2000 年情報論的学習理論ワークショップ
2000 Workshop on Information-Based Induction Sciences (IBIS2000)
Izu, Japan, July 17-18, 2000.

パラメータ空間が特異点を持つ場合の漸近理論について

On Asymptotic theory for parameter family with singular points

林 正人* Masahito Hayashi

Abstract: It is pointed out that the expected Kullbak-Leibler divergence, which is used for asymptotic evaluation of estimation in regular models, is not suitable as a risk when a parametric space has singular points. Three new criteria are introduced for such a case. It is proved that the MLE is the optimal under these criteria in a toy model with a singular point.

キーワード: 漸近論, 特異点, 基準の取り方, 大偏差型評価, 収束の一様性

1 はじめに

近年ニューラルネットの研究に関連してパラメータ空間 $\mathcal{Y}$ が特異点を持つ場合のパラメータ推定の必要となりつつある [1, 2, 3]. 正則なモデルについては 1 次漸近論は漸近的 Cramér-Rao 不等式に要約されるが, 特異点を持つ場合については十分に議論されていない. 本稿では, 指数型分布族 $\mathcal{X}$ の部分分布族 $\mathcal{Y}$ のパラメータ推定の 1 次漸近論について $\mathcal{Y}$ が特異点を持つ場合を含めて議論する. 以下本稿の構成を述べる. $\mathcal{Y}$ が特異点を持つときは $\mathcal{Y}$ の座標を考えることが出来ず, 平均 2 乗誤差行列やその対角和に代って Kullbak-Leibler divergence $D(p_X\|p_y)$ (以下 $D(X\|y)$ と略す.) を用いて真の分布が p_y であるときの推定量 X $(X(\omega) \in \mathcal{Y})$ の誤差を $\mathrm{E}_y(D(X\|y))$ で評価することが多い. (E_y は確率分布 p_y または p_y^n についての期待値を表す.) そして 1 次漸近論を扱うには推定量の無限列 $\vec{X} := \{X_n\}$ (以下特に誤解が無い限り推定量の無限列を単に推定量と記す.) に対して極限値

$$\mathcal{R}(\vec{X}, y) := \lim_{n\to\infty} n\mathrm{E}_y(D(X_n\|y)) \qquad (1)$$

を基準量に取ることが多い. しかし, §2 で議論するように超有効性の問題が生じ, これを基準量とすることは妥当ではない. なお, 正則なモデルでも Hodges 推定量のような超有効な推定量は存在するが推定量に適切な一致性条件を課すことにより超有効な推定量を排除することができる. しかし, 特異点を持つ場合に, 同様な一致性条件を推定量に課すと, そのような条件を満たす推定量の存在は確認できていない. 少なくとも特異点を持つ場合では最尤推定量はそのような一致性条件を満たさない. (おそらく存在しないと思われる.) そのため, 基準量として極限値 $\mathcal{R}(\vec{X}, y)$ を取り, 特異点を持つ場合では超有効な推定量は排除し, 最尤推定量のような性質の良い推定量を許容する一致性条件は見つかっていない.

このような状況の中で取るべき戦略は 2 つ考えられる. 1 つは, あくまでも基準量 $\mathcal{R}(\vec{X}, y)$ にこだわり, 上記の条件を満たす一致性条件を探す方法である. しかし, 何らかの意味で基準量 $\mathcal{R}(\vec{X}, y)$ の下で, そのような一致性条件が存在しないことが示されると, この戦略は放棄せざるを得ない.

もう 1 つの戦略は基準量 $\mathcal{R}(\vec{X}, y)$ 以外の適切な基準量を探すことである. 本稿ではこの戦略を採用し, 3 つの基準量を提案する. そのうちはじめの 2 つは特異点 $y_0 \in \mathcal{Y}$ と推定量 $\vec{X}$ について分布 $p^n_{y_0+\frac{x}{\sqrt{n}}} \circ X_n^{-1}$ の極限を考え, 分布 $p^n_{y_0} \circ X_n^{-1}$ の極限との相違に注目する方法である. 1 つ目ではこの 2 つの分布の間の Kullbak-Leibler divergence で測った情報量に注目し, 2 つ目ではこの 2 つの分布の識別の難易度に注目する. そして 3 つ目の基準として, 大偏差に注目する方法を提案する.

2 特異点がある場合の Kullbak-Leibler divergence 誤差

特異点を持つモデルの典型的な例として $\mathcal{X}$ が共分散行列が単位行列であり, $\mathcal{Y}$ が x 軸, y 軸の和集合からなるモデルが知られており十字モデルと呼ばれる. このモデルは原点 $(0, 0)$ (以下誤解の恐れが無い限り 0 と記す.) で特異点を持つ. このモデルを例に取りながらこの

*理化学研究所 脳科学総合研究センター 脳数理研究チーム〒 351-0198 埼玉県和光市広沢 2-1 tel. 048-467-9664(共用),fax 048-467-9693(共用) e-mail: masahito@brain.riken.go.jp,
Lab. for Mathematical Neuroscience, Brain-Style Information Systems Research Group, Brain Science Institute, RIKEN 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan

節では先に挙げた極限値 $\mathcal{R}(\vec{X}, y)$ を基準に取ることが適切ではないことを議論する.

このモデルで最尤推定量 $\vec{X}$ に注目すると,

$$\mathcal{R}(\vec{X}, y)\begin{cases} = 1 & y \neq 0 \\ \leq 2 & y = 0 \end{cases}$$

となることが知られている. このため, このような特異点を持つモデルでは y が特異点以外のところでは $\mathcal{R}(\vec{X}, y)$ が x の近傍での次元と一致する条件の下で, x が特異点であるときに $\mathcal{R}(\vec{X}, y)$ がどこまで小さくできるかという問題設定を採用しがちである. しかし, 以下に述べるようにこの問題設定の下では, 原理的に超有効性の排除が不可能である. 例えば上の十字モデルではサンプル平均 $\overline{\omega}_n$ に対して

$$X_n^{\epsilon}(\overline{\omega}_n) := \begin{cases} 0 & \tan(\frac{\pi}{4} - \epsilon) \leq \left|\frac{\overline{\omega}_{n,2}}{\overline{\omega}_{n,1}}\right| \leq \tan(\frac{\pi}{4} + \epsilon) \\ \hat{X}_n(\overline{\omega}_n) & \text{otherwise} \end{cases}$$

という推定量を考えると,

$$\mathcal{R}(\vec{X}^{\epsilon}, y)\begin{cases} = 1 & y \neq 0 \\ \leq 2\frac{\pi - 4\epsilon}{\pi} & y = 0 \end{cases}$$

となり, ϵ を $\frac{\pi}{4}$ に近づければいくらでも, $\mathcal{R}(\vec{X}^{\epsilon}, 0)$ を小さくできる. 正則な 1 次元モデルであっても原点以外では $\mathcal{R}(\vec{X}, 0)$ が 1 でるという条件の下で $\mathcal{R}(\vec{X}, 0)$ をいくらでも小さくすることができる. (Hodges 推定量) このような推定量は超有効推定量と呼ばれ通常はそのような超有効推定量を排除した上で漸近論を扱うことになる. しかし, 超有効推定量を Cramér-Rao 不等式が与える下限よりも誤差が小さくなるからという理由で排除するわけにはいかない.

通常, 統計学ではいわゆる弱一致性 ($p_y^n\{\|X_n - y\| > \epsilon\} \to 0$) だけではなく, より強い一致性条件を推定量に課すことによって, 超有効推定量を排除する. 例えば, (1) の右辺の収束の局所一様性を課すと上記の様な超有効推定量は除外でき, 漸近的 Cramér-Rao 不等式が成立する. これはあくまでも正則な場合の事情である. では特異点を持つ場合はどうであろうか. 同様に (1) の右辺の収束の局所一様性を推定量 $\vec{X}$ に課すことにすると, $y \mapsto \mathcal{R}(\vec{X}, y)$ の連続性が導ける. 十字モデルの場合に当てはめて考えると最尤推定量 $\vec{X}$ はこの条件を満たさないことが確かめられる. それどころか, 収束の局所一様性を課した段階で, 特異点以外で $\mathcal{R}(\vec{X}, y) = 1$ となる条件の下で, $\mathcal{R}(\vec{x}, 0)$ がどれだけ小さくできるかという問題設定が, 無意味になる. さらに議論を進めると, 上述の収束の局所一様性を満たす推定量の存在が問題である. (著者は存在しないと予想している.) このような理由で y が特異点であるときに, 極限値 $\mathcal{R}(\vec{X}, y)$ を評価の基準として採用することは妥当性を欠く.

3 極限分布に注目した基準の提案

では特異点があるときの評価の基準量として, どのようなものを考えれば良いであろうか. 以下では, 全く一般の指数型分布族 $\mathcal{X}$ の部分集合 $\mathcal{Y}$ を扱うことは困難であるので, 以下に定義する漸近的安定性を満たす $\mathcal{Y}$ について考えることにする.

定義 **1** $\mathcal{Y}$ が以下の条件を満たすとき $y_0 \in \mathcal{Y}$ で漸近的に安定であるという; 任意の $x_0 \in \mathcal{X}$ に対して, $\inf\{\|x_0 - x\| | x \in \mathcal{T}_{y_0,n}\mathcal{Y}\}$ が収束する. ただし, $\|x_0 - x\|$ は x, x_0 を自然パラメータとして見たときのベクトルとしてのノルムを表し, $\mathcal{T}_{y_0,n}\mathcal{Y} := \{x \in \mathcal{X} | y_0 + \frac{x}{\sqrt{n}} \in \mathcal{Y}\}$ とする. さらに, 上記の値が 0 に収束する点 x の集合を $\mathcal{T}_{y_0}\mathcal{Y}$ で表す. このとき $\mathcal{T}_{y_0}\mathcal{X}$ を *Fisher* 情報行列 J_{y_0} の逆行列 $J_{y_0}^{-1}$ を共分散行列として持つ正規分布族と考え, $\mathcal{T}_{y_0}\mathcal{Y}$ をその部分分布族とみなす. 特に, $\mathcal{Y}$ の任意の点 y_0 について $\mathcal{Y}$ が漸近的安定であるとき $\mathcal{Y}$ は漸近的安定であるという.

モデルが正則であるときは漸近的安定であり, $\mathcal{T}_{y_0}\mathcal{Y}$ はベクトル空間になる. 十字モデルも漸近的安定であり $\mathcal{T}_{y_0}\mathcal{Y}$ は特異点以外でベクトル空間になり, 特異点である原点では $\mathcal{T}_{y_0}\mathcal{Y}$ は元のモデル $\mathcal{Y}$ と一致する.

例えば次のようなモデルを考えると漸近安定性を満たさない. $\mathcal{X}$ を先の十字モデルと同様に取り, $\mathcal{Y} := \{(x_1, x_2) \in \mathcal{X} | x_2 = x_1 \sin\frac{1}{x_1}, x_1 \neq 0\} \cup \{(0,0)\}$ とする $x = (0,0)$ で特異点になり上述の漸近的安定性を満たさない.

漸近的安定なモデル $\mathcal{Y}$ の推定量 $\vec{X}$ について以下の量を考える.

$$\begin{aligned} &\mathcal{C}_1(\vec{X}, y_0) \\ &:= \lim_{\epsilon \to 0} \frac{1}{\epsilon^2} \inf_{\|x\| = \epsilon} \lim_{n \to \infty} D(p_{y_0}^n \circ X_n^{-1} \| p_{y_0 + \frac{x}{\sqrt{n}}}^n \circ X_n^{-1}). \end{aligned}$$

もし, 推定量 X_n に情報量の損失が無ければ, $\frac{J_{y_0}}{2}$ に一致するが, 十字モデルのように特異点がある場合には一般には, 情報量の損失があり, $\frac{J_{y_0}}{2}$ より小さくなる.

ここで推定量 $\vec{X}$ に対して $\mathcal{T}_{y_0}\mathcal{X}$ の部分状態族 $\mathcal{T}_{y_0}\mathcal{Y}$ に対する推定量として極限推定量 $\tilde{X}$ を $\tilde{X}(x) := \lim_{n \to \infty} \sqrt{n}\, X_n(y_0 + \frac{x}{\sqrt{n}} - y_0)$ で定義する. 以下では, この極限推定量の存在を推定量 $\vec{X}$ に課す. すると以下の式が成立

する.

$$\lim_{n\to\infty} D(p^n_{y_0}\circ X_n^{-1}\|p^n_{y_0+\frac{x}{\sqrt{n}}}\circ X_n^{-1}) = D(P_0\circ\tilde{X}^{-1}\|P_x\circ\tilde{X}^{-1}).$$

ここで右辺の P_0, P_x は $\mathcal{T}_{y_0}\mathcal{X}$ 内の分布を表す.

定理 2 十字モデルの場合には基準 $\mathcal{C}_1(\vec{X}, y_0)$ の下で最尤推定量 $\vec{X}$ の最適となり, その値は $\frac{1}{2}-\frac{1}{4\pi}\int_{-\infty}^{\infty}\int_{-|b|}^{|b|} a^2 e^{-\frac{a^2+b^2}{2}}\,da\,db$ となる.

次に 2 つの分布の識別の容易度に注目する. すなわち, 推定値 $X_n(\omega)$ から 2 つの分布 $p^n_{y_0}, p^n_{y_0+\frac{x}{\sqrt{n}}}$ の識別を行う問題を考える. このとき二種の誤りが考えられるがそれぞれの誤りを犯す確率の和を考える. その最小値はネイマン-ピアソンの補題を用いると

$$p^n_{y_0}\left\{X_n(x')\,\middle|\,p^n_{y_0}(X_n(x')) \geq p^n_{y_0+\frac{x}{\sqrt{n}}}(X_n(x'))\right\} + p^n_{y_0+\frac{x}{\sqrt{n}}}\left\{y\,\middle|\,p^n_{y_0}(X_n(x')) < p^n_{y_0+\frac{x}{\sqrt{n}}}(X_n(x'))\right\}$$

で与えられ, その極限値は

$$D_{y_0}(\tilde{X}, x) := P_0\left\{\tilde{X}(x')\,\middle|\,P_0(\tilde{X}(x')) \geq P_x(\tilde{X}(x'))\right\} + P_x\left\{\tilde{X}(x')\,\middle|\,P_0(\tilde{X}(x')) < P_x(\tilde{X}(x'))\right\} \tag{2}$$

で与えられる. 本稿では

$$\mathcal{C}_2(\vec{X}, y_0) := \lim_{\epsilon\to 0}\frac{-1}{\epsilon}\log\sup_{\|x\|=\epsilon} D_{y_0}(\tilde{X}, x)$$

を誤差評価の基準に取って考える.

定理 3 十字モデルについては最尤推定量 $\vec{X}$ は基準 $\mathcal{C}_2(\vec{X}, y_0)$ に関して最適となり, そのときの値は $\frac{1}{\sqrt{2}}$ となる.

4 大偏差に注目した基準

そもそも, 正則な場合に極限値 $\mathcal{R}(\vec{X}, y)$ を基準量に取るという根拠は Kullbak-Leibler divergence 誤差 $E_y(D(X_n\|y))$ が 0 に収束する早さを極限値 $\mathcal{R}(\vec{X}, y)$ が表しているからである. しかし,(1) の右辺の収束が一様 (または局所一様) でないときには極限値 $\mathcal{R}(\vec{X}, y)$ は $E_y(D(X_n\|y))$ が 0 に収束する早さを表しているとは言えない. その意味でも特異点のあるときに極限値 $\mathcal{R}(\vec{X}, y)$ を基準量として扱うことは妥当性を欠く.

他に $E_y(D(X_n\|y))$ が 0 に収束する早さを見る方法が見つからないので, 代わりに裾確率が 0 に収束する早さに注目することにする. すなわち, 大偏差を基準量として扱うことになる.

始めに極限値

$$\beta(\vec{X}, y, \epsilon) := \lim_{n\to\infty}\frac{-1}{n}\log p^n_y\{D(X_n\|y) \geq \epsilon\} \tag{3}$$

を考える. たとえモデル $\mathcal{Y}$ が特異点を持っても, 少なくとも最尤推定量 $\vec{X}$ の場合には (3) の右辺は変数 y, ϵ について局所一様収束することが容易に確かめられる. そして, その極限値

$$\alpha(\vec{X}, y) := \lim_{\epsilon\to 0}\frac{\beta(\vec{X}, y, \epsilon)}{\epsilon} \tag{4}$$

を基準量として扱うことが多い. しかし, (4) の右辺は特異点を持つときには, 最尤推定量 $\vec{X}$ の場合には y について一様に収束するとは限らない. 例えば十字モデルの場合には

$$\beta(\vec{X}, (a,0), \epsilon) = \begin{cases} \epsilon & \epsilon \leq \frac{a^2}{4} \\ \frac{a^2}{4} & \frac{a^2}{4} \leq \epsilon \leq \frac{5a^2}{8} \\ 2\epsilon - \frac{a^2}{2} - \sqrt{2}a\sqrt{\epsilon - \frac{a^2}{2}} & \frac{5a^2}{8} \leq \epsilon \leq a^2 \\ \epsilon - \frac{a^2}{2} & a^2 \leq \epsilon \end{cases}$$

となる. 極限値 $\alpha(\vec{X}, y)$ は y に関わらず 1 となるが, 原点の近傍では一様収束ではない. 一般に特異点の近傍では (4) が一様収束しないことが示せる. したがって, この場合にも ϵ が十分小さいときの $\beta(\vec{X}, y, \epsilon)$ の挙動を見るために極限値 $\alpha(\vec{X}, y)$ を見ることは適切ではない.

なお, 最尤推定量 $\vec{X}$ に注目する限り, $\alpha(\vec{X}, y)$ は常に 1 になることから, $\frac{\beta(\vec{X}, y, \epsilon)}{\epsilon}$ が 1 に収束する早さを見るのが妥当と思われる. 特に特異点があることによってその推定精度の悪化に興味があるので, その特異点の近傍での収束の遅さを以下の方法で測ることにする.

$$\gamma(\vec{X}, y, \delta) := \inf\left\{\epsilon\,\middle|\,\frac{\beta(\vec{X}, y, \epsilon)}{\epsilon} < 1-\delta\right\}.$$

(4) の右辺が特異点の近傍で一様収束しない原因は y を特異点に近づけたときに $\gamma(\vec{X}, y, \delta)$ が 0 に収束することにある. 例えば,$\gamma(\vec{X}, y, \delta)$ が 0 に収束する早さを見るために,

$$\mathcal{C}_3(\vec{X}, y_0) := \inf_{\delta>0}\liminf_{b\to 0}\left\{\frac{\gamma(\vec{X}, y, \delta)}{D(y_0\|y)}\,\middle|\,D(y_0\|y) = b\right\}$$

を 1 つの基準量として考えることは可能である.

定理 4 十字モデルの場合, 基準量 $\mathcal{C}_3(\vec{X}, y_0)$ に関して最尤推定量 $\vec{X}$ は最適となり, そのときの基準量 $\mathcal{C}_3(\vec{X}, y_0)$ の値は $\frac{1}{2}$ となる.

5　今後の課題

パラメータ空間に特異点がある場合について, 正則な場合に使われてきた基準量 $\mathcal{R}(\vec{X}, x)$ を用いることが適切ではないことを指摘した. その上でそれに代る 3 つの基準量 $\mathcal{C}_1(\vec{X}, y_0), \mathcal{C}_2(\vec{X}, y_0), \mathcal{C}_3(\vec{X}, y_0)$ を提案し, 十字モデルの場合についてそれらの下で, 最尤推定量が最適になることを示した. しかしながら, これらの基準量がどのような意味を持つかについては十分に考察できていない. 特に, 基準量 $\mathcal{C}_1(\vec{X}, y_0)$ については情報量損失に注目したわけであるが, このような状況の下で情報量損失が如何なる操作的意味を持つかについて全く議論していない. 今後,$\mathcal{C}_1(\vec{X}, y_0)$ の定義で与えたような極限の取り方で情報量の損失がどのような意味を持つか考察する必要がある.

また, 十字モデル以外の一般のモデルとりわけ本稿で定義した漸近的安定性を満たすモデルでもこれらの基準の下で, 最尤推定量の最適性が示せると思われ, これらについては今後の課題となる. その他, Watanabe[1] によってなされた, 特異点を持つモデルに関するベイズ推定量の枠組みでの研究との関連や本稿で扱った議論の幾何学的考察は今後の研究課題となる. 本稿で扱った内容とニューラルネットワークとの関連は興味深いところでありこれについても今後の課題となる.

謝辞

甘利俊一教授には本研究を進める上で色々と御議論して頂きました.

A　定理 2 の証明

まず, 極限推定量 $\tilde{X}$ はモデルの持つ対称性から 90 度回転,x 軸,y 軸反転について共変なものに制限できる.

したがって $D_1(a) := \tilde{X}^{-1}((a, 0))$ は y 軸対象 $D_2(b) := \tilde{X}^{-1}((0, b))$ は x 軸対象になる.

$$
\begin{aligned}
& D(P_{(0,0)} \circ \tilde{X}^{-1} \| P_{(0,\epsilon)} \circ \tilde{X}^{-1}) \\
=& \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\int_{D_2(b)} e^{-\frac{|x|^2}{2}} dx \right) \left(\log \left(\int_{D_2(b)} e^{-\frac{|x|^2}{2}} dx \right) \right. \\
& \left. - \log \left(\int_{D_2(b)} e^{-\frac{|x|^2 - 2x_1\epsilon + \epsilon^2}{2}} dx \right) \right) db \\
& + \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\int_{D_1(a)} e^{-\frac{|x|^2}{2}} dx \right) \left(\log \left(\int_{D_1(a)} e^{-\frac{|x|^2}{2}} dx \right) \right. \\
& \left. - \log \left(\int_{D_1(a)} e^{-\frac{|x|^2 - 2x_1\epsilon + \epsilon^2}{2}} dx \right) \right) da. \qquad (5)
\end{aligned}
$$

ここで x は $\mathcal{T}_{y_0}\mathcal{X}$ の要素を表すことに注意されたい. (5) の第 2 項は $\tilde{X}$ が最尤推定量のとき $\frac{\epsilon}{4}$ になり, 一般には

$$
\begin{aligned}
& \frac{\epsilon}{4} \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-|y|}^{|y|} e^{-\frac{|x|^2}{2}} \left(\log e^{-\frac{|x|^2}{2}} - \log e^{-\frac{|x|^2 - 2x_1\epsilon + \epsilon^2}{2}} \right) dx \\
&\geq \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\int_{D_1(a)} e^{-\frac{|x|^2}{2}} dx \right) \left(\log \left(\int_{D_1(a)} e^{-\frac{|x|^2}{2}} dx \right) \right. \\
& \quad \left. - \log \left(\int_{D_1(a)} e^{-\frac{|x|^2 - 2x_1\epsilon + \epsilon^2}{2}} dx \right) \right) da
\end{aligned}
$$

となる. 一方 (5) の第 1 項については ϵ について展開すると,

$$
\begin{aligned}
& \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\left(\int_{D_2(b)} x_1 e^{-\frac{|x|^2}{2}} dx \right) \epsilon \right. \\
& \quad + \left(\int_{D_2(b)} (1 - x_1^2) e^{-\frac{|x|^2}{2}} dx \right. \\
& \quad \left. \left. + \left(\int_{D_2(b)} e^{-\frac{|x|^2}{2}} dx \right)^{-1} \left(\int_{D_2(b)} x_1 e^{-\frac{|x|^2}{2}} dx \right)^2 \right) \frac{\epsilon^2}{2} \right) db
\end{aligned}
$$

となり,$D_2(b)$ の y 軸に関する対称性から以下の式に変形できる.

$$
\begin{aligned}
& \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{D_2(b)} (1 - x_1^2) e^{-\frac{|x|^2}{2}} dx\, db \frac{\epsilon^2}{2} \\
&= \left(\frac{1}{4} - \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-|b|}^{|b|} a^2 e^{-\frac{a^2 + b^2}{2}} da\, db \right) \epsilon^2.
\end{aligned}
$$

したがって (5) の第 1 項を ϵ^2 で割った値の極限値は極限推定量 $\tilde{X}$ が上述の対称性を満たす限り, $\tilde{X}$ に依存しない.

一方,(5) の第 2 項を ϵ^2 で割った値の極限値は先の議論から, $\frac{1}{4}$ で上から押さえられ, 最尤推定量のとき $\frac{1}{4}$ となることから, 最尤推定量の最適性が示され, そのときの最適値は $\frac{1}{2} - \frac{1}{4\pi} \int_{-\infty}^{\infty} \int_{-|b|}^{|b|} a^2 e^{-\frac{a^2 + b^2}{2}} da\, db$ となり, 情報量損失は $\frac{1}{4\pi} \int_{-\infty}^{\infty} \int_{-|b|}^{|b|} a^2 e^{-\frac{a^2 + b^2}{2}} da\, db$ となる.

B　定理 3 の証明

定理 2 の証明と同様に, $D_1(a), D_2(b)$ を定義する. そして先に定理 2 の証明で述べたような対称性を満たす極限推定量 $\tilde{X}$ に制限して議論しても良い. $D_2(b)$ の y 軸に対する対称性から以下の不等式が導ける. (証明は後ほど行う.)

$$
\int_{D_2(b)} e^{-\frac{|x|^2}{2}} dx \geq \int_{D_2(b)} e^{-\frac{|x|^2 - 2x_1\epsilon + \epsilon^2}{2}} dx \quad \forall b \in \mathbb{R}. \qquad (6)
$$

したがって, 適当な $a(\epsilon)$ を取ると (2) の右辺の第 1 項第 2 項はそれぞれ以下のように表せる.

$$
\begin{aligned}
& P_0\left\{\tilde{X}(x') \,\middle|\, P_0(\tilde{X}(x')) \geq P_{(\epsilon,0)}(\tilde{X}(x'))\right\} \\
&= \int_{a(\epsilon)}^{\infty} \int_{D_1(a)} e^{-\frac{|x|^2}{2}}\, dx\, da, \qquad (7) \\
& P_{(\epsilon,0)}\left\{\tilde{X}(x') \,\middle|\, P_0(\tilde{X}(x')) < P_{(\epsilon,0)}(\tilde{X}(x'))\right\} \\
&= \int_{-\infty}^{a(\epsilon)} \int_{D_1(a)} e^{-\frac{|x|^2-2x_1\epsilon+\epsilon^2}{2}}\, dx\, da \\
&\quad + \int_{|x_2|>|x_1|} e^{-\frac{|x|^2-2x_1\epsilon+\epsilon^2}{2}}\, dx. \qquad (8)
\end{aligned}
$$

そして (7) の右辺及び (8) の右辺の第 2 項の和については以下の不等式が成り立つ.

$$
\begin{aligned}
& \int_{a(\epsilon)}^{\infty} \int_{D_1(a)} e^{-\frac{|x|^2}{2}}\, dx\, da \\
&\quad + \int_{-\infty}^{a(\epsilon)} \int_{D_1(a)} e^{-\frac{|x|^2-2x_1\epsilon+\epsilon^2}{2}}\, dx\, da \\
&\geq \int_{|x_2|\leq|x_1|, x_1>\frac{\epsilon}{2}} e^{-\frac{|x|^2}{2}}\, dx \\
&\quad + \int_{|x_2|\leq|x_1|, x_1\leq\frac{\epsilon}{2}} e^{-\frac{|x|^2-2x_1\epsilon+\epsilon^2}{2}}\, dx.
\end{aligned}
$$

特に,$\tilde{X}$ が最尤推定量のときは等号が成立する. この事実と (8) の右辺の第 1 項が極限推定量 $\tilde{X}$ に依存しないことから, 基準 $\mathcal{C}_2(\vec{X}, 0)$ の下での最尤推定量の最適性が確認できる. このとき $D_0(\vec{\tilde{X}}, (\epsilon, 0))$ は以下で与えられる.

$$
\begin{aligned}
& 1 - P_0\left(\left\{(x_1,x_2) \,\middle|\, x_2 \geq \frac{\epsilon}{2}, x_2 + \frac{\epsilon}{2} > x_1 > x_2 - \frac{\epsilon}{2}\right\}\right) \\
& + P_0\left(\left\{(x_1,x_2) \,\middle|\, x_2 \leq \frac{\epsilon}{2}, -x_2 + \frac{\epsilon}{2} > x_1 > -x_2 - \frac{\epsilon}{2}\right\}\right) \\
& + P_0\left(\left\{(x_1,x_2) \,\middle|\, \frac{\epsilon}{2} > x_1 > -\frac{\epsilon}{2}, \frac{\epsilon}{2} > x_2 > -\frac{\epsilon}{2}\right\}\right).
\end{aligned}
$$

したがってこれを計算すると,

$$
\lim_{\epsilon\to 0} \frac{-1}{\epsilon} \log D_0(\vec{\tilde{X}}, (\epsilon, 0)) = \frac{1}{\sqrt{2}}
$$

となる.

以下 (6) を示す. $D_2(b)$ 上の確率測度

$$
\mu(\, dx) := \frac{e^{-\frac{|x|^2-2x_1\epsilon+\epsilon^2}{2}}}{\int_{D_2(b)} e^{-\frac{|x|^2-2x_1\epsilon+\epsilon^2}{2}}\, dx}\, dx
$$

と凸関数 $z \mapsto e^z$ について Jensen の不等式を適用すると,

$$
\int_{D_2(b)} e^{\frac{-2x_1\epsilon+\epsilon^2}{2}} \mu(\, dx) \geq e^{\mathrm{E}_\mu\left(\frac{-2x_1\epsilon+\epsilon^2}{2}\right)} = e^{\frac{\epsilon^2}{2}} > 1
$$

となる. ここで $D_2(b)$ の y 軸についての対称性を用いた. この式と $\mu(\, dx)$ の定義から (6) が得られる.

C 定理 4 の証明

定理 2 の証明で述べたように各推定量 X_n が対称性を満たすとして議論しても良い.

したがって

$$
\begin{aligned}
& \{\overline{\omega}_n \,|X_n(\overline{\omega}_n) = (x_1, 0), x_1 > 0\} \\
& \subset \{(x_1, x_2) | x_1 > 0, |x_2| < x_1\}
\end{aligned}
$$

となり,

$$
\lim_{n\to\infty} \frac{-1}{n} \log p^n_{(\epsilon,0)}\left(\{\overline{\omega}_n \,|X_n(\overline{\omega}_n) = (x_1, 0), x_1 > 0\}^c\right) \leq \frac{\epsilon^2}{4}
$$

が示せる. 一方

$$
\begin{aligned}
& \{\overline{\omega}_n \,|X_n(\overline{\omega}_n) = (x_1, 0), x_1 > 0\}^c \\
& \subset \left\{\overline{\omega}_n \,\middle|\, D(X_n(\overline{\omega}_n) \| (\epsilon, 0)) > \frac{\epsilon^2}{2}\right\}
\end{aligned}
$$

となることから,

$$
\gamma(\vec{X}, (\epsilon, 0), \delta) \leq \frac{\frac{\epsilon^2}{4}}{1-\delta}
$$

が得られる. $D((0,0)\|(\epsilon,0)) = \frac{\epsilon^2}{2}$ となることから,

$$
\mathcal{C}_3(\vec{X}, (0,0)) \leq \frac{1}{2}
$$

を得る. ここで各 X_n として最尤推定量 $\hat{X}_n$ を用いると, 上記の全ての不等号のところで等号が成立し, 最尤推定量の最適性が示せる.

参考文献

[1] S. Watanabe, "Algebraic analysis for non-identifiable learning machines", Neural Computation, to appear.

[2] K. Fukumizu and S. Amari, "Local minima and plateaus in hierarchical structures of multilayer perceptorons", Neural Networks, to appear.

[3] S. Amari, T. Ozeki and H. Park, "Information geometry of adaptive systems", Symposium 2000 on Adaptive Systems for Signal Processing, Communications, and Control.