

スパースな統計モデルの周辺尤度に対する漸近展開 について

高崎経済大学・経済学部* 宮田 庸一

Yoichi Miyata

Faculty of Economics,

Takasaki City University of Economics

概要

チューニングパラメーターが LASSO などの罰則付き推定量の振る舞いに大きな影響を与えることはよく知られた事実である。階層的なベイズアプローチにおいては、標本が生成されるモデルに対応する周辺尤度の近似を最大にするチューニングパラメーターを選ぶ。本稿においては、いくつかの階層的なベイズアプローチを説明するとともに、L1 罰則に対応する事前分布を用いたときの周辺尤度に対するラプラス近似に関する結果を紹介する。さらに線形モデルの下での周辺尤度に対するラプラス近似の誤差を、シミュレーションにより確認する。

1 はじめに

応答変数 $\mathbf{Y} = (Y_1, \dots, Y_n)'$ は、以下の線形回帰モデル

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon} \quad (1)$$

から生成されているものとする。ただし \mathbf{X} を $n \times p$ 計画行列とし、 $\boldsymbol{\beta}^* = (\beta_1^*, \dots, \beta_p^*)' \in \mathbb{R}^p$ を真のパラメーターのベクトルとし、 $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)'$ において $E(\boldsymbol{\epsilon}) = \mathbf{0}$ 、 $V(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$ とする。また $\boldsymbol{\beta}^*$ の非ゼロの要素の個数は p_0 個あり、 $\gamma_0 = \{i \in \{1, \dots, p\} | \beta_i^* \neq 0\}$ を非ゼロの要素に対応する添え字の集合とする。ここでは、観測データ (\mathbf{X}, \mathbf{Y}) から、 $\boldsymbol{\beta}^*$ および γ_0 を推定することを考える。パラメーターの次元 p が小さい場合には、候補となる 2^p 個のモ

* 〒370-0801 群馬県高崎市上並榎町 1300

デルそれぞれに対して適当な情報量規準を用いて評価することができるが、 p が大きい場合には、計算時間の観点から困難になる。それを回避するための手法の1つとして、以下の LASSO 推定 (Tibshirani 1996) が知られている。

$$\hat{\beta}_\lambda = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \{ \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_1 \} \quad (2)$$

ただし $\beta = (\beta_1, \dots, \beta_p)'$ とし、 $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ を L1 ノルムとする。推定量 $\hat{\beta}_\lambda$ のパフォーマンスはチューニングパラメーター λ に依存しており、 λ をどのように決めるのが興味の対象となる。 λ の推定方法に関しては、交差検証法 (Stone 1974), C_p 規準 (Zou et al. 2007) 等の研究がある。本稿ではベイズアプローチによる λ の推定方法を紹介し、その導出に関わる問題点に対する考察を考える。

2 ベイズ型モデル選択

ここでは λ のベイズ型モデル選択について説明を行う。候補となるモデルを表す 2 値の確率変数 $\gamma_i \in \{0, 1\}$ のベクトルを $\gamma = (\gamma_1, \dots, \gamma_p)'$ で表す。ただし、モデルに i 番目の回帰係数が含まれる場合には $\gamma_i = 1$, 含まれない場合には $\gamma_i = 0$ となるように定める。このとき、線形モデル (1) を推定するために、

$$\mathbf{Y} | \gamma, \beta \sim N(\mathbf{X}(\gamma)\beta(\gamma), \sigma^2 \mathbf{I}_n)$$

であること仮定する。ただし $|\gamma|$ は、ベクトル γ において $\gamma_i = 1$ である要素の個数とし、 $\mathbf{X}(\gamma)$ は、 $\mathbf{X} = (\mathbf{x}^{(1)} \dots \mathbf{x}^{(p)})$ において、 $\gamma_i = 0$ となる i 番目の列を全て取り除いた $n \times |\gamma|$ 行列であるとする。 $\beta(\gamma)$ についても同様に定義する。次に、 γ が与えられたもとの β の条件付き事前分布を定義する。いくつかの候補が考えられるが、Yuan and Lin (2005) においては、以下の事前分布が提案されている。

$$\pi(\beta_i | \gamma_i) = (1 - \gamma_i) \delta_{\beta_i}(0) + \gamma_i g(\beta_i | \lambda, \sigma^2) \quad (3)$$

ただし、 $\delta_{\beta_i}(0)$ は β_i のデルタ関数とし、 $g(\beta_i | \lambda, \sigma^2) = (\lambda/4\sigma^2) \exp\{-(\lambda/2\sigma^2)|\beta_i|\}$ を両側指数分布における p.d.f. とする。またモデル γ に対する事前分布としては

$$P(\gamma) \propto q_1^{|\gamma|} (1 - q_1)^{p-|\gamma|} \det(\mathbf{X}(\gamma)' \mathbf{X}(\gamma))^{q_2} \quad (4)$$

が考えられる。ただし $q_1 \in (0, 1)$, $q_2 \geq 0$ は事前に決めておく定数とする。George and Foster (2000, p.733) では、 $q_2 = 0$ を提案しており、Yuan and Lin (2005) では $q_2 = 1/2$

を提案している.*¹

このとき、 $(\gamma, \beta, \mathbf{Y})$ の結合確率分布は、以下の式で与えられる。

$$\begin{aligned} P(\gamma, \beta, \mathbf{Y}|\sigma^2, \lambda) &= p(\mathbf{Y}|\gamma, \beta, \sigma^2)\pi(\beta|\gamma, \sigma^2, \lambda)P(\gamma) \\ &\propto \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \frac{\sqrt{\det(\mathbf{X}(\gamma)'\mathbf{X}(\gamma))}}{(\sqrt{2\pi\sigma^2})^{|\gamma|}} \\ &\quad \times \exp\left\{-\frac{\|\mathbf{Y} - \mathbf{X}(\gamma)\beta(\gamma)\|^2 + \lambda \sum_{i \in \gamma} |\beta_i|}{2\sigma^2}\right\} (1 - q_1)^p w^{|\gamma|}, \end{aligned}$$

ただし

$$w = \left(\frac{q_1}{1 - q_1} \frac{\lambda}{2\sigma^2} \sqrt{2\pi\sigma^2}\right) \quad (5)$$

とする。ここで経験ベイズ法の観点からは、周辺尤度

$$p(\mathbf{Y}|\sigma^2, \lambda) = \sum_{\gamma} P(\gamma) \underbrace{\int_{\mathbb{R}^{|\gamma|}} P(\beta(\gamma), \mathbf{Y}|\gamma, \sigma^2, \lambda) d\beta(\gamma)}_{p(\mathbf{Y}|\gamma, \sigma^2, \lambda)} \quad (6)$$

を最大にする λ , σ^2 を選べばよいが、説明変数の次元 p が大きくなるにつれ、上記の総和の計算は指数的に増加する。このため、そのような場合には計算を行うことが困難になる。それに代わるハイパーパラメーターの選択法として、条件付き最大尤度規準 (Conditional maximum likelihood criterion, 以後 CML 規準という) を考える。LASSO 推定 $\hat{\beta}_\lambda$ から得られた $\hat{\gamma}_\lambda := \{j \in \{1, \dots, p\} | \hat{\beta}_{\lambda, j} \neq 0\}$ を、 $p(\mathbf{Y}|\gamma, \sigma^2, \lambda)$ の γ に代入して $p(\mathbf{Y}|\hat{\gamma}_\lambda, \sigma^2, \lambda)$ を得て、それに対して以下の近似を行う：

$$p(\mathbf{Y}|\hat{\gamma}_\lambda, \sigma^2, \lambda) = \int_{\mathbb{R}^p} p(\mathbf{Y}|\gamma, \beta, \sigma^2)\pi(\beta|\gamma, \sigma^2, \lambda) d\beta \Big|_{\gamma=\hat{\gamma}_\lambda} \quad (7)$$

$$\begin{aligned} &\approx \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n-|\hat{\gamma}_\lambda|} \left(\frac{\lambda}{4\sigma^2}\right)^{|\hat{\gamma}_\lambda|} \{ \det(\mathbf{X}(\hat{\gamma}_\lambda)'\mathbf{X}(\hat{\gamma}_\lambda)) \}^{-1/2} \\ &\quad \times \exp\left(-\frac{\min_{\beta} (\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \sum_{i=1}^p |\beta_i|)}{2\sigma^2}\right) \end{aligned} \quad (8)$$

(8) 式を σ^2 に関して最大化することで、以下の CML 規準

$$\begin{aligned} \text{CML}(\lambda) &\equiv (n + |\hat{\gamma}_\lambda|) \left[\log \left(\frac{\min_{\beta} (\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \sum_{i=1}^p |\beta_i|)}{n + |\hat{\gamma}_\lambda|} \right) \right] \\ &\quad + \log \{ \det(\mathbf{X}(\hat{\gamma}_\lambda)'\mathbf{X}(\hat{\gamma}_\lambda)) \} - 2|\hat{\gamma}_\lambda| \log(\sqrt{2\pi}\lambda/4) \end{aligned} \quad (9)$$

*¹ $\mathbf{X}(\gamma)$ において多重共線性が成り立つ場合、 $\det(\mathbf{X}(\gamma)'\mathbf{X}(\gamma)) = 0$ となる。つまりそのようなモデルが選ばれる事前確率は 0 になる。

が導出される。これより (9) 式を最小にする λ を選ばばよい。尚、導出の詳細については、Yuan and Lin (2005) を参照されたい。もう一つの考え方としては、(7) 式において、 γ にフルモデルを表す $\mathbf{1}_p = (1, \dots, 1)'$ を代入することである。この場合、 $p(\mathbf{Y}|\mathbf{1}_p, \sigma^2, \lambda) = \int_{\mathbb{R}^p} p(\mathbf{Y}|\mathbf{1}_p, \boldsymbol{\beta}, \sigma^2) \pi(\boldsymbol{\beta}|\mathbf{1}_p, \sigma^2, \lambda) d\boldsymbol{\beta}$,

$$\pi(\boldsymbol{\beta}|\mathbf{1}_p, \sigma^2, \lambda) = \prod_{j=1}^p \left\{ \frac{\lambda}{4\sigma^2} \exp\left(-\frac{\lambda}{2\sigma^2} |\beta_j|\right) \right\} \quad (10)$$

となる。事前分布 (10) は L1 罰則に対応しているが、L2 罰則に対応する事前分布 $N(\mathbf{0}, (\sigma^2/\lambda)\mathbf{I}_p)$ を考え、 $\mathbf{1}_p = (1, \dots, 1)'$ を代入したものについては、Konishi et al. (2004) により研究されている。以後、事前分布 (10) は $\mathbf{1}_p$ を省略して $\pi(\boldsymbol{\beta}|\sigma^2, \lambda)$ とする。上記の 2 つのアプローチにおいては、 γ の選択によらず

$$\mathcal{I}_n(\lambda) = \int_{\mathbb{R}^p} n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n) \pi(\boldsymbol{\beta}|\sigma^2, \lambda) d\boldsymbol{\beta} \quad (11)$$

の形の積分をいかにして展開するかということが問題となる。ただし、 $n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ は期待値 $\mathbf{X}\boldsymbol{\beta}$ 、共分散 $\sigma^2\mathbf{I}_n$ の正規分布における p.d.f. とする。このため今後の説明においては、あるモデルを表す γ の記号は省略する。また本稿で考えるのは、下記の問題である：

- **$p := p_n \rightarrow \infty$ の問題:** (11) 式に対するラプラス近似 $\hat{\mathcal{I}}_n(\lambda)$ が妥当性 (即ち $\mathcal{I}_n(\lambda) = \hat{\mathcal{I}}_n(\lambda)\{1 + o_p(1)\}$) を持つためには、標本のサイズ n およびパラメーターの個数 $p \equiv p_n$ に対してどのような条件が課されるべきか？ この問題は積分 (7) の近似においても現れる。
- **被積分関数の微分可能性:** 事前分布 (10) は、 $\beta_i = 0$ で微分不可能なため、MLE の代わりに LASSO 推定 $\hat{\boldsymbol{\beta}}_\lambda$ を用いたラプラス近似の導出は可能かどうか？ 尚、積分 (7) の近似においては $\hat{\beta}_{\lambda,i} = 0$ となる点が現れないため、この問題は生じない。一方で $p(\mathbf{Y}|\mathbf{1}_p, \sigma^2, \lambda)$ に対してラプラス近似を用いるときには、この問題が生じる。

3 ラプラス近似

前の章で与えた周辺尤度 (11) に対する漸近展開を考える。ここでは、式 (11) における設定を少しだけ一般化して、被説明変数 Y_1, \dots, Y_n は互いに独立に一般化線形モデル (Nelder and Wedderburn 1972)

$$p(y_i|\mathbf{x}_i, \boldsymbol{\beta}) \propto \exp\{y_i \mathbf{x}_i' \boldsymbol{\beta} - b(\mathbf{x}_i \boldsymbol{\beta})\} \quad (12)$$

に従うとする。ただし $b(\theta)$ は滑らかで凸な実関数とし、 $\mathbf{x}_i \in \mathbb{R}^p$ は説明変数を表すベクトルとする。そして $H_n(\boldsymbol{\beta}) = -\sum_{i=1}^n \log p(y_i | \mathbf{x}_i, \boldsymbol{\beta})$ とおき、 $\mathcal{I}_n^*(\lambda) = \int_{\mathbb{R}^{p_n}} \pi(\boldsymbol{\beta} | \sigma^2, \lambda) \exp\{-H_n(\boldsymbol{\beta})\} d\boldsymbol{\beta}$ とする。まず、最尤推定を用いたラプラス近似を紹介する。 $H_n(\boldsymbol{\beta})$, $\pi(\boldsymbol{\beta} | \sigma^2, \lambda)$ には $\boldsymbol{\beta}$ に関する滑らかさの条件が必要になるが、ここでは省略する。

定理 1 (Barber et al. 2016) q は真のモデルの大きさの上限、即ち $|\gamma_0| \leq q$ と満たすある正の数とする。任意の λ に対して、ある定数 $c_{\text{sample}} > 0$, $\nu > 0$, $c_{\text{Laplace}} > 0$ が存在して、 $n \leq c_{\text{sample}} q^3 \max\{\log(p), (\log n)^3\}$ ならば

$$P \left(\left| \frac{\mathcal{I}_n^*(\lambda)}{\tilde{\mathcal{I}}_n^*(\lambda)} - 1 \right| \leq c_{\text{Laplace}} \sqrt{\frac{p^3 (\log n)^3}{n}} \right) \geq 1 - p^{-\nu}.$$

ただし、 $\hat{\boldsymbol{\beta}}_{ML}$ は $\boldsymbol{\beta}$ に対する最尤推定とし、 $\tilde{\mathcal{I}}_n^*(\lambda) = (2\pi)^{p_n/2} \det(D^2 H_n(\hat{\boldsymbol{\beta}}_{ML}))^{-1/2} \exp\{-H_n(\hat{\boldsymbol{\beta}}_{ML})\} \pi(\hat{\boldsymbol{\beta}}_{ML} | \sigma^2, \lambda)$ とする.*2

次に LASSO 推定を用いたラプラス近似を紹介する。 $\hat{\mathbf{A}} := \hat{\mathbf{A}}_n$ は正値定符号の $p_n \times p_n$ 行列の列とし、 $\hat{B}_n := \{\boldsymbol{\beta} \in \mathbb{R}^{p_n} | (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)' \hat{\mathbf{A}} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda) \leq \delta_n^2\}$ を \mathbb{R}^{p_n} における開球とする。また $\|\mathbf{A}\|_{sp}$ を行列 \mathbf{A} のスペクトルノルムとする。 $\lambda_{\min}(\mathbf{A})$, $\lambda_{\max}(\mathbf{A})$ をそれぞれ行列 \mathbf{A} の最小、最大固有値とする。また記号を簡略化するために $DH_n := (\partial/\partial \boldsymbol{\beta}) H_n(\hat{\boldsymbol{\beta}}_\lambda)$, $D^2 H_n := (\partial^2/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}') H_n(\hat{\boldsymbol{\beta}}_\lambda)$, $\pi_n^\lambda := \pi(\hat{\boldsymbol{\beta}}_\lambda | \sigma^2, \lambda)$ とする。定理 1 と同様にして、 $H_n(\boldsymbol{\beta})$, $\pi(\boldsymbol{\beta} | \sigma^2, \lambda)$ には $\boldsymbol{\beta}$ に関する滑らかさの条件が必要になるが、ここでは省略する。ただし下記の 2 つの条件は、 p_n のオーダーに関わる重要な条件のため、書いておく。 $Q_n(\boldsymbol{\beta}) = H_n - H_n(\boldsymbol{\beta}) + \log\{\pi(\boldsymbol{\beta} | \sigma^2, \lambda)/\pi_n^\lambda\}$ とする。このとき任意の $\epsilon > 0$ に対して、実数の列 $\{a_n\}$, 定数 $M > 0$ が存在して、

$$(i) \quad P \left[\int_{\mathbb{R}^{p_n} \setminus \hat{B}_n} \exp\{n^{-1} Q_n(\boldsymbol{\beta})\} d\boldsymbol{\beta} \leq \exp(a_n) M \right] \rightarrow 1,$$

$$(ii) \quad P \left[\sup_{\boldsymbol{\beta} \in \mathbb{R}^{p_n} \setminus \hat{B}_n} Q_n(\boldsymbol{\beta}) - \frac{n}{2(n-1)} \log \det(D^2 H_n) + \frac{n}{n-1} a_n < -\frac{1}{\epsilon} \right] \rightarrow 1, \quad (13)$$

($n \rightarrow \infty$). この 2 つの条件は、 $\int_{\mathbb{R}^{p_n} \setminus \hat{B}_n} \pi(\boldsymbol{\beta} | \sigma^2, \lambda) \exp\{-H_n(\boldsymbol{\beta})\} d\boldsymbol{\beta}$ があるレートで 0 に収束することを示すために必要となる。

*2 ただし Barber et al. (2016) においては、このラプラス近似に補正項を加えた情報量規準を提案している。

定理 2 (Miyata 2017b) $\lambda > 0$ を任意の数とする. 条件 (i), (ii) および, いくつかの条件の下で

$$\mathcal{I}_n^*(\lambda) = (2\pi)^{p_n/2} \det(D^2 H_n)^{-1/2} \exp\{-H_n\} \pi_n^\lambda \{C_n + o_p(1)\}, \quad (14)$$

ただし $C_n = \int_{\hat{B}_n} \exp\{-DH'_n(\beta - \hat{\beta}_\lambda)\} n(\beta|\hat{\beta}_\lambda, (D^2 H_n)^{-1}) d\beta$ とする.

尚, 弱い条件の下で $C_n = \exp\{(1/2)DH'_n(D^2 H_n)^{-1}DH_n\} + o_p(1) = 1 + o_p(1)$ となる. 定理 1 においては, 対数尤度関数を最尤推定のまわりで展開を行ったが, 定理 2 においては LASSO 推定のまわりで展開を行っている. 最尤推定と LASSO 推定の”ずれ”が C_n に現れている. さて周辺尤度 (11) の近似に話を戻す. 即ち, モデル (1) から生成されている \mathbf{Y} に対して, 推定に用いるモデルを $n(\mathbf{y}|\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$ とし, β の事前分布を両側指数分布 (10) とする. $\Lambda_n = [0, \bar{\lambda}_n]$ とし, さらに以下を仮定する:

- (B1) ある定数 $\underline{c} > 0$, $\bar{c} > 0$ が存在して, 大きな n それぞれに対して, $\lambda_{\min}(n^{-1}\mathbf{X}'\mathbf{X}) \geq \underline{c}$, かつ $\lambda_{\max}(n^{-1}\mathbf{X}'\mathbf{X}) \leq \bar{c}$.
- (B2) $\bar{\lambda}_n \rightarrow \infty$, ($n \rightarrow \infty$).
- (B3) $\frac{\bar{\lambda}_n^2 p_n (\log n)^{1/2}}{n^{1/2}} \rightarrow 0$ ($n \rightarrow \infty$).

ここで $\hat{\mathbf{A}} = \mathbf{I}_{p_n}$, $\delta_n = p_n^{1/2} \bar{\lambda}_n (\log n)^{1/2} n^{-1/2} \rightarrow 0$ ($n \rightarrow \infty$) とする列とすると, 定理 2 から以下が成り立つ.*³

系 3 (B1)-(B3) の下で,

$$\sup_{\lambda \in \Lambda_n} \left| \frac{\mathcal{I}_n(\lambda)}{(2\pi\sigma^2)^{p_n/2} \det(\mathbf{X}'\mathbf{X})^{-1/2} n(\mathbf{y}|\mathbf{X}\hat{\beta}_\lambda, \sigma^2 \mathbf{I}_n) \pi_n^\lambda} - C_n \right| = o_p(1). \quad (15)$$

4 シミュレーション

ここでは, (15) 式で与えられた近似がどれほどの精度を持っているのかをシミュレーションにて確認する. \mathbf{X} のそれぞれの行 \mathbf{x}_i は p 次元正規分布 $N(\mathbf{0}, \Sigma)$ から生成し, 攪乱項 \mathbf{u} は標準正規分布 $N(\mathbf{0}, \mathbf{I}_n)$ に従うとして, $\beta^* = (1, -1.25, 0.75, -0.95, 1.5, 0, 0, 0, 0, 0)'$ とした線形回帰モデル (1) から生成する. ただし $\Sigma = (\rho^{|i-j|})$, $\rho \in [0, 1]$ とする. 次に (11) 式の $\mathcal{I}_n(\lambda)$ に対して, 3 つの近似を考える:

*³ 定理 2 においては, $\lambda > 0$ は固定された値であったが, これは容易に一様性を持つものに拡張できる. 例えば条件 (i), (ii) においては, $P[\dots]$ を $P[\sup_{\lambda \in \Lambda_n} \dots]$ に置き換えればよい.

- (i) (定理 1) $\hat{\mathcal{I}}_{ML}(\lambda) = (2\pi)^{p/2} |\mathbf{X}'\mathbf{X}|^{-1/2} n(\mathbf{y} | \mathbf{X}\hat{\beta}_{ML}, \mathbf{I}_n) \pi(\hat{\beta}_{ML} | \mathbf{1}_p, \lambda)$.
(ii) (定理 2) $\hat{\mathcal{I}}_{LASSO1}(\lambda) = (2\pi)^{p/2} |\mathbf{X}'\mathbf{X}|^{-1/2} n(\mathbf{y} | \mathbf{X}\hat{\beta}_\lambda, \mathbf{I}_n) \pi(\hat{\beta}_\lambda | \mathbf{1}_p, \lambda)$
(iii) (定理 2) $\hat{\mathcal{I}}_{LASSO2}(\lambda) = \hat{\mathcal{I}}_{LASSO1}(\lambda) \tilde{C}_n$,
ただし $\tilde{C}_n = \exp\{(1/2)DH'_n(D^2H_n)^{-1}DH_n\}$ とする.

このとき, $\Lambda = \{0.2j | (j = 1, \dots, 100)\}$ のそれぞれの要素 λ に対して, $\mathcal{I}_n(\lambda)$ と近似式との相対誤差 $r.e. := \sup_{\lambda \in \Lambda} |\log(\mathcal{I}_n(\lambda)) - \log(\hat{\mathcal{I}}_s(\lambda))|$, $s \in \{ML, LASSO, LASSO2\}$ を評価する. このシミュレーションを 100 回繰り返す, そこから得られた 100 個の相対誤差 (r.e) の平均を求めたものが図 3 と図 4 にプロットしてある. 尚, 縦軸は 100 個の相対誤差の平均, 横軸は標本のサイズ n とする. 図 1, 図 2 は, $n = 60$ としたときの, $\mathcal{I}_n(\lambda)$ の正確な値 (—), 近似 (i) (- - -), 近似 (ii) (⋯⋯⋯), 近似 (iii) (- · - · -) の λ の値による変化を表している. 縦軸は近似された周辺尤度の値, 横軸は Λ における j を表している. 例えば横軸の 20 は $\lambda = 0.2 \times 20 = 4$ となる. 尚, $\mathcal{I}_n(\lambda)$ の正確な値は 10^6 個の標本を用いたモンテカルロ近似で計算を行い, トレースプロットにより十分に正確に近似されていることを確かめた.

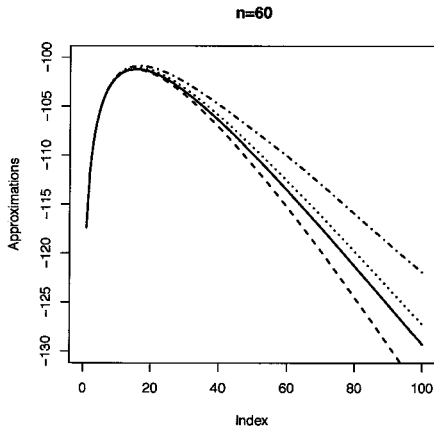


図 1 $n = 60, \rho = 0$.

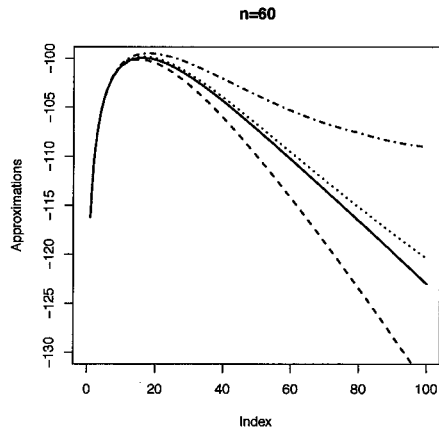
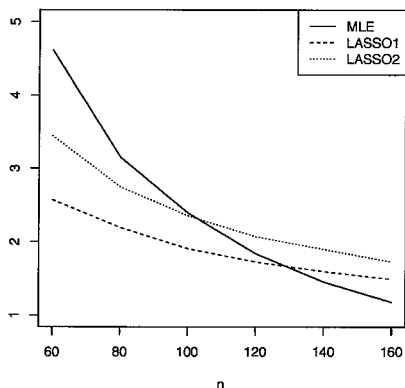
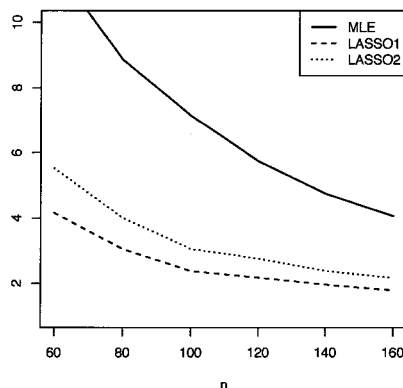


図 2 $n = 60, \rho = 0.5$.

図3 $\rho = 0$.図4 $\rho = 0.5$.

5 まとめ

今回、標本を生成するモデルが線形モデルであるときの周辺尤度に対するラプラス近似の妥当性に関する条件を紹介した. 条件の (B3) からわかるように, $\bar{\lambda}_n = O(1)$ であったとしても, 近似の妥当性を担保するためには $p_n = o(n^{1/2}(\log n)^{-1/2})$ である必要がある. このことは, $p_n > n$ の場合には, 近似の妥当性が崩れてしまうことを意味しており, このような場合にベイズアプローチを適用するためには更なる研究が必要になる. 一方で, β に関して微分不可能な点を持つ両側指数分布の下でも妥当なラプラス近似が導出可能であり, 少なくとも今回行ったシミュレーションにおいては, 説明変数に相関がある場合, 最尤推定を用いたラプラス近似よりも小さい一様誤差を持つことがわかった. しかし定理 2 おける補正項 C_n は一般的にうまく機能していない.

参考文献

- [1] Barber, R. F., Drton, M. and Tan, K. M. (2016). *Laplace approximation in high-dimensional Bayesian regression*: in *Statistical Analysis for High-Dimensional Data*, 15–36, Springer.
- [2] George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection, *Biometrika*, **87**, 731–747.

- [3] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models, *J. R. Statist. Soc. Ser. A (General)*, **135**, 370-384.
- [4] Konishi, S., Ando, T. and Imoto, S., (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks, *Biometrika*, **91**, 27-43.
- [5] Stone, M. (1974). Cross-validated choice and assessment of statistical predictions, *J. R. Statist. Soc. B*, **36**, 111-147.
- [6] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B.*, **58**, 267-288.
- [7] Miyata, Y. (2017a). Laplace approximations and Bayesian Information criteria in possibly misspecified models, *Commun. Stat. A-Theor.* (to appear)
- [8] Miyata, Y. (2017b). Laplace approximations in sparse statistical models. (in preparation for submission)
- [9] Yuan, M., and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *J. Am. Stat. Assoc.*, **100**, 1215-1225.
- [10] Zou, H. and Hastie, T. and Tibshirani, R. (2007). On the “degrees of freedom” of the lasso, *Ann. Stat.*, **35**, 2173-2192.

付録 A 系 3 の a_n の導出について

補題 4 $c_1 > 0, c_2 > 0$ を正の定数とする。このとき、 $-c_1 x^2 + c_2 |x| \leq \frac{c_2^2}{4c_1}$ が成り立つ。

証明 簡単に示せるため、省略する。

ここでは $\hat{\mathbf{A}} = \mathbf{I}_{p_n}$, $\delta_n = p_n^{1/2} \bar{\lambda}_n (\log n)^{1/2} n^{-1/2}$ として定理 2 の (i) を確認する。 $Q_n(\boldsymbol{\beta}) = H_n - H_n(\boldsymbol{\beta}) + \log\{\pi(\boldsymbol{\beta}|\sigma^2, \lambda)/\pi_n^\lambda\}$ は凹関数であることを利用すると、条件 (ii) は Miyata (2017a) と同様の方法で示すことができる。このため (i) の条件を満たす a_n の導出についてのみ説明する。

$$\begin{aligned} \log \frac{\pi(\boldsymbol{\beta}|\sigma^2, \lambda)}{\pi_n^\lambda} &= \frac{\lambda}{2\sigma^2} \sum_{j=1}^{p_n} (|\hat{\beta}_j| - |\beta_j|) \\ &= \frac{\lambda}{2\sigma^2} \|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|_1 \\ &\leq \frac{\lambda}{2\sigma^2} p_n^{1/2} \|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|. \end{aligned} \tag{16}$$

凸関数の最小値に関する劣微分の標準的な定理を用いて LASSO 推定量を評価すると

$$\begin{aligned}
& -\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)'(n^{-1}\mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda) - \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)'\mathbf{X}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}) \\
& \leq \frac{1}{2\sigma^2}\{\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)\}'(n^{-1}\mathbf{X}'\mathbf{X})^{-1}\{\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)\} \\
& \leq \frac{1}{2c\sigma^2}\|\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)\|^2 \\
& \leq \frac{1}{8c\sigma^2}p_n\lambda^2.
\end{aligned} \tag{17}$$

記号の簡略化のため, $g_n = 1 - n^{-1}$ とおく. (16) 式を利用すると

$$\begin{aligned}
& H_n - H_n(\boldsymbol{\beta}) + \log \frac{\pi(\boldsymbol{\beta}|\sigma^2, \lambda)}{\pi_n^\lambda} \\
& \leq -\frac{1}{2\sigma^2}g_n(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda) \\
& + \left\{ -\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)'(n^{-1}\mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda) - \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)'\mathbf{X}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}) \right\} + \frac{\lambda p_n^{1/2}}{2\sigma^2}\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|
\end{aligned}$$

(17) より

$$\begin{aligned}
& \leq -\frac{n}{2\sigma^2}g_n(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)'(n^{-1}\mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda) + \frac{p_n\lambda^2}{8c\sigma^2} + \frac{\lambda}{2\sigma^2}p_n^{1/2}\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\| \\
& \leq -\frac{cn}{4\sigma^2}g_n\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda\|^2 + \frac{1}{8c\sigma^2}(p_n\lambda^2) - \frac{cn}{4\sigma^2}g_n\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda\|^2 + \frac{\lambda}{2\sigma^2}p_n^{1/2}\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\| \tag{18}
\end{aligned}$$

$$\leq -\frac{cn}{4\sigma^2}g_n\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda\|^2 + K(p_n\lambda^2). \tag{19}$$

ただし $K > 0$ はある正の数とする. 最後の不等式 (19) は, (18) 式の最後の 2 つの項に対して, 補題 4 を適用すると導出することができる. これより

$$\begin{aligned}
& \sup_{\lambda \in \Lambda_n} \int_{\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda\| \geq \delta_n} \exp\{n^{-1}Q_n(\boldsymbol{\beta})\} d\boldsymbol{\beta} \\
& \leq \int_{\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda\| \geq \delta_n} \exp\left\{-\frac{c}{4\sigma^2}g_n\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda\|^2\right\} d\boldsymbol{\beta} \sup_{\lambda \in \Lambda_n} \exp\left\{K\left(\frac{p_n\lambda^2}{n}\right)\right\}. \tag{20}
\end{aligned}$$

これより $a_n = K(p_n\lambda_n^2/n)$ となる.