汚れのあるデータとロバスト推測

M2011MM055 沖翔太
指導教員：木村美善

1 はじめに

データの汚れを記述するための近傍として Kakiuch and Kimura(2012) が提案した (c_1, c_2, γ)-近傍 $\mathscr{P}_{c_1,c_2,\gamma}(F^\circ)$ を用いる．この近傍は，3つのパラメータにより決まる特殊容量から生成される新しい近傍である．パラメータ γ は汚染の大きさを表し，c_1, c_2 はモデル分布 F° からのずれの大きさを表現している．この近傍がロバスト推測にどのように応用されるのか見るために，具体的な問題に対して具体的な分布を適用して考察する．一般的な分布をモデル分布とし，(c_1, c_2, γ)-近傍がどのような近傍であるかを考察する．また，この 3 つのパラメータを持つ近傍を用いることにより，モデル分布からの「ずれ」があった場合にどのように信頼性が保証されるのかについて，シュミレーションにより考察する．

2 ロバスト推測

$\mathbb{R}$ 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する．$X_1, X_2, ..., X_p$ を分布Gに従う互いに独立な標本とする．Gが特定の分布 F° に近似的に等しいことがわかっているとき，F° の未知の特性値に関する統計的推測を行う．

ロバスト推測で達成すべきこととして，Huber は以下のことを主張している．

Efficienncy ： 仮定された分布近傍のもとで，高い効率をもつ．

Stability ： 仮定されたモデル分布からのわずかな「ずれ」や「乖離」があっても，安定した振る舞いをする．

Breakdown ： 仮定されたモデル分布からある程度大きな「ずれ」や「乖離」があっても極端に効率が下がらず，破綻しない．

3 分布近傍

3.1 様々な分布近傍

これまでに提案された分布近傍は多くあるが，その中で代表的な近傍を列挙する．

ε -近傍:

$$\mathscr{P}_\epsilon(F^{\mathrm{o}}) = \{G \in \mathscr{M} | G\{A\} \leq (1-\epsilon)F^{\mathrm{o}}\{A\} + \epsilon, \ {}^\forall \mathrm{A} \in \mathscr{B}\}, 0 \leq \epsilon \leq 1$$

全変動近傍:

$$\mathscr{P}_{0,\delta}(F^{\mathrm{o}}) = \{G \in \mathscr{M} | G\{A\} \leq \min\{F^{\mathrm{o}}\{\mathrm{A}\} + \delta, 1\}, \ {}^\forall \mathrm{A} \in \mathscr{B}\}, 0 \leq \delta \leq 1$$

(c, γ)-近傍 (Ando，Kakiuchi and Kimura，2009):

$$\mathscr{P}_{c,\gamma}(F^{\mathrm{o}}) = \{G \in \mathscr{M} | G\{A\} \leq \min\{cF^{\mathrm{o}}\{A\} + \gamma, 1\}, \ {}^\forall \mathrm{A} \in \mathscr{B}\}, 0 \leq \gamma \leq 1, 1-\gamma \leq c \leq \infty$$

他にも Rider 近傍などの近傍もある．また (c, γ)-近傍は，c と γ に特別な値を入れることで多くの新しい近傍を得ることができる．これらの近傍の定義では上限しか示されていないので下限も示す．

[(c, γ)-近傍の場合]

$$\begin{aligned} 1 - \mathrm{G(A)} &\leq \min\{c(1 - F(A)) + \gamma, 1\} \\ \mathrm{G(A)} &\geq \max\{c(F(A) - 1) - \gamma, -1\} + 1 \\ &\geq \max\{c(F(A) - 1) - \gamma + 1, 0\} \end{aligned}$$

となる．他の近傍でも同様にして示すことができる．

3.2 特殊容量

Bednaruski(1981) は，連続な凹関数 $h : [0, 1] \to [0, 1]$ で $h(1) = 1$ を満たすものに対して，集合関数 $v_h : \mathscr{B} \to [0, 1]$ を

$$v_h\{A\} = \begin{cases} h(F^\circ\{A\}) & \phi \neq A \in \mathscr{B} \text{ のとき} \\ 0 & A = \phi \text{ のとき} \end{cases} \quad (1)$$

により定義し，これを特殊容量 (special capacity) と呼んだ．($h(0) = 0$ が成り立つとき 2-alternating Choquet capacity になる) この確率測度を一般化した特殊容量 v_h を用いて得られる F° の分布近傍

$$\mathscr{P}_h(F^{\mathrm{o}}) = \left\{G \in \mathscr{M} | G\{A\} \leq v_h\{A\}, {}^\forall \mathrm{A} \in \mathscr{B}\right\}$$

となる．

4 (c_1, c_2, γ)-近傍について

次のような 3 つのパラメータを持つ分布近傍を定義する．

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

ただし，$0 \leq c_1 \leq 1 - \gamma \leq c_2 \leq \infty, c_1 \neq c_2, 0 \leq \gamma \leq 1$ とし，この h に対応する F° の分布近傍

$$\begin{aligned} \mathscr{P}_{c_1,c_2,\gamma}(F^{\mathrm{o}}) = \{G \in \mathscr{M} | G\{A\} \leq \min\{c_2 F^{\mathrm{o}}\{\mathrm{A}\} \\ + \gamma, c_1 F^{\mathrm{o}}\{\mathrm{A}\} + 1 - c_1\}, {}^\forall \mathrm{A} \in \mathscr{B}\} \end{aligned} \quad (2)$$

を，(c_1, c_2, γ)-近傍と呼ぶ．また，注意することとして

1. $h(x) = x$ のとき，$v_h = F^\circ$．すなわち，確率測度は特殊容量であり，$\mathscr{P}_h(F^\circ) = F^\circ$.
2. (c_1, c_2, γ)-近傍は，

$$\begin{aligned} \mathscr{P}_{c_1,c_2,\gamma}(F^{\mathrm{o}}) = \{G \in \mathscr{M} | c_1 F^{\mathrm{o}}\{A\} \leq G\{A\} \\ \leq c_2 F^{\mathrm{o}}\{A\} + \gamma, {}^\forall \mathrm{A} \in \mathscr{B}\} \end{aligned}$$

とも表せること．

(c_1, c_2, γ)-近傍は，3 つのパラメータ c_1, c_2, γ を変化させることで多様な近傍を得ることができる．(c, γ)-近傍より変数を増やすことでさらに多くの近傍が得られる．実際に c_1, c_2, γ に値を代入することでどのような近傍が得られるか調べる．

・$c_1 = 1 - \epsilon, \gamma = \epsilon$ または $c_2 = 1 - \epsilon, \gamma = \epsilon$ のとき，ϵ-近傍に等しい．

・$c_1 = 0, c_2 = 1, \gamma = \delta$ のとき，全変動近傍に等しい．

・$c_1 = 0, c_2 = 1 - \epsilon, \gamma = \epsilon + \delta$ のとき，Rieder 近傍に等しい．

・$c_1 = 0, c_2 = c, \gamma = \gamma$ のとき，(c,γ)-近傍に等しい．

・$\gamma = 0$ のとき，中心の汚れ近傍となる．この場合には

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

であり，局外値などを含まない中心部のみが汚れた近傍である．

・ϵ-近傍と全変動近傍を組み合わせることによって $TN\epsilon$-近傍ができる．$c_1 = 1 - \epsilon$, $c_2 = 1$, $\gamma = \delta (\epsilon > \delta)$ のとき

$$h(x) = \min\{x + \delta, (1 - \epsilon)x + \epsilon\}, 0 \le \delta \le \epsilon \le 1$$

となる．これは，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分布近傍となる．この分布近傍を汚染と全変動の混合近傍と呼ぶ．

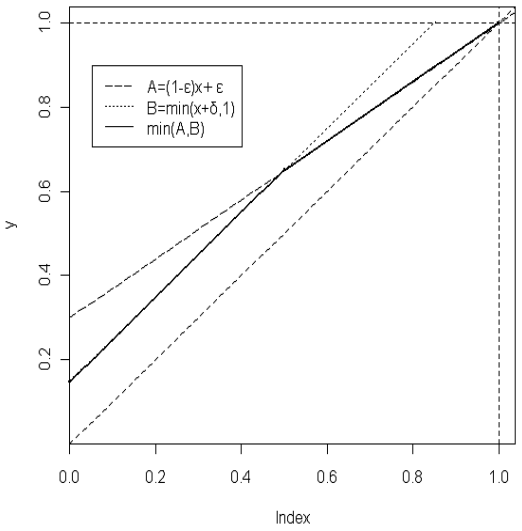

図 1 混合近傍のグラフ

近傍のまとめ

図 2 をみると，2 つの直線の最小値を通る直線を選んでいるものが (c_1, c_2, γ)-近傍であることがわかる．

c_1, c_2, γ は，それぞれの値の取り方により多くの近傍が得られる．汚れの近傍，全変動近傍,Rider 近傍，(c, γ)-近傍などを特別な場合としてすべてを含む分布近傍になる．図 1 における関数 $h(x)$ の切片 ($h(0)$ の値) は空集合を除いた集合に対するその上限の確率を示している．すなわち，極端な値 (局外値, 大きな値や小さな値) を取る確率を持つという分布を示している．したがって，切片の値が大きいほど，外れ値を取る確率が高くなる．また (special capacity) の $h(x)$ が 1 となる x の値が小さいほど，正となる確率をもつ集合さえ無視されやすくなることを意味する．その意味でリーダー近傍は，汚れの近傍や全変動近傍を含む大きな分布近傍を表していると言える．

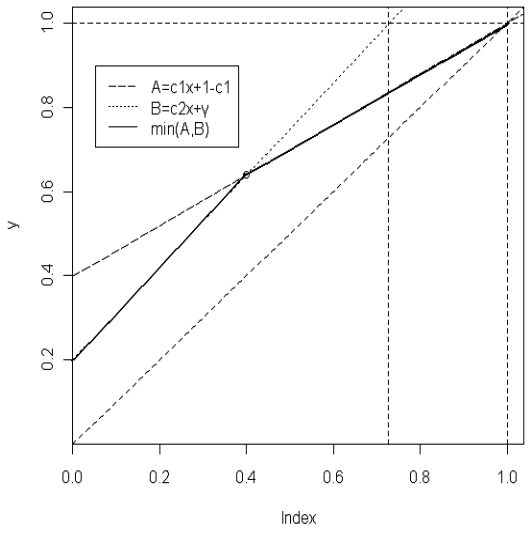

図 2 (c_1, c_2, γ)-近傍のグラフ

5 (c_1, c_2, γ)-近傍の特徴

5.1 密度関数による特徴

Kakiuchi and Kimura(2012) の Theorem 3.2 は式 (2) により定義された (c_1, c_2, γ)-近傍の密度関数による特徴づけを以下のように与えている．

(c_1, c_2, γ)-近傍は

$$\mathscr{P}_{c_1,c_2,\gamma}(F^{\circ}) = \{G = (1-\gamma)F + \gamma K | F \in \mathscr{F}_{c_1,c_2,\gamma}(F^{\circ}), K \in \mathscr{M}\} \quad (3)$$

として表される．ここで，連続な分布 F° の密度関数を f° とし

$$\mathscr{F}_{c_1,c_2,\gamma}(F^{\circ}) = \left\{ F \in \mathscr{M}_c | \frac{c_1}{1-\gamma} f^{\circ} \le f \le \frac{c_2}{1-\gamma} f^{\circ} \right\} \quad (4)$$

とする．ただし，f は F の密度関数とし，$\mathscr{M}_c$ は $(\mathbb{R}, \mathscr{B})$ 上の連続な確率分布からなる集合である．

5.2 (c_1, c_2, γ)-近傍の限界近傍

F_L, F_R を $\mathscr{F}_{c_1,c_2,\gamma}(F^{\circ})$ に含まれる以下の分布として定義する．

$$F_L(x) = \begin{cases} \dfrac{c_2}{1-\gamma} F^{\circ}(x) & (x \le x_L) \\ \dfrac{c_1}{1-\gamma} F^{\circ}(x) + (1 - \dfrac{c_1}{1-\gamma}) & (x > x_L) \end{cases}$$

$$F_R(x) = \begin{cases} \dfrac{c_1}{1-\gamma} F^{\circ}(x) & (x \le x_R) \\ \dfrac{c_2}{1-\gamma} F^{\circ}(x) + (1 - \dfrac{c_2}{1-\gamma}) & (x > x_R) \end{cases}$$

ここで，それぞれ x_L, x_R は

$$x_L = (F^\circ)^{-1}\left(\frac{1-\gamma-c_1}{c_2-c_1}\right),$$
$$x_R = (F^\circ)^{-1}\left(\frac{c_2+\gamma-1}{c_2-c_1}\right)$$

である．

6 (c_1, c_2, γ)-近傍の実用化

6.1 背景

分布近傍を用いたロバスト推測の結果が実際にどのように応用されるのかみるために具体的な分布に対して具体的な近傍を適用することによりその有効性を探る．また，この近傍内の分布を用いて検定の信頼性を考察する．

6.2 (c_1, c_2, γ)-近傍の問題

具体的な分布として標準正規分布と t-分布を用いる．モデル分布を標準正規分布とする．標準正規分布を何倍かすることによって t-分布の全ての自由度の密度関数が近傍内に含まれるようにし，そのときの c_1, c_2, γ の値がどうなるか考察していく．

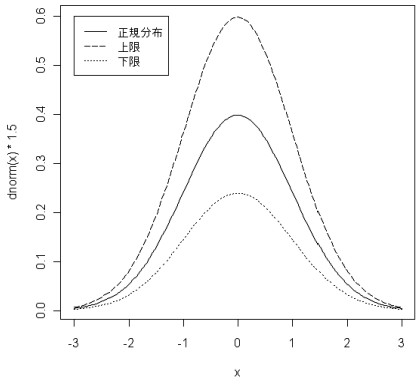

図 3 (c_1, c_2, γ)-近傍の密度関数

図 3 の上限と下限の中に t-分布を挟み込む問題である．

6.3 考え方

標準正規分布と t-分布の分散は異なっている．そこで，標準正規分布と t-分布の分散を合わせるために次の二つの方法を用いる．

1. 標準正規分布の分散に変換する
2. t-分布の分散に変換する

6.4 計算方法

標準正規分布と t-分布の裾の部分を比べると t-分布の方が確率が大きくなる．そのため，ある範囲からは t-分布の密度関数を用いる．このとき，標準正規分布と t-分布の密度関数 f° と t_m を $\pm x_a$ でつないで連続にするために k を以下で定める．

$$f(x) = \begin{cases} t_n(x), & (x > |x_a|) \\ kf^\circ(x), & (-x_a \le x \le x_a) \end{cases}$$

ただし k は，$kf^\circ(x_a) = t_n(x_a)$ となる定数である．次に，密度関数 $f^*(x) = cf(x)$ と定義する．ただし $c = 1/\int_{-\infty}^{\infty} f(x)dx$.

この $f^*(x)$ を用いて c_1, c_2, γ 値を求める．

6.5 結果

- 標準正規分布の分散に合わせた場合

表 1 挟み込む最小の自由度による違い

	n=3 以上	n=5 以上	n=10 以上
k	0.464	0.779	0.949
c	2.065	1.260	1.046
$\frac{c_2}{1-\gamma}f^*$(上限)	1.662	1.249	1.097
$\frac{c_1}{1-\gamma}f^*$(下限)	0.465	0.700	0.853
$c_1(\gamma$=0.01)	0.461	0.693	0.844
$c_2(\gamma$=0.01)	1.645	1.237	1.086
$c_1(\gamma$=0.05)	0.442	0.665	0.810
$c_2(\gamma$=0.05)	1.579	1.187	1.042

表 1 は, 挟み込む最小の自由度を変化させたときの表である．自由度の最小の値が大きくなればなるほど上限・下限の間は狭くなる．

表 2 範囲による違い（n=3 以上）

x_a	1.96	2.3	2.58	3.0
k	0.464	0.568	0.619	1.436
c	2.065	1.711	1.576	0.691
$\frac{c_2}{1-\gamma}f^*$(上限)	1.662	1.641	1.633	1.739
$\frac{c_1}{1-\gamma}f^*$(下限)	0.465	0.460	0.457	0.450
$c_1(\gamma$=0.01)	0.461	0.455	0.453	0.445
$c_2(\gamma$=0.01)	1.645	1.624	1.617	1.722
$c_1(\gamma$=0.05)	0.442	0.437	0.435	0.427
$c_2(\gamma$=0.05)	1.579	1.559	1.551	1.652

表 2 より範囲 x_a を 2.58 まで広げていくと上限と下限は小さくなっていることがわかる．また γ の値に観点を置くと，γ を増やしていくと c_1,c_2 の値は汚染がない値より小さい値になる．x_a の範囲が 3.0 になると上限が大きくなっている．この理由は, 図より明らかとなる．

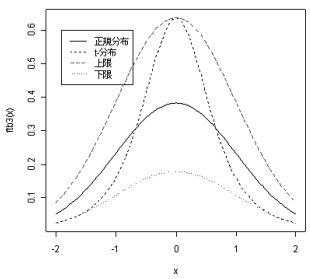

図 4 範囲 x_a=1.96

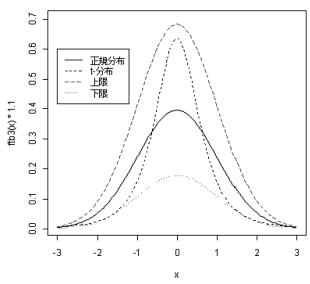

図 5 範囲 x_a=3.0

図 4,5 を比べると，t-分布と上限の接点が異なっている

ことがわかる．接点が範囲の端になっていることで上限の値が大きくなった．このことから範囲が大きくなるにつれて上限の値は大きくなることがわかる．

- t-分布の分散に合わせた場合

表 3　t-分布の分散に変換した場合

$\alpha = 0.05$	n=3	n=5	n=7	n=10
x_a	5.840	4.032	3.499	3.169
k	3.072	2.098	1.585	1.322
c	0.324	0.475	0.628	0.753
$\frac{c_2}{1-\gamma}f^*$(上限)	3.076	2.104	1.589	1.326
$\frac{c_1}{1-\gamma}f^*$(下限)	0.488	0.690	0.783	0.850
$c_1(\gamma = 0.01)$	0.444	0.683	0.775	0.842
$c_2(\gamma = 0.01)$	3.049	2.083	1.573	1.313
$c_1(\gamma = 0.05)$	0.426	0.655	0.744	0.807
$c_2(\gamma = 0.05)$	2.926	1.999	1.509	1.260

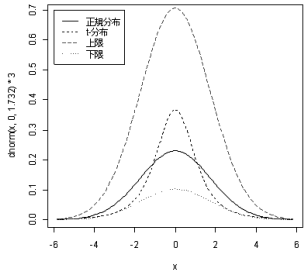

図 6　自由度 n=3

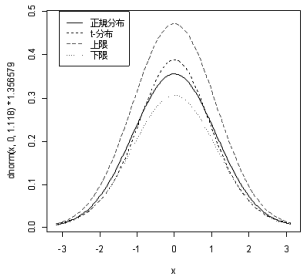

図 7　自由度 n=10

自由度が大きくなるにつれて上限と下限の幅が小さくなっていることがわかる．この方法に関しても $-x_a, +x_a$ が t-分布と上限との接点になっている．よって，範囲 x_a が大きくなると上限の値も大きくなる．

これらの結果より，正規分布の (c_1, c_2, γ)-近傍により t-分布を覆うことができる．$\frac{c_1}{1-\gamma}f^*, \frac{c_2}{1-\gamma}f^*$ の値は，範囲が大きくなることで幅が広くなり，自由度が大きくなることで狭くなる．

6.6　検定

(c_1, c_2, γ)-近傍によるモデル分布からの「ずれ」の影響を次の検定問題で調べる．$X_1, X_2, ..., X_n$を $N(\mu, 1)$ からの無作為標本とするとき，次の検定問題を有意水準 5% で考える．

帰無仮説を $H_0 : \mu = 0$，対立仮説を $H_1 : \mu > 0$ として，有意水準を λ=0.05，第二種の過誤を犯す確率を β，検出力を $1 - \beta$ とする．

この検定問題に対する一様最強力検定の棄却域は

$$R = \{\sqrt{n}\bar{x} > 1.645\}$$

となる．有意水準は

$$0.05 = \lambda = P(\sqrt{n}\bar{x} > 1.645|\mu = 0) \qquad (5)$$

となり，検出力は

$$1 - \beta(\mu) = P(\sqrt{n}\bar{x} > 1.645|\mu) \qquad (6)$$

となる．この一様最強力検定がモデル分布である正規分布からずれたときにどういう影響を受けるかについて調べる．具体的にずれた分布は上限と下限を用い，確率が一番右側に寄った分布 (最大分布) と左側に寄った分布 (最小分布) である．

ここでは μ=2 の場合の検定を行い，標本数を変化させたときの比較を行う．また，標準正規分布の分散に変換した場合の上限と下限を用いる (挟み込む自由度の最小が 3 以上，x_a=1.96 のとき)．

表 4　式 (5)，式 (6) の比較

標本数	3	5	7	10
有意水準 (ずれなし)	0.049	0.050	0.051	0.050
有意水準 (ずれあり)	0.529	0.717	0.836	0.932
検出力 (ずれなし)	0.965	0.998	1.000	1.000
検出力 (ずれあり)	0.550	0.737	0.855	0.943

表 4 より，ずれがありとなしでは有意水準と検出力に大きな差がある．ずれがある場合には標本数が増えていくと，有意水準と検出力はともに大きくなることが分かる．またずれがある場合に有意水準 λ を 0.05 とするためには式 (5) の 1.645 の値をより大きい値にしなければならない．μ の値を 0 に近くすると検出力は低くなり，μ の値を大きくすると検出力が高くなる．

7　おわりに

本研究では,(c_1, c_2, γ)-近傍の実用化の一歩として正規分布によって t-分布を格納することを研究した．正規分布と t-分布の密度関数をつないだことによって,$\frac{c_1}{1-\gamma}f^*$と $\frac{c_2}{1-\gamma}f^*$ の値の幅が狭くなった．一様最強力検定のずれがあり，なしの場合の有意水準と検出力を調べ，ずれによる影響を考察した．今後の発展として, 他の分布で同様な研究を行うことや平均値の差の検定などの考察を行うことが考えられる．

参考文献

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

[2] Bednarski, T. (1981). On solutions of minmax tests problems for special capacities. Z. Wahreck. verw Gebiete. 10, 269-278.

[3] Kakiuchi, I. and Kimura, M. (2012). Robust non-parametric inference for the median under a new neighborhood of distributions, Technical Report of the Nanzan Academic Society Information Sciences and Engineering.

[4] Maronna, R. A., Martin, R. D., Yohai, V. J. (2006). *Robust Statistics Theory and Methods.*