Robustní odhady vícerozměrného parametru polohy

Příklad (jednorozměrný)

Nechť $X_1, \dots, X_n$ je náh. výběr z $N(\theta, \sigma^2)$. $\theta \in \mathbb{R}$ je neznámý parametr (parametr polohy), $\sigma^2 > 0$ je rušivý parametr (parametr měřítka)

$$X_i = \theta + e_i, \text{ kde } e_i \sim N(0, \sigma^2)$$

Nechť $\hat{\theta}$ je odhad parametru θ. Jako měřítko jeho kvality bereme střední čtvercovou chybu: $E(\hat{\theta} - \theta)^2$. Chceme najít odhad $\hat{\theta}$, který tuto chybu minimalizuje (pro všechny hodnoty parametru θ, tj. stejnoměrně).

$\hat{\theta} = \overline{X}$ je hledaný odhad.

vícerozměrný případ: $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ je p-rozměrný parametr $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$.

$X_1, \dots, X_n$ je náh. výběr z p-rozměrného rozdělení

Nechť $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(X_1, \dots, X_n)$ je odhad parametru $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_p)^T$$

$L: \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ je ztrátová (riziková) funkce, jestliže $L(x, y) \geq 0$ a $L(x, x) = 0$

$L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ ... je ztráta, kterou utrpíme, když parametr $\boldsymbol{\theta}$ odhadneme pomocí $\hat{\boldsymbol{\theta}}$.

$$L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = \sum_{i=1}^{p} (\hat{\theta}_i - \theta_i)^2 \quad \text{... kvadratická ztrátová funkce}$$

$$R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = E\,L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = E\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = \sum_{i=1}^{p} E(\hat{\theta}_i - \theta_i)^2 \quad \text{.... riziko}$$

Poznámka

$p = 1$, pak riziko $R(\hat{\theta}, \theta) = E(\hat{\theta} - \theta)^2$ ... je střední čtvercová chyba

$\hat{\theta}_1$ a $\hat{\theta}_2$ dva odhady parametru θ. Řekneme, že $\hat{\theta}_1$ dominuje $\hat{\theta}_2$, jestliže $R(\hat{\theta}_1, \theta) \leq R(\hat{\theta}_2, \theta)$, $\forall \theta \in \Theta$ a existuje $\theta_0 \in \Theta$: $R(\hat{\theta}_1, \theta_0) < R(\hat{\theta}_2, \theta_0)$.

Definice

Odhad $\hat{\theta}$ parametru θ je přípustný (admissible), jestliže neexistuje žádný jiný odhad parametru θ, který by jej dominoval.

$X_1, \dots, X_n$ je náh. výběr z p-rozměrného normálního rozdělení se střední hodnotou $\theta \in \mathbb{R}^p$.

James, Stein (1961):

- $p = 1, 2$, pak $\hat{\theta} = \overline{X}$ je přípustný odhad parametru θ
- $p \geq 3$, pak $\hat{\theta} = \overline{X}$ není přípustný

Příklad (vícerozměrný)

$X_1, \dots, X_n$ je náh. výběr z p-rozměrného normálního rozdělení $N_p(\theta, \sigma^2 \cdot \mathbb{I})$, kde $p \geq 3$ a $\sigma^2 > 0$ je (pro jednoduchost) známé.

$X_i = \theta + e_i$, kde $e_i \sim N_p(\theta, \sigma^2 \cdot \mathbb{I})$

Odhadujeme parametr $\theta = (\theta_1, \dots, \theta_p)^T$.

$\hat{\theta} = \overline{X} = (\overline{X^1}, \dots, \overline{X^p})^T$ je výběrový průměr

$$R(\hat{\theta}, \theta) = \mathbb{E}\|\hat{\theta} - \theta\|^2 = \sum_{i=1}^{p} E(\hat{\theta}_i - \theta_i)^2 = \sum_{i=1}^{p} E(\overline{X^i} - \theta_i)^2 = \sum_{i=1}^{p} D(\overline{X^i}) = p \cdot \frac{\sigma^2}{n}.$$

James a Stein : $\hat{\Theta}_{JS} = \left(1 - \dfrac{(p-2)\frac{\sigma^2}{n}}{\|\bar{X}\|^2}\right)\cdot \bar{X}$.

ale ani tento odhad není přípustný

$$R(\hat{\Theta}_{JS}, \Theta) = p\cdot\frac{\sigma^2}{n} - \underbrace{(p-2)\cdot E\left(\frac{1}{\|\bar{X}\|^2}\right)}_{>0} < p\cdot\frac{\sigma^2}{n}$$

positive part Jamesův - Steinův odhad: $\hat{\Theta}_{JS}^{+} = \left(1 - \dfrac{(p-2)\cdot\frac{\sigma^2}{n}}{\|\bar{X}\|^2}\right)_{+}\cdot \bar{X}$.

Poznámka

↑ shrinkage odhady: snížení rozptylu odhadu na úkor vychýlenosti.

Jak definovat medián pro p-rozměrná data?

a) marginální medián - složky jsou příslušné jednorozměrné mediány

b) geometrický (spatial, L_1, ...) medián

data: $x_1, \dots, x_n$ (p-rozměrná)

$\hat{\Theta} = \operatorname{argmin}_{y\in\mathbb{R}^p} \sum_{i=1}^{n} \|x_i - y\|$, kde $\|.\|$ je eukleidovská norma v $\mathbb{R}^p$

bod selhání $\varepsilon^* = 0{,}5$.

$p=1$ ... $\hat{\Theta} = \operatorname{argmin}_{y\in\mathbb{R}} \sum_{i=1}^{n} |x_i - y|$ ... klasický medián

c) medián založený na koncepci hloubky dat

Hloubková funkce: hloubka bodu $x \in \mathbb{R}^p$ vzhledem k datovému souboru $\mathbb{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$ je funkce $d(x, \mathbb{X})$ splňující následující vlastnosti:

(i) hloubka je afinně invariantní

(ii) nulová v nekonečnu $d(x, \mathbb{X}) \xrightarrow[\|x\| \to \infty]{} 0$.

(iii) je maximální v centru symetrie

(iv) je monotónní od nejhlubšího bodu

Tukeyho hloubka dat $\mathrm{depth}(x, \mathbb{X}) = \frac{1}{n} \cdot \min_{\substack{\|u\|=1 \\ u \in \mathbb{R}^p}} \{ i : u^T X_i \geq u^T x \} = \frac{1}{n} \{$ nejmenší počet pozorování X_i obsažený v libovolném uzavřeném poloprostoru v $\mathbb{R}^p$ s hranicí procházející bodem $x \}$

Tukeyho (hloubkový) medián: $\hat{\theta} = \operatorname*{argmax}_{x \in \mathbb{R}^p} \mathrm{depth}(x, \mathbb{X})$.

Poznámka

Je-li $\mathbb{X}$ v obecné pozici, pak $\frac{1}{p+1} \leq \mathrm{depth}(\hat{\theta}, \mathbb{X}) \leq \frac{1}{2}$.

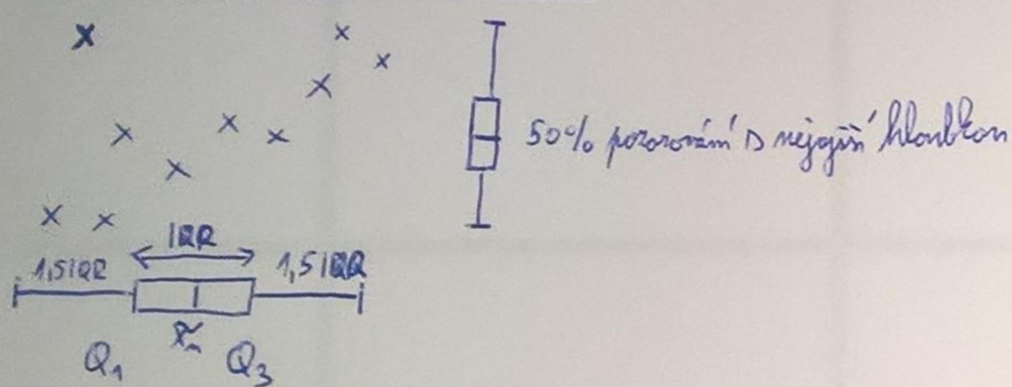
$\mathrm{depth}(\hat{\theta}, \mathbb{X}) = \frac{1}{2} \iff \mathbb{X}$ pochází z úhlově symetrického rozdělení $\frac{X - \hat{\theta}}{\|X - \hat{\theta}\|} \stackrel{d}{=} \frac{\hat{\theta} - X}{\|X - \hat{\theta}\|}$.

$\varepsilon^* \geq \frac{1}{p+1}$

α-useknutý průměr $\hat{\Theta}_\alpha = \sum\limits_{i:\, depth(x_i, \mathbb{X}) \ge \alpha} X_i \cdot \frac{1}{|i:\, depth(x_i, \mathbb{X})|}$.

$\varepsilon^* = \alpha$

Odlehlá pozorování ve více dimenzích



Bagplot - dvourozměrný boxplot (Rousseeuw, 1999)

1) Najdeme oblast s největší hloubkou a definujeme hloubkový (Tukeyho) medián T^* jako těžiště této oblasti.

2) Najdeme bag = oblast obsahující 50% pozorování s největší hloubkou

definujeme úrovňové množiny $D_k = \{ x : depth(x, \mathbb{X}) \ge \frac{k}{n} \}$ pro $k = 1, 2, \dots$

$D_1 \supset D_2 \supset D_3 \supset \dots$ $\qquad |D_1| \ge |D_2| \ge |D_3| \ge \dots$

označme $|D_k|$ = počet pozorování x_i obsažených v D_k

Najdeme k takové, že $|D_k| \le \left\lfloor \frac{n}{2} \right\rfloor < |D_{k-1}|$

a sestrojíme bag B: konvexní množina mezi D_k a D_{k-1} (lineární interpolace relativně vzhledem k T^*)

3) Určíme fence tak, že bag 3× zvětšíme vzhledem k T^* (0,5% pozorování z dvourozměrného normálního rozdělení leží mimo fence)

4) Sestrojíme loop – konvexní obal všech bodů mezi fence a bag.

5.) Ostatní pozorování jsou označena jako odlehlá (outliers).