SMĚROVÉ KVANTILY PRO VÍCEROZMĚRNÁ DATA

Lukáš Kotík
kotik@utia.cas.cz
Ústav teorie informace a automatizace, AV ČR, v. v. i.
Oddělení stochastické informatiky

Abstrakt: *Máme-li jednoznačně určený parametr polohy (střed) pro vícerozměrná data, můžeme přejít k hypersférickým souřadnicím s počátkem v tomto středu. Směrový kvantil v daném směru pak definujeme jako jednorozměrný kvantil rozdělení rádiusu podmíněného úhly, které určují směr ze středu. To můžeme provést pro každý směr a přechodem zpět do kartézských souřadnic získáme konfidenční region. Pro určení středu zpravidla použijeme nějaké vícerozměrné rozšíření jednorozměrného mediánu (např. založeného na hloubce). Výběrový směrový kvantil můžeme odhadnout po přetransformování výběru do hypersférických souřadnic pomocí kvantilové regrese nebo jádrového vyhlazování. Příspěvek je zaměřen na základní vlastnosti konfidenčních množin získaných pomocí směrových kvantilů a na možnosti jejich odhadu.*

SMĚROVÝ KVANTIL PRO ABS. SPOJITÁ ROZDĚLENÍ

Směrový kvantil se poprvé objevil v [1] a později v [2]. Postup je založen na určení jednorozměrných kvantilů na polopřímkách začínajících v jednom bodě. Prakticky to bude provedeno přechodem k hypersférickým souřadnicím, jednotlivé polopřímky budou určeny úhly v těchto souřadnicích. K tomu ale nejdříve potřebujeme určit bod $\boldsymbol{\theta}$, v kterém budou tyto polopřímky začínat. Měl by ležet v nějakém smyslu „co nejvíce ve středu dat“. Uveďme si nejprve možnosti volby tohoto bodu.

Nechť $\boldsymbol{X} \in \mathbb{R}^k$ je náhodný vektor s absolutně spojitým rozdělením s hustotou f.

Parametry polohy

Volba bodu $\boldsymbol{\theta}$ bude mít vliv i na vlastnosti směrových kvantilů. Je ho proto třeba volit s rozvahou. Nejvhodnější se jeví body získané pomocí hloubky. Např. vážená poloprostorová hloubka pro váhu, která zaručí optimální rozvržení pravděpodobnosti ve všech „směrech“ od nejhlubšího bodu. Nebo např. nejhlubší bod pro hloubkovou funkci: $D(\boldsymbol{a}) = \inf\{\mathrm{P}(\boldsymbol{X} \in K) : \ K \in \mathcal{K}_{\boldsymbol{a}}\}$, kde $\mathcal{K}_{\boldsymbol{a}}$ je množina všech kuželů s vrcholem v bodě $\boldsymbol{a}$ a předem daným vrcholovým úhlem .

MOŽNOSTI VOLBY STŘEDU:

Nejhlubší bod - hloubky: *poloprostorová, vážená poloprostorová, L_1, simplexová, ...*, viz např. [3]; Medián po složkách, Maximálně věrohodný bod (modus), Střední hodnota, ...

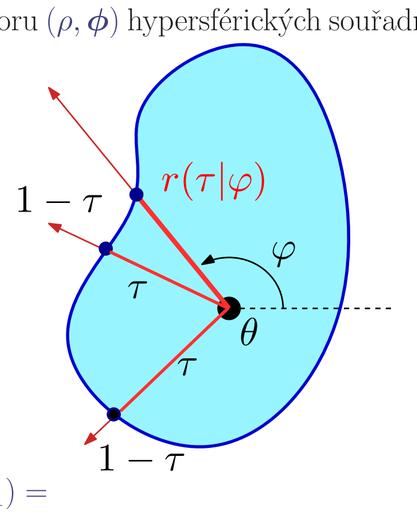
Transformace

Pomocí věty o transformaci přejdeme k náhodnému vektoru $(\rho, \boldsymbol{\phi})$ hypersférických souřadnic.

HYPERSFÉRICKÉ SOUŘADNICE:

$$\begin{aligned}
x_1 &= \theta_1 + r\sin\varphi_1\sin\varphi_2\cdots\sin\varphi_{k-2}\sin\varphi_{k-1},\\
x_2 &= \theta_2 + r\sin\varphi_1\sin\varphi_2\cdots\sin\varphi_{k-2}\cos\varphi_{k-1},\\
&\vdots\\
x_{k-1} &= \theta_{k-1} + r\sin\varphi_1\cos\varphi_2,\\
x_k &= \theta_k + r\cos\varphi_1,
\end{aligned}$$

$\varphi_i \in (0,\pi),\ i = 1,\dots,k-2,$
$\varphi_{k-1} \in (0,2\pi),\ r > 0$



Hustota po transformaci bude: $p(r,\varphi_1,\dots,\varphi_{k-1}) =$

$$r^{k-1}\,|\sin^{k-2}\varphi_1\cdots\sin\varphi_{k-2}|\,f(\theta_1 + r\sin\varphi_1\sin\varphi_2\cdots\sin\varphi_{k-2}\sin\varphi_{k-1},\dots,\theta_k + r\cos\varphi_1).$$

Odsud snadno určíme podmíněnou hustotu vzdálenosti od středu při dané hodnotě vektoru úhlů určujících směr ze středu.

$$q(r|\boldsymbol{\varphi}) = \begin{cases} \dfrac{p(r,\boldsymbol{\varphi})}{s(\boldsymbol{\varphi})} & \text{pro } s(\boldsymbol{\varphi}) \neq 0,\\ 0 & \text{pro } s(\boldsymbol{\varphi}) = 0, \end{cases} \qquad \text{kde } s(\boldsymbol{\varphi}) = \int_0^{+\infty} p(r,\boldsymbol{\varphi})\,\mathrm{d}r.$$

Pomocí podmíněné hustoty můžeme definovat směrový kvantil jako jednorozměrný kvantil rozdělení vzdálenosti od středu podmíněného volbou úhlů.

SMĚROVÝ KVANTIL:

Pro zvolený směr (daný vektor úhlových veličin $\boldsymbol{\varphi}$) definujeme τ-směrový kvantil $r(\tau|\boldsymbol{\varphi})$ jako řešení vyhovující

$$\tau = \mathrm{P}(\rho \le r(\tau|\boldsymbol{\varphi})|\boldsymbol{\phi} = \boldsymbol{\varphi}) = \int_0^{r(\tau|\boldsymbol{\varphi})} q(r|\boldsymbol{\varphi})\mathrm{d}r.$$

Vlastnosti

Předpokládejme, že nosič hustoty $\mathcal{M} = \{\boldsymbol{x} : \ f(\boldsymbol{x}) > 0\}$ je hvězdicovitý kolem bodu $\boldsymbol{\theta}$, který představuje střed a je jejím vnitřním bodem. Dále, že $\mathsf{int}(\mathcal{M})$ je souvislá a že průnik libovolné přímky procházející bodem $\boldsymbol{\theta}$ s $\partial\mathcal{M}$ má nulovou délku. Označme $\mathcal{K}_{\boldsymbol{X}}(\tau)$ „výplň“ množiny, která vznikne transformací funkce $r(\tau|\cdot)$ zpět do kartézských souřadnic (viz obr. 1 dole) pro náhodný vektor $\boldsymbol{X}$.

- $r(\tau|\boldsymbol{\varphi})$ je spojitou funkcí proměnné $\boldsymbol{\varphi}$.

  Dk: *Lebesgueova věta zaručuje spojitost funkce $\boldsymbol{\varphi} \mapsto \int_0^u q(r|\boldsymbol{\varphi})\,\mathrm{d}r$ pro všechna u. To je postačující podmínka pro lemma 8.3.1 v [4], které zaručuje spojitost $r(\tau|\cdot)$.*

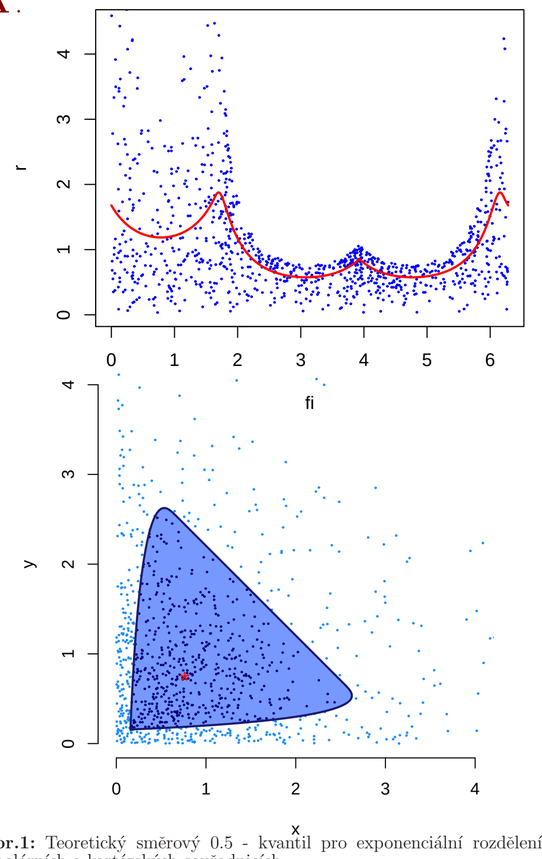
- Ekvivariance vůči afinním transformacím, pokud je střed ekvivariantní vůči af. transf. (např. Nejhlubší bod pro poloprostorovou, ekvivariantní váženou poloprostorovou, simplexovou hloubku), tj.

  $$\mathcal{K}_{\mathbf{A}\boldsymbol{X}}(\tau) = \mathbf{A}\mathcal{K}_{\boldsymbol{X}}(\tau).$$

  Dk: *Afinní transformací lib. přímka přechází na přímku a rozdělení na ní je jen „natažené“, či „zúžené“ (ve smyslu vynásobení konstantou) oproti původnímu rozdělení.*

- $\mathrm{P}(\boldsymbol{X} \in \mathcal{K}(\tau)) =$

  $$\int_0^{2\pi}\int_0^{\pi}\cdots\int_0^{\pi}\int_0^{r(\tau|\boldsymbol{\varphi})} q(r|\boldsymbol{\varphi})s(\boldsymbol{\varphi})\,\mathrm{d}r\,\mathrm{d}\varphi_1 \ldots \mathrm{d}\varphi_{k-1} = \tau.$$



Obr.1: Teoretický směrový 0.5 - kvantil pro exponenciální rozdělení v polárních a kartézských souřadnicích.

ODHADY

Nechť $\boldsymbol{X}_i,\ i = 1,\dots,n$ je k-rozměrný náhodný výběr. Tento výběr v hypersférických souřadnicích označme $(R_i, \boldsymbol{F}_i),\ i = 1,\dots,n$.

Odhad pomocí Fourierových řad

$r(\tau|\cdot)$ je periodická funkce. Za dosti obecných podmínek je funkcí spojitou, lebesgueovsky integrovatelnou na omezených intervalech a po částech hladkou. Pak existuje Fourierův rozvoj této funkce, který k ní konverguje stejnoměrně. V **dvourozměrném případě** můžeme tento rozvoj odhadnout trigonometrickou řadou

$$r_p(\tau|\varphi) = a_0 + \sum_{j=1}^{p}(a_j\cos j\varphi + b_j\sin j\varphi),$$

jejíž koeficienty získáme pomocí kvantilové regrese.

ODEZVA A REGRESORY PRO KVANTILOVOU REGRESI:

$R_i, \quad 1,\ \cos F_i,\ \cos 2F_i,\ \dots,\ \cos pF_i,\ \sin F_i,\ \sin 2F_i,\ \dots,\ \sin pF_i, \quad i = 1,\dots,n$

Ve **vícerozměrném případě** je situace trochu složitější. Postup ukážeme pro třírozměrný výběr. K odhadu použijeme řadu

$$\sum_{m=0}^{p}\sum_{n=0}^{q}(a_{mn}\cos m\varphi_1\cos n\varphi_2 + b_{mn}\cos m\varphi_1\sin n\varphi_2 + c_{mn}\sin m\varphi_1\cos n\varphi_2 + d_{mn}\sin m\varphi_1\sin n\varphi_2).$$

Na její koeficienty musíme ale klást dodatečné podmínky. Bod se sférickými souřadnicemi $(r, \pi+\xi, \varphi_2)$ má po přechodu do kartézských souřadnic stejné koordináty jako bod $(r, \pi-\xi, \varphi_2+\pi)$ pro $\xi \in (0,\pi)$. Dále body se sférickými souřadnicemi $(r, 0, \varphi_2),\ \varphi_2 \in [0,\pi)$ určují v kartézských souřadnicích právě jeden bod bez ohledu na hodnotu φ_2. Stejně tak (r, π, φ_2). Jedná se o body na ose x_3. Z toho dostáváme **vlastnosti $r(\tau|\varphi_1,\varphi_2)$, které musíme brát na zřetel při většině odhadů pro tří a vícerozměrné výběry!** Odsud, výpočtem Fourierových koeficientů (viz [2]), získáme následující vztahy. Regresory pro kvantilovou regresy pak získáme snadno.

VLASTNOSTI $r(\tau|\varphi_1,\varphi_2)$:

- $r(\tau|\varphi_1,\varphi_2) = r(\tau|2\pi-\varphi_1,\pi+\varphi_2)$
- $r(\tau|0,\varphi_2) = konst, \quad \forall\varphi_2$
- $r(\tau|\pi,\varphi_2) = konst, \quad \forall\varphi_2$

POŽADAVKY NA KOEFICIENTY ODHADU:

$$a_{mn} = b_{mn} = \begin{cases} 0, & \text{pro } n \text{ liché}\\ \neq 0, & \text{pro } n \text{ sudé a rovné } 0 \end{cases} \qquad c_{mn} = d_{mn} = \begin{cases} 0, & \text{pro } n \text{ sudé a rovné } 0\\ \neq 0, & \text{pro } n \text{ liché} \end{cases}$$

$$\sum_{m=0}^{p} a_{mn} = \sum_{m=0}^{p} b_{mn} = \sum_{m=0}^{p}(-1)^m a_{mn} = \sum_{m=0}^{p}(-1)^m b_{mn} = 0$$

S rostoucím rozsahem výběru je vhodné zvyšovat i řád rozvoje p. Označme $\mathcal{K}_n(\tau)$ odhad množiny $\mathcal{K}(\tau)$.

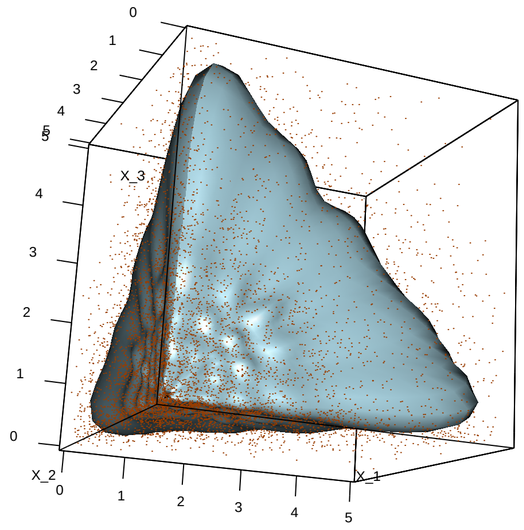
VLASTNOSTI ODHADU:

- Přibližně $n\tau$ bodů leží v $\mathcal{K}_n(\tau)$.

  Dk: *Plyne z vlastnosti kvantilové regrese: počet bodů pod regresní plochou se dá omezit zespoda hodnotou ($n\tau$ - počet regresorů) a zeshora $n\tau$.*

- Ekvivariance vzhledem k posunutí, stejné změně měřítka všech os a otočení kolem osy x_k (v úhlové veličině φ_{k-1}).

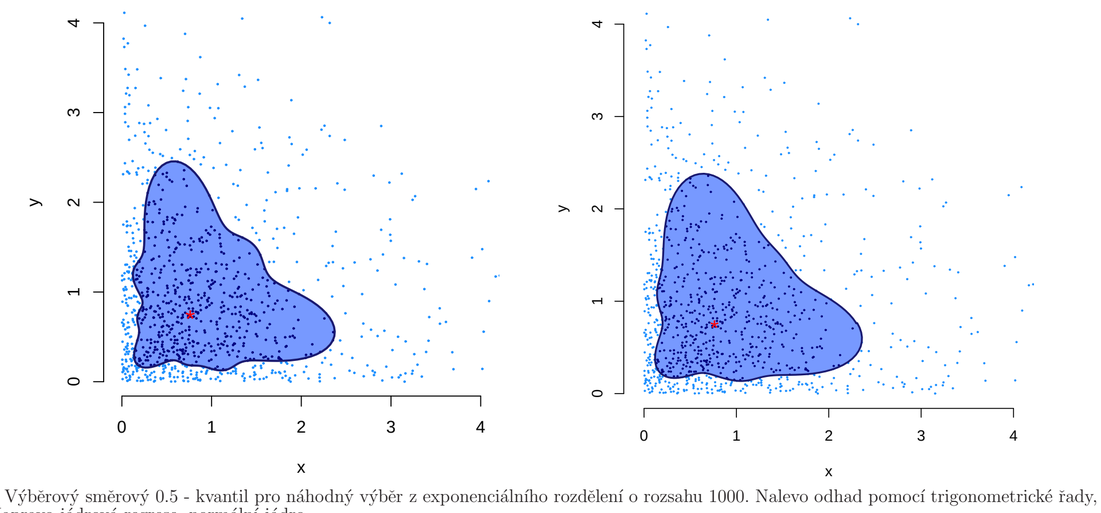
  Dk: *Dosazením souřadnic otočeného výběru do $r_p(\tau|\boldsymbol{\varphi})$, použitím součtových vzorců, a vhodným přepsáním minimalizační úlohy v kvantilové regresy.*



Obr. 2: Výběrový směrový 0.75 - kvantil pro exponenciální rozdělení a rozsah 50000. Řády $p = q = 25$.

Jádrová (neparametrická) kvantilová regrese

Další možností je odhadnout funkci $r(\tau|\cdot)$ pomocí neparametrické regrese. Ve dvourozměrném případě stačí zkopírovat část nebo celý výběr ležící před mezí 2π před bod 0 (tj. k výběru vyjádřenému v polárních přidáme výběr $(R_i, F_i - 2\pi),\ i = 1,\dots,n$). Analogicky postupujeme za mezí 2π. Tím si při rozumné volbě jádra zaručíme, že bude splněno $r(\tau|0) = r(\tau|2\pi)$. Ve vícerozměrném případě opět nastává problém vyplývající z vlastností funkce $r(\tau|\cdot)$ (resp. z vlastností polárních souřadnic). Řešení tohoto problému může být předmětem budoucí práce.



Obr. 3: Výběrový směrový 0.5 - kvantil pro náhodný výběr z exponenciálního rozdělení o rozsahu 1000. Nalevo odhad pomocí trigonometrické řady, $p = 9$. Napravo jádrová regrese, normální jádro.

Další možnosti

Je možné zvolit obdobný postup jako u odhadu pomocí trigonometrických řad, stačí místo báze trigonometrických funkcí zvolit bázi funkcí (např. splajnovou), která zajistí, že u odhadu budou splněny vlastnosti funkce $r(\tau|\cdot)$ (periodicita, ...). Další možností, jak získat výběrové směrové kvantily je při výpočtu podle definice nahradit hustotu f nějakým jejím odhadem.

Poznámka o hloubce

Máme-li určený bod, který budeme považovat za nejhlubší, pak je možné definovat směrovou hloubku bodu s hypersférickými souřadnicemi $(r_0, \boldsymbol{\varphi}_0)$ jako $\mathrm{P}(\rho > r_0|\boldsymbol{\phi} = \boldsymbol{\varphi}_0)$.

Poděkování. Autor děkuje doc. Danielovi Hlubinkovi, jenž má nemalý podíl na formování myšlenek obsažených na tomto posteru. Poster je financován granty AV0Z10750506 a MSM0021620839.

Literatura.

[1] Hasil J. (2004), *Problém kvantilu ve více rozměrech*, Diplomová práce, MFF UK, Praha.
[2] Kotík L. (2007), *Periodické regresní kvantily*, Diplomová práce, MFF UK, Praha.
[3] Liu R. Y., Serfling R., Souvaine D. L. (2006). *DIMACS; Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*, American Mathematical Society.
[4] Resnick S. I. (1998), *A probability path*, Boston: Birkhauser.