BOD SELHÁNÍ NĚKTERÝCH ODHADŮ PARAMETRU POLOHY

Jan Hanousek, MFF UK Praha

1. ÚVOD

Většina výsledků je formulována pro Lévyho okolí distribuční funkce F_o

$$P_\varepsilon(F_o) = \left[F \mid (\forall t)\ F_o(t-\varepsilon) - \varepsilon \leq F(t) < F_o(t+\varepsilon) + \varepsilon\right].$$

Často se toto okolí vyjadřuje pomocí Lévyho metriky d_ε

$$d_\varepsilon(F;G) = \inf\left[\varepsilon \mid (\forall t) F(t-\varepsilon) - \varepsilon \leq G(t) < F(t+\varepsilon) + \varepsilon\right].$$

Potom

$$P_\varepsilon(F_o) = \left[F \mid d_\varepsilon(F;F_o) \leq \varepsilon\right].$$

Snadno odvodíme, že stochasticky největším prvkem okolí $P_\varepsilon(F_o)$ je funkce F_1 tvaru

$$F_1(x) = \left[F_o(x-\varepsilon) - \varepsilon\right]^+ = \begin{cases} 0 & x \leq x_o + \varepsilon \\ F_o(x-\varepsilon) - \varepsilon & x > x_o + \varepsilon, \end{cases}$$

kde $x_o = F_o^{-1}(\varepsilon)$. Analogicky najdeme stochasticky nejmenší funkci F_2

$$F_2(x) = \min\left[F_o(x+\varepsilon) + \varepsilon;\ 1\right] = \begin{cases} F_o(x+\varepsilon) + \varepsilon & x \leq x_o - \varepsilon \\ 1 & x > x_o - \varepsilon, \end{cases}$$

zde $x_o = F_o^{-1}(1-\varepsilon)$.

(Připomeňme si, že F je stochasticky větší než G, pokud pro všechna x platí, že $F(x) < G(x)$).
Vzhledem k častému používání uveďme ještě tvar F_1^{-1} a F_2^{-1};
Zde jako obvykle

$$F^{-1}(y) = \sup\left[x: F(x) \leq y\right].$$

Tedy

$$F_1^{-1}(x) = \begin{cases} \varepsilon + F_o^{-1} & 0 \leq x \leq 1-\varepsilon \\ +\infty & 1-\varepsilon < x \leq 1 \end{cases}$$

a

$$F_2^{-1}(x) = \begin{cases} -\varepsilon + F_o^{-1}(x-\varepsilon) & \varepsilon \leq x \leq 1 \\ -\infty & 0 \leq x < \varepsilon. \end{cases}$$

2. DEFINICE A PŘÍKLADY

2.1. Definice bodu selhání

Bodem selhání odhadu T pro rozdělení F_o nazveme

$$\varepsilon^* = \varepsilon^*(F_o,T) = \sup\left[\varepsilon \mid b(\varepsilon) < b(1)\right],$$

kde

$$b(\varepsilon) = \lim_{n\to\infty} \sup_{F\in P_\varepsilon} \left|M(F,T_n)\right|.$$

$M(F,T_n)$ je medián rozdělení $L_F[T_n - T(F_0)]$.
Tuto obecnou definici budeme modifikovat na naše případy.
Zavedeme veličiny

$$b^+(\varepsilon) = \sup\left[T(F) \mid d_g(F,F_0) \le \varepsilon \right],$$
$$b^-(\varepsilon) = \inf\left[T(F) \mid d_g(F,F) \le \varepsilon \right]$$
$$b_1(\varepsilon) = \max\left[b^+(\varepsilon) ; - b^-(\varepsilon) \right].$$

a využijeme vět 1.4.1 a 1.4.2 Huber (1981), ze kterých plyne, že $b_1(\varepsilon) = b(\varepsilon)$ v bodech spojistosti b_1.

Další úvahy budou patřit výběrovému bodu selhání.
Nechť $X = (x_1,\dots,x_n)$ je libovolný soubor o pevném rozsahu n. Soubor můžeme znehodnotit několika způsoby:

a) $\underline{\varepsilon}$ - znečištění
Připojíme m libovolných hodnot $Y = (y_1,\dots,y_m)$ k souboru. Znehodnocený soubor $X' = X \cup Y$ je třídy $n+m$ a relativní obsah špatných hodnot je roven $\varepsilon = \frac{m}{n+m}$.

b) $\underline{\varepsilon}$ - vrácení
Do souboru znovu zařadíme libovolnou podmnožinu m prvků základního souboru s hodnotami $y_1,\dots,y_m$. Znehodnocený soubor je třídy n a relativní obsah špatných hodnot je $\varepsilon = \frac{m}{n}$.

c) $\underline{\varepsilon}$ - modifikace
Nechť $\rho(\cdot)$ je libovolná metrika definovaná na prostoru empirických měr. Nechť F_n je empirická míra odpovídající danému souboru X'. Nechť X je nějaký jiný soubor s empirickou mírou G_n takovou, že $\rho(F_n,G_n) \le \varepsilon$.

V našem případě je nejvhodnější užívat případ a). Nyní definujme výběrový bod selhání odhadu T.
Nechť je dán soubor $X = (x_1,\dots,x_n)$ a k němu příslušný ε- znečištěný soubor X'.
Maximální spád odhadu T při ε-znečištění je

$$b(\varepsilon, X, T) = \sup \left| T(X') - T(X) \right| ,$$

kde supremum se bere přes množinu všech ε-znečištěných souborů X'. Potom výběrový bod selhání odhadu T definujeme jako

$$\varepsilon^*(X,T) = \inf\left[\varepsilon \mid b(\varepsilon, X, T) = +\infty \right].$$

Bod selhání je tedy, nepřesně řečeno, nejmenší ε-znečištění, na kterém může odhad nabýt libovolně velké hodnoty.

2.2. JEDNOROZMĚRNÉ M-ODHADY

VĚTA 2.2

Nechť $\psi(\cdot)$ je neklesající, ale ne nutně spojitá omezená funkce nabývající kladných i záporných hodnot. Uvažujme M-odhad polohy $T(F)$ definovaný rovností

$$\int \psi(x - T(F))\, dF(x) = 0.$$

Bod selhání odhadu T se definuje vztahy

$$\varepsilon^* = \varepsilon^*(T,\psi) = \frac{\eta}{1+\eta}, \quad \text{kde}$$

$$\eta = \min\left[- \frac{\psi(-\infty)}{\psi(+\infty)} ; \frac{-\psi(+\infty)}{\psi(-\infty)} \right]$$

a dosahuje své maximální hodnoty $\varepsilon^* = 1/2$ za podmínky $\psi(-\infty) = -\psi(+\infty)$.

Důkaz

Označme $\lambda(t,F) = \int \psi(x-t)\, dF(x)$. Všimněme si, že $\lambda(.,.)$ je nerostoucí funkcí v parametru t; v F je tato funkce neklesající, jestliže F je stochasticky větší.

Platí tedy nerovnost

$$\lambda(t,F) \le \lambda(t,F_1) = \int_R \psi(x-t)\, dF_1(x) = \int_{x_0}^{\infty} (x-t+\varepsilon)\, dF_0(x) + \varepsilon \cdot \psi(\infty),$$

$$x_0 = F_0^{-1}(\varepsilon).$$

Z vlastností funkce λ a F_1 plyne, že

$$b^+(\varepsilon) = \inf \left[t \mid \lambda(t,F_1) < 0 \right].$$

Je vidět, že

$$b^+(\varepsilon) < \infty \Rightarrow \lim_{t\to\infty} \lambda(t,F_1) < 0 .$$

$$\lim_{t\to+\infty} \lambda(t,F_1) = (1-\varepsilon)\times\psi(-\infty) + \varepsilon\,\psi(+\infty)$$

Odhad neselže, pokud

$$(1-\varepsilon)\times\psi(-\infty) + \varepsilon\times\psi(+\infty) < 0$$

tedy

$$\frac{\varepsilon}{1-\varepsilon} < -\frac{\psi(-\infty)}{\psi(+\infty)} .$$

Obdobně, bereme-li funkci F_2-stochasticky nejmenší, je

$$b^-(\varepsilon) = \sup \left[t \mid \lambda(t,F_2) > 0 \right] \quad \text{a}$$

$$b^-(\varepsilon) < -\infty \Rightarrow \lim_{t\to-\infty} \lambda(t,F_2) < 0.$$

Výpočtem zjistíme, že

$$\lambda(t,F_2) = \int_{-\infty}^{x_0} \psi(x-t-\varepsilon)\, dF_0(x) + \varepsilon\times\psi(-\infty)$$

a

$$\lim_{t\to-\infty} \lambda(t,F_2) = (1-\varepsilon)\,\psi(+\infty) + \varepsilon\times\psi(-\infty).$$

Z podmínky

$$(1-\varepsilon)\,\psi(+\infty) + \varepsilon\times\psi(-\infty) < 0$$

vyjde

$$\frac{\varepsilon}{1-\varepsilon} < -\frac{\psi(+\infty)}{\psi(-\infty)} .$$

Tím už snadno dojdeme k tvrzení věty 2.2.

2.3. L-ODHADY

VĚTA 2.3

Nechť $M = M^+ - M^-$ je konečná znaménková míra na $(0,1)$.
(Z Jordanova rozkladu dostaneme, že M^+ a M^- jsou konečné, kladné míry na $(0,1)$) a $T(F) = \int F^{-1}(s)\, dM(s)$. Nechť d je největší reálné číslo, při kterém interval $(d, 1-d)$ obsahuje nosič měr M^+ a M^-. Dále nechť v bodech nespojitosti F_0^{-1} má M nulovou míru. Potom bod selhání $\varepsilon^* \geq d$. Jestliže je M kladná míra, potom $\varepsilon^* = d$.

Důkaz

Vzhledem k rozkladu $M = M^+ - M^-$ dostaneme $T = T^+ - T^-$, kde

$$T^+(F) = \int F^{-1}(s)\, dM^+(s)$$

a

$$T^-(F) = \int F^{-1}(s)\, dM^-(s).$$

Jestliže body selhání T^+ i T^- jsou větší nebo rovny α, pak bod selhání T je zřejmě také větší nebo roven α. Hledejme zatím podmínky, které platí pro bod selhání odhadu $T = \int F^{-1}(s)\, dM(s)$, kde M je konečná, kladná míra na $(0,1)$. Nejprve předpokládejme, že $0 < \varepsilon < \alpha$. Znovu využijeme veličin

$$b^+(\varepsilon) = \sup \left[T(F) \mid d_g(F_o,F) \le \varepsilon \right]$$

a

$$b^-(\varepsilon) = \inf \left[T(F) \mid d_g(F_o,F) \le \varepsilon \right].$$

Z vlastností funkcí F_1 a F_2 snadno spočteme, že

$$b^+(\varepsilon) = \int F_1^{-1}(s)\, dM(s) = \varepsilon + \int_{\alpha}^{1-\alpha} F_o^{-1}(s+\varepsilon)\, dM(s)$$

a

$$b^-(\varepsilon) = \int F_2^{-1}(s)\, dM(s) = -\varepsilon + \int_{\alpha}^{1-\alpha} F_o^{-1}(s-\varepsilon)\, dM(s).$$

Víme, že

$$b_1(\varepsilon) = \max \left[b^+(\varepsilon); -b^-(\varepsilon) \right] \le b^+(\varepsilon) - b^-(\varepsilon) \downarrow 0$$

pro $\varepsilon \to 0$ neboť body nespojitosti F_o^{-1} mají při M nulovou míru (Funkcionál T je spojitý). Tedy veličina $b_1(\varepsilon)$ je konečná při $\varepsilon < \alpha$. Odtud plyne, že $\varepsilon^* \ge \alpha$. Pokud M je konečná kladná míra, je vidět, že bod selhání $\varepsilon^* \le \alpha$. Tím je věta 2.3 dokázána.

2.4. R-ODHADY

VĚTA 2.4

Nechť funkce J, sloužící k výpočtům vah u R- odhadů, je neklesající, integrovatelná a nechť J splňuje podmínku

$$J(1-t) = -J(t), \qquad 0 < t < 1.$$

R- odhad $T(F)$ definujme rovností

$$\int J \left[1/2 \left(s + 1 - F\left(2 T(F) - F^{-1}(s)\right)\right)\right] ds = 0.$$

Potom bod selhání $\varepsilon^*(T,F_o)$ je řešením rovnice

$$\int_{1/2}^{1-\varepsilon^*/2} J(s)\, ds = \int_{1-\varepsilon^*/2}^{1} J(s)\, ds.$$

Důkaz

Platí, že funkce

$$\lambda(t,F) = \int J \left[1/2 \left(s+1 - F\left(2t - F^{-1}(s)\right)\right)\right] ds$$

je nerostoucí v t a neklesající v F, pokud F je stochasticky větší. Znovu využijeme vlastností funkcí F_1 a F_2, respektive F_1^{-1} a F_2^{-1}. Odtud plyne, že

$$\lambda(t;F_2) \le \lambda(t;F) \le \lambda(t;F_1).$$

Spočteme tedy $\lambda(t;F_1)$ a $\lambda(t;F_2)$. Jestliže jsou splněny podmínky

$$0 \le s \le 1-\varepsilon \qquad \text{a} \qquad 2t - F_1^{-1}(s) \ge x_0 + \varepsilon$$

čili
$$0 \le s \le 1-\varepsilon \qquad \text{a} \qquad s \le F_0(2(t-\varepsilon) - x_0) - \varepsilon ,$$

dostaneme, že

$$F_1\left[2t - F_1^{-1}(s)\right] = F_0\left[2(t-\varepsilon) - F_0^{-1}(s+\varepsilon)\right] - \varepsilon.$$

Označíme-li bod, ve kterém nastává zlom

$$s_0 = \left[F_0(2(t-\varepsilon) - x_0) - \varepsilon\right]^+,$$

můžeme vyjádřit

$$\lambda(t;F_1) = \int_0^{s_0} J\left[1/2\left(s + \varepsilon + 1 - F_0\left(2(t-\varepsilon) - F_0^{-1}(s+\varepsilon)\right)\right)\right] ds + \int_{s_0}^{1} J\left(1/2(s+1)\right) ds.$$

Snadno se ukáže, že

$$b^+(\varepsilon) = \inf\left[t \mid \lambda(t;F_1) < 0\right]$$

a že

$$b^+(\varepsilon) < \infty \Leftrightarrow \lim_{t\to+\infty} \lambda(t;F_1) < 0.$$

Limitním přechodem zjistíme

$$\lim_{t\to\infty} \lambda(t;F_1) = \int_0^{1-\varepsilon} J\left[1/2(s+\varepsilon)\right] ds + \int_{1-\varepsilon}^{1} J\left[1/2(s+1)\right] ds.$$

Jednoduchou substitucí a s využitím vlastností J máme

$$\lim_{t\to+\infty} \lambda(t;F_1) = 2\left(\int_{1-\varepsilon/2}^{1} J(s)\, ds - \int_{1/2}^{1-1/2\varepsilon} J(s)\, ds\right).$$

V tomto případě odhad selže, bude-li

$$\int_{1-\varepsilon/2}^{1} J(s)\, ds = \int_{1/2}^{1-\varepsilon/2} J(s)\, ds.$$

Podobně budeme postupovat pro stochasticky nejmenší funkci F_2. Vyjdou tyto výsledky

$$\lambda(t;F_2) = \int_{s_0}^{1} J\left[1/2\left(s + 1 - \varepsilon - F_0\left(2(t+\varepsilon) - F_0^{-1}(s-\varepsilon)\right)\right)\right] ds + \int_0^{s_0} J\left(\tfrac{s}{2}\right) ds$$

kde

$$s_0 = F_0\left(2(t+\varepsilon) - x_0\right) + \varepsilon \vee 1.$$

Dále spočteme

$$\lim_{t\to-\infty} \lambda(t;F_2) = \int_{\varepsilon}^{1} J\left(1/2[s+1-\varepsilon]\right) ds + \int_0^{\varepsilon} J\left(\tfrac{s}{2}\right) =$$

$$= 2\left(\int_{1/2}^{1-\varepsilon/2} J(s)\, ds - \int_{1-\varepsilon/2}^{1} J(s)\, ds\right).$$

Opět $b^{-}(\varepsilon)$ je konečné, pokud $\lim_{t\to-\infty} \lambda(t;F_2) > 0$.
Tedy i v tomto případě dostaneme pro bod selhání ε podmínku

$$\int_{1/2}^{1-\varepsilon/2} J(s)\, ds = \int_{1-\varepsilon/2}^{1} J(s)\, ds ,$$

což je tvrzení věty 2.4.

2.5. REDESCENDNÍ M-ODHADY S OMEZENOU FUNKCÍ

Předpokládejme, že $\rho(\cdot)$ má minimum v 0, $\rho(0) = -1$ a že ρ je neklesající na obou stranách $(k +\infty$ a $k -\infty)$ a $\lim_{x\to\infty} \rho(x) = 0$.

VĚTA 2.5

Položí-li

$$\sum_{x} \rho(x - T(X)) = -A,$$

pak (výběrový) bod selhání T v modelu ε-znečištění je roven

$$\varepsilon(X,T) = \frac{m'}{n + m'} ,$$

kde m' je přirozené číslo, vyhovující vztahu $\lceil A \rceil \le m' < \lfloor A \rfloor + 1$.
Když existují nějaké $c < \infty$ takové, že $\rho(x) = 0$ pro $x \ge c$, potom $m' = \lceil A \rceil$.

2.6. REDESCENDNÍ M-ODHADY S NEOMEZENOU FUNKCÍ

Předpokládejme, že ρ je sudá, $\rho(0) = 0$ a ρ je neklesající na $(0,\infty)$.
Dále je

$$\lim_{x\to\pm\infty} \rho(x) = +\infty$$

a

$$\lim_{x\to\infty} \frac{\rho(x)}{x} = 0.$$

Předpokládejme, že $\psi = \rho'$ je spojitá a že existují x_0 tak, že ψ je neklesající pro $0 < x < x_0$ a nerostoucí pro $x_0 < x < \infty$.

VĚTA 1.6

Za výše uvedených podmínek je bod selhání M-odhadu v modelu ε-znečištění roven $\varepsilon^{*} = 1/2$.

<u>Důkaz</u>

Viz Huber (1984).

2.7. P-ODHADY

Odhad Pitmanova typu, neboli p-odhad polohy se definuje

$$T_p = \frac{\int \exp\left[-\sum \rho(x_i - \theta)\right] \theta\, d\theta}{\int \exp\left[-\sum \rho(x_i - \theta)\right] d\theta} ,$$

pokud má výraz smysl.

VĚTA 2.7

Předpokládejme, že ρ je sudá a konvexní a $0 < \lim_{x\to\pm\infty} \rho(x) / x = c < \infty$.

Potom bod selhání P-odhadu je $\varepsilon^* = 1/2$.

Důkaz

Viz Huber (1984).

2.8. PŘÍKLADY

1. Hodges - Lehmannův odhad

H-L odhad $T = \operatorname{med}_{i>j} \left[(x_i + x_j)/2\right]$ můžeme dostat také jako R-odhad s funkcí $J(t) = t - 1/2$. Spočteme jeho bod selhání.

a) pomocí věty 2.4

Hledáme řešení rovnice

$$\int_{1/2}^{1-\delta/2} (t-1/2)\,dt = \int_{1-\delta/2}^{1} (t-1/2)\,dt.$$

Spočteme, že δ musí splňovat rovnici

$$2\delta^2 - 4\delta + 1 = 0$$

a z ní plyne $\underline{\varepsilon^* = 1 - \frac{\sqrt{2}}{2} \cong 0,239}$

b) z definice bodu selhání

Připojíme-li k výběru m hodnot, odhad selže, bude-li

$$\binom{n}{2} < \frac{1}{2}\binom{n+m}{2}$$

odtud vypočteme

$$m = \frac{1 - 2n + \sqrt{(8n^2 - 8n + 1)}}{2}$$

$$\varepsilon^* = \frac{m}{n+m} = \frac{1/2 - n + \sqrt{(2n^2 - 2n + 1/4)}}{1/2 + \sqrt{(2n^2 - 2n + 1/4)}} \to \frac{\sqrt{2} - 1}{\sqrt{2}} = \underline{1 - \frac{\sqrt{2}}{2}}\ .$$

2. Výběrový průměr

Odhad selže už připojením jediného špatného porovnání.

$$\varepsilon^* = \frac{1}{n+1}$$

3. g - uříznutý průměr

Tj. zanedbáme g největších a nejmenších pozorování. Snadno spočteme

$$\varepsilon^* = \frac{g+1}{n+g+1}$$

4. medián

Pomocí věty 2.2 - $\psi = \operatorname{sgn}(x)$ dostáváme, že $\varepsilon^* = 1/2$

5. α- uřiznutý průměr

Pomocí věty 2.3 lze snadno ukázat, že $\varepsilon^* = \alpha$.

6. R-odhad s normálními vahami

Tj $J(t) = \Phi^{-1}(t)$. Opět hledáme řešení rovnice

$$\int_{1/2}^{1-1/2\delta} \Phi^{-1}(s)\,ds = \int_{1-\varepsilon/2}^{1} \Phi^{-1}(s)\,ds.$$

Nejprve použijeme substituci $s = \Phi(x)$, potom $\frac{x^2}{2} = t$ a dojdeme k rovnici

$$\ln 2 - \frac{[\Phi^{-1}(1 - 1/2)]^2}{2} = 0.$$

Dopočteme a máme

$$\varepsilon^* = 2\,(1 - \Phi(\sqrt{\ln 4})) = 2\,\Phi(-\sqrt{\ln 4})$$

$$\varepsilon^* \doteq 0{,}239$$

POUŽITÁ LITERATURA

Donoho, D. L. and Huber, P. J. (1982). The notion of breakdown point. In: Festschrift in Honor of Erich Lehmann, Ed. by K. Doksum and J. L. Hodges, Wadsworth, Belmont, CA.

Huber, P. J. (1981). Robust Statistics. Wiley, New York.

Huber, P. J. (1984). Finite sample breakdown of M- and P-estimators. The Annals of Statistics, 1984, Vol. 12, No. 1, 119-126.