SCAD

Kazeto Fukasawa

平成 27 年 9 月 14 日

第 I 部
準備

1　フィッシャー情報量

$$I = E\Big[\Big(\frac{\partial}{\partial\theta}\log f(X,\theta)\Big)^2|\theta\Big] = -E\Big[\frac{\partial^2}{\partial\theta^2}\log f(X,\theta)|\theta\Big] \tag{1}$$

定理 1

$X_1,\ldots,X_n$ が互いに独立に母数 θ をふくむ密度関数 $f(x,\theta)$ を持つ分布に従うとする。このとき、分布の背即例に関する適当な条件のもとで、θ の任意の不偏推定量 $\hat{\theta}$ について

$$V(\hat{\theta}) \geq \frac{1}{nI} \tag{2}$$

が成り立つ。これをクラメール・ラオ不等式という。

ちなみに、正規分布の場合 $I = 1/\sigma^2$ になる。

2　正規直交行列

X_i に正規直交行列 (orthonotmal) を仮定すると、β の推定値は $\mathbf{z} = X^T\mathbf{y}$ になる。X を正規直交行列とすると、$X^TX = E$ となるので。

第 II 部
SCAD

3 SCAD 推定値

$$p_{\lambda}^{SCAD}(\beta_j) = \begin{cases} \lambda|\beta_j| & \text{if} \quad |\beta_j| \leq \lambda, \\ -\left(\frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}\right) & \text{if} \quad \lambda < |\beta_j| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if} \quad |\beta_j| > a\lambda \end{cases} \tag{3}$$

この一回微分は、$a > 2$,$\beta > 0$ のもとで

$$p'_{\lambda}(\beta) = \lambda\{I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda)\} \tag{4}$$

となり、β の推定値は

$$\hat{\beta}_j^{SCAD} = \begin{cases} (|\hat{\beta}_j| - \lambda)_+ \text{sign}(\hat{\beta}_j) & \text{if} \quad |\hat{\beta}_j| < 2\lambda, \\ \{(a-1)\hat{\beta}_j - \text{sign}(\hat{\beta}_j)a\lambda\}/(a-2) & \text{if} \quad 2\lambda < |\hat{\beta}_j| \leq a\lambda, \\ \hat{\beta}_j & \text{if} \quad |\hat{\beta}_j| > a\lambda \end{cases} \tag{5}$$

1 階微分と推定値
lasso($|\beta| > \lambda$ のケース)

3.0.1 連続であった方がよい理由

(5) を $\hat{\beta}$ を入力して $\hat{\beta}_j^{SCAD}$ を出力する関数 $f(\hat{\beta})$ とみると、hard thresholding では $\lim_{\hat{\beta}\to\lambda+} f(\hat{\beta}) = \hat{\beta}$, $\lim_{\hat{\beta}\to\lambda-} f(\hat{\beta}) = 0$ というように、同じ値への関数の極限値が左右で変わってしまう。つまり $\hat{\beta} = \lambda$ の点では、関数の値が定まらない。そうなると推定量として、$\hat{\beta} = \lambda$ だった時にこまるので、連続の方が良い。

3.1 SCAD 推定値の仮定

仮定 1 観測データは独立である。

仮定 2 対数尤度関数を $l(\beta)$ とする。このとき、対数尤度関数の 1 階微分は以下を満たす。

$$E\Big[\frac{\partial l(\beta)}{\partial \beta_j}\Big] = 0, \quad j = 1, \ldots, p$$

仮定 3 対数尤度関数 $l(\beta)$ の 2 階微分は以下を満たす。

$$E\Big[-\frac{\partial^2 l(\beta)}{\partial\beta_j\partial\beta_k}\Big] = E\Big[\frac{\partial l(\beta)}{\partial\beta_j}\frac{\partial l(\beta)}{\partial\beta_k}\Big], \quad j,k = 1,\dots,p$$

仮定 4 フィッシャー情報行列

$$I(\beta) = E\Big[\frac{\partial l(\beta)}{\partial\beta}\frac{\partial l(\beta)}{\partial\beta'}\Big]$$

は有限であり、真のパラメータ $\beta = \beta_0$ において $I(\beta_0)$ は正定値行列である。

仮定 5 パラメータ空間内 $\mathbb{R}^p$ の部分集合 ω は真のパラメータ $\beta = \beta_0$ を含み、対数尤度関数の 3 階微分の絶対値

$$\Big|\frac{\partial^3 l(\beta)}{\partial\beta_j\partial\beta_k\partial\beta_l}\Big|$$

は全ての $\beta \in \omega$ について有限である。

仮定 6 正則化パラメータは以下の漸近的性質を持つ。

$$\lambda \to 0, \quad \sqrt{n}\lambda \to \infty$$

3.2 Oracle Procedure

A, A^* を以下のように定義する。

$$A = \{j; \beta_{0j} \neq 0\}, \quad A^* = \{j; \hat{\beta}_j \neq 0\} \tag{6}$$

つまり A は目的変数と真に関連する集合であり、A^* は目的変数と関連すると推定された集合である。
Oracle procedure は以下の性質である。

$$\lim_{n\to\infty} P(A^* = A) = 1, \tag{7}$$

$$\sqrt{n}(\beta_{01} - \hat{\beta}_1) \to N(\mathbf{0}, \Sigma^*) \tag{8}$$

ここで、β_{01} は目的変数と真に関連する説明変数の集合 A についての回帰係数ベクトル、$\hat{\beta}_1$ は対応する推定された回帰係数ベクトルである。
つまり、(7) はスパース性、(8) は漸近正規性を表す。
上の仮定のもとで、SCAD 推定量は oracle procedure になる。

3.3　一致性

一致性を示すには、任意の $\epsilon > 0$ に対してある定数 C が存在し、確率 $1-\epsilon$ 以上で $Q(\beta)$ を局所最大化する β が存在することを示せば、$||\hat{\beta}-\beta_0|| = O_p(\alpha_n)$ が成立する。
(？　局所最大化する β が存在 $\rightarrow$ $||\hat{\beta}-\beta_0|| = O_p(\alpha_n)$ が成立　がよくわからない。関数が連続であれば、$f(\beta)-f(\beta_0)=0$ になれば $\beta-\beta_0=0$ になっていく気もするが、これで示しているのは不等号であって等号ではない。任意の確率で成り立つからいいのか？)
(？　任意の確率で、それが 0 に近いことを言いたいのだとすると、extreem estimater と同じ話に見えるが、だとするとなぜ $1-\epsilon$ の形になっている？)

$$P\Big[\sup_{||\mathbf{u}||=C} Q(\beta_{\mathbf{0}}+\alpha_n\mathbf{u}) < Q(\beta_{\mathbf{0}})\Big] \geq 1-\epsilon, \tag{9}$$

$$\alpha_n = n^{-\frac{1}{2}} + d_n, \tag{10}$$

$$d_n = \max\{p'_{\lambda_n}(\beta_{0k}) \;\;;\;\; \beta_{0k} \neq 0\}, \tag{11}$$

$Q(\beta)$ は SCAD 罰則付き対数尤度関数で、以下のように定義される。

$$Q(\beta) = L(\beta) - n\sum_{j=1}^{p} p(|\beta_j|) = \sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i,\beta) - n\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|), \tag{12}$$

(？　なぜ罰則項に n がある？あといつのまにか λ_n になってる。n/n をかけていて、n が外にあり、1/n が関数 p の中に入った？)

P の中に着目して、$Q(\beta_{\mathbf{0}}+\alpha_n\mathbf{u}) - Q(\beta_{\mathbf{0}})$ が常に 0 以下となればよいので、

$$\begin{aligned} D_n(\mathbf{u}) &= Q(\beta_{\mathbf{0}}+\alpha_n\mathbf{u}) - Q(\beta_{\mathbf{0}}) \\ &\leq L(\beta_0+\alpha_n\mathbf{u}) - L(\beta_0) - n\sum_{j=1}^{s}\Big[p_{\lambda_n}(|\beta_{j0}+\alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)\Big] \end{aligned} \tag{13}$$

ここで、s は β_{10} の数である。$L(\beta_0+\alpha_n\mathbf{u}), p_{\lambda_n}(|\beta_{j0}+\alpha_n u_j|)$ をそれぞれ β_0 の周りでテイラー展開して、

$$\begin{aligned} D_n(\mathbf{u}) \leq\; &\alpha_n L'(\beta_0)^T\mathbf{u} \\ &-\frac{1}{2}\mathbf{u}^T I(\beta_0)\mathbf{u} n\alpha^2\{1+o_p(1)\} \\ &-\sum_{j=1}^{s}\Big[n\alpha_n p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})u_j + n\alpha_n^2 p''_{\lambda_n}(|\beta_{j0}|)u_j^2\{1+o(1)\}\Big] \end{aligned}$$

第 1 項目と第 2 項目は $O_p(n\alpha_n^2)$ であり、定数 C を十分大きく取ると第 1 項目と第 2 項目の和は負になり、第 3 項目は第 2 項目よりも小さくなるので、右辺は負になる。
（order については hancenP125）
（？　$o_p(1)$ の部分はテイラー展開の 3 次の項だとは思うが、2 次の項でこうやってくくれるのか？3 次の項を行列表記するとどうなる？行列表記できない気がする）

第 3 項目は、

$$\sqrt{s}n\alpha_n d_n||\mathbf{u}|| + n\alpha_n^2 \max\{|p''_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\}||\mathbf{u}||^2 \tag{14}$$

で抑えられるため、第 2 項目が dominate する。
（？　dominateっていまいちわからない。収束のオーダーの違いから、ある n_0 以上の n をとると、それだけ考えればいいということ？）
3 項目を詳しくみると、まず

$$\sum_{j=1}^{s} n\alpha_n p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})u_j \tag{15}$$

があり、p'_{λ_n} の max をとると、d_n になり、max をとっているので常に正になり $sign$ の部分もいらなくなる。
すると j に依存する部分が u_j だけになり、$\sum_{j=1}^{s} u_j$ を $||\mathbf{u}||^2 = \sum_{j=1}^{s} u_j^2$ で置き換える。
そうすると $n\alpha_n d_n||\mathbf{u}||^2$ になるが、$\sqrt{s}$ はどこから出てきた？あと $||\mathbf{u}||^2$ になってる。むりやり出すとしたら u_j を $||\mathbf{u}||^2$ で置き換えて、$s||\mathbf{u}||^2$ になり、これのルートをとると $\sqrt{s}||\mathbf{u}||$ になって、$\sqrt{s}n\alpha_n d_n||\mathbf{u}||$ に一致するが。
次に

$$n\alpha_n^2 p''_{\lambda_n}(|\beta_{j0}|)u_j^2\{1 + o(1)\} \tag{16}$$

は、$p''_{\lambda_n}(|\beta_{j0}|)$ の max をとって、$o(1)$ 項を無視し、$\sum_{j=1}^{s} u_j^2 = ||\mathbf{u}||^2$ と表記を変えると一致する。
$||\hat{\beta} - \beta_0|| = O_p(\alpha_n)$ については、

$$\hat{\beta} = \beta_0 + \alpha_n u \tag{17}$$

と定義すると、

$$\begin{aligned}
||\hat{\beta} - \beta_0|| &= ||\beta_0 + \alpha_n u - \beta_0|| \\
&= ||\alpha_n u|| = (\alpha_n u)^T(\alpha_n u) = \alpha_n^2 u^T u
\end{aligned}$$

従ってこれを α_n で割ると、$\alpha_n u^T u$ になる。$\alpha_n = n^{-\frac{1}{2}} + d_n$ であり、$n \to \infty$ で 0 になる。

3.4 スパース性

新たに以下を仮定する。

$$\liminf_{n\to\infty}\liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0 \tag{18}$$

ある小さい $\epsilon_n = C/\sqrt{n}$、そして $j = s+1,\dots d$ について、

$$\frac{\partial Q(\beta)}{\partial \beta_j} < 0 \quad \text{for} \quad 0 < \beta_j < \epsilon_n \tag{19}$$

$$> 0 \quad \text{for} \quad -\epsilon_n < \beta_j < 0 \tag{20}$$

を示せばよい。
(？　最終的に $P(\hat{\beta_{20}} = \mathbf{0}) \to 1$ が示されたとあるが、これが成り立つことを示すと、ϵ_n の定義より、$n \to \infty$ で ϵ_n は 0 になり、$0 < \beta_j < 0$ となって、$\beta_j = 0$ が証明できるということだろうか？だとすると、$n \to \infty$ で $\beta_j = 0$ と、$P(\hat{\beta_{20}} = \mathbf{0}) \to 1$ の違いはなんだ？）

$$\frac{\partial Q(\beta)}{\partial \beta_j} = \frac{\partial L(\beta)}{\partial \beta_j} - np'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j) \tag{21}$$

第 1 項目を β_0 の周りでテイラー展開して、

$$\begin{aligned}\frac{\partial Q(\beta)}{\partial \beta_j} &= \frac{\partial L(\beta)}{\partial \beta_j} - np'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j)\\ &= \frac{\partial L(\beta_0)}{\partial \beta_j} + \sum_{l=1}^{d}\frac{\partial^2 L(\beta_0)}{\partial \beta_j \partial \beta_l}(\beta_l - \beta_{l0}) + \sum_{l=1}^{d}\sum_{k=1}^{d}\frac{\partial^3}{\partial \beta_j \partial \beta_l \partial \beta_k}(\beta_l - \beta_{l0})(\beta_k - \beta_{k0}) - np'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j)\end{aligned}$$

これを $n\lambda_n$ でくくると、

$$\begin{aligned}\frac{\partial Q(\beta)}{\partial \beta_j} = n\lambda_n\Big[&\lambda_n^{-1}n^{-1}\frac{\partial L(\beta_0)}{\partial \beta_j}\\ &+ \lambda_n^{-1}n^{-1}\sum_{l=1}^{d}\frac{\partial^2 L(\beta_0)}{\partial \beta_j \partial \beta_l}(\beta_l - \beta_{l0})\\ &+ \lambda_n^{-1}n^{-1}\sum_{l=1}^{d}\sum_{k=1}^{d}\frac{\partial^3}{\partial \beta_j \partial \beta_l \partial \beta_k}(\beta_l - \beta_{l0})(\beta_k - \beta_{k0})\\ &- \lambda_n^{-1}p'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j)\Big]\end{aligned}$$

標準的な議論より、

$$n^{-1}\frac{\partial L(\beta_0)}{\partial \beta_j} = O_p(n^{-\frac{1}{2}}) \tag{22}$$

また、

$$\frac{1}{n}\frac{\partial^2 L(\beta_0)}{\partial\beta_j\partial\beta_l} = E\Big\{\frac{L(\beta_0)}{\partial\beta_j\partial\beta_l}\Big\} + o_p(1) \tag{23}$$

より

$$\frac{\partial L(\beta_0)}{\partial\beta_j} = n\lambda_n\Big\{-\lambda_n^{-1}p'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j) + O_p(n^{-\frac{1}{2}}/\lambda_n)\Big\} \tag{24}$$

ここで、仮定より $\liminf_{n\to\infty}\liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0$ なので、1 階微分の符号は β_j によって決定される。

3.5 漸近正規性

局所解は β_1 の関数としてみることができ、以下を満たす。

$$\frac{\partial Q(\beta)}{\beta_j}\Big|_{\beta=\binom{\beta_1}{0}} = 0 \quad \text{for} \quad j = 1, \ldots, s \tag{25}$$

$$\frac{\partial Q(\beta)}{\beta_j}\Big|_{\beta=\binom{\beta_1}{0}} = \frac{\partial L(\beta)}{\beta_j}\Big|_{\beta=\binom{\beta_1}{0}} - np'_{\lambda_n}(|\hat{\beta}_j|)\mathrm{sgn}(\hat{\beta}_j) \tag{26}$$

$$= \frac{\partial L(\beta_0)}{\partial\beta_j} + \sum_{l=1}^{s}\Big\{\frac{\partial^2 L(\beta_0^2)}{\partial\beta_j\partial\beta_l} + o_p(1)\Big\}(\hat{\beta}_l - \beta_{l0}) - n\Big(p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0}) + \{p''_{\lambda_n}(|\beta_{j0}|) + o_p(1)\}(\hat{\beta}_j - \beta_{j0})\Big) \tag{27}$$

これは Stlutsky's Theorem と中心極限定理に従い、

$$\sqrt{n}(I_1(\beta_{\mathbf{10}}) + \Sigma)\{\hat{\beta_{\mathbf{10}}} - \beta_{\mathbf{10}} + (I_1(\beta_{\mathbf{10}}) + \Sigma)^{-1}\mathbf{b} \to N\{\mathbf{0}, I_1(\beta_{\mathbf{10}})\} \tag{28}$$

となる。

(27) を 0 と置いたものの、$o_p(1)$ の項を無視する。

$$\frac{\partial L(\beta_0)}{\partial \beta_j}+\sum_{l=1}^{s}\frac{\partial^2 L(\beta_0)}{\partial \beta_j \partial \beta_l}(\hat{\beta}_l-\beta_{l0})-np'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})-np''_{\lambda_n}(|\beta_{j0}|)(\hat{\beta}_j-\beta_{j0})=0$$

$$\frac{\partial L(\beta_0)}{\partial \beta_j}=-\sum_{l=1}^{s}\frac{\partial^2 L(\beta_0)}{\partial \beta_j \partial \beta_l}(\hat{\beta}_l-\beta_{l0})+np'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})+np''_{\lambda_n}(|\beta_{j0}|)(\hat{\beta}_j-\beta_{j0})$$

$$\frac{\partial L(\beta_0)}{\partial \beta_j}=n\Big\{-\frac{1}{n}\sum_{l=1}^{s}\frac{\partial^2 L(\beta_0)}{\partial \beta_j \partial \beta_l}(\hat{\beta}_l-\beta_{l0})+p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})+p''_{\lambda_n}(|\beta_{j0}|)(\hat{\beta}_j-\beta_{j0})\Big\}$$

$$\frac{1}{n}\frac{\partial L(\beta_0)}{\partial \beta_j}=-\frac{1}{n}\sum_{l=1}^{s}\frac{\partial^2 L(\beta_0)}{\partial \beta_j \partial \beta_l}(\hat{\beta}_l-\beta_{l0})+p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})+p''_{\lambda_n}(|\beta_{j0}|)(\hat{\beta}_j-\beta_{j0})$$

$$\frac{1}{n}\frac{\partial L(\beta_0)}{\partial \beta_j}=-\frac{1}{n}\sum_{l\neq j}\frac{\partial^2 L(\beta_0)}{\partial \beta_j \partial \beta_l}(\hat{\beta}_l-\beta_{l0})-\frac{1}{n}\frac{\partial^2 L(\beta_0)}{\partial \beta_j^2}(\hat{\beta}_j-\beta_{j0})+p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})+p''_{\lambda_n}(|\beta_{j0}|)(\hat{\beta}_j-\beta_{j0})$$

$$\frac{1}{n}\frac{\partial L(\beta_0)}{\partial \beta_j}=\Big\{p''_{\lambda_n}(|\beta_{j0}|)-\frac{1}{n}\frac{\partial^2 L(\beta_0)}{\partial \beta_j^2}\Big\}(\hat{\beta}_j-\beta_{j0})-\frac{1}{n}\sum_{l\neq j}\frac{\partial^2 L(\beta_0)}{\partial \beta_j \partial \beta_l}(\hat{\beta}_l-\beta_{l0})+p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})$$

$$\frac{1}{n}\frac{\partial L(\beta_0)}{\partial \beta_j}=\Big\{p''_{\lambda_n}(|\beta_{j0}|)-\frac{1}{n}\frac{\partial^2 L(\beta_0)}{\partial \beta_j^2}\Big\}\Big\{(\hat{\beta}_j-\beta_{j0})+\frac{p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})}{p''_{\lambda_n}(|\beta_{j0}|)-\frac{1}{n}\frac{\partial^2 L(\beta_0)}{\partial \beta_j^2}}-\frac{\frac{1}{n}\sum_{l\neq j}\frac{\partial^2 L(\beta_0)}{\partial \beta_j \partial \beta_l}(\hat{\beta}_l-\beta_{l0})}{p''_{\lambda_n}(|\beta_{j0}|)-\frac{1}{n}\frac{\partial^2 L(\beta_0)}{\partial \beta_j^2}}\Big\}$$

右辺の最後の項、

$$-\frac{1}{n}\sum_{l\neq j}\frac{\partial^2 L(\beta_0)}{\partial \beta_j \partial \beta_l}(\hat{\beta}_l-\beta_{l0})$$

については、$\hat{\beta}_l-\beta_{l0}$ が $n^{\frac{1}{2}}$ のオーダーで 0 にいき、$\sum_{l\neq j}$ は $s-1$ に関する有限和であるため、この項は $n\to\infty$ で 0 になる。すると

$$\frac{1}{n}\frac{\partial L(\beta_0)}{\partial \beta_j}=\Big\{p''_{\lambda_n}(|\beta_{j0}|)-\frac{1}{n}\frac{\partial^2 L(\beta_0)}{\partial \beta_j^2}\Big\}\Big\{(\hat{\beta}_j-\beta_{j0})+\frac{p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})}{p''_{\lambda_n}(|\beta_{j0}|)-\frac{1}{n}\frac{\partial^2 L(\beta_0)}{\partial \beta_j^2}}\Big\}$$

となる。これの左辺が中心極限定理より $n\to\infty$ で正規分布に従うため、右辺も正規分布に従う。

$$\Sigma=\mathrm{diag}\{p''_{\lambda_n}(|\beta_{10}|),\ldots,p''_{\lambda_n}(|\beta_{s0}|)\} \tag{29}$$

$$\mathbf{b}=\Big(p'_{\lambda_n}(|\beta_{10}|)\mathrm{sgn}(\beta_{10}),\ldots,p'_{\lambda_n}(|\beta_{s0}|)\mathrm{sgn}(\beta_{s0})\Big) \tag{30}$$

$$I(\beta_1)=-\frac{1}{n}\frac{\partial^2 L(\beta_0)}{\partial \beta_j^2} \tag{31}$$

とおくと、(28) となる。