（3）「統計的推測理論とその情報論的側面」に関する研究報告

A Class of Robust Principal Component Vectors

The Institute of Statistical Mathematics
Hidehiko Kamiya

1 Problem

Suppose we are given an i.i.d. sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from $N_p(\mathbf{0}, \Sigma)$. We consider the problem of estimating the first principal component vector γ_1, i.e., the eigenvector of Σ corresponding to the largest eigenvalue.

2 Estimator

We propose an estimator of γ_1 defined in the following way.

Let

$$z(\gamma, \boldsymbol{x}) = \frac{1}{2}\left\{\|\boldsymbol{x}\|^2 - (\gamma^T \boldsymbol{x})^2\right\}$$

for $\gamma \in \mathcal{S}^{p-1}$ (the unit sphere in $\mathcal{R}^p$) and $\boldsymbol{x} \in \mathcal{R}^p$. Then, for a nondecreasing, concave function ρ satisfying $\rho(0) = 0$, $\psi(0) = 1$, $\psi := \partial\rho/\partial z$, we define a functional $T_\rho(G)$ of distributions G on $\mathcal{R}^p$ as follows:

$$T_\rho(G) = \arg \min_{\gamma \in \mathcal{S}^{p-1}} L_G(\gamma),$$

where

$$L_G(\gamma) = E_G\left[\rho\left\{z(\gamma, \boldsymbol{x} - \boldsymbol{\mu}_G)\right\}\right]$$

and $\boldsymbol{\mu}_G = E_G(\boldsymbol{x})$.

Now, denoting the empirical distribution of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ by $\hat{F}_n$, we define an estimator $\hat{\gamma}_*$ of γ_1 by

$$\hat{\gamma}_* = T_\rho(\hat{F}_n).$$

Examples of ρ : (i) $\rho_0(z; \beta, \eta) = -\frac{1+\exp(-\beta\eta)}{\beta} \log \frac{1+\exp\{-\beta(z-\eta)\}}{1+\exp(\beta\eta)}$, $\beta > 0$, $\eta > 0$; (ii) $\rho_1(z) = \lim_{\beta \to 0} \rho_0(z; \beta, \eta) = \lim_{\eta \to \infty} \rho(z; \beta, \eta) = z$; (iii) $\rho_2(z; \eta) = \lim_{\beta \to \infty} \rho_0(z; \beta, \eta) = \min\{z, \eta\}$, $\eta > 0$.

The choice $\rho(z) = \rho_0(z; \beta, \eta)$ leads to Xu and Yuille's self-organizing rule for robust principal component analysis ([2]), whereas $\rho(z) = \rho_1(z)$ yields the classical estimator.

3 Main results

Theorem 3.1 *Functional T_ρ is Fisher consistent:*

$$T_\rho\{N_p(\mathbf{0}, \Sigma)\} = \gamma_1.$$

The influence function $IF\{\boldsymbol{x}; T_\rho, N_p(\mathbf{0}, \Sigma)\} := \frac{d}{d\epsilon} T_\rho \{(1-\epsilon)N_p(\mathbf{0}, \Sigma) + \epsilon\delta_{\boldsymbol{x}}\}\Big|_{\epsilon=0+}$, where $\delta_{\boldsymbol{x}}$ is the point mass 1 at $\boldsymbol{x} \in \mathcal{R}^p$, is given as follows.

Theorem 3.2

$$IF\{\boldsymbol{x}; T_\rho, N_p(\mathbf{0}, \Sigma)\} = \psi\left\{\frac{1}{2}\|\boldsymbol{a}_2(\boldsymbol{x})\|^2\right\} a_1(\boldsymbol{x}) \sum_{j=2}^{p} \frac{\lambda_j a_j(\boldsymbol{x})}{\lambda_j^*(\lambda_1 - \lambda_j)} \boldsymbol{\gamma}_j,$$

where $a_1(\boldsymbol{x}) = \boldsymbol{\gamma}_1^T(\boldsymbol{x} - \boldsymbol{\mu})$, $\boldsymbol{a}_2(\boldsymbol{x}) = \{a_2(\boldsymbol{x}), \ldots, a_p(\boldsymbol{x})\}^T = \Gamma_2^T(\boldsymbol{x} - \boldsymbol{\mu})$; $\Sigma = \Gamma\Lambda\Gamma^T$, $\Gamma = (\boldsymbol{\gamma}_1, \Gamma_2) = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_p) \in \mathcal{O}(p)$(the orthogonal group), $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, $\lambda_1 > \cdots > \lambda_p > 0$; $\lambda_j^* = E\left[\psi\left\{\frac{1}{2}\|\boldsymbol{a}_2(\boldsymbol{x})\|^2\right\} a_j(\boldsymbol{x})^2\right]$, $j = 2, \ldots, p$.

This explicit expression of $IF\{\boldsymbol{x}; T_\rho, N_p(\mathbf{0}, \Sigma)\}$ allows us to investigate robustness and efficiency of our estimator.

From the point of view of robustness, we propose taking ρ satisfying

$$\sup_{z \geq 0}\left\{z^{1/2}\frac{\partial \rho(z)}{\partial z}\right\} < \infty.$$

This condition is obtained by considering gross-error sensitivity of T_ρ. Note that ρ_0 and ρ_2 satisfy this condition, while ρ_1, the classical case, does not.

On the other hand, concerning efficiency, we have by Theorems 3.1 and 3.2 that

Corollary 3.1

$$n^{1/2}(\hat{\boldsymbol{\gamma}}_* - \boldsymbol{\gamma}_1) \to N_p[\mathbf{0}, V\{T_\rho, N_p(\mathbf{0}, \Sigma)\}] \text{ in distribution,}$$

where

$$V\{T_\rho, N_p(\mathbf{0}, \Sigma)\} = \lambda_1 \sum_{j=2}^{p} \frac{\lambda_j^{**}\lambda_j^2}{\lambda_j^{*2}(\lambda_1 - \lambda_j)^2} \boldsymbol{\gamma}_j \boldsymbol{\gamma}_j^T$$

and $\lambda_j^{**} = E\left[\left\{\psi\left(\frac{1}{2}\|\boldsymbol{a}_2(\boldsymbol{x})\|^2\right)\right\}^2 a_j(\boldsymbol{x})^2\right]$, $j = 2, \ldots, p$.

Define the asymptotic relative efficiency of $\hat{\boldsymbol{\gamma}}_* = T_\rho(\hat{F}_n)$ by $\mathrm{Eff}(\hat{\boldsymbol{\gamma}}_*) = \overline{\det} V\{T_{\rho_1}, N_p(\mathbf{0}, \Sigma)\} / \overline{\det} V\{T_\rho, N_p(\mathbf{0}, \Sigma)\}$, where $\overline{\det} V$ denotes the product of nonzero eigenvalues of V. Then

$$\mathrm{Eff}(\hat{\boldsymbol{\gamma}}_*) = \prod_{j=2}^{p} \frac{\lambda_j^{*2}}{\lambda_j \lambda_j^{**}}.$$

Remark 3.1 *Essentially the same argument is possible for general elliptically contoured distributions.*

This talk is based on Kamiya and Eguchi [1].

References

[1] Kamiya, H. and Eguchi, S. (1998). A class of robust principal component vectors. Research Memorandum No.699, The Institute of Statistical Mathematics.

[2] Xu, L. and Yuille, A. L. (1995). Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Trans. on Neural Networks*, **6**, 131–143.

Simultaneous Confidence Intervals for Linear Contrasts of Means in Repeated Measures with Missing Data

東京理科大学.理工　瀬尾隆
University of Toronto Muni S. Srivastava

分散共分散行列が一様な構造を持つ Intraclass correlation model のもとでの繰り返し測定データにおいて，与えられたデータに欠測値が生じた場合の平均に関する同時信頼区間について考える．

問題を定式化するために次のような two-components mixed linear model を考える．

$$x_{ij} = \mu_j + \alpha_i + \varepsilon_{ij}, \quad i = 1, \cdots, n_j,\ j = 1, \cdots p,$$

ここに，α_i と ε_{ij} は互いに独立にそれぞれ平均 0，分散 σ_α^2 の正規分布と平均 0，分散 σ_ε^2 の正規分布に従うものとする．このとき，$\mathrm{E}[x_{ij}] = \mu_j,\ j = 1, \cdots, p$ であり，$\mathrm{Var}(x_{ij}) = \sigma_\alpha^2 + \sigma_\varepsilon^2$, $\mathrm{Cov}(x_{ij}, x_{ij'}) = \sigma_\alpha^2, j \neq j'$ そして，$\mathrm{Cov}(x_{ij}, x_{i'j}) = 0, i \neq i'$ となる．ここで，$n_j = n$ のとき，$\boldsymbol{x}_j = (x_{j1}, \cdots, x_{jp})'$, $j = 1, \cdots, n$ とすると，$\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ は，平均ベクトル$\boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)'$，分散共分散行列$\boldsymbol{\Sigma} = \sigma^2[(1-\rho)\boldsymbol{I} + \rho \boldsymbol{1}\boldsymbol{1}']$ の p 変量正規分布 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ に従う．ただし，$\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$, $0 \leq \rho = \sigma_\alpha^2/\sigma^2 \leq 1$．分散共分散行列 $\boldsymbol{\Sigma}$ がこのような構造をもつようなモデルは，intraclass correlation model と呼ばれ，Bhargava and Srivastava [3] は，ρ が既知の場合の Tukey 法と Scheffé 法 (Miller[4]，Scheffé[6] 参照) を拡張した平均のコントラスト，すなわち $\boldsymbol{a}'\boldsymbol{\mu}$ (ここに $\boldsymbol{a}$ は，$\boldsymbol{a} \in R^p - \{\boldsymbol{0}\}$, $\boldsymbol{a}'\boldsymbol{1} = 0$ を満足するものである) に対する正確な同時信頼区間を与えた．

本報告では，このようなモデルのもとでデータに欠測が生じた場合の平均のコントラストに対する同時信頼区間について議論する．

まず，Rao[5]，Anderson[1] そして Bhargava[2] の中で議論されている単調欠測データ，すなわち，$n_1 \geq n_2 \geq \cdots \geq n_p$ の場合について議論し，同時信頼係数が正確

に $1-\alpha$ となる平均のコントラストである Scheffé 型の同時信頼区間とボンフェロニ型の同時信頼区間を導出した.

さらに, 単調欠測データではなく, 欠測がランダムであるより一般的な欠測データの場合についても議論し, 漸近的な議論であるが, Srivastava and Carter[7] による反復法のアイデアを用いた分散共分散行列の最尤推定量を利用して平均のコントラストに対する同時信頼区間を導出した. 最後に, これらの方法をわかりやすく説明するために実際の繰り返し測定データにこれらの方法を適用した数値例を与えた.

参考文献

[1] Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *J. Amer. Statist. Assoc.*, **52,** 200–203.

[2] Bhargava, R. P. (1962). Multivariate tests of hypothesis with incomplete data. *Tech. Rep.* NO.3, Stanford University.

[3] Bhargava, R. P. and Srivastava, M. S. (1973). On Tukey's confidence intervals for the contrasts in the means of the intraclass correlation model. *J. Royal Statist. Soc.*, **B35,** 147–152.

[4] Miller, R. G. (1966). *Simultaneous Statistical Inference.* McGraw-Hill: New York.

[5] Rao, C. R. (1956). Analysis of dispersion with missing observations. *J. Royal Statist. Soc.*, **B18,** 259–264.

[6] Scheffé, H. (1959). *The Analysis of Variance.* Wiley: New York.

[7] Srivastava, M. S. and Carter, E. M. (1986). The maximum likelihood method for non-response in sample survey. *Survey Methodology,* **12,** 61–72.

INFORMATION CRITERIA IN MODEL SELECTION FOR STOCHASTIC PROCESSES (II)

統計数理研究所 内田 雅之
東京大学・数理 吉田 朋広

1 Introduction

Akaike's information criterion AIC (Akaike(1973, 1974)) is a model evaluation-selection tool in terms of estimating the Kullback and Leibler information of the true model with respect to the fitted model. It can be derived under the assumptions that

(i) the data is a random sample from an unknown distribution,

(ii) estimation is done by the maximum likelihood method, and

(iii) it is carried out in a parametric family of distributions including the true model.

Takeuchi (1976) derived Takeuchi's information criterion TIC from assumptions (i) and (ii) only, relaxing the assumption (iii) when model classes used can be incorrect. Konishi and Kitagawa (1996) proposed generalized information criteria GIC by using assumption (i) and functional-type estimators instead of assumption (ii), and showed the simplifications resulting from assumptions (i) and (ii) only, and then from (i), (ii) and (iii).

In our work, we will propose information criteria for evaluating models constructed by various estimation procedures for stochastic processes. This paper is organized as follows. In Section 2, we state our main results. Using the asymptotic expansion of the distribution of the estimators for stochastic processes, a general theory is developed and two information criteria are proposed.

2 General theory

Let $(\mathcal{X}_T, \mathcal{A}_T)$ be a measurable space for each $T > 0$. Given $(\Omega, \mathcal{F}, P)$ a probability space, let $\boldsymbol{X}_T$ denote a $\mathcal{X}_T$-valued random variable with unknown distribution $Q_T(\cdot) = P(\boldsymbol{X}_T^{-1}(\cdot))$ having probability density function $q_T(\cdot)$ with respect to a reference measure. Let $\hat{\theta}_T : (\mathcal{X}_T, \mathcal{A}_T) \to \Theta$ be a measurable function, where $\Theta \subset \mathbf{R}^p$. The Borel σ-field of $\mathbf{R}^p$ is denoted by $\mathcal{B}^p$. Estimation is done within a parametric family of distributions $\{P_{T,\theta}(\cdot); \theta \in \Theta\}$ with densities $\{f_T(\cdot, \theta); \theta \in \Theta\}$, which may or may not contain $q_T(\cdot)$. The predictive density function $f_T(z, \hat{\theta})$ for a future observation $\boldsymbol{X}_T(\omega') = z$ (for $\omega' \in \Omega$) can constructed by replacing the unknown parameter vector θ by $\hat{\theta}_T$. Let l_T be an $\mathbf{R}^1$-valued function on $\mathcal{X}_T \times \Theta$. Set

$$\begin{aligned}
\bar{Z}_T^{(0)} &:= r_T \left\{ l_T(\boldsymbol{X}_T(\omega), \theta_0) - \int l_T(\boldsymbol{X}_T(\omega'), \theta_0) P(d\omega') \right\}, \\
\bar{Z}_T^{(1)} &:= r_T \left\{ \partial_\theta l_T(\boldsymbol{X}_T(\omega), \theta_0) - \int \partial_\theta l_T(\boldsymbol{X}_T(\omega'), \theta_0) P(d\omega') \right\}, \quad \partial_\theta = \frac{\partial}{\partial \theta},
\end{aligned}$$

where $r_T = 1/\sqrt{T}$.

First of all, we assume that there exists a parameter $\theta_0 \in \Theta$ such that

$$r_T{}^{-1}(\hat{\theta}_T - \theta_0) = \bar{\zeta}_T^{(0)} + o_p(1).$$

Let us prepare some notations. Let $\bar{Z}_T = (\bar{Z}_T^{(0)}, \bar{Z}_T^{(1)}, \bar{\zeta}_T^{(0)})$ and $Z_T = r_T^{-1}\bar{Z}_T$. Divide $\bar{Z}_T$ corresponding to the three subvectors $\bar{Z}_T^{(0)}, \bar{Z}_T^{(1)}$ and $\bar{\zeta}_T^{(0)}$ of $\bar{Z}_T$, i.e.,

$$\Sigma_T = \mathrm{Cov}[\bar{Z}_T^{(0)}, \bar{Z}_T^{(1)}, \bar{\zeta}_T^{(0)}] = \begin{bmatrix} \Sigma_T^{(00)} & (\Sigma_T^{(10)})' & (\Sigma_T^{(20)})' \\ \Sigma_T^{(10)} & \Sigma_T^{(11)} & \Sigma_T^{(12)} \\ \Sigma_T^{(20)} & (\Sigma_T^{(12)})' & \Sigma_T^{(22)} \end{bmatrix} \text{(say)}.$$

We then propose an information criterion based on the asymptotically expectation bias corrected log likelihood as follows:

Information criterion 1 (in the sense of AEU).

$$IC_1(\boldsymbol{X}_T(\omega)) = r_T l_T(\boldsymbol{X}_T(\omega), \hat{\theta}_T(\boldsymbol{X}_T(\omega))) - r_T b_1(\hat{\theta}_T) + o(r_T),$$

where

$$b_1(\theta_0) = \mathrm{tr}\Sigma_T^{(12)}.$$

We also propose another information criterion based on the asymptotically median bias corrected log likelihood as follows:

Information criterion 2 (in the sense of the second order AMU).

$$IC_2(\boldsymbol{X}_T(\omega)) = r_T l_T(\boldsymbol{X}_T(\omega), \hat{\theta}_T(\boldsymbol{X}_T(\omega))) - r_T b_2(\hat{\theta}_T) + o(r_T),$$

where

$$b_2(\theta_0) = -\frac{1}{6} r_T{}^{-1} \lambda_T^{000} \frac{1}{\Sigma_T^{(00)}} + \left[\mathrm{tr}\Sigma_T^{(12)} - \frac{(\Sigma_T^{(10)})'\Sigma_T^{(20)}}{\Sigma_T^{(00)}} \right].$$

References

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle", *2nd International Symposium in Information Theory*, Petrov, B.N. and Csaki, F. eds., Akademiai Kaido, Budapest, 267-281.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Auto. Control*, **AC-19**, 716-723.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection, *Biometrika*, **83**, 875-890.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models, *Mathematical Sciences*, **153**, 12-18, (in Japanese).

分割表における周辺同等性からの隔たりを測る一般化尺度

東京理科大学・理工　槙井 剛志
東京理科大学・理工　富澤 貞男

1.＜2次元分割表における尺度について＞

行と列が順序の付いていない同じ分類からなる$r \times r$正方分割表において(i,j)セル確率をp_{ij}とする$(i=1,2,\ldots,r; j=1,2,\ldots,r)$. 周辺同等(MH)モデルは次のように定義される:

$$p_{i\cdot} = p_{\cdot i} \quad (i=1,2,\ldots,r),$$

ただし，$p_{i\cdot} = \sum_{j=1}^{r} p_{ij}$，$p_{\cdot i} = \sum_{j=1}^{r} p_{ji}$. このモデルはまた次のように表されてもよい：

$$p_{i\cdot}^{c} = p_{\cdot i}^{c} \quad (i=1,2,\ldots,r),$$

ただし，$p_{i\cdot}^{c} = (p_{i\cdot} - p_{ii})/\delta$，$p_{\cdot i}^{c} = (p_{\cdot i} - p_{ii})/\delta$，$\delta = \sum\sum_{i \neq j} p_{ij}$. Tomizawa(1995)はMHモデルからの隔たりを測る2種類の尺度を提案した. 一つは，(i)周辺確率$\{p_{i\cdot}\}$と$\{p_{\cdot i}\}$に基づく尺度であり，もう一つは，(ii)観測値が正方分割表の非対角セルの一つに入るという条件のもとでの条件付き周辺確率$\{p_{i\cdot}^{c}\}$と$\{p_{\cdot i}^{c}\}$に基づく尺度である.

本報告では，Tomizawaの尺度を含む一般化した尺度を導入する. はじめに(i)の場合の一般化した尺度を考える. $p_{i\cdot} + p_{\cdot i} > 0$ $(i=1,2,\ldots,r)$を仮定して，$p_{1(i)} = p_{i\cdot}/(p_{i\cdot} + p_{\cdot i})$，$p_{2(i)} = p_{\cdot i}/(p_{i\cdot} + p_{\cdot i})$，$\pi_i^{*} = (p_{i\cdot} + p_{\cdot i})/2$とおき，尺度を次のように導入する:$\lambda > -1$に対して

$$\Psi_{MH}^{(\lambda)} = \frac{\lambda(\lambda+1)}{2(2^{\lambda}-1)} \left[I^{(\lambda)}\left(\{p_{i\cdot}\};\{\pi_i^{*}\}\right) + I^{(\lambda)}\left(\{p_{\cdot i}\};\{\pi_i^{*}\}\right) \right],$$

ただし

$$I^{(\lambda)}(\{a_i\};\{b_i\}) = \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{r} a_i \left[\left(\frac{a_i}{b_i} \right)^{\lambda} - 1 \right], \quad \Psi_{MH}^{(0)} = \lim_{\lambda \to 0} \Psi_{MH}^{(\lambda)},$$

ここに，$I^{(\lambda)}(\{a_i\};\{b_i\})$は二つの分布$\{a_i\}$と$\{b_i\}$との間のpower-divergence（Read and Cressie,1988）である. 尺度$\Psi_{MH}^{(\lambda)}$は$\lambda = 0, 1$のときがTomizawa(1995)の尺度であり，また次のように表されてもよい：$\lambda > -1$に対して

$$\Psi_{MH}^{(\lambda)} = \frac{\lambda(\lambda+1)}{2^{\lambda}-1} \sum_{i=1}^{r} \pi_i^{*} I_i^{(\lambda)} \left(\left\{ p_{k(i)} \right\}; \left\{ \frac{1}{2} \right\} \right),$$

ただし，

$$I_i^{(\lambda)}(\cdot;\cdot) = \frac{1}{\lambda(\lambda+1)} \left[p_{1(i)} \left\{ \left(\frac{p_{1(i)}}{1/2} \right)^{\lambda} - 1 \right\} + p_{2(i)} \left\{ \left(\frac{p_{2(i)}}{1/2} \right)^{\lambda} - 1 \right\} \right], \quad \Psi_{MH}^{(0)} = \lim_{\lambda \to 0} \Psi_{MH}^{(\lambda)},$$

更に，次のように表されてもよい：$\lambda > -1$ に対して

$$\Psi_{MH}^{(\lambda)} = \sum_{i=1}^{r} \pi_i^{*} \left[1 - \frac{\lambda 2^{\lambda}}{2^{\lambda}-1} H_i^{(\lambda)} \left(\left\{ p_{k(i)} \right\} \right) \right],$$

ただし，

$$H_i^{(\lambda)}(\cdot) = \frac{1}{\lambda} \left[1 - (p_{1(i)})^{\lambda+1} - (p_{2(i)})^{\lambda+1} \right], \quad \Psi_{MH}^{(0)} = \lim_{\lambda \to 0} \Psi_{MH}^{(\lambda)}.$$

ここに,$H_i^{(\lambda)}(\cdot)$ は $\{p_{1(i)}, p_{2(i)}\}$ に対する Patil and Taillie(1982) の degree-λ の diversity index である．また，$0 \leq \Psi_{MH}^{(\lambda)} \leq 1$ であり，任意のλに対して，(1)MH モデルが成り立つための必要十分条件は $\Psi_{MH}^{(\lambda)} = 0$ であり，(2)MH モデルからの隔たりが最大[これは $p_{i\cdot} = 0$ または $p_{\cdot i} = 0 (i = 1, 2, \ldots, r)$ のときと定義する]であるための必要十分条件は$\Psi_{MH}^{(\lambda)} = 1$のときである．次に，(ii) の場合の一般化した尺度として，上記の尺度で $\{p_{i\cdot}\}$ と $\{p_{\cdot i}\}$ をそれぞれ $\{p_{i\cdot}^c\}$ と $\{p_{\cdot i}^c\}$ に置き換えたものを導入する．

2. ＜多次元分割表における尺度について＞

カテゴリに順序の付いていない同じ分類からなる $r^T -$ 分割表 $(T \geq 3)$ において，観測値が分割表 $(i_k = 1, 2, \ldots, r; k = 1, 2, \ldots, T)$ のセル $(i_1, i_2, \ldots, i_T)$ に入る確率を $p_{i_1 i_2 \ldots i_T}$ とし，$X_k (k = 1, 2, \ldots, T)$ を第k変数とする．

この時，次数1の周辺同等 (MH) モデルは次のように定義される:

$$p_i^{[1]} = p_i^{[2]} = \cdots = p_i^{[T]} \quad (i = 1, 2, \ldots, r)$$

ただし，$p_i^{[k]} = \Pr(X_k = i) \quad (i = 1, 2, \ldots, r; k = 1, 2, \ldots, T)$.

Tomizawa(1995) は多次元分割表における MH モデルからの隔たりを測る尺度を提案した．本報告では，Tomizawa の尺度を含む一般化した尺度を導入する．

$\sum_{k=1}^T p_i^{[k]} > 0 \ (i = 1, 2, \ldots, r)$ を仮定して，$p_{k(i)} = p_i^{[k]} / \sum_{u=1}^T p_i^{[u]}$, $\pi_i^{**} = \sum_{k=1}^T p_i^{[k]}$ とおき，尺度を次のように導入する:$\lambda > -1$ に対して

$$\Psi_{[T]}^{(\lambda)} = \frac{\lambda(\lambda+1)}{T(T^\lambda - 1)} \sum_{k=1}^{T} I_k^{(\lambda)} \left(\left\{ p_i^{[k]} \right\}; \{\pi_i^{**}\} \right),$$

ただし，

$$I_k^{(\lambda)}(\cdot;\cdot) = \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{r} p_i^{[k]} \left\{ \left(\frac{p_i^{[k]}}{\pi_i^{**}} \right)^{\lambda} - 1 \right\}, \quad \Psi_{[T]}^{(0)} = \lim_{\lambda \to 0} \Psi_{[T]}^{(\lambda)}.$$

ここに，$I_k^{(\lambda)}(\cdot;\cdot)$ は二つの分布 $\{p_i^{[k]}\}$ と $\{\pi_i^{**}\}$ との間の power-divergence (Read and Cressie, 1988) である．特に，尺度$\Psi_{[T]}^{(\lambda)}$は$\lambda = 0, 1$のときがTomizawa (1995) の尺度である．尺度はまた2次元のときと同様に2種類の別表現が出来る．また，$0 \leq \Psi_{[T]}^{(\lambda)} \leq 1$であり，任意の$\lambda$に対して，(1)MH モデルが成り立つための必要十分条件は $\Psi_{[T]}^{(\lambda)} = 0$ であり，(2)MH モデルからの隔たりが最大[これは 任意の i=1,2,...,r に対して，$p_i^{[k_i]} > 0 (1, 2, \ldots, T$の中にある$k_i$に対して) かつ $p_i^{[k]} = 0$(すべての$k = 1, 2, \ldots, T, k \neq k_i$に対して) のときと定義する]であるための必要十分条件は$\Psi_{MH}^{[T](\lambda)} = 1$のときである．

この尺度の信頼区間，例などを報告し、その考察を行った．

参考文献

[1] Patil and Taillie(1982). *J.Amer.Statist.Assoc.*, **77**,548-561.

[2] Read and Cressie(1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer-Verlag.

[3] Tomizawa,S.(1995). *Journal of the Royal Statistical Society, Series D: The Statistician*, **44**,425-439.

2×2×2不完備分割表における準独立かつ条件付き対称モデル

東京理科大学・理工　鈴木 淳一
東京理科大学・理工　富澤 貞男

母集団における全個体数を推定する方法の1つとして捕獲・再捕獲法がある. ここで, この方法で得られたデータを解析し全個体数を推定する際, そのデータに多項標本モデルを仮定することによって分割表を利用しての解析が考えられる. 今, 3次元の捕獲・再捕獲法を考えると, その観測データは $(2,2,2)$ セルが欠測した $2\times2\times2$ 不完備分割表で表現することができる. 本報告では, この不完備分割表における全個体数の推定法について検討する.

$2\times2\times2$ 不完備分割表において, (i,j,k) セル確率を p_{ijk} ($i=1,2$; $j=1,2$; $k=1,2$) とする. また, (i,j,k) セル観測度数を x_{ijk} ($(i,j,k)\neq(2,2,2)$ に対して) とし, $n=\sum\sum\sum_{(i,j,k)\neq(2,2,2)} x_{ijk}$ とおく. 更に, 母集団における全個体数を N とする. ここに N は未知であり, また, $m^*_{222}=N-n$ は $(2,2,2)$ セルに分類される個体数であり観測不能である. 今, 条件付確率を $Q_{ijk}=p_{ijk}/(1-p_{222}), ((i,j,k)\neq(2,2,2))$ とし, 対応する期待度数を $m_{ijk}=nQ_{ijk}$ とする.

Bishop *et al.* (1975) は, $(i,j,k)\neq(2,2,2)$ に対して $\{m_{ijk}\}$ (または $\{Q_{ijk}\}$) の構造を示す 4 種類の準独立モデル ($H_{(X,Y,Z)}, H_{(XY,Z)}, H_{(XY,YZ)}, H_{(XY,YZ,XZ)}$ と略記する) をあてはめ, 更に $\{p_{ijk}\}$ の構造に三因子交互作用が存在しないことを仮定して, 各モデルの下での $\{m_{ijk}\}$ の推定値を用いて m^*_{222} の推定値を求めることにより, N の推定値 $\widehat{N}$ とその漸近分散の推定値 $\widehat{V}(\widehat{N})$ を求めている. 本報告では, 上記の 4 種類の準独立モデルに更に種々の条件付対称性の構造を入れたモデルを提案し, N の推定値とその漸近分散を導出する.

対数線型モデル

$$\log m_{ijk}=u+u_{1(i)}+u_{2(j)}+u_{3(k)}+u_{12(ij)}+u_{23(jk)}+u_{13(ik)},\qquad (i,j,k)\neq(2,2,2),$$

において, 次の制約を入れた種々のモデルを提案する:

$H_{CS(X,Y,Z)[XY]}: u_{12(ij)} = u_{23(jk)} = u_{13(ik)} = 0, u_{1(i)} = u_{2(i)}.$

[同様に $H_{CS(X,Y,Z)[YZ]}, H_{CS(X,Y,Z)[XZ]}$ を定義する (詳細略)]

$H_{CS(XY,Z)[XY]}: u_{23(jk)} = u_{13(ik)} = 0, u_{1(i)} = u_{2(i)}, u_{12(ij)} = u_{12(ji)}.$

[同様に $H_{CS(YZ,X)[YZ]}, H_{CS(XZ,Y)[XZ]}$ を定義する (詳細略)]

$H_{CS(XY,YZ)[XZ]}: u_{13(ik)} = 0, u_{1(i)} = u_{3(i)}, u_{12(ij)} = u_{23(ji)}.$

[同様に $H_{CS(XY,XZ)[YZ]}, H_{CS(XZ,YZ)[XY]}$ を定義する (詳細略)]

$H_{CS(XY,YZ,XZ)[XY]}: u_{1(i)} = u_{2(i)}, u_{12(ij)} = u_{12(ji)}, u_{23(jk)} = u_{13(jk)}.$

[同様に $H_{CS(XY,YZ,XZ)[YZ]}, H_{CS(XY,YZ,XZ)[XZ]}$ を定義する (詳細略)]

上記各モデルにおいて, Bishop *et al.* (1975) と同様の考えによって, N の推定値 $\widehat{N}$ とその漸近分散の推定値 $\widehat{V}(\widehat{N})$ を導出することができる. (結果の詳細は, 当日報告した.)

さて, 各モデルの下での $\{m_{ijk}\}$ の最尤推定値を用いて求めた N の推定値 $\widehat{N}$ に対する漸近分散の推定値について, たとえば $H_{CS(X,Y,Z)[XY]}$ モデルに対しては $V_{CS(X,Y,Z)[XY]}$ と略記すると, 次の定理が成り立つ.

定理 ある条件の下で次の関係式が成り立つ:

(1) $V_{CS(XY,YZ,XZ)[XZ]} > V_{CS(XY,YZ)[XZ]}$, $V_{CS(XY,YZ,XZ)[XY]} > V_{CS(XY,Z)[XY]}$,

(2) $V_{CS(XY,YZ)[XZ]} > V_{CS(X,Y,Z)[XZ]}$, $V_{CS(XY,Z)[XY]} > V_{CS(X,Y,Z)[XY]}$,

(3) $V_{(XY,YZ,XZ)} \geq V_{CS(XY,YZ,XZ)[XY]}$, $V_{(XY,YZ)} \geq V_{CS(XY,YZ)[XZ]}$,
$V_{(XY,Z)} = V_{CS(XY,Z)[XY]}$, $V_{(X,Y,Z)} = V_{CS(X,Y,Z)[XY]}$.

この定理によって, Bishop *et al.* (1975) が主張した "よりパラメータが少ないモデルの下では漸近分散の推定値もより小さくなる" ということが, さらに確かめられたことになる.

参考文献

Bishop,Y.M.M., Fienberg,S.E., and Holland,P.W. (1975).

Discrete Multivariate Analysis. Cambridge, MA: MIT Press.

Strong Converse to the Quantum Channel Coding Theorem

電気通信大学大学院情報システム学研究科　小川 朋宏, 長岡 浩司

1　はじめに

近年完成された量子通信路符号化定理は，Direct Part と Converse Part の 2 つの部分に分けられる．Direct Part は Hausladen *et al.* の後に，Holevo [3] により (独立に Schumacher and Westmoreland [4]) 完成された．一方で Converse Part は 1970 年台の Holevo の仕事による．彼は Holevo bound と呼ばれる不等式と古典的な Fano の不等式を組み合わせることにより Converse Part を示した．Holevo bound は今日では量子相対エントロピーの単調性の特殊ケースであるとみなすことができる．このようにして得られた Converse Part は，通信路容量を越えたレートでは，誤り確率が漸近的に 0 にはならないということを示しており，Strong Converse との対比において Weak Converse と呼ばれる．Strong Converse の示す所は，通信路容量を越えたレートでは誤り確率が漸近的に 1 になるということである．

最近 Burnashev and Holevo [2] は受信状態が pure state ときに限った状況で，誤り確率の upper bound を導いた．これは古典情報理論における Gallager のランダムコーディング bound に相当するものであり，これから受信状態が pure state ときの量子通信路符号化定理の別証明が得られる．また，これより信頼性関数の重要な lower bound が導かれる．本報告では誤り確率の lower bound を導き，そこから量子通信路に対する強逆定理を示す．これは古典情報理論における Arimoto [1] の bound に相当する．

2　定義と量子通信路符号化定理

$\mathcal{H}$ を信号の物理系を表す Hilbert 空間 ($\dim\mathcal{H}<\infty$) とする．ここでの量子通信路は，写像 $i\in\mathcal{X}\ \mapsto\ \rho_i\,(i=1,\cdots,a)$ であるとする．ただし，$\mathcal{X}=\{1,\cdots,a\}$ は入力アルファベットで，$\rho_i\,(i=1,\cdots,a)$ は密度作用素 (エルミート非負定値でトレースが 1) である．

この通信路を n 回使用して以下のようにメッセージの送信を行う．各メッセージ $k\in\{1,\cdots,M_n\}$ はあらかじめ定められた符号語 $u^k=i_1^k\cdots i_n^k$ に符号化される．符号語の集合 $\mathcal{C}^{(n)}=\{u^1,\cdots,u^{M_n}\}\subset\mathcal{X}^n$ をコードブックと呼ぶ．各メッセージは通信路を通して $\rho_{u^k}=\rho_{i_1^k}\otimes\cdots\otimes\rho_{i_n^k}$ に写像される．復号化は $\mathcal{H}^{\otimes n}$ 上の $\{0,1,\cdots,M_n\}$ 値量子測定 $X^{(n)}=\{X_0,X_1,\cdots,X_{M_n}\}$ により行われる．ここで 0 は復号の失敗を示すものとし，各 X_k は $\mathcal{H}^{\otimes n}$ 上のエルミート非負定値作用素で，$\sum_{k=0}^{M_n}X_k=I$ をみたすものとする．符号化と復号化を合わせた $(\mathcal{C}^{(n)},X^{(n)})$ をメッセージ数 M_n の符号と呼ぶ．また $R_n=\log M_n/n$ を符号 $(\mathcal{C}^{(n)},X^{(n)})$ の伝送レートと呼ぶ．以下では n は省略することがある．

メッセージ k が送信されたもとで，メッセージ l が復号される条件付き確率は $P(k|l)=\mathrm{Tr}\,\rho_{u^l}X_k$ で与えられる．それぞれのメッセージが一様分布で発生するとみなすと，平均誤り確率は

$$\mathrm{Pe}(\mathcal{C},X)=1-\frac{1}{M}\sum_{k=1}^{M}\mathrm{Tr}\,\rho_{u^k}X_k,$$

となる．符号化と復号化を最適化した後の誤り確率を

$$\mathrm{Pe}(M_n,n)=\min_{\mathcal{C}}\min_{X}\mathrm{Pe}(\mathcal{C},X),\tag{1}$$

とおく．通信路容量は $0 \leq R < C$ なるレートならば $\mathrm{Pe}(e^{nR}, n)$ が漸近的に 0 に近づき，$R > C$ なるレートならば $\mathrm{Pe}(e^{nR}, n)$ が漸近的に 0 にならない，というような実数 C であると定義できる．

$\pi = \{\pi_i\}_{i=1}^{a}$ を $\mathcal{X}$ 上の確率分布として，量子相互情報量を

$$I(\pi) = H(\overline{\rho}_\pi) - \sum_{i=1}^{a} \pi_i H(\rho_i),$$

で定義する．ここで，$\overline{\rho}_\pi = \sum_{i=1}^{a} \pi_i \rho_i$ であり，$H(\rho) = -\mathrm{Tr}\,\rho \log \rho$ は von Neumann entropy である．量子通信路符号化定理は $\max_\pi I(\pi)$ が先に定義した通信路容量 C に等しいことを表す．

3 量子通信路に対する強逆定理

レート R の任意の符号 $(\mathcal{C}, X)$ の平均誤り確率に対して，次の不等式が成立する．

$$\mathrm{Pe}(\mathcal{C}, X) \geq 1 - \exp\left[-n\left[-sR + \min_{\pi \in \mathcal{P}_\mathcal{X}} E_0(s, \pi)\right]\right], \quad (-1 < \forall s \leq 0). \tag{2}$$

ただし，

$$E_0(s, \pi) = -\log\left(\mathrm{Tr}\left(\sum_{i=1}^{a} \pi_i \rho_i^{\frac{1}{s+1}}\right)^{s+1}\right). \tag{3}$$

これは古典論における Arimoto の方法とほぼ同様の方針で示されるが，非可換性による慎重な扱いが必要である．

(2) 式において $E_0(s, \pi)$ のグラフを考えることで，$R > C$ の場合には exp の中身が厳密に負である $-1 < s \leq 0$ が存在することが示される．したがって，次の定理が成立する．

Theorem 1 伝送レート $R > C$ の任意の符号 $(\mathcal{C}, X)$ に対して，$\mathrm{Pe}(\mathcal{C}, X)$ は漸近的に指数関数的に *1* に近づく．

古典論においては $E_0(s, \pi)$ は s に関して単調非減少で上に凸な関数である．量子通信路においては単調非減少であることは証明できるが，凸性については未解決である．

参考文献

[1] S. Arimoto, "On the converse to the coding theorem for discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 357–359, 1973.

[2] M. V. Burnashev and A. S. Holevo, "On reliability function of quantum communication channel." LANL Rep, quant-ph/9703013, 1997.

[3] A. S. Holevo, "The capacity of the quantum channel with genaral signal states," *IEEE Trans. Inform. Theory*, vol. IT-44, pp. 269–273, 1998.

[4] B. Schumacher and M. D. Westmoreland, "Sending classical information via noisy quantum channels," *Phys. Rev. A*, vol. 56, pp. 131–138, 1997.

量子仮説検定の漸近論について

電気通信大学・情報システム学研究科　長岡浩司

古典情報理論の多くの問題においてその最も一般的な結果は、韓-Verdú に創まる「情報スペクトル的方法」([1]) によって与えられる。以下では、量子仮説検定の漸近論のうち最も基本的な問題について、同様な観点からの考察を試みる。

1　量子仮説検定の漸近特性

ヒルベルト空間 $\mathcal{H}$ 上の密度作用素ρ,σ が与えられたとする。対象とする量子力学系の状態が ρ か σ のどちらかで表されると仮定し、実際にどちらであるかを何らかの測定をもとに判断する仮説検定問題を考える。このような仮説検定は一般に $0 \le T \le I$ を満たすエルミート作用素 T によって表わすことができる。このとき、真の状態がρ であるのに間違って σ であると判断してしまう確率（第一種の誤り確率）は $1-\mathrm{Tr}[\rho T]$, 逆に真の状態がσ であるのに ρ と判断してしまう確率（第二種の誤り確率）は $\mathrm{Tr}[\sigma T]$ となる。

今、状態を要素とする無限列 $\vec{\rho}=(\rho_1,\rho_2,\ldots)$ および$\vec{\sigma}=(\sigma_1,\sigma_2,\ldots)$ が与えられたとする。ここで、各 n に対し ρ_n と σ_n は共通のヒルベルト空間 $\mathcal{H}_n$ 上の密度作用素であるとするが、異なる n に対するρ_n, σ_n, $\mathcal{H}_n$ の間にはどんな関係も仮定しない。検定の無限列 $\vec{T}=(T_1,T_2,\ldots)$ （ T_n は $0 \le T_n \le I$ を満たす $\mathcal{H}_n$ 上のエルミート作用素）に対する第一種誤り確率 $1-\mathrm{Tr}[\rho_n T_n]$ および第二種誤り確率 $\mathrm{Tr}[\sigma_n T_n]$ の $n\to\infty$ における漸近的挙動に関して、次のような2種類の量を考える。

$$B(\vec{\rho}\|\vec{\sigma}) \stackrel{\mathrm{def}}{=} \sup\{R \mid \exists \vec{T},\ \lim_{n\to\infty}\mathrm{Tr}[\rho_n T_n]=1 \text{ かつ } \limsup_{n\to\infty}\frac{1}{n}\log\mathrm{Tr}[\sigma_n T_n] \le -R\} \tag{1}$$

$$C(\vec{\rho}\|\vec{\sigma}) \stackrel{\mathrm{def}}{=} \inf\{R \mid \forall \vec{T},\ \limsup_{n\to\infty}\frac{1}{n}\log\mathrm{Tr}[\sigma_n T_n] \le -R \text{ ならば } \lim_{n\to\infty}\mathrm{Tr}[\rho_n T_n]=0\} \tag{2}$$

明らかに $B(\vec{\rho}\|\vec{\sigma}) \le C(\vec{\rho}\|\vec{\sigma})$ が成り立つ。

2　情報スペクトル的特徴付け

与えられた二つの状態列 $\vec{\rho}=(\rho_n)$, $\vec{\sigma}=(\sigma_n)$ に対し

$$\underline{D}(\vec{\rho}\|\vec{\sigma}) \stackrel{\mathrm{def}}{=} \sup\{\lambda \mid \lim_{n\to\infty}\mathrm{Tr}[\rho_n\{\rho_n - e^{n\lambda}\sigma_n > 0\}]=1\} \tag{3}$$

$$\overline{D}(\vec{\rho}\|\vec{\sigma}) \stackrel{\mathrm{def}}{=} \inf\{\lambda \mid \lim_{n\to\infty}\mathrm{Tr}[\rho_n\{\rho_n - e^{n\lambda}\sigma_n > 0\}]=0\} \tag{4}$$

とおく。ただし $\{\rho_n - e^{n\lambda}\sigma_n > 0\}$ は、エルミート作用素 $\rho_n - e^{n\lambda}\sigma_n$ の正の固有値に対応する固有ベクトルの張る部分空間の上への正射影作用素を表わす。このとき、常に次式が成立する（証明略）。

$$B(\vec{\rho}\|\vec{\sigma}) = \underline{D}(\vec{\rho}\|\vec{\sigma}), \quad C(\vec{\rho}\|\vec{\sigma}) = \overline{D}(\vec{\rho}\|\vec{\sigma}). \tag{5}$$

3　i.i.d. の場合

本節では、$\mathcal{H}_n, \rho_n, \sigma_n$ が単一のヒルベルト空間 $\mathcal{H}$ とその上の密度作用素 ρ,σ によって

$$\mathcal{H}_n = \mathcal{H}^{\otimes n}, \quad \rho_n = \rho^{\otimes n}, \quad \sigma_n = \sigma^{\otimes n}$$

と表わされる場合を考える。$\mathcal{H}^{\otimes n}$ 上の任意の測定（簡単のため離散値とする）$M^{(n)}$ に対し、状態 $\rho^{\otimes n}$ および $\sigma^{\otimes n}$ のもとでの測定値の確率分布は、それぞれ $p_n(x)=\mathrm{Tr}[\rho^{\otimes n}M^{(n)}(x)]$ および $q_n(x)=\mathrm{Tr}[\sigma^{\otimes n}M^{(n)}(x)]$

で与えられる。これらの分布の古典相対エントロピー (Kullback divergence) を $D_{M^{(n)}}(\rho^{\otimes n}\|\sigma^{\otimes n}) \stackrel{\rm def}{=} D(p_n\|q_n) = \sum_x p_n(x)\log\frac{p_n(x)}{q_n(x)}$ とおく。また、ρ と σ の量子相対エントロピーを $D(\rho\|\sigma) \stackrel{\rm def}{=} {\rm Tr}\,[\rho(\log\rho - \log\sigma)]$ で定義する。このとき、日合-Petz の定理 ([2], [3])

$$\lim_{n\to\infty}\frac{1}{n}\sup_{M^{(n)}} D_{M^{(n)}}(\rho^{\otimes n}\|\sigma^{\otimes n}) = D(\rho\|\sigma) \tag{6}$$

が成り立ち、これを用いて

$$\underline{D}(\vec{\rho}\|\vec{\sigma}) = D(\rho\|\sigma) \tag{7}$$

を示すことができる。さらに最近、次の等式も証明された ([4])。

$$\overline{D}(\vec{\rho}\|\vec{\sigma}) = D(\rho\|\sigma) \tag{8}$$

4 古典論との比較

古典論の場合、確率分布の列 $\vec{p} = (p_n)$, $\vec{q} = (q_n)$ に対し、仮説検定の漸近特性として $B(\vec{p}\|\vec{q})$ および $C(\vec{p}\|\vec{q})$ が式 (1) (2) と同様にして定義される。そして、X_n を分布 p_n に従う確率変数とし、

$$\underline{D}(\vec{p}\|\vec{q}) \stackrel{\rm def}{=} \sup\{\lambda \mid \lim_{n\to\infty}{\rm Pr}\{\frac{1}{n}\log\frac{p_n(X_n)}{q_n(X_n)} > \lambda\} = 1\} \tag{9}$$

$$\overline{D}(\vec{\rho}\|\vec{\sigma}) \stackrel{\rm def}{=} \inf\{\lambda \mid \lim_{n\to\infty}{\rm Pr}\{\frac{1}{n}\log\frac{p_n(X_n)}{q_n(X_n)} > \lambda\} = 0\} \tag{10}$$

とおけば、$B(\vec{p}\|\vec{q}) = \underline{D}(\vec{p}\|\vec{q})$ および $C(\vec{p}\|\vec{q}) = \overline{D}(\vec{p}\|\vec{q})$ が成り立つ ([5], [1]) 。式 (5) はこれらの等式の量子版である。両者の類似性は

$$\frac{1}{n}\log\frac{p_n(X_n)}{q_n(X_n)} > \lambda \iff p_n(X_n) - e^{n\lambda}q_n(X_n) > 0$$

と書き直してみればより明確になる。ただし量子の場合、

$$\{\rho_n - e^n\lambda\sigma_n > 0\} = \{\frac{1}{n}(\log\rho_n - \log\sigma_n) - \lambda > 0\}$$

は一般には成り立たないことに注意。

古典的な i.i.d. の場合を考えよう。p_n, q_n を分布 p, q の n 次 i.i.d. 拡張であるとすると、確率変数 $\frac{1}{n}\log\frac{p_n(X_n)}{q_n(X_n)}$ に関する大数の法則より

$$\underline{D}(\vec{p}\|\vec{q}) = \overline{D}(\vec{p}\|\vec{q}) = D(p\|q)$$

が示される。量子 i.i.d. の場合の式 (7) (8) を大数の法則のある種の量子版とみなす観点は大変興味深い。

参考文献

[1] 韓 太舜: 「情報理論における情報スペクトル的方法」, 培風館, 1998.

[2] F. Hiai and D. Petz: "The proper formula for relative entropy and its asymptotics in quantum probability," Commun. Math. Phys., vol.143, pp.99-114, 1991.

[3] M. Hayashi: "Asymptotic attainment for quantum relative entropy," e-print quant-ph/9704040, 1997.

[4] T. Ogawa and H. Nagaoka: "Strong converse and Stein's lemma in the quantum hypothesis testing," 準備中.

[5] S. Verdú: private communication, 1994.

量子推定理論における漸近的大偏差型評価について

林 正人 [1]　　京都大学 理学研究科 数学教室

本研究では, 量子状態の推定理論を大偏差型評価に注目して扱った.

まずここで量子系の測定に関する基本事項を触れておくことにする. 量子系で測定を行うと, 測定器の構造と測定対象である状態の双方に依存して測定値が得られる. しかし, 一般にはその測定器の構造と準備された状態から, 測定値の値を予言することはできない. 両者からは個々の測定値が得られる確率しか予言できない. そこで, 測定器 M で状態 ρ を測定したときに得られる確率分布を P_ρ^M で表すことにする.

本研究では測定対象である状態が未知で, なおかつ十分たくさんにその未知状態が準備できるとの仮定の下で, その状態を推定するにはどのような性質をもつ測定器を用いて測定すればよいか考察することにある. なお, ここで測定器と呼んでいるものは, 単にビームなどの測定対象を測定して, データを与える測定装置だけではなく, そのデータを変換する計算機と測定装置を組み合わせたものを意味することもある.

以下で量子系での状態に関する数学的記述をまとめておく. 量子力学では, 状態はある複素 Hilbert 空間 $\mathcal{H}$ 上の self-adjoint 非負定値作用素でトレースが 1 のもので表される. どのようなヒルベルト空間 $\mathcal{H}$ をとるかは対象となる系によって異なり, その系の性質によって決まる. なおこの Hilbert 空間 $\mathcal{H}$ は量子系の表現空間と呼ばれる. 表現空間が $\mathcal{H}$ となる量子系の状態の集合を $\mathcal{S}(\mathcal{H})$ で表すことにする. そしてパラメトライズされた $\mathcal{S}(\mathcal{H})$ の部分集合 $S := \{\rho_\theta | \theta \in \Theta\}$ を状態族と呼ぶ.

量子力学では状態をノルム 1 の Hilbert 空間 $\mathcal{H}$ の元 $|f\rangle$ で表す流儀があるが, それは $|f\rangle$ で張られる一次元部分空間への射影 $|f\rangle\langle f|$ [2] を考えることにより, $\mathcal{S}(\mathcal{H})$ の元を表しているとみなせる. なお, ノルム 1 の Hilbert 空間の元で状態を表す流儀では $|f\rangle$ と $e^{i\phi}|f\rangle$ を同一視しているので, ここで述べている対応関係には問題ない. 状態集合 $\mathcal{S}(\mathcal{H})$ は作用素としての凸結合を考えると, 凸集合になっている. このような Hilbert 空間 $\mathcal{H}$ の 1 次元部分空間への射影で表される状態はこの凸結合に関する状態集合 $\mathcal{S}(\mathcal{H})$ の端点[3] になっており, 純粋状態と呼ぶことにする.

そして量子系で測定を考えるときに不可欠なのが正作用素値測度 (Positive Operator-Valued Measure,POVM) である POVM は σ-field $\mathcal{F}$ から $\mathcal{H}$ 上の非負定値な自己共役作用素への写像 M' で与えられる.

- $M'(\emptyset) = 0,\ M'(\Omega) = \mathrm{Id},$
- $B_i \cap B_j = \emptyset (i \neq j)$ を満たす加算個の集合列 $\{B_j\} \subset \mathcal{B}(\Omega)$ に対し $\sum_j M'(B_j) = M'\left(\bigcup_j B_j\right).$

なお, 今後 表現空間が $\mathcal{H}$ となる σ-field $\mathcal{F}$ 上の POVM の集合を $\mathcal{M}(\mathcal{F}, \mathcal{H})$ で表すことにする. さらに, 量子力学では測定 M に対して以下の凸結合則を要請する. なお M の測定値集合を Ω

[1] e-mail address: masahito@kusm.kyoto-u.ac.jp

[2] 量子力学では $\mathcal{H}$ の元 $|f\rangle$ を Hilbert 空間の内積を用いて自然に $\mathcal{H}$ の双対空間の元と見なしたものを $\langle f|$ で表す. そして $\mathcal{H}$ 上の線形写像 $|f\rangle \otimes \langle f|$ を $|f\rangle\langle f|$ で表す.

[3] 凸集合の元で他の異なる 2 つの元の凸結合で表せないものを端点と呼ぶ.

で表し, Ω は可分かつ完備 (すなわちポーランド空間) であるとする. そして可測集合族は Borel 集合族 $\mathcal{B}(\Omega)$ を考えることにする.

$$P^M_{\lambda\rho_1+(1-\lambda)\rho_2}(B) = \lambda P^M_{\rho_1}(B) + (1-\lambda)P^M_{\rho_2}(B), \quad \forall B \in \mathcal{B}(\Omega), \forall \lambda \in [0,1], \forall \rho_1, \rho_2 \in \mathcal{S}(\mathcal{H}). \tag{1}$$

(1) の条件を満たす $\{P^M_\rho | \rho \in \mathcal{S}(\mathcal{H})\}$ は適当な正作用素値測度 (Positive Operator-Valued Measure,POVM) M' を用いて $\mathrm{tr}\, M'(B)\rho = P^M_\rho(B)$ と表すことができる. なお, このとき M' の σ-field としては Borel 集合族 $\mathcal{B}(\Omega)$ を対応させると良い.

このことは測定値の確率分布にのみ注目するのであれば実際に構成される測定器 M は POVM で記述できることを主張している. しかし, 任意の POVM に対応する測定装置が実際にできるかについては今のところ十分には分っていない.

次に独立同一分布のの量子的対応物を考える. 同一な量子状態 ρ が n 個独立に準備されたときそれは n-テンソル空間 $\mathcal{H}^{(n)} := \underbrace{\mathcal{H}\otimes\cdots\otimes\mathcal{H}}_{n}$ 上の量子状態 $\rho^{(n)} := \underbrace{\rho\otimes\cdots\otimes\rho}_{n}$ で表される. したがってこのような仮定の下で n 個のサンプルに対する推定量は n-テンソル空間 $\mathcal{H}^{(n)}$ 上の POVM $M \in \mathcal{M}(\mathcal{B}(\Theta), \mathcal{H}^{(n)})$ で与えられる.

以下では POVM の範囲で測定誤差の下限に関する議論を行う. 量子状態族 $\mathcal{S} := \{\rho_\theta | \theta \in \Theta\}$ に対して推定量の系列 $\{M_n\}$ (各 M_n は $\mathcal{M}(\mathcal{B}(\Theta), \mathcal{H}^{(n)})$ の元.) が以下の条件を満たすとき弱一致という.

$$\mathrm{tr}\, \rho^{(n)}_\theta M_n(\{\hat\theta \in \Theta | \|\theta - \hat\theta\| > \epsilon\}) \to 0, \quad \forall \epsilon > 0 \quad \forall \theta \in \Theta. \tag{2}$$

本研究では上記の一致性の下で大偏差型の評価を行う. 確率分布族のパラメータ推定での大偏差型評価では Kullback-Leibler の divergence が重要であったが, その量子的対応物である梅垣によって導入された量子相対エントロピー (梅垣エントロピー) $D(\rho\|\sigma) := \mathrm{tr}\,\rho(\log\rho - \sigma)$ が重要な役割を果たす. 弱一致性を満たす推定量の系列 $\{M_n\}$ に対して以下の不等式が成立する.

$$\liminf_{n\to\infty} \frac{1}{n} \log\left(\mathrm{tr}\, \rho^{(n)}_\theta M_n(\{\hat\theta \in \Theta | \|\theta - \hat\theta\| > \epsilon\})\right) \geq -\inf\{D(\rho_{\theta'}\|\rho_\theta) | \|\theta' - \theta\| \geq \epsilon\}, \quad \forall \epsilon > 0, \forall \theta \in \Theta.$$

さらに極限 $\epsilon \to 0$ を考えることにより以下の不等式も成立する.

$$\liminf_{\epsilon\to 0} \liminf_{n\to\infty} \frac{1}{n\epsilon^2} \log\left(\mathrm{tr}\, \rho^{(n)}_\theta M_n(\{\hat\theta \in \Theta | \|\theta - \hat\theta\| > \epsilon\})\right) \geq -\frac{1}{2}\tilde J_\theta. \tag{3}$$

ここで

$$\tilde J_\theta := \lim_{\epsilon\to 0} \frac{1}{\epsilon^2} \inf\{D(\rho_{\theta'}\|\rho_\theta) | \|\theta' - \theta\| \geq \epsilon\}.$$

と定義した. 今のところ式 (refaee) の右辺を達成し, 弱一致性を満たす推定量の系列を任意の量子状態族に対して構成することはできていない. いまのところ, 量子性があらわれるモデルで (refaee) の右辺が達成されることが示されているモデルは以下に定義する熱ノイズを受けたコヒーレント状態族 (Thermaly-Noised Coherent States Family) $\{\rho_{\theta,N} | \theta \in \mathbb{C}\}$ のみである.

$$\rho_{\theta,N} := \frac{1}{\pi N} \int_{\mathbb{C}} e^{-\frac{|\theta-\beta|^2}{N}} \|\beta\rangle\langle\beta\|\, d^2\beta.$$

詳しくは予稿を参照のこと.

Independent Component Analysis の信号処理への応用

理化学研究所 脳科学総合研究センター　村田 昇
さきがけ研究 21「情報と知」領域　池田 思朗

1 Independent Component Analysis

Independent Component Analysis (ICA, 独立成分分析) では線形作用によって混合されたいくつかの独立成分の復元が目的であり，一般的には以下のように定式化される．平均 0，互いに独立な信号ベクトルを $\boldsymbol{s}(t) = (s_1(t), \ldots, s_n(t))^T$, $t = 0, 1, 2, \ldots$ で表し，観測を $\boldsymbol{x}(t) = (x_1(t), \ldots, x_n(t))^T$, $t = 0, 1, 2, \ldots$ とする．A を線形作用素として $\boldsymbol{s}(t)$ と $\boldsymbol{x}(t)$ との間には $\boldsymbol{x}(t) = A\boldsymbol{s}(t)$ なる関係が成り立つと仮定する．典型的な A としては定数行列や convolution filter が用いられる．ここで問題は信号 $\boldsymbol{s}(t)$ や A に関する知識を持たず観測 $\boldsymbol{x}(t)$ のみを用いて，$\boldsymbol{y}(t) = B\boldsymbol{x}(t)$ で定まる $\boldsymbol{y}(t)$ の各成分が互いに独立となるような B を探すことである．B は理想的には A^{-1} となれば良いわけだが，容易にわかるように要素の順番の入れ違いとその大きさの任意性は残る．

2 解法

この問題の解法は大きく分けると独立性の定義に基づいて信号の確率分布を用いるものと，信号の時系列としての性質を利用して時間相関を用いるものがある．

2.1 確率分布の独立条件に基づく分離法

各 $s_i(t)$ が強定常で non-Gaussian であると仮定する．このとき $\boldsymbol{y}(t)$ は強定常過程となりその同時分布が存在するので，各要素間の独立性は同時分布と周辺分布の積が一致することと特徴付けられる．したがってこれらの間の統計的距離を最小化する B を求めれば良い．応用上良く用いられるのは Kullback-Leibler 情報量であり，これは $\boldsymbol{y}$ の相互情報量と呼ばれる量となる．最も単純なのは A が定数行列の場合であるが，この場合も一般に B は解析的に求めることができず，繰り返し演算を用いることになる．この時更新則は y_i の周辺分布を $p_i(y_i)$ として

$$\Delta B \propto \left(I - \frac{1}{T} \sum_{t=1}^{T} \varphi(\boldsymbol{y}(t)) \boldsymbol{y}(t)^T \right) B, \quad \varphi(\boldsymbol{y}) = \left(-\frac{d \log p_1(y_1)}{dy_1}, \ldots, -\frac{d \log p_n(y_n)}{dy_n} \right) \tag{1}$$

という形にまとめられる．周辺分布 $p_i(y_i)$ は未知なので通常はパラメトリックな非線形関数や統計的な展開法を用いて近似される (例えば [2] 参照)．

2.2 相関関数に基づく分離法

信号 $\boldsymbol{s}(t)$ は弱定常で，各要素は異なる相関関数を持つと仮定する．信号の独立性よりその相関関数行列 $\langle \boldsymbol{s}(t)\boldsymbol{s}(t+\tau)^T \rangle$ は任意の τ において対角行列となるので，観測値の相関関数行列 $\langle \boldsymbol{x}(t)\boldsymbol{x}(t+\tau)^T \rangle = A \langle \boldsymbol{s}(t)\boldsymbol{s}(t+\tau)^T \rangle A^T$ を任意の τ において同時に対角化する B を求める問題に帰着される．

特に A が定数行列，したがって B も定数行列となる場合は対称行列の同時対角化問題とし代数的に解く方法が提案されている [5]．このアルゴリズムではまず観測信号の分散行列 $V = \langle \boldsymbol{x}(t)\boldsymbol{x}(t)^T \rangle$ により定義される $\sqrt{V^{-1}}$ を用いて "sphering" を行ない無相関な信号 $\boldsymbol{x}'(t) = \sqrt{V^{-1}}\boldsymbol{x}(t)$ を得る．ただし S, A をそれぞれ直交行列，対角行列とし，分散行列を $V = SAS^T$ のように分解し，$\sqrt{V^{-1}} = \sqrt{A^{-1}}S^T$ と定義する．この後 Givens の unitary rotation を拡張した方法 (Cardoso and Souloumiac [3]) により "rotation" を行なう．これらの 2 つの操作を順次適用して，求める行列は $B = C\sqrt{V^{-1}}$ と表される．

この手法は一定数の演算で B が求まり，また 2 次の統計量しか用いていないため数値的に安定した解が得られる．以下で報告する実験においては主にこの手法を用いている．

3 信号処理への応用

3.1 MEG データの解析

MEG (Magnetoencephalography，脳磁計) は時間分解能にして 500Hz-1000Hz 程度，空間分解能にして 5mm-1cm 程度の情報が得られるため，非侵襲で脳内の神経活動を捉える方法として現在注目を集めている．脳から発生される磁場は極めて微弱 (地磁気の数億分の一程度) であるため，MEG の計測においては地磁気，商用電源，センサーの量子力学的雑音，また注目していない脳部位の神経活動などの雑音

除去が重要な問題となる．従来はシールドルームの利用，フィルタ操作，別系統のセンサーによる補償などを経た上で，同一条件下で複数回 (100-200 回程度) 計測されたデータを加算平均することにより信号雑音比を上げている．また脳内の活動部位を推定するためには，脳内に仮想的に電流双極子が存在すると仮定し，各センサーの位置と向き，強度分布を基に双極子の位置，向きと強度を推定するなどの手法が用いられている．

現在我々は島津製作所の協力の下，同社製 MEG による観測データに 2.2 で示した手法を適用し，計測された信号の中から着目する信号成分を分離・抽出し解析する可能性を探っている．予備的な実験としてファントム (生理食塩水に満たされた球内に置かれた電流双極子) のデータに ICA を適用することにより雑音と双極子による信号成分を分離し，従来に比べ少ない加算平均回数で正確な双極子位置推定が行なえることを確認している．また被験者を用いた実験では平常時の計測磁場から加算平均を行なうことなしにα波やまばたきなどによる信号成分を抽出することができ，分離成分に基づいた双極子の推定を行っている [1].

3.2 音声信号の分離

我々が対象とするのは，例えば二人の会話を二つのマイクで録音し，それを分離するという問題で，カクテルパーティー効果で扱われるのと同じ状況である．この問題では信号の到達時間の遅れや，部屋の壁等での反射があるため，線形作用としては Finite Impulse Response (FIR) フィルタの convolution を考えなければならない．こうした観測データから信号を復元するのにはいくつかの手法が提案されているが，代表的なものはこの FIR フィルタの係数を推定し，逆フィルタを構成して分離する手法である．このような解法は Blind Source Deconvolution と呼ばれている．

我々の提案する方法では，windowed Fourier 変換を利用し，元の信号と観測の間の関係を

$$\hat{\boldsymbol{x}}(\omega, t_s) = \hat{A}(\omega)\hat{\boldsymbol{s}}(\omega, t_s), \tag{2}$$

と時間周波数表現に書き換えることにより，分離の問題を各周波数帯での non-convolutive な問題に分解して，計算を単純化している．

具体的には複素数体上に拡張した 2.2 節の手法を用いて各周波数において ICA を実行し，音声信号の周波数成分間にある大域的な相関を利用して，同一信号源から得られたと推定される周波数成分の分類を行い，逆 Fourier 変換により最終的に信号の再構成を行っている．

現在のところ，信号源の数を 2 ないし 3 として，人工的なデータおよび実験室で録音したデータを用いた数値実験を行い，概良好な結果を得ている [4].

4 まとめ

今後の課題としては，センサーの数と信号源の数が一致しない場合やノイズのある状況下での理論解析が挙げられる．応用例として挙げた MEG データ解析に対しては，脳のデータの双極子位置推定を含めた解析を精密化していく必要がある．また，音声分離の手法としては on-line アルゴリズムの開発，ハードウェアでの実現を考えていきたい．

References

[1] 池田思朗, 村田昇. Independent component analysis を用いた MEG データの解析. 信学技報, NC98, 1998.

[2] S. Amari, A. Cichocki, and H. H. Yang. A new learning algorightm for blind signal separation. In *Advances in Neural Information Processing Systems 8*, pages 757–763. MIT Press, Cambridge MA, 1996.

[3] J.-F. Cardoso and A. Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM J. Mat. Anal. Appl.*, 17(1):161–164, jan 1996.

[4] N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. submitted to IEEE trans. on Signal Processing.

[5] A. Ziehe, K.-R. Müller, G. Nolte, B.-M. Mackert, and G. Curio. ICA analysis of MEG data. NIPS 97: Functional brain imaging workshop (workshop talk), 1997.

Fisher 情報行列が縮退する場合のニューラルネットの学習誤差と汎化誤差について

三重大学 工学部 物理工学科 萩原克幸

1 はじめに

階層型ニューラルネットを用いた非線形回帰モデルのモデル選択においては, ネットワークの族が真の関数を含んでいて最小でない場合に, 結合重み (ネットワーク・パラメータ) が識別不能となり Fisher 情報行列が縮退するため AIC(Akaike Information Criterion)[1] などのモデル選択規準が漸近展開により導けないという問題がある. モデル選択規準は, 学習誤差に基づく汎化誤差の不偏推定量であるため, 階層型ニューラルネットのモデル選択の問題を解決するためには, その汎化誤差および学習誤差のデータの分布に関する期待値を知る必要がある. 本研究では, 階層型のニューラルネットの与える関数をパラメトライズされた基底関数の線形結合と見做し, その基底関数の学習における選択性に着目し, 学習誤差の期待値の上界および汎化誤差の期待値の下界を導く.

2 問題設定

確率密度 $q(x,y) = q(y|x)q(x)$ をもつ確率分布 $Q(x,y) = Q(y|x)Q(x)$ からの独立な T個のデータ $(x_t, y_t), t = 1, \ldots, T$の集まりを $D_T = \{(x_t, y_t) : x_t \in \mathbf{R}^m,\ y_t \in \mathbf{R},\ 1 \leq t \leq T\}$ と表し, y_tは x_tに対し, $y_t = h(x_t) + \xi_t,\ t = 1, \ldots, T$という規則により生成されるものとする. ただし, ξ_tは平均が 0 で有限な分散σ_*^2をもつ確率分布からの独立なサンプルとする. 一方, 関数$\psi_{\varepsilon_{n,i}, b_{n,i}} : \mathbf{R}^m \to \mathbf{R}, i = 1, \ldots, n$ 達の線形結合 $f_{w_n}(x) = \sum_i c_{n,i} \psi_{b_{n,i}, \varepsilon_{n,i}}(x),\ x \in \mathbf{R}^m$により D_Tをフィッティングすることを考える. ただし, $w_n = (c_n, b_n, \varepsilon_n) \in W_n$はパラメータであり, $c_n = (c_{n,1}, \ldots, c_{n,n}), c_{n,i} \in \mathbf{R}, b_n = (b_{n,1}, \ldots, b_{n,n}), b_{n,i} \in \mathrm{B}, \varepsilon_n = (\varepsilon_{n,1}, \ldots, \varepsilon_{n,n}), \varepsilon_{n,i} \in \mathrm{E}$ とする. ここで, $W_n = \mathbf{R}^n \times \mathrm{B}^n \times \mathrm{E}^n$ は関数のパラメータ空間である. 損失を二乗誤差とし, $s_{D_T}(w_n) = s_{D_T}(c_n, b_n, \varepsilon_n) = \frac{1}{T}\sum_t (y_t - f_{w_n}(x_t))^2$と定義する. これを経験二乗誤差と呼ぶ. $b_n \in \mathrm{B}^n$および$\varepsilon_n \in \mathrm{E}^n$を任意に固定したとき, $\min_{c_n} s_{D_T}(c_n, b_n, \varepsilon_n) = s_{D_T}(\widehat{c}_n(b_n, \varepsilon_n), b_n, \varepsilon_n)$ と定義し, $\inf_{w_n} s_{D_T}(w_n) = \inf_{b_n, \varepsilon_n} s_{D_T}(\widehat{c}_n(b_n, \varepsilon_n), b_n, \varepsilon_n)$ と定義する. これを学習誤差と呼ぶ. この定義の下で, $s_{D_T}(\widehat{w}_n) = \inf_{w_n} s_{D_T}(w_n),\ \widehat{w}_n = (\widehat{c}_n, \widehat{b}_n, \widehat{\varepsilon}_n)$ とする. $\widehat{w}_n = \widehat{w}_n(D_T)$ は上の定義の下での最小二乗推定量である. 学習誤差の D_Tの確率分布に関する期待値を $E_{D_T}\{s_{D_T}(\widehat{w}_n)\} = \int \cdots \int s_{D_T}(\widehat{w}_n) \prod_{t=1}^T q(x_t, y_t) dy_t dx_t$で定義する. これを学習誤差の ($D_T$の確率分布に関する) 期待値と呼ぶ. 一方, $D_T^+ = \{(u_t, z_t) : 1 \leq t \leq T\}$ を, 互いに独立かつ各 (u_t, z_t) がどの $(x_t, y_t) \in D_T$とも独立になるように$Q(x,y)$ からサンプルされたデータの集まりとする. このとき, 汎化誤差を $E_{D_T^+}\left\{s_{D_T^+}(\widehat{w}_n)\right\} = \int \cdots \int s_{D_T^+}(\widehat{w}_n) \prod_{t=1}^T q(z_t, u_t) dz_t du_t$と定義する. ただし, $s_{D_T^+}(\widehat{w}_n) = \frac{1}{T}\sum_t \left(z_t - f_{\widehat{w}_n}(u_t)\right)^2$である. 通常, モデル選択の枠組では, 汎化誤差は D_Tに依存するから, $E_{D_T}\left\{E_{D_T^+}\left\{s_{D_T^+}(\widehat{w}_n)\right\}\right\}$によるネットワークの評価を考える. これを汎化誤差の (D_Tの確率分布に関する) 期待値と呼ぶ.

3 n 点のみをフィッティングする関数

いま, パラメータ$\beta \in \mathbf{R}^m$により決まる指標関数を$\chi_\beta(x) = 1\ (\beta = x)\ ;\ 0\ (\beta \neq x),\ x \in \mathbf{R}^m$と定義し, $f_{\omega_n}(x) = \sum_i \alpha_{n,i} \chi_{\beta_{n,i}}(x)$ と定義する. ただし, $\omega_n = (\alpha_n, \beta_n),\ \alpha_n = (\alpha_{n,1}, \ldots, \alpha_{n,n}), \alpha_{n,i} \in \mathbf{R}, \beta_n = (\beta_{n,1}, \ldots, \beta_{n,n}), \beta_{n,i} \in \mathrm{B} \subset \mathbf{R}^m, \mathrm{B} = \{x_t : x_t \in \mathbf{R}^m, 1 \leq t \leq T\}$ とする. $s_{D_T}(\omega_n) := \frac{1}{T}\sum_t (y_t - f_{\omega_n})^2$ とし, $s_{D_T}(\widehat{\omega}_n) = \min_{\omega_n} s_{D_T}(\omega_n)$ とする. このとき$\widehat{\omega}_n = (\widehat{\alpha}_n, \widehat{\beta}_n)$ はパラメータの最小二乗推定量である. いま, データについて次のことを仮定する.

(A) $q(y|x) = 1/\sqrt{2\sigma_*^2}\exp\{-y^2/(2\sigma_*^2)\}$ とする. すなわち, y_tの確率分布を正規分布 $N(0, \sigma_*^2)$ とする. $q(x)$ は任意の確率分布とする. これは, 与えるデータが正規雑音のみ, すなわち, データの生成機構において, $\xi_t \sim N(0, \sigma_*^2)$ であり, $h(x) = 0,\ x$ a.e.であることを意味する.

この仮定の下では, パラメトライズされた基底関数の線形結合による非線形回帰モデルのパラメータ値は常に識別不能となり, Fisher 情報行列が縮退するため, 通常の漸近展開により汎化誤差と学習誤差の関係を導くことはできない. 仮定 (A) の下で, [2] の結果を使うと, 次のことが言える.

補題 1 仮定 *(A)* の下で, 十分大きな $T \gg n$ に対し,

$$E_{D_T}\{s_{D_T}(\widehat{\omega}_n)\} \sim \sigma_*^2 - \sigma_*^2 \frac{2n}{T} \log T.$$

4 主要な結果

$\psi_{b_{n,i},\varepsilon_{n,i}}, i = 1, \ldots, n$ について次の仮定をおく.

(B) 任意の $\bar{\beta}_{n,i} \in \mathbf{R}^m, i = 1, \ldots, n$ に対し, $\bar{b}_{n,i} \in \mathrm{B}$ と a_i が存在して, 次式を満たすものとする.

$$\lim_{\varepsilon_{n,i} \to a_i} \psi_{\bar{b}_{n,i},\varepsilon_{n,i}}(x) = \chi_{\bar{\beta}_{n,i}}(x), \ x \in \mathbf{R}^m, \ \text{a.e.}$$

この仮定の下で, $\bar{\beta}_n = \widehat{\beta}_n$ とおいて, $\lim_{\varepsilon_n \to a} s_{D_T}(\widehat{c}_n, \bar{b}_n, \varepsilon_n) = s_{D_T}(\widehat{\omega}_n)$ が示される. ただし, $a = (a_1, \ldots, a_n)$ である. これと補題 1 より, 学習誤差の期待値について次のことが言える.

定理 1 仮定 *(A)* および *(B)* の下で, 十分大きな $T \gg n$ に対し,

$$E_{D_T}\{s_{D_T}(\widehat{w}_n)\} \leq \sigma_*^2 - \sigma_*^2 \frac{2n}{T} \log T.$$

次に汎化誤差の期待値を解析する. ここでは, 次のことを仮定する.

(C) 入力データの確率分布 $Q(x)$ について, $Q(x) = \frac{1}{T}\sum_t \delta(x - r_t)$ とする. だだし, $\delta(x - r_t), r_t \in \mathbf{R}^m$ は Dirac の δ 関数である. これは, 各入力データ x_t が確定的に r_t をとることを意味する.

仮定 (C) の下では, 定理 1 の結果を使うと, 汎化誤差の期待値について, 直ちに次のことが知れる.

定理 2 仮定 *(A),(B),(C)* の下で, 十分大きな $T \gg n$ に対し,

$$E_{D_T}\left\{E_{D_T^+}\left\{s_{D_T^+}(\widehat{w}_n)\right\}\right\} \geq \sigma_*^2 + \sigma_*^2 \frac{2n}{T} \log T.$$

定理 1 および定理 2 は, Gaussian Radial Basis Function や中間層に bell 型の活性化関数をもつ 3 層階層型ニューラルネットおよび中間層に sigmoid 型の活性化関数をもつ 3 層階層型ニューラルネットなどのニューラルネットのクラスに対して成り立つことが容易に示される.

5 結論

本稿では, データを正規雑音のみとし, 定義した最小二乗誤差規範の下で, 任意の入力分布の下での学習誤差の期待値の上界および入力が確定的な場合の汎化誤差の期待値の下界を導き, これらが成り立つニューラルネットの例を示した. データ数がある程度大きい場合には, 定理 1 の学習誤差の期待値の上界は, [3] において Fisher 情報行列が縮退しない下で漸近展開により導かれる学習誤差の期待値より小さく, 定理 2 の汎化誤差の期待値の下界は, [3] で得られている汎化誤差の期待値より大きい. 今後は, 学習誤差の期待値の下界および汎化誤差の期待値の上界を導く必要がある.

参考文献

[1] Akaike H. : In *2nd International Symposium on Information Theory*, B.N.Petrov and F.Csáki eds., Akadémia Kiado, Budapest, 267–281, (1973).

[2] 早坂太一, 萩原克幸, 戸田尚宏, 臼井支朗 : 日本神経回路学会誌 , **4**, pp.18–26 (1997).

[3] Murata N., Yoshizawa S. and Amari S. : *IEEE Trans. on Neural Networks*, **5**, 6, pp.865–872 (1994).

真のパラメータ集合が特異点を持つ確率モデルの統計的推測

東京工業大学・PI Lab 渡辺澄夫

1 はじめに

入力 $x \in R^M$ に対して出力 $y \in R^N$ を確率的に推論するシステム $p(y|x,w)$ を考える。ここで $w \in W \subset R^d$ であり、W 上に確率密度関数 $\varphi(w) \in C_0^\infty$ が定義されているとする。$q(x)$ を R^M 上のコンパクトサポートかつ連続な確率密度関数とし、同時確率 $p(y|x,w_0)q(x)$ から独立に得られた n 個のサンプルの組を $\{(x_i,y_i); i=1,2,...,n\}$ とする。このサンプルより推測されたシステム $p_n(y|x)$ をギブス分布に従って構成する。

$$f_n(w) = \frac{1}{n}\sum_{i=1}^{n}\{\log p(y_i|x_i,w_0) - \log p(y_i|x_i,w)\}$$

$$p_n(y|x) = \int p(y|x,w)\exp(-nf_n(w))\varphi(w)dw \;/\; \int \exp(-nf_n(w))\varphi(w)dw$$

カルバック学習精度 $K(n)$ と自由エネルギー $F(n)$ とを次式で定義する。

$$K(n) = E_n\{\int \log\frac{q(y|x)}{p_n(y|x)}q(x,y)dxdy\}, \quad F(n) = -E_n\{\log\int \exp(-nf_n(w))\varphi(w)dw\}$$

ここで $E_n\{\cdot\}$ は n 個のサンプルの出方に関する平均を表す。定義より $K(n) = F(n+1) - F(n)$ であるから、学習精度 $K(n)$ を知るためには自由エネルギー $F(n)$ を求めれば良い。平均損失関数を

$$f(w) = \int\{\log p(y|x,w_0) - \log p(y|x,w)\}\, q(x,y)dxdy$$

と定義し、この関数が $w \in W$ について解析的な場合を考える。もし $f(w)$ の零点集合が 1 点であり、その点のヘシアン（フィッシャー情報行列）が正定値であれば最尤推定量の漸近正規性を経由して $F(n) = (d/2)\log n + O(1)$ が成り立つ。しかしながら、一般にデータから構造を推定しようとするモデルでは、真のパラメータ集合 W_0 は実代数多様体であり、従来の方法は適用できない。これは混合モデル・階層的神経回路網・双極子推定・クラスタリング・木構造抽出などの情報処理において頻出するにも関わらず統計的には未解決の問題である。本論ではこの問題の数学的基礎を確立することを目的とする。

2 自由エネルギーの上からの評価

W として W_0 を含む十分小さい開集合 $\{w \in W; f(w) < \epsilon\}$ を考えれば十分なので、以下、この集合を W とかく。まず Jensen の不等式より

$$F(n) \leq -\log\int_W \exp(-nf(w))\varphi(w)dw \tag{1}$$

が成り立つ。次に佐藤 - Bernstein - Björk - 柏原の定理から、λ について多項式の微分作用素 $P(\lambda,w,\partial w)$ と多項式 $b(\lambda)$ (零点は実数で負の有理数) が存在して

$$P(\lambda,w,\partial w)f(w)^{\lambda+1} = b(\lambda)f(w)^\lambda \quad ({}^\forall w \in W\setminus W_0,\ {}^\forall\lambda \in C)$$

を満たすことが証明されている。このことより、1 変数複素数 λ の関数

$$J(\lambda) = \int_W f(w)^\lambda\varphi(w)dw \quad (\lambda \in C)$$

は、C 全体に解析接続され、有理型関数となる ($|\lambda| < \infty$ において高々極のみを持つ)。そこで $J(\lambda)$ の極を絶対値が小さいものから$-\lambda_1, -\lambda_2, -\lambda_3, \cdots$, ($\lambda_k > 0$) と定義し、その位数を $m_1, m_2, m_3, \cdots$, とおく ($m_k \geq 1$)。正則モデルでは $\lambda_1 = d/2$, $m_1 = 1$ である。この極と位数を用いると　$t \to 0$ における漸近展開

$$\int_W \delta(t - f(w))\varphi(w)dw \quad \cong \quad \sum_{k=1}^{\infty} \sum_{m=1}^{m_k} c_{km} t^{\lambda_k - 1} (-\log t)^{m-1}$$

を示すことができ、このことと式 (1) から $n \to \infty$ のとき次の漸近評価が成り立つ。

$$F(n) \leq \lambda_1 \log n - (m_1 - 1) \log \log n - \log c_{1m_1} - \log \Gamma(\lambda_1) \tag{2}$$

3 自由エネルギーの下からの評価

関数 $\log p(y|x, w) - \log p(y|x, w_0)$ が w について解析的で、(x, y) について連続ならば、Weierstrass の予備定理から、$\{h_j(x, y, w_0)\}$ が一次独立な複素数値関数 $\{h_j\}$ と $g_j(w_0) = 0$ となる複素数値関数 $\{g_j(w)\}$ とがあって

$$\log p(y|x, w) - \log p(y|x, w_0) = \sum_{j=1}^{J} g_j(w) h_j(x, y, w)$$

と書くことができる。このことより、

$$\alpha_n = \sup_{w \in W \setminus W_0} \frac{\sqrt{n}(f(w) - f_n(w))}{\sqrt{f(w)}} \quad \text{と定義すると} \quad E_n\{|\alpha_n|^2\} < Const.$$

を示すことができ、上からの評価と同じ方法を用いて

$$\begin{aligned} F(n) &\geq -E_n\{\log \int_W \exp(-nf(w) + \alpha_n \sqrt{nf(w)})\varphi(w)dw\} \\ &\geq \lambda_1 \log n - (m_1 - 1) \log \log n - \log c_{1m_1} - Const. \end{aligned} \tag{3}$$

となる。式 (2) と式 (3) とから結論の式が得られる。

4 結論

自由エネルギー $F(n)$ についてある定数 Const. が存在して

$$|F(n) - \lambda_1 \log n - (m_1 - 1) \log \log n| < Const.$$

が成立することを示した。$F(n)$ は確率的複雑さ・ベイズファクタ・ABIC などと密接な関係があり、統計的推測・情報理論・ベイズ推定において役立つことが期待される。

参考文献

[1] E.Levin, N.Tishby, S.A.Solla, *Proc. of IEEE*, Vol.78, No.10, pp.1568-1674, 1990.
[2] S. Amari, N. Fujita, S. Shinomoto, *Neural Comp.*, Vol.4, No.4, pp.608-618, 1992.
[3] M. Sato, T.Shintani, *Anals. of Math.*, Vol.100, pp.131-170, 1974.
[4] I.N. Bernstein, *Functional Anal. Appl.* Vol.6, pp. 26-40,1972.
[5] M.Kashiwara, *Inventions Math.*, Vol.38, pp.33-53, 1976.
[6] 渡辺澄夫, 信学技報、NC98-64, pp73-80, 1998.

Generalized amount of information and estimation for a family of non-regular distributions

筑波大・数学　赤平 昌文

1. はじめに

統計的推測理論において良く知られている情報量にはFisher情報量, Kullback-Leibler情報量などがあるが, これらは正則な場合, すなわち分布に正則条件が仮定されている場合には有用であるが, そうでない場合すなわち非正則な場合には必ずしも有用とはいえない. では, 非正則の場合に有用な情報量はあるだろうか.

そこで, 標本空間 $(\mathfrak{X}, \mathfrak{B})$ 上の確率測度 P, Q が, ある σ-有限測度 μ に関して絶対連続であると仮定するとき, P, Q の間の情報量を

$$I(P,Q) := -8\log\int_{\mathfrak{X}}\left(\frac{dP}{d\mu}\cdot\frac{dQ}{d\mu}\right)^{1/2}d\mu$$

によって定義した ([AT91], [AT95]). ここで右辺の積分値は類似度 (affinity) と呼ばれている. この情報量は非正則な推定問題に有用であることが確かめられている. その後, 上の $I(P,Q)$ をさらに一般化 (した) 情報量として, 任意の α ($|\alpha|<1$) について

$$I^{(\alpha)}(P,Q) := -\frac{8}{1-\alpha^2}\log\int_{\mathfrak{X}}\left(\frac{dP}{d\mu}\right)^{(1-\alpha)/2}\left(\frac{dQ}{d\mu}\right)^{(1+\alpha)/2}d\mu$$

を定義した ([A96]). これは Rényi 型測度になっている. 特に $\alpha=0$ とすれば一般化情報量は情報量 $I(P,Q)$ に一致する, すなわち $I^{(0)}(P,Q) = I(P,Q)$ である.

本論では, 統計量の一般化情報量損失の概念とその非正則分布族の場合への応用について述べ ([A96]), また, 一般ベイズ推定量の2次の漸近的性質についても論じる.

2. 一般化情報量損失の概念

まず, $X_1, \cdots, X_n, \cdots$ を互いに独立にいずれも (σ-有限測度 μ に関する) 密度関数 $f(x,\theta)$ $(\theta\in\Theta)$ をもつ分布に従う確率変数列とする. ただし Θ は母数空間で, $\mathbf{R}^1$ の開区間とする. このとき, 任意の θ_1, $\theta_2\in\Theta$ に対して $f(\cdot,\theta_1)$, $f(\cdot,\theta_2)$ の間の X_1 の一般化情報量 (generalized amount of information) を各 α ($|\alpha|<1$) に対して

$$I^{(\alpha)}_{X_1}(\theta_1,\theta_2) := -\frac{8}{1-\alpha^2}\log\int_{-\infty}^{\infty} f(x,\theta_1)^{(1-\alpha)/2}f(x,\theta_2)^{(1+\alpha)/2}d\mu(x)$$

で表わす. 同様にして, $f(\cdot,\theta_1)$ と $f(\cdot,\theta_2)$ の間の $\boldsymbol{X}=(X_1,\cdots,X_n)$ の一般化情報量を $I^{(\alpha)}_{\boldsymbol{X}}(\theta_1,\theta_2)$ で表わせば, $I^{(\alpha)}_{\boldsymbol{X}}(\theta_1,\theta_2) = nI^{(\alpha)}_{X_1}(\theta_1,\theta_2)$ になる. 一般に, 統計量 $T_n := T_n(\boldsymbol{X})$ の一般化情報量も同様に定義して, それを $I^{(\alpha)}_{T_n}(\cdot,\cdot)$ で表わせば, 適当な正則条件の下で, $I^{(\alpha)}_{T_n}(\theta_1,\theta_2)\le I^{(\alpha)}_{\boldsymbol{X}}(\theta_1,\theta_2)$ が成り立つ. そこで, 任意の α ($|\alpha|<1$) について統計量 T_n の一般化情報量損失を $I^{(\alpha)}_{\boldsymbol{X}}(\theta_1,\theta_2) - I^{(\alpha)}_{T_n}(\theta_1,\theta_2)$ で定義する. 次節で $|\theta_1-\theta_2| = O(n^{-1})$ のときに, その一般化情報量損失を $o(n^{-1})$ の次数まで考察する.

なお, 一般化情報量と Fisher 情報量の関係については次のようになる. 適当な正則条件の下で, 任意の α ($|\alpha|<1$) と十分小さい $|\Delta\theta|$ に対して, $I^{(\alpha)}_{X_1}(\theta,\theta+\Delta\theta) = I_{X_1}(\theta)(\Delta\theta)^2 + o\left((\Delta\theta)^2\right)$ になる. ただし $I_{X_1}(\theta) = E_\theta\left[\{(\partial/\partial\theta)\log f(X_1,\theta)\}^2\right]$ (Fisher 情報量) とする.

3. 非正則分布族における統計量の2次の一般化情報量損失

まず, $X_1, \cdots, X_n, \cdots$ を互いに独立にいずれも (ルベーグ測度に関する) 密度関数 $f(x,\theta)$ $(\theta\in\Theta)$ をもつ分布に従う実確率変数列とする. ただし $\Theta=\mathbf{R}^1$ とする. このとき θ が位置母数, すなわち $f(x,\theta) = f_0(x-\theta)$ の場合を考える. また f_0 に次の条件を仮定する.

(A1) $f_0(x)>0$ $(a<x<b)$; $f_0(x)=0$ $(x\le a, x\ge b)$. ただし a, b は有限とする.

(A2) $f_0(x)$ は開区間 (a,b) において 2 回連続微分可能で, $\lim_{x\to a+0} f_0(x) = \lim_{x\to b-0} f_0(x) = c$, $\lim_{x\to b-0} f_0'(x) = -\lim_{x\to a+0} f_0'(x) = h$ である. ただし c は正の定数で, h は定数とする.

(A3) $0 < I_0 := \displaystyle\int_a^b \{f_0'(x)\}^2 / f_0(x)dx < \infty.$

上記の条件の下では, 一致性の次数は n であることが知られている.

次に, 極値統計量 $\bar{\theta}$ と $\underline{\theta}$ を $\bar{\theta} := \min_{1\le i\le n} X_i - a$, $\underline{\theta} := \max_{1\le i\le n} X_i - b$ とし, $Z_1(\theta) := -(1/\sqrt{n})\sum_{i=1}^n f_0'(X_i-\theta)/f_0(X_i-\theta)$ $(\underline{\theta} < \theta < \bar{\theta})$ とする. また $\hat{\theta}^* = (\underline{\theta}+\bar{\theta})/2$ とおくと $\hat{\theta}^*$ は θ の一致推定量になる. さらに $Z_1^* := Z_1(\hat{\theta}^*)$ とおくと, Z_1^* は漸近補助統計量になる.

一般に, 統計量 $T_n := T_n(\boldsymbol{X})$ の 2 次の一般化情報量損失を, 任意の α $(|\alpha| < 1)$ に対して

$$L_n^{(\alpha)}(T_n) := \frac{1}{n\Delta^2}\left\{I_{\boldsymbol{X}}^{(\alpha)}(\theta,\theta+\Delta) - I_{T_n}^{(\alpha)}(\theta,\theta+\Delta)\right\} + o(1)$$

で定義する. ただし $\Delta = O(1/n)$ とする. このとき統計量 $T_n^* = \left(Z_1^*/(\sqrt{n}I_0), \bar{\theta}, \underline{\theta}\right)$ の一般化情報量を求めると, 次の結果を得る ([A96]).

定理 1. 条件 (A1) ～ (A3) の下で, $\Delta = O(1/n)$ とすれば, 統計量 T_n^* の 2 次の一般化情報量損失は, 任意の α $(|\alpha| < 1)$ に対して, $L_n^{(\alpha)}(T_n^*) = o(1)$ $(n\to\infty)$ であり, また, 統計量 $(\bar{\theta},\underline{\theta})$ の 2 次の一般化情報量損失は, 任意の $\alpha(|\alpha| < 1)$ に対して, $L_n^{(\alpha)}(\bar{\theta},\underline{\theta}) = I_0 + o(1)$ $(n\to\infty)$ である.

上の定理から, 任意の α $(|\alpha| < 1)$ に対して, $I_{\boldsymbol{X}}^{(\alpha)}(\theta,\theta+\Delta) - I_{T_n^*}^{(\alpha)}(\theta,\theta+\Delta) = o(1/n)$ $(n\to\infty)$ になり, 統計量 T_n^* の一般化情報量損失は 2 次の次数まで, すなわち $o(1/n)$ まで 0 になる. このことは, 一方向型分布族に対して T_n^* が 2 次の漸近十分統計量になるという結果とも符合している. また, 上の定理の結果が α について不変であることに注意.

4. 一般ベイズ推定量の 2 次の漸近展開

前節の設定の下で, 損失関数 $L(u)$ を 3 回連続微分可能で, $|u|$ の単調増加な非負値関数とする. 損失 L とルベーグ測度に関する (θ の) 一般ベイズ推定量は, $a.a.$ $(x_1,\cdots,x_n)$ について $\int_{\underline{\theta}}^{\bar{\theta}} L(\hat{\theta}-\theta)\prod_{j=1}^n f(x_j-\theta)d\theta$ を最小にする $\hat{\theta}$ になる ([A88]). そこで, それを $\hat{\theta}_{GB} = \hat{\theta}_{GB}(\boldsymbol{X})$ で表わす. また, $L^{(k)}(u) = (d^k/du^k)L(u)$ $(k = 1,2,3)$ として, $L^{(1)}(0) = 0$, $L^{(3)}(0) = 0$ と仮定する. さらに, 条件 (A2) において $h \le 0$ とし, $f_0(x)$ は $x = (a+b)/2$ に関して対称であるとした条件を (A2)′ とする. このとき, 一般推定量 $\hat{\theta}_{GB}$ は Pitman 推定量と漸近的に同等と考えられる.

定理 2. 条件 (A1), (A2)′, (A3) の下で, $\hat{\theta}_{GB}$ は漸近展開

$$n(\hat{\theta}_{GB}-\theta) = n(\hat{\theta}^*-\theta) + \frac{1}{3\sqrt{n}}Z_1(\theta)T^2 - \frac{I}{3n}n(\hat{\theta}^*-\theta)T^2 + o_p\left(\frac{1}{n}\right) \qquad (n\to\infty)$$

をもつ. ただし, $T := n(\bar{\theta}-\underline{\theta})/2$, $I := I_0 - 2h$ とする.

上記の定理 1, 2 の例としては切断正規分布等の場合 ([AT95]) が考えられる.

参考文献

[A88] Akahira, M. (1988). Second order asymptotic properties of the generalized Bayes estimators for a family of non-regular distributions. In: *Statistical Theory and Data Analysis II*, (K. Matusita, Ed.), North-Holland, Amsterdam, 87–100.

[A96] Akahira, M. (1996). Loss of information of a statistic for a family of non-regular distributions. *Ann. Inst. Statist. Math.*, **48**, 349–364.

[AT91] Akahira, M. and Takeuchi, K. (1991). A definition of information amount applicable to non-regular cases. *Journal of Computing and Information*, **2**, 71–92.

[AT95] Akahira, M. and Takeuchi, K. (1995). *Non-Regular Statistical Estimation.* Lecture Notes in Statistics **107**, Springer, New York.

応答変数が2値変数である一般化線形モデル における初期推定量のベイズ的構成法

岡山理科大学　中村　忠
岡山大学　平井安久
中国短期大学　奥村英則

1．はじめに

応答変数が2値変数である一般化線形モデル(GLM)を考える：

1. $Y_1,...,Y_K$ を K 個の独立な確率変数とし，各々が未知母数 $p_i \in (0, 1)$ をもつ二項分布 $B(N_i, p_i)$ に従うものとする．
2. 未知母数からなるベクトル $\theta = (\theta_1,..., \theta_r)'$ および既知の説明変数からなるベクトル $x_i = (x_{i1},..., x_{ir})'$ によって線形予測子 $\eta_i = x_i'\theta$ が構成される．
3. $\eta_i = F^{-1}(p_i)$ である．ただし，$F(\cdot)$ は既知の確率分布関数(CDF)である．

真の母ベクトル θ_0 を推定する方法として最尤法，非線形または非線形（重み付き）最小二乗法などが知られている．最尤推定量(MLE)あるいは線形（重み付き）最小二乗推定量(N(W)LSE)は一般にはデータの関数として明確な形で表せない．この種の推定量の値を計算する場合は逐次計算の技法が用いられる．このとき良い初期値を与える推定量が必要である．初期推定量とは次の3つの条件を満たすときにいう：(i) 計算が容易である；(ii) 使用する逐次計算法において十分な初期値を与える；(iii) 統計的意味で真の母数の良い推定量となっている．本論では，初期推定量のベイズ的アプローチによる構成法を与える。

2．ベイズ的バイアス修正推定量の構成

Y を二項分布 $B(N, p)$ に従う確率変数とし，$g(p)$ をいくつかの条件(Nakamura and Hirai 1994)を満たす $(0, 1)$ 上のなめらかな関数とする．ここで，$p \in (0, 1)$ は未知母数である．関数 $g(p)$ を推定することを考える．$g(p)$ の推定量として p のMLE $p^* = Y/N$ を用いた $g(p)$ のMLE $g(p^*)$ がしばしば採用される．しかし，MLE $g(p^*)$ にはいくつかの欠点がある．この問題を避けるために，p の推定量としてBerksonの $2N$ ルールによって構成されたもの，すなわち $a + b > -N$ に対して

$$\hat{p}(Y; N, a, b) = \begin{cases} \dfrac{1}{2N}, & 0 \le Y \le -a, \\ \dfrac{Y + a}{N + a + b}, & -a < Y < N + b, \\ \dfrac{2N - 1}{2N}, & N + b \le Y \le N \end{cases}$$

(cf. Gart, Pettigrew and Thomas 1985; Nakamura and Hirai 1994)．$g(p)$ がCDF $F(\cdot)$ の逆関数のとき $g(\hat{p}(Y; N, a, b))$ は修正経験変換とよばれる(cf. Cox and Snell 1989)．a と b を決めるために $g(\hat{p}(Y; N, a, b))$ のバイアスの漸近表示の N^{-1} の係数

$$I_g^*(p; a, b) = g'(p)(a - (a + b)p) + g''(p)p(1 - p)/2$$

を考える．Nakamura and Hirai(1994)はこの $I_g^*(p; a, b)$ が $g(\hat{p}(Y; N, a, b))$ の漸近表示の1次の項であることを示している．ここで p は非情報事前密度関数 $\pi(p)$ をもつ確率変数とみなす．この

分布の下側（または上側）10%点をδ_1（またはδ_2）とする．次の最小化の問題を考える：

(I) 関数$\int_{\delta_1}^{\delta_2} I_g^*(p; a, b)^2 \pi(p)dp$を集合$\{(a, b); a+b>-N\}$上で最小にするような$\hat{a}, \hat{b}$を求めよ．

ある条件のもとで，最適解$\hat{a}=\hat{a}(\pi, g)$と$\hat{b}=\hat{b}(\pi, g)$が存在する．$g(\hat{p}(Y; N, \hat{a}, \hat{b}))$をバイアス修正推定量とよぶ．

3．初期推定量の選択

逐次計算法で，MLE θ^*，非線形最小2乗推定量$\hat{\theta}_{\mathrm{NLSE}}$，非線形重み付き最小2乗推定量$\hat{\theta}_{\mathrm{NWLSE}}$などの推定量を計算するとき，初期値が必要である．本節では，そのような初期値を与える初期推定量を提案する．$Z=(F^{-1}(p_1),\ldots, F^{-1}(p_K))'$，$X=(x_1',\ldots, x_K')'$とおくとGLMは$Z=X\theta$と書ける．逐次計算法を使わないで得られる推定量として，線形最小2乗推定量$\hat{\theta}_{\mathrm{OLSE}}(\pi, F^{-1})$，線形重み付き最小2乗推定量$\hat{\theta}_{\mathrm{WLSE}}(\pi, F^{-1})$がある．前節で得られた結果を利用すると，$\hat{\theta}_{\mathrm{OLSE}}(\pi, F^{-1})$，$\hat{\theta}_{\mathrm{WLSE}}(\pi, F^{-1})$はそれぞれ次のような明確な形で書ける．

$$\hat{\theta}_{\mathrm{OLSE}}(\pi, F^{-1})=(X'X)^{-1}X'\hat{Z}(\pi, F^{-1}),\quad \hat{\theta}_{\mathrm{WLSE}}(\pi, F^{-1})=(X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}\hat{Z}(\pi, F^{-1}).$$

ただし，$\hat{Z}(\pi, F^{-1})=(F^{-1}(\hat{p}(Y_1; N_1, \hat{a}(\pi, F^{-1}), \hat{b}(\pi, F^{-1}))),\ldots, F^{-1}(\hat{p}(Y_K; N_K, \hat{a}(\pi, F^{-1}), \hat{b}(\pi, F^{-1}))))'$，$h(p)=f(F^{-1}(p))^{-2}p(1-p)/N$，$a^*=\hat{a}(\pi, h)$，$b^*=\hat{b}(\pi, h)$で，$\hat{V}$は対角行列で，その対角成分は$f(F^{-1}(\hat{p}(Y_i; N_i, a^*, b^*)))^{-2}\hat{p}(Y_i; N_i, a^*, b^*)(1-\hat{p}(Y_i; N_i, a^*, b^*))/N_i$である．$\hat{\theta}_{\mathrm{OLSE}}(\pi, F^{-1})$，$\hat{\theta}_{\mathrm{WLSE}}(\pi, F^{-1})$は事前分布によって決定される推定量である．種々の非情報事前分布が考えられる．ベイズ解析ではJeffreys's priorが優れた性質をもっていることが知られている(cf. Ibrahim and Laud 1991)．Jeffreys's prior $\pi_J(p)=\frac{1}{\pi}p^{-1/2}(1-p)^{-1/2}$に対応するOLSE $\hat{\theta}=\hat{\theta}_{\mathrm{OLSE}}(\pi, F^{-1})$を初期推定量として提案する．

4．初期推定量の評価

われわれの提案する推定量が良い推定量であることを数値実験によって示した．

参考文献

[1] Cox, D. R. and Snell, E. J.(1989) "Analysis of binary data." Chapman and Hall, London.

[2] Gart, J. J., Pettigrew, H. M. and Thomas, D. G.(1985) The effect of bias, variance estimation, skewness and kurtosis of the empirical logit on weighted least squares analyses. Biometrika, 72, 179 - 190.

[3] Ibrahim, J. B. and Laud, P. W.(1991) On Bayesian analysis of generalized linear models using Jeffreys's prior. J. Amer. Statist. Assoc., 86, 981 - 986.

[4] Nakamura, T. and Hirai, Y.(1994) An expression of asymptotic bias of an estimator based on binomial sample and its application to dose-reponse model. Fifth Japan-China Simposium on Statistics (Okayama University of Science, Okayama, Japan), 199 - 202.

[5] Zellner, A.(1996) Models, prior information, and Bayesian analysis. J. Econometrics, 75, 51 - 68.

統計量の強収束について

和歌山大学経済学部　松田　忠之
早稲田大学理工学部　鈴木　武

1　はじめに

Edward W. Barankin の予想：

$\mathcal{P}_0 = \{p_0(\cdot,\theta);\theta \in \Theta\}$ に対する十分統計量は、適当な条件のもとで、$p_n(x,\theta) \to p_0(x,\theta)$ を満足する分布族 $\mathcal{P}_n = \{p_n(\cdot,\theta);\theta \in \Theta\}$ に対する十分統計量の極限として求めることができる

が正しいことを示すために、Kudo (1970) は、次の統計量の強収束 (strong convergence) の概念を用いた。

定義 (Neveu (1965))　λ を可測空間 $(\mathcal{X},\mathcal{A})$ 上の確率測度とし、$\mathcal{B}_n$ $(n \geq 1)$ と $\mathcal{B}$ を $\mathcal{A}$ の sub-σ-algebra とする。すべての λ-可積分関数 f に対して

$$\lim_{n\to\infty} \int |E_\lambda[f:\mathcal{B}_n] - E_\lambda[f:\mathcal{B}]|\, d\lambda \;=\; 0$$

が成立するとき、$\mathcal{B}_n$ は $\mathcal{B}$ に強収束するという。同様に、統計量の強収束は、統計量によって induce された σ-algebra の強収束によって定義される。

Kudo (1970) は、統計量 T_n が T_0 に強収束するための十分条件を与えるために、統計量に対して微分可能性と一様収束性を仮定した。講演の前半では、これらの条件を仮定するのではなく、確率密度関数に対する適当な条件のもとで、Barankin の予想が成立することを示す。

2　極限統計量の十分性

k 次元ユークリッド空間 R^k のルベーグ可測部分集合を $\mathcal{X}$ で表し、その上のルベーグ可測集合の全体を $\mathcal{A}$ で表す。この可測空間 $(\mathcal{X},\mathcal{A})$ 上で定義された確率測度の族を $\mathcal{P}_n = \{P_{\theta,n};\theta \in \Theta\}, n = 0,1,2,\cdots,$ とする。ここで Θ は母数空間を表し、有限次元ユークリッド空間の開集合であると仮定する。

確率分布族$\mathcal{P}_n$は $(\mathcal{X},\mathcal{A})$ 上のルベーグ測度 μ に関して絶対連続と仮定する。このとき，$P_{\theta,n}$の μ に関する確率密度関数を $p_n(x,\theta)$ で表し，その台 (carrier) を $S_{n,\theta} \equiv \{x \in \mathcal{X};p_n(x,\theta) > 0\}$ とおく。さらに、すべての $n \geq 0$ に対して、$S_n = \bigcup_{\theta\in\Theta} S_{n,\theta}$ とおく。

統計量の列 $\{T_n;n \geq 0\}$ は $\mathcal{X}$ 上で定義され、あるユークリッド空間のルベーグ可測部分集合 $\mathcal{Y}$ に値をとると仮定する。$\mathcal{Y}$ 上の測度として、ルベーグ測度 ν を考える。

以後、可側性あるいはほとんどすべて (a.e.) という言葉を使うときは、とくに断らない限りルベーグ測度 μ に対して用いると解釈する。以下の議論では、確率密度関数および統計量に対して次の条件を仮定する。

仮定１　すべての $\theta \in \Theta$に対して、$\lim_{n\to\infty} p_n(x,\theta) = p_0(x,\theta)$　a.e. $x \in \mathcal{X}$ 。

仮定２　すべての $n \geq 1$ に対して、$T_n(X)$ は$\mathcal{P}_n$に対して十分統計量である。

仮定３　$\lim_{n\to\infty} T_n(x) = T_0(x)$　a.e. $x \in \mathcal{X}$.

仮定２より、因子分解定理を適用すれば、すべての$n \geq 1$に対して $p_n(x,\theta)$を $p_n(x,\theta) = f_n(T_n(x),\theta)g_n(x)$ a.e. $x \in S_n$, for each $\theta \in \Theta$と分解できる。ここで、$f_n(T_n(\cdot),\theta)$ と g_n は $\mathcal{X}$ 上で定義された非負可測関数である。さらに、$f_n(y,\theta)$ と $g_n(x)$ に対して、

仮定4　任意の $\theta \in \Theta$に対して、次の条件 (a)-(c) を満足する $\mathcal{Y}$ の部分集合 $\mathcal{Y}_\theta$ を選ぶことができる。
(a)　$\nu\{\mathcal{Y} - \mathcal{Y}_\theta\} = 0$ かつ $\mu\{\mathcal{X} - T_0^{-1}(\mathcal{Y}_\theta)\} = 0$.
(b)　すべての $y \in \mathcal{Y}_\theta$ に対して、$\lim_{n\to\infty} f_n(y,\theta) = f_0(y,\theta)$。
(c)　関数列 $\{f_n(\cdot,\theta);\ n = 1,2,\cdots,\}$ は $y(\in \mathcal{Y}_\theta)$ に関して同程度連続である。
仮定5　$0 < \liminf_{n\to\infty} g_n(x) \leq \limsup_{n\to\infty} g_n(x) < \infty$　a.e. $x \in \mathcal{X}$.

定理　仮定1から仮定5が成立するとき、統計量 $T_0(X)$ は$\mathcal{P}_0$に対して十分統計量である。

3　漸近理論の枠組みにおける強収束

Jeganathan (1987, 1988) は漸近理論の枠組みにおいて，Kudo (1970) とは異なる視点から統計量の強収束について論じている．本報告では，Kudo (1970) で導入された強収束の概念に基づく結果を述べる．標本 $(X_1,\ldots,X_n)$ は $(\mathcal{X}^{(n)},\mathcal{A}_n)$ 上の分布 $P_{\theta,n}$ $(\theta \in \Theta, \Theta$ は$\mathcal{R}^p$ の開部分集合 $)$ に従うとする．$\theta_0 \in \Theta$, $h \in \mathcal{R}^p$ に対し，$\theta_n(h) = \theta_0 + \delta_n h$ ($\delta_n = \delta_n(\theta_0)$ は $p\times p$ 正値対称) とおく．$\{P_{\theta,n}\}$ は θ_0 でLAN（局所漸近正規）であるとする．つまり，p 次元確率ベクトル $Y_n(\theta_0), p\times p$ 正値対称行列 $I(\cdot)$ が存在して，$\forall h \in \mathcal{R}^p$ に対して，(i) $\log\{dP_{\theta_n(h),n}/dP_{\theta_0,n}\} = h'Y_n(\theta_0) - \frac{1}{2}h'I(\theta_0)h + \varepsilon_n$, $\varepsilon_n \to 0$ in $P_{\theta_0,n}-$probability, (ii) $\mathcal{L}(Y_n(\theta_0)|P_{\theta_0,n}) \Rightarrow N(0, I(\theta_0))$, (iii) $I(\theta)$ は θ_0 で連続, を満たすとする．統計量 $T_n : \mathcal{X}^{(n)} \to \mathcal{R}^p$ ($\mathcal{A}_n-$可測) の系列 $\{T_n\}$ は θ_0 で正則であるとする．即ち，$\mathcal{L}\left(\delta_n^{-1}(T_n - \theta_n(h))|P_{\theta_n(h),n}\right) \Rightarrow H(\theta_0)\ (\forall h \in \mathcal{R}^p)$ を満たすとする．$S_n = \delta_n^{-1}(T_n - \theta_0)$, $Y_n = Y_n(\theta_0)$, $\bar{S}(h) = S + h\,(S = \mathcal{L}(H(\theta_0))), \bar{Y}(h) = Y + I(\theta_0)h\,(\mathcal{L}(Y) = N(0, I(\theta_0)))$ とおく．S_n を与えた時の Y_n の（$P_{\theta_n(h),n}$ の下での）特性関数，及び $\bar{S}(h)$ を与えた時の $\bar{Y}(h)$ の特性関数をそれぞれ $\varphi_{Y_n|S_n}(t;\theta_n)$ 及び $\varphi_{\bar{Y}(h)|\bar{S}}(t)$ とする．この時

定理

$$\varphi_{\bar{Y}(h)|\bar{S}(h)=s}(t) = \lim_{n\to+\infty} \varphi_{Y_n|S_n=s}(t;\theta_n),\quad \forall\, t \in \mathcal{R}^p .$$

但し，(Y_n, S_n) の $P_{\theta_n(h),n}$ の下での特性関数 $\varphi_{(Y_n,S_n)}(t,v;\theta_n)$ につき，各 t を固定した時，vに関して $\{n\}$ 一一様可積分とする．

参考文献

[1] Bartlett, M.S. (1938): The characteristic function of a conditional statistic. *J. London Math. Soc.* **13**, 62–67.

[2] Beran, R. (1997): Diagnosing bootstrap success. *Ann. Inst. Statist. Math.* **49**, No.1, 1–24.

[3] Jeganathan, P. (1987): Strong convergence of distributions of estimators. *Ann. Statist.* **15**, No.4, 1699–1708.

[4] Jeganathan, P. (1988): On the strong approximation of the distributions of estimators in linear stochastic models, I and II : Stationary and explosive AR models. *Ann. Statist.* **16**, No.3, 1283–1314.

[5] Kudō, H. (1970): On an approximation to a sufficient statistic including a concept of asymptotic sufficiency. *J. Fac. Sci., Univ. of Tokyo, Sec I* **17**,273–290.

[6] Neveu, Jacques (1965): Mathematical Foundations of the Calculus of Probability. Holden-Day, Inc.

情報量の定義と統計的推測における意味

明治学院大学・国際学部　竹内　啓

1　情報量の意味

1.1　抽象的、一般的な考え方

情報量と統計的推測理論の方法の効率とは本来表裏の関係にある。すなわち

「あるデータの持つ情報量」
　　＝　「ある原因系に関する事前の不確実性」　－　「事後の不確実性」

「統計的推測」　＝　「データの変換」

「統計的推測の効率」
　　＝　「変換されたデータの持つ情報量」　÷　「データの持つ情報量」

と定義できよう。

1.2　「不確実性」の定義

ここで第一の問題は「不確実性」をどのように数量的に定義するかである。

不確実性にはデータそのものにかかわるもの（ばらつき）とデータの表す構造(未知母数）にかかわるものとがある。統計的推測と関係するのは後者である。従って次のようになる。

既知母数の場合　—　Shannon；統計的推測とは（とりあえず）無関係

離散母数の場合　—　Kullback-Leibler

連続母数の場合　—　Fisher

無限次元母数の場合　—　？

2　2つの分布の場合

2.1　不確実性＝分布の距離

とりあえず可能性として2つの分布F, Gが考えられ、それが密度関数f, gを持つとしよう。この仮定は完全に一般的であって

$$f = \frac{2dF}{d(f+g)} \qquad g = \frac{2dG}{d(F+G)}$$

とすればよい。

そしてこの2つの分布の間の不確実性は2つの分布の距離$\rho(f,g)$として定義できる。それは次のような性質を持つことが望ましい。

1. $\rho(f,g) \geq 0$
2. $\rho(f,g) = \rho(g,f)$
3. $\rho(f,h) \leq \rho(f,g) + \rho(g,h)$
4. $\rho(f_1 \times f_2, g_1 \times g_2) = \rho(f_1,g_1) + \rho(f_2,g_2)$
5. $\rho(f,g) \geq \rho(f^T, g^T)$
6. $\rho^* \sup \rho = \rho(f^*,g^*) \Longleftrightarrow f^*g^* \equiv 0$
   $(f^*g^* \not\equiv 0$ならば$\rho(f,g) < \rho^* \leq \infty)$

そうすると次のことが分かる。
・K-L情報量 $I_K = \int f \log \frac{f}{g}$　は2,3,6を満たさない。
・1~6を満たすもの　$I_H = -c \log \left(\int f^{1/2} g^{1/2} \right)$

2.2　推測

2つの分布に関する推測は、真の分布がそのどちらであるかを決めること（2決定問題)、あるいは単純仮説を単純対立仮説に対して検定することと考えられる。

それは標本に対して2つの点のいずれかを対応させることを意味する。すなわち Xに対して$0, 1$をとる統計量Tに変換することになる。f, gに対応する Tの分布を $(1-\alpha, \alpha), (\beta, 1-\beta)$とすれば、その情報量は$\rho(f^T, g^T)$は

$$I_K^T = \alpha \log \frac{\alpha}{1-\beta} + (1-\alpha) \log \frac{1-\alpha}{\beta} \leq I_K$$
$$\sqrt{\alpha(1-\beta)} + \sqrt{(1-\alpha)\beta} \geq \exp -cI$$

という不等式が得られる。$X_n = (X_1, X_2, \ldots, X_n)$に対応する$T$を $T_n = T(X_n)$として、nが大きいとき漸近的有効性を

$$\rho(f^{Tn}, g^{Tn}) / \rho(f^n, g^n) \to 1$$

で定義できる。

3　連続母数の場合

3.1　$\theta = \theta_0$における情報量

連続母数の場合、情報量は母数の特定の値$\theta = \theta_0$に対して

$$I_{\theta_0} = \lim_{\Delta\theta \to 0} \frac{1}{(\Delta\theta)^2} \left(\rho \left(f_{\theta_0}, f_{\theta_0 + \Delta\theta} \right) \right)$$
$$= \frac{\partial^2}{\partial\theta^2} \rho \left(f_{\theta_0}, f_\theta \right) |_{\theta=\theta_0}$$
$$\left(= \frac{\partial^2}{\partial\theta \partial\theta'} \rho \left(f_{\theta_0}, f_{\theta_0 + \Delta\theta} \right) |_{\theta=\theta_0} \right)$$

と定義される。

3.2　正則な場合

正則条件が満たされる場合には、2母数の場合のいくつかの情報量の定義から出発して Fisher情報量が得られる。

$$I_K, I_H \to \text{Fisher情報量} I_F$$
$$I_F = -\int f_\theta \frac{\partial^2}{\partial\theta^2} \log f_\theta = \int \left(\frac{\partial}{\partial\theta} \log f_\theta \right)^2 f_\theta$$

実母数θの推定はXを実確立変数$T = \hat{\theta}(X)$
に変換することである。その時不偏性を要求すれば
束縛条件　$E_\theta(T) = \theta \to \frac{\partial}{\partial\theta} E_\theta(T) = 1$
となり、この下で

$$V_\theta(T) \geq 1/I_F$$

が成り立つ。
検定問題　$H: \theta = \theta_0$　　$K: \theta > \theta_0$　　$0 \leq \Phi \leq 1$　　$E(\Phi_{\theta_0}) = \alpha$

については
局所最強力検定　$\frac{\partial}{\partial\theta}E_\theta(\Phi)\mid_{\theta=\theta_0}\to \max$
とすると

$$\Phi=1 \longleftrightarrow \frac{\partial}{\partial\theta}\log f_\theta \mid_{\theta=\theta_0} > \lambda$$

となり

$$n\to\infty\text{ならば}\quad \frac{\partial}{\partial\theta}E_\theta(\Phi)\mid_{\theta=\theta_0}\sim c\sqrt{nI_0}$$

が成り立つ。

4　局所情報量の定義

4.1　非正則な場合

非正則な場合（すなわちF, Gが互いに絶対連続でない場合）にはFisher情報量は定義できないが、局所情報量は

$$I^*_\Delta(\theta)=\int\frac{(f_{\theta+\Delta\theta}-f_\theta)^2}{\Delta\theta^2(f_{\theta+\Delta\theta}+f_\theta)}d\theta$$

で定義できる。そうすると

$$V_\theta(T)+V_{\theta+\Delta\theta}(T)\geq\frac{1}{I^*_\Delta(\theta)}-\frac{\Delta\theta^2}{2}$$

が成り立ち、
一様分布の場合

$$V_\theta(T)+V_{\theta+\Delta\theta}(T)\geq\frac{\Delta\theta^2}{2}\left(\frac{1}{1-(1-\Delta\theta)^n}-1\right)\approx\frac{1}{2n^2}\frac{t^2}{e^t-1}$$

となるので

$$\lim_{\epsilon\to0}\sup_{\theta':|\theta'-\theta|<\epsilon}n^2\left(V_\theta(T)+V_{\theta'}(T)\right)\geq\sup_t\frac{t^2}{e^t-1}=0.64761$$

となる。一様分布の場合もう一つの定義は

$$I^{**}_\Delta(\theta)=\begin{cases}-\left(\log\int(f_\theta+f_{\theta+\Delta\theta})^{1/2}\right)/\Delta\theta\\ -n\log(1-\Delta\theta)/\Delta\theta\simeq n\end{cases}$$

となる。

5　高次漸近理論と情報空間

情報量概念から出発して「距離」のほかに「曲率」も導入することによって、推定の高次の漸近理論を甘利俊一氏が構築した。その詳細については、ここではふれない。

甘利は曲指数分布族（有限次元母数指数分布）について考察しているが、母数空間の微分幾何学的構造を定義することは、一般の正則な場合について全く同様に可能である。

6 局外母数と情報量

6.1 分布が母数$\theta(\in R^1)$および$\xi(\in R^n)$で定義される場合

分布を規定する母数が、関心の対象となる実母数θと局外母数ξ (一般に多次元)である場合、θについてのみの情報量を考えて θに関する部分情報量partial informationを

$$I^p(\theta_1,\theta_2 \mid \xi) = \begin{cases} I((\theta_1,\xi),(\theta_2,\xi)) & \text{または} \\ \inf_{\xi'} I((\theta_1,\xi),(\theta_2,\xi')) \end{cases}$$

と定義する。Fisher情報量については

$$I_F^p(\theta \mid \xi) = (I_{\theta\theta} - I_{\theta\xi} I_{\xi\xi}^{-1} I_{\xi\theta})$$

と定義する。ただし$I_{\theta\theta}, I_{\theta\xi}$等は Fisher情報量行列の対応する成分を表す。またξが無限次元でも定義可能である。

6.2 非正則な場合

一つの例として

$$I^p(\theta_1,\theta_2 \mid \xi) = -\frac{n}{2}\log\frac{\mid \theta_1-\theta_2 \mid}{\xi+\mid \theta_1-\theta_2 \mid}$$

一様分布$(\theta-\xi/2, \theta+\xi/2)$について

$$I((\theta_1,\xi_1),(\theta_2,\xi_2)) = -\frac{n}{2}\log(1-\frac{l}{\xi_1})(1-\frac{l}{\xi_2})$$

lは$(\theta_1 \pm \frac{\xi_1}{2})$と $(\theta_2 \pm \frac{\xi_2}{2})$の共通部分の長さ

7 むすび

非正則な場合の情報量の定義と統計的推測の効率との関係については、まだ研究が進んでいない。それについて困難もあるが、また面白い結果も得られそうである。

MDL原理とその周辺

電気通信大学　情報システム学研究科　　韓太舜

1. データ解析とは

情報理論では，複合情報源モデル（確率モデル族）の候補 $\{p^n_{\theta^{(i)}}\}_{\theta^{(i)}\in\Theta_i}$ がいくつもある場合 $(i=1,2,\cdots,M)$ のユニバーサル符号化問題が重要性を有する．そこでの基本的考え方は，各複合情報源モデル i に対するユニバーサル符号語長 $l^{(i)}(x^n)$ （n はデータ長）を計算してその最小値 $l_{\mathrm{MDL}}(x^n)$ を達成するようなモデル $i=\hat{i}_0(x^n)$ を選択するというものである．

このような仕組みは，データ解析や統計学において長い歴史をもつモデル選択（model selection）問題の基本的枠組みそのものである．一般に，データ解析とは，あるデータ x^n が与えられたときその x^n が固有に持っている“内在的構造”を抽出するためのさまざまな手法やさまざまな方法論の総称のことであるが，ここでの最も重要な問題は，“内在的構造避’とは一体何かということである．データとは，一般に，それ自身の中にある固有の内在的構造と不規則な雑音（や歪み）とが重ね合わされた結果としてわれわれの前に提示されたり観測されたりするものであるが，何を内在的構造とし何を雑音とするかはそのデータを眺める我々の立場によってまったく異なってくる．世の中の出来事はどんな些末事でもすべてが意味を持っていると考える人もいれば，不確実性の時代にあっては確かな意味を持っているものなど何もないと考える人もいる．前者の立場からはデータのすべてが内在的構造であり，後者の立場からはデータは雑音の塊りそのものに過ぎない．それでは，そのような主観に左右されやすい“立場”というものをどう客観的に定式化すればよいのか．

2. MDL原理とその応用分野

実は，情報理論で考えられている上記のような複合情報源モデルのユニバーサル符号化の考え方は，この問題に1つの答えを与えているのである．まず，最初に用意する複数の情報源モデル $\{p^n_{\theta^{(i)}}\}_{\theta^{(i)}\in\Theta_i}$ $(i=1,2,\cdots,M)$ の各々がデータを眺めるためのそれぞれの“立場”を定量的に表現していると考えることができる．したがって複数の情報源モデルを用意しておくということは，あらかじめ1つの“立場”に限定しないで，出てきたデータ”x^nを最もよく説明できるような‘立場”をデータに合わせて選択していくという“柔軟性”を意味している．しかも，“立場”の選択が「ユニバーサル符号語長を最小にする」という単純明解な基準だけによって客観的に実行されるという仕組みが極めて魅力的なのである．この方式はその単純明解さのゆえに，原理的には，広範なデータ解析一般にも適用し得る“普遍性”をも備えている．このように，データ圧縮という固有の文脈を一旦離れて，データ解析一般の立場に立って眺めるとき，われわれはユニバーサル最適符号語長 $l_{\mathrm{MDL}}(x^n)$ を達成する最適情報源モデル $p^n_{\hat{\theta}^{(i_0)}}$ で指定される情報源モ

デルの番号$\hat{i}_0$，モデル次数$k_{\hat{i}_0}$，最尤推定量$\hat{\theta}^{(\hat{i}_0)}$の3つの組$(\hat{i}_0, k_{\hat{i}_0}, \hat{\theta}^{(\hat{i}_0)})$をデータ$x^n$の**内在的構造**と呼ぶことにする．そして，“雑音”で汚されているデータx^nを解析して，それ自身に固有の$(\hat{i}_0, k_{\hat{i}_0}, \hat{\theta}^{(\hat{i}_0)})$を抽出することが“データ解析”にほかならないと考える．その際，$p^n_{\hat{\theta}^{(\hat{i}_0)}}$はデータ$x^n$に含まれる“雑音”の確率分布を与えているものとみなす．データ解析に対するこのような考え方をRissanenはMDL原理（Minimum Description Length Principle）と呼び，最小のユニバーサル符号語長$l_{\mathrm{MDL}}(x^n)$をMDL基準（MDL Criterion）と呼んだ．MDL原理に関して特に強調しておかなければならないことは，それがデータに基づく“モデルの自動生成横構”を内蔵しているという点である．もともとはユニバーサル符号の研究に端を発したMDL原理は，この“自動生成機構’という特質のゆえにこそ，人工知能，機械学習，統計解析，パターン認識，画像認識など情報関連の諸分野における自己組織化，自動学習，自動認識，自動クラスタリング問題にも適用され，その有効性が実証されつつあるといえる．

ここで,「データx^nの符号語長（記述長）を最小にする」ことがすなわち与えられたデータx^nを説明する最良の“理論”を構成することにほかならないという思想が，人類の見果てぬ“夢”として古くから語り伝えられてきたことを指摘しておこう．例えば，次のような音葉からそれがうかがえる．MDL原理はこれらの夢を定式化したものと言うことができよう．

The shortest complete description is the best understanding (Ockham)

If I had more time I could write a shorter letter (B. Pascal)

Make everything as simple as possible—but not simpler (A. Einstein)

3. MDL原理による空間図形認識問題

従来のMDL原理では，複合情報源モデルの候補$\{p^n_{\theta^{(i)}}\}_{\theta^{(i)}\in\Theta_i}$ $(i=1,2,\cdots,M)$のパラメータ$\theta^{(i)}$がすべて意味のある「構造母数」である場合を考えていたが，ロボットが観測データやセンサーデータなどから3次元環境を構築するといったような幾何学的図形推定問題を考えると，複合情報源モデルはデータの内在的構造としての意味を有する「構造母数」以外に，“推定”したくない（あるいは，“推定”しても意味のない）「撹乱母数」を多数（データ長nに比例する程度）含まざるを得ないことになる．したがって，このような幾何学的図形推定問題にも適用できるようにMDL原理を拡張する必要が出て来る．このように拡張したMDL原理を仮に「幾何学的MDL原理」と呼ぼう．自然な拡張の1つの方法は,「撹乱母数」を適当な「事前分布」で平均して，もともとの「構造母数」と「事前分布」を指定するための新たな「構造母数」だけを含む複合情報源モデル$\{p^n_{\theta^{(i)}}\}_{\theta^{(i)}\in\Theta_i}$ $(i=1,2,\cdots)$を再構成した上で従来のMDL原理の考え方を適用することである．

1例として，データx^nが従う分布（定常無記憶）も「事前分布」も共に正規分布である場合を考え，膨大なシミュレーションによって,「幾何学的MDL原理」の考え方の有効性を検証した．

ベイズプライアーと情報量

電通大・電子情報　久保木　久孝

1. はじめに

ベイズ統計学におけるプライアー（事前情報）の決定問題は，古くから論争のつきないテーマである．最近は，プライアーを決める構造的な規則を探そうという考えが主流になって来ており，さまざまなスキームが提案されている．本稿の試みは，プライアーを決めるためのフォーマルなルールを概観する事である．

ところで，そのような規則で作られるプライアーには，さまざまなラベルが付けられている．例えば，“noninformative prior”, “conventional prior”, “default prior”, “generic prior” 等々．しかし最近は，リファレンスプライアー（a reference prior—a prior as a “standard of reference”）という用語が使われることが多くなった．ただし，このラベルは，Berger と Bernardo による情報理論的なスキームにもとづくルールによって選ばれたプライアーを指してに限定して使われることがあるので，注意を要する．

ルールによってプライアーを選ぶという考えは，Jeffreys による．このようなルール作りは，‘know nothing’ あるいは ‘ignorance’ という概念の定式化という問題を必然的に誘導する．したがって，何について知らないのかということの峻別が必要となる．事実，Jeffreys はパラメータの推定問題におけるルールを，仮説検定問題に対するルールとははっきり区別して考えている．同様な意味で，我々は予測問題におけるルールをパラメータ推定問題とは別に作る必要がある．これについては，論文 (Kuboki, 1998) を参照していただきたい．本稿では，パラメータの推定におけるプライアーの選択ルールに限定して議論する．特に情報量と関係する Berger-Bernardo method に重点を置く．

しかし，さまざまな議論も Jeffreys によって示唆された，いわゆる，Jeffreys prior あるいはそのバージョンに至る．このプライアーについては，本人の著書 (Jeffreys, 1961) に詳しい解説がある．

2. Jeffreys's general rule

パラメトリックモデルの Fisher 情報量行列を $I(\theta)$ とする．ここで，

$$I(\theta)_{ij} = E\left(-\frac{\partial^2 l}{\partial\theta_i \partial\theta_j}\right),$$

そして，l は対数尤度である．そのとき，Jeffreys のルールはプライアーを

$$\pi_\theta(\theta) \propto \det(I(\theta))^{1/2}$$

と定めるものである．このルールの一つの正当性は，パラメータの取り方に関し不変であるという点である．さらにこのプライアーは，パラメータ空間が群 G と同一視できるような変換群モデルでは，左 Haar 測度になっているという意味の不変性を持つ．

ところで，上のルールを特に Jeffreys's nonlocation rule と言うこともある．これに対し，もしモデルが θ の他にロケーションパラメータ $\mu_1, \ldots \mu_k$ を含む場合，Jeffreys は

$$\pi(\mu_1, \ldots, \mu_k, \theta) \propto \det(I(\theta))^{1/2}$$

というルールでプライアーを決めることを推奨している．ここで，左辺の $I(\theta)$ は $\mu_1, \ldots \mu_k$ を固定して計算されるものである．

3. The Berger-Bernardo Method

情報論的なスキームでプライアーを決めるという方法は，Bernardo (1979) によって革新的な発展の第一歩が与えられ，Berger と Bernardo の一連の論文で，理論と応用の飛躍的な展開がなされた．それ故，この方法を Berger-Bernardo method という．また，この方法で決定されるプライアーを Berger-Bernardo prior というが，リファレンスプライアーということもある．

今，$X_1^n = (X_1, \ldots, X_n)$ を i.i.d. 確率変数とする．プライアーを $\pi(\theta)$ とし，それを $X_1^n = x_1^n$ で $\pi(\theta|x_1^n)$ にアップデートしたとき，両者の違いを Kullback-Leibler distance あるいは相対エントロピー

$$K_n(\pi(\theta|x_1^n), \pi(\theta)) = \int \pi(\theta|x_1^n) \log \frac{\pi(\theta|x_1^n)}{\pi(\theta)}\, d\theta$$

で見る．これは，実験による情報の獲得を表している．この期待値，すなわち，X_1^n の周辺分布 $m(x_1^n) = \int p(x_1^n|\theta)\pi(\theta)\, d\theta$ に関する積分

$$K_n^\pi = E[K_n(\pi(\theta|x_1^n), \pi(\theta))]$$

は，情報の期待獲得と解釈される．サンプルサイズ n が十分大きいとき，ポステリアー $\pi(\cdot|x_1^n)$ はピークのある尤度関数に支配されて，プライアー π の形に依存しないことが期待される．Bernardo は，$K_\infty^\pi = \lim_{n\to\infty} K_n^\pi$ を missing information の測度と考え，それを最大にするプライアーを見つけることを提案した．そして，そのプライアーを "the" reference prior と呼んだ．

しかし，普通 K_∞^π は発散するので，この手続きは不可能である．Bernardo はそれを次のようなアルゴリズムで回避した：(i) K_n^π を最大にする π_n を求める；(ii) ポステリアー $\pi_n(\theta|x)$ の極限を求める；(iii) Bayes の定理を使い，その極限ポステリアーに対応するプライアーを求め，それを reference prior と定義する．十分な正則条件の下では，このプライアーは Jeffreys's nonlocation rule と一致する．

Bernardo のアプローチが威力をはっきするのは，局外 (nuisance) パラメータ問題である．今，$\theta = (\omega, \lambda)$，ここで，$\omega$ は関心のあるパラメータ，λ は局外パラメータとする．この場合，Bernardo は次のように彼の手続きを変更している：(i) ω を固定して λ に関する Berger-Bernardo prior $\pi(\lambda|\omega)$ を求める；(ii) 周辺モデル $p(x|\omega) = \int p(x|\omega, \lambda)\pi(\lambda|\omega)\, d\lambda$ を計算する；(iii) 周辺モデル $p(x|\omega)$ にもとづき，Berger-Bernardo prior $\pi(\omega)$ を求める；(iv) 最終的な Berger-Bernardo prior は $\pi(\omega, \lambda) = \pi(\omega)\pi(\lambda|\omega)$ で定義する．適当な正則条件の下では，Berger-Bernardo prior は

$$\pi(\omega, \lambda) \propto j_\omega(\lambda) \exp\left\{ \int j_\omega(\lambda) \log S(\omega, \lambda)\, d\lambda \right\}$$

となる．ここで，$j_\omega(\lambda)$ は ω が固定されたときの λ に関する nonlocation Jeffreys prior，そして $S = \sqrt{\det I / \det I_{22}}$ である．なお，I は Fisher 情報量行列，I_{22} は局外パラメータに対応する I の部分行列を表す．

Berger-Bernardo method は，局外パラメータ問題ばかりでなく，パラメータが順位の付いた任意の個数のグループに分割されている場合が取り扱えるよう，さらに拡張されている．また応用も急速に拡大している．これらは Berger and Bernardo (1992) に詳しい．

References

Berger, J.O. and Bernardo, J.M. (1992). On the development of reference priors (with discussion). In: *Bayesian Statistics 4*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford Univ. Press, Oxford, pp. 35–60. A.M.

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc.*, Ser. B **41**, 113–147.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford Univ. Press, London.

Kuboki, H. (1998). Reference priors for prediction. *J. Statist. Planning and Inference* **62**, 295–317.