High-Dimensional Sparse Additive Hazards Regression

Wei LIN and Jinchi LV

High-dimensional sparse modeling with censored survival data is of great practical importance, as exemplified by modern applications in high-throughput genomic data analysis and credit risk analysis. In this article, we propose a class of regularization methods for simultaneous variable selection and estimation in the additive hazards model, by combining the nonconcave penalized likelihood approach and the pseudoscore method. In a high-dimensional setting where the dimensionality can grow fast, polynomially or nonpolynomially, with the sample size, we establish the weak oracle property and oracle property under mild, interpretable conditions, thus providing strong performance guarantees for the proposed methodology. Moreover, we show that the regularity conditions required by the L_1 method are substantially relaxed by a certain class of sparsity-inducing concave penalties. As a result, concave penalties such as the smoothly clipped absolute deviation, minimax concave penalty, and smooth integration of counting and absolute deviation can significantly improve on the L_1 method and yield sparser models with better prediction performance. We present a coordinate descent algorithm for efficient implementation and rigorously investigate its convergence properties. The practical use and effectiveness of the proposed methods are demonstrated by simulation studies and a real data example.

KEY WORDS: Empirical process; Oracle property; Regularization; Risk difference; Survival data; Variable selection; Weak oracle property.

1. INTRODUCTION

Advances in experimental technologies in molecular biology during the past decade have brought in a wealth of biomedical data. For instance, DNA microarrays now can be used to measure the expression of tens of thousands of genes in a sample of cells or to identify hundreds of thousands of single nucleotide polymorphisms for an individual at the same time. Data of this kind pose tremendous challenges to effective statistical inference, since the number of features, p, is large compared with the number of observations, n, in which case many classical inference methods can easily fail or become inapplicable. Variable selection, a powerful tool for sparse modeling, is a fundamental task in high-dimensional regression problems, which aims to select only a small set of important variables from a huge number of features, in the hope of alleviating the overfitting problem in high dimensions and improving the predictive power and interpretability of the resulting model. See, for example, Fan and Lv (2010) for a review of recent developments in high-dimensional variable selection.

When clinical data on patient survivals are also available, it would be informative to link high-dimensional biomedical data to survival outcomes. A number of efforts have recently been made in this direction. Regularization methods, which can yield sparse models and hence perform simultaneous variable selection and estimation, are particularly useful and have gained increasing popularity. Several regularization methods originally developed for linear regression have been adapted to survival models. For example, Tibshirani (1997) and Fan and Li (2002) extended the Lasso and nonconcave penalized likelihood, respectively, to the Cox model, while Zhang and Lu (2007) and Zou (2008) developed weighted L_1 methods for the Cox model. Cai et al. (2005) were the first to study regularization methods for survival models in a framework with p growing with n; they demonstrated that the nonconcave penalized pseudopartial likelihood estimator for multivariate failure time data enjoys the oracle property when p grows slowly with n. Antoniadis, Fryzlewicz, and Letué (2010) studied the Dantzig selector for the Cox model in a high-dimensional setting, but did not address the issue of model selection consistency. In addition to their classical applications for survival analysis in public health, survival models have been widely used to model time-to-event data for credit risk analysis in finance and economics (Jarrow 2009; Fan, Lv, and Qi 2011). Identifying important risk factors and quantifying their contributions are crucial aims of these problems.

Variable selection techniques for survival data have also been extended beyond the Cox model. As a useful alternative to the Cox model, the additive hazards model assumes that the hazard function of a failure time T conditional on a p-vector of possibly time-dependent covariates $\mathbf{Z}(\cdot)$ takes the form

$$\lambda(t \mid \mathbf{Z}) = \lambda_0(t) + \boldsymbol{\beta}_0^T \mathbf{Z}(t), \tag{1}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function and $\boldsymbol{\beta}_0$ is a p-vector of regression coefficients (Cox and Oakes 1984, p. 74; Breslow and Day 1987, p. 182; Lin and Ying 1994). The additive models provide a characterization of the regression effects different than the multiplicative models and have some remarkable features that are not shared by the latter. In particular, model (1) pertains to the risk difference, or excess risk, a measure that is especially relevant and informative in epidemiological and clinical studies. Variable selection in model (1) has been studied by a number of authors; for example, Leng and

Wei Lin is Postdoctoral Researcher, Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *weilin1@mail.med.upenn.edu*). Jinchi Lv is Assistant Professor, Information and Operations Management Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089 (E-mail: *jinchilv@marshall.usc.edu*). This research was partially supported by NSF CAREER Award DMS-0955316 and Grant DMS-0806030. The authors sincerely thank the Co-Editor, an Associate Editor, and two referees for their valuable comments that led to substantial improvement of the article.

Journal of the American Statistical Association
March 2013, Vol. 108, No. 501, Theory and Methods
DOI: 10.1080/01621459.2012.746068

Ma (2007) proposed a weighted L_1 approach and Martinussen and Scheike (2009) considered several regularization methods including the Lasso and Dantzig selector.

Despite the aforementioned developments, a rigorous high-dimensional theory that can provide strong performance guarantees for regularization-based variable selection methods in the survival setting is still lacking. Specifically, it is unclear how high dimensionality these methods can handle and what conditions are required for obtaining model selection consistency and nice sampling properties. The need for the development of such a theory is urgent, in view of the recent advances in understanding the performance of regularization methods in the linear regression and classification contexts (e.g., Zhao and Yu 2006; Fan and Fan 2008; Lv and Fan 2009; Wainwright 2009).

In this article, we propose a class of regularization methods for simultaneous variable selection and estimation in model (1), by combining the nonconcave penalized likelihood approach (Fan and Li 2001) and the pseudoscore method (Lin and Ying 1994). To justify the superior performance of the proposed methodology, we consider a high-dimensional setting where the dimension of covariates can grow fast, possibly nonpolynomially, with the sample size. Under mild, interpretable conditions, we establish the weak oracle property (Lv and Fan 2009) and oracle property (Fan and Li 2001) of the proposed regularized estimators. Our high-dimensional analysis is innovative in that it involves empirical process techniques that, to the best of our knowledge, have not been previously used in the survival analysis literature, and provides new insights into the model selection properties of regularization methods for survival data. In particular, we show that the regularity conditions required by the L_1 method are substantially relaxed by a certain class of sparsity-inducing concave penalties, which includes some commonly used concave penalties as special cases (see Section 2.2). Furthermore, we present a coordinate descent algorithm for efficient implementation and rigorously investigate its convergence properties. The practical use and effectiveness of the proposed methodology are demonstrated by both simulated and real data.

In an independent work closely related to this article, Bradic, Fan, and Jiang (2011) studied regularized estimation for variable selection in the Cox model and obtained important oracle-type theoretical results in which the dimension of covariates may grow nonpolynomially with the sample size. Besides model assumptions, a critical difference from our results, however, is that they imposed a *random* condition on a large *empirical* covariance matrix; see their condition 8. Thus, it is natural to ask the question whether the regularized estimators still enjoy the desired properties if a similar condition is imposed on the population version of the matrix. Since the empirical covariance matrix involves the outcome variables, as is generally the case for survival models, a *nonrandom* condition on the *population* covariance matrix seems more natural and will provide more confirmative performance guarantees. Such conditions will also have the benefit that they can be viewed as high-dimensional extensions of the classical asymptotic regularity conditions in the low-dimensional setting, which are imposed on the population covariance matrix. We will provide an affirmative answer to this important question.

The remainder of this article is organized as follows. In Section 2, we propose a class of regularization methods and discuss choices of the penalty function. The theoretical properties of these regularized estimators are studied in Section 3. In Section 4, we describe a coordinate descent algorithm, study its convergence properties, and discuss selection of tuning parameters. Simulation studies and a real data example are presented in Sections 5 and 6, respectively. Some discussion is provided in Section 7, and all proofs and technical details are relegated to the Appendices.

2. REGULARIZATION METHODOLOGY

2.1 Regularized Estimation

We begin with the problem setup. Let T be the failure time and C the censoring time. Denote the censored failure time by $X = T \wedge C$ and the failure indicator by $\Delta = I(T \le C)$, where $I(\cdot)$ is the indicator function. Let $\mathbf{Z}(\cdot) = (Z_1(\cdot), \ldots, Z_p(\cdot))$ be a vector of predictable covariate processes and assume that T and C are conditionally independent given $\mathbf{Z}(\cdot)$. The observed data consist of $(X_i, \Delta_i, \mathbf{Z}_i(\cdot))$, $i = 1, \ldots, n$, which are independent copies of $(X, \Delta, \mathbf{Z}(\cdot))$. We assume that the conditional hazard function is given by model (1).

We adopt the usual counting process notation. Define the observed-failure counting process $N_i(t) = I(X_i \le t, \Delta_i = 1)$, the at-risk indicator $Y_i(t) = I(X_i \ge t)$, and the counting process martingale

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\{\lambda_0(s) + \boldsymbol{\beta}_0^T \mathbf{Z}_i(s)\}\, ds.$$

We will also use $N(t)$, $Y(t)$, and $M(t)$ to denote the generic versions of these processes.

The pseudoscore function of Lin and Ying (1994) is defined as

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \{\mathbf{Z}_i(t) - \overline{\mathbf{Z}}(t)\}\{dN_i(t) - Y_i(t)\boldsymbol{\beta}^T \mathbf{Z}_i(t)\, dt\},$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$, $\overline{\mathbf{Z}}(t) = \sum_{j=1}^n Y_j(t)\mathbf{Z}_j(t) / \sum_{j=1}^n Y_j(t)$, and τ is the maximum follow-up time. This estimating function is linear in $\boldsymbol{\beta}$; through some algebraic manipulation, we can write $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{b} - \mathbf{V}\boldsymbol{\beta}$ with

$$\mathbf{b} = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \{\mathbf{Z}_i(t) - \overline{\mathbf{Z}}(t)\}\, dN_i(t)$$

and

$$\mathbf{V} = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau Y_i(t)\{\mathbf{Z}_i(t) - \overline{\mathbf{Z}}(t)\}^{\otimes 2}\, dt, \tag{2}$$

where $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^T$ for any vector $\mathbf{v}$. Since $\mathbf{V}$ is positive semidefinite, integrating $-\mathbf{U}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ leads to the least-squares-type loss function

$$L(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^T \mathbf{V}\boldsymbol{\beta} - \mathbf{b}^T\boldsymbol{\beta}. \tag{3}$$

Using this loss function for regularized estimation in model (1) has been suggested by a number of authors, including Leng and Ma (2007) and Martinussen and Scheike (2009); the latter authors noted also that $L(\boldsymbol{\beta})$ can be interpreted as an empirical prediction error, up to a constant, for the part of the model orthogonal to the at-risk indicator.

We now define the regularized estimator $\widehat{\boldsymbol{\beta}}$ as a solution to the regularization problem

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ Q(\boldsymbol{\beta}) \equiv L(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \right\}, \tag{4}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $p_\lambda(\theta), \theta \geq 0$, is a penalty function that depends on the regularization parameter $\lambda \geq 0$. When the minimization problem (4) is nonconvex, we will consider a local minimizer, as is common in the literature. It is often convenient to rewrite the penalty function as $p_\lambda(\cdot) = \lambda\rho_\lambda(\cdot)$; we write $\rho_\lambda(\cdot)$ as $\rho(\cdot)$ when it is free of λ. Without the penalty term, $\widehat{\boldsymbol{\beta}}$ reduces to the pseudoscore estimator of Lin and Ying (1994). When the dimensionality is high, however, some form of regularization is needed to guard against overfitting, and the performance of the regularized estimator depends critically on the choice of the penalty function. Thus, in what follows we will first define a general class of penalty functions and discuss several popular choices among the class, and then present some theory to gain further insight into these choices.

2.2 Penalty Function

To answer the question on what kind of penalty functions are ideal for model selection, Fan and Li (2001) advocated penalty functions giving rise to estimators with three desired properties: sparsity, unbiasedness, and continuity. These properties motivate consideration of a class of penalty functions that satisfies the following condition.

Condition 1. The function $\rho_\lambda(\theta)$ is increasing and concave in $\theta \in [0, \infty)$, and has a continuous derivative $\rho'_\lambda(\theta)$ on $(0, \infty)$. In addition, $\rho'_\lambda(\theta)$ is increasing in λ and $\rho'_\lambda(0+) \equiv \rho'(0+) > 0$ is independent of λ.

Some intuition for Condition 1 is as follows. The singularity at the origin encourages sparsity; the concavity assumption aims to reduce the estimation bias; the requirement that $\rho'_\lambda(\theta)$ is increasing in λ allows λ to effectively control the overall strength of the penalty. It should be noted that we do *not* require *strict* concavity or monotonicity, so that a wide range of penalty functions, including those that do not lead to all of the aforementioned three properties, are included in this class, which will facilitate our comparisons among different penalty functions. In the contexts of (generalized) linear models, this class of penalty functions has been studied by Lv and Fan (2009) and Fan and Lv (2011). Of particular interest are the following examples.

- The Lasso (Tibshirani 1996) uses the L_1-penalty, that is, $\rho(\theta) = \theta, \theta \geq 0$.
- The smoothly clipped absolute deviation (SCAD) penalty (Fan 1997; Fan and Li 2001) is given by the derivative

$$\rho'_\lambda(\theta) = I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda), \qquad \theta \geq 0,$$

  where $a > 2$ is a shape parameter. This penalty function takes off at the origin as the L_1-penalty and then gradually levels off until its derivative reaches zero.
- The minimax concave penalty (MCP) proposed by Zhang (2010) has the derivative

$$\rho'_\lambda(\theta) = \frac{(a\lambda - \theta)_+}{a\lambda}, \qquad \theta \geq 0,$$

  where $a > 1$ is a shape parameter. In a similar spirit to SCAD, this penalty function gradually decreases its derivative to zero, except that it drops the L_1 part of SCAD.
- The smooth integration of counting and absolute deviation (SICA) penalty (Lv and Fan 2009) takes the form

$$\rho(\theta) = \frac{(a+1)\theta}{a+\theta}, \qquad \theta \geq 0, \tag{5}$$

  where $a > 0$ is a shape parameter. With a varying from 0 to ∞, this family provides a smooth homotopy between the L_0- and L_1-penalties. Each penalty function starts with slope $1 + a^{-1}$ at the origin, passes through the point (1, 1), and decreases its slope toward zero over the interval $[0, \infty)$.

The L_1-penalty is a convex relaxation of the L_0-penalty and falls at the boundary of the class of penalty functions that satisfies Condition 1. Although the L_1-regularization method enjoys the advantage of computational simplicity, it can suffer from several drawbacks that have motivated a number of improvements. The SCAD penalty was originally proposed to alleviate the bias caused by the L_1 approach, and has been shown to possess the oracle property, that is, the resulting estimator performs asymptotically as well as the oracle estimator, which knew the true sparse model in advance. The estimation bias of the Lasso can also interfere with variable selection; as a result, more stringent conditions such as the irrepresentable condition (Zhao and Yu 2006) are typically required for consistent variable selection. The advantages of concave penalties regarding model selection consistency have recently been revealed and justified by a number of authors. Zhang (2010) showed that the MCP penalty has certain minimax optimality, which enables it to strike a balance between the superior theoretical properties of concave penalties and the computational cost of nonconvex regularization problems. By investigating a nonasymptotic weak oracle property, Lv and Fan (2009) showed that the regularity conditions needed for the L_1 approach can be substantially relaxed by using concave penalties. The SICA family proposed in that article has the remarkable feature that it can be used to define a sequence of regularization problems with varying theoretical performance and computational complexity.

3. THEORETICAL PROPERTIES

Besides the choice of the penalty function, the performance of the regularized estimators depends on a variety of factors, such as the dimensionality of the model, the dependency among the covariates, and the choice of the regularization parameter. To determine how these factors interact with each other and together affect the performance of the proposed estimators, in this section we rigorously develop a high-dimensional theory and discuss some of its implications.

We begin by introducing some notation to be used in our theoretical results. For any vector $\mathbf{v}$, recall that $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^T$ and for

notational convenience, we write $\mathbf{v}^{\otimes 0} = 1$ and $\mathbf{v}^{\otimes 1} = \mathbf{v}$. Define

$$\begin{aligned} \mathbf{s}^{(k)}(t) &= E\{Y(t)\mathbf{Z}(t)^{\otimes k}\}, \qquad k = 0, 1, 2, \\ \mathbf{e}(t) &= \mathbf{s}^{(1)}(t)/s^{(0)}(t), \\ \mathbf{D} &= E\left[\int_0^\tau Y(t)\{\mathbf{Z}(t) - \mathbf{e}(t)\}^{\otimes 2}\, dt\right], \end{aligned} \tag{6}$$

and

$$\mathbf{\Sigma} = E\left[\int_0^\tau \{\mathbf{Z}(t) - \mathbf{e}(t)\}^{\otimes 2}\, dN(t)\right].$$

It is worthwhile to note that $\mathbf{D}$ is the population counterpart of the matrix $\mathbf{V}$ defined in (2), while $\mathbf{\Sigma}$ is the population counterpart of the matrix

$$\mathbf{W} = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \{\mathbf{Z}_i(t) - \overline{\mathbf{Z}}(t)\}^{\otimes 2}\, dN_i(t).$$

These matrices characterize the covariance structure of the model and will play a key role in our high-dimensional analysis.

Furthermore, define the *active set* $A = \{j : \beta_{0j} \neq 0\}$, where β_{0j}, $1 \leq j \leq p$, is the jth component of the true regression coefficient vector $\boldsymbol{\beta}_0$. Let $s = |A|$, that is, the number of nonzero coefficients in $\boldsymbol{\beta}_0$, and we allow the dimension triple (s, n, p) to vary freely. Similarly, define the *estimated active set* $\widehat{A} = \{j : \widehat{\beta}_j \neq 0\}$, where $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^T$. Denote the complement of a set B by B^c. We will use sets to index vectors and matrices; for example, $\boldsymbol{\beta}_{0A}$ is the vector formed by the components β_{0j} of $\boldsymbol{\beta}_0$ with $j \in A$, and $\mathbf{D}_{A^cA}$ is the matrix formed by the entries D_{ij} of $\mathbf{D}$ with $i \in A^c$ and $j \in A$. Define the (half) *minimum signal*

$$d = \frac{1}{2}\min_{j\in A} |\beta_{0j}|.$$

For any $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^T \in \mathbb{R}^q$ with $\theta_j \neq 0$ for all j, following Lv and Fan (2009), define the *local concavity* of the penalty function $\rho_\lambda(\cdot)$ at point $\boldsymbol{\theta}$ as

$$\kappa(\rho_\lambda; \boldsymbol{\theta}) = \lim_{\varepsilon\to 0+} \max_{1\leq j\leq q} \sup_{|\theta_j|-\varepsilon<t_1<t_2<|\theta_j|+\varepsilon} \left\{-\frac{\rho'_\lambda(t_2) - \rho'_\lambda(t_1)}{t_2 - t_1}\right\}.$$

Finally, define

$$\begin{aligned} \kappa_0 &= \sup\{\kappa(\rho_\lambda; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^s, \|\boldsymbol{\theta} - \boldsymbol{\beta}_{0A}\|_\infty \leq d\}, \\ \varphi &= \left\|\mathbf{D}_{AA}^{-1}\right\|_\infty, \end{aligned}$$

and

$$\mu = \Lambda_{\min}(\mathbf{D}_{AA}) - \lambda\kappa_0,$$

where $\Lambda_{\min}(\cdot)$ denotes the minimum eigenvalue. It is important to note that all the quantities defined above can depend on the sample size n, and we have suppressed this dependency for notational simplicity.

3.1 Weak Oracle Property

Lv and Fan (2009) introduced the concept of weak oracle property for comparing different regularization methods. An estimator is said to have the weak oracle property if it is both consistent in model selection and uniformly consistent in estimation. This notion is weaker than the oracle property introduced by Fan and Li (2001) and hence can be satisfied by a broader class of estimators. To derive a nonasymptotic result regarding the weak oracle property of the proposed estimators, we need to impose the following conditions.

Condition 2. (1) $\int_0^\tau \lambda_0(t)\, dt < \infty$. (2) $P\{Y(\tau) = 1\} > 0$. (3) There exist constants $D, K, r > 0$ such that

$$P\left(\sup_{t\in[0,\tau]} |Z_j(t)| > x\right) \leq D\exp(-Kx^r)$$

for all $x > 0$ and $j = 1, \ldots, p$. (4) The sample paths of $Z_j(\cdot)$, $j = 1, \ldots, p$, are of uniformly bounded variation.

Condition 3. There exist constants $\alpha \in (0, 1]$, $\gamma \in [0, 1/2]$, and $c > 0$ such that

$$\left\|\mathbf{D}_{A^cA}\mathbf{D}_{AA}^{-1}\right\|_\infty \leq \left\{(1-\alpha)\frac{\rho'(0+)}{\rho'_\lambda(d)}\right\} \wedge (cn^\gamma).$$

In Condition 2, parts (1) and (2) are standard for survival models; part (3) controls the tail behavior of the covariates and is trivially satisfied for bounded covariates; part (4) is a very mild technical condition that will facilitate entropy calculations.

Condition 3 is an analog of condition (35) in Lv and Fan (2009) for penalized least squares, which is in turn a generalization of condition (15) in Wainwright (2009) for the Lasso. Often for linear regression, such conditions are first imposed on the deterministic Gram matrix, and then a variety of random design matrices such as Gaussian ensembles can be further considered. For survival models such as model (1), however, there is no exact analog of the deterministic Gram matrix; here the matrix $\mathbf{V}$, which plays the same role as the Gram matrix in linear regression, involves the at-risk indicators and hence, is *nondeterministic*. Thus, Condition 3 is imposed on the population version $\mathbf{D}$ of $\mathbf{V}$. Note also that we are not restricted to the cases where the covariates are bounded or Gaussian.

The right-hand side of Condition 3 consists of two parts: the first part is an upper bound that reflects the intrinsic capability of the penalty function for variable selection; the second part is at most $O(\sqrt{n})$, where the parameter γ needs to be determined by other conditions to be presented later. For the L_1-penalty, the first part is bounded by constant 1, which is stringent; for concave penalties, the upper bound is generally relaxed, since concavity implies that $\rho'_\lambda(\theta)$ is decreasing in θ and thus $\rho'(0+)/\rho'_\lambda(d)$ can diverge asymptotically. When signals are fairly strong so that $d \gg \lambda$, the first part imposes no constraint for the SCAD and MCP penalties, since $\rho'_\lambda(d) = 0$ in that case. Also, the upper bound for the SICA penalty can be substantially relaxed by choosing a small value of a.

Since Condition 3 and definitions of φ and μ involve the matrices $\mathbf{D}_{A^cA}\mathbf{D}_{AA}^{-1}$, $\mathbf{D}_{AA}^{-1}$, and $\mathbf{D}_{AA}$, a key step to establishing the weak oracle property is to show that the empirical counterparts of these matrices are close to them in some sense. This intermediate result is provided by the following lemma, which gives explicit probability bounds for similar conditions to hold for the

empirical matrices. In what follows, let Ω_L denote the event that $\max_{j=1}^{p} \sup_{t\in[0,\tau]} |Z_j(t)| \le L$ for $L > 0$.

Lemma 1 (Concentration of empirical matrices). Under Conditions 1–3, if $\mu > 0$ and $\varphi \vee \mu^{-1} = O(\sqrt{n}/s)$, then there exist constants $D, K > 0$ such that

$$P\big(\|\mathbf{V}_{AA}^{-1}\|_\infty \ge 2\varphi \mid \Omega_L\big) \le s^2 D \exp\left\{-K\frac{n}{L^4}\left(\frac{1}{\varphi^2 s^2} \wedge 1\right)\right\}, \tag{7}$$

$$\begin{aligned} P\left[\|\mathbf{V}_{A^cA}\mathbf{V}_{AA}^{-1}\|_\infty \ge \left\{\left(1-\frac{\alpha}{2}\right)\frac{\rho'(0+)}{\rho'_\lambda(d)}\right\} \wedge (2cn^\gamma) \mid \Omega_L\right] \\ \le (p-s)sD\exp\left[-K\frac{n}{L^4}\left\{\frac{\big(\rho'_\lambda(d)^{-1}\wedge n^\gamma\big)^2}{\varphi^2 s^2}\wedge 1\right\}\right] \\ + s^2 D\exp\left\{-K\frac{n}{L^4}\left(\frac{1}{\varphi^2 s^2}\wedge 1\right)\right\}, \end{aligned} \tag{8}$$

and

$$P(\Lambda_{\min}(\mathbf{V}_{AA}) \le \lambda\kappa_0 \mid \Omega_L) \le s^2 D\exp\left\{-K\frac{n}{L^4}\left(\frac{\mu^2}{s^2}\wedge 1\right)\right\}. \tag{9}$$

Inequalities (7) and (8) show that there would not be much difference if we had defined the quantity φ or imposed Condition 3 on the empirical matrices. The eigenvalue condition $\Lambda_{\min}(\mathbf{V}_{AA}) > \lambda\kappa_0$ is needed for identification of a strict local minimizer of problem (4); inequality (9) says that this condition holds with high probability if $\Lambda_{\min}(\mathbf{D}_{AA})$ and $\lambda\kappa_0$ have a positive gap μ that does not shrink to zero too fast.

We now state our main theoretical result regarding the weak oracle property of the proposed estimators.

Theorem 1 (Weak oracle property). In addition to Conditions 1–3, assume that the following conditions hold:

$$\frac{n\big(\rho'_\lambda(d)^{-1}\wedge n^\gamma\big)^2}{\varphi^2 s^2(\log p)^{r_1}} \to \infty, \qquad \frac{n(\varphi^{-1}\wedge\mu)^2}{s^2(\log s)^{r_1}} \to \infty, \tag{10}$$

$$\frac{n\lambda^2}{(\log p)^{r_1}} \to \infty, \qquad \frac{n^{1-2\gamma}\lambda^2}{(\log s)^{r_1}} \to \infty, \tag{11}$$

and

$$d \ge c_1\varphi\lambda\rho'(0+), \tag{12}$$

where $\mu > 0$, $r_1 = (r+4)/r$, and $c_1 = 2 + 1/(4c)$. Then, for some constants $D, K > 0$, with probability at least

$$\begin{aligned} 1 - D\exp\left[-Kn^{1/r_1}\left\{\frac{(\varphi^{-1}\wedge\mu)^2}{s^2}\wedge 1\right\}^{1/r_1}\right] \\ -D\exp\left\{-Kn^{1/r_1}\left(\frac{\lambda^2}{n^{2\gamma}}\wedge 1\right)^{1/r_1}\right\} \to 1, \end{aligned}$$

there exists a regularized estimator $\widehat{\boldsymbol{\beta}}$ that satisfies the following properties:

(a) (Sparsity) $\widehat{\boldsymbol{\beta}}_{A^c} = \mathbf{0}$.
(b) (L_∞-loss) $\|\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_\infty \le c_1\varphi\lambda\rho'(0+)$.

To develop intuition for the two conditions in (10), we consider some simplified cases. First, concavity implies that $\rho'_\lambda(d) \le \rho'(0+)$; thus, a sufficient condition for the first condition in (10) to hold is

$$\frac{n}{\varphi^2 s^2(\log p)^{r_1}} \to \infty. \tag{13}$$

Consider the second condition in (10) and recall that $\mu = \Lambda_{\min}(\mathbf{D}_{AA}) - \lambda\kappa_0$. For the L_1-penalty, $\kappa_0 = 0$; for SCAD and MCP, $\lambda\kappa_0 = (a-1)^{-1}$ and a^{-1}, respectively. Thus, for this condition to hold for these penalties, it suffices to assume that $\Lambda_{\min}(\mathbf{D}_{AA})$ is bounded away from zero and that

$$\frac{n}{\varphi^2 s^2(\log s)^{r_1}} \to \infty,$$

where the latter is implied by (13). Therefore, conditions in (10) are primarily constraints on the growth rates of the model dimensions p and s and certain matrix norms of $\mathbf{D}_{AA}^{-1}$. On the other hand, if we assume, for simplicity, that φ is constant, then (13) gives a lower bound for the number of observations that are needed for guaranteed sparse recovery, $n \gg s^2(\log p)^{r_1}$. This is an interesting setting, since it shows that the proposed estimators can handle a nonpolynomially growing dimension of covariates as high as $\log p = o(n^{1/r_1})$, while the dimension of the true sparse model grows as $s = o(\sqrt{n})$. In particular, for bounded covariates, we can take $r_1 = 1$ by letting $r \to \infty$ and thus allow $\log p = o(n)$.

For simplicity, consider the case of bounded covariates, that is, $r_1 = 1$. The two conditions in (11) give a lower bound for the regularization parameter λ,

$$\lambda \gg \sqrt{\frac{\log p}{n}} \vee \sqrt{\frac{\log s}{n^{1-2\gamma}}}.$$

Thus, in view of (12), we see that λ should be chosen to satisfy

$$\sqrt{\frac{\log p}{n}} \vee \sqrt{\frac{\log s}{n^{1-2\gamma}}} \ll \lambda \le \frac{d}{c_1\varphi\rho'(0+)}.$$

For such choices of λ to exist, the minimum signal d must satisfy

$$d \gg \varphi\left(\sqrt{\frac{\log p}{n}} \vee \sqrt{\frac{\log s}{n^{1-2\gamma}}}\right). \tag{14}$$

Recall that $\gamma \in [0, 1/2]$ has appeared in Condition 3. More insight can be gained by comparing the two parts on the right-hand side of (14): the first part will dominate if $n^\gamma \ll \sqrt{(\log p)/(\log s)}$, and in this case, Theorem 1 guarantees recovery of signals that satisfy $d \gg \varphi\sqrt{(\log p)/n}$, independent of γ; otherwise, the second part will dominate, and the weakest recoverable signal will depend on the correlation between the two sets of true variables and noise variables as reflected by the value of γ. Of course, for the L_1-penalty, since the first part in Condition 3 always dominates the second part, we can simply take $\gamma = 0$, and thus, the value of γ plays no role in determining the lower bound for d.

Despite the similarities between the results presented here and those for (generalized) linear models in Lv and Fan (2009) and Fan and Lv (2011), it is worthwhile to note some important differences. The restriction on the correlation structure described in Condition 3 has a more complex form, in that the matrix $\mathbf{D}$ involves not only the covariates but also the failure process and censoring mechanism. This complexity arises from the semiparametric nature of survival models and the presence

of censoring, and as a consequence, the ability of performing variable selection is affected by the interplay of several factors, including the covariates, baseline hazard, and censoring mechanism. Moreover, although our results allow the dimensions to grow at rates comparable to those available for linear regression models, our proofs suggest that the necessary sample size for observing the effects of these growth rates, which is determined by the constants, may be significantly larger. In addition, there is a rate loss in characterizing the convergence of the matrix $\mathbf{V}$ to its population counterpart $\mathbf{D}$. These facts call for a need to increase the sample size for making reliable inference.

3.2 Oracle Property

In addition to model selection consistency, the oracle property requires the regularized estimator to be asymptotically as efficient as the oracle estimator with the true sparse model known a priori. For this purpose, some extra eigenvalue conditions are needed. Define $\Lambda_1 = \Lambda_{\min}(\mathbf{D}_{AA})$, $\Lambda_2 = \Lambda_{\min}(\mathbf{\Sigma}_{AA})$, and $\Lambda_3 = \Lambda_{\min}(\mathbf{D}_{AA}^{-1}\mathbf{\Sigma}_{AA}\mathbf{D}_{AA}^{-1})$. The oracle property of the proposed regularized estimators is stated in the following result.

Theorem 2 (Oracle property). Assume that all conditions of Theorem 1 hold. In addition, assume that

$$\frac{n\Lambda_1^2}{s^2(\log s)^{r_1}} \to \infty, \quad \frac{n\Lambda_2^2}{s^2} \to \infty, \quad \frac{n\Lambda_1^4\Lambda_3}{s^3} \to \infty, \quad (15)$$

and

$$\frac{ns\lambda^2\rho'_\lambda(d)^2}{\Lambda_1^2\Lambda_3} \to 0, \quad (16)$$

where $r_1 = (r+4)/r$. Then, for some constants D, $K > 0$, with probability at least

$$1 - D\exp\left[-Kn^{1/r_1}\left\{\frac{(\varphi^{-1} \wedge \mu \wedge \Lambda_1)^2}{s^2} \wedge 1\right\}^{1/r_1}\right] - D\exp\left\{-Kn^{1/r_1}\left(\frac{\lambda^2}{n^{2\gamma}} \wedge 1\right)^{1/r_1}\right\} \to 1,$$

there exists a regularized estimator $\widehat{\boldsymbol{\beta}}$ that satisfies the following properties:

(a) (Sparsity) $\widehat{\boldsymbol{\beta}}_{A^c} = \mathbf{0}$.
(b) (Asymptotic normality) For every $\mathbf{u} \in \mathbb{R}^s$ with $\|\mathbf{u}\|_2 = 1$,

$$\sqrt{n}\mathbf{u}^T\mathbf{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}) \xrightarrow{\mathcal{D}} N(0,1).$$

The three conditions in (15) relate the sparsity dimension s to eigenvalue bounds for the matrices $\mathbf{D}_{AA}$, $\mathbf{\Sigma}_{AA}$, and $\mathbf{D}_{AA}^{-1}\mathbf{\Sigma}_{AA}\mathbf{D}_{AA}^{-1}$. If we consider the special case where the eigenvalues of these matrices are all bounded away from zero, then these conditions are trivially satisfied for $s = o(n^{1/3})$. The form of (15), however, deals with much more general situations and is very meaningful in that the eigenvalue bounds are closely related to the difficulty of the estimation problem.

In the context of linear regression, it is well known that the L_1-penalty does not have the oracle property (Zou 2006; Wainwright 2009). It is clear from (16) that this is also the case for the problem considered here. To see this, consider the case where Λ_1 and Λ_3 are fixed, and note that $\rho'(d) \equiv 1$ for the L_1-penalty. Conditions in (11) imply that $n\lambda^2 \to \infty$, and hence, (16) cannot hold. For the SCAD and MCP penalties, (16) is trivially satisfied for $d \gg \lambda$, since $\rho'_\lambda(d) = 0$ in that case. For the SICA penalty, we have $\rho'(d) = a(a+1)/(a+d)^2$; thus, to obtain the oracle property, we should take $a \to 0+$ at a rate such that $d \gg a$ and $\sqrt{ns}\lambda a/d^2 \to 0$. This result is reasonable since the SICA penalty approaches the L_0-penalty as $a \to 0+$.

4. IMPLEMENTATION

In this section, we describe an efficient coordinate descent algorithm for the implementation of the proposed methodology, analyze its convergence properties, and discuss the selection of tuning parameters.

4.1 Coordinate Descent Algorithm

The idea of coordinate optimization for penalized least-squares problems was proposed by Fu (1998) and Daubechies, Defrise, and De Mol (2004), and was demonstrated by Friedman et al. (2007) and Wu and Lange (2008) to be exceptionally efficient for large-scale sparse problems. Recently, a number of authors, including Fan and Lv (2011), Breheny and Huang (2011), and Mazumder, Friedman, and Hastie (2011), generalized this idea to regularized regression with concave penalties and showed that it is an attractive alternative to earlier proposals such as the local quadratic approximation (Fan and Li 2001) and local linear approximation (Zou and Li 2008).

To balance the regularization strengths on different components of $\boldsymbol{\beta}$, we minimize the weighted version of the objective function,

$$\widetilde{Q}(\boldsymbol{\beta};\lambda) \equiv L(\boldsymbol{\beta}) + \sum_{j=1}^{p} V_{jj}\, p_\lambda(|\beta_j|),$$

where, for simplicity, we assume that $V_{jj} \neq 0$ for all j. The coordinate descent method minimizes the objective function in one coordinate at a time and cycles through all coordinates until convergence. To produce a solution path over a grid of values of the regularization parameter λ, at each grid point the solution from a neighboring point is used as a warm start to accelerate convergence and, for concave penalties, to avoid suboptimal local solutions. Pick $\lambda_{\max}$ sufficiently large such that $\widehat{\boldsymbol{\beta}} = \mathbf{0}$ is a (local) solution and take a decreasing sequence $(\lambda_1, \ldots, \lambda_K)$ with $\lambda_1 = \lambda_{\max}$; for the Lasso, SCAD, MCP, and SICA, one can take $\lambda_{\max} = \max_{j=1}^{p}(|b_j|/V_{jj})$, where b_j is the jth component of $\mathbf{b}$. The following algorithm produces a solution path $\widehat{\boldsymbol{\beta}}^{\lambda_k}$ over a grid of points λ_k, $k = 1, \ldots, K$.

Coordinate Descent Algorithm.

1. Initialize $\widehat{\boldsymbol{\beta}} = \mathbf{0}$ and set $k = 1$.
2. Cyclically for $j = 1, \ldots, p$, update the jth component $\widehat{\beta}_j$ of $\widehat{\boldsymbol{\beta}}$ by the univariate (global) minimizer of $\widetilde{Q}(\widehat{\boldsymbol{\beta}};\lambda_k)$ with respect to $\widehat{\beta}_j$ until convergence.
3. Set $\widehat{\boldsymbol{\beta}}^{\lambda_k} = \widehat{\boldsymbol{\beta}}$ and $k \leftarrow k+1$.
4. Repeat Steps 2 and 3 until $k = K+1$.

The above algorithm requires efficiently solving a univariate regularization problem in Step 2. Since $L(\boldsymbol{\beta})$ is quadratic, closed-form solutions to this univariate problem exist for commonly used penalty functions, including the Lasso, SCAD,

MCP, and SICA. The solutions for the first three penalties have been given in the aforementioned references, and we provide the solution for the SICA penalty in Appendix B. Coordinate descent may be slow if the resulting model is not sparse; hence, to save computation time, one can set a level of sparsity for early stopping if only models up to a certain size are desired.

4.2 Convergence Analysis

Denote by $\boldsymbol{\beta}^m = (\beta_1^m, \ldots, \beta_p^m)^T$ the mth update vector generated by coordinate descent, that is,

$$\beta_j^m = \arg\min_{\theta \in \mathbb{R}} \widetilde{Q}\big(\beta_1^m, \ldots, \beta_{j-1}^m, \theta, \beta_{j+1}^{m-1}, \ldots, \beta_p^{m-1}\big)$$

for $j = 1, \ldots, p$ and $m = 1, 2, \ldots$, where $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_p^0)^T$ is the starting point and we have suppressed the dependence of $\widetilde{Q}(\boldsymbol{\beta})$ on λ. Define the *maximum concavity* of the penalty function $p_\lambda(\cdot)$ by

$$\kappa(p_\lambda) = \sup_{0<t_1<t_2<\infty} \left\{ -\frac{p_\lambda'(t_2) - p_\lambda'(t_1)}{t_2 - t_1} \right\}.$$

For the L_1-penalty, SCAD, MCP, and SICA, we have $\kappa(p_\lambda) = 0$, $(a-1)^{-1}$, a^{-1}, and $2\lambda(a^{-1} + a^{-2})$, respectively. Due to the concavity of $p_\lambda(\cdot)$, the objective function $\widetilde{Q}(\boldsymbol{\beta})$ is generally nonconvex. However, under certain conditions such that the concavity of $p_\lambda(\cdot)$ is dominated by the convexity of the quadratic loss function $L(\boldsymbol{\beta})$ componentwise (subject to normalization by V_{jj}), $\widetilde{Q}(\boldsymbol{\beta})$ can still be strictly convex along each coordinate. This key observation leads to the following convergence result.

Theorem 3 (Convergence of coordinate descent). Under Condition 1, the sequence $\{\boldsymbol{\beta}^m\}$ generated by the coordinate descent algorithm is bounded. Moreover, the following statements hold:

(a) If the penalty function $p_\lambda(\cdot)$ satisfies $\kappa(p_\lambda) < 1$, then every cluster point of $\{\boldsymbol{\beta}^m\}$ is a local minimizer of $\widetilde{Q}(\boldsymbol{\beta})$.
(b) If in addition, the sequence $\{\boldsymbol{\beta}^m\}$ eventually lies in a compact neighborhood $\mathcal{K}$ of $\boldsymbol{\beta}^*$ such that $\boldsymbol{\beta}^*$ is the unique local minimizer of $\widetilde{Q}(\boldsymbol{\beta})$ in $\mathcal{K}$, then the sequence $\{\boldsymbol{\beta}^m\}$ converges to $\boldsymbol{\beta}^*$.

Note that the condition $\kappa(p_\lambda) < 1$ is always satisfied by the L_1-penalty, SCAD ($a > 2$), and MCP ($a > 1$). For the SICA penalty, the condition is satisfied if $\lambda(a^{-1} + a^{-2}) < 1/2$. This latter condition suggests that no matter what value λ takes, one can adjust the value of a to ensure computational stability. Since the optimal λ is often small, one can set a to a small value to achieve good performance of the SICA method.

It is often useful to determine when a local minimizer $\widehat{\boldsymbol{\beta}}$ of $\widetilde{Q}(\boldsymbol{\beta})$ is also a global minimizer and when the sequence $\{\boldsymbol{\beta}^m\}$ converges to it. The (restricted) global optimality of nonconcave penalized likelihood estimators was characterized by Fan and Lv (2011). The following result gives sufficient conditions for the (restricted) global optimality of $\widehat{\boldsymbol{\beta}}$ on some subspace of $\mathbb{R}^p$ and the convergence of $\{\boldsymbol{\beta}^m\}$ to the global optimum.

Theorem 4 (Restricted global optimality). Let $\widehat{\boldsymbol{\beta}}$ be a local minimizer of $\widetilde{Q}(\boldsymbol{\beta})$. For any subset S of $\{1, \ldots, p\}$, denote by $\mathcal{B}_S$ the $|S|$-dimensional subspace $\{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_j = 0, j \notin S\}$. Under Condition 1, the following statements hold:

(a) If $\widehat{\boldsymbol{\beta}}$ lies in $\mathcal{B}_S$ and $\Lambda_{\min}(\mathbf{V}_{SS}) \geq \kappa(p_\lambda) \max_{j \in S} V_{jj}$, then $\widehat{\boldsymbol{\beta}}$ is a global minimizer of $\widetilde{Q}(\boldsymbol{\beta})$ in $\mathcal{B}_S$.
(b) If the sequence $\{\boldsymbol{\beta}^m\}$ generated by the coordinate descent algorithm eventually lies in $\mathcal{B}_S$ and $\Lambda_{\min}(\mathbf{V}_{SS}) > \kappa(p_\lambda) \max_{j \in S} V_{jj}$, then $\widetilde{Q}(\boldsymbol{\beta})$ has a unique global minimizer $\boldsymbol{\beta}^*$ in $\mathcal{B}_S$ and the sequence $\{\boldsymbol{\beta}^m\}$ converges to $\boldsymbol{\beta}^*$.

The condition in part (a) is trivially satisfied for the L_1-penalty. For the SCAD, MCP, and SICA, the condition can be satisfied with some S if the correlation among covariates is not too strong and the concavity of the penalty function is not too large. Following Fan and Lv (2011), under some mild regularity conditions we can further establish the global optimality of $\widehat{\boldsymbol{\beta}}$ on the union of all $|S|$-dimensional coordinate subspaces of $\mathbb{R}^p$. Although the global optimality is somewhat hard to guarantee for concave penalties, we remark that it is not required for achieving practically good performance of the regularized estimator. In fact, our theoretical arguments suggest that a sparse local solution should possess nice statistical properties, while the coordinate descent algorithm is likely to find such a solution by carefully following a solution path from the sparsest end.

4.3 Selection of Tuning Parameters

After a solution path has been produced, the optimal regularization parameter λ can be chosen by M-fold cross-validation. The cross-validation score is defined as

$$\mathrm{CV}(\lambda) = \frac{1}{M} \sum_{m=1}^{M} L^{(m)}\big(\widehat{\boldsymbol{\beta}}^{(-m)}(\lambda)\big),$$

where $L^{(m)}(\cdot)$ is the loss function given in (3) computed from the mth part of the data, and $\widehat{\boldsymbol{\beta}}^{(-m)}(\lambda)$ is the estimate from the data with the mth part removed. The concave penalties have one additional parameter a to be tuned. For the SCAD penalty, $a = 3.7$ was suggested by Fan and Li (2001) from a Bayesian perspective and is commonly used in the literature. Since the MCP is similar to SCAD, we also take $a = 3.7$. The choice of a for the SICA penalty requires a little more caution, since a small a that is often needed to yield a superior theoretical performance can sometimes cause the problem of computational instability. This can be clearly seen from the discussion following Theorem 3; a too small a would cause the convexity condition $\lambda(a^{-1} + a^{-2}) < 1/2$ to be violated. To solve this dilemma, we first compute a pilot solution path with a larger a suggested by the convexity condition, and then set a to a smaller value and recompute the solution path by taking the pilot solutions as warm starts. If needed, the above process can be repeated several times to gain further stability. Often we take $a = 1$ for the pilot solution and one or a few values toward zero for recomputing the solution. We find that, in the settings we have considered, the practical performance of our methods is not sensitive to these choices. In the more difficult settings and if computing resources allow, a fine tuning on a may yield additional benefits; see Mazumder, Friedman, and Hastie (2011) for more discussion on producing a solution surface along (λ, a).

5. SIMULATION STUDIES

We conducted three simulation studies to evaluate the finite-sample performance of the proposed regularization methods

Table 1. Results for various methods in the first simulation study with $n = 200$, $s = 6$, and censoring rate about 25%. Values shown are means (standard deviations) of each performance measure over 100 replicates

Setting	Method	PE1	PE2	L_2-loss	L_1-loss	#S	#FN
$p = 50$	Lasso	0.191 (0.055)	22.73 (6.38)	1.061 (0.311)	3.341 (0.838)	19.6 (4.5)	0.0 (0.2)
$\rho = 0.1$	SCAD	0.135 (0.044)	10.90 (5.24)	0.498 (0.250)	1.266 (0.658)	10.7 (2.5)	0.0 (0.2)
	MCP	0.138 (0.047)	11.34 (5.83)	0.518 (0.278)	1.271 (0.716)	8.9 (2.1)	0.0 (0.3)
	SICA	0.130 (0.048)	10.30 (5.77)	0.471 (0.273)	1.015 (0.665)	6.2 (1.0)	0.1 (0.4)
	Enet	0.192 (0.055)	22.75 (6.32)	1.062 (0.308)	3.367 (0.896)	19.9 (4.9)	0.0 (0.1)
	Oracle	0.121 (0.030)	9.26 (3.62)	0.424 (0.172)	0.894 (0.398)	6.0 (0.0)	0.0 (0.0)
$p = 50$	Lasso	0.199 (0.052)	18.33 (4.58)	1.078 (0.310)	3.504 (0.916)	21.2 (5.3)	0.1 (0.5)
$\rho = 0.5$	SCAD	0.134 (0.071)	9.98 (5.63)	0.522 (0.354)	1.370 (1.089)	10.7 (3.0)	0.0 (0.2)
	MCP	0.130 (0.057)	9.92 (5.77)	0.522 (0.374)	1.295 (1.044)	8.6 (2.4)	0.1 (0.7)
	SICA	0.121 (0.040)	8.87 (4.17)	0.461 (0.246)	1.058 (0.788)	6.9 (2.8)	0.0 (0.0)
	Enet	0.199 (0.052)	18.32 (4.44)	1.078 (0.300)	3.548 (1.049)	21.8 (5.9)	0.0 (0.0)
	Oracle	0.108 (0.019)	7.56 (2.96)	0.398 (0.187)	0.850 (0.452)	6.0 (0.0)	0.0 (0.0)
$p = 100$	Lasso	0.297 (0.060)	25.44 (5.31)	1.289 (0.284)	4.312 (0.932)	25.2 (6.2)	0.0 (0.1)
$\rho = 0.1$	SCAD	0.253 (0.054)	11.72 (5.01)	0.558 (0.257)	1.601 (0.740)	15.1 (4.0)	0.0 (0.2)
	MCP	0.253 (0.059)	11.83 (5.39)	0.563 (0.275)	1.507 (0.801)	11.5 (3.1)	0.0 (0.3)
	SICA	0.236 (0.062)	10.13 (5.44)	0.483 (0.270)	1.070 (0.751)	6.7 (2.5)	0.0 (0.2)
	Enet	0.300 (0.066)	25.70 (5.79)	1.302 (0.307)	4.403 (1.137)	26.0 (7.7)	0.0 (0.1)
	Oracle	0.219 (0.041)	8.77 (3.90)	0.423 (0.203)	0.892 (0.472)	6.0 (0.0)	0.0 (0.0)
$p = 100$	Lasso	0.275 (0.057)	22.98 (4.82)	1.379 (0.334)	4.716 (0.931)	27.5 (7.0)	0.2 (0.8)
$\rho = 0.5$	SCAD	0.207 (0.053)	11.24 (5.87)	0.589 (0.393)	1.697 (1.055)	15.3 (4.2)	0.1 (0.8)
	MCP	0.209 (0.054)	11.48 (6.29)	0.606 (0.420)	1.643 (1.115)	11.5 (3.1)	0.2 (1.0)
	SICA	0.198 (0.071)	10.03 (6.75)	0.542 (0.417)	1.264 (1.151)	6.7 (2.6)	0.2 (0.8)
	Enet	0.275 (0.058)	23.07 (4.96)	1.385 (0.343)	4.733 (0.974)	27.6 (7.1)	0.2 (0.8)
	Oracle	0.169 (0.023)	7.42 (2.71)	0.394 (0.179)	0.838 (0.427)	6.0 (0.0)	0.0 (0.0)

with the Lasso, SCAD, MCP, and SICA penalties, and compare them with the oracle estimator, which knew the true sparse model in advance. Since the elastic net (Enet; Zou and Hastie 2005) is also a common choice for the penalty function and is known to outperform the Lasso when the important variables are highly correlated, we also include it in our comparisons.

The first simulation study aims to examine the case where p is comparable to but smaller than n. We generated data from the model

$$\lambda(t \mid \mathbf{Z}) = 1 + \boldsymbol{\beta}_0^T \mathbf{Z},$$

where $\mathbf{Z}$ has a multivariate normal distribution with mean zero and covariance matrix $(\rho^{|i-j|})_{i,j=1}^p$ and subject to $\boldsymbol{\beta}_0^T \mathbf{Z} > -1$, and $\boldsymbol{\beta}_0 = (\mathbf{v}^T, \ldots, \mathbf{v}^T, \mathbf{0})^T$ with the pattern $\mathbf{v} = (1, 0, -1, 0, 0, 0)^T$ repeated q times. We set $\rho = 0.1$ and 0.5, and $q = 3$ so that the sparsity dimension $s = 6$. We considered $p = 50$ and 100, and $n = 200$. The censoring time C has a uniform distribution $U(0, c_0)$ with c_0 chosen to obtain a censoring rate about 25%.

We first choose the optimal regularization parameter λ by ten-fold cross-validation and evaluate the performance of the resulting estimators by six measures. The first two measures quantify prediction performance: PE1 is the prediction error based on the loss function $L(\cdot)$ (up to a constant), and PE2 is the L_2 prediction error in the excess risk, $\|\mathbf{Z}^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2$, both computed from an independent test sample of size 500. For estimation accuracy, we report the L_2-loss $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ and L_1-loss $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1$. The other two measures pertain to model selection consistency: #S and #FN refer to the number of selected variables and the number of incorrectly excluded variables (false negatives), respectively. The means and standard deviations of each measure over 100 replicates are summarized in Table 1.

From Table 1, we see that the Lasso and Enet had very close performance in this setting and the concave penalties outperformed the Lasso and Enet in that they selected a sparser model with better prediction and estimation performance. As expected from their similarity, the SCAD and MCP had comparable performance, of which the latter selected a slightly sparser model than the former. The SICA performed best among the five methods, with a performance very close to that of the oracle estimator; in most cases, it was able to identify exactly the six important variables.

It has been noted that prediction-based criteria such as cross-validation tend to include too many irrelevant variables in Lasso-type procedures (Meinshausen and Bühlmann 2006; Leng, Lin, and Wahba 2006), which was also observed in our simulations. Thus, it is natural to wonder whether the concave penalties really improve the variable selection performance over the Lasso and Enet, if the most appropriate amount of regularization for each method can be selected. In other words, it is sensible to compare the performance of the best model that ever exists on the solution path, assuming that we were assisted by an oracle, which knew the true sparse model. Since our goal is to identify a parsimonious model that includes as many relevant variables as possible, we recorded the maximum number of correctly selected variables among all models up to a certain size on the solution path and averaged it over all replicates. By definition, this measure is an increasing function of the (maximum) model size and characterizes the best possible performance that can be achieved by any model selection criterion.

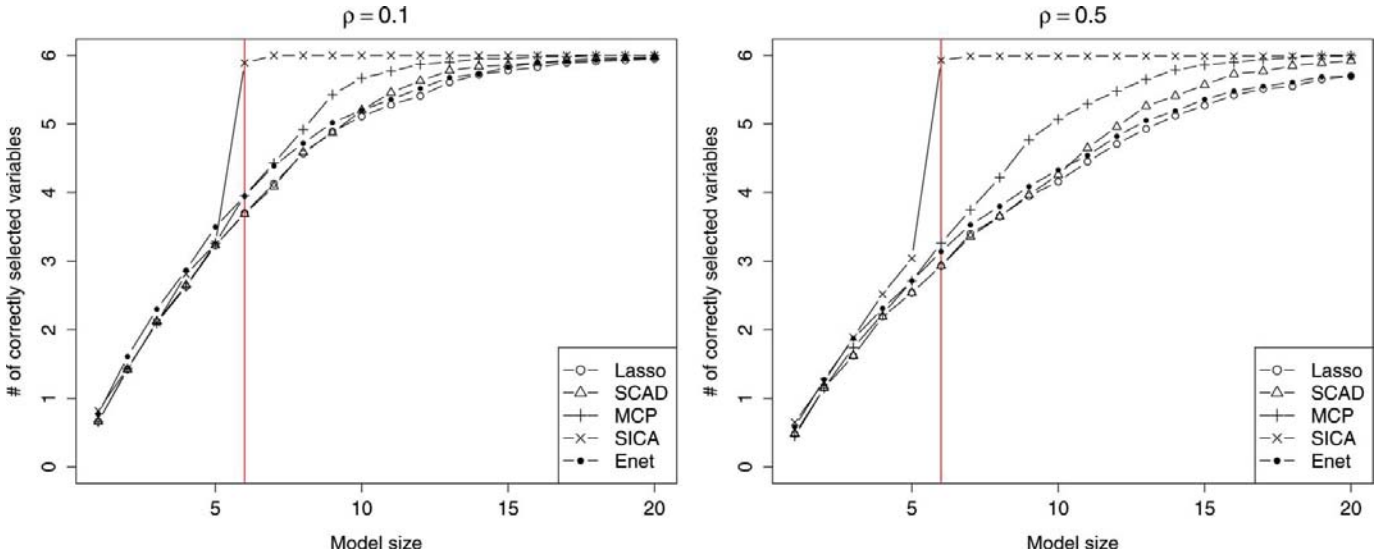

Figure 1. Variable selection performance for various methods in the first simulation study with $p = 100$. The vertical line indicates the true sparsity dimension.

The results on the above-defined performance measure for the first simulation study are shown in Figure 1. It is clear from the figure that the SICA outperformed the other methods in that it had a high chance to identify all six important variables immediately after the model size reached the true sparsity dimension. The MCP also exhibited a performance boost above the true sparsity level, while the boost for SCAD was more subtle and occurred at a later stage. These trends are magnified when the correlation among covariates increases. The Enet improved on the performance of the Lasso only marginally in this case.

In the second simulation study, we compare the performance of various methods in the setting where p is much larger than n. We used the same model setup as before, except that $p = 2000$ and 5000, and $n = 500$. The performance measures over 50 replicates with ten-fold cross-validation to choose the optimal λ are shown in Table 2, which indicates the same trends as in

Table 2. Results for various methods in the second simulation study with $n = 500$, $s = 6$, and censoring rate about 25%. Values shown are means (standard deviations) of each performance measure over 50 replicates

Setting	Method	PE1	PE2	L_2-loss	L_1-loss	#S	#FN
$p = 2000$	Lasso	10.673 (0.038)	27.43 (3.99)	1.396 (0.208)	5.036 (0.780)	58.4 (14.8)	0.0 (0.0)
$\rho = 0.1$	SCAD	10.558 (0.017)	6.48 (1.95)	0.301 (0.096)	0.972 (0.435)	21.0 (9.3)	0.0 (0.0)
	MCP	10.556 (0.020)	6.29 (2.56)	0.293 (0.123)	0.772 (0.492)	11.5 (4.9)	0.0 (0.0)
	SICA	10.551 (0.014)	5.64 (2.48)	0.265 (0.120)	0.582 (0.376)	6.7 (3.6)	0.0 (0.0)
	Enet	10.673 (0.038)	27.43 (3.99)	1.396 (0.208)	5.036 (0.780)	58.4 (14.8)	0.0 (0.0)
	Oracle	10.548 (0.010)	5.16 (1.87)	0.244 (0.097)	0.511 (0.223)	6.0 (0.0)	0.0 (0.0)
$p = 2000$	Lasso	11.913 (0.051)	27.24 (3.85)	1.638 (0.250)	5.809 (0.747)	65.4 (18.2)	0.1 (0.3)
$\rho = 0.5$	SCAD	11.743 (0.022)	7.30 (2.21)	0.362 (0.117)	1.409 (0.498)	34.3 (12.8)	0.0 (0.0)
	MCP	11.738 (0.019)	6.35 (2.19)	0.322 (0.119)	0.956 (0.408)	15.9 (5.9)	0.0 (0.0)
	SICA	11.732 (0.022)	5.19 (2.32)	0.274 (0.123)	0.609 (0.339)	6.8 (4.0)	0.0 (0.0)
	Enet	11.913 (0.051)	27.24 (3.85)	1.638 (0.250)	5.809 (0.747)	65.4 (18.2)	0.1 (0.3)
	Oracle	11.728 (0.009)	4.92 (1.79)	0.264 (0.108)	0.558 (0.238)	6.0 (0.0)	0.0 (0.0)
$p = 5000$	Lasso	11.320 (0.047)	30.62 (4.21)	1.610 (0.223)	5.559 (0.790)	59.2 (18.2)	0.0 (0.1)
$\rho = 0.1$	SCAD	11.138 (0.016)	8.23 (2.36)	0.391 (0.122)	1.611 (0.570)	38.2 (14.5)	0.0 (0.0)
	MCP	11.132 (0.017)	7.23 (2.44)	0.348 (0.126)	1.029 (0.435)	16.5 (7.1)	0.0 (0.0)
	SICA	11.123 (0.011)	6.06 (2.52)	0.298 (0.133)	0.648 (0.351)	6.2 (1.0)	0.0 (0.0)
	Enet	11.321 (0.048)	30.62 (4.22)	1.610 (0.223)	5.571 (0.824)	59.4 (18.6)	0.0 (0.1)
	Oracle	11.122 (0.010)	5.78 (2.25)	0.283 (0.121)	0.608 (0.291)	6.0 (0.0)	0.0 (0.0)
$p = 5000$	Lasso	10.364 (0.053)	31.81 (4.30)	1.997 (0.293)	6.390 (0.553)	56.9 (28.7)	1.1 (1.8)
$\rho = 0.5$	SCAD	10.120 (0.018)	9.24 (2.39)	0.464 (0.151)	2.424 (0.839)	64.2 (20.1)	0.0 (0.0)
	MCP	10.111 (0.019)	7.69 (2.38)	0.388 (0.136)	1.471 (0.643)	27.8 (10.8)	0.0 (0.0)
	SICA	10.096 (0.010)	4.91 (1.90)	0.267 (0.115)	0.573 (0.272)	6.1 (0.2)	0.0 (0.0)
	Enet	10.364 (0.053)	31.81 (4.30)	1.997 (0.293)	6.387 (0.532)	56.9 (28.0)	1.0 (1.7)
	Oracle	10.094 (0.007)	4.65 (1.65)	0.255 (0.110)	0.552 (0.270)	6.0 (0.0)	0.0 (0.0)

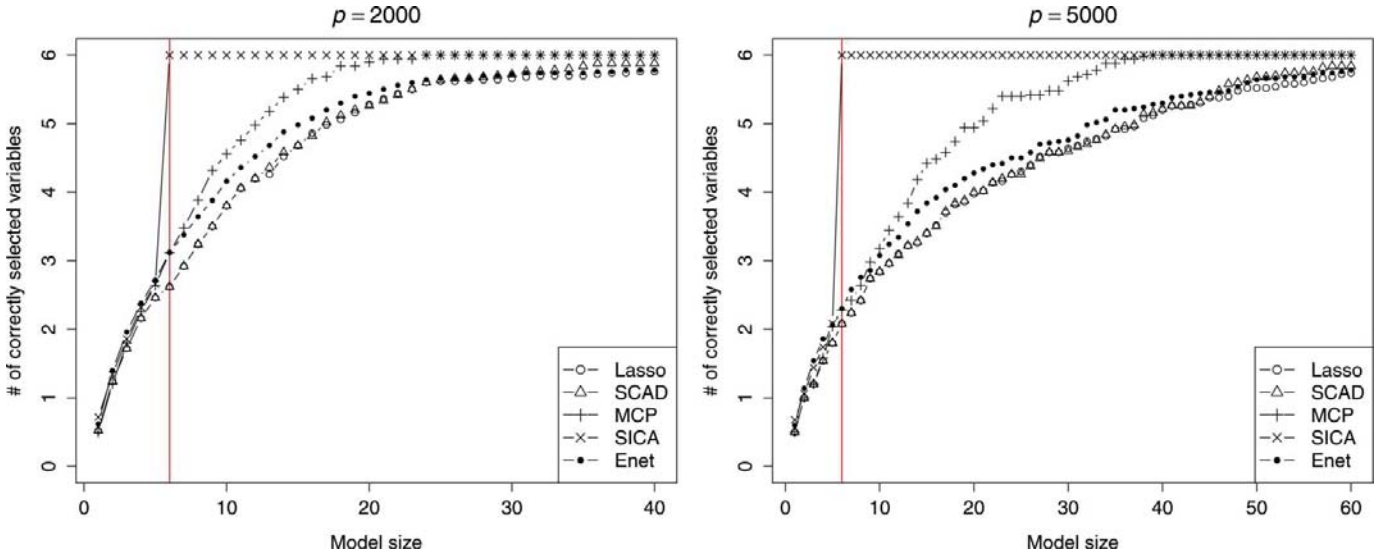

Figure 2. Variable selection performance for various methods in the second simulation study with $\rho = 0.5$. The vertical line indicates the true sparsity dimension.

Table 1. Note that in the most difficult setting, $p = 5000$ and $\rho = 0.5$, both the Lasso and Enet missed on average at least one important variable, whereas the concave penalties still included all six important variables. The variable selection performance assisted by an oracle is then compared in Figure 2. The SICA was still the winner in this case, followed by MCP. It is interesting to note that the Enet clearly outperformed the Lasso with moderate model size; this difference, however, was not notable in the cross-validated results. Since cross-validation seems to have less tolerance on false negatives than on false positives in these scenarios, its behavior is better explained by the tail of the curve: the SCAD had a slight performance boost on the tail, whereas the Lasso and Enet were very close to each other and both below the SCAD.

The third simulation study reflects the more challenging case where the true model is only approximately sparse and we wish

Table 3. Results for various methods in the third simulation study with $n = 500$, $p = 2000$, and censoring rate about 25%. Values shown are means (standard deviations) of each performance measure over 50 replicates. The oracle method is based on all nonzero (strong and weak) effects

Setting	Method	PE1	PE2	L_2-loss	L_1-loss	#S	#FN	#FN-S
$s = 50$	Lasso	9.743 (0.039)	29.80 (3.28)	1.808 (0.214)	8.461 (0.675)	63.8 (17.5)	42.5 (1.4)	0.1 (0.6)
$\varepsilon = 0.1$	SCAD	9.597 (0.018)	12.07 (1.26)	0.552 (0.073)	4.039 (0.692)	43.5 (14.5)	41.6 (1.4)	0.0 (0.0)
	MCP	9.591 (0.017)	11.74 (1.25)	0.546 (0.082)	3.568 (0.575)	20.4 (8.6)	43.0 (0.9)	0.0 (0.0)
	SICA	9.582 (0.009)	11.16 (1.04)	0.536 (0.090)	3.168 (0.352)	6.0 (0.0)	44.0 (0.0)	0.0 (0.0)
	Enet	9.743 (0.040)	29.80 (3.28)	1.809 (0.215)	8.463 (0.676)	63.9 (17.3)	42.5 (1.4)	0.1 (0.6)
	Oracle	9.615 (0.046)	14.59 (2.72)	0.692 (0.145)	3.905 (0.794)	50.0 (0.0)	0.0 (0.0)	0.0 (0.0)
$s = 50$	Lasso	10.517 (0.040)	38.02 (2.95)	2.288 (0.205)	11.531 (0.749)	47.4 (29.9)	43.5 (3.4)	1.4 (2.0)
$\varepsilon = 0.2$	SCAD	10.359 (0.057)	22.17 (4.90)	1.097 (0.360)	7.986 (1.175)	66.3 (17.7)	38.0 (3.3)	0.1 (0.8)
	MCP	10.364 (0.071)	23.33 (6.30)	1.200 (0.460)	7.701 (1.494)	33.4 (10.5)	40.7 (2.9)	0.3 (1.1)
	SICA	10.327 (0.023)	21.63 (1.28)	1.121 (0.110)	6.755 (0.607)	7.1 (6.1)	43.9 (0.4)	0.0 (0.0)
	Enet	10.518 (0.040)	38.04 (2.94)	2.289 (0.205)	11.547 (0.769)	48.4 (30.5)	43.4 (3.4)	1.3 (2.0)
	Oracle	10.252 (0.045)	15.22 (2.46)	0.717 (0.129)	4.005 (0.649)	50.0 (0.0)	0.0 (0.0)	0.0 (0.0)
$s = 100$	Lasso	12.630 (0.045)	33.34 (3.47)	2.026 (0.229)	10.942 (0.689)	57.0 (23.1)	91.6 (2.7)	0.6 (1.3)
$\varepsilon = 0.1$	SCAD	12.460 (0.023)	14.87 (1.17)	0.712 (0.073)	6.706 (0.641)	53.7 (14.8)	89.1 (2.5)	0.0 (0.0)
	MCP	12.456 (0.024)	14.83 (1.40)	0.727 (0.090)	6.313 (0.617)	27.2 (9.5)	91.5 (1.7)	0.0 (0.0)
	SICA	12.438 (0.015)	14.27 (1.24)	0.733 (0.098)	5.768 (0.382)	6.1 (0.3)	94.0 (0.0)	0.0 (0.0)
	Enet	12.630 (0.045)	33.34 (3.47)	2.026 (0.229)	10.942 (0.689)	57.0 (23.1)	91.6 (2.7)	0.6 (1.3)
	Oracle	12.703 (0.111)	26.11 (3.88)	1.219 (0.185)	9.679 (1.407)	100.0 (0.0)	0.0 (0.0)	0.0 (0.0)
$s = 100$	Lasso	13.652 (0.037)	41.79 (2.54)	2.514 (0.175)	15.707 (0.605)	31.2 (27.9)	94.4 (4.3)	2.6 (2.3)
$\varepsilon = 0.2$	SCAD	13.581 (0.068)	30.94 (7.51)	1.659 (0.582)	13.556 (1.437)	60.6 (26.6)	88.1 (5.7)	1.0 (2.0)
	MCP	13.578 (0.070)	31.79 (7.33)	1.738 (0.560)	13.308 (1.570)	30.2 (15.0)	91.8 (4.3)	1.1 (2.1)
	SICA	13.523 (0.058)	28.93 (4.88)	1.569 (0.353)	12.061 (1.313)	6.5 (4.2)	94.4 (1.2)	0.5 (1.4)
	Enet	13.652 (0.038)	41.80 (2.54)	2.514 (0.175)	15.744 (0.727)	33.0 (30.4)	94.4 (4.3)	2.6 (2.2)
	Oracle	13.587 (0.121)	27.41 (3.87)	1.296 (0.201)	10.307 (1.490)	100.0 (0.0)	0.0 (0.0)	0.0 (0.0)

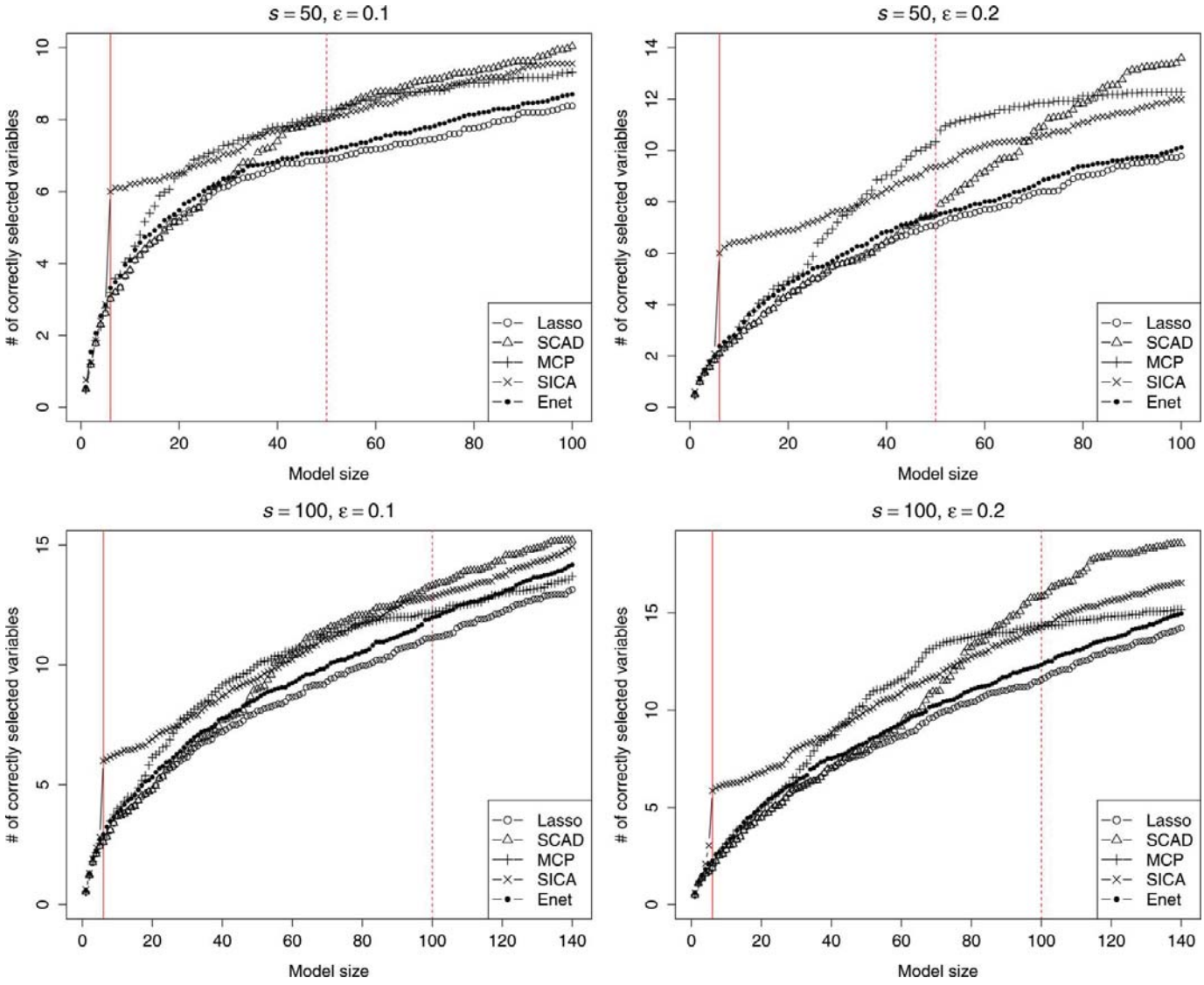

Figure 3. Variable selection performance for various methods in the third simulation study. The vertical lines indicate the sparsity level with strong effects (solid) and the sparsity level with all nonzero effects (dashed).

to analyze the robustness of our methods against departure from the sparsity assumption. To this end, in the previous model with $p = 2000$ and $n = 500$, we randomly chose 44 or 94 of the zero coefficients and perturbed them by $U(0, \varepsilon)$ with signs randomly placed, so that $s = 50$ or 100, respectively. This setting is intended to mimic a possible scenario in genetic studies where a relatively large number of genes have nonzero but weak effects, while a few key genes with strong effects stand out. To describe different levels of weak effects, we set $\varepsilon = 0.1$ and 0.2. The dimensionality apparently exceeds the limit for exact recovery of all, strong and weak, signals; recall from condition (13) that we require $n \gg s^2 \log p$ with constant φ and $r_1 = 1$. In this case, even the oracle estimator, which is based on all covariates with nonzero effects, does not perform well. Thus, our target here is to identify the few key variables and a small portion of the variables with weak effects.

The results with ten-fold cross-validation over 50 replicates are shown in Table 3, where the added performance measure #FN-S is the number of incorrectly excluded variables with strong effects. These results indicate that the SICA had the best performance in identifying the strong effects, because it aggressively seeks a sparse representation of the data and treats the variables with weak effects as noise variables. If the weak effects are also of interest, however, the SCAD and MCP performed better in that they selected more variables with nonzero effects at the expense of picking a larger model. It is worthwhile to note that when $\varepsilon = 0.1$, the SCAD, MCP, and SICA all had better prediction and estimation accuracy than the oracle estimator, which demonstrates the advantages of sparse modeling. The variable selection performance of all methods assisted by an oracle is depicted in Figure 3, from which we see that the benefits of the concave penalties remain intact. Moreover, as the perturbation threshold ε increases, that is, the model deviates further from sparsity, the SCAD and MCP showed less aggressive behavior than the SICA and were able to identify more variables with weak effects.

Finally, we point out that there are also scenarios where the Enet tends to outperform the other methods; for instance, when many highly correlated variables have comparable effects and the sparsity dimension is intrinsically high, the L_2 term in the

Table 4. Results for various methods applied to the DLBCL data

Method	No. of selected genes	Prediction error	p-value
Lasso	24	0.3047	0.016
SCAD	21	0.3048	0.013
MCP	13	0.3083	0.012
SICA	11	0.3038	0.003
Enet	26	0.3046	0.013

Enet penalty will play a pivotal role in regularization. These situations, however, are not our focus in sparse modeling and we do not pursue further in this article.

6. REAL DATA ANALYSIS

We illustrate the proposed methods by an application to the diffuse large-B-cell lymphoma (DLBCL) data analyzed by Rosenwald et al. (2002). This dataset consists of gene expression measurements for 7399 genes and survival outcomes after chemotherapy on 240 patients. The median follow-up time was 2.8 years and 138 patients died during the study period. The aim of the study was to formulate a molecular predictor of survival after chemotherapy for the disease. As in Rosenwald et al. (2002), the dataset was randomly divided into a training set of 160 patients and a test set of 80 patients. We then applied the Lasso, SCAD, MCP, SICA, and Enet methods to the training set and used ten-fold cross-validation to choose the optimal regularization parameter.

To assess the prediction performance of the resulting model, in addition to computing the prediction error based on the loss function $L(\cdot)$, we classified the test patients into a low-risk group and a high-risk group of equal size, according to the individual predicted excess risk $\widehat{\boldsymbol{\beta}}^T\mathbf{Z}$, and performed a two-sample log-rank test. Table 4 reports the number of selected genes, prediction error, and p-value of the log-rank test for each method, and the estimated coefficients for selected genes are given in Table 5. These results suggest that all methods performed reasonably well in prediction, but the concave penalties selected sparser models than the Lasso and Enet. To see the similarity of the prediction results, we note that 60 of the 80 test patients received the same risk classification from all five methods. The selected genes also exhibit some consistency across different methods, although variability was observed in the cross-validation procedure due to the small sample size and ultra-high dimensionality. We remark that, to gain further stability, the idea of the sure independence screening (Fan and Lv 2008) can be used to reduce the dimensionality to a more manageable scale before applying the regularization methods, in which case the prediction performance of all methods tends to improve.

7. DISCUSSION

We have proposed a class of regularization methods for simultaneous variable selection and estimation in the additive hazards model. The main message of this article is that regularization methods can be used for variable selection with censored survival data in high-dimensional settings under conditions that are parallel to those in the linear regression context. Moreover, we have shown that concave penalties can substantially improve on the L_1 method and yield sparser models with better prediction performance, as evident from our theoretical and numerical results. Although we have focused on the additive hazards model, the empirical process techniques used here are fairly general and can be easily adapted to other survival models; for example, a key step to establishing a high-dimensional theory for the Cox model under similar conditions is to characterize the concentration of the empirical information matrix around its population counterpart.

Table 5. Estimated coefficients for selected genes in the DLBCL data by various methods

Gene ID	Lasso	SCAD	MCP	SICA	Enet
25054	0.0066	0.0064	0.0035	0.0102	0.0065
33791	−0.0041	−0.0021			−0.0048
27181	−0.0090	−0.0084	−0.0019		−0.0092
28654	0.0039	0.0035			0.0037
31242	0.0112	0.0101	0.0005	0.0129	0.0113
31981	0.0223	0.0200	0.0222	0.0410	0.0214
27718	0.0015	0.0015			0.0024
24725	0.0095	0.0085	0.0133		0.0092
27218	0.0148	0.0153	0.0335	0.0015	0.0149
33014	0.0067	0.0054			0.0075
16006	0.0021				0.0029
33974					0.0003
27731	−0.0252	−0.0206		−0.0551	−0.0242
24394	−0.0190	−0.0198	−0.0557		−0.0197
24400	0.0035	0.0013	0.0002		0.0045
24203	0.0002				0.0012
24271	−0.0065	−0.0038	−0.0001	−0.0114	−0.0069
25977	−0.0197	−0.0168	−0.0157	−0.0221	−0.0200
34344	0.0321	0.0302	0.0400	0.0348	0.0312
24530	0.0004				0.0011
27191					−0.0002
33358	0.0012	0.0017			0.0019
26470	0.0067	0.0036	0.0048	0.0074	0.0063
26524	0.0030	0.0021			0.0038
34376	0.0278	0.0243	0.0311	0.0296	0.0273
32679	−0.0075	−0.0054		−0.0053	−0.0081

The fact that our theoretical results allow the dimensionality to grow exponentially with the sample size has important implications. In practice, how high dimensionality the proposed methods can handle depends critically on the sample size and the correlation structure of the covariates. Variable selection in regression, in general, is a very difficult problem, and is even more challenging in the survival context. As a consequence, a relatively large sample size is essential to making reliable inference. In situations where the proposed methods may fail, it would be desirable to explore strategies that combine the strengths of a variety of approaches, and regularization methods can then be used as building blocks in such more powerful procedures.

We have partly based our performance comparisons on the best model selected by an oracle. In practice, how to choose the optimal regularization parameter remains a challenging issue for the regularization methodology. Although cross-validation is often the choice we have to resort to and sometimes performs better than Akaike information criterion (AIC) or Bayesian information criterion (BIC) type criteria, it suffers from several drawbacks and limitations. When the dimensionality is

exceedingly high and the amount of regularization is necessarily large, the cross-validation curve can easily blow up and result in selecting too few variables. If variable selection is the sole purpose, stability selection (Meinshausen and Bühlmann 2010) and related methods could potentially provide more accurate error control. These problems will be interesting topics for future research.

APPENDIX A: PROOFS

For clarity and readability, before proceeding to the proofs of our main results, we first present a lemma that provides optimality conditions for the regularization problem (4), and establish a series of concentration inequalities that are essential to the main proofs. These results are also of independent interest. For notational simplicity, constants that are used in our proofs may vary from line to line.

A.1 Optimality Conditions

The following lemma provides optimality conditions that characterize a local solution to the regularization problem (4) and will be needed in the proof of Theorem 1.

Lemma A.1. (Characterization of the regularized estimator). Under Condition 1, $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is a strict local minimizer of problem (4) if the following conditions hold:

$$\mathbf{U}_{\widehat{A}}(\widehat{\boldsymbol{\beta}}) - \lambda \rho'_\lambda(|\widehat{\boldsymbol{\beta}}_{\widehat{A}}|) \circ \operatorname{sgn}(\widehat{\boldsymbol{\beta}}_{\widehat{A}}) = \mathbf{0}, \tag{A.1}$$

$$\|\mathbf{U}_{\widehat{A}^c}(\widehat{\boldsymbol{\beta}})\|_\infty < \lambda \rho'(0+), \tag{A.2}$$

and

$$\Lambda_{\min}(\mathbf{V}_{\widehat{A}\widehat{A}}) > \lambda \kappa(\rho_\lambda; \widehat{\boldsymbol{\beta}}_{\widehat{A}}), \tag{A.3}$$

where $\circ$ is the Hadamard (entrywise) product and the functions $|\cdot|$, $\rho'_\lambda(\cdot)$, and $\operatorname{sgn}(\cdot)$ are applied componentwise.

Proof. We first consider the $|\widehat{A}|$-dimensional subspace $\mathcal{B} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta}_{\widehat{A}^c} = \mathbf{0}\}$. Condition (A.3) implies that the objective function $Q(\boldsymbol{\beta})$ in (4) is strictly convex in a neighborhood of $\widehat{\boldsymbol{\beta}}$ in $\mathcal{B}$. Then condition (A.1) implies that $\widehat{\boldsymbol{\beta}}$ is a stationary point and hence a strict local minimizer of $Q(\boldsymbol{\beta})$ in the subspace $\mathcal{B}$.

It remains to show that, for any $\boldsymbol{\beta}_1 \in \mathbb{R}^p \setminus \mathcal{B}$ that lies in a sufficiently small neighborhood of $\widehat{\boldsymbol{\beta}}$, we have $Q(\boldsymbol{\beta}_1) > Q(\widehat{\boldsymbol{\beta}})$. To this end, let $\boldsymbol{\beta}_2$ be the projection of $\boldsymbol{\beta}_1$ onto the subspace $\mathcal{B}$. Since $Q(\boldsymbol{\beta}_2) \geq Q(\widehat{\boldsymbol{\beta}})$ from the preceding paragraph, it suffices to show that $Q(\boldsymbol{\beta}_1) > Q(\boldsymbol{\beta}_2)$. By the mean value theorem, we have

$$\begin{aligned} Q(\boldsymbol{\beta}_1) - Q(\boldsymbol{\beta}_2) &= \sum_{j \in \widehat{A}^c : \beta_{1j} \neq 0} \frac{\partial Q(\boldsymbol{\beta}^*)}{\partial \beta_j} \beta_{1j} \\ &= \sum_{j \in \widehat{A}^c : \beta_{1j} \neq 0} \{-U_j(\boldsymbol{\beta}^*) + \lambda \rho'_\lambda(|\beta_j^*|) \operatorname{sgn}(\beta_j^*)\} \beta_{1j}, \end{aligned} \tag{A.4}$$

where $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T$ is a point on the line segment between $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1p})^T$ and $\boldsymbol{\beta}_2$. It follows from condition (A.2) and continuity that $|U_j(\boldsymbol{\beta}^*)| < \lambda \rho'_\lambda(|\beta_j^*|) \operatorname{sgn}(\beta_j^*)$ for all $j \in \widehat{A}^c$, provided that $\boldsymbol{\beta}_1$, and hence $\boldsymbol{\beta}^*$, is sufficiently close to $\widehat{\boldsymbol{\beta}}$. Using this fact and that $\operatorname{sgn}(\beta_j^*) = \operatorname{sgn}(\beta_{1j})$, we see that each term in (A.4) is positive, and thus $Q(\boldsymbol{\beta}_1) > Q(\boldsymbol{\beta}_2)$. This completes the proof.

A.2 Concentration Inequalities

The major complexity in our proofs lies in characterizing the concentration of the large matrices $\mathbf{V}$ and $\mathbf{W}$ and vector $\mathbf{U}(\boldsymbol{\beta}_0)$. Due to the high dimensionality and dependency among entries, this is not a direct consequence of classical random matrix theory. Also, since each entry of the stochastic matrices or vector is not a sum of independent terms, the usual Hoeffding's inequality is not applicable. Hence, to establish the needed concentration inequalities, we rely on a functional Hoeffding-type inequality and some maximal inequalities for empirical processes as our primary mathematical tools.

We adopt the standard empirical process notation. For any measurable function f, we denote by $\mathbb{P}_n f$ and Pf the expectations of f under the empirical measure $\mathbb{P}_n$ and the probability measure P, respectively. Let $\|\cdot\|_{P,r}$ denote the usual $L_r(P)$-norm. The "size" of a class $\mathcal{F}$ of functions is measured by the *bracketing number* $N_{[]}(\varepsilon, \mathcal{F}, L_r(P))$, the minimum number of ε-brackets in $L_r(P)$ needed to cover $\mathcal{F}$, and the *covering number* $N(\varepsilon, \mathcal{F}, L_2(Q))$, the minimum number of $L_2(Q)$-balls of radius ε needed to cover $\mathcal{F}$. The logarithms of the bracketing number and covering number are called *entropy with bracketing* and *entropy*, respectively. The *bracketing integral* and *uniform entropy integral* are defined as

$$J_{[]}(\delta, \mathcal{F}, L_2(P)) = \int_0^\delta \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, L_2(P))}\, d\varepsilon$$

and

$$J(\delta, \mathcal{F}, L_2) = \int_0^\delta \sqrt{\log \sup_Q N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))}\, d\varepsilon,$$

respectively, where F is an envelope of $\mathcal{F}$, that is, $|f| \leq F$ for all $f \in \mathcal{F}$, and the supremum is taken over all probability measures Q with $\|F\|_{Q,r} > 0$. We refer the unfamiliar reader to van der Vaart (1998, chap. 19) or Kosorok (2008, chap. 2) for more definitions and concepts.

Define $\mathbf{S}^{(k)}(t) = n^{-1} \sum_{j=1}^n Y_j(t) \mathbf{Z}_j(t)^{\otimes k}$, $k = 0, 1, 2$, which are the sample counterparts of $\mathbf{s}^{(k)}(t)$ defined in (6), and recall that $\overline{\mathbf{Z}}(t) = \mathbf{S}^{(1)}(t)/\mathbf{S}^{(0)}(t)$. We begin with the following lemma, on which the other inequalities will be based.

Lemma A.2. (Concentration of $\mathbf{S}^{(k)}(\cdot)$, $k = 0, 1, 2$). Under Condition 2, there exist constants $C, K > 0$ such that

$$P\left(\sup_{t \in [0,\tau]} |S^{(0)}(t) - s^{(0)}(t)| \geq Cn^{-1/2}(1+x) \right) \leq \exp(-Kx^2), \tag{A.5}$$

$$P\left(\sup_{t \in [0,\tau]} \left|S_j^{(1)}(t) - s_j^{(1)}(t)\right| \geq Cn^{-1/2}(1+x) \,|\, \Omega_L \right) \leq \exp(-Kx^2/L^2), \tag{A.6}$$

and

$$P\left(\sup_{t \in [0,\tau]} \left|S_{ij}^{(2)}(t) - s_{ij}^{(2)}(t)\right| \geq Cn^{-1/2}(1+x) \,|\, \Omega_L \right) \leq \exp(-Kx^2/L^4), \tag{A.7}$$

for all $x > 0$ and $i, j = 1, \ldots, p$, where $S_j^{(1)}(\cdot)$ is the jth component of $\mathbf{S}^{(1)}(\cdot)$ and $S_{ij}^{(2)}(\cdot)$ is the (i, j)th entry of the matrix $\mathbf{S}^{(2)}(\cdot)$.

Proof. We only show (A.6), and the other two inequalities follow similarly. Write $R_j = \sup_{t \in [0,\tau]} |S_j^{(1)}(t) - s_j^{(1)}(t)|$. The main idea is to apply a functional Hoeffding-type inequality, Theorem 9 of Massart (2000). To this end, we need to control the term ER_j.

We first show that the class of functions $\{Y(t)Z_j(t) : t \in [0, \tau]\}$ has bounded uniform entropy integral. Since a function of bounded variation can be expressed as the difference of two increasing functions, it follows from Lemma 9.10 of Kosorok (2008) that $\mathcal{Z}_j \equiv \{Z_j(t) : t \in [0, \tau]\}$ is a VC-hull class associated with a VC class of index 2. Then, by Corollary 2.6.12 of van der Vaart and Wellner (1996), the entropy of $\mathcal{Z}_j$ satisfies $\log N(\varepsilon \|F\|_{Q,2}, \mathcal{Z}_j, L_2(Q)) \leq K'(1/\varepsilon)$ for some constant $K' > 0$, and hence, $\mathcal{Z}_j$ has the uniform entropy integral

$$J(1, \mathcal{Z}_j, L_2) \leq \int_0^1 \sqrt{K'(1/\varepsilon)}\, d\varepsilon < \infty.$$

Also, by Example 19.16 of van der Vaart (1998), $\mathcal{Y} \equiv \{Y(t) : t \in [0, \tau]\}$ is a VC class and hence has bounded uniform entropy

integral. Thus, by Theorem 9.15 of Kosorok (2008), $\mathcal{Y}\mathcal{Z}_j$ has bounded uniform entropy integral.

Now an application of Lemma 19.38 of van der Vaart (1998) gives

$$ER_j \le C'n^{-1/2}J(1, \mathcal{Y}\mathcal{Z}_j, L_2)\|F\|_{P,2} \le Cn^{-1/2}$$

for some constants $C', C > 0$, where the envelope F is taken as $\sup_{t\in[0,\tau]} Y(t)|Z_j(t)|$. It follows from Theorem 9 of Massart (2000) that

$$\begin{aligned} P(R_j \ge Cn^{-1/2}(1+x) \,|\, \Omega_L) &\le P(R_j \ge ER_j + Cn^{-1/2}x \,|\, \Omega_L) \\ &\le \exp(-Kx^2/L^2) \end{aligned}$$

for some constant $K > 0$, which concludes the proof.

Lemma A.3. (Concentration of $\mathbf{U}(\boldsymbol{\beta}_0)$). Under Conditions 2, there exist constants $C, D, K > 0$ such that

$$P(|U_j(\boldsymbol{\beta}_0)| \ge Cn^{-1/2}(1+x) \,|\, \Omega_L) \le D\exp\left(-K\frac{x^2 \wedge n}{L^4}\right)$$

for all $x > 0$ and $j = 1, \ldots, p$, where $U_j(\boldsymbol{\beta}_0)$ is the jth component of $\mathbf{U}(\boldsymbol{\beta}_0)$.

Proof. We first write

$$\begin{aligned} U_j(\boldsymbol{\beta}_0) &= \mathbb{P}_n \int_0^\tau \{Z_j(t) - \bar{Z}_j(t)\}\, dM(t) \\ &= \mathbb{P}_n \int_0^\tau Z_j(t)\, dM(t) - \mathbb{P}_n \int_0^\tau \bar{Z}_j(t)\, dM(t) \equiv T_1 - T_2, \end{aligned}$$

where $\bar{Z}_j(\cdot)$ is the jth component of $\overline{\mathbf{Z}}(\cdot)$. Since term T_1 is an iid sum of mean-zero random variables, an application of Hoeffding's (1963) inequality gives $P(|T_1| \ge n^{-1/2}x \,|\, \Omega_L) \le 2\exp(-Kx^2/L^4)$ for some constant $K > 0$.

We will apply Theorem 9 of Massart (2000) to bound term T_2. Note that, from (A.5) and (A.6) in Lemma A.2, we have

$$P\left(\sup_{t\in[0,\tau]} \left|S^{(0)}(t) - s^{(0)}(t)\right| \ge \delta\right) \le \exp(-Kn)$$

and

$$P\left(\sup_{t\in[0,\tau]} \left|S_j^{(1)}(t) - s_j^{(1)}(t)\right| \ge \delta \,|\, \Omega_L\right) \le \exp(-Kn/L^2),$$

for some constant $\delta > 0$ and $j = 1, \ldots, p$. Since these two tail probabilities are bounded by $\exp(-Kn/L^4)$ for $L \ge 1$, it suffices to consider the case where $\sup_{t\in[0,\tau]} |S^{(0)}(t) - s^{(0)}(t)| \le \delta$ and $\sup_{t\in[0,\tau]} |S_j^{(1)}(t) - s_j^{(1)}(t)| \le \delta$ for $j = 1, \ldots, p$. Write

$$\begin{aligned} \bar{Z}_j(t) - e_j(t) = {} & \frac{1}{S^{(0)}(t)}\left\{S_j^{(1)}(t) - s_j^{(1)}(t)\right\} \\ & - \frac{s_j^{(1)}(t)}{S^{(0)}(t)s^{(0)}(t)}\{S^{(0)}(t) - s^{(0)}(t)\}. \end{aligned} \tag{A.8}$$

Since $S^{(0)}(\cdot)$ and $s^{(0)}(\cdot)$ are bounded away from zero by Condition 2(2), the above representation implies that $\sup_{t\in[0,\tau]} |\bar{Z}_j(t) - e_j(t)| \le \delta'$ for some constant $\delta' > 0$. Let $\mathcal{F}_j$ denote the class of functions $f : [0, \tau] \to \mathbb{R}$ that are of uniformly bounded variation and satisfy $\sup_{t\in[0,\tau]} |f(t) - e_j(t)| \le \delta'$. Define the class of functions $\mathcal{G}_j = \{\int_0^\tau f(t)\, dM(t) : f \in \mathcal{F}_j\}$ and $G_j = \sup_{g\in\mathcal{G}_j} |(\mathbb{P}_n - P)g| = \sup_{g\in\mathcal{G}_j} |\mathbb{P}_n g|$. We need to control EG_j.

By constructing $\|\cdot\|_\infty$-balls centered at piecewise constant functions on a regular grid, one can show that the covering number of the class $\mathcal{F}_j$ satisfies $N(\varepsilon, \mathcal{F}_j, \|\cdot\|_\infty) \le (K/\varepsilon)^{K'/\varepsilon}$ for some constants $K, K' > 0$. Note also that, for any $f_1, f_2 \in \mathcal{F}_j$,

$$\begin{aligned} &\left|\int_0^\tau f_1(t)\, dM(t) - \int_0^\tau f_2(t)\, dM(t)\right| \\ &\qquad \le \sup_{s\in[0,\tau]} |f_1(s) - f_2(s)| \int_0^\tau |dM(t)|. \end{aligned}$$

By Theorem 2.7.11 of van der Vaart and Wellner (1996), the bracketing number of the class $\mathcal{G}_j$ satisfies $N_{[]}(2\varepsilon\|F\|_{P,2}, \mathcal{G}_j, L_2(P)) \le N(\varepsilon, \mathcal{F}_j, \|\cdot\|_\infty) \le (K/\varepsilon)^{K'/\varepsilon}$, where $F = \int_0^\tau |dM(t)|$. Hence, $\mathcal{G}_j$ has bounded bracketing integral.

An application of Corollary 19.35 of van der Vaart (1998) yields that

$$EG_j \le C'n^{-1/2}J_{[]}(\|G\|_{P,2}, \mathcal{G}_j, L_2(P)) \le Cn^{-1/2}$$

for some constants $C', C > 0$, where G is an envelope of $\mathcal{G}_j$. We then apply Theorem 9 of Massart (2000) to obtain $P(|T_2| \ge Cn^{-1/2}(1 + x) \,|\, \Omega_L) \le \exp(-Kx^2/L^4)$ for some constant $K > 0$. Putting the bounds for T_1 and T_2 together leads to the desired inequality, thus completing the proof.

Lemma A.4. (Concentration of $\mathbf{V}$). Under Condition 2, there exist constants $C, D, K > 0$ such that

$$P(|V_{ij} - D_{ij}| \ge Cn^{-1/2}(1+x) \,|\, \Omega_L) \le D\exp\left(-K\frac{x^2 \wedge n}{L^4}\right)$$

for all $x > 0$ and $i, j = 1, \ldots, p$, where V_{ij} and D_{ij} are the (i, j)th entries of the matrices $\mathbf{V}$ and $\mathbf{D}$, respectively.

Proof. We first write

$$\begin{aligned} V_{ij} - D_{ij} = {} & \int_0^\tau \left\{S_{ij}^{(2)}(t) - s_{ij}^{(2)}(t)\right\} dt \\ & + \int_0^\tau \left\{\frac{S_i^{(1)}(t)S_j^{(1)}(t)}{S^{(0)}(t)} - \frac{s_i^{(1)}(t)s_j^{(1)}(t)}{s^{(0)}(t)}\right\} dt \equiv T_1 + T_2. \end{aligned}$$

Clearly, (A.7) in Lemma A.2 implies that $P(|T_1| \ge Cn^{-1/2}(1 + x) \,|\, \Omega_L) \le \exp(-Kx^2/L^4)$. To bound term T_2, write

$$\begin{aligned} \frac{S_i^{(1)}(t)S_j^{(1)}(t)}{S^{(0)}(t)} - \frac{s_i^{(1)}(t)s_j^{(1)}(t)}{s^{(0)}(t)} &= \frac{S_j^{(1)}(t)}{S^{(0)}(t)}\left\{S_i^{(1)}(t) - s_i^{(1)}(t)\right\} \\ + \frac{s_i^{(1)}(t)}{S^{(0)}(t)}\left\{S_j^{(1)}(t) - s_j^{(1)}(t)\right\} &- \frac{s_i^{(1)}(t)s_j^{(1)}(t)}{S^{(0)}(t)s^{(0)}(t)}\left\{S^{(0)}(t) - s^{(0)}(t)\right\}. \end{aligned}$$

By the same arguments as in the proof of Lemma A.3, it suffices to consider the case where $\sup_{t\in[0,\tau]} |S^{(0)}(t) - s^{(0)}(t)| \le \delta$ and $\sup_{t\in[0,\tau]} |S_j^{(1)}(t) - s_j^{(1)}(t)| \le \delta$ for some constant $\delta > 0$ and $j = 1, \ldots, p$. From the above representation and (A.5) and (A.6) in Lemma A.2, it follows that $P(|T_2| \ge Cn^{-1/2}(1 + x) \,|\, \Omega_L) \le 3\exp(-Kx^2/L^2)$. Combining the bounds for T_1 and T_2 yields the desired inequality and concludes the proof.

Lemma A.5. (Concentration of $\mathbf{W}$). Under Condition 2, there exist constants $C, D, K > 0$ such that

$$P(|W_{ij} - \Sigma_{ij}| \ge Cn^{-1/2}(1+x) \,|\, \Omega_L) \le D\exp\left(-K\frac{x^2 \wedge n}{L^4}\right)$$

for all $x > 0$ and $i, j = 1, \ldots, p$, where W_{ij} and Σ_{ij} are the (i, j)th entries of the matrices $\mathbf{W}$ and $\boldsymbol{\Sigma}$, respectively.

Proof. We first write

$$\begin{aligned} W_{ij} - \Sigma_{ij} &= (\mathbb{P}_n - P)\int_0^\tau Z_i(t)Z_j(t)\,dN(t) \\ &\quad - \left\{\mathbb{P}_n \int_0^\tau \bar{Z}_i(t)Z_j(t)\,dN(t) - P\int_0^\tau e_i(t)Z_j(t)\,dN(t)\right\} \\ &\quad - \left\{\mathbb{P}_n \int_0^\tau Z_i(t)\bar{Z}_j(t)\,dN(t) - P\int_0^\tau Z_i(t)e_j(t)\,dN(t)\right\} \\ &\quad + \left\{\mathbb{P}_n \int_0^\tau \bar{Z}_i(t)\bar{Z}_j(t)\,dN(t) - P\int_0^\tau e_i(t)e_j(t)\,dN(t)\right\} \\ &\equiv T_1 - T_2 - T_3 + T_4. \end{aligned}$$

Since term T_1 is an iid sum, an application of Hoeffding's inequality gives $P(|T_1| \ge n^{-1/2}x \mid \Omega_L) \le 2\exp(-Kx^2/L^4)$. To bound term T_2, write

$$\begin{aligned} T_2 &= (\mathbb{P}_n - P)\int_0^\tau \bar{Z}_i(t)Z_j(t)\,dN(t) \\ &\quad + P\int_0^\tau \{\bar{Z}_i(t) - e_i(t)\}Z_j(t)\,dN(t) \equiv T_{21} + T_{22}. \end{aligned}$$

Term T_{21} can be bounded similarly as term T_2 in the proof of Lemma A.3. Also, note that

$$|T_{22}| \le \sup_{s\in[0,\tau]} |\bar{Z}_i(s) - e_i(s)| P\int_0^\tau |Z_j(t)|\,dN(t).$$

Then it follows from (A.8) and Lemma A.2 that

$$P(|T_{22}| \ge Cn^{-1/2}(1+x) \mid \Omega_L) \le D\exp\left(-K\frac{x^2\wedge n}{L^4}\right).$$

We can bound terms T_3 and T_4 similarly and thus obtain the desired inequality. This completes the proof.

A.3 Proof of Lemma 1

By the union bound and Lemma A.4, we have

$$\begin{aligned} &P\left(\|\mathbf{V}_{AA} - \mathbf{D}_{AA}\|_\infty \ge \frac{1}{2\varphi} \,\middle|\, \Omega_L\right) \\ &= P\left(\max_{i\in A}\sum_{j\in A}|V_{ij} - D_{ij}| \ge \frac{1}{2\varphi} \,\middle|\, \Omega_L\right) \\ &\le \sum_{i\in A} P\left(\sum_{j\in A}|V_{ij} - D_{ij}| \ge \frac{1}{2\varphi} \,\middle|\, \Omega_L\right) \\ &\le \sum_{i\in A}\sum_{j\in A} P\left(|V_{ij} - D_{ij}| \ge \frac{1}{2\varphi s} \,\middle|\, \Omega_L\right) \\ &\le s^2 D\exp\left\{-K\frac{n}{L^4}\left(\frac{1}{\varphi^2 s^2}\wedge 1\right)\right\}. \end{aligned}$$

By an error bound for matrix inversion (Horn and Johnson 1985, p. 336), if $\|\mathbf{V}_{AA} - \mathbf{D}_{AA}\|_\infty < 1/(2\varphi)$, then

$$\frac{\|\mathbf{V}_{AA}^{-1} - \mathbf{D}_{AA}^{-1}\|_\infty}{\varphi} \le \frac{\varphi\|\mathbf{V}_{AA} - \mathbf{D}_{AA}\|_\infty}{1 - \varphi\|\mathbf{V}_{AA} - \mathbf{D}_{AA}\|_\infty} < 1,$$

and hence, $\|\mathbf{V}_{AA}^{-1}\|_\infty \le \|\mathbf{D}_{AA}^{-1}\|_\infty + \|\mathbf{V}_{AA}^{-1} - \mathbf{D}_{AA}^{-1}\|_\infty = \varphi + \|\mathbf{V}_{AA}^{-1} - \mathbf{D}_{AA}^{-1}\|_\infty < 2\varphi$. Then probability bound (7) follows.

To show probability bound (8), we write

$$\begin{aligned} \mathbf{V}_{A^cA}\mathbf{V}_{AA}^{-1} - \mathbf{D}_{A^cA}\mathbf{D}_{AA}^{-1} &= (\mathbf{V}_{A^cA} - \mathbf{D}_{A^cA})\mathbf{V}_{AA}^{-1} + \mathbf{D}_{A^cA}(\mathbf{V}_{AA}^{-1} - \mathbf{D}_{AA}^{-1}) \\ &= (\mathbf{V}_{A^cA} - \mathbf{D}_{A^cA})\mathbf{V}_{AA}^{-1} \\ &\quad - \mathbf{D}_{A^cA}\mathbf{D}_{AA}^{-1}(\mathbf{V}_{AA} - \mathbf{D}_{AA})\mathbf{V}_{AA}^{-1} \equiv T_1 - T_2. \end{aligned}$$

Similarly as before, by the union bound and Lemma A.4, we have

$$\begin{aligned} &P\left[\|\mathbf{V}_{A^cA} - \mathbf{D}_{A^cA}\|_\infty \ge \frac{1}{2\varphi}\left\{\frac{\alpha}{4}\frac{\rho'(0+)}{\rho'_\lambda(d)}\right\}\wedge\left(\frac{c}{2}n^\gamma\right) \,\middle|\, \Omega_L\right] \\ &\quad \le (p-s)sD\exp\left[-K\frac{n}{L^4}\left\{\frac{\left(\rho'_\lambda(d)^{-1}\wedge n^\gamma\right)^2}{\varphi^2 s^2}\wedge 1\right\}\right]. \end{aligned}$$

This, along with (7), gives

$$\begin{aligned} &P\left[\|T_1\|_\infty \ge \left\{\frac{\alpha}{4}\frac{\rho'(0+)}{\rho'_\lambda(d)}\right\}\wedge\left(\frac{c}{2}n^\gamma\right) \,\middle|\, \Omega_L\right] \\ &\le P\left[\|\mathbf{V}_{A^cA} - \mathbf{D}_{A^cA}\|_\infty \ge \frac{1}{2\varphi}\left\{\frac{\alpha}{4}\frac{\rho'(0+)}{\rho'_\lambda(d)}\right\}\wedge\left(\frac{c}{2}n^\gamma\right) \,\middle|\, \Omega_L\right] \\ &\quad + P\left(\left\|\mathbf{V}_{AA}^{-1}\right\|_\infty \ge 2\varphi \mid \Omega_L\right) \le (p-s)sD \\ &\quad \times \exp\left[-K\frac{n}{L^4}\left\{\frac{\left(\rho'_\lambda(d)^{-1}\wedge n^\gamma\right)^2}{\varphi^2 s^2}\wedge 1\right\}\right] \\ &\quad + s^2 D\exp\left\{-K\frac{n}{L^4}\left(\frac{1}{\varphi^2 s^2}\wedge 1\right)\right\}. \end{aligned}$$

Also, by Condition 3 and (7), we have

$$\begin{aligned} &P\left[\|T_2\|_\infty \ge \left\{\frac{\alpha}{4}\frac{\rho'(0+)}{\rho'_\lambda(d)}\right\}\wedge\left(\frac{c}{2}n^\gamma\right) \,\middle|\, \Omega_L\right] \\ &\le P\left[\|\mathbf{V}_{AA} - \mathbf{D}_{AA}\|_\infty \ge \frac{1}{2\varphi}\left\{\frac{\alpha}{4(1-\alpha)}\wedge\frac{1}{2}\right\} \,\middle|\, \Omega_L\right] \\ &\quad + P\left(\left\|\mathbf{V}_{AA}^{-1}\right\|_\infty \ge 2\varphi \mid \Omega_L\right) \\ &\le s^2 D\exp\left\{-K\frac{n}{L^4}\left(\frac{1}{\varphi^2 s^2}\wedge 1\right)\right\}. \end{aligned}$$

Putting the bounds for T_1 and T_2 together, we obtain

$$\begin{aligned} &P\left[\left\|\mathbf{V}_{A^cA}\mathbf{V}_{AA}^{-1} - \mathbf{D}_{A^cA}\mathbf{D}_{AA}^{-1}\right\|_\infty \ge \left\{\frac{\alpha}{2}\frac{\rho'(0+)}{\rho'_\lambda(d)}\right\}\wedge(cn^\gamma) \mid \Omega_L\right] \\ &\le (p-s)sD\exp\left[-K\frac{n}{L^4}\left\{\frac{\left(\rho'_\lambda(d)^{-1}\wedge n^\gamma\right)^2}{\varphi^2 s^2}\wedge 1\right\}\right] \\ &\quad + s^2 D\exp\left\{-K\frac{n}{L^4}\left(\frac{1}{\varphi^2 s^2}\wedge 1\right)\right\}. \end{aligned}$$

Then probability bound (8) follows from Condition 3 and the triangle inequality.

Finally, to show probability bound (9), by the Hoffman–Wielandt inequality (Horn and Johnson 1985), we have

$$\begin{aligned} |\Lambda_{\min}(\mathbf{V}_{AA}) - \Lambda_{\min}(\mathbf{D}_{AA})| &\le \left\{\sum_{j=1}^{s}|\Lambda_{(j)}(\mathbf{V}_{AA}) - \Lambda_{(j)}(\mathbf{D}_{AA})|^2\right\}^{1/2} \\ &\le \|\mathbf{V}_{AA} - \mathbf{D}_{AA}\|_F, \end{aligned}$$

where $\Lambda_{(j)}(\cdot)$ denotes the jth smallest eigenvalue and $\|\cdot\|_F$ is the Frobenius norm. Then it follows from the union bound and Lemma A.4 that

$$\begin{aligned} &P(|\Lambda_{\min}(\mathbf{V}_{AA}) - \Lambda_{\min}(\mathbf{D}_{AA})| \ge \mu \mid \Omega_L) \\ &\le P(\|\mathbf{V}_{AA} - \mathbf{D}_{AA}\|_F \ge \mu \mid \Omega_L) = P\left(\sum_{i,j\in A}|V_{ij} - D_{ij}|^2 \ge \mu^2 \mid \Omega_L\right) \\ &\le \sum_{i,j\in A} P\left(|V_{ij} - D_{ij}| \ge \frac{\mu}{s} \,\middle|\, \Omega_L\right) \le s^2 D\exp\left\{-K\frac{n}{L^4}\left(\frac{\mu^2}{s^2}\wedge 1\right)\right\}, \end{aligned}$$

which, by the definition of μ, implies (9). This concludes the proof.

A.4 Proof of Theorem 1

The main idea of the proof is to first define an event with a high probability and then analyze the behavior of the regularized estimator

$\widehat{\boldsymbol{\beta}}$ conditional on that event. The first part involves concentration inequalities developed in Section A.2 and Lemma 1, while the second part involves nonprobabilistic arguments based on Lemma A.1.

First, by the union bound and Lemma A.3, we have

$$\begin{aligned} P\biggl(\|\mathbf{U}_A(\boldsymbol{\beta}_0)\|_\infty \geq \frac{1}{2cn^\gamma}\frac{\alpha}{4}\lambda\rho'(0+) \mid \Omega_L\biggr) &\leq \sum_{j\in A} P\biggl(|U_j(\boldsymbol{\beta}_0)| \geq \frac{1}{2cn^\gamma}\frac{\alpha}{4}\lambda\rho'(0+) \mid \Omega_L\biggr) \\ &\leq sD\exp\biggl\{-K\frac{n}{L^4}\biggl(\frac{\lambda^2}{n^{2\gamma}}\wedge 1\biggr)\biggr\}. \end{aligned} \tag{A.9}$$

Similarly, we have

$$\begin{aligned} P\Bigl(\|\mathbf{U}_{A^c}(\boldsymbol{\beta}_0)\|_\infty \geq \frac{\alpha}{4}\lambda\rho'(0+) \mid \Omega_L\Bigr) &\leq (p-s)D\exp\biggl\{-K\frac{n}{L^4}(\lambda^2\wedge 1)\biggr\}. \end{aligned} \tag{A.10}$$

Also, Condition 2(3) and the union bound imply that

$$P\bigl(\Omega_L^c\bigr) \leq \sum_{j=1}^{p} P\biggl(\sup_{t\in[0,\tau]}|Z_j(t)| > L\biggr) \leq pD\exp(-KL^r). \tag{A.11}$$

It follows from (A.9)–(A.11) and Lemma 1 that with probability at least

$$\begin{aligned} 1-(p-s)sD\exp\biggl[-K\frac{n}{L^4}\biggl\{\frac{\bigl(\rho'_\lambda(d)^{-1}\wedge n^\gamma\bigr)^2}{\varphi^2 s^2}\wedge 1\biggr\}\biggr] \\ -s^2D\exp\biggl[-K\frac{n}{L^4}\biggl\{\frac{(\varphi^{-1}\wedge\mu)^2}{s^2}\wedge 1\biggr\}\biggr] \\ -sD\exp\biggl\{-K\frac{n}{L^4}\biggl(\frac{\lambda^2}{n^{2\gamma}}\wedge 1\biggr)\biggr\} \\ -(p-s)D\exp\biggl\{-K\frac{n}{L^4}(\lambda^2\wedge 1)\biggr\}-pD\exp(-KL^r), \end{aligned} \tag{A.12}$$

the following inequalities hold:

$$\|\mathbf{U}_A(\boldsymbol{\beta}_0)\|_\infty < \frac{1}{2cn^\gamma}\frac{\alpha}{4}\lambda\rho'(0+), \qquad \|\mathbf{U}_{A^c}(\boldsymbol{\beta}_0)\|_\infty < \frac{\alpha}{4}\lambda\rho'(0+),$$

$$\bigl\|\mathbf{V}_{AA}^{-1}\bigr\|_\infty < 2\varphi, \qquad \bigl\|\mathbf{V}_{A^cA}\mathbf{V}_{AA}^{-1}\bigr\|_\infty < \biggl\{\biggl(1-\frac{\alpha}{2}\biggr)\frac{\rho'(0+)}{\rho'_\lambda(d)}\biggr\}\wedge(2cn^\gamma),$$

and

$$\Lambda_{\min}(\mathbf{V}_{AA}) > \lambda\kappa_0. \tag{A.13}$$

From now on, we condition on the event that these inequalities hold. It suffices to find a $\widehat{\boldsymbol{\beta}}\in\mathbb{R}^p$ that satisfies all the optimality conditions in Lemma A.1 and the desired properties. To this end, take $\widehat{\boldsymbol{\beta}}_{A^c}=\mathbf{0}$, and we will first determine $\widehat{\boldsymbol{\beta}}_A$ by solving Equation (A.1). Since $\mathbf{U}(\boldsymbol{\beta})=\mathbf{b}-\mathbf{V}\boldsymbol{\beta}$, we have $\mathbf{U}_A(\widehat{\boldsymbol{\beta}})=\mathbf{U}_A(\boldsymbol{\beta}_0)-\mathbf{V}_{AA}(\widehat{\boldsymbol{\beta}}_A-\boldsymbol{\beta}_{0A})$. Substituting this into the equation $\mathbf{U}_A(\widehat{\boldsymbol{\beta}})-\lambda\rho'_\lambda(|\widehat{\boldsymbol{\beta}}_A|)\circ\operatorname{sgn}(\widehat{\boldsymbol{\beta}}_A)=\mathbf{0}$ gives

$$\widehat{\boldsymbol{\beta}}_A-\boldsymbol{\beta}_{0A}=\mathbf{V}_{AA}^{-1}\{\mathbf{U}_A(\boldsymbol{\beta}_0)-\lambda\rho'_\lambda(|\widehat{\boldsymbol{\beta}}_A|)\circ\operatorname{sgn}(\widehat{\boldsymbol{\beta}}_A)\}. \tag{A.14}$$

Define the function $f:\mathbb{R}^s\to\mathbb{R}^s$ by $f(\boldsymbol{\theta})=\boldsymbol{\beta}_{0A}+\mathbf{V}_{AA}^{-1}\{\mathbf{U}_A(\boldsymbol{\beta}_0)-\lambda\rho'_\lambda(|\boldsymbol{\theta}|)\circ\operatorname{sgn}(\boldsymbol{\theta})\}$, and let $\mathcal{K}$ denote the hypercube $\{\boldsymbol{\theta}\in\mathbb{R}^s:\|\boldsymbol{\theta}-\boldsymbol{\beta}_{0A}\|_\infty\leq c_1\varphi\lambda\rho'(0+)\}$. By the inequalities established before, for $\boldsymbol{\theta}\in\mathcal{K}$, we have

$$\begin{aligned} \|f(\boldsymbol{\theta})-\boldsymbol{\beta}_{0A}\|_\infty &\leq \bigl\|\mathbf{V}_{AA}^{-1}\bigr\|_\infty\{\|\mathbf{U}_A(\boldsymbol{\beta}_0)\|_\infty+\lambda\rho'(0+)\} \\ &\leq 2\varphi\biggl\{\frac{1}{2cn^\gamma}\frac{\alpha}{4}\lambda\rho'(0+)+\lambda\rho'(0+)\biggr\}\leq c_1\varphi\lambda\rho'(0+), \end{aligned}$$

that is, $f(\mathcal{K})\subset\mathcal{K}$. Also, condition (12) implies that for $\boldsymbol{\theta}\in\mathcal{K}$, $\|\boldsymbol{\theta}-\boldsymbol{\beta}_{0A}\|_\infty\leq d$, so that $\operatorname{sgn}(\boldsymbol{\theta})=\operatorname{sgn}(\boldsymbol{\beta}_{0A})$. Thus, in view of Condition 1, f is a continuous function on the convex, compact hypercube $\mathcal{K}$. An application of Brouwer's fixed point theorem yields that equation (A.14) has a solution $\widehat{\boldsymbol{\beta}}_A$ in $\mathcal{K}$. Moreover, $\operatorname{sgn}(\widehat{\boldsymbol{\beta}}_A)=\operatorname{sgn}(\boldsymbol{\beta}_{0A})$ and hence $\widehat{A}=A$. Thus, we have found a $\widehat{\boldsymbol{\beta}}\in\mathbb{R}^p$ that satisfies (A.1) and the desired properties. It remains to check conditions (A.2) and (A.3).

To verify that $\widehat{\boldsymbol{\beta}}$ satisfies (A.2), we write

$$\begin{aligned} \mathbf{U}_{A^c}(\widehat{\boldsymbol{\beta}}) &= \mathbf{U}_{A^c}(\boldsymbol{\beta}_0)-\mathbf{V}_{A^cA}(\widehat{\boldsymbol{\beta}}_A-\boldsymbol{\beta}_{0A}) \\ &= \mathbf{U}_{A^c}(\boldsymbol{\beta}_0)-\mathbf{V}_{A^cA}\mathbf{V}_{AA}^{-1}\{\mathbf{U}_A(\boldsymbol{\beta}_0)-\lambda\rho'_\lambda(|\widehat{\boldsymbol{\beta}}_A|)\circ\operatorname{sgn}(\widehat{\boldsymbol{\beta}}_A)\}, \end{aligned}$$

where we have substituted (A.14). Since $\|\widehat{\boldsymbol{\beta}}_A-\boldsymbol{\beta}_{0A}\|_\infty\leq d$, we have $\|\widehat{\boldsymbol{\beta}}_A\|_\infty=\|\boldsymbol{\beta}_{0A}+(\widehat{\boldsymbol{\beta}}_A-\boldsymbol{\beta}_{0A})\|_\infty\geq\|\boldsymbol{\beta}_{0A}\|_\infty-\|\widehat{\boldsymbol{\beta}}_A-\boldsymbol{\beta}_{0A}\|_\infty\geq 2d-d=d$. The triangle inequality, concavity of $\rho_\lambda(\cdot)$, and the inequalities established before together imply that

$$\begin{aligned} \bigl\|\mathbf{U}_{A^c}(\widehat{\boldsymbol{\beta}})\bigr\|_\infty &\leq \|\mathbf{U}_{A^c}(\boldsymbol{\beta}_0)\|_\infty+\bigl\|\mathbf{V}_{A^cA}\mathbf{V}_{AA}^{-1}\bigr\|_\infty\{\|\mathbf{U}_A(\boldsymbol{\beta}_0)\|_\infty+\lambda\rho'_\lambda(d)\} \\ &< \frac{\alpha}{4}\lambda\rho'(0+)+2cn^\gamma\frac{1}{2cn^\gamma}\frac{\alpha}{4}\lambda\rho'(0+) \\ &\quad+\biggl(1-\frac{\alpha}{2}\biggr)\frac{\rho'(0+)}{\rho'_\lambda(d)}\lambda\rho'_\lambda(d) \\ &= \frac{\alpha}{4}\lambda\rho'(0+)+\frac{\alpha}{4}\lambda\rho'(0+)+\biggl(1-\frac{\alpha}{2}\biggr)\lambda\rho'(0+) \\ &= \lambda\rho'(0+). \end{aligned}$$

Since $\|\widehat{\boldsymbol{\beta}}_A-\boldsymbol{\beta}_{0A}\|_\infty\leq d$, it follows from (A.13) that $\Lambda_{\min}(\mathbf{V}_{AA})>\lambda\kappa_0\geq\lambda\kappa(\rho_\lambda;\widehat{\boldsymbol{\beta}}_A)$, and hence, (A.3) is satisfied. Finally, we choose L by matching the exponential terms in (A.12) and note that the probability tends to 1 by conditions (10) and (11). This completes the proof.

A.5 Proof of Theorem 2

The proof Theorem 2 is based on the proof of Theorem 1 in Section A.4. First, by the same arguments as for deriving probability bound (9) in Lemma 1, one can obtain

$$\begin{aligned} &P(\Lambda_{\min}(\mathbf{V}_{AA})\leq\Lambda_1/2\mid\Omega_L) \\ &\quad= P(|\Lambda_{\min}(\mathbf{V}_{AA})-\Lambda_1|\geq\Lambda_1/2\mid\Omega_L) \\ &\quad\leq s^2D\exp\biggl\{-K\frac{n}{L^4}\biggl(\frac{\Lambda_1^2}{s^2}\wedge 1\biggr)\biggr\}. \end{aligned}$$

Thus, with probability at least

$$1-s^2D\exp\biggl\{-K\frac{n}{L^4}\biggl(\frac{\Lambda_1^2}{s^2}\wedge 1\biggr)\biggr\}-pD\exp(-KL^r), \tag{A.15}$$

it holds that $\Lambda_{\min}(\mathbf{V}_{AA})>\Lambda_1/2$, and hence

$$\bigl\|\mathbf{V}_{AA}^{-1}\bigr\|_2=1/\Lambda_{\min}(\mathbf{V}_{AA})<2/\Lambda_1. \tag{A.16}$$

From now on, we condition on the intersection of the event that (A.16) holds and the event defined in the proof of Theorem 1; such an event still has a high probability. Since sparsity has been established in Theorem 1, we only need to show the asymptotic normality. By substituting (A.14), we can write

$$\begin{aligned} \sqrt{n}\mathbf{u}^T\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}(\widehat{\boldsymbol{\beta}}_A-\boldsymbol{\beta}_{0A}) &= \sqrt{n}\mathbf{u}^T\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}\mathbf{V}_{AA}^{-1}\{\mathbf{U}_A(\boldsymbol{\beta}_0) \\ &\quad-\lambda\rho'_\lambda(|\widehat{\boldsymbol{\beta}}_A|)\circ\operatorname{sgn}(\widehat{\boldsymbol{\beta}}_A)\} \\ &= \sqrt{n}\mathbf{u}^T\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{U}_A(\boldsymbol{\beta}_0) \\ &\quad+\sqrt{n}\mathbf{u}^T\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}\bigl(\mathbf{V}_{AA}^{-1}-\mathbf{D}_{AA}^{-1}\bigr)\mathbf{U}_A(\boldsymbol{\beta}_0) \\ &\quad-\sqrt{n}\mathbf{u}^T\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}\mathbf{V}_{AA}^{-1}\lambda \\ &\quad\times\rho'_\lambda(|\widehat{\boldsymbol{\beta}}_A|)\circ\operatorname{sgn}(\widehat{\boldsymbol{\beta}}_A) \\ &\equiv T_1+T_2-T_3. \end{aligned}$$

We first consider term T_2. Since $\|\mathbf{u}\|_2=1$, we have

$$|T_2|\leq\sqrt{n}\|\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}\|_2\bigl\|\mathbf{D}_{AA}^{-1}\bigr\|_2\|\mathbf{V}_{AA}-\mathbf{D}_{AA}\|_2\bigl\|\mathbf{V}_{AA}^{-1}\bigr\|_2\|\mathbf{U}_A(\boldsymbol{\beta}_0)\|_2.$$

It follows from Lemma A.4 that

$$\begin{aligned} \|\mathbf{V}_{AA}-\mathbf{D}_{AA}\|_2\leq\|\mathbf{V}_{AA}-\mathbf{D}_{AA}\|_F &\leq\Bigl(s^2\max_{i,j\in A}|V_{ij}-D_{ij}|\Bigr)^{1/2} \\ &= sO_p(n^{-1/2}), \end{aligned}$$

and similarly, by Lemma A.3, $\|\mathbf{U}_A(\boldsymbol{\beta}_0)\|_2 = \sqrt{s}O_p(n^{-1/2})$. Using also $\|\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}\|_2 = \Lambda_3^{-1/2}$, $\|\mathbf{D}_{AA}^{-1}\|_2 = 1/\Lambda_1$, and (A.16), we obtain

$$\begin{aligned}|T_2| &\le \sqrt{n}\Lambda_3^{-1/2}\Lambda_1^{-1}sO_p(n^{-1/2})2\Lambda_1^{-1}\sqrt{s}O_p(n^{-1/2})\\ &= \frac{2s^{3/2}}{\Lambda_1^2\Lambda_3^{1/2}}O_p(n^{-1/2}) = o_p(1),\end{aligned}$$

by the third condition in (15). We then consider term T_3. The concavity of $\rho_\lambda(\cdot)$, (A.16), and condition (16) lead to

$$|T_3| \le \sqrt{n}\|\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}\|_2\|\mathbf{V}_{AA}^{-1}\|_2\lambda\sqrt{s}\rho'_\lambda(d) < \frac{2\sqrt{ns}\lambda\rho'_\lambda(d)}{\Lambda_1\Lambda_3^{1/2}} \to 0.$$

It remains to show that term T_1 is asymptotically normal. Note that

$$\mathbf{u}^T\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{W}_{AA}\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{u} = 1 + \mathbf{u}^T\boldsymbol{\Sigma}_{AA}^{-1/2}(\mathbf{W}_{AA} - \boldsymbol{\Sigma}_{AA})\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{u}.$$

It follows from Lemma A.5 that $\|\mathbf{W}_{AA} - \boldsymbol{\Sigma}_{AA}\|_2 = sO_p(n^{-1/2})$. Then the second term in the above display is bounded by

$$\begin{aligned}&\|\boldsymbol{\Sigma}_{AA}^{-1/2}\|_2\|\mathbf{W}_{AA} - \boldsymbol{\Sigma}_{AA}\|_2\|\boldsymbol{\Sigma}_{AA}^{-1/2}\|_2 = \Lambda_2^{-1/2}sO_p(n^{-1/2})\Lambda_2^{-1/2}\\ &\quad= \frac{s}{\Lambda_2}O_p(n^{-1/2}) = o_p(1),\end{aligned}$$

in view of the second condition in (15). An application of the martingale central limit theorem (Andersen and Gill 1982) yields that T_1 is asymptotically standard normal. Finally, we choose the optimal L in (A.15), note that the probability tends to 1 by the first condition in (15), and combine it with the probability in Theorem 1. This concludes the proof.

A.6 Proof of Theorem 3

Since the sequence of objective functions $\{\widetilde{Q}(\boldsymbol{\beta}^m)\}$ is decreasing, we see that the sequence $\{\widetilde{Q}(\boldsymbol{\beta}^m)\}$ lies in a compact set of $\mathbb{R}$. This entails that $\{\boldsymbol{\beta}^m\}$ is bounded, since any $|\beta_j| \to \infty$ would imply that $\widetilde{Q}(\boldsymbol{\beta}) \to \infty$ by the assumption $V_{jj} \ne 0$ for all j. Denote by $\boldsymbol{\alpha}_j^m = (\beta_1^m, \ldots, \beta_{j-1}^m, \beta_j^m, \beta_{j+1}^{m-1}, \ldots, \beta_p^{m-1})^T$.

To show part (a), let $\boldsymbol{\beta}^*$ be a cluster point of $\{\boldsymbol{\beta}^m\}$ and $\{\boldsymbol{\beta}^{m_k}\}$ a subsequence of $\{\boldsymbol{\beta}^m\}$ that converges to $\boldsymbol{\beta}^*$. We first prove a claim that if $\boldsymbol{\beta}^{m_k} - \boldsymbol{\beta}^{m_k-1} \to \mathbf{0}$, then $\boldsymbol{\beta}^*$ is a local minimizer of $\widetilde{Q}(\boldsymbol{\beta})$. Denote by $\partial f(\boldsymbol{\beta};\boldsymbol{\alpha})$ the directional derivative of a function f along the direction $\boldsymbol{\alpha}$. For any $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$, we have

$$\partial\widetilde{Q}(\boldsymbol{\beta}^*;\boldsymbol{\theta}) = \sum_{j=1}^p \frac{\partial L(\boldsymbol{\beta}^*)}{\partial\beta_j}\theta_j + \sum_{j=1}^p V_{jj}\partial p_\lambda(\beta_j^*;\theta_j) = \sum_{j=1}^p \partial\widetilde{Q}(\boldsymbol{\beta}^*;\theta_j\mathbf{e}_j), \tag{A.17}$$

where β_j^* is the jth component of $\boldsymbol{\beta}^*$ and $\mathbf{e}_j$ is the p-vector with 1 at the jth component and 0 elsewhere. Since $\beta_j^{m_k}$ minimizes $\widetilde{Q}(\beta_1^{m_k}, \ldots, \beta_{j-1}^{m_k}, \cdot, \beta_{j+1}^{m_k-1}, \ldots, \beta_p^{m_k-1})$, we have $\partial\widetilde{Q}(\boldsymbol{\alpha}_j^{m_k};\theta_j\mathbf{e}_j) \ge 0$. Since $\boldsymbol{\beta}^{m_k} - \boldsymbol{\beta}^{m_k-1} \to \mathbf{0}$, we have $\lim_{k\to\infty}\boldsymbol{\beta}^{m_k-1} = \lim_{k\to\infty}\boldsymbol{\beta}^{m_k} = \boldsymbol{\beta}^*$ and hence $\lim_{k\to\infty}\boldsymbol{\alpha}_j^{m_k} = \boldsymbol{\beta}^*$. It then follows from the upper semicontinuity of directional derivatives (Bertsekas 1999) that

$$\partial\widetilde{Q}(\boldsymbol{\beta}^*;\theta_j\mathbf{e}_j) \ge \limsup_{k\to\infty}\partial\widetilde{Q}\big(\boldsymbol{\alpha}_j^{m_k};\theta_j\mathbf{e}_j\big) \ge 0$$

for all j. In view of (A.17), we have $\partial\widetilde{Q}(\boldsymbol{\beta}^*;\boldsymbol{\theta}) \ge 0$ for all $\boldsymbol{\theta} \in \mathbb{R}^p$, so that $\boldsymbol{\beta}^*$ is a local minimizer of $\widetilde{Q}(\boldsymbol{\beta})$. It remains to show that $\boldsymbol{\beta}^{m_k} - \boldsymbol{\beta}^{m_k-1} \to \mathbf{0}$.

In fact, we will show that $\boldsymbol{\beta}^m - \boldsymbol{\beta}^{m-1} \to \mathbf{0}$. Consider the update from $\boldsymbol{\alpha}_{j-1}^m$ to $\boldsymbol{\alpha}_j^m$. The condition $\kappa(p_\lambda) < 1$ implies that $\widetilde{Q}(\boldsymbol{\beta})$ is strictly convex in β_j, so that

$$\widetilde{Q}\big(\boldsymbol{\alpha}_{j-1}^m\big) - \widetilde{Q}\big(\boldsymbol{\alpha}_j^m\big) \ge \theta\big(\beta_j^{m-1} - \beta_j^m\big) + \frac{c_0}{2}\big(\beta_j^{m-1} - \beta_j^m\big)^2,$$

for every subgradient θ of $\widetilde{Q}(\beta_1^m, \ldots, \beta_{j-1}^m, \cdot, \beta_{j+1}^{m-1}, \ldots, \beta_p^{m-1})$ at β_j^m, where $c_0 = 1 - \kappa(p_\lambda) > 0$ (Dem'yanov and Vasil'ev 1985). The optimality of β_j^m entails that 0 is a subgradient and hence

$$\widetilde{Q}\big(\boldsymbol{\alpha}_{j-1}^m\big) - \widetilde{Q}\big(\boldsymbol{\alpha}_j^m\big) \ge \frac{c_0}{2}\big(\beta_j^{m-1} - \beta_j^m\big)^2.$$

Adding both sides over $j = 1, \ldots, p$ and $m = 1, \ldots, M$ yields

$$\widetilde{Q}(\boldsymbol{\beta}^0) - \widetilde{Q}(\boldsymbol{\beta}^M) \ge \frac{c_0}{2}\sum_{m=1}^M\|\boldsymbol{\beta}^{m-1} - \boldsymbol{\beta}^m\|_2^2.$$

Noting that the sequence $\{\widetilde{Q}(\boldsymbol{\beta}^m)\}$ is bounded, we have $\sum_{m=1}^\infty\|\boldsymbol{\beta}^{m-1} - \boldsymbol{\beta}^m\|_2^2 < \infty$, so that $\boldsymbol{\beta}^m - \boldsymbol{\beta}^{m-1} \to \mathbf{0}$. Part (a) is proved.

To show part (b), we use a contradiction argument. Suppose that the sequence $\{\boldsymbol{\beta}^m\}$ does not converge to $\boldsymbol{\beta}^*$. Then there exists a subsequence $\{\boldsymbol{\beta}^{m_k}\}$ of $\{\boldsymbol{\beta}^m\}$ such that $\|\boldsymbol{\beta}^{m_k} - \boldsymbol{\beta}^*\|_2 \ge \varepsilon$ for some $\varepsilon > 0$. Since $\{\boldsymbol{\beta}^{m_k}\}$ is bounded, by taking a further subsequence if necessary, we can assume that $\{\boldsymbol{\beta}^{m_k}\}$ converges to a point $\boldsymbol{\beta}^{**}$. Clearly $\boldsymbol{\beta}^{**} \in \mathcal{K}$ since $\mathcal{K}$ is closed, and $\|\boldsymbol{\beta}^{**} - \boldsymbol{\beta}^*\|_2 \ge \varepsilon$. It follows from part (a) that $\boldsymbol{\beta}^{**}$ is a local minimizer of $\widetilde{Q}(\boldsymbol{\beta})$. This contradicts the assumption that $\boldsymbol{\beta}^*$ is the unique local minimizer in $\mathcal{K}$ and proves part (b).

A.7 Proof of Theorem 4

The inequality $\Lambda_{\min}(\mathbf{V}_{SS}) \ge \kappa(p_\lambda)\max_{j\in S}V_{jj}$ in part (a) ensures that $\widetilde{Q}(\boldsymbol{\beta})$ is convex on the subspace $\mathcal{B}_S$, from which the restricted global optimality follows. Part (a) is proved.

To show part (b), note first that the strict inequality $\Lambda_{\min}(\mathbf{V}_{SS}) > \kappa(p_\lambda)\max_{j\in S}V_{jj}$ implies strict convexity and hence the existence of a unique global minimizer $\boldsymbol{\beta}^*$ of $\widetilde{Q}(\boldsymbol{\beta})$ on $\mathcal{B}_S$. Also, since $\Lambda_{\min}(\mathbf{V}_{SS}) \le \min_{j\in S}V_{jj} \le \max_{j\in S}V_{jj}$, we have $\kappa(p_\lambda) < 1$. It then follows from Theorem 3 and the boundedness of the sequence $\{\boldsymbol{\beta}^m\}$ that the sequence $\{\boldsymbol{\beta}^m\}$ converges to $\boldsymbol{\beta}^*$. This proves part (b).

APPENDIX B: SICA-PENALIZED LEAST SQUARES IN ONE DIMENSION

In this appendix, we present an analytic form of the SICA-regularized estimator in one dimension. Consider the one-dimensional SICA-penalized least-squares problem

$$\widehat{\theta} = \arg\min_{\theta\in\mathbb{R}}\left\{\frac{1}{2}(\theta - \theta_0)^2 + p_\lambda(|\theta|)\right\},$$

where $\theta_0 \in \mathbb{R}$ and $p_\lambda(\cdot) = \lambda\rho(\cdot)$ with the SICA penalty $\rho(\cdot)$ given by (5). To find the nonzero stationary points of the objective function, it is easy to derive from the derivative equation that we need to solve the cubic equation $t^3 + c_2t^2 + c_1t + c_0 = 0$ for the positive roots, where $c_2 = 2a - |\theta_0|$, $c_1 = a^2 - 2a|\theta_0|$, and $c_0 = \lambda a(a+1) - a^2|\theta_0|$. Let $Q = (c_2^2 - 3c_1)/9$ and $R = (2c_2^3 - 9c_1c_2 + 27c_0)/54$. If $Q^3 \le R^2$, the cubic equation either has a unique, negative root or, in addition, has a real root corresponding to an inflection point; hence, $\widehat{\theta} = 0$. Otherwise, compute the two largest roots

$$t_1 = -2\sqrt{Q}\cos\left(\frac{\alpha - 2\pi}{3}\right) - \frac{c_2}{3} < t_2 = -2\sqrt{Q}\cos\left(\frac{\alpha + 2\pi}{3}\right) - \frac{c_2}{3},$$

where $\alpha = \arccos(R/\sqrt{Q^3})$. If $t_1 > 0$, then t_1 is a local maximum and t_2 is a local minimum; by comparing the values of the objective function at 0 and t_2, we have that $\widehat{\theta} = \mathrm{sgn}(\theta_0)t_2$ if $t_2/2 + \lambda(a+1)/(a+t_2) < \theta_0$, and $\widehat{\theta} = 0$ otherwise. If $t_1 < 0$ and $t_2 > 0$, then t_2 is a local minimum and $\widehat{\theta} = \mathrm{sgn}(\theta_0)t_2$. Otherwise, the cubic equation has no positive roots and $\widehat{\theta} = 0$.

In summary, $\widehat{\theta} = \mathrm{sgn}(\theta_0)t_2$ if $Q^3 > R^2$, $t_1 > 0$, and $t_2/2 + \lambda(a+1)/(a+t_2) < \theta_0$, or $Q^3 > R^2$, $t_1 < 0$, and $t_2 > 0$; otherwise $\widehat{\theta} = 0$.

This gives a complete characterization of the SICA solution $\widehat{\theta}$ in one dimension.

[Received August 2011. Revised June 2012.]

REFERENCES

Andersen, P. K., and Gill, R. D. (1982), "Cox's Regression Model for Counting Processes: A Large Sample Study," *The Annals of Statistics*, 10, 1100–1120. [263]

Antoniadis, A., Fryzlewicz, P., and Letué, F. (2010), "The Dantzig Selector in Cox's Proportional Hazards Model," *Scandinavian Journal of Statistics*, 37, 531–552. [247]

Bertsekas, D. P. (1999), *Nonlinear Programming* (2nd ed.), Belmont, MA: Athena Scientific. [263]

Bradic, J., Fan, J., and Jiang, J. (2011), "Regularization for Cox's Proportional Hazards Model With NP-Dimensionality," *The Annals of Statistics*, 39, 3092–3120. [248]

Breheny, P., and Huang, J. (2011), "Coordinate Descent Algorithms for Nonconvex Penalized Regression, With Applications to Biological Feature Selection," *The Annals of Applied Statistics*, 5, 232–253. [252]

Breslow, N. E., and Day, N. E. (1987), *Statistical Models in Cancer Research, 2: The Design and Analysis of Cohort Studies*, Lyon: IARC. [247]

Cai, J., Fan, J., Li, R., and Zhou, H. (2005), "Variable Selection for Multivariate Failure Time Data," *Biometrika*, 92, 303–316. [247]

Cox, D. R., and Oakes, D. (1984), *Analysis of Survival Data*, London: Chapman & Hall. [247]

Daubechies, I., Defrise, M., and De Mol, C. (2004), "An Iterative Thresholding Algorithm for Linear Inverse Problems With a Sparsity Constraint," *Communications on Pure and Applied Mathematics*, 57, 1413–1457. [252]

Dem'yanov, V. F., and Vasil'ev, L. V. (1985), *Nondifferentiable Optimization*, New York: Springer. [263]

Fan, J. (1997), Comments on "Wavelets in Statistics: A Review," by A. Antoniadis, *Journal of the Italian Statistical Society*, 6, 131–138. [249]

Fan, J., and Fan, Y. (2008), "High-Dimensional Classification Using Features Annealed Independence Rules," *The Annals of Statistics*, 36, 2605–2637. [248]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [248,249,250,252,253]

——— (2002), "Variable Selection for Cox's Proportional Hazards Model and Frailty Model," *The Annals of Statistics*, 30, 74–99. [247]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space" (with discussion), *Journal of the Royal Statistical Society,* Series B, 70, 849–911. [258]

——— (2010), "A Selective Overview of Variable Selection in High Dimensional Feature Space (invited review article)," *Statistica Sinica*, 20, 101–148. [247]

——— (2011), "Nonconcave Penalized Likelihood With NP-Dimensionality," *IEEE Transactions on Information Theory*, 57, 5467–5484. [249,251,252,253]

Fan, J., Lv, J., and Qi, L. (2011), "Sparse High-Dimensional Models in Economics," *Annual Review of Economics*, 3, 291–317. [247]

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *The Annals of Applied Statistics*, 1, 302–332. [252]

Fu, W. J. (1998), "Penalized Regressions: The Bridge Versus the Lasso," *Journal of Computational and Graphical Statistics*, 7, 397–416. [252]

Hoeffding, W. (1963), "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association*, 58, 13–30. [260]

Horn, R. A., and Johnson, C. R. (1985), *Matrix Analysis*, New York: Cambridge University Press. [261]

Jarrow, R. A. (2009), "Credit Risk Models," *Annual Review of Financial Economics*, 1, 37–68. [247]

Kosorok, M. R. (2008), *Introduction to Empirical Processes and Semiparametric Inference*, New York: Springer. [259]

Leng, C., Lin, Y., and Wahba, G. (2006), "A Note on the Lasso and Related Procedures in Model Selection," *Statistica Sinica*, 16, 1273–1284. [254]

Leng, C., and Ma, S. (2007), "Path Consistent Model Selection in Additive Risk Model via Lasso," *Statistics in Medicine*, 26, 3753–3770. [248]

Lin, D. Y., and Ying, Z. (1994), "Semiparametric Analysis of the Additive Risk Model," *Biometrika*, 81, 61–71. [247,248,249]

Lv, J., and Fan, Y. (2009), "A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares," *The Annals of Statistics*, 37, 3498–3528. [248,249,250,251]

Martinussen, T., and Scheike, T. H. (2009), "Covariate Selection for the Semiparametric Additive Risk Model," *Scandinavian Journal of Statistics*, 36, 602–619. [248]

Massart, P. (2000), "About the Constants in Talagrand's Concentration Inequalities for Empirical Processes," *The Annals of Probability*, 28, 863–884. [259,260]

Mazumder, R., Friedman, J. H., and Hastie, T. (2011), "*SparseNet*: Coordinate Descent With Nonconvex Penalties," *Journal of the American Statistical Association*, 106, 1125–1138. [252,253]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 34, 1436–1462. [254]

——— (2010), "Stability Selection" (with discussion), *Journal of the Royal Statistical Society,* Series B, 72, 417–473. [259]

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., and Staudt, L. M. (2002), "The Use of Molecular Profiling to Predict Survival After Chemotherapy for Diffuse Large-B-Cell Lymphoma," *The New England Journal of Medicine*, 346, 1937–1947. [258]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [249]

——— (1997), "The Lasso Method for Variable Selection in the Cox Model," *Statistics in Medicine*, 16, 385–395. [247]

van der Vaart, A. W. (1998), *Asymptotic Statistics*, New York: Cambridge University Press. [259,260]

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York: Springer. [259,260]

Wainwright, M. J. (2009), "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using ℓ_1-Constrained Quadratic Programming (Lasso)," *IEEE Transactions on Information Theory*, 55, 2183–2202. [248,250,252]

Wu, T. T., and Lange, K. (2008), "Coordinate Descent Algorithms for Lasso Penalized Regression," *The Annals of Applied Statistics*, 2, 224–244. [252]

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [249]

Zhang, H. H., and Lu, W. (2007), "Adaptive Lasso for Cox's Proportional Hazards Model," *Biometrika*, 94, 691–703. [247]

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563. [248,249]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [252]

——— (2008), "A Note on Path-Based Variable Selection in the Penalized Proportional Hazards Model," *Biometrika*, 95, 241–247. [247]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society,* Series B, 67, 301–320. [254]

Zou, H., and Li, R. (2008), "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models" (with discussion), *The Annals of Statistics*, 36, 1509–1566. [252]