# High-dimensional Variable Selection in Cox Model with Generalized Lasso-type Convex Penalty

Yi Yu

*University of Cambridge*

## 1 Introduction

Survival analysis is a commonly-used method for the analysis of time to event data. This kind of data arises in a number of applied fields, such as medicine, biology, public health, epidemiology, engineering, economics, and demography. A common feature of these data sets is they contain either censored of truncated observations. Censored data arises when an individual's life length is known to occur only in a certain period of time. To analyze right censored data, a powerful tool is Cox Model (1972), often called the proportional hazards model. The model assumes that hazard function satisfying

$$\lambda(t|z) = \lambda_0(t)\exp(\beta' z).$$

Here the baseline hazard function $\lambda_0(t)$ is typically completely unspecied and needs to be estimated nonparametrically. $\beta = (\beta_1, \cdots, \beta_p)'$ is the regression parameter vector. While conducting survival analysis, we not only need to estimate $\beta$, but also have to estimate the baseline hazard function $\lambda_0(t)$ nonparametrically.

Up to now, much work have been in variable selection in Cox model. For example, in Tibshirani (1997) and in Fan and Li (2002), Lasso penalized Cox model and SCAD penalized Cox model are used respectively to do variable selection in fixed dimension case. In Gui and Li (2005) and and Engler and Li (2009), they proposed EN-Cox model to do variable selection in high dimension case, $p = O(\exp(n^{\alpha}))$. But the theoretical results are not that much as in linear regression case. So far in high dimensional setting, what we have is just the results for Lasso-Cox model (Zhang et al) and for SCAD-Cox model (Bradic et al). From Gui and Li (2005) and and Engler and Li (2009) we can know, EN-Cox is a very powerful tool to analyze high-dimensional case. But no theoretical results available in previous work.

Elastic Net is proposed in Regularization and Variable Selection via the Elastic Net (Zou and Hastie, 2005). In this paper, they imposed both $\ell_1$ and $\ell_2$ penalties to linear regression.

$$L(\beta; \lambda_1, \lambda_2) = \frac{1}{2n}\|y - X\beta\|^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2,$$

$\hat{\beta} = \text{argmin}_\beta\{L(\beta; \lambda_1, \lambda_2)\}$ is called Naive EN-estimator. They solved the minimization problem by expanding the design matrix. So the naive elastic net can be viewed as a two-stage procedure: a ridge-type direct shrinkage followed by a lasso-type thresholding. They thought this is double shrinkage, which not only does not help to reduce the variances much but also introduces unnecessary extra bias. So they defined elastic net estimator as $\hat{\beta}$(elastic net) $= (1+\lambda_2)\hat{\beta}$(naive elastic net). Grouping Effect is discussed in this paper. Due to $\ell_2$ penalty, the regression coefficients of a group of highly correlated variables tend to be equal. Theorem 1 of this paper shows grouping effect under the condition $\hat{\beta}_i(\lambda_1)\hat{\beta}_j(\lambda_2) > 0$.

For fixed $p$ case, Yuan and Lin (2007) proved the elastic net consistently selects the true model for suitable $\lambda_1$ and $\lambda_2$. In this paper they proposed "path consistency". That means the solution path contains at least one "desirable" candidate estimate. An estimate $\hat{\beta}$ is considered desirable if it's consistent in terms of both coefficient estimate and variable selection. They proposed a necessary and sufficient condition for the elastic net to be path consistent, which is a version of Elastic Irrepresentable Condition. $\max_{j\notin\mathcal{I}}\left(\liminf_{c_1,c_2\to 0^+}\left[cov(X_j, X_\mathcal{I})\right.\right.$ $\left.\left.\{cov(X_\mathcal{I}) + c_1 I\}^{-1}\big(s_\mathcal{I} + \frac{c_1}{c_2}\beta_\mathcal{I}\big)\right]\right) < 1$. The key point of this is the nearly orthogonality of oracle part and the complement. The added ridge penalty just shrinks the information.

For $p$ and $n$ with relationship of $\lim_{n\to\infty}\frac{\log(p)}{\log(n)} = \nu,\ 0 \le \nu < 1$, the oracle properties of elastic net is discussed in On the Adaptive Elastic-Net with a Diverging Number of Parameters (Zou and Zhang, 2009). Since $p < n$ in this paper, they assume $0 < b \le \lambda_{\min}\big(\frac{1}{n}X'X\big) \le \lambda_{\max}\big(\frac{1}{n}X'X\big) \le B$. We have two skeptical points. Firstly, the conditions on $\lambda_1$ and $\lambda_2$ are $\lim_{n\to\infty}\frac{\lambda_2}{n} = 0$, $\lim_{n\to\infty}\frac{\lambda_1}{n} = 0$. This will easily lead to LASSO in large sample, i.e., ridge penalty is degenerated. Secondly, $\lim_{n\to\infty}\frac{\lambda_2}{\sqrt{n}}\sqrt{\Sigma_{j\in\mathcal{A}}\beta_j^{*2}} = 0$ and $\lim_{n\to\infty}\frac{n}{\lambda_1\sqrt{p}}(\min_{j\in\mathcal{A}}|\beta_j^*|) \to \infty$. If we use our notation, this actually means $\lambda_2\sqrt{n}\|\beta^*\| \to 0$ and $\beta_* = \min_{j\in\mathcal{A}}|\beta_j^*| >> \lambda_1\sqrt{p}$. These conditions are quite strong. There are two key points in the proofs of this paper. One is using the oracle ridge estimator as a bridge between the true and the elastic net estimator; the other is treating the ridge penalty as part of the target function, i.e., expanding the design matrix.

Denote $q = \#\{j; \beta_j^o \neq 0\}$, for $p$, $q$, $n$ satisfying $q = o(n^{1/2})$ and $q\log(p-q) = o(n)$, model selection consistency is showed by Jia and Yu (2008). The basic condition in this paper is Elastic Irrepresentable Condition (EIC), which is elastic version of Irrepresentable Condition for Lasso (IC). Proposition 1 of this paper shows EIC is weaker than IC. Similar to the previous papers, this one also treats the ridge penalty as part of the target function. The key point of EIC and IC are the same, which is the nearly orthogonality of the oracle part and the complement.

Cox-based Elastic Net is discussed in Survival Analysis with High-Dimensional Covariates: An Application in Microarray Studies (Engler and Li, 2009). In this paper, they did quadratic approximation to log-partial likelihood. They also proved the grouping effect under the condition $\hat{\beta}_j(\lambda-1, \lambda_2)\hat{\beta}_k(\lambda-1, \lambda_2) > 0$. But they didn't prove asymptotical results and didn't point out the proper relationship between $p$ and $n$.

Besides Elastic Net penalty, there're some other combinative penalties. One of them is Mnet,

which is a combination of MCP and $\ell_2$ penalty. In this paper, they proved grouping effect, just by standardizing the coefficients. This is an improvement, comparing to the previous results under the condition $\text{sgn}(\hat{\beta}_i(\lambda))\text{sgn}(\hat{\beta}_j(\lambda)) > 0$. Mnet estimator's selection consistency is proved in this paper. There are three key points in their proofs. Firstly, this proof is mainly based on the proof of selection consistency of MCP estimator. Secondly, they expand the design matrix, which is the same as what previous Elastic Net paper did. Thirdly, they introduced oracle ridge estimator as a bridge of the true parameter and the Mnet estimator.

From the above review, we can see the main method to deal with ridge penalty part is augmenting the design matrix. For linear regression, augmenting can easily turn the problem into LASSO. But for Cox-based Elastic Net, due to the implicit solution, even if we treat the ridge penalty as part of the target function, complexity of the problem can't be easily taken down. So we can also put the ridge penalty into the error part. The cost is we have to bound the upper bound of $\|\beta\|_1$. In this paper, we want to prove the prediction consistency, selection consistency and grouping effect for Cox-based Elastic Net, with $p = O(\exp(n\alpha))$ for small $\alpha > 0$, under sparse Riesz condition for Cox model.

To make things more general, based on Elastic-Net, we proposed Generalized Lasso-type Convex Penalty. Elastic-net is a special case of it. One of the benefit of generalization is we can loose the bounded condition on $\|\beta^o\|_\infty$. In Section 2, we use on KKT conditions to derive oracle inequalities. These inequalities give us estimation error bounds. And these bounds are the analyzing foundation. In Section 3, we proposes an irrepresentable condition for GLCP-Cox model, which is in the same spirit of Zhao and Yu (2006). It gaurantees selection consistency. In Section 4, we also proof the grouping effect of certain kind of GLCP-Cox. What we give is a necessary condition of penalties satisfying grouping effect.

All previous work about high dimensional research in Cox model is focused on variable selection. But an important application of Cox model is estimation of survival probability. It's interesting and important in medical statistic and finance. So we use Breslow estimator and prove its consistency in Section 5.

## 2 KKT Conditions and Oracle Inequalities

We define the Cox-based generalized Lasso-type convex penalty(GLCP) criterion.

$$L(\beta; \lambda_1, \lambda_2) = \ell(\beta) + \lambda_1\|\beta\|_1 + \lambda_2 g(\beta),$$

where $\ell(\beta)$ is the negative log-partial likelihood devided by n, $g(\cdot)$ is a convex and nonnegative function, and equals to zero at zero points. $\ell_2$-norm is a special case. $\lambda_1$ and $\lambda_2$ are positive constants. They are tuning parameters here. The GLCP estimator is $\hat{\beta}(\lambda_1, \lambda_2) = \text{argmin}_\beta\{L(\beta; \lambda_1, \lambda_2)\}$. Considering it's a convex minimization problem, its corresponding KKT condition is

$$\begin{cases} \dot{\ell}_j(\hat{\beta}) + \lambda_2\dot{g}_j(\hat{\beta}) = -\lambda_1\text{sgn}(\hat{\beta}_j), & \hat{\beta}_j \neq 0 \\ |\dot{\ell}_j(\hat{\beta}) + \lambda_2\dot{g}_j(\hat{\beta})| \leq \lambda_1, & \hat{\beta}_j = 0 \end{cases}$$

And the solution of KKT conditions is the global minimizer of $L(\beta;\lambda_1,\lambda_2)$. Thus we have the following theorem which gives basic inequality.

**Theorem 1** *Let* $z^* = \|\dot{\ell}(\beta^o) + \lambda_2\dot{g}(\beta^o)\|_\infty$, *then we have*

$$(\lambda_1 - z^*)\|\hat{\theta}_{\mathcal{O}^c}\|_1 \le \hat{\theta}'\{\dot{\ell}(\beta^0+\hat{\theta}) + \lambda_2\dot{g}(\beta^o+\hat{\theta}) - \dot{\ell}(\beta^0) - \lambda_2\dot{g}(\beta^o)\} + (\lambda_1 - z^*)\|\hat{\theta}_{\mathcal{O}^c}\|_1 \le (\lambda_1 + z^*)\|\hat{\theta}_{\mathcal{O}}\|_1.$$

**Proof.**

Since $\ell(\beta) + \lambda_2 g(\beta)$ is a convex function, $\hat{\theta}'\{\dot{\ell}(\beta^o+\hat{\theta}) + \lambda_2\dot{g}(\beta^o+\hat{\theta}) - \dot{\ell}(\beta^o) - \lambda_2\dot{g}(\beta^o)\} \ge 0$, so that this first inequality holds. For the second inequality, the KKT conditions imply

$$\begin{aligned}
& \hat{\theta}'\{\dot{\ell}(\beta^o+\hat{\theta}) + \lambda_2\dot{g}(\beta^o+\hat{\theta}) - \dot{\ell}(\beta^o) - \lambda_2\dot{g}(\beta^o)\} \\
\le\ & \sum_{j=1}^{p} \hat{\theta}_j(\dot{\ell}_j(\beta^o+\hat{\theta}) + \lambda_2\dot{g}_j(\beta^o+\hat{\theta})) + z^*\|\hat{\theta}_{\mathcal{O}}\|_1 + z^*\|\hat{\theta}_{\mathcal{O}^c}\|_1 \\
\le\ & \sum_{j\in\mathcal{O}} \hat{\theta}_j(\dot{\ell}_j(\beta^o+\hat{\theta}) + \lambda_2\dot{g}_j(\beta^o+\hat{\theta})) + \sum_{j\in\mathcal{O}^c} \hat{\theta}_j(\dot{\ell}_j(\beta^o+\hat{\theta}) + \lambda_2\dot{g}_j(\beta^o+\hat{\theta})) + z^*\|\hat{\theta}_{\mathcal{O}}\|_1 + z^*\|\hat{\theta}_{\mathcal{O}^c}\|_1 \\
\le\ & \lambda_1\|\hat{\theta}_{\mathcal{O}}\|_1 - \lambda_1\|\hat{\theta}_{\mathcal{O}^c}\|_1 + z^*\|\hat{\theta}_{\mathcal{O}}\|_1 + z^*\|\hat{\theta}_{\mathcal{O}^c}\|_1 \\
\le\ & (z^* - \lambda_1)\|\hat{\theta}_{\mathcal{O}^c}\|_1 + (\lambda_1 + z^*)\|\hat{\theta}_{\mathcal{O}}\|_1
\end{aligned}$$

due to $\hat{\theta}_j = \hat{\beta}_j$ for $j \in \mathcal{O}^c$.

This leads to conclusion. □

In the event $\{z^* \le (1-2/\zeta)\lambda_1\}$, the estimation error $\hat{\theta} = \hat{\beta} - \beta^o$ is belong to the cone

$$\mathcal{B} = \mathcal{B}(\zeta,\mathcal{O}) = \{b \in \mathbb{R}^p;\ \|b\|_1 \le \zeta\|b_{\mathcal{O}}\|_1\}.$$

We will give detailed proof later, that due to $\dot{\ell}(\beta^o)$ is a mean-zero martingale, for all $\lambda_1 > 0$, $\|\dot{\ell}(\beta^o)\|_\infty < \lambda_1$ holds with exponential rates. So if we define $\|\dot{g}(\beta^o)\|_\infty = K$, then we should require $\lambda_2 K < (1-2/\zeta)\lambda_1$. If $K = O(1)$, then $\lambda_2$ is of the same order as $\lambda_1$, which means it won't degenerate to Lasso case. Compared to previous work, it's an improvement.

A famous example of GLCP is elastic-net penalty. For elastic-net, we have the following corollary.

**Corollary 1** ***(Basic inequality for EN-Cox)***

$$(\lambda_1 - z^*)\|\hat{\theta}_{\mathcal{O}^c}\|_1 \le \hat{\theta}'\{\dot{\ell}(\beta^0+\hat{\theta}) - \dot{\ell}(\beta^0) + \lambda_2\hat{\theta}\} + (\lambda_1 - z^*)\|\hat{\theta}_{\mathcal{O}^c}\|_1 \le (\lambda_1 + z^*)\|\hat{\theta}_{\mathcal{O}}\|_1.$$

Notice $z^* = \|\dot{\ell}(\beta^o) + \lambda_2\dot{g}(\beta^o)\|_\infty \le \|\dot{\ell}(\beta^o)\|_\infty + \lambda_2\|\dot{g}(\beta^o)\|_\infty$. Denote $\|\dot{\ell}(\beta^o)\|_\infty = \tilde{z}^*$, under the condition $\|\dot{g}(\beta^o)\|_\infty \le K$, we get $(\lambda_1 - \lambda_2 K - \tilde{z}^*)\|\hat{\theta}_{\mathcal{O}^c}\|_1 \le (\lambda_1 + \lambda_2 K + \tilde{z}^*)\|\hat{\theta}_{\mathcal{O}}\|_1$. Then in the event $\{\tilde{z}^* \le (1-2/\zeta)\lambda_1 - (K/\zeta)\lambda_2\}$, $\hat{\theta}$ is belong to the cone $\mathcal{B} = \mathcal{B}(\zeta,\mathcal{O}) = \{b \in \mathbb{R}^p :\ \|b\|_1 \le \zeta\|b_{\mathcal{O}}\|_1\}$.

In the same spirit of Lasso-Cox paper, we define

$$RE_1 = RE_1(\zeta,\mathcal{O}) = \inf_{b\in\mathcal{B}} \frac{\{b'\ddot{\ell}(\beta)b\}^{1/2}}{\|b_{\mathcal{O}}\|_1/q^{1/2}},$$

and

$$RE_{1,2} = RE_{1,2}(\zeta, \mathcal{O}) = \inf_{b \in \mathcal{B}} \frac{\{b' \ddot{\ell}(\beta) b\}}{(\|b_{\mathcal{O}}\|_1 / q^{1/2}) \|b\|},$$

And get the following theorems, which extends the oracle inequalities from the linear regression model, Lasso penalized Cox model to the generalized lasso-type convex penalized Cox model.

**Theorem 2** *Suppose* $\|Z_i\|_\infty \le K_1$ *for all* $i$, $\|\beta^o\|_0 = q$, *and* $\tau = 2K_1(1-1/\zeta)(2\lambda_1 + K\lambda_2)q / RE_1^2(\zeta, \mathcal{O}) \le 1/e$. *Then in the event* $\{\tilde{z}^* \le \big(1 - 2/\zeta\big)\lambda_1 - (K/\zeta)\lambda_2\}$,

$$\|\hat{\beta} - \beta^o\|_1 \le \frac{e^\eta (1 - 1/\zeta)(2\lambda_1 + K\lambda_2) q}{RE_1^2(\zeta, \mathcal{O})}, \quad \|\hat{\beta} - \beta^o\|_2 \le \frac{e^\eta (1 - 1/\zeta)(2\lambda_1 + K\lambda_2)\sqrt{q}}{\zeta RE_{1,2}(\zeta, \mathcal{O})},$$

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

**Proof.**

Suppose $\hat{\theta} = \hat{\beta} - \beta^o \neq 0$. Let $b = \hat{\theta} / \|\hat{\theta}\|_1$. It follows from Theorem 1 that $b \in \mathcal{B}(\zeta, \mathcal{O})$ and

$$\hat{\theta}' \{\dot{\ell}(\beta^o + \hat{\theta}) - \dot{\ell}(\beta^o)\} \le (\lambda_1 + \lambda_2 K + \tilde{z}^*) \|\hat{\theta}_{\mathcal{O}}\|_1,$$

and the event , then

$$b' \bigg( \dot{\ell}(\beta^o + xb) - \dot{\ell}(\beta^o) \bigg) \le (1 - 1/\zeta)(2\lambda_1 + K\lambda_2) \|b_{\mathcal{O}}\|_1, \quad x \le \|\hat{\theta}\|_1. \tag{1}$$

Considering a fixed $b \in \mathcal{B}(\zeta, \mathcal{O})$ with $\|b\|_1 = 1$ and all $x$ satisfying (1). We first prove $2K_1 x \le \eta \le 1$.

Let $a_i = b' \{Z_i(s) - \bar{Z}_n(s, \beta^o)\}$ and $w_i^o = Y_i(s) \exp[(\beta^o)' Z_i(s)]$. We have

$$\begin{aligned}
& b' \{\bar{Z}_n(s, \beta^o + xb) - \bar{Z}_n(s, \beta^o)\} \\
= \; & \Sigma_{ij} \omega_i^o \omega_j^o a_i (e^{x a_i} - e^{x a_j}) \big/ \Sigma_{ij} \omega_i^o \omega_j^o e^{x a_i} \\
= \; & \Sigma_{ij} \omega_i^o \omega_j^o (a_i - a_j)(e^{x a_i} - e^{x a_j}) \big/ 2\Sigma_{ij} \omega_i^o \omega_j^o e^{x a_i} \\
\ge \; & x \exp(-2x \max_i |a_i|) \Sigma_{ij} \omega_i^o \omega_j^o (a_i - a_j)^2 \big/ \Sigma_{ij} 2\omega_i^o \omega_j^o \\
= \; & x \exp(-2x \max_i |a_i|) \Sigma_i \omega_i^o a_i^2 \big/ \Sigma_i \omega_i^o,
\end{aligned}$$

due to $\Sigma_{ij} \omega_i^o \omega_j^o (a_i - a_j)^2 = 2\Sigma_i \omega_i^o \Sigma_i \omega_i^o a_i^2$. Since

$$\begin{aligned}
b' \ddot{\ell}(\beta^o) b &= \frac{1}{n} \int_0^1 b' V_n(s, \beta^o) b \; d\bar{N}(s) \\
&= \frac{1}{n} \int_0^1 \sum_{i=1}^n \frac{\omega_i^o}{\Sigma_j \omega_j^o} b' (Z_i(s) - \bar{Z}_n(s, \beta^o))^{\otimes 2} b \; d\bar{N}(s) \\
&= \frac{1}{n} \int_0^1 \sum_{i=1}^n \omega_i^o a_i^2 \big/ \Sigma_i \omega_i^o \; d\bar{N}(s),
\end{aligned}$$

we have a lower bound for $b'\{\dot{\ell}(\beta^o + xb) - \dot{\ell}(\beta^o)\}$,

$$\begin{aligned} b'\{\dot{\ell}(\beta^o + xb) - \dot{\ell}(\beta^o)\} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 b'\{\bar{Z}_n(s,\beta^o + xb) - \bar{Z}_n(s,\beta^o)\}\ dN_i(s) \\ &= \frac{1}{n}\int_0^1 b'\{\bar{Z}_n(s,\beta^o + xb) - \bar{Z}_n(s,\beta^o)\}\ d\bar{N}(s) \\ &\geq \frac{1}{n}\int_0^1 x\exp(-2x\max_i|a_i|)\Sigma_i\omega_i^o a_i^2/\Sigma_i\omega_i^o\ d\bar{N}(s) \\ &= x\exp(-2x\max_i|a_i|)b'\ddot{\ell}(\beta^o)b \\ &= x\exp(-2K_1x)b'\ddot{\ell}(\beta^o)b. \end{aligned}$$

Then it follows (1) and the definition of $\mathrm{RE}_1(\zeta,\mathcal{O})$ that

$$x\exp(-2K_1x) \leq \frac{\big(1-\frac{1}{\zeta}\big)(2\lambda_1 + K\lambda_2)\|b_{\mathcal{O}}\|_1}{b'\ddot{\ell}(\beta^o)b} \leq \frac{\big(1-1/\zeta\big)(2\lambda_1 + K\lambda_2)q}{\mathrm{RE}_1^2(\zeta,\mathcal{O})}$$

Thus, all $x$ satisfying (1) has to satisfy

$$2K_1x\exp(-2K_1x) \leq \frac{2K_1(1-1/\zeta)(2\lambda_1 + K\lambda_2)q}{\mathrm{RE}_1^2(\zeta,\mathcal{O})}, \quad \forall x \leq \|\hat{\theta}\|_1. \tag{2}$$

Since $b'\{\dot{\ell}(\beta^o + xb) - \dot{\ell}(\beta^o)\}$ is an increasing function of $x$ due to the convexity of $\ell$, the set of all $x$ satisfying (1) is a closed interval with 0 at the left end, i.e. $[0,\tilde{x}]$. It follows from (2) that $2K_1\tilde{x}$ has to be $\eta$, which is the smaller solution of $\eta e^{-\eta} = \tau$. Then we have $2K_1x \leq \eta$. Therefore,

$$\|\hat{\theta}\|_1 = x \leq \frac{\eta}{2K_1} = \frac{e^\eta\tau}{2K_1} = \frac{e^\eta(1-1/\zeta)(2\lambda_1 + K\lambda_2)q}{\mathrm{RE}_1^2(\zeta,\mathcal{O})}.$$

By the definition of $RE_{1,2}(\zeta,\mathcal{O})$, we have

$$\begin{aligned} x\exp(-2K_1x) &\leq \frac{(1-1/\zeta)(2\lambda_1 + K\lambda_2)\|b_{\mathcal{O}}\|_1}{RE_{1,2}(\zeta,\mathcal{O})(\|b_{\mathcal{O}}\|_1/q^{1/2})\|b\|} \\ &\leq \frac{(1-1/\zeta)(2\lambda_1 + K\lambda_2)\|\sqrt{q}}{RE_{1,2}(\zeta,\mathcal{O})\|b\|} \end{aligned}$$

similarly we get

$$\|\hat{\theta}\|_2 \leq x\|b\|_2 \leq \frac{e^\eta(\zeta-1)(2\lambda_1 + K\lambda_2)\sqrt{q}}{\zeta\mathrm{RE}_{1,2}(\zeta,\mathcal{O})}.$$

□

This theorem gave us the bound of $\ell_1$ and $\ell_2$ norm of estimation error. It follows from Lasso-Cox paper, by SRC condition for Cox model, we know $\mathrm{RE}_{1,2}(\zeta,\mathcal{O})$ is bounded away from 0. Next we prove this holds with probability close to 1.

**Lemma 1** *Suppose $\|Z_i\|_\infty \leq K$ for all $i$. Then, for all $\epsilon > 0$,*

$$\mathrm{P}(\|\dot{\ell}(\beta^o)\|_\infty > \epsilon) \leq 2p\exp\Big(-n\epsilon^2/2\Big)$$

**Proof.**

Since $\dot{\ell}(\beta^o)$ has $p$ components, it suffices to prove the lemma for $p = K = 1$. In this case, $\dot{\ell}(\beta^o)$ can be written as

$$\dot{\ell}(\beta^o) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 a_i(s)dN_i(s) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 a_i(s)dM_i(s)$$

with a centered predictable $\{a_i(t) = Z_i(t) - \bar{Z}_n(t, \beta^o)), 0 \le t < 1\}$. Due to the boundness of these terms, there exits a constant $C_0$, such that $\langle\dot{\ell}(\beta^o)\rangle < \frac{C_0}{n}$. Let $\alpha = \frac{1}{2n}$, $\beta = \frac{1}{2C_0}$, $x = n\sqrt{C_0}\epsilon$, we can get

$$\mathrm{P}\big\{\big|\dot{\ell}(\beta^o)\big| > \epsilon\big\} \le \mathrm{P}\big\{\big|\dot{\ell}(\beta^o)\big| > (\alpha + \beta\langle\dot{\ell}(\beta^o)\rangle)x\big\} \le 2\exp(-n\epsilon^2/2).$$

Then we can get the conclusion. □

**Theorem 3** *Suppose* $\|\beta^o\|_\infty \le K$, $\|\beta^o\|_0 = q$, $\|Z_i\|_\infty \le K$ *for all* $i$. *For a certain* $\zeta > 2$, $\lambda_1$ *and* $\lambda_2$ *satisfy*

$$(1 - 2/\zeta)\lambda_1 - (K/\zeta)\lambda_2 = \sqrt{2\log(2p/\xi)/n},$$

*where* $\xi > 0$. *Suppose* $\tau = 2K_1(\zeta - 1)(2\lambda_1 + K\lambda_2)q/RE_1^2(\zeta, \mathcal{O}) \le 1/e$. *Let* $\eta \le 1$ *be the smaller solution of* $\eta e^{-\eta} = \tau$. *Then, for all* $c_1 > 0$, $c_{1,2} > 0$,

$$\begin{aligned} & \mathrm{P}\Big\{\|\hat{\beta} - \beta^o\|_1 \le \frac{e^{\eta_1}(\zeta - 1)(2\lambda_1 + K\lambda_2)q}{c_1^2}, \quad \|\hat{\beta} - \beta^o\|_2 \le \frac{e^{\eta_2}(\zeta - 1)(2\lambda_1 + K\lambda_2)q}{\zeta c_{1,2}}\Big\} \\ \ge\ & P\Big\{RE_1(\zeta, \mathcal{O}) > c_1, RE_{1,2}(\zeta, \mathcal{O}) > c_{1,2}\Big\} - \xi. \end{aligned}$$

The condition on the $\lambda_1$ and $\lambda_2$ is weaker than the one in Zou and Zhang(2009). And for this theorem, the relationship of $p$ and $n$ can be $p = O(\exp(an))$, for small a. The only remaining thing is to bound $RE_1(\zeta, \mathcal{O})$ and $RE_{1,2}(\zeta, \mathcal{O})$. This is identical to the discussion of our previous paper, under the Sparse Riesz Condition for Cox model. So under the corresponding SRC-Cox condition, we have the following corollary.

**Corollary 2** *Assume SRC-Cox(q,* $c_*$, $c^*$*) holds, where* $c_*$ *and* $c^*$ *satisfying* $RE_1(\zeta, \mathcal{O}) > c_1, RE_{1,2}(\zeta, \mathcal{O}) > c_{1,2}$. *And the conditions in Theorem 3 hold. Then for all* $\xi_1$ *in SRC-Cox context, we have*

$$\begin{aligned} & \mathrm{P}\Big\{\|\hat{\beta} - \beta^o\|_1 \le \frac{e^{\eta_1}(\zeta - 1)(2\lambda_1 + K\lambda_2)q}{c_1^2}, \quad \|\hat{\beta} - \beta^o\|_2 \le \frac{e^{\eta_2}(\zeta - 1)(2\lambda_1 + K\lambda_2)q}{\zeta c_{1,2}}\Big\} \\ \ge\ & 1 - \xi_1 - \xi. \end{aligned}$$

# 3 Selection Consistency

## 3.1 Irrepresentable Condition For Cox Model

In P. Zhao and B. Yu (2006), they proposed an irrepresentable condition to get the selection consistency for linear regression under LASSO penalty. Their basic idea is KKT condition is the necessary and sufficient condition for global minimizer of target function. And the spirit of irrepresentable condition is to claim the oracle part is nearly uncorrelated with the complement.

In the same spirit, we prove the selection consistency for generalized Lasso-type convex penalized Cox model. To prove it, we introduce two conditions, which are common in computation science.

**Condition 1** ***Strong Convexity for Penalty Function***

*There exists a constant $L$, such that for all $u,\ v \in \mathbb{R}^p$,*

$$g(u) \geq g(v) + g'(v)\langle u, v\rangle + \frac{1}{2}L\|u - v\|^2.$$

In elastic-net case, $L = 1$ and the equality holds. Strong convexity gives us gaurantee to invert second derivative matrix.

**Condition 2** ***Strong Smoothness for Penalty Function***

*There exists a constant $L'$, such that for all $u,\ v \in \mathbb{R}^p$,*

$$g(u) \leq g(v) + g'(v)\langle u, v\rangle + \frac{1}{2}L'\|u - v\|^2.$$

In elastic-net case, $L' = 1$ and the equality holds.

**Definition 1** ***Irrepresentable Condition for GLCP-Cox***

$$\max_{k\notin\mathcal{O}} \sum_{\beta^*\in O(\beta^o,r)} \sup_{s\in[0,1]} \left| \frac{\big(\sum_{i=1}^n Z_{ik}\omega_i(s)\big)\big(\sum_{i=1}^n a_j(\beta^*)\omega_j(s)\big)}{\big(\sum_{i=1}^n \omega_i(s)\big)^2} - \frac{\sum_{i=1}^n Z_{ik}a_i(\beta^*)\omega_i(s)}{\sum_{i=1}^n \omega_i(s)} \right| \leq C\lambda_1,$$

*where $\omega_i(s) = Y_i(s)\exp(\beta^{o\prime}Z_i)$, $a_i(\beta^*) = \exp(\tilde{\theta}(\beta^*)'Z_i)$, $\tilde{\theta}(\beta^*) = -\big(\ddot{\ell}(\beta^*) + \lambda_2 I\big)_{\mathcal{O}}^{-1}\big(\lambda_1 sgn(\beta^o) + \dot{\ell}(\beta^o)\big)_{\mathcal{O}}$, $r = e^{\eta}(\zeta - 1)(2\lambda_1 + K\lambda_2)q/RE_1^2(\zeta, \mathcal{O})$, and $\frac{C}{\exp(KK_1)} \leq (1-\eta)\lambda_1$, for some $0 < \eta < 1$.*

**Theorem 4** *Suppose the irrepresentable condition for EN-Cox holds, 1, 2 and the conditions of Corollary 2 are satisfied. $\beta_* = \min_{j\in\mathcal{O}}\beta_j^o > \frac{e^{\eta_1}(\zeta-1)(2\lambda_1+K\lambda_2)q}{RE_1^2(\zeta,\mathcal{O})}$. $q = o(\sqrt{n/\log p})$. Then the estimator of EN-Cox has selection consistency, i.e.,*

$$\mathrm{P}\big\{sgn(\hat{\beta}) = sgn(\beta^o)\big\} > 1 - \xi.$$

**Proof.**

Recall the KKT conditions,

$$\begin{cases} \dot{\ell}_j(\hat{\beta}) + \lambda_2\dot{g}_j(\hat{\beta}) = -\lambda_1 \mathrm{sgn}(\hat{\beta}_j), & \hat{\beta}_j \neq 0 \\ |\dot{\ell}_j(\hat{\beta}) + \lambda_2\dot{g}_j(\hat{\beta})| \leq \lambda_1, & \hat{\beta}_j = 0 \end{cases}$$

Notice it's the sufficient and necessary condition for the global minimizer. Let $\hat{\beta}_{\mathcal{O}} = \big(\ddot{\ell}(\beta^*)\big)_{\mathcal{O}}^{-1}\big(\dot{\ell}(\hat{\beta}) - \dot{\ell}(\beta^o)\big)_{\mathcal{O}} + \beta^o_{\mathcal{O}}$, if $\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^o)$, then $p$-dimensional vector $\hat{\beta} = (\hat{\beta}'_{\mathcal{O}}, 0')'$ should be the solution to the KKT condition.

So that $\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^o)$ if

$$\begin{cases} |\hat{\beta}_j - \beta_j^o| < |\beta_j^o|, & j \in \mathcal{O} \\ |\dot{\ell}_j(\hat{\beta}) + \lambda_2\dot{g}_j(\hat{\beta})| \leq \lambda_1, & j \notin \mathcal{O} \end{cases}$$

And these hold if

$$\begin{cases} \left\|\hat{\theta}\right\|_1 < \beta_* \equiv \min_{j\in\mathcal{O}} \left|\beta_j^o\right| \\ \left|\dot{\ell}_j(\beta^o) + \lambda_2 \dot{g}_j(\beta^o)\right| = \left|\dot{\ell}_j(\beta^o)\right| \le \eta\lambda_1, & j \notin \mathcal{O} \\ \left|\dot{\ell}_j(\hat{\beta}) - \dot{\ell}_j(\beta^o) + \lambda_2(\dot{g}_j(\hat{\beta}) - \dot{g}_j(\beta^o))\right| \le (1-\eta)\lambda_1, & j \notin \mathcal{O}. \end{cases}$$

It follows from Theorem 3 and conditions on $\beta_*$, that $\mathrm{P}\{\left\|\hat{\theta}\right\|_1 < \beta_* \equiv \min_{j\in\mathcal{O}} \left|\beta_j^o\right|\} \to 1$ holds. And considering $\dot{\ell}_j(\beta^o)$ is a martingale, it follows from de la Pena (1999), $\mathrm{P}\{\left\|\dot{\ell}(\beta^o)_{\mathcal{O}^c}\right\|_\infty \le \eta\lambda_1\} \to 1$.

So it suffies to prove $\left|\dot{\ell}_j(\hat{\beta}) - \dot{\ell}_j(\beta^o) + \lambda_2(\dot{g}_j(\hat{\beta}) - \dot{g}_j(\beta^o))\right| \le (1-\eta)\lambda_1$, $j \notin \mathcal{O}$ now.

We notice that by strong smoothness condition,

$$\left|\dot{\ell}_j(\hat{\beta}) - \dot{\ell}_j(\beta^o) + \lambda_2(\dot{g}_j(\hat{\beta}) - \dot{g}_j(\beta^o))\right| \le \left|\dot{\ell}_j(\hat{\beta}) - \dot{\ell}_j(\beta^o)\right| + \frac{1}{2}\lambda_2 L' r.$$

For all $k \in \mathcal{O}^c$, we have

$$\begin{aligned} \left|\dot{\ell}_k(\hat{\beta}) - \dot{\ell}_k(\beta^o)\right| &= \Big| -\frac{1}{n}\int_0^1 \{\frac{\sum_{i=1}^n Z_{ik} Y_i(s)\exp(\hat{\beta}' Z_i)}{\sum_{i=1}^n Y_i(s)\exp(\hat{\beta}' Z_i)} - \frac{\sum_{i=1}^n Z_{ik} Y_i(s)\exp(\beta^{o\prime} Z_i)}{\sum_{i=1}^n Y_i(s)\exp(\beta^{o\prime} Z_i)}\}\, dN_i(s)\Big| \\ &\le \sup_{s\in[0,1]} \left| \frac{\sum_{i=1}^n Z_{ik}\omega_i(s)\exp(\hat{\theta}' Z_i)}{\sum_{i=1}^n \exp(\hat{\theta}' Z_i)\omega_i(s)} - \frac{\sum_{i=1}^n Z_{ik}\omega_i(s)}{\sum_{i=1}^n \omega_i(s)} \right| \\ &= \sup_{s\in[0,1]} \left| \frac{(\sum_{i=1}^n Z_{ik}\omega_i(s))(\sum_{j=1}^n a_j\omega_j(s)) - (\sum_{i=1}^n Z_{ik}a_i\omega_i(s))(\sum_{i=1}^n \omega_i(s))}{(\sum_{i=1}^n a_i\omega_i(s))(\sum_{i=1}^n \omega_i(s))} \right| \end{aligned}$$

Since $\mathrm{P}\{\|\hat{\beta} - \beta^o\| < r\} > 1-\xi$, with prob. tending to $1-\xi$,

$$\begin{aligned} \left|\dot{\ell}_k(\hat{\beta}) - \dot{\ell}_k(\beta^o)\right| &\le \exp(rK_1) \sup_{\beta^*\in O(\beta^o, r)} \sup_{s\in[0,1]} \Bigg| \frac{\big(\sum_{i=1}^n Z_{ik}\omega_i(s)\big)\big(\sum_{i=1}^n a_j(\beta^*)\omega_j(s)\big)}{\big(\sum_{i=1}^n \omega_i(s)\big)^2} \\ &\quad - \frac{\sum_{i=1}^n Z_{ik}a_i(\beta^*)\omega_i(s)}{\sum_{i=1}^n \omega_i(s)} \Bigg|. \end{aligned}$$

Since $q = o(\sqrt{n/\log p})$ and $\lambda_1 = O(\sqrt{\log p/n})$, then $r \to 0$. So from the irrrepresentable condition for Cox and Theorem 3, we can get the conclusion. □

**Remark 1** *In this theorem, we just require $\beta_* > \|\hat{\theta}\|_1$, which converges to 0. So it's definitely an improvment comparing to the previous work.*

### 3.2 Upper Bound For False Positive

Irrepresentable condition is the nearly The selection consistency provided in last subsection is under an irrepresentable condition. Here, under some weaker conditions, we get the upper bound of false positive.

**Theorem 5** *Suppose that the conditions of Corollary 2 are satisfied, suppose*

$$b'(\ddot{\ell}^2(\beta^o))b \le c_2^*\|b\|^2, \quad \forall b \in \mathcal{B} \equiv \{b \in \mathbb{R}^p : \|b\|_1 \le \zeta\|b_{\mathcal{O}}\|_1\}.$$

*On the set* $\{\|\dot{\ell}(\beta^o)\|_\infty \leq \big(1-\frac{2}{\zeta}\big)\lambda_1 - \frac{K}{\zeta}\lambda_2\}$, $\lambda_2 = \frac{2a\lambda_1}{(\zeta-1)K}$, $0 < a < 1$. $\|\beta^o\|_0 = q = q_n = o\bigg(\big(\frac{\log p}{n}\big)^{1/4}\bigg)$. *Then there is a constant* $M_1$ *depend on* $K$, $K_1$, $c_2^*$, *a such that*

$$|\hat{\mathcal{O}}| \leq M_1|\mathcal{O}|.$$

**Proof.**

Let $\hat{\mathcal{O}} = \{j : \hat{\beta}_j \neq 0\}$. Let $A = \{j : |\dot{\ell}_j(\hat{\beta}) + \lambda_2\dot{g}_j(\hat{\beta})| = \lambda_1\}$ be the active set. Then $\hat{\mathcal{O}} \subset A$. Let $\hat{s}_A = (\hat{s}_j : j \in A)'$, where $\hat{s}_j = \text{sgn}(\dot{\ell}_j(\hat{\beta}) + \lambda_2\dot{g}_j(\hat{\beta}))$. We have $\|\hat{s}_A\|_1 = |A|$, $\|\hat{s}_A\| = \sqrt{|A|}$. Then $\hat{s}_A'(\dot{\ell}_A(\hat{\beta}) + \lambda_2\dot{g}_A(\hat{\beta})) = \lambda_1\hat{s}_A'\hat{s}_A = \lambda_1|A|$. Thus on the set $\{z^* \leq \big(1-\frac{2}{\zeta}\big)\lambda_1 - \frac{K}{\zeta}\lambda_2\}$, $\lambda_2 = \frac{2a\lambda_1}{(\zeta-1)K}$, for $0 < a < 1$.

$$\begin{aligned}
\lambda_1|A| &\leq |\hat{s}_A'(\dot{\ell}_A(\hat{\beta}) + \lambda_2\dot{g}_A(\hat{\beta}) - \dot{\ell}_A(\beta^o) - \lambda_2\dot{g}_A(\beta^o))| + |\hat{s}_A'(\dot{\ell}_A(\beta^o) + \lambda_2\dot{g}_A(\beta^o))| \\
&\leq \|\hat{s}_A\|\|\dot{\ell}_A(\hat{\beta}) - \dot{\ell}_A(\beta^o) + \lambda_2(\dot{g}_A(\hat{\beta}) - \dot{g}_A(\beta^o))\| + \bigg\{(1-\frac{2}{\zeta})\lambda_1 + (1-\frac{1}{\zeta})K\lambda_2\bigg\}|A|.
\end{aligned}$$

Then,

$$(1-a)\frac{2}{\zeta}\lambda_1|A| \leq \sqrt{|A|}\|\dot{\ell}_A(\hat{\beta}) - \dot{\ell}_A(\beta^o) + \lambda_2(\dot{g}_A(\hat{\beta}) - \dot{g}_A(\beta^o))\|,$$

thus,

$$|A|\lambda_1^2 \leq \frac{\zeta^2}{4(1-a)^2}\|\dot{\ell}_A(\hat{\beta}) - \dot{\ell}_A(\beta^o) + \lambda_2(\dot{g}_A(\hat{\beta}) - \dot{g}_A(\beta^o))\|^2.$$

So we need to find an upper bound for $\|\dot{\ell}_A(\hat{\beta}) - \dot{\ell}_A(\beta^o) + \lambda_2(\dot{g}_A(\hat{\beta}) - \dot{g}_A(\beta^o))\|$. By the mean-valued theorem, for $T = \{1, ..., p\}$, and with $\hat{\theta} = \hat{\beta} - \beta^o$,

$$\dot{\ell}_A(\hat{\beta}) - \dot{\ell}_A(\beta^o) = \bigg(\int_0^1 \ddot{\ell}_{AT}(\beta^o + t\hat{\theta})\ dt\bigg)\hat{\theta},$$

where $\ddot{\ell}_{AT}(\beta) = (\ddot{\ell}_{jk}(\beta),\ j \in A,\ k \in T)$, which is a $|A| \times p$ matrix. Let $D_{AT} = \int_0^1\{\ddot{\ell}_{AT}(\beta^o + t\hat{\theta}) - \ddot{\ell}_{AT}(\beta^o)\}\ dt$, and $\delta_* = \max\{d_{jk} : j \in A,\ k \in T\}$, where $d_{jk}$ is the $(j,k)$th element of $D_{AT}$. Then

$$\|\dot{\ell}_A(\hat{\beta}) - \dot{\ell}_A(\beta^o) + \lambda_2(\dot{g}_A(\hat{\beta}) - \dot{g}_A(\beta^o))\|^2 \leq 3\|D_{AT}\hat{\theta}\|^2 + 3\|\ddot{\ell}_{AT}(\beta^o)\hat{\theta}\|^2 + 3\lambda_2^2(\frac{L'}{2})^2\|\hat{\theta}\|^2.$$

We have

$$\|D_{AT}\hat{\theta}\|^2 \leq |A|\delta_*^2\|\hat{\theta}\|_1^2 \leq |A|\zeta^2\delta_*^2\|\hat{\theta}_{\mathcal{O}}\|_1 \leq |A|\zeta^2\delta_*^2 q\|\hat{\theta}\|^2.$$

Assume an upper restricted eigenvalue condition on $\ddot{\ell}^2(\beta^o)$.

$$\|\ddot{\ell}_{AT}(\beta^o)\hat{\theta}\|^2 \leq c_2^*\|\hat{\theta}\|^2.$$

Then we get

$$|A| \leq \frac{\zeta^2}{4\lambda_1^2(1-a)^2}3|A|\zeta^2\delta_*^2 q\|\hat{\theta}\|^2 + \frac{3\zeta^2}{4\lambda_1^2(1-a)^2}(c_2^* + (\frac{\lambda_2 L'}{2})^2)\|\hat{\theta}\|^2.$$

It follows from Theorem 2, there exits a constant $C_1$ such that $\|\hat{\theta}\|_1 \leq \frac{e^\eta(\zeta-1)(2\lambda_1+K\lambda_2)\sqrt{q}}{\zeta\text{RE}_{1,2}(\zeta,\mathcal{O})} = \frac{e^\eta(\zeta-1+a)2\lambda_1\sqrt{q}}{\zeta\text{RE}_{1,2}(\zeta,\mathcal{O})} \equiv C_1\sqrt{q}\lambda_1$. Then we get

$$|A| \leq \frac{3\zeta^2(c_2^* + (\frac{\lambda_2 L'}{2})^2)C_1^2}{4(1-a)^2 - 3\zeta^4(\delta_*^2 q^2)C_1^2}q.$$

If $\delta_*^2 q^2$ is small, then there is a constant $M_1$ such that

$$|A| \le M_1 q.$$

Since $|\mathcal{O}| \le |A|$, the only thing left is to prove $\delta_*^2 q^2$. We noticed $4d_{jk} = 4e_j' D_{AT} e_k = (e_j + e_k)' D_{AT}(e_j + e_k) - (e_j - e_k)' D_{AT}(e_j - e_k)$. And considering the definition of $\ddot{\ell}(\beta)$, we know for all $u \in \mathbb{R}^p$,

$$\big(e^{-Kt\|\hat{\theta}\|_1} - 1\big) u' \ddot{\ell}(\beta) u \lessapprox u' D_{AT} u \lessapprox \big(e^{Kt\|\hat{\theta}\|_1} - 1\big) u' \ddot{\ell}(\beta) u.$$

Since $\|e_j\|_0 = 1 < q$, so $(e_j \pm e_k)' \ddot{\ell}(\beta)(e_j \pm e_k) \le \sqrt{2} c^*$, for $j \ne k$. So $\delta_* \asymp K\|\hat{\theta}\|_1 \asymp K\lambda_1 q$. So what we need is $q^4 \lambda_1^2 \to 0$. And it holds, due to $\lambda_1 = o(\sqrt{\log p/n})$ and $q = O\Big(\big(\frac{\log p}{n}\big)^{1/4}\Big)$.

□

# 4 Grouping Effect

In Engler and Li (2009), they showed that EN-Cox has grouping effect, i.e. if two covariates have high correlation, they tend to be selected or not selected together. In this section, we first prove the grouping effect for EN-Cox, then discuss the set of penalties with grouping effect.

**Theorem 6** *For EN-Cox model, assume* $\sum_{i=1}^n Z_{i\ell}^2 = n$, $\ell = 1, .., p$. *Let* $\rho_{jk} = \frac{1}{n}\sum_{i=1}^n Z_{ij} Z_{ik}$ *be the correlation coefficient. Suppose* $\lambda_2 > 0$. *Assume the conditions of Corollary 2 hold. If* $sgn(\hat{\beta}_j) sgn(\hat{\beta}_k) > 0$, *there exists a constant* $M_2$, *such that*

$$\mathrm{P}\{|\hat{\beta}_j - \hat{\beta}_k| \le M_2 \sqrt{1 - \rho_{jk}}\} \to 1.$$

**Proof.**

According to KKT conditions, if $\hat{\beta}_j \ne 0$, $\hat{\beta}_j = -\frac{1}{\lambda_2}\dot{\ell}_j(\hat{\beta}) - \frac{\lambda_1}{\lambda_2}\mathrm{sgn}(\hat{\beta}_j)$. Then for $\mathrm{sgn}(\hat{\beta}_j)\mathrm{sgn}(\hat{\beta}_k) > 0$, we can get $|\hat{\beta}_j - \hat{\beta}_k| = \frac{1}{\lambda_2}|\dot{\ell}_j(\hat{\beta}) - \dot{\ell}_k(\hat{\beta})|$.

Notice,

$$\begin{aligned}
|\dot{\ell}_j(\hat{\beta}) - \dot{\ell}_k(\hat{\beta})| &= \big| - \frac{1}{n}\sum_{i=1}^n \int_0^1 \{(Z_{ij} - Z_{ij}) - (\bar{Z}_j(s,\hat{\beta}) - \bar{Z}_k(s,\hat{\beta}))\}\, dN_i(s)\big| \\
&= |\frac{1}{n}\sum_{i=1}^n \int_0^1 \{(Z_{ij} - Z_{ik}) - \frac{\sum_{i=1}^n (Z_{ij} - Z_{ik}) Y_i(s) \exp(\hat{\beta}' Z_i)}{\sum_{i=1}^n Y_i(s)\exp(\hat{\beta}' Z_i)}\}\, dN_i(s)| \\
&\le |\frac{1}{n}\sum_{i=1}^n \int_0^1 \{(Z_{ij} - Z_{ik}) - \frac{\sum_{i=1}^n (Z_{ij} - Z_{ik}) Y_i(s) \exp(\hat{\beta}' Z_i)}{\sum_{i=1}^n Y_i(s)\exp(\hat{\beta}' Z_i)}\}\, dM_i(s)| \\
&+ |\frac{1}{n}\sum_{i=1}^n \int_0^1 \{(Z_{ij} - Z_{ik}) - \frac{\sum_{i=1}^n (Z_{ij} - Z_{ik}) Y_i(s) \exp(\hat{\beta}' Z_i)}{\sum_{i=1}^n Y_i(s)\exp(\hat{\beta}' Z_i)}\}\, Y_i(s)\exp(\beta^{o\prime} Z_i)\lambda_0(s)\, ds| \\
&\equiv I_1 + I_2.
\end{aligned}$$

For $I_1$, we can use de la Pena's theorem, and get for all $0 < \alpha < 1/2$,

$$\mathrm{P}\{|I_1| \le \frac{1 - \rho_{jk}}{n^{1-\alpha}}\} \to 1.$$

For $I_2$, due to the boundness of the relevant terms, there exists a constant $M_2$, such that

$$\begin{aligned} I_2 &= |\frac{1}{n}\int_0^1 \sum_{i=1}^n (Z_{ij}-Z_{ik})Y_i(s)\lambda_0(s)\{\exp(\hat{\beta}'Z_i(s))\frac{\sum Y_i(s)\exp(\beta^{o'}Z_i)}{\sum Y_i(s)\exp(\hat{\beta}'Z_i)} - \exp(\beta^{o'}Z_i)\}\ ds| \\ &\leq \sqrt{2(1-\rho_{jk})}\int_0^1 \lambda_o(s)\ ds \max_{s,i} \left|\{\exp(\hat{\beta}'Z_i(s))\frac{\sum Y_i(s)\exp(\beta^{o'}Z_i)}{\sum Y_i(s)\exp(\hat{\beta}'Z_i)} - \exp(\beta^{o'}Z_i)\}\right| \\ &= \sqrt{2(1-\rho_{jk})}\int_0^1 \lambda_o(s)\ ds \max_{s,i} \left|\exp(\beta^{o'}Z_i(s))\{\frac{\exp(\hat{\theta}'Z_i)\sum Y_i(s)}{\sum Y_i(s)\exp(\hat{\theta}'Z_i)} - 1\}\right| \\ &\leq M_2\sqrt{2(1-\rho_{jk})} \end{aligned}$$

Combining the previous two statements, we get the conclusion. □

From the proof we can realize why Lasso doesn't have grouping effect. The derivate of target function for Lasso-Cox is $\dot{\ell}_j(\hat{\beta}) + \lambda_1 \text{sgn}(\beta_j)$, but for EN-Cox is $\dot{\ell}_j(\hat{\beta}) + \lambda_1 \text{sgn}(\beta_j) + \lambda_2\hat{\beta}_j$. When $\text{sgn}(\hat{\beta}_j) = \text{sgn}(\hat{\beta}_k)$, deviance of ridge penalty part must be coincidence with the deviance of the partial likelihood part. This derives grouping effect.

For GLCP here, if we still want to reserve grouping effect, a necessary condition is $|\dot{g}_j(\hat{\beta}) - \dot{g}_k(\hat{\beta})|$ is an increasing function of $|\hat{\beta}_j - \hat{\beta}_k|$. So elastic-net penalty is a special case. And the proof of the GLCP's grouping effect is straight forward after the proof we provided here.

**Corollary 3** *For GLCP-Cox model, assume* $\sum_{i=1}^n Z_{i\ell}^2 = n$, $\ell = 1, .., p$. *Let* $\rho_{jk} = \frac{1}{n}\sum_{i=1}^n Z_{ij}Z_{ik}$ *be the correlation coefficient. Suppose* $\lambda_2 > 0$. *Assume the conditions of Corollary 2 hold. If* $|\dot{g}_j(\hat{\beta}) - \dot{g}_k(\hat{\beta})|$ *is an increasing function of* $|\hat{\beta}_j - \hat{\beta}_k|$, *and* $sgn(\hat{\beta}_j)sgn(\hat{\beta}_k) > 0$, *there exists a constant* $M_2$, *such that*

$$\text{P}\{|\hat{\beta}_j - \hat{\beta}_k| \leq M_2\sqrt{1-\rho_{jk}}\} \to 1.$$

# 5 Asymptotics distribution of $\hat{\Lambda}_0$

All previous work of high dimensional research in Cox model is focused on variable selection and estimation error bounding. In fact, an important application of Cox model in reality is estimation of survival probability. This involves estimation of baseline hazard. So in this section we use Breslow estimator of baseline hazard and give its asymptotical properties. This will easily lead to estimation of survival probability in Cox model.

## 5.1 Formulations, notations and conditions

Using Breslow estimator(1972), the estimated cumulated hazard function is

$$\hat{\Lambda}_0 = \sum_{T_i \leq t} \frac{\delta_i}{\sum_{j\in\mathcal{R}_t}\exp(\hat{\beta}'Z_j)}.$$

Formulated by means of counting processes, the estimated $\hat{\Lambda}_0$ has the form

$$\hat{\Lambda}_0(t) = \int_0^t \frac{d\bar{N}(s)}{\sum_{i=1}^n Y_i(s)\exp(\hat{\beta}'Z_i)},$$

and hence

$$\begin{aligned} n^{1/2}\{\hat{\Lambda}_0(t)-\Lambda_0(t)\} &= n^{1/2}\int_0^1 \left\{\frac{1}{\sum_{i=1}^n Y_i(s)\exp(\hat{\beta}' Z_i)}-\frac{1}{\sum_{i=1}^n Y_i(s)\exp(\beta_0' Z_i)}\right\} d\bar{N}(s) \\ &+ n^{1/2}\left\{\int_0^t \frac{d\bar{N}}{\sum_{i=1}^n Y_i(s)\exp(\hat{\beta}_0' Z_i)}-\Lambda_0^*(t)\right\} \\ &+ n^{1/2}\{\Lambda_0^*(t)-\Lambda_0(t)\}, \end{aligned}$$

where

$$\Lambda_0^*(t)=\int_0^t \lambda_0(s) I\{\sum_{i=1}^n Y_i(s)>0\}\, ds.$$

In this section we will prove the third term is asymptotically negligible, the second term is a local martingale, and is asymptotically orthogonal to the Taylor expansion of the first term. So $n^{1/2}\{\hat{\Lambda}_0(t)-\Lambda_0(t)\}$ is weakly convergent to the asymptotical distribution of the sum of the first and the second term.

Directly form asymptotic stability and asymptotic regularity conditions,

$$\mathbb{P}(\Lambda_0^*=\Lambda_0 \text{ on } [0,1])\to 1.$$

So we need not consider the term $n^{1/2}(\Lambda_0^*-\Lambda_0)$ in the equality.

We have the following notations,

$$S_n^{(\ell)}(\beta,t)=n^{-1}\sum_{i=1}^n Y_i(t)\{Z_i(t)\}^{\otimes \ell}\exp(\beta' Z_i(t)), \quad \ell=0,1,2.$$

Let $\beta_1$ be a subvector of $\beta$ formed by all nonzero components. So under the conditions of Theorem 4, we have $\mathbb{P}\{\text{sgn}(\hat{\beta})=\text{sgn}(\beta^o)\}>1-\xi$. So with probability at least $1-\xi$, we have $\dim(\hat{\beta}_1)=\dim(\beta_1^o)$. Considering $\xi$ is an arbitrary positive number, so in the following, we can treat them of the same dimension.

Here are some conditions used in this chapter.

(i) (**Finite interval**) $\int_0^1 \lambda_0(t)\, dt<\infty$.

(ii) (**Asymptotic stability**) There exists a compact neighborhood $\mathcal{B}$ of $\beta^o$ and scalar, vector and matrix functions $s^{(0)}$, $s^{(1)}$ and $s^{(2)}$ defined on $\mathcal{B}\times[0,1]$ such that in probability as $n\to\infty$ for $j=0,1,2$,

$$\sup_{t\in[0,1],\beta\in\mathcal{B}} \|S_n^{(j)}(\beta,t)-s^{(j)}(\beta,t)\|_2\to 0.$$

(iii) (**Lindeberg condition**) There exists $\delta>0$ such that in probability as $n\to\infty$

$$n^{-1/2}\sup_{i,t}|Z_i|Y_i(t)I\{\beta^{o\prime} Z_i>-\delta|Z_i|\}\to 0.$$

(iv) (**Asymptotic regularity conditions**) Let $\mathcal{B}$, $s^{(0)}$, $s^{(1)}$ and $s^{(2)}$ be as in condition (ii) and define $e(\beta,t)=s^{(1)}(\beta,t)/s^{(0)}(\beta,t)$, $v(\beta,t)=s^{(2)}(\beta,t)/s^{(0)}(\beta,t)-e(\beta,t)^{\otimes 2}$ and $\Sigma_\beta(t)=\int_0^t v(\beta,u)s^{(0)}(\beta^o,u)\, d\Lambda_o(u)$. Let

$$\Sigma_{\beta_1}(t)=\int_0^t v(\beta_1,u)s^{(0)}(\beta_1^o,u)\, d\Lambda_o(u)$$

and $\Sigma_{\beta_1} = \Sigma_{\beta_1}(1)$. Assume that the $q \times q$ matrix $\Sigma_{\beta_1^o}$ is positive definite for all $n$ and $\Lambda_0(1) < \infty$. The functions $s^{(j)}$ are bounded and $s^{(0)}$ is bounded away from 0 on$\mathcal{B} \times [0,1]$; for $j = 0, 1, 2$, the family of functions $s^{(j)}(\cdot, t)$, $0 \le t \le 1$, is an equicontinuous family at $\beta^o$.

(v) The process $Y(t) = (Y_1(t), \ldots, Y_n(t))'$ is left continuous with right hand limits and satisfies $\mathbb{P}(Y(t) = 1, 0 \le t \le 1) > 0$.

(vi) Let $c_n = \sup_{t\in[0,1]} \|E_n(\beta^o, t) - e(\beta^o, t)\|_\infty$ and $d_n = \sup_{t\in[0,1]} |S_n^{(0)}(\beta^o, t) - s^{(0)}(\beta, t)|$. The random sequences $c_n$ and $d_n$ are bounded almost surely.

(vii) For all matrix $A$, let $r_\sigma(A) = \max\{|\lambda| : \lambda \in \sigma(A)\}$. $r_\sigma(\Sigma_{\beta_1^o}) = O(1)$ and $r_\sigma(\Sigma_{\beta_1^o}^{-1}) = O(1)$.

(viii) $E\{\sup_{0\le t\le 1} Y(t)\|S(t)\|_2^2 \exp(\beta_1^{o'} S(t))\} = O(q)$.

(ix) $\|\dot{g}(x)\|_\infty$ is of the polynomial order of $\|x\|_\infty$.

(x) Conditions of Corollary 2

These conditions are mainly from Bradic et al. In fact, these are modification version of the conditions used in Anderson and Gill(1982).

## 5.2 Normality

In this subsection, we mainly follow the proof of Theorem 4.6 in Bradic et al.(to be appeared). Since the loss functions are identical, and the only difference is the penalty part.

Since $\hat{\beta}_2 = 0$, $\hat{\beta}_1$ is a maximizer of

$$C(\beta, 1) = \sum_{i=1}^n \int_0^1 \{\beta' Z_i - \log\left(S_n^{(0)}(\beta, t)\right)\} \, dN_i(t) - n\lambda_1 \sum_{j=1}^p |\beta_j| - \frac{n}{2}\lambda_2 g(\beta).$$

So $\hat{\beta}_1$ satisfies the equation

$$\frac{1}{n} U_n(\hat{\beta}_1) = \lambda_1 \mathrm{sgn}(\hat{\beta}_1) + \lambda_2 \dot{g}_1(\hat{\beta}),$$

where

$$U_n(\beta) = U_n(\beta, 1) = \sum_{i=1}^n \int_0^1 \{Z_i - \bar{Z}(s, \beta)\} \, dN_i(s).$$

Using Taylor expansion, we obtain that

$$U_n(\hat{\beta}_1) = U_n(\beta_1^o) + \partial U_n(\beta_1^o)(\hat{\beta}_1 - \beta_1^o) + r_n(\beta_1^*),$$

where $\beta^*$ is on the line segmant between $\hat{\beta}$ and $\beta^o$, and the remainder term $r_n(\beta_1^*)$ is equal to

$$\frac{1}{2}\Sigma_{j,k}(\beta_{1j}^* - \beta_{ij}^o)(\beta_{1k}^* - \beta_{1k}^o)\frac{\partial^2 U_n(\beta_1^*)}{\partial\beta_{1j}\partial\beta_{1k}},$$

hence $\|n^{-1} r_n(\beta_1^*)\|_2 = O_p(\sqrt{q}\|\beta_1^* - \beta_1^o\|_2^2)$. Since we have

$$\|\beta_1^* - \beta_1^o\|_2 \le \|\hat{\beta}_1 - \beta_1^o\|_2 \le \frac{e^\eta(\zeta - 1)(2\lambda_1 + K\lambda_2)\sqrt{q}}{\zeta \mathrm{RE}_{1,2}(\zeta, \mathcal{O})},$$

and $\lambda_1, \ \lambda_2 = O_p(\sqrt{\frac{q\log p}{n}})$, hence $\|n^{-1} r_n(\beta_1^*)\|_2 = O_p(q^{5/2}\log p/n)$.

Notice that

$$
\begin{aligned}
n^{1/2}(\hat{\beta}_1 - \beta_1^o) &= \big(-\frac{1}{n}\frac{\partial}{\partial\beta}U_n(\beta_1^o)\big)^{-1} n^{-1/2} U_n(\beta_1^o) && (3)\\
&+ \big(-\frac{1}{n}\frac{\partial}{\partial\beta}U_n(\beta_1^o)\big)^{-1} n^{-1/2} r_n(\beta_1^*) && (4)\\
&- \big(-\frac{1}{n}\frac{\partial}{\partial\beta}U_n(\beta_1^o)\big)^{-1} n^{1/2}(\lambda_1 \mathrm{sgn}(\hat{\beta}_1) + \dot{g}_1(\hat{\beta})). && (5)
\end{aligned}
$$

Lemma A.2, Lemma A.4 and Lemma A.5 in Bradic et al. (to be appeared) are only related to the log-partial likelihood part, which is identical to the log-partial likelihood we used here. So we can directly use them. With the same notation we have

$$
\begin{aligned}
-n^{-1}\partial U_n(\beta_1) &= n^{-1}\sum_{i=1}^{n}\int_0^1 V(\beta_1, t)\; dN_i(t)\\
&= \int_0^1 V(\beta_1,t) S_n^{(0)}(\beta_1^*, t)\lambda_0(t)\; dt + n^{-1}\int_0^1 V(\beta_1,t)\; d\bar{M}(t)\\
&\equiv \mathcal{I}_{\beta_1} + \mathcal{W}_{\beta_1}.
\end{aligned}
$$

From Lemma A.2 and Lemma .4, under the conditions we have, $\sup_{\beta_1\in\mathcal{B}}\|\mathcal{I}_{\beta_1}\|_2 = O_p(1)$, $\|\mathcal{I}_{\beta_1^*}^{-1}\|_2 = O_p(1)$, and $\sup_{\beta_1\in\mathcal{B}}\|\mathcal{I}_{\beta_1} - \Sigma_{\beta_1}\|_2 = o_p(1)$. $\|\mathcal{W}_{\beta_1^*}\|_2 = O_p(s/\sqrt{n})$, and for any consistent estimator $\hat{\beta}_1$ of $\beta_1$, $\|\mathcal{W}_{\hat{\beta}_1}\|_2 = o_p(1)$.

For all $q\times 1$ unit vector $b$, if $q^{5/2}\log p = o_p(n^{1/2})$,

$$
\begin{aligned}
|b'\Sigma_{\beta_1^o}^{1/2}(-\frac{1}{n}\frac{\partial}{\partial\beta}U_n(\beta_1^o))^{-1}n^{-1/2}r_n(\beta_1^*)| &\le n^{1/2}\|\Sigma_{\beta_1^o}^{1/2}\|_2\|(-\frac{1}{n}\frac{\partial}{\partial\beta}U_n(\beta_1^o)^{-1}\|_2\|\frac{1}{n}r_n(\beta_1^*)\|_2\\
&= n^{1/2}O_p(1)O_p(\frac{q^{5/2}\log p}{n}) = o_p(1)
\end{aligned}
$$

For the penalty term,

$$
\begin{aligned}
&\|b'\Sigma_{\beta_1^o}^{1/2}(-\frac{1}{n}\frac{\partial}{\partial\beta}U_n(\beta_1^o))^{-1}n^{1/2}(\lambda_1\mathrm{sgn}(\hat{\beta}_1) + \lambda_2\dot{g}_1(\hat{\beta}))\|\\
\le\ &\|\Sigma_{\beta_1^o}^{1/2}\|_2 O_p(1) n^{1/2} O_p(\sqrt{q\log p/n})O_p(1) = o_p(1)
\end{aligned}
$$

These two inequalities demonstrate that (4) and (5), when multiplied by the vector $b'\Sigma_{\beta_1^o}^{1/2}$, are of order $o_p(1)$. Combined with Lemma A.5 in Bradic et al.(to be appeared), we get

$$
\begin{aligned}
\sqrt{n}b'\Sigma_{\beta_1^o}^{1/2}(\hat{\beta}_1 - \beta_1^o) &= b'\Sigma_{\beta_1^o}^{1/2}(-\frac{1}{n}\frac{\partial}{\partial\beta}U_n(\beta_1^o))^{-1}n^{-1/2}U_n(\beta_1^o) + o_p(1)\\
&= b'\Sigma_{\beta_1^o}^{-1/2}n^{-1/2}U_n(\beta_1^o) + o_p(1) \equiv \phi_{n1} + o_p(1).
\end{aligned}
$$

It suffices to show the asymptotic normality of $\phi_{n1}$, because then the asymptotic normality of $\hat{\beta}_1$ can be obtained by Slutsky's Theorem. Directly from Theorem 4.6 in Bradic et al., $\phi_{n1}(t)$ is asymptotically standard normal.

From the asymptotic stability condition, Linderberg condition, and asymptotic regularity conditions, using martingale central limit theorem (Anderson and Gill, 1982), we can get the second term is asymptotically distributed as a Gaussian martingale with variance function

$$\int_0^t \frac{\lambda_0(u)}{s^{(0)(\beta_1^o,u)}\; du}.$$

### 5.3 Asymptotical orthogonality

Let's look back at the counting process form Breslow estimator. Taylor expansion of the first term is

$$\begin{aligned} & n^{1/2}\int_0^t \left\{\frac{1}{\sum_{i=1}^n Y_i(s)\exp(\hat{\beta}_1' Z_i)} - \frac{1}{\sum_{i=1}^n Y_i(s)\exp(\beta_1^{o\prime} Z_i)}\right\} d\bar{N}(s) \\ = & \left[-\int_0^t \frac{\sum_{i=1}^n Y_i(s) Z_i e^{\beta_1^* Z_i}}{\{\sum_{i=1}^n Y_i(s) e^{\beta_1^* Z_i}\}^2}\, d\bar{N}(s)\right]' n^{1/2}(\hat{\beta}_1 - \beta_1^o). \end{aligned}$$

The second term is a local martingale,

$$n^{1/2}\left\{\int_0^t \frac{d\bar{N}(s)}{\sum_{i=1}^n Y_i(s) e^{\beta_1^{o\prime} Z_i} - \Lambda_0^*(t)}\right\} = \int_0^t \frac{n^{1/2}\, d\bar{M}(s)}{\sum_{i=1}^n Y_i(s) e^{\beta_1^{o\prime} Z_i}}.$$

Recall $n^{1/2}(\hat{\beta}_1 - \beta_1^o)$ is the linear combination of functions of $U_n(\beta_1^o)$, $r_n(\beta^*)$ and $\lambda_1 \mathrm{sgn}(\hat{\beta}_1) + \lambda_2 \dot{g}_1(\hat{\beta})$. So we prove the asymptotical orthogonality respectively. The score part is the following,

$$\langle U_n(\beta_1^o, \cdot), \int \frac{d\bar{M}}{\sum_{i=1}^n Y_i(s) e^{\beta_1^{o\prime} Z_i}} \rangle = \sum_{i=1}^n \int_0^t \frac{Z_i - \bar{Z}(s, \beta_1^o)}{\sum_{i=1}^n Y_i(s) e^{\beta_1^{o\prime} Z_i}} Y_i(s) e^{\beta_1^{o\prime} Z_i} \lambda_o(s)\, ds = 0$$

The penalty part is the following,

$$\begin{aligned} & \langle \lambda_1 \mathrm{sgn}(\hat{\beta}_1) + \lambda_2 \dot{g}_1(\hat{\beta}), \int \frac{d\bar{M}(s)}{\sum_{i=1}^n Y_i(s) e^{\beta_1^{o\prime} Z_i}} \rangle \\ = & \langle \frac{1}{n}\sum_{i=1}^n \int_0^1 (Z_i - \bar{Z}(s, \hat{\beta}_1))\, dM_i(s), \int \frac{d\bar{M}(s)}{\sum_{i=1}^n Y_i(s) e^{\beta_1^{o\prime} Z_i}} \rangle \\ = & \frac{1}{n}\sum_{i=1}^n \int_0^1 \frac{(\bar{Z}(s, \beta_1^o) - \bar{Z}(s, \hat{\beta}_1)) Y_i(s) \beta_1^{o\prime} Z_i}{\sum_{i=1}^n Y_i(s) e^{\beta_1^{o\prime} Z_i}} \lambda_0(s)\, ds = o_p(1) \end{aligned}$$

the first equality is due to the orthogonality of predictable process and martingale, and the consistency of $\hat{\beta}$ respectively. The proof is similar to the proof in Theorem 2.

The remainder part is of the order $o_p(1)$, due to the order of $\|r_n(\beta_1^*)\|_2$ and the order of $q$. Hence, we get the asymptotical orthogonality of the first term and the second term.

### 5.4 Conclusion

So we can get the weak convergence of $n^{1/2}(\hat{\Lambda} - \Lambda_0)$,

**Theorem 7** *Assume Conditions (i) - (x) hold.* $n^{1/2}(\hat{\beta}_1 - \beta_1^o)$ *and the process equal in the point* $t$ *to*

$$n^{1/2}\{\hat{\Lambda}(t) - \Lambda_0(t)\} + n^{1/2}(\hat{\beta}_1 - \beta_1^o)' \int_0^t e(\beta_1^o, u)\lambda_0(u)\, du$$

*are asymptotically independent, the latter being asymptotically distributed as a Gaussian martingale with variance function*

$$\int_0^t \frac{\lambda_0(u)}{s^{(0)}(\beta_1^o, u)}\, du.$$

With the weak convergence of baseline hazard and the consistency of $\hat{\beta}$ we got in the previous chapters, we can directly get estimator of the survival probability, and its asymptotical properties.

**Corollary 4** ***Estimator of survival probability*** *Assume Conditions (i) - (x) hold. Let* $\hat{S}_n(t|z) = \exp(-\hat{\Lambda}(t)\exp(\hat{\beta}_1' z))$ *be the estimator of survival function given covariate* $z$*, where* $\hat{\Lambda}$ *and* $\hat{\beta}_1$ *are given in the previous context. Then* $\hat{S}_n(t|z)$ *is a consistent estimator of* $S(t|z)$ *for all* $t \in [0,1]$ *and* $z \in \mathbb{R}^p$*.*

# 6 Discussion

We proposed a class of combinative penalties for Cox model here, which are Lasso-type convex penalties. The representative of them is EN-Cox, which is proposed in Gui and Li (2005). But none of the previous papers gave us rigorous proofs of its theoretical properties.

In this paper the relationship between $n,\ p,\ q$ and the relationship between $\lambda_1$ and $\lambda_2$ are sharp enough. $p$ can be order of $\exp(n^\alpha)$, this has achieved the exponential rate, which is the fastest we can handle now. And $q$ can go to infinite as $n$ goes to infinite in a certain rate. The requirement is different in different purposes. $\lambda_1$ and $\lambda_2$ can be of same order, which is an improvement compared to previous work. This won't let the penalty degenerate to Lasso case.

KKT conditions is the necessary and sufficient condition in covex problem for a global minimizer. Based on KKT conditions, we give basic inequality. Considering the second penalty is convex, so we just take the target function and the second penalty as one part. In fact this makes the inequality not as sharp as it is in Lasso case. But based on this inequality, we get the estimation error bound is as sharp as it is in Lasso case. Still we need Sparse Riesz Condition for Cox model here.

Zhao and Yu (2006) proposed irrepresentable condition to prove selection consistency in Lasso penalized linear regression. In this paper, we borrow the same spirit to propose irrepresentable condition to prove selection consistency in our model, and two more conditions which are common in computation science – that is strong convexity and strong smoothness. We admit that this irrepresentable condition is irrepresentable, although it's nearly necessary and sufficient for selection consistency. So if we just need to give an upper bound for false positive, we don't need such strong condition. Under some regularity conditions and certain relationship between $q$ and $n$, also relationship between $\lambda_1$ and $\lambda_2$, we give an upper bound.

Grouping effect is an attracting feature in EN-Cox, in this paper we proved it theoretically. And we also a necessary condition for a penalty with grouping effect. We have to say, the condition in proving it is not that satisfactory to us. Although it's used frequently in previous work.

Estimating as survival probability is very important in reality application of survival analysis. So in this paper, we used Breslow estimator and proved its consistency, under bunch of conditions. But all these conditions are common in survival analysis or variable selection, and not hard to be satisfied.

# References


Anderson, 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

Bradic, J. et al. Regularization for Cox's proportional hazards model with NP-dimensionality. (to be appeared on *JRSS B* )

de la Pena, V. (1999) A general class of exponential inequalities for martingales and ratio. *Annals of Probability* **27** 537-564

Engler, D. and Li, Y.(2009) Survival analysis with high-dimensional covariates: an application in microarray studies. *Statistical Applications in Genetics and Molecular Biology*

Gui, J. and Li, H. (2005) Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with application to microarray gene expression data. *Bioinformatics* **21** 3001-3008

Klein, J. and Moeschberger, M. (2005) Survival Analysis, second edition

Tibshirani, R. (1997) The LASSO method for variable selection in the Cox model. *Statistics in Medicine* **16** 385-395

Yu, B. and Jia, J. On model selection consistency of the elastic net when $p >> n$.

Yuan, M. and Lin, Y. (2007) On the non-negative garrotte estimator. *J. R. Statist. Soc. B* **69** 143-161.

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

Zhang, C-H. and Huang, J. (2008) The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Stat* **36** 1567-1594

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

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* **67** 301-320.

Zou, H. and Zhang, H. H. (2009) On the adaptive elastic-net with a diverging number of parameters *Ann. Stat.* **37(4)** 1733-1751.