統計数理 (1999)
第47巻 第1号 29-48

特集「統計基礎理論」
[研究詳解]

概パラメトリック推測
── 柔らかなモデルの構築 ──

統計数理研究所　江　口　真　透

(受付 1998年10月1日；改訂 1999年2月18日)

要　旨

パラメトリックな方法論とノンパラメトリックな方法論のギャップを埋めるためにセミパラメトリック推測の理論が発展しつつある．本論では，そのようなアプローチの一つである概パラメトリック推測について紹介する．その方法は，最尤法とパラメトリックモデルの強固な関係を緩和させるためにモデルをそのチューブ近傍の中に拡げ，データへの仮定を緩和させる．特長はパラメトリック推測の有効性を失わずに，より柔らかな仮定の下でも性能が保たれる点にある．その適用例として次の3例を紹介する：(1)密度推定のための局所尤度法，(2)自己組織化ルールによる尤度法，特に主成分分析法のロバスト推論，(3)観測バイアスの感度分析のための選択性パラメーターの尤度解析．いくつかの数値例を通してこの概パラメトリック推測論の良さを示す．

キーワード：概パラメトリックス，観測バイアス，局所尤度，自己組織化ルール，主成分分析，選択性パラメーター．

1. はじめに

Fisher (1922) によって提案された「最大尤度法」(最尤法) は，広範な適用可能性，理論的妥当性，科学の多様な分野への貢献を考えると今世紀最大の知的財産の一つに数え上げられるだろう．統計学の基本原理として最尤法は不動の位置を占め，理論的に重要な概念が構築された．一致性，有効性，十分性，不変性などの抽象化された概念としての壮麗さは，数理科学全体の中でも輝き続けている．しかしながら一方では，現代の統計科学の中で数々の観点から最尤法を改良するための，もしくは最尤法に対抗する方法論が展開されている．このような流れの中に最尤法の最適性を巡る理論展開の中で万能主義的な帰結に対する批判が含まれているようだ．データをより柔軟に解析するための方向とその精神が統計学に鼓舞された影響も大きい．(Tukey (1977, 1990), Hoaglin et al. (1985) and Huber (1981))．方法論の提示という視点から眺めると現代の統計学の流れは例えば次のような方向が注目される．

- セミパラメトリックスの実装：Wang and Zhou (1996)；Taylor (1995)；Roeder et al. (1996)；Robins et al. (1995) など．
- ノンパラメトリックスの新しい手法：Donoho and Johnstone (1995)；Donoho et al. (1995)；Hall and Patil (1995)；Nason (1996) など．
- Bayes統計学の再興：Chib (1995)；Good (1996)；Dawid and Mortera (1996)；Young and Pettit (1996)；Kass and Wasserman (1996)；Efron (1996) など．

本論では「概パラメトリックス」(near-parametrics) という観点から最尤法の再考察のひとつのアプローチを紹介する．最尤法を実行するための基本設定は

(1.1) ランダムネスの指定：データ $x_1,\ldots, x_n$ は独立で共通な確率分布 $p(x)$ から生じる．
(1.2) 分布モデルの指定：有限個のパラメーター θ によって (1.1) の確率分布が $p(x, \theta)$ と書かれる．

ここでは確率分布と密度を区別しないで使用する．(1.1) の確率の仮定である，「独立同一性」はマルコフ性へ，また n は連続時間へ緩めることが可能である．また (1.2) も θ は共変量 x とパラメーター β の回帰形に読み替えれば，より多くの統計学の実際を含むだろう．

この二つの仮定から，対数尤度関数が

$$L(\theta) = \sum_{i=1}^{n} \log p(x_i, \theta)$$

と定義される．これよりただちに $L(\theta)$ を最大にする θ の値を最尤推定値とし，帰無仮説 $H_0: \theta = \theta_0$ に対して対立する仮説 $H_1: \theta = \theta_1$ の検定は，それぞれにおける尤度の比の対数をとって $2\{L(\theta_0) - L(\theta_1)\}$ によって実行される．

このように仮定 (1.1) と (1.2) の自然な反映である $L(\theta)$ によって θ の推測が自然に行われる．もし仮定 (1.1) 又は (1.2) が成立しない状況からデータが観測されたら何が起こるだろうか？ 仮定 (1.1) の崩壊は観測のバイアスをもたらし，ランダム割り付けの失敗や欠損データに無視出来ない情報の生成を意味する．同様に仮定 (1.2) の崩れはモデリングの失敗を意味し，外れ値の混入が生じる．仮定の成立が疑わしい時は仮定の自然な反映である尤度関数 $L(\theta)$ もまた大きく狂ってしまうだろう．というよりむしろ仮定 (1.1) または (1.2) の非成立は最尤法を考察する前に，θ の推測自体の意味を失わせるだろう．しかも困ったことに現実のデータに対して，仮定 (1.1) と (1.2) の成立を期待できることは希である．もっと正確な表現は「有限個のデータで (1.1) と (1.2) が成立するという実証は不可能である」．

現実のデータが与えられたとき，分布がパラメトリックモデルに従っていると確信できることは，めったにない．無限個のデータが得られたと想定した時に初めて，仮定 (1.1) の $p(x)$ が仮定 (1.2) の形に記述されていると確かめられるのが精一杯であろう．むしろ，モデルとは，数学的な仮定であり，現実のデータに対しては完全ではないけれど，ほぼモデルに従っていると仮定する方が自然な場合が多い．このように仮定 (1.1) と (1.2) は実証不可能な数学的な設定と考えるべきで，この状況を反映させるため柔らかなモデルを考察しよう．これによって仮定 (1.1) と (1.2) によって定められるモデルと尤度関数 $L(\theta)$ の強固にすぎる関係を柔らかくすることを意図する．

概パラメトリック推測のキーアイディアは，(1.1) と (1.2) を緩め，分布モデルを包むチューブ近傍

$$U_\varepsilon = \left\{ p(x) : \min_{\theta \in \Theta} D(p, p_\theta) < \varepsilon^2 \right\}$$

の中に概モデル $p^*_\theta(x, \omega)$ を構築し，その尤度関数からの最尤法の実行にある．(参照 Eguchi (1997))．ここで ω はデータのランダム性に対して仮定 (1.1) と (1.2) への分布 $p(x)$ の乖離の程度を記述するパラメーターである．$\omega = 0$ のとき，

$$p^*_\theta(x, \omega) = p_\theta(x)$$

とする．実際には，この概モデルが U_ε に入るためには $|\omega| < O(\varepsilon)$ と仮定される．このように提案される尤度関数の形は

$$L^*(\theta, \omega) = \sum_{i=1}^{n} \log p^*(x_i, \omega) = \sum_{i=1}^{n} \{z_i \log p(x_i, \theta) + \lambda(z_i, \theta, \omega)\} \tag{1.3}$$

と与えられる．具体的な $L^*(\theta, \omega)$ の形は目的，設定によっていろいろなバージョンが考えられる．以下で与えられるそれぞれの文脈の中で z_i はデータ x_i の観測状態を示す操作変数を表す．その基本設定では $\omega = 0$ のとき，及び，自明な $\{z_i\}$ に対しては $L^*(\theta, \omega)$ はもとの $L(\theta)$ に帰着される．

この論説では概パラメトリックスの典型的な応用を以下の節で紹介する．2節では局所尤度法を密度推定へ適用する．(参照 Eguchi and Copas (1998))．データ空間の点 x の廻りの情報が知りたい時，例えば，点 x に於ける密度 $p(x)$ の推定を考えよう．この時，(1.3) 式の z_i を核関数 K によって $z_i = K(\omega(x_i - x))$ と与え，

$$\lambda(z_i, \theta, \omega) = z_i \log EK(\omega(X - x))$$

と決めよう．こうして定義された $L^*(\theta, \omega)$ は x を中心とする半径 $1/\omega$ の円外のデータが全て打ち切られた時の尤度に近似される．このように K の滑らかさに応じて，データの $L^*(\theta, \omega)$ への貢献は x への近さに比例して与えられる．その最尤推定値は x と ω に依存するので $\theta(x, \omega)$ と書くとき，これを使って x における密度推定値を

$$\frac{p(x, \theta(x, \omega))}{Z} \qquad (Z \text{は規格化定数})$$

と提案する．この推定値のパラメトリック/ノンパラメトリックの振る舞いの良さが示されるだろう．

3節はニューラル計算の中で自己組織化ルールの話題を概パラメトリックスの観点から紹介する．最尤法は最小 Kullback-Leibler 法と近似的に同値になる．この観点から自己組織化ルールによる Kullback-Leibler ダイバージェンスの変形を提案し，その最小化を考えよう．(1.3) 式の z_i はこの文脈ではデータ x_i が仮定 (1.2) に従っているかどうかを表す潜在変量である．x_i が外れ値の時は，$z_i = 0$ とおき，$z_i = 0$ の確率をニューラルネット理論における自己組織化と関連付けた．具体的な応用として一般化線形モデルの疑最尤法と逸脱関数の自己組織化ルールを考察する．最後に，多変量解析の中で主成分分析についての自己組織化ルールによる方法を紹介する (Higuchi and Eguchi (1998) and Kamiya and Eguchi (1998))．

4節では観測バイアスについての概パラメトリックスについて紹介する (Copas and Eguchi (1998))．選択性の下での尤度の拡張が考えられた．データの中にランダムでない欠損値や割り付けの疑いがある時，(1.3) 式の z_i は観測状態，例えば，欠損であるか，あるいは，どのグループに割り付けられたかなどを表し，λ は z_i と x_i の因果性のモデリングから定められる．特に欠損データがある場合は観測値 x_i が得られた時は $z_i = 1$，失われたときは $z_i = 0$ と定め，$L^*(\theta, \omega)$ の第2項 λ は

$$z_i \log P(z=1|x_i, \omega) + (1 - z_i) \log P(z = 0|x_i, \omega)$$

となる．観測の選択性を記述する ω を 0 の廻りで摂動させると，$L^*(\theta)$ からの最尤推定値の挙動からデータが受けた欠損が θ の推測にどのくらい影響を及ぼすかを知ることができる．

以上の3つの節を通して尤度関数 (1.3) の典型例が紹介された．このようにデータ $\{x_i\}$ に基づく θ の推測に付随する $\{z_i, \omega\}$ を導入してデータの現実にモデルを適合させる方法を，概パラメトリックスの推測と呼ぶ．

2. 局所尤度法

2.1 パラメトリックかノンパラメトリックか？

最初に1変量の密度推定について考察しよう．簡単のため，推定したい密度関数 $p(x)$ は実軸上の滑らかな関数とする．この $p(x)$ をデータ $x_1,\ldots, x_n$ から推定しよう．最も単純なケースは p がパラメトリックなモデルに属する場合だろう：

$$p \in \{p(\cdot, \theta);\ \theta \in \Theta\}. \tag{2.0}$$

例えば p が正規分布 $N(\mu, \sigma^2)$ であれば $\theta = (\mu, \sigma^2)$ の最尤推定値 $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$ によって

$$\hat{p}_0(x) = \frac{1}{\hat{\sigma}}\phi\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right) \tag{2.1}$$

と推定できる．ここで ϕ は標準正規密度関数である．一般形は，$p(x, \hat{\theta})$ で与えられる (cf. Efron (1982))．一方でモデルが仮定できない場合でもノンパラメトリックな方法が提案され，同様に密度推定が可能である．例えば，正規核関数 $\phi(x)$ によって密度推定量が

$$\hat{p}_1(x) = \frac{1}{nh}\sum_{i=1}^{n}\phi\left(\frac{x-x_i}{h}\right) \tag{2.2}$$

と定められる．ここで h はバンド幅を表す．

さて，試験的に6個のデータ $\{-4.2, -3.2, 0.5, 1.2, 2.4, 3.9\}$ について，推定方法 (2.1) と (2.2) による密度関数をグラフに書いてみよう (図1参照)．6個のデータに対して最適な h を決める方法は知られていない．ここでは発見的に $h=1$ と採用した．このようにかなり異なる密度関数が推定されてしまう．しかしデータが与えられたとき，基礎分布が正規分布であるかどうかを明確に決めることはデリケートで困難な問題である．厳密な意味において，データが正規乱数装置からのアウトプット以外では，正規性の仮定は正しくないだろう．現実には，正規性を仮定しても実用的に問題ないと期待できる場合に適用される．データに対する正規性の仮定は本来的には便宜的であるので，ほとんど信頼できる場合もあれば，かなり怪しい場合までいろいろなケースが考えられる．一方で手法の選択肢は (2.1) と (2.2) で代表されるパラメトリック法とノンパラメトリック法しかなく，同じデータからまったく違う結論を導く危険がある．

ちなみに，このデータの正規性の適合度のための Shapiro-Wilk 統計量は .9271 で observed value は50パーセントを少し超える値になるので通常では正規性は棄却しないことになる．適合度検定の宿命として，この標本数では曖昧な結論しか下せない．このようにこの検定の結果から方法 (2.1) か (2.2) のどちらが良いかは，明確な答えが得られない．

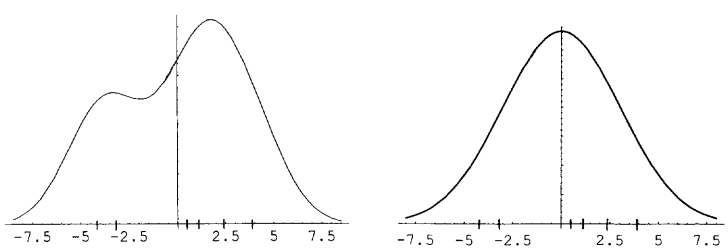

図1． ノンパラメトリック vs パラメトリック密度推定．

2.2 局所尤度の密度推定

パラメトリックな手法とノンパラメトリックな手法を柔らかにつなぐ方法論の提案をしたい(参照 Eguchi and Copas (1998))．その直感的な説明は次のようになる．そもそも密度関数の推定とは 無限個のパラメーター $\{p(x);\ x \in \boldsymbol{R}\}$ の同時推定を意味する．最尤推定値の代入：$p(x, \hat{\theta})$ のように，これを x に対して一様に推定するのは賢いやり方ではないのではないか？点 x における密度 $p(x)$ を各々の x ごとに推定すればより柔軟な方法が考えられるだろう(参照 Hjort and Johns (1996))．

この直感から区間 $I(x, h) = (x-h, x+h]$ のデータだけに制限した，打ち切り尤度を前節の問題に適用しよう．データに対して滑らかにするために $I(x, h)$ の定義関数の代わりに正規核関数 $K(y) = \phi((y-x)/h)$ を使う．後で展開される理論では核関数 K は $K(y) = 1-(y-x)^2/2h^2 + o(h^{-4})$ を満たすものであればよい．数値的な実験を通して通常のノンパラメトリックな核型密度推定と比較して核関数の選択には結果があまり影響されないことが確かめられた．ここでは簡便のため正規核関数を採用した．この時，この打ち切り尤度の最尤推定は方程式系

$$\begin{cases} \dfrac{\sum K(x_i)(x_i-\mu)}{\sum K(x_i)} = \dfrac{EK(X)(X-\mu)}{EK(X)} \\ \dfrac{\sum K(x_i)(x_i-\mu)^2}{\sum K(x_i)} = \dfrac{EK(X)(X-\mu)^2}{EK(X)} \end{cases}$$

の μ と σ^2 に関する解 $\hat{\mu}(x, h)$ と $\hat{\sigma}^2(x, h)$ として与えられる．これを具体的に解くと，

$$(\hat{\mu}(x, h),\ \hat{\sigma}^2(x, h)) = \left(\frac{\tilde{\mu}_{x,h} - x\tilde{\sigma}^2_{x,h}\, h^{-2}}{1-\tilde{\sigma}^2_{x,h}\, h^{-2}},\ \frac{\tilde{\sigma}^2_{x,h}}{1-\tilde{\sigma}^2_{x,h}\, h^{-2}}\right)$$

この解は重み $K(x_i)$ による平均と分散の Bayes 推定の形になることに注意する．

提案する密度推定を，この x における局所最尤推定値の代入によって

$$\hat{p}_h(x) = \frac{1}{Z(h)\,\hat{\sigma}(x, h)}\,\phi\left(\frac{x-\hat{\mu}(x, h)}{\hat{\sigma}(x, h)}\right) \tag{2.3}$$

と与えよう．ここで，$Z(h)$ は規格化定数を表す．作り方から，h が ∞ のとき，重み関数が $K(x_i) \equiv 1$ なので (2.3) は (2.1) に還元される．また，図 2 に示されるように h を小さくすると (2.2) と同じ挙動を示す．このように操作パラメータ h によって滑らかにパラメトリック法とノンパ

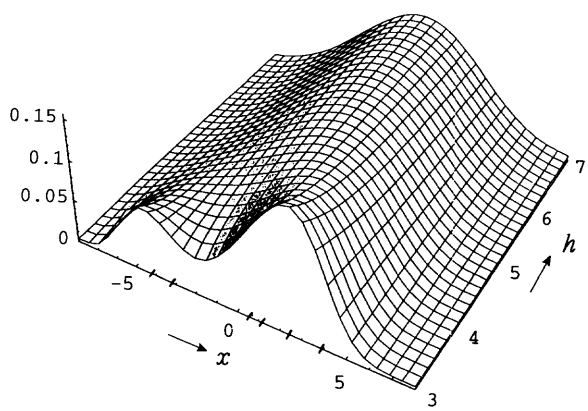

図 2. 局所尤度密度推定 ($3 \le h \le 7$).

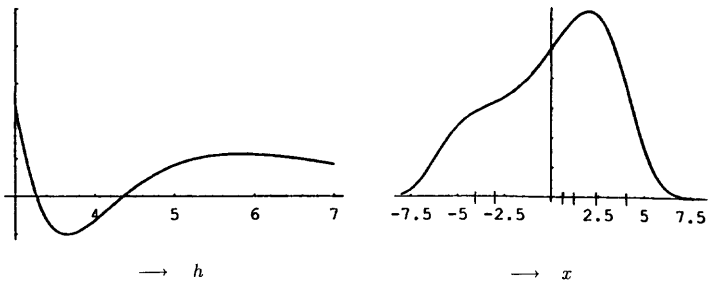

図3. リスク vs バンド幅.

ラ法をつなげた方法になっている.

残された問題点は，どの h を選択すべきか？ という点である．この解決のために，良さの基準を Kullback-Leibler リスク：

$$R(\hat{p}, p) = \boldsymbol{E}_p D(p, \hat{p})$$

に採ろう．ここで，$\boldsymbol{E}_p$ はデータが基礎分布 p からのランダムサンプルであるときの期待値を表し，

$$D(p, \hat{p}) = E\left(\log \frac{p(X)}{\hat{p}(X)}\right)$$

である．これを cross validation によって，リスク $R(\hat{p}, p)$ を推定すると図3の左のようになる．これより，$h_{\text{opt}} = 3.65$ のとき最小リスクとなることが数値的最小化で求められた．帰結される密度推定値は図3の右で与えられる．このような発見的なやり方の性能を考察するための理論を次節で与える．

2.3 概パラメトリックモデル

一般のモデル (2.0) の設定に戻ろう．局所尤度のクラスを推定方程式系

$$\sum K(x_i) \frac{\partial \log p(x_i, \theta)}{\partial \theta} = \sum \xi(K(x_i), EK)\, E\left\{K \frac{\partial \log p(X, \theta)}{\partial \theta}\right\}$$

と定める．ここで，ξ は $\xi(u, u) = 1$ をみたす局所尤度のバージョンを与える関数とする．もし，$\xi(u, v) = u/v$ を採用すれば，局所打ち切り尤度に帰着される．この局所最尤推定量 $\hat{\theta}(x, h)$ を代入し，(一般化) 局所尤度密度推定量 $\hat{p}_h = p(x, \hat{\theta}(x, h))/Z(h)$ が得られる．

帰無仮説 $H: p \in \{p(\cdot, \theta);\ \theta \in \Theta\}$ に対して対立仮説列を漸近的に，

$$A_n: \min_{\theta \in \Theta} D(p, p(\cdot, \theta)) = o(n^{-1-\alpha}) \tag{2.4}$$

と定める検定の問題を考えよう．$\alpha > 0$ の時すべての検定は漸近的に検出力が0となるという意味で検定不可能である．このように，漸近的にさえ，データが，パラメトリックモデル (2.1) に従っているかどうかは検出できない領域がある．この領域はモデルを包むチューブをなすことに注意しながら，$\alpha > 0$ に対して

$$\mathcal{N}_\alpha = \left\{p: \min_{\theta \in \Theta} D(p, p(\cdot, \theta)) = O(n^{-1-\alpha})\right\}$$

を α-order の概パラメトリックモデルと呼ぼう.

命題 1. 仮定：$\xi(1+ah^{-2}, 1+bh^{-2}) = 1+O(h^{-2})$ を置く. このとき, 基礎分布 p が α-order の概パラメトリックモデルに従うとき, Kullback-Leibler リスクは

$$R(\hat{p}_h, p) = R(p(\cdot, \hat{\theta}), p) - \frac{b^2(p, \theta)}{h^2} + \frac{v(\theta)}{nh^4} + o\left(\frac{1}{nh^4}\right)$$

と評価される.

このようにこのリスクはバイアスの 2 乗項と分散項のジレンマの妥協点として放物線の最下点が支持される. 局所尤度密度推定量 $\hat{p}_h$ はリスク最小にするバンド幅 h_{opt} を一意に持つことが証明された. これより 2 節の概正規モデルの例題における $h_{\text{opt}} = 3.65$ の考察は理論的にも支持される (図 3 参照).

多変量データへの拡張は核関数 $K(y) = \phi((y-x)\, S^{-1}(y-x)/h)$ で同様に定義されるが, その性能に対する理論や数値的実験は完成されていない. ここで S は標本分散行列を表す.

3. 自己組織化法則による統計解析

ニューラルネットワークの理論の中で自己組織化法則による算法の提案がなされている. この方法論を尤度解析に適用しよう.

データ $x_1, \ldots, x_n$ による経験分布関数を $\bar{G}_n$ としよう. データに仮定されたモデルを $\{p_\theta(x) : \theta \in \Theta\}$ と書く. 適当なノンパラの密度推定法により推定された密度関数を $\hat{g}_n(x)$ とする時 $p(x, \theta)$ へ Kullback-Leibler ダイバージェンスは

$$D(\hat{g}_n, p_\theta) \propto -\int \hat{g}_n(x) \log p(x, \theta)\, d\mu(x) \simeq -\int \log p(x, \theta)\, d\bar{G}_n(x) = L(\theta)$$

を満たす. このように

$$\tilde{\theta}_n = \operatorname*{argmin}_{\theta \in \Theta} D(\hat{g}_n, p_\theta)$$

は近似的に最尤推定量と一致する.

潜在変数 z_i を

$$z_i = \begin{cases} 1 & (x_i \text{が仮定に従っている時}) \\ 0 & (x_i \text{が仮定から外れている時}) \end{cases}$$

と定めよう. 現実のデータを得る時に $\{z_i\,;\, i = 1, \ldots, n\}$ は観測不能であるが仮に (x_i, z_i) に基づく対数尤度を考えよう：

$$L^\dagger(\theta\,;\, \beta, \eta) = -\beta \sum_{i=1}^{n} \{z_i \zeta(x_i, \theta) + (1 - z_i)\, \eta\}.$$

ここで β, η は正の定数を表し,

$$\zeta(y, \theta) = \log \frac{\hat{g}_n(y)}{f_\theta(y)} + \frac{f_\theta(y) - \hat{g}_n(y)}{\hat{g}_n(y)}.$$

実際には z_i は観測できないので周辺化すると

$$L^*(\theta)=\sum_{i=1}^{n}\log\left(e^{-\beta\eta}+e^{-\beta\xi(x_i,\theta)}\right) \tag{3.1}$$

が得られる．x_i が与えられた時 $z_i=1$ の条件付き確率は

$$w_i(\theta)=P(z_i=1|x_i)=\frac{1}{1+e^{\beta\{\xi(x_i,\theta)-\eta\}}}$$

となることを注意すると式 (3.1) の推定方程式は近似的に

$$\frac{1}{n}\sum_{i=1}^{n}w_i(\theta)\left\{\frac{\partial}{\partial\theta}\log f(x_i,\theta)-\frac{\partial}{\partial\theta}f(x_i,\theta)\right\}=0$$

が得られる．このように条件確率 $w_i(\theta)$ の大きさに応じて重み付けられた最尤法が考えられる．つまり各データ x_i は θ の推測に対して自分の影響を計り，その結果，尤度への貢献度 w_i を決めている．この意味で最尤法の自己組織化バージョンと言える．この文脈において β は逆温度，η はしきい値と呼ばれ $\beta\to 0$ または $\eta\to\infty$ の時に上の方法は通常の最尤法に帰属される．尤度 $L(\theta)$ と Kullback-Leibler ダイバージェンスの同値な関係から自己組織化された尤度 $L^*(\theta;\beta,\eta)$ に対応するダイバージェンスは

$$D_\rho(g,f)=E_g\rho\left(\log\frac{g(x)}{f(x)}+\frac{f(x)-g(x)}{g(x)}\right),$$

で与えられることに注意する．ここで

$$\rho(z)=\frac{1}{\beta}\log\frac{1+e^{\beta\eta}}{1+\exp\{-\beta(z-\eta)\}}. \tag{3.2}$$

このように逆温度 β が 0 に近づくにつれて D_ρ は Kullback-Leibler に近づくことが分かる．

ここで $\hat{g}_n(x)$ について必要なのは x がデータ点 $\{x_i\}$ 上だけであることに注意すると，より良い方法を考察すべきであるが，未解決である．$\hat{g}_n(x)$ が介在した考察なのでこれ以上の一般的な議論は止めよう．

さて我々はもっと具体的なモデルで上のアイディアを実現してみよう．最初に一般化線形モデルについて考察するために次の復習をしよう．モデルは

$$f(y,\theta)=\exp\left\{\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)\right\}$$

と書かれる (McCullagh and Nelder (1989))．x_i を y_i の共変量ベクトルとすると線形予測子

$$\eta_i=\alpha^{\mathrm{T}}x_i\qquad(i=1,\ldots,n) \tag{3.3}$$

がリンク関数 g を通して y_i の期待値 μ_i と結ばれている $\mu_i=g(\eta_i)$ と仮定する．この時偏差関数は

$$D(\mu)=\frac{1}{n}\Sigma d(y_i;\mu_i)$$

で導入される．ここで $d(y;\mu)=2\{\ell(y,y)-\ell(y,\mu)\}$，ただし $\ell(y,\mu)$ は標準パラメーター θ と期待値パラメーター μ の一対一関係を通して $\ell(y,\mu)=\log f(y,\mu)$ と定める．推定方程式は次のように与えられる．

$$X^{\mathrm{T}}V(\alpha)\{y-\mu(\alpha)\}=0$$

ここで $X=(x_1,\ldots,x_n)$, $V(\alpha)=\mathrm{diag}\,(\partial g/\partial\mu(\mu_1)\cdots\partial g/\partial\mu(\mu_n))$ とする．ここで擬似最尤推定量は線形モデル (3.3) の下での偏差関数Dの最小化によって得られることに着目する．$D(\mu)$ は y_i と $g^{-1}(\alpha^{\mathrm{T}}x_i)$ の隔たり $d(y_i,\mu_i)$ の総和をとったものである．$D(\mu)$ の自己組織化バージョンを得るには式 (3.2) で定義された ρ を使って

$$D_\rho(\mu)=\frac{1}{n}\sum_{i=1}^{n}\rho(d(y_i,\mu_i))$$

とする．推定方程式は

$$X^{\mathrm{T}}V(\alpha)\,W(\alpha)\{y-\mu(\beta)\}=0$$

となる．ここで，

$$W(\beta)=\mathrm{diag}\left(\frac{1}{1+\exp[\beta\{d(y_i,\mu_i)-\eta\}]}\right)_{i=1,\ldots,n}$$

となる条件確率行列である．各々の対角成分は大きな隔たり $d(y_i,\mu_i)$ に対して小さい値を返す．ここで提案された方法は Huber の M-推定と密接な関係にある．しかしながら M-推定と本質的に異なる点は対数尤度関数 $(1/n)\sum\ell(y_i,\mu)$ でなく，偏差関数Dに基づく変形を行った点にある．これにより尤度関数の外れ値の影響を相対的に抑制できるので高い性能が期待されるが，実証は今後の課題である．ただし正規回帰モデルの下では上の方法と M-推定は一致することに注意する．

3.1 自己組織化による主成分分析

次に主成分分析の自己組織化について紹介しよう (参照 Amari (1977), Oja (1982), Xu and Yuille (1995), Higuchi and Eguchi (1998), Kamiya and Eguchi (1998))．p 変量のデータ $x_1,\ldots,x_n$ が得られた時主成分ベクトル $\hat{\gamma}$ は

$$\sum_{i=1}^{n}\{\gamma^{\mathrm{T}}(x_i-\bar{x})\}^2$$

の γ の単位球面上における最大化によって求められる．ここで $\bar{x}$ は標本平均ベクトルとする．統計学においては，標本分散行列

$$S=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^{\mathrm{T}}$$

とおいて

$$n\sum_{i=1}^{n}\{\gamma^{\mathrm{T}}(x_i-\bar{x})\}^2=\gamma^{\mathrm{T}}S\gamma$$

の関係から直ちに $\hat{\gamma}$ がSの第1固有ベクトル (最大固有値 λ_1 に対応する固有ベクトル) と導かれ，k-主成分ベクトルも同様に導出される．一方でニューラルネット理論では各入力 $(x_i-\bar{x})$ に対して結合係数ベクトル γ の線形和 $\gamma_j^{\mathrm{T}}(x_i-\bar{x})$ を出力すると解釈する．

$$z(x_i,\gamma)=\|x_i-\bar{x}\|^2-\{\gamma^{\mathrm{T}}(x_i-\bar{x})\}^2$$

と置くと

$$\hat{\gamma}=\underset{\gamma:\,\gamma=1}{\mathrm{argmin}}\frac{1}{n}\sum_{i=1}^{n}z(x_i,\gamma)$$

と書けることに注目しよう．更に $z(x_i, \gamma)=0$ は入力ベクトル $x_i-\bar{x}$ が γ と同一直線上にあることを意味する．以上の準備から自己組織化バージョンは，式 (3.2) で与えられる ρ を使って

$$\hat{\gamma}_* = \underset{\gamma:\gamma=1}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\rho(z(x_i, \gamma))$$

と定められる (参照 Xu and Yuille (1995))．この最小化を求めるために次のアルゴリズムを提案する．

$$\gamma_{t+1} = \text{1st eigenvalue of } S(\gamma_t),$$

ここで $S(\gamma)$ は次で定められる重み付き分散行列である：

$$S(\gamma)=\sum_{i=1}^{n} w(x_i-\bar{x}, \gamma)(x_i-\bar{x})(x_i-\bar{x})^{\mathrm{T}},$$

ただし重み関数は

$$w(x, \gamma)=\frac{1}{1+e^{\beta\{z(x,\gamma)-\eta\}}}.$$

この $\rho(z)$ が凹関数であることから任意の初期ベクトル γ_0 に対して目的関数 $L(\gamma)=\sum\rho(z(x_i, \gamma))/n$ を一様に各ステップ毎に減少させていることが示せる．あとで紹介される数値実験の中では10回程度で収束が認められた．結局，$\hat{\gamma}^*$ を求める推定方程式は

表1．土壌データ．

土壌番号	泥 (%)	粘土 (%)	有機物 (%)	ペーハー
1	13.0	9.7	1.5	6.4
2	10.0	7.5	1.5	6.5
3	20.6	12.5	2.3	7.0
4	33.8	19.0	2.8	5.8
5	20.5	14.2	1.9	6.9
6	10.0	6.7	2.2	7.0
7	12.7	5.7	2.9	6.7
8	36.5	15.7	2.3	7.2
9	37.1	14.3	2.1	7.2
10	25.5	12.9	1.9	7.3
11	26.5	14.9	2.4	6.7
12	22.3	8.4	4.0	7.0
13	30.8	7.4	2.7	6.4
14	25.3	7.0	4.8	7.3
15	31.2	11.6	2.4	6.5
16	22.7	10.1	3.3	6.2
17	31.2	9.6	2.4	6.0
18	13.2	6.6	2.0	5.8
19	11.1	6.7	2.2	7.2
20	20.7	9.6	3.1	5.9

$$S(\gamma)\,\gamma = \lambda\gamma, \qquad \lambda = \gamma^{\mathrm{T}} S(\gamma)\,\gamma$$

なので上のアルゴリズムは一般化線形モデルの擬似最尤推定量を求める反復再重み付け最小2乗法 (IRWLS) と全く同様なアイディアである．各ステップで古典的な主成分を求め，反復していることになっている．以上がバッチ形式のデータに対する紹介であるが，一方で，オンライン形式のデータでは，学習アルゴリズム

$$\gamma_{t+1} = \gamma_t + \alpha_t w(x_t, \gamma_t)\,\gamma_t^{\mathrm{T}}(x_t - \bar{x}_t)\{x_t - \bar{x}_t - \gamma_t^{\mathrm{T}}(x_t - \bar{x}_t)\,\gamma_t\}$$

で与えられる．ここで α_t は学習比 (ステップサイズ) を表す．さて古典的な主成分ベクトル $\hat{\gamma}$ と自己組織化による $\hat{\gamma}^*$ の比較のためバッチデータに注目しよう．最初に Kendall の多変量解析の土壌のデータについて考察しよう．データは 40 地点の土質の泥，粘土，有機物，ペーハーの 4 項目について与えられている (表 1，Kendall (1975) の Tables 2.1, 2.4 参照)．第 1 主成分ベクトルは

古典 PCA = {0.955785, 0.293681, 0.0149718, 0.00131652}
ニューラル PCA = {0.957297, 0.288627, 0.0121217, 0.0113762}

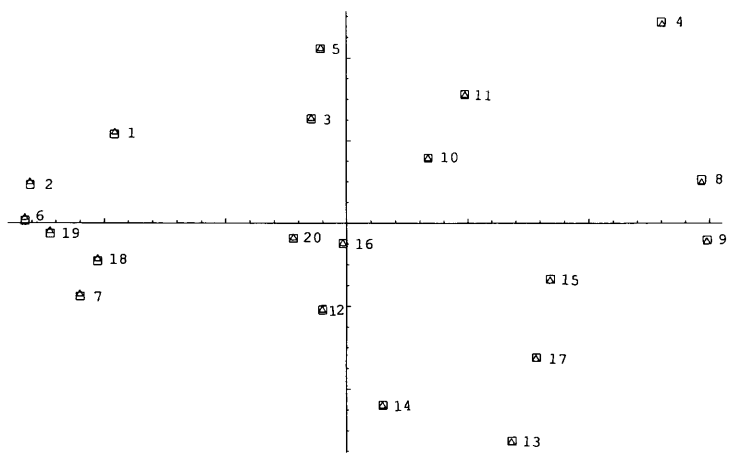

図 4. 土壌データの主成分得点の 2 次元プロット (□古典的 PCA，△ニューラル PCA)．

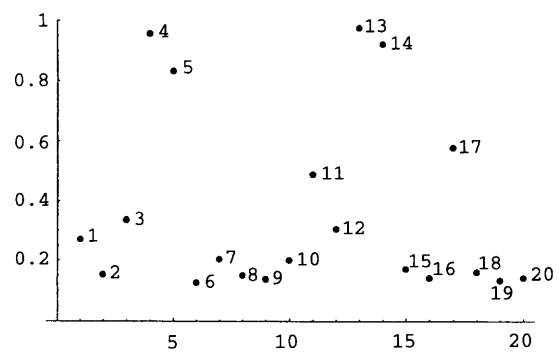

図 5. (a) 重み $w(x_i, \gamma)\,(i = 1,\ldots, 20)$ のプロット．

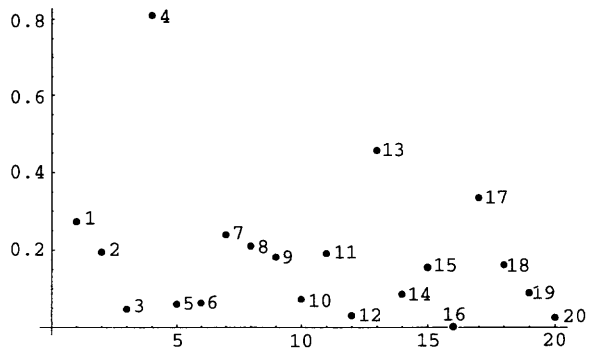

図5. (b) 影響統計量のプロット.

寄与率は第1主成分で91.7%, 第2主成分までとれば98.6%になる. 主成分分析得点が次のようになる (図4参照). このようにニューラル版も従来の方法も殆ど同じ主成分得点を与えている. しかしながら重み関数の図5(a)を見ると土壌番号4, 13, 14, 5, 17, 11の順で外れ値である確率 $(1-w_i)$ が大きいことを示す. 一方でCritchley (1985) の影響解析によると感度統計量は図5(b)のように14と5に対して低い値を示した. 全ての主成分得点を土壌5と14について計算すると

$$(-1.060,\ -4.223,\ -0.068,\ 0.253),\ (1.453,\ 4.401,\ -1.658,\ 0.416)$$

となり, ほとんど原点対称の位置にある. これはお互いの影響を消し合ってしまうマスク効果が働いていると考えられる. これよりニューラルPCAはこのようなケースもうまく特定していると言える.

3.2 理論

この節では自己組織化ルールによる主成分分析を理論的側面から考察する (参照 Kamiya and Eguchi (1998)).

非負値 z 上で定義された単調増加の凹関数 $\rho(z)$ で $\rho(0)=0$, $\rho'(0)=1$ を満たすクラスを $\mathcal{P}$ と置く. この $\rho(z)$ を使って

$$L_\rho(\gamma;\ G)=E_G\{\rho(z_\gamma(X-\mu_G))\}$$

と定義しよう. ここで μ_G は G の平均ベクトル,

$$z_\gamma(y)=\|y\|^2-(\gamma^{\mathrm{T}}y)^2.$$

この L_ρ を使い, 球面上への統計汎関数を

$$T_\rho(G):=\operatorname{argmin}\{L_\rho(\gamma,\ G):\|\gamma\|=1\}$$

と定義する. 統計量 $T_\rho(\bar{G}_n)$ について ρ をクラス $\mathcal{P}$ に制限して考察を進めよう. ただし, $\bar{G}_n$ はデータ $x_1,\ldots,x_n$ からの経験分布関数とする. 特に $\rho(z)=z$ の時は $T_\rho(\bar{G}_n)$ は古典的主成分ベクトルを与え, (3.2) 式で定める $\rho(z)$ は自己組織化バージョンを与える. 次の命題では一般 ρ についてFisher一致性を示し, 影響関数

$$\mathrm{IF}(x,\ T_\rho,\ G):=\lim_{\varepsilon\searrow 0}\frac{T_\rho(G_\varepsilon)-T_\rho(G)}{\varepsilon}$$

を具体的に与える．ここで $G_\varepsilon = (1-\varepsilon)\,G + \varepsilon\delta_x$ (δ_x は点 x で重さ 1 を持つ分布)．

命題 2. 基礎分布 G が p 変量正規分布 $N(\mu, \Sigma)$ と仮定する．ここで分散行列 Σ のスペクトル分解を

$$\Sigma = (\gamma_1 \cdots \gamma_p)\,\mathrm{diag}\,(\lambda_1, \ldots, \lambda_p)\,(\gamma_1 \cdots \gamma_p)^{\mathrm{T}}$$

と書く．ここで

$$\lambda_1 > \cdots > \lambda_p, \quad \gamma_i^{\mathrm{T}}\gamma_j = \delta_{ij}(\text{Kronecker's delta}).$$

この時，勝手な $\rho \in \mathcal{P}$ に対して

(1) $T_\rho(N(\mu, \Sigma)) = \gamma_1$

(2) $\mathrm{IF}\,(x, T_\rho, N(\mu, \Sigma)) = \psi\left(\frac{1}{2}\sum_{j=2}^{p}\|a_j(x)\|^2\right) a_1(x) \sum_{j=2}^{p} \frac{\lambda_j a_j(x)}{\lambda_j^*(\lambda_1 - \lambda_j)}\,\gamma_j$.

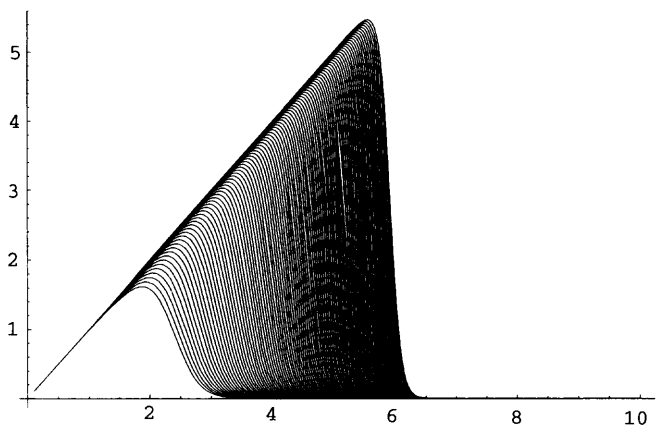

図 6. (a) $\sqrt{z}\,\psi(z;\beta,\eta)$; $\beta = 1$, $0.1 \le \eta \le 15$.

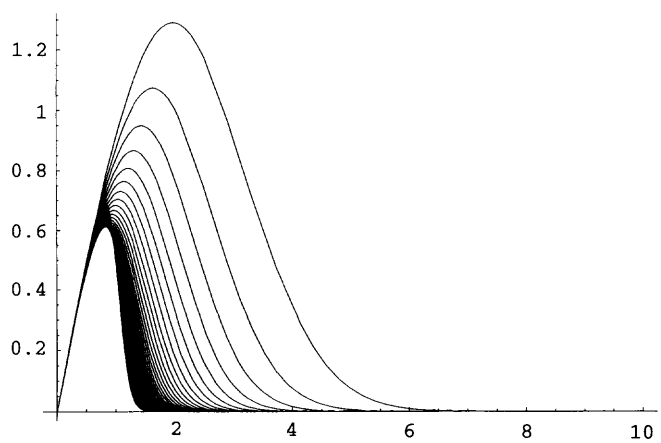

図 6. (b) $\sqrt{z}\,\psi(z;\beta,\eta)$; $\eta = 1$, $0.1 \le \beta \le 10$.

ここで $a_j(x)=\gamma_j^T x$,

$$\lambda_j^*=E_{N(\mu,\Sigma)}\left\{\psi\left(\frac{1}{2}\sum_{j=2}^{p}\|a_j(x)\|^2\right)a_j^2(x)\right\}$$

とする.

この命題から直ちにロバスト性と相対効率の考察が導かれる. もし $\sup_{z>0}\sqrt{z}\,\psi(z)=c_0$ ならば

$$\sup_{x:a_1(x)^2\le c_1^2}\|\mathrm{IF}(x,\ T_\rho,\ N(\mu,\ \Sigma))\|^2\le c_1^2c_0^2\sum\frac{\lambda_j^2}{\lambda_j^{*2}(\lambda_1-\lambda_j)^2}$$

である. これより T_ρ がロバストである必要十分条件として

$$\sup_{z>0}\sqrt{z}\,\psi(z)<\infty$$

が得られた. (3.2) 式から作られる $\psi(z)$ に対して, 任意の $\beta>0$, $\eta>0$ に対して上の条件は満たされている (図 6(a), (b)を参照).

4. 不完全観測のバイアスの感度分析法

4.1 ミッシングデータの問題

大規模な標本調査においてデータがミッシングを伴うことは避けられない現象である. 近年, アンケート調査において回答の回収率の低下が問題になっているが, これは複雑化した社会の構造や多様化する価値観や高度な情報化による影響が原因として考えられる. しかしながら, 予め回収率を上げるための工夫や未回収データの 再調査など Follow-up 解析は困難であったり, 調査の設計の骨格が崩れる危険が生じる. こんな状況で, 「不完全データの解析」がますます重要となっている.

“ミッシングデータ”は我々に何を語りかけているのだろうか? もしミッシングデータが

純粋なランダム性

から生じたのであれば, 得られたデータだけに基づいて解析を実行しても問題は生じない. このとき, ミッシングデータは「私を完全に無視していいのよ.」と語りかけている. しかし, 何らかの無視できない要因でミッシングデータが

偏ったランダム性

から生じたのであれば, 観測されたデータだけに基づいた解析は大きな偏りを生じるだろう. このときはミッシングデータは「私を無視すると痛い目に会うわよ.」と警告しているのだ.

この問題は, 量的であり同時に質的でもある. 厳密には, ミッシングデータがたった一つでも発生した時点で観測のランダム性に対する仮定は再考されるべきだろう. しかし現実には, 調査が大規模で長期にわたるほど, その変更に関わるコストが法外に高価だったり, 原理的に不可能であることが多い. このようにミッシングの伴うデータの解析は避けては通れない問題であり, さまざまな困難な点が報告されている (参照 Little and Rubin (1987)).

この節を通して, 観測される変数の中で共変量は完全に観測されているが, 反応変数の観測は不完全であると仮定する. この設定のもとでミッシングデータの発生するメカニズムをモデル化する

選択パラメーター

を導入する．しかし，このパラメーターを推定する如何なる方法も misspecification に対して誤った結論を導く危険が高いことを示している．この考察からミッシングの観測選択性に対する感度分析アプローチを提案する．これによりミッシングデータを無視することの影響を定量的に計ることができる．

4.2 観測の選択パラメーター

観測変数 y の分布を $f(y, \theta)$ とし，観測値 $y_1,\ldots, y_N$ を得て θ の推測を目的としよう．多くの場合は θ は共変量 x とパラメーター β によって $\theta = \theta(x, \beta)$ とデザインされている．このように共変数 $x_1,\ldots, x_N$ は得られたが，しばしば 観測のプロセスに於いて $y_1,\ldots, y_N$ の中からミッシングデータが生じ，その際にランダム性の仮定が壊された恐れのある状況について考察しよう．

y の観測状態を表す量 r は $r = 1$ のとき y が観測され，$r = 0$ のとき y の欠測を表す．このように，我々の観測データは $(y_1, r_1),\ldots, (y_N, r_N)$ の形で得られる．もし r が θ の情報を持っていないなら，ミッシングデータを無視して，観測できたデータだけに基づいて θ を推測すれば良い．以後，このデータを並び替えて $y_1,\ldots, y_n$ と書く $(n = \sum r_i)$．

いまから，(y, r) の分布が $f(y, \theta)\, p(r, \phi_y)$ と書かれたとき，

$$\phi_y = \phi + I_\phi^{-1/2} \Omega I_\theta^{-1/2} \frac{\partial}{\partial \theta} \log f(y, \theta)$$

と仮定しよう．ここで Ω は θ と ϕ パラメーター空間を結ぶ行列である．もし $\Omega = 0$ ならば θ と ϕ は独立に推測できるのでミッシングデータを無視してよいが，$\Omega \neq 0$ ならミッシングデータに無視できない情報が生じる恐れがある．

この無視できない情報を正しく評価するためには Ω を正しく目盛り付ける必要がある．このためには

$$p(r, \phi_y) \text{ がどの位 } y \text{ に関連しているか？}$$

を測ればよいだろう．このようにデータがミッシングされるときの選択指数を次で与える：

$$\eta^2(\Omega) = \mathrm{Var}\left(\log \frac{p(r, \phi_y)}{Ep(r, \phi_y)}\right)$$

ここでこの節を通して E は y の周辺分布 $f(y, \theta)$ に関する期待値を表す．選択指数 η は，

$\eta = 0$ は “missing at random”
$\eta = \log 2 \simeq 0.7$ は 2 倍のオッヅ
$\eta = 1$ は厳しい選択力がサンプルに伴った

ことを意味するだろう．

観測状態が完全のときの条件付き分布

$$f(y, \theta) \frac{p(1, \phi_y)}{Ep(1, \phi_y)} \tag{4.1}$$

からのランダムサンプル $(y_1,\ldots, y_n)$ に対して，Ω を固定したプロファイル尤度関数は

$$L(\theta, \Omega) \propto \sum_{i=1}^{n} \log f(y_i, \theta) + \sum_{i=1}^{n} \log p(1, \phi_{y_i}) \tag{4.2}$$

となる．これから求められた最尤推定量 $\hat{\theta}(\Omega)$ を (4.2) に代入するとプロファイル尤度関数は

$$\frac{1}{n}\frac{\partial}{\partial\Omega}L(\hat{\theta}(\Omega),\Omega)|_{\Omega=0}=0,$$
$$\frac{1}{n}E\left(\frac{\partial^2}{\partial\Omega\partial\Omega}L(\hat{\theta}(\Omega),\Omega)|_{\Omega=0}\right)=0$$

を満たすので，プロファイル尤度関数は $\Omega\simeq 0$ において平坦となり，モデル化に依って非常に敏感に結果が変ってしまう．このように，如何なる (y,r) のモデル化に対しても観測バイアスが生じる恐れがある場合において Ω の点推定は最尤法では大きな誤りを犯す危険性がある．この考察から Ω に対する影響分析アプローチを考えよう．

4.3 情報を持つ欠損データの影響

反応変数 y に同程度の選択力が加わったとき，ミッシングデータの存在を無視したら，どの位，影響を受けるか調べてみよう．前節で与えた選択指数関数 $\eta(\Omega)$ を使って

$$\mathcal{S}_\eta=\{\Omega|\eta(\Omega)=\eta\}$$

と定め，選択等位面と呼ぼう．このように，適当に η を選び，$\mathcal{S}_\eta$ のすべての要素 Ω に対して実際に観測されたデータ $(y_1,\ldots,y_n)$ があたかも $r=1$ の条件付き分布からのランダムサンプルと考えたときの方程式 (4.1) からのプロファイル最尤推定量を $\hat{\theta}(\Omega)$ としよう．このように θ のパラメーター空間のなかに部分空間

$$\Theta_\eta=\{\hat{\theta}(\Omega)|\Omega\in\mathcal{S}_\eta\}$$

が張られる．この Θ_η が選択指数が η のときの θ の推測にミッシングデータを無視することの影響のすべての可能性を与えている．これより，我々がミッシングデータを無視して得た結論 $\hat{\theta}(\Omega)|_{\Omega=0}$ の危険性の程度を知ることができる．具体的な詳細を次の例題で調べよう．

4.4 日本人の国民性調査

第9回の日本人の国民性調査 (Sakamoto (1993)) の中で性別及びどの地方に住んでいるか分かっている人たちに次の質問がなされた．

社会の慣行に対して自分の考え方との関係を次の3つから選んで下さい

$j=1$：自分の考えを押し通す
$j=2$：慣習に従う
$j=3$：状況による

共変量は性 $u=1$，2，地方 $l=1,\ldots,9$ で表2の分割表が得られた．

このミッシングデータの影響を解析しよう．この分割表に対して $2\times 9\times 3$ の3相対数線形モデルに記述したとき，性，地方は加法的であることが支持された．これから，性が u，住居は l 地方の人が項目 j と答える確率 θ_{ulj} に対して

$$\log\theta_{ulj}=\mu+\lambda_u+\lambda_l+\lambda_j+\lambda_{uj}+\lambda_{lj} \tag{4.3}$$

と書かれる．一方で選択パラメーターを導入したモデルは

$$\log\theta_{ulj}(\Omega)=\log\theta_{ulj}+\frac{\tilde{\pi}_{ul}}{\sqrt{\hat{\theta}_{ulj}}}\omega_j \tag{4.4}$$

表2. 日本人の国民性調査.

地方	男性 1	2	3	欠落	女性 1	2	3	欠落
北海道	12	16	20	23	10	19	20	27
東北	22	22	21	39	13	29	41	39
関東	68	82	77	235	65	93	106	203
中部(東)	19	42	20	31	22	48	32	22
中部(西)	26	28	25	44	23	46	31	44
近畿	40	41	45	122	55	53	67	111
中国	21	18	29	31	10	22	26	26
四国	17	11	9	17	14	18	12	11
九州	30	32	27	67	34	49	34	63

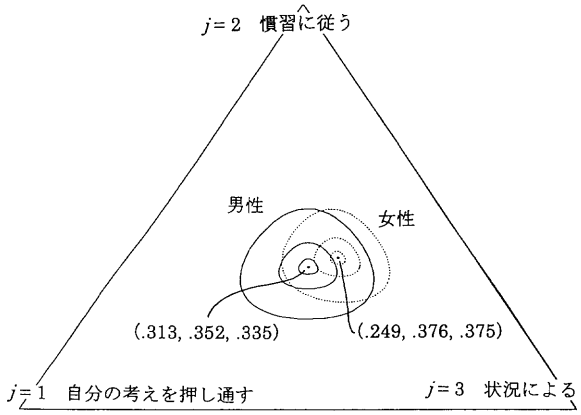

図7. 第9回日本人の国民性調査.

ここで $\Omega=(\omega_1, \omega_2, \omega_3)$, $\hat{\pi}_{ul}$ はミッシング確率の推定値, $\hat{\theta}_{ulj}$ はモデル (4.3) での最尤推定値を表す. これは, 選択プロセスが指数分布に従うことを仮定して導出される (図7参照). 性 u の群が項目 j と答える確率は

$$p_{uj}=\frac{\sum_{l=1}^{9}N_{ul}\theta_{ulj}}{\sum_{j=1}^{3}\sum_{l=1}^{9}N_{ul}\theta_{ulj}}$$

$(u=1, 2, j=1, 2, 3)$ なので, θ_{ulj} に推定値を代入して $\hat{p}_{uj}$ を求めよう. 次に, それぞれ総合選択指数 $\eta=0.1$, 0.3, 0.7 を満たす選択パラメーター Ω のつくる $\theta_{ulj}(\Omega)$ の推定値を代入してトレースされる等高線が3つできる. プログラムの実行は S-plus で行われた. 球面 $\{\Omega=(\omega_1, \omega_2, \omega_3)\,|\,\omega_1^2+\omega_2^2+\omega_3^2=\eta^2\}$ 上のすべての点で $\theta_{ulj}(\Omega)$ を評価することは大変そうであるが, 実際にはモデル (4.3) から (4.4) への変更は対数線形モデルの形を保つので容易に実行可能である.

$\eta=0$ の時の結論は男性と女性を比較した時, ほぼ似た傾向を示す. 相違点は"自分の考えを押し通す"と答えた確率が男性グループは .313, 女性グループは .249 となり, より男性の方が高い傾向が見られる. $\eta=0.1$ 以内の範囲ではこのミッシングデータの影響は前述の結論を覆すほど重大ではないことが分かった. 従ってこのケースは, 我々は幸運にもこれ以上ミッシングデータの影響は考慮しなくて良いことを示唆するだろう. この示唆を支持する更にいろいろ

な経験的証拠をつみ重ねる必要がある.

参 考 文 献

Amari, S.-I. (1977). Neural theory of association and concept formation, *Biol. Cybernetics*, **26**, 175-185.

Chib, S. (1995). Marginal likelihood from the Gibbs output, *J. Amer. Statist. Assoc.*, **90**, 1313-1321.

Copas, B. J. and Eguchi, S. (1998). Sensitivity approximations for selectivity bias in observational data analysis, Research Memo., No. 660, The Institute of Statistical Mathematics, Tokyo.

Critchley, F. (1985). Influence in principal components analysis, *Biometrika*, **72**, 627-636.

Dawid, A. P. and Mortera, J. (1996). Coherent analysis of forensic identification evidence, *J. Roy. Statist. Soc. Ser. B*, **58**, 425-443.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage, *J. Amer. Statist. Assoc.*, **90**, 1200-1224.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymptopia ?, *J. Roy. Statist. Soc. Ser. B*, **57**, 301-337.

Efron, B. (1982). Maximum likelihood and decision theory, *Ann. Statist.*, **10**, 340-356.

Efron, B. (1996). Empirical Bayes methods for combining likelihoods, *J. Amer. Statist. Assoc.*, **91**, 538-550.

Eguchi, S. (1997). Near-parametric inference, Workshop organised by C. M. Bishop at Newton Institute, Cambridge University.

Eguchi, S. and Copas, B. J. (1998). A class of local likelihood methods and near-parametric asymptotics, *J. Roy. Statist. Soc. Ser. B*, **60**, 709-724.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics, *Philos. Trans. Roy. Soc. London Ser. A*, **222**, 309-368.

Good, I. J. (1996). When batterer becomes murderer, *Nature*, **381**, p. 481.

Hall, P. and Patil, P. (1995). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators, *Ann. Statist.*, **23**, 905-928.

Higuchi, I. and Eguchi, S. (1998). The influence function of principal component analysis by self-organizing rule, *Neural Computation*, **10**, 1435-1444.

Hjort, N.L. and Jones, M. C. (1996). Locally parametric nonparametric density estimation, *Ann. Statist.*, **24**, 1619-1647.

Hoaglin, D. C., Mosteller, F. and Tukey, J. W. (ed.) (1985). *Exploring Data Tables, Trends, and Shapes*, Wiley, New York.

Huber, P. J. (1981). *Robust Statistics*, Wiley, New York.

Kamiya, H. and Eguchi, S. (1998). A class of principal component vectors (in preparation).

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules, *J. Amer. Statist. Assoc.*, **91**, 1343-1370.

Kendall, M. G. (1975). *Multivariate Analysis*, Griffin, London.

Little, R. J. A. and Rubin, D. A. (1987). *Statistical Analysis with Missing Data*, Wiley, New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed., Chapman and Hall, London.

Nason, G. P. (1996). Wavelet shrinkage using cross-validation, *J. Roy. Statist. Soc. Ser. B*, **58**, 463-479.

Oja, E. (1982). A simplified neuron model as a principal component analyzer, *J. Math. Biol.*, **15**, 267-273.

Robins, J. M., Hsieh, F. and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates, *J. Roy. Statist. Soc. Ser. B*, **57**, 409-424.

Roeder, K., Carroll, R. J. and Lindsay, B. G. (1996). A semiparametric mixture approach to case-control studies with errors in covariables, *J. Amer. Statist. Assoc.*, **91**, 722-732.

Sakamoto, Y. (1993). A study of the Japanese national character: The ninth nationwide survey, Research Memo., No. 572, The Institute of Statistical Mathematics, Tokyo.

Taylor, J. M. G. (1995). Semi-parametric estimation in failure time mixture models, *Biometrics*, **51**, 899-907.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison Wesley, Massachusetts.

Tukey, J. W. (1990). Data-based graphics: Visual display in the decades to come, *Statist. Sci.*, **5**, 327-339.

Wang, W. and Zhou, M. (1996). Semi-parametric estimation of disequilibrium models, *Econometric*

Rev., **15**, 445-462.

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

Young, K. D. S. and Pettit, L. I. (1996). Measuring discordancy between prior and data, *J. Roy. Statist. Soc. Ser. B*, **58**, 679-689.

Proceedings of the Institute of Statistical Mathematics Vol. 47, No. 1, 29-48 (1999)

Near Parametric Inference — Towards Flexible Modeling —

Shinto Eguchi

(The Institute of Statistical Mathematics)

This paper introduces a near-parametric inference to extend a working area of the usual likelihood method to a wider area where the proposed method performs well against a slight departure from assumptions for a parametric model with possible directions. A diversity of semiparametric approaches has been established in order to bridge a gap between parametric and nonparametric methods. In this approach along semiparametrics the key idea is to enlarge a parametric model into the tube neighborhood so that it may relax the inflexible relation of the parametric model with the likelihood function.

Three typical applications to near-parametric inference are given as follows: (1) Density estimation by local likelihood method is discussed, where a given model is enlarged according to a data point of which density is to be estimated. In effect a structure of incomplete observation is mounted by kernel function. In this context the structure becomes vanishing as the bandwidth becomes infinity. A large bandwidth asymptotics is discussed under near parametric situation where the underlying distribution is asymptotically reduced to the parametric one. (2) In neural computational algorithm we introduce a self-organizing rule to likelihood method by considering a latent variable indexing whether each observation comes from the assumptions in the parametric setting. In particular we present a special application to principal component analysis. The proposed algorithm is of EM-type, where the conditional probability that the respective observation is well controlled given the observation is imputed in the E step ; the principal component vector on the sample covariance matrix by weighting the conditional probabilities is calculated in the M step. (3) We introduce a sensitivity approach to observational bias by modeling a selectivity parameter. The key point is that the selectivity parameter is not estimated but assessed the influence against the observational possible bias deviate from pure randomness assumption under missing or allocation sampling. A selectivity index invariant with the selectivity parametrization gives a reasonable assessment whether the observational assumption is broken down.

Through these applications an advantageous point is commonly addressed such that near parametric inference keeps the same efficiency as the parametric inference reasonably, and performs well against the departure from parametric setting.

Key words : Local likelihood, near parametrics, observational bias, principal component analysis, selectivity parameter.