ROBUST'2002, 143 – 152

O KVANTILECH VE VÍCE ROZMĚRECH

DANIEL HLUBINKA

Abstract. In the paper we study possible generalization of quantiles to multi-dimensional data. We show that our definition is correct for infinite dimensional Hilbert spaces as well. The presented theory is illustrated on several special cases using simulated two dimensional samples.

Резюме. В статье мы изучаем возможность общего определения многорозмерных квантилов. Мы покажем что наще определение справедливо даже для бесконечномерных гильбертовских пространств. Для иллюстрации теории в статье мы используем моделирования двухмерных случайных выборов.

1. Úvod

V matematické statistice se pro centrální míry polohy jednorozměrných dat často používají dvě míry – střední hodnota a medián. Obě tyto míry lze snadno zobecnit i do vícerozměrných prostorů. Máme-li posoudit polohu nějaké hodnoty vůči datům či rozdělení, používáme jinou míru polohy charakterizující dané rozdělení. Za tyto hodnoty se nejčastěji používají kvantily.

Definice kvantilu pro jednorozměrná data je snadná a názorná. Bohužel však zcela závisí na lineárním uspořádání jednorozměrných reálných dat. V následujících odstavcích se pokusíme navrhnout jinou možnou definici necentrální míry polohy, tedy obdoby kvantilu, pro mnohorozměrné prostory a ukážeme si, že tato definice je použitelná i na širokou třídu neeuklidovských prostorů, minimálně na Hilbertovy prostory. Ukážeme si odhadnuté kvantily pro simulovaná dvourozměrná data, zkusíme spočítat nejjednodušší případy přesně a v závěrečné diskusi nastíníme možná použití, ale i mnoho problémů a otevřených otázek.

2. Pokusy o mnohorozměrné kvantily

Uveďme zde některé existující návrhy definice mnohorozměrných kvantilů. Poté se pokusíme hledat kritéria, která nás později dovedou k naší definici kvantilu.

Z mnoha existujících definic mnohorozměrných kvantilů připomeňme v tuto chvíli dvě. První z nich je takzvaná hloubka dat. Jedná se vlastně o konvexní obálky které musíme „projít“ cestou ke středu dat. Podle toho, jak hluboko se data nacházejí, stanovujeme hodnotu hloubky daného pozorování. Pozorování s největší hloubkou je medián. Tato představa je velmi názorná a již poměrně dobře diskutovaná v literatuře. Zájemce o ni odkazujeme například na Rousseeuwovy práce. Povšimněme si však, že již pro dvourozměrná spojitá rozdělení je nemožné definovat hloubku

2000 *Mathematics Subject Classification.* Primary: 62E10, 62H05; Secondary: 60B11, 60E05.

Klíčová slova. Mnohorozměrné kvantily.

Tato práce vznikla za podpory výzkumného záměru MŠMT ČR MSM 113200008 Matematické metody ve stochastice. Autor by rád poděkoval všem spolupracovníkům na MFF UK, kteří trpělivě naslouchali vývoji jeho pohledů na diskutovanou problematiku a svými poznámkami přispěli ke vzniku tohoto článku.

dat korektně – například pro dvourozměrné normální rozdělení. Tím se tato metoda omezuje, alespoň z našeho pohledu, na metodu určenou k analýze dat, nikoliv možnost, jak definovat teoretické centrální oblasti pro mnohorozměrná rozdělení.

Druhá možnost, kterou zde zmíníme, je založena na jednorozměrných kvantilech a projekcích. Je použitelná jak pro data, tak i pro rozdělení, bohužel však pouze dvourozměrná. Podstatou je projekce dvourozměrného prostoru na přímku a spočítání kvantilů na této přímce. Provedení tohoto postupu na všechny možné směry (přímky procházející počátkem) pak dává požadovaný výsledek. Nezanedbatelnou výhodou tohoto postupu je jeho snadný výpočet i pro dostatečné množství různých směrů.

Již z uvedených příkladů je zřejmé, že při hledání definice mnohorozměrného kvantilu musíme zapomenout na interpretaci jednorozměrného kvantilu coby hodnoty rozdělující data na větší a menší. V mnohorozměrném prostoru nám chybí přirozené lineární uspořádání. Ačkoliv si jistě můžeme zavést lineární uspořádání (například lexikografické na Euklidově konečně rozměrném prostoru), nemusí takové uspořádání umožňovat rozumné pohledy na daný prostor a proto se tudy dále pouštět nebudeme.

Pokusy použít distribuční funkci k definici mnohorozměrného kvantilu obdobně jako se používá pro jednorozměrný kvantil bohužel selhávají z podobného důvodu.

Je zde ovšem jiná interpretace kvantilu. Medián chápeme jako centrální míru polohy a symetrické kvantily jsou pak hranicemi intervalů (ve více rozměrech raději oblastí) „spolehlivosti“ oddělující centrum dat od vnějších oblastí. Takto pojímaný kvantil již umožňuje hledat příhodnou definici vedoucí k jeho určení.

3. Mnohorozměrný kvantil jako funkce směru a vychýlenosti

V této části vyjdeme z analogie s regresními kvantily a dospějeme k obecné definici kvantilu v Hilbertově prostoru. Ukážeme, že kvantil je proces indexovaný prvky jednotkové koule (zde je budeme nazývat směry) a vychýleností, hodnotou z intervalu $[0,1]$. Vychýlenost 0 patří mediánu, extrémní kvantily mají vychýlenost blízkou 1.

Regresní kvantily jsou do hloubky rozpracovanou statistickou metodou. Problém kvantilů je zde řešen podmíněně, ale zároveň současně pro všechny podmínky. Pro naše úvahy je však podstatná ztrátová funkce vedoucí k odhadu kvantilu. Tak, jako střední hodnota minimalizuje kvadratickou ztrátovou funkci a medián absolutní odchylky, lze určit kvantil pomocí speciální ztrátové funkce. Pro nejjednodušší regresní model $Y = \beta_0 + \beta_1 X + \epsilon$ dostáváme například odhad regresního α-kvantilu řešením úlohy

$$\min_{a,b} \sum_{i=1}^{n} \rho(Y_i - a\ - bX_i); \quad \rho(z,\alpha) = \mathrm{I}[z<0](1-\alpha)|z| + \mathrm{I}[z>0]\alpha|z|. \tag{1}$$

Ztrátová funkce tedy připomíná nakloněnou funkci absolutní hodnoty. Hodnota α je číslo z oblasti $[0,1]$ a medián dostáváme pro $\alpha = 1/2$.

Chceme-li interpretovat kvantil jako vychýlenou míru polohy v kladném nebo záporném směru ($s = 1$ nebo $s = -1$), pak lze funkci ρ přepsat do tvaru

$$\rho(z,\alpha,s) = |z|\{\mathrm{I}[zs>0](1-2\alpha)+\alpha\}, \tag{2}$$

kde $\alpha \in [0,1/2]$. Hodnota $1-2\alpha \in [0,1]$ reprezentuje vychýlenost míry polohy.

3.1. **Dvourozměrná data.** Zde již začíná být zřejmý další krok. Proveďme jej nejprve pro dvourozměrný reálný prostor. Ztrátová funkce závislá na směru a vychýlenosti bude mít tvar nakloněného „trychtýře“ zobecňujíc tak ztrátovou funkci z jednorozměrného případu. Indikátor $\mathrm{I}[zs > 0]$ nám říkal, díváme-li se souhlasným, či protivným směrem ke směru s. V $\mathbb{R}^2$ však musíme rozlišovat více než dvě možnosti a proto zavedeme *směrovou funkci* $\gamma(z, s) : \mathbb{R}^2 \times \mathbb{R}^2 \to [0, 1]$. Na tuto funkci klademe následující základní podmínky pro daný směr $s \in \mathbb{S}_1 = \{x \in \mathbb{R}^2 : \|x\| = 1\}$

(1) Definujme $\gamma(0, s) = 1$.
(2) $\gamma(z, s) = \gamma(z/\|z\|, s)$, neboli směrová funkce nezávisí na velikosti z.
(3) $\gamma(-z, s) = 1 - \gamma(z, s)$, takže směrová funkce opačného vektoru je doplňkem směrové funkce vektoru do 1.
(4) $\gamma(s, s) = 1$, směrová funkce ve svém směru je 1.

Na směrovou funkci lze klást další doplňkové požadavky. Z nich nejdůležitější je, aby funkce $\gamma(z, s)$ byla nerostoucí se zvětšujícím se úhlem mezi z a s. Druhý doplňkový požadavek je invariance vůči rotaci, tedy $\gamma(r(z), r(s)) = \gamma(z, s)$, kde $r(\cdot)$ je rotací vektoru. Oba tyto doplňkové požadavky jsou natolik intuitivní, že v dále uvažovaných příkladech budou vždy splněny.

Nyní již můžeme napsat ztrátovou funkci vedoucí k definici kvantilu o daném směru a vychýlenosti. Touto funkcí je

$$\rho(z, \alpha, s) = \|z\|\{(1 - 2\alpha)\gamma(z, s) + \alpha\}, \tag{3}$$

kde opět $1 - 2\alpha$ je index necentrality, vychýlenost. Nyní definujeme α-kvantil pro náhodný výběr $X_1, \ldots, X_n$ ze dvourozměrného rozdělení jako

$$Q_{\alpha,s} = \arg\min_{u \in \mathbb{R}^2} \sum_{i=1}^{n} \rho(X_i - u, \alpha, s). \tag{4}$$

a jeho populační protějšek jako

$$q_{\alpha,s} = \arg\min_{u \in \mathbb{R}^2} \int_{\mathbb{R}^2} \rho(x - u, \alpha, s)\mathrm{P}(dx). \tag{5}$$

3.2. **Ještě více rozměrů.** Když se podíváme na předchozí definice, snadno zjistíme, že omezení prostorem $\mathbb{R}^2$ vlastně nepotřebujeme. Mluvíme-li o směru, stačí nám umět popsat jednotkovou sféru v daném prostoru. Pokud navíc umíme přiřadit dvěma prvkům „úhel“ mezi nimi, pak můžeme snadno volit i funkci směru γ. Oboje je velmi přirozeně splněno v Hilbertově prostoru H se skalárním součinem $\langle \cdot, \cdot \rangle$. Pak hranicí jednotkové koule je množina $\mathbb{S}_1 = \{s \in H : \langle s, s \rangle = 1\}$ a funkci $\gamma(z, s)$ lze volit jako funkci $\gamma(\langle z/\|z\|, s \rangle)$ závislou na skalárním součinu. Tím pádem se uvedená definice stává velmi univerzální. Vzhledem k aplikovatelnosti však budeme dále vesměs míti na paměti prostory $\mathbb{R}^d$.

3.3. **Alternativní zápis definice.** Uvedli jsme si, že mnohorozměrný kvantil je hodnota indexovaná směrem s (prvkem jednotkové sféry) a vychýleností (hodnotou $1 - 2\alpha$, kde $\alpha \in [0, 1/2]$). Lze však uvažovat i indexaci jediným indexem u, prvkem jednotkové koule. Libovolné $u, \|u\| \leq 1$ jednoznačně určuje α a s vztahy $\alpha = 1/2(1 - \|u\|)$ a $s = u/\|u\|$. Alternativní zápis ztrátové funkce pak je

$$\rho(z, u) = \|z\| \left\{ \|u\| \left(\gamma(z, u) - \frac{1}{2} \right) + \frac{1}{2} \right\}; \; z \in \mathbb{R}^d, u \in \mathbb{B}_1, \tag{6}$$

kde pro směrovou funkci platí $\gamma(z, u) = \gamma(z/\|z\|, u/\|u\|)$, $\mathbb{B}_1 = \{u \in H : \|u\| \leq 1\}$ je jednotková koule Hilbertova prostoru H.

4. Volba směrové funkce a funcke vzdálenosti

Ze zavedení ztrátové funkce γ předpisem (3) je zřejmé, že máme k dispozici dvě volné volby umožňující upravovat funkci γ. Ani jedna z těchto voleb neovlivní ztrátovou funkci v případě jednorozměrného kvantilu, tedy ani regresního kvantilu, a proto se s nimi setkáváme až nyní. Omezme se na prostory $\mathbb{R}^d$.

První otázkou je volba normy $\|z\|$. Na první pohled vypadá zřejmá volba L_1 normy, neboť ta odpovídá absolutní vzdálenosti. Zde je dobré si uvědomit, že i L_2, L_∞ a ostatní normy při zúžení na jednorozměrný prostor vedou k absolutní vzdálenosti. Všechny tyto normy tedy můžeme uvažovat jako zobecnění absolutní hodnoty.

Volba směrové funkce je asi nejdůležitějším krokem pro mnohorozměrné kvantily. V reálných prostorech máme širokou nabídku možností pomocí skalárních součinů. Snadno přijdeme na volbu

$$\gamma_S(z,s) = \frac{1}{2}\left\langle \frac{z}{\|z\|}, s \right\rangle;\ s\ \in \mathbb{S}_1, z\ \in \mathbb{R}^d \setminus \{0\}. \tag{7}$$

Tato speciální volba směrové funkce byla použita ve článku [1] k definici geometrického kvantilu. Pro mnohorozměrná data se jedná o směrovou funkci, která klesá při zvětšujícím se úhlu mezi předepsaným směrem s a směrem $z/\|z\|$ jako kosinus stažený do intervalu $[0,1]$. Nazvěme tuto základní volbu jako *skalární*. Jinou volbou je *lineární* směrová funkce. Zde volíme

$$\gamma_L(z,s) = \frac{1}{\pi}\left(\pi - \arccos\left\langle \frac{z}{\|z\|}, s \right\rangle\right);\ s\ \in \mathbb{S}_1, z\ \in \mathbb{R}^d \setminus \{0\}. \tag{8}$$

Je mnoho dalších možností jak volit směrovou funkci. Uveďme několik dalších definic, všechny pro $z \in \mathbb{R}^d \setminus \{0\}$ a $s \in \mathbb{S}_1$.

$$\begin{aligned}
\gamma_A(z,s) &= 2\gamma_L(z,s) - \gamma_S(z,s)\\
\gamma_l(z,s) &= \mathrm{I}\left[\langle z,s\rangle > \sqrt{2}/2\right]\{2\gamma_L(z,s) - 1\} + 1/2\mathrm{I}\left[|\langle z,s\rangle| \le \sqrt{2}/2\right] +\\
&\quad + \mathrm{I}\left[\langle z,s\rangle < -\sqrt{2}/2\right] 2\gamma_L(z,s)\\
\gamma_2(z,s) &= \mathrm{I}[\langle z,s\rangle > 0] + \frac{1}{2}\mathrm{I}[\langle z,s\rangle = 0]\\
\gamma_{3a}(z,s) &= \mathrm{I}[\langle z,s\rangle > a] + \frac{1}{2}\mathrm{I}[|\langle z,s\rangle| \le a]
\end{aligned} \tag{9}$$

Směrová funkce opravdu velmi výrazně ovlivní polohu kvantilu. Není však jasné, má-li některá volba přednost před ostatními.

5. Příklady

Zkusme si alespoň na nejjednodušším příkladu ukázat, jak vypadají teoretické kvantily dvourozměrného rozdělení. Dále si ukážeme kvantily určené ze simulovaných normálně a rovnoměrně rozložených dat.

Uvažujme dvourozměrné rozdělení s hustotou $f(x,y)$, kde nosič rozdělení je $\mathbb{R}^2$. Za směrovou funkci zvolme γ_2 ze vzorce (9) s možnými hodnotami 0, 1/2 a 1. Všimněme si, že hodnoty 1/2 je nabýváno pouze na jedné přímce, tedy množině nulové míry a proto ji nyní lze pominout. Jako funkci vzdálenosti zde volme L_1 normu. Směr kvantilu budeme volit $(1,0)$, tedy podél osy x. Kvantil s vychýleností $1-2\alpha$ ve směru $(1,0)$ je vektor (a,b) minimalizující

(10)
$$\alpha \int_{-\infty}^{a} \int_{-\infty}^{\infty} (a - x + |y - b|) f(x, y) dy dx + (1 - \alpha) \int_{a}^{\infty} \int_{-\infty}^{\infty} (x - a + |y - b|) f(x, y) dy dx.$$

Vektory $(x - a, y - b)$ s kladnou první složkou mají směrovou funkci 1, tyto vektory se zápornou první složkou mají směrovou funkci 0. Derivací uvedené ztrátové funkce podle a a podle b dostaneme dvojici rovnic

$$\begin{aligned} 0 &= (2\alpha - 1) \int_{-\infty}^{\infty} |y - b| f(a, y) dy + F_X(a) - 1 + \alpha \\ 0 &= (2\alpha - 1)[2F(a, b) - F_X(a)] + (1 - \alpha)[2F_Y(b) - 1] \end{aligned} \tag{11}$$

a tuto řešíme pro a a b. Ve speciálním případě nezávislosti dostáváme snadno, že b musí být mediánem marginálního rozdělení F_Y. Stejně tak snadno dostaneme i dvourozměrný medián $(\tilde{X}, \tilde{Y})$ volbou $\alpha = 1/2$ v tomto speciálním případě.

Z definice ztrátové funkce ρ též okamžitě vidíme, že mnohorozměrný medián závisí na volbě vzdálenosti, nikoliv na volbě směrové funkce. Toto triviální zjištění je plně v souladu s naší představou kladenou na mnohorozměrný medián.

Na obrázcích v dodatku vidíme možné volby ztrátových funkcí a vybrané výsledky simulací. Data byla simulována a posléze zpracována v programu Matlab. Při optimalizačních úlohách nebyla použita žádná speciální volba přesnosti, ponechali jsme základní nastavení funkce `fminsearch`. Na první pohled je zřetelné, že při některých volbách směrové funkce (skalární) dostáváme oblasti podobné kruhům, zatímco pro jiné volby (antiskalární – γ_A) dostáváme oblasti více závislé na původním tvaru dat. Výsledky nespojitých směrových funkcí jsou při zvolené základní přesnosti výpočtu neuspokojvé. Rozdíly si vysvětlujeme malou, případně naopak velkou, změnou penalizace vzdálenosti při malé odchylce od směru kvantilu.

6. Aplikace a další možná práce

Domníváme se, že navržená definice necentrální míry polohy pro mnohorozměrná data má široké použití. Na druhou stranu nelze popřít, že výpočet kvantilů z dat je sice možný, ale při absenci teoretických výsledků nelze dostatečně kvalifikovaně interpretovat získané odhady.

Z možných aplikací naznačme například data o výšce a hmotnosti dospělých mužů. Muž s hmotností přesahující nějaký kvantil by při současném posouzení hmotnosti a výšky nemusel vybočovat z normálu, neboli jeho hmotnost a výška by byla uvnitř dvourozměrného intervalu spolehlivosti. Lze jistě namítnout, že podobný závěr lze učinit i na základě regresní analýzy a regresních kvantilů. Zde je ovšem výhodou našeho postupu jeho nezávislost na souřadnicích, neboť v regresní analýze nedostaneme stejný výsledek pro podmínku danou hmotností a pro podmínku danou výškou. Podobně pro růst dětí je možné stanovit, zda průběh růstu dítěte kontrolovaný v několika význačných časech je či není výjimečný na základě celkového porovnání výšek v kontrolních bodech, místo na základě překročení nějaké kritické hodnoty v jediném časovém okamžiku.

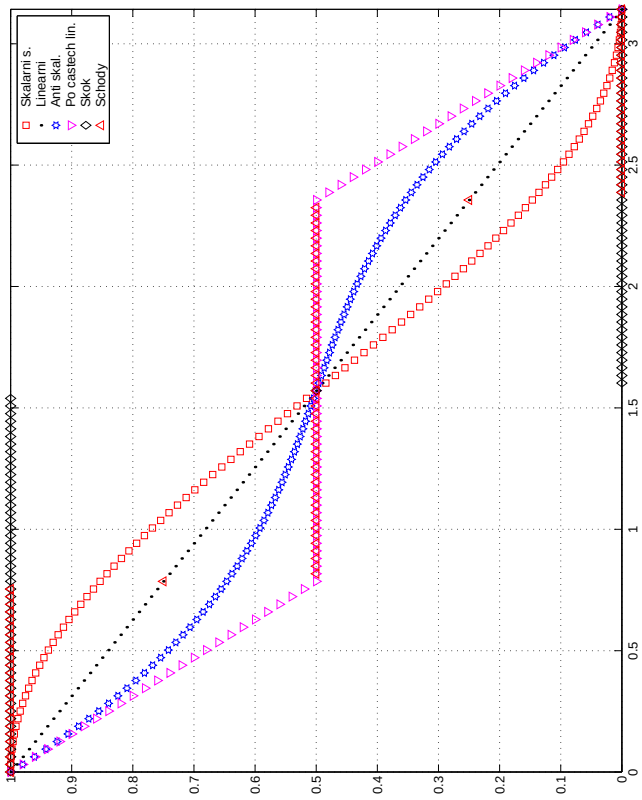
Otevřenou otázkou je vztah geometrie dat a volby kvantilu. Například pro data rovnoměrně rozložená na mezikruží $0 < a \leq x^2 + y^2 \leq b$ je definice kvantilu problematická, neboť centrální kvantily budou obsahovat 0 coby medián, ale nemusí obsahovat jediný bod nosiče. Tento příklad sice může být poněkud umělý, ale v zásadě ilustruje meze mnohorozměrných oblastí spolehlivosti (pro úplnost dodejme,

že i těch založených na parametrických modelech). Dobrý význam má mnohorozměrný kvantil pro konvexní nosiče (v přirozeně zobecněném smyslu při diskrétním rozdělení), případně ještě pro nosiče hvězdicovitého tvaru. U jiných rozdělení se nabízí myšlenka transformace, výpočtu kvantilu a zpětné transformace, ovšem tento postup nemusí být již vůbec možný.

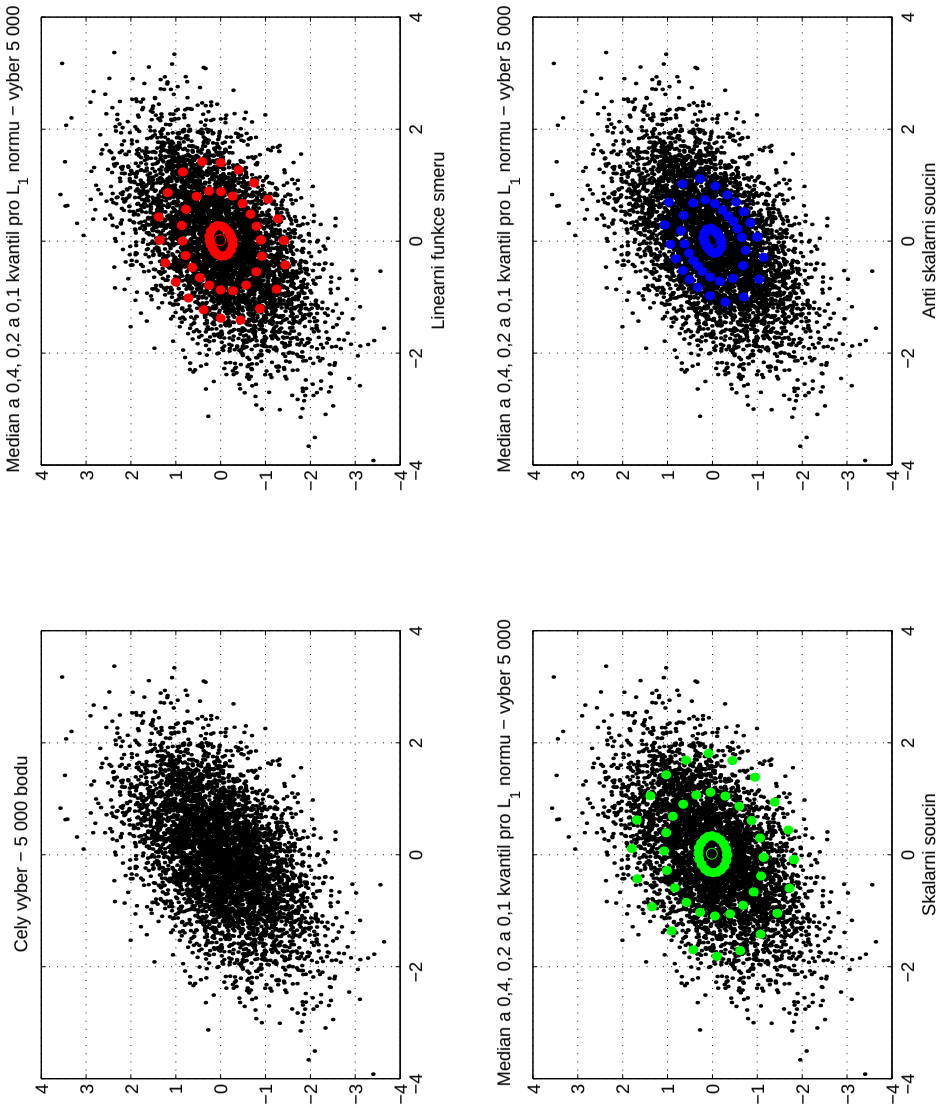
Poslední poznámkou je volba vychýlenosti. Při rovnoměrném rozdělení na intervalu $[-1,1]$ pokrývá interval $[-0,8,0,8]$ přesně 80% nosiče. Pro jednotkovou kouli $\mathbb{B}_1 \subset \mathbb{R}^2$ však koule o poloměru 0,8 pokrývá jen 64% nosiče. K pokrytí 80% potřebujeme kouli o poloměru 0,894, pro pokrytí 95% pak kouli o poloměru 0,975. Pro třírozměrný prostor se požadované poloměry ještě zvětší. Z toho vidíme, že při rostoucím rozměru prostoru se kvantily „vytlačují“ směrem od centra dat a pro spolehlivý odhad extrémního kvantilu značně narůstá potřebný počet pozorování, zejména v porovnání s jednorozměrnými daty.

Poznámka: Uvedené a několik dalších obrázků je možno najít ve formátu eps na autorových stránkách `http://www.karlin.mff.cuni.cz/ hlubinka/robust02`.

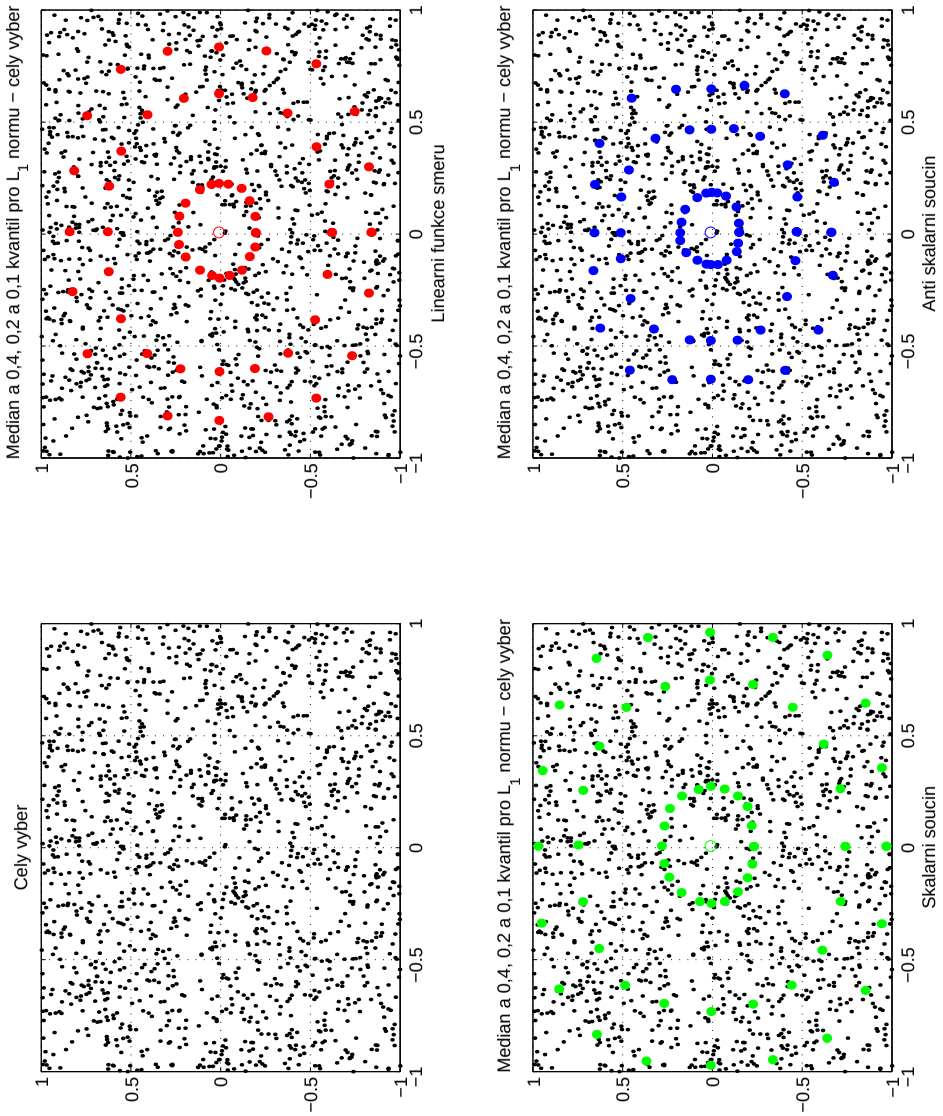
7. Dodatek



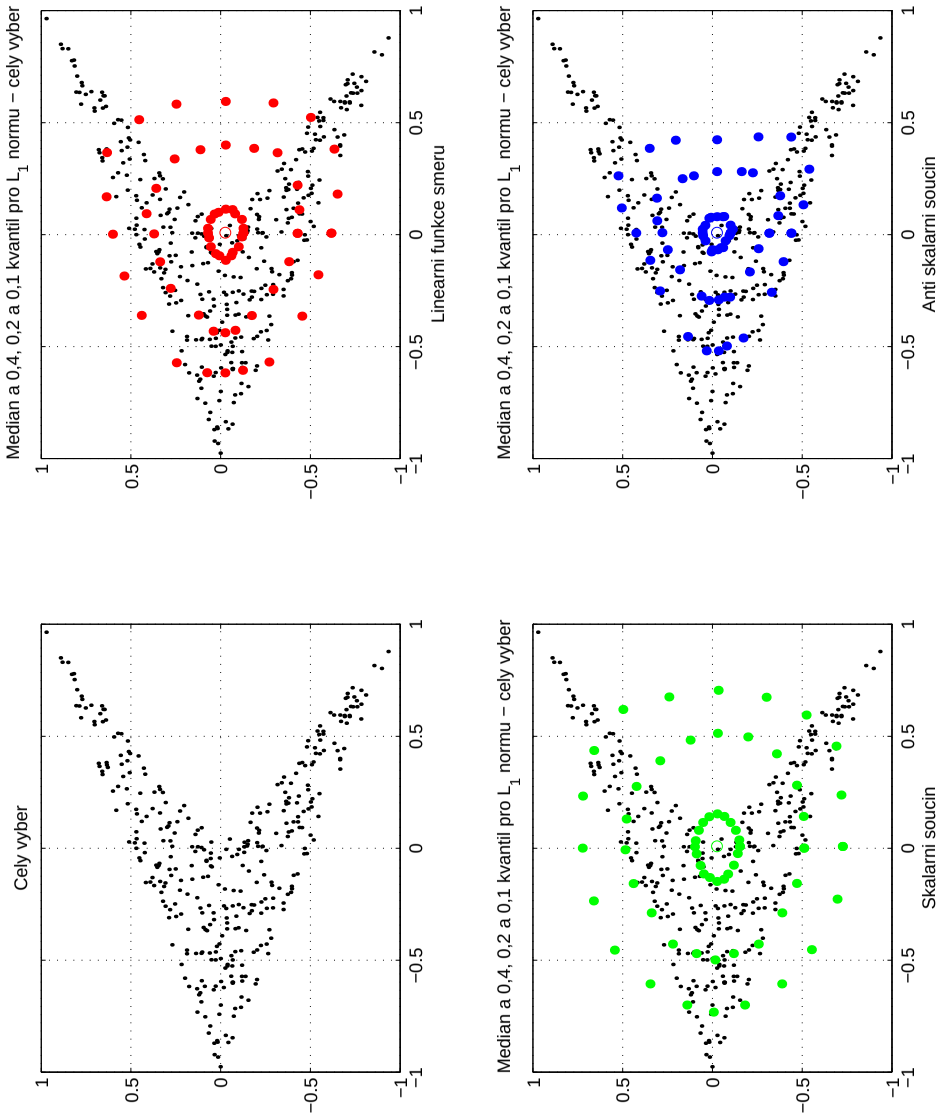
Obr. 1 *Na obrázku vidíme postupně (v pořadí podle legendy) funkce směru* γ_S, γ_L, γ_A, γ_l, γ_2 *a* γ_{3a}.



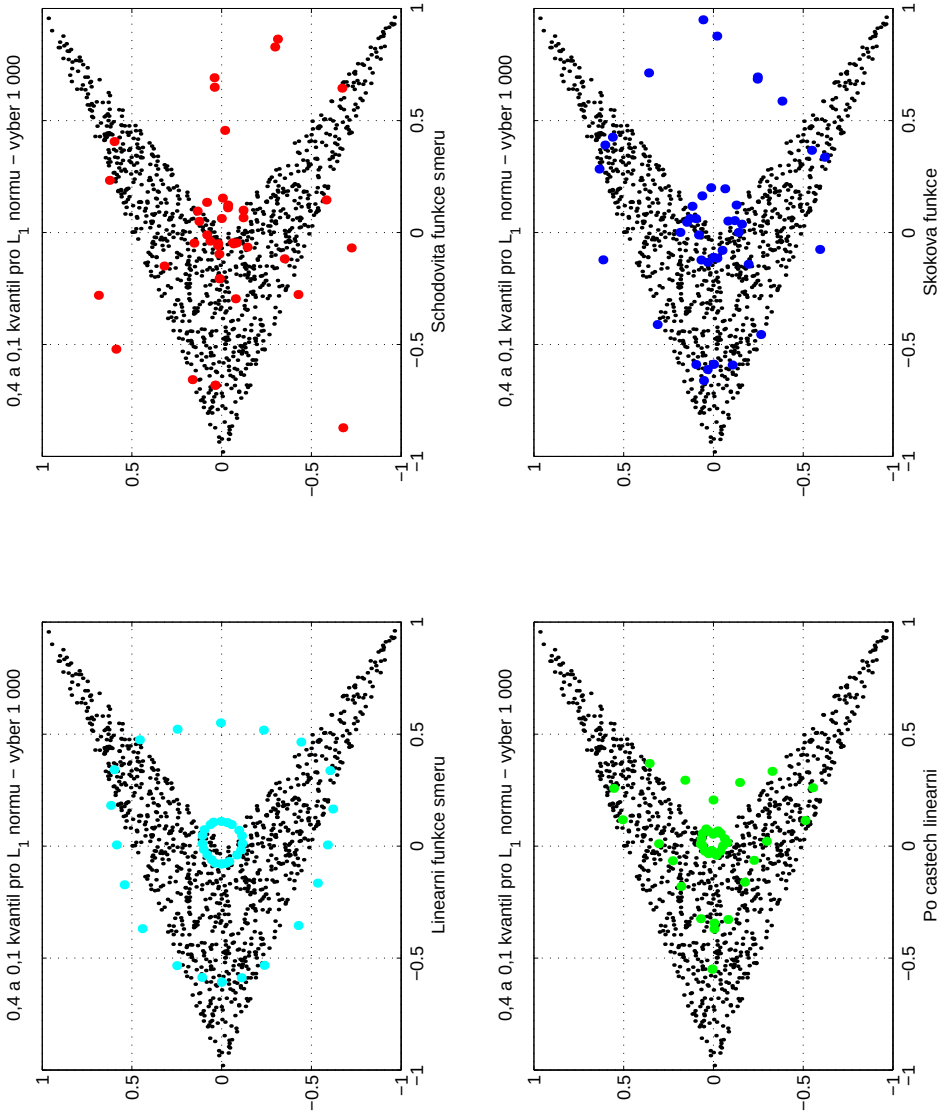
Obr. 2 *Výběrové kvantily pro simulovaný náhodný výběr z dvourozměrného normálního rozdělení s korelačním koeficientem* $1/2$.



Obr. 3 *Výběrové kvantily pro simulovaný náhodný výběr z rovnoměrného rozdělení na čtverci.*



Obr. 4 *Výběrové kvantily pro simulovaný náhodný výběr z rovnoměrného rozdělení na nehladké konkávní oblasti.*



Obr. 5 *Při použití běžných procedur programu MATLAB s obvyklými kritérii optimality dostáváme pro nespojité funkce směru velmi neuspokojivé výsledky.*

Literatura

[1] Chaudhuri, P. (1996): On a Geometric Notion of Quantiles for Multivariate Data, *Jour. Am. Stat. Assoc.*, **91**.434, pp. 862–872.

[2] Rousseeuw, P.J., Ruts, I., Tukey, J.W. (1999): The Bagplot: A Bivariate Boxplot, *The American Stat.*, **53**.4, pp 382–387.

UK MFF, KPMS, Sokolovská 83, 186 75, Praha 8
E-mail: daniel.hlubinka@mff.cuni.cz