裾の重い分布に対するボンフェロニ法とチューブ法

栗木 哲　　数理・推論研究系 教授

本発表は Evgeny Spodarev ウルム大教授 (統計数理研究所2022年度外国人客員教授) との共同研究に基づきます．

1　はじめに

検定統計量 $T_1, \dots, T_n$ に基づく多重検定の多重性 p 値は，帰無仮説の下での最大値分布で定義される．最大値分布を近似するための最も簡便な方法は，ボンフェロニ法

$$\mathbb{P}\Big(\max_i T_i > c\Big) \lessapprox \sum_i \mathbb{P}(T_i > c) \tag{1}$$

である．図1は T_i が多変量ガウス分布 $\mathcal{N}_3(0, (\rho_{ij}))$, $\rho_{ij} = 1$ $(i = j)$, $1/4$ $(i \neq j)$ に従う場合と，T_i をさらに共通の誤差推定量 $\widehat{\sigma} \sim \chi_\nu/\sqrt{\nu}$ $(\nu = 3)$ で除した場合の上側裾確率 (1) の対数 $\log \mathbb{P}(\max\{T_1, T_2, T_3\} > c)$ のプロットである．ボンフェロニ法は前者で非常に良い近似を与えるが，後者では必ずしもそうではない．本発表では多変量 t 分布を含む裾の重い分布に対するボンフェロニ法の誤差を解析する．

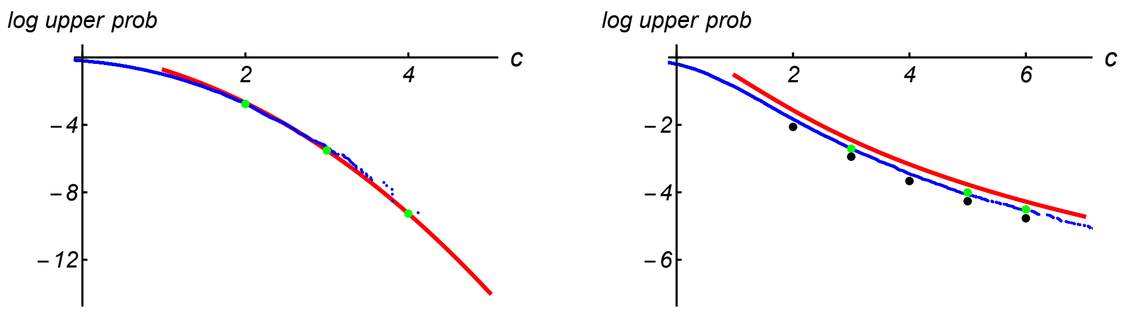

図1. 上側裾確率の対数，多変量ガウス (左) と多変量 t (右)
(青：10000回乱数，赤：ボンフェロニ法，緑：正確な上側確率，
紫：バイアス補正 (2)，黒：$\bar{\Delta}$ による下界)

2　ボンフェロニ近似の誤差評価

$x_1, \dots, x_n \in \mathbb{S}^{n-1}$ ($\mathbb{R}^n$ の単位球面) を $\langle x_i, x_j \rangle = \rho_{ij}$ をみたす点とする．

標準ガウス $\mathcal{N}_n(0, I_n)$ または多変量 t 分布 $\mathcal{N}_n(0, I_n)/(\chi_\nu/\sqrt{\nu})$ に従う確率ベクトル ξ を用いて $T_i =^d \langle x_i, \xi \rangle = \langle x_i, \eta \rangle \|\xi\|$, $\eta = \xi/\|\xi\|$, $i = 1, \dots, n$ と表現する．ここで η は $\|\xi\|$ と独立に $\mathbb{S}^{n-1}$ 上一様分布する．独立性より

$$\mathbb{P}\Big(\max_i T_i > c\Big) = \mathbb{E}\Big[\bar{B}(c/\|\xi\|)\Big], \quad \bar{B}(c) = \mathbb{P}\Big(\max_i \langle x_i, \eta \rangle > c\Big)$$

期待値は $\|\xi\|^2 \sim \chi^2_n$ または $\chi^2_n/(\chi^2_\nu/\nu) = nF^n_\nu$ (F 分布) についてとる．

$\bar{B}(c)$ の評価法を $n = 3$ で説明する．球帽を $A_i = \{\eta \in \mathbb{S}^{3-1} \mid \langle x_i, \eta \rangle > c\}$, $i = 1, 2, 3$ とおくと $\bar{B}(c) = \mathbb{P}\big(\eta \in A_1 \cup A_2 \cup A_3\big)$, $\eta \sim \mathrm{Unif}\big(\mathbb{S}^{3-1}\big)$ である．球面ボロノイを用いて $A_1 \cup A_2 \cup A_3$ の面積を重複なく計算する．

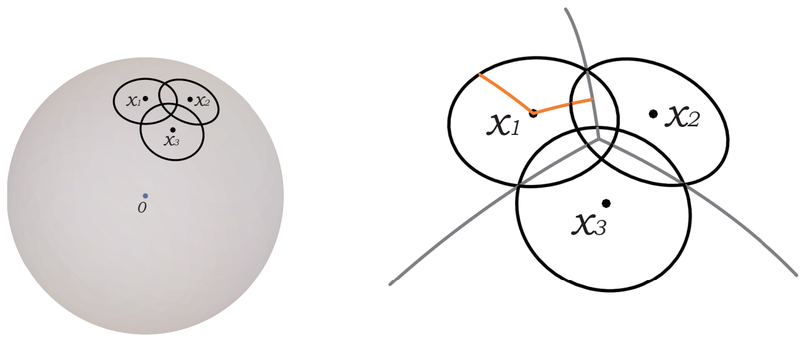

図2. 球帽の和 (左) と勢力圏 (右)

点 x_i を出発点とし方向 v (v は x_i と直交) にすすむ大円

$$\varphi(s) = x_i \cos s + v \sin s, \quad 0 \le s < 2\pi$$

は $s < \cos^{-1} c$, $\langle x_i, \varphi(s) \rangle > \langle x_j, \varphi(s) \rangle$ $(\forall j \neq i)$ の範囲で x_i の勢力圏内である．円弧 $\big\{\varphi(s) \mid 0 < s < \min\{\cos^{-1} c, \theta_i(v)\}\big\}$，ただし

$$\theta_i(v) = \tan^{-1} \min_{j \neq i} \frac{1 - \langle x_i, x_j \rangle}{\langle x_j, v \rangle}$$

をすべての v (s.t. $\langle v, x_i \rangle = 0$) について足し合わせる（積分する）と，$x_i$ の勢力圏の面積が得られる．

定理 2.1 (ボンフェロニ法の相対誤差). $B \sim B_{\frac{1}{2}, \frac{n-1}{2}}$, $V_i \sim \mathrm{Unif}(\{v \in \mathbb{S}^{n-1} \mid \langle v, x_i \rangle = 0\})$ とするとき

$$\Delta(c) = 1 - \frac{\mathbb{P}(\max_i T_i > c)}{\sum_i \mathbb{P}(T_i > c)} = \frac{\sum_{i=1}^n \mathbb{P}\big(c^2/\|\xi\|^2 < B < \cos^2 \theta_i(V_i)\big)}{n\mathbb{P}\big(c^2/\|\xi\|^2 < B\big)}$$

3　裾の重い分布

確率変数の分布関数を F，上側確率を $\bar{F} = 1 - F$ で表す．

裾の重い分布はその裾の重さに応じて，

$$\mathcal{H}\ (\text{Heavy-tailed}) \supset \mathcal{L}\ (\text{Long-tailed}) \supset \mathcal{S}\ (\text{劣指数}) \supset \mathrm{RV}\ (\text{正則変動})$$

と階層的にクラス分けされる．特に $F \in \mathcal{S}$ (劣指数, Subexponential) とは $\lim_{x\to\infty} \frac{\overline{F*F}(x)}{\bar{F}(x)} = 2$ であること ($F * F$ は独立な2つの確率変数の和の分布関数) である．また $F \in \mathrm{RV}_{-\alpha}$ (指数 α の正則変動分布) とは ℓ を $\lim_{x\to\infty} \frac{\ell(\lambda x)}{\ell(x)} = 1, \forall \lambda > 0$ をみたす緩慢変動関数とするとき $\bar{F}(x) = \frac{\ell(x)}{x^\alpha}$ をいう．たとえば自由度 ν_1, ν_2 の F 分布は $\mathrm{RV}_{-\nu_2}$，また劣指数であるが正則変動でない例に対数正規分布がある．

(β, ℓ) $(\beta \le 1, \ell \in \mathrm{RV}_0)$ で指定される分布族を導入する．

$$\bar{F}(x) \sim C \exp\left(-\int_{x_0}^x q(t)\mathrm{d}t\right), \quad q(t) = \frac{\ell(t)}{t^\beta}$$

ただし ℓ の収束先 $\lim_{x\to\infty} \ell(x) = \gamma \in (0, \infty) \cup \{\infty\}$ が存在するとする．

表1. (β, ℓ) で指定される分布族

	$\beta = 0$	$\beta \in (0, 1)$	$\beta = 1$
$\gamma < \infty$	指数的	$\mathcal{S}$	$\mathrm{RV}_{-\gamma}$
$\gamma = \infty$	超指数的	$\mathcal{S}$	$\mathcal{S}$

4　主結果：誤差解析と近似の補正

T_i の上側確率を $\bar{F}(x)$ とおく．

定理 4.1 (正則変動分布でない場合). $\beta < 1$ または $\gamma = \infty$ とする．

$$\log \Delta(c) \sim -c^{2(1-\beta)} \ell(c^2) g_\beta(\cos^2 \theta_{\mathrm{cri}}), \quad c \to \infty$$

ただし $g_\beta(y) = \frac{y^{\beta-1}-1}{1-\beta}$ $(\beta < 1)$, $-\log y$ $(\beta = 1)$，$\theta_{\mathrm{cri}} = \min_i \min_{v:\, \|v\|=1,\, \langle v, x_i \rangle = 0} \theta_i(v) = \min_{i<j} \cos^{-1} \rho_{ij}/2$ は臨界半径．

カイ2乗分布 ($\beta = 0$, $\gamma = 1/2$)，対数正規分布 ($\beta = 1$, $\ell(x) = \log x$, $\gamma = \infty$) のときはそれぞれ

$$\log \Delta(c) \sim -(1/2) c^2 \tan^2 \theta_{\mathrm{cri}}, \quad -\log(c^2)(-\log \cos^2 \theta_{\mathrm{cri}})$$

定理 4.2 (正則変動分布の場合). $\beta = 1$, $\gamma < \infty$ とする．

$$\lim_{c\to\infty} \Delta(c) = \frac{1}{n} \sum_{i=1}^n \mathbb{P}\Big(\widetilde{B} < \cos^2 \theta_i(V_i)\Big), \quad \widetilde{B} \sim B_{\gamma + \frac{1}{2}, \frac{n-1}{2}} \tag{2}$$

さらに $0 < \lim_{c\to\infty} \Delta(c) < \bar{\Delta} := \mathbb{P}\big(\widetilde{B} < \cos^2 \theta_{\mathrm{cri}}\big)$

正則変動分布の場合は $c \to \infty$ であっても相対誤差は消えない．その誤差は $\bar{\Delta}$ で見積ることができる (図1右)．以上の解析は，ボンフェロニ法の一般化であるチューブ法についても行うことができる．