ORACLE INEQUALITIES AND SELECTION CONSISTENCY FOR WEIGHTED LASSO IN HIGH-DIMENSIONAL ADDITIVE HAZARDS MODEL

Haixiang Zhang[1], Liuquan Sun[2], Yong Zhou[3] and Jian Huang[4]

Tianjin University[1], *Chinese Academy of Sciences*[2],

Shanghai University of Finance and Economics[3] *and University of Iowa*[4]

Abstract: The additive hazards model has many applications in high-throughput genomic data analysis and clinical studies. In this article, we study the weighted Lasso estimator for the additive hazards model in sparse, high-dimensional settings where the number of time-dependent covariates is much larger than the sample size. Based on compatibility, cone invertibility factors, and restricted eigenvalues of the Hessian matrix, we establish some non-asymptotic oracle inequalities for the weighted Lasso. Under mild conditions, we show that these quantities are bounded from below by positive constants, thus the compatibility and cone invertibility factors can be treated as positive constants in the oracle inequalities. A multistage adaptive method with weights recursively generated from a concave penalty is presented. We prove a selection consistency theorem and establish an upper bound for dimension of the weighted Lasso estimator.

Key words and phrases: High-dimensional covariates, oracle inequalities, sign consistency, survival analysis, variable selection.

1 Introduction

Censored survival data arises in such fields as epidemiological studies and clinical trials. The additive hazards (AH) model is an important alternative to the Cox (1972) proportional hazards model for studying the association between such data and risk factors (Cox and Oakes (1984)). In a traditional biomedical study, the number of covariates p is usually relatively small compared to the sample size n. Theoretical properties of the AH model in the fixed p and large n setting have been well established. For example, Lin and Ying (1994) proposed a least-squares type estimator of regression parameter in the AH model and studied its asymptotic properties using martingale techniques; Kulich and Lin (2000) studied the AH model when covariates are subject to measurement error; Martinussen and Scheike (2002) proposed an efficient estimation approach in AH regression with current status data.

In recent years, advances in experimental technologies have brought in a wealth of high-throughput and high-dimensional genomic data, where an important task is to find genetic risk factors related to clinical outcomes, such as survival and age of disease onset. In such high-dimensional settings, the standard approach to the AH model is not applicable, since the number of potential genetic risk factors is typically much larger than the sample size, and regularized methods that can do variable selection and estimation have been proposed. Examples include the Lasso (Tibshirani (1996)), SCAD (Fan and Li (2001)) and MCP (Zhang (2010)). Much of the work on the theoretical properties of these methods has focused on linear and generalized linear regression models; see Bühlmann and van de Geer (2011), Fan and Lv (2010), Zhang and Zhang (2012), and the references therein. Several authors have studied these methods for the Cox regression model in sparse, high-dimensional settings.

In particular, oracle inequalities for the prediction and estimation error of the Lasso in the Cox model (Kong and Nan (2014); Lemler (2012); Huang et al. (2013)); Bradic, Fan and Jiang (2011) extended the results of Fan and Li (2002) to a class of concave penalties in the high-dimensional Cox model under certain sparsity and regularity conditions.

Variable selection for survival data has also been extended to the AH model. In fixed dimensional settings, Leng and Ma (2007) proposed a weighted Lasso approach, and Martinussen and Scheike (2009) considered several regularization methods, including the Lasso and the Dantzig selector. In high-dimensional settings, Gaïffas and Guilloux (2012) considered a general AH model in a non-asymptotic setting; Lin and Lv (2013) studied a class of regularization methods for simultaneous variable selection and estimation in this model. In view of the important role of the AH model in survival analysis and the basic importance of the Lasso as a regularization method, it is of interest to understand the properties of the weighted Lasso for this model in the $p \gg n$ setting.

In this paper we establish the theoretical properties of the weighted Lasso in the high-dimensional AH model concerning estimation error bounds, selection consistency, and s-parsity. We obtain some non-asymptotic oracle inequalities for the weighted Lasso in the high-dimensional AH model, extending the oracle inequalities for the Lasso in Cox regression (Huang et al. (2013)) to the AH model. Under mild conditions, we prove that the compatibility and cone invertibility factors, and the corresponding restricted eigenvalue are greater than a fixed positive constant. We provide sufficient conditions under which the weighted Lasso is sign consistent in the AH model, generalizing the irrepresentable condition for the sign consistence of the Lasso in linear regression (Zhao and Yu (2006)). The sparsity property of the weighted Lasso in AH model is also proved.

The remainder of this article is organized as follows. In Section 2, we describe the AH model and introduce the weighted Lasso penalty. In Section 3, we establish some oracle inequalities for the weighted Lasso in the high-dimensional AH model. The compatibility and cone invertibility factors and the corresponding restricted eigenvalue of the Hessian matrix are presented. In Section 4, a multistage adaptive method is provided, we give some sufficient conditions for selection consistency, and provide an upper bound on the dimension of the weighted Lasso estimator. Section 5 includes some concluding remarks. Proofs are in the Appendix.

2 AH model with the weighted ℓ_1 penalty

We adopt the counting process framework for the AH model (Lin and Ying (1994)). Consider a set of n independent subjects such that the counting process $\{N_i(t); t \geq 0\}$ is the number of observed events for the ith individual in time interval $[0,\ t]$. Assume that the intensity function for $N_i(t)$ is given by

$$d\Lambda_i(t) = Y_i(t)\{d\Lambda_0(t) + \boldsymbol{\beta}_0' \boldsymbol{Z}_i(t)dt\}, \tag{1}$$

where $\boldsymbol{\beta}_0 = (\beta_{01}, \cdots, \beta_{0p})'$ is a p-vector of true regression coefficients, $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ denotes the cumulative baseline hazard function, $Y_i(t) \in \{0, 1\}$ is a predictable at-risk indicator process for the ith individual, and $\boldsymbol{Z}(\cdot) = (Z_1(\cdot), \cdots, Z_p(\cdot))'$ is a predicable covariate process. In the $p \gg n$ setting, let S be any set of indices with $S \supseteq \{j : \beta_{0j} \neq 0\}$, with S^c the complement of S in $\{1, \cdots, p\}$. Let $d_0 = |S|$ be the number of elements in S. Here we are interested in the case where d_0 is much smaller than the dimension of $\boldsymbol{\beta}_0$.

Following Lin and Ying (1994), we introduce the pseudoscore estimating function

$$U(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_{0}^{\tau}\{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}_n(t)\}\{dN_i(t) - Y_i(t)\boldsymbol{\beta}'\boldsymbol{Z}_i(t)dt\},$$

where $\bar{\boldsymbol{Z}}_n(t) = \sum_{j=1}^{n} Y_j(t)\boldsymbol{Z}_j(t)/\sum_{j=1}^{n} Y_j(t)$, and τ is the maximum follow-up time. After some algebra, we can get that $U(\boldsymbol{\beta}) = \boldsymbol{a} - \boldsymbol{A}\boldsymbol{\beta}$ with $\boldsymbol{a} = n^{-1}\sum_{i=1}^{n}\int_{0}^{\tau}\{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}_n(t)\}dN_i(t)$ and

$$\boldsymbol{A} = \frac{1}{n}\sum_{i=1}^{n}\int_{0}^{\tau} Y_i(t)\{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}_n(t)\}^{\otimes 2}dt, \tag{2}$$

where $\mathbf{c}^{\otimes 2} = \mathbf{c}\mathbf{c}'$ for any vector $\mathbf{c}$. For technical convenience, we rewrite the estimating function $U(\boldsymbol{\beta})$ in terms of a martingale, as suggested by Lin and Ying (1994),

$$U(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_{0}^{\tau}\{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}_n(t)\}dM_i(t),$$

where $M_i(t) = N_i(t) - \int_0^t Y_i(u)\{\lambda_0(u) + \boldsymbol{\beta}_0'\boldsymbol{Z}_i(u)\}du$ is a martingale. By integrating $-U(\boldsymbol{\beta})$ with respective to $\boldsymbol{\beta}$, we obtain a least-squares-type loss function (Martinussen and Scheike (2009)),

$$L(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}'\mathbf{A}\boldsymbol{\beta} - \boldsymbol{a}'\boldsymbol{\beta}. \tag{3}$$

The gradient of $L(\boldsymbol{\beta})$ is $\dot{L}(\boldsymbol{\beta}) = \partial L(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{a}$, and the Hessian matrix of $L(\boldsymbol{\beta})$ is $\ddot{L}(\boldsymbol{\beta}) = \boldsymbol{A}$. Here $\mathbf{A}$ is free of $\boldsymbol{\beta}$, which is a major difference with the theory for Cox model (Huang et al. (2013)).

Since $\mathbf{A}$ is singular in the $p \gg n$ setting, it is difficult to derive the estimator for $\boldsymbol{\beta}_0$ by minimizing (3) directly, so we employ the regularized approach. Let $\hat{w} \in \mathbb{R}^p$ be a (possibly estimated) weight vector with nonnegative elements $\hat{w}_j$, $1 \le j \le p$, and $\hat{\mathbf{W}} = \text{diag}(\hat{w})$. We consider the weighted ℓ_1-penalized least-squares-type loss criterion

$$Q(\boldsymbol{\beta};\lambda) = L(\boldsymbol{\beta}) + \lambda|\hat{\mathbf{W}}\boldsymbol{\beta}|_1, \tag{4}$$

where $\lambda \geq 0$ is a penalty parameter. Hereafter, we use the notation $|\boldsymbol{v}|_q = \{\sum_{i=1}^p |\boldsymbol{v}_j|^q\}^{1/q}$ for $1 \leq q < \infty$, and $|\boldsymbol{v}|_\infty = \max_{1\leq j\leq p} |\boldsymbol{v}_j|$ for any $\boldsymbol{v} \in \mathbb{R}^p$. For a given λ, the weighted ℓ_1-penalized estimator, or the weighted Lasso estimator is

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}; \lambda). \tag{5}$$

The weighted Lasso estimator can be characterized by the Karush-Kuhn-Tucker (KKT) conditions. Since $L(\boldsymbol{\beta})$ is convex, a vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)'$ is a solution to (5) if and only if

$$\begin{cases} \dot{L}_j(\hat{\boldsymbol{\beta}}) = -\lambda \hat{w}_j \mathrm{sgn}(\hat{\beta}_j), \text{ if } \hat{\beta}_j \neq 0, \\ |\dot{L}_j(\hat{\boldsymbol{\beta}})| \leq \lambda \hat{w}_j, \text{ if } \hat{\beta}_j = 0, \end{cases} \tag{6}$$

where $\dot{L}(\boldsymbol{\beta}) = (\dot{L}_1(\boldsymbol{\beta}), \cdots, \dot{L}_p(\boldsymbol{\beta}))' = \partial L(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ is the gradient of $\mathcal{L}(\boldsymbol{\beta})$. The (unweighted) Lasso is a special case of (5), with the choice $\hat{w}_j = 1$, $1 \leq j \leq p$.

3 Non-asymptotic oracle inequalities

In this section, we establish some non-asymptotic oracle inequalities for the estimation error of weighted Lasso in the high-dimensional AH model. Let $\mathbf{W} = \mathrm{diag}(w)$ for a possibly unknown vector $w \in \mathbb{R}^p$ with elements $w_j \geq 0$. As in Huang and Zhang (2012), we define

$$\begin{aligned} z^* &= \max\{|\dot{L}(\boldsymbol{\beta}_0)_S|_\infty,\ |W_{S^c}^{-1}\dot{L}(\boldsymbol{\beta}_0)_{S^c}|_\infty\}, \\ \Omega_0 &= \{\hat{w}_j \leq w_j, \forall j \in S\} \cap \{w_j \leq \hat{w}_j, \forall j \in S^c\}. \end{aligned}$$

Hereafter, for any p-vector $\mathbf{v} = (v_1, \cdots, v_p)'$ and sets $\mathcal{A}$ and $\mathcal{C}$, $\mathbf{v}_{\mathcal{A}} = (v_j : j \in \mathcal{A})'$, $M_{\mathcal{AC}}$ denotes the $\mathcal{A} \times \mathcal{C}$ subblock of a matrix M and $M_{\mathcal{A}} = M_{\mathcal{AA}}$.

Lemma 1 *Let $\hat{\boldsymbol{\beta}}$ be the weighted Lasso estimator, and $\hat{\mathbf{e}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$. Then in the event Ω_0,*

$$(\lambda - z^*)|\mathbf{W}_{S^c}\hat{\mathbf{e}}_{S^c}|_1 \leq D(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) + (\lambda - z^*)|\mathbf{W}_{S^c}\hat{\mathbf{e}}_{S^c}|_1 \leq (\lambda|w_S|_\infty + z^*)|\hat{\mathbf{e}}_S|_1.$$

Furthermore, for any $\xi > |w_S|_\infty$, $|\mathbf{W}_{S^c}\hat{\mathbf{e}}_{S^c}|_1 \leq \xi|\hat{\mathbf{e}}_S|_1$ in the event $\Omega_0 \cap \{z^* \leq \lambda(\xi - |w_S|_\infty)/(\xi + 1)\}$, where $D(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\{\dot{L}(\hat{\boldsymbol{\beta}}) - \dot{L}(\boldsymbol{\beta})\} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is the Bregman divergence(Gaïffas and Guilloux (2012)) and $\mathbf{A}$ is defined in (2).

It follows from Lemma 1 that in the event $\Omega_0 \cap \{z^* \leq \lambda(\xi - |w_S|_\infty)/(\xi + 1)\}$, for any $\xi > |w_S|_\infty$, the estimation error $\hat{\mathbf{e}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ belongs to the cone

$$\Theta(\xi, S) = \{\boldsymbol{b} \in \mathbb{R}^p : |\mathbf{W}_{S^c}\boldsymbol{b}_{S^c}|_1 \leq \xi|\boldsymbol{b}_S|_1\}. \tag{7}$$

To establish some useful oracle inequalities, for the cone in (7) and the Hessian matrix $\boldsymbol{A}$ in (2), we set

$$\kappa(\xi, S; \mathbf{A}) = \inf_{0 \neq \boldsymbol{b} \in \Theta(\xi, S)} \frac{d_0^{1/2}(\boldsymbol{b}'\mathbf{A}\boldsymbol{b})^{1/2}}{|\boldsymbol{b}_S|_1}$$

as the compatibility factor (van de Geer (2007); van de Geer and Bühlmann (2009)), and

$$F_q(\xi, S; \mathbf{A}) = \inf_{0 \neq \boldsymbol{b} \in \Theta(\xi, S)} \frac{d_0^{1/q}\boldsymbol{b}'\boldsymbol{A}\boldsymbol{b}}{|\boldsymbol{b}_S|_1|\boldsymbol{b}|_q} \tag{8}$$

as the weak cone invertibility factor (Ye and Zhang (2010)). The two quantities are closely related to the restricted eigenvalue (Bickel, Ritov and Tsybakov (2009); Koltchinskii (2009)), defined as

$$\mathrm{RE}(\xi, S; \mathbf{A}) = \inf_{0 \neq \boldsymbol{b} \in \Theta(\xi, S)} \frac{(\boldsymbol{b}'\boldsymbol{A}\boldsymbol{b})^{1/2}}{|\boldsymbol{b}|_2}.$$

According to Ye and Zhang (2010), the compatibility and cone invertibility factors are greater than the restricted eigenvalue. Therefore, using $\kappa(\xi, S; \boldsymbol{A})$ and $F_q(\xi, S; \boldsymbol{A})$ can yield shaper oracle inequalities than the restricted eigenvalue.

Theorem 1 *If $|\boldsymbol{Z}_i(t) - \boldsymbol{Z}_j(t)|_\infty \leq K$ uniformly in $\{t, i, j\}$ for a finite $K > 0$, and $\hat{\boldsymbol{\beta}}$ be the weighted Lasso estimator as (5), in the event $\Omega_0 \cap \{z^* \leq \lambda(\xi - |w_S|_\infty)/(\xi + 1)\}$,*

$$D(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) \leq \frac{\xi^2\lambda^2 d_0(1 + |w_S|_\infty)^2}{(\xi+1)^2\kappa^2(\xi, S; \mathbf{A})}, \ |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|_1 \leq \frac{\lambda d_0(1 + |w_S|_\infty)(\xi + \min\{w_{S^c}\})^2}{4\min\{w_{S^c}\}\kappa^2(\xi, S; \mathbf{A})(\xi + 1)}, \tag{9}$$

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|_q \leq \frac{d_0^{1/q}(\lambda|w_S|_\infty + z^*)}{F_q(\xi, S; \mathbf{A})}, \quad q \geq 1. \tag{10}$$

Remark 1 For $w_j = 1$, $1 \leq j \leq p$, the established error bounds for the AH model have the same form as those for the linear model (Huang et al. (2013)), except for an improved factor of $4\xi/(1+\xi) \geq 2$ for the ℓ_1 oracle inequality as (9).

The oracle inequalities in Theorem 1 hold only in the event $\Omega_0 \cap \{z^* \leq \lambda(\xi - |w_S|_\infty)/(\xi + 1)\}$, so a probabilistic upper bound for z^* is needed. We have $N_i(\infty) \leq 1$ and $\dot{L}(\boldsymbol{\beta}_0) = -n^{-1}\sum_{i=1}^n \int_0^\tau \{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}_n(t)\}dM_i(t)$. Without loss of generality, the martingale difference generated by $\{M_i(t), t > 0\}$ is bounded by 1. Then by martingale version of the Hoeffding inequality (Azuma (1967)) and Lemma 3.3 of Huang et al. (2013), we can get that $P\{z^* > Kx\} \leq 2pe^{-nx^2/2}$.

Theorem 2 *Suppose the conditions in Theorem 1 hold. Let $\xi > |w_S|_\infty$ and $\lambda = \{(\xi+1)/(\xi - |w_S|_\infty)\}K\sqrt{(2/n)\log(2p/\epsilon)}$ with a small $\epsilon > 0$. Then in the event Ω_0, for any $C_\kappa > 0$ and $C_{F,q} > 0$, we have*

$$D(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) \leq \frac{\xi^2\lambda^2 d_0(1 + |w_S|_\infty)^2}{(\xi+1)^2 C_\kappa^2}, \qquad |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|_1 \leq \frac{\lambda d_0(1 + |w_S|_\infty)(\xi + \min\{w_{S^c}\})^2}{4\min\{w_{S^c}\}C_\kappa^2(\xi + 1)},$$

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|_q \leq \frac{\xi d_0^{1/q}\lambda(|w_S|_\infty + 1)}{(\xi+1)C_{F,q}}, \quad q \geq 1,$$

all hold with probability at least $P\{\kappa(\xi, S; \mathbf{A}) \geq C_\kappa, F_q(\xi, S; \mathbf{A}) \geq C_{F,q}\} - \epsilon$.

Remark 2 By Theorem 2, to ensure the error $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|_q$ is small with high probability, it is required that $p = \exp\{o(n/d_0^{1/q})\}$. If d_0 is bounded, then p can be as large as $\exp(o(n))$.

We have established non-asymptotic oracle inequalities expressed in terms of compatibility and weak cone invertibility factors. As the Hessian matrix is based on the cross-products of time-dependent covariates in censored risk sets, these quantities are random variables. We provide some sufficient conditions under which they can be treated as constants, and since these factors appear in the denominator of the error bounds, it suffices to bound them from below. To simplify the statement of the results, we use $\Phi(\xi, S; \boldsymbol{A})$ to denote any of the quantities:

$$\Phi(\xi, S; \mathbf{A}) = \kappa^2(\xi, S; \boldsymbol{A}), F_q(\xi, S; \boldsymbol{A}), \text{ and } \mathrm{RE}^2(\xi, S; \boldsymbol{A}). \tag{11}$$

If we make a claim about $\Phi(\xi, S; \boldsymbol{A})$, then the claim holds for any quantity in (11).

Lemma 2 *Let $\kappa^2(\xi, S; \boldsymbol{A}), F_q(\xi, S; \boldsymbol{A}), \mathrm{RE}^2(\xi, S; \boldsymbol{A})$ and $\Phi(\xi, S; \mathbf{A})$ be defined in (11). Denote A_{ij} as the elements of $\mathbf{A}$ and let $\mathbf{B}$ is another nonnegative-definite matrix with elements B_{ij}, then*

(i) for $1 \le q \le 2$,

$$\min\{\kappa^2(\xi, S; \boldsymbol{A}), (1 + \min\{w_{S^c}\}^{-1}\xi)^{2/q-1} F_q(\xi, S; \boldsymbol{A})\} \ge \quad \mathrm{RE}^2(\xi, S; \boldsymbol{A}) \ge \Lambda_{\min}(\mathbf{A}),$$

where $\Lambda_{\min}(\cdot)$ denotes the smallest eigenvalue,

(ii) $\Phi(\xi, S; \boldsymbol{A}) \ge \Phi(\xi, S; \boldsymbol{B}) - d_0(1 + \min\{w_{S^c}\}^{-1}\xi)^2 \max_{1 \le i \le j \le p} |A_{ij} - B_{ij}|$,

(iii) if $\mathbf{A} \geq \mathbf{B}$*, then* $\Phi(\xi, S; \boldsymbol{A}) \geq \Phi(\xi, S; \boldsymbol{B})$*, where* $\mathbf{A} \geq \mathbf{B}$ *means* $\mathbf{A} - \mathbf{B}$ *is nonnegative definite.*

As in Huang et al. (2013), we can bound the quantities of type $\Phi(\xi, S; \boldsymbol{A})$ from below in two ways: bound the matrix $\boldsymbol{A}$ from below, or approximate $\boldsymbol{A}$ under the supreme norm for its elements. Here we choose a suitable truncation of $\boldsymbol{A} = \ddot{L}(\boldsymbol{\beta}_0)$ as a lower bound of the matrix. This is done by truncating the maximum event time under consideration. Since $\ddot{L}(\boldsymbol{\beta}_0) = n^{-1}\sum_{i=1}^n \int_0^\tau Y_i(t)\{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}_n(t)\}^{\otimes 2}dt$, then $\ddot{L}(\boldsymbol{\beta}_0) \geq \bar{\boldsymbol{A}}(t^*)$ with $\bar{\boldsymbol{A}}(t^*) = \int_0^{t^*} \bar{\Sigma}_n(t)dt$, where $\bar{\Sigma}_n(t) = n^{-1}\sum_{i=1}^n Y_i(t)\{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}_n(t)\}^{\otimes 2}$, and $t^* > 0$. Suppose that $\{Y_i(t), \boldsymbol{Z}_i(t), t > 0\}$ are i.i.d. stochastic processes of $\{Y(t), \boldsymbol{Z}(t), t > 0\}$. The population version of $\bar{\boldsymbol{A}}(t^*)$ is $\boldsymbol{A}(t^*) = E(\int_0^{t^*} \Sigma_n(t)dt)$, where $\Sigma_n(t) = n^{-1}\sum_{i=1}^n Y_i(t)\{\boldsymbol{Z}_i(t) - \boldsymbol{\mu}(t)\}^{\otimes 2}$ with $\boldsymbol{\mu}(t) = E\{Y(t)\boldsymbol{Z}(t)\}/E(Y(t))$. Let $F_n(t) = \sqrt{(2/n)\log t}$, then we have the following results.

Theorem 3 *Suppose that* $\{Y_i(t), \boldsymbol{Z}_i(t), t \geq 0\}$ *are i.i.d. processes as* $\{Y(t), \boldsymbol{Z}(t), t \geq 0\}$ *with* $\sup_t P\{|\boldsymbol{Z}_i(t) - \boldsymbol{Z}(t)|_\infty \leq K\} = 1$*. If* t^* *be a positive constant and* $r_* = EY(t^*)$*, then*

$$\Phi(\xi, S; \ddot{L}(\boldsymbol{\beta}_0)) \geq \Phi(\xi, S; \boldsymbol{A}(t^*)) - d_0(1 + \min\{w_{S^c}\}^{-1}\xi)^2 K^2 t^* \{F_n(p(p+1)/\epsilon) + (2/r_*)t_{n,p,\epsilon}^2\}$$

with probability at least $1 - 2\epsilon$*, where* $t_{n,p,\epsilon}$ *is the solution of* $p(p+1)\exp\{-nt_{n,p,\epsilon}^2/(2 + 2t_{n,p,\epsilon}/3)\} = \epsilon/2.221$*. Furthermore, for* $1 \leq q \leq 2$*,*

$$\begin{aligned} & \min\{\kappa^2(\xi, S; \mathbf{A}), (1 + \min\{w_{S^c}\}^{-1}\xi)^{2/q-1} F_q(\xi, S; \mathbf{A})\} \\ \geq\ & \mathrm{RE}^2(\xi, S; \ddot{L}(\boldsymbol{\beta}_0)) \\ \geq\ & \Lambda_{\min}(\boldsymbol{A}(t^*)) - d_0(1 + \min\{w_{S^c}\}^{-1}\xi)^2 K^2 t^* \{F_n(p(p+1)/\epsilon) + (2/r_*)t_{n,p,\epsilon}^2\} \end{aligned}$$

with probability greater than $1 - 2\epsilon$*, where* $\Lambda_{\min}(\cdot)$ *denotes the smallest eigenvalue.*

Accordingly, the compatibility and cone invertibility factors and the restricted eigenvalue can be treated as constants in the high-dimensional AH model with time-dependent covariates. Our discussion focuses on the quantities in $\Phi(\xi, S; \boldsymbol{A})$ for the Hessian matrix $\boldsymbol{A}$. But, since $\ddot{L}(\boldsymbol{\beta}_0 + \tilde{\boldsymbol{b}}) = \ddot{L}(\boldsymbol{\beta}_0) = \boldsymbol{A}$, for any $\tilde{\boldsymbol{b}} \in \mathbb{R}^p$, Theorem 3 provides lower bounds for these quantities at any $\boldsymbol{\beta}$. This conclusion is different from those for Cox regression model (Huang et al. (2013)), which only provide lower bounds for these quantities with $\boldsymbol{\beta}$ not far from $\boldsymbol{\beta}_0$ in terms of ℓ_1-distance.

An earlier result on oracle inequalities for the high-dimensional AH model is due to Gaïffas and Guilloux (2012), who considered a data-driven ℓ_1 penalization and proved oracle inequalities for a more general non-parametric AH model. They only focused on the time-independent covariates case. Lin and Lv (2013) studied the properties of a class of concave penalties, including the Lasso for the AH model. They obtained ℓ_∞ error bounds and asymptotic oracle properties for the regression coefficient under different conditions from what we assumed here. A key assumption in their results is a strong version of the irrepresentable condition, which is not required in our results on the error bounds.

4 Multistage adaptive method and selection consistency

In this section, we consider how to choose the weights $\hat{w}_j$ in (4), for $j = 1, \cdots, p$. A multistage adaptive approach is proposed with weights recursively generated from a concave penalty function, e.g. SCAD (Fan and Li (2001)) and MCP (Zhang (2010a)). Let $P_\lambda(t)$ be a concave penalty with $\dot{P}_\lambda(0+) = \lambda$. The maximum concavity of this penalty is

$$\varpi = \sup_{0<t_1<t_2} \frac{|\dot{P}_\lambda(t_2) - \dot{P}_\lambda(t_1)|}{t_2 - t_1}, \tag{12}$$

where $\dot{P}_\lambda(t) = (\partial/\partial t)P_\lambda(t)$.

Theorem 4 *If $\phi > 1$, $\xi \geq (\phi+1)/(\phi-1)$, $\tilde{\boldsymbol{\beta}}$ is an initial estimator of $\boldsymbol{\beta}_0$, and $\hat{\boldsymbol{\beta}}$ is the weighted Lasso estimator in (5) with weights $\hat{w}_j = \dot{P}_\lambda(|\tilde{\beta}_j|)/\lambda$, for $j = 1, \cdots, p$. Then in the event $\Omega_0 \cap \{z^* \leq \lambda/\phi\}$,*

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|_1 \leq \frac{d_0}{F_1(\xi, S; \mathbf{A})} \left\{ |\dot{P}_\lambda(|\boldsymbol{\beta}_{0S}|)|_1 + \frac{d_0 \lambda}{\phi} + \varpi |\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|_1 \right\}, \tag{13}$$

where $P_\lambda(\cdot)$ is a concave penalty, and $F_1(\xi, S; \mathbf{A})$ is defined in (8) with $q = 1$.

Thus the weighted Lasso $\hat{\boldsymbol{\beta}}$ improves its initial estimator $\tilde{\boldsymbol{\beta}}$, and we can repeatedly apply this procedure with the multistage algorithm (Zhang (2010b)),

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \arg\min_{\beta} \left\{ L(\boldsymbol{\beta}) + \sum_{j=1}^{p} \dot{P}_\lambda(\hat{\beta}_j^{(k)})|\beta_j| \right\}, \quad k = 0, 1, \cdots,$$

where $L(\boldsymbol{\beta})$ is defined in (3).

Define $\|M\|_\infty = \max_{|u|_\infty \leq 1} |Mu|_\infty$ as the ℓ_∞ to ℓ_∞ norm of a matrix M. We have the following results on selection consistency and sparsity for the weighted Lasso estimator $\hat{\boldsymbol{\beta}}$ in (5).

Theorem 5 *(i) If $\mathfrak{B}_0^* = \{\boldsymbol{\beta} : \boldsymbol{\beta}_{S^c} = 0\}$ and $S_{\boldsymbol{\beta}} = \{j : \beta_j \neq 0\}$, and if*

$$\sup_{\boldsymbol{\beta} \in \mathfrak{B}_0^*} |\hat{\mathbf{W}}_{S^c}^{-1} \mathbf{A}_{S^c S_{\boldsymbol{\beta}}} \mathbf{A}_{S_{\boldsymbol{\beta}}}^{-1} \hat{\mathbf{W}}_{S_{\boldsymbol{\beta}}} \mathrm{sgn}(\boldsymbol{\beta}_{S_{\boldsymbol{\beta}}})|_\infty \leq \kappa_0 < 1, \tag{14}$$

$$\sup_{\boldsymbol{\beta} \in \mathfrak{B}_0^*} \| \hat{\mathbf{W}}_{S^c}^{-1} \mathbf{A}_{S^c S_{\boldsymbol{\beta}}} \mathbf{A}_{S_{\boldsymbol{\beta}}}^{-1} \|_\infty \leq \kappa_1 \tag{15}$$

hold, then $\{j : \hat{\beta}_j \neq 0\} \subseteq S$ *in the event*

$$\Omega_1 = \Omega_0 \cap \{z^*(1+\kappa_1) < (1-\kappa_0)\lambda\}. \tag{16}$$

(ii) If $\mathfrak{B}_0 = \{\boldsymbol{\beta} : \mathrm{sgn}(\boldsymbol{\beta}) = \mathrm{sgn}(\boldsymbol{\beta}_0)\}$*, and (14) and (15) hold with* $\mathfrak{B}_0^*$ *replaced by* $\mathfrak{B}_0$*, then* $\mathrm{sgn}(\hat{\boldsymbol{\beta}}) = \mathrm{sgn}(\boldsymbol{\beta}_0)$ *in the event*

$$\Omega_1 \cap \left\{ \sup_{\boldsymbol{\beta}\in\mathfrak{B}_0} \parallel \mathbf{A}_S^{-1} \parallel_\infty (|\hat{w}_S|_\infty \lambda + z^*) < \min_{j\in S} |\beta_{j0}| \right\}. \tag{17}$$

By Theorem 5 and the probabilistic upper bound for z^*, we have the following.

Corollary 1 *(i) If* $\mathfrak{B}_0^* = \{\boldsymbol{\beta} : \boldsymbol{\beta}_{S^c} = 0\}$*,* $S_{\boldsymbol{\beta}} = \{j : \beta_j \neq 0\}$*,* $\lambda = \{(1+\kappa_1)/(1-\kappa_0)\}K\sqrt{(2/n)\log(2p/\epsilon)}$ *with a small* $\epsilon > 0$ *(e.g.* $\epsilon = 0.01$*), and (14) and (15) hold, then in the event* Ω_0*,* $\{j : \hat{\beta}_j \neq 0\} \subseteq S$ *hold with at least probability* $1-\epsilon$*.*

(ii) If $\mathfrak{B}_0 = \{\boldsymbol{\beta} : \mathrm{sgn}(\boldsymbol{\beta}) = \mathrm{sgn}(\boldsymbol{\beta}_0)\}$*, (14) and (15) hold with* $\mathfrak{B}_0^*$ *replaced by* $\mathfrak{B}_0$*, and* $\min\left\{(1-\kappa_0)/(1+\kappa_1)\lambda,\ (\sup_{\boldsymbol{\beta}\in\mathfrak{B}_0} \parallel \mathbf{A}_S^{-1} \parallel_\infty)^{-1} \min_{j\in S} |\beta_{j0}| - |\hat{w}_S|_\infty \lambda\right\} = K\sqrt{(2/n)\log(2p/\epsilon)}$*, then* $\mathrm{sgn}(\hat{\boldsymbol{\beta}}) = \mathrm{sgn}(\boldsymbol{\beta}_0)$ *in the event* Ω_0 *hold with at least probability* $1-\epsilon$*.*

The proof of this corollary is similar to that of Theorem 2, so we omit the details. These conditions of the Corollary 1 can be regarded as an extension of the irrepresentable condition for Lasso in the linear regression model (Meinshausen and Bühlmann (2006); Zhao and Yu (2006)) to the current setting.

We now derive an upper bound for the dimension of $\hat{\boldsymbol{\beta}}$. Take

$$\kappa_+(m) = \sup_{|\mathcal{B}|=m} \{\Lambda_{\max}(\mathbf{W}_{\mathcal{B}}^{-2}\mathbf{A}_{\mathcal{B}}) : \mathcal{B} \cap S = \varnothing\} \tag{18}$$

as a restricted upper eigenvalue, where $\Lambda_{\max}(\cdot)$ denotes the largest eigenvalue, $\mathcal{B} \subseteq \{1, \cdots, p\}$, $\mathbf{A}_{\mathcal{B}}$ and $\mathbf{W}_{\mathcal{B}}$ are the restrictions of the Hessian of (3) and the weight $\mathbf{W} = \mathrm{diag}\{w\}$ to $\mathbb{R}^{\mathcal{B}}$.

Theorem 6 *If $\hat{\boldsymbol{\beta}}$ is the weighted Lasso estimator (5) and $\xi > |w_S|_\infty$, then in the event $\Omega_0 \cap \{z^* \leq (\xi - |w_S|_\infty)/(\xi+1)\lambda\}$, we have*

$$\#\{j : \hat{\beta}_j \neq 0, j \notin S\} < d_1 = \min\left\{m \geq 1 : \frac{m}{\kappa_+(m)} > \frac{\xi^2\lambda^2 d_0(1+|w_S|_\infty)^2}{(\lambda - z^*)^2(\xi+1)^2\kappa^2(\xi, S; \mathbf{A})}\right\}.$$

Corollary 2 *If $\hat{\boldsymbol{\beta}}$ is the weighted Lasso estimator (5), $\xi > |w_S|_\infty$, and $\lambda = \{(\xi+1)/(\xi - |w_S|_\infty)\}K\sqrt{(2/n)\log(2p/\epsilon)}$ with a small $\epsilon > 0$, then in the event Ω_0, for any $C_\kappa > 0$, we have*

$$\#\{j : \hat{\beta}_j \neq 0, j \notin S\} < \tilde{d}_1 = \min\left\{m \geq 1 : \frac{m}{\kappa_+(m)} > \frac{\xi^2 d_0}{C_\kappa^2}\right\}$$

holds with probability no less than $P\{\kappa(\xi, S; \mathbf{A}) \geq C_\kappa\} - \epsilon$.

A direct consequence of this corollary is that $\#\{j : \hat{\beta}_j \neq 0\} \leq d_1 + d_0$. In particular, under the condition $\kappa_+(m) < k_+^*$ for all m, we have

$$\#\{j : \hat{\beta}_j \neq 0\} \leq (1 + \kappa_+^*\xi^2/C_\kappa^2)d_0.$$

This is an upper bound for the number of nonzero components of the weighted Lasso in the high-dimensional AH model.

5 Concluding remarks

There exist several directions for research in the future. One reviewer suggests that it would be useful to consider tests for individual coefficients and error control such as false discovery rate control in the high-dimensional AH model (Zhong, Hu, and Li (2015)); some treatments of this topic with the weighted Lasso would be interesting, and have practical implications. The established results assume that the sequence of penalty parameters is fixed, which is not applicable to the case where the penalty parameters are selected based on data-driven procedures, such as cross validation. This problem deserves further study, but is beyond the scope of the current paper. It would be interesting to consider the more general form of the AH model: $d\Lambda_i(t) = Y_i(t)\{d\Lambda_0(t) + h(\boldsymbol{Z}_i(t))dt\}$, where $h : \mathbb{R}^p \to \mathbb{R}_+$ is a nonparametric function. A particular case of interest is when h is an additive function, $d\Lambda_i(t) = Y_i(t)\{d\Lambda_0(t) + \sum_{j=1}^p h_j(Z_j(t))dt\}$. The linear AH model (1) is the parametric case with $h(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$. We expect that our methods would be useful for studying the properties of the weighted Lasso in these models.

Acknowledgments

We would like to thank the Editor, an Associate Editor and two anonymous reviewers for their helpful comments and suggestions, which helped us improve the article substantially. The work of Haixiang Zhang is partially supported by the National Natural Science Foundation of China (Nos. 11301212 and 11401146) and China Postdoctoral Science Foundation (No. 2014M550861). The work of Liuquan Sun is partially supported by the National Natural Science Foundation of China Grants (Nos. 11231010 and 11690015) and Key Laboratory of RCSDS, CAS (No. 2008DP173182). The work of Yong Zhou is partially supported by the State Key Program of National Natural Science Foundation of China (71331006), the State Key Program in the Major Research Plan of National Natural Science Foundation of China (91546202), and Innovative Research Team of Shanghai University of Finance and Economics (IRTSHUFE13122402).

Appendix

Here we prove Lemmas 1 - 2, Theorems 1-6, and Corollary 2.

Proof of Lemma 1. Since $L(\boldsymbol{\beta})$ is a convex function, and $D(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) = \hat{\mathbf{e}}'\{\dot{L}(\boldsymbol{\beta}_0 + \hat{\mathbf{e}}) - \dot{L}(\boldsymbol{\beta}_0)\} \geq 0$, the first inequality holds. With $\hat{e}_j = \hat{\beta}_j$ for $j \in S^c$,

$$\begin{aligned}
& \hat{\mathbf{e}}'\{\dot{L}(\boldsymbol{\beta}_0 + \hat{\mathbf{e}}) - \dot{L}(\boldsymbol{\beta}_0)\} \\
= \; & \sum_{j\in S^c} \hat{e}_j \dot{L}(\boldsymbol{\beta}_0 + \hat{\mathbf{e}})_j + \sum_{j\in S} \hat{e}_j \dot{L}(\boldsymbol{\beta}_0 + \hat{\mathbf{e}})_j + \hat{\mathbf{e}}'(-\dot{L}(\boldsymbol{\beta}_0)) \\
\leq \; & \sum_{j\in S^c} \hat{\beta}_j\big(-\lambda\hat{w}_j \mathrm{sgn}(\hat{\beta}_j)\big) + \sum_{j\in S} |\hat{e}_j|\lambda\hat{w}_j + \hat{\mathbf{e}}'_{S^c}(-\dot{L}(\boldsymbol{\beta}_0)_{S^c}) + \hat{\mathbf{e}}'_S(-\dot{L}(\boldsymbol{\beta}_0)_S) \\
\leq \; & -\lambda|\mathbf{W}_{S^c}\hat{\mathbf{e}}_{S^c}|_1 + \lambda|\mathbf{W}_S\hat{\mathbf{e}}_S|_1 + (\mathbf{W}_{S^c}\hat{\mathbf{e}}_{S^c})'\big(-\mathbf{W}_{S^c}^{-1}\dot{L}(\beta_0)_{S^c}\big) + \hat{\mathbf{e}}'_S(-\dot{L}(\boldsymbol{\beta}_0)_S) \\
\leq \; & (z^* - \lambda)|\mathbf{W}_{S^c}\hat{\mathbf{e}}_{S^c}|_1 + (z^* + \lambda|w_S|_\infty)|\hat{\mathbf{e}}_S|_1.
\end{aligned}$$

The first inequality here requires $\dot{L}(\boldsymbol{\beta}_0 + \hat{\mathbf{e}})_j = -\lambda\hat{w}_j\mathrm{sgn}(\hat{\beta}_j)$ only in the set $S^c \cap \{j : \hat{\beta}_j \neq 0\}$, since $\hat{e}_j = \hat{\beta}_j - \beta_{0j} = 0$ when $j \in S^c$ and $\hat{\beta}_j = 0$. This completes the proof of Lemma 1. $\square$

Proof of Lemma 2. (i) By the Hölder inequality, $|\mathbf{b}|_q \leq |\mathbf{b}|_1^{2/q-1}|\mathbf{b}|_2^{2-2/q}$. It follows from $|\mathbf{b}|_1 \leq (1 + \min\{w_{S^c}\}^{-1}\xi)|\mathbf{b}_S|_1$ in the cone, and $|\mathbf{b}_S|_1 \leq d_0^{1/2}|\mathbf{b}|_2$, that

$$|\mathbf{b}_S|_1|\mathbf{b}|_q/d_0^{1/q} \leq (1 + \min\{w_{S^c}\}^{-1}\xi)^{2/q-1}|\mathbf{b}_S|_1^{2/q}|\mathbf{b}|_2^{2-2/q}/d_0^{1/q} \leq (1 + \min\{w_{S^c}\}^{-1}\xi)^{2/q-1}|\mathbf{b}|_2^2.$$

Then, since $|\mathbf{b}_S|_1 \leq d_0^{1/2}|\mathbf{b}|_2$, (i) holds.

(ii) From $|\mathbf{b}'\mathbf{A}\mathbf{b} - \mathbf{b}'\mathbf{B}\mathbf{b}| \leq |\mathbf{b}|_1^2 \max_{i,j}|A_{ij} - B_{ij}|$ and

$$|\mathbf{b}|_1 \leq (1 + \min\{w_{S^c}\}^{-1}\xi)|\mathbf{b}_S|_1 \leq (1 + \min\{w_{S^c}\}^{-1}\xi)d_0^{1/q}|\mathbf{b}|_q,$$

it is easy to obtain the desired result.

(iii) The conclusion immediately follows from (11). This completes the proof of Lemma 2. $\square$

Proof of Theorem 1. Let $\hat{\mathbf{e}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \neq 0$ and $\boldsymbol{b} = \hat{\mathbf{e}}/|\hat{\mathbf{e}}|_1$. Because of the convexity of $L(\boldsymbol{\beta})$,

$$x^{-1}D(\boldsymbol{\beta}_0 + x\mathbf{b}, \boldsymbol{\beta}_0) = \frac{\partial}{\partial x}\{L(\boldsymbol{\beta}_0 + x\mathbf{b}) - x\mathbf{b}'\dot{L}(\boldsymbol{\beta}_0)\}$$

is an increasing function of x. Thus, in the event $\Omega_0 \cap z^* \leq \lambda(\xi - |w_S|_\infty)/(\xi+1)$, by Lemma 1 we have

$$\boldsymbol{b}'\{\dot{L}(\boldsymbol{\beta}_0 + x\boldsymbol{b}) - \dot{L}(\boldsymbol{\beta}_0)\} + \frac{\lambda(1+|w_S|_\infty)}{\xi+1}|\mathbf{W}_{S^c}\boldsymbol{b}_{S^c}|_1 \leq \frac{\xi\lambda(1+|w_S|_\infty)}{\xi+1}|\boldsymbol{b}_S|_1, \tag{19}$$

where $x \in [0, |\hat{\mathbf{e}}|_1]$, and $\boldsymbol{b} \in \Theta(\xi, S)$ which is defined in (7). Then for all nonnegative x, it follows from $x\boldsymbol{b}'\{\dot{L}(\boldsymbol{\beta}_0 + x\boldsymbol{b}) - \dot{L}(\boldsymbol{\beta}_0)\} = x^2\boldsymbol{b}'\ddot{L}(\boldsymbol{\beta}_0)\boldsymbol{b}$, the definition of $\kappa(\xi, S; \mathbf{A})$, and (19) that

$$\begin{aligned}
x\kappa^2(\xi, S; \mathbf{A})|\boldsymbol{b}_S|_1^2/d_0 &\leq x\boldsymbol{b}'\ddot{L}(\boldsymbol{\beta}_0)\boldsymbol{b} \\
&\leq \frac{\xi\lambda(1+|w_S|_\infty)}{\xi+1}|\boldsymbol{b}_S|_1 - \frac{\lambda(1+|w_S|_\infty)}{\xi+1}|\mathbf{W}_{S^c}\boldsymbol{b}_{S^c}|_1 \\
&\leq \frac{\lambda(1+|w_S|_\infty)(\xi+\min\{w_{S^c}\})}{\xi+1}|\boldsymbol{b}_S|_1 - \frac{\lambda\min\{w_{S^c}\}(1+|w_S|_\infty)}{\xi+1} \\
&\leq \frac{\lambda(1+|w_S|_\infty)(\xi+\min\{w_{S^c}\})^2}{4\min\{w_{S^c}\}(\xi+1)}|\boldsymbol{b}_S|_1^2.
\end{aligned}$$

Therefore, for all x satisfying (19), we have

$$x \leq \frac{\lambda d_0(1+|w_S|_\infty)(\xi+\min\{w_{S^c}\})^2}{4\min\{w_{S^c}\}\kappa^2(\xi, S; \mathbf{A})(\xi+1)}. \tag{20}$$

Since L is convex, $\boldsymbol{b}'\{\dot{L}(\boldsymbol{\beta}_0 + x\boldsymbol{b}) - \dot{L}(\boldsymbol{\beta}_0)\}$ is an increasing function of x, the set of all nonnegative x satisfying(19) is a closed interval $[0, \tilde{x}]$ for some $\tilde{x}$. Thus, (20) yields

$$|\hat{\mathbf{e}}|_1 \leq |\tilde{x}| \leq \frac{\lambda d_0(1+|w_S|_\infty)(\xi+\min\{w_{S^c}\})^2}{4\min\{w_{S^c}\}\kappa^2(\xi, S; \mathbf{A})(\xi+1)},$$

which is the second part of (9). Furthermore, by Lemma 1 we have

$$\kappa^2(\xi, S; \mathbf{A})|\hat{\mathbf{e}}_S|_1^2/d_0 \leq \hat{\mathbf{e}}'\ddot{L}(\boldsymbol{\beta}_0)\hat{\mathbf{e}} = D(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) \leq \frac{\xi\lambda(1+|w_S|_\infty)|\hat{\mathbf{e}}_S|_1}{\xi+1}.$$

Thus, the first part of (9) holds.

Lastly, from the definition of $F_q(\xi, S; \mathbf{A})$ and Lemma 1, we can derive that

$$|\hat{\mathbf{e}}|_q \leq \frac{d_0^{1/q}\hat{\mathbf{e}}'\mathbf{A}\hat{\mathbf{e}}}{|\hat{e}_S|_1 F_q(\xi, S; \mathbf{A})} = \frac{d_0^{1/q} D(\boldsymbol{\beta}_0 + \hat{\mathbf{e}}, \boldsymbol{\beta}_0)}{|\hat{e}_S|_1 F_q(\xi, S; \mathbf{A})} \leq \frac{d_0^{1/q}(\lambda |w_S|_\infty + z^*)}{F_q(\xi, S; \mathbf{A})},$$

so (10) holds. This completes the proof of Theorem 1. □

Proof of Theorem 2. Let $x = \lambda(\xi - |w|_\infty)/\{K(\xi+1)\} = \sqrt{(2/n)\log(2p/\epsilon)}$ in the probability bound $P\{z^* > Kx\} \leq 2pe^{-nx^2/2}$, then it can be verified that the probability of the event $z^* > (\xi - |w|_\infty)/(\xi+1)\lambda$ is at most ϵ. Then it follows from Theorem 1 that the desired results hold. This completes the proof of Theorem 2. □

Proof of Theorem 3. By the definition of $\bar{\boldsymbol{A}}(t^*)$ and Lemma 2 (iii), we have

$$\Phi(\xi, S; \ddot{L}(\boldsymbol{\beta}_0)) \geq \Phi(\xi, S; \bar{\boldsymbol{A}}(t^*)). \tag{21}$$

From the definition of $\Sigma_n(t)$ and $\bar{\Sigma}_n(t)$, we have

$$\Sigma_n(t) = \bar{\Sigma}_n(t) + n^{-1}\sum_{i=1}^{n} Y_i(t)\{\bar{\boldsymbol{Z}}_n(t) - \boldsymbol{\mu}(t)\}^{\otimes 2}.$$

Thus,

$$\bar{\boldsymbol{A}}(t^*) = \int_0^{t^*} \Sigma_n(t)dt - \int_0^{t^*} n^{-1}\sum_{i=1}^{n} Y_i(t)\{\bar{\boldsymbol{Z}}_n(t) - \boldsymbol{\mu}(t)\}^{\otimes 2}dt. \tag{22}$$

Take $\bar{Y}_n(t) = n^{-1}\sum_{i=1}^n Y_i(t)$ and $\Gamma(t) = \bar{Y}_n(t)\{\bar{\boldsymbol{Z}}_n(t) - \boldsymbol{\mu}(t)\} = n^{-1}\sum_{i=1}^n Y_i(t)\{\boldsymbol{Z}_i(t) - \boldsymbol{\mu}(t)\}$. Since $Y_i(t)$ is a non-increasing function in t, we have

$$0 \leq \int_0^{t^*} \bar{Y}_n(t)\{\bar{\boldsymbol{Z}}_n(t) - \boldsymbol{\mu}(t)\}^{\otimes 2}dt \leq \frac{\int_0^{t^*}\Gamma^{\otimes 2}(t)dt}{\bar{Y}_n(t^*)}. \tag{23}$$

Because $\bar{Y}_n(t^*)$ is an average of i.i.d. random variables taking values 0 or 1 and $E\bar{Y}_n(t^*) = r_*$, by the Hoeffding (1963) inequality, we have

$$P\{\bar{Y}_n(t^*) < r_*/2\} \leq e^{-nr_*^2/2}.$$

Since $\Gamma(t)$ is an average of i.i.d. mean-zero random vectors, $(n^2 \int_0^{t^*} \Gamma^{\otimes 2}(t)dt)_{i,j}$ is a degenerate V-statistic for each (i,j), and the summands of these V-statistic are all bounded by K^2t^*, by Lemma 4.2 of Huang et al. (2013), we have

$$P\left\{\pm\left(\int_0^{t^*} \Gamma^{\otimes 2}(t)dt\right)_{i,j} > (K^2t^*)t^2\right\} \leq 2.221\exp\left(\frac{-nt^2/2}{1+t/3}\right).$$

By (22), (23), the two above probability bounds and Lemma 2 (ii), we can derive that

$$\Phi(\xi, S; \bar{\boldsymbol{A}}(t^*)) \geq \Phi(\xi, S; \int_0^{t^*} \Sigma_n(t)dt) - d_0(1+\min\{w_{S^c}\}^{-1}\xi)^2K^2t^*(2/r_*)t^2_{n,p,\epsilon} \tag{24}$$

with at least probability $1 - e^{-nr_*^2/2} - \epsilon$.

Moreover, since $\int_0^{t^*} \Sigma_n(t)dt$ is an average of i.i.d. matrices with mean $\boldsymbol{A}(t^*)$ and the summands of $(\int_0^{t^*} \Sigma_n(t)dt)_{i,j}$ are uniformly bounded by K^2t^*, thus by the Hoeffding (1963) inequality, we get

$$P\left\{\max_{i,j}\left|\left(\int_0^{t^*} \Sigma_n(t)dt - \boldsymbol{A}(t^*)\right)_{i,j}\right| \geq K^2t^*t\right\} \leq p(p+1)e^{-nt^2/2}.$$

Then, it follows from (21), (24), the above inequality with $t = F_n(p(p+1)/\epsilon)$, and Lemma 2 (ii) that

$$\begin{aligned}\Phi(\xi, S; \ddot{L}(\boldsymbol{\beta}_0)) &\geq \Phi(\xi, S; \int_0^{t^*} \Sigma_n(t)dt) - d_0(1+\min\{w_{S^c}\}^{-1}\xi)^2K^2t^*(2/r_*)t^2_{n,p,\epsilon} \\ &\geq \Phi(\xi, S; \boldsymbol{A}(t^*)) - d_0(1+\min\{w_{S^c}\}^{-1}\xi)^2K^2t^*\{F_n(p(p+1)/\epsilon) + (2/r_*)t^2_{n,p,\epsilon}\}\end{aligned}$$

with at least probability $1 - e^{-nr_*^2/2} - 2\epsilon$.

From Lemma 2 that

$$\Phi(\xi, S; \boldsymbol{A}(t^*) \geq \mathrm{RE}^2(\xi, S; \boldsymbol{A}(t^*)) \geq \Lambda_{\min}(\boldsymbol{A}(t^*)),$$

and the desired results follow. This completes the proof of Theorem 3. □

Proof of Theorem 4. Let $\hat{\mathbf{e}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$, $w_j = \hat{w}_j$. Since $|\hat{w}|_\infty \leq 1$, we have

$$\frac{|\hat{w}|_\infty \lambda + z^*}{\lambda - z^*} \leq \frac{\lambda + \frac{\lambda}{\phi}}{\lambda - \frac{\lambda}{\phi}} = \frac{\phi + 1}{\phi - 1} \leq \xi.$$

Thus, from the KKT condition (6) and the proof of Lemma 1, we can show that $\hat{\mathbf{e}} \in \Theta(\xi, S)$ and $D(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) \leq |\hat{\mathbf{e}}_S|_1(|\hat{w}_S|_1 + |\dot{L}(\boldsymbol{\beta}_0)_S|_1)$. By the definition of $F_1(\xi, S; \mathbf{A})$ in (8), we get that

$$d_0^{-1} F_1(\xi, S; \mathbf{A}) |\hat{\mathbf{e}}_S|_1 |\hat{\mathbf{e}}|_1 \leq D(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) \leq |\hat{\mathbf{e}}_S|_1(|\hat{w}_S|_1 + |\dot{L}(\boldsymbol{\beta}_0)_S|_1).$$

Since $|\hat{\mathbf{e}}_S|_1 = 0$ implies $\hat{\mathbf{e}} = 0$ for $\hat{\mathbf{e}} \in \Theta(\xi, S)$,

$$d_0^{-1} F_1(\xi, S; \mathbf{A}) |\hat{\mathbf{e}}|_1 \leq |\hat{w}_S|_1 + |\dot{L}(\boldsymbol{\beta}_0)_S|_1. \tag{25}$$

It follows from $\hat{w}_j \lambda = \dot{P}_\lambda(|\tilde{\beta}_j|) \leq \dot{P}_\lambda(|\beta_{j0}|) + \varpi \cdot |\tilde{\beta}_j - \beta_{j0}|$ that

$$|\hat{w}_S|_1 \lambda \leq |\dot{P}_\lambda(|\boldsymbol{\beta}_{0S}|)|_1 + \varpi |\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|_1. \tag{26}$$

From (25) and (26), $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|_1 \leq \frac{d_0}{F_1(\xi, S; \mathbf{A})} \left\{ |\dot{P}_\lambda(|\beta_{0S}|)|_1 + |\dot{L}(\boldsymbol{\beta}_0)_S|_1 + \varpi |\tilde{\beta} - \beta_0|_1 \right\}$. Moreover, $|\dot{L}(\boldsymbol{\beta}_0)_S|_1 \leq z^* \leq \phi/\lambda$ and $|S| = d_0$ lead to the desired results. This ends the proof of Theorem 4. □

Proof of Theorem 5. (i) Let $\tilde{\mathbf{a}} = \mathbf{a} - \mathbf{A}\boldsymbol{\beta}_0$ and λ be fixed. Take

$$\hat{\boldsymbol{\beta}}(\lambda, t) = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \boldsymbol{\beta}' \mathbf{A} \boldsymbol{\beta} - \boldsymbol{\beta}'(t\tilde{\mathbf{a}} + \mathbf{A}\boldsymbol{\beta}_0) + t\lambda \sum_{j=1}^{p} \hat{w}_j |\beta|_j : \boldsymbol{\beta}_{S^c} = 0 \right\}$$

as an artificial path for $0 \leq t \leq 1$. Then for each t, the KKT conditions for $\hat{\boldsymbol{\beta}}(\lambda, t)$ are:

$$g_S(\lambda, t) = t\lambda \hat{\mathbf{W}}_S \mu_S(\lambda, t), \;\; \mu_j(\lambda, t) \begin{cases} = \operatorname{sgn}(\hat{\beta}_j(\lambda, t)), & \text{if } \hat{\beta}_j(\lambda, t) \neq 0, \\ \in [-1, 1], & \text{if } \hat{\beta}_j(\lambda, t) = 0, \end{cases}$$

where $g(\lambda, t) = -\mathbf{A}\hat{\boldsymbol{\beta}}(\lambda, t) + \mathbf{A}\boldsymbol{\beta}_0 + t\tilde{\mathbf{a}}$.

Let $S_t = \{j : \hat{\beta}_j(\lambda, t) \neq 0\}$. By applying differentiation $D = (\partial/\partial t)$ to the KKT conditions, it follows that almost everywhere in t,

$$(Dg)_{S_t}(\lambda, t) = \tilde{\mathbf{a}}_{S_t} - \mathbf{A}_{S_t}\{(D\hat{\boldsymbol{\beta}})_{S_t}(\lambda, t)\} = \lambda \hat{\mathbf{W}}_{S_t}\mu_{S_t}(\lambda, t).$$

Then we have

$$(D\hat{\boldsymbol{\beta}})_{S_t}(\lambda, t) = \mathbf{A}_{S_t}^{-1}\{\tilde{\mathbf{a}}_{S_t} - \lambda \hat{\mathbf{W}}_{S_t}\mu_{S_t}(\lambda, t)\}. \tag{27}$$

An application of the chain rule leads to

$$(Dg)_{S^c}(\lambda, t) = \tilde{\mathbf{a}}_{S^c} - \mathbf{A}_{S^c S_t}\mathbf{A}_{S_t}^{-1}\{\tilde{\mathbf{a}}_{S_t} - \lambda \hat{\mathbf{W}}_{S_t}\mu_{S_t}(\lambda, t)\}.$$

As $g(\lambda, t)$ is almost differentiable and $\hat{\boldsymbol{\beta}}(\lambda, 0+) = \boldsymbol{\beta}_0$, we have $g(\lambda, 0+) = 0$ and $g_{S^c}(\lambda, 1-) = \int_0^1 [\tilde{\mathbf{a}}_{S^c} - \mathbf{A}_{S^c S_t}\mathbf{A}_{S_t}^{-1}\{\tilde{\mathbf{a}}_{S_t} - \lambda \hat{\mathbf{W}}_{S_t}\mu_{S_t}(\lambda, t)\}]dt$. Thus, by (14) and (15), $|\hat{\mathbf{W}}_{S^c}^{-1} g_{S^c}(\lambda, 1-)|_\infty \leq |\hat{\mathbf{W}}_{S^c}^{-1}\tilde{\mathbf{a}}_{S^c}|_\infty + \kappa_1 |\tilde{\mathbf{a}}_{S^c}|_\infty + \kappa_0 \lambda |\mu_{S_t}(\lambda, t)|_\infty$, which is smaller than λ in the event (16). Then $\hat{\boldsymbol{\beta}}(\lambda, 1-)$ is the unique solution of the KKT condition (6) for $\hat{\boldsymbol{\beta}}$. This ends the proof of part (i).

(ii) We note that (17) implies that $S = \{j : \beta_{j0} \neq 0\}$. Because $\hat{\boldsymbol{\beta}}(\lambda, 0+) = \boldsymbol{\beta}_0$, there exists $t_1 > 0$, $\mu_S(\lambda, t) = \text{sgn}(\boldsymbol{\beta}_{0S})$ for $0 < t < t_1$. By (27) and (17), for $0 < t < t_1$ and some $\epsilon > 0$, we have

$$|(D\hat{\boldsymbol{\beta}})_S(\lambda, t)|_\infty \leq \| \mathbf{A}_{S_t}^{-1} \|_\infty |\tilde{\mathbf{a}}_S - \lambda \text{sgn}(\boldsymbol{\beta}_{0S})\hat{\mathbf{W}}_S|_\infty < \min_{j \in S} |\beta_{0j}| - \epsilon.$$

Due to $\hat{\boldsymbol{\beta}}(\lambda, 0+)$, $|\hat{\boldsymbol{\beta}}_S(\lambda, t) - \boldsymbol{\beta}_{0S}|_\infty < \min_{j \in S} |\beta_{0j}| - \epsilon$, for all $0 < t < \min\{t_1, 1\}$. Furthermore, by the continuity of $\hat{\boldsymbol{\beta}}(\lambda, t)$ in t, we know that $\text{sgn}(\hat{\boldsymbol{\beta}}(\lambda, t)) = \text{sgn}(\beta_0)$ for $0 < t \leq 1$. Then, (14) and (15) are only needed for the smaller $\mathfrak{B}_0$ in the proof of (i). Thus, $\hat{\boldsymbol{\beta}}(\lambda, 1) = \hat{\boldsymbol{\beta}}$. This completes the proof of Theorem 5. □

Proof of Theorem 6. Let $\hat{\mathbf{e}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$, then it follows from (3) that $\mathbf{A}\hat{\mathbf{e}} = \dot{L}(\hat{\boldsymbol{\beta}}) - \dot{L}(\boldsymbol{\beta}_0)$. By the KKT conditions (6), we have

$$|(\mathbf{A}\hat{\mathbf{e}})_j| = |(\dot{L}(\hat{\boldsymbol{\beta}}) - \dot{L}(\boldsymbol{\beta}_0))_j| \geq \hat{w}_j\lambda - |\dot{L}(\boldsymbol{\beta}_0)_j| \geq w_j(\lambda - z^*) > 0, \quad j \notin S.$$

If $\mathcal{B} \subseteq \{j \notin S : \hat{\beta}_j \neq 0\}$ with $|\mathcal{B}| \leq d_1$, (18) implies that

$$\max_{|u|_2=1} |(\mathbf{W}^{-1}\mathbf{A}^{1/2}u)_{\mathcal{B}}|_2^2 = \Lambda_{\max}(\mathbf{W}_{\mathcal{B}}^{-2}\mathbf{A}_{\mathcal{B}}) \leq \kappa_+(d_1).$$

Thus,

$$(\lambda - z^*)^2|\mathcal{B}| \leq |(\mathbf{W}^{-1}\mathbf{A}\hat{\mathbf{e}})_{\mathcal{B}}|_2^2 \leq \kappa_+(d_1)\hat{\mathbf{e}}'A\hat{\mathbf{e}} = \kappa_+(d_1)D(\boldsymbol{\beta}_0 + \hat{\mathbf{e}}, \boldsymbol{\beta}_0).$$

From the predication bound in Theorem 2, we get

$$|\mathcal{B}| \leq \frac{\kappa_+(d_1)D(\boldsymbol{\beta}_0 + \hat{\mathbf{e}}, \boldsymbol{\beta}_0)}{(\lambda - z^*)^2} \leq \frac{\kappa_+(d_1)\xi^2\lambda^2 d_0(1 + |w_S|_\infty)^2}{(\lambda - z^*)^2(\xi + 1)^2\kappa^2(\xi, S; \mathbf{A})} < d_1. \tag{28}$$

All subsets $\mathcal{B} \subseteq \{j \notin S : \hat{\beta}_j \neq 0\}$ with $|\mathcal{B}| \leq d_1$ satisfy $|\mathcal{B}| < d_1$, so $\#\{j \notin S : \hat{\beta}_j \neq 0\} < d_1$. This completes the proof of Theorem 6. □

Proof of Corollary 2. For $\widetilde{\mathcal{B}} \subseteq \{j \notin S : \hat{\beta}_j \neq 0\}$ with $|\widetilde{\mathcal{B}}| \leq \tilde{d}_1$, since $\lambda - z^* \geq (|w_S|_\infty + 1)/(\xi + 1)\lambda$, similar to (28), we get

$$\begin{aligned} |\widetilde{\mathcal{B}}| &\leq \frac{\kappa_+(\tilde{d}_1)D(\boldsymbol{\beta}_0 + \hat{\mathbf{e}}, \boldsymbol{\beta}_0)}{(\lambda - z^*)^2} \leq \frac{\kappa_+(\tilde{d}_1)\xi^2\lambda^2 d_0(1 + |w_S|_\infty)^2}{(\lambda - z^*)^2(\xi + 1)^2\kappa^2(\xi, S; \mathbf{A})} \\ &\leq \frac{\kappa_+(\tilde{d}_1)\xi^2 d_0}{\kappa^2(\xi, S; \mathbf{A})} < \tilde{d}_1. \end{aligned} \tag{29}$$

Let $x = \lambda(\xi - |w|_\infty)/\{K(\xi + 1)\} = \sqrt{(2/n)\log(2p/\epsilon)}$. By the probability bound $P\{z^* > Kx\} \leq 2pe^{-nx^2/2}$, we see that the probability of the event $z^* > (\xi - |w|_\infty)/(\xi + 1)\lambda$ is at most ϵ. Thus, by replacing $\kappa(\xi, S; \mathbf{A})$ in (29) with C_κ , we have the desired result. This completes the proof of Corollary 2. □

References

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal* **19**, 357-367.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37**, 1705-1732.

Bradic, J., Fan, J. and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-Dimensionality. *Ann. Statist.* **39**, 3092-3120.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, New York.

Cox, D. (1972). Regression models and life-tables (with discussions). *J. Roy. Statist. Soc. Ser. B* **34**, 187-220.

Cox, D. and Oakes, D. (1984). *Analysis of Survival Data.* London: Chapman & Hall.

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 Coxs proportional hazards model and frailty model. *Ann. Statist.* **30**, 74-99.

Fan, J. and Lv, J. (2010). A Selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20**, 101-148.

Gaïffas, S. and Guilloux, A. (2012). High dimensional additive hazards models and the Lasso. *Electronic Journal of Statistics* **6**, 522-546.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13-30.

Huang, J., Sun, T., Ying, Z., Yu, Y. and Zhang, C.-H. (2013). Oracle inequalities for the Lasso in the Cox model. *Ann. Statist.* **41**, 1142-1165.

Huang, J. and Zhang, C.-H. (2012). Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *J. Machine Learning Research* **13**, 1839-1864.

Koltchinskii, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15**, 799-828.

Kong, S. and Nan, B. (2014). Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso. *Statist. Sinica* **24**, 25-42.

Kulich, M. and Lin, D. Y. (2000). Additive hazards regression with covariate measurement error. *J. Amer. Statist. Assoc.* **95**, 238-248.

Lemler, S. (2012). Oracle inequalities for the Lasso for the conditional hazard rate in a high-dimensional setting. **arXiv**:1206.5628.

Leng, C., and Ma, S. (2007). Path consistent model selection in additive risk model via Lasso. *Statist. Medicine* **26**, 3753-3770.

Lin, D. Y., and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81**, 61-71.

Lin, W. and Lv, J. (2013). High-dimensional sparse additive hazards regression. *J. Amer. Statist. Assoc.* **108**, 247-264.

Martinussen, T., and Scheike, T. (2002). Efficient estimation in additive current status data. *Biometrika* **89**, 649-658.

Martinussen, T., and Scheike, T. (2009). Covariate selection for the semiparametric additive risk model. *Scand. J. Statist.* **36**, 602-619.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34**, 1436-1462.

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

van de Geer, S. (2007). On non-asymptotic bounds for estimation in generalized linear models with highly correlated design. *Lecture Notes-Monograph Series* **55**, 121-134.

van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics* **3**, 1360-1392.

Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Machine Learning Research* **11**, 3519-3540.

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

Zhang, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *J. Machine Learning Research* **11**, 1087-1107.

Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.* **27**, 576-593.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Machine Learning Research* **7**, 2541-2563.

Zhong, P., Hu, T. and Li, J. (2015). Tests for coefficients in high-dimensional additive hazard models. *Scand. J. Statist.* **42**, 649-664.

Center for Applied Mathematics, Tianjin University, Tianjin, 300072, China

E-mail: haixiang.zhang@tju.edu.cn

Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China

E-mail: slq@amt.ac.cn

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, 200433, China

E-mail: yzhou@amss.ac.cn

Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, USA

E-mail: jian-huang@uiowa.edu