2008 © JČMF

Kvantily v sekvenční analýze bodu změny

Ondřej Chochola

chochola@karlin.mff.cuni.cz

Katedra pravděpodobnosti a matematické statistiky,
Matematicko-fyzikální fakulta Univerzity Karlovy v Praze

Poster prezentuje možné využití kvantilů v detekci bodu změny v sekvenčně pozorovaných datech. Uvažují se jak výběrové kvantily v modelu polohy, tak i regresní kvantily. Teoretická část je modifikací práce [1], která je doplněna výsledky simulační studie.

ÚVOD A MOTIVACE

Práce vychází ze schematu navrženém v [2] a rozpracovaném v [1] na tzv. L_1 odhady parametru polohy a regresního parametru. Zde se zaměřujeme na testování změny kvantilu distribuce pozorovaní, případně na změnu regresních kvantilů.

Zkoumání kvantilů bylo motivováno možnou aplikací na testování změny Value at Risk (VaR). VaR se často používá v ekonometrii jako míra nejrůznějších rizik, především pak tržního rizika daného portfolia. VaR odpovídá ztrátě, kterou portfolio s danou pravděpodobností nepřekročí v daném časovém úseku. Z matematického hlediska je tedy kvantilem distribuce výnosů.

MODEL POLOHY

Uvažujeme sekvenčně přicházející data splňující model polohy

$$Y_i = \mu_i + e_i, \tag{1}$$

kde Y_i jsou pozorovaná data, μ_i je parametr polohy a e_i jsou iid náhodné chyby s distribuční funkcí F specifikovanou níže. Změna τ kvantilu t_i^τ pozorování Y_i je tedy způsobena změnou parametru polohy z hodnoty μ_0 na μ_*.

Můžeme však uvažovat i model

$$Y_i = \mu + e_i, \tag{2}$$

kde e_i jsou také nezávislé, avšak jejich distribuční funkce se může v neznámém čase změnit z F na F_1. Kvantily jsou proto ovlivněny změnou celé distribuce, tedy model (1) je zřejmě speciálním případem modelu (2).

O distribučních funkcích předpokládáme, že jsou spojité, přičemž nosič hustoty je interval a na jeho vnitřku existuje její derivace. U F předpokládáme navíc symetrii kolem 0, neboť se jedná o distribuční funkci chyb.

Dále předpokládáme, že máme k dispozici historická (tréninková) data beze změny o délce m (tzv. podmínka stability), tj. $t_1^\tau = \ldots = t_m^\tau = t_0^\tau$.

Testujeme tedy nulovou hypotézu, že se τ kvantil pozorování nemění:

$$H_0 : t_i^\tau = t_0^\tau, \quad 1 \le i < \infty \tag{3}$$

proti alternativě, že se v neznámém bodě $m + k^*$ změní:

$$H_1 : \text{existuje } k^* \ge 1 \text{ takové, že } t_i^\tau = t_0^\tau,\ 1 \le i < m + k^*,\ t_i^\tau = t_*^\tau,\ m + k^* \le i < \infty, \quad t_0^\tau \ne t_*^\tau. \tag{4}$$

Testová statistika

Testová statistika využívá analogie L_1 reziduí. Definujeme

$$\tilde\varepsilon_i = \tau - I[Y_i < \tilde t_m^\tau],$$

kde $I[\ldots]$ značí indikátor jevu a $\tilde t_m^\tau$ je výběrový τ kvantil historických pozorování získaný minimalizací $\min_a \sum_{i=1}^m \rho_\tau(Y_i - a)$ pro $\rho_\tau(u) = u(\tau - I[u < 0])$.

Populační analogie je $\varepsilon_i = \tau - I[Y_i < t_0^\tau]$, pro které za H_0 platí: $\mathrm{E}\,\varepsilon_i = 0,\ \mathrm{var}\,\varepsilon_i = \tau(1-\tau)$. Proto velké hodnoty testové statistiky

$$Q(m,k) = \frac{1}{\sqrt{\tau(1-\tau)}} \sum_{i=m+1}^{m+k} \tilde\varepsilon_i$$

zřejmě vedou k zamítnutí nulové hypotézy. Statistika je normována obvyklou hraniční funkcí

$$g(m,k,\gamma) = \sqrt{m} \left(\frac{m+k}{m}\right) \left(\frac{k}{m+k}\right)^\gamma$$

s dolaďující konstantou $\gamma \in [0, 1/2)$.

Pro asymptotické chování statistiky platí stejné vztahy jako pro statistiky v [2] či [1]. Za H_0 platí (5)

$$\lim_{m\to\infty} P\left(\sup_{1\le k<\infty} \frac{|Q(m,k)|}{g(m,k,\gamma)} \le c\right) = P\left(\sup_{0\le t\le 1} \frac{|W(t)|}{t^\gamma} \le c\right)$$

pro všechny $c > 0$, kde $\{W(t),\ t \in [0,1]\}$ je Wienerův proces. Tento vztah nám umožňuje aproximovat kritické hodnoty pro testovací proceduru. Za H_1 platí

$$\sup_{1\le k<\infty} \frac{|Q(m,k)|}{g(m,k,\gamma)} \xrightarrow{P} \infty, \quad m \to \infty, \tag{6}$$

což zaručuje, že nastalá změna bude detekována s pravděpodobností jdoucí k 1.

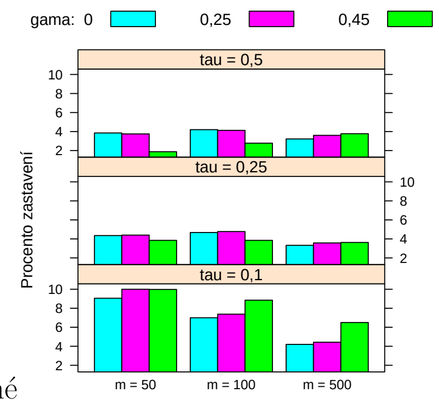
Simulace

Byly provedeny simulace pro ohodnocení výkonnosti procedury s konečným tréninkovým obdobím při použití asymptotických kritických hodnot, tj. kritických hodnot pro uvedený funkcionál Wienerova procesu. Uvažovaly se vlivy

- délky historického období ($m = 50; 100; 500$)
- okamžiku změny ($k^* = 50; 100; 500$)
- dolaďovací konstanty ($\gamma = 0; 0{,}25; 0{,}45$)
- rozdělení chyb (normální a Laplaceovo)
- velikosti změny $\delta = \mu_* - \mu_0$ (příp. typu alternativy)

pro různé kvantily ($\tau = 0{,}1; 0{,}25; 0{,}5; 0{,}75; 0{,}9$). To vše na 5% hladině spolehlivosti.

Jelikož se v testové statistice neobjevují přímo velikosti pozorování, není procedura citlivá na rozdělení chyb jako to bývá u procedur založených na L_2 odhadech. Na obrázku jsou shrnuty výsledky za platnosti H_0 (procenta chybných zastavení). K překročení předepsané hladiny dochází tedy jen pro extrémní kvantily a zejména malé m.



Za alternativní hypotézy procedura vykazuje stejné rysy jako obdobné procedury, a to pro všechny kvantily (specielně pro medián jsou výsledky shodné s [1]) To znamená:

- dolaďovací konstanta γ - pro k^* malé je nejvhodnější γ blízké 1/2, naopak pro pozdní změnu $\gamma = 0$. Konstanta 0,25 je nevhodnější pro k^* srovnatelné s m
- zhoršení rychlosti detekce s nárůstem k^* - typický problém CUSUM procedur
- prodloužení historického období - zlepšení rychlosti detekce; výrazné pro malé změny ($\delta = 1/2 \cdot \mathrm{sd}(Y_i)$)
- větší změny (nad $2{\cdot}\mathrm{sd}(Y_i)$) se již neprojevují zrychlením detekce - typické pro „L_1" procedury

Oproti jiným procedurám se však musí rozlišovat znaménko změny. Zatímto pro medián je zpozdění detekce pro změnu $+c$ resp. $-c$ prakticky stejné, u jiných kvantilů je situace rozdílná. V tabulce jsou uvedeny mediány času zastavení pro různé velikosti změny (pro $k^* = 100$). Pokud změna nastala „ve směru kvantilu", tj. např. záporná změna pro 1. kvartil, dochází ke zrychlení detekce, naproti tomu při opačné změně dochází ke zpomalení. Pro extrémnější kvantily je situace ještě výraznější. Tento fakt je způsoben tvarem definice $\tilde\varepsilon_i$. Také můžeme pozorovat zrychlení detekce s růstem m u malé změny.

τ	$\delta \backslash$ m	50	100	500
0,5	0,5	360	246	206
	2	176	156	151
	-2	173	156	151
0,25	0,5	1000	782	335
	2	322	228	199
	-2	135	129	128
0,75	0,5	249	207	187
	2	135	129	128
	-2	322	235	199

Následující tabulka obsahuje mediány času zastavení pro model (2), kdy se distribuce chyb měnila z N(0,1) na N(0,9), a různé kvantily (pro $m = k^* = 100,\ \gamma = 0{,}25$). Medián distribuce se nezměnil, 1000 odpovídá max. monitorovacímu období.

0,1	0,25	0,5	0,75	0,9
168	376	1000	365	167

LINEÁRNÍ MODEL

Problém je analogický jako v případě modelu polohy. Nyní uvažovaný model je

$$Y_i = \mathbf{X}_i^T \boldsymbol\beta_i + e_i,$$

kde e_i mají opět distribuční funkci F, $\mathbf{X}_i$ jsou s nimi nezávislé a splňují další předpoklady uvedené v [1]. Zajímáme se o regresní kvantily $\boldsymbol\beta_i^\tau$, tzn. podmínka stability má tvar $\boldsymbol\beta_1^\tau = \ldots = \boldsymbol\beta_m^\tau = \boldsymbol\beta_0^\tau$ a testové hypotézy jsou pro ně stejné jako (3), (4).

Taktéž testová statistika je analogická. Definujeme

$$\tilde\varepsilon_i^R = \tau - I[Y_i - \mathbf{X}_i^T \tilde{\boldsymbol\beta}_m^\tau < 0],$$

kde $\tilde{\boldsymbol\beta}_m^\tau$ je τ regresní kvantil historických pozorování získaný minimalizací $\min_{\mathbf{b}} \sum_{i=1}^m \rho_\tau(Y_i - \mathbf{X}_i^T \mathbf{b})$.

Populační analogie je $\varepsilon_i^R = \tau - I[Y_i - \mathbf{X}_i^T \boldsymbol\beta_0^\tau]$, pro které za H_0 opět platí: $\mathrm{E}\,\varepsilon_i^R = 0,\ \mathrm{var}\,\varepsilon_i^R = \tau(1-\tau)$. Testová statistika je tedy definována jako

$$Q^R(m,k) = \frac{1}{\sqrt{\tau(1-\tau)}} \sum_{i=m+1}^{m+k} \tilde\varepsilon_i^R.$$

Za trochu zúžených předpokladů (v závislosti na regresorech $\mathbf{X}_i$ - viz. [1]) pro statistiku $Q^R(m,k)$ platí obdobné vztahy jako (5), (6), které zaručují teoretickou použitelnost procedury.

Byly provedeny simulace pro model lineární regrese, kde $\mathbf{X}_i = (1, X_i)^T$ a rozdělení X_i bylo buď normální N(0,16) nebo stejnoměrné U(0,4). Zbylé parametry se shodovaly s modelem polohy. Taktéž výsledky byly obdobné.

Zdůrazněme pouze vztah regresorů X_i a znaménka velikosti změny $\boldsymbol\delta = \boldsymbol\beta_* - \boldsymbol\beta_0$. U krajních kvantilů opět dochází ke zlepšení/zhoršení rychlosti detekce dle směru posunu. Pro centrované regresory však druh změny směrnice nemá tento vliv. Avšak v tomto případě neplatí pro medián analogie vztahu (6), což se projevilo neschopností detekce (1000 je max. pozorovací období, $m = k^* = 100,\ \gamma = 0{,}25$).

$\mathcal{L}(X_i)$	$\tau \backslash \boldsymbol\delta$	$(1,0)^T$	$(0,1)^T$	$(-1,0)^T$	$(0,-1)^T$
N(0,16)	0,1	1000	151	145	151
	0,25	304	264	142	263
	0,5	175	1000	178	1000
U(0,4)	0,1	1000	1000	143	121
	0,25	304	269	142	135
	0,5	178	170	175	167

Poděkování

Autor by rád poděkoval paní prof. Huškové za věcné připomínky a odbornou pomoc a ČSOB za umožnění účasti na konferenci. Práce byla částečně podporována grantem GACR 201/06/0186.

Literatura

[1] A. Koubková. *Sequential Change-Point Analysis.* PhD thesis, Faculty of Mathematics and Physics, Charles University Prague, 2006.

[2] L. Horváth, M. Hušková, P. Kokoszka, and J. Steinebach. Monitoring changes in linear models. *J. Stat. Plann. Inference*, 126:225–251, 2004.