIC	100	100	100	100	100	100	100	100	99	78
	SCAD-5CV	100	100	100	100	100	100	100	100	94	53

One important problem that we did not address in this paper is how to partition the covariates into linear and nonlinear ones. In practice, one might use the strategy as in [17], which used low dimensional clinical covariates in the nonparametric part and high dimensional gene expressions in the parametric part. One could also adopt a strategy similar to [35] that used

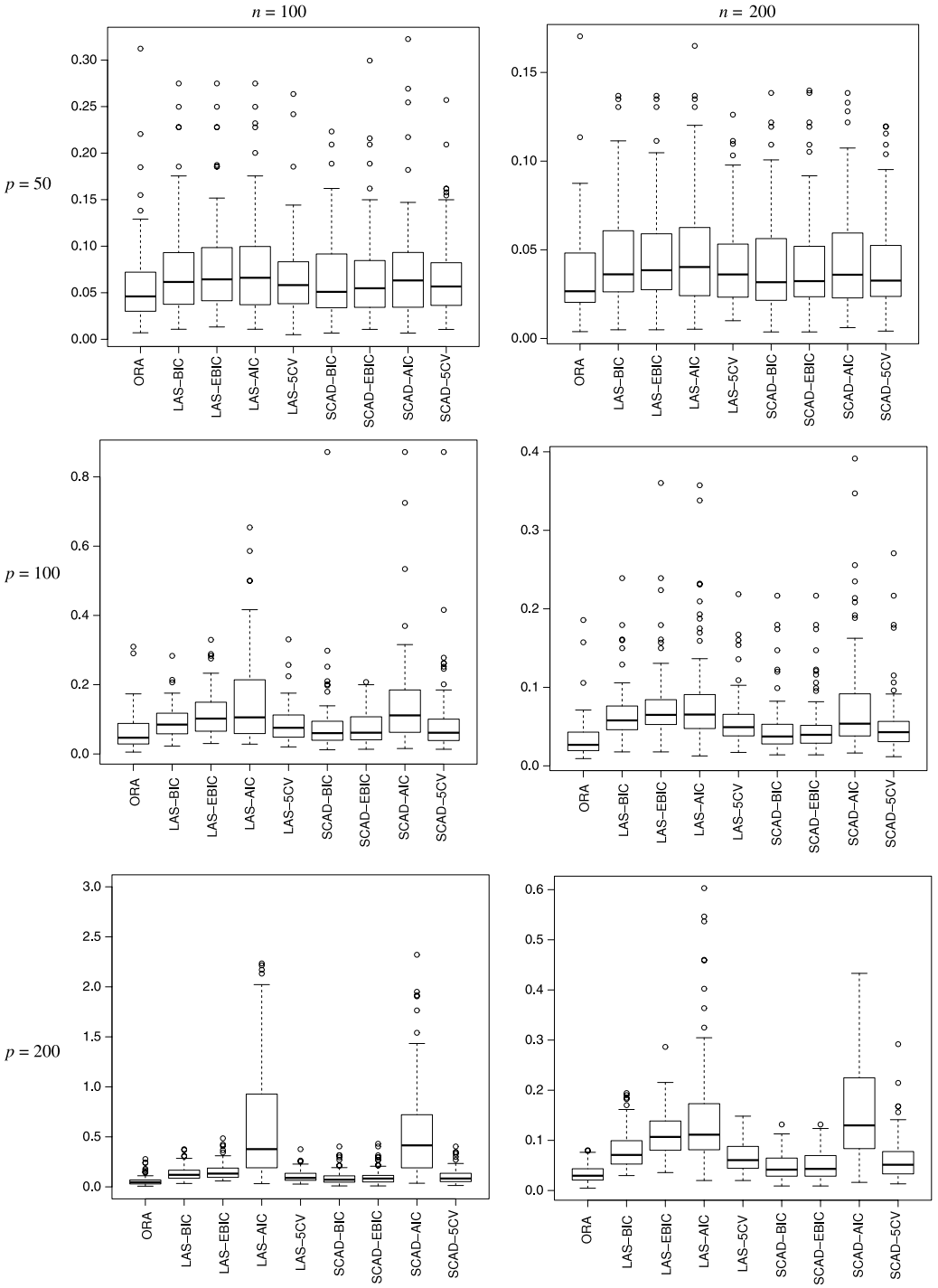

Fig. 1. Boxplots showing the estimation errors.

penalization to determine automatically the nonparametric and the parametric part, but it remains to be seen how well it works in the high dimensional setting.

Acknowledgments

The authors thank anonymous reviewers for their constructive comments that greatly improved the paper. The research of Heng Lian is supported by National Natural Science Foundation of China (Grant Nos. 11271241 and 11301279).

Appendix A

Proof of Theorem 1. The strategy of proof is similar to [1] with the main difference being that the nonparametric components need to be appropriately dealt with in spline approximation. Let $a_0 = (a_{01}^T, \ldots, a_{0p}^T)^T$ be a qK dimensional vector that satisfies $\|\phi_{0j} - a_{0j}^T B_j\|_\infty = O(K^{-d})$, $1 \le j \le q$ (such approximation rates are possible due to our smoothness assumption (C2) and well-known approximation properties of B-splines). Denote $b = (a, \beta)$ and $b_0 = (a_0, \beta_0)$. Let $\gamma_n = C(\sqrt{(K+p)/n} + K^{-d})$ and $u \in R^{qK+p}$ with $\|u\| = 1$.

It is sufficient to show that for any $\epsilon > 0$, there exists a large enough C (in the definition of γ_n) such that

$$P\left\{\sup_{\|u\|=1} l((a_0, \beta_0) + \gamma_n u) < l(a_0, \beta_0)\right\} \ge 1 - \epsilon, \tag{8}$$

when n is big enough.

We have

$$l(b_0 + \gamma_n u) - l(b_0) = \gamma_n U(b_0)^T u + \frac{1}{2}\gamma_n^2 u^T \partial U(b_0) u + r_n, \tag{9}$$

where r_n is equal to

$$\frac{1}{6}\sum_{j,k,l}(b_j - b_{0j})(b_k - b_{0k})(b_l - b_{0l})\frac{\partial^2 U_l(\tilde{b})}{\partial b_j \partial b_k},$$

U_l is the l-th component of U, and $\tilde{b}$ is a value between b_0 and $b = b_0 + \gamma_n u$.

We first consider

$$U(b_0) = \sum_i \int_0^\tau Q_i - \frac{S_n^{(1)}(m_{0n}, t)[Q]}{S_n^{(0)}(m_{0n}, t)} dN_i(t),$$

where $m_{0n}(R) = Za_0 + X^T\beta_0$.

Similar to Lemma 5.3 of [10], we have

$$P_{\Delta n}\frac{S_n^{(1)}(m_{0n}, t)[Q]}{S_n^{(0)}(m_{0n}, t)} - \frac{S_n^{(1)}(m_0, t)[Q]}{S_n^{(0)}(m_0, t)} = P_\Delta \frac{S^{(1)}(m_{0n}, t)[Q]}{S^{(0)}(m_{0n}, t)} - \frac{S^{(1)}(m_0, t)[Q]}{S^{(0)}(m_0, t)} + o_p(n^{-1/2}) \tag{10}$$

where $m_0(R) = \phi_0(W) + X^T\beta_0$. Using Lemma 7.4 of [10] and that $\|m_{0n} - m_0\| = O(K^{-d})$, we have

$$P_\Delta \frac{S^{(1)}(m_{0n}, t)[Q]}{S^{(0)}(m_{0n}, t)} - \frac{S^{(1)}(m_0, t)[Q]}{S^{(0)}(m_0, t)} = O(K^{-d}). \tag{11}$$

Thus

$$U(b_0) = \sum_i \int_0^\tau Q_i - \frac{S_n^{(1)}(m_0, t)[Q]}{S_n^{(0)}(m_0, t)} dN_i(t) + O_p(\sqrt{n} + nK^{-d}).$$

Let ξ_n denote the first term on the right hand side above, direct algebraic calculations show that

$$\begin{aligned} E(\xi_n^T \xi_n) &= \mathrm{tr}(E[\xi_n \xi_n^T]) \\ &= n\mathrm{tr}\left(E\int_0^\tau V_n(m_0, t)[Q] S_n^{(0)}(m_0, t)\lambda_0(t)dt\right). \end{aligned}$$

Since

$$\begin{aligned} \mathrm{tr}(E[V_n(m_0, t)[Q]S_n^{(0)}(m_0, t)]) &= \mathrm{tr}\left(E\left[\sum_i (Q_i - G_n(m_0, t)[Q])^{\otimes 2} Y_i(t)\exp\{m_0(R_i)\}\right]\right) \\ &\le E[\mathrm{tr}(Q_i^{\otimes 2} Y_i \exp\{m_0(R_i)\})] = O(K + p), \end{aligned}$$

we have $\|\xi_n\|_2 = O_p(\sqrt{n(K+p)})$ and thus

$$\|U(b_0)\| = O_p(\sqrt{n(K+p)} + nK^{-d}). \tag{12}$$

Next, we have

$$-\partial U(b_0) = \sum_i \int_0^\tau \frac{S_n^{(2)}(m_{0n},t)[Q]S_n^{(0)}(m_{0n},t) - (S_n^{(1)}(m_{0n},t)[Q])^{\otimes 2}}{(S_n^{(0)}(m_{0n},t))^2} dN_i(t)$$
$$= \sum_i \int_0^\tau \frac{S_n^{(2)}(m_0,t)[Q]S_n^{(0)}(m_0,t) - (S_n^{(1)}(m_0,t)[Q])^{\otimes 2}}{(S_n^{(0)}(m_0,t))^2} dN_i(t) + O(nK^{-d}) \quad (13)$$

where again we used Lemma 7.4 of [10]. Similar to Lemmas A.2 and A.4 of [1], we can show that

$$\sum_i \int_0^\tau \frac{S_n^{(2)}(m_0,t)[Q]S_n^{(0)}(m_0,t) - (S_n^{(1)}(m_0,t)[Q])^{\otimes 2}}{(S_n^{(0)}(m_0,t))^2} dN_i(t) = n\int_0^\tau V(m_0,t)[Q]S^{(0)}(m_0,t)\lambda_0(t)dt + o_p(n), \quad (14)$$

and thus the minimum eigenvalue of $-\partial U(b_0)/n$ is bounded away from zero.

Then, as in the proof of Theorem 4.2 in [1],

$$r_n = O_p(n\gamma_n^3). \quad (15)$$

Combining (9)–(15), we get $\|\hat{b} - b_0\| = O_p(\gamma_n)$, which implies the statement of the theorem. □

Proof of Theorem 2. Let $h_{njl}^* \in S_j^0$ be the spline functions that approximates h_{jl}^* with $\|h_{njl}^* - h_{jl}^*\|_\infty = O(K^{-d})$. Let $h_n^*(w) = h_{n1}^*(w_1) + \cdots + h_{nq}^*(w_q)$ where $h_{nj}^* = (h_{nj1}^*, \ldots, h_{njp}^*)^T$. Since $\hat{b} = (\hat{a}, \hat{\beta})$ maximizes the partial likelihood (3), it is easy to see that $v = 0$ maximizes

$$\sum_{i=1}^n \int_0^\tau \hat{m}(R_i) + (X_i - h_n^*(W_i))^T v - \log\left(\sum_k Y_k(t)\exp\{\hat{m}(R_i) + (X_i - h_n^*(W_i))^T v\}\right) dN_i(t),$$

where $\hat{m}(R) = Z^T\hat{a} + X^T\hat{\beta}$. The first order condition gives

$$\sum_i \int_0^\tau (X_i - h_n^*(W_i)) - \frac{S_n^{(1)}(\hat{m},t)[X - h_n^*(W)]}{S_n^{(0)}(\hat{m},t)} dN_i(t) = 0.$$

Consider the difference

$$n^{-1}\sum_i \int_0^\tau (X_i - h^*(W_i)) - \frac{S_n^{(1)}(\hat{m},t)[X - h^*(W)]}{S_n^{(0)}(\hat{m},t)} dN_i(t)$$
$$- n^{-1}\sum_i \int_0^\tau (X_i - h_n^*(W_i)) - \frac{S_n^{(1)}(\hat{m},t)[X - h_n^*(W)]}{S_n^{(0)}(\hat{m},t)} dN_i(t)$$
$$= n^{-1}\sum_i \int_0^\tau (h_n^*(W_i) - h^*(W_i)) - \frac{S_n^{(1)}(\hat{m},t)[h_n^*(W) - h^*(W)]}{S_n^{(0)}(\hat{m},t)} dN_i(t)$$
$$\equiv A_{1n} + A_{2n} + A_{3n} + A_{4n}$$

where

$$A_{1n} = (P_{\Delta n} - P_\Delta)\left\{h_n^*(W_i) - h^*(W_i) - \frac{S^{(1)}(\hat{m},t)[h_n^*(W) - h^*(W)]}{S^{(0)}(\hat{m},t)}\right\},$$
$$A_{2n} = P_{\Delta n}\left\{\frac{S^{(1)}(\hat{m},t)[h_n^*(W) - h^*(W)]}{S^{(0)}(\hat{m},t)} - \frac{S_n^{(1)}(\hat{m},t)[h_n^*(W) - h^*(W)]}{S_n^{(0)}(\hat{m},t)}\right\},$$
$$A_{3n} = P_\Delta\left\{\frac{S^{(1)}(m_0,t)[h_n^*(W) - h^*(W)]}{S^{(0)}(m_0,t)} - \frac{S^{(1)}(\hat{m},t)[h_n^*(W) - h^*(W)]}{S^{(0)}(\hat{m},t)}\right\},$$
$$A_{4n} = P_\Delta\left(h_n^*(W_i) - h^*(W_i) - \frac{S^{(1)}(m_0,t)[h_n^*(W) - h^*(W)]}{S^{(0)}(m_0,t)}\right).$$

By the maximal inequality and the entropy calculations in Lemma 7.1 and Corollary 7.1 of [10], we have $A_{1n} = o_p(n^{-1/2})$. Similar to Lemma 7.3 of [10], $A_{2n} = o_p(n^{-1/2})$. Similar to (11) and using that $\|h_{njl}^* - h_{jl}^*\| = O(K^{-d})$, $A_{3n} = o_p(n^{-1/2})$, finally, we note that since for any function h, $S^{(1)}(m_0,t)[h]/S^{(0)}(m_0,t) = E[h(R)|T = t, \Delta = 1]$ [20, Lemma 2], $A_{4n} = 0$.

Thus we have that

$$n^{-1}\sum_i \int_0^\tau (X_i - h^*(W_i)) - \frac{S_n^{(1)}(\hat{m}, t)[X - h^*(W)]}{S_n^{(0)}(\hat{m}, t)} dN_i(t) = o_p(n^{-1/2}). \tag{16}$$

Similar to (10),

$$\begin{aligned}
&n^{-1}\sum_i \int_0^\tau (X_i - h^*(W_i)) - \frac{S_n^{(1)}(\hat{m}, t)[X - h^*(W)]}{S_n^{(0)}(\hat{m}, t)} dN_i(t) \\
&= n^{-1}\sum_i \int_0^\tau (X_i - h^*(W_i)) - \frac{S_n^{(1)}(m_0, t)[X - h^*(W)]}{S_n^{(0)}(m_0, t)} dN_i(t) \\
&\quad - P_\Delta\left(\frac{S^{(1)}(\hat{m}, t)[X - h^*(W)]}{S^{(0)}(\hat{m}, t)} - \frac{S^{(1)}(m_0, t)[X - h^*(W)]}{S^{(0)}(m_0, t)}\right) + o_p(n^{-1/2}).
\end{aligned} \tag{17}$$

Direct Taylor expansion shows that

$$\begin{aligned}
&P_\Delta\left(\frac{S^{(1)}(\hat{m}, t)[X - h^*(W)]}{S^{(0)}(\hat{m}, t)} - \frac{S^{(1)}(m_0, t)[X - h^*(W)]}{S^{(0)}(m_0, t)}\right) \\
&= P_\Delta\left(X - h^*(W) - \frac{S^{(1)}(m_0, t)[X - h^*(W)]}{S^{(0)}(m_0, t)}\right)\left(X - \frac{S^{(1)}(m_0, t)[X]}{S^{(0)}(m_0, t)}\right)(\hat{\beta} - \beta_0) \\
&\quad + P_\Delta\left(X - h^*(W) - \frac{S^{(1)}(m_0, t)[X - h^*(W)]}{S^{(0)}(m_0, t)}\right)\left(\hat{\phi}(W) - \phi_0(W) - \frac{S^{(1)}(m_0, t)[\hat{\phi} - \phi_0]}{S^{(0)}(m_0, t)}\right) + O_p(\|\hat{m} - m_0\|^2) \\
&= P_\Delta\left(X - h^*(W) - \frac{S^{(1)}(m_0, t)[X - h^*(W)]}{S^{(0)}(m_0, t)}\right)^{\otimes 2}(\hat{\beta} - \beta_0) + o_p(n^{-1/2}),
\end{aligned} \tag{18}$$

where the last step used that

$$P_\Delta\left(X - h^*(W) - \frac{S^{(1)}(m_0, t)[X - h^*(W)]}{S^{(0)}(m_0, t)}\right)\left(\hat{\phi}(W) - \phi_0(W) - \frac{S^{(1)}(m_0, t)[\hat{\phi} - \phi_0]}{S^{(0)}(m_0, t)}\right) = 0,$$

by the definition of h^*.

Furthermore, similar to A_{2n} above,

$$\begin{aligned}
&n^{-1}\sum_i \int_0^\tau (X_i - h^*(W_i)) - \frac{S_n^{(1)}(m_0, t)[X - h^*(W)]}{S_n^{(0)}(m_0, t)} dN_i(t) \\
&= n^{-1}\sum_i \int_0^\tau (X_i - h^*(W_i)) - \frac{S_n^{(1)}(m_0, t)[X - h^*(W)]}{S_n^{(0)}(m_0, t)} dM_i(t) \\
&= n^{-1}\sum_i \int_0^\tau (X_i - h^*(W_i)) - \frac{S^{(1)}(m_0, t)[X - h^*(W)]}{S^{(0)}(m_0, t)} dM_i(t) + o_p(n^{-1/2})
\end{aligned} \tag{19}$$

and can be seen to be asymptotically normal, where $M_i(t) = N_i(t) - \int_0^t Y_i(t)\exp[m_0(R_i)]\lambda_0(t)dt$. Combining (16)–(19), the asymptotic normality of $\hat{\beta}$ follows. We also note that

$$\begin{aligned}
P_\Delta\left(X - h^*(W) - \frac{S^{(1)}(m_0, t)[X - h^*(W)]}{S^{(0)}(m_0, t)}\right)^{\otimes 2} &= P_\Delta(X - h^*(W) - E[X - h^*(W)]|T = t, \Delta = 1)^{\otimes 2} \\
&= P_\Delta(X - h^*(W) - a^*(T))^{\otimes 2}. \quad \square
\end{aligned}$$

Proof of Theorems 3 and 4. We first show the convergence rate (6) in Theorem 4. This only requires a minor modification of the proof of Theorem 1. We will show that with probability approaching 1, uniformly over the unit vector $u = (u_1, u_2)$, where u_1 is a qK-vector and u_2 is a p-vector,

$$l((a_0, \beta_0) + \gamma_n u) - n\sum_{j=1}^p p_\lambda(|\beta_{0j} + \gamma_n u_{2j}|) - l(a_0, \beta_0) + n\sum_{j=1}^p p_\lambda(|\beta_{0j}|) < 0. \tag{20}$$

The term $l((a_0, \beta_0) + \gamma_n u) - l(a_0, \beta_0)$ is obviously the same as in the proof of Theorem 1. When $j > s$, $-np_\lambda(|\beta_{0j} + \gamma_n u_{2j}|) + np_\lambda(|\beta_{0j}|) = -np_\lambda(|\beta_{0j} + \gamma_n u_{2j}|) \le 0$. When $j \le s$, we have $p_\lambda(|\beta_{0j}|) = 0$ and $p_\lambda(|\beta_{0j} + \gamma_n u_{2j}|) = 0$ by condition (C6). Thus the penalty terms do not affect the negativity of (20).

Next, we proceed to show the variable selection consistency (Theorem 3). By way of contradiction, suppose for some $j > s$, $\hat\beta_j \neq 0$. Let $\hat\beta^*$ be the same as $\hat\beta$ except that its jth component is replaced by the true parameter value 0. Then by Taylor expansion we have

$$pl(\hat a, \hat\beta) - pl(\hat a, \hat\beta^*) = U_{nj}(\hat a, \tilde\beta^*)(\hat\beta_j) - np_\lambda(|\hat\beta_j|) \tag{21}$$

where $\tilde\beta^*$ lies between $\hat\beta$ and $\hat\beta^*$. By the convergence rate (6), $|\hat\beta_j| = O_p(\sqrt{(K+p)/n} + K^{-d}) = o_p(\lambda)$ which implies $np_\lambda(|\hat\beta_j|) = n\lambda|\hat\beta_j|$ by the definition of the SCAD penalty. Similar to (12), $|U_{nj}(\hat a, \tilde\beta^*)| = O_p(\sqrt{n(K+s)} + nK^{-d})$. By condition (C6), (21) will be negative leading to a contradiction.

Given that $\hat\beta^{(2)} = 0$ with probability tending to 1, the proof of asymptotic normality of $\hat\beta^{(1)}$ is basically the same as the proof of Theorem 2. □

Proof of Theorem 5. The proof of Theorem 5 is based on the following proposition, which is a direct extension of Theorem 1 in [8] to the case of partial likelihood (but specialized to the SCAD penalty instead of the more general ones considered there). A similar condition was also used in [13] in linear models (see the proof of their Theorem 1). Thus the proof of the following proposition is omitted.

Proposition 1. *$(a, \beta) \in R^{qK+p}$ is a local maximizer of the SCAD-penalized partial likelihood (5) if*

$$\sum_i \delta_i \left\{ Z_{ij} - \frac{\sum_k Y_k(T_i) Z_{kj} e^{m(R_k)}}{\sum_k Y_k(T_i) e^{m(R_k)}} \right\} = 0 \tag{22}$$

$$\sum_i \delta_i \left\{ X_{ij} - \frac{\sum_k Y_k(T_i) X_{kj} e^{m(R_k)}}{\sum_k Y_k(T_i) e^{m(R_k)}} \right\} = 0 \quad \text{and} \quad |\beta_j| \ge c\lambda \quad \text{for } j \le s, \tag{23}$$

$$\max_{s+1\le j\le p} \left| \sum_i \delta_i \left\{ X_{ij} - \frac{\sum_k Y_k(T_i) X_{kj} e^{m(R_k)}}{\sum_k Y_k(T_i) e^{m(R_k)}} \right\} \right| \le n\lambda \quad \text{and} \quad \max_{s+1\le j\le p} |\beta_j| < \lambda, \tag{24}$$

where $Z_{ij} = (B_{j1}(W_{ij}), \ldots, B_{jK}(W_{ij}))^T \in R^K$ and $m(R_i) = \sum_{j=1}^q a_{jk} B_{jk}(W_{ij}) + X_i^T \beta$.

Let $(\hat a^o, \hat\beta^o)$ be the oracle estimator as defined in the main text. The rest of the proof consists of showing that $(\hat a = \hat a^o, \hat\beta^{(1)} = \hat\beta^o, \hat\beta^{(2)} = 0)$ satisfies (22)–(24). It can be seen that (22) and (23) hold trivially by the definition of the oracle estimator (they are just the first order conditions for the oracle estimation problem). Note that $|\hat\beta_j| \ge c\lambda$, $j \le s$ is implied by

$$\min_{1\le j\le s} |\beta_{0j}| \gg \lambda,$$
$$|\hat\beta_j - \beta_{0j}| \ll \lambda,$$

and both equations above are implied by (C6) as well as the convergence rates of the oracle estimator.

For $j = s+1, \ldots, p$, that $|\hat\beta_j| < \lambda$ is trivial since $\hat\beta_j = 0$. Furthermore, letting $\hat m(R) = Z^T \hat a^o + X^{(1)} \hat\beta^o$, we have

$$\begin{aligned} \left| \sum_i \delta_i \left\{ X_{ij} - \frac{\sum_k Y_k(T_i) X_{kj} e^{\hat m(R_k)}}{\sum_k Y_k(T_i) e^{\hat m(R_k)}} \right\} \right| \le{}& \left| \sum_i \delta_i \left\{ \frac{\sum_k Y_k(T_i) X_{kj} e^{\hat m(R_k)}}{\sum_k Y_k(T_i) e^{\hat m(R_k)}} - \frac{\sum_k Y_k(T_i) X_{kj} e^{m_0(R_k)}}{\sum_k Y_k(T_i) e^{m_0(R_k)}} \right\} \right| \\ &+ \left| \sum_i \left\{ \frac{\sum_k Y_k(T_i) X_{kj} e^{m_0(R_k)}}{\sum_k Y_k(T_i) e^{m_0(R_k)}} - \frac{S_j^{(1)}(m_0, T_i)[X]}{S_0(m_0, T_i)} \right\} \right| \\ &+ \left| \sum_i \delta_i \left\{ X_{ij} - \frac{S_j^{(1)}(m_0, T_i)[X]}{S_0(m_0, T_i)} \right\} \right| \end{aligned} \tag{25}$$

where $S_j^{(1)}(m_0, t)[X]$ is the jth component of $S^{(1)}(m_0, t)[X]$.

Using Lemma A.3(ii), and similar to Lemma A.6 in [10], we have that the first term above is of order $O_p(\sqrt{n(K+s)}+nK^{-d})$. Similarly, the second term is $o_p(\sqrt{n})$. Using assumption (C8), by Bernstein's inequality (for example Lemma 5.7 in [24]), we have

$$P\left(\max_{s+1\le j\le p}\left|\sum_i \delta_i\left\{X_{ij}-\frac{S_j^{(1)}(m_0,T_i)[X]}{S^{(0)}(m_0,T_i)}\right\}\right|>b\right)\le Cp\exp\left\{-\frac{b^2/n}{2(bJ/n+D^2)}\right\}.$$

Taking $b=C\sqrt{n\max\{\log(p),\log(n)\}}$ for C sufficiently large will make the above converge to zero, which implies

$$\max_{s+1\le j\le p}\left|\sum_i \delta_i\left\{X_{ij}-\frac{S_j^{(1)}(m_0,T_i)[X]}{S^{(0)}(m_0,T_i)}\right\}\right|=O_p(\sqrt{n\max\{\log(p),\log(n)\}}).$$

Thus (25) is of order $\sqrt{n(K+s)}+nK^{-d}+\sqrt{n\max\{\log(p),\log(n)\}}\ll n\lambda$ by assumption (C7), which verifies (24) and completes the proof. □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmva.2013.12.002.

References

[1] J. Bradic, J. Fan, J. Jiang, Regularization for Cox's proportional hazards model with NP-dimensionality, Ann. Statist. 39 (6) (2011) 3092–3120.
[2] L. Breiman, Heuristics of instability and stabilization in model selection, Ann. Statist. 24 (6) (1996) 2350–2383.
[3] J. Chen, Z. Chen, Extended Bayesian information criteria for model selection with large model spaces, Biometrika 95 (3) (2008) 759–771.
[4] P. Du, S. Ma, H. Liang, Penalized variable selection procedure for Cox models with semiparametric relative risk, Ann. Statist. 38 (4) (2010) 2092–2117.
[5] J. Fan, I. Gijbels, M. King, Local likelihood and local partial likelihood in hazard regression, Ann. Statist. 25 (4) (1997) 1661–1690.
[6] J.Q. Fan, R.Z. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (456) (2001) 1348–1360.
[7] J. Fan, R. Li, Variable selection for Cox's proportional hazards model and frailty model, Ann. Statist. 30 (1) (2002) 74–99.
[8] J.Q. Fan, J. Lv, Non-concave penalized likelihood with NP-dimensionality, IEEE Trans. Inform. Theory 57 (8) (2011) 5467–5484.
[9] Y. Hu, H. Lian, Variable selection in partially linear proportional hazards model with a diverging dimensionality, Statist. Probab. Lett. 83 (1) (2013) 61–69.
[10] J. Huang, Efficient estimation of the partly linear additive Cox model, Ann. Statist. 27 (5) (1999) 1536–1563.
[11] J. Huang, J.L. Horowitz, S.G. Ma, Asymptotic properties of bridge estimators in sparse high-dimensional regression models, Ann. Statist. 36 (2) (2008) 587–613.
[12] J. Huang, J.L. Horowitz, F. Wei, Variable selection in nonparametric additive models, Ann. Statist. 38 (4) (2010) 2282–2313.
[13] Y. Kim, H. Choi, H. Oh, Smoothly clipped absolute deviation on high dimensions, J. Amer. Statist. Assoc. 103 (484) (2008) 1665–1673.
[14] R. Li, H. Liang, Variable selection in semiparametric regression modeling, Ann. Statist. 36 (1) (2008) 261–286.
[15] H. Lian, Variable selection in high-dimensional generalized varying-coefficient models, Statist. Sinica 22 (2012) 1563–1588.
[16] X. Liu, L. Wang, H. Liang, Estimation and variable selection for semiparametric additive partial linear models, Statist. Sinica 21 (2011) 1225–1248.
[17] S. Ma, P. Du, Variable selection in partly linear regression model with diverging dimensions for right censored data, Statist. Sinica 22 (2012) 1003–1020.
[18] L. Meier, S. Van de Geer, P. Buhlmann, High-dimensional additive modeling, Ann. Statist. 37 (6B) (2009) 3779–3821.
[19] F. O'Sullivan, Nonparametric estimation in the Cox model, Ann. Statist. 21 (1993) 124–145.
[20] P. Sasieni, Non-orthogonal projections and their application to calculating the information in a partly linear Cox model, Scand. J. Stat. 19 (1992) 215–233.
[21] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for Cox's proportional hazards model via coordinate descent, J. Stat. Softw. 39 (5) (2011) 1–13.
[22] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1) (1996) 267–288.
[23] R. Tibshirani, The lasso method for variable selection in the Cox model, Stat. Med. 16 (4) (1997) 385–395.
[24] S.A. van der Geer, Applications of Empirical Process Theory, Cambridge University Press, Cambridge, 2000.
[25] L.F. Wang, H.Z. Li, J.H.Z. Huang, Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements, J. Amer. Statist. Assoc. 103 (484) (2008) 1556–1569.
[26] H.S. Wang, B. Li, C.L. Leng, Shrinkage tuning parameter selection with a diverging number of parameters, J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (2009) 671–683.
[27] H. Wang, R. Li, C.L. Tsai, Tuning parameter selectors for the smoothly clipped absolute deviation method, Biometrika 94 (3) (2007) 553–568.
[28] L. Wang, X. Liu, H. Liang, R. Carroll, Estimation and variable selection for generalized additive partially linear models, Ann. Statist. 39 (4) (2011) 1827–1851.
[29] H.S. Wang, Y.C. Xia, Shrinkage estimation of the varying coefficient model, J. Amer. Statist. Assoc. 104 (486) (2009) 747–757.
[30] L. Wang, J. Zhou, A. Qu, Penalized generalized estimating equations for high-dimensional longitudinal data analysis, Biometrics 68 (2) (2012) 353–360.
[31] F. Wei, J. Huang, H. Li, Variable selection in high-dimensional varying-coefficient models, Statist. Sinica 21 (2011) 1515–1540.
[32] H.L. Xie, J. Huang, SCAD-penalized regression in high-dimensional partially linear models, Ann. Statist. 37 (2) (2009) 673–696.
[33] L. Xue, Consistent variable selection in additive models, Statist. Sinica 19 (2009) 1281–1296.
[34] C. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist. 38 (2) (2010) 894–942.
[35] H. Zhang, G. Cheng, Y. Liu, Linear or nonlinear? Automatic structure discovery for partially linear models, J. Amer. Statist. Assoc. 106 (495) (2011) 1099–1112.
[36] H. Zhang, W. Lu, Adaptive lasso for Cox's proportional hazards model, Biometrika 94 (3) (2007) 691–703.
[37] L. Zhu, L. Zhu, Nonconcave penalized inverse regression in single-index models with high dimensional predictors, J. Multivariate Anal. 100 (5) (2009) 862–875.
[38] H. Zou, The adaptive lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (476) (2006) 1418–1429.
[39] H. Zou, R.Z. Li, One-step sparse estimates in nonconcave penalized likelihood models, Ann. Statist. 36 (4) (2008) 1509–1533.