研究集会「大規模統計モデリングと計算統計 III」

スパース推定に対する AIC 導出のための漸近論について

二宮 嘉行

九州大学 マス・フォア・インダストリ研究所

2016 年 9 月 27 日

目的および流れ

目的:

非凹罰則付き最尤法において，罰則の強弱を決めるチューニングパラメータ $\lambda\ (>0)$ の選択のための情報量規準を開発すること

流れ:

1. 非凹罰則のチューニングパラメータ選択問題の背景
2. SURE 理論に基づく AICc の紹介
3. 二種類の漸近設定のもとでの AIC の導出
4. LASSO を用いたときの数値実験と実データ解析

本内容は川野秀一氏や梅津佑太氏との共同研究の成果を基とする

非凹罰則付き最尤法

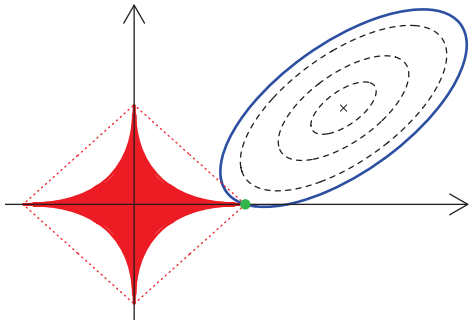
取り扱う正則化法：
パラメータ $\boldsymbol{\beta}$ の非凹罰則項 $\eta_\lambda(\boldsymbol{\beta})$ を推定関数の中に加えるもの

- 推定関数（の $-n$ 倍）としては対数尤度 $l(\boldsymbol{\beta})$ が普通であり，例えば自然リンク関数を用いた一般化線形モデルならば
$$-\sum_{i=1}^{n}\{\boldsymbol{y}_i^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{\beta}-a(\boldsymbol{X}_i\boldsymbol{\beta})+b(\boldsymbol{y}_i)\}=-\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}-A(\boldsymbol{X}\boldsymbol{\beta})+B(\boldsymbol{y})$$
  - 分散既知の正規線形回帰ならば $-\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}+\|\boldsymbol{X}\boldsymbol{\beta}\|^2/2+\|\boldsymbol{y}\|^2/2$

$\eta_\lambda(\boldsymbol{\beta})=\lambda\sum_{j=1}^{p}|\beta_j|^q$ なる Bridge (Frank & Friedman '93),
$\eta_\lambda(\boldsymbol{\beta})=\sum_{j=1}^{p}[9\lambda^2-(3\lambda-|\beta|)^2 1_{\{|\beta|\le r\lambda\}}]/6$ なる MCP (Zhang '10),
$\eta_\lambda(\boldsymbol{\beta})=\sum_{j=1}^{p}[\lambda|\beta|1_{\{|\beta|\le 4.7\lambda\}}-(|\beta|-\lambda)^2 1_{\{\lambda<|\beta|\le 4.7\lambda\}}/7.4+2.85\lambda^2 1_{\{|\beta|>4.7\lambda\}}]$ なる SCAD (Fan & Li '01) を具体的には考える

Bridge 推定量のスパース性

$$\underset{\boldsymbol{\beta}\in\mathcal{B}}{\operatorname{argmin}}\{-l(\boldsymbol{\beta})/n+\lambda\|\boldsymbol{\beta}\|_q^q\}=\underset{\boldsymbol{\beta}\in\mathcal{B};\ \|\boldsymbol{\beta}\|_q^q\le\eta}{\operatorname{argmin}}\{-l(\boldsymbol{\beta})/n\}$$



チューニングパラメータ λ の選択問題 (1)

計算機的手法:

- 通常はクロスバリデーションを用いる（参考: glmnet）
- サブサンプリングで測った Stability に基づき，λ の選択への依存度を減らした方法 (Meinshausen & Bühlmann '10) も注目を浴びているが，やはり計算負荷は高い

情報量規準を用いない話:

- Fan & Li '01 や Huang et al. '08: 「$\lambda=h(p_n,q_n,n)$ ならばオラクル性あり」という h の式や $p_n\ (=|\{j:\beta_j^*\neq 0\}|)$, $q_n\ (=|\{j:\beta_j^*=0\}|)$ の条件を求める話
- 上の場合「$\lambda=c\times h(p_n,q_n,n)$ ならばオラクル性あり」となるわけだが，c を決める話がないので意義が小さい
  - いかようにでもモデル選択の結果を変えられる

チューニングパラメータ λ の選択問題 (2)

厳密な誤差評価をする話:

- Zhang '10 などでは，$\mathrm{P}(\text{推定誤差}\le\delta_\lambda)\ge 1-\epsilon_\lambda$ という形式の評価を導いている
  - δ_λ と ϵ_λ を同時に小さくできないので，最適な λ は決まらない

情報量規準を用いる話:

- Wang et al. '07 '09，Zhang et al. '10，Fan & Tang '13 などでは，$-2l(\hat{\boldsymbol{\beta}}_\lambda)+h(n)(|\hat{\mathcal{J}}|\text{ or }\hat{\mathrm{DF}})$ という形の基準がモデル選択の一致性などをもつことを示している（$\hat{\mathcal{J}}=\{j:\hat{\beta}_{\lambda,j}\neq 0\}$）
  - $h(n)(|\hat{\mathcal{J}}|\text{ or }\hat{\mathrm{DF}})$ の係数が 1 であることの妥当性はない，つまり $c\times h(n)(|\hat{\mathcal{J}}|\text{ or }\hat{\mathrm{DF}})$ としても一致性などは成り立つが c を決める話がない

チューニングパラメータ λ の選択問題 (3)

問題点のない話:

- 分散既知の正規線形回帰の枠組みで，Efron et al. '04 と Zou et al. '07 が LASSO に対する平均二乗誤差（マイナス定数）の不偏推定量，つまり Cp タイプの情報量規準を得ている
- 換言すると次のような AICc（有限補正した AIC）を得ている

$$-2l(\hat{\boldsymbol{\beta}}_\lambda) + 2|\hat{\mathcal{J}}| \qquad (\hat{\mathcal{J}} = \{j : \hat{\beta}_{\lambda,j} \neq 0\})$$

- 自明な形っぽいが実は不思議（Lockhart et al. '13）
  - $|\hat{\mathcal{J}}|$ 個の変数を適応的に選んでいるぶんのバイアス増加は？
  - 縮小推定しているぶんのバイアス減少は？

分散既知の正規線形回帰+非凹罰則付き最尤法

枠組み:

- $\boldsymbol{y} = \boldsymbol{\mu}_0 + \boldsymbol{\epsilon}$ $(\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}))$ という真の構造に対し，$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ とモデルを仮定し，$\hat{\boldsymbol{\mu}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda$ で $\boldsymbol{\mu}_0$ を推定する

$$\hat{\boldsymbol{\beta}}_\lambda = \operatorname*{argmin}_{\boldsymbol{\beta}} \{-\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta} + \|\boldsymbol{X}\boldsymbol{\beta}\|^2/2 + n\eta_\lambda(\boldsymbol{\beta})\}$$

  - $\hat{\mathcal{J}} \equiv \{j \mid \hat{\beta}_{\lambda,j} \neq 0\}$: アクティブセット（$\lambda \nearrow$ で $|\hat{\mathcal{J}}| \searrow$）
  - Λ^{TP}: $|\hat{\mathcal{J}}|$ が変わる λ（transition point）の集合

$\hat{\boldsymbol{\mu}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda$ の表現（$\lambda \notin \Lambda^{\mathrm{TP}}$）:

- $j \in \hat{\mathcal{J}}$ ならば $-\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda + n\eta'_\lambda(\hat{\boldsymbol{\beta}}_\lambda)_j = 0$，つまり $-\boldsymbol{X}_{\hat{\mathcal{J}}}{}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{X}_{\hat{\mathcal{J}}}{}^{\mathrm{T}}\boldsymbol{X}_{\hat{\mathcal{J}}}\hat{\boldsymbol{\beta}}_{\lambda,\hat{\mathcal{J}}} + n\eta'_\lambda(\hat{\boldsymbol{\beta}}_\lambda)_{\hat{\mathcal{J}}} = \boldsymbol{0}$ なので，

$$\hat{\boldsymbol{\mu}} = \boldsymbol{X}_{\hat{\mathcal{J}}}(\boldsymbol{X}_{\hat{\mathcal{J}}}{}^{\mathrm{T}}\boldsymbol{X}_{\hat{\mathcal{J}}})^{-1}\{\boldsymbol{X}_{\hat{\mathcal{J}}}{}^{\mathrm{T}}\boldsymbol{y} - n\eta'_\lambda(\hat{\boldsymbol{\beta}}_\lambda)_{\hat{\mathcal{J}}}\}$$

  - $\boldsymbol{X}_{\hat{\mathcal{J}}} = (\boldsymbol{x}_j)_{j\in\hat{\mathcal{J}}}$, $\hat{\boldsymbol{\beta}}_{\lambda,\hat{\mathcal{J}}} = (\hat{\beta}_{\lambda,j})_{j\in\hat{\mathcal{J}}}$, $\eta'_\lambda(\hat{\boldsymbol{\beta}}_\lambda)_{\hat{\mathcal{J}}} = (\eta'_\lambda(\hat{\boldsymbol{\beta}}_\lambda)_j)_{j\in\hat{\mathcal{J}}}$

一般化 Cp 基準

平均二乗誤差:

$$\mathrm{E}(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|^2) = \mathrm{E}(\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2) + 2\mathrm{E}\{\hat{\boldsymbol{\mu}}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu}_0)\} + \|\boldsymbol{\mu}_0\|^2 - \mathrm{E}(\|\boldsymbol{y}\|^2)$$

- $\mathrm{df}(\hat{\boldsymbol{\mu}}) \equiv \mathrm{E}\{\hat{\boldsymbol{\mu}}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu}_0)\}$: 一般化自由度
- $\|\boldsymbol{\mu}_0\|^2 - \mathrm{E}(\|\boldsymbol{y}\|^2)$ はモデルによらない定数項

一般化 Cp 基準:

$$\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2 + 2\hat{\mathrm{df}}(\hat{\boldsymbol{\mu}})$$

- $\hat{\mathrm{df}}(\hat{\boldsymbol{\mu}})$: 一般化自由度 $\mathrm{df}(\hat{\boldsymbol{\mu}})$ の不偏推定量
- スパース推定では $\mathrm{df}(\hat{\boldsymbol{\mu}})$ は陽に求まらない
- これを最小にするモデルが最適とみなされる

SURE 理論とその非凹罰則付き最尤法への適用

Stein's unbiased risk estimation:

$$\hat{\mathrm{df}}(\hat{\boldsymbol{\mu}}) = \mathrm{tr}\Big(\frac{\partial\hat{\boldsymbol{\mu}}}{\partial\boldsymbol{y}^{\mathrm{T}}}\Big)$$

- $\mathrm{df}(\hat{\boldsymbol{\mu}}) = \mathrm{E}\{\hat{\boldsymbol{\mu}}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{\mu}_0)\} = \sum_{i=1}^n \mathrm{E}(\partial\hat{\mu}_i/\partial y_i) = \mathrm{E}\{\mathrm{tr}(\partial\hat{\boldsymbol{\mu}}/\partial\boldsymbol{y}^{\mathrm{T}})\}$

スパース推定への適用（λ: fix，Vaite et al. '14）:

$$\mathrm{tr}\Big(\frac{\partial\hat{\boldsymbol{\mu}}}{\partial\boldsymbol{y}^{\mathrm{T}}}\Big) = \mathrm{tr}[\boldsymbol{X}_{\hat{\mathcal{J}}}\{\boldsymbol{X}_{\hat{\mathcal{J}}}{}^{\mathrm{T}}\boldsymbol{X}_{\hat{\mathcal{J}}} + n\eta''_\lambda(\hat{\boldsymbol{\beta}}_\lambda)_{\hat{\mathcal{J}}\hat{\mathcal{J}}}\}^{-1}\boldsymbol{X}_{\hat{\mathcal{J}}}{}^{\mathrm{T}}]$$

- $\boldsymbol{X}_{\hat{\mathcal{J}}}{}^{\mathrm{T}} - \boldsymbol{X}_{\hat{\mathcal{J}}}{}^{\mathrm{T}}\boldsymbol{X}_{\hat{\mathcal{J}}}\dfrac{\partial\hat{\boldsymbol{\beta}}_{\lambda,\hat{\mathcal{J}}}}{\partial\boldsymbol{y}^{\mathrm{T}}} - n\eta''_\lambda(\hat{\boldsymbol{\beta}}_\lambda)_{\hat{\mathcal{J}}\hat{\mathcal{J}}}\dfrac{\partial\hat{\boldsymbol{\beta}}_{\lambda,\hat{\mathcal{J}}}}{\partial\boldsymbol{y}^{\mathrm{T}}} = \boldsymbol{0}$
- $\hat{\boldsymbol{\mu}}$ は $\boldsymbol{y}$ に関して連続，またほとんどの $\boldsymbol{y}$ で $\lambda \notin \Lambda^{\mathrm{TP}}$ で $\hat{\mathcal{J}}$ は局所的に不変だから微分可能，つまり SURE を適用可能（？）

非凹罰則付き最尤法における AICc

問題点のない話:

- 分散既知の正規線形回帰の枠組みでは以下が得られる（？）

$$\mathrm{AICc} = -2l(\hat{\boldsymbol{\beta}}_\lambda) + 2\mathrm{tr}[\boldsymbol{X}_{\hat{\mathcal{J}}}{}^{\mathrm{T}}\boldsymbol{X}_{\hat{\mathcal{J}}}\{\boldsymbol{X}_{\hat{\mathcal{J}}}{}^{\mathrm{T}}\boldsymbol{X}_{\hat{\mathcal{J}}} + n\eta''_\lambda(\hat{\boldsymbol{\beta}}_\lambda)_{\hat{\mathcal{J}}\hat{\mathcal{J}}}\}^{-1}]$$

この AICc を一般化線形回帰の枠組みで導くことはできないので，漸近論に頼って AIC を導くことを考える

漸近論のタイプ:

- λ を $n^\xi\lambda$（MCP や SCAD で $-1/2 < \xi < 0$，Bridge で $\gamma/2 - 1 < \xi < -1/2$）に置き換える → オラクル性保証
- λ を定数のままで考える → GIC と同じ設定

分布

自然な指数型分布族:

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

- $\boldsymbol{\theta}$ $(\in \Theta \subset \mathbb{R}^r)$: 自然パラメータ
- Θ: 自然パラメータ空間
  - $\boldsymbol{\theta} \in \Theta^{\mathrm{O}}$ において $\mathrm{E}_{\boldsymbol{\theta}}(\boldsymbol{y}) = a'(\boldsymbol{\theta})$ や $\mathrm{V}_{\boldsymbol{\theta}}(\boldsymbol{y}) = a''(\boldsymbol{\theta})$ など
- $\mathrm{V}_{\boldsymbol{\theta}}(\boldsymbol{y}) = a''(\boldsymbol{\theta})$ は正定値とする
  - $-\log f(\boldsymbol{y};\boldsymbol{\theta})$ は $\boldsymbol{\theta}$ について強凸関数となる

一般化線形モデル

データ $\{(\boldsymbol{y}_i, \boldsymbol{X}_i) \mid 1 \leq i \leq n\}$:

- 目的変数 $\boldsymbol{y}_i$: 独立な r 次元確率ベクトル
- 説明変数 $\boldsymbol{X}_i$: $r \times p$ 定数行列

自然リンク関数:

$$f(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta}) = \exp\{\boldsymbol{y}_i^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{\beta} - a(\boldsymbol{X}_i\boldsymbol{\beta}) + b(\boldsymbol{y}_i)\}$$

- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ $(\in \mathcal{B})$: パラメータベクトル
- $\boldsymbol{\beta}$ の空間 $\mathcal{B}$: 開凸集合
- $\boldsymbol{\beta}^*$ $(\in \mathcal{B})$: $\boldsymbol{\beta}$ の真値

説明変数に関する仮定

説明変数の空間に関する仮定:

(C1) $\{\boldsymbol{X}_i\}$ の空間 $\mathcal{X}$ 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，$\boldsymbol{X}\boldsymbol{\beta} \in \Theta^{\mathrm{O}}$ を満たす

- 漸近論は「有限サンプルの話」の近似のためにあると考え，$\boldsymbol{X}_i$ が発散するような漸近論は扱わない

$\boldsymbol{X}_i$ の漸近挙動に関する仮定:

(C2) 各 $\boldsymbol{\beta}$ に対し，$\sum_{i=1}^n a(\boldsymbol{X}_i\boldsymbol{\beta})/n$，$\sum_{i=1}^n \boldsymbol{X}_i^{\mathrm{T}} a'(\boldsymbol{X}_i\boldsymbol{\beta})/n$，$\sum_{i=1}^n \boldsymbol{X}_i^{\mathrm{T}} a''(\boldsymbol{X}_i\boldsymbol{\beta})\boldsymbol{X}_i/n$ はレート $\mathrm{o}(n^{-1/2})$ で収束し，特に $\sum_{i=1}^n \boldsymbol{X}_i^{\mathrm{T}} a''(\boldsymbol{X}_i\boldsymbol{\beta})\boldsymbol{X}_i/n$ の極限は正定値となる

- $\boldsymbol{X}_i$ をその n 個の実現値の一様分布からの iid サンプルと考えると，レートは $\mathrm{o}(n^{-1/2})$ にならないが，そのような実現値を得られることは稀であり不自然と考える

対数尤度に関する基本的な漸近的性質

$g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}) = \log f(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta})$ に関する性質:

(R1) 各 $\boldsymbol{\beta}$ に対し，$\sum_{i=1}^n \{g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^*) - g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})\}/n \xrightarrow{\mathrm{P}} h(\boldsymbol{\beta})$ を満たすような微分可能な凸関数 $h(\boldsymbol{\beta})$ が存在する

(R2) $\sum_{i=1}^n \mathrm{E}\{g'_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})\}/n - \partial h(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \mathrm{o}(n^{-1/2})$

(R3) $\boldsymbol{J}_n(\boldsymbol{\beta}) \equiv \sum_{i=1}^n \mathrm{E}\{-g''_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})\}/n \to \boldsymbol{J}(\boldsymbol{\beta})$ を満たすような正定値行列 $\boldsymbol{J}(\boldsymbol{\beta})$ が存在する

(R4) $\sum_{i=1}^n [g'_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}) - \mathrm{E}\{g'_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})\}]/\sqrt{n} \xrightarrow{\mathrm{d}} \mathrm{N}(\boldsymbol{0}, \boldsymbol{J}(\boldsymbol{\beta}^*))$

- 凸解析における基本定理（Rockafellar '70）や GEE 推定量の漸近正規性を示した Xie & Yang '03 と同じ方法より

推定量とその極限を調べるための準備

非凹罰則付き最尤推定量（λ: チューニングパラメータ）:

$$\hat{\boldsymbol{\beta}}_\lambda \equiv \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathcal{B}} \Big\{ -\frac{1}{n}\sum_{i=1}^n g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}) + \eta_\lambda(\boldsymbol{\beta}) \Big\}$$

罰則項を $\eta_{\lambda n^\xi}(\boldsymbol{\beta})$ としてもよいが，漸近論は「有限サンプルの話」の近似のためにあるとすると，$\hat{\boldsymbol{\beta}}_\lambda$ が一致性をもつような漸近論は妥当でないと考えられるため，ここでは $\eta_\lambda(\boldsymbol{\beta})$ にしている

- 罰則項を $\lambda n^\xi ||\boldsymbol{\beta}||_q^q$ $(\xi < 0)$ として推定量の漸近的性質を導いた Radchenko '05 の方法を，以降では一般化させて用いる

最後の仮定と推定量の極限

$\mathcal{B}$ に関する仮定:

(C3) $\mathcal{B}$ は唯一の $\operatorname{argmin}_{\boldsymbol{\beta}\in\mathcal{B}}\{h(\boldsymbol{\beta}) + \eta_\lambda(\boldsymbol{\beta})\}$ をもつ

- その argmin を $\boldsymbol{\beta}^{**} = (\beta_1^{**}, \ldots, \beta_p^{**})$ と書くことにする

補題: 非凹罰則付き最尤推定量の極限

$$\hat{\boldsymbol{\beta}}_\lambda \xrightarrow{\mathrm{P}} \boldsymbol{\beta}^{**} \equiv \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathcal{B}}\{h(\boldsymbol{\beta}) + \eta_\lambda(\boldsymbol{\beta})\}$$

- Knight & Fu '00 より

推定量の極限の場合分け

極限の場合分け:

$$\mathcal{J}^{(1)} \equiv \{j : \beta_j^{**} = 0\} \quad \text{と} \quad \mathcal{J}^{(2)} \equiv \{j : \beta_j^{**} \neq 0\}$$

- 記法の準備（$h, k \in \{1, 2\}$）:
  - p 次元ベクトル $\boldsymbol{\beta}$ に対して $\boldsymbol{\beta}^{(h)} = (\beta_j)_{j\in\mathcal{J}^{(h)}}$
  - $p \times p$ 行列 $\boldsymbol{J}$ に対して $\boldsymbol{J}^{(hk)} = (\boldsymbol{J}_{ij})_{i\in\mathcal{J}^{(h)},\ j\in\mathcal{J}^{(k)}}$
- 以降，例えば $\boldsymbol{\beta}$ を $(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})$ と書いたりすることにする

推定量の漸近的性質 (1)

準備:

- Radchenko '05 を参考にして以下のランダム関数を考える

$$\mathbb{G}_n(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**}+\boldsymbol{u}) - \eta_\lambda(\boldsymbol{\beta}^{**}+\boldsymbol{u})$$

- $0 \geq \mathbb{G}_n(\boldsymbol{0}) - \mathbb{G}_n(\hat{\boldsymbol{u}})$ を展開すると

$$0 \geq \Big[\sum_{j\in\mathcal{J}^{(1)}} \{h'(\boldsymbol{\beta}^{**})_j \hat{u}_j + \eta_\lambda(\hat{u}_j)\} + \frac{1}{\sqrt{n}}\boldsymbol{s}_n^{(2)\mathrm{T}}\hat{\boldsymbol{u}}^{(2)} + \frac{1}{2}\hat{\boldsymbol{u}}^{\mathrm{T}}\boldsymbol{J}_n(\boldsymbol{\beta}^{**})\hat{\boldsymbol{u}}\Big]\{1+\mathrm{o_P}(1)\}$$

であり，第一 & 三項目は 1 に近づく確率で正なので $\hat{\boldsymbol{u}} = \mathrm{O_P}(1/\sqrt{n})$ がわかる

- (R2) (R4) より $\boldsymbol{s}_n^{(2)} \xrightarrow{\mathrm{d}} \boldsymbol{s}^{(2)} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**}))$

推定量の漸近的性質 (2)

スパース性 & 漸近正規性:

- $0 \geq \mathbb{G}_n(\boldsymbol{0}, \hat{\boldsymbol{u}}^{(2)}) - \mathbb{G}_n(\hat{\boldsymbol{u}}^{(1)}, \hat{\boldsymbol{u}}^{(2)})$ を展開すると

$$0 \geq \Big[\sum_{j\in\mathcal{J}^{(1)}} \{h'(\boldsymbol{\beta}^{**})_j \hat{u}_j + \eta_\lambda(\hat{u}_j)\} + \frac{1}{2}\boldsymbol{u}^{(1)\mathrm{T}}\boldsymbol{J}_n^{(11)}(\boldsymbol{\beta}^{**})\hat{\boldsymbol{u}}^{(1)} + \hat{\boldsymbol{u}}^{(1)\mathrm{T}}\boldsymbol{J}_n^{(12)}(\boldsymbol{\beta}^{**})\hat{\boldsymbol{u}}^{(2)}\Big]\{1+\mathrm{o_P}(1)\}$$

が得られ，$cn^\epsilon \|\sqrt{n}\hat{\boldsymbol{u}}^{(1)}\|_\gamma^\gamma \leq \mathrm{O_P}(\|\sqrt{n}\hat{\boldsymbol{u}}^{(1)}\|_\gamma^\gamma)$ がいえるので

$$\mathrm{P}(\hat{\boldsymbol{u}}^{(1)} = \boldsymbol{0}) \to 1$$

- $\boldsymbol{0} = \partial\mathbb{G}_n/\partial\boldsymbol{u}^{(2)}(\hat{\boldsymbol{u}})$ を展開して $\hat{\boldsymbol{u}}^{(1)} = \mathrm{o_P}(1/\sqrt{n})$ を使うと

$$0 = -\boldsymbol{s}_n^{(2)} + \{\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**}) + \eta_\lambda''^{(22)}(\boldsymbol{\beta}^{**})\}\sqrt{n}\hat{\boldsymbol{u}}^{(2)} + \mathrm{o_P}(1)$$

が得られ，第二項目の行列を $\boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})$ と書けば

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{**(2)})) \xrightarrow{\mathrm{d}} \boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})\boldsymbol{s}^{(2)}$$

AIC タイプの情報量規準

KL ダイバージェンスの不偏推定量:

$$-2\sum_{i=1}^{n} g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) + 2\mathrm{E}\Big[\sum_{i=1}^{n} g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) - \tilde{\mathrm{E}}\Big\{\sum_{i=1}^{n} g_{\tilde{\boldsymbol{y}}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda)\Big\}\Big]$$

- $(\tilde{\boldsymbol{y}}_1,\ldots,\tilde{\boldsymbol{y}}_n)$ は $(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n)$ のコピー，$\tilde{\mathrm{E}}$ は $(\tilde{\boldsymbol{y}}_1,\ldots,\tilde{\boldsymbol{y}}_n)$ だけについての期待値を表すとする

AIC と同様の漸近評価に基づく基準:

$$-2\sum_{i=1}^{n} g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) + 2\mathrm{E}(z^{\mathrm{limit}})$$

- z^{limit} は以下の弱極限とする

$$\sum_{i=1}^{n}\{g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) - g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**})\} + \sum_{i=1}^{n}\{g_{\tilde{\boldsymbol{y}}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) - g_{\tilde{\boldsymbol{y}}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**})\}$$

漸近バイアスの導出

漸近バイアスの素:

$$z^{\mathrm{limit}} = -\boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})^{-1}\tilde{\boldsymbol{s}}^{(2)} + \boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})^{-1}\boldsymbol{s}^{(2)}$$

- $\tilde{\boldsymbol{s}}^{(2)}$ は $\boldsymbol{s}^{(2)}$ のコピー，つまり $\tilde{\boldsymbol{s}}^{(2)} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{*}))$
- $\hat{\boldsymbol{\beta}}_\lambda$ の極限定理を $\sum_{i=1}^{n}\{g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) - g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**})\} + \sum_{i=1}^{n}\{g_{\tilde{\boldsymbol{y}}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) - g_{\tilde{\boldsymbol{y}}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**})\}$ のテイラー近似に用いて得る
  - $\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \mathrm{o_P}(1/\sqrt{n})$ より，これに関する項は消える
  - 上記の形を得るためには対数尤度に関する性質 (R2)，つまり説明変数の挙動に関する性質 (C2) が必要

定理: 漸近バイアス

$$\mathrm{E}(z^{\mathrm{limit}}) = \mathrm{tr}\{\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{*})\boldsymbol{K}^{(22)}(\boldsymbol{\beta}^{**})^{-1}\}$$

非凹罰則付き最尤法に対する AIC

$\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{*})$ & $\boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})$ の一致推定量:

$$\boldsymbol{J}_n(\hat{\boldsymbol{\beta}}_0)_{\hat{\mathcal{J}}\hat{\mathcal{J}}} \quad \& \quad \boldsymbol{J}_n(\hat{\boldsymbol{\beta}}_\lambda)_{\hat{\mathcal{J}}\hat{\mathcal{J}}} + \eta_\lambda''(\hat{\boldsymbol{\beta}}_\lambda)_{\hat{\mathcal{J}}\hat{\mathcal{J}}}$$

- $\hat{\mathcal{J}} = \{j : \hat{\beta}_{\lambda,j} \neq 0\}$: アクティブセット
- $\boldsymbol{J}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{X}_i^{\mathrm{T}} a''(\boldsymbol{X}_i\boldsymbol{\beta})\boldsymbol{X}_i/n$
- $\hat{\boldsymbol{\beta}}_0$ & $\hat{\boldsymbol{\beta}}_\lambda$: 最尤推定量 & 非凹罰則付き最尤推定量

提案: 非凹罰則付き最尤法に対する情報量規準 AIC

$$-2\sum_{i=1}^{n} g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) + 2\mathrm{tr}[\boldsymbol{J}_n(\hat{\boldsymbol{\beta}}_0)_{\hat{\mathcal{J}}\hat{\mathcal{J}}}\{\boldsymbol{J}_n(\hat{\boldsymbol{\beta}}_\lambda)_{\hat{\mathcal{J}}\hat{\mathcal{J}}} + \eta_\lambda''(\hat{\boldsymbol{\beta}}_\lambda)_{\hat{\mathcal{J}}\hat{\mathcal{J}}}\}^{-1}]$$

正規分布モデル（線形回帰モデル）の場合

$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ $(\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}))$ というモデルのケース:

- $a''(\boldsymbol{X}_i\boldsymbol{\beta}) = \boldsymbol{I}$ より提案の情報量規準は次の AICc となる

$$-2\sum_{i=1}^{n} g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) + 2\mathrm{tr}[\boldsymbol{X}_{\hat{\mathcal{J}}}{}^{\mathrm{T}}\boldsymbol{X}_{\hat{\mathcal{J}}}\{\boldsymbol{X}_{\hat{\mathcal{J}}}{}^{\mathrm{T}}\boldsymbol{X}_{\hat{\mathcal{J}}} + n\eta_\lambda''(\hat{\boldsymbol{\beta}}_\lambda)_{\hat{\mathcal{J}}\hat{\mathcal{J}}}\}^{-1}]$$

実は，オラクル性を保証する漸近設定のもとでも AIC を同様のアプローチで導出でき，その罰則項は $2|\hat{\mathcal{J}}|$ である

- 有限サンプルのときの話との整合性がない
- オラクル性を与える漸近設定は将来の話のためのものであり，有限サンプルのときの話を近似するための漸近設定とは別
- 欲しいのは今あるサンプルに対する真の構造

Logistic LASSO における数値実験 (1)

$\boldsymbol{\beta}^* = (\beta_1^*, \beta_1^*, \beta_2^*, \beta_2^*, 0, \ldots, 0) \in \mathbb{R}^p$, n: サンプルサイズ

$(n, p, \beta_1^*, \beta_2^*)$		KLav	KLsd	FPR	FNR
$(100, 8, 6.0, 0.5)$	CV	0.161	0.010	0.42	0.18
	AIC	0.158	——	0.38	0.20
	$\mathrm{AIC}_{\mathrm{naive}}$	0.166	0.028	0.48	0.16
$(200, 8, 6.0, 0.5)$	CV	0.139	0.003	0.44	0.09
	AIC	0.138	——	0.40	0.10
	$\mathrm{AIC}_{\mathrm{naive}}$	0.140	0.007	0.49	0.09
$(100, 16, 6.0, 0.5)$	CV	0.173	0.017	0.39	0.21
	AIC	0.171	——	0.38	0.22
	$\mathrm{AIC}_{\mathrm{naive}}$	0.194	0.041	0.46	0.19
$(100, 8, 6.5, 1.0)$	CV	0.149	0.009	0.42	0.08
	AIC	0.147	——	0.38	0.10
	$\mathrm{AIC}_{\mathrm{naive}}$	0.150	0.016	0.48	0.08

Logistic LASSO における数値実験 (2)

$\boldsymbol{\beta}^* = (\beta_1^* \times 5, \beta_2^* \times 5, 0, \ldots, 0) \in \mathbb{R}^p$, n: サンプルサイズ

$(n, p, \beta_1^*, \beta_2^*)$		KLav	KLsd	FPR	FNR
$(100, 500, 6.0, 0.5)$	CV	0.263	0.015	0.05	0.43
	AIC	0.254	——	0.05	0.43
	$\mathrm{AIC}_{\mathrm{naive}}$	0.293	0.026	0.01	0.47
$(200, 500, 6.0, 0.5)$	CV	0.190	0.009	0.08	0.32
	AIC	0.187	——	0.06	0.34
	$\mathrm{AIC}_{\mathrm{naive}}$	0.216	0.018	0.02	0.40
$(100, 1000, 6.0, 0.5)$	CV	0.292	0.015	0.06	0.44
	AIC	0.283	——	0.06	0.44
	$\mathrm{AIC}_{\mathrm{naive}}$	0.316	0.027	0.02	0.48
$(100, 500, 6.5, 1.0)$	CV	0.277	0.016	0.05	0.36
	AIC	0.267	——	0.05	0.36
	$\mathrm{AIC}_{\mathrm{naive}}$	0.297	0.026	0.02	0.41

実データ解析（8 種）

「訓練データとして m サンプルとって推定した分布と残りのテストデータから推定した分布との KL 距離を測る」を繰り返す

data	(n, p)	m	model	CV	AIC	$\mathrm{AIC}_{\mathrm{naive}}$
pima	(200,7)	100	logistic	0.494 (0.030)	0.481	0.494 (0.037)
biod.	(1055,41)	100	logistic	0.448 (0.036)	0.444	0.480 (0.045)
colon	(62,2000)	40	logistic	0.530 (0.151)	0.464	0.494 (0.067)
leuk.	(38,3051)	20	logistic	0.373 (0.101)	0.286	0.335 (0.108)
take.	(126,14)	50	Poisson	1.63 (0.03)	1.61	1.62 (0.01)
doct.	(5190,11)	100	Poisson	0.717 (0.028)	0.713	0.719 (0.020)
flow.	(7466,11)	100	GGM	12.7 (0.2)	12.6	12.6 (0.0)
math.	(55,5)	50	GGM	3.40 (0.01)	3.39	3.39 (0.00)

数値は KL 距離の平均（括弧内の数値は KL 距離の差の標準偏差）

結論と課題

1. AIC 元来の定義に基づき，一般化線形モデルにおける非凹罰則付き最尤法に対する情報量規準を導いた
   - 恣意的に決めなければならない値はない
2. AIC 導出のための漸近設定は複数種あるが，漸近論を用いずに得られる AICc の一般形として与えられるものを選択した
   - オラクル性をいうための漸近設定とは異なる
3. 選んだ漸近設定のもとでの AIC が良い性質をもつこと，CV さえも上回る性能をもつことを LASSO の数値実験で示唆した
   - 次元 p が大きいときも許容できる
4. 課題: 次元 p が大きいときの妥当性を理論的に保証すること