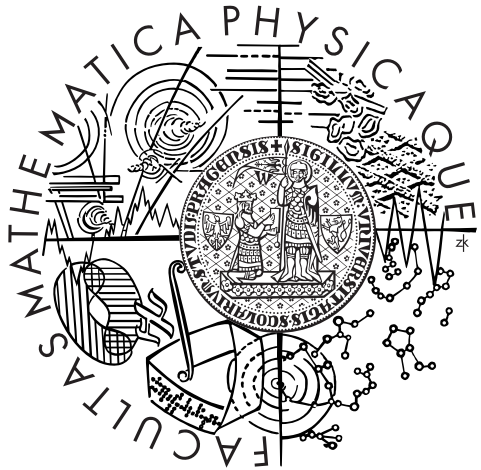
Asymptotika versus konečně mnoho pozorování

J. Jurečková



Obsah

Těžké chvosty robustních odhadů při konečném počtu pozorování

Hlavní argument pro nahrazování klasických odhadů robustními v 60. a 70. letech byl, že klasické odhady jsou citlivé k odlehlým pozorováním a k rozdělením s těžkými chvosty. Jako hlavní míra citlivosti odhadu k odlehlým pozorovaním byla zavedena influenční funkce. Podívejme se na to blíže.
Uvažujme náhodný výběr $X_1, \ldots, X_n$ z rozdělení s distribuční funkcí $F(x-\theta)$, obecně neznámou. Může se jednat o měření fyzikální charakteristiky θ, kterou chceme odhadnout. Pro jednoduchost předpokládejme, že F je symetrická kolem 0, tedy odhadujeme střed symetrie. Distribuční funkce F může mít těžké chvosty v tom smyslu, že

$$\lim_{x\to\infty} \frac{-\ln(1-F(x))}{m \ln x} = 1,\ m > 0;$$

a tedy podle von Misesovy podmínky splňuje

$$1 - F(x) = x^{-m} L(x), \quad m > 0, \tag{1}$$

kde $L(x)$ je funkce, pomalu se měnící v nekonečnu.
V tom případě použijeme raději některý robustní odhad T_n pro θ.
I robustní odhad má často asymptoticky normální rozdělení, tj.

$$P_\theta \left\{ \frac{T_n - \theta}{\sqrt{\operatorname{var} T_n}} \leq x \right\} \to \Phi(x), \; n \to \infty.$$

Limitní rozdělení má lehké chvosty, protože splňuje $\lim_{x\to\infty} \frac{-\ln(1-F(x))}{bx^2} = 1, \; b > 0$. Na druhou stranu, je-li n konečné, má T_n těžké chvosty, protože pro mnoho ekvivariantních odhadů platí

$$1 \leq \liminf_{x\to\infty} \frac{-\ln P_\theta\{|T_n - \theta| > x\}}{-\ln(1 - F(x))} \leq \limsup_{x\to\infty} \frac{-\ln P_\theta\{|T_n - \theta| > x\}}{-\ln(1 - F(x))} \leq n.$$

Rozdělení odhadu T_n je tedy stále těžké pro každé konečné n; sice exponent m odhadu T_n v (1) roste s rostoucím n, ale je stále konečný $\forall n$.

Poznámka: *Studentovo t_m rozdělení má hustotu*

$$f_m(x) = \frac{1}{\sqrt{m}} \frac{\left(1 + \frac{x^2}{m}\right)^{-(m+1)/2}}{B\left(\frac{m}{2}, \frac{1}{2}\right)}, \; x \in \mathbb{R}.$$

Pak

$$\lim_{x\to\infty} x^m(1 - F(x)) = \lim_{x\to\infty} L_m(x) = A_m = \frac{m^{\frac{m}{2}} - 1}{B\left(\frac{m}{2}, \frac{1}{2}\right)}$$

a $\lim_{m\to\infty} A_m = \infty$; *tedy* $\lim_{x\to\infty} x^m(1 - F_m(x/\sigma_m)) = 1$ *pro* $\sigma_m = A_m^{-1/m}$, *a* $\lim_{m\to\infty} \sigma_m = 0$. *Hodnoty* A_m *a* σ_m *pro některá m jsou v tabulce:*

m	1	2	3	4	5	6
A_m	.32	.50	1.10	3.00	9.50	16.90
σ_m	3.14	1.41	.97	.76	.64	.62

Nepřípustnost robustních odhadů při konečném počtu pozorování

Uvažujme chování ekvivariantních odhadů vzhledem k L_1 ztrátové funkci, $L_1(t,\theta) = |t-\theta|$ a k riziku odhadu $R(T_n,\theta) = E_\theta|T_n - \theta|$. Odhad T_n se nazývá *přípustný* vzhledem k L_1, jestliže neexistuje jiný odhad T_n^* takový, že $R(T_n^*,\theta) \leq R(T_n,\theta)$ pro vš. $\theta \in \mathbb{R}$, a $R(T_n^*,\theta_0) < R(T_n,\theta_0)$ pro nějaké θ_0. Odhad T_n, který usekne krajní pozorování, je nepřípustný. Platí následující věta:

VĚTA 1. *Nechť $X_1,\ldots,X_n, \quad n \geq 3$, jsou nezávislá pozorování s hustotou $f(x-\theta)$, kde $f(x) > 0$ je unimodální.*

(i) Nechť $T_n = T_n(X_1,\ldots,X_n)$ je ekvivariantní odhad, spojitý v každém argumentu, konstantní pro $X_{n:1}, X_{n:2}, X_{n:n-1}$ a $X_{n:n}$, ale jednoznačně určený jako funkce $X_{n:3},\ldots,X_{n:n-2}$, kde $X_{n:1} \leq X_{n:2} \leq \ldots \leq X_{n:n}$ jsou pořádkové statistiky. Pak T_n je nepřípustný jako odhad θ vzhledem k L_1 ztrátové funkci.

(ii) Nechť M_n je M-odhad generovaný spojitou neklesající funkcí ψ takovou, že $\psi(x) = \psi(c_1)$ pro $x \leq c_1$ a $\psi(x) = \psi(c_2)$ pro $x \geq c_2$, $c_1 < 0 < c_2$ ve tvaru

$$M_n = \frac{1}{2}(M_n^+ + M_n^-) \quad \text{kde}$$

$$M_n^- = \sup\{t : \sum_{i=1}^{n} \psi(X_i - t) > 0\}, \quad M_n^+ = \inf\{t : \sum_{i=1}^{n} \psi(X_i - t) < 0\}.$$

Pak M_n je nepřípustný jako odhad θ vzhledem k L_1 ztrátové funkci.
Důkaz: [3]
Např. medián a Huberův M-odhad nejsou přípustné při konečném n pro žádnou hustotu f, ačkoliv medián je maximálně věrohodným odhadem pro Laplaceovo rozdělení a Huberův odhad pro rozdělení normální uvnitř a exponenciální vně intervalu.

Nyní nás zajímá otázka, zdali zobecněný bayesovský odhad θ^* (včetně např Pitmanova odhadu) může být nezávislý na obou extrémech $X_{1:n}$ a $X_{n:n}$. Odpověď je za obecných podmínek záporná:

VĚTA 2.
Nechť $\theta^* = \theta^*(X_1, \ldots, X_n)$ *je zobecněný bayesovský odhad* θ *za apriorního rozdělení* Π, *který má spojitou zobecněnou hustotu* $f(x, \theta)$ *[nemusí být nutně pravděpodobnostní hustota] vzhledem k Lebesgueově míře. Nechť* $L(t - \theta) \geq 0$ *je ztrátová funkce taková, že* L *je spojitě diferencovatelná, ryze konvexní a integrovatelná se čtvercem vzhledem k mírám* $f(x, \theta)d\Pi(\theta)$, $x \in \mathbb{R}$. *Pak, je-li* $f(x, \theta)$ *kladná a spojitá v* x *a v* θ, $x \in \mathbb{R}$, $\theta \in \Theta$, *a lineární prostor nad funkcemi* $f(x, .)$ *je hustý v prostoru všech funkcí* θ, *integrovatelných vzhledem k míře* Π, *musí odhad* θ^* *být funkcí aspoň jednoho extrému* $X_{n:1}, X_{n:n}$.
Důkaz: [4]

DŮSLEDEK:
Nechť $X_1, \ldots, X_n$ jsou nezávislá pozorování se spojitou hustotou $f(x-\theta) > 0,\ \theta \in \mathbb{R}$, jejíž charakteristická funkce je všude různá od nuly. Pak Pitmanův odhad θ (ekvivariantní odhad s minimálním L_2 rizikem) je funkcí aspoň jednoho z extrémů $X_{1:n}$ a $X_{n:n}$.

Lineární regresní model: useknuté odhady a nepřípustnost

Uvažujme lineární regresní model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, kde $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, matice $\mathbf{X}$ je řádu $(n \times p)$ a plné hodnosti p, s řádky $\mathbf{x}_i^\top$, $x_{i1} = 1$, $i = 1, \ldots, n$, s neznámým parametrem $\beta \in \mathbb{R}^p$. Nezávislé chyby $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ mají hustotu f. Pak pozorování jsou v "základní poloze" *(general position)* s pravděpodobností 1, tj. libovolný vektor $(\mathbf{x}_{i_1}^\top, y_{i_1}, \ldots, \mathbf{x}_{i_p}^\top, y_{i_p})$, kde $1 \le i_1 < \ldots < i_p \le n$, vede k jedinému řešení $\mathbf{b}$ soustavy rovnic $y_{i_\nu} = \mathbf{x}_{i_\nu}^T \mathbf{b}$, $\nu = 1, \ldots, p$. α-regresní kvantil $0 < \alpha < 1$, je definován jako

$$\hat{\beta}(\alpha) = \operatorname{argmin}\{\sum_{i=1}^{n} \rho_\alpha(Y_i - \mathbf{x}_i^\top \mathbf{b}),\ \mathbf{b} \in \mathbb{R}^p\},$$

kde $\rho_\alpha(z) = |z|\{\alpha I[z > 0] + (1 - \alpha) I[z < 0]\}$, $z \in \mathbb{R}^1$.

Robustní L-odhady typicky useknou Y_i, pro které

$$\text{buď} \quad Y_i \leq \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\alpha_1) \quad \text{nebo} \quad Y_i \geq \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\alpha_2),\ i = 1, \ldots, n$$

pro zvolené $0 < \alpha_1 < \alpha_2 < 1$. Známý Koenkerův-Bassetův *useknutý odhad metodou nejmenších čtverců* je odhad metodou nejmenších čtverců vypočtený z neuseknutých pozorování.

Useknuté L-odhady v lineárním modelu jsou nepřípustné ve třídě ekvivariantních odhadů $\boldsymbol{\beta}$ vzhledem ke každé ryze konvexní, nezáporné a spojitě diferencovatelné ztrátové funkci. Z toho plyne, že i v modelu s parametrem posunutí jsou useknuté odhady nepřípustné i vzhledem k L_p normě, $p = 1, 2, \ldots$.

Hustota M-odhadu

Uvažujme výběr $X_1, \ldots, X_n$ z rozdělení s distribuční funkcí $F(x-\theta)$, kde F má absolutně spojitou hustotu f a konečnou Fisherovu informaci. Nechť T_n je M-odhad θ, který je řešením rovnice $\sum_{i=1}^{n} \psi(X_i - t) = 0$ s monotonní ψ, a $g_\theta(t)$ je jeho hustota, kterou chceme odhadnout. Známé jsou výsledky Hampela-Fielda-Ronchetiho, které velmi dobře aproximují hodnotu $g_\theta(t)$ v daném t numericky pro velmi malé výběry (small sample asymptotics).
Hustotu $g_\theta(t)$ lze (přesně) vyjádřit v následujícím tvaru, odkud můžeme odvodit další vlastnosti T_n :

$$g_\theta(t) = E_0\Big\{ \sum_{i=1}^{n} \frac{f'(X_i)}{f(X_i)}\, I\Big[\sum_{i=1}^{n} \psi(X_i - (t-\theta)) \le 0 \Big] \Big\}.$$

Za určitých podmínek lze pro T_n dokázat momentovou konvergenci (konvergenci momentů k momentům asymptoticky normálního rozdělení T_n.) Chceme-li srovnat asymptotické momenty s momenty při konečném n, můžeme využít následující věty. Zde $\gamma^{(\nu)}(\theta) = E_\theta(T_n)^\nu$ značí ν-tý moment T_n, a předpokládáme, že je konečný a diferencovatelný v θ; $\nu > 0$ může být i zlomek.

VĚTA 4.

Za uvedených podmínek platí pro vš. $\theta \in \mathbb{R}$

$$\dot{\gamma}^{(\nu)}(\theta) = E_0\Big[(T_n(\mathbf{X}) + \theta)^\nu \Big(- \sum_{i=1}^{n} \frac{f'(X_i)}{f(X_i)} \Big)\Big],$$

kde

$$\dot{\gamma}^{(\nu)}(\theta) = \frac{d\gamma^{(\nu)}(\theta)}{d\theta}.$$

Chvosty kontaminovaného rozdělení

Známý Tukeyho-Huberův model kontaminace předpokládá, že v průměru zlomek $(1-\varepsilon)$ dat pochází z normálního rozdělení, zatímco zbývající data mohou být poškozena abnormálním šumem; ε je rušivý parametr, obvykle se uvažuje $0 < \varepsilon < 0.25$. Robustní postupy chtějí provést inferenci o dominující části směsi, a filtrovat vliv kontaminující složky. Tento model vedl k základním pojmům robustnosti, jako influenční funkce, globální citlivost, maxbias a bod selhání. Dá se popsat jako

$$X = (1-B)Y + BZ,$$

kde X je pozorovatelná veličina a Y, Z a B jsou nepozorovatelné a nezávislé, $Y \sim F$ [hladké rozdělení s polohou a měřítkem, např. $N(\mu, \sigma^2)$], $Z \sim G$ (neznámé rozdělení odlehlých pozorování) a B je náhodný indikátor kontaminace, $P(B = 1) = \varepsilon$. Pak rozdělení pozorované veličiny X je směs $(1-\varepsilon)F + \varepsilon G$, a jeho chvosty jsou určeny těžším rozdělením z F, G, nehledě na velikost $\varepsilon > 0$.

To ilustruje následující věta:

VĚTA 3.
Nechť Y, Z jsou nezávislé náhodné veličiny se spojitými distribučními funkcemi F, G, symmetrickými kolem 0, s nedegenerovanými pravými chvosty. Nechť F nebo G má těžký pravý chvost, tj. např. $\lim_{x\to\infty} \frac{-\ln(1-F(x))}{m \ln x} = 1$, $m > 0$. *Pak pravý chvost součtu $X = Y + Z$ je stejně těžký jako těžší z F, G.*

Vliv počátečního odhadu na jednokrokovou iteraci M-odhadu

Nechť $\{X_i,\ i \geq 1\}$ jsou nezávislé náhodné veličiny s distribuční funkcí $F(x, \theta_0)$. Hledáme M-odhad θ_0, který je řešením minimalizace $\sum_{i=1}^{n} \rho(X_i, t) = \min,\ t \in \mathbf{\Theta}$, otevřený interval v $\mathbb{R}^1$, nebo řešením rovnice $\sum_{i=1}^{n} \psi(X_i, t) = 0$, kde $\psi(x, t) = \frac{\partial \rho(x,t)}{\partial t}$. Pak za určitých podmínek na F a ρ, že pro každou posloupnost $\{M_n\}$ kořenů rovnice, která splňuje $\sqrt{n}(M_n - \theta_0) = \mathcal{O}_p(1)$, platí reprezentace

$$M_n - \theta_0 - (n\gamma(\theta_0)))^{-1} \sum_{i=1}^{n} \psi(X_i, \theta_0) + R_n, \quad \text{kde} \quad R_n = \mathcal{O}_p(n^{-1}).$$

Zde $\gamma(\theta) = E\dot{\psi}(X, \theta),\ \dot{\psi}(X, \theta) = \frac{\partial \psi(x,\theta)}{\partial \theta}$. Jednokroková (Newton-Raphsonova) iterace odhadu M_n je definována jako

$$M_n^1 = M_n^0 - (n\hat{\gamma}_n)^{-1} \sum_{i=1}^{n} \psi(X_i, M_n^0) \quad \text{pokud} \quad \hat{\gamma}_n \neq 0,$$

$$M_n^1 = M_n^0 \quad \text{pokud} \quad \hat{\gamma}_n = 0,$$

kde $\hat{\gamma}_n = \frac{1}{n} \sum_{i=1}^{n} \dot{\psi}(X_i, M_n^0)$. Podobně definujeme 2.,3.,... iteraci. Pak

$n(M_n^1 - M_n) = \mathcal{O}_p(1)$ a $n(M_n^2 - M_n) = o_p(1)$, pokud $\sqrt{n}(M_n^0 - \theta_0) = \mathcal{O}_p(1)$.

Numerické výpočty však ukazují, že iterace není stejně dobrá pro různé volby M_n^0. Dá se dokázat, že nejlepší iterace (už ve smyslu asymptotiky 2. řádu) je pro takové M_n^0, které má stejnou influenční funkci jako M_n. Je tedy lépe udělat alespoň 2. iteraci M_n^2. V lineárním regresním modelu lze volit počáteční odhad, který má bod selhání 1/2 (i když má řád konsistence $< n^{-1/2}$), ale už 1. iterace je asymptoticky ekvivalentní neiterovanému odhadu. Bohužel tato iterace pomalu konverguje a dědí jiné slabé vlastnosti od počátečního odhadu.

Literatura

[1] (1985) P.Janssen, J. Jurečková and N.Veraverbeke: ”Rate of convergence of one- and two-step M-estimators with applications to maximum likelihood and Pitman estimators.” *Ann. Statist.* 13, 1222-1229.
[2] (1987) J. Jurečková and S.Portnoy: ”Asymptotics for one-step M-estimators in regression with application to combining efficiency and high breakdown point.” *Comm. Statist. A* 16, 2187-2199.
[3] (1997) J. Jurečková and L. Klebanov: ”Inadmissibility of robust estimators with respect to L_1 norm.” *L_1-Statistical Procedures and Related Topics* (Y. Dodge, ed.). *IMS Lecture Notes - Monographs Series* 31, 71-78.
[4] (1998) J. Jurečková and L. Klebanov: ”Trimmed, Bayesian and admissible estimators.” *Statist. and Probab. Letters* 42, 47–51.
[5] (1990) J. Jurečková and P.K.Sen: ”Effect of the initial estimator on the asymptotic behavior of one-step M-estimator.” *Ann. Inst. Statist. Math.* 42, 345-357.
[6] (2010) J. Jurečková and J. Picek: “Finite-sample behavior of robust estimators.” Submitted.