Matematické nástroje robustnosti

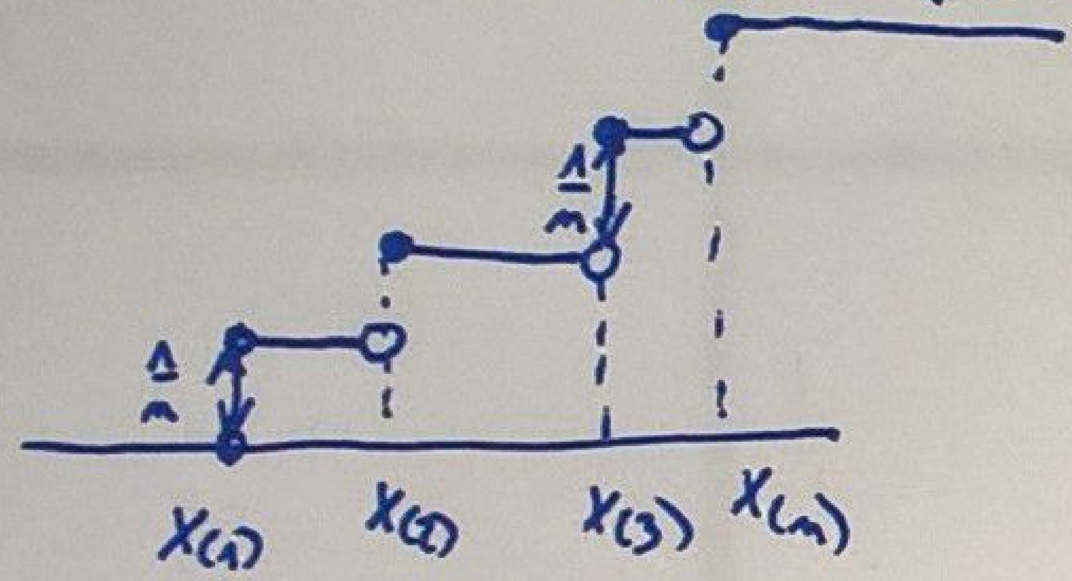
model: $X_1,\dots,X_n$ je náhodný výběr, kde n.v. X_i má rozdělení pravděpodobnosti $P = P_\theta$, $\theta \in \mathbb{R}$ je neznámý parametr.

Ten budeme chápat jako funkcionál P, tj. $\theta = T(P)$.

Jeho přirozeným (empirickým) odhadem je funkcionál $T(P_n)$, kde P_n je empirické rozdělení pravděpodobnosti náh. veličin $X_1,\dots,X_n$.

P_n je diskrétní rovnoměrné rozdělení na množině $\{X_1,\dots,X_n\}$, tj. $P_n(\{X_i\}) = \frac{1}{n}$ a příslušná distr. fce je empirická distribuční funkce.

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \le x\}$$



Příklady

1.) Střední hodnota $EX_1 = T(P) = \int_{\mathbb{R}} x_1 \, dP(x_1)$

odhad $T(P_n) = \int_{\mathbb{R}} x_1 \, dP_n(x_1) = \frac{1}{n}x_{(1)} + \frac{1}{n}x_{(2)} + \dots + \frac{1}{n}x_{(n)} = \overline{X}$

výběrový průměr

2.) Rozptyl $DX_1 = T(P) = \int_{\mathbb{R}} (x_1 - EX_1)^2 \, dP(x_1)$

jeho odhad je $T(P_n) = \int_{\mathbb{R}} (x_1 - EX_1)^2 \, dP_n(x_1) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$

max. věrohodný odhad rozptylu

Poznámka

Jeden parametr může být vyjádřen více funkcionály, např. pro symetrické, unimodální rozdělení je střední hodnota stejná jako medián a modus.

Definice

Nechť $\mathcal{P}$ je množina pravd. rozdělení na $(\mathbb{R}, \mathcal{B})$. Nechť T je statistický funkcionál na $\mathcal{P}$. Dále nechť $P, Q \in \mathcal{P}$ a $0 \le t \le 1$.

Rozdělení pravděpodobnosti $P_t(Q) = (1-t)P + t\cdot Q$ nazýváme kontaminací P rozdělením Q v poměru t.

Pozn.

$t=0$ ... žádná kontaminace

$t=1$ ... úplná kontaminace

Definice

Řekneme, že funkcionál T je diferencovatelný (v Gâteauxově smyslu) podle P ve směru Q, jestliže existuje limita:

$$T'_Q(P) = \lim_{t\to 0_+} \frac{T((1-t)P + tQ) - T(P)}{t}$$ ... Gâteauxova derivace T podle P ve směru Q

Poznámka

$\varphi(t) = T((1-t)\cdot P + t\cdot Q)$ pro $0 \le t \le 1$

Taylorův rozvoj φ v bodě 0:

$$\underbrace{\varphi(1)}_{T(Q)} = \underbrace{\varphi(0)}_{T(P)} + (1-0)\cdot \underbrace{\varphi'(0_+)}_{T'_Q(P)} + \underbrace{\frac{(1-0)^2}{2}\cdot \varphi''(u)}_{R\ \dots\ \text{chybový člen}}, \text{ kde } u\in(0,1).$$

$$T(Q) = T(P) + T'_Q(P) + R$$

Zvolme nyní $Q = P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$, kde δ_{x_i} je Diracova míra v x_i, tj. $\delta_{x_i}(x_i) = 1$, $\delta_{x_i}(x) = 0,\ x \neq x_i$

$$T(P_n) - T(P) = T'_{P_n}(P) + o\left(\frac{1}{\sqrt{n}}\right) = \frac{1}{n}\sum_{i=1}^{n} T'_{\delta_{x_i}}(P) + o\left(\frac{1}{\sqrt{n}}\right) = \frac{1}{n}\sum_{i=1}^{n} \overbrace{T'_{x_i}(P)}^{\text{vliv } x_i \text{ na chybu odhadu}} + \underbrace{o_P\left(\frac{1}{\sqrt{n}}\right)}_{\text{zanedbatelně malé}}$$

... chyba odhadu $T(P)$, když ho odhadujeme pomocí $T(P_n)$

Definice

Influenční funkce funkcionálu T v bodě P nazveme derivaci T podle P ve směru δ_x, $x \in \mathbb{R}$, tj.

$$IF(x, T, P) = T'_x(P) = \lim_{t \to 0+} \frac{T((1-t)P + t \cdot \delta_x) - T(P)}{t}.$$

Poznámka

IF popisuje efekt kontaminace našeho rozdělení jedním bodem x na odhad, který hledáme.

Má-li být odhad robustní, IF by měla být omezená.

Příklad

1) výběrový průměr jakožto odhad střední hodnoty

$$T(P) = EX_1 = \int_{\mathbb{R}} x_1 \, dP(x_1)$$

$$IF(x, T, P) = \lim_{t \to 0+} \frac{\int_{\mathbb{R}} x_1 \, d((1-t)P + t \cdot \delta_x) - \int_{\mathbb{R}} x_1 \, dP}{t} = \lim_{t \to 0+} \frac{(\cancel{1}-t)\int_{\mathbb{R}} x_1 \, dP + t \cdot \int_{\mathbb{R}} x_1 \, d\delta_x - \cancel{\int_{\mathbb{R}} x_1 \, dP}}{t} = \lim_{t \to 0+} -\int_{\mathbb{R}} x_1 \, dP + \int_{\mathbb{R}} x_1 \, d\delta_x = x - EX_1.$$

2) max. věrohodný odhad rozptylu jakožto odhad rozptylu

$$T(P) = DX_1 = \int_{\mathbb{R}} (x_1 - EX_1)^2 \, dP$$

$$IF(x, T, P) = \lim_{t \to 0+} \frac{\int_{\mathbb{R}} (x_1 - EX_1)^2 \, d((1-t)P + t\delta_x) - \int_{\mathbb{R}} (x_1 - EX_1)^2 \, dP}{t} = \lim_{t \to 0+} \frac{(\cancel{1}-t)\int_{\mathbb{R}} (x_1 - EX_1)^2 \, dP + t \int_{\mathbb{R}} (x_1 - EX_1)^2 \, d\delta_x - \cancel{\int_{\mathbb{R}} (x_1 - EX_1)^2 \, dP}}{t} = -\int_{\mathbb{R}} (x_1 - EX_1)^2 \, dP + (x - EX_1)^2 =$$

$$(x - EX_1)^2 - DX_1.$$

Kvantitativní charakteristiky robustnosti

a) globální citlivost funkcionálu T pro rozdělení pravděpodobnosti P: $\gamma^* = \sup_{x\in\mathbb{R}} |IF(x,T,P)|$

b) lokální citlivost -||- : $\lambda^* = \sup_{\substack{x,y\in\mathbb{R}\\ x\neq y}} \left| \dfrac{IF(y,T,P)-IF(x,T,P)}{y-x} \right|$

Příklady

1) Výběrový průměr (střední hodnota)

$IF(x,T,P) = x - EX_1$

$$\gamma^* = \sup_{x\in\mathbb{R}} |IF(x,T,P)| = \sup_{x\in\mathbb{R}} |x-EX_1| = \underline{\infty}$$

$$\lambda^* = \sup_{\substack{x,y\in\mathbb{R}\\ x\neq y}} \left| \frac{y-EX_1-(x-EX_1)}{y-x} \right| = \sup_{\substack{x,y\in\mathbb{R}\\ x\neq y}} 1 = \underline{1}.$$

2) Max. věr. odhad rozptylu (rozptyl)

$IF(x,T,P) = (x-EX_1)^2 - DX_1$

$$\gamma^* = \sup_{x\in\mathbb{R}} |(x-EX_1)^2 - DX_1| = \infty.$$

$$\lambda^* = \sup_{\substack{x,y\in\mathbb{R}\\ x\neq y}} \left| \frac{\overset{y^2-2yEX_1+(EX_1)^2}{(y-EX_1)^2} - DX_1 - \{\overset{x^2-2xEX_1+(EX_1)^2}{(x-EX_1)^2} - DX_1\}}{y-x} \right| = \sup_{\substack{x,y\in\mathbb{R}\\ x\neq y}} \left| \frac{y^2-x^2-2yEX_1+2xEX_1}{y-x} \right| = \sup_{\substack{x,y\in\mathbb{R}\\ x\neq y}} \left| \frac{(y-x)(y+x)-2EX_1(y-x)}{y-x} \right| = \sup_{\substack{x,y\in\mathbb{R}\\ x\neq y}} |y+x-2EX_1| = \infty.$$

c) bod selhání

označme $\mathbf{x}^0 = (x_1^0, \dots, x_n^0)$ realizace počátečního náh. výběru

$T_n(\mathbf{x}^0)$ … příslušný odhad funkcionálu T

v $\mathbf{x}^0$ nahradíme $\underline{m}$ jeho složek co nejnepříznivějšími hodnotami (příp. $\pm\infty$), označme jej $\mathbf{x}^{(m)}$ a $T_n(\mathbf{x}^{(m)})$ příslušný odhad

Bod selhání odhadu T_n ve výběru $\mathbf{x}^0$ definujeme jako $\varepsilon_n^*(T_n, \mathbf{x}^0) = \dfrac{m^*(\mathbf{x}^0)}{n}$, kde $m^*(\mathbf{x}^0)$ je nejmenší celé číslo, pro které $\sup\limits_{\mathbf{x}^{(m)}} \left| T_n(\mathbf{x}^{(m)}) - T_n(\mathbf{x}^0) \right| = \infty$

Pokud $\varepsilon_n^*(T_n, \mathbf{x}^0)$ nezávisí na $\mathbf{x}^0$, můžeme definovat limitu: $\varepsilon^*(T_n) = \lim\limits_{n\to\infty} \varepsilon_n^*(T_n, \mathbf{x}^0)$. … nejmenší podíl pozorování, které po nahrazení libovolnými hodnotami převedou T_n k nekonečným hodnotám

<u>Příklady</u>

1.) Výběrový průměr: $T_n = \overline{X}$

$$\varepsilon_n^*(T_n, \mathbf{x}^0) = \frac{1}{n}, \quad \varepsilon^*(T_n) = \lim_{n\to\infty} \frac{1}{n} = 0.$$

2.) Výběrový medián: $T_n = \tilde{X} = X_{\left(\frac{n+1}{2}\right)}$

$$\left.\begin{array}{ll} \varepsilon_n^*(T_n, \mathbf{x}^0) = \dfrac{n+1}{2n} & \text{pro } n \text{ liché} \\[2em] \varepsilon_n^*(T_n, \mathbf{x}^0) = \dfrac{n}{2n} & \text{pro } n \text{ sudé} \end{array}\right\} \varepsilon^*(T_n) = \lim_{n\to\infty} \varepsilon^*(T_n, \mathbf{x}^0) = \frac{1}{2}.$$