VYHLEDÁVÁNÍ ZMĚN ROZPTÝLENOSTI

DANIELA JARUŠKOVÁ

FSv ČVUT

Můj zájem o vyhledávání změn rozptýlenosti v řadě pozorování byl motivován představou některých hydrologů, že díky odlesňování a zhoršování kvalit půdy lze očekávat změnu charakteristik průtokové řady. Změna rozptýlenosti může být způsobena menší schopností rostlinstva a půdy zadržovat vodu. V literatuře jsou však uváděny i jiné aplikace, např. změna v chování cen akcií po aféře Watergate, viz Hsu (1977).

V příspěvku se předpokládá, že pozorované veličiny $Y_1, \ldots, Y_n$ jsou normálně rozdělené se známou konstantní střední hodnotou, kterou můžeme bez újmy na obecnosti považovat za nulovou. Problém vyhledávání změn v rozptýlenosti řady můžeme v rámci testování hypotéz zformulovat následně:

(1) H_0: $Y_1, \ldots, Y_n \sim N(0, \sigma^2)$,
(2) A: $\exists k \in \{1, \ldots, n-1\}$, $Y_1, \ldots, Y_k \sim N(0, \sigma_1^2)$, $Y_{k+1}, \ldots, Y_n \sim N(0, \sigma_2^2)$,
(3) $\sigma_1^2 \neq \sigma_2^2$.

Náhodné veličiny $X_i = Y_i^2$ mají gamma rozdělení s parametrem tvaru $\alpha = 1/2$ a parametrem měřítka $\Theta = 2\sigma^2$. Shora popsaný problém je tedy speciálním případem vyhledávání změn v parametru měřítka náhodných veličin s gamma rozdělením při známém konstantním parametru tvaru.

Pokud bychom předem znali čas k, ve kterém může dojít ke změně rozptylu řady, pak bychom použili k testování nulové hypotézy 2·logaritmický poměr věrohodnosti, který má v našem případě tvar

$$Z_k = k \log \frac{\bar{x}_n}{\bar{x}_k} + (n-k) \log \frac{\bar{x}_n}{\bar{x}_k^*} = -k \log \left(\frac{S_k}{S_n} \frac{n}{k} \right) - (n-k) \log \left(\left(1 - \frac{S_k}{S_n} \right) \frac{n}{n-k} \right),$$

kde $S_k = \sum_{i=1}^{k} X_i, S_k^* = \sum_{i=k+1}^{n} X_i, \bar{x}_k = S_k/k, \bar{x}_k^* = S_k^*/(n-k)$. Při hledání rozdělení náhodné veličiny Z_k je třeba si uvědomit, že jev $\{Z_k < t\}$ je totožný s jevem $\{a_k(t) < S_k/S_n < b_k(t)\}$, kde $a_k(t)$ a $b_k(t)$ jsou řešením rovnice

$$-k \log \left(x \frac{n}{k} \right) - (n-k) \log \left((1-x) \frac{n}{n-k} \right) = t.$$

Odtud pravděpodobnost $P(Z_k < t) = P(a_k(t) < S_k/S_n < b_k(t))$, kde náhodná veličina S_k/S_n má beta rozdělení o parametrech $k/2$ a $(n-k)/2$, resp. náhodná veličina $\frac{S_k/k}{S_k^*/(n-k)}$ má F-rozdělení o k a $n-k$ stupních volnosti.

Jestliže čas k, v kterém může dojít ke změně rozptylu předem neznáme, pak lze k testování nulové hypotézy použít statistiku $\max_{k=1,\dots,n-1} Z_k$. Zřejmě platí

$$P\left(\max_{k=1,\dots,n-1} Z_k < t\right) = P\left(a_1(t) < \frac{S_1}{S_n} < b_1(t), \dots, a_{n-1}(t) < \frac{S_{n-1}}{S_n} < b_{n-1}(t)\right),$$

přičemž sdružená hustota vektoru $\left(\frac{S_1}{S_n}, \dots, \frac{S_{n-1}}{S_n}\right)$ má tvar

$$f(s_1, \dots, s_{n-1}) = \frac{const}{\sqrt{s_1}} \prod_{i=2}^{n-2} \frac{1}{\sqrt{s_i - s_{i-1}}} \frac{1}{\sqrt{1 - s_{n-1}}}, \quad 0 \le s_1 \le \dots \le s_{n-1} \le 1.$$

Avšak najít rozdělení náhodné veličiny $\max_{k=1,\dots,n-1} Z_k$ numerickou integrací se zdá být bohužel neschůdné, viz Worsley (1986).

Při odhadu shora p-hodnoty můžeme použít Bonferoniho nerovnost

$$P\left(\max_{k=1,\dots,n-1} Z_k > t\right) \le \sum_{k=1}^{n-1} P(Z_k > t).$$

Najdeme-li t takové, pro které je pro určitou hladinu významnosti α splněn vztah $\sum_{k=1}^{n-1} P(Z_k > t) = \alpha$, pak toto t lze použít jako kritickou hodnotu konzervativního testu. Pomocí programu *Mathematika* jsme získali následující kritické hodnoty odvozené z Bonferoniho nerovnosti pro různé hodnoty n.

n	10	20	30	40	50
t	8.65625	9.789	10.437	10.9062	11.2688
n	60	70	80	90	100
t	11.5625	11.8234	12.0453	12.2422	12.4219

Pro velké hodnoty n lze použít asymptotické rozdělení

$$P\left(\max_{k=1,\dots,n-1} (2\log\log n) Z_k \le \left(x + 2\log\log n + \frac{1}{2}\log\log\log n - \frac{1}{2}\log\pi\right)^2\right) \to e^{-2e^{-x}}.$$

Je známo, že statistika $\max_{k=1,\dots,n-1} Z_k$ nabývá velkých hodnot pro k blízko krajům, a proto se hodí k vyhledávání změn, které nastanou na jednom nebo druhém konci řady, viz Gombay a Horvath (1990). Chceme-li detekovat změnu spíše uprostřed, používáme statistik $\max_{\lambda_1 n \le k \le \lambda_2 n} Z_k$ nebo $\max_{k=1,\dots,n-1} Z_k \frac{k(n-k)}{n^2}$, viz Deshayes a Picard (1986), přičemž

$$\max_{\lambda_1 n \le k \le \lambda_2 n} Z_k \xrightarrow{\mathcal{D}} \sup_{\lambda_1 \le s \le \lambda_2} \frac{B^2(s)}{s(1-s)},$$

$$\max_{k=1,\dots,n-1} Z_k \frac{k(n-k)}{n^2} \xrightarrow{\mathcal{D}} \sup_s B^2(s).$$

Dalším přístupem k vyhledávání změny v parametru rozdělení je pseudobayesovský přístup zavedený Chernoffem a Zackem (1964) a pro vyhledávání změn v rozptylu aplikovaný Hsuem. Jestliže apriorní rozdělení času, v kterém dochází ke změně můžeme považovat za rovnoměrné, pak lze k testování nulové hypotézy použít statistiku

$$T = \frac{\sum_{i=1}^{n-1} iX_i}{\sum_{i=1}^{n-1} X_i} = \sum_{i=1}^{n-1} \frac{S_i^*}{S_n},$$

kde $E\,T = \frac{n-1}{2}$ a $Var\,T = \frac{(n-1)(n+1)}{6(n+2)}$. Statistika T má asymptoticky normální rozdělení, přičemž pro malá n lze použít k výpočtu kritických hodnot Edgeworthova rozvoje. Hsu tabeloval kritické hodnoty pro některá n.

Hsu také navrhl pro testování další statistiku, která se rovná průměru p-hodnot statistik S_k/S_n, tj.

$$G = \frac{1}{n-1} \sum_{k=1}^{n-1} B^{-1}_{k/2,(n-k)/2}\left(\frac{S_k}{S_n}\right),$$

kde $B^{-1}_{\alpha,\beta}(x)$ je inverzní funkce k distribuční funkci beta rozdělení o parametrech α, β. Statistika G má limitní rozdělení, které Hsu navrhl aproximovat symetrickým beta rozdělením o parametrech (β, β), kde

$$\beta = \frac{1 - 4\sigma_n^2(G)}{8\sigma_n^2(G)},$$

$$\sigma_n^2(G) = 0,0393 + 0,0206/(n-1) + 0,0999/(n-1)^2 - 0,1445/(n-1)^3 + 0,0662/(n-1)^4.$$

Koeficienty v Taylorově rozvoji rozptylu $\sigma_n^2(G)$ nalezl pomocí metody Monte Carlo. Zároveň pro některá n tabeloval kritické hodnoty statistiky G.

References

1. Deshayes J., Picard D., *Off-line statistical analysis of change-point models*, *Lecture Notes in Control and Information Sciences*, vol. 77, Detection of abrupt changes in signals and dynamics systems, 1986, pp. 103–168.
2. Gombay E., Horvath L., *Asymptotic distribution of maximum likelihood tests for change in the mean*, Biometrika 77 (1990), 411–414.
3. Hsu D.A., *Tests for variance shift at an unknown time point*, Applied Statistics 26 (1977), 279–284.
4. Chernoff H., Zack S., *Estimating the current mean of a normal distribution which is subjected to changes in time*, Annals of Mathematical Statistics 35 (1964), 999–1018.
5. Kander Z., Zack S., *Test procedures for possible changes in parameters of statistical distributions occuring at unknown time points*, Annals of Mathematical Statistics 37 (1966), 1196–1210.
6. Worsley K.J., *Confidence regions and tests for a change-point in a sequence of exponential family random variables*, Biometrika 73 (1986), 91–104.

FSv ČVUT, KATEDRA MATEMATIKY, THÁKUROVA 7, 160 00 PRAHA 6

E-mail: JARUS at CSEARN.BITNET