一橋大学経済学研究科

ディスカッションペーパー No. 2022-01

高次元 Cox 回帰モデルの統計的推測について

本田敏雄

2022 年 3 月

高次元 Cox 回帰モデルの統計的推測について

本田敏雄*

Statistical inferences on high-dimensional Cox regression models

Toshio Honda*

データ収集技術の飛躍的進歩により，説明変数の数 p の非常に多い高次元データが得られるようになっており，その統計解析が重要な話題となって久しく，代表的な手法である Lasso などは，学部生向けのテキストにも紹介されるようになっている．またさらに，説明変数の数 p は標本数 n の指数オーダーと考えて差し支えないような超高次元データも，統計解析の対象になっている．本論文では，生存時間解析でもっともよく使われているといって差し支えない Cox 回帰モデルを中心に，（超）高次元の説明変数がある生存時間の扱いについて，最近の研究について，著者の研究自身の研究の観点から紹介する．

High-dimensional data with many and many covariates are available because of drastic progress in data-collecting technology. Hence statistical analysis of such high-dimensional data has been a very important issue for many years. We can find typical methods such as the Lasso even in textbooks for undergraduate students. In additon, in some areas we usually use ultra-high dimensional data such that $p \sim \exp(n^c)$, where p is the number of covariates, n is the sample size, c is a constant. In this paper, we review important studies on survival time data with (ultra-)high dimensional covariates from a perpective of my own studies on high-dimensional data. We focus on the Cox regression model which is one of the most important models for survival analysis. In addition, we refer to related topics.

キーワード: Lasso, SCAD, オラクル不等式, オラクル性, feature screening, sure screening property

1. Cox 回帰モデルとその拡張について

第 1 節では準備として，右側打ち切りをもつ生存時間データについて説明し，Cox 回帰モデルとその部分尤度推定法について述べる．ついで Cox 回帰モデルの拡張として，加法モデル，変動係数モデル，部分線形モデルを簡単に紹介する．第 2 節では，高次元線形回帰モデルに関する基本的な結果を紹介する．そして第 2 節の結果に対応する形で，第 3 節で (超) 高次元 Cox 回帰モデルを中心に説明し，第 4 節では超高次元生存時間データの場合によく用いられる feature screening などと呼ばれる，事前にある程度まで説明変数の数を

* 一橋大学経済学研究科：〒 186-8601 国立市中 2-1

減らしておく手法について説明する．説明変数の数 p が標本数 n の指数オーダーと考えて差し支えないような超高次元データでは，Lasso などの標準的手法に関する計算が実行できない場合も多いため，その場合にはこの feature screening のような処理が必要となる．また Lasso 推定量などの計算に使われる R パッケージについても簡単に触れる．以上のように，タイトルにある Cox 回帰モデル以外の，高次元生存時間データに関する話題にも触れる．ただし関連する論文は非常に多くなっており，すべてを紹介できてはいないことに注意されたい．

1.1 生存時間データと Cox 回帰モデル

Cox 回帰モデルとその部分尤度推定法の説明から始める．詳しくは，Kalbfleisch と Prentice(2002) などの標準的な教科書をを参照されたい．ここで T_0 を生存時間，C を右側打ち切り時間とする．この場合，実際に観測されるのは T_0 ではなく，$T = \min\{T_0, C\}$ であり，$\delta = I\{T_0 \leq C\}$ も同時に観測されるとする．この T と δ 以外に，p 次元の共変量 $\boldsymbol{X}$ があり，この $\boldsymbol{X}$ が T_0 にどのように影響するかをデータから調べたいとする．これから紹介するいくつかの話題では，時間変化する共変量も扱うことができるが，簡単のため本稿では $\boldsymbol{X}$ は時間変化しないとする．

T_0 の $\boldsymbol{X}$ に関する条件付き密度関数を $f(t|\boldsymbol{X})$，条件付き分布関数を $F(t|\boldsymbol{X})$ とする．ここで生存時間は非負であることに注意しておく．このとき T_0 の条件付きハザード関数 $\lambda(t|\boldsymbol{X})$ は

$$\lambda(t|\boldsymbol{X}) = \frac{f(t|\boldsymbol{X})}{1 - F(t|\boldsymbol{X})}$$

で定義される．Cox 回帰モデル (Cox(1972)) では，$\lambda(t|\boldsymbol{X})$ は以下の形にモデル化される．

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t) \exp(\boldsymbol{\beta}_0^T \boldsymbol{X}) \tag{1.1}$$

$\lambda_0(t)$ と $\boldsymbol{\beta}_0$ は，それぞ真のベースラインハザード関数と真の回帰係数である．$\lambda_0(t)$ は無次元の攪乱母数である．ここで $\boldsymbol{v}^T$ はベクトル $\boldsymbol{v}$ の転置を表す．

しかしながら実際の観測されるのは，$(T, \delta, \boldsymbol{X})$ であり，T_0 は観測されない．このとき，T_0 と C が $\boldsymbol{X}$ に関して条件付き独立であれば，

$$\mathrm{P}(\delta = 1, t < T \leq t + \Delta | T > t, \boldsymbol{X}) \approx \frac{f(t|\boldsymbol{X})\Delta \mathrm{P}(C > t + \Delta)}{\mathrm{P}(T_0 > t)\mathrm{P}(C > t)} \approx \frac{f(t|\boldsymbol{X})\Delta}{1 - F(t|\boldsymbol{X})}$$

となり，右側打ち切りの影響はないことになり，(1.1) の $\boldsymbol{\beta}_0$ は，適切な方法により $\delta = 1$ の観測値から推定できることになる．そして Cox 回帰モデルの場合には，部分尤度推定量により $\boldsymbol{\beta}_0$ を推定する．

部分尤度の定義のため，点過程 $N(t)$ と at-risk 過程 $Y(t)$ を定義する．

$$N(t) = \delta I\{T \leq t\} \quad \text{かつ } Y(t) = I\{T \geq t\}$$

このとき，(1.1) の仮定の下，

$$N(t) - \int_0^t Y(s)\lambda_0(s)\exp(\boldsymbol{X}^T\boldsymbol{\beta}_0)ds \tag{1.2}$$

は，適当なフィルトレーション $\{\mathcal{F}_t\}$ の下で，マルチンゲールとなる．このマルチンゲール性が，これから定義する部分尤度の最大化による推定量 $\widehat{\boldsymbol{\beta}}$ の漸近正規性の証明に重要である．

簡単のため観測は有限の時間 τ で打ち切られるとし，そして $(T, \delta, \boldsymbol{X})$ の独立同一分布に従う n 個の標本 $(T_i, \delta_i, \boldsymbol{X}_i)$ があるとする．i 番目の観測値の T_0 は T_{0i} と書く．このとき対数部分尤度 $L(\boldsymbol{\beta})$ と部分尤度最大化推定量 $\widehat{\boldsymbol{\beta}}$ は，

$$L(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^\tau \boldsymbol{\beta}^T\boldsymbol{X}_i dN_i(s) - \int_0^\tau \log\Big\{\sum_{i=1}^n Y_i(s)\exp(\boldsymbol{\beta}^T\boldsymbol{X}_i)\Big\}d\overline{N}(s), \tag{1.3}$$

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\,\ell(\boldsymbol{\beta}), \quad \text{ここで } \ell(\boldsymbol{\beta}) = -\frac{1}{n}L(\boldsymbol{\beta}) \text{ かつ} \overline{N}(s) = \frac{1}{n}\sum_{i=1}^n N_i(s) \tag{1.4}$$

のように定義される．部分尤度の導出にはいくつかの方法があるが，ベースラインハザード関数を区分的定数関数として尤度を最大化するノンパラメトリック最尤法の考え方によるものが，数学的には明解かもしれない．この導出も，$\widehat{\boldsymbol{\beta}}$ の漸近正規性などと合わせて Kalbfleisch と Prentice(2002) などを参照されたい．

1.2 Cox 回帰モデルの拡張

本小節では Cox 回帰モデルの拡張について述べる．線形回帰モデルが，データの特性に合わせて，あるいはモデルの適合度を向上させるために，ノンパラメトリックモデル，セミパラメトリックモデル，構造を持つノンパラメトリックモデルに拡張されているように，(1.1) の指数関数の中の線形関数の部分も，様々な形に拡張されている．説明変数が $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ のみの場合と，これに $\boldsymbol{Z} = (Z_1, \ldots, Z_q)^T$ が加わる場合の二通りある．後者は，年齢，時間，年収などの，重要な鍵となる変数の場合が多く，$q = 1$ のことが多いので $q = 1$ として，これ以降 Z と書く．以下に指数関数の中の線形関数をどのように拡張するか，注意すべきこと，関連する文献をまとめる．またノンパラメトリック成分の推定には，局所線形回帰 (Fan と Gijbels(1996)) やスプライン回帰などの方法があるが，スプライン回帰の場合，第 2 節，第 3 節で紹介される，既存の高次元データ解析に関する理論や R パッケージを容易に適用できる．スプライン回帰を実行するためには，説明変数に何らかの変換をおこない，その台を $[0, 1]$ などにする必要がある．スプライン関数については，Schumaker(2007) などを参照されたい．以下 (1.1) の指数関数の中がどのようにモデル化されるかを述べる．

ノンパラメトリックモデル $g(\boldsymbol{X})$: $g(\boldsymbol{x})$ は十分になめらかな未知の関数．このときベースラインハザード関数のため，$g(\boldsymbol{x})$ と $g(\boldsymbol{x}) + \alpha$ は識別できないので，推定できるのは基準

点 $\boldsymbol{x}_0$ を固定した上での $g(\boldsymbol{x}) - g(\boldsymbol{x}_0)$ や平均を引いて中心化したものである．前者の場合の局所線形回帰推定については，Honda(2004) あるいは Chen と Zhou(2007)，加えてそれらの引用文献を参照されたい．

加法モデル $\sum_{j=1}^{p} g_j(X_j)$：各成分関数 $g_j(x_j)$ は十分になめらかな未知の関数．このとき各要素の識別のため

$$\mathrm{E}\{g_j(X_j)\} = 0 \quad \text{あるいは} \int_0^1 g_j(x_j)dx_j = 0 \tag{1.5}$$

などとする必要がある．例えば後者を用いるとする．まず $(L+1)$ 次元の $[0,1]$ 上の区分的 k 次関数である B スプライン基底をとり (今後は $k=2$ としておく)，この基底に適当な正則行列をかけて L 個の要素が後者の基準化を満たし，残りの 1 個の要素は積分して 1 にならないとする．満たすほうの L 個の要素を並べて $\boldsymbol{B}(t) = (B_1(t), \ldots, B_L(t))^T$ とする．これを新しいスプライン基底と用いる．そして $\boldsymbol{X_i} = (X_{i1}, \ldots, X_{ip})^T$ から，新たな pL 次元の説明変数 $\boldsymbol{W_i} = (\boldsymbol{B}(X_{i1})^T, \ldots, \boldsymbol{B}(X_{ip})^T)^T$ を生成し，部分尤度の最大化により加法モデルの各要素関数 $g_j(x_j)$ を推定する．詳しくは Huang 他 (2000) あるいは Honda と Yabe(2017) などを参照されたい．

変動係数モデル $g_0(Z) + \sum_{j=1}^{p} X_j g_j(\boldsymbol{Z})$：各係数関数 $g_j(\boldsymbol{z})$ は十分になめらかな未知の関数であるが，(1.5) のような制約は必要ない．ただし Z の台は $[0,1]$ とする．また Z が時間であっても本質的な違いはない．従ってここでは，$\boldsymbol{B}(t) = (B_1(t), \ldots, B_L(t))^T$ を $[0,1]$ の区分的 2 次関数である L 次元の B スプライン基底とする．Z_i を i 番目の Z の標本として，新たな pL 次元の説明変数 $\boldsymbol{W_i} = (X_{i1}\boldsymbol{B}(Z_i)^T, \ldots, \quad X_{ip}\boldsymbol{B}(Z_i)^T)^T$ を生成し，部分尤度の最大化により変動係数モデルの各係数関数 $g_j(z)$ を推定する．局所線形回帰による推定については，Cai と Sun(2003) などを参照されたい．

部分線形モデル $\sum_{j=1}^{p'} \beta_j X_j + \sum_{j=p'+1}^{p} g_j(X_j)$ または $\sum_{j=1}^{p'} \beta_j X_j + \sum_{j=p'+1}^{p} X_j g_j(Z) + g_0(Z)$: 加法モデル，変動係数モデルの関数の一部が定数になったものである．推定については Huang(1999) などを参照されたい．このようなモデルの場合，事前情報等により構造が特定できない限り，線形部分と非線形部分をどう分けるかという構造の特定化が重要な問題となる．高次元 Cox 回帰の場合のその問題については，第 3 節で Honda と Yabe(2017) を紹介する．その他その引用文献も参照されたい．

2. 高次元線形回帰モデルに関する基本的な結果

2.1 Lasso と SCAD

ここでは，Lasso(Tibshirani(1996)) と SCAD(Fan と Li(2001)) を中心に，高次元線形モデルに関する基本的な結果を説明する．Lasso と類似する Dantzing selector(Candes と Tao(2007)) と SCAD とおおよそ同様の性質を持つ MCP(Zhang(2010)) についてここでは触れない．また Fan 他 (2020) には高次元データ解析に関する結果がまとめられている．

$p > n$ という高次元の設定では実際に有効な説明変数は少数でも，通常の最小 2 乗法による回帰係数の推定は不可能であり，以下のペナルティー付きの最小 2 乗法を考えるのが標準的な手法となっている．

まず最初に高次元線形モデルの記号を与える．ここで $\boldsymbol{X} = (X_{ij})$ は説明変数の代表値ではなく，$n \times p$ の計画行列とする．$\boldsymbol{Y}$ は観測値を並べた n 次元ベクトルとする．$\|\boldsymbol{v}\|$，$\|\boldsymbol{v}\|_1$，$\|\boldsymbol{v}\|_\infty$，$\|\boldsymbol{v}\|_0$ を，それぞれベクトル $\boldsymbol{v}$ のユークリッドノルム，L_1 ノルム，sup ノルム，非ゼロ要素数とする．高次元線形モデルは以下の通りである．

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}, \qquad \boldsymbol{\beta}_0 \in \mathbb{R}^p, \tag{2.1}$$

ここで $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ は被説明変数，$\boldsymbol{\beta}_0$ は少数の非ゼロ (有効な) 要素を持つ真の回帰係数，$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ は攪乱項とする．繰り返すが，この節では $\boldsymbol{X}$ は計画行列とする．

このときペナルティー付きの最小 2 乗推定量 $\widehat{\boldsymbol{\beta}}$ は次のように定義される．

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \Big\{ n^{-1} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|) \Big\}, \tag{2.2}$$

ここで $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$，$p_{\lambda_j}(|\beta_j|)$ はペナルティー，λ_j はチューニングパラメータである．λ_j を共通として，$p_{\lambda_j}(|\beta_j|) = 2\lambda|\beta_j|$ のときが Lasso である．

以下 Lasso の基本的な結果を述べ, Lasso を拡張したあるいは発展させた, group Lasso(Meier 他 (2008))，adaptive Lasso(Zou(2006))，de-biased または de-sparcified Lasso(Javanmard と Montanari(2014), Zhang と Zhang(2014), van de Geer 他 (2014)) を紹介する．

Lasso の性質に関しては，Bickel 他 (2009) により初めて明確で理解しやすい形で与えられた．計画行列 $\boldsymbol{X}$ は定数で，攪乱項は独立に同一分布に従い適当な条件を満たすとする．まず $\mathcal{S}$ を有効な説明変数の添え字集合とし，s をその要素数とする．$\widehat{\Sigma} = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$ とおくと，この $\widehat{\Sigma}$ の最小固有値は高次元 $(p > n)$ の設定では 0 である．Bickel 他 (2009) は，Lasso がこの問題を自然な形で回避していることを明らかにした．

ここで重要なのが，以下に定義する Deviation 条件と RE(restricted eigenvalue) 条件である．それぞれを，偏差条件と制限固有値条件と呼んでも差し支えないと思われる．

Deviation 条件：$\lambda \sim \sigma\sqrt{\log p/n}$ かつ $(\sigma^2 = \mathrm{E}\{\epsilon_1^2\})$ で，以下の不等式が成立する．

$$\|n^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}\|_\infty \leq \lambda/2 \tag{2.3}$$

この Deviation 条件は，高い確率で成立することが様々な文献で証明されている．

そして $\widehat{\boldsymbol{\beta}}$ の最適性と Deviation 条件より次の不等式が得られる．

$$\|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)_{\mathcal{S}^c}\|_1 \leq 3\|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)_{\mathcal{S}}\|_1, \tag{2.4}$$

ここで $\boldsymbol{\beta}_D$ は，$\boldsymbol{\beta}$ から添え字集合 D に対応する部分だけを取り出した，$\boldsymbol{\beta}$ の部分ベクトルである．そして最小固有値的なものについては，以下の $A(3)$ でのみ考えればよい．

RE 条件：$A(c_0) = \{\boldsymbol{\delta} \in \mathbb{R}^p \,|\, \|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 \leq c_0 \|\boldsymbol{\delta}_{\mathcal{S}}\|_1,\ \boldsymbol{\delta} \neq 0\}$ の記号の下で，ある正の数 κ^2 に対し以下の不等式が成立する．

$$\min_{\boldsymbol{\delta} \in A(3)} \frac{\boldsymbol{\delta}^T \widehat{\Sigma} \boldsymbol{\delta}}{\|\boldsymbol{\delta}\|^2} \geq \kappa^2 > 0. \tag{2.5}$$

この条件も説明変数が独立に同一分布に従い，適当な条件をみたすとき，高い確率で成立することが，様々な文献で証明されている．

そして Deviation 条件と RE 条件が成立するとき，以下のオラクル不等式が成立する．C は適当な定数である．

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq C \frac{s}{\kappa^2} \lambda \quad \text{および} \quad \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \leq C \frac{\sqrt{s}}{\kappa^2} \lambda \tag{2.6}$$

$$n^{-1/2} \|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\| \leq C \sqrt{\frac{s}{\kappa^2}} \lambda \tag{2.7}$$

Lasso 推定量は，画期的かつ大変有用な推定量であるが，推定量の非ゼロ要素を取り出す形の変数選択については，その一致性のためには非常に強い条件が必要 (Zou(2006) など) であり，また大きなバイアスをもつことも知られている．従って，最終的な推定量というよりは，SCAD などの何らかの推定の初期値などに用いられることが多い．

これから，group Lasso，adaptive Lasso，de-biased Lasso を紹介する．

group Lasso：計画行列の列，言い換えれば説明変数を，$\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_G)$ かつ $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_G^T)^T$ のように自然にグループ化できるとする．このときグループごとのペナルティ—を考えるのが group Lasso であり，group Lasso 推定量は以下のように定義される．

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ n^{-1} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{g=1}^{G} \lambda_g \|\boldsymbol{\beta}_g\| \right\}$$

group Lasso は，先に紹介した変動係数モデル，加法モデルなどの構造をもつノンパラメトリックモデルなどにも自然な形で適用できる．

adaptive Lasso： 変数選択の一致性の問題を解決するために提案され，以下の重み w_j は Lasso 推定量 $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_p)^T$ 等より，$w_j = 1/|\widetilde{\beta}_j|^\gamma$ のように計算する．γ は適当な定数であり，adaptive Lasso 推定量は以下のように定義される．

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ n^{-1} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + 2\lambda \sum_{j=1}^{p} w_j |\beta_j| \right\} \tag{2.8}$$

group Lasso に対しても，同様な形で adaptive group Lasso を定義することができる．

de-biased Lasso：推定量のゼロ要素の部分を除くような変数選択は行わなわず，元の高次元モデルのままで推定を行う手法である．この de-biased Lasso 推定量は漸近分布を扱

うことができ，信頼区間の構成などに用いることができる．Lasso 推定量 $\widehat{\boldsymbol{\beta}}$ は，以下の最適化の条件式を満たすことに注意する．この $\lambda\boldsymbol{\kappa}$ がバイアスの原因であり，その $\lambda\boldsymbol{\kappa}$ はペナルティーの劣勾配である．

$$n^{-1}\boldsymbol{X}^T\boldsymbol{Y} = n^{-1}\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \lambda\boldsymbol{\kappa}$$

$\widehat{\Sigma} = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$ は高次元の設定では退化しているが，その逆行列の代わりとなる $\widehat{\Theta}$ を構成し，以下のようにバイアスを除く．

$$\widehat{\boldsymbol{b}} = \widehat{\boldsymbol{\beta}} + \widehat{\Theta}\lambda\boldsymbol{\kappa} = \widehat{\boldsymbol{\beta}} + \frac{1}{n}\widehat{\Theta}\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \tag{2.9}$$

$$= \boldsymbol{\beta}_0 + \frac{1}{n}\widehat{\Theta}\boldsymbol{X}^T\epsilon + (\widehat{\Theta}\widehat{\Sigma} - I)(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}) \tag{2.10}$$

(2.10) の第三項は無視可能である．また (2.10) の $(\widehat{\Theta}\widehat{\Sigma} - I)$ に注意されたい．$\widehat{\Theta}$ の構成には，node-wise Lasso という形で Lasso を用いる場合や CLIME(Cai 他 (2011)) という手法を用いる場合などがあるが，必要な条件さえ満たせばいずれの方法でも特に問題はない．さらに (2.9) で $-\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ がスコア関数であることに注意することにより，de-biased Lasso 推定量は，one-step 推定量の形であることも知られており，線形回帰モデルに限らない一般的な設定の下でも定義可能である．

SCAD の説明の前に，便利な R パッケージとチューニングパラメータ選択に簡単に触れておく．glment，ncvreg，grpreg(group Lasson などの説明変数がグループ分けされる場合に有用) などの R パッケージで，線形回帰モデルだけでなく，GLM，Cox 回帰モデルについて Lasso，SCAD，MCP などの推定量を計算できる．文献その他で最もよく目にするの glmnet である．チューニングパラメータは，BIC，EBIC(Chen と Chen(2008))，CV を用いて選択されることが多い．上記の R パッケージなどには CV によるチューニングパラメータ選択機能が備わっている．ただ R と R パッケージの使用については，完全自己責任である．

この小節の最後に SCAD について述べる．SCAD のペナルティーは以下で与えられる (導関数で定義)．SCAD のペナルティーは非凸関数で，$|\beta_j| > a\lambda$ では定数 (導関数は 0) である．a については $a = 3.7$ がよく用いられる．

$$p'_\lambda(|\beta_j|) = \lambda I\{|\beta_j| \leq \lambda\} + \frac{(a\lambda - |\beta_j|)_+}{a-1} I\{|\beta_j| > \lambda\}, \tag{2.11}$$

ここで $a_+ = \max\{0, a\}$ である．

SCAD 推定量はある種の閾値推定量であり，オラクル性を持つ．即ち，真の疎なモデルあるいは非ゼロ係数についての情報を用いて推定されたオラクル推定量 $\widehat{\boldsymbol{\beta}}_{oracle}$ と同様の挙動を持つ．MCP(Zhang(2010)) も同様である．SCAD の場合には，

$$\mathrm{P}(\widehat{\boldsymbol{\beta}}_{oracle} = \widehat{\boldsymbol{\beta}}) \to 1 \tag{2.12}$$

という結果が成り立つ．

SCAD のペナルティーは非凸なので，p が非常に大きいときには適切な初期値が必要であり，その場合は Lasso 推定量が用いられることが多い．あるいは次小節でのべるような何等かのスクリーニング法を用いて，事前に説明変数の数を十分に減らすことが行われる．

2.2 説明変数のスクリーングについて

feature screening と言われる，説明変数のスクリーングについて簡単に述べる．

説明変数の数 p が非常に多いとき，グラム行列 $n^{-1}\boldsymbol{X}^T\boldsymbol{X}$ は $p \times p$ で非常に大きくなるため Lasso 推定量ですら計算できない場合がある．特に線形モデル以外ではその可能性が十分にある．従って予めある程度説明変数を絞りこむ，あるいは明らかに不要なものを除く手法が必要である．Liu 他 (2015) はわかりやすいサーベイ論文である．

$\mathcal{S}$ を有効な説明変数の添え字集合で，$\widehat{S}$ をスクリーニングによって選ばれた添え字集合とする．スクリーニングには，

$$\mathrm{P}(\mathcal{S} \subset \widehat{S}) \to 1$$

という，sure screening property が必須であり，$\widehat{S}$ は小さいほうが望ましいが，変数選択に関する一致性は期待しない．あくまでスクリーニングである．

スクリーニングの方法には，周辺モデルによる方法，モデルによらない方法，モデルを用いた前進型の方法がなどある．モデルによらない方法は，変数間の関係を示す適切な指標をとり，被説明変数と個々の説明変数に関してその指標の標本値を計算する．結果として p 個，説明変数ごとに指標値が計算される．そしてその指標値に基づいて，スクリーニングを行う．詳細は Liu 他 (2015) を参照されたい．前進型については，もっとも有名な Wang(2009) と最新の文献を含んだ Honda と Lin(2021) を挙げるにとどめる．

以下周辺モデルによる SIS(sure independence screening) と呼ばれる方法 (Fan と Lv(2008), Fan と Song(2010) など) について簡単に説明する．

$$Y_i = (X_{i1}, \ldots, X_{ip})\boldsymbol{\beta}_0 + \epsilon_i \tag{2.13}$$

という，被説明変数，説明変数ともに中心化され，説明変数も分散 1 に基準化された高次元 p 変数線形回帰モデルを考える．このモデルに対して，p 個の以下のような 1 変数の周辺モデルを考える．

$$Y_i = X_{ij}\gamma_j + v_{ij} \tag{2.14}$$

そして標本も中心化，基準化して p 個の回帰係数の推定値 $\widehat{\gamma}_j$ を求め，$|\widehat{\gamma}_j|$ により説明変数を順位付けし，大きい方から適当な数の説明変数を選ぶという方法が，Fan と Lv(2008) で提案されている．周辺モデル (2.14) が真のモデル (2.13) を，適切に反映していることが何等かの形で仮定されていると考えてよい．

Fan と Song(2010) では，高次元の一般化線形モデル (GLM) に対し，1 次元の周辺モデルを考えることにより SIS を行うことが提案されている．GLM などの標準的なパラメトリックモデルだけでなく，加法モデルなどの構造を持つノンパラメトリックモデルでも同様のスクリーニング法が提案されている (Fan 他 (2011) など).

3. 高次元 Cox 回帰モデルに関する研究

この節では，高次元 Cox 回帰モデルに関する，Lasso，SCAD，adaptive Lasso，de-biased Lasso，加法モデル，変動係数モデルに関する結果を紹介する．定理の詳細を与えるのは困難なので，第 2 節との関連について述べる形とする．以下 $p > n$ のような状況を想定する．

(1.1) の高次元 Cox 回帰モデルに対する Lasso 推定量 $\widehat{\boldsymbol{\beta}}$ は以下に与えられる．

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\{\ell(\boldsymbol{\beta}) + 2\lambda\|\boldsymbol{\beta}\|_1\}, \tag{3.1}$$

ここで $\ell(\boldsymbol{\beta})$ は (1.4) に定義されたものである．Deviation 条件および RE 条件は，それぞれ

$$\|\dot{\ell}(\boldsymbol{\beta}_0)\|_\infty, \quad \text{ここで}\dot{\ell}(\boldsymbol{\beta}) = \frac{\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}, \quad \text{および} \quad \ddot{\ell}(\boldsymbol{\beta}_0) = \frac{\ell(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$$

について，多少の定数倍の違いを除いて，第 2 節と同じ形で考えることになる．Lasso 推定量の (2.6) と同様なオラクル不等式は，Huang 他 (2013) において正しく導出されており，. Deviation 条件および RE 条件も，高い確率で成立することが証明されている．

Cox 回帰モデルに対する adaptive Lasso は，Zhang と Lu(2007) において初めて提案されているが，その後の研究の飛躍的発展もあり，Zhang と Lu(2007) よりは一般的な設定で考えることができる．adaptive Lasso 推定量の定義と性質については第 2 節と同様であり，(2.8) において $n^{-1}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ を $\ell(\boldsymbol{\beta})$ に置き換えるだけである．やはり変動係数モデルなどに対して，adaptive group Lasso を考えることもできる (Yan と Huang(2012)，Honda と Härdle(2014) など).

ついで de-biased Lasso に関する結果を紹介する．詳しくは Kong 他 (2021), Yu 他 (2021) を参照されたい．Cox 回帰モデルに対する de-biased Lasso 推定量は，Lasso 推定量から以下の式で定義される．(2.9) の表現と対比されるとよい．

$$\widehat{\boldsymbol{b}} = \widehat{\boldsymbol{\beta}} - \widehat{\Theta}\dot{\ell}(\widehat{\boldsymbol{\beta}}), \tag{3.2}$$

ここで $\widehat{\Theta}$ は，$\ddot{\ell}(\widehat{\boldsymbol{\beta}})$ の逆行列の代わりになる $p \times p$ の行列である．Yu 他 (2021) ではこの計算に CLIME を修正した手法を用い，Kong 他 (2021) では node-wise Lasso という van de Geer 他 (2014) で提案されている手法を用いている．

SCAD については，Bradic 他 (2011) で考えられている．SCAD 推定量は (2.2) において， $n^{-1}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ を (1.4) の $\ell(\boldsymbol{\beta})$ に置き換え，(2.11) の SCAD ペナルティーを用いて

定義される．SCAD 推定量については，高次元線形回帰モデルと同じ形のオラクル性が成立する．繰り返しになるが，Lasso についての正しい結果は Huang 他 (2013) に与えられている．

この節の最後に，Honda と Yabe(2017) の結果について紹介する．そこでは，超高次元の部分変動係数モデルと部分線形加法モデルについて，group Lasso による変数選択と構造の決定を統一的に扱っている．ここでは部分変動係数モデルの場合について説明する．

まず係数関数 $g_j(z)$ を

$$g_j(z) = g_{cj} + g_{nj}(z), \tag{3.3}$$

ここで $\int_0^1 g_{nj}(z)dz = 0$，のように分解する．それぞれについて有効な添え字集合を

$$\mathcal{S}_c = \{j \,|\, g_{cj} \neq 0\} \quad \text{かつ} \quad \mathcal{S}_n = \{j \,|\, g_{nj}(z) \not\equiv 0\} \tag{3.4}$$

とする．j が前者にのみ含まれていれば，それは部分線形モデルの線形部分に対応し，後者にも含まれていれば変動係数部分に対応する．

そして (3.3) の分解に対応して，$[0,1]$ 上の区分的 2 次関数である L 次元の B スプライン基底 $\boldsymbol{B}_0(z)$ に対して適当な行列 A により，以下のような正規直交化を行う．

$$\overline{\boldsymbol{B}}(z) = \begin{pmatrix} 1/\sqrt{L} \\ \boldsymbol{B}(z) \end{pmatrix} = A\boldsymbol{B}_0(z) \quad \text{かつ} \quad \int_0^1 \overline{\boldsymbol{B}}(z)\overline{\boldsymbol{B}}^T(z)dz = \frac{1}{L}I. \tag{3.5}$$

$1/\sqrt{L}$ が g_{cj} に対応し，$\boldsymbol{B}(z)$ が $g_{nj}(z)$ に対応する．新しい回帰係数 $\boldsymbol{W}_i$ は 1.2 節の変動係数の項のように，$\overline{\boldsymbol{B}}(Z_i)$ と $\boldsymbol{X}_i \otimes \overline{\boldsymbol{B}}(Z_i)$ を合わせたものである．ここで $\otimes$ はクロネッカー積である．対応する回帰係数を $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0^T, \ldots, \boldsymbol{\gamma}_p^T)^T \in R^{(p+1)L}$ として，$\ell(\boldsymbol{\gamma})$ は

$$\ell(\boldsymbol{\gamma}) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \boldsymbol{\gamma}^T\boldsymbol{W}_i dN_i(t) + \int_0^\tau \log\Big\{\sum_{i=1}^{n} Y_i(t)\exp(\boldsymbol{\gamma}^T\boldsymbol{W}_i)\Big\}d\overline{N}(t) \tag{3.6}$$

となる．$\boldsymbol{\gamma}_j$ を $1/\sqrt{L}(g_{cj}$に対応) と $\boldsymbol{B}(z)(g_{nj}(z)$ に対応) に合わせて，$\boldsymbol{\gamma}_j = (\gamma_{1j}, \boldsymbol{\gamma}_{-1j}^T)^T$ と分け，

$$P_1(\boldsymbol{\gamma}) = \lambda\sum_{j=0}^{p}(|\gamma_{1j}| + \|\boldsymbol{\gamma}_{-1j}\|). \tag{3.7}$$

と group Lasso のペナルティーを定義する．部分線形モデルの構造の特定化に興味がなく，変動係数モデルの変数選択のみを扱うときは，$\sum_{j=0}^{p}\lambda\|\boldsymbol{\gamma}_j\|$ が group Lasso のペナルティーとなる．これらについては，Huang 他 (2013) に倣ってオラクル不等式などが証明できる．加法モデルについても多少の修正を行うだけで同様に扱うことができる．Lasso 推定量には変数選択の一致性はないので，この後に SCAD を行うあるいは統計的検定を行う (変数選択後ではあるが) 等の必要がある．

4. 高次元生存時間データのスクリーニングに関する研究

高次元生存時間データのスクリーニングに関する研究は多くあるが，その中からいくつか選んで紹介する．Hong と Li(2017) は，高次元生存時間データのスクリーニングに関するサーベイ論文である．ここでは Cox 回帰モデルに密接に関連した手法の説明から始める．

Zhao と Li(2012)：周辺モデルによる SIS 法である．まず $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ から j 番目の説明変数 X_j を取り出して，説明変数一つで (1.4) の $\ell(\boldsymbol{\beta})$ を定義するが，これを $\ell_j(\beta_j)$ と書くことにする．この $\ell_j(\beta_j)$ の最小化により，この周辺モデルの最大部分尤度推定量 $\widehat{\beta}_j$ を定義する．疑似最大部分尤度推定量と呼んでもよいかもしれない．この論文では，

$$I_j(\widehat{\beta}_j)^{1/2}|\widehat{\beta}_j| \geq \gamma_n,$$

ここで $I_j(\beta_j) = \ell_j''(\beta_j)$，により説明変数のスクリーニングを行う．$\gamma_n$ については論文を参照されたい．Hong 他 (2018) では，事前情報により添え字集合 $\mathcal{C}$ を回帰に含めることとして，$(\boldsymbol{X}_{\mathcal{C}}, X_j)$ による部分尤度を考え，X_j の係数によるスクリーニングを考察している．

Yang 他 (2016)：有効な添え字集合 $\mathcal{S}$ の要素数 s の上限 m が所与とする．ここでは，

$$\widetilde{\boldsymbol{\beta}}_m = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\, \ell(\boldsymbol{\beta}), \quad \text{ただし } \|\boldsymbol{\beta}\|_0 \leq m$$

の $\ell(\boldsymbol{\beta})$ かわりに，

$$g(\boldsymbol{\gamma}|\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^T\dot{\ell}(\boldsymbol{\beta}) - \frac{u}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^T W(\boldsymbol{\gamma} - \boldsymbol{\beta}), \tag{4.1}$$

ここで W は $-\ddot{\ell}(\boldsymbol{\beta})$ の対角要素からなる対角行列，を考え，Yang 他 (2016) に与えられるアルゴリズムにより最小化を行う (元論文は最大化．少し記号が異なることに注意)．アルゴリズムの中で u と $\boldsymbol{\beta}$ は更新される．Yang 他 (2019) は同様の方法で変動係数モデルも扱っている．

その他，Cox 回帰モデルに関連したスクリーニング法についての論文としては，

$$\boldsymbol{d} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \{\boldsymbol{X}_i - \overline{\boldsymbol{X}}(t)\}dN_i(t), \quad \text{ここで } \overline{\boldsymbol{X}}(t) = \sum_{i=1}^{n}\boldsymbol{X}_i Y_i(t) / \sum_{i=1}^{n} Y_i(t),$$

を用いる Gorst-Rasmussen と Scheike(2013)，あるいは前進型スクリーニング法を扱った Hong 他 (2019) などがある．

最後に，右側打ち切りを考慮した，生存時間と説明変数の関係の指標によるスクリーニング法を紹介する．指標が一つあれば論文が一つ書けるという時期はさすがに過ぎたが，ややアドホックな手法であるという印象を与えるものも多いので，代表的な手法を一つ紹介するにとどめる．

Song 他 (2014)：打ち切り時間 C_i は，T_{0i} や $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ とは独立とする．また $S(t) = \mathrm{P}(C_i \geq t)$ とおく．このとき，

$$\mathrm{E}\Big[\frac{\delta_j}{S^2(T_j)} I\{X_{ik} > X_{jk}, T_i > T_j\}\Big] = \mathrm{P}(X_{ik} > X_{jk}, T_{0i} > T_{0j}) \tag{4.2}$$

であり，X_{ik} と T_{0i} が独立かつタイ等がなければ，(4.2) は 1/4 である．$\tau_k = \mathrm{P}(X_{ik} > X_{jk}, T_{0i} > T_{0j}) - 1/4$ は，

$$\widehat{\tau}_k = \binom{n}{2}^{-1} \sum_{i<j} \frac{\delta_j}{\widehat{S}^2(T_j)} I\{X_{ik} > X_{jk}, T_i > T_j\} - \frac{1}{4} \tag{4.3}$$

で推定できる．そして $|\widehat{\tau}_k|$ により，説明変数のスクリーニングを行う．

この他に，周辺分位点回帰モデルによる He 他 (2013)，Li 他 (2016)，Zhong 他 (2021) などがある．

謝辞

本論文の執筆をお勧めくださった，久留米大学江村剛志先生に深く感謝いたします．また本論文作成にあたり，日本学術振興会科学研究費 (課題番号 20K11705) の補助を受けています．

参 考 文 献

Bickel, P. J., Ritov, Y. A. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37, 1705–1732.

Bradic, J., Fan, J., and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Annals of Statistics*, 39, 3092–3120.

Cai, T., Liu, W. and Luo, X. (2011) A constrained $\ell 1$ minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106, 594–607.

Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics*, 30, 93–111.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35, 2313–2351.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95, 759–771.

Chen, S. and Zhou, L. (2007). Local partial likelihood estimation in proportional hazards regression. *Annals of Statistics*, 35, 888–916.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34, 187–202.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications, Chapman and Hall/CRC*, London.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106, 544–557.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96, 1348–1360.

Fan, J., Li, R., Zhang, C. H. and Zou, H. (2020). *Statistical Foundations of Data Science, Chapman and Hall/CRC*, Boca Raton.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 70, 849–911.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Annals of Statistics*, 38, 3567–3604.

Gorst-Rasmussen, A. and Scheike, T. (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *Journal of the Royal Statistical Society: Series B*, 75, 217–245.

He, X., Wang, L. and Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Annals of Statistics*, 41, 342–369.

Honda, T. (2004). Nonparametric regression in proportional hazards models. *Journal of the Japan Statistical Society*, 34, 1–17.

Honda, T. and Härdle, W. K. (2014). *Journal of Statistical Planning and Inference*, 148, 67–81.

Honda, T. and Lin, C. T. (2021). Forward variable selection for sparse ultra-high-dimensional generalized varying coefficient models. *Japanese Journal of Statistics and Data Science*, 4, 151–179.

Honda, T. and Yabe, R. (2017). Variable selection and structure identification for varying coefficient Cox models *Journal of Multivariate Analysis*, 161, 103–122

Hong, H. G., Kang, J. and Li, Y. (2018). Conditional screening for ultra-high dimensional covariates with survival outcomes. *Lifetime Data Analysis*, 24, 45–71.

Hong, H. G. and Li, Y. (2017). Feature selection of ultrahigh-dimensional covariates with survival outcomes: a selective review. *Applied Mathematics-A Journal of Chinese Universities*, 32, 379–396.

Hong, H. G., Zheng, Q. and Li, Y. (2019). Forward regression for Cox models with high-dimensional covariates. *Journal of Multivariate Analysis*, 173, 268–290.

Huang, J. Z., Kooperberg, C., Stone, C. J., and Truong, Y. K. (2000). Functional ANOVA modeling for proportional hazards regression. *Annals of Statistics*, 28, 961–999.

Huang, J. (1999). Efficient estimation of the partly linear additive Cox model. *Annals of Statistics*, 27, 1536–1563.

Huang, J., Sun, T., Ying, Z., Yu, Y. and Zhang, C. H. (2013). Oracle inequalities for the lasso in the Cox model. *Annals of statistics*, 41, 1142–1165.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15, 2869–2909.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data, John Wiley & Sons*, Hoboken.

Kong, S., Yu, Z., Zhang, X. and Cheng, G. (2021). High-dimensional robust inference for Cox regression models using desparsified Lasso. *Scandinavian Journal of Statistics*, 48, 1068–1095.

Li, J., Zheng, Q., Peng, L. and Huang, Z. (2016). Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes. *Biometrics*, 72, 1145–1154.

Liu, J., Zhong, W. and Li, R. (2015). A selective overview of feature screening for ultrahigh-dimensional data. *Science China Mathematics*, 58, 1–22.

Meier, L., van de Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*, 70, 53–71.

Schumaker L. L. (2007). *Spline Functions: Basic Theory, Cambridge University Press*, New York.

Song, R., Lu, W., Ma, S. and Jeng, J. X. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika*, 101, 799–814.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58, 267–288.

van de Geer, S., Bühlmann, P., Ritov, Y. A., & Dezeure, R. (2014). On asymptotically optimal confidenc-e regions and tests for high-dimensional models. *Annals of Statistics*, 42, 1166–1202.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104, 1512–1524.

Yan, J. and Huang, J. (2012). Model selection for Cox models with time-varying coefficients. *Biometrics*, 68, 419–428.

Yang, G., Yu, Y., Li, R. and Buu, A. (2016). Feature screening in ultrahigh dimensional Cox's model. *Statistica Sinica*, 26, 881–901.

Yang, G., Zhang, L., Li, R. and Huang, Y. (2019). Feature screening in ultrahigh-dimensional varying-coefficient Cox model. *Journal of Multivariate Analysis*, 171, 284–297.

Yu, Y., Bradic, J. and Samworth, R. J. (2021) Confidence intervals for high-dimensional Cox models. *Statistica Sinica*, 31. 243–267.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38, 894–942.

Zhang, C. H. & Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, 76, 217–242.

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94, 691–703.

Zhao, S. D. and Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, 105, 397–411.

Zhong, W., Wang, J. and Chen, X. (2021). Censored mean variance sure independence screening for ultrahigh dimensional survival data. *Computational Statistics & Data Analysis*, 159, 107206.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.