3つのパラメータをもつ分布近傍とロバスト推測

垣内逸郎　（神戸大・システム情報学研究科）
木村美善　（南山大・情報理工）

1 要旨

ロバスト推測理論においては，モデル分布からの「ずれ」や「乖離」を表現するためにモデル分布の近傍が用いられ，ε-汚染近傍や全変動近傍などがよく用いられている．本報告では，ある容量により定義される3つのパラメータをもつ新たな分布近傍を導入し，その特徴付けを与える．また，ロバスト推測への応用として，近傍を定義する分布のメディアンに対するロバスト・ノンパラメトリック信頼区間と検定の構成，およびそのロバストネスについて論じる．導入した分布近傍は，ϵ-汚染近傍や全変動近傍を一般化した Rieder 近傍や (c,γ)-近傍を特別の場合として含むものであり，得られた結果は，Yohai and Zamar (2004) や Ando, Kakiuchi and Kimura (2009) の結果を含むものである．

2 $(\mathbf{c_1}, \mathbf{c_2}, \boldsymbol{\gamma})$-近傍

$\mathbb{R}$ を実数直線, $\mathscr{B}$ を $\mathbb{R}$ の部分集合からなるボレル集合族, $\mathscr{M}$ を $(\mathbb{R}, \mathscr{B})$ 上の確率分布の全体とし， F° を連続な分布とする．このとき， $0 \le \gamma < 1,\ 0 \le c_1 \le 1-\gamma \le c_2 < \infty,\ c_1 \ne c_2$ に対し， F° の分布近傍 $\mathscr{P}_{c_1,c_2,\gamma}(F^\circ)$ を次のように定義する．

$$h(x) = \min\{c_2 x + \gamma, c_1 x + 1 - c_1\}, \qquad 0 \le x \le 1$$

とし，$\phi \ne {}^\forall A \in \mathscr{B}$ に対し $v_h\{A\} = h(F^\circ\{A\})$ かつ $v_h\{\phi\} = 0$ とすると v_h は特殊容量となり，これを用いて

$$\mathscr{P}_{c_1,c_2,\gamma}(F^\circ) = \{G \in \mathscr{M} \mid G\{A\} \le v_h\{A\}, {}^\forall A \in \mathscr{B}\}$$

とする．これを (c_1, c_2, γ)-近傍と呼ぶ．(c_1, c_2, γ)-近傍は3つのパラメータ c_1, c_2, γ を持ち，これらの値を変えることにより，ε-汚染近傍や全変動近傍，また上限の確率が ε-汚染近傍と全変動近傍の小さい方で抑えられる近傍などが得られる．(c_1, c_2, γ)-近傍は，直感的によりわかりやすい形で，次のように密度関数を用いることによって与えられることが分かる． F° の密度関数を f° とし，

$$\mathscr{F}_{c_1,c_2,\gamma}(F^\circ) = \left\{ F \in \mathscr{M}_c \mid \frac{c_1}{1-\gamma} f^\circ \le f \le \frac{c_2}{1-\gamma} f^\circ \right\}$$

とする．ここで，$\mathscr{M}_c$ は $(\mathbb{R}, \mathscr{B})$ 上の連続な確率分布全体からなる集合であり，F の密度関数を f とする．このとき，

$$\mathscr{P}_{c_1,c_2,\gamma}(F^\circ) = \left\{ G \in \mathscr{M} \mid G = (1-\gamma)F + \gamma K,\ {}^\forall F \in \mathscr{F}_{c_1,c_2,\gamma}(F^\circ),\ {}^\forall K \in \mathscr{M} \right\}.$$

3 ロバスト推測への応用

$\boldsymbol{X}_n = (X_1, \cdots, X_n)$ を $G \in \mathscr{P}_{c_1,c_2,\gamma}(F^\circ)$ に従う大きさ n の無作為標本とし，$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ をその順序統計量とする．ここで，F° は未知とし，ただ一つのメディアン $\theta = (F^\circ)^{-1}(1/2)$ をもつとする．また，$0 \leq \gamma < 1/2,\ c_2 < 2(1-\gamma)$ と仮定する．このとき，$\mathscr{P}_{c_1,c_2,\gamma}(F^\circ)$ に属する任意の分布 G の下で，F° のメディアンに関し，与えられた信頼係数や有意水準を満足する信頼区間と検定を構成し，そのロバストネスを与える．

符号検定統計量に基づく θ のロバスト・ノンパラメトリック信頼区間 I_n を，次のように構成する．

Z_n を 2 項分布 $B(n, (1-\lambda)/2)$ に従う確率変数とし, λ $(0 \leq \lambda < 1)$ を

$$\lambda = \min\{(1-\gamma) - c_1,\ c_2 - (1-\gamma)\} + \gamma$$

とする．また，$\alpha^*(n,k,\lambda) = 1 - P(k < Z_n < n-k)$ とし，$n, \alpha\,(0 \leq \alpha < 1)$ に対して, 非負の整数 k_n を $k_n = \arg\min_k |\alpha^*(n,k,\lambda) - \alpha|$ により定義する． このとき，信頼区間 $I_n = [X_{(k_n+1)}, X_{(n-k_n)})$ は，(c_1,c_2,γ) - robust nonparametric coverage $1-\alpha^*$ である．すなわち，

$$\inf_{G \in \mathscr{P}_{c_1,c_2,\gamma}(F_0)} P_G(X_{(k_n+1)} \leq \theta < X_{(n-k_n)}) = 1 - \alpha^*.$$

次に，λ に基づいて構成された区間 I_n に対し，実汚染 $(\tilde{c}_1, \tilde{c}_2, \tilde{\gamma})$ の下でのロバストネスを考える．区間列 $\{I_n\}$ の F° における $(\tilde{c}_1, \tilde{c}_2, \tilde{\gamma})$ の下での最大漸近幅を

$$L\{I_n, F^\circ, (\tilde{c}_1, \tilde{c}_2, \tilde{\gamma})\} = \sup_{G \in \mathcal{P}_{\tilde{c}_1,\tilde{c}_2,\tilde{\gamma}}(F^\circ)} \operatorname{essup} \limsup_{n\to\infty} (X_{(n-k_n)} - X_{(k_n+1)}),$$

と定義し，$L\{I_n, F^\circ, (\tilde{c}_1, \tilde{c}_2, \tilde{\gamma})\} < \infty$ のとき，I_n は F° で $(\tilde{c}_1, \tilde{c}_2, \tilde{\gamma})$ - robust length をもつという．また，Length breakdown size λ^* を

$$\lambda^*\{I_n, F^\circ, (\tilde{c}_1, \tilde{c}_2, \tilde{\gamma})\} = \sup\{\lambda \mid L\{I_n, F^\circ, (\tilde{c}_1, \tilde{c}_2, \tilde{\gamma})\} < \infty\},$$

により定義する．次の結果について報告する．

(1) $L\{I_n, F^\circ, (\tilde{c}_1, \tilde{c}_2, \tilde{\gamma})\}$ の導出.

(2) $\lambda^*\{I_n, F^\circ, (\tilde{c}_1, \tilde{c}_2, \tilde{\gamma})\} = 1 - 2\tilde{\gamma}$.

(3) I_n の F° の下での Efficiency.

さらに，(c_1,c_2,γ) - 近傍の下での検定とそのロバストネスについても報告する．

参考文献

[1] Ando, M., Kakiuchi, I. and Kimura, M. (2009) Robust nonparametric confidence intervals and tests for the median in the presence of (c,γ)-contamination, *J. Statist. Plann. Inference.*, **139**, 1836-1846.

[2] Yohai, V. J. and Zamar, R. H. (2004) Robust nonparametric inference for the median, *Ann. Statist.*, **32**, 1841-1857.