36-702 Final Project: Estimation and Inference for High-dimensional Proportional Hazards Model

Wanshan Li*

Department of Statistics& Data Science
Carnegie Mellon University
Pittsburgh, PA 15213
wanshanl@andrew.cmu.edu

Abstract

Cox model plays a central role in survival analysis and has widespread applications. With the development of high-dimensional statistics, the estimation and inference problem for Cox model in high-dimensional cases become necessary. Some theoretical results were established. The basic intuitions under the theory follow that of linear model. However, the intrinsic complexity of Cox model makes the theoretical analysis more involved. This report focuses on this important and interesting problem.

1 Introduction

The proportional hazards model proposed by David Cox in his famous paper in 1972 is perhaps the most important model specifically focusing on time to event data. Due to the wide-spread application in epidemiology, medicine, economics, sociology, and industry, it is one of the most-cited paper ever (Actually, according to a survey by Nature in 2014, it is the number 24). The classic theory of Cox model is based on fixed dimensions. Two representative works on the classic theory of Cox model are [1] and [11]. [1], the foundation of almost all theoretical works from then on, established the large sample theory for Cox model based on counting process and martingale theory. While [11] rebuilt the whole theoretical framework in terms of empirical processes and extend the Cox model to model recurrent events. They both extend the original version of Cox model by formalizing it under the framework of counting process. Such formalization makes Cox model even more flexible for applications.

The theory for variable selection in Cox model is established much later, and much less than the linear model. For fixed dimension, [5] and [12] study the Lasso penalized and SCAD penalized Cox model respectively and give some theoretical results. They prove that by properly choosing the tuning parameter in the penalization terms, one can get a $\sqrt{n}$-consistent estimator for the model parameter β and the nonzeros part of β is $\sqrt{n}$-asymptotically normal. For high-dimension, [7] and [4] propose the EN-Cox model to do variable selection for $p = O(\exp(n^{\alpha}))$, but do not do much in theory. The EN-Cox model is just to use Elastic Net to do variable selection for Cox model.

Recently several papers, including [3], [8], and [10], developed the theory for the variable selection of Cox model in high-dimensional cases. [3] and [8] analyse the theoretical properties of general penalties (e.g., SCAD) and Lasso on Cox model, under different conditions. [8] only prove the oracle inequalities, or bounding the estimation error of the estimator $\hat{\beta}$, while [3] also study the asymptotic distribution of the estimator. [8] use two key quantities, the compatibility and cone invertibility factors

*

xxst Conference on Neural Information Processing Systems (NIPS 20xx), xxxx, xx, xxx.

of the Hessian of the negative log-partial likelihood, in the proof. Thus in spirit [8] follows the idea of [15] and [16]. [3] considers a more general case, where the penalized log-partial likelihood can be non-concave, but just have some weaker "local concavity". Since they do not assume concavity, some careful study on the identification of β and choice of local maximizer is necessary, which makes the proof more difficult. Also, this needs more and stronger conditions than [8]. [10] study the non-asymptotic oracle inequality for Lasso-Cox model, under different conditions from that of [8]. They follow an approach of [13] and get the prediction and ℓ_1 error bounds.

There are much less work in high-dimensional Cox model than linear models, and even less is known about the inference theory of Cox model in high-dimensional cases. [6] study this problem and propose a unified likelihood ratio inferential framework, applicable for score, Wald and partial likelihood ratio statistics. They derive the asymptotic distributions of these statistics, even without the assumption of model selection consistency, and prove that they are semiparametrically optimal. They also provide confidence interval for the baseline hazard function as well as conditional hazard function. They show the control of type I error of all tests by simulation studies. In words, they developed a all-round framework for testing of high-dimensional Cox model.

2 Notations and Assumptions

Consider an n dimensional counting process $\mathbf{N}^{(n)}(t) = (N_1(t), \cdots, N_n(t)),\ t \geq 0$, where $N_i(t)$ counts the number of observed events for the i-th individual in the time interval $[0, t]$. Suppose we have $Y_i(t) \in \{0, 1\}$, a predictable at risk indicator process, and $\mathbf{Z}_i(t) = (Z_{i,1}(t), \cdots, Z_{i,p}(t))'$, a p-dimensional predictable covariate process. For $t \geq 0$, let $\mathcal{F}_t = \sigma\{N_i(s), Y_i(s), \mathbf{Z}_i(s) : s \leq t, i = 1, \cdots, n\}$. Assume that for $\{\mathcal{F}_t, t \geq 0\}$, $\mathbf{N}^{(n)}$ has predictable compensator $\mathbf{\Lambda} = (\Lambda_1, \cdots, \Lambda_n)$ with

$$d\Lambda_i(t) = Y_i(t) \exp\{\mathbf{Z}_i(s)'\boldsymbol{\beta}^o\} d\Lambda_0(t), \tag{1}$$

where $\boldsymbol{\beta}^0$ is a p-vector of true regression coefficients, and Λ_0 is an unknown baseline cumulative hazard function. Denote the log-partial likelihood for the survival experience at time t as $C(\beta, t) = \log L(\beta)$, given by

$$C(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} \int_0^t \mathbf{Z}_i(s)'\boldsymbol{\beta} dN_i(s) - \int_0^t \log \left\{ \sum_{i=1}^{n} Y_i(s) e^{\mathbf{Z}_i(s)'\boldsymbol{\beta}} \right\} d\bar{N}(s), \tag{2}$$

where $\bar{N}(s) = \sum_{i=1}^{n} N_i(s)$. The log-partial likelihood function is defined as $C(\boldsymbol{\beta}, \infty) = \lim_{t\to\infty} C(\boldsymbol{\beta}, t)$. Denote $\ell(\boldsymbol{\beta}) = -C(\boldsymbol{\beta}, \infty)/n$. As people usually do([1],[9],[11], [8]), we introduce following notations (For a vector v, $v^{\otimes 0} = 1$, $v^{\otimes 1} = v$, $v^{\otimes 2} = v'v$):

$$\begin{aligned}
S^{(k)}(t, \boldsymbol{\beta}) &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i(t)^{\otimes k} Y_i(t) \exp\{\mathbf{Z}_i(t)'\boldsymbol{\beta}\}, \quad k = 0, 1, 2, \\
R_n(t, \boldsymbol{\beta}) &= \frac{1}{n} \sum_{i=1}^{n} Y_i(t) \exp\{\mathbf{Z}_i(t)'\boldsymbol{\beta}\}, \quad \bar{\mathbf{Z}}_n(\boldsymbol{\beta}, t) = \frac{S^{(1)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)}, \\
V_n(t, \boldsymbol{\beta}) &= \sum_{i=1}^{n} w_{ni}(t, \boldsymbol{\beta})(\mathbf{Z}_i(t) - \bar{\mathbf{Z}}_n(\boldsymbol{\beta}, t))^{\otimes 2} = \frac{S^{(2)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)} - \bar{\mathbf{Z}}_n(\boldsymbol{\beta}, t)^{\otimes 2},
\end{aligned} \tag{3}$$

where $w_{ni}(t, \boldsymbol{\beta}) = Y_i(t) \exp\{\mathbf{Z}_i(t)'\boldsymbol{\beta}\}/[nS^{(0)}(t, \boldsymbol{\beta})]$. Notice that $S^{(0)} \in \mathbb{R}$,$S^{(0)} \in \mathbb{R}^p$,$S^{(0)} \in \mathbb{R}^{p\times p}$. $\bar{\mathbf{Z}}_n(\boldsymbol{\beta}, t)$ is a weighted average of $\mathbf{Z}_i(\boldsymbol{\beta}, t)$. By basic calculations, the score function and the Hessian matrix of $\ell()$ can be written as

$$\begin{aligned}
\dot{\ell}(\boldsymbol{\beta}, t) &\triangleq \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{1}{n} \sum_{i=1}^{n} \{ \int_0^{\infty} Z_i(s) - \bar{\mathbf{Z}}_n(\boldsymbol{\beta}, t) \} dN_i(s), \\
\ddot{\ell}(\boldsymbol{\beta}, t) &\triangleq \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \frac{1}{n} \int_0^{\infty} V_n(s, \boldsymbol{\beta}) d\bar{N}(s).
\end{aligned}$$

A powerful tool that is frequently used in the theory of Cox model is the martingale theory. By definition, Λ_i are compensators, thus

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) \exp\{\mathbf{Z}_i(s)'\boldsymbol{\beta}^o\} d\Lambda_0(s), \quad 1 \leq i \leq n, t \geq 0, \tag{4}$$

are (local) martingales with predictable variation/covariation processes

$$\langle M_i, M_i\rangle(t) = \int_0^t Y_i(s)\exp\{\mathbf{Z}_i(s)'\boldsymbol{\beta}^o\}d\Lambda_0(s) \text{ and } \langle M_i, M_j\rangle = 0,\ i \neq j. \tag{5}$$

The main idea of variable selection is to add a penalization function $p_\lambda(\boldsymbol{\beta})$ when minimizing $\ell(\boldsymbol{\beta})$. Therefore, the estimator $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, L(\boldsymbol{\beta}, \lambda) \triangleq \ell(\boldsymbol{\beta}) + p_\lambda(\boldsymbol{\beta}). \tag{6}$$

3 Main Results for Estimation and Inference

3.1 Lasso-Cox Model

I will only discuss [8] in this report. For lasso-Cox model, $p_\lambda(\boldsymbol{\beta}) = \lambda|\boldsymbol{\beta}|_1$ is convex, and (6) is a convex optimization problem. Therefore we can use the Karush-Kuhn-Tucker conditions and get that a vector $\hat{\boldsymbol{\beta}}$ is a solution to (6)if and only if

$$\begin{cases} \dot{\ell}_j(\hat{\boldsymbol{\beta}}) = -\lambda \mathrm{sgn}(\hat{\beta}_j), & \text{if} \hat{\beta}_j \neq 0, \\ |\dot{\ell}_j(\hat{\boldsymbol{\beta}})| \leq \lambda, & \text{if} \hat{\beta}_j = 0. \end{cases} \tag{7}$$

Such KKT conditions make it possible to analyse the lasso in the Cox model in a similar way to the lasso in the linear regression model. This is the main idea of [8].

Their main results are Theorem 3.1, Theorem 3.2 and Theorem 4.1. Let $\xi > 1$, and $\mathcal{O} = \{j : \boldsymbol{\beta}_j \neq 0\}$. Define the compatibility factor ([15]) $\kappa(\xi, \mathcal{O}; \bar{\Sigma}) = \inf_{0\neq\mathbf{b}\in\mathscr{C}(\xi,\mathcal{O})} \frac{d_o^{1/2}(\mathbf{b}'\bar{\Sigma}\mathbf{b})^{1/2}}{|\mathbf{b}_\mathcal{O}|_1}$ and $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o|_q$ in terms of the weak cone invertibility factor ([16]) $F_q(\xi, \mathcal{O}; \bar{\Sigma}) = \inf_{0\neq\mathbf{b}\in\mathscr{C}(\xi,\mathcal{O})} \frac{d_o^{1/q}(\mathbf{b}'\bar{\Sigma}\mathbf{b})}{|\mathbf{b}_\mathcal{O}|_1|\mathbf{b}|_q}$. These two concepts are intimately related to the restricted eigenvalue ([2])

$$\mathrm{RE}(\xi, \mathcal{O}; \bar{\Sigma}) = \inf_{0\neq\mathbf{b}\in\mathscr{C}(\xi,\mathcal{O})} \frac{(\mathbf{b}'\bar{\Sigma}\mathbf{b})^{1/2}}{|\mathbf{b}|_2}. \tag{8}$$

The quantity RE was used to establish oracle inequalities in linear regression ([2]). In linear regression $\bar{\Sigma}$ is the Hessian of the squared loss $|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|_2^2/(2n)$. Such idea can be generalized to Cox model, where $\bar{\Sigma}$ is the Hessian of the log-partial likelihood (2), as we will show now.

Theorem 3.1 and 3.2 establish the oracle inequalities bounding the error measured by the ℓ_q norm $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o|_1$ and the symmetric Bregman divergence

$$\mathbf{D}^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^o) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\dot{\ell}(\hat{\boldsymbol{\beta}}) - \dot{\ell}(\boldsymbol{\beta}))$$

in terms of $\kappa(\xi, \mathcal{O}) = \kappa(\xi, \mathcal{O}; \ddot{\ell}(\boldsymbol{\beta}^o))$ and $F_q(\xi, \mathcal{O}) = F_q(\xi, \mathcal{O}; \ddot{\ell}(\boldsymbol{\beta}^o))$. The reason for using κ and F is that they can yield sharper oracle inequalities. They also need a condition for the covariate process $Z(t)$:

$$\max_{i<i'\leq n} \sup_{0\leq t<\infty} \max_{j\leq p} |Z_{i,j}(t) - Z_{i',j}(t)| \leq K. \tag{9}$$

Theorem 3.1 (Theorem 3.1)**.** *Let $\tau = K(\xi+1)d_o\lambda/\{2\kappa^2(\xi, \mathcal{O})\}$ with a certain $K > 0$. Suppose condition* (9) *holds and $\tau \leq 1/e$. Then, in the event $|\dot{\ell}(\boldsymbol{\beta}^o)|_\infty \leq (\xi-1)/(\xi+1)\lambda$,*

$$D^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \leq \frac{4e^\eta(1+1/\xi)^{-2}\lambda^2 d_o}{\kappa^2(\xi, \mathcal{O})}, \quad |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1 \leq \frac{e^\eta(1+\xi)\lambda d_o}{2\kappa^2(\xi, \mathcal{O})} \tag{10}$$

and

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o|_q \leq \frac{2e^\eta(1+1/\xi)\lambda d_o^{1/q}}{F_q(\xi, \mathcal{O})}, \quad q \geq 1, \tag{11}$$

where $\eta \leq 1$ is the smaller solution of $\eta e^\eta = \tau$.

Theorem 3.1 bounds the error on a specific event. Theorem 3.2 further write the bounds in a probability version. Also, one should notice that $\kappa(\xi, \mathcal{O})$ and $\mathbf{F}_q(\xi, \mathcal{O})$ appearing in the right-hand-side of bounds are random variables. In theorem 3.3 we will find lower bound for them and make the control the error by constants.

Theorem 3.2 (Theorem 3.2)**.** *Suppose condition* (9) *holds and* $N_i(\infty) \le 1$ *for all* $i \le n$ *and* $t \ge 0$*. Let* $\xi > 1$ *and* $\lambda = \{(\xi+1)/(\xi-1)\}K\sqrt{(2/n)\log(2p/\varepsilon)}$ *with a small* $\varepsilon > 0$ *(e.g.,* $\varepsilon = 1\%$*). Let* $C_\kappa > 0$ *satisfying* $\tau = K(\xi+1)d_o\lambda/(2C_\kappa^2) \le 1/e$*. Let* $\eta \le 1$ *be the smaller solution of* $\eta e^{-\eta} = \tau$*. Then, for any* $C_{F,q} > 0$*,*

$$D^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \le \frac{4e^\eta \xi^2 \lambda^2 d_o}{(1+1/\xi)^2 C_\kappa^2}, \quad |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1 \le \frac{e^\eta(1+\xi)\lambda d_o}{2C_\kappa^2} \tag{12}$$

and

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o|_q \le \frac{2e^\eta \xi \lambda d_o^{1/q}}{(1+1/\xi)C_{F,q}}, \quad q \ge 1 \tag{13}$$

all hold with at least probability $\mathbb{P}\{\kappa(\xi, \mathcal{O}) \ge C_\kappa, F_q(\xi, \mathcal{O}) \ge C_{F,q}\} - \varepsilon$*.*

Theorem 4.1 establishes bounds on $\kappa(\xi, \mathcal{O})$ and $F_q(\xi, \mathcal{O})$ (both are random variables) so that they can be treated as constants in the bounds in Theorem 3.1 and 3.2. To state the result, let's first introduce the matrix

$$\Sigma(t^*; M) = \mathbb{E}\int_0^{t^*} \mathbf{G}_n(s; M) d\Lambda_0(s), \tag{14}$$

where $\mathbf{G}_n(t; M) = n^{-1}\sum_{i=1}^n \{\mathbf{Z}_i - \boldsymbol{\mu}(t; M)\}^{\otimes 2} Y_i(t) \min\{M, \exp(\mathbf{Z}_i'(t)\boldsymbol{\beta}^o)\}$ with

$$\boldsymbol{\mu}(t; M) = \frac{\mathbb{E}[\mathbf{Z}(t)Y(t)\min\{M, \exp(\mathbf{Z}'\boldsymbol{\beta}^o)\}]}{\mathbb{E}[Y(t)\min\{M, \exp(\mathbf{Z}'\boldsymbol{\beta}^o)\}]}.$$

Theorem 3.3 (Theorem 4.1)**.** *Suppose* $\xi \ge 1$*,* $\mathcal{O} \subset \{1, \cdots, p\}$ *with* $|\mathcal{O}| = d_o$*. Let* $\{Y_i(t), \mathbf{Z}_i(t), t \ge 0\}$ *be i.i.d. processes form* $\{Y(t), \mathbf{Z}(t), t \ge 0\}$ *with* $\sup_t \mathbb{P}\{|\mathbf{Z}_i(t) - \mathbf{Z}(t)|_\infty \le K\} = \mathbb{P}\{\max_i N_i(\infty) \le 1\} = 1$*. Let* $\{t^*, M\}$ *be positive constants,* $r_* = \mathbb{E}Y(t^*)\min\{M, \exp(\mathbf{Z}'(t)\boldsymbol{\beta}^o)\}$*, and* $L_n(t) = \sqrt{(2/n)\log t}$*. Denote* $\phi(\xi, \mathcal{O}; \ddot{\boldsymbol{\beta}}^o)$ *for either* $\kappa^2(\xi, \mathcal{O})$ *or* $F_q(\xi, \mathcal{O})$*. Then,*

$$\phi(\xi, \mathcal{O}; \ddot{\boldsymbol{\beta}}^o) \ge \phi(\xi, \mathcal{O}; \Sigma(t^*; M)) - d_o(\xi+1)^2 K^2 \{C_1 L_n(p(p+1)/\varepsilon) + C_2 t_{n,p,\varepsilon}^2\} \tag{15}$$

with probability at least $1 - 3\varepsilon$*, where* $C_1 = 1 + \Lambda_0(t^*)$*,* $C_2 = (2/r_*)\Lambda_0(t^*)$ *and* $t_{n,p,\varepsilon}$ *is the solution of* $p(p+1)\exp\{-nt_{n,p,\varepsilon}^2/(2 + 2t_{n,p,\varepsilon}/3)\} = \varepsilon/2.221$*. Consequently, for* $1 \le q \le 2$*,*

$$\begin{aligned} &\min\{\kappa^2(\xi, \mathcal{O}), (1+\xi)^{2/q-1} F_q(\xi, \mathcal{O})\} \\ &\quad \ge \mathrm{RE}^2(\xi, \mathcal{O}; \ddot{\ell}(\boldsymbol{\beta}^o)) \\ &\quad \ge \rho_* - d_o(\xi+1)^2 K^2 \{C_1 L_n(p(p+1)/\varepsilon) + C_2 t_{n,p,\varepsilon}^2\} \end{aligned} \tag{16}$$

with probability at least $1 - 3\varepsilon$*, where* $\rho_* = \phi_{\min}(\Sigma(t^*; M))$ *with the matrix* $\Sigma(t^*; M)$ *in* (14)

Remark In the right-hand-side of (16), $C_2 t_{n,p,\varepsilon}^2$ is of smaller order than $L_n(p(p+1)/\varepsilon)$. By theorem 3.2, when $p = \exp(o(n/d_o^2))$, we have $d_o\lambda \to 0$ as $n \to \infty$, which means $d_o\sqrt{(\log p)/n}$ is sufficiently small. Therefore when $p = \exp(o(n/d_o^2))$, the right-hand-side of (16) can be treated as $\rho_*/2$. Since ρ_* is the smallest eigenvalue of the population integrated covariance matrix in (14), we can treat it as a constant. Then we can choose constant $C_\kappa, C_{F,q}$ to make $\mathbb{P}\{\kappa(\xi, \mathcal{O}) \ge C_\kappa, F_q(\xi, \mathcal{O}) \ge C_{F,q}\} \ge 1 - 3\varepsilon$ and hence make the error bound (13) holds with probability at least $1 - 4\varepsilon$. So theorem 3.2 and 3.3 together give a high-probability upper bound for the estimation error.

3.2 SCAD-Cox Model

[3] study the problem (6) with general $p_\lambda(\boldsymbol{\beta})$, e.g., the SCAD penalty, defined via its derivative $p'(\lambda) = \lambda\{I(t \le \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda}\}I(t > \lambda)$, $t \ge 0$, for some $a > 2$. Such a general $p_\lambda(\boldsymbol{\beta})$ may not be convex globally, and consequently (6) may not have a global minimizer. To deal with this issue, they loose the condition of (global) convexity to local convexity and prove a set of conditions for local minimizers under the local convexity in Theorem 2.1, as a counterpart of KKT conditions in the case of global convexity. For convenience denote

$$\mathcal{C}(\boldsymbol{\beta}, \tau) = \sum_{i=1}^n \int_0^\tau \mathbf{Z}_i(s)'\boldsymbol{\beta} dN_i(s) - \int_0^\tau \log\left\{\sum_{i=1}^n Y_i(s) e^{\mathbf{Z}_i(s)'\boldsymbol{\beta}}\right\} d\bar{N}(s), \tag{17}$$

Suppose that $\boldsymbol{\beta}^o = ((\boldsymbol{\beta}_1^o)', \mathbf{0}')'$, and $|\boldsymbol{\beta}_{1j}^o| > 0$, $j = 1, \cdots, s$. Thus there are s signals in $\boldsymbol{\beta}^o$. As we do in linear model, we introduce $\hat{\boldsymbol{\beta}}^{o} = ((\hat{\boldsymbol{\beta}}_1^{o})', \mathbf{0}')'$, called the oracle estimator, in which $\hat{\boldsymbol{\beta}}_1^{o}$ is defined as a local minimizer of (17) over a subspace $\Omega_s = \{\boldsymbol{\beta} : \beta_j = 0, j = s+1, \cdots, p\}$. In words,

$$\hat{\boldsymbol{\beta}}_1^{o} = \text{argmax}\{\mathcal{C}(\boldsymbol{\beta}_1, \tau) : \boldsymbol{\beta}_1 \in \Omega_s\} \text{ with } \mathcal{C}(\boldsymbol{\beta}_1, \tau) = \mathcal{C}((\boldsymbol{\beta}_1, \mathbf{0}), \tau).$$

Theorem 4.1 proves that under some conditions $\hat{\boldsymbol{\beta}}^{o}$ is the unique global minimizer of the penalized log-partial likelihood $\mathcal{C}(\boldsymbol{\beta}_1, \tau)$ in Ω_s.

In Theorem 4.2 they prove a bound for the error $\|\hat{\boldsymbol{\beta}}^{o} - \boldsymbol{\beta}^o\|$ of the oracle estimator $\hat{\boldsymbol{\beta}}^{o}$.

Theorem 3.4 (Theorem 4.2 (Estimation loss))**.** *Under Condition 1, 2 and 4, 5 in [3], with probability tending to one, there exists an oracle estimator $\hat{\boldsymbol{\beta}}^{o}$ such that*

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

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

In Theorem 4.3 they prove the strong oracle inequality, stating that there is a local minimizer in (6) such that $P(\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{o}) \geq 1 - c_0(p-s)\exp\{-c_1 n^{c_2}\}$ for some positive constants c_0, c_1, c_2. Therefore, these three theorems together give us a description of the estimator given in (6).

Theorem 3.5 (Theorem 4.3 (Strong oracle))**.** *Let the oracle estimator $\hat{\boldsymbol{\beta}}^{o}$ be a local maximizer of $\mathcal{C}(\boldsymbol{\beta}_1, \tau)$ given by theorem 3.4. If $\max_j(\sigma_j^2) = O(n^{(0.5\alpha+\alpha_1-1)_+ + \alpha_2})$, and Conditions $1-8$ in [3] hold, then with probability tending to one, there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $\mathcal{C}(\boldsymbol{\beta}, \tau)$ such that*

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

where c_0 and c_1 are positive constants.

Remark The result above is stated as the existence of a nice local minimizer $\hat{\boldsymbol{\beta}}$ of (6). Such uncertainty is the price to pay for the more general choice of $p_\lambda(t)$. Theorem 3.4 and 3.5 together guarantee that there is an estimator $\hat{\boldsymbol{\beta}}$ such that with high probability (tending to 1 exponentially fast), the entries of $\hat{\boldsymbol{\beta}}$ for $j = s+1, \cdots, p$ are all zeros and the first s entries, the signal part, has error bound

$$\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 = O_P\{\sqrt{s}(n^{-1/2} + \lambda_n \rho'(\beta_n^*))\}. \tag{18}$$

Notice that they do not give error bound for the whole estimator $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|$, and I think the reason for that are 1. when dealing with the non-signal part, apart from the probability bound of $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{o}$, we also need some bounds for the entries of $\hat{\boldsymbol{\beta}}$, which can be restrictive; 2. Even if we add some boundedness conditions, the bounds for the non-signal part may not be written in a same form as (18). This is a drawback of their paper.

In Theorem 4.4 and 4.5 they give oracle inequalities for Lasso and SCAD on Cox model, respectively. Briefly speaking, for the lasso estimator, the ℓ_2 error for the signal part is

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

for λ_n s.t. $\sqrt{s}\lambda_n \to 0$, $\lambda_n \gg n^{-1/2+\alpha_2}$(for α_2, see 3.5), while for the SCAD estimator, and $\lambda \gg n^{-1/2+(0.5\alpha_2+\alpha_1-1)_+ + \alpha_2}$ the error is

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

In Theorem 4.6 they prove the asymptotic distribution of $\hat{\boldsymbol{\beta}}_1$.

Theorem 3.6 (Theorem 4.6)**.** *Under Conditions 1-8 in [3], and for $\lambda_n \rho'(\beta_n^*) = o((sn)^{-1/2})$ for any $s \times 1$ unit vector $\mathbf{b}_n$, if $s = o(n^{1/3})$, the penalized partial likelihood estimator $\hat{\boldsymbol{\beta}}_1$ from* (18) *satisfies*

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

3.3 Results for Inference

[6] establish three testing procedures, as high-dimensional counterparts of the Wald, conventional score, and partial likelihood ratio tests. In the paper they just show the results for testing a single component of $\boldsymbol{\beta}$, and they give the results for testing multi-dimensional parameter in the supplementary materials. For convenience I will just write about the single component. Suppose $\boldsymbol{\beta} = (\alpha, \boldsymbol{\theta}')' \in BR^d$, where $\alpha = \beta_1 \in \mathbb{R}$ is the parameter of interest. They consider testing $H_0 : \alpha^* = 0$ vs. $H_1 : \alpha^* \neq 0$. Define

$$\mathbf{w}^* = \underset{\mathbf{w}\in\mathbb{R}^{d-1}}{\operatorname{argmin}} \mathbb{E}\{\nabla_\alpha \ell(0, \boldsymbol{\theta}^*) - \mathbf{w}'\nabla_{\boldsymbol{\theta}}\ell(0, \boldsymbol{\theta}^*)\}^2, \tag{19}$$

and let $\hat{\mathbf{w}}$ be a (lasso-type) estimator of $\mathbf{w}^*$, with tuning parameter λ' (corresponds to the penalization term $\lambda'\|w\|_1$). Also denote the lasso-type estimator of $\boldsymbol{\beta}^*$ with tuning parameter λ as $\hat{\boldsymbol{\beta}}$. The decorrelated score function and its derivative w.r.t. α are defines as

$$\hat{U}(\alpha, \hat{\boldsymbol{\theta}}) = \nabla_\alpha \ell(0, \hat{\boldsymbol{\theta}}) - \hat{\mathbf{w}}\nabla_{\boldsymbol{\theta}}\ell(0, \hat{\boldsymbol{\theta}}),\ \hat{H}_{\alpha|\boldsymbol{\theta}} = \nabla^2_{\alpha\alpha}\ell(\hat{\alpha}, \hat{\boldsymbol{\theta}}) - \hat{\mathbf{w}}\nabla_{\boldsymbol{\theta}\alpha}\ell(\hat{\alpha}, \hat{\boldsymbol{\theta}}). \tag{20}$$

Then we can define three test statistics.

1. The decorrelated score test statistic: $\hat{S}_n = n\hat{H}_{\alpha|\boldsymbol{\theta}}\hat{U}^2(0, \hat{\boldsymbol{\theta}})$.
2. The decorrelated Wald test statistic: $\hat{W}_n = n\hat{H}_{\alpha|\boldsymbol{\theta}}\tilde{\alpha}$, where $\tilde{\alpha} = \hat{\alpha} - \{\frac{\partial\hat{U}(\hat{\alpha},\hat{\boldsymbol{\theta}})}{\partial\alpha}\}^{-1}\hat{U}(\hat{\alpha}, \hat{\boldsymbol{\theta}})$.
3. The decorrelated partial likelihood ration test statistic: $\hat{L}_n = 2n\{\ell_{decor}(0) - \ell_{decor}(\tilde{\alpha})\}$, where $\ell_{decor}(\alpha) = \ell(\alpha, \hat{\boldsymbol{\theta}} - \alpha\hat{\mathbf{w}})$.

The asymptotic property of these three statistics are proved in theorem 1-3 in their paper. In the following statement, the quantity s is the number of signals among d entries of $\boldsymbol{\beta}^*$ (true parameter), or $s = \|\boldsymbol{\beta}^*\|_0$, under assumptions 1-2 in their paper. Besides, let $\mathbf{H}(\boldsymbol{\beta})$ be the Fisher information matrix based on the partial likelihood, given by

$$\mathbf{H}(\boldsymbol{\beta}) = \mathbb{E}\left[\int_0^\tau \left\{\frac{S^{(2)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)} - \mathbf{e}(\boldsymbol{\beta}, t)^{\otimes 2}\right\} dN(t)\right], \tag{21}$$

and let $\mathbf{H}^* = \mathbf{H}(\boldsymbol{\beta}^*)$. Also let $\mathbf{H}^*_{\alpha\boldsymbol{\theta}}, \mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}}, \mathbf{H}^*_{\boldsymbol{\theta}\alpha}$ be the partitions of $\mathbf{H}^*$ corresponds to the index of α and $\boldsymbol{\theta}$.

Theorem 3.7 (Brief Summary of Theorem 1,2,3)**.** *Suppose assumptions 1-5 in [6] hold. If* $\lambda \asymp \sqrt{n^{-1}\log(d)}$, $\lambda' \asymp \sqrt{n^{-1}\log(d)}$ *and* $n^{-1/2}s\log(d) = o(1)$, *under the null hypothesis* $\alpha^* = 0$, *we have*

$$\sqrt{n}\hat{U}(0, \hat{\theta}) \xrightarrow{d} Z, \quad Z \sim N(0, H_{\alpha|\theta}), \tag{22}$$

$$\sqrt{n}\tilde{\alpha} \xrightarrow{d} Z, \quad Z \sim N(0, H_{\alpha|\theta}), \tag{23}$$

$$\hat{L}_n \xrightarrow{d} Z_\chi, \quad Z_\chi \sim \chi^2_1, \tag{24}$$

where $H_{\alpha|\theta} = \mathbf{H}^*_{\alpha\alpha} - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha}$.

Remark These results are impressive because the requirement of (n, s, d) in the theorems, $n^{-1/2}s\log(d) = o(1)$, agrees with the existing work for the linear model and the generalized linear model, which means that the testing for high-dimensional Cox model is not essentially more difficult than the linear model. Also their choice of λ and λ' is the typical rate $\sqrt{n^{-1}\log(d)}$ in high-dimensional estimation and inference.

In theorem 4 they establish the asymptotic property (limiting distribution) of the baseline hazard function Λ_0 and the baseline survival function $S_0(t)$. Due to the limit of length I omit the result here.

4 Proof Outlines for Main Results

4.1 Proof Outlines for Section 3.1

The key points of their proof are 1). $|\dot{\ell}_j(\hat{\boldsymbol{\beta}})|_\infty$ is sufficiently small, shown in lemma 3.3, and 2). the Hessian of $\ell(\boldsymbol{\beta})$ does not vanish for a sparse $\boldsymbol{\beta}$ at the true value $\boldsymbol{\beta}^o$, shown in the proof of theorem 3.3. The local martingales for the counting process are the fundamental tool to ensure these two points.

The proof of theorem 3.1 is based on lemma 3.1 and lemma 3.2. Lemma 3.1 proves that in the event $z^* \leq (\xi-1)/(\xi+1)\lambda$, the estimation error $\tilde{\theta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o$ is in the cone $\mathscr{C}(\xi, \mathcal{O}) = \{\mathbf{b} \in \mathbb{R}^p : |\mathbf{b}_{\mathcal{O}^c}|_1 \leq \xi |\mathbf{b}_{\mathcal{O}}|_1\}$. This fact is useful in controlling the estimation error of the lasso in the linear case. Lemma 3.2 controls the symmetric Bregman-divergence and Hessian of the log-partial likelihood in a neighborhood of $\boldsymbol{\beta}$, in terms of $\eta_{\mathbf{b}} = \max_{s\geq 0} \max_{i,j} |\mathbf{b}'\mathbf{Z}_i(s) - \mathbf{b}'\mathbf{Z}_j(s)|$. Under the conditions of theorem 3.1, the factors in lemma 3.2 lead to the factor e^{η} in the bounds.

Theorem 3.2 is a direct corollary of theorem 3.1 and lemma 3.3. Lemma 3.3 establishes a probability upper bound for $|\dot{\ell}_j(\hat{\boldsymbol{\beta}})|_\infty$, given by

$$\mathbb{P}\left\{ |\dot{\ell}_j(\hat{\boldsymbol{\beta}})|_\infty > C_0 K x, \sum_{i=1}^n \int_0^\infty Y_i(t) dN_i(t) \leq C_0^2 n \right\} \leq 2p \exp(-nx^2/2).$$

The key point of the proof for this inequality is to relate $\dot{\ell}_j(\hat{\boldsymbol{\beta}})$ to $\{X_m\}$, a martingale with $|X_j - X_{j-1}| \leq 1$. Then we can apply the martingale version of Hoeffding's inequality and get the bound.

Theorem 3.3 gives lower bounds for $\kappa(\xi, \mathcal{O})$ and $\mathbf{F}_q(\xi, \mathcal{O})$. To find such lower bound for $\phi(\xi, \mathcal{O}; \bar{\Sigma})$ (which is either κ or F_q), they prove lemma 4.1, showing two approaches. One is to bound the matrix $\bar{\Sigma}$ from below and the second is to approximate $\bar{\Sigma}$. In the proof of theorem 3.3, they first lower bound $\ddot{\ell}(\boldsymbol{\beta}^o)$ by a finite time version $\ddot{\ell}(\boldsymbol{\beta}^o; t^*) = \frac{1}{n}\int_0^{t^*} V_n(s, \boldsymbol{\beta}) d\bar{N}(s)$. Then they approximate $\ddot{\ell}(\boldsymbol{\beta}^o; t^*)$ by the truncated population version $\Sigma(t^*; M)$ given in (14).

4.2 Proof Outlines for Section 3.2

Denote the true parameter as $\boldsymbol{\beta}^*$. The score vector of the log-partial likelihood (2) is

$$\boldsymbol{\xi} = \sum_{i=1}^n \int_0^\tau (\mathbf{X}_i(t) - \bar{\mathbf{Z}}_n(\boldsymbol{\beta}^*, t)) dN_i(t). \tag{25}$$

By the property of counting process, $M_i(t) \equiv N_i(t) - \Lambda_i(t)$ is a martingale with compensator $\Lambda_i(t)$. Denote $\mathbf{Z}_{nj}(\boldsymbol{\beta}^*, t)$ as the jth component of $\bar{\mathbf{Z}}_n(\boldsymbol{\beta}^*, t)$. Since $\sum_{i=1}^n \int_0^\tau (X_{ij}(t) - \mathbf{Z}_{nj}(\boldsymbol{\beta}^*, t)) d\Lambda_i(t) = 0$, we can rewrite $\boldsymbol{\xi}_j$ as

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

The following theorem, theorem 3.1 in [3], proves a large deviation inequality for ξ_j and is the key basement of the proof of theorem 3.5. The power of this theorem is that, it gives a uniform control for all p components of $\boldsymbol{\xi}$, without p appearing in the bounds. Such a dimension-free uniform control is useful when $p \gg n$.

Theorem 4.1 (Theorem 3.1)**.** *Under condition 2 and 3, for any positive sequence $\{u_n\}$ bounded away from zero there exist positive constants c_0 and c_1 such that*

$$\mathbb{P}(|\xi_j| > \sqrt{n} u_n) \leq c_0 \exp(-c_1 u_n) \tag{26}$$

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

The key point in the proof of theorem 4.1 is to decompose ξ_j into two parts and deal with them separately. Condition 2 (a set of typical conditions for the analysis of Cox model) in their paper ensures that there exists bounded functions $s^{(j)}$ defined on $\mathcal{B} \times [0, \tau]$ such that in probability $\sup_{t\in[0,\tau], \beta_1 \in \mathcal{B}_1} \|S_n^{(j)}(\boldsymbol{\beta}_1, t) - s^{(j)}(\boldsymbol{\beta}_1, t)\|_2 \to 0$, for $j = 1, 2, 3$. Let $\mathbf{e}(\boldsymbol{\beta}, t) = s^{(1)}(\boldsymbol{\beta}, t)/s^{(0)}(\boldsymbol{\beta}, t)$, then we can decompose ξ_j as

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

Let's denote the first and second part in the right-hand-size as $\xi_{j1}(\tau)$ and $\xi_{j2}(\tau)$. Condition 3 ensures that we can apply Bernstein inequality to $\xi_{j1}(\tau)$ and get $\mathbb{P}(|\xi_{j1}| > a) \leq 2\exp\{-a^2/2(n\sigma_j^2 + Ma)\}$. For the second part, since M_i's are martingales with respect to $\mathcal{F}_t$, the integration $\xi_{j2}(t)$

is also a martingale w.r.t. $\mathcal{F}_t$. Condition 2(ii),(iii),(iv) ensure that $|\Delta(n^{-1/2}\xi_{j2}(t))| \leq K$ and $\langle n^{-1/2}\xi_{j2}(t)\rangle \leq b^2$, where $0 \leq K < \infty$, $0 < b < \infty$ are constants independent of j. Then by the exponential inequality for martingales with bounded jumps (Lemma 2.1 in [14]), we can get that for $u_n > 0$, $\mathbb{P}(|\xi_{j1}(\tau)| > \sqrt{n}u_n) \leq 2\exp\{-cu_n\}$ uniformly over j. The two inequality together prove theorem 4.1.

The proof of theorem 3.4 is straight forward, which controls the error of the estimator (a local maximizer) by controlling the variation in the value of $\mathcal{C}(\boldsymbol{\beta}, \tau)$ in a neighborhood of $\boldsymbol{\beta}^* = ((\boldsymbol{\beta}_1^*)', \mathbf{0})$. Formally, it proves for large n, $\mathbb{P}\{\sup_{\|u\|_2=1} \mathcal{C}(\boldsymbol{\beta}_1^* + \gamma_n \mathbf{u}, \mathbf{0}) < \mathcal{C}(\boldsymbol{\beta}_1^*, \mathbf{0})\}$, where $\gamma_n = B\{\sqrt{s}(n^{-1/2} + \lambda_n \rho'(\beta_n^*))\}$, where for short $\mathcal{C}(\boldsymbol{\beta})$ denotes $\mathcal{C}(\boldsymbol{\beta}, \tau)$. To prove it, we just apply Taylor expansion to $\boldsymbol{\beta}$ and bound each term in the expansion separately.

To prove theorem 3.5, it suffices to show that $\hat{\boldsymbol{\beta}}^o$ is a local maximizer of $\mathcal{C}(\boldsymbol{\beta}, \tau)$ on set Ω_n with $\mathbb{P}(\Omega_n) \to 1$. To show this, by theorem 2.1 in the paper, we just need to verify 3 optimality conditions, and that's where 4.1 and 3.4 are used.

The proof of theorem 3.6 mainly uses the Taylor expansion of $U_n(\hat{\boldsymbol{\beta}}_1)$ and the condition for the maximizer given in theorem 2.1 of the paper. The goal is to decompose $\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)$ into three terms. Among these 3 terms, 2 are $o_p(1)$ and the last one is a martingale at time τ, and one can use the martingale central limit theorem to prove the weak convergence of that term.

5 Conclusion

I mainly summarize the results in three papers [8],[3], and [6] in this report. Comparing [8] and [3], we conclude that [3] consider a far more general setting where even the global maximizer may not be guaranteed, and therefore require more assumptions and conditions than [8]. In terms of the error bound for the lasso estimator, [8] prove that if $\|\boldsymbol{\beta}^*\|_0 = s$ and $\lambda \asymp \sqrt{n^{-1}\log(d)}$, then $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O_P(s\lambda)$, while [3] prove for the true signal part that $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 = O_P(\sqrt{s}\lambda_n)$ for λ_n s.t. $\sqrt{s}\lambda_n \to 0$, $\lambda_n \gg n^{-1/2+\alpha_2}$(for α_2, see 3.5). It's not straight forward to compare the bounds because the former cares the error of the whole estimator but the latter concerns that error of the signal part, but I would say that the result in [8] is better in some way because it's both easier to apply to other problems (e.g., [6]) and more interpretable. But in the theoretical aspect [3] establish some really interesting results, showing deep insight into the problem, and may be improved in the future to get more interpretable results (for example, give the error bound for the whole estimator without imposing much more restrictions). Also, an interesting but very difficult problem for future work is that how we can find the nice local maximizer used in [3]. For [8], a very interesting problem for future work is about the implementations, especially for the selection of tuning parameter λ.

For the inference problem, [6] establish a complete framework for testing and establishing pointwise confidence intervals. Their methods, in terms of the numerical performance, is more powerful than other tests. Their method is also very robust because it does not require the model selection consistency.

More comments are already included during the statement of the main results.

Acknowledgments

Group members: Wanshan Li.

References

[1] P.K. Andersen and R.D. Gill. Cox's regression model for couting processes: a large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

[2] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 08 2009.

[3] Jelena Bradic, Jianqing Fan, and Jiancheng Jiang. Regularization for cox's proportional hazards model with np-dimensionality. *Ann. Statist.*, 39(6):3092–3120, 12 2011.

[4] David Engler and Yi Li. Survival analysis with high-dimensional covariates: An application in microarray studies. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–22, 2009.

[5] Jianqing Fan and Runze Li. Variable selection for cox's proportional hazards model and frailty model. *Ann. Statist.*, 30(1):74–99, 02 2002.

[6] Ethan X. Fang, Yang Ning, and Han Liu. Testing and confidence intervals for high dimensional proportional hazards models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1415–1437, 2016.

[7] Jiang Gui and Hongzhe Li. Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21(13):3001–3008, 2005.

[8] Jian Huang, Tingni Sun, Zhiliang Ying, Yi Yu, and Cun-Hui Zhang. Oracle inequalities for the lasso in the cox model. *Ann. Statist.*, 41(3):1142–1165, 06 2013.

[9] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, New York, 2nd edition, 2002.

[10] Shengchun Kong and Bin Nan. Non-asymptotic oracle inequalities for the high-dimensional cox regression via lasso. *Stat Sinica*, 24(1):25–42, 2014.

[11] D.Y. Lin, L. J. Wei, I. Yang, and Z. Ying. Semiparametric regression for the mean and rate functions of recurrent events. *J. R. Statist. Soc: B*, 62:711–730, 2000.

[12] Robert Tibshirani. The lasso method for variable selection in the cox model. In *Statistics in Medicine*, pages 385–395, 1997.

[13] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645, 04 2008.

[14] Sara van de Geer. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist.*, 23(5):1779–1801, 10 1995.

[15] Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.*, 3:1360–1392, 2009.

[16] Fei Ye and Cun hui Zhang. Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *JOURNAL OF MACHINE LEARNING RESEARCH*, 11:3519–3540, 2010.