Rozdělení zpoždění detekce bodu změny v sekvenčním přístupu

Alena Koubková
Katedra pravděpodobnosti a matematické statistiky, Matematicko-fyzikální fakulta Univerzity Karlovy v Praze,
Sokolovská 83, 186 75 Praha 8
koubkova@karlin.mff.cuni.cz

Poděkování. Tato je podporována granty GAČR 201/05/H007 a GAČR 201/06/0186.

Abstrakt. V Koubková [2004a] byl navržen postup detekce bodu změny v sekvenčně pozorovaných datech založený na L_1-reziduích. Zde je odvozeno rozdělení zpoždění takto detekované změny oproti času skutečné změny za některých speciálních podmínek.

Definice problému a předpoklady

Uvažujme sekvenčně přicházející data splňující model polohy

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

kde Y_i jsou pozorovaná data, μ_i parametry polohy a e_i náhodné chyby.

Předpokládejme, že v datech může nastat změna v parametru polohy. Úkolem je odhalit tuto změnu co nejdříve a omezit přitom pravděpodobnost falešného poplachu.

O datech a jednotlivých parametrech modelu předpokládáme

(A) K dispozici jsou tréninková data beze změny o velikosti m, tj.

$$\mu_1 = \ldots = \mu_m.$$

(B) e_i, $1 \le i < \infty$ jsou nezávislé stejně rozdělené náhodné veličiny s distribuční funkcí F symetrickou kolem nuly, mající druhou derivaci v okolí nuly, a takovou, že platí $F'(0) = f(0) > 0$.

Řešení problému

Problém se řeší testováním nulové hypotézy, že změna nenastala, tj.

$$H_0 : \mu_i = \mu_0, \quad m+1 \le i < \infty$$

proti alternativě, že změna nastala

$$H_A : \text{existuje } k^* \ge 1 \text{ a } \mu_* \ne \mu_0 \text{ tak, že}$$
$$\mu_i = \mu_0,\ m+1 \le i < m+k^*, \quad \mu_i = \mu_*,\ m+k^* \le i < \infty,$$

za podmínek

$$\lim_{m\to\infty} P_{H_0}(\tau(m) < \infty) = \alpha, \tag{2}$$
$$\lim_{m\to\infty} P_{H_A}(\tau(m) < \infty) = 1, \tag{3}$$

kde $\tau(m)$ je čas detekce změny.

Podmínka (2) zaručuje, že pravděpodobnost falešného poplachu je omezena číslem α. Podmínka (3) říká, že pravděpodobnost odhalení změny, pokud nastala, je v limitě rovna jedné (tj. test je konzistentní).

V Koubková [2004a] je pro tento případ navržena testová statistika

$$\widetilde{Q}(m,k) = \left| \sum_{i=m+1}^{m+k} \operatorname{sign}(Y_i - \widetilde{\mu}_m) \right|,$$

která je založena na kumulativních součtech L_1-reziduí $\widetilde{e}_i = \operatorname{sign}(Y_i - \widetilde{\mu}_m)$, kde $\widetilde{\mu}_m$ je medián z pozorování $Y_1, \ldots, Y_m$.

Změna je detekována (tj. H_0 zamítnuta) a pozorování zataveno, pokud hodnota testové statistiky přesáhne ktitickou hranici. Definujeme čas detekce změny

$$\tau(m) = \begin{cases} \inf\{k \ge 1 : \widetilde{Q}(m,k)/g(m,k,\gamma) \ge c(\alpha)\} \\ 0, \qquad \text{pokud } \widetilde{Q}(m,k)/g(m,k,\gamma) < c(\alpha), \ \text{ pro všechna } k, \end{cases}$$

kde $g(m,k,\gamma)$ je hraniční funkce ve tvaru

$$g(m,k,\gamma) = \sqrt{m}\left(1+\frac{k}{m}\right)\left(\frac{k}{m+k}\right)^{\gamma}$$

s dolad'ující konstantou $\gamma \in [0, 1/2)$ a $c(\alpha)$ je konstanta zaručující, že podmínky (2) a (3) jsou splněny.

V Koubková [2004a] je odvozeno limitní chování testové statistiky za platnosti nulové i alternativní hypotézy. Za platnosti H_0 platí

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

pro každé $c > 0$, kde $\{W(t),\ 0 \le t < \infty\}$ značí Wienerův proces. A za platnosti H_A bylo dokázáno, že

$$\sup_{1\le k<\infty} \frac{\left|\widetilde{Q}(m,k)\right|}{g(m,k,\gamma)} \xrightarrow{P} \infty.$$

Pro $\gamma > 0$ není znám přesný tvar limitního rozdělení za H_0, a kritické hodnoty je třeba aproximovat pomocí simulací, viz. Horváth a col. [2001] a Koubková [2004b].

Reference

1. **Horváth a col. [2001]** Horváth L., Hušková M., Kokoszka P. a Steinebach J., Monitoring changes in linear models. *J. of Statistical Planning and Inference* **126**, 225 – 251, 2004.

2. **Koubková [2004a]** Koubková A. Sequential change point analysis in location model. *Sborník WDS'04*, 24 – 29, 2004.

3. **Koubková [2004b]** Koubková A., Critical values for changes in sequential regression models. Proceedings of COMPSTAT 2004, Antoch J. ed., Physica-Verlag/Springer, Heidelberg, 1345 – 1352, 2004.

Rozdělení času detekce změny

Předpokládejme následující speciální případ.

(C)

$$(i) \quad \delta_m \to 0, \quad \wedge \quad \sqrt{m}\delta_m \to \infty$$
$$(ii) \quad k^* = O(m^{\theta}) \text{ with some } 0 \le \theta < \left(\frac{1-2\gamma}{2(1-\gamma)}\right)^2.$$

Uvažujeme tedy velikost změny klesající k nule, spolu s m jdoucím do nekonečna, ale ne příliš rychle, aby změnu ještě bylo možno detekovat. Zároveň předpokládáme, že změna nenastala příliš brzy. Pro jednoduchost dále předpokládejme, že $\mu_0 = 0$. Za těchto podmínek platí mnásledující tvrzení.

Věta 1 *Necht' platí model (1) a jsou splněny předpoklady (A), (B) a (C). Necht' $\gamma \in [0, 1/2)$. Pak za platnosti alternativní hypotézy H_A platí*

$$\lim_{m\to\infty} P\left\{\frac{\tau(m) - a(m)}{b(m)} \le x\right\} = \Phi(x),$$

kde $\Phi(x)$ značí distribuční funkci standardního normálního rozdělení,

$$a(m) = \left(\frac{c_m(\alpha) m^{1/2-\gamma}}{2f(0)\delta_m}\right)^{1/(1-\gamma)}, \quad b(m) = \frac{\sqrt{a(m)}}{(1-\gamma)2f(0)\delta_m}$$

a $c_m(\alpha)$ je konstanta zaručující splnění podmínek (2) a (3).

Pro obecnější případy není přesně známo rozdělení času $\tau(m)$, ale je možno alespoň řádově odhadnout jeho zpoždění oproti skutečnému času změny. Předpokládejme nyní následující situaci.

(D) Změna nastává v čase $m + k^*$, kde $k^* = cm^{\beta}, \beta \in (0, \infty)$, $c > 0$ je konstanta a velikost změny je konstantní, tedy $\delta_m = \delta$ pro všechna m.

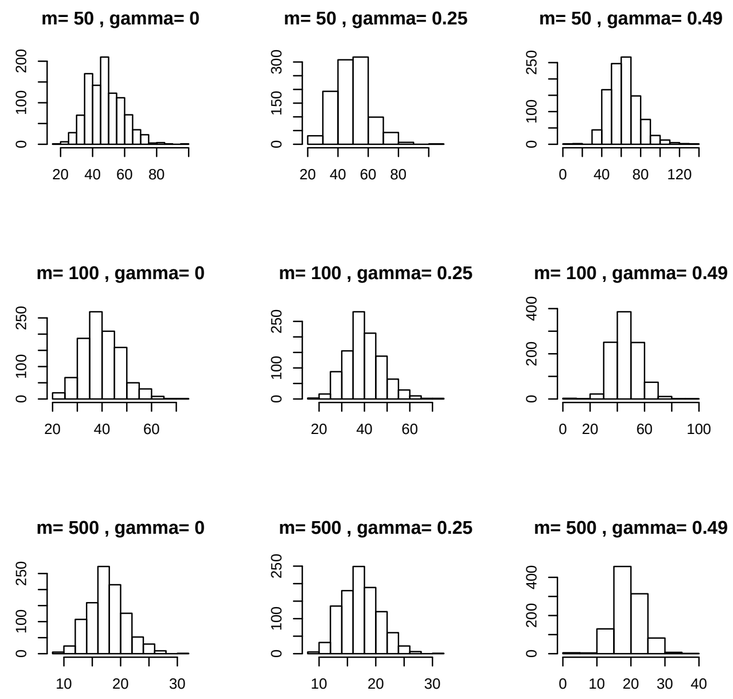
V takovém případě platí následující věta.

Věta 2 *Necht' platí model (1) a jsou splněny podmínky (A), (B) a (D). Pak za platnosti alternativy H_A, platí*

1. *je-li $0 < \beta < (1-2\gamma)/2(1-\gamma)$ pak $\tau(m) - k^* = O(m^{\frac{1-2\gamma}{2(1-\gamma)}})$,*
2. *je-li $(1-2\gamma)/2(1-\gamma) \le \beta < 1$ pak $\tau(m) - k^* = O(m^{1/2+\gamma(\beta-1)})$*
3. *je-li $\beta \ge 1$ pak $\tau(m) - k^* = O(m^{\beta-1/2})$.*

Výsledky simulací

Pro situaci popsanou ve Větě 1 byly provedeny simulace s hodnotami $\gamma = 0, 0.25, 0.49$, $m = 50, 100, 500$, $\delta = 5m^{-1/4}$ a $\theta = ((1-2\gamma)/(2(1-\gamma)))^2/2$. Výsledky jsou shrnuty v grafech a v tabulce popisující chování náhodné veličiny $(\tau(m) - a(m))/b(m)$.



Shrnutí výsledků pro $(\tau(m) - a(m))/b(m)$ s hodnotou $\gamma = 0.25$

	min	1^{st}Q	median	3^{th}Q	max	mean
$m = 50$	23.16	40.71	48.23	55.75	101.5	48.92
$m = 100$	17.14	34.29	39.73	44.33	73.2	40.02
$m = 500$	8.428	15.14	16.94	19.23	31.01	17.3

Z tabulky i z grafů je patrno, že s rostoucím m se střední hodnota veličiny $(\tau(m) - a(m))/b(m)$ blíží k nule a klesá i její rozptyl. Pro zkoumané hodnoty m však byla zamítnuta normalita této veličiny. Konvergence k cílovému normálnímu rozdělení je tedy poměrně pomalá.