数理解析研究所講究録
第2018巻 2017年 116-130



# スパース推定における情報量規準

名古屋工業大学　梅津 佑太
Yuta Umezu
Nagoya Institute of Technology

**概要**

スパース推定は推定関数に適切な罰則項を付加することでパラメータ推定と変数選択を同時に実行できる手法であり，生命科学や機械学習，統計解析などで広く利用されている．これまで，スパース推定により得られる推定量の統計的性質やパラメータ推定のためのアルゴリズムなどの観点からSCADやMCPなどの様々な罰則項が提案されてきた．本稿では，これらを含むスパース推定法により得られる推定量の漸近的性質について解説する．また，一般化線形モデルを用いた際のスパース推定において，情報量規準AICを用いた調整パラメータの選択法を紹介する．

## 1 はじめに

スパース推定とは，推定関数に原点で微分不可能な罰則項を付加してパラメータ推定を行う手法である．これにより，スパースな解 (いくつかの成分が正確に0であるような解) が得られやすく，結果としてパラメータ推定とともに変数選択を実行することができる．スパース推定で最も基本的なものは $\ell_1$ 罰則を用いるLasso (Tibshirani 1996) である．推定関数が凸関数であればLassoは凸最適化問題として定式化できるため実用上非常に有用であるにもかかわらず，推定量を縮小しすぎてしまうため推定の有効性が高くないなどの問題がある．このような問題点を解消するためSCAD (smoothly clipped absolute deviation: Fan & Li 2001) やMCP (minimax concave penalty: Zhang 2010) など様々な罰則が提案されている．また，提案された背景はSCADやMCPと異なるが，Frank & Friedman (1993) によって提案された $\ell_\gamma$ ($\gamma \in (0,1)$) 罰則を用いるBridge推定もLassoの問題を解消するものとして知られている．通常，変数選択に必要な計算コストはパラメータの次元に対して指数的に増加するため，パラメータ推定のための計算コストのみで変数選択を実行できるスパース推定は有効な手法であるといえる．スパース推定はサンプルサイズが少なくても有効に働くことがあるため，近年では高次元ベクトルや量子トモグラフィーにおける密度行列の推定などでも用いられている．

スパース推定では，最小化すべき関数に罰則の強さを制御する調整パラメータが含まれており，推定量に対する統計的漸近理論を構築する上ではその性質が重要となる．また，実際の問題に適用する際には，調整パラメータを客観的に選択しなければならないことも重要な課題として知られている．CV (Stone 1974) などの計算機的手法を用いて調整パラメータを選択することが多いが，一般にこのような選択法は計算コストが大きくなりがちである．一方，AIC (Akaike 1973) やBIC (Schwarz 1978) などの情報量規準を用いた選択法も発展しており，上記の手法よりも計算コストの小さなものである (例えばEfron et al. 2004; Wang et al. 2007; Ninomiya & Kawano 2014; Umezu et al. 2015).

本稿では，一般化線形モデル (McCullagh & Nelder 1989) に対するスパース推定において，推定

量の漸近的な性質と情報量規準 AIC について Umezu et al. (2015); Umezu & Ninomiya (2016) に沿って紹介する.

## 2 モデルと仮定

確率ベクトル $\boldsymbol{y} \in \mathbb{R}^r$ に対して，自然パラメータ $\boldsymbol{\theta} \in \Theta\ (\subset \mathbb{R}^r)$ を持つ指数型分布族を考える．このとき，ある $\sigma$-有限測度に関して $\boldsymbol{y}$ の確率密度関数は

$$f(\boldsymbol{y};\boldsymbol{\theta}) = \exp(\boldsymbol{y}^\top\boldsymbol{\theta} - a(\boldsymbol{\theta}) + b(\boldsymbol{y}))$$

で与えられる．$\Theta$ は自然パラメータ空間，つまり，$\boldsymbol{\theta} \in \Theta$ に対して $0 < \int \exp(\boldsymbol{y}^\top\boldsymbol{\theta} + b(\boldsymbol{y}))d\boldsymbol{y} < \infty$ を仮定する．このとき，$\Theta$ の内部 $\Theta^\circ$ で $\boldsymbol{y}$ の任意の次数のモーメントが存在し，特に $\mathrm{E}[\boldsymbol{y}] = a'(\boldsymbol{\theta})]$ や $\mathrm{V}[\boldsymbol{y}] = a''(\boldsymbol{\theta})$ と表すことができる[*1]．また，$V[\boldsymbol{y}]$ の正定値性，つまり $\boldsymbol{\theta}$ に関して $-\log f(\boldsymbol{y};\boldsymbol{\theta})$ は狭義凸関数であると仮定する．

さて，観測データを $\{(\boldsymbol{y}_i, \boldsymbol{X}_i) \in \mathbb{R}^r \times \mathcal{X};\ i = 1,\ldots,n\}$ $(\mathcal{X} \subset \mathbb{R}^{r\times p})$ とし $\boldsymbol{y}_i$ は独立な目的変数ベクトル，$\boldsymbol{X}_i$ は既知の非確率的な説明変数行列とする．いま，自然連結関数をもつ一般化線形モデル，つまり，確率密度関数のクラスとして $\{f(\boldsymbol{y};\boldsymbol{X}\boldsymbol{\beta});\ \boldsymbol{\beta} \in \mathcal{B}\}$ であるものを考える．ここで，開凸集合 $\mathcal{B} \subset \mathbb{R}^p$ に対して $\boldsymbol{\beta} \in \mathcal{B}$ は推定すべきパラメータである．また，$g_i(\boldsymbol{\beta}) = \log f(\boldsymbol{y}_i;\boldsymbol{X}_i\boldsymbol{\beta})$ を $\boldsymbol{y}_i$ の対数尤度関数とする．このモデルに対する漸近理論を構築するため，$\{\boldsymbol{X}_i\}$ の振る舞いに関して次の仮定をおく：

(C1) $\mathcal{X}$ はコンパクトであり，任意の $\boldsymbol{X} \in \mathcal{X}$ と $\boldsymbol{\beta} \in \mathcal{B}$ に対して $\boldsymbol{X}\boldsymbol{\beta} \in \Theta^\circ$ である
(C2) $\mathcal{X}$ 上の不変分布 $\mu$ が存在する．特に，$n^{-1}\sum_{i=1}^n \boldsymbol{X}_i^\top a''(\boldsymbol{X}_i\boldsymbol{\beta})\boldsymbol{X}_i$ は正定値行列

$$\boldsymbol{J}(\boldsymbol{\beta}) = \int_{\mathcal{X}} \boldsymbol{X}^\top a''(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{X}\mu(d\boldsymbol{X})$$

に収束する．

このとき，次の補題が成り立つ．証明は Ninomiya & Kawano (2014) や Umezu et al. (2015) を参照してほしい．

**補題 1.** (C1) および (C2) を仮定する．$\boldsymbol{\beta}^* \in \mathcal{B}$ を $\boldsymbol{\beta}$ の真値とする．このとき，次が成り立つ：

(R1) 任意の $\boldsymbol{\beta}$ に対して $n^{-1}\sum_{i=1}^n\{g_i(\boldsymbol{\beta}^*) - g_i(\boldsymbol{\beta})\} \xrightarrow{\mathrm{p}} h(\boldsymbol{\beta})$ であるような凸関数 $h(\boldsymbol{\beta})$ が存在する．
(R2) $\boldsymbol{J}_n(\boldsymbol{\beta}) \equiv -n^{-1}\sum_{i=1}^n g_i''(\boldsymbol{\beta})$ は $\boldsymbol{J}(\boldsymbol{\beta})$ に収束する．
(R3) $\boldsymbol{s}_n \equiv n^{-1/2}\sum_{i=1}^n g_i'(\boldsymbol{\beta}^*) \xrightarrow{\mathrm{d}} \boldsymbol{s} \sim \mathrm{N}(0, \boldsymbol{J}(\boldsymbol{\beta}^*))$.

(C2) より，$h(\boldsymbol{\beta})$ は

$$h(\boldsymbol{\beta}) = \int_{\mathcal{X}} \{a'(\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{X}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) - (a(\boldsymbol{X}\boldsymbol{\beta}^*) - a(\boldsymbol{X}\boldsymbol{\beta}))\mu(d\boldsymbol{X})$$

[*1] 関数 $f$ に対して $f'$ および $f''$ で $f$ の 1 階微分および 2 階微分を表す．

と陽に表現でき，唯一の最小化点 $\beta^*$ を持つことに注意する．以下では，次の罰則付き最尤法を考える：

$$\hat{\beta}_\lambda = \underset{\beta \in \mathcal{B}}{\operatorname{argmin}} \; -\frac{1}{n}\sum_{i=1}^{n} g_i(\beta) + \sum_{j=1}^{p} \eta_{\lambda_n}(\beta_j). \tag{1}$$

ここで，$\eta_{\lambda_n}(\cdot)$ はパラメータに関して凸とは限らない非負の罰則項であり，$\lambda_n$ $(>0)$ は調整パラメータである．特に，Bridge 罰則は正数 $\gamma$ を用いて $\eta_{\lambda_n}(\cdot) = \lambda_n|\beta|^\gamma$ と表される．数学的には $\gamma$ は正であれば問題ないが，スパースな解を得るためには $0 < \gamma \leq 1$ でなければならない．$\gamma = 1$ の場合は Lasso そのものである．図 1 は Bridge，SCAD，MCP のグラフを示している．SCAD，MCP で用いられる罰則の関数形については Fan & Li (2001) や Zhang (2010) を参照してほしい．SCAD や MCP は原点近傍で Lasso と同様に振る舞うため，$\ell_1$-型罰則とよばれ，これを用いて得られる (1) の推定量を $\ell_1$-型正則化推定量とよぶことにする．また，Lasso とは異なり SCAD，MCP は一様に有界な罰則項である．このことにより，SCAD や MCP は大きな値を持つ推定量を過剰に縮小することを防ぐことができる．

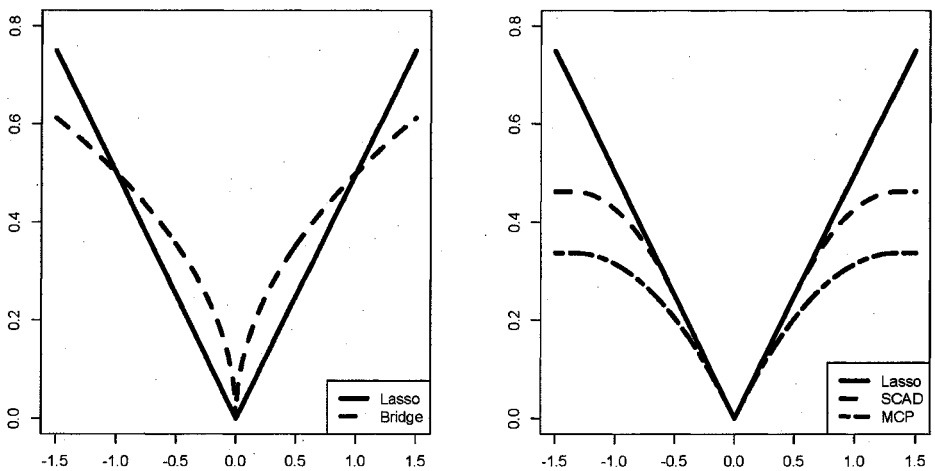


図 1　(左)：Bridge 罰則のグラフ．Lasso とは異なり原点での微分が発散する．(右)：SCAD，MCP で用いられる罰則のグラフ．原点近傍では Lasso と同様に振る舞うが，一様有界である．

さて，Bridge 罰則と $\ell_1$-型罰則の振る舞いの違いにより，推定量が良い漸近的性質を持つための $\lambda_n$ のオーダーはやはり異なることに注意する．具体的には，以下では $\lambda_n = n^{(\gamma_0-2)/2}\lambda$ とし，特に断らない限り，Bridge 罰則を考える場合は $\gamma < \gamma_0 \leq 1$，$\ell_1$-型罰則を考える場合は $1 \leq \gamma_0 < 2$ とする．いずれの罰則を用いた場合でも $\lambda_n \to 0$ であることに注意する．さらに，$\ell_1$-型罰則には以下の条件を仮定する：

(P1) $\eta_{\lambda_n}(\beta)$ は原点でのみ微分不可能であり，$\beta=0$ に関して対称かつ $|\beta|$ に関して単調非減少.
(P2) 任意の $\beta$ に対して $\lim_{n\to\infty}\eta_{\lambda_n}(\beta)=0$.
(P3) $\lim_{\beta\to 0}\eta_{\lambda_n}(\beta)/|\beta|=\lambda_n$.
(P4) ある $\tau>0$ が存在して任意の $\beta\geq\tau\lambda_n$ に対して $\eta'_{\lambda_n}(\beta)=0$.
(P5) $\beta\neq 0$ に対して $\lim_{n\to\infty}\eta''_{\lambda_n}(\beta)=0$.

(P1) はスパースな解を得るための基本的な条件であり，(P2) は漸近的に罰則が消えることを表している．また，(P3) は $\ell_1$-型罰則が原点近傍で Lasso と同じように振る舞うことを表している．なお，(P3) より $\eta_{\lambda_n}(0)=0$ が成り立ち，$\beta\neq 0$ に対して $\mathrm{sgn}(\beta)\neq 0$ であることと (P2) から

$$\lim_{\beta\to 0}\eta'_{\lambda_n}(\beta)/\mathrm{sgn}(\beta)=\lambda_n \tag{2}$$

が成り立つ (L'Hospital の定理)．ただし，$\mathrm{sgn}(\beta)$ は $\beta>0\,(<0)$ ならば $\mathrm{sgn}(\beta)=+1\,(-1)$，$\beta=0$ ならば $\mathrm{sgn}(\beta)=0$ を返す符号関数である．(P4) により罰則項は一様に有界であり，(P5) はややテクニカルであるが漸近分布の導出に必要となる条件である．これらの条件は SCAD や MCP など多くの罰則でみたされる．

## 3 推定量の漸近的性質

### 3.1 準備

以下では，$\mathcal{J}^{(1)}=\{j;\ \beta^*_j=0\}$ および $\mathcal{J}^{(2)}=\{j;\ \beta^*_j\neq 0\}$ とする．また，混乱のない限り $\boldsymbol{J}=\boldsymbol{J}(\boldsymbol{\beta}^*)$ とし，ベクトル $(\beta_j)_{j\in\mathcal{J}^{(k)}}$ や行列 $(\boldsymbol{J}_{ij})_{i\in\mathcal{J}^{(k)},j\in\mathcal{J}^{(l)}}$ を $\boldsymbol{\beta}^{(k)}$ や $\boldsymbol{J}^{(kl)}$ と表すことにする．さらに，$\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}^{(1)},\hat{\boldsymbol{\beta}}^{(2)})$ などと表記する．

本節では，(1) で与えられる推定量に対して次の 3 つの性質を考える：

(スパース性)： $\mathrm{P}(\hat{\boldsymbol{\beta}}^{(1)}_\lambda=\boldsymbol{0})\to 1$
(漸近正規性)： $\sqrt{n}(\hat{\boldsymbol{\beta}}^{(2)}_\lambda-\boldsymbol{\beta}^{*(2)})\xrightarrow{\mathrm{d}}\mathrm{N}(\boldsymbol{0},\boldsymbol{J}^{(22)})$
(変数選択の一致性)： $\mathrm{P}(\hat{\boldsymbol{\beta}}^{(2)}_\lambda\neq\boldsymbol{0})\to 1$

上記の漸近正規性および，スパース性または変数選択の一致性を有する推定量は oracle property を持つと呼ばれる (例えば Fan & Li 2001; Zou 2006)．まず，Knight & Fu (2000) や Umezu et al. (2015) と同様の議論により推定量の一致性が成立する．

**補題 2.** $\gamma_0<2$ とする．このとき，(C1)，(C2) のもと Bridge 推定量は $\boldsymbol{\beta}^*$ の一致推定量である．さらに，(P1)，(P2) を仮定する．このとき，$\ell_1$-型正則化推定量は $\boldsymbol{\beta}^*$ の一致推定量である．

詳細は省くが，補題 2 はランダム関数

$$\mathbb{G}_n(\boldsymbol{u})=\frac{1}{n}\sum_{i=1}^{n}(g_i(\boldsymbol{\beta}^*)-g_i(\boldsymbol{\beta}^*+\boldsymbol{u}))+\sum_{j=1}^{p}(\eta_{\lambda_n}(\beta^*_j+u_j)-\eta_{\lambda_n}(\beta^*_j)) \tag{3}$$

の最小化点 $\hat{\boldsymbol{u}}_n=\hat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*$ が $\mathrm{O_p}(1)$ であること，$\mathbb{G}_n(\boldsymbol{u})$ が $\boldsymbol{u}$ のコンパクト集合上で一様に $h(\boldsymbol{\beta}^*+\boldsymbol{u})$ へ収束すること，および $h(\boldsymbol{\beta}^*+\boldsymbol{u})$ の最小化点が $\boldsymbol{0}$ であることから示される．

## 3.2 推定量のスパース性

Radchenko (2005) の結果を一般化線形モデルへ拡張することで，一致性よりも強い結果を示すことができる．そのための準備として，推定量の収束レートを導出しておく．Taylor の定理と補題 2 より，$\mathbb{G}_n(\hat{\boldsymbol{u}}_n)$ は

$$\begin{aligned}\mathbb{G}_n(\hat{\boldsymbol{u}}_n) = &- n^{-1/2}\boldsymbol{s}_n^\top \hat{\boldsymbol{u}}_n + 2^{-1}\hat{\boldsymbol{u}}_n^\top \boldsymbol{J}_n(\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{u}}_n \\ &+ \sum_{j\in\mathcal{J}^{(1)}} \eta_{\lambda_n}(\hat{u}_{n,j}) + \sum_{j\in\mathcal{J}^{(2)}} \eta'_{\lambda_n}(\beta_j^*)\hat{u}_{n,j}[1+\mathrm{o_p}(1)]\end{aligned}$$

と展開できる．ここで，$\tilde{\boldsymbol{\beta}}$ は $\hat{\boldsymbol{\beta}}_\lambda$ と $\boldsymbol{\beta}^*$ の間にあるベクトルである．第 1 項と第 3 項はそれぞれ $\mathrm{O_p}(n^{-1/2}\|\hat{\boldsymbol{u}}_n\|_2)$ および非負である．$\eta_{\lambda_n}(\cdot)$ として Bridge 罰則を用いる場合，第 4 項は $\mathrm{O_p}(\lambda_n\|\hat{\boldsymbol{u}}_n\|_2) = \mathrm{O_p}(n^{(\gamma_0-2)/2}\|\hat{\boldsymbol{u}}_n\|_2)$ となる．したがって，$\hat{\boldsymbol{u}}_n$ が $\mathbb{G}_n(\boldsymbol{u})$ の最小化点であることから

$$0 \geq \mathbb{G}_n(\hat{\boldsymbol{u}}_n) - \mathbb{G}_n(\boldsymbol{0}) \geq 2^{-1}\hat{\boldsymbol{u}}_n^\top \boldsymbol{J}_n(\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{u}}_n + \mathrm{O_p}(n^{-1/2}\|\hat{\boldsymbol{u}}_n\|_2) + \mathrm{O_p}(n^{(\gamma_0-2)/2}\|\hat{\boldsymbol{u}}_n\|_2)$$

となる．$\gamma_0 \leq 1$ および (R2) から十分大きな $n$ に対して $\boldsymbol{J}_n(\tilde{\boldsymbol{\beta}})$ は正定値行列であるため，

$$\hat{\boldsymbol{u}}_n = \mathrm{O_p}(\max\{n^{-1/2}, n^{(\gamma_0-2)/2}\}) = \mathrm{O_p}(n^{-1/2})$$

が成り立つ．一方，$\ell_1$-型罰則を用いる場合，(P4) より十分大きな $n$ に対して $\eta'_{\lambda_n}(\beta_j^*) = 0$, $j \in \mathcal{J}^{(2)}$ となることに注意すれば，同様の議論により $\hat{\boldsymbol{u}}_n = \mathrm{O_p}(n^{-1/2})$ が得られる．

**定理 1** (スパース性)．(C1)，(C2) のもと Bridge 推定量 $\hat{\boldsymbol{\beta}}_\lambda$ に対して $\mathrm{P}(\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \boldsymbol{0}) \to 1$ が成り立つ．さらに，(P1)–(P4) および $1 < \gamma_0 < 2$ を仮定する．このとき，$\ell_1$-型正則化推定量 $\hat{\boldsymbol{\beta}}_\lambda$ に対して $\mathrm{P}(\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \boldsymbol{0}) \to 1$ が成り立つ．

*Proof.* (3) を改めて $\mathbb{G}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ とすれば，

$$\begin{aligned}0 &\geq \mathbb{G}_n(\hat{\boldsymbol{u}}_n^{(1)}, \hat{\boldsymbol{u}}_n^{(2)}) - \mathbb{G}_n(0, \hat{\boldsymbol{u}}_n^{(2)}) \\ &= -n^{-1/2}\boldsymbol{s}_n^{(1)\top}\hat{\boldsymbol{u}}_n^{(1)} + 2^{-1}\hat{\boldsymbol{u}}_n^{(1)\top}\boldsymbol{J}_n^{(11)}(\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{u}}_n^{(1)} + \hat{\boldsymbol{u}}_n^{(1)\top}\boldsymbol{J}_n^{(12)}(\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{u}}_n^{(2)} + \sum_{j\in\mathcal{J}^{(1)}} \eta_{\lambda_n}(\hat{u}_{n,j})\end{aligned}$$

となる．ここで，$\tilde{\boldsymbol{\beta}}$ は $\hat{\boldsymbol{\beta}}_\lambda$ と $\boldsymbol{\beta}^*$ の間にあるベクトルである．いま，右辺第 1 項は $\mathrm{O_p}(n^{-1/2}\|\hat{\boldsymbol{u}}_n\|_2)$ であり，$\hat{\boldsymbol{u}}_n = \mathrm{O_p}(n^{-1/2})$ および (R2) から右辺第 3 項は

$$|\hat{\boldsymbol{u}}_n^{(1)\top}\boldsymbol{J}_n^{(12)}(\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{u}}_n^{(2)}| = \mathrm{O_p}(n^{-1/2}\|\hat{\boldsymbol{u}}_n^{(1)}\|_2)$$

である．また，(P3) より右辺第 4 項は Bridge 罰則と $\ell_1$-型罰則に対してそれぞれ

$$\sum_{j\in\mathcal{J}^{(1)}} \eta_{\lambda_n}(\hat{u}_{n,j}) = \lambda_n\|\hat{\boldsymbol{u}}_n^{(1)}\|_\gamma^\gamma, \qquad \sum_{j\in\mathcal{J}^{(1)}} \eta_{\lambda_n}(\hat{u}_{n,j}) = \lambda_n\|\hat{\boldsymbol{u}}_n^{(1)}\|_1[1+\mathrm{o_p}(1)]$$

となることに注意する．Bridge 罰則を用いる場合，(R2) から十分大きな $n$ に対して $\boldsymbol{J}_n^{(12)}(\tilde{\boldsymbol{\beta}})$ は正定値行列であるため，

$$\|\hat{\boldsymbol{u}}_n^{(1)}\|_2^2 + \lambda_n\|\hat{\boldsymbol{u}}_n^{(1)}\|_\gamma^\gamma \leq \mathrm{O_p}(n^{-1/2}\|\hat{\boldsymbol{u}}_n^{(1)}\|_2)$$

より，

$$\|n^{1/2}\hat{\boldsymbol{u}}_n^{(1)}\|_2^2 + \lambda_n n^{(2-\gamma)/2}\|n^{1/2}\hat{\boldsymbol{u}}_n^{(1)}\|_\gamma^\gamma \leq \mathrm{O_p}(\|n^{1/2}\hat{\boldsymbol{u}}_n^{(1)}\|_2)$$

を得る．$\gamma < \gamma_0 \leq 1$ より $\lambda_n n^{(2-\gamma)/2} = n^{(\gamma_0-\gamma)/2}\lambda \to \infty$ であるが，$n^{1/2}\hat{\boldsymbol{u}}_n = \mathrm{O_p}(1)$ であるため $\mathrm{P}(\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \mathbf{0}) \to 1$ が成り立つ．同様に，$\ell_1$-型罰則に対しては

$$\|n^{1/2}\hat{\boldsymbol{u}}_n^{(1)}\|_2^2 + \lambda_n n^{1/2}\|n^{1/2}\hat{\boldsymbol{u}}_n^{(1)}\|_1[1+\mathrm{o_p}(1)] \leq \mathrm{O_p}(\|n^{1/2}\hat{\boldsymbol{u}}_n^{(1)}\|_2)$$

であるが，$1 < \gamma_0 < 2$ より $\lambda_n n^{1/2} = n^{(\gamma_0-1)/2}\lambda \to \infty$ であるため Bridge 罰則の場合と同様に $\mathrm{P}(\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \mathbf{0}) \to 1$ が成り立つ． □

$\ell_1$-型罰則において $\gamma_0 = 1$ の場合，定理 1 は成立しないことに注意する．結果として，$\gamma_0 = 1$ の場合の $\ell_1$-型正則化推定量に対して $\hat{\boldsymbol{\beta}}_\lambda^{(1)}$ の漸近分布にはランダムなバイアスが残ることが示される．一方，Bridge 推定量に対しても $\gamma_0 = 1$ ならば非確率的なバイアスが残る．そのため，以下では $\gamma_0 \neq 1$ と $\gamma_0 = 1$ の場合でそれぞれ漸近分布と変数選択の一致性について議論する．

## 3.3 $\gamma_0 \neq 1$ の場合

ランダムな関数

$$\mathbb{H}_n(\boldsymbol{\beta}) = -\sum_{i=1}^n g_i(\boldsymbol{\beta}) + n\sum_{j=1}^p \eta_n(\beta_j) \tag{4}$$

を考える．補題 2 および定理 1 より，$\hat{\boldsymbol{\beta}} = \mathrm{O_p}(n^{-1/2})$ であり，十分大きな $n$ に対して $\hat{\boldsymbol{\beta}}_\lambda^{(2)}$ は $\mathbf{0}$ ではない．従って，1 に収束する確率で $\hat{\boldsymbol{\beta}}_\lambda^{(2)}$ は尤度方程式の解となる：

$$\left.\frac{\partial \mathbb{H}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(2)}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_\lambda} = -\sum_{i=1}^n g_i'^{(2)}(\hat{\boldsymbol{\beta}}_\lambda) + n\boldsymbol{\eta}_n'(\hat{\boldsymbol{\beta}}_\lambda^{(2)}) = \mathbf{0} \tag{5}$$

ただし，$\boldsymbol{\eta}_n'(\hat{\boldsymbol{\beta}}_\lambda^{(2)}) = (\eta_{\lambda_n}'(\hat{\beta}_{\lambda,j}))_{j\in\mathcal{J}^{(2)}}$ である．Taylor の定理より，$j \in \mathcal{J}^{(2)}$ に対して $\hat{\beta}_{\lambda,j}$ と $\beta_j^*$ の間にある実数 $\beta_j^\dagger$ が存在して

$$\boldsymbol{\eta}_n'(\hat{\boldsymbol{\beta}}_\lambda^{(2)}) = \boldsymbol{\eta}_n'(\hat{\boldsymbol{\beta}}^{*(2)}) + H_n(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)})$$

が成り立つ．ただし，$H_n$ は第 $j$-対角成分が $\eta_{\lambda_n}''(\beta_j^\dagger)$ であるような対角行列である．よって，$\hat{\boldsymbol{\beta}}_\lambda^{(2)} = \mathrm{O_p}(n^{-1/2})$，(P4) および (P5) から，十分大きな $n$ に対して $\boldsymbol{\eta}_n'(\hat{\boldsymbol{\beta}}_\lambda^{(2)}) = \mathrm{O_p}(n^{1/2})$ を得る．さらに，Taylor 展開により

$$g_i'^{(2)}(\hat{\boldsymbol{\beta}}_\lambda) = g_i'^{(2)}(\boldsymbol{\beta}^*) + g_i''^{(21)}(\boldsymbol{\beta}^*)\hat{\boldsymbol{\beta}}_\lambda^{(1)} + g_i''^{(22)}(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) + \mathrm{o_p}(1)$$

である．定理 1 より 1 に収束する確率で右辺第 2 項は $\mathbf{0}$ なので，(5) は

$$-n^{1/2}\boldsymbol{s}_n^{(2)} + n\boldsymbol{J}_n^{(22)}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) + \mathrm{o_p}(1) + \mathrm{o_p}(n^{1/2}) = \mathbf{0}$$

となることが分かる．したがって，(R2) から十分大きな $n$ に対して $\boldsymbol{J}_n^{(22)}$ は正定値なので

$$n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) = \boldsymbol{J}_n^{(22)-1}\boldsymbol{s}_n^{(2)} + \mathrm{o_p}(1)$$

であり，(R3) より $n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)})$ の漸近正規性が示される．Bridge 推定量の場合，$\gamma < \gamma_0 < 1$ ならば $\lambda_n = \mathrm{o}(1)$ であることと補題 2 から，$j \in \mathcal{J}^{(2)}$ に対して

$$\eta'_{\lambda_n}(\hat{\beta}_{\lambda,j}) = \lambda_n \gamma \mathrm{sgn}(\hat{\beta}_{\lambda,j})|\hat{\beta}_{\lambda,j}|^{\gamma-1} = \lambda_n \gamma \mathrm{sgn}(\beta_j^*)|\beta_j^*|^{\gamma-1}[1 + \mathrm{o_p}(1)] = \mathrm{o_p}(1)$$

に注意すれば同様に漸近正規性が示される．

**定理 2** (漸近正規性)．(C1)，(C2) のもと，$\gamma < \gamma_0 < 1$ ならば Bridge 推定量 $\hat{\boldsymbol{\beta}}_\lambda$ に対して

$$n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) = \boldsymbol{J}_n^{(22)-1}\boldsymbol{s}_n^{(2)} + \mathrm{o_p}(1)$$

が成り立つ．さらに，(P1)–(P5) および $1 < \gamma_0 < 2$ を仮定する．このとき，$\ell_1$-型正則化推定量 $\hat{\boldsymbol{\beta}}_\lambda$ に対して

$$n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) = \boldsymbol{J}_n^{(22)-1}\boldsymbol{s}_n^{(2)} + \mathrm{o_p}(1)$$

が成り立つ．

定理 2 より，Bridge 推定量と $\ell_1$-型正則化推定量は同じ漸近分布を持つが，$\gamma_0$ のとりうる値の範囲が異なる．これは，定理 1 において Bridge 推定量が原点近傍でスパース性を保証している一方で，$\ell_1$-型正則化推定量は原点から離れたところでスパース性を保証しているためであると解釈できる．

次に，変数選択の一致性について考える．これは $\mathrm{P}(\hat{\mathcal{J}}^{(2)} = \mathcal{J}^{(2)}) \to 1$ と同値である．ただし，$\hat{\mathcal{J}}^{(2)} = \{j;\ \hat{\beta}_{\lambda,j} \neq 0\}$ は $\hat{\boldsymbol{\beta}}_\lambda$ のアクティブセットと呼ばれる集合である．補題 2 より，$\gamma_0 < 2$ ならば (1) から得られる推定量は一致性を持つので，任意の $j \in \mathcal{J}^{(2)}$ に対して $\mathrm{P}(j \in \hat{\mathcal{J}}^{(2)}) \to 1$，つまり $\mathrm{P}(\hat{\mathcal{J}}^{(2)} \supset \mathcal{J}^{(2)}) \to 1$ が成り立つ．従って，変数選択の一致性の成立のためには次が成り立つことを確認すればよい：

$$\text{任意の } j \in \mathcal{J}^{(1)} \text{に対して } \mathrm{P}(j \in \hat{\mathcal{J}}^{(2)}) \to 0 \tag{6}$$

まず，$\ell_1$-型正則化推定量を考える．先に述べたように，$\ell_1$-型正則化推定量は 1 に収束する確率で (5) を満たすので，任意の $j \in \hat{\mathcal{J}}^{(2)}$ に対して

$$-\sum_{i=1}^n \frac{\partial g_i(\boldsymbol{\beta})}{\partial \beta_j}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_\lambda} + n\eta'_{\lambda_n}(\hat{\beta}_{\lambda,j}) = 0$$

が成立する．Taylor の定理より $\hat{\boldsymbol{\beta}}_\lambda$ と $\boldsymbol{\beta}^*$ の間にあるベクトル $\boldsymbol{\beta}^\dagger$ が存在して

$$-\sum_{i=1}^n \frac{\partial g_i(\boldsymbol{\beta})}{\partial \beta_j}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_\lambda} = -n^{1/2}s_{n,j} + n\sum_{k=1}^p \boldsymbol{J}_n(\boldsymbol{\beta}^\dagger)_{jk}(\hat{\beta}_{\lambda,k} - \beta_k^*)$$

となる．したがって，

$$s_{n,j} - \sum_{k=1}^p \boldsymbol{J}_n(\boldsymbol{\beta}^\dagger)_{jk}\{n^{1/2}(\hat{\beta}_{\lambda,k} - \beta_k^*)\} = n^{1/2}\eta'_{\lambda_n}(\hat{\beta}_{\lambda,j}) \tag{7}$$

が成り立つ．(R2) および $\boldsymbol{s}_n = \mathrm{O_p}(1)$，補題 2 より，(7) の左辺は $\mathrm{O_p}(1)$ である．さらに，$j \in \mathcal{J}^{(1)}$ と $n^{1/2}\lambda_n \to \infty$，(2) より (7) の左辺は確率的に発散する．以上より，任意の $j \in \mathcal{J}^{(1)}$ に対して

$$\mathrm{P}(j \in \hat{\mathcal{J}}^{(2)}) \leq \mathrm{P}\left(s_{n,j} - \sum_{k=1}^p \boldsymbol{J}_n(\boldsymbol{\beta}^\dagger)_{jk}\{n^{1/2}(\hat{\beta}_{\lambda,k} - \beta_k^*)\} = n^{1/2}\eta'_{\lambda_n}(\hat{\beta}_{\lambda,j})\right) \to 0$$

が成り立つ．Bridge 推定量を考える場合，(7) の代わりに

$$s_{n,j} - \sum_{k=1}^{p} \boldsymbol{J}_n(\boldsymbol{\beta}^\dagger)_{jk}\{n^{1/2}(\hat{\beta}_{\lambda,k} - \beta_k^*)\} = \gamma\lambda n^{(\gamma_0-\gamma)/2}\mathrm{sgn}(\hat{\beta}_{\lambda,j})|n^{1/2}\hat{\beta}_{\lambda,j}|^{\gamma-1}$$

を考えれば同様の結果が得られる．

**定理 3** (変数選択の一致性)．(C1), (C2) のもと，$\gamma < \gamma_0 < 1$ ならば Bridge 推定量 $\hat{\boldsymbol{\beta}}_\lambda$ に対して $\mathrm{P}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} \neq 0) \to 1$ が成り立つ．さらに，(P1)–(P5) および $1 < \gamma_0 < 2$ を仮定する．このとき，$\ell_1$-型正則化推定量 $\hat{\boldsymbol{\beta}}_\lambda$ に対して $\mathrm{P}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} \neq 0) \to 1$ が成り立つ．

## 3.4 $\gamma_0 = 1$ の場合

### 3.4.1 $\ell_1$-型正則化推定量の漸近分布

$\gamma_0 = 1$ の場合，$\ell_1$-型正則化推定量はスパース性を持たないため，$\gamma_0 \neq 1$ の場合と同様の議論を行うことができない．以下，推定量の漸近分布を導出するため，Hjort & Pollard (1993) を拡張した次の補題を用いる：

**補題 3.** $\phi_n(\boldsymbol{u})$ をランダムな狭義凸関数，$\tilde{\phi}_n(\boldsymbol{u})$ をその近似とする．非確率的な関数 $\psi_n(\boldsymbol{u})$ は $\boldsymbol{u}$ のコンパクト集合上で一様に凸関数 $\psi(\boldsymbol{u})$ へ収束するものとする．さらに，

$$\nu_n(\boldsymbol{u}) = \phi_n(\boldsymbol{u}) + \psi_n(\boldsymbol{u}) \quad \text{および} \quad \tilde{\nu}_n(\boldsymbol{u}) = \tilde{\phi}_n(\boldsymbol{u}) + \psi(\boldsymbol{u})$$

に対して，$\nu_n(\boldsymbol{u})$, $\tilde{\nu}_n(\boldsymbol{u})$ の最小化点を $\boldsymbol{u}_n$, $\tilde{\boldsymbol{u}}_n$ とする．このとき，任意の $\varepsilon\ (>0)$ と $\delta\ (>0)$, $\xi\ (>\delta)$ に対して

$$\mathrm{P}(|\boldsymbol{u}_n - \tilde{\boldsymbol{u}}_n| \geq \delta) \leq \mathrm{P}\left(2\Delta_n(\delta) + \varepsilon \geq \Upsilon_n(\delta)\right) + \mathrm{P}(|\boldsymbol{u}_n - \tilde{\boldsymbol{u}}_n| \geq \xi) \tag{8}$$

が成り立つ．ただし，

$$\Delta_n(\delta) = \sup_{|\boldsymbol{u}-\tilde{\boldsymbol{u}}_n|\leq\delta} |\nu_n(\boldsymbol{u}) - \tilde{\nu}_n(\boldsymbol{u})| \quad \text{および} \quad \Upsilon_n(\delta) = \inf_{|\boldsymbol{u}-\tilde{\boldsymbol{u}}_n|=\delta} \tilde{\nu}_n(\boldsymbol{u}) - \tilde{\nu}_n(\tilde{\boldsymbol{u}}_n) \tag{9}$$

である．

証明は Umezu et al. (2015); Umezu & Ninomiya (2016) を参照してほしい．

さて，補題 3 を用いて $\ell_1$-型正則化推定量の漸近分布を導出しよう．ランダムな関数

$$\nu_n(\boldsymbol{u}) = \phi_n(\boldsymbol{u}) + \psi_n(\boldsymbol{u})$$

を考える．ただし，

$$\phi_n(\boldsymbol{u}) = \sum_{i=1}^{n}\{g_i(\boldsymbol{\beta}^*) - g_i(\boldsymbol{\beta}^* + n^{-1/2}\boldsymbol{u})\}$$

および

$$\psi_n(\boldsymbol{u}) = n\sum_{j=1}^{p}\{\eta_{\lambda_n}(\beta_j^* + n^{-1/2}u_j) - \eta_{\lambda_n}(\beta_j^*)\}$$

とする．$\nu_n(\boldsymbol{u})$ の最小化点は $\boldsymbol{u}_n = (\boldsymbol{u}_n^{(1)}, \boldsymbol{u}_n^{(2)}) = (n^{1/2}\hat{\boldsymbol{\beta}}_\lambda^{(1)}, n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}))$ で与えられることに注意する．Taylor の定理より，$\phi_n(\boldsymbol{u})$ は

$$\tilde{\phi}_n(\boldsymbol{u}) = -\boldsymbol{u}^\top \boldsymbol{s}_n + \boldsymbol{u}^\top \boldsymbol{J}\boldsymbol{u}/2$$

で近似できる．一方，(P3) と (P4) より $\psi_n(\boldsymbol{u})$ は $\boldsymbol{u}$ のコンパクト集合上で一様に $\psi(\boldsymbol{u}) = \lambda\|\boldsymbol{u}^{(1)}\|_1$ に収束することが分かる．

以下，$\nu_n(\boldsymbol{u})$ などを改めて $\nu_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ などと表すことにする．このとき，$\tilde{\nu}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \tilde{\phi}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \psi(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ は

$$\begin{aligned}\tilde{\nu}(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = &\left\{\boldsymbol{u}^{(2)} - \boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{J}^{(21)}\boldsymbol{u}^{(1)})\right\}^\top \boldsymbol{J}^{(22)} \left\{\boldsymbol{u}^{(2)} - \boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{J}^{(21)}\boldsymbol{u}^{(1)})\right\}/2 \\ &+ \boldsymbol{u}^{(1)\top}\boldsymbol{J}^{(1|2)}\boldsymbol{u}^{(1)}/2 - \boldsymbol{u}^{(1)\top}\boldsymbol{s}_n^{(1|2)} + \lambda\|\boldsymbol{u}^{(1)}\|_1 - \boldsymbol{s}_n^{(2)\top}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)}/2\end{aligned}$$

とかけるので，$\tilde{\nu}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ の最小化点は $\tilde{\boldsymbol{u}}_n = (\tilde{\boldsymbol{u}}_n^{(1)}, \tilde{\boldsymbol{u}}_n^{(2)}) = (\hat{\boldsymbol{u}}_n^{(1)}, \boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{J}^{(21)}\hat{\boldsymbol{u}}_n^{(1)}))$ で与えられる．ただし，

$$\hat{\boldsymbol{u}}_n^{(1)} = \underset{\boldsymbol{u}^{(1)}}{\operatorname{argmin}}\{\boldsymbol{u}^{(1)\top}\boldsymbol{J}^{(1|2)}\boldsymbol{u}^{(1)}/2 - \boldsymbol{u}^{(1)\top}\boldsymbol{s}_n^{(1|2)} + \lambda\|\boldsymbol{u}^{(1)}\|_1\} \tag{10}$$

および，$\boldsymbol{J}^{(1|2)} = \boldsymbol{J}^{(11)} - \boldsymbol{J}^{(12)}\boldsymbol{J}^{(22)-1}\boldsymbol{J}^{(21)}$，$\boldsymbol{s}_n^{(1|2)} = \boldsymbol{s}_n^{(1)} - \boldsymbol{J}^{(12)}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)}$ である．三角不等式と $\phi_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \boldsymbol{u}^{(2)\top}\boldsymbol{s}_n^{(2)}$ の凸性，$\psi_n(\boldsymbol{u})$ の一様収束性より (9) の $\Delta_n(\delta)$ は 0 に確率収束する．また，(10) より，$\tilde{\boldsymbol{u}}_n^{(1)}$ は

$$\boldsymbol{J}^{(1|2)}\tilde{\boldsymbol{u}}_n^{(1)} - \boldsymbol{\tau}_\lambda(\boldsymbol{s}_n) + \lambda\boldsymbol{\gamma} = 0,$$

を満たす．ただし，$\boldsymbol{\gamma}$ は $\|\tilde{\boldsymbol{u}}_n^{(1)}\|_1$ の劣勾配である．したがって，$\tilde{\boldsymbol{u}}_n^{(1)\top}\boldsymbol{\gamma} = \|\tilde{\boldsymbol{u}}_n^{(1)}\|_1$ に注意すれば，$\tilde{\nu}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) - \tilde{\nu}_n(\tilde{\boldsymbol{u}}_n^{(1)}, \tilde{\boldsymbol{u}}_n^{(2)})$ は

$$\begin{aligned}&(\boldsymbol{u}^{(1)} - \tilde{\boldsymbol{u}}_n^{(1)})^\top \boldsymbol{J}^{(1|2)}(\boldsymbol{u}^{(1)} - \tilde{\boldsymbol{u}}_n^{(1)})/2 + \lambda \sum_{j\in\mathcal{J}^{(1)}} (|u_j| - \gamma_j u_j) \\ &+ \left\{\boldsymbol{u}^{(2)} - \boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{J}^{(21)}\boldsymbol{u}^{(1)})\right\}^\top \boldsymbol{J}^{(22)} \left\{\boldsymbol{u}^{(2)} - \boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{J}^{(21)}\boldsymbol{u}^{(1)})\right\}/2\end{aligned} \tag{11}$$

と書き換えることができる．$0 \le \zeta \le \delta$ に対して，$\boldsymbol{w}_1$ と $\boldsymbol{w}_2$ を $\boldsymbol{u}^{(1)} = \tilde{\boldsymbol{u}}_n^{(1)} + \zeta\boldsymbol{w}_1$ and $\boldsymbol{u}^{(2)} = \tilde{\boldsymbol{u}}_n^{(2)} + (\delta^2 - \zeta^2)^{1/2}\boldsymbol{w}_2$ であるような単位ベクトルとする．また，$\boldsymbol{J}^{(22)}$ と $\boldsymbol{J}^{(1|2)}$ の最小固有値の半分をそれぞれ $\rho^{(22)}$, $\rho^{(1|2)}$ とする．このとき，(11) の第 2 項が非負であることから

$$\Upsilon_n(\delta) \ge \min_{0\le\zeta\le\delta} \left\{\rho^{(1|2)}\zeta^2 + \rho^{(22)}|(\delta^2 - \zeta^2)^{1/2}\boldsymbol{w}_2 + \zeta\boldsymbol{J}^{(22)-1}\boldsymbol{J}^{(21)}\boldsymbol{w}_1|^2\right\} > 0$$

が成り立つ．したがって，十分小さな $\varepsilon$ と十分大きな $n$ に対して (8) の右辺第 1 項は任意に小さくすることができる．さらに，$\boldsymbol{u}_n$, $\tilde{\boldsymbol{u}}$ は $\mathrm{O_p}(1)$ なので，十分大きな $\xi$ に対して (8) の右辺第 2 項も任意に小さくすることができる．以上をまとめると，$\ell_1$-型正則化推定量の漸近分布は次で与えられる．

**定理 4** ($\ell_1$-型正則化推定量の漸近分布)．$\hat{\boldsymbol{u}}^{(1)}$ を (10) のものとする．このとき，(C1), (C2) および (P1)–(P5) のもと，$\ell_1$-型正則化推定量 $\hat{\boldsymbol{\beta}}_\lambda$ に対して

$$n^{1/2}\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \hat{\boldsymbol{u}}_n^{(1)} + \mathrm{o_p}(1) \tag{12}$$

および

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{\lambda}^{(2)} - \boldsymbol{\beta}^{*(2)}) = \boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{J}^{(21)}\hat{\boldsymbol{u}}_n^{(1)}) + \mathrm{o_p}(1) \tag{13}$$

が成り立つ.

定理 4 より，$\hat{\boldsymbol{\beta}}_{\lambda}^{(1)}$ はスパース性を持たず，その結果，$n^{1/2}(\hat{\boldsymbol{\beta}}_{\lambda}^{(2)} - \boldsymbol{\beta}^{*(2)})$ の漸近分布に対してもランダムなバイアスとして影響することが分かる.

### 3.4.2 Bridge 推定量の漸近分布と変数選択の一致性

定理 1 より，$\gamma_0 = 1$ の場合でも Bridge 推定量はスパース性を有するため，$\gamma_0 \neq 1$ の場合と同様の議論ができる．$\gamma_0 = 1$ のとき，$\lambda_n = n^{-1/2}\lambda$ であるため，3.3 節と同様に $\hat{\boldsymbol{\beta}}_{\lambda}^{(2)}$ は 1 に収束する確率で

$$\left.\frac{\partial \mathbb{H}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(2)}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\lambda}} = -\sum_{i=1}^{n} g_i^{\prime(2)}(\hat{\boldsymbol{\beta}}_{\lambda}) + n^{1/2}\lambda \boldsymbol{\eta}^{\prime}(\hat{\boldsymbol{\beta}}_{\lambda}^{(2)}) = \boldsymbol{0} \tag{14}$$

を満たす．ただし，$\boldsymbol{\eta}^{\prime}(\hat{\boldsymbol{\beta}}_{\lambda}^{(2)}) = (\gamma \mathrm{sgn}(\hat{\beta}_{\lambda,j})|\hat{\beta}_{\lambda,j}|^{\gamma-1})_{j\in\mathcal{J}^{(2)}}$ である．このとき，$j \in \mathcal{J}^{(2)}$ に対して，

$$\eta^{\prime}(\hat{\beta}_{\lambda,j}) = \gamma \mathrm{sgn}(\hat{\beta}_{\lambda,j})|\hat{\beta}_{\lambda,j}|^{\gamma-1} = \gamma \mathrm{sgn}(\beta_j^*)|\beta_j^*|^{\gamma-1}[1 + \mathrm{o_p}(1)]$$

であるから，(14) は

$$-n^{1/2}\boldsymbol{s}_n^{(2)} + n\boldsymbol{J}_n^{(22)}(\hat{\boldsymbol{\beta}}_{\lambda}^{(2)} - \boldsymbol{\beta}^{*(2)}) + n^{1/2}\lambda\boldsymbol{\eta}^{\prime(2)} + \mathrm{o_p}(n^{1/2}) = \boldsymbol{0}$$

と書くことができ，(R2) より

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{\lambda}^{(2)} - \boldsymbol{\beta}^{*(2)}) = \boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \lambda\boldsymbol{\eta}^{\prime(2)}) + \mathrm{o_p}(1)$$

が得られる.

**定理 5** (Bridge 推定量の漸近分布). このとき，(C1), (C2) のもと，Bridge 推定量 $\hat{\boldsymbol{\beta}}_{\lambda}$ に対して

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{\lambda}^{(2)} - \boldsymbol{\beta}^{*(2)}) = \boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \lambda\boldsymbol{\eta}^{\prime(2)}) + \mathrm{o_p}(1) \tag{15}$$

が成り立つ.

定理 4 とは異なり，Bridge 推定量のスパース性により $\hat{\boldsymbol{\beta}}_{\lambda}^{(1)}$ の漸近分布が $n^{1/2}(\hat{\boldsymbol{\beta}}_{\lambda}^{(2)} - \boldsymbol{\beta}^{*(2)})$ の漸近分布に影響を与えることはないが，代わりに非確率的なバイアスが生じていることが分かる.

また，3.3 節にある，Bridge 推定量の変数選択の一致性に対する議論は $\gamma_0 = 1$ でもそのまま成立するので，Bridge 推定量は変数選択の一致性を持つことが分かる.

## 4 情報量規準

3 章で得られた推定量の漸近的な性質を用いて，調整パラメータを選択するための AIC 型の情報量規準を導出する．具体的には，予測の観点から真の分布と推定された分布の Kullback-Leibler 情

報量の2倍

$$2\tilde{\mathrm{E}}\left[\sum_{i=1}^{n}\tilde{g}_i(\boldsymbol{\beta}^*)\right]-2\tilde{\mathrm{E}}\left[\sum_{i=1}^{n}\tilde{g}_i(\hat{\boldsymbol{\beta}}_\lambda)\right]$$

を漸近的に最小にすることで AIC によるモデル選択を行う．ここで，$(\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2,\ldots,\tilde{\boldsymbol{y}}_n)$ は $(\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots,\boldsymbol{y}_n)$ のコピー，つまり，$(\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2,\ldots,\tilde{\boldsymbol{y}}_n)$ は $(\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots,\boldsymbol{y}_n)$ と独立であり，$(\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots,\boldsymbol{y}_n)$ 同じ分布を持つとする．また，$\tilde{g}_i(\boldsymbol{\beta})$ および $\tilde{\mathrm{E}}$ はそれぞれ $\tilde{\boldsymbol{y}}_i$ に基づく対数尤度関数 $\log f(\tilde{\boldsymbol{y}}_i;\boldsymbol{X}_i\boldsymbol{\beta})$ と $(\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2,\ldots,\tilde{\boldsymbol{y}}_n)$ のみに関する期待値を表すものとする．第1項はモデル選択に依存しない定数項であるため，第2項の漸近不偏推定量として AIC は定義される (Akaike 1973)．いまの場合，第2項の自然な推定量は $-2\sum_{i=1}^{n}g_i(\hat{\boldsymbol{\beta}}_\lambda)$ であるが，これは第2項を過小評価することが知られている．そこで，AIC 型の情報量規準として，バイアス補正した

$$-2\sum_{i=1}^{n}g_i(\hat{\boldsymbol{\beta}}_\lambda)+2\mathrm{E}\left[\sum_{i=1}^{n}g_i(\hat{\boldsymbol{\beta}}_\lambda)-\tilde{\mathrm{E}}\left[\sum_{i=1}^{n}\tilde{g}_i(\hat{\boldsymbol{\beta}}_\lambda)\right]\right] \tag{16}$$

を考える (例えば Konishi & Kitagawa 2008)．(16) の期待値，つまりバイアス項は真の分布に依存しているため，一般には陽に評価することは困難である．したがって，ここではオリジナルの AIC と同じ方法により，漸近的にバイアス項を評価する．

(16) のバイアス項は

$$\sum_{i=1}^{n}\{g_i(\hat{\boldsymbol{\beta}}_\lambda)-g_i(\boldsymbol{\beta}^*)\}-\sum_{i=1}^{n}\{\tilde{g}_i(\hat{\boldsymbol{\beta}}_\lambda)-\tilde{g}_i(\boldsymbol{\beta}^*)\} \tag{17}$$

の期待値として書き換えることができるため，(17) の分布収束先 $z^{\mathrm{limit}}$ の期待値 $\mathrm{E}[z^{\mathrm{limit}}]$ を評価することで AIC を導出する．また，この $\mathrm{E}[z^{\mathrm{limit}}]$ を漸近バイアスと呼ぶ．

## 4.1 $\gamma_0\neq 1$ の場合

Taylor の定理より，(17) の第1項は

$$(\hat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*)^\top\sum_{i=1}^{n}g_i'(\boldsymbol{\beta}^*)+(\hat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*)^\top\sum_{i=1}^{n}g_i''(\boldsymbol{\beta}^\dagger)(\hat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*)/2 \tag{18}$$

と書き換えることができる．ここで，$\boldsymbol{\beta}^\dagger$ は $\hat{\boldsymbol{\beta}}_\lambda$ と $\boldsymbol{\beta}^*$ の間にあるベクトルである．補題2と (R2) より $-n^{-1}\sum_{i=1}^{n}g_i''(\boldsymbol{\beta}^\dagger)$ は $\boldsymbol{J}$ に確率収束する．定理1より十分大きな $n$ に対して $\hat{\boldsymbol{\beta}}_\lambda^{(1)}=\mathbf{0}$ であり，定理2より $n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)}-\boldsymbol{\beta}^{*(2)})=\boldsymbol{J}_n^{(22)-1}\boldsymbol{s}_n^{(2)}+\mathrm{o_p}(1)$ なので，

$$\boldsymbol{s}_n^{(2)\top}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)}-\boldsymbol{s}_n^{(2)\top}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)}/2+\mathrm{o_p}(1)$$

が成り立ち，(R3) よりこれは

$$\boldsymbol{s}^{(2)\top}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)}-\boldsymbol{s}^{(2)\top}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)}/2$$

に分布収束する．同様に，(17) の第2項は

$$(\hat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*)^\top\sum_{i=1}^{n}\tilde{g}_i'(\boldsymbol{\beta}^*)+(\hat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*)^\top\sum_{i=1}^{n}\tilde{g}_i''(\boldsymbol{\beta}^\ddagger)(\hat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*)/2 \tag{19}$$

とかける．ここで，$\boldsymbol{\beta}^{\ddagger}$ は $\hat{\boldsymbol{\beta}}_\lambda$ と $\boldsymbol{\beta}^*$ の間にあるベクトルである．そして，これは

$$\tilde{\boldsymbol{s}}_n^{(2)\top}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)} - \boldsymbol{s}_n^{(2)\top}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)}/2 + \mathrm{o_p}(1)$$

となり，

$$\tilde{\boldsymbol{s}}^{(2)\top}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)} - \boldsymbol{s}^{(2)\top}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)}/2$$

に分布収束する．ただし，$\tilde{\boldsymbol{s}}_n^{(2)}$ と $\tilde{\boldsymbol{s}}^{(2)}$ はそれぞれ $\boldsymbol{s}_n^{(2)}$ と $\boldsymbol{s}^{(2)}$ のコピーである．したがって，

$$z^{\mathrm{limit}} = \boldsymbol{s}^{(2)\top}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)} - \tilde{\boldsymbol{s}}^{(2)\top}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)}$$

が得られる．$\boldsymbol{s}^{(2)}$ と $\tilde{\boldsymbol{s}}^{(2)}$ が独立に平均 0，分散共分散行列 $\boldsymbol{J}^{(22)}$ の正規分布に従うことから，次の定理が得られる．

**定理 6** (漸近バイアス: $\gamma \neq 1$)．定理 2 と同じ条件を仮定する．このとき，$\ell_1$-型正則化推定量と Bridge 推定量に対して，(17) の漸近バイアスは $\mathrm{E}[z^{\mathrm{limit}}] = |\mathcal{J}^{(2)}|$ で与えられる．

定理 6 の漸近バイアスはパラメータの真値を含むため，推定量で置き換える必要がある．漸近バイアスの自然な推定量は $|\hat{\mathcal{J}}^{(2)}| = |\{j;\ \hat{\beta}_{\lambda,j} \neq 0\}|$ であるが，これが一致推定量となっていることを確認しておく．$\gamma_0 \neq 1$ なので，いまの場合，推定量は変数選択の一致性を有する．したがって，

$$\mathrm{P}(|\hat{\mathcal{J}}^{(2)}| = |\mathcal{J}^{(2)}|) \geq \mathrm{P}(\hat{\mathcal{J}}^{(2)} = \mathcal{J}^{(2)}) \to 1$$

なので，$|\hat{\mathcal{J}}^{(2)}|$ は漸近バイアス $\mathrm{E}[z^{\mathrm{limit}}] = |\mathcal{J}^{(2)}|$ の一致推定量である．

以上より，AIC 型の情報量規準を定義できる：

$$\mathrm{AIC} = -2\sum_{i=1}^{n} g_i(\hat{\boldsymbol{\beta}}_\lambda) + 2|\hat{\mathcal{J}}^{(2)}|$$

そこで，この基準が最小にする $\lambda$ を調整パラメータとして選択することでモデル選択を行うことができる．

## 4.2 $\gamma_0 = 1$ の場合

$\ell_1$-型正則化推定量に対して，Taylor の定理と定理 4 より (17) の第 1 項は

$$\hat{\boldsymbol{u}}_n^{(1)\top}\boldsymbol{s}_n^{(1|2)} + \boldsymbol{s}_n^{(2)\top}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)} - \hat{\boldsymbol{u}}_n^{(1)\top}\boldsymbol{J}^{(1|2)}\hat{\boldsymbol{u}}_n^{(1)}/2 - \boldsymbol{s}_n^{(2)\top}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)}/2 + \mathrm{o_p}(1) \tag{20}$$

となる．ただし，$\boldsymbol{u}_n^{(1)}$ は (10) で定義される確率ベクトル，つまり

$$\hat{\boldsymbol{u}}_n^{(1)} = \underset{\boldsymbol{u}^{(1)}}{\mathrm{argmin}}\{\boldsymbol{u}^{(1)\top}\boldsymbol{J}^{(1|2)}\boldsymbol{u}^{(1)}/2 - \boldsymbol{u}^{(1)\top}\boldsymbol{s}_n^{(1|2)} + \lambda\|\boldsymbol{u}^{(1)}\|_1\} \tag{21}$$

であり，$\boldsymbol{s}_n^{(1|2)} = \boldsymbol{s}_n^{(1)} - \boldsymbol{J}^{(12)}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)}$，$\boldsymbol{J}^{(1|2)} = \boldsymbol{J}^{(11)} - \boldsymbol{J}^{(12)}\boldsymbol{J}^{(22)-1}\boldsymbol{J}^{(21)}$ である．いま，(R3) より $\boldsymbol{s}_n^{(1|2)} \xrightarrow{\mathrm{d}} \boldsymbol{s}^{(1|2)}$，$\boldsymbol{s}_n^{(2)} \xrightarrow{\mathrm{d}} \boldsymbol{s}^{(2)}$ である．また，(10) の最適化問題の目的関数は凸なので convexity lemma (Hjort & Pollard 1993) より，

$$\hat{\boldsymbol{u}}_n^{(1)} \xrightarrow{\mathrm{d}} \hat{\boldsymbol{u}}^{(1)} = \underset{\boldsymbol{u}^{(1)}}{\mathrm{argmin}}\{\boldsymbol{u}^{(1)\top}\boldsymbol{J}^{(1|2)}\boldsymbol{u}^{(1)}/2 - \boldsymbol{u}^{(1)\top}\boldsymbol{s}^{(1|2)} + \lambda\|\boldsymbol{u}^{(1)}\|_1\}$$

が成り立つ．よって，(21) は

$$\hat{\boldsymbol{u}}^{(1)\top}\boldsymbol{s}^{(1|2)}+\boldsymbol{s}^{(2)\top}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)}-\hat{\boldsymbol{u}}^{(1)\top}\boldsymbol{J}^{(1|2)}\hat{\boldsymbol{u}}^{(1)}/2-\boldsymbol{s}^{(2)\top}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)}/2+\mathrm{o_p}(1)$$

に分布収束する．同様に，(17) の第 2 項は

$$\hat{\boldsymbol{u}}^{(1)\top}\tilde{\boldsymbol{s}}^{(1|2)}+\boldsymbol{s}^{(2)\top}\boldsymbol{J}^{(22)-1}\tilde{\boldsymbol{s}}^{(2)}-\hat{\boldsymbol{u}}^{(1)\top}\boldsymbol{J}^{(1|2)}\hat{\boldsymbol{u}}^{(1)}/2-\boldsymbol{s}^{(2)\top}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)}/2$$

に分布収束し，したがって

$$z^{\mathrm{limit}}=\hat{\boldsymbol{u}}^{(1)\top}\boldsymbol{s}^{(1|2)}+\boldsymbol{s}^{(2)\top}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)}-\hat{\boldsymbol{u}}^{(1)\top}\tilde{\boldsymbol{s}}^{(1|2)}-\boldsymbol{s}^{(2)\top}\boldsymbol{J}^{(22)-1}\tilde{\boldsymbol{s}}^{(2)}$$

が得られる．ただし，$\tilde{\boldsymbol{s}}^{(1|2)}$ と $\tilde{\boldsymbol{s}}^{(2)}$ はそれぞれ $\boldsymbol{s}^{(1|2)}$ と $\boldsymbol{s}^{(2)}$ のコピーである．よって，$\tilde{\boldsymbol{s}}$ と $\boldsymbol{s}$ が独立に $\mathrm{N}(\boldsymbol{0},\boldsymbol{J})$ に従うことから，

$$\mathrm{E}[z^{\mathrm{limit}}]=\mathrm{E}[\hat{\boldsymbol{u}}^{(1)\top}\boldsymbol{s}^{(1|2)}]+|\mathcal{J}^{(2)}|$$

となる．

Bridge 推定量に対しては $\gamma_0\neq 1$ の場合と同様に $\mathrm{E}[z^{\mathrm{limit}}]=|\mathcal{J}^{(2)}|$ となる．

**定理 7** (漸近バイアス: $\gamma=1$)．定理 4 および定理 5 と同じ条件を仮定する．このとき，$\ell_1$-型正則化推定量と Bridge 推定量に対して，(17) の漸近バイアスはそれぞれ

$$\mathrm{E}[z^{\mathrm{limit}}]=\begin{cases}|\mathcal{J}^{(2)}|+K, & \ell_1\text{-型正則化推定量}\\ |\mathcal{J}^{(2)}|, & \text{Bridge 推定量}\end{cases}$$

となる．ただし，$K=\mathrm{E}[\hat{\boldsymbol{u}}^{(1)\top}\boldsymbol{s}^{(1|2)}]$ である．

定理 7 より，$\ell_1$-型正則化推定量に対しては期待値の評価が必要となることが分かる．これは，$\ell_1$-型正則化推定量がスパース性を持たないためであると考えられる．そこで，$K$ を $\mathrm{N}(\boldsymbol{0},\boldsymbol{J})$ からの標本による経験平均 $\hat{K}$ で，$|\mathcal{J}^{(2)}|$ を $|\hat{\mathcal{J}}^{(2)}|$ で置き換えることで，AIC 型の情報量規準を定義する．

$$\mathrm{AIC}=\begin{cases}-2\sum_{i=1}^{n}g_i(\hat{\boldsymbol{\beta}}_\lambda)+2|\hat{\mathcal{J}}^{(2)}|+2\hat{K}, & \ell_1\text{-型正則化推定量}\\ -2\sum_{i=1}^{n}g_i(\hat{\boldsymbol{\beta}}_\lambda)+2|\hat{\mathcal{J}}^{(2)}|, & \text{Bridge 推定量}\end{cases}$$

$\gamma_0\neq 1$ の場合と同様に，この基準が最小にする $\lambda$ を調整パラメータとして選択することでモデル選択を行うことができる．変数選択の一致性より，Bridge 推定量に対しては $|\hat{\mathcal{J}}^{(2)}|$ が漸近バイアスの一致推定量になっている．一方で，$\ell_1$-型正則化推定量に対しては $2|\hat{\mathcal{J}}^{(2)}|+2\hat{K}$ が漸近バイアスの一致推定量とは言えないことに注意する．

## 5 まとめ

スパース推定などの正則化法において，正則化項の強さを制御する調整パラメータを選択することは非常に重要な問題である．これは，恣意的な調整パラメータの選択が恣意的なモデル選択を実行してしまうことにつながり，誤った結果を招きかねないためである．本稿では，Umezu et al. (2015);

Umezu & Ninomiya (2016) に沿って，SCAD, MCP などを含む $\ell_1$-型正則化法と Bridge 推定法に対して，それらの漸近的な性質と情報量規準 AIC について紹介した．具体的には，推定量の漸近的な性質は，罰則の形状や $\gamma_0$ の値によって異なることをみた．結果として，AIC の漸近バイアス項がパラメータの真値に対するアクティブセットとなるためには，推定量のスパース性が重要であった．このとき，推定量が変数選択の一致性を持てば，推定量のアクティブセットが漸近バイアスの一致推定量となることが分かった．

本稿では，パラメータの次元を固定した元での漸近理論により AIC を導出したが，近年の量子状態の推定などの大規模行列や高次元データ解析に対する課題として，サンプルサイズの増加とともにパラメータの次元も増加する高次元枠組みでも AIC を導出することは重要な課題であると考えられる．

## 参考文献


Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, In *Proc. 2nd International Symposium on Information Theory*, eds. Petrov, B. N. and Csaki, F, Akademiai Kiado, 267–281.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression, *The Annals of Statistics*, **32**, 407–499.

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.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools, *Technometrics*, **35**, 109–135.

Hjort, N. L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes, *arXiv preprint arXiv:1107.3806*.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators, *The Annals of Statistics*, **28**, 1356–1378.

Konishi, S. and Kitagawa, G. (2008). *Information criteria and statistical modeling*, Springer Series in Statistics: Springer, New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, Monographs on Monographs on Statistics and Applied Probability: Chapman & Hall, London.

Ninomiya, Y. and Kawano, S. (2014). AIC for the LASSO in generalized linear models, In *ISM Research Memorandum*, 1187.

Radchenko, P. (2005). Reweighting the lasso, In *2005 Proceedings of the American Statistical Association [CD-ROM]*.

Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**, 461–464.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society: Series B*, **36**, 111–147.

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

Umezu, Y., Shimizu, Y., Masuda, H., and Ninomiya, Y. (2015). AIC for the Non-concave Penalized Likelihood Method, *arXiv preprint arXiv:1509.01688*.

Umezu, Y. and Ninomiya, Y. (2016). On the Consistency of the Bias Correction Term of the AIC for the Non-Concave Penalized Likelihood Method, *arXiv preprint arXiv:1603.07843*.

Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method, *Biometrika*, **94**, 553–568.

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

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