# パラメータ推定問題における MDL 原理とベイズ符号について

竹内 純一

RWCP[1] 理論 NEC 研究室 [2]

〒 216 神奈川県川崎市宮前区宮崎 4-1-1

電子メール tak@sbl.cl.nec.co.jp

[1]RWCP: Real World Computing Partnership (新情報処理開発機構)

[2]NEC C&C 研究所内

あらまし　確率分布の逐次型推定問題におけるベイズ符号と MDL 原理の関係を調べた. ベイズ符号は, ジェフリーズの事前分布を用いたとき, ユニバーサル符号として漸近的に (minimax の意味で) 最適になることが, 最近示された. これをベルヌーイ系列の推定量として用いた場合, ラプラス推定量 $(k+0.5)/(N+1)$ を, KL 情報量の意味で最適な推定量として導く. 一方で MDL 原理からも同一の推定量が得られるが, その導き方には恣意性がある. しかし, 一般にはベイズ符号はその像がもとのクラスに属さないという問題がある. ここに着目し, 一般の指数型分布族の推定問題において, MDL 原理とベイズ符号の関係を調べた. 結果として, MDL 原理にある幾何学的に解釈出来る前提を加えて得られる推定量が, ある条件のもとで, 最適なベイズ符号の元のクラスへの射影に $O(1/N)$ の項まで一致することが分かった. また, 得られる推定量は自然座標に関してバイアス補正した推定量とも $O(1/N)$ まで一致した. ベイズ符号の最適性から, これらの推定量はその像がもとのクラスに属するものとしては最良であると予想出来る.

和文キーワード　ベイズ符号, 記述長最小原理, ラプラス推定量, ジェフリーズの事前分布, 指数型分布族

# The MDL principle and the Bayes code in Parameter Estimation

Jun-ichi Takeuchi

Theory NEC Laboratory[3], RWCP[4]

4-1-1 Miyazaki, Miyamae-ku, Kawasaki, Kanagawa 216, Japan

e-mail tak@sbl.cl.nec.co.jp

[3] c/o C&C Research Laboratories, NEC Corporation

[4] RWCP: Real World Computing Partnership

Abstract　We analyze the relation between the MDL principle and the Bayes code when used for on-line parameter estimation. The Bayes code is a recently proposed scheme of universal coding, and it has been shown that the Bayes code with Jeffreys prior is asymptotically optimal as universal code. When applied to the parameter estimation problem for a Bernoulli trial, the Bayes code translates to the Laplace estimator $(k+0.5)/(N+1)$ as the optimal estimator w.r.t. the KL divergence. The MDL principle derives the same estimator, but there is subjectivity in the choice of coding. The Bayes code has the defect, on the other hand, that its image does not generally belong to the original class of sources. Inspired by this point, we analyzed the relation between the two estimators in parameter estimation problems for the general exponential family. As a result, we have shown that the estimator derived from the MDL principle together with a certain assumption which has a natural geometrical interpretation agrees with the projection of the optimal Bayes code to the original class up to the $O(1/N)$ term. Moreover, the two estimators agree with the bias corrected m.l.e. w.r.t. the natural parameter up to the $O(1/N)$ term. We conjecture that these estimators are optimal among those whose image belonging to the original class.

英文 Keywords　Bayes code, MDL principle, Laplace estimator, Jeffreys prior, exponential family

# 1 まえがき

i. i. d. 系列からのパラメータ推定問題において, MDL (minimum description length) 推定量と新しいユニバーサル符号として注目されているベイズ符号との関係を調べた.

ベイズ符号の考え方は, 二つの重要な点を含んでいる. 一つはリファレンスプライアの方法を適用している点であり, もう一つは, 二段階符号化はベイズ符号に本質的に劣るという点である.

二段階符号化とは, データから情報源 (確率モデル) の構造を推定してそれを符号化し, さらにそのモデルのもとでデータを符号化するという方法である. これは MDL 原理 [Riss78] の根拠となっている (Rissanen は predictive MDL という二段階でないものも提唱している [Riss87] が, 本稿では MDL 原理で二段階のもののみを指すものとする). すなわち MDL 原理は, データが与えられたとき, 二段階符号化で最小の符号長 (MDL) を与えるモデルが情報源の推定値として最良であると主張する. こうして作られる推定量を MDL 推定量と呼ぶが, 特に情報源のクラスの複雑さ (パラメータ数) が分からない次数選択問題において, 理論的にその有効性が証明されている [BC91], [Yam92], [竹内 93].

一方で, ベイズ符号は本質的に二段階符号化とは異なり, モデルを記述するという考えがない. ベイズ符号とは, 事前分布 $w(\theta)$ について, 平均冗長度

$$R_N(\theta, q) \equiv \int p(x^N|\theta) \ln \frac{p(x^N|\theta)}{q(x^N)} dx^N$$

の期待値 $R_N(w, q) \equiv \int R_N(\theta, q) w(\theta) d\theta$ ($q$ の平均符号長のベイズリスクと呼ぶ) を最小にする確率モデル $q(x^N) = p_w(x^N) \equiv \int p(x^N|\theta) w(\theta) d\theta$ でデータ $x^N$ を記述するという方法である. データの各実現値に対するベイズ符号の符号長は Rissanen の stochastic complexity に一致するが, それが MDL を下界することが [Riss87] で既に指摘されている. すなわち, 二段階符号化は本質的にベイズ符号に劣るのである.

ベイズ符号自体は, 従来からあるベイズ統計の枠組をユニバーサルデータ圧縮に用いたに過ぎない. 従ってその性能は仮定する事前分布に依存する. 事前分布があらかじめ分かっていれば, それを用いるのが最良であるのは勿論だが, ユニバーサルデータ圧縮に望まれる性質は, 通常, 平均冗長度の最悪値 $\max_\theta R_N(\theta, q)$ が最小になることであろう. この性質をもった符号 $q$ を minimax 符号と呼ぶが, 情報理論的な考察により, minimax 符号は, $w$ のベイズ符号 $p_w$ のベイズリスク $R_N(w, p_w)$ を最大にする $w = w^*$(the least favorable prior と呼ぶ) によるベイズ符号 $p_{w^*}$ に一致することが分かっている [DL80, MIH91]. また, このことが知られるより早く, $\theta$ と $x^N$ の間の相互情報量 ($R_N(w, p_w)$ に一致する) を最大にする $w$ の近似形が漸近的にジェフリーズのプライア $w_J(\theta) = c(\det |g_{ij}(\theta)|)^{1/2}$ ($g_{ij}$ はフィッシャー情報行列) [Jeff61] に近付くことから, $w_J$ が (他の事前分布を比較する基準という意味で) リファレンスプライアであると予想されていた [Ber79]. さらに最近になり, Clarke & Barron がベイズリスク $R_N(w, p_w)$ を漸近的に評価し, これを最大にする $w$ がユニークに $w_J$ になる, すなわち, Jeffreys のプライアが漸近的な意味で the least favorable prior となることを証明した [CB92]. ここにいたり, ベイズ統計学ははじめて客観的なリファレンスプライアを手にしたことになる. ただし, 符号長を最小にするという目的に限定されてはいる. しかし次のように, それは確率分布間の距離として最も一般性が高い Kullback-Leibler 情報量を最小にすることと等価である. すなわち, $q(x_{i+1}|x^i) \equiv \frac{q(x^{i+1})}{q(x^i)}$ とおくと, $q$ の平均冗長度は

$$R_N(\theta, q) = \sum_{i=1}^{N} \int D(p(x_i|\theta)||q(x_i|x^{i-1})) p(x^{i-1}|\theta) dx^{i-1}$$

と書ける. ここで, $D(p||q)$ は $q$ の $p$ に対する KL 情報量である. さらに言えば, $R_N(\theta, q)$ をなるべく小さくする問題は Yamanishi の (log loss に関する)Loss Bound Model[Yam91] と等価である. すなわち, minimax 符号 $p_{w^*}$ は, Loss Bound Model に関する minimax 解にもなっている. 例えば, ベルヌーイ系列の場合 ($x = 1$ or $0$, $p(1|r) = r$) は, $p_{w^*}(1|x^N) = \frac{k+0.5}{N+1}$ ($k$ は $N$ 回中の 1 の回数) となり, ラプラス推定量が minimax な推定量であることが分かる.

一方, MDL 原理からも同じ推定量が導出される [竹内他 93] が, 実はそれは, MDL 原理だけから決定されるのではなく, パラメータ $r$ の値の二進小数展開を符号とするという前提をおいていた (それは, 事前分布を用いる [WF87] の MML 推定で, $r$ のレンジに一様分布を仮定するのと同等と考えられる). このことを, 二段階符号化は最良の圧縮法ではないという事実と併せると, パラメータ推定問題に限っては, MDL 原理は有効でないような印象を受ける.

しかしながらベイズ符号にも, 推定量として見たときには問題点がある. ベルヌーイ系列の場合は, 情報源のクラスが情報源の混合の操作について閉じているため, $p_w$ は再びもとのクラスの要素となったが, それは一般のクラスについては成り立たない. すなわち, ベイズ符号は厳密には推定量とはならないのである. これに対し, MDL 推定量は必ずもとのクラスに属する.

このことと, ベルヌーイ系列の場合には MDL 推定量が minimax ベイズ符号に一致したこととからヒントを得て, 情報源のクラスを一般の指数型分布族に拡張した場合に, minimax ベイズ符号 $p_{w^*}$ と MDL 推定量との関係がどうなっているのかを調べてみた. このとき, 一般に $p_{w^*}$ の像は指数型分布族の要素とはならないので, もとのクラスへの射影をとることにした. これは, 指数型分布族の期待値パラメータ (または $\eta$ 座標. ベルヌーイ系列の場合は $r$ が相当する) の事後分布での期待値で指定されるものと一致する. すなわち, ベルヌーイ系列の場合のラプラス推定量の一般化になっているので, これを一般化ラプラス推定量 ($p_L$ と書く) と呼ぶことにする. また, MDL 推定量のパラメータの符号化をどうするかが問題となるが, ベルヌーイ系列の場合か

ら類推して, $\eta$ 座標の値を符号化すればよいだろうと予想した (ただし, 座標系が直交していない場合には工夫がいる).

結果として, ある条件のもと, $\eta$ 座標をパラメータの符号化に用いる MDL 推定量 ($p_{mdl}$ と書く) は $p_L$ にオーダ $O(\frac{1}{N})$ の項まで一致することが分かった. さらに, これらの推定量は, 指数型分布族の自然パラメータ ($\theta$ 座標) に関してバイアス補正した最尤推定量 ($p_{bc}$ と書く) にオーダ $O(\frac{1}{N})$ の項まで一致することがわかった.

以上ことと $p_{w^*}$ が minimax 符号であることとから, $p_{mdl}$, $p_L$, および $p_{bc}$ は, KL 情報量を基準としたときに, 厳密な意味での推定量としては最良のものであることを示唆していると考えられる.

以下, 2 節は準備にあて, 3 節から 5 節で, 三つの推定量を説明し, 6 節で結果を述べ, 7 節で (概) 証明をつける.

## 2 準備

$p(x|\theta)$ をレンジ $\mathcal{X}$ 上の確率分布とする. すなわち, $p(x) > 0$ かつ $\int_{x\in\mathcal{X}} p(x|\theta)dx = 1$ とする. ただし, $\mathcal{X}$ が可算の場合は積分のかわりに和をとる. ここで, $\theta$ はパラメータで, $\Re^n$ の体積 0 でない部分集合 $\Theta$ の要素とする. また, $S = \{p(x|\theta)|\theta\in\Theta\}$ とおく.

本稿では, $\mathcal{X}^N$ から $S$ への写像 (の列) を推定量とよぶ. したがって, 不偏推定量という言葉は意味をなさない. 本稿では, 例えば $\theta$ について不偏な推定量と言う. $p$ が推定量であるとき, $p$ による $x^N\in\mathcal{X}^N$ の像を $p[x^N]$ と書く (確率密度としての値 $p(x^N)$ と区別するため).

本稿では $S$ として, 次の指数型分布族のみを考える. ここでの定義および, いくつかの性質は [Amari90],[甘利他 93] による.

**定義 1 (指数型分布族)**

$$\begin{aligned} S &= \{p(x|\theta)|\theta\in\Theta\} \\ p(x|\theta) &= \exp\bigl(C(x)+\theta^i F_i(x)-\psi(\theta)\bigr) \end{aligned} \tag{1}$$

ここで, 上下に現れた同じ指標について 1 から $n$ にわたって総和をとるという, いわゆる和の規約を用いている. $\theta$ を自然パラメータ (または $\theta$ 座標) と呼ぶ. また, $\eta_i \equiv E_\theta(F_i(x))$ で定める $\eta$ を期待値パラメータ (または $\eta$ 座標) と呼ぶ. $E_\theta$ は $p(x|\theta)$ での平均を意味するものとする. 例えば $\Re^+$ 上の分布 $p(x|\theta) = \theta\exp(-\theta x)$ は指数型分布族の例である (以後これは指数分布と呼ぶ). また, 有限離散確率分布, 正規分布, ポアソン分布などは指数型分布族である.

$F_i(x)$ を第 $i$ 成分とするベクトルを新しい確率変数と考え, これを $x_i$ と書き直し, さらにベクトル $x$ の確率密度を $d\mu(x) = \exp(C(x))dx$ という測度のもとで定義すれば, (1) は $p(x|\theta) = \exp\left(\theta^i x_i - \psi(\theta)\right)$ と書ける. 以下ではこの形で議論する.

後で必要になる性質を述べておく. まず,

$$\begin{aligned} \frac{\partial\psi}{\partial\theta^i} &= \eta_i \\ \frac{\partial^2\psi}{\partial\theta^i\partial\theta^j} &= g_{ij} \\ \frac{\partial^3\psi}{\partial\theta^i\partial\theta^j\partial\theta^k} &= T_{ijk} \end{aligned}$$

が成り立つ. ただし, $g_{ij}$ は $\theta$ 座標に関するフィッシャー情報行列で, $g_{ij} = E_\theta((x_i-\eta_i)(x_j-\eta_j))$ が成り立つ. また, $T_{ijk} = E_\theta((x_i-\eta_i)(x_j-\eta_j)(x_k-\eta_k))$ である. $g_{ij}$ は正定値であるから, $\psi(\theta)$ は下に凸, 従って, $\theta^i x_i - \psi(\theta)$ は $\theta$ の関数として上に凸である. また, $\eta$ 座標に関するフィッシャー情報行列を $g^{ij}$ と書くが, このとき $g_{ij}g^{jk} = \delta_i^k$ が成り立つ. すなわち, $g_{ij}$ と $g^{ij}$ は互いに逆行列になっている. また, 次が成り立つ.

$$\frac{\partial\ln\sqrt{\det|g_{ij}|}}{\partial\theta^i} = \frac{1}{2}T_{ijk}g^{jk}$$

これは, 微分幾何で良く知られた公式, $\frac{\partial}{\partial\theta^i}\ln\sqrt{\det|g_{ij}|} = \Gamma^{(0)}_{ijk}g^{jk}$ から出る. ここで $\Gamma^{(0)}$ は $g_{ij}$ に関するリーマン接続で, 指数型分布族については, $\Gamma^{(0)}_{ijk} = \frac{1}{2}T_{ijk}$ である.

## 3 ベイズ符号

ジェフリーズのプライアを $w_J(\theta) \equiv c\sqrt{\det|g_{ij}(\theta)|}$ で定義する. $c$ は規格化定数である. ベイズ符号 $p_{w_J}(x^N)$ は,

$$p_{w_J}(x^N) \equiv \int_\Theta p(x^N|\theta)c\sqrt{\det|g_{ij}|}d\theta$$

で定義される. これの平均冗長度は, 漸近的に $R(\theta, p_{w_J}) = \frac{n}{2}\log\frac{N}{2\pi e} - \log c + o(1)$ となることが分かっている [CB92]. $\theta$ に依存しないことに注意.

逐次型ベイズ符号 $\frac{p_{w_J}(x^{N+1})}{p_{w_J}(x^N)}$ を求めると,

$$p_{w_J}(x_{N+1}|x^N) = \frac{\int p(x_{N+1}|\theta)p(x^N|\theta)\sqrt{\det|g_{ij}|}d\theta}{\int p(x^N|\theta)\sqrt{\det|g_{ij}|}d\theta} \tag{2}$$

となり $c$ がキャンセルする. すると, $c$ が有限でない場合でも計算出来るので (2) を定義として採用する. ただし, この表式の分母が収束することは仮定する. ここで,

$$w_J(\theta|x^N) = \frac{p(x^N|\theta)\sqrt{\det g_{ij}}}{\int p(x^N|\theta)\sqrt{\det g_{ij}}d\theta}$$

とおくと, $p_B^N(x) = \int p(x|\theta)w_J(\theta|x^N)d\theta$ と書ける.

今, $p_B^N(x) \equiv p_{w_J}(x|x^N)$ と書く. $p_B^N$ を $S$ へ射影したときの射影の足 $p_L^N$ を次で定める.

$$p_L^N \equiv \arg\min_{p\in S} D(p_B^N||p).$$

$p_L^N$ の $\eta$ 座標を $\eta_L$ と書く. 次の命題が成り立つ.

**命題 1** $p(x^N|\theta)\sqrt{\det|g_{ij}|}$ と $\eta_i p(x^N|\theta)\sqrt{\det|g_{ij}|}$ が $\Theta$ 上で積分可能のとき, 次式が成り立つ.

$$(\eta_L)_i = \int_\Theta \eta_i(\theta)w_J(\theta|x^N)d\theta.$$

証明は略す.

# 4 MDL 推定量

ここでは情報源のクラス $\{p(x|\xi)|\xi \in \Xi\}$ について, MDL 推定量を求める. ここでの議論は, 後半で変分法を用いるところ以外は, [Riss78] と [WF87] を参考にしている.

今, $\Xi$ は有界であると仮定し, $\Re^n$ 中での $\Xi$ の体積を $V$ とする. パラメータの符号化は $\Xi$ のなかに高々可算個の量子点をうち, それらに番号をつけることに相当する. 最適な符号化はデータ数に依存するので, 量子点の集合を $\Xi_N$ と書く. 量子点 $\bar{\xi}$ で代表される領域を $r(\bar{\xi})$ とし, その $\Re^n$ 中での体積を $v(\bar{\xi})$ とする. さらに, 量子点 $\bar{\xi}$ に符号長 $l_N(\bar{\xi}) = -\ln \frac{v(\bar{\xi})}{V}$ を割りふるとする[1] . このような符号化全てがつくる集合を $\mathcal{L}_N$ とする. MDL 推定量を定義しよう.

**定義 2 (座標系 $\xi$ に関する MDL 推定量)** 次の様に $\xi_{mdl}$ を定義し, $x^N$ から $p_{mdl}[x^N] \equiv p(\cdot|\xi_{mdl}(x^N))$ への写像を座標系 $\xi$ に関する *MDL* 推定量と呼ぶ.

$$\begin{aligned}
\hat{\xi}_{l_N} &\equiv \arg\min_{\bar{\xi}\in\Xi_N}(-\ln p(x^N|\bar{\xi}) + l_N(\bar{\xi})) \\
l_N^* &\equiv \arg\min_{l\in\mathcal{L}_N} E_\xi(-\ln p(x^N|\hat{\xi}_l) + l_N(\hat{\xi}_{l_N})) \\
\xi_{mdl}(x^N) &\equiv \arg\min_{\bar{\xi}}(-\ln p(x^N|\bar{\xi}) + l_N^*(\bar{\xi}))
\end{aligned}$$

後で分かるように $l_N^*$ は $\xi$ に依存せずに定まる. $l_N^*$ を求めるが, まず $r(\bar{\xi})$ は点 $\bar{\xi}$ におけるフィッシャー情報行列の主軸方向を向いた直方体であるとしてよい (そうでないと結果的に記述長が長くなる. [WF87] 参照). 点 $\xi$ における $\alpha$ 番目の主軸方向を向いた単位接ベクトルを $X_\alpha$ とし, 対応する固有値を $\lambda_\alpha$ とする. $\xi$ の近傍にある $X_\alpha$ 方向に直交する直方体の辺の長さが平均 $d_\alpha$ とすると, 点 $\xi$ の近傍においては, $v(\xi) = \prod_{\alpha=1}^n d_\alpha(\xi)$ となる. ここで, $\hat{\xi}_l \in r(\bar{\xi})$ であるという条件のもとでの全記述長の期待値 $DL(\bar{\xi})$ を求めよう. $P(\hat{\xi})$ で最尤推定量 $\hat{\xi}$ の確率分布を表し, $r(\bar{\xi})$ 中ではそれを一様分布で近似する. また, $r(\bar{\xi})$ 中では $X_\alpha, \lambda_\alpha$ は $\bar{\xi}$ での値で近似する. すると,

$$DL(\bar{\xi}) \sim \frac{N}{24}\sum_\alpha \lambda_\alpha d_\alpha^2 - \ln v(\bar{\xi}) + C \tag{3}$$

となる. $\hat{\xi}_l \in r(\bar{\xi})$ となる確率は $P(\bar{\xi})v(\bar{\xi})$ で近似出来るから全記述長 $L$ は,

$$\begin{aligned}
L &\sim \sum_{\bar{\xi}} \big(DL(\bar{\xi})v(\bar{\xi}) + C\big)P(\bar{\xi}) \\
&\sim \int_\Xi P(\hat{\xi})\big(\frac{N}{24}\sum_\alpha(\lambda_\alpha(\hat{\xi})d_\alpha(\hat{\xi})^2 - \ln d_\alpha(\hat{\xi})\big)d\hat{\xi} + C
\end{aligned}$$

これを最小化する $d_\alpha$ を求めるために $d(\hat{\xi})_\alpha \to d(\hat{\xi})_\alpha + \delta(\hat{\xi})_\alpha$ と変分をとると,

$$\delta L = \int_\Xi P(\hat{\xi})\sum_\alpha\Big(\frac{N\lambda_\alpha(\xi)d_\alpha}{12} - \frac{1}{d_\alpha(\hat{\xi})}\Big)\delta(\hat{\xi})_\alpha d\hat{\xi}$$

となる. $p(x^N|\xi) > 0$ より $\forall\xi \in \Xi$ $P(\hat{\xi}) > 0$ であるから, 様々な変分に対し $\delta L = 0$ であるためには, 各 $\alpha$ とすべての $\hat{\xi}$ に対し, $\frac{N\lambda_\alpha(\xi)d_\alpha}{12} - \frac{1}{d_\alpha(\hat{\xi})} = 0$ でなければならない. よって, $d_\alpha = \sqrt{\frac{12}{N\lambda_\alpha}}$. すなわち

$$v(\xi) = \prod_{\alpha=1}^n d_\alpha(\xi) = 12^{\frac{n}{2}}N^{-\frac{n}{2}}(\det|g^{ab}|)^{-\frac{1}{2}}$$

となる. $g^{ab}$ は $\xi$ のフィッシャー情報行列. これから,

$$l^*(\bar{\xi}) \sim \frac{1}{2}\ln\det|g(\bar{\xi})^{ab}| + \frac{n}{2}\ln N + \ln V - \frac{n\ln 12}{2}$$

と求められ, 次を得る.

$$\xi_{mdl} = \arg\min_{\bar{\xi}\in\Xi_N}\big(-\ln p(x^N|\bar{\xi}) + \frac{1}{2}\ln\det|g^{ab}(\bar{\xi})|\big).$$

ここで, $v(\xi) \propto (\det|g^{ab}|)^{-\frac{1}{2}}$ に注意しよう. $\xi$ における量子点の密度を $\rho(\xi)$ と書くと, $\rho \propto (\det|g^{ab}|)^{\frac{1}{2}}$ である. $(\det|g^{ab}|)^{\frac{1}{2}}d\xi$ は多様体 $S$ の体積素片であるから, 量子点の分布は多様体 $S$ 上の一様分布になっている[2] .

以下では, 量子化誤差を無視する. また, $\Xi$ は有界と仮定したが, その仮定も外す. すなわち, 次の式を MDL 推定量の定義として採用し, その背景に定義 2 のような解釈が存在すると考える.

$$\hat{\xi}_{mdl} \equiv \arg\min_{\xi\in\Xi}\big(-\ln p(x^N|\xi) + \frac{1}{2}\ln\det|g^{ab}(\xi)|\big).$$

# 5 バイアス補正した最尤推定量

バイアス補正した最尤推定量を定義する. 指数型分布族について, $p(x^N|\theta) = \exp(N(\theta^i\bar{x}_i - \psi(\theta)))$ と書ける. ただし, $\bar{x} = \frac{1}{N}\sum_{t=1}^N x_t$ である. $x_t$ は $t$ 番目の観測値のベクトルを表す (第 $t$ 成分ではない). このとき, $\eta$ 座標の最尤推定量を $\hat{\eta}$ とすると, $\hat{\eta}_i = \bar{x}_i$ となる [Amari90]. $\eta$ 座標の定義から, $E_\eta(\hat{\eta}) = \eta$ である. すなわち $\hat{\eta}$ は不偏推定量である.

次に, 他の座標 $u$ について考える. 今, 真のパラメータを $u_0, \eta_0$ とする. $\hat{u} = u(\hat{\eta})$ をテーラー展開すると,

$$\begin{aligned}
\hat{u}_i - u_{0i} &= \Big(\frac{\partial u_i}{\partial\eta_k}\Big)_0(\hat{\eta}_k - \eta_{0k}) \\
&+\frac{1}{2}\Big(\frac{\partial^2 u_i}{\partial\eta_l\partial\eta_k}\Big)_0(\hat{\eta}_k - \eta_{0k})(\hat{\eta}_l - \eta_{0l}) + O(||\hat{\eta} - \eta^*||^3).
\end{aligned}$$

両辺の期待値をとると,

$$E_\eta(\hat{u}_i - u_{0i}) = \Big(\frac{\partial^2 u_i}{\partial\eta_l\partial\eta_k}\Big)_0\frac{g_{kl0}}{2N} + O\big(\frac{1}{N\sqrt{N}}\big)$$

となる. すなわち, $\hat{u}$ は $\frac{1}{N}$ オーダのバイアスを含むが,

$$\tilde{u}_i \equiv \hat{u}_i - \frac{\partial^2 u_i}{\partial\eta_l\partial\eta_k}(\hat{\eta})\frac{g_{kl}(\hat{\eta})}{2N} \tag{4}$$

とすると $\tilde{u}$ のバイアスは $O(\frac{1}{N\sqrt{N}})$ となることが知られている [Amari90]. このとき, 写像 $p_{bc} : x^N \mapsto p(x|\tilde{u}) \in S$ を $u$ に関してバイアス補正した最尤推定量と呼ぶ.

[1] この条件が, 座標系 $\xi$ に関する MDL 推定量を特徴づける. 座標系が直交していれば, $\xi_i$ の値の小数展開を符号とすることと等価である. また, ベイズ的に解釈すると, $\Xi \subseteq \Re^n$ 上の一様分布を仮定していることになる.

[2] それは, ジェフリーズのプライア $w_J$ に比例する. もし, ベイズ的な解釈をとり, $\Xi$ 上にジェフリーズの事前分布を仮定すると, 最尤推定量を得る.

## 6 主要結果

この節では本稿の主要結果を述べる. そのために必要となる仮定を箇条書しよう.

**仮定 1** $\hat{\theta}$ は $\Theta$ の内点である.

**仮定 2** $N$ が, ある有限の値 $N_1$ 以上のとき, $p(x^N|\theta)\sqrt{\det|g_{ij}|}$ と $\eta(\theta)p(x^N|\theta)\sqrt{\det|g_{ij}|}$ は $\Theta$ で積分可能である.

**仮定 3** $T_{ijk}g^{jk}\cdot p(x^N|\theta)\sqrt{\det|g_{ij}|}$ は, $N$ がある有限の値 $N_2$ 以上のとき, $\Theta$ で積分可能である.

**仮定 4** $\Theta$ は有限個の凸集合に分解出来る.

以下で $\hat{T}_{ijk}$, $\hat{g}^{jk}$ は $\hat{\eta}$ での値を意味するものとする. $p_L$, $p_{mdl}$, $p_{bc}$ を評価し, 以下の三つの結果を得た.

**補題 1 (一般化ラプラス推定量)** 指数型分布族 $S$ と観測データ $x^N$ について, 仮定 *1, 2, 3, 4* が成り立つとき, 一般化ラプラス推定量の像 $p_L[x^N]$ の $\eta$ 座標の値を $\eta_L$ とすると次が成り立つ.

$$(\eta_L)_i = \bar{x}_i + \frac{\hat{T}_{ijk}\hat{g}^{jk}}{2N} + O(\frac{\sqrt{\ln N}}{N\sqrt{N}})$$

**補題 2 ($\eta$ 座標に関する MDL 推定量)** $\eta$ 座標に関する *MDL* 推定量の像 $p_{mdl}[x^N]$ の $\eta$ 座標の値を $\eta_{mdl}$ とすると, 次が成り立つ.

$$(\eta_{mdl})_i = \bar{x}_i + \frac{\hat{T}_{ijk}\hat{g}^{jk}}{2N} + O(\frac{1}{N^2})$$

**補題 3 ($\theta$ 座標に関するバイアス補正した最尤推定量)** $\theta$ 座標に関するバイアス補正した最尤推定量の像 $p_{bc}[x^N]$ の $\eta$ 座標の値を $\eta_{bc}$ とすると, 次が成り立つ.

$$(\eta_{bc})_i = \bar{x}_i + \frac{\hat{T}_{ijk}\hat{g}^{jk}}{2N} + O(\frac{1}{N^2})$$

以上の補題において, $O(\frac{1}{N^2})$ などの意味は, $\bar{x}$ を固定して $N\to\infty$ としたときのオーダである. 補題 1, 2, 3 を併せて, 次の定理を得る.

**定理 1 ($p_{mdl}, p_L, p_{bc}$)** 指数型分布族 $S$ と観測データ $x^N$ について, 仮定 *1, 2, 3, 4* が成り立つとき, $\eta$ 座標に関する *MDL* 推定量の像 $p_{mdl}[x^N]$ と, 一般化ラプラス推定量の像 $p_L[x^N]$ と, $\theta$ 座標に関するバイアス補正した最尤推定量の像 $p_{bc}[x^N]$ との座標の値の差は $O(\frac{\sqrt{\ln N}}{N\sqrt{N}})$ である.

具体例を見てみよう.

**例 1 (指数分布 $p(x|\xi)=\xi\exp(-\xi x)$ $(\xi>0, x>0)$)** この場合, $\theta=-\xi$, $\psi=-\ln(-\theta)$, $\Theta=\Re^-$, $\eta=\frac{-1}{\theta}$ である. よって, $g_{11}=\frac{\partial\eta}{\partial\theta}=\frac{1}{\theta^2}$.

まず一般化ラプラス推定量を求める. $p(x^N|\theta)\sqrt{g(\theta)}=(-\theta)^{N-1}\exp(N\theta\bar{x})$. これから $\int_0^{-\infty}p(x^N|\theta)\sqrt{g(\theta)}d\theta=\frac{\Gamma(N)}{(\bar{x}N)^N}$ と $\int_0^{-\infty}\eta p(x^N|\theta)\sqrt{g(\theta)}d\theta=\frac{\Gamma(N-1)}{(\bar{x}N)^{N-1}}$ を得る. よって, $\eta_L=\frac{N\bar{x}}{N-1}$ となる *(*これは $N=1$ のときだけ収束しない*)*.

次に $\eta$ 座標に関する *MDL* 推定量を求める. $g^{11}=\eta^{-2}$ より, 全記述長は $-\ln p(x^N|\eta)-\ln\eta=(N-1)\ln\eta+\frac{N\bar{x}}{\eta}$ これを微分して 0 とおいて解くと, $\eta_{mdl}=\frac{N\bar{x}}{N-1}$ となり, この場合厳密に $\eta_L$ に一致する.

次にこれらを $\theta$ で表現すると, $\theta_{mdl}=\frac{N-1}{N\bar{x}}$ となる. これの期待値をとると, $\int\theta_{mdl}(\bar{x})(-\theta)^N\exp(-\theta N\bar{x})dx^N=\theta$ となり, $\theta$ に関して不偏であることが確かめられる.

**例 2 (正規分布)** $p(z|\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp(\frac{-(z-\mu)^2}{2\sigma^2})$ であるが, $x_1=z$, $x_2=z^2$ というベクトル値の確率変数を定義すると, $\theta^1=\frac{\mu}{\sigma^2}$, $\theta^2=-\frac{1}{2\sigma^2}$, $\psi(\theta)=-\frac{(\theta^1)^2}{4\theta^2}+\frac{1}{2}\log\left(-\frac{\pi}{\theta^2}\right)$ として, 指数型分布族になる. ただし $-\infty<\theta^1<\infty, \theta^2<0$ とする. $\det|g_{ij}|=\frac{-2}{(\theta^2)^3}=2\sigma^6$ となる. $\mu$ の推定については, 一般化ラプラス推定量と *MDL* 推定量はいずれも最尤推定と一致するので, $\bar{x_1}=\bar{z}=0$ の場合を考えれば十分. そこで, 座標 $\sigma^2$ の推定だけを考える. 煩雑なので結果だけ述べると, $\hat{\sigma^2}_L=\frac{(N+1)v^2}{N-2}$, $\hat{\sigma^2}_{mdl}=\frac{Nv^2}{N-3}$. となり, その差は $O(\frac{1}{N^2})$ である. ただし, $v^2=N^{-1}\sum_{t=1}^N z_t^2$. いわゆる不偏分散 *($\sigma^2$ に関して不偏な推定量)* は $\frac{Nv^2}{N-1}$ である.

## 7 補題の証明

この節では補題の (概) 証明を与える.

**(補題 1 の概証明)** $p(x^N|\theta)\sqrt{g}=\exp\left(N(\theta^i\bar{x}_i-\psi(\theta))\right)\sqrt{g}$ を $\theta^k$ で偏微分すると ($g\equiv\det|g_{ij}|$ とおいた),

$$\begin{aligned}&\frac{\partial}{\partial\theta^k}\left(\exp\left(N(\theta^i\bar{x}_i-\psi(\theta))\right)\sqrt{g}\right)\\ =\;& N(\bar{x}_k-\eta_k)p(x^N|\theta)\sqrt{g}+\frac{1}{2}T_{kij}g^{ij}p(x^N|\theta)\sqrt{g}\end{aligned}$$

となる. $N$ が十分大きいとき, 仮定 2,3 より右辺は全て積分可能, よって左辺も積分可能. よって, 両辺を $\Theta$ にわたって積分して変形すると, 次を得る.

$$\tilde{\eta}_k=\bar{x}_k+\frac{1}{2N}\int T_{kij}g^{ij}w_N(\theta)d\theta-\frac{1}{N}\int\frac{\partial w_N(\theta)}{\partial\theta^k}d\theta.$$

ただし, $w_N(\theta)$ は $\theta$ の事後分布. そして,

$$\int T_{kij}g^{ij}w_N(\theta)d\theta = \hat{T}_{kij}\hat{g}^{ij}+O(\frac{\sqrt{\ln N}}{\sqrt{N}}) \quad (5)$$

$$\int\frac{\partial w_N(\theta)}{\partial\theta^k}d\theta = O(\exp(-CN)) \quad (6)$$

が示せる. (5) では $\hat{\theta}$ の周りで正規分布近似を用いる. (6) ではフビニの定理を用い, 境界上の積分に帰着させる. これにより, 補題の主張を得る. (概証明終わり)

**(補題 2 の証明)** $\eta$ 座標に関する MDL の場合の全記述長は,

$$\begin{aligned}&-N(\theta^i\bar{x}_i-\psi(\theta))+\ln\sqrt{\det|g^{ij}|}\\ =\;& -N(\theta^i\bar{x}_i-\psi(\theta))-\ln\sqrt{\det|g_{ij}|}.\end{aligned}$$

右辺を $\eta_l$ で微分して 0 とおくと,

$$-g^{il}(\bar{x}_i - \eta_i) - \frac{1}{2N} T_{ijk} g^{jk} g^{il} = 0$$

両辺に $g_{kl}$ をかけて縮約をとり, さらに変形すると,

$$\eta_k = \bar{x}_k + \frac{1}{2N} T_{ijk} g^{jk}.$$

両辺をテーラー展開してこの解を求めると,

$$(\eta_{mdl})_k = \bar{x}_k + \frac{1}{2N} \hat{T}_{ijk} \hat{g}^{jk} + O(\frac{1}{N^2})$$

を得る. (証明終わり)

**(補題 3 の証明)**

(4) で $u$ に $\theta$ を代入すると,

$$(\theta_{bc})^i - \hat{\theta}^i = -\frac{\partial^2 \theta^i}{\partial \eta_l \partial \eta_k}(\hat{\eta}) \frac{g_{kl}(\hat{\eta})}{2N} = -\frac{\partial g^{ik}}{\partial \eta_l}(\hat{\eta}) \frac{g_{kl}(\hat{\eta})}{2N}$$

よって,

$$\begin{aligned} & (\eta_{bc})_j - \hat{\eta}_j \\ = & -\frac{\partial \eta_j}{\partial \theta^i} \frac{\partial g^{ik}}{\partial \eta_l}(\hat{\eta}) \frac{g_{kl}(\hat{\eta})}{2N} + O(\frac{1}{N^2}) \\ = & -g_{ji}(\hat{\eta}) \frac{\partial g^{ik}}{\partial \eta_l}(\hat{\eta}) \frac{g_{kl}(\hat{\eta})}{2N} + O(\frac{1}{N^2}). \end{aligned} \quad (7)$$

ここで, $g_{ji} g^{ik} = \delta^k_j$ を $\eta_l$ で微分して,

$$-g_{ji} \frac{\partial g^{ik}}{\partial \eta_l} = \frac{\partial g_{ji}}{\partial \eta_l} g^{ik} = \frac{\partial g_{ji}}{\partial \theta^m} \frac{\partial \theta^m}{\partial \eta_l} g^{ik} = T_{jmi} g^{ml} g^{ik}$$

よって,

$$-g_{ji} \frac{\partial g^{ik}}{\partial \eta_l} g_{kl} = T_{jmi} g^{ml} g^{ik} g_{kl} = T_{jmi} g^{ml} \delta^i_l = T_{jmi} g^{mi}$$

これと (7) とから補題の主張を得る. (証明終わり)

## 8 むすび

定理 1 の主張は, グローバルに定まる推定量 $p_L$ とローカルに定まる推定量 $p_{mdl}$, $p_{bc}$ とを関連づけるものである. 特に, $p_{w_J}$ と $p_L$ の差が問題にならない程度だと仮定すれば, $p_L$, $p_{mdl}$, $p_{bc}$ いずれもが, KL 情報量の意味で最適に近い性能をもった推定量であると考えられる.

すると, $p_L$ と $p_{mdl}$ の関係からは, 次ような考察が出来る. MDL 推定量は最尤法にペナルティ $l(\eta)$ をつけたものと解釈出来るが, $\eta$ 座標に関する MDL 推定量では, $l(\eta) = \frac{1}{2} \ln \det |g^{ij}|$ であった. すなわち, 確率変数の期待値が少し変わるときに, 確率分布自体が大きく変わる様な領域は, KL の意味では推定しにくいことを表していると考えられる.

また, $p_L$ と $p_{bc}$ の関係は, $\theta$ 座標について不偏な推定量が, KL の意味で良い推定量になるということを主張していると考えられる. $\theta$ 座標は, 1− 接続に関するアファインパラメータという幾何学的な意味を持っているが, これがどのように関係しているのかは興味深い問題である.

こうした解釈の他に, 定理 1 は, ほぼ同一の推定量を求めるための異なる三つの方法を与えており, 計算時間の節約等に役立つ可能性がある.

もとより, 以上の考察に意味をもたせるためには, $p_{w_J}$ と $p_L$ の差を評価すること, また, $p_L$, $p_{mdl}$, $p_{bc}$ らの性能を直接評価することは重要な課題である.

**謝辞** 本研究をはじめるヒントを与えてくれた長岡浩司氏, MDL 推定量の導出法の不備を指摘し, またベイズ符号について詳しく教えてくれた川端勉氏, 議論し, 有益な助言を与えてくれた安倍直樹氏, 議論してくれた岡村利彦氏, 久保克維氏に感謝します. また, 日頃御指導頂く中村勝洋氏に感謝します.

## 参考文献

[Amari90] Amari, S. (1990). Differential-geometrical methods in statistics (2nd pr.), Lecture Notes in Statistics, Vol.28, *Springer-Verlag.*

[BC91] Barron, A. & Cover, T. (1991). Minimum complexity density estimation. *IEEE trans. on IT*, Vol. 37, No. 4, July 1991.

[Ber79] Bernardo, J. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B.* 41. pp.113-147.

[CB92] Clarke, B. & Barron, A. (1992). Jeffreys prior is asymptotically least favorable under entropy risk. *Submitted to the JSPI.*

[DL80] Davison, L. & Leon-Garcia, A. (1980). A source matching approach to finding minimax codes. *IEEE trans. on IT*, Vol. 26, No. 2, Mar 1980.

[Jeff61] Jeffreys, H. (1961). Theory of probability, 3rd ed., Univ. of California Press, Barkeley, Cal.

[MIH91] Matsushima, T., Inazumi, H., & Hirasawa, S. (1991). A class of distortionless codes designed by Bayes dicision theory. *IEEE Trans. on IT.* Vol. 37. No. 5, pp1288-1293.

[Riss78] Rissanen, J. (1978). Modeling by shortest data description. *Automatica.* 14, pp. 465-471.

[Riss87] Rissanen, J. (1987). Stochastic complexity. *J. Roy. Statis. Soc. B.*49. No.3 pp.223-239 & 252-265.

[WF87] Wallace, C.S. & Freeman, P.R. (1987). Estimating and inference by compact coding. *J. Roy. Statist. Soc. B.* 49. No.3 pp.240-265.

[Yam91] Yamanishi, K. (1991). A loss bound model for online stochastic prediction strategies. *Proc. of COLT '91*, pp. 290-302.

[Yam92] Yamanishi, K. (1992). A learning criterion for stochastic rules. *Machine Learning, a special issue for COLT '90*, 9(2/3).

[甘利他 93] 甘利 俊一, 長岡 浩司. (1993). 情報幾何の方法 (岩波講座, 応用数学). 岩波書店.

[竹内他 93] 竹内 純一, 安倍 直樹. (1993). Laplace 型推定量の確率的 PAC 学習モデルによる性能評価. 信学技報. IT92-128. 1993-03. pp. 1-6.

[竹内 93] 竹内 純一. (1993). MDL 推定量の Kullback-Leibler 情報量に関する収束速度について. SITA93 予稿集. pp. 161-164.