2014年9月14日

Density Power Divergenceを用いたロバストな情報量規準の提案とその応用

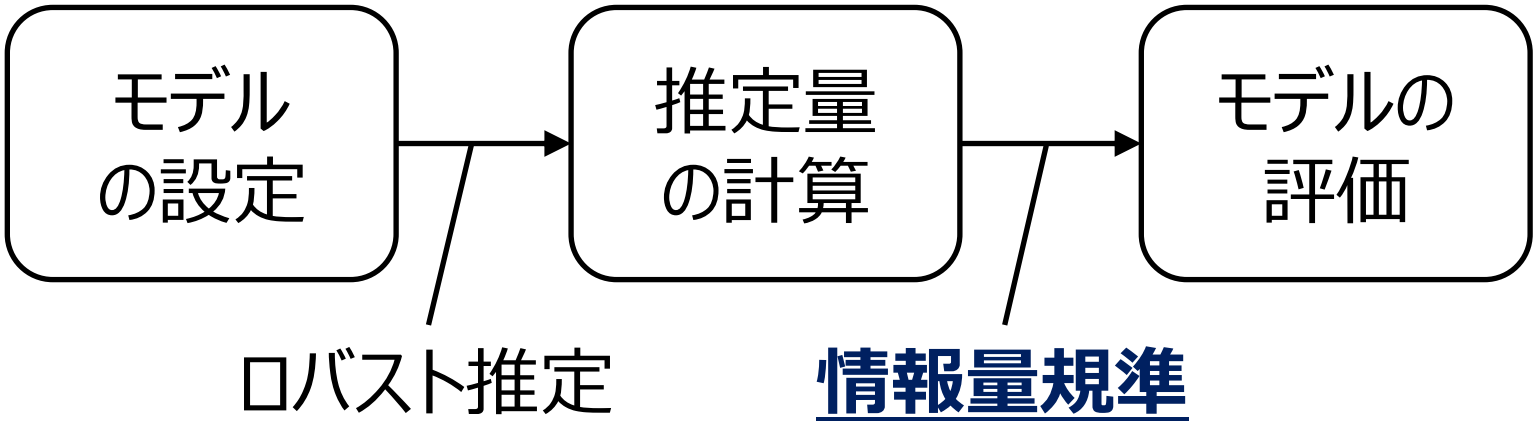
大阪大学大学院基礎工学研究科

奥野 彰文

下平 英寿

研究の目的

外れ値を含むデータを分析する



ロバストな情報量規準を導出する

Density **P**ower **D**ivergence (Basu et al. (1998))を用いる.

1. DPDとβ-推定の紹介

Kullback **L**eibler Divergence (KL)

$$d(q, p_\theta) := -\int q(x) \log p(x \mid \theta) \mathrm{d}x$$

$$\hat{\theta}_{\mathrm{MLE}} = \arg\min_{\theta \in \Theta} d(\hat{q}, p_\theta)$$

Density **P**ower **D**ivergence (DPD)
(Basu et al.(1998). 別名β-divergence)

$$d_\beta(q, p_\theta) := -\int q(x) \frac{p(x \mid \theta)^\beta - 1}{\beta} \mathrm{d}x \quad + b_{1+\beta}(\theta) \qquad (\beta > 0)$$

β-推定量　$\hat{\theta}_\beta := \arg\min_{\theta \in \Theta} d_\beta(\hat{q}, p_\theta)$

$\hat{\theta}_\beta \to \hat{\theta}_{\mathrm{MLE}}, \ (\beta \downarrow 0)$　　β-推定量はM-推定量である

重み　　スコア関数

$$\frac{\partial d_\beta(\hat{q}, p_\theta)}{\partial \theta} = -\frac{1}{n}\sum_{t=1}^{n} p(x_t \mid \theta)^\beta \frac{\partial \log p(x_t \mid \theta)}{\partial \theta} + \frac{\partial b_{1+\beta}(\theta)}{\partial \theta} = 0$$

β が
- 小さい⇒(データの)有効性が高い
- 大きい⇒ロバスト

βは通常 0~1 の値を取るが,
外れ値によっては β=2 程度が必要となる.

2. IC_Bの提案

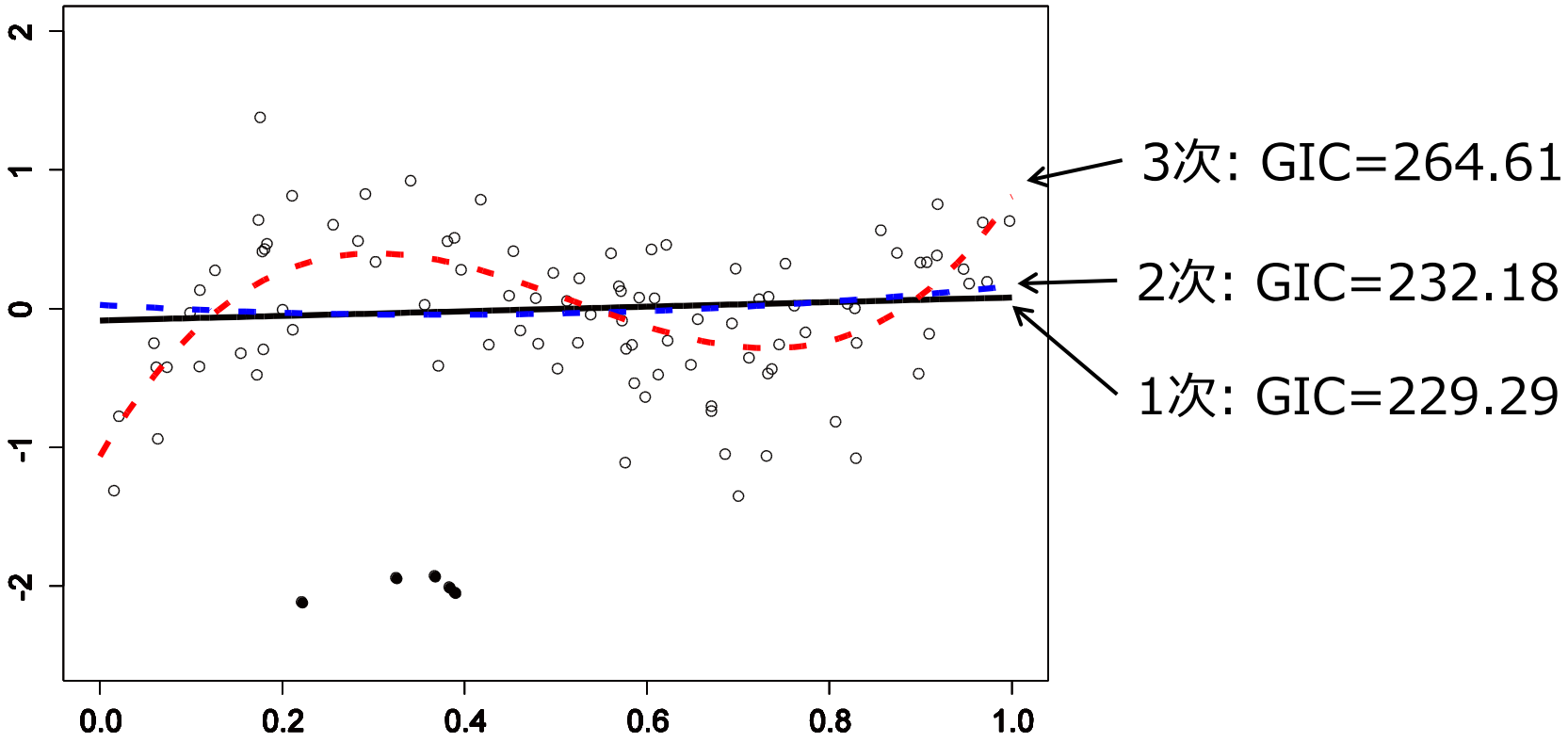
β-推定はM-推定である

GICはM-推定されたモデルを評価できる.

$\mathrm{GIC}_\beta \Leftrightarrow d(q, \underline{\hat{p}_\beta})$ **β-推定量**をKLで評価

GICは外れ値に弱い

ロバスト推定された1次,2次,3次のモデルをGICで評価



GICは**1次**のモデルを選択する. (当てはまりが良くない)

Konishi and Kitagawa(1996) $\mathrm{GIC}_\beta \Leftrightarrow d(q, \hat{p}_\beta)$

KL情報量はロバストでない

Ronchetti(1982) $\mathrm{AICR}_\beta \Leftrightarrow d_\beta(q, \hat{p}_\beta)$

βに依存したモデル評価

本研究 $\mathrm{IC}_\mathrm{B} \Leftrightarrow d_B(q, \hat{p}_\beta)$

ロバスト かつ **βに依存しない** モデル評価

今回提案する情報量規準IC_B

B>0 :fixed,

$$\mathrm{IC}_B := 2n\, d_B(\hat{q}, \hat{p}_\beta) + 2\mathrm{tr} G_{\beta,B} H_\beta^{-1}$$

$$\rho_\beta(x \mid \theta) := -\frac{1}{\beta} p(x \mid \theta)^\beta + \frac{1}{1+\beta} \int p(x \mid \theta)^{1+\beta} \mathrm{d}x$$

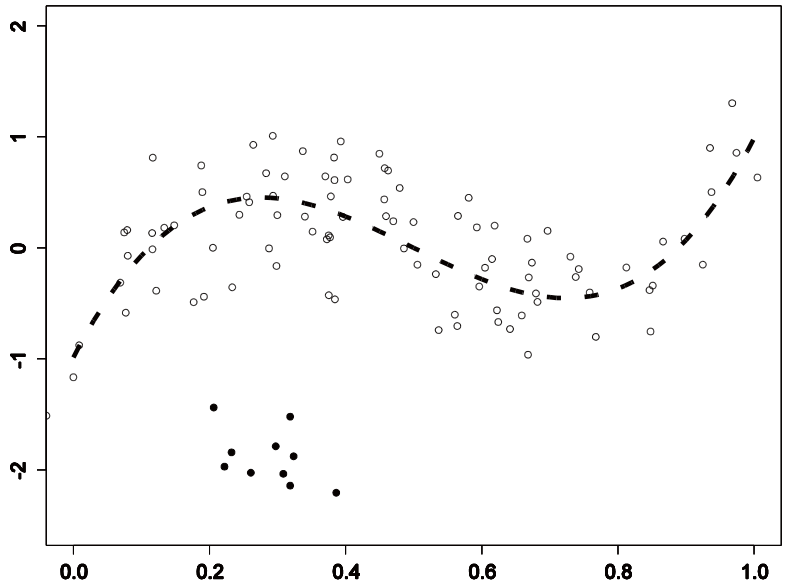
$$G_{\beta,B} := -\frac{1}{n} \sum_{t=1}^{n} \left. \frac{\partial \rho_\beta(x_t \mid \theta)}{\partial \theta} \right|_{\theta=\hat{\theta}_\beta} \left. \frac{\partial \rho_B(x_t \mid \theta)}{\partial \theta'} \right|_{\theta=\hat{\theta}_\beta}$$

$$H_\beta := -\frac{1}{n} \sum_{t=1}^{n} \left. \frac{\partial^2 \rho_\beta(x_t \mid \theta)}{\partial \theta \partial \theta'} \right|_{\theta=\hat{\theta}_\beta}$$

3. 数値実験 と まとめ

多項式回帰によるモデル選択(平均予測二乗誤差)

真のモデル: 3次



モデルの候補

1. $y = a_0 + a_1 x$
2. $y = a_0 + a_1 x + a_2 x^2$

$\vdots$

5. $y = a_0 + a_1 x + \cdots + a_5 x^5$

GIC, IC_B は $\beta = 0.1, 0.3, \cdots, 1.9$ も選択する

結果

既存の方法　　新しい

	GIC_{β}	$AICR_{0.01}$	$AICR_{0.5}$	$AICR_{1.0}$	IC_B
平均予測二乗誤差	0.1758	0.1951	0.0471	0.0518	0.0464

ただしB＝1.0とし，PMSEは1000回の平均を取った

まとめ

今回提案したIC_Bは

ロバスト性 を持ち

βの選択 を含めて

良い予測を示す.

今後の展望:

DPDは外れ値が多いとバイアスが大きい

$q(x) := (1-\varepsilon)\underbrace{f(x)}_{\text{真の分布}} + \varepsilon\xi(x)$ とすると

$$\arg\min_{p\in\mathcal{D}} d_\beta(q;p) \approx \arg\min_{p\in\mathcal{D}} d_\beta((1-\varepsilon)f,p) \neq f$$

Gamma-divergence (Fujisawa and Eguchi 2008) に置き換える

$$\arg\min_{p\in\mathcal{D}} d_\gamma(q;p) \approx \arg\min_{p\in\mathcal{D}} d_\gamma((1-\varepsilon)f,p) \approx f$$

参考文献

1. Ronchetti, Elvezio. (1982). Robust alternatives to the F-test for the linear model. *Springer Netherlands*.
2. Konishi, Sadanori. and Genshiro, Kitagawa. (1996). Generalised information criteria in model selection. *Biometrika* **83,4** 875-890.
3. Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference.* **90,2** 227–244.
4. Basu, A. et al. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*. **85,3** 549-559
5. Mattheou, Kyriacos, and Alex Karagrigoriou. (2006). Robust model selection criteria with applications.
6. Fujisawa,H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *J. Stat. Plan. And Inf.*
7. Yata, Kazuyoshi. et al. (2011). Note on robust model selection by density power divergence in a contaminated regression model. *数理解析研究所講究録*. **1758** 150-159

補足スライド

β-推定量は反復的に計算できる:

① $\hat{\theta}^{(0)} := \hat{\theta}_{\mathrm{MLE}}$ とする.

② $E_{\hat{q}}[p(X_t \mid \hat{\theta}^{(i)})^{\beta} S(X_t, \theta)]$

$$= E_{p_\theta}[p(X \mid \theta)^{\beta} S(X, \theta)]\Big|_{\theta=\hat{\theta}^{(i)}}$$

の解を $\hat{\theta}^{(i+1)}$ と定義する.

③ $\|\theta^{(i+1)} - \theta^{(i)}\| < \varepsilon$ となるまで②を繰り返す

重回帰での反復計算の例

$X_i \overset{\text{i.i.d.}}{\sim} U_d(0,1)$ かつ $Y_i \overset{\text{i.i.d.}}{\sim} N(X_i^T\beta, 1^2)$ を仮定する.

確率モデルを $p(\{\mathbf{x}, y\}, \beta) := 1_{[0,1]^d}(\mathbf{x})N(y; \mathbf{x}^T\beta, 1^2)$

とすると, パラメータの更新式は陽的に解けて,

$$\beta^{(i+1)} = (X^T W^{(i)} X)^{-1} X^T W^{(i)} y$$

ただし $W^{(i)} = (p(x_1 \mid \theta^{(i)}), \cdots, p(x_N \mid \theta^{(i)}))$

他の情報量規準との関係

$$\mathrm{IC}_B = \mathrm{AICR}_\beta,\ (B = \beta)$$

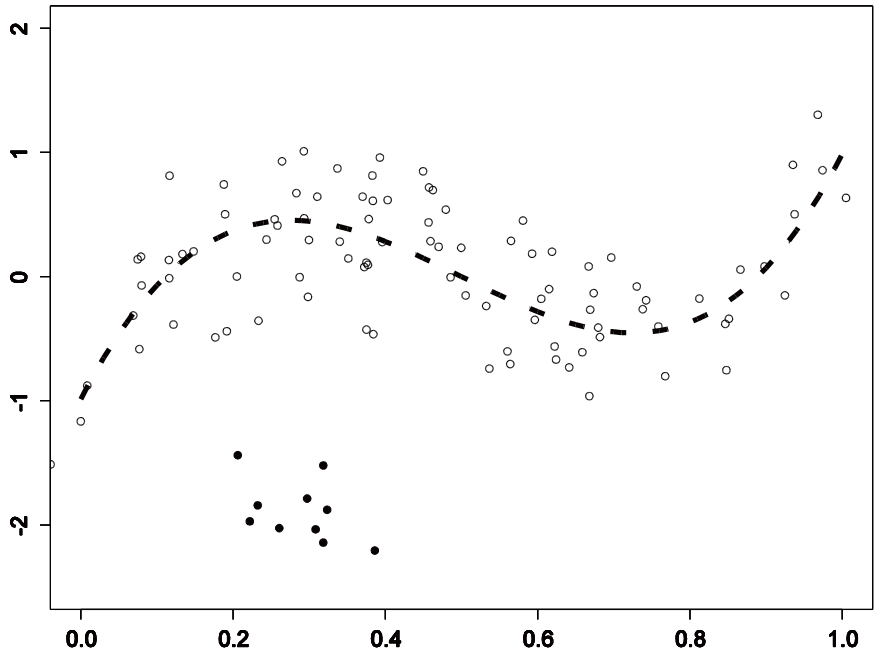
$$\mathrm{IC}_B \to \mathrm{GIC}_\beta,\ (B \to 0)$$

$$\mathrm{IC}_B \to \mathrm{TIC},\ (B, \beta \to 0)$$

$q(x) = p(x \mid \bar{\theta}_\beta)$ の仮定の下で

$$\mathrm{IC}_B \to \mathrm{AIC},\ (B, \beta \to 0)$$

モデル選択 (真のモデルを含まない)



データの設定

X: N(0.5,0.3^2)

Y: Xの3次曲線

N=100 (外れ値10%)

誤差: $Y - X^T\beta \sim N(0, 0.2^2)$

確率モデルの設定

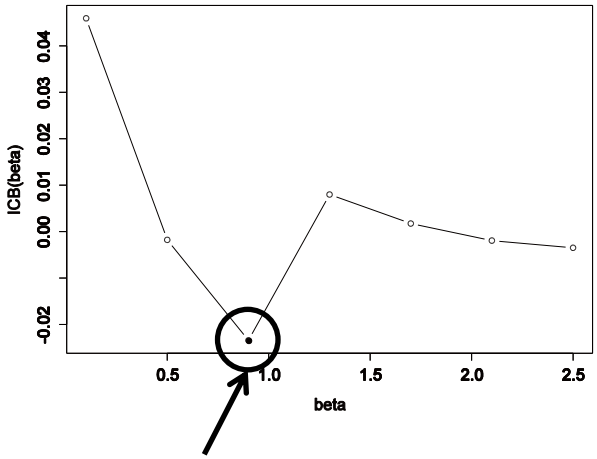
X: 一様に分布

Y: Xのd次曲線

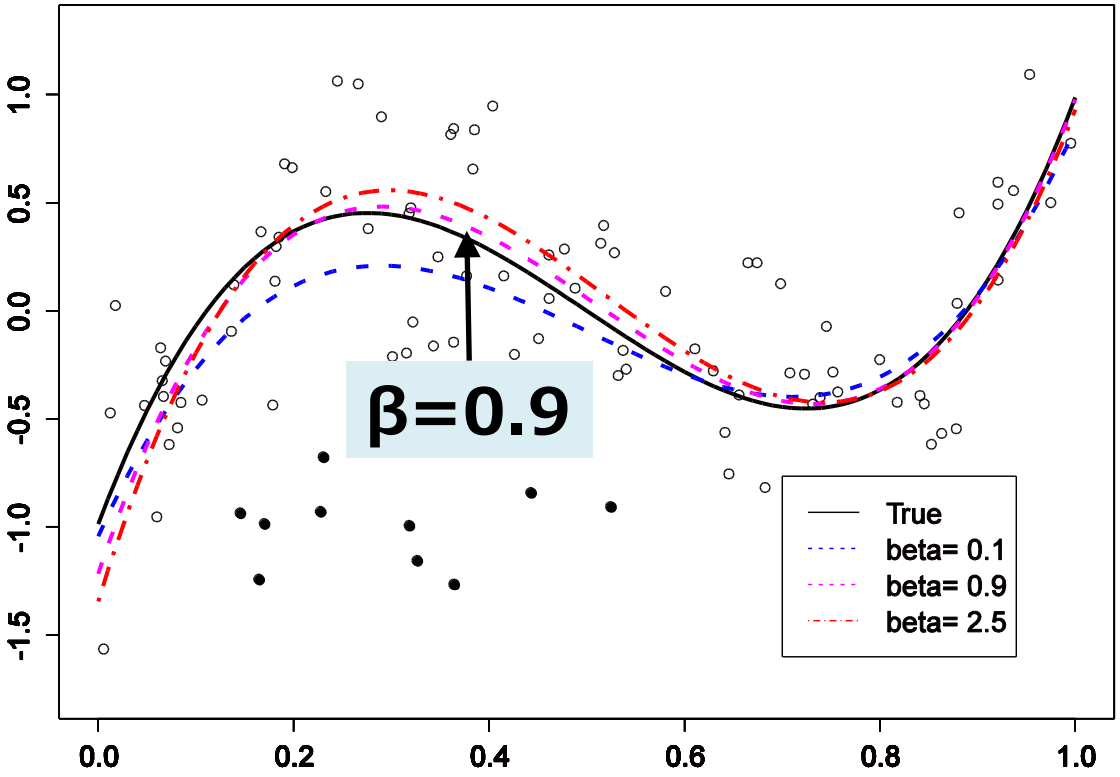
誤差: $Y - X^T\beta \sim N(0, \sigma^2)$

IC_Bを使うと，βを選択できる．

例)3次多項式に限定，IC_Bが最小となるβを選択する．



β=0.9で最小



beta=＊：外れ値を含むデータで，beta=＊としてbeta推定．