数理解析研究所講究録
第1703巻 2010年 159-179

Note on robust estimation and model selection in a contaminated mixture model

筑波大学大学院・数理物質科学研究科　小林　裕子 (Yuko Kobayashi)
Graduate School of Pure and Applied Sciences
University of Tsukuba

筑波大学・数学系　矢田　和善 (Kazuyoshi Yata)
Institute of Mathematics
University of Tsukuba

筑波大学・数学系　青嶋　誠 (Makoto Aoshima)
Institute of Mathematics
University of Tsukuba

要旨

データを生み出す未知の分布と想定するモデルとの距離を Kullback-Leibler 情報量で測るとき，データの個数が十分であれば，モデルのパラメータは最尤推定量で与えられる．得られたモデルを予測の意味で評価するには，将来観測されるデータをモデル構築に用いたデータで代用することによる偏りを，適切に評価して補正する必要がある．こうして Akaike (1973) は，モデル選択の一つの基準として赤池情報量規準 (AIC) を提唱した．

一般に，標本数が十分ではないとき，最尤推定量は異常値の影響を受けやすく，よい推定ができないことが知られている．本論文では，異常値に対して頑健なパラメータ推定を考える．データを生み出す未知の分布とモデルとの距離を，K-L 情報量に替えて β-ダイバージェンスで測る．真の分布は異常値と潜在分布の混合分布であると考え，異常値に影響を受けずに潜在分布を推定することを考える．潜在分布には多次元混合正規分布モデルを仮定して，パラメータを推定するための更新アルゴリズムを与える．得られたモデルを予測の意味で評価するために，偏りを補正した新しいモデル選択の基準を提案する．最後に，クラスタリングへの応用を考える．

1　はじめに

データから本質的な情報を抽出し，知識獲得を目指すデータ解析や予測・制御において，モデル選択が重要な役割を果たす．モデルが特定されると，予測・制御，情報抽出，検定，リスク評価，意思決定など，様々な形式の推論を演繹の枠組みで論じることができるようになる．したがって，直面する現実の問題を解く

鍵は，観測されたデータを生み出す真の分布とモデルとの距離を如何に測るか，そして，予測の意味で合理的なモデルを如何に選ぶかということにある．

Akaike (1973) は，データを生み出す真の分布が想定するモデルの族に含まれるという仮定の下で，真の分布からのモデルの近さを Kullback-Leibler 情報量 (K-L 情報量) で測り，最尤推定量 (MLE) を用いて構築されるモデルのモデル選択基準として赤池情報量規準 (AIC) を提唱した．竹内 (1976) は，真の分布が想定するモデルの族に含まれるという仮定をはずして，AIC に替わるモデル選択基準として TIC を提案した．AIC と TIC は共に，モデルを規定するパラメータに MLE を代入することでモデルが構築され，K-L 情報量に基づいてモデルが選択されるという共通点をもつ．Konishi and Kitagawa (1996) は，MLE の枠組みをはずした一般的なモデル構築を考え，それらから K-L 情報量に基づいてモデル選択するための GIC とよばれる基準を提案した．一方，Akaike (1977), Schwarz (1978) は，モデルの近さを測る方法ではなく，MLE で構築される候補モデルのなかで周辺尤度が最大になるモデルを選ぶための BIC とよばれる基準を提案した．また，Ishiguro, Sakamoto and Konishi (1997) は，ブートストラップ法を用いた数値的なアプローチに基づいてモデルを選択するための EIC とよばれる基準を提案した．

一般に，MLE は異常値の影響を受けやすく，K-L 情報量はこの欠点をあわせもつ．この欠点に対して，Huber (1964) は，MLE に代わる頑健な推定量として M-推定量を提案した．これは MLE そのものを改良するのではなく，仮定するモデル自体を改良するアプローチである．また，Hampel (1968, 1974) は，影響関数を考えることにより，異常値に対する推定量の頑健性を捉えた．Basu et al. (1998) は，密度のべき乗をスコア関数にかけることで異常値の影響を小さくして，β-ダイバージェンスとよばれる新しい距離を提案した．Scott (2001) は，β-ダイバージェンスの特別な場合として，密度の 2 乗である L_2 距離を推定に適用した．Eguchi (2006) は，β-ダイバージェンスを拡張させ，ある関数 U によって定義される距離として U-ダイバージェンスを提案した．Fujisawa and Eguchi (2006) は，β-ダイバージェンスに基づいて，正則な場合の一次元混合正規分布のパラメータ推定のためのアルゴリズムを提案した．ここで，正則な場合とは，真の分布がモデルに含まれる場合のことをいう．Fujisawa and Eguchi (2008) は，β-ダイバージェンスに基づいて，異常値の混入確率が高い場合の 1 次元正規分布のパラメータ推定を考え，そのためのアルゴリズムを提案した．

K-L 情報量で測るとき，想定するモデルを真の分布に最も近づけるパラメータの推定は MLE で与えられる．言い換えれば，MLE 以外の推定量は，K-L 情報量を用いる場合に最適な推定量とはいえない．本論文では，異常値に対して頑健なモデル構築とモデル選択を目指し，β-ダイバージェンスを考える．まず，2 節では，K-L 情報量を拡張した U-ダイバージェンスを導入する．次に 3 節で，U-ダイバージェンスを異常値に対する頑健性という立場で考えることにより β-ダイバージェンスを構成し，それに基づくモデル選択基準を提案する．4 節では，真の分

布に異常値と潜在分布の混合分布を仮定し，β-ダイバージェンスに基づいて，異常値に影響を受けずに潜在分布を推定することを考える．さらに，クラスタリングへ適用させるために，潜在分布に多次元混合正規分布を仮定し，そのパラメータの最適な推定量を得るためのアルゴリズムを与える．そして，β-ダイバージェンスに基づくモデル選択基準について，多次元混合正規分布における最適なクラスター数を決定する．5節では，異常値の影響をMLEと比較し，また，3節で与えたモデル選択基準と4節で提案した推定アルゴリズムの性能をシミュレーション実験で検証する．6節で，データ解析の実例を示す．本論文で得られる定理の証明は，最後に纏める．

2 U-ダイバージェンス

観測されたデータ $\boldsymbol{x}_n = \{x_1, x_2, \cdots, x_n\}$ が，密度関数 $g(x)$ をもつ未知の分布関数 $G(x)$ から生成されたとする．そのとき，$g(x)$ あるいは $G(x)$ を，真の分布という．真の分布 $g(x)$ を近似する密度関数 $f(x)$ をモデルといい，特に，データに基づいて構成されたモデルを統計的モデルという．モデルが p 次元パラメータ $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_p)$ によって規定される場合，モデルは $f(x|\boldsymbol{\theta})$ と表される．$\boldsymbol{\theta}$ が適当な空間 Θ の一点として与えられるとき，$\{f(x|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$ をモデル族という．

Akaike (1973) は，真の分布 $g(x)$ からみたモデル $f(x|\boldsymbol{\theta})$ の乖離を測る基準として，K-L情報量を用いることを考えた．対数尤度を最大にさせる $\boldsymbol{\theta}$ の推定量 $\hat{\boldsymbol{\theta}}(\boldsymbol{x}_n)$ がMLEである．すなわち，K-L情報量の意味で真の分布 $g(x)$ に最も近いモデルを $\{f(x|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$ から探すとき，データ数 n が十分大きいときの漸近的に最適なモデルが $f(x|\hat{\boldsymbol{\theta}}(\boldsymbol{x}_n))$ である．

構築されたモデル $f(x|\hat{\boldsymbol{\theta}}(\boldsymbol{x}_n))$ の良さは，あくまで，予測の観点から評価される．その意味で，現在観測されているデータ $\boldsymbol{x}_n$ だけで構築されたモデル $f(x|\hat{\boldsymbol{\theta}}(\boldsymbol{x}_n))$ を，将来観測されるデータを考慮に入れてバイアス補正する必要がある．このバイアスは，次の式で評価される．

$$b(G) = E_{G(\boldsymbol{x}_n)}\left[\sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\boldsymbol{\theta}}(\boldsymbol{x}_n)) - nE_{G(z)}\left\{\log f(z|\hat{\boldsymbol{\theta}}(\boldsymbol{x}_n))\right\}\right]$$

実際には，データを生み出す真の分布 $G(x)$ は未知なので，経験分布関数 $\hat{G}(x)$ でこれを置き換えて得られるバイアスの推定量 $b(\hat{G})$ を用いる．エントロピーの観点から $-2\log f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}(\boldsymbol{x}_n))$ を考え，これのバイアス補正をしたものとして，次のように情報量規準が定義される．

$$\mathrm{IC} = -2\log f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}(\boldsymbol{x}_n)) + 2b(\hat{G})$$

ICが小さいモデルを，予測の意味で，真の分布に漸近的に近い良いモデルと考える．

K-L 情報量の拡張として U-ダイバージェンスを定義し，U-ダイバージェンスの意味でモデルの最適なパラメータを定義する．U を狭義凸関数とし，その導関数 $u(\geq 0)$ が逆関数をもつことを仮定する．そのとき，$f(x|\boldsymbol{\theta})$ の $g(x)$ に対する U-ダイバージェンスは，次の式で定義される．

$$D_U(g(x); f(x|\boldsymbol{\theta})) = \int_{-\infty}^{\infty} \left\{U(u^{-1}(f(x|\boldsymbol{\theta}))) - U(u^{-1}(g(x))\right\} dx - \int_{-\infty}^{\infty} g(x) \left\{u^{-1}(f(x|\boldsymbol{\theta})) - u^{-1}(g(x))\right\} dx \tag{2.1}$$

U-ダイバージェンスは，$U(x) = e^x$ とおくと，K-L 情報量と一致する．

性質 1　U-ダイバージェンスは，次の性質をもつ．

(i) $D_U(g(x); f(x|\boldsymbol{\theta})) \geq 0$

(ii) $D_U(g(x); f(x|\boldsymbol{\theta})) = 0 \iff g(x) = f(x|\boldsymbol{\theta}),\ \forall x \in \mathbb{R}$

U-ダイバージェンスの値が小さいほど，真の分布 $g(x)$ にモデル $f(x|\boldsymbol{\theta})$ が近いと考えられる．$f(x|\boldsymbol{\theta})$ の $G(x)$ における U-平均尤度は，次の式で定義される．

$$C_U(g(x); f(x|\boldsymbol{\theta})) = \int_{-\infty}^{\infty} u^{-1}\left(f(x|\boldsymbol{\theta})\right) dG(x)dx - \int_{-\infty}^{\infty} U\left(u^{-1}\left(f(x|\boldsymbol{\theta})\right)\right) dx$$

U-ダイバージェンスと U-平均尤度には，

$$D_U(g(x); f(x|\boldsymbol{\theta})) = C_U(g(x); g(x)) - C_U(g(x); f(x|\boldsymbol{\theta})) \tag{2.2}$$

という関係があり，第 1 項はモデルに依存しない．従って，U-ダイバージェンスの意味で最適なパラメータは，U-平均尤度を最大にするパラメータとなる．U-平均尤度の第 1 項における $G(x)$ が未知なので，経験分布関数 $\hat{G}(x)$ でこれを置き換えた U-尤度を考える．観測データを $\boldsymbol{x}_n = \{x_1, x_2, \cdots, x_n\}$ とするとき，$f(x|\boldsymbol{\theta})$ の U-尤度は次の式で定義される．

$$\ell_U(\boldsymbol{\theta}) = \frac{1}{n}\sum_{\alpha=1}^{n} u^{-1}\left(f(x_\alpha|\boldsymbol{\theta})\right) - c_U(\boldsymbol{\theta})$$

ただし，$c_U(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} U\left(u^{-1}\left(f(x|\boldsymbol{\theta})\right)\right) dx$ である．

性質 2　$\boldsymbol{\theta}_0$ と $\hat{\boldsymbol{\theta}}_U$ は，$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}}\, C_U(g(x); f(x|\boldsymbol{\theta}))$，$\hat{\boldsymbol{\theta}}_U = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}}\, \ell_U(\boldsymbol{\theta})$ を満たすパラメータ空間の点とする．いま，$q(x|\boldsymbol{\theta}) = u^{-1}(f(x|\boldsymbol{\theta})) - c_U(\boldsymbol{\theta})$ とおき，これが次の正則条件 (I)-(II) を満たすと仮定する．

(I)　$q(x|\boldsymbol{\theta})$ は $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_p)^T$ に関して 3 階連続微分可能である．

(II) $\mathbb{R}$ 上で積分可能な $F_1(x)$, $F_2(x)$, $F_3(x)$ に対して,

$$\left|\frac{\partial q(x|\boldsymbol{\theta})}{\partial \theta_i}\right| < F_1(x), \quad \left|\frac{\partial^2 q(x|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right| < F_2(x), \quad \left|\frac{\partial^3 q(x|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j \partial \theta_k}\right| < F_3(x);$$
$$i,j,k = 1,2,\cdots,p$$

が任意の $\boldsymbol{\theta} \in \Theta$ に対して成り立つ. そのとき, 次の (i)-(ii) が主張できる.

(i) $\hat{\boldsymbol{\theta}}_U$ は $n \to \infty$ のとき $\boldsymbol{\theta}_0$ に確率収束する.

(ii) $n \to \infty$ のとき,

$$E_{G(x)}[(\hat{\boldsymbol{\theta}}_U - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_U - \boldsymbol{\theta}_0)^T] = \boldsymbol{J}_U(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}_U(\boldsymbol{\theta}_0)\boldsymbol{J}_U(\boldsymbol{\theta}_0)^{-1}/n + o(n^{-1})$$

が成り立つ.

(iii) $\hat{\boldsymbol{\theta}}_U$ は漸近正規性を有する. すなわち, $n \to \infty$ のとき,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_U - \boldsymbol{\theta}_0) \to N_p(\boldsymbol{0}, \boldsymbol{J}_U(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}_U(\boldsymbol{\theta}_0)\boldsymbol{J}_U(\boldsymbol{\theta}_0)^{-1})$$

に法則収束する. ただし, $\boldsymbol{I}_U(\boldsymbol{\theta}_0)$ と $\boldsymbol{J}_U(\boldsymbol{\theta}_0)$ は次で定義される.

$$\boldsymbol{I}_U(\boldsymbol{\theta}_0) = \int_{-\infty}^{\infty} \frac{\partial q(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial q(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} dG(x)$$
$$\boldsymbol{J}_U(\boldsymbol{\theta}_0) = -\int_{-\infty}^{\infty} \frac{\partial^2 q(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} dG(x)$$

3 U-ダイバージェンスと頑健性

ここでは, U-ダイバージェンスの特別な場合として, 異常値に対して頑健性をもつものを考える.

3.1 β-ダイバージェンス

性質 2 の $\hat{\boldsymbol{\theta}}_U$ は, $\dfrac{\partial}{\partial \boldsymbol{\theta}}\ell_U(\boldsymbol{\theta}) = \boldsymbol{0}$ の解になっている. いま,

$$s(x|\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f(x|\boldsymbol{\theta}), \quad w(x|\boldsymbol{\theta}) = f(x|\boldsymbol{\theta}) \frac{\partial}{\partial y} u^{-1}(y)\Big|_{y=f(x|\boldsymbol{\theta})} \tag{3.1}$$

とおくと,

$$\frac{\partial}{\partial \boldsymbol{\theta}}\ell_U(\boldsymbol{\theta}) = \frac{1}{n}\sum_{\alpha=1}^{n} w(x_\alpha|\boldsymbol{\theta}) s(x_\alpha|\boldsymbol{\theta}) - E_{f(x|\boldsymbol{\theta})}\left[w(x|\boldsymbol{\theta}) s(x|\boldsymbol{\theta})\right] \tag{3.2}$$

と表せることが分かる．このことから，$w(x|\boldsymbol{\theta}) = f(x|\boldsymbol{\theta})^\beta$ $(\beta > 0)$ を考える．観測されたデータに含まれる異常値を x_* とすると，スコア関数に掛け算される $w(x_*|\boldsymbol{\theta})$ の値が小さくなり，$\hat{\boldsymbol{\theta}}_U$ が異常値の影響を受けにくくなる．したがって，$w(x|\boldsymbol{\theta}) = f(x|\boldsymbol{\theta})^\beta$ となるように U を定めることによって，異常値に対して頑健なパラメータ推定量 $\hat{\boldsymbol{\theta}}_U$ を得ることができると考えられる．

(3.1) において $w(x|\boldsymbol{\theta}) = f(x|\boldsymbol{\theta})^\beta$ $(\beta > 0)$ とおくと，$u^{-1}(y) = (y^\beta - 1)/\beta$ $(y > 0)$，すなわち

$$U(x) = \frac{1}{1+\beta}(1+\beta x)^{\frac{1+\beta}{\beta}} \quad (-\infty < x < \infty) \tag{3.3}$$

と定まる．そのときの U-ダイバージェンスを，β-ダイバージェンスという．

$f(x|\boldsymbol{\theta})$ の $g(x)$ に対する β-ダイバージェンスは，次の式で定義される．

$$\begin{aligned} D_\beta(g(x); f(x|\boldsymbol{\theta})) =& \frac{1}{\beta(1+\beta)} \int_{-\infty}^{\infty} g(x)^{1+\beta} dx \\ & - \int_{-\infty}^{\infty} \frac{f(x|\boldsymbol{\theta})^\beta}{\beta} dG(x) + \int_{-\infty}^{\infty} \frac{f(x|\boldsymbol{\theta})^{1+\beta}}{1+\beta} dx \quad (\beta > 0) \end{aligned} \tag{3.4}$$

β-ダイバージェンスは，$\beta \to 0$ のとき $U(x) \to e^x$ となり，そのとき K-L 情報量に一致する．

$f(x|\boldsymbol{\theta})$ の $G(x)$ における β-平均尤度は，次の式で定義される．

$$C_\beta(g(x); f(x|\boldsymbol{\theta})) = \int_{-\infty}^{\infty} \frac{f(x|\boldsymbol{\theta})^\beta - 1}{\beta} dG(x) - c_\beta(\boldsymbol{\theta}) \quad (\beta > 0) \tag{3.5}$$

ただし，$c_\beta(\boldsymbol{\theta}) = \displaystyle\int_{-\infty}^{\infty} \frac{f(x|\boldsymbol{\theta})^{\beta+1}}{\beta+1} dx$ である．β-ダイバージェンスと β-平均尤度には，

$$D_\beta(g(x); f(x|\boldsymbol{\theta})) = C_\beta(g(x); g(x)) - C_\beta(g(x); f(x|\boldsymbol{\theta})) \tag{3.6}$$

という関係がある．β-ダイバージェンスの意味で最適なパラメータは，β-平均尤度を最大にするパラメータ $\boldsymbol{\theta}_{\beta 0} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}} C_\beta(g(x); f(x|\boldsymbol{\theta}))$ となる．β-平均尤度の第 1 項における $G(x)$ が未知なので，経験分布関数 $\hat{G}(x)$ でこれを置き換えた β-尤度を考える．観測データを $\boldsymbol{x}_n = \{x_1, x_2, \cdots, x_n\}$ とするとき，$f(x|\boldsymbol{\theta})$ の β-尤度は次の式で定義される．

$$\ell_\beta(\boldsymbol{\theta}; f(x|\boldsymbol{\theta})) = \frac{1}{n}\sum_{\alpha=1}^{n} \frac{f(x_\alpha|\boldsymbol{\theta})^\beta - 1}{\beta} - c_\beta(\boldsymbol{\theta}) \quad (\beta > 0) \tag{3.7}$$

β-尤度を最大にさせるパラメータ $\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n) = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}} \ell_\beta(\boldsymbol{\theta}; f(x|\boldsymbol{\theta}))$ を最大 β-尤度推定量という．β-ダイバージェンスの意味で真の分布 $g(x)$ に最も近いモデルを $\{f(x|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$ から探すとき，$f(x|\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n))$ がデータ数 n が大きいときの漸近的に最適なモデルである．

3.2 β-ダイバージェンスに基づくモデル選択基準

構築されたモデル $f(x|\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n))$ の良さは，あくまで，予測の観点から評価される．その意味で，現在観測されているデータ $\boldsymbol{x}_n$ だけで構築されたモデル $f(x|\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n))$ を，将来観測されるデータを考慮に入れてバイアス補正する必要がある．

β-ダイバージェンスのバイアスは次の式で評価される．

$$b_\beta(G) = E_{G(\boldsymbol{x}_n)}\left[n\ell_\beta(\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n); f(x|\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n))) - nC_\beta(g(x); f(x|\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n)))\right] \tag{3.8}$$

定理1　β-ダイバージェンスのバイアスは，$n \to \infty$ のとき，次の値で漸近的に評価される．

$$b_\beta(G) = \mathrm{tr}\{\boldsymbol{I}(\hat{\boldsymbol{\theta}}_\beta)\boldsymbol{J}(\hat{\boldsymbol{\theta}}_\beta)^{-1}\} + o(1)$$

ここで，$\boldsymbol{I}(\hat{\boldsymbol{\theta}}_\beta)$ と $\boldsymbol{J}(\hat{\boldsymbol{\theta}}_\beta)$ は次の式で計算される．

$$\boldsymbol{I}(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}}\left(\frac{f(x|\boldsymbol{\theta})^\beta - 1}{\beta} - c_\beta(\boldsymbol{\theta})\right)\frac{\partial}{\partial \boldsymbol{\theta}^T}\left(\frac{f(x|\boldsymbol{\theta})^\beta - 1}{\beta} - c_\beta(\boldsymbol{\theta})\right) dG(x)$$

$$\boldsymbol{J}(\boldsymbol{\theta}) = -\int_{-\infty}^{\infty} \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\left(\frac{f(x|\boldsymbol{\theta})^\beta - 1}{\beta} - c_\beta(\boldsymbol{\theta})\right) dG(x)$$

真の分布 $G(x)$ は未知なので，経験分布関数 $\hat{G}(x)$ で置き換えて，バイアスを $b_\beta(\hat{G})$ で推定する．バイアス補正を施して，次の情報量規準を提案する．

[β-ダイバージェンスに基づく情報量規準]

次で定義される IC_β が小さいモデルを，予測の意味で真の分布に漸近的に近い良いモデルと考える．

$$\begin{aligned}\mathrm{IC}_\beta &= -2\sum_{\alpha=1}^{n}\frac{f(x_\alpha|\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n))^\beta - 1}{\beta} + 2nc_\beta(\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n)) + 2b_\beta(\hat{G}) \\ &= -2n\ell_\beta(\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n); f(x|\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n))) + 2b_\beta(\hat{G})\end{aligned} \tag{3.9}$$

ここで，$b_\beta(\hat{G})$ は次の式で計算される．

$$b_\beta(\hat{G}) = \mathrm{tr}\{\boldsymbol{I}(\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n))\boldsymbol{J}(\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n))^{-1}\}$$

4 異常値混入モデルへの応用

異常値 $\boldsymbol{x}_\star$ が領域 $\mathbb{R}_o^p$ に確率 $\tau \in (0,1)$ で混入すると仮定する．ただし，領域 $\mathbb{R}_o^p$ は開集合とする．ここで，$\boldsymbol{x}_\star$ の密度関数を

$$\int_{\mathbb{R}_o^p} \psi(\boldsymbol{x})d\boldsymbol{x} = 1, \quad \sup_{\boldsymbol{x}\in\mathbb{R}_o^p} \psi(\boldsymbol{x}) < \infty$$

をみたす $\psi(\boldsymbol{x})$ とする．クラス数 $k_\star$ とパラメータ $\boldsymbol{\theta}_\star = (w_{1\star}, ..., w_{k\star}, \boldsymbol{\mu}_{1\star}, ..., \boldsymbol{\mu}_{k\star}, \boldsymbol{V}_{1\star}, ..., \boldsymbol{V}_{k\star})$ をもつ混合分布からなる潜在分布 $f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star)$ と，異常値の分布 $\psi(\boldsymbol{x})$ の混合分布として，真の分布は

$$g(\boldsymbol{x}) = (1-\tau)f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star) + \tau\psi(\boldsymbol{x}) \tag{4.1}$$

で表されるとする．そのとき

$$\int_{\mathbb{R}_o^p} f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star)d\boldsymbol{x} = \delta \ (> 0)$$

とおく．一般に，異常値が混入する確率は低く，また，異常値が混入する領域における潜在分布 $f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star)$ の確率は十分に低いと考えられるので，τ と δ は十分に小さい値を想定する．

4.1 β-ダイバージェンスによる潜在分布のモデル選択

潜在分布を推定するためのモデルとして $f(\boldsymbol{x}, k|\boldsymbol{\theta})$ を考える．異常値が混入する真の分布において，β-ダイバージェンスを用いて潜在分布を扱うときの精度は，真の分布と潜在分布の距離として，次の定理で評価することができる．

定理2 $f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star)$ の $g(\boldsymbol{x})$ に対する β-ダイバージェンスについて，$\tau \to 0$, $\delta \to 0$ のもとで

$$D_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star)) = O(\tau^2) + O(\tau^{1+\beta}) + O(\delta^{1+\beta})$$

が成り立つ．

注意1 $k < k_\star$ のとき，$D_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k|\boldsymbol{\theta}_{\beta 0})) > 0$ となる．

いま，β-ダイバージェンスによるバイアス $b_\beta(G)$ は定理1で与えられ，モデルのクラス数 k を考慮して，$b_\beta(G) = b_\beta(G, k)$ と表記する．ここで，$\tau = o(n^{-1/2})$ と仮定する．そのとき，次の定理が成り立つ．

定理3 $f(\boldsymbol{x}, k_\star|\hat{\boldsymbol{\theta}}_\beta)$ の $g(\boldsymbol{x})$ に対する β-ダイバージェンスについて，

$$E_{G(\boldsymbol{x}_n)}[D_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k_\star|\hat{\boldsymbol{\theta}}_\beta))] = (2n)^{-1}b_\beta(G, k_\star) + o(n^{-1})$$

が収束条件 $\tau^{1+\beta} = o(n^{-1})$, $\delta^{1+\beta} = o(n^{-1})$ のもとで成り立つ.

通常，バイアス $b_\beta(G,k)$ はパラメータ数に伴い増加する関数であるので，$k > k_\star$ において $0 < b_\beta(G,k_\star) < b_\beta(G,k)$ を仮定する．いま，モデルのクラス数 k を考慮して，β-ダイバージェンスに基づく情報量基準 IC_β を $\mathrm{IC}_\beta(k)$ と表記する．そのとき，次の定理が成り立つ.

定理 4　β-ダイバージェンスに基づく情報量基準 $\mathrm{IC}_\beta(k)$ について,

$$k_\star = \underset{k}{\mathrm{argmin}} \lim_{n\to\infty} \frac{1}{n} E_{G(\boldsymbol{x}_n)}[\mathrm{IC}_\beta(k)]$$

が定理 3 の収束条件のもとで成り立つ.

注意 2　τ の上限 τ_u $(\geq \tau)$ と δ の上限 δ_u $(\geq \delta)$ を仮定する．そのとき定理 3 より，$O(\tau_u^{1+\beta}) << n^{-1}$, $O(\delta_u^{1+\beta}) << n^{-1}$ を満たすように β を決めれば，定理 4 が成り立つ．定理 4 より，$f(\boldsymbol{x}, k|\boldsymbol{\theta})$ のモデル選択において，定理 3 の収束条件のもとで，異常値の影響を受けずに $k = k_\star$ となるモデルを平均的に選択できる．なお，τ と β の値による収束条件 $\tau^{1+\beta}$ の値を表 4.1 に纏める.

表 4.1　$\tau^{1+\beta}$ の値

$\beta \setminus \tau$	0.01	0.03	0.05	0.07	0.09
0.2	0.0040	0.015	0.027	0.041	0.056
0.4	0.0016	0.0074	0.015	0.024	0.034
0.6	0.00063	0.0037	0.0083	0.014	0.021
0.8	0.00025	0.0018	0.0046	0.0083	0.013

4.2　異常値に頑健なクラスタリングによるモデル構築

潜在分布を推定するモデルとして，多次元混合正規分布を考える．そのときに，モデルのパラメータを最大 β-尤度推定で与えるためのアルゴリズムを提案する．多次元混合正規分布モデルは次式で与えられる：

$$f(\boldsymbol{x}, k|\boldsymbol{\theta}) = \sum_{j=1}^{k} w_j \phi\left(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{V}_j\right), \ \boldsymbol{x} \in \mathbb{R}^p; \quad w_j \in (0,1), \ \sum_{j=1}^{k} w_j = 1, \ k \geq 1 \quad (4.2)$$

ここで，$\boldsymbol{\theta} = (w_1, ..., w_k, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k, \boldsymbol{V}_1, ..., \boldsymbol{V}_k)$ は未知で，$\phi(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{V}_j)$ は $N_p(\boldsymbol{\mu}_j, \boldsymbol{V}_j)$ の密度関数である．(4.2) の $f(\boldsymbol{x}, k|\boldsymbol{\theta})$ に対して，最大 β-尤度推定量を陽に求めることは困難であるため，何らかの反復解法が必要になる．ここでは，目的関数を $(\boldsymbol{\mu}_j, \boldsymbol{V}_j)$ で直接に微分して最大値を求めるのではなく，これらを一旦，$(\boldsymbol{s}_j, \boldsymbol{T}_j) =$

$(\boldsymbol{V}_j^{-1}\boldsymbol{\mu}_j, (-1/2)\boldsymbol{V}_j^{-1})$ で微分してから $(\boldsymbol{\mu}_j, \boldsymbol{V}_j)$ に戻すというアプローチを考える．最大値を与える $\hat{\boldsymbol{\theta}}_\beta$ は，次の反復解法により求める．

[多次元混合正規分布のパラメータ更新式]

パラメータの推定値 $\hat{\boldsymbol{\theta}}_\beta$ は，次の反復計算による収束値で与えられる．

$$\boldsymbol{\mu}_j^{\text{new}} = \left(\frac{1}{n}\sum_{\alpha=1}^{n} \boldsymbol{x}_\alpha w_\beta(j|\boldsymbol{x}_\alpha, \boldsymbol{\theta}) - \frac{\partial c_\beta(\boldsymbol{\theta})}{\partial \boldsymbol{s}_j} \right) \Big/ \frac{1}{n}\sum_{\alpha=1}^{n} w_\beta(j|\boldsymbol{x}_\alpha, \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\text{old}}}$$

$$(\boldsymbol{\mu}_j\boldsymbol{\mu}_j^T + \boldsymbol{V}_j)^{\text{new}} = \left(\frac{1}{n}\sum_{\alpha=1}^{n} \boldsymbol{x}_\alpha\boldsymbol{x}_\alpha^T w_\beta(j|\boldsymbol{x}_\alpha, \boldsymbol{\theta}) - \frac{\partial c_\beta(\boldsymbol{\theta})}{\partial \boldsymbol{T}_j} \right) \Big/ \frac{1}{n}\sum_{\alpha=1}^{n} w_\beta(j|\boldsymbol{x}_\alpha, \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\text{old}}}$$

$$w_j^{\text{new}} = \left(\frac{1}{n}\sum_{\alpha=1}^{n} w_\beta(j|\boldsymbol{x}_\alpha; \boldsymbol{\theta}) - w_j\frac{\partial c_\beta(\boldsymbol{\theta})}{\partial w_j} + w_j\sum_{i=1}^{k} w_i\frac{\partial c_\beta(\boldsymbol{\theta})}{\partial w_i} \right) \Big/ \frac{1}{n}\sum_{\alpha=1}^{n}\sum_{i=1}^{k} w_\beta(i|\boldsymbol{x}_\alpha; \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\text{old}}}$$

ここで，更新式の中にある演算は，次のように定義される．

$$w_\beta(j|\boldsymbol{x}; \boldsymbol{\theta}) = w_j\phi(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{V}_j) f(\boldsymbol{x}, k|\boldsymbol{\theta})^{\beta-1} \tag{4.3}$$

$$\frac{\partial c_\beta(\boldsymbol{\theta})}{\partial \boldsymbol{s}_j} = \int_{\mathbb{R}^p} w_j(\boldsymbol{x} - \boldsymbol{\mu}_j)\phi(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{V}_j) f(\boldsymbol{x}, k|\boldsymbol{\theta})^\beta d\boldsymbol{x}$$

$$\frac{\partial c_\beta(\boldsymbol{\theta})}{\partial \boldsymbol{T}_j} = \int_{\mathbb{R}^p} w_j(\boldsymbol{x}\boldsymbol{x}^T - (\boldsymbol{\mu}_j\boldsymbol{\mu}_j^T + \boldsymbol{V}_j))\phi(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{V}_j) f(\boldsymbol{x}, k|\boldsymbol{\theta})^\beta d\boldsymbol{x}$$

$$\frac{\partial b_\beta(\boldsymbol{\theta})}{\partial w_j} = \int_{\mathbb{R}^p} \phi(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{V}_j) f(\boldsymbol{x}, k|\boldsymbol{\theta})^\beta d\boldsymbol{x}$$

ここで提案するアルゴリズムは，特別な場合 $(p = 1)$ として，Fujisawa and Eguchi (2006) が開発した 1 次元混合正規分布のパラメータ推定を含む．また，EM アルゴリズム $(\beta = 0)$ の自然な拡張 $(\beta > 0)$ にもなっている．更新式において重要な役割を果たすのが，データがもつ負担率に関する (4.3) である．データ $\boldsymbol{x}$ が異常値である場合に，密度のべき関数が掛けられることで負担率は小さく見積もられる．したがって，(4.3) の値がどのクラスに対しても閾値を下回るような $\boldsymbol{x}$ は，異常値の候補と考えることができるであろう．(4.3) の値が閾値を超える場合には，その値が最大となるクラス k に個体 $\boldsymbol{x}$ を所属させる，というクラスタリング手法が考えられる．

5 シミュレーション

最大β-尤度推定量の異常値に対する頑健性を，シミュレーション実験で検証する．真の分布には，次のような混合分布を仮定する．

$$g(x) = 0.97\phi(x|0,1) + 0.03\psi(x|4 < x < 7) \tag{5.1}$$

ここで，$\phi(x|0,1)$ は潜在分布を表し，標準正規分布 $N(0,1)$ を仮定する．一方，$\psi(x|4 < x < 7)$ は異常値の分布を表し，区間 (4,7) の一様分布を仮定する．このとき，真の分布の平均と分散は，$E(x) = 0.165$, $V(x) = 1.873$ である．

いま，真の分布から独立にデータを100個発生させ，平均と分散をMLE ($\beta = 0$) と最大β-尤度推定 (β=0.2, 0.4, 0.6) で推定した．この実験を独立に100回繰り返して，推定値の平均と標準誤差（括弧内）の値を，表5.1に纏めた．異常値が正の領域に混入しているために，MLE ($\beta = 0$) による平均の推定は $N(0,1)$ の平均よりも正の方向に偏り，それに伴い分散は大きめに推定されている．MLEは，異常値の影響を忠実に反映して，異常値を含んだ真の分布の平均と分散を推定していることが見て取れる．それに対して最大β-尤度推定は，異常値の影響を受けることなく，$N(0,1)$ の平均と分散を正しく推定していることが見て取れる．最大β-尤度推定においてβの値を上げ過ぎると，異常値だけでなく潜在分布の裾部分の影響までつぶすことになり，その結果，平均と分散の推定を小さめに見積もる傾向がある．このことから，最大β-尤度推定において，βの値の選択には注意が必要である．

表5.1　MLE ($\beta = 0$) と最大β-尤度推定 (β=0.2, 0.4, 0.6) による平均と分散の推定値と標準誤差 (括弧内)

$\theta \setminus \beta$	0	0.2	0.4	0.6
$\mu = 0$	0.170	0.006	0.006	-0.020
	(0.013)	(0.014)	(0.011)	(0.010)
$\sigma^2 = 1$	1.890	0.913	0.808	0.756
	(0.049)	(0.017)	(0.011)	(0.011)

次に，3節で提案したモデル選択基準 IC_β の性能を，シミュレーション実験で検証する．真の分布には，次のような混合分布を仮定する．

$$g(\boldsymbol{x}) = 0.97f(\boldsymbol{x}, k_\star = 2|\boldsymbol{\theta}_\star) + 0.03\psi(\boldsymbol{x}) \tag{5.2}$$

ここで，$f(\boldsymbol{x}, k_\star = 2|\boldsymbol{\theta}_\star)$ は潜在分布を表し，次で定義される2次元正規分布 $\phi(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ の混合正規分布を仮定する．

$$f(\boldsymbol{x}, k_\star = 2|\boldsymbol{\theta}_\star) = 0.4\phi(\boldsymbol{x}|(-2,-2), \boldsymbol{I}_2) + 0.6\phi(\boldsymbol{x}|(2,2), \boldsymbol{I}_2) \tag{5.3}$$

また，$\psi(\boldsymbol{x})$ は異常値の分布を表し，$\mathbb{R}_o^2 = \{(x_1, x_2)|5^2 < x_1^2 + x_2^2 < 8^2\}$ 上の一様分布を仮定する．潜在分布を推定するために，モデルにはクラス数 $k = 1(1)5$ の混合正規分布を考える．

いま，真の分布から独立にデータを 100 個発生させ，潜在分布を推定する $k = 1(1)5$ の場合のモデルを MLE ($\beta = 0$) と最大 β-尤度推定 (β=0.2, 0.4) で構築した．その後，TIC ($\beta = 0$) と IC_β ($\beta = 0.2, 0.4$) を用いて，$k = 1(1)5$ の中から最適なモデルを選択した．(ここでは真の分布が想定するモデルの族に含まれないため，AIC の替わりに TIC を用いた．) この実験を独立に 100 回繰り返して，TIC ($\beta = 0$) と IC_β ($\beta = 0.2, 0.4$) による各モデルの選択回数と k の期待値を，表 5.2 に纏めた．TIC は異常値の影響を受けて，潜在クラス数である $k = 2$ よりも多いクラス数のモデルを選びやすく，それに対して IC_β は，潜在分布のクラス数を正しく特定していることが見て取れる．なお，表 5.2 では，異常値が無い場合の実験結果も纏めている．IC_β は，異常値が無い場合にも正常に機能していることが分かる．

表 5.2　100 回の実験における TIC ($\beta = 0$) と IC_β ($\beta = 0.2, 0.4$) による各モデルの選択回数と k の期待値

	異常値あり						異常値なし					
$\beta \setminus k$	1	2	3	4	5	$E(k)$	1	2	3	4	5	$E(k)$
0	0	42	37	16	5	2.84	0	61	27	7	5	2.56
0.2	2	75	17	5	1	2.28	1	81	11	5	2	2.26
0.4	13	73	7	7	0	2.08	5	87	5	3	0	2.06

6　データ解析

米国ワイオミング州に位置するイエローストーン国立公園 (Yellowstone National Park) 内の Old Faithful 間欠泉は，観光ポイントとして有名な熱水間欠泉である．この名前は，噴出におけるある種の規則性に由来している．図 6.1 は，1978 年 8 月 1 日から 7 日までの連続した $n = 94$ 個のデータをプロットしたものである．各データは 1 回の噴出を表し，各噴出における噴出継続時間（分）と次の噴出までの待ち時間（分）という 2 変数からなる．次の噴出までの待ち時間には大きなばらつきがあるが，直近の噴出の継続時間を知れば，より精確に予測ができるであろうことを示唆している．

いま，データを発生させている真の分布に，(4.1) の分布を仮定する．この潜在分布 $f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star)$ に 2 次元混合正規分布モデル (4.2) を仮定し，最適なクラス数のモデルを推定したい．この仮定のもとで，潜在分布は全体における主要な部分を表現していると考えられる．したがって，潜在分布で形成する各クラスの混合比率は，異常値の混入確率よりもある程度大きいと考えることは自然であろう．そ

こで，モデルのパラメータは，

$$(1-\tau_u)w_j \geq \tau_u,\ j=1,...,k \tag{6.1}$$

の条件を満たすものと仮定する．つまり，$(1-\tau_u)w_j < \tau_u\ j=1,...,k$ となるモデルは，潜在分布に異常値よりも小さいクラスがあることを意味しているので，モデルとしては不適切であると考える．いま，異常値の混入確率の上限を $\tau_u = 0.05$ とし，異常値の混入する領域での潜在分布の確率の上限 δ_u は τ_u よりも小さいと仮定する．モデルの選択には，TIC と IC_β を用い，IC_β には，注意 2 の収束条件をみたすものとして $\beta = 0.4$ を採用する．クラス数が $k=1$ のモデルから順に，それぞれのパラメータとモデル選択基準を計算した．ただし，(6.1) の条件をみたさないときのモデルのクラス数 k は不適切であるので，そのときのモデルは選ばないものとする．

まず，TIC で (6.1) の条件をみたすモデルのクラス数と，そのときのモデル選択基準の値は 906.1 ($k=1$), 824.3 ($k=2$), 817.3 ($k=3$), 816.8 ($k=4$), 818.0 ($k=5$) となった．したがって，TIC の推奨するモデルはクラス数が $k=4$ のモデルであり，(4.3) の値によってデータをクラスタリングしたものが図 6.2 である．それに対して，IC_β で (6.1) の条件をみたすモデルのクラス数と，そのときのモデル選択基準の値は 420.9 ($k=1$), 387.7 ($k=2$), 388.2 ($k=3$) となった．ここで，クラス数が $k=4$ を超えたとき，パラメータに関する条件 (6.1) をみたさなかったため，そのモデルは不適切であると判断し採用しなかった．このとき，IC_β の推奨するモデルはクラス数が $k=2$ のモデルであり，(4.3) の値によってデータをクラスタリングしたものが図 6.3 である．なお，(4.3) の値がどのクラスにおいても特に小さい値を示しているものを異常値として検出し，黒の点でプロットした．図 6.2 のモデルに比べて図 6.3 のモデルは，データのクラスターをより特定していることが見て取れる．また，クラスターから離れたデータは，異常値であることを示唆していることもわかる．

このデータに対する調査は古くから続いているが，噴出に関する傾向としては「長い」噴出 (3 分 30 秒以上) と極めて「短い」噴出を繰り返していて，その中間の長さの噴出はほとんどないことが報告されている．これが欠泉の特徴であり，IC_β はその特徴をより的確に捉えたモデルを推奨しているといえる．

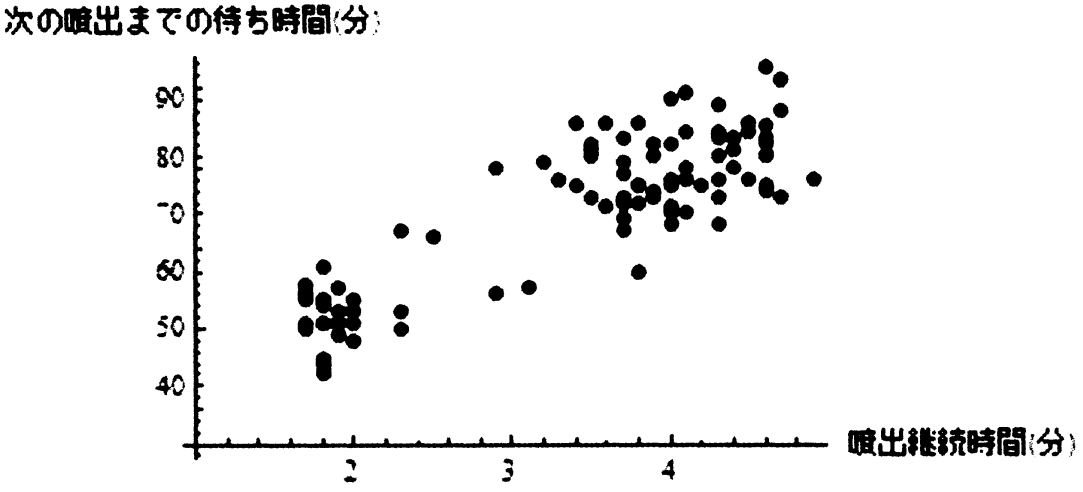

図 6.1　噴出の継続時間と次の噴出までの待ち時間のデータ ($n = 94$)

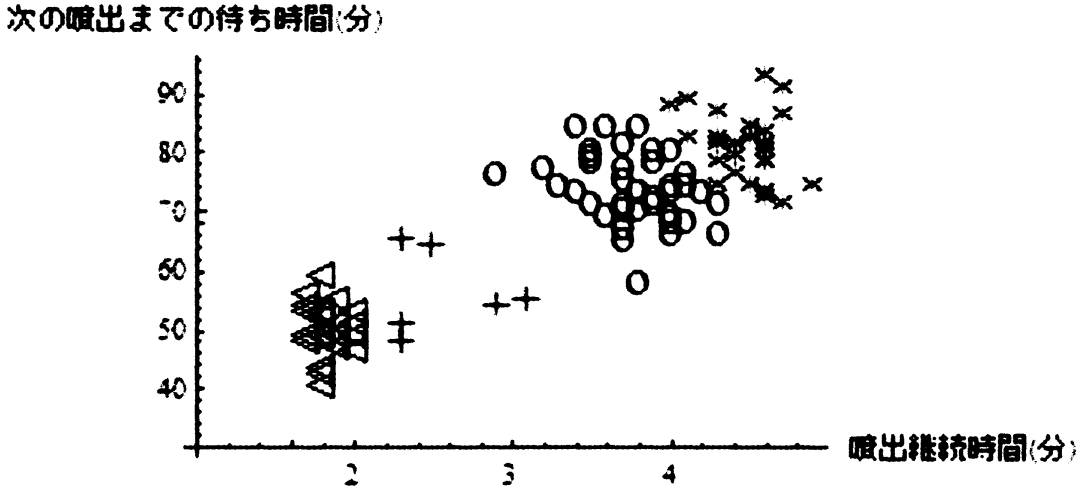

図 6.2　TIC の推奨モデル (クラス数 $k = 4$)

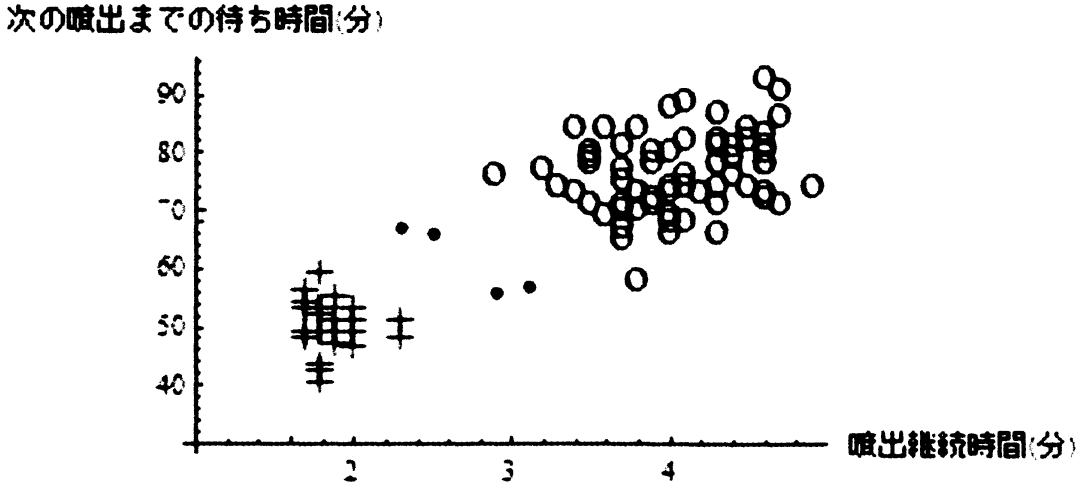

図 6.3　IC_β の推奨モデル (クラス数 $k = 2$, 黒の 4 点が異常値と検出された)

Appendix

定理1の証明 いま,

$$h_1(\boldsymbol{\theta}) = n\ell_\beta(\boldsymbol{\theta}; f(x|\boldsymbol{\theta})),$$
$$h_2(\boldsymbol{\theta}) = nC_\beta(g(x); f(x|\boldsymbol{\theta}))$$

とおく．そのとき，(3.8) より

$$\begin{aligned} b_\beta(G) =& E_{G(\boldsymbol{x}_n)}[h_1(\hat{\boldsymbol{\theta}}_\beta) - h_1(\boldsymbol{\theta}_{\beta 0})] + E_{G(\boldsymbol{x}_n)}[h_1(\boldsymbol{\theta}_{\beta 0}) - h_2(\boldsymbol{\theta}_{\beta 0})] \\ &+ E_{G(\boldsymbol{x}_n)}[h_2(\boldsymbol{\theta}_{\beta 0}) - h_2(\hat{\boldsymbol{\theta}}_\beta)] \end{aligned} \tag{A.1}$$

と表される．ここで，(A.1) の第2項は

$$\begin{aligned} & E_{G(\boldsymbol{x}_n)}[h_1(\boldsymbol{\theta}_{\beta 0}) - h_2(\boldsymbol{\theta}_{\beta 0})] \\ &= E_{G(\boldsymbol{x}_n)}\left[\sum_{\alpha=1}^{n} \frac{f(x_\alpha|\boldsymbol{\theta}_{\beta 0})^\beta - 1}{\beta} - \int_{-\infty}^{\infty} \frac{f(x|\boldsymbol{\theta}_{\beta 0})^\beta - 1}{\beta} g(x)dx\right] \\ &= \frac{1}{\beta} E_{G(\boldsymbol{x}_n)}\left[\sum_{\alpha=1}^{n} f(x_\alpha|\boldsymbol{\theta}_{\beta 0})^\beta\right] - \frac{1}{\beta} n E_{G(\boldsymbol{x}_n)}[f(x|\boldsymbol{\theta}_{\beta 0})^\beta] \\ &= 0 \end{aligned}$$

となるので,

$$b_\beta(G) = E_{G(\boldsymbol{x}_n)}[h_1(\hat{\boldsymbol{\theta}}_\beta) - h_1(\boldsymbol{\theta}_{\beta 0})] + E_{G(\boldsymbol{x}_n)}[h_2(\boldsymbol{\theta}_{\beta 0}) - h_2(\hat{\boldsymbol{\theta}}_\beta)] \tag{A.2}$$

を評価する．いま，$\left.\dfrac{\partial h_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}_\beta} = \mathbf{0}$ に注意し，$n \to \infty$ のもとで $h_1(\boldsymbol{\theta}_{\beta 0})$ を $\hat{\boldsymbol{\theta}}_\beta$ の周りでテイラー展開すると

$$E_{G(\boldsymbol{x}_n)}[h_1(\boldsymbol{\theta}_{\beta 0})] = E_{G(\boldsymbol{x}_n)}\left[h_1(\hat{\boldsymbol{\theta}}_\beta) + \frac{1}{2}(\boldsymbol{\theta}_{\beta 0} - \hat{\boldsymbol{\theta}}_\beta)^T \left.\frac{\partial^2 h_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\hat{\boldsymbol{\theta}}_\beta} (\boldsymbol{\theta}_{\beta 0} - \hat{\boldsymbol{\theta}}_\beta)\right] + o(1)$$

となるので，(A.2) の第1項は次のように式変形される．

$$\begin{aligned} & E_{G(\boldsymbol{x}_n)}[h_1(\hat{\boldsymbol{\theta}}_\beta) - h_1(\boldsymbol{\theta}_{\beta 0})] \\ &= -\frac{1}{2} E_{G(\boldsymbol{x}_n)}\left[(\boldsymbol{\theta}_{\beta 0} - \hat{\boldsymbol{\theta}}_\beta)^T \left.\frac{\partial^2 h_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\hat{\boldsymbol{\theta}}_\beta} (\boldsymbol{\theta}_{\beta 0} - \hat{\boldsymbol{\theta}}_\beta)\right] + o(1) \\ &= -\frac{1}{2} \mathrm{tr}\left\{ E_{G(\boldsymbol{x}_n)}\left[\left.\frac{\partial^2 h_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\hat{\boldsymbol{\theta}}_\beta} (\boldsymbol{\theta}_{\beta 0} - \hat{\boldsymbol{\theta}}_\beta)(\boldsymbol{\theta}_{\beta 0} - \hat{\boldsymbol{\theta}}_\beta)^T\right]\right\} + o(1) \\ &= -\frac{1}{2} \mathrm{tr}\left\{ E_{G(\boldsymbol{x}_n)}\left[\left.\frac{\partial^2 h_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_{\beta 0}} (\boldsymbol{\theta}_{\beta 0} - \hat{\boldsymbol{\theta}}_\beta)(\boldsymbol{\theta}_{\beta 0} - \hat{\boldsymbol{\theta}}_\beta)^T\right]\right\} + o(1) \end{aligned} \tag{A.3}$$

次に，$\left.\dfrac{\partial h_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_{\beta 0}} = \mathbf{0}$ に注意し，$n \to \infty$ のもとで $h_2(\hat{\boldsymbol{\theta}}_\beta)$ を $\boldsymbol{\theta}_{\beta 0}$ の周りでテイラー展開すると

$$E_{G(\boldsymbol{x}_n)}[h_2(\hat{\boldsymbol{\theta}}_\beta)] = E_{G(\boldsymbol{x}_n)}\left[h_2(\boldsymbol{\theta}_{\beta 0}) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_{\beta 0})^T \left.\frac{\partial^2 h_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_{\beta 0}} (\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_{\beta 0})\right] + o(1)$$

となるので，(A.2) の第 2 項は次のように式変形される．

$$\begin{aligned} &E_{G(\boldsymbol{x}_n)}[h_2(\boldsymbol{\theta}_{\beta 0}) - h_2(\hat{\boldsymbol{\theta}}_\beta)] \\ &= -\frac{1}{2}\mathrm{tr}\left\{E_{G(\boldsymbol{x}_n)}\left[\left.\frac{\partial^2 h_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_{\beta 0}} (\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_{\beta 0})(\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_{\beta 0})^T\right]\right\} + o(1) \end{aligned} \tag{A.4}$$

したがって，(A.2) に (A.3) と (A.4) を代入して，さらに性質 2(ii) を用いることで，

$$\begin{aligned} b_\beta(G) &= -\mathrm{tr}\left\{E_{G(\boldsymbol{x}_n)}\left[\left.\frac{\partial^2 h_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_{\beta 0}} (\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_{\beta 0})(\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_{\beta 0})^T\right]\right\} + o(1) \\ &= \mathrm{tr}\left\{n\boldsymbol{J}(\boldsymbol{\theta}_{\beta 0})E_{G(\boldsymbol{x}_n)}\left[(\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_{\beta 0})(\hat{\boldsymbol{\theta}}_\beta - \boldsymbol{\theta}_{\beta 0})^T\right]\right\} + o(1) \\ &= \mathrm{tr}\{\boldsymbol{I}(\boldsymbol{\theta}_{\beta 0})\boldsymbol{J}^{-1}(\boldsymbol{\theta}_{\beta 0})\} + o(1) \end{aligned}$$

を得る．以上より，定理が示された． □

定理 2 の証明　β-ダイバージェンスについて，$f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star)$ を含む項を積分範囲によって分けると，

$$\begin{aligned} &D_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star)) \\ &= \frac{1}{\beta(1+\beta)}\int_{\mathbb{R}^p} g(\boldsymbol{x})^{1+\beta}d\boldsymbol{x} + \int_{\mathbb{R}^p_o}\left(\frac{f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star)^{1+\beta}}{1+\beta} - \frac{f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star)^\beta}{\beta}g(\boldsymbol{x})\right)d\boldsymbol{x} \\ &\quad + \int_{\mathbb{R}^p\setminus\mathbb{R}^p_o}\left(\frac{f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star)^{1+\beta}}{1+\beta} - \frac{f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star)^\beta}{\beta}g(\boldsymbol{x})\right)d\boldsymbol{x} \end{aligned} \tag{A.5}$$

と表される．まず，(A.5) の第 1 項について，$\sup_{\boldsymbol{x}\in\mathbb{R}^p_o}\psi(\boldsymbol{x}) < \infty$ より，$\boldsymbol{x} \in \mathbb{R}^p_o$ において $g(\boldsymbol{x})^\beta = O(((1-\tau)\delta + \tau)^\beta) = O((\delta+\tau)^\beta)$ がいえるので，

$$\begin{aligned} &\frac{1}{\beta(1+\beta)}\int_{\mathbb{R}^p} g(\boldsymbol{x})^{1+\beta}d\boldsymbol{x} \\ &= \frac{1}{\beta(1+\beta)}\int_{\mathbb{R}^p_o} g(\boldsymbol{x})^{1+\beta}d\boldsymbol{x} + \frac{1}{\beta(1+\beta)}\int_{\mathbb{R}^p\setminus\mathbb{R}^p_o} g(\boldsymbol{x})^{1+\beta}d\boldsymbol{x} \\ &= O((\delta+\tau)^\beta) \times \frac{1}{\beta(1+\beta)}\int_{\mathbb{R}^p_o} g(\boldsymbol{x})d\boldsymbol{x} + \frac{1}{\beta(1+\beta)}\int_{\mathbb{R}^p\setminus\mathbb{R}^p_o} g(\boldsymbol{x})^{1+\beta}d\boldsymbol{x} \\ &= O((\delta+\tau)^{1+\beta}) + \frac{1}{\beta(1+\beta)}\int_{\mathbb{R}^p\setminus\mathbb{R}^p_o} g(\boldsymbol{x})^{1+\beta}d\boldsymbol{x} \end{aligned} \tag{A.6}$$

を得る．次に，(A.5) の第 2 項について考えるために，$\max_{\boldsymbol{x}\in\mathbb{R}_o^p} f(\boldsymbol{x},k_\star|\boldsymbol{\theta}_\star) = f(\boldsymbol{x}_M,k_\star|\boldsymbol{\theta}_\star)(>0)$ とおく．このとき，$\boldsymbol{x}_M+\boldsymbol{\epsilon}\in\mathbb{R}_o^p$，かつ，$\omega f(\boldsymbol{x}_M+\boldsymbol{\epsilon},k_\star|\boldsymbol{\theta}_\star)\le\delta$ をみたすような定数ベクトル $\boldsymbol{\epsilon}$ $(||\boldsymbol{\epsilon}||>0)$ と定数 $\omega>0$ が存在する．さらに，ある定数 $\nu(0<\nu<1)$ を用いて，$f(\boldsymbol{x}_M+\boldsymbol{\epsilon},k_\star|\boldsymbol{\theta}_\star) = \nu f(\boldsymbol{x}_M,k_\star|\boldsymbol{\theta}_\star)$ と表せるので，$\omega\nu f(\boldsymbol{x}_M,k_\star|\boldsymbol{\theta}_\star)\le\delta$ を得る．したがって，ω,ν は定数であったから，$f(\boldsymbol{x}_M+\boldsymbol{\epsilon},k_\star|\boldsymbol{\theta}_\star) = O(\delta)$ $(\delta\to 0)$ がいえる．よって，(A.5) の第 2 項については，$\delta\to 0$, $\tau\to 0$ のとき，

$$\begin{aligned}
&\int_{\mathbb{R}_o^p}\left(\frac{f(\boldsymbol{x},k_\star|\boldsymbol{\theta}_\star)^{1+\beta}}{1+\beta}-\frac{f(\boldsymbol{x},k_\star|\boldsymbol{\theta}_\star)^{\beta}}{\beta}g(\boldsymbol{x})\right)d\boldsymbol{x}\\
&<\int_{\mathbb{R}_o^p}\left(\frac{\delta^\beta}{1+\beta}f(\boldsymbol{x},k_\star|\boldsymbol{\theta}_\star)+\frac{\delta^\beta}{\beta}g(\boldsymbol{x})\right)d\boldsymbol{x}\\
&<\frac{\delta^\beta}{\beta}\int_{\mathbb{R}_o^p}\{f(\boldsymbol{x},k_\star|\boldsymbol{\theta}_\star)+g(\boldsymbol{x})\}d\boldsymbol{x}\\
&<O(\delta^\beta(\delta+(1-\tau)\delta+\tau)\ =O(\delta^{1+\beta})+O(\tau\delta^\beta) \qquad \text{(A.7)}
\end{aligned}$$

を得る．最後に，(A.5) の第 3 項について，$\boldsymbol{x}\in\mathbb{R}^p\backslash\mathbb{R}_o^p$ においては異常値が発生しないので，$g(\boldsymbol{x}) = (1-\tau)f(\boldsymbol{x},k_\star|\boldsymbol{\theta}_\star)$ がいえる．したがって，

$$\begin{aligned}
&\int_{\mathbb{R}^p\backslash\mathbb{R}_o^p}\left(\frac{f(\boldsymbol{x},k_\star|\boldsymbol{\theta}_\star)^{1+\beta}}{1+\beta}-\frac{f(\boldsymbol{x},k_\star|\boldsymbol{\theta}_\star)^{\beta}}{\beta}g(\boldsymbol{x})\right)d\boldsymbol{x}\\
&=\int_{\mathbb{R}^p\backslash\mathbb{R}_o^p}\left((1-\tau)^{-1-\beta}\frac{g(\boldsymbol{x})^{1+\beta}}{1+\beta}-(1-\tau)^{-\beta}\frac{g(\boldsymbol{x})^{1+\beta}}{\beta}\right)d\boldsymbol{x}
\end{aligned}$$

と表わされる．さらにこの被積分関数の $(1-\tau)^{-1-\beta}$ と $(1-\tau)^{-\beta}$ について，それぞれ 0 の周りでテイラー展開すると

$$\begin{aligned}
&\int_{\mathbb{R}^p\backslash\mathbb{R}_o^p}\left(\frac{f(\boldsymbol{x},k_\star|\boldsymbol{\theta}_\star)^{1+\beta}}{1+\beta}-\frac{f(\boldsymbol{x},k_\star|\boldsymbol{\theta}_\star)^{\beta}}{\beta}g(\boldsymbol{x})\right)d\boldsymbol{x}\\
&=\int_{\mathbb{R}^p\backslash\mathbb{R}_o^p}\left\{\left(1+(1+\beta)\tau+O(\tau^2)\right)\frac{g(\boldsymbol{x})^{1+\beta}}{1+\beta}-\left(1+\beta\tau+O(\tau^2)\right)\frac{g(\boldsymbol{x})^{1+\beta}}{\beta}\right\}d\boldsymbol{x}\\
&=-\frac{1}{\beta(1+\beta)}\int_{\mathbb{R}^p\backslash\mathbb{R}_o^p}g(\boldsymbol{x})^{1+\beta}d\boldsymbol{x}+O(\tau^2) \qquad \text{(A.8)}
\end{aligned}$$

を得る．したがって，(A.5)-(A.8) より，$\delta\to 0$, $\tau\to 0$ のとき，

$$\begin{aligned}
&D_\beta(g(\boldsymbol{x});f(\boldsymbol{x},k_\star|\boldsymbol{\theta}_\star))\\
&=O(\tau^2)+O((\delta+\tau)^{1+\beta})+O(\delta^{1+\beta})+O(\tau\delta^\beta)\\
&=O(\tau^2)+O((\delta+\tau)^{1+\beta})
\end{aligned}$$

となる．ここで，$\tau\ge\delta$ のとき $O((\delta+\tau)^{1+\beta}) = O(\tau^{1+\beta})$ となり，$\tau<\delta$ のとき $O((\delta+\tau)^{1+\beta}) = O(\delta^{1+\beta})$ となるので，以上を纏めて

$$D_\beta(g(\boldsymbol{x});f(\boldsymbol{x},k_\star|\boldsymbol{\theta}_\star)) = O(\tau^2)+O(\tau^{1+\beta})+O(\delta^{1+\beta})$$

を得る．よって，定理が示された． □

定理3の証明 いま，(A.4) と性質2(ii) より，

$$
\begin{aligned}
&E_{G(\boldsymbol{x}_n)}[D_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k_\star|\hat{\boldsymbol{\theta}}_\beta))] \\
&= \frac{1}{\beta(1+\beta)} \int_{\mathbb{R}^p} g(\boldsymbol{x})^{1+\beta} d\boldsymbol{x} \\
&\quad + E_{G(\boldsymbol{x}_n)} \left[\int_{\mathbb{R}^p} \left(\frac{f(\boldsymbol{x}, k_\star|\hat{\boldsymbol{\theta}}_\beta)^{1+\beta}}{1+\beta} - \frac{f(\boldsymbol{x}, k_\star|\hat{\boldsymbol{\theta}}_\beta)^{\beta}}{\beta} g(\boldsymbol{x}) \right) d\boldsymbol{x} \right] \\
&= \frac{1}{\beta(1+\beta)} \int_{\mathbb{R}^p} g(\boldsymbol{x})^{1+\beta} d\boldsymbol{x} \\
&\quad + E_{G(\boldsymbol{x}_n)} \left[\int_{\mathbb{R}^p} \left(\frac{f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_{\beta 0})^{1+\beta}}{1+\beta} - \frac{f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_{\beta 0})^{\beta}}{\beta} g(\boldsymbol{x}) \right) d\boldsymbol{x} \right] + \frac{1}{2n} b_\beta(G, k_\star) \\
&= E_{G(\boldsymbol{x}_n)}[D_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_{\beta 0}))] + \frac{1}{2n} b_\beta(G, k_\star) \qquad \text{(A.9)}
\end{aligned}
$$

を得る．ここで，$\tau \to 0$ のとき，

$$
\begin{aligned}
&C_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k_\star|\boldsymbol{\theta})) \\
&= \int_{\mathbb{R}^p} \frac{f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}))^\beta - 1}{\beta} g(\boldsymbol{x}) d\boldsymbol{x} - \int_{\mathbb{R}^p} \frac{f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}))^{1+\beta}}{1+\beta} d\boldsymbol{x} \\
&= \int_{\mathbb{R}^p} \frac{f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}))^\beta - 1}{\beta} f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star) d\boldsymbol{x} - \int_{\mathbb{R}^p} \frac{f(\boldsymbol{x}, k_\star|\boldsymbol{\theta})^{1+\beta}}{1+\beta} d\boldsymbol{x} + O(\tau) \\
&= C_\beta(f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star); f(\boldsymbol{x}, k_\star|\boldsymbol{\theta})) + O(\tau)
\end{aligned}
$$

と表せるので，

$$
\frac{\partial}{\partial \boldsymbol{\theta}} C_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}_\star} = \frac{\partial}{\partial \boldsymbol{\theta}} C_\beta(f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star); f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}_\star} + \boldsymbol{O}(\tau) = \boldsymbol{O}(\tau) \qquad \text{(A.10)}
$$

から $\dfrac{\partial}{\partial \boldsymbol{\theta}} D_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}_\star} = \boldsymbol{O}(\tau)$ を得る．さらに，(A.10) の左辺を $\boldsymbol{\theta}_{\beta 0}$ のまわりでテイラー展開すると

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\theta}} C_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}_\star} \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} C_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}_{\beta 0}} + \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} C_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}))\Big|_{\ddot{\boldsymbol{\theta}}} (\boldsymbol{\theta}_\star - \boldsymbol{\theta}_{\beta 0}) \\
&= \boldsymbol{J}(\ddot{\boldsymbol{\theta}})(\boldsymbol{\theta}_{\beta 0} - \boldsymbol{\theta}_\star)
\end{aligned}
$$

となる．ここで，$\ddot{\boldsymbol{\theta}}$ は $\boldsymbol{\theta}_\star$ と $\boldsymbol{\theta}_{\beta 0}$ との間の線分上の点の値をとるベクトルであり，$\boldsymbol{\theta}_\star$ と $\boldsymbol{\theta}_{\beta 0}$ の一致性から $\boldsymbol{J}(\ddot{\boldsymbol{\theta}})$ は逆行列が存在するので，$\boldsymbol{\theta}_{\beta 0} - \boldsymbol{\theta}_\star = \boldsymbol{O}(\tau)$ を得る．

したがって，定理2と$\tau = o(n^{-1/2})$, $\tau^{1+\beta} = o(n^{-1})$, $\delta^{1+\beta} = o(n^{-1})$の条件から，$D_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_{\beta 0}))$を$\boldsymbol{\theta}_\star$のまわりでテイラー展開すると

$$\begin{aligned}
&D_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_{\beta 0})) \\
&= D_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}_\star)) + (\boldsymbol{\theta}_{\beta 0} - \boldsymbol{\theta}_\star)^T \frac{\partial}{\partial \boldsymbol{\theta}} D_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k_\star|\boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}_\star} + O(\tau^2) \\
&= O(\tau^2) + O(\tau^{1+\beta}) + O(\delta^{1+\beta}) \\
&= o(n^{-1}) \qquad \text{(A.11)}
\end{aligned}$$

を得る．したがって，(A.9)と(A.11)より，

$$E_{G(\boldsymbol{x}_n)}[D_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k_\star|\hat{\boldsymbol{\theta}}_\beta))] = o(n^{-1}) + \frac{1}{2n} b_\beta(G, k_\star)$$

となり，定理が示された． □

定理4の証明　まず，$E_{G(\boldsymbol{x}_n)}[\mathrm{IC}_\beta(k)]$は次のように式変形できる．

$$\begin{aligned}
E_{G(\boldsymbol{x}_n)}[\mathrm{IC}_\beta(k)] &= -2E_{G(\boldsymbol{x}_n)}\left[n\ell_\beta(\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n); f(\boldsymbol{x}, k|\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n)))\right] + 2b_\beta(G, k) + o(1) \\
&= -2E_{G(\boldsymbol{x}_n)}\left[n\ell_\beta(\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n); f(\boldsymbol{x}, k|\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n))) - nC_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k|\boldsymbol{\theta}_{\beta 0}))\right] \\
&\quad - 2nC_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k|\boldsymbol{\theta}_{\beta 0})) + 2b_\beta(G, k) + o(1)
\end{aligned}$$

この第1項について，(A.4)と性質2(ii)より

$$E_{G(\boldsymbol{x}_n)}\left[n\ell_\beta(\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n); f(\boldsymbol{x}, k|\hat{\boldsymbol{\theta}}_\beta(\boldsymbol{x}_n))) - nC_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k|\boldsymbol{\theta}_{\beta 0}))\right] = \frac{1}{2} b_\beta(G, k) + o(1)$$

であるから，

$$\frac{1}{n} E_{G(\boldsymbol{x}_n)}[\mathrm{IC}_\beta(k)] = -2C_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k|\boldsymbol{\theta}_{\beta 0})) + \frac{1}{n} b_\beta(G, k) + o(n^{-1}) \qquad \text{(A.12)}$$

を得る．ここで，$k = k_\star$のとき，(3.6)と(A.11)より

$$\frac{1}{n} E_{G(\boldsymbol{x}_n)}[\mathrm{IC}_\beta(k_\star)] = -2C_\beta(g(\boldsymbol{x}); g(\boldsymbol{x})) + \frac{1}{n} b_\beta(G, k_\star) + o(n^{-1}) \qquad \text{(A.13)}$$

を得る．次に，$k < k_\star$のとき，注意1より$C_\beta(g(\boldsymbol{x}); g(\boldsymbol{x})) > C_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k|\boldsymbol{\theta}_{\beta 0}))$なので

$$\frac{1}{n} E_{G(\boldsymbol{x}_n)}[\mathrm{IC}_\beta(k)] > -2C_\beta(g(\boldsymbol{x}); g(\boldsymbol{x})) + \frac{1}{n} b_\beta(G, k) + o(n^{-1}) \qquad \text{(A.14)}$$

を得る．最後に，$k > k_\star$のとき，$\sup C_\beta(g(\boldsymbol{x}); f(\boldsymbol{x}, k|\boldsymbol{\theta}_{\beta 0})) = C_\beta(g(\boldsymbol{x}); g(\boldsymbol{x}))$より

$$\frac{1}{n} E_{G(\boldsymbol{x}_n)}[\mathrm{IC}_\beta(k)] \geq -2C_\beta(g(\boldsymbol{x}); g(\boldsymbol{x})) + \frac{1}{n} b_\beta(G, k) + o(n^{-1}) \qquad \text{(A.15)}$$

を得る．そのとき，(A.13)-(A.15) と $k > k_\star$ において，$0 < b_\beta(G, k_\star) < b_\beta(G, k)$ の仮定より，$b_\beta(G, k)$ が最小となるのは $k = k_\star$ のときである．したがって，$n \to \infty$ のもとで，(A.12) を最小にするのは $k = k_\star$ のときであるから，定理が示された．
□

謝辞 本研究は，科学研究費補助金 基盤研究 (B) 22300094 研究代表者: 青嶋 誠「高次元データの理論と方法論の総合的研究」から，研究助成を受けています．

参考文献

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd Inter. Symp. on Information Theory* (Petrov, B. N. and Csaki, F., eds.), Akademiai Kiado, 267–281 (Reproduced in *Breakthroughs in Statistics*, 1 (Kotz, S. and Johnson, N. L., eds.), Springer-Verlag (1992)).

Akaike, H. (1977). On entropy maximization principle. *Applications of Statistics* (Krishnaiah, P. R., ed.), North-Holland, **27** (41).

Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85**, 549–559.

Eguchi, S. (2006). Prediction and discovery: Towards novel methodology for genome data analysis (in Japanese). *Proceedings of the Institute of Statistical Mathematics.* **54**, 375–403.

Fujisawa, H. and Eguchi, S. (2006). Robust estimation in the normal mixture model. *J. Statist. Plan. Infer.* **136**, 3989–4011.

Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *J. Multivariate Analysis.* **99**, 2053–2081.

Hampel, F. R. (1968). *Contributions to the theory of robust estimation.* Ph. D. Thesis, University of California, Berkeley.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **62**, 1179–1186.

Huber, P. J. (1964). Robust estimetion of a location parameter. *Ann. Math. Statist.* **35**, 73–101.

Ishiguro, M., Sakamoto, Y. and Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. *Ann. Inst. Statist. Math.* **49**, 411–434.

Jones, M. C., Hjort, N. L., Harris, I. R. and Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika* **88**, 865–873.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statistics* **6**, 461–464.

Scott, D. W. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics* **43**, 274–285.

竹内啓 (1976). 情報統計量の分布とモデルの適切さの規準. 数理科学 **153**, 12–18.