ROBUST'2002, 95 – 106

INFORMACE VE VÝBĚRU Z ROZDĚLENÍ

ZDENĚK FABIÁN

Abstract. Core functions of regular continuous probability distributions are re-introduced and their moments are studied. The first core moment appears to be a center of gravity of the distribution, the second moment has a direct relation to the information of the distribution. Their estimates are discussed.

Резюме. Вводяця коре функции непрерывных регулярных распределений правдоподобия и изучаются их моменты. Первая момента - ето центер тяжести распределения, вторая момента имеет прямое отношение к информации распределения. Изучаются их оценки.

1. Úvod

Ve sborníku Robust'96 byly zavedeny geometrické momenty. Jejich modifikaci uvedené ve sborníku Robust'2000 říkám core momenty. V tomto článku, v němž se na oba předchozí odkazuji jako na Robust'96 a Robust'00, se pokusím vysvětlit, proč si myslím, že by mohly být užitečné.

Obvykle se intuitivně očekává, že pozorovaný datový soubor by měla charakterizovat (nejméně) dvě čísla: jedno určující polohu či centrum dat (Wilcox (2001) tomu říká „measure of central tendency"), druhé variabilitu či rozptýlenost bodů souboru kolem centra („measure of dispersion").

Pozoruhodným rysem klasické statistiky je však zvláštní dvoukolejnost postupů vedoucích k odhadům charakteristik datových souborů.

Na jedné koleji se nalézají obecné a centrální momenty jakožto statistické funkcionály převzaté z teorie pravděpodobnosti. Jejich empirickými protějšky jsou výběrové momenty - budeme mluvit jen o výběrové střední hodnotě a výběrovém rozptylu - kterými lze popsat jakákoli data, která je možno považovat za náhodný výběr z rozdělení F. Popsat jednoduše a ve shodě s řečenou intuicí, ale bohužel velice často špatně: průměr ani rozptyl jednak nejsou dobrými charakteristikami nesymetrických datových souborů, jednak se někdy ani předem neví, zda příslušný teoretický protějšek existuje. Existovat nemusí, viz známý případ *symetrického* Cauchyova rozdělení, které nemá střední hodnotu. Podívejte se, prosím, na tabulku 1. Jsou v ní uvedeny hustoty f_θ a vzorce pro k-tý obecný moment

$$m_k(\theta) = E_\theta x^k = \int_S x^k dF_\theta(x) \qquad k = 1, 2... \tag{1}$$

tří parametrických rodin definovaných na $S = (0, \infty)$. V posledním sloupci tabulky je vyznačen obor hodnot k, pro které integrál (1) existuje (jedná se o rodiny s těžkými chvosty, jejich rodiči jsou rozdělení (1) extrémní hodnoty II, (2) log-logistické a (3) log-Cauchyovo).

2000 *Mathematics Subject Classification.* 62A01 62E10.
Klíčová slova. Core function, basic characteristic, point estimates.
Vzniklo za podpory grantu GA AVČR A1075101.

F_θ	f_θ	$m_k(\theta)$	platí pro
1	$\frac{\alpha^\alpha}{\Gamma(\alpha)}\frac{\beta}{x}\left(\frac{x}{\tau}\right)^{-\beta\alpha}e^{-\alpha\left(\frac{x}{\tau}\right)^{-\beta}}$	$\left(\tau\alpha^{1/\beta}\right)^k\frac{\Gamma(\alpha-k/\beta)}{\Gamma(\alpha)}$	$k<\beta\alpha$
2	$\frac{\beta}{x}\frac{\left(\frac{x}{\tau}\right)^{\beta\alpha}}{\left(\left(\frac{x}{\tau}\right)^{\beta}\right)^2+1/\nu)^{\alpha+\nu}}$	$\left(\frac{\tau}{\nu^{1/\beta}}\right)^k\frac{\Gamma(\alpha\nu+k/\beta)\Gamma(\alpha-k/\beta)}{\Gamma(\nu\alpha)\Gamma(\alpha)}$	$k<\beta\alpha$
3	$\frac{\beta}{B(\frac{1}{2},\alpha-\frac{1}{2})}\frac{1}{[1+\ln^2(\frac{x}{\tau})^{\beta}]^\alpha}$	$\frac{\Gamma(k+1/2)\Gamma(\alpha-k-1/2)}{\Gamma(1/2)\Gamma(k-1/2)}$	$k<\alpha-\frac{1}{2}$

Tab. 1 *Absolutní momenty některých rodin rozdělení*

Všechny tři rodiny jsou regulární. Z tabulky 1 je patrné, že při určitém rozsahu parametrů integrály (1) nabývají bez jakéhokoli zřejmého důvodu nekonečných hodnot. Bez důvodu z toho hlediska, že charakterizují funkci, pod kterou je jednotková plocha.

Na druhé koleji je metoda maximální věrohodnosti, v níž jsou naopak statistické funkcionály extrapolovány z konečné věrohodnostní funkce dat. Elegantní teorie zaručuje za poměrně slabých předpokladů ty nejlepší odhady parametrů rozdělení, odhadnuté parametry však často nenaplňují naše intuitivní očekávání, že budou centrem a mírou rozptýlenosti datového souboru: pro řadu rozdělení to jsou prostě jen nějaká čísla, která pozorovaný výběr „viditelně“ necharakterizují, se kterými jsme však zvyklí pracovat. Věcně je vše v pořádku, esteticky už méně. Bickel a Lehmann (1975) to komentují slovy: „Výběr vhodných charakteristik odhadovaných z dat velice závisí na předpokládané rodině rozdělení“.

Příjemné je, když jsou data z normálního rozdělení, pro které obě koleje splývají: střední hodnota a rozptyl pozorovaných souborů jsou zároveň odhady parametrů rozdělení, z něhož soubory pocházejí.

Tento trend pokračuje i při výstavbě moderní - robustní - statistiky. Na první koleji se nacházejí trochu nematematické, ale dobře použitelné charakteristiky zvané useknutý průměr a Winsorizovaný rozptyl, na druhé jsou M-odhady. I na této druhé koleji se však nejčastěji mluví o parametrech polohy a měřítka souborů, jako by univerzálním předobrazem modelů s kontaminací bylo kontaminované normální rozdělení. Kontaminované ovšem může být každé rozdělení, i to, jehož parametry „viditelně“ necharakterizují pozorovaná data.

Situace je důsledkem volby číselných charakteristik rozdělení ve tvaru obecných (1) a centrálních momentů. V článcích publikovaných v Robustu'96 a Robustu'00 jsem ukázal, že místo nich je možné zavést core momenty, které pro každé spojité regulární rozdělení existují a výstižně rozdělení charakterizují, i když trochu jinak, než jsme zvyklí. V tomto článku budeme studovat první dva core momenty. Uvidíme, že definují dvě zajímavé číselné charakteristiky: *těžiště* a *informaci* rozdělení. Těžiště lze zvolit za parametr rozdělení, odhadnout parametry těžiště a měřítka a jejich pomocí zkonstruovat odhad informace. „Výběrové těžiště“ a „informace obsažená ve výběru“ představují pak odhad těžiště a informace daného rozdělení podobně jako výběrová střední hodnota a výběrový rozptyl jsou odhadem střední hodnoty a rozptylu normálního rozdělení.

Za takovou „estetickou konzistenci“ je ovšem nutno zaplatit určitou cenu: parametr těžiště není pro rozdělení s nosičem $S \neq R$ parametrem polohy. Dále, parametr měřítka pro rozdělení s nosičem $S \neq R$ není tím parametrem měřítka, na který jsme zvyklí. Kromě toho je některá rozdělení třeba reparametrizovat.

Základním stavebním kamenem popisovaného přístupu je core funkce absolutně spojitého rozdělení, která je znovu zavedena v kapitole 2. Nestandardním prvkem teorie je, že k určení core funkce obecného rozdělení F je třeba nalézt „nejjednodušší“

rozklad $F = G\psi$, což je zdánlivě velice vágní, ale pro většinu konkrétních rozdělení kupodivu jednoznačné, požadujeme-li, aby core funkce matematicky jednoduše vyjádřených rozdělení byly taky jednoduché. Podstatné je, že pro přibližně normálně rozdělená data a asymptoticky normální odhady se nově navržené charakteristiky (a případné postupy z nich vyplývající) shodují s tradičními: core funkce normálního rozdělení je totiž $T_{\mu,\sigma}(x) = (x - \mu)/\sigma$.

2. Core funkce

2.1. **Core funkce jednoduchých rozdělení.** Buď $\emptyset \neq S = (a, b) \subseteq R$ a Π_S třída rozdělení na borelovských podmnožinách reálné přímky R, absolutně spojitých vzhledem k Lebesgueově míře λ na R, s distribuční funkcí F a hustotou $f = dF/d\lambda$, pro niž platí

$$f(x) = \begin{cases} > 0 & \text{pro} \quad x \in S \\ = 0 & \text{pro} \quad x \in R - S \end{cases}$$

regulární v Hájkově a Šidákově smyslu, což znamená že integrál

$$I_F = \int_S \left(\frac{f'(x)}{f(x)} \right)^2 f(x)\, dx \tag{2}$$

je konečný a kladný. Značíme $f'(x) = df(x)/dx$. I když je S otevřená, říkejme jí nosič. S je tedy nosičem F a výběrovým prostorem náhodné veličiny X s rozdělením F.

Budiž Y náhodná veličina s rozdělením $G \in \Pi_R$ a hustotou g. Její *core funkce* je prostě skórová funkce

$$T_G(y) = -\frac{g'(y)}{g(y)}. \tag{3}$$

Vyberme několik funkcí s definičním oborem reálná osa, které reprezentují různá možná chování funkce v nekonečnu: neomezené (označíme je N), omezené (O) a smíšené typy OB a BO.

N1	N2	NO	ON	O1	O2
$\sinh y$	y	$1 - e^{-y}$	$e^y - 1$	$\tanh \frac{y}{2}$	$\frac{2y}{1+y^2}$

Tab. 2 *Reálné funkce s definičním oborem $S = R$*

Typ	$T_G(y)$	$g(y)$	Rozdělení
$U1$	$\sinh y$	$\frac{1}{2K_0(1)} e^{-\cosh y}$	neužívané
$U2$	y	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$	normální
UB	$1 - e^{-y}$	$e^{-y} e^{-e^{-y}}$	extremí hodnoty
BU	$e^y - 1$	$e^y e^{-e^y}$	Gumbelovo
$B1$	$\tanh \frac{y}{2}$	$\frac{e^y}{(1+e^y)^2}$	logistické
$B2$	$\frac{2y}{1+y^2}$	$\frac{1}{\pi(1+y^2)}$	Cauchyho

Tab. 3 *Core funkce a hustoty jednoduchých rozdělení*

Považujme funkce v tabulce 2 za core funkce jednoduchých rozdělení. Pomocí (3) určíme příslušné hustoty. Výsledek operace je zachycen v tabulce 3 (shodné s tabulkou 1 v Robust'96). Zde K_0 je Besselova funkce třetího druhu. S výjimkou prvého jsou to všechno známá standardizovaná rozdělení.

2.2. **Core funkce složených rozdělení.** Buď $\Phi_S = \{\varphi : R \to S\ \}$ množina homeomorfních zobrazení a $X = \varphi(Y)$ transformovaná náhodná veličina. Distribuční funkce X je $F = G\varphi^{-1}$ a hustota

$$f(x) = g(\psi(x))\ \psi'(x), \tag{4}$$

kde jsme označili $\psi = \varphi^{-1}$. Rozdělení, které lze zapsat ve tvaru $F = G\psi$, kde $G \in \Pi_R$ a $\varphi \in \Phi_S$, budeme říkat *složené*. Core funkce složeného rozdělení byla v R00 (tam jsme mu říkali indukované) definována vztahem

$$T_F(x) = T_G(\psi(x)) = -\frac{g'(\psi(x))}{g(\psi(x))}. \tag{5}$$

Pomocí (4) se okamžitě odvodí vzorec (Robust'00, Věta 1), v němž core funkce složeného rozdělení $F = G\psi$ už nezávisí na G, ale pouze na hustotě f a zobrazení ψ:

$$T_F(x) = \frac{1}{f(x)}\frac{d}{dx}\left(-\frac{1}{\psi'(x)}f(x)\right). \tag{6}$$

Skutečně, položíme-li $u = \psi(x)$, platí

$$T_F(x) = -\frac{1}{g(u)}\frac{dg(u)}{du} = \frac{\psi'(x)}{f(x)}\frac{d}{dx}\left(-\frac{f(x)}{\psi'(x)}\right)\frac{dx}{du}$$

a $du/dx = \psi'(x)$.
V tabulce 4 jsou uvedeny obecné výrazy pro core funkce a hustoty složených rozdělení $F = G\psi$, kde G jsou jednoduchá rozdělení z tabulky 3.

Typ	$T_F(x)$	$f(x)$
$N1$	$\frac{1}{2}(e^{\psi(x)} - e^{-\psi(x)})$	$\frac{1}{2K_0(1)}e^{-\frac{1}{2}(e^{\psi(x)}+e^{-\psi(x)})}\ \psi'(x)$
$N2$	$\psi(x)$	$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\psi^2(x)}\ \psi'(x)$
NO	$1 - e^{-\psi(x)}$	$e^{-\psi(x)}e^{-e^{-\psi(x)}}\ \psi'(x)$
ON	$e^{\psi(x)} - 1$	$e^{\psi(x)}e^{-e^{\psi(x)}}\ \psi'(x)$
$O1$	$\frac{e^{\psi(x)}-1}{e^{\psi(x)}+1}$	$\frac{e^{\psi(x)}}{(1+e^{\psi(x)})^2}\ \psi'(x)$
$O2$	$\frac{2\psi(x)}{1+\psi^2(x)}$	$\frac{1}{\pi(1+\psi^2(x))}\ \psi'(x)$

Tab. 4 *Core funkce a hustoty složených rozdělení*

Z porovnání tabulek 3 a 4 je patrné, že typ core funkce složeného rozdělení F je stejný jako typ core funkce jeho „prototypu" G nezávisle na konkrétní $\psi = \varphi^{-1}, \varphi \in \Phi_S$.

Tabulka 5 je verzí tabulky 4 pro speciální volbu $\psi(x) = \ln x$ pro $S = (0, \infty)$ a je shodná s tabulkou 2 v Robust'96. Zde GIG značí 'generalized inverse Gaussian', viz Johnson, Kotz a Ballakrishnan (1994).

Jiným možným zobrazením $\psi : (0, \infty) \to R$ je například $\psi(x) = Ei(x)$. Žádné z užívaných modelových rozdělení však nemá hustotu ve tvaru (4), ve kterém by vystupovala funkce Ei. Dlouholeté úsilí autorů modelových statistických rozdělení

o popis hustot jednoduchými vzorci přineslo ovoce: transformace $\psi(x) = \ln(x)$ má pro rozdělení na nosiči $S = (0, \infty)$ monopolní postavení. Hustota každého rozdělení F na $S = (0, \infty)$ je buď přímo tvaru $f(x) = h(x)\frac{1}{x}$ nebo lze zapsat jako $f(x) = xf(x)\frac{1}{x}$ a položit $g(y) = e^y f(e^y)$. Vzorec (6) má v případě $S = (0, \infty)$ tvar

$$T_F(x) = -1 - x\frac{f'(x)}{f(x)}.$$

Typ	$T_F(x)$	$f(x)$	Rozdělení
$N1$	$\frac{1}{2}(x - 1/x)$	$\frac{1}{2K_0(1)x}e^{-\frac{1}{2}(x+1/x)}$	GIG
$N2$	$\ln x$	$\frac{1}{\sqrt{2\pi}x}e^{-\frac{1}{2}\ln^2 x}$	lognormální
NO	$1 - 1/x$	$\frac{1}{x^2}e^{-1/x}$	extrémní hodnoty II
ON	$x - 1$	e^{-x}	exponenciální
$O1$	$\frac{x-1}{x+1}$	$\frac{1}{(1+x)^2}$	log-logistické
$O2$	$\frac{2\ln x}{1+\ln^2 x}$	$\frac{1}{\pi x(1+\ln^2 x)}$	log-Cauchyovo

Tab. 5 *Core funkce a hustoty rozdělení na* $S = (0, \infty)$ *při* $\psi(x) = \ln(x)$

2.3. **Struktura parametrických rozdělení.** Buďte $\mu \in R$ a $\sigma \in (0, \infty)$ obvyklé parametry polohy a měřítka. Náhodná veličina Y s rozdělením $G \in \Pi_R$ indukuje rodinu

$$\sigma Y + \mu = Y_{(\mu,\sigma)}, \qquad \mu \in R, \sigma > 0$$

s parametry polohy a měřítka. Díky vlastnostem zobrazení $\psi(\cdot)$ existuje náhodná veličina $X_{(\mu,\sigma)}$ daná vztahem

$$X_{(\mu,\sigma)} = \psi^{-1}(Y_{(\mu,\sigma)}).$$

Zřejmě $X = X_{(0,1)}$ a platí

$$X_{(\mu,\sigma)} = \psi^{-1}(\sigma\psi(X) + \mu).$$

Položíme $\tau = \psi^{-1}(\mu)$, takže $X_{(\mu,\sigma)} = X_{(\psi(\tau),\sigma)}$. Parametr τ, kterému jsme v Robust'00 říkali *transformovaný parametr polohy,* ovšem není parametrem polohy v obvyklém smyslu slova [např. Bickel a Lehmann (1975)), Jurečková (2001)]. Hustotu a core funkci náhodné veličiny $X_{(\tau,\sigma)} = X_{(\psi(\tau),\sigma)}$ můžeme podle (4) a (5) zapsat jako

$$f_{\tau,\sigma}(x) = \frac{1}{\sigma}g(u)\psi'(x), \tag{7}$$

$$T_{F_{\tau,\sigma}}(x) = T_G(u), \tag{8}$$

kde pro u platí

$$u = \frac{\psi(x) - \psi(\tau)}{\sigma}. \tag{9}$$

Některé možné transformace $\psi : S \to R$ pro nejběžnější S a tvary příslušných u jsou uvedeny v tabulce 6.

S	$\psi(x)$	$\psi'(x)$	u
R	x	1	$\frac{x-\mu}{\sigma}$
R	$\sinh x$	$\cosh x$	$\frac{2}{\sigma}\sinh\frac{x-\tau}{2}\cosh\frac{x+\tau}{2}$
$(0,\infty)$	$\ln x$	$\frac{1}{x}$	$\ln\left(\frac{x}{\tau}\right)^{1/\sigma}$
$(0,1)$	$\ln\frac{x}{1-x}$	$\frac{1}{x(1-x)}$	$\ln\left(\frac{x(1-\tau)}{(1-x)\tau}\right)^{1/\sigma}$
$(0,1)$	$-\ln(-\ln x)$	$\frac{-1}{x\ln x}$	$\ln\left(\frac{\ln\tau}{\ln x}\right)^{1/\sigma}$
$(-1,1)$	$\tanh^{-1}x$	$\frac{1}{1-x^2}$	$\frac{1}{\sigma}\tanh^{-1}\frac{x-\tau}{1-x\tau}$
$(-1,1)$	$\tan\frac{\pi}{2}x$	$\frac{\pi}{2}(\cos\frac{\pi}{2}x)^{-2}$	$\frac{1}{\sigma}\frac{\sin\frac{\pi}{2}(x-\tau)}{\cos\frac{\pi}{2}x\cos\frac{\pi}{2}\tau}$

Tab. 6 *Transformace $\psi : S \to R$ a příslušná u*

V parametrickém prostoru $\Theta \subseteq R^m$ jsme tedy detekovali určitou pevnou strukturu, která je pro složená rozdělení obrazem struktury $u = \frac{x-\mu}{\sigma}$ jednoduchých rozdělení. Pro dané $S = (a, b) \subseteq R$ můžeme Θ uvažovat ve tvaru

$$\Theta = \{(t, \sigma, \mathbf{c}) : t \in S, \sigma \in (0,\infty), \mathbf{c} \in (0,\infty)^{m-2}\}$$

kde

$$t = \begin{cases} \mu & \text{pro jednoduchá rozdělení } G \\ \tau\ \{= \psi^{-1}(\mu)\} & \text{pro } F = G\psi, \end{cases}$$

σ je parametr měřítka (rozdělení na $S = (0,\infty)$ mívají obvykle parametr $\beta = 1/\sigma$) a $\mathbf{c}$ představuje $m-2$ tvarových parametrů. Kterýkoli z parametrů může samozřejmě chybět.

Tím jsme vlastně rozdělili obecný parametrický prostor Θ na dvě části. V jedné jsou parametry t a σ, které pro jednoduchá rozdělení vytvářejí strukturu $u = \frac{y-\mu}{\sigma}$, pro složená transformovanou strukturu $u = (\psi(x) - \psi(\tau))/\sigma$. Ve druhé části jsou tvarové parametry, které při našich úvahách můžeme zahrnout do vzorce pro hustotu rodiče (které je ale obecně nutno odhadovat podobně jako parametry první části). Příklady rozdělení s tvarovými parametry vidíme v tabulce 7. Rodiny $f_{\alpha,\nu}$ s $\alpha \in (0,\infty)$ a $\nu \in (0,\infty)$ mají rodiče $f = f_{1,1}$ v tabulce 5.

Typ	$T_{F_{\alpha,\nu}}(w)$	$f_{\alpha,\nu}(w)$
$N1$	$\frac{\alpha}{2}[w - \nu/w + \nu - 1]$	$\frac{\nu^{\rho/2}w^{-\rho}}{2K_\rho(\alpha\sqrt{\nu})}e^{-\frac{\alpha}{2}(w+\nu/w)}$
NO	$\alpha(1-1/w)$	$\frac{\alpha^\alpha}{\Gamma(\alpha)}w^{-\alpha}e^{-\alpha/w}$
ON	$\alpha(w-1)$	$\frac{\alpha^\alpha}{\Gamma(\alpha)}w^{\alpha}e^{-\alpha w}$
$O1$	$\alpha\frac{w-1}{w+1/\nu}$	$\frac{1}{\nu^\alpha B(\nu\alpha,\alpha)}\frac{w^{\nu\alpha}}{(w+1/\nu)^{[1+\nu]\alpha}}$
$O2$	$\frac{2\alpha\ln w}{1+\ln^2 w}$	$\frac{1}{B(\frac{1}{2},\alpha-\frac{1}{2})}\frac{1}{(1+\ln^2 w)^\alpha}$

Tab. 7 *Core funkce a hustoty rodin s tvarovými parametry*

V tabulce je $B(p,q)$ beta funkce, Γ gamma funkce a $\rho = \frac{\alpha}{2}(\nu - 1)$. Za w je možno dosadit „strukturu“ $w = e^u = (\frac{x}{\tau})^\beta, \beta = 1/\sigma$ (v tom případě jsou $N1$ a $O1$

reparametrizovaná rozdělení GIG a transformed beta, viz Johnson (1994) a Klugman (1998), NO a ON jsou zobecněné rozdělení extrémní hodnoty II a general gamma, Klugman (1998), a $O2$ je zobecněné log-Cauchyovo rozdělení).

2.4. **Základní věta.** Připomeňme, že věrohodnostní skór parametru $\gamma = \theta_j$ rozdělení F_θ je $s_\gamma(x) = \frac{\partial}{\partial\gamma} \ln f_\theta(x)$. Teď už snadno dokážeme větu (Robust'00, Věta 2), ze které vyplývá interpretace core funkce.

Věta 1. *Pro $F_\theta \in \Pi_S$, $\theta \in \Theta$ platí*

$$s_t(x) = \frac{1}{\sigma}\psi'(t)\ T_{F_\theta}(x).$$

Důkaz. Parametr **c** zahrneme do rodiče F (a „prototypu“ G). Platí

$$\frac{\partial}{\partial t} \ln f_\theta(x) = \frac{1}{f_\theta(x)} \frac{d}{du} f_\theta(x) \frac{\partial u}{\partial t}.$$

Podle (7) je $f_\theta(x) = f_{t,\sigma}(x) = \frac{1}{\sigma} g(u)\psi'(x)$, podle (9) je $\frac{\partial u}{\partial t} = \sigma^{-1}\psi'(t)$. Užitím (3) a (8) dostáváme

$$\frac{\partial}{\partial t} \ln f_\theta(x) = \frac{1}{\sigma}\psi'(t) T_G(u) = \frac{1}{\sigma}\psi'(t) T_{F_{t,\sigma}}(x).$$

□

Pro rozdělení, která mají parametr t (např. všechna rozdělení na $S = R$, Weibullovo, log-logistické, log-Cauchyho, Johnsonovo U_B a spousta dalších) je tedy core funkce *vnitřní částí věrohodnostního skóru pro t*. Jiná parametrická rozdělení na $S \neq R$ parametr t nemají (gamma, beta, Lomaxovo, rovnoměrné a některá další). Ta je často možné reparametrizovat.

3. Core momenty

V Robust'00 jsme definovali core momenty rozdělení F s core funkcí T_F jako

$$M_k(F) = \int_S T_F(x)^k dF(x), \qquad k = 1, 2, ...$$

Core momenty jsou překvapivě jednoduché. Platí pro ně

(i) Buď $G \in \Pi_R$ a $F = G\psi$. Pak $M_k(F) = M_k(G)$ (Robust'96, věta 4).
(ii) Buď $F \in \Pi_S$. Pak $M_k(F)$ existuje pro libovolné přirozené k (Robust'96, věta 5).

Následující věta tvrdí, že core momenty parametrických rozdělení nezávisí na „strukturních“ parametrech, ale jen na tvaru (tvarových parametrech) rodiče.

Věta 2. *$M_k(F)$ parametrického rozdělení nezávisí na parametrech t a σ.*

Důkaz: Podle (7) je $M_k(F_\theta) = \int_S T_G(u)^k g(u) \frac{1}{\sigma}\dot{\psi}(x)\ dx = \int_R T_G(u)^k g(u)\ du = M_k(G) = M_k(F)$. □

V tabulce 8 jsou uvedeny první čtyři core momenty rodin z tabulky 1 (obecné vzorce jsou rekurentní). Core momenty rozdělení s těžkými chvosty nejenže existují, ale jsou vyjádřeny pouze pomocí parametrů (mezi nimiž ve shodě s Větou 2 chybí t a σ). V tabulce jsme označili $\rho = (\nu + 1)\alpha$.

F_θ	M_1	M_2	M_3	M_4
1	0	α	-2α	$3\alpha(\alpha+2)$
2	0	$\frac{\nu}{\rho+1}\alpha^2$	$\frac{2\nu(1-\nu)}{(\rho+1)(\rho+2)}\alpha^3$	$\frac{3\nu[\nu\rho+2(\nu^2-\nu+1)]}{(\rho+1)(\rho+2)(\rho+3)}\alpha^4$
3	0	$\frac{\alpha(2\alpha-1)}{\alpha+1}$	0	$\frac{12\alpha^3(2\alpha-1)(2\alpha+1)}{(\alpha+1)(2\alpha+4)(2\alpha+6)}$

Tab. 8 *Core momenty rodin z tabulky 1*

4. Těžiště a informace rozdělení

Vysvětlíme smysl prvních dvou core momentů.

Věta 3. *Pro $F \in \Pi_S$ platí*

$$M_1(F) = \int_S T_F(x)\, dF(x) = 0. \tag{10}$$

Důkaz. Větu už jsme sice dokázali v (Robust'00, Věta 2), ale ještě jednou. Podle Definice 1 a (7) platí

$$M_1(F) = \int_S T_F(x) f(x)\, dx = \int_S -\frac{g'(u)}{g(u)}\frac{1}{\sigma} g(u)\psi'(x)\, dx = -\int_{-\infty}^{\infty} g'(y)\, dy = 0.$$

□

Core momenty jsou tedy centrálními momenty kolem bodu $x^* : T_F(x^*) = 0$.

Definice. *Buď $F \in \Pi_S$. Bod x^*, pro který platí $T_F(x^*) = 0$, nazveme těžištěm rozdělení F.*

Pro rozdělení G s nosičem $S = R$ je těžištěm mód rozdělení, který označíme y^* a který pro $G = G_\theta$ vystupuje jako parametr polohy μ. Těžištěm parametrického rozdělení $F_\theta = G_\theta\psi$ s nosičem $S \neq R$ je podle předešlého „obraz" parametru polohy rozdělení G_θ, což je transformovaný parametr polohy $\tau = \psi^{-1}(\mu)$. Těžištěm rozdělení s nosičem $S \neq R$ bez parametrů (nebo s parametry, mezi kterými t chybí) je tedy zřejmě bod $x^* = \psi^{-1}(y^*)$. Ten je vždy různý od módu rozdělení a obecně i od jeho střední hodnoty (pokud existuje) a mediánu. Parametr t, který jsme z praktických důvodů zavedli už dříve, je tedy *parametrem těžiště*.

Teď už také dokážeme přesně říct, co vlastně je core funkce spojitého rozdělení: je to věrohodnostní skór vzhledem k těžišti rozdělení (*ať už toto je parametrem rozdělení nebo ne*). Z robustní statistiky víme, že vlivová funkce odhadovaného parametru je úměrná věrohodnostnímu skóru pro tento parametr. Konečná interpretace core funkce tedy zní:

Core funkce $T_F(x)$ regulárního rozdělení je vlivová funkce těžiště a popisuje relativní vliv hodnoty $x \in S$ na polohu těžiště rozdělení.

Považuji za dobré znamení, že i čtverec core funkce má rozumnou interpretaci. Střední hodnota čtverce core funkce rozdělení G s nosičem $S = R$ je dána vztahem (2), který Cover a Thomas (1991, str.494) interpretují jako Fisherovu informaci rozdělení G (bez obvyklého vázání pojmu Fisherovy informace k určitému parametru).

Co se vlastně od informace obsažené v pozorování $x \in S$ požaduje? Především by měla být nezáporná (této základní podmínce nevyhovuje funkce $-\ln f(x)$, o jejíž střední hodnotě, známé jako diferenciální entropie, se v teorii informace uvažuje jako o veličině se vztahem k informaci). Dále: relativní informace, obsažená v pozorování

x, které se dá očekávat, by měla být nízká, zatímco nečekané pozorování by se mělo projevit vysokou hodnotou informace.

V Robust'96 bylo dokázáno, že bod x^*, pro který platí $T_F(x^*) = 0$, je nejméně informativním bodem rozdělení F. Připomeňme proč. Hustotu F lze podle (4) vyjádřit jako $f(x) = g(\psi(x))\psi'(x)$, ovšem člen $\psi'(x)$ je společný celé třídě rozdělení na daném nosiči S a nenese tedy o rozdělení žádnou informaci. Řešením rovnice $\frac{d}{dx}g(\psi(x)) = 0$ (zde jsme ještě jednou použili (4)) je bod $x : \frac{d}{dx}(\frac{1}{\psi'(x)}f(x)) = 0$, neboli (použijeme (6)) bod $x^* : T_F(x^*) = 0$.

Funkce $i_F = T_F^2$ je nezáporná funkce, která nabývá minima v nejméně informativním bodě rozdělení. Rozdělení s lehkými chvosty mají neomezenou T_F, takže odlehlá pozorování x_i poskytují vysoké hodnoty $T_F^2(x_i)$ - upozorňující, že je třeba změnit model - zatímco v případě rozdělení s těžkými konci, která mají omezenou T_F, jsou hodnoty $T_F^2(x_i)$ pro odlehlá pozorování x_i nízké - při takovém rozdělení se odlehlá pozorování dají očekávat. Jak už bylo řečeno, střední hodnota funkce i_F je považována za informaci.

Mám tedy zato, že reálná funkce $i_F(x) = T_F(x)^2$ vyjadřuje *relativní informaci o těžišti rozdělení* F, kterou nese hodnota $x \in S$. Střední hodnota funkce i_F,

$$M_2(F) = \int_S T_F^2(x) f(x)\, dx,$$

je *střední informací o těžišti rozdělení* F a lze ji chápat jako *střední informaci rozdělení* F.

Pro parametrické rozdělení $F_\theta \in \Pi_S$, kde $\theta \in \Theta$, je Fisherova informace o parametru t dána vztahem $I_{tt}(\theta) = E_\theta s_t^2$ a podle Věty 1 tedy platí

$$I_{tt}(\theta) = (\sigma^{-1}\psi'(t))^2 M_2(F). \tag{11}$$

Za střední informaci J_{F_θ} parametrického rozdělení F_θ lze tedy považovat Fisherovu informaci o těžišti,

$$J_{F_\theta} = I_{tt}(\theta). \tag{12}$$

Pro $S = (0, \infty)$, (11) znamená $J_{F_\theta} = (\sigma\tau)^{-2} M_2(F)$.

5. Odhady

Buď $X_1, \ldots, X_n$ náhodný výběr z rozdělení $F_{t,\sigma}$ s nosičem S a neznámými parametry t, σ. Hledáme „výběrové těžiště“ a informaci výběru. Ve vzorci (11) figuruje parametr měřítka a úloha se tedy redukuje na odhad parametrů těžiště a měřítka rozdělení $F_{t,\sigma}$.

5.1. **Výběr z normálního rozdělení.** Předpokládejme, že $X_1, \ldots, X_n$ je výběr z normálního rozdělení Φ s hustotou $\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$. Střední hodnota a rozptyl normálního rozdělení jsou μ a

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \phi(x)\, dx,$$

zatímco jeho Fisherova informace o μ je

$$I_{\mu\mu}(\theta) = \int_{-\infty}^{\infty} \left(\frac{x-\mu}{\sigma^2}\right)^2 \phi(x)\, dx = \frac{1}{\sigma^2}. \tag{13}$$

Core funkce normálního rozdělení je $T_\Phi(x) = (x-\mu)/\sigma$, druhý core moment je

$$M_2(\Phi) = \int_{-\infty}^{\infty} T_\Phi(x)^2 \phi(x)\, dx = \int_{-\infty}^{\infty} \left(\frac{x-\mu}{\sigma}\right)^2 \phi(x)\, dx = 1.$$

Podle (11) je informace normálního rozdělení $J_\Phi = 1/\sigma^2$, což souhlasí s (13). Soustava rovnic pro odhad parametrů core momentovou metodou (viz 5.2.2) je totožná se soustavou věrohodnostních rovnic. Odhadem těžiště a informace normálního rozdělení jsou tedy hodnoty $\hat\mu$ a $\hat J_\Phi = 1/s^2$, kde s je odhad rozptylu výběru $X_1, ..., X_n$.

5.2. **Výběr z obecného rozdělení.** Buď $X_1, \dots, X_n$ výběr z obecného rozdělení $F_{t,\sigma}$ (tvarové parametry předpokládáme dané).

5.2.1. *Těžiště.* Pro některá rozdělení lze odhadovat těžiště bez znalosti parametru měřítka přímo z empirického protějšku funkcionálu (10), t.j. rovnice

$$\hat t_n : \qquad \frac{1}{n}\sum_{i=1}^{n} T_F(u_i) = 0, \tag{14}$$

kde u_i je „struktura“ příslušná danému S a ψ (tabulka 6), do které jsou dosazeny pozorované hodnoty, tedy např.

$$u_i = \begin{cases} \frac{x_i-\mu}{\sigma} & \text{pro} \quad S = R \ \text{ a } \ \psi(u) = u \\ \ln\left(\frac{x_i}{\tau}\right)^{1/\sigma} & \text{pro} \quad S = (0,\infty) \ \text{ a } \ \psi(u) = \ln u. \end{cases}$$

V Tabulce 9 jsou uvedeny hustoty, core funkce a těžiště t některých rozdělení a vzorec pro *výběrové těžiště* $\hat t_n$ (odhad těžiště souboru i rozdělení).

Rozdělení	$f_\theta(x)$	$T_{F_\theta}(x)$	t	$\hat t_n$
Normální	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$	$\frac{x-\mu}{\sigma}$	μ	$\bar x = \frac{1}{n}\sum x_i$
Lognormální	$\frac{1}{\sqrt{2\pi}x} e^{-\frac{1}{2}\log^2(x/\tau)}$	$\ln x/\tau$	τ	$\bar x_G = \sqrt[n]{x_1 ... x_n}$
Exponenciální	$\tau^{-1} e^{-x/\tau}$	$x/\tau - 1$	τ	$\bar x$
Extr. hodn. II	$\frac{\tau}{x^2} e^{-\tau/x}$	$1 - \tau/x$	τ	$\bar x_H = \frac{n}{\sum 1/x_i}$
GIG	$\frac{1}{2K_0(1)x} e^{-\frac{1}{2}(x/\tau+\tau/x)}$	$\frac{1}{2}(x/\tau - \tau/x)$	τ	$\bar x - \bar x_H$
Gumbelovo	$e^{x-\mu} e^{-e^{x-\mu}}$	$e^{x-\mu} - 1$	μ	$\ln(\frac{1}{n}\sum e^{x_i})$
Lomaxovo	$\frac{\alpha}{(1+x)^{\alpha+1}}$	$\frac{x-1/\alpha}{x+1}$	$1/\alpha$	$\frac{\sum x_i/(1+x_i)}{\sum 1/(1+x_i)}$
Gamma	$\frac{\gamma^\alpha}{\Gamma(\alpha)x} x^{\alpha-1} e^{-\gamma x}$	$\alpha(\frac{x}{\alpha/\gamma} - 1)$	α/γ	$\bar x$
Beta	$\frac{1}{B(p,q)} x^{p-1}(1-x)^{q-1}$	$p(\frac{x}{p/(p+q)} - 1)$	$\frac{p}{p+q}$	$\bar x$

Tab. 9 *Core funkce a odhady těžiště*

Odhadem těžiště souboru s lognormálním rozdělením je geometrický průměr, souboru s rozdělením extremní hodnoty harmonický průměr. Core funkce exponenciálního rozdělení a rozdělení gamma a beta jsou lineární. Stejně jako u normálního rozdělení, jejich těžištěm je střední hodnota a odhadem těžiště aritmetický průměr naměřených hodnot. Z tabulky je i patrné, jak je nutno reparametrizovat hustoty rozdělení gamma a beta, aby měly (podobně jako v tabulce jejich core funkce) explicitně vyjádřený parametr těžiště.

5.2.2. *Odhad těžiště a měřítka.* Pokud má rozdělení F těžiště vyjádřené jako parametr, je rovnice (14) věrohodnostní rovnicí pro t (Věta 1). Na rozdíl od rozdělení z tabulky 9, odhady těžiště většiny ostatních rozdělení nejsou invariantní vůči změně měřítka a rovnici (14) je nutno řešit při vhodně volené hodnotě s parametru σ.

Pro současný odhad parametrů těžiště t a měřítka σ (respektive $\beta = 1/\sigma$) je k dispozici řešení věrohodnostních rovnic a jejich robustních modifikací. Příslušné odhady označme $(\hat{t}_n)_{ML}$ a $(\hat{\sigma}_n)_{ML}$. V R00 byla představena jiná možnost: metoda core momentů. Empirickým protějškem funkcionálů pro první dva core momenty je soustava rovnic pro MM-odhady $(\hat{t}_n)_{MM}$ a $(\hat{\sigma}_n)_{MM}$,

$$\begin{aligned} \sum_{i=1}^{n} T_F(u_i) &= 0 \\ \frac{1}{n}\sum_{i=1}^{n} T_F^2(u_i) &= M_2(F) \end{aligned}$$

(do rovnic (20) v Robust'00 se vloudila chybička: chybí tam ten moment). I MM-odhady jsou konsistentní a asymptoticky normální, nejsou efficientní, ale jejich robustní varianty představují dobrou alternativu k obvykle používaným robustním metodám, viz Robust'00 a Fabián (2001b).

5.2.3. *Informace.* Označme J_n informaci o těžišti rozdělení obsaženou ve výběru $X_1, \ldots, X_n$ z rozdělení F_θ a J_{F_θ} střední informaci F_θ. Podle (12) je $J_{F_\Phi} = I_{tt}(\theta)$ kde $I_{tt}(\theta)$ je Fisherova informace o t. Podle Cramér-Raovy věty je rozptyl asymptoticky normálního odhadu $\hat{t}_n$ roven

$$Var(\hat{t}_n) = \frac{1}{nI_{tt}(\theta)},$$

a pro normální rozdělení platí (viz 5.1)

$$J_n = 1/Var(\hat{t}_n).$$

Máme tedy

$$J_n = nJ_{F_\theta}.$$

Cramér-Raovu větu lze tedy interpretovat ve smyslu, že informace J_n o těžišti ve výběru $X_1, \ldots, X_n$ z rozdělení je n-krát informace rozdělení, z něhož výběr pochází. Podle (11) pak např. pro rozdělení s nosičem $S = (0, \infty)$ platí

$$J_n = \frac{n}{(\hat{\tau}_n\hat{\sigma}_n)^2} M_2(F),$$

z něhož lze odhadovat nutnou velikost výběru n pro předpokládaný typ F rodiče rozdělení.

Příklad. Buď $X_1, \ldots, X_n$ výběr z gamma rozdělení (tabulka 9). Podle tabulky $\tau = \alpha/\gamma$, podle Fabián (2001) $M_2(\text{gamma}) = \alpha$ a $J_n = \frac{n}{\alpha/\gamma}\alpha = n\gamma$. Buď n_1 velikost výběru z exponenciálního rozdělení ($\tau = 1, \sigma = 1, M_2 = 1$), pro kterou už přibližně platí asymptotické vztahy (lze odhadnout pomocí simulací). Pak $J_{n_1} = n_1$ a z požadavku $J_n = J_{n_1}$ plyne $n = n_1/\gamma$.

Literatura

[1] Antoch, J., Vorlíčková, D. (1992): *Vybrané metody statistické analýzy dat*, Academia Praha.

[2] Bickel, P.J. a Lehmann, E.L. (1975): Descriptive statistics for nonparametric models I, II, *Ann. Statist.*, **3**, 1038-69.

[3] Cover, T.M. a Thomas, J.A. (1991): *Elements of Information Theory,* Wiley.
[4] Fabián, Z. (1997): Geometrické momenty, sb. *Robust'96*, 49-62.
[5] Fabián, Z. (2001): MM-odhady, sb. *Robust'2000*, 33-41.
[6] Fabián, Z. (2001a): Vzdálenost pozorovaných hodnot, *Informační bulletin ČSS*, **3**, 17-22.
[7] Fabián, Z. (2001b): Induced cores and their use in robust parametric estimation. *Commun. in statistics, Theory methods*, 30, **3**, 537-556.
[8] Johnson, N. L. (1949): Systems of frequency curves generated by methods of translations, *Biometrika,* 36, 149-176.
[9] Johnson, N. L., Kotz, S. a Balakrishnan, N. (1994, 95): *Continuous univariate distributions 1, 2*, Wiley.
[10] Jurečková, J. (2001): *Robustní statistické metody*, Karolinum.
[11] Klugmann, S.A., Panjer, H.H. a Willmott, G.E. (1998): *Loss models. From data to decisions*, Wiley.
[12] Lachout, P. (2001): Proces odhadování parametrů modelu, sb. *Robust'2000*, 149-163.
[13] Wilcock,R.R. (2001): *Fundamentals of Modern Statistical Methods,* Springer.

Ústav informatiky AV ČR, Pod Vodárenskou věží 2, 182 07 Praha
E-mail: zdenek@cs.cas.cz