The Annals of Statistics
2011, Vol. 39, No. 6, 3092–3120
DOI: 10.1214/11-AOS911

REGULARIZATION FOR COX'S PROPORTIONAL HAZARDS MODEL WITH NP-DIMENSIONALITY[1]

BY JELENA BRADIC, JIANQING FAN AND JIANCHENG JIANG

University of California, San Diego, Princeton University and University of North Carolina at Charlotte

High throughput genetic sequencing arrays with thousands of measurements per sample and a great amount of related censored clinical data have increased demanding need for better measurement specific model selection. In this paper we establish strong oracle properties of nonconcave penalized methods for *nonpolynomial* (NP) dimensional data with censoring in the framework of Cox's proportional hazards model. A class of folded-concave penalties are employed and both LASSO and SCAD are discussed specifically. We unveil the question under which dimensionality and correlation restrictions can an oracle estimator be constructed and grasped. It is demonstrated that nonconcave penalties lead to significant reduction of the "irrepresentable condition" needed for LASSO model selection consistency. The large deviation result for martingales, bearing interests of its own, is developed for characterizing the strong oracle property. Moreover, the nonconcave regularized estimator, is shown to achieve asymptotically the information bound of the oracle estimator. A coordinate-wise algorithm is developed for finding the grid of solution paths for penalized hazard regression problems, and its performance is evaluated on simulated and gene association study examples.

1. Introduction. A central theme in high-dimensional data analysis is efficient discovery of sparsity patterns. For such data, where dimensionality possibly grows exponentially faster than the sample size, sparsity structures are imposed as means of recovering important signals. Under the linear regression model framework, various methods ranging from regularized to marginal regressions and graphical models have been effectively proposed for identification, reconstruction and estimation of the unknown sparse regression parameters.

With increasing understanding of sparse recovery in these novel high-dimensional spaces, more and more attention is paid to efficient discovery of sparsity patterns for ultra-high dimensional data and great progress has been made in the least squares setting. For example, Meinshausen and Bühlmann (2006), Zhao and

Received October 2010; revised March 2011.

[1]Supported in part by NSF Grants DMS-07-04337, DMS-07-14554 and DMS-09-06482 and NIH Grant R01-GM072611. Majority of the work was done when Jelena Bradic was a graduate student at Princeton University.

MSC2010 subject classifications. Primary 62N02, 60G44; secondary 62F12, 60F10.

Key words and phrases. Hazard rate, LASSO, SCAD, large deviation, oracle.

Yu (2006) and Zhang and Huang (2008) investigated model selection consistency of LASSO when the number of variables is of a greater order than the sample size and Candes and Tao (2007) introduced the Dantzig selector specifically to handle the NP-dimensional variable selection problem, and Bunea, Tsybakov and Wegkamp (2007), Bickel, Ritov and Tsybakov (2009), van de Geer and Bühlmann (2009), Koltchinskii (2009), Meinshausen and Yu (2009), Massart and Meynet (2010), among others, showed their asymptotic or finite sample oracle risk properties for fixed or random ill-posed designs. Various versions of the "restricted eigenvalue condition," "sparse Riesz condition" or "incoherence condition" that exclude high correlations among variables play a key role here. On the other hand, when the LASSO estimator does not satisfy some of these conditions, it often selects a model which is overly dense in its effort to relax the penalty on the relevant coefficients [Fan and Li (2001), Zhang (2010), Zhang and Huang (2008)]. Hence, nonconvex penalties [Fan and Li (2001)] are proposed where Zhang (2010) pioneered the work with NP-dimensionality and demonstrated its sign consistency for $p \gg n$ and its advantages over LASSO in the sense of attaining minimax convergence rates. Lv and Fan (2009) and Fan and Lv (2011) made important connections between finite sample and asymptotic oracle properties using folded-concave penalties for the penalized least squares estimator with NP-dimensionality. Although extensive work has been done for linear regression models, censored survival data have been left greatly unexplored for $p \gg n$.

Extending oracle results to censored data with NP-dimensionality presents a tremendous novel challenge, and, to the best of our knowledge, there is no previous work on this topic. The extensions to LASSO and SCAD algorithms for survival data were successfully proposed by Tibshirani (1997) and Fan and Li (2002), respectively, but both algorithms were theoretically tested only when $p \ll n$. In recent papers, Johnson (2009), Wang et al. (2009) and Du, Ma and Liang (2010) addressed the problem in accelerated failure time models, Cox's model and semiparametric relative risk models by combining the LASSO, group LASSO and adaptive LASSO penalties, but, likewise, they only discussed the case of $p \ll n$.

Motivated by the growing importance of gene selection problems, in this paper we go one step further and address the problem of existence of an oracle estimator and regularization estimator under an ultra-high dimensionality setting, where the full dimensionality might grow exponentially or nonpolynomially fast with the sample size, in order of $\log p = O(n^{\delta})$ for some $\delta > 0$, and the intrinsic dimensionality goes to infinity, in order of $s = O(n^{\alpha})$ for $\alpha \in (0, 1)$. We develop a strong oracle argument, which shares the spirit of Fan and Li (2002), but guarantees that the folded-concave penalized partial likelihood estimator is equal to the oracle one, with probability tending to 1. A similar strong oracle argument was developed by Kim, Choi and Oh (2008) and Bradic, Fan and Wang (2011) in the contexts of linear regression models. Extending such results to Cox's proportional hazards model is a new exceptional challenge due to its nature of censoring and NP-dimensionality.

1.1. *Model setup*. We consider multivariate data $\{(\mathbf{X}_i, T_i)\}_{i=1}^n$, which form an i.i.d. sample from the population $(\mathbf{X}, \mathbf{T})$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ is a column vector of covariates for the ith individual. For a variety of reasons not all survival times $(T_i)_{i=1}^n$ are fully observable. The independent right censoring scheme is considered where i.i.d. censoring times $(C_i)_{i=1}^n$ are conditionally independent of survival times given covariates $\{\mathbf{X}_i\}_{i=1}^n$. Hence, we work with i.i.d. sample $\{(\mathbf{X}_i, Z_i, \delta_i)\}_{i=1}^n$, where $Z_i = \min(T_i, C_i)$ and $\delta_i = \mathbf{1}\{T_i \le C_i\}$ are event times and censoring indicator, respectively.

The conditional hazard rate function of T given $\mathbf{X} = \mathbf{x}$ is denoted by $\lambda(t|\mathbf{x})$. Cox's proportional hazards model assumes that

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

where the baseline hazard rate $\lambda_0(t)$ is a nuisance function. Let $t_1 < \cdots < t_N$ denote the ordered failure times and (j) denote the label of the item failing at t_j. Denote by $\mathcal{R}_j = \{i \in \{1, \ldots, n\} : Z_i \ge t_j\}$ the risk set at time t_j and by $\Lambda_0(t) = \int_0^t \lambda_0(u)\,du$ the cumulative baseline hazard function.

Following the approach of nonparametric maximum likelihood estimation, the "least informative" nonparametric modeling of $\Lambda_0(t)$ assumes that $\Lambda_0(t)$ has a jump of size θ_j at the failure time t_j: $\Lambda_0(t;\theta) = \sum_{j=1}^N \theta_j \mathbf{1}\{t_j \le t\}$. If we use the Breslow MLE $\hat{\theta}_j^{-1} = \sum_{i\in\mathcal{R}_j}\exp(\boldsymbol{\beta}^T\mathbf{X}_i)$, then the penalized Cox's log partial likelihood becomes [Fan and Li (2002)]

$$Q_n(\boldsymbol{\beta}) - n\sum_{k=1}^p p_{\lambda_n}(|\beta_k|), \tag{2}$$

where $Q_n(\boldsymbol{\beta}) = \sum_{j=1}^N \{\boldsymbol{\beta}^T\mathbf{X}_{(j)} - \log(\sum_{i\in\mathcal{R}_j}\exp(\boldsymbol{\beta}^T\mathbf{X}_i))\}$, $p_{\lambda_n}(\cdot)$ is a penalty function, and λ_n is a nonnegative regularization parameter. Note that the covariate vector $\mathbf{X}$ may be time dependent and incorporated in the standard way in model (1) through

$$\lambda(t|\mathbf{X}(t)) = \lim_{\Delta t\to 0} P\{t \le T \le t + \Delta t | T \ge t, \mathbf{X}(t)\}/\Delta t = \lambda_0(t)\exp(\boldsymbol{\beta}^T\mathbf{X}(t)).$$

Note that from hereon we will be working with the time-dependent left continuous covariate vector $\mathbf{X}(t)$.

1.2. *Counting process representation*. Let $N_i(t) = 1\{Z_i \le t, \delta_i = 1\}$, $\bar{N}(t) = \sum_{i=1}^n N_i(t)$ and $Y_i(t) = 1\{Z_i \ge t\}$. Note that the process $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_n(t))^T$ is assumed to be left continuous with right-hand limits and satisfies $P(\mathbf{Y}(t) = 1, 0 \le t \le \tau) > 0$. Using the counting process notation, one can rewrite the log partial likelihood $Q_n(\boldsymbol{\beta})$ for model (2) as

$$Q_n(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^\tau \{\boldsymbol{\beta}^T\mathbf{X}_i(t) - \log(S_n^{(0)}(\boldsymbol{\beta}, t))\}\,dN_i(t),$$

where and hereafter τ is the study ending time, and

$$S_n^{(\ell)}(\boldsymbol{\beta}, t) = n^{-1} \sum_{i=1}^{n} Y_i(t)\{\mathbf{X}_i(t)\}^{\otimes \ell} \exp(\boldsymbol{\beta}^T \mathbf{X}_i(t)), \qquad \ell = 0, 1, 2,$$

with $\otimes$ denoting the outer product. Thus, the penalized log partial likelihood becomes

$$\mathcal{C}(\boldsymbol{\beta}, \tau) \equiv \sum_{i=1}^{n} \int_0^{\tau} \{\boldsymbol{\beta}^T \mathbf{X}_i(t) - \log(S_n^{(0)}(\boldsymbol{\beta}, t))\} \, dN_i(t) - n \sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|). \tag{3}$$

Define the sparse estimator $\hat{\boldsymbol{\beta}}$ as the maximizer of $\mathcal{C}(\boldsymbol{\beta}, \tau)$ over $\boldsymbol{\beta} \in \Omega_p$, where Ω_p is the parameter space which is a compact subset of R^p and contains the true value of $\boldsymbol{\beta}$. Note that $N_i(t)$ is a counting process with intensity process $\lambda_i(t, \boldsymbol{\beta}) = \lambda_0(t) Y_i(t) \exp\{\boldsymbol{\beta}^T X_i(t)\}$, which does not admit jumps at the same time as $N_j(t)$ for $j \neq i$. Denote by $\boldsymbol{\beta}^*$ the true value of $\boldsymbol{\beta}$ and $\Lambda_i(t) = \int_0^t \lambda_i(u, \boldsymbol{\beta}^*) \, du$. Then $M_i(t) = N_i(t) - \Lambda_i(t)$ is an orthogonal local square integrable martingale with respect to filtration

$$\mathcal{F}_{t,i} = \sigma\{N_i(u), \mathbf{X}_i(u^+), Y_i(u^+), 0 \leq u \leq t\},$$

that is, $\langle M_i(t), M_j(t)\rangle = 0$ for $i \neq j$. Let $\mathcal{F}_t = \bigcup_{i=1}^n \mathcal{F}_{t,i}$ be the smallest σ-algebra containing $\mathcal{F}_{t,i}$. Then $\bar{M}(t) = \sum_{i=1}^n M_i(t)$ is a martingale with respect to $\mathcal{F}_t$.

1.3. *Choice of the penalty function.* There are many commonly used penalties in the literature, for example, the L_2 penalty used in ridge regression; the nonnegative garrote as a shrinkage estimation [Yuan and Lin (2007)]; the L_0 penalty for the best subset selection; the L_1 penalty LASSO [Tibshirani (1996)] as a convex relaxation of the L_0 penalty; the SCAD penalty [Fan and Li (2001)], defined via its derivative $p'_\lambda(t) = \lambda\{I(t \leq \lambda) + \frac{(a\lambda - t)+}{(a-1)\lambda} I(t > \lambda)\}, t \geq 0$, for some $a > 2$, as a folded-concave relaxation of L_0 penalty; the MCP [Zhang (2010)] penalty. Recently, a class of penalties bridging L_0 and L_1 penalties was introduced in Lv and Fan (2009). All of these penalties are folded concave penalties, as noted in Fan and Li (2001) and Fan and Lv (2011). As a collection of nonconvex relaxations of the L_0 penalty, they serve as a tool of allowing bigger correlations among covariates (see Condition 8) and hence relax significantly the standard "incoherence condition" and control the tail bias of the resulting penalized estimator (see Theorem 4.2). For any penalty function $p_{\lambda_n}(\cdot)$, let $\rho(t; \lambda_n) = \lambda_n^{-1} p_{\lambda_n}(t)$ and write $\rho(t; \lambda_n)$ as $\rho(t)$ for simplicity when there is no confusion. According to Fan and Lv (2011), the folded concave penalties are defined through the following Condition 1.

CONDITION 1. $\rho(t; \lambda_n)$ is increasing and concave in $t \in [0, \infty)$ and has a continuous derivative $\rho'(t; \lambda_n)$ with $\rho'(0+; \lambda_n) > 0$. In addition, $\rho'(t; \lambda_n)$ is increasing in $\lambda_n \in (0, \infty)$ and $\rho'(0+; \lambda_n)$ is independent of λ_n.

Note that most commonly used nonconvex penalties, including SCAD and MCP ($a \geq 1$), satisfy Condition 1. We will employ the folded concave penalties to increase flexibility of our method. LASSO penalty as a convex function falls at the boundary of penalties in Condition 1, and our results will be applicable for LASSO penalty as well.

The rest of the paper is organized as follows. In Section 2 we deal with identification problem of the penalized estimator $\hat{\boldsymbol{\beta}}$, which is key to the proof of oracle results. A compelling large deviation result is derived for divergence of a martingale from its compensator in Section 3. In Section 4 we work out the new strong oracle property and its implications for LASSO and SCAD and asymptotic properties of the proposed estimator. In Section 5 we propose an iterative coordinate ascent algorithm (ICA) and examine a thorough simulation example; see Section 5.1. The gene association study is done in Section 5.2 where the non-Hodgkin's lymphoma dataset of Dave et al. (2004) is analyzed. Technical lemmas and proofs are collected in the Appendix and in the supplementary material [Bradic, Fan and Jiang (2011)].

2. Identification. This section gives the appropriate necessary and sufficient conditions on the existence of estimator $\hat{\boldsymbol{\beta}}$. We can always assume that the true parameter $\boldsymbol{\beta}^*$ can be arranged as $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*T}, \mathbf{0}^T)^T$, with $\boldsymbol{\beta}_1^* \in \Omega_s$ being a vector of nonvanishing elements of $\boldsymbol{\beta}^*$, where $\Omega_s = \Omega_p \cap R^s$.

Throughout the paper the following notation on a vector/matrix norm is used. Denote by $\lambda_{\min}(\mathbf{B})$ and $\lambda_{\max}(\mathbf{B})$ the minimum and maximum eigenvalues of a symmetric matrix $\mathbf{B}$, respectively. We also use $\lambda(\mathbf{B})$ to denote any eigenvalue of $\mathbf{B}$. Let $\|\cdot\|_q$ be the L_q norm of a vector or matrix. Then for a $s \times s$ matrix $\mathbf{A}$, $\|\mathbf{A}\|_\infty = \max\{\sum_{k=1}^s |(\mathbf{A})_{jk}| : 1 \leq j \leq s\}$ and $\|\mathbf{A}\|_2 = \{\lambda_{\max}(\mathbf{A}^T\mathbf{A})\}^{1/2}$. We also let $\sigma(\mathbf{A})$ be the set consisting of all of eigenvalues of $\mathbf{A}$, and let $r_\sigma(\mathbf{A}) = \max\{|\lambda| : \lambda \in \sigma(\mathbf{A})\}$ be the spectral radius of $\mathbf{A}$. If $\mathbf{A}$ is symmetric, then $r_\sigma(\mathbf{A}) = \|\mathbf{A}\|_2$.

Since no concavity is assumed for the penalized log partial likelihood (3), it is difficult, in general, to study the global maximizer of the penalized likelihood. One useful index controlling the convexity of the whole optimization problem (3) is the following "local concavity" of the penalty function $\rho(\cdot)$ at $\mathbf{v} = (v_1, \dots, v_s)^T \in R^s$ with $\|\mathbf{v}\|_0 = s$,

$$\kappa(\rho, \mathbf{v}) = \lim_{\varepsilon \to 0+} \max_{1 \leq j \leq s} \sup_{t_1 < t_2 \in (|v_j| - \varepsilon, |v_j| + \varepsilon)} -\frac{\rho'(t_2) - \rho'(t_1)}{t_2 - t_1}, \tag{4}$$

which is defined in Lv and Fan (2009) and shares similar spirit to the "maximum concavity" of ρ in Zhang (2010). Since ρ is concave on $(0, \infty)$, $\kappa(\rho, \mathbf{v}) \geq 0$. For LASSO penalty $\kappa(\rho, \mathbf{v}) = 0$, whereas for the SCAD penalty

$$\kappa(\rho, \mathbf{v}) = \begin{cases} (a-1)^{-1}\lambda^{-1}, & \text{if there exists a } v_j \text{ such that } \lambda \leq |v_j| \leq a\lambda; \\ 0, & \text{otherwise.} \end{cases}$$

Let $\boldsymbol{\beta}_1$ be a subvector of $\boldsymbol{\beta}$ formed by all nonzero components and $s = \dim(\boldsymbol{\beta}_1)$. Denote by $\mathbf{S}_i$ the subvector of $\mathbf{X}_i$ with same indexes as $\hat{\boldsymbol{\beta}}_1$ in $\hat{\boldsymbol{\beta}}$ and by $\mathbf{Q}_i$ the complement to $\mathbf{S}_i$. For $\mathbf{v} = (v_1, \dots, v_s)^T \in R^s$, let $\boldsymbol{\rho}'(\mathbf{v}) = (\rho'(v_1), \dots, \rho'(v_s))^T$ and $\operatorname{sgn}(\mathbf{v}) = (\operatorname{sgn}(v_1), \dots, \operatorname{sgn}(v_s))^T$. Partition $S_n^{(1)}(\boldsymbol{\beta}, t) = [S_{n1}^{(1)}(\boldsymbol{\beta}, t), S_{n2}^{(1)}(\boldsymbol{\beta}, t)]$ and

$$S_n^{(2)}(\boldsymbol{\beta}, t) = \begin{pmatrix} S_{n11}^{(2)}(\boldsymbol{\beta}, t) & S_{n12}^{(2)}(\boldsymbol{\beta}, t) \\ S_{n21}^{(2)}(\boldsymbol{\beta}, t) & S_{n22}^{(2)}(\boldsymbol{\beta}, t) \end{pmatrix}$$

according to the partition of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, so that $S_{n1}^{(1)}(\boldsymbol{\beta}, t)$ is a $\dim(\boldsymbol{\beta}_1) \times 1$ vector and $S_{n11}^{(2)}(\boldsymbol{\beta}, t)$ is a $\dim(\boldsymbol{\beta}_1) \times \dim(\boldsymbol{\beta}_1)$ matrix. Let $\mathbf{E}_n^{(1)}(\boldsymbol{\beta}, t) = S_{n1}^{(1)}(\boldsymbol{\beta}, t)/S_n^{(0)}(\boldsymbol{\beta}, t)$, $\mathbf{E}_n^{(2)}(\boldsymbol{\beta}, t) = S_{n2}^{(1)}(\boldsymbol{\beta}, t)/S_n^{(0)}(\boldsymbol{\beta}, t)$, $\mathbf{E}_n(\boldsymbol{\beta}, t) = S_n^{(1)}(\boldsymbol{\beta}, t)/S_n^{(0)}(\boldsymbol{\beta}, t)$, $\mathbf{V}(\boldsymbol{\beta}, t) = S_{n11}^{(2)}(\boldsymbol{\beta}, t)/S_n^{(0)}(\boldsymbol{\beta}, t) - (S_{n1}^{(1)}(\boldsymbol{\beta}, t)/S_n^{(0)}(\boldsymbol{\beta}, t))^{\otimes 2}$ and $\mathbf{V}(\boldsymbol{\beta}_1, t) = \mathbf{V}((\boldsymbol{\beta}_1, \mathbf{0}), t)$.

The following theorem provides a sufficient condition on the strict local maximizer of $\mathcal{C}(\boldsymbol{\beta}, \tau)$. Proof is relegated to the supplementary material [Bradic, Fan and Jiang (2011)].

THEOREM 2.1. *If Condition 1 is satisfied, then an estimate $\hat{\boldsymbol{\beta}} \in R^p$ is a strict local maximizer of the nonconcave penalized log partial likelihood* (3) *if*

$$\sum_{i=1}^{n} \int_0^{\tau} (\mathbf{S}_i(t) - \mathbf{E}_n^{(1)}(\hat{\boldsymbol{\beta}}, t))\, dN_i(t) - n\lambda_n \boldsymbol{\rho}'(|\hat{\boldsymbol{\beta}}_1|) \circ \operatorname{sgn}(\hat{\boldsymbol{\beta}}_1) = \mathbf{0}, \tag{5}$$

$$\|\mathbf{z}(\hat{\boldsymbol{\beta}})\|_\infty \equiv \left\| \sum_{i=1}^{n} \int_0^{\tau} (\mathbf{Q}_i(t) - \mathbf{E}_n^{(2)}(\hat{\boldsymbol{\beta}}, t))\, dN_i(t) \right\|_\infty < n\lambda_n \rho'(0+), \tag{6}$$

$$\lambda_{\min}\left\{ \int_0^{\tau} \mathbf{V}(\hat{\boldsymbol{\beta}}, t)\, d\bar{N}(t) \right\} > n\lambda_n \kappa(\rho, \hat{\boldsymbol{\beta}}_1), \tag{7}$$

where $\circ$ is the Hadamard product. Conversely, if $\hat{\boldsymbol{\beta}}$ is a local maximizer of $\mathcal{C}(\boldsymbol{\beta}, \tau)$, then it must satisfy (5)–(7) *with strict inequalities replaced by nonstrict inequalities.*

When the LASSO penalty is used, $\kappa(\rho, \mathbf{v}) = 0$, hence the condition of nonsingularity for the matrix in (7) is automatically satisfied with a nonstrict inequality. For the SCAD penalty, $\kappa(\rho, \hat{\boldsymbol{\beta}}_1) = 0$; that is, (7) holds with nonstrict inequality, unless there are some j such that $\lambda_n < |\hat{\beta}_j| < a\lambda_n$, which usually has a small chance. In the latter case, $\kappa(\rho, \hat{\boldsymbol{\beta}}_1) = (a-1)^{-1}\lambda_n^{-1}$, and the condition in (7) reduces to

$$\lambda_{\min}\left\{ \int_0^{\tau} n^{-1}\mathbf{V}(\hat{\boldsymbol{\beta}}, t)\, d\bar{N}(t) \right\} > 1/(a-1).$$

This will hold if a large a is used, due to nonsingularity of the matrix.

It is natural to ask if the penalized nonconcave Cox's log partial likelihood has a global maximizer. Since $p \gg n$, it is hard to show the global optimality of a local maximizer. Theorem 4.1 in Section 4 suggests a condition for $\hat{\boldsymbol{\beta}}$ to be unique and global. Once the unique maximizer is available, it will be equal to the oracle one with probability tending to one exponentially fast, when the effective dimensionality s is bounded by $O(n^{\alpha})$ for $\alpha < 1$ (see Theorem 4.3). In this way Theorems 2.1, 4.1 and 4.3 address uniqueness of the solution and provide methods for finding the global maximizer among potentially many. Methodological innovations among others consist of using equations (5) and (6) as an identification tool to surpass the absence of analytical form of an estimator $\hat{\boldsymbol{\beta}}$.

3. A large deviation result. In view of (5) and (6), to study a nonconcave penalized Cox's partial likelihood estimator $\hat{\boldsymbol{\beta}}$, we need to analyze the deviation of p-dimensional counting process $\int_0^t \{\mathbf{X}_i(u) - \mathbf{E}_n(\boldsymbol{\beta}^*, u)\}\, dN_i(u)$ from its compensator $\mathbf{A}_i = \int_0^t \{\mathbf{X}_i(u) - \mathbf{E}_n(\boldsymbol{\beta}^*, u)\}\, d\Lambda_i(u)$. In other words, we need to simultaneously analyze the deviation of marginal score vectors from their compensators. Some conditions are needed for this purpose.

CONDITION 2. There exists a compact neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}^*$ that satisfies each of the following conditions:

(i) There exist scalar, vector and matrix functions $s^{(j)}$ defined on $\mathcal{B} \times [0, \tau]$ such that, in probability as $n \to \infty$ for $j = 0, 1, 2$, $\sup_{t \in [0,\tau], \beta_1 \in \mathcal{B}_1} \|S_n^{(j)}(\boldsymbol{\beta}_1, t) - s^{(j)}(\boldsymbol{\beta}_1, t)\|_2 \to 0$, for $\mathcal{B}_1 \in \mathcal{R}^s, \mathcal{B}_1 \subset \mathcal{B}$.

(ii) The functions $s^{(j)}$ are bounded and $s^{(0)}$ is bounded away from 0 on $\mathcal{B} \times [0, \tau]$; for $j = 0, 1, 2$, the family of functions $s^{(j)}(\cdot, t)$, $0 \le t \le \tau$, is an equicontinuous family at $\boldsymbol{\beta}^*$.

(iii) Let $\mathbf{e}(\boldsymbol{\beta}, t) = s^{(1)}(\boldsymbol{\beta}, t)/s^{(0)}(\boldsymbol{\beta}, t)$, $\mathbf{v}(\boldsymbol{\beta}, t) = s^{(2)}(\boldsymbol{\beta}, t)/s^{(0)}(\boldsymbol{\beta}, t) - \{\mathbf{e}(\boldsymbol{\beta}, t)\}^{\otimes 2}$ and $\boldsymbol{\Sigma}_\beta(t) = \int_0^t \mathbf{v}(\boldsymbol{\beta}, u) s^{(0)}(\boldsymbol{\beta}^*, u)\, d\Lambda_0(u)$. Define $\mathbf{v}(\boldsymbol{\beta}_1, t)$ in the same way as for $\mathbf{V}(\boldsymbol{\beta}_1, t)$ but with $S_n^{(\ell)}$ replaced by $s^{(\ell)}$. Let

$$\boldsymbol{\Sigma}_{\beta_1}(t) = \int_0^t \mathbf{v}(\boldsymbol{\beta}_1, u) s^{(0)}(\boldsymbol{\beta}_1^*, u)\, d\Lambda_0(u)$$

and $\boldsymbol{\Sigma}_{\beta_1} = \boldsymbol{\Sigma}_{\beta_1}(\tau)$. Assume that the $s \times s$ matrix $\boldsymbol{\Sigma}_{\beta_1^*}$ is positive definite for all n and $\Lambda_0(\tau) < \infty$.

(iv) Let $c_n = \sup_{t \in [0,\tau]} \|\mathbf{E}_n(\boldsymbol{\beta}^*, t) - \mathbf{e}(\boldsymbol{\beta}^*, t)\|_\infty$ and $d_n = \sup_{t \in [0,\tau]} |S_n^{(0)}(\boldsymbol{\beta}^*, t) - s^{(0)}(\boldsymbol{\beta}^*, t)|$. The random sequences c_n and d_n are bounded almost surely.

The above conditions, (i)–(iii), agree with the conditions in Section 8.2 of Fleming and Harrington (1991) for fixed p and in Cai et al. (2005) for diverging p. Condition (iii) is restricted to hold on the s instead of usually assumed p-dimensional subspace. This is a counterpart of the similar conditions imposed on the covariance matrix $\mathbf{X}$ in the linear regression models [see, e.g., Bunea, Tsybakov and Wegkamp (2007), van de Geer and Bühlmann (2009), Zhang (2010)]. Nonsingularity of the matrix $\boldsymbol{\Sigma}_{\beta_1^*}$ in (iii) could have been relaxed toward restricted eigenvalue properties like those for linear models [Bickel, Ritov and Tsybakov (2009), Koltchinskii (2009)] but for easier composure we impose a bit stronger condition. Condition (iv) is used to ensure that the score vector of the log partial likelihood, which is a martingale, has bounded jumps and quadratic variation. By following the discussion on pages 305 and 306 of Fleming and Harrington (1991), this condition is not stringent for i.i.d. samples.

The following Condition 3 is coming as a consequence of martingale representation of the score function for the Cox model, and it is valuable in analyzing large deviations of counting processes.

CONDITION 3. Let $\varepsilon_{ij} = \int_0^\tau (X_{ij}(t) - e_j(\boldsymbol{\beta}^*, t))\, dM_i(t)$, where $e_j(\boldsymbol{\beta}^*, t)$ is the jth component of $\mathbf{e}(\boldsymbol{\beta}^*, t)$. Suppose the Cramér condition holds for ε_{ij}, that is,

$$E|\varepsilon_{ij}|^m \le m! M^{m-2} \sigma_j^2 / 2$$

for all j, where M is a positive constant, $m \ge 2$ and $\sigma_j^2 = \operatorname{var}(\varepsilon_{ij}) < \infty$.

In linear regression models, the large deviation is established upon the Cramér condition for the covariates. Condition 3 takes a similar role here and can be regarded as an extension to the classical Cramér condition. Moreover, it is trivially fulfilled if the covariates are bounded. In that sense it represents a relaxation of typical assumption of bounded covariates. Since ε_{ij} is a mean zero martingale, it can be shown that $\sigma_j^2 = E(\varepsilon_{ij}^2) = (\boldsymbol{\Sigma}_{\beta^*})_{jj}$ is the jth diagonal entry of $\boldsymbol{\Sigma}_{\beta^*}$. Define $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)^T$ to be the score vector of the log partial likelihood function $Q_n(\boldsymbol{\beta})$,

$$\boldsymbol{\xi} = \sum_{i=1}^n \int_0^\tau \big(\mathbf{X}_i(t) - \mathbf{E}_n(\boldsymbol{\beta}^*, t)\big)\, dN_i(t).$$

Since $M_i(t) = N_i(t) - \Lambda_i(t)$ is a martingale with compensator $\Lambda_i(t) = \int_0^t \lambda_i(u, \boldsymbol{\beta}^*)\, du$, we can rewrite ξ_j as $\sum_{i=1}^n \int_0^\tau \{X_{ij}(t) - E_{nj}(\boldsymbol{\beta}^*, t)\}(dM_i(t) + d\Lambda_i(t))$, where $E_{nj}(\boldsymbol{\beta}^*, t)$ is the jth component of $\mathbf{E}_n(\boldsymbol{\beta}^*, t)$. Note that $\sum_{i=1}^n \int_0^\tau \{X_{ij}(t) - E_{nj}(\boldsymbol{\beta}^*, t)\}\, d\Lambda_i(t) = 0$, leading to the representation of the form

$$\xi_j = \sum_{i=1}^n \int_0^\tau \{X_{ij}(t) - E_{nj}(\boldsymbol{\beta}^*, t)\}\, dM_i(t).$$

The following theorem characterizes the uniform deviation of the score vector $\boldsymbol{\xi}$ and is critical in obtaining strong oracle property; see Theorem 4.3 in Section 4. To the best of our knowledge there is no similar result in the literature.

THEOREM 3.1. *Under Conditions 2 and 3, for any positive sequence* $\{u_n\}$ *bounded away from zero there exist positive constants* c_0 *and* c_1 *such that*

$$P\big(|\xi_j| > \sqrt{n}u_n\big) \le c_0 \exp(-c_1 u_n) \tag{8}$$

uniformly over j, *if* $v_n = \max_j \sigma_j^2/u_n$ *is bounded.*

PROOF. Denote by $E_{nj}(\boldsymbol{\beta}^*, t)$ and $e_j(\boldsymbol{\beta}^*, t)$ the jth components of $\mathbf{E}_n(\boldsymbol{\beta}^*, t)$ and $\mathbf{e}(\boldsymbol{\beta}^*, t)$, respectively. Then ξ_j can be written as

$$\begin{aligned}\xi_j &= \sum_{i=1}^n \int_0^\tau \big(X_{ij}(t) - e_j(\boldsymbol{\beta}^*, t)\big)\,dM_i(t)\\ &\quad - \sum_{i=1}^n \int_0^\tau \big(E_{nj}(\boldsymbol{\beta}^*, t) - e_j(\boldsymbol{\beta}^*, t)\big)\,dM_i(t)\\ &\equiv \xi_{j1}(\tau) - \xi_{j2}(\tau).\end{aligned}$$

To establish the exponential inequality about ξ_j, in the following we will establish the exponential inequalities about $\xi_{j1}(\tau)$ and $\xi_{j2}(\tau)$.

Note that $\xi_{j1}(\tau) = \sum_{i=1}^n \varepsilon_{ij}$, where $\{\varepsilon_{ij}\}_{i=1}^n$ is a sequence of i.i.d. random variables with mean zero and satisfying Condition 3. It follows from the Bernstein exponential inequality that

$$P(|\xi_{j1}| > a) \le 2\exp\{-a^2/2(n\sigma_j^2 + Ma)\}. \tag{9}$$

Note that $\bar{M}(t)$ is a martingale with respect to $\mathcal{F}_t$; it follows that $\xi_{j2}(t)$ is also a martingale with respect to $\mathcal{F}_t$. Let $\bar{N}(t) = \sum_{i=1}^n N_i(t)$. Then $\Delta\bar{N}(t) = \sum_{i=1}^n \Delta N_i(t)$, where and thereafter $\Delta N_i(t) = N_i(t) - N_i(t^-)$ denotes the jump of $N_i(\cdot)$ at time t. Since no two counting processes N_i jump at the same time, we have $|\Delta\bar{N}(t)| \le 1$. Let $\bar{\Lambda}(t) = \sum_{i=1}^n \Lambda_i(t)$. By continuity of the compensator $\Lambda_i(t) = \int_0^t \lambda_i(u, \boldsymbol{\beta}^*)\,du$, $|\Delta\bar{\Lambda}(t)| = 0$. Since $\bar{M}(t) = \bar{N}(t) - \bar{\Lambda}(t)$, $|\Delta\bar{M}(t)| = |\Delta\bar{N}(t)| \le 1$. Note that $\mathbf{Y}(t)$ and $\mathbf{X}(t)$ are left continuous in t. It is easy to see that

$$\begin{aligned}|\Delta(n^{-1/2}\xi_{j2}(t))| &= n^{-1/2}|E_{nj}(\boldsymbol{\beta}^*, t) - e_j(\boldsymbol{\beta}^*, t)|\\ &\le n^{-1/2} \sup_{t\in[0,\tau]} \|\mathbf{E}_n(\boldsymbol{\beta}^*, t) - \mathbf{e}(\boldsymbol{\beta}^*, t)\|_\infty\\ &\equiv n^{-1/2} c_n,\end{aligned}$$

which is bounded almost surely by Condition 3(vi). Note that the predictable quadratic variation of $n^{-1/2}\xi_{j2}(t)$, denoted by $\langle n^{-1/2}\xi_{j2}(t)\rangle$, is bilinear and satisfies that

$$\begin{aligned}\langle n^{-1/2}\xi_{j2}(t)\rangle &= n^{-1}\int_0^t \big(E_{nj}(\boldsymbol{\beta}^*,u)-e_j(\boldsymbol{\beta}^*,u)\big)^2\,d\langle\bar{M}(u)\rangle\\ &= \int_0^t \{E_{nj}(\boldsymbol{\beta}^*,u)-e_j(\boldsymbol{\beta}^*,u)\}^2 S_n^{(0)}(\boldsymbol{\beta}^*,u)\,d\Lambda_0(u)\\ &\le \int_0^t \|\mathbf{E}_n(\boldsymbol{\beta}^*,u)-\mathbf{e}(\boldsymbol{\beta}^*,u)\|_\infty^2 S_n^{(0)}(\boldsymbol{\beta}^*,u)\,d\Lambda_0(u)\equiv b_n^2(t).\end{aligned}$$

Obviously, $b_n^2(t)\le b_n^2(\tau)\le c_n^2\int_0^\tau S_n^{(0)}(\boldsymbol{\beta}^*,t)\,d\Lambda_0(t)$. Note that

$$\int_0^\tau S_n^{(0)}(\boldsymbol{\beta}^*,t)\,d\Lambda_0(t)\le\int_0^\tau s^{(0)}(\boldsymbol{\beta}^*,t)\,d\Lambda_0(t)+d_n\Lambda_0(\tau).$$

By Condition 2(ii), (iii) and (vi), there exist constants $0\le K<\infty$ and $0<b<\infty$, independent of j, such that $|\Delta(n^{-1/2}\xi_{j2}(t))|\le K$ and $\langle n^{-1/2}\xi_{j2}(t)\rangle\le b^2$. It follows from the exponential inequality for martingales with bounded jumps [see Lemma 2.1 of van de Geer (1995)] that, for $u_n>0$,

$$P\{|\xi_{j2}(\tau)|>\sqrt{n}u_n\}=P\{|n^{-1/2}\xi_{j2}(\tau)|>u_n\}\le 2\exp\left\{-\frac{u_n^2}{2(Ku_n+b^2)}\right\}.$$

Therefore, by Condition 3(iv), there exists a constant $c>0$ such that

$$P\{|\xi_{j2}(\tau)|>\sqrt{n}u_n\}\le 2\exp\{-cu_n\}\tag{10}$$

uniformly over j. Note that

$$P\{|\xi_j(\tau)|>\sqrt{n}u_n\}\le P\{|\xi_{j1}(\tau)|>0.5\sqrt{n}u_n\}+P\{|\xi_{j2}(\tau)|>0.5\sqrt{n}u_n\}.$$

It follows from (9) and (10) that $P\{|\xi_j(\tau)|>\sqrt{n}u_n\}$ is bounded by

$$2\exp\left\{-\frac{u_n}{4(2\sigma_j^2u_n^{-1}+Mn^{-1/2})}\right\}+2\exp(-0.5cu_n).\tag{11}$$

Then there exist positive constants c_0 and c_1 such that $P\{|\xi_j(\tau)|>\sqrt{n}u_n\}<c_0\exp(-c_1u_n)$ uniformly over j, if $\max_j\sigma_j^2=O(u_n)$. □

Theorem 3.1 represents a uniform, nonasymptotic exponential inequality for martingales. Compared with other exponential inequalities [de la Peña (1999), Juditsky and Nemirovski (2011), van de Geer (1995)], it is uniform over all components j. Moreover, its independence of dimensionality p proves to be invaluable for NP variable selection.

4. Strong oracle property. In this section we will prove a strong oracle property result, that is, that $\hat{\boldsymbol{\beta}}$ is an oracle estimator with overwhelming probability, and not that it behaves like an oracle estimator [Fan and Li (2002)]. We assume that the effective and full dimensionality satisfy $s = O(n^\alpha)$ and $\log p = O(n^\delta)$, for some $\alpha \in (0, 1)$ and $\delta > 0$, respectively. This notion of strong oracle property requires a definition of biased oracle estimator as it was defined in Bradic, Fan and Wang (2011) for the linear regression problem.

Let us define the biased oracle estimator $\hat{\boldsymbol{\beta}}^{\mathbf{o}} = (\hat{\boldsymbol{\beta}}_1^{\mathbf{o}T}, \mathbf{0}^T)^T$ where $\hat{\boldsymbol{\beta}}_1^{\mathbf{o}}$ is a solution to the s dimensional sub-problem

$$\arg\max_{\beta_1 \in \Omega_s} \sum_{i=1}^n \int_0^\tau [\boldsymbol{\beta}_1^T \mathbf{S}_i(t) - \log(S_n^{(0)}((\boldsymbol{\beta}_1, \mathbf{0}), t))] \, dN_i(t) - n\lambda_n \sum_{j=1}^s \rho(|\beta_j|; \lambda_n).$$

That is, $\hat{\boldsymbol{\beta}}_1^{\mathbf{o}} = \arg\max\{\mathcal{C}(\boldsymbol{\beta}_1, \tau) : \boldsymbol{\beta}_1 \in \Omega_s\}$ with $\mathcal{C}(\boldsymbol{\beta}_1, \tau) = \mathcal{C}((\boldsymbol{\beta}_1, \mathbf{0}), \tau)$. The estimator $\hat{\boldsymbol{\beta}}^{\mathbf{o}}$ is called the biased oracle estimator, since the oracle knows the true submodel $\mathcal{M}_* = \{j : \beta_j^* \neq 0\}$, but still applies a penalized method to estimate the nonvanishing coefficients.

THEOREM 4.1 (Global optimality). *Suppose that* $\min_{\beta_1 \in \Omega_s} \lambda_{\min}\{\int_0^\tau \mathbf{V}(\boldsymbol{\beta}_1, t) \, d\bar{N}(t)\} > n\lambda_n \kappa(\rho, \boldsymbol{\beta}_1)$ *holds almost surely. Then* $\hat{\boldsymbol{\beta}}_1^{\mathbf{o}}$ *is a unique global maximizer of the penalized log-likelihood* $\mathcal{C}(\boldsymbol{\beta}_1, \tau)$ *in* Ω_s.

The above theorem could be relaxed to a minimum over the level sets of Cox's partial likelihood in a similar manner to Proposition 1 of Fan and Lv (2011). Its proof is left for the supplementary material [Bradic, Fan and Jiang (2011)]. For LASSO penalty, $\hat{\boldsymbol{\beta}}_1^{\mathbf{o}}$ is unique and is the global maximizer, since $\mathcal{C}(\boldsymbol{\beta}_1, \tau)$ is strictly concave. In general, global maximizers are available for SCAD and MCP penalties, if one uses a large parameter a. In this setting, the biased oracle estimator is unique as a solution to strictly concave optimization problem. Note that it still depends on the penalty function. The biased oracle estimator, by its definition, satisfies only equation (5) in Theorem 2.1. Since the vanishing component does not need any penalty, the smaller the penalty the less the bias. In this sense, the biased oracle estimator with the SCAD penalty has a better performance than the biased oracle estimator with the LASSO penalty. The former is asymptotically unbiased [the second term in (5) is zero], while the latter is not (see Theorems 4.4 and 4.5).

In order to establish asymptotic properties of $\hat{\boldsymbol{\beta}}_1$, we need to govern the conditioning number of the $s \times s$ information matrix $\boldsymbol{\Sigma}_{\beta_1^*}$ through its eigenvalues. This is done in the following condition.

CONDITION 4. $r_\sigma(\boldsymbol{\Sigma}_{\beta_1^*}) = O(1)$ and $r_\sigma(\boldsymbol{\Sigma}_{\beta_1^*}^{-1}) = O(1)$.

Concerning Condition 2, positive definiteness of $\mathbf{\Sigma}_{\beta_1^*}$ is not enough and further bound on its spectrum is needed. Condition 4 is in the same spirit as the partial Riesz condition and is weaker than Condition A3 of Cai et al. (2005), where Condition 4 holds for $\mathbf{\Sigma}_{\beta*}$. In respect to Theorem 3.1, Condition A3 of Cai et al. (2005), ensures that $\max_j \sigma_j^2$ is bounded, therefore satisfying $\max_j \sigma_j^2 = O(u_n)$ for any positive sequence u_n bounded away from zero.

The following lemma controls the difference between the empirical information matrix with

$$\mathcal{I}_{\beta_1} = \int_0^T \mathbf{V}(\boldsymbol{\beta}_1, t) S_n^{(0)} \lambda_0(t)\, dt$$

and its population counterpart Σ_{β_1}, and plays a crucial part in the theoretical developments of this section.

LEMMA 4.1. *Assume that Conditions 2 and 4 hold. Then* $\sup_{\beta_1 \in \mathcal{B}} \|\mathcal{I}_{\beta_1}\|_2 = O_p(1)$, $\|\mathcal{I}_{\beta_1^*}^{-1}\|_2 = O_p(1)$ *and* $\sup_{\beta_1 \in \mathcal{B}} \|\mathcal{I}_{\beta_1} - \mathbf{\Sigma}_{\beta_1}\|_2 = o_p(1)$.

PROOF. We prove the statement in the following three steps:

(i) For any $s \times 1$ vector function $\mathbf{a}(t)$ on $[0, \tau]$, we have

$$\left\| \int_0^\tau \mathbf{a}(t)\lambda_0(t)\, dt \right\|_2^2 \le \Lambda_0(\tau) \int_0^\tau \|\mathbf{a}(t)\|_2^2 \lambda_0(t)\, dt.$$

In fact, by definition, $\| \int_0^\tau \mathbf{a}(t)\lambda_0(t)\, dt\|_2^2 = \sum_{i=1}^s (\int_0^\tau a_i(t)\lambda_0(t)\, dt)^2$, where $a_i(t)$ is the ith component function of $\mathbf{a}(t)$. Using the Hölder inequality, we obtain that

$$\begin{aligned}\left\| \int_0^\tau \mathbf{a}(t)\lambda_0(t)\, dt \right\|_2^2 &\le \sum_{i=1}^s \Lambda_0(\tau) \int_0^\tau a_i^2(t)\lambda_0(t)\, dt \\ &= \Lambda_0(\tau) \int_0^\tau \|\mathbf{a}(t)\|_2^2 \lambda_0(t)\, dt.\end{aligned}$$

(ii) For any matrix function $\mathbf{A}(t)$ on $[0, \tau]$, we have

$$\left\| \int_0^\tau \mathbf{A}(t)\lambda_0(t)\, dt \right\|_2^2 \le \Lambda_0(\tau) \int_0^\tau \|\mathbf{A}(t)\|_2^2 \lambda_0(t)\, dt.$$

In fact,

$$\begin{aligned}\left\| \int_0^\tau \mathbf{A}(t)\lambda_0(t)\, dt \right\|_2^2 &= \sup_{\|\mathbf{u}\|_2 = 1} \left\| \left(\int_0^\tau \mathbf{A}(t)\lambda_0(t)\, dt \right) \mathbf{u} \right\|_2^2 \\ &= \sup_{\|\mathbf{u}\|_2 = 1} \left\| \int_0^\tau \mathbf{a}_u(t)\lambda_0(t)\, dt \right\|_2^2,\end{aligned}$$

where $\mathbf{a}_u(t) = \mathbf{A}(t)\mathbf{u}$. Then

$$\begin{aligned}\int_0^\tau \|\mathbf{A}(t)\|_2^2 \lambda_0(t)\,dt &= \int_0^\tau \sup_{\|\mathbf{u}\|=1} \mathbf{u}^T \mathbf{A}(t)^{\otimes 2}\mathbf{u}\lambda_0(t)\,dt \\ &= \int_0^\tau \sup_{\|\mathbf{u}\|=1} \|\mathbf{a}_u(t)\|_2^2 \lambda_0(t)\,dt \\ &\geq \sup_{\|\mathbf{u}\|=1} \int_0^\tau \|\mathbf{a}_u(t)\|_2^2 \lambda_0(t)\,dt.\end{aligned}$$

Therefore, by (i), the result holds.

(iii) By definition, we have

$$\begin{aligned}\mathcal{I}_{\beta_1} - \mathbf{\Sigma}_{\beta_1} &= \int_0^\tau \{\mathbf{V}(\boldsymbol{\beta}_1, t) - \mathbf{v}(\boldsymbol{\beta}_1, t)\} s^{(0)}(\boldsymbol{\beta}_1^*, t)\lambda_0(t)\,dt \\ &\quad + \int_0^\tau \mathbf{V}(\boldsymbol{\beta}_1, t)\{S_n^{(0)}(\boldsymbol{\beta}_1^*, t) - s^{(0)}(\boldsymbol{\beta}_1^*, t)\}\lambda_0(t)\,dt \\ &\equiv \mathbf{A}_{n1}(\boldsymbol{\beta}_1) + \mathbf{A}_{n2}(\boldsymbol{\beta}_1).\end{aligned}$$

Using (ii), we obtain that

$$\|\mathbf{A}_{n1}(\boldsymbol{\beta}_1)\|_2^2 \leq \Lambda_0(\tau) \int_0^\tau \|\mathbf{V}(\boldsymbol{\beta}_1, t) - \mathbf{v}(\boldsymbol{\beta}_1, t)\|_2^2 \big(s^{(0)}(\boldsymbol{\beta}_1^*, t)\big)^2 \lambda_0(t)\,dt.$$

Then, by Condition 2, $\sup_{\beta_1 \in \mathcal{B}} \|\mathbf{A}_{n1}(\boldsymbol{\beta}_1)\|_2^2 = o_p(1)$. Similarly,

$$\sup_{\beta_1 \in \mathcal{B}} \|\mathbf{A}_{n2}(\boldsymbol{\beta}_1)\|_2^2 = o_p(1).$$

Therefore,

$$\sup_{\beta_1 \in \mathcal{B}} \|\mathcal{I}_{\beta_1} - \mathbf{\Sigma}_{\beta_1}\|_2 \leq \sup_{\beta_1 \in \mathcal{B}} \|\mathbf{A}_{n1}(\boldsymbol{\beta}_1)\|_2 + \sup_{\beta_1 \in \mathcal{B}} \|\mathbf{A}_{n2}(\boldsymbol{\beta}_1)\|_2 = o_p(1). \tag{12}$$

By Condition 2(ii), we have

$$\sup_{\beta_1 \in \mathcal{B}} \|\mathbf{\Sigma}_{\beta_1}\|_2 \leq \int_0^\tau \sup_{\beta_1 \in \mathcal{B}, t \in [0,\tau]} \|\mathbf{v}(\boldsymbol{\beta}_1, u)\|_2 s^{(0)}(\boldsymbol{\beta}_1^*, u)\,d\Lambda_0(u) = O_p(1).$$

This combining with (12) leads to

$$\sup_{\beta_1 \in \mathcal{B}} \|\mathcal{I}_{\beta_1}\|_2 \leq \sup_{\beta_1 \in \mathcal{B}} \|\mathbf{\Sigma}_{\beta_1}\|_2 + \sup_{\beta_1 \in \mathcal{B}} \|\mathcal{I}_{\beta_1} - \mathbf{\Sigma}_{\beta_1}\|_2 = O_p(1).$$

Decompose $\mathcal{I}_{\beta_1^*}^{-1}$ as

$$\mathcal{I}_{\beta_1^*}^{-1} = \mathbf{\Sigma}_{\beta_1^*}^{-1/2}\{I + \mathbf{\Sigma}_{\beta_1^*}^{-1/2}(\mathcal{I}_{\beta_1^*} - \mathbf{\Sigma}_{\beta_1^*})\mathbf{\Sigma}_{\beta_1^*}^{-1/2}\}^{-1}\mathbf{\Sigma}_{\beta_1^*}^{-1/2}$$

and let $\mathcal{A} = I + \boldsymbol{\Sigma}_{\beta_1^*}^{-1/2}(\mathcal{I}_{\beta_1^*} - \boldsymbol{\Sigma}_{\beta_1^*})\boldsymbol{\Sigma}_{\beta_1^*}^{-1/2}$. Then $\mathcal{I}_{\beta_1^*}^{-1} = \boldsymbol{\Sigma}_{\beta_1^*}^{-1/2}\mathcal{A}^{-1}\boldsymbol{\Sigma}_{\beta_1^*}^{-1/2}$. Using the Bauer–Fike inequality [Bhatia (1997)], we obtain that

$$|\lambda(\mathcal{A}) - 1| \le \|\boldsymbol{\Sigma}_{\beta_1^*}^{-1/2}(\mathcal{I}_{\beta_1^*} - \boldsymbol{\Sigma}_{\beta_1^*})\boldsymbol{\Sigma}_{\beta_1^*}^{-1/2}\|_2 \le \|\boldsymbol{\Sigma}_{\beta_1^*}^{-1/2}\|_2\|\mathcal{I}_{\beta_1^*} - \boldsymbol{\Sigma}_{\beta_1^*}\|_2\|\boldsymbol{\Sigma}_{\beta_1^*}^{-1/2}\|_2.$$

Then by (12) and Condition 4, $|\lambda(\mathcal{A}) - 1| = o_p(1)$. Hence, $\lambda(\mathcal{A}^{-1}) = 1 + o_p(1)$. Since $\mathcal{A}$ is symmetrical, $\|\mathcal{A}^{-1}\|_2 = O_p(1)$. This together with Condition 4 yield that $\|\mathcal{I}_{\beta_1^*}^{-1}\|_2 \le \|\boldsymbol{\Sigma}_{\beta_1^*}^{-1/2}\|_2\|\mathcal{A}^{-1}\|_2\|\boldsymbol{\Sigma}_{\beta_1^*}^{-1/2}\|_2 = O_p(1)$. □

The following tail condition is needed as a technicality in establishing estimation loss results on the oracle estimator $\hat{\boldsymbol{\beta}}^{\mathbf{o}}$.

CONDITION 5. $E\{\sup_{0\le t\le\tau} Y(t)\|\mathbf{S}(t)\|_2^2 \exp(\boldsymbol{\beta}_1^{*T}\mathbf{S}(t))\} = O(s)$.

For a fixed effective dimensionality s, Condition 5 is implied by the following condition from Andersen and Gill (1982):

$$E\Big\{\sup_{0\le t\le\tau,\beta_1\in\mathcal{B}} Y(t)\|\mathbf{S}(t)\|_2^2 \exp(\boldsymbol{\beta}_1^T\mathbf{S}(t))\Big\} < \infty. \tag{13}$$

However, we deal with diverging s, the above condition (13) is obviously too tight to be satisfied. For example, when all variables in $\mathcal{M}_*$ are bounded, we have $\|\mathbf{S}(t)\|_2^2 = O(s)$. In general, if each $S_k(t)$ in $\mathbf{S}(t)$ satisfies (13), then Condition 5 holds. Now we are ready to state the result on the existence of the biased oracle estimator.

THEOREM 4.2 (Estimation loss). *Under Conditions* 1, 2 *and* 4, 5, *with probability tending to one*, *there exists an oracle estimator* $\hat{\boldsymbol{\beta}}^{\mathbf{o}}$ *such that*

$$\|\hat{\boldsymbol{\beta}}^{\mathbf{o}} - \boldsymbol{\beta}^*\|_2 = O_P\{\sqrt{s}(n^{-1/2} + \lambda_n\rho'(\beta_n^*))\},$$

where $\beta_n^* = \min\{|\beta_j^*|, j \in \mathcal{M}_*\}$ *is the minimum signal strength*.

PROOF. Since $\hat{\boldsymbol{\beta}}_2^{\mathbf{o}} = \boldsymbol{\beta}_2^* = \mathbf{0}$, we only need to consider the subvector in the first s components, that is, we can restrict our attention to the s-dimensional subspace $\{\boldsymbol{\beta}_1 \in \mathbb{R}^s : \boldsymbol{\beta}_{\mathcal{M}_*^c} = 0\}$. It suffices to show that, for any $\varepsilon > 0$, there exists a large constant B and $\gamma_n = B\{\sqrt{s}(n^{-1/2} + \lambda_n\rho'(\beta_n^*)\}$ such that

$$P\Big\{\sup_{\|\mathbf{u}\|_2=1} \mathcal{C}(\boldsymbol{\beta}_1^* + \gamma_n\mathbf{u}, \mathbf{0}) < \mathcal{C}(\boldsymbol{\beta}_1^*, \mathbf{0})\Big\} \ge 1 - \varepsilon,$$

when n is big enough, where for short $\mathcal{C}(\boldsymbol{\beta})$ denotes $\mathcal{C}(\boldsymbol{\beta}, \tau)$, and in particular $\mathcal{C}(\boldsymbol{\beta}_1, \mathbf{0})$ represents $\mathcal{C}((\boldsymbol{\beta}_1, \mathbf{0}), \tau)$. This indicates that, with probability tending to one, there exists a local maximizer such that $\|\hat{\boldsymbol{\beta}}^{\mathbf{o}} - \boldsymbol{\beta}^*\|_2 = O_p\{\sqrt{s}(n^{-1/2} + \lambda_n\rho'(\beta_n^*))\}$.

Let $\mathbf{E}_n^{(1)}(\boldsymbol{\beta}_1, t) = \mathbf{E}_n^{(1)}((\boldsymbol{\beta}_1, \mathbf{0}), t)$, $\mathcal{P}_n(\boldsymbol{\beta}_1) = n\lambda_n \sum_{j=1}^{s} \rho(|\beta_j|; \lambda_n)$ and

$$U_n(\boldsymbol{\beta}_1) = \partial\mathcal{L}(\boldsymbol{\beta}_1) = \sum_{i=1}^{n} \int_0^\tau \{\mathbf{S}_i(t) - \mathbf{E}_n^{(1)}(\boldsymbol{\beta}_1, t)\}\, dN_i(t).$$

By the Taylor expansion at $\gamma_n = 0$,

$$\begin{aligned} &\mathcal{C}(\boldsymbol{\beta}_1^* + \gamma_n \mathbf{u}, 0) - \mathcal{C}(\boldsymbol{\beta}_1^*, 0) \\ &\quad = \mathbf{u}^T U_n(\boldsymbol{\beta}_1^*)\gamma_n + 0.5\gamma_n^2 \mathbf{u}^T\, \partial U_n(\boldsymbol{\beta}_1^*)\mathbf{u} + r_n(\boldsymbol{\beta}_1) \\ &\qquad - \mathcal{P}_n(\boldsymbol{\beta}_1^* + \gamma_n\mathbf{u}, 0) + \mathcal{P}_n(\boldsymbol{\beta}_1^*), \end{aligned} \tag{14}$$

where the remainder term $r_n(\boldsymbol{\beta}_1)$ is equal to

$$\frac{1}{6}\sum_{j,k}(\beta_{1j} - \beta_{1j}^*)(\beta_{1k} - \beta_{1k}^*)(\beta_{1\ell} - \beta_{1\ell}^*)\frac{\partial^2 U_{n\ell}(\boldsymbol{\beta}_1)}{\partial\beta_{1j}\,\partial\beta_{1k}}$$

with $U_{n\ell}$ being the ℓth component of U_n and $\boldsymbol{\beta}_1$ lying between $\boldsymbol{\beta}_1^* + \gamma_n\mathbf{u}$ and $\boldsymbol{\beta}_1^*$. By Lemma 2.2 in the supplementary material [Bradic, Fan and Jiang (2011)] we have $\|U_n(\boldsymbol{\beta}_1^*)\|_2 = O_p(\sqrt{ns})$. It follows that

$$|\mathbf{u}^T U_n(\boldsymbol{\beta}_1^*)\gamma_n| = O_p(\sqrt{ns}\gamma_n). \tag{15}$$

By simple decomposition, we have $\partial U_n(\boldsymbol{\beta}_1^*) = -n(\mathcal{I}_{\beta_1} + \mathcal{W}_{\beta_1})$, where $\mathcal{I}_{\beta_1}$ was defined in Lemma 4.2 and $\mathcal{W}_{\beta_1} = n^{-1}\int_0^\tau \mathbf{V}(\boldsymbol{\beta}_1, t)\, d\bar{M}(t)$. Hence,

$$\begin{aligned} \gamma_n^2 \mathbf{u}^T\, \partial U_n(\boldsymbol{\beta}_1^*)\mathbf{u} &= -n\gamma_n^2\{\mathbf{u}^T(-n^{-1}\,\partial U_n(\boldsymbol{\beta}_1^*))\mathbf{u}\} \\ &= -n\gamma_n^2\{\mathbf{u}^T\boldsymbol{\Sigma}_{\beta_1^*}\mathbf{u} + \mathbf{u}^T[(\mathcal{I}_{\beta_1^*} - \boldsymbol{\Sigma}_{\beta_1^*}) + \mathcal{W}_{\beta_1^*}]\mathbf{u}\}. \end{aligned}$$

By Lemma 2.3 in the supplementary material [Bradic, Fan and Jiang (2011)] and Lemma 4.1,

$$\|(\mathcal{I}_{\beta_1^*} - \boldsymbol{\Sigma}_{\beta_1^*}) + \mathcal{W}_{\beta_1^*}\|_2 \le \|\mathcal{I}_{\beta_1^*} - \boldsymbol{\Sigma}_{\beta_1^*}\|_2 + \|\mathcal{W}_{\beta_1^*}\|_2 = o_p(1).$$

Therefore, by Condition 4, there exists a constant $c > 0$ such that

$$\gamma_n^2\mathbf{u}^T\,\partial U_n(\boldsymbol{\beta}_1^*)\mathbf{u} \le -cn\gamma_n^2(1 + o_p(1)). \tag{16}$$

Since $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \le \gamma_n$ and the average of i.i.d. terms, $n^{-1}\frac{\partial^2 U_{n\ell}(\boldsymbol{\beta}_1)}{\partial\beta_{1j}\,\partial\beta_{1k}}$, is of order $O_p(1)$, we have $r_n(\boldsymbol{\beta}_1) = O_p(n\gamma_n^3)$. By concavity of ρ and decreasing property of ρ' from Condition 1,

$$\begin{aligned} |\mathcal{P}_n(\boldsymbol{\beta}_1^* + \gamma_n\mathbf{u}, 0) - \mathcal{P}_n(\boldsymbol{\beta}_1^*)| &= n\lambda_n\sum_{j=1}^{s}\big|\rho(|\beta_j^* + \gamma_n u_j|; \lambda_n) - \rho(|\beta_j^*|; \lambda_n)\big| \\ &\le n\lambda_n\gamma_n\|\boldsymbol{\rho}_0'(\beta_n^*)\|\|\mathbf{u}\|_2(1 + o_p(1)), \end{aligned}$$

where β_n^* is the minimal signal length and $\boldsymbol{\rho}_0'(\cdot)$ is the subvector of $\boldsymbol{\rho}(\cdot)$, consisting of its first s elements. Then

$$|\mathcal{P}_n(\boldsymbol{\beta}_1^* + \gamma_n \mathbf{u}, 0) - \mathcal{P}_n(\boldsymbol{\beta}_1^*)| = O_p\big(n\lambda_n\sqrt{s}\gamma_n\rho'(\beta_n^*)\big). \tag{17}$$

Combining (14)–(17) leads to

$$\mathcal{C}(\boldsymbol{\beta}_1^* + \gamma_n \mathbf{u}, 0) - \mathcal{C}(\boldsymbol{\beta}_1^*, 0) < n\gamma_n\big\{O_p\big(\sqrt{s/n} + \sqrt{s}\lambda_n\rho'(\beta_n^*)\big) - c\gamma_n\big(1 + o_p(1)\big)\big\},$$

where with probability tending to one, the RHS is smaller then zero when $\gamma_n = B(\sqrt{s/n} + s\lambda_n\rho'(\beta_n^*))$ for a sufficiently large B. □

A simple corollary of this theorem is that the L_1, L_∞ estimation losses of the oracle estimator are bounded by $s(n^{-1/2} + \lambda_n\rho'(\beta_n^*))\}$ and by $\sqrt{s}(n^{-1/2} + \lambda_n\rho'(\beta_n^*))$, respectively. Hence, L_1 loss can have a chance to be close to zero only if the sparsity parameter $\alpha < 1/2$, whereas L_∞ loss will converge to zero with no restrictions on α.

To make the bias in the penalized estimation negligible, $\rho'(\beta_n^*)$ needs to converge to zero at a specific rate controlled by the next condition.

CONDITION 6. The regularization parameter λ_n satisfies that $\sqrt{s}\lambda_n\rho'(\beta_n^*; \lambda_n) \to 0$ and $\lambda_n \gg n^{-0.5+(0.5\alpha+\alpha_1-1)_+ + \alpha_2}$, where α_1 is defined in Condition 8, and α_2 is a positive constant.

Condition 6 regulates the behavior of the regularization parameter λ_n around 0 and ∞. From the result of Theorem 4.2, we see that for different penalties, the "extra term" $\sqrt{s}\lambda_n\rho'(\beta_n^*; \lambda_n)$ in the L_2 estimation loss will require either extra conditions on the λ_n or extra conditions on the minimum signal strength β_n^* (see Theorems 4.3–4.5 for further details) and can govern estimation efficiency of the penalized estimators.

CONDITION 7. Let $\kappa_0 = \max_{\delta\in\mathcal{N}_0}\kappa(\rho, \delta)$, where $\mathcal{N}_0 = \{\delta \in R^s : \|\delta - \boldsymbol{\beta}_1^*\|_\infty \leq \beta_n^*\}$. Assume that λ_n and β_n^* satisfy that (i) $\beta_n^* \gg \sqrt{s}(n^{-1/2} + \lambda_n\rho'(\beta_n^*))$ and (ii) $\lambda_{\min}(\boldsymbol{\Sigma}_{\beta_1^*}) > \lambda_n\kappa_0$.

Condition 7(i) is employed to make $\hat{\boldsymbol{\beta}}_1^{\mathbf{o}}$ fall in $\mathcal{N}_0$ with probability tending to one. For LASSO, since $\rho'(\beta_n^*) = 1$, it means that $\beta_n^* \gg \sqrt{s}\lambda_n$. By Condition 6, it reduces to $\beta_n^* \gg \sqrt{s}n^{-0.5+(0.5\alpha+\alpha_1-1)_+ + \alpha_2}$. For SCAD, if $\beta_n^* \gg \lambda_n$, then $\rho'(\beta_n^*) = 0$ when n is large enough and hence it requires that $\beta_n^* \gg \sqrt{s}n^{-0.5}$. Therefore, Condition 7(i) is less restrictive for SCAD-like penalties. Condition 7(ii) is used to ensure the condition in (7) holds with probability tending to one (see the proof of Theorem 4.3). It always holds when $\kappa_0 = 0$ (e.g., for the LASSO penalty) and is satisfied for the SCAD type of penalty when $\beta_n^* \gg \lambda_n$.

CONDITION 8. For $\alpha_1 > 0$ and $0 < C < \infty$,

$$\sup_{0\le t\le \tau}\sup_{\mathbf{v}_1\in\mathcal{B}(\boldsymbol{\beta}_1^*,\beta_n^*)} \|\tilde{\mathbf{V}}(t,\mathbf{v})\|_{2,\infty} = \min\left(C\frac{\rho'(0+)}{\rho'(\beta_n^*)}, O_p(n^{\alpha_1})\right),$$

where $\mathcal{B}(\boldsymbol{\beta}_1^*, \beta_n^*)$ is an s-dimensional ball centered at $\boldsymbol{\beta}_1^*$ with radius β_n^*, for $\mathbf{v} = (\mathbf{v}_1^T, \mathbf{0}^T)^T$,

$$\tilde{\mathbf{V}}(t,\mathbf{v}) = \frac{S_n^{(0)}(\mathbf{v},t)S_{n21}^{(2)}(\mathbf{v},t) - S_{n2}^{(1)}(\mathbf{v},t)(S_{n1}^{(1)}(\mathbf{v},t))^T}{\{S_n^{(0)}(\mathbf{v},t)\}^2} \in \mathbb{R}^{(p-s)\times s}$$

and $\|\tilde{\mathbf{V}}(t,\mathbf{v})\|_{2,\infty} = \max_{\|\mathbf{x}\|_2=1}\|\tilde{\mathbf{V}}(t,\mathbf{v})\mathbf{x}\|_\infty$.

As noted in Fleming and Harrington [(1991), page 149], $\tilde{\mathbf{V}}(\boldsymbol{\beta}, t)$ is an empirical covariance matrix of $\mathbf{X}_i(t)$ computed with weights proportional to $Y_i(t) \times \exp\{\boldsymbol{\beta}^T\mathbf{X}_i(t)\}$. Hence, $\tilde{\mathbf{V}}$ is the empirical covariance matrix between the important variables $\mathbf{S}_i(t)$ and unimportant variables $\mathbf{Q}_i(t)$. Condition 8 controls the uniform growth rate of the norm of these covariance matrices, a notion of weak correlation between $\mathbf{S}_i(t)$ and $\mathbf{Q}_i(t)$. For the L_1 penalty, $\rho'(\beta_n^*) = 1$, and Condition 8 becomes a version of "strong irrepresentable" condition [Zhao and Yu (2006)] for censored data. It is very stringent as the right-hand side has to be bounded by $O(1)$. On the other hand for the SCAD penalty, if $\beta_n^* \gg \lambda_n$, then $\rho'(\beta_n^*) = 0$ when n is large enough. Therefore, Condition 8 is significantly relaxed to $O(n^{\alpha_1})$. In general, when a folded concave penalty is employed, the upper bound on the right-hand side in Condition 8 can grow to infinity at polynomial rate. This was also noted in the work of Fan and Lv (2011) in the context of generalized linear models.

THEOREM 4.3 (Strong oracle). *Let the oracle estimator $\hat{\boldsymbol{\beta}}^{\mathbf{o}}$ be a local maximizer of $\mathcal{C}(\boldsymbol{\beta}_1, \tau)$ given by Theorem 4.2. If $\max_j(\sigma_j^2) = O(n^{(0.5\alpha+\alpha_1-1)_+ +\alpha_2})$, and Conditions 1–8 hold, then with probability tending to one, there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $\mathcal{C}(\boldsymbol{\beta}, \tau)$ such that*

$$P(\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\mathbf{o}}) \ge 1 - c_0(p-s)\exp\{-c_1 n^{(0.5\alpha+\alpha_1-1)_+ +\alpha_2}\},$$

where c_0 and c_1 are positive constants.

PROOF. It suffices to show that $\hat{\boldsymbol{\beta}}^{\mathbf{o}}$ is a local maximizer of $\mathcal{C}(\boldsymbol{\beta}, \tau)$ on a set Ω_n which has a probability tending to one. By Theorem 2.1, we need to show that, with probability tending to one, $\hat{\boldsymbol{\beta}}^{\mathbf{o}}$ satisfies (5)–(7). Since $\hat{\boldsymbol{\beta}}^{\mathbf{o}}$ already satisfies (5) by definition, we are left to check (6) and (7).

Define $\Omega_n = \{\boldsymbol{\xi} : \|\boldsymbol{\xi}_{\mathcal{M}_*^c}\|_\infty \le \sqrt{n}u_n\}$ for some diverging sequence u_n to be chosen later, where $\boldsymbol{\xi}_{\mathcal{M}_*^c}$ is the subvector of $\boldsymbol{\xi}$ with indices in $\mathcal{M}_*^c$. By Theorem 3.1, there exist positive constants c_0 and c_1 such that

$$P(|\xi_j| > \sqrt{n}u_n) \le c_0\exp\{-c_1 u_n\}$$

uniformly over j. Then using the Bonferroni union bound, we obtain that

$$P(\Omega_n) \geq 1 - \sum_{j \in \mathcal{M}_*^c} P(|\xi_j| > \sqrt{n} u_n) \tag{18}$$
$$\geq 1 - c_0(p-s)e^{-c_1 u_n} \to 1 \qquad \text{as } n \to \infty,$$

where u_n can be chosen later to make $(p-s)e^{-c_1 u_n} \to 0$. We now check if (6) holds for $\hat{\boldsymbol{\beta}}^{\mathbf{o}}$ on the set Ω_n. Denote by $\boldsymbol{\rho}'_{\mathcal{M}_*^c}$ the subvector of $\boldsymbol{\rho}'(|\hat{\boldsymbol{\beta}}^{\mathbf{o}}|)$ with indexes in $\mathcal{M}_*^c$. Let $\gamma(\boldsymbol{\beta}) = \int_0^t S_n^{(1)}(\boldsymbol{\beta}, u)/S_n^{(0)}(\boldsymbol{\beta}, u)\, d\bar{N}(u)$ and

$$\mathbf{z}(\hat{\beta}^{\mathbf{o}}) = \sum_{i=1}^n \int_0^\tau \{\mathbf{Q}_i(t) - \mathbf{E}_n^{(2)}(\hat{\boldsymbol{\beta}}^{\mathbf{o}}, t)\}\, dN_i(t),$$

where $\mathbf{E}_n^{(2)}(\boldsymbol{\beta}, t) = S_{n2}^{(1)}(\boldsymbol{\beta}, t)/S_n^{(0)}(\boldsymbol{\beta}, t)$. Then by Condition 1, we have

$$\|\mathbf{z}(\hat{\boldsymbol{\beta}}^{\mathbf{o}})\|_\infty \leq \|\boldsymbol{\xi}_{\mathcal{M}_*^c}\|_\infty + \|\gamma_{\mathcal{M}_*^c}(\boldsymbol{\beta}^*) - \gamma_{\mathcal{M}_*^c}(\hat{\boldsymbol{\beta}}^{\mathbf{o}})\|_\infty$$
$$= O\biggl(\sqrt{n} u_n + \biggl\| \int_0^t \tilde{V}(u, \mathbf{v}_1)(\hat{\boldsymbol{\beta}}_1^{\mathbf{o}} - \boldsymbol{\beta}_1^*)\, d\bar{N}(u) \biggr\|_\infty \biggr) \tag{19}$$
$$= O\Bigl(\sqrt{n} u_n + \sup_{0 \leq u \leq \tau} \sup_{\mathbf{v}_1 \in \mathcal{B}(\boldsymbol{\beta}_1, \beta_n^*)} \|\tilde{V}(u, \mathbf{v}_1)\|_{2,\infty} \|\hat{\boldsymbol{\beta}}_1^{\mathbf{o}} - \boldsymbol{\beta}_1^*\|_2\Bigr),$$

where $\mathbf{v} = (\mathbf{v}_1^T, \mathbf{0}^T)^T$ with $\mathbf{v}_1$ being between $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_1^{\mathbf{o}}$, and $\tilde{V}(u, \mathbf{v}_1)$ is defined in Condition 8. By Theorem 4.2 and Condition 8, we obtain that $(n\lambda_n \times \rho'(0+))^{-1}\|\mathbf{z}(\hat{\boldsymbol{\beta}}^{\mathbf{o}})\|_\infty$ is bounded by

$$n^{-1}\lambda_n^{-1} O_p\Bigl\{\sqrt{n} u_n + \sup_{0 \leq u \leq 1} \|\tilde{V}(u, \mathbf{v}_1)\|_{2,\infty} \sqrt{s}\bigl(n^{-1/2} + \lambda_n \rho'(\beta_n^*)\bigr)\Bigr\}$$
$$= O_p\{n^{-1/2}\lambda_n^{-1}(u_n + n^{0.5\alpha + \alpha_1 - 1}) + n^{-1+0.5\alpha}\rho'(0+)\} \to 0,$$

if we take $u_n = n^{(0.5\alpha + \alpha_1 - 1)_+ + \alpha_2}$ and $\lambda_n \gg n^{-0.5 + (0.5\alpha + \alpha_1 - 1)_+ + \alpha_2}$. Therefore, (6) holds on Ω_n. Once $\delta < (0.5\alpha + \alpha_1 - 1)_+ + \alpha_2$, (18) holds.

We are now left to show that (7) holds with probability tending to one, that is, $\lambda_{\min}\{n^{-1}\int_0^\tau \mathbf{V}(\hat{\boldsymbol{\beta}}^{\mathbf{o}}, t)\, d\bar{N}(t)\} > \lambda_n \kappa(\rho, \hat{\boldsymbol{\beta}}_1^{\mathbf{o}})$, which is guaranteed by Condition 7. In fact, by Theorem 4.2 and Condition 7(i), with probability tending to one, $\hat{\boldsymbol{\beta}}_1^{\mathbf{o}}$ falls in $\mathcal{N}_0$ as $n \to \infty$, so that $\kappa(\rho, \hat{\boldsymbol{\beta}}_1^{\mathbf{o}}) \leq \kappa_0$. Hence, by Condition 7(ii), with probability tending to one,

$$\lambda_{\min}(\boldsymbol{\Sigma}_{\beta_1^*}) > \lambda_n \kappa(\rho, \hat{\boldsymbol{\beta}}_1^{\mathbf{o}}). \tag{20}$$

Recall that

$$n^{-1}\int_0^\tau \mathbf{V}(\hat{\boldsymbol{\beta}}_1^{\mathbf{o}}, t)\, d\bar{N}(t) = \mathcal{I}_{\hat{\beta}_1^{\mathbf{o}}} + \mathcal{W}_{\hat{\beta}_1^{\mathbf{o}}} = \boldsymbol{\Sigma}_{\hat{\beta}_1^{\mathbf{o}}} + (\mathcal{I}_{\hat{\beta}_1^{\mathbf{o}}} - \boldsymbol{\Sigma}_{\hat{\beta}_1^{\mathbf{o}}}) + \mathcal{W}_{\hat{\beta}_1^{\mathbf{o}}}.$$

By Theorem 4.2, as $n \to \infty$, $\hat{\boldsymbol\beta}_1^{\mathbf{o}} \in \mathcal{B}$ with probability tending to one. This combining with Lemma 4.1 and Lemma 2.3 in the supplementary material [Bradic, Fan and Jiang (2011)] leads to

$$n^{-1}\int_0^\tau \mathbf{V}(\hat{\boldsymbol\beta}_1^{\mathbf{o}}, t)\, d\bar{N}(t) = \boldsymbol\Sigma_{\hat\beta_1^{\mathbf{o}}} + E,$$

where $\|E\|_2 = o_p(1)$. By Condition 2(i), (iii), with probability tending to one,

$$\|\boldsymbol\Sigma_{\hat\beta_1^{\mathbf{o}}} - \boldsymbol\Sigma_{\beta_1^*}\|_2 = o_p(1).$$

Let $E^* = n^{-1}\int_0^\tau \mathbf{V}(\hat{\boldsymbol\beta}_1^{\mathbf{o}}, t)\, d\bar{N}(t) - \boldsymbol\Sigma_{\beta_1^*}$. Then $\|E^*\|_2 = o_p(1)$. Using Weyl's pertubation theorem [Bhatia (1997)], we obtain that

$$\min_{1\le k\le s}\left|\lambda_k\left\{n^{-1}\int_0^\tau \mathbf{V}(\hat{\boldsymbol\beta}_1^{\mathbf{o}}, t)\, d\bar{N}(t)\right\} - \lambda_k(\boldsymbol\Sigma_{\beta_1^*})\right| \le \|E^*\|_2,$$

where $\lambda_k(\boldsymbol\Sigma_{\beta_1^*})$ is the kth largest eigenvalue of $\boldsymbol\Sigma_{\beta_1^*}$. Therefore,

$$\lambda_{\min}\left\{n^{-1}\int_0^\tau \mathbf{V}(\hat{\boldsymbol\beta}_1^{\mathbf{o}}, t)\, d\bar{N}(t)\right\} = \lambda_{\min}(\boldsymbol\Sigma_{\beta_1^*}) + o_p(1).$$

This combining with (20) yields that with probability tending to one

$$\lambda_{\min}\left\{n^{-1}\int_0^\tau \mathbf{V}(\hat{\boldsymbol\beta}_1^{\mathbf{o}}, t)\, d\bar{N}(t)\right\} > \lambda_n\kappa(\rho, \hat{\boldsymbol\beta}_1^{\mathbf{o}}). \qquad \square$$

The theorem becomes nontrivial if $\delta < (0.5\alpha + \alpha_1 - 1)_+ + \alpha_2$, since $\log p = O(n^\delta)$. Apart from the work of Bradic, Fan and Wang (2011), no formal work explicitly relates the oracle property and the full and effective dimensionalities. Theorem 4.3 shows that $\hat{\boldsymbol\beta}$ becomes the biased oracle with probability tending to one exponentially fast. Then combining Theorems 4.2 and 4.3 leads to the following L_2 estimation loss:

$$\|\hat{\boldsymbol\beta}_1 - \boldsymbol\beta_1^*\|_2 = O_P\{\sqrt{s}(n^{-1/2} + \lambda_n\rho'(\beta_n^*))\}. \tag{21}$$

This theorem tells us that the resulting estimator behaves as if the true set of "important variables" (i.e., as oracle estimator) were known with probability converging to 1 as both p and n go to ∞. The previous notions of oracle were that the estimator behaves like the oracle rather than an actual oracle itself. Classical oracle property of Fan and Li (2002) or sign consistency of Bickel, Ritov and Tsybakov (2009) are both corollaries of this result. In this sense Theorem 4.3 introduces a tighter notion of an oracle property. It was first mentioned in Kim, Choi and Oh (2008) for the SCAD estimator of the linear model with polynomial dimensionality and then extended by Bradic, Fan and Wang (2011) to the penalized M-estimators under the ultra-high dimensionality setting. Extending their work to Cox's model was exceptionally challenging because of martingale and censoring structures.

THEOREM 4.4 (LASSO). *Under Conditions 2–5, if* $\max_j(\sigma_j^2) = O(n^{\alpha_2})$, $\sqrt{s}\lambda_n \to 0$, $\lambda_n \gg n^{-0.5+\alpha_2}$ *and*

$$\sup_{0 \le t \le \tau} \sup_{\mathbf{v}_1 \in \mathcal{B}(\boldsymbol{\beta}_1^*, \beta_n^*)} \|\tilde{\mathbf{V}}(t, \mathbf{v})\|_{2,\infty} = O_p(1),$$

then the result in Theorem 4.3 holds for LASSO *estimator with probability being at least* $1 - c_0(p - s)\exp\{-c_1 n^{\alpha_2}\}$. *Furthermore*,

$$\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 = O_P(\sqrt{s}\lambda_n).$$

The proof of this theorem is relegated to the supplementary material [Bradic, Fan and Jiang (2011)]. For the LASSO, the rate of convergence for nonvanishing components is dominated by the bias term $\lambda_n \gg n^{-1/2}$. In addition, since $s = n^{\alpha}$, the condition $\sqrt{s}\lambda_n \to 0$ indicates that $\alpha < 1 - 2\alpha_2$, where $\alpha_2 \in [0, 1/2)$. That is, the bigger is α_2, and the smaller sparsity dimension s can be recovered using LASSO. Moreover, LASSO with $\alpha_2 < 1/2$ requires $p \ll \exp\{c_1 n^{\alpha_2}\}$ to achieve the strong oracle property. Hence, as p (or α_2) gets bigger, s (or α) should get smaller. This means that, as data dimensionality gets higher, recoverable problems get sparser. This is a new discovery and has not been documented in the literature. On the other hand for folded concave penalties, faster rates of convergence are obtained with fewer restrictions on p and s. This can be seen from the following result, which is a straightforward corollary of Theorem 4.3 and whose proof is left for the supplementary material [Bradic, Fan and Jiang (2011)].

THEOREM 4.5 (SCAD). *Under Conditions 1–5, if* $\beta_n^* \gg \lambda_n$, $\max_j(\sigma_j^2) = O(n^{(0.5\alpha+\alpha_1-1)_+ + \alpha_2})$, $\lambda_n \gg n^{-0.5+(0.5\alpha+\alpha_1-1)_+ + \alpha_2}$ *and*

$$\sup_{0 \le t \le \tau} \sup_{\mathbf{v}_1 \in \mathcal{B}(\boldsymbol{\beta}_1^*, \beta_n^*)} \|\tilde{\mathbf{V}}(t, \mathbf{v})\|_{2,\infty} = O_p(n^{\alpha_1}),$$

then the result in Theorem 4.3 holds for SCAD *estimator with probability being at least* $1 - c_0(p - s)\exp\{-c_1 n^{(0.5\alpha+\alpha_1-1)_+ + \alpha_2}\}$. *Furthermore*,

$$\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 = O_P(\sqrt{s/n}).$$

Note that the proof of Theorem 4.3 shows $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ on a set whose probability measure is going to one exponentially fast. For statistical inference about $\boldsymbol{\beta}$, asymptotic properties of $\hat{\boldsymbol{\beta}}_1$ are needed to be explored. To be able to construct confidence intervals of $\boldsymbol{\beta}_1$ we need to derive its asymptotic distribution. This was done in Fan and Li (2002) for fixed p and in Cai et al. (2005) for $p = o(n^{1/4})$. Here, we allow p to diverge at exponential rate $O(\exp\{n^{\delta}\})$ and the effective dimensionality s to diverge at rate of $o(n^{1/3})$. To the best of our knowledge there is no work available for such a setting. Extending the previous work to such a NP-dimensional setting is not trivial and requires complicated eigenvalue results.

Moreover, the large deviation result in Section 3, the strong oracle result in Theorem 4.3 and Lemmas 2.1–2.3 in the supplementary material [Bradic, Fan and Jiang (2011)] are essential for establishing the desired asymptotics. Moreover, the following Lemma 4.2 is an important extension of the classical asymptotic Taylor expansion results when the number of parameters is diverging with the sample size.

LEMMA 4.2. *For any* $s \times 1$ *unit vector* $\mathbf{b}_n$, *let*

$$\phi_n = \mathbf{b}_n^T \boldsymbol{\Sigma}_{\beta_1^*}^{1/2} (-n^{-1}\,\partial U_n(\boldsymbol{\beta}_1^*))^{-1} n^{-1/2} U_n(\boldsymbol{\beta}_1^*)$$

and

$$\phi_{n1} = \mathbf{b}_n^T \boldsymbol{\Sigma}_{\beta_1^*}^{-1/2} n^{-1/2} U_n(\boldsymbol{\beta}_1^*).$$

If Conditions 2, 4 *and* 5 *hold and if* $s = o(n^{1/3})$, *then* $\phi_n = \phi_{n1} + o_p(1)$.

PROOF. Let $\mathcal{B} = I + \mathcal{I}_{\beta_1^*}^{-1/2} \mathcal{W}_{\beta_1^*} \mathcal{I}_{\beta_1^*}^{-1/2}$, where I is an $s \times s$ identity matrix. Using the Bauer–Fike inequality [Bhatia (1997)], we obtain that

$$|\lambda(\mathcal{B}) - 1| \le \|\mathcal{I}_{\beta_1^*}^{-1/2} \mathcal{W}_{\beta_1^*} \mathcal{I}_{\beta_1^*}^{-1/2}\|_2.$$

Then by the Hölder inequality we have $|\lambda(\mathcal{B}) - 1| \le \|\mathcal{I}_{\beta_1^*}^{-1/2}\|_2^2 \|\mathcal{W}_{\beta_1^*}\|_2$. Applying Condition 4 and Lemma 4.1 and Lemma 2.3 of the supplementary material [Bradic, Fan and Jiang (2011)], we establish that

$$\lambda(\mathcal{B}) = 1 + O_p\bigl(s/\sqrt{n}\bigr) \tag{22}$$

uniformly for all eigenvalues of $\mathcal{B}$. Note that

$$(-n^{-1}\,\partial U_n(\boldsymbol{\beta}_1^*))^{-1} = (\mathcal{I}_{\beta_1^*} + \mathcal{W}_{\beta_1^*})^{-1} = \mathcal{I}_{\beta_1^*}^{-1} - \mathcal{I}_{\beta_1^*}^{-1/2} \{I - \mathcal{B}^{-1}\} \mathcal{I}_{\beta_1^*}^{-1/2}.$$

It follows that

$$\begin{aligned}\phi_n &= \mathbf{b}_n^T \boldsymbol{\Sigma}_{\beta_1^*}^{1/2} \mathcal{I}_{\beta_1^*}^{-1} n^{-1/2} U_n(\boldsymbol{\beta}_1^*) - \mathbf{b}_n^T \boldsymbol{\Sigma}_{\beta_1^*}^{1/2} \mathcal{I}_{\beta_1^*}^{-1/2} \{I - \mathcal{B}^{-1}\} \mathcal{I}_{\beta_1^*}^{-1/2} n^{-1/2} U_n(\boldsymbol{\beta}_1^*) \\ &\equiv \phi_{n1} - \phi_{n2}.\end{aligned}$$

Since $I - \mathcal{B}^{-1}$ is symmetrical, $r_\sigma(I - \mathcal{B}^{-1}) = \|I - \mathcal{B}^{-1}\|_2$. Recall that $\|\mathbf{b}_n\|_2 = 1$; it follows that

$$|\phi_{n2}| \le r_\sigma(I - \mathcal{B}^{-1}) \|\boldsymbol{\Sigma}_{\beta_1^*}^{1/2}\|_2 \|\mathcal{I}_{\beta_1^*}^{-1/2}\|_2^2 \|n^{-1/2} U_n(\boldsymbol{\beta}_1^*)\|_2.$$

By Condition 4, $\|\boldsymbol{\Sigma}_{\beta_1^*}^{1/2}\|_2 = O_p(1)$. From Lemma 4.1, we have $\|\mathcal{I}_{\beta_1^*}^{-1/2}\|_2 = O_p(1)$. By Lemma 2.2 in the supplementary material [Bradic, Fan and Jiang (2011)], $\|n^{-1/2} U_n(\boldsymbol{\beta}_1^*)\|_2 = O_p(\sqrt{s})$. Therefore,

$$|\phi_{n2}| = r_\sigma(I - \mathcal{B}^{-1}) O_p\bigl(\sqrt{s}\bigr). \tag{23}$$

By definition, it is easy to see that

$$r_\sigma(I - \mathcal{B}^{-1}) = \max\{|1 - \lambda| : \lambda \in \sigma(\mathcal{B}^{-1})\} = \max\{|1 - \lambda^{-1}| : \lambda \in \sigma(\mathcal{B})\},$$

which, combined with (22), leads to $r_\sigma(I - \mathcal{B}^{-1}) = O_p(s/\sqrt{n})$. This together with (23) yields that $\phi_{n2} = O_p(\sqrt{s^3/n}) = o_p(1)$, if $s = o(n^{1/3})$. Hence, $\phi_n = \phi_{n1} + o_p(1)$. □

With the Lemma 4.2 and technical lemmas presented in the supplementary material [Bradic, Fan and Jiang (2011)] we are ready to state the results on the asymptotic behavior of the penalized estimator. Detailed proof is included in the supplementary material [Bradic, Fan and Jiang (2011)].

THEOREM 4.6. *Under Conditions 1–8, and for $\lambda_n \rho'(\beta_n^*) = o((sn)^{-1/2})$ for any $s \times 1$ unit vector $\mathbf{b}_n$, if $s = o(n^{1/3})$, the penalized partial likelihood estimator $\hat{\boldsymbol\beta}_1$ from (21) satisfies*

$$\sqrt{n}\mathbf{b}_n^T \boldsymbol\Sigma_{\beta_1^*}^{1/2}(\hat{\boldsymbol\beta}_1 - \boldsymbol\beta_1^*) \to \mathcal{N}(0, 1).$$

Theorems 4.3 and 4.6 claim that $\hat{\boldsymbol\beta}$ enjoys model selection consistency and achieves the information bound mimicking that of the oracle estimator $\hat{\boldsymbol\beta}^{\mathbf{o}}$.

5. Iterative coordinate ascent algorithm (ICA). Coordinate-wise algorithms are especially attractive for $p \gg n$ and have been previously introduced for penalized least-squares with the L_q-penalty by Daubechies, Defrise and De Mol (2004), Friedman et al. (2007), Wu and Lange (2008) and for generalized linear models with the folded concave penalty by Friedman, Hastie and Tibshirani (2010) and Fan and Lv (2011). By Condition 1 and Proposition 2.7.1 in Bertsekas (2003), the coordinate-wise maximization algorithm in each iteration provides limits that are stationary points of the overall optimization (3). Therefore, each output of ICA algorithm will give a stationary point. We will adapt the algorithm in Fan and Lv (2011) to the censored data.

First, let us, with slight abuse in notation, denote by $Q_n(\boldsymbol\beta) = L_n(\boldsymbol\beta) - P_n(\boldsymbol\beta)$, where $L_n(\cdot)$ and $P_n(\cdot)$ stand for the loss and penalty parts, respectively. Let $l_n(\boldsymbol\beta, \zeta, j)$ be the *partial quadratic approximation* of $L_n(\boldsymbol\beta)$ at $\zeta \in R^p$ along the jth coordinate, where $\{\beta_k = \zeta_k, k \neq j\}$ are held fixed, but β_j is allowed to vary

$$q_n(\boldsymbol\beta_j, \zeta, j) = l_n(\boldsymbol\beta, \zeta, j) - np_{\lambda_n}(|\beta_j|).$$

Because of the complex likelihood function we need an additional loop to compute the partial quadratic approximation.

This penalized quadratic optimization problem can be solved analytically, avoiding the challenges of nonconcave optimization. It updates each coordinate if the maximizer of the penalized univariate optimization strictly increases the objective function $Q_n(\boldsymbol{\beta})$ and if it satisfies $\{j : |z_j| > \rho'(0+)\}$. The algorithm stops when two values of the objective function $Q_n(\boldsymbol{\beta})$ are not different by more than 10^{-8}, say. Details of the algorithm are presented in the supplementary material [Bradic, Fan and Jiang (2011)].

5.1. *Simulated examples.* To show good model selection and estimation properties of the proposed methodology, we simulated 100 standard Toeplitz ensembles of size 100 with population correlation $\rho(X_i, X_j) = \rho^{|i-j|}$ with ρ ranging from 0.25, 0.5, 0.75 and 0.9. The distribution of censoring time C is exponential with mean $U * \exp\{\mathbf{X}_i^T\boldsymbol{\beta}\}$, where U is randomly generated from uniform distribution over $[1, 3]$ for each simulated data set. This censoring was used in Fan and Li (2002), which makes about 30% of the data censored. The full and effective dimensionalities of the true parameter $\boldsymbol{\beta}$ are taken as $\{100, 4\}$, $\{1{,}000, 4\}$, $\{5{,}000, 4\}$ and $\{1{,}000, 25\}$, respectively, with values ± 1 randomly placed (the rest is set as zero). The penalties employed are LASSO [Tibshirani (1996)], SCAD [Fan and Li (2001)], SIC*a* [Lv and Fan (2009)] with $p_\lambda(|\beta_j|) = (\lambda + 1)|\beta_j|/(\lambda + |\beta_j|)$ and MCP+ [Zhang (2010)] with all regularization parameters being computed with 5-fold sparse generalized cross validation; see Section 5.2 and Table 2 therein for detailed discussion on the choice of cross validation statistics.

The results of the simulations are summarized into three tables (see Table 1 in the main text and Tables 2 and 3 in the supplementary material [Bradic, Fan and Jiang (2011)]) where we reported the median prediction error (PE)

$$\mathbb{P}_n[\exp\{-\boldsymbol{\beta}^{*T}\mathbf{X}\} - \exp\{-\hat{\boldsymbol{\beta}}^T\mathbf{X}\}]^2, \tag{24}$$

where $\mathbb{P}_n$ stands for the empirical probability measure. We also report the median number of nonzero parameters estimated in the set $\mathcal{M}_*$ as the number of true positives TP. Furthermore, we summarize the median number of nonzero estimates of the set $\mathcal{M}_*^c$ as the number of false positives FP.

Table 1 summarizes three $p \geq n$ examples, where especially the last two stress the strengths of the methods when $p \gg n$ and spectra of the design matrix is high; see Table 1 in the supplementary material [Bradic, Fan and Jiang (2011)]. All four methods work quite well, where LASSO has higher PE than the rest, with SCAD and MCP+ performing quite closely to each other. SICa performs worse than others, always loosing a number of TPs. The case of $\rho = 0.90$ affects all methods in bigger prediction error and smaller number of TP, where the jump is the largest in LASSO penalty. SCAD and MCP keep their performance similarly to the oracle one through all examples, hence verifying the strengths of nonconvex penalties. For more detailed discussions and results when the oracle estimator fails, when the censoring rate is too high and assessing the relative estimation efficiency of LASSO estimator with respect to SCAD, SICa and MCP+, we direct you to the supplementary material [Bradic, Fan and Jiang (2011)] for this paper.

TABLE 1
Simulation results for $p \geq n$ under correlation settings ranging from 0.25 to 0.90 with medium prediction error (MPE), # of true positives (TP), # of false positives (FP) and standard deviation in parenthesis of each estimate

	MPE	**TP**	**FP**	**MPE**	**TP**	**FP**
			Settings of $n = 100$, $p = 100$, $s = 4$			
	Case $\rho = 0.25$			Case $\rho = 0.5$		
Oracle	0.0154 (1.27$^\sharp$)	4	0	0.0215 (1.97$^\sharp$)	4	0
LASSO	0.0178 (1.26)	4 (1.61)	2 (33.34)	0.0284 (2.12)	4 (1.52)	13 (33.02)
SCAD	0.0161 (1.24)	4 (1.61)	2 (34.21)	0.0223 (2.03)	4 (1.52)	13 (35.56)
SICa	0.0190 (1.27)	3 (1.48)	2 (26.11)	0.0275 (2.43)	3 (1.44)	9 (21.54)
MCP+	0.0166 (1.22)	3 (1.71)	2 (32.49)	0.0271 (2.33)	4 (1.54)	24 (34.62)
	Case $\rho = 0.75$			Case $\rho = 0.9$		
Oracle	0.0322 (2.05$^\sharp$)	4	0	0.0538 (4.43$^\sharp$)	4	0
LASSO	0.0371 (2.42)	3 (1.14)	12 (31.21)	0.0665 (4.62)	2 (1.48)	13 (32.16)
SCAD	0.0326 (2.12)	4 (1.14)	12 (31.53)	0.0549 (3.36)	3.5 (1.49)	8 (31.11)
SICa	0.0343 (2.27)	2 (1.30)	3 (18.41)	0.0566 (3.26)	2 (1.32)	6 (24.42)
MCP+	0.0326 (2.21)	3.5 (1.22)	12 (32.31)	0.0558 (3.44)	2.5 (1.29)	15 (29.68)
			Settings of $n = 100$, $p = 1{,}000$, $s = 4$			
	Case $\rho = 0.25$			Case $\rho = 0.5$		
Oracle	0.0154 (1.27$^\sharp$)	4	0	0.0215 (1.97$^\sharp$)	4	0
LASSO	0.0201 (1.38)	4 (0.85)	23 (371.8)	0.0383 (3.16)	3.5 (1.23)	45 (532.1)
SCAD	0.0162 (1.25)	4 (0.83)	15 (323.4)	0.0281 (2.12)	4 (1.12)	36 (430.3)
SICa	0.0189 (1.17)	3.5 (0.54)	9 (120.5)	0.0492 (3.18)	3 (1.43)	15 (319.4)
MCP+	0.0192 (1.23)	4 (0.83)	17 (345.5)	0.0281 (2.15)	4 (1.12)	36 (409.2)
	Case $\rho = 0.75$			Case $\rho = 0.9$		
Oracle	0.0322 (2.05$^\sharp$)	4	0	0.0538 (4.43$^\sharp$)	4	0
LASSO	0.0497 (3.16)	3 (0.44)	96 (306.5)	0.0703 (4.24)	3 (1.54)	97 (411.5)
SCAD	0.0358 (2.45)	4 (0.34)	85 (250.7)	0.0583 (4.13)	4 (1.51)	67 (380.9)
SICa	0.0372 (2.15)	2 (1.30)	90.5 (90.3)	0.0546 (3.98)	1 (1.78)	30 (354.1)
MCP+	0.0361 (2.77)	3.5 (1.14)	90 (320.4)	0.0592 (4.25)	3.5 (1.58)	98 (402.3)
			Settings of $n = 100$, $p = 5{,}000$, $s = 4$			
	Case $\rho = 0.25$			Case $\rho = 0.5$		
Oracle	0.0154 (1.27$^\sharp$)	4	0	0.0215 (1.97$^\sharp$)	4	0
LASSO	0.0220 (1.49)	4 (1.05)	68 (398.1)	0.0462 (4.05)	3.5 (1.64)	33 (206.8)
SCAD	0.0170 (1.28)	4 (1.05)	67 (298.2)	0.0328 (3.15)	3.5 (1.56)	21.5 (205.4)
SICa	0.0195 (1.19)	2.5 (1.17)	14 (345.7)	0.0285 (3.35)	4 (1.41)	30 (323.3)
MCP+	0.0188 (1.29)	3 (1.10)	67 (298.2)	0.0358 (2.85)	3.5 (1.51)	73.5 (348.7)
	Case $\rho = 0.75$			Case $\rho = 0.9$		
Oracle	0.0322 (2.05$^\sharp$)	4	0	0.0538 (4.43$^\sharp$)	4	0
LASSO	0.0567 (5.02)	3 (1.73)	23 (250.5)	0.0865 (4.52)	2 (1.23)	59 (208.8)
SCAD	0.0360 (2.31)	4 (1.51)	18 (234.7)	0.0596 (4.12)	4 (0.89)	49 (105.4)
SICa	0.0385 (2.13)	2.5 (1.30)	3 (225.2)	0.0602 (4.92)	3 (0.45)	46 (90.3)
MCP+	0.0392 (2.82)	4 (1.74)	4 (326.2)	0.0578 (4.33)	4 (0.89)	11 (217.1)

$^\sharp$stands for column of standard deviation $\times$ 100.

5.2. *Real data example*. To demonstrate the strength of the proposed methodology, in this section, we present gene association study with respect to the survival time of non-Hodgkin's lymphoma. Genetic mechanisms responsible for the clinical heterogeneity of follicular lymphoma are still unknown. Dave et al. (2004) have collected gene expression data on 191 biopsy specimens obtained from patients with untreated follicular lymphoma. RNA was extracted from fresh-frozen tumor-biopsy specimens and survival times, from 191 patients, who had received a diagnosis between 1974 and 2001, which were obtained from seven institutions and examined for gene expression with the use of Affymetrix U133A and U133B microarrays. The median age at diagnosis was 51 years (range, 23 to 81), and the median follow-up time was 6.6 years (range, less than 1.0 to 28.2). The dataset was obtained from http://llmpp.nih.gov/FL.

The full cohort study included 44,187 probe expressions values out which only 34,188 were properly annotated. Among these, some received multiple (2–7) measurements per gene. We took the median value as a unique representative and were left with 17,118 different genes presented. We separated the dataset into training and testing sets with 80% and 20% of censored samples, respectively. The censoring rate of 50% was kept in each of the training and testing samples. Recorded for each individual are follow up time, indicator of the status at the follow up time and measurements of expression value for each Affymetrix probe set.

The classical L fold cross-validation is defined as

$$\mathrm{CV}(\lambda) = \sum_{k=1}^{L} \{ l(\hat{\boldsymbol{\beta}}_{\lambda}^{(-k)}) - l^{(-k)}(\hat{\boldsymbol{\beta}}^{(-k)}{}_{\lambda}) \},$$

where l stands for the partial likelihood and $l^{(-k)}$ for the partial likelihood evaluated without the kth subset and similarly $\hat{\boldsymbol{\beta}}_{\lambda}^{(-k)}$ for the penalized estimator derived without using the kth subset. The measure of information contained in the full Cox partial likelihood is biased with respect to the number of nonzero elements and proper normalization is needed. The method of generalized cross validation proposed by Fan and Li (2002) works very well for small p but fails for large p because of its dependence on the inverse of the Hessian matrix of the partial likelihood. This inspired us to define a sparse approximation to the generalized cross-validation as

$$\mathrm{SGCV}(\lambda) = \sum_{k=1}^{L} \left(\frac{l(\hat{\boldsymbol{\beta}}_{\lambda}^{(-k)})}{n\{1 - \hat{s}_{\lambda}/n\}^2} - \frac{l^{(-k)}(\hat{\boldsymbol{\beta}}_{\lambda}^{(-k)})}{n^{(-k)}\{1 - \hat{s}_{\lambda}/n^{(-k)}\}^2} \right),$$

where $\hat{s}_{\lambda} = \|\hat{\beta}_{\lambda}^{(-k)}\|_0$ and $n^{(-k)}$ stands for the sample size of the whole set without the kth subset. Then, we choose the regularization parameter as

$$\hat{\lambda} = \operatorname*{arg\,min}_{\lambda : \hat{s}_{\lambda} < n} \mathrm{SGCV}(\lambda).$$

We applied 5-fold cross validation on the test set and evaluated its performance on the training set. The Nelson–Aalen estimate of the cumulative hazard rate function was used. The results are summarized in Table 2 and show a big difference

TABLE 2
Data summary with number of nonzero elements reported on the whole data set and prediction error and its standard deviation $\times$ 100 *comparisons reported on the training set* [*Dave et al.* (2004)]

	LASSO	**SCAD**	**SICA**	**MCP+**
CV				
# of nonzeros	2145	653	0	154
Prediction error	0.1516 (1.51)	0.1276 (1.60)	0.1898 (–)	0.1743 (1.45)
SGCV				
# of nonzeros	24	26	0	13
Prediction error	0.0812 (1.03)	0.0643 (1.02)	0.1898 (–)	0.1043 (0.78)

between the classical CV statistics and generalized one. The CV, being not scaled to the number of nonzero elements always prefers models with bigger number of nonzeros. Note that $\hat{s} > n$, for small λ, is caused by the artifact of ICA algorithm.

The SICa penalty completely fails in this example. It detects nonzeros only in 3 grid points with the number of nonzeros as 2, 3 and 879. Both CV methods fail to pick up the optimal one among the three points and choose the fourth one, which lead to no signal detection. This is not unexpected, since in all simulations SICa was always picking the least number of TP+FP; see Table 1.

Table 3 depicts the estimation results of the sparse generalized cross validation method with LASSO, SCAD and MCP+ penalties. All three penalties yield sign consistency of estimated coefficients among the selected gene sets. Note that the relative rankings of estimated corresponding coefficients are different among all methods. For example, gene FLJ40298 has the biggest absolute size in the SCAD penalty, it is ranked number 5 among those coefficients produced by LASSO penalty and it is not even selected in the MCP+ penalty. Interestingly, the common set of genes selected by LASSO and SCAD has very consistent estimated coefficients. For most genes MCP results in smaller estimated values than SCAD and LASSO.

6. Discussion. We have studied penalized log partial likelihood methods for ultra-high dimensional variable selection for Cox's regression models. With nonconcave penalties, we have shown that such methods have model selection consistency with oracle properties even for NP-dimensionality. We have established that oracle properties hold with probability converging to one exponentially fast, and that the rate explicitly depends on the real and intrinsic dimensionality p and s, respectively. We have also developed an exponential inequality for deviations of a counting process from its compensator. Results for LASSO penalty were obtained as a special case. It confirms explicitly that folded concave penalties allow for far weaker correlation structure than LASSO penalty. Furthermore, the asymptotic normality was proved, results of which can be used to construct confidence intervals of the estimated coefficients.

TABLE 3
Data estimation summary of the genes selected by the sparse generalized cross validation with standard deviation × 100‡ reported in the parenthesis

Gene annotation	**LASSO**	**SCAD**	**MCP+**
FOSB (BC036724)	−0.0093 (2.34‡)	×	−0.0027 (1.54‡)
GABRA6 (AK090735)	0.0070 (0.56)*	0.0150 (1.00‡)***	×
GHRH (AW_134884)	×	−0.0489 (1.39)**	×
GNGT_1 (BC030956)	×	−0.0041 (0.46)	×
HIST1H1E (BU603483)	−0.0026 (1.98)	−0.0032 (1.41)	×
HIST1H2AE (BE741093)	×	−0.0137 (0.41)**	×
IFNA2 (NM_000605)	×	0.0095 (1.29)	×
IMPG1 (NM_001563)	−0.0168 (2.56)	−0.0116 (0.81)	×
MATN3 (NM_002381)	0.0206 (0.89)**	0.0301 (0.36)***	0.0065 (1.25)
RTH (NM_000315)	−0.0032 (0.85)	−0.0177 (0.74)**	×
RAG2 (NM_000536)	−8.8728e–05 (1.56)	×	×
SCN9A (NM_002977)	0.0049 (0.87)	×	8.5785e–04 (1.56)
CXCL5 (NM_002994)	×	0.0026 (1.69)	×
SH3BGR (BM725357)	−0.0125 (0.25)***	×	×
HIST1H3B (NM_006770)	×	−0.0029 (0.81)	×
MARCO (BP872375)	0.0013 (2.54)	×	×
CLCA3 (NM_004921)	0.0172 (0.85)*	0.0170 (0.71)*	0.0171 (0.54)**
SEMA3A (XM_376647)	−0.0049 (1.15)	×	−3.8781e–05 (0.76)
KIAA0861 (BX694003)	−0.0261 (0.96)*	−0.0181 (0.74)**	−0.0170 (0.58)**
FSCN2 (NM_012418)	0.0136 (1.25)	0.0194 (1.44)	0.0058 (1.12)
DKFZP566K0 (ALO50040)	−0.0025 (1.36)	×	×
MORC (BC050307)	0.0204 (0.75)*	0.0204 (1.00)*	0.0165 (0.94)
C14orf105 (ALO1512)	0.0021 (1.47)	×	×
SAGE1 (NM_018667)	×	0.0012 (2.56)	×
C6orf103 (AL832192)	×	0.0023 (1.20)	×
FLJ13841 (AK023903)	0.0146 (0.35)**	0.0146 (0.49)**	0.0129 (0.47)*
FLJ22655 (BC042888)	×	0.0028 (0.64)	×
FLJ21934 (AY358727)	−0.0127 (0.55)*	−0.0125 (0.49)**	−0.0079 (0.32)*
KIAA1912 (AB067499)	×	−0.0013 (0.65)	×
FLJ40298 (NM_173486)	0.0307 (0.98)***	0.0316 (1.20)**	×
MGC33951 (BC029537)	×	−0.0042 (1.44)	×
NALP4 (AF479747)	−0.0059 (0.56)	−0.0059 (0.76)	−0.0062 (0.23)*
FLJ46154 (AK128035)	0.0185 (0.89)*	0.0185 (0.94)*	0.0141 (1.58)
MGC50372 (BX647272)	5.3676e–04 (1.68)	×	×
LOC285016 (XM_211736)	0.0182 (2.56)	0.0182 (2.25)	0.0147 (2.15)

Superscripts ***, **, * are decodings of significance values.

SUPPLEMENTARY MATERIAL

Supplementary material for "Regularization for Cox's proportional hazards model with NP-dimensionality" (DOI: 10.1214/11-AOS911SUPP; .pdf). In the Supplementary Material [Bradic, Fan and Jiang (2011)] we give additional results of our simulation study, we specify the statements and detailed proofs of technical Lemmas 2.1–2.3 and give complete proofs of Theorems 2.1, 4.1, 4.4–4.6. We present the details of the ICA algorithm of the Section 5 together with new simulation settings were we increased the censoring rate and/or increased the number of significant variables s, and with discussion on the relative estimation efficiency of the penalized methods. We develop results on the growth of the L_2 norm of the score vector $U_n(\boldsymbol{\beta}_1^*)$ and of the matrix $\int_0^\tau \mathbf{V}(\boldsymbol{\beta}_1^*, t)\, d\bar{M}(t)$. Moreover we establish a result on the asymptotic behavior of vector $\hat{\boldsymbol{\beta}}_1^*$ when $s = o(n^{1/3})$ diverging with n. The main tools used are the theory of martingales [Fleming and Harrington (1991)] and the results of various matrix norms of Lemmas 4.1, 4.2 and 2.1–2.3.

REFERENCES

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10** 1100–1120. MR0673646

BERTSEKAS, D. P. (2003). Nonlinear programming. Athena Scientific, Nashua, NH.

BHATIA, R. (1997). *Matrix Analysis. Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

BRADIC, J., FAN, J. and WANG, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73** 325–349. MR2815779

BRADIC, J., FAN, J. and JIANG, J. (2011). Supplement to "Regularization for Cox's proportional hazards model with NP-dimensionality." DOI:10.1214/11-AOS911SUPP.

BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194 (electronic). MR2312149

CAI, J., FAN, J., LI, R. and ZHOU, H. (2005). Variable selection for multivariate failure time data. *Biometrika* **92** 303–316. MR2201361

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

DAUBECHIES, I., DEFRISE, M. and DE MOL, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.* **57** 1413–1457. MR2077704

DAVE, S. S. ET AL. (2004). Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *N. Engl. J. Med.* **351** 2159–2169.

DE LA PEÑA, V. H. (1999). A general class of exponential inequalities for martingales and ratios. *Ann. Probab.* **27** 537–564. MR1681153

DU, P., MA, S. and LIANG, H. (2010). Penalized variable selection procedure for Cox models with semiparametric relative risk. *Ann. Statist.* **38** 2092–2117. MR2676884

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** 74–99. MR1892656

FAN, J. and LV, J. (2011). Non-concave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* **57** 5467–5484.

FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York. MR1100924

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** 1–22.

FRIEDMAN, J., HASTIE, T., HÖFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1** 302–332. MR2415737

JOHNSON, B. A. (2009). On lasso for censored data. *Electron. J. Stat.* **3** 485–506. MR2507457

JUDITSKY, A. B. and NEMIROVSKI, A. S. (2011). Large deviations of vector-valued martingales in 2-smooth normed spaces. *Ann. Appl. Probab.* To appear. Available at arXiv:0809.0813.

KIM, Y., CHOI, H. and OH, H. (2008). Smoothly clipped absolute deviation on high dimensions. *J. Amer. Statist. Assoc.* **103** 1656–1673.

KOLTCHINSKII, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799–828. MR2555200

LV, J. and FAN, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37** 3498–3528. MR2549567

MASSART, P. and MEYNET, C. (2010). An l_1 oracle inequality for the LASSO. Available at arXiv:1007.4791.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TIBSHIRANI, R. (1997). The LASSO method for variable selection in the Cox model. *Stat. Med.* **16** 385–395.

VAN DE GEER, S. (1995). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist.* **23** 1779–1801. MR1370307

VAN DE GEER, S. and BÜHLMANN, P. (2009). On conditions used to prove oracle results for the LASSO. *Electron. J. Stat.* **3** 1360–1392.

WANG, S., NAN, B., ZHOU, N. and ZHU, J. (2009). Hierarchically penalized Cox regression with grouped variables. *Biometrika* **96** 307–322. MR2507145

WU, T. T. and LANGE, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.* **2** 224–244. MR2415601

YUAN, M. and LIN, Y. (2007). On the non-negative garrote estimator. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 143–161. MR2325269

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

J. BRADIC
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, SAN DIEGO
LA JOLLA, CALIFORNIA 92093-0112
USA
E-MAIL: jbradic@math.ucsd.edu

J. FAN
DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
BENDHEIM CENTER FOR FINANCE
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: jqfan@princeton.edu

J. JIANG
DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF NORTH CAROLINA AT CHARLOTTE
CHARLOTTE, NORTH CAROLINA 28223
USA
E-MAIL: jjiang1@uncc.edu