Viacrozmerný test o parametroch polohy nevyžadujúci rovnaký typ rozdelení v súboroch

Ján Somorčík

somorcik@fmph.uniba.sk
Katedra aplikovanej matematiky a štatistiky, Univerzita Komenského, Bratislava

Mnohé testy o rovnosti parametrov polohy viacerých rozdelení sú vhodné len pre situáciou, keď sa tieto rozdelenia od seba líšia nanajvýš posunutím. Tu je predstavený pomerne robustný test, ktorý rieši i všeobecnejšiu situáciu, a je porovnaný so svojimi „predkami“ i konkurenciou.

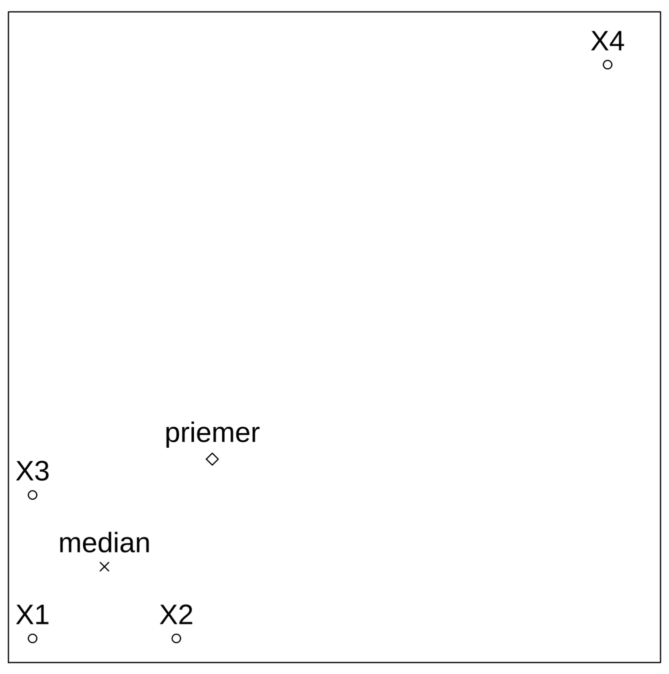
ÚVOD

Uvažujme q náhodných výberov d-rozmerných dát, ktoré pochádzajú z rozdelení s parametrami polohy $\mu_1, \ldots, \mu_q$. Rozdelenia sa od seba obyčajne líšia iba posunutím a rovnosť parametrov polohy potom znamená totožnosť týchto rozdelení. Na testovanie hypotézy

$$H_0 : \mu_1 = \ldots = \mu_q$$

existuje mnoho testov (viac v [1]).

My sme boli inšpirovaní Lawley-Hotellingovou testovacou štatistikou T^2 založenou na aritmetických priemeroch v jednotlivých súboroch. Na jej základe sme v [2] skonštruovali testovacie štatistiky M_1 a M_2, ktoré namiesto aritmetických priemerov používajú priestorové mediány (*priestorový medián dát* = bod v priestore, od ktorého je súčet Euklidovských vzdialeností k jednotlivým dátam najmenší). Dôvodom na použitie priestorových mediánov bola ich väčšia robustnosť voči odľahlým pozorovaniam, ktorú ilustruje obrázok.



Naše testovacie štatistiky mali tvar

$$M_1 := \sum_{a=1}^{q} n_a(\widehat{\mu}_a - \boxed{\widehat{\mu}})^T \widehat{V}^{-1}(\widehat{\mu}_a - \boxed{\widehat{\mu}}),$$
$$M_2 := \sum_{a=1}^{q} n_a(\widehat{\mu}_a - \boxed{\bar{\mu}})^T \widehat{V}^{-1}(\widehat{\mu}_a - \boxed{\bar{\mu}}),$$

kde n_a sú počty dát v súboroch a $\widehat{\mu}_a$ sú priestorové mediány súborov. $\widehat{V}$ je odhad aymptotickej kovariančnej matice V priestorového mediánu, ktorá je rovnaká pre všetky rozdelenia súborov.

M_1 a M_2 sa od seba líšia iba veličinami $\widehat{\mu}$ (=medián získaný zo všetkých dát) a $\bar{\mu}$ (=vážený priemer mediánov v súboroch). Vlastnosti M_1 a M_2 a porovnanie s konkurentmi sú v [2]. Spolu s Lawley-Hotellingovou štatistikou však majú nevýhodu, že na ich použitie sa vyžadujú až na posun rovnaké rozdelenia pravdepodobnosti v súboroch.

NOVÁ TESTOVACIA ŠTATISTIKA

Naším cieľom preto bolo upraviť M_1 resp. M_2, aby vzniknutá testovacia štatistika bola vhodná aj pre situácie, keď sa rozdelenia jednotlivých súborov líšia aj inak než len parametrom polohy. Problém M_1 a M_2 spočíva v odhade akéhosi „spoločného parametra polohy“ všetkých rozdelení. Zavádzame preto novú testovaciu štatistiku

$$M_3 := \sum_{a=1}^{q} n_a(\widehat{\mu}_a - \boxed{\widetilde{\mu}})^T \widehat{V}_a^{-1}(\widehat{\mu}_a - \boxed{\widetilde{\mu}}),$$

kde odhad spoločného parametra polohy nahrádza

$$\widetilde{\mu} := \widehat{W}^{-1} \sum_{a=1}^{q} n_a \widehat{V}_a^{-1} \widehat{\mu}_a.$$

Je to vážený priemer priestorových mediánov v jednotlivých súboroch. Ako váhy používame počty dát (v „pozitívnom“ zmysle) a odhady asymptotických kovariančných matíc V_a priestorových mediánov $\widehat{\mu}_a$ v súboroch (v „negatívnom“ zmysle). Takáto idea váženia sa v inej viacvýberovej situácii objavila napr. v [3].

Vlastnosti M_3

(I) za platnosti H_0 má asymptoticky rozdelenie $\chi^2_{(q-1)d}$

(II) za platnosti H_0 + rovnosti rozdelení v súboroch sa M_3 asymptoticky rovná M_1 aj M_2 (v zmysle konvergencie podľa pravdepodobnosti).

(III) za platnosti H_0 + rovnosti rozdelení v súboroch + sférickej symetrie sa M_3 asymptoticky rovná niektorým štatistikám využívajúcim priestorové znamienka (v zmysle konvergencie podľa pravdepodobnosti).

(IV) parameter necentrálnosti za platnosti Pitmanových alternatív (t.j. priestorový medián $\mu + \frac{h_a}{\sqrt{n}}$ v a-tom súbore) je $\sum_{a=1}^{q} p_a h_a^T V_a^{-1} h_a$. Za podmienok bodu (II) je to rovnaká hodnota ako u M_1 a M_2.

Čo vlastne testuje M_3?

Ak sa rozdelenia v súboroch líšia aj inak ako posunutím, tak ťažko už hovoriť o parametroch polohy. Test hypotézy H_0 sa stáva iba testom rovnosti priestorových mediánov rozdelení a výsledok testovania tým stráca prirodzenú interpretáciu. Ak však predpokladáme nejaký druh symetrie rozdelení súborov, tak stredy jednotlivých symetrií sú priestorovými mediánmi rozdelení. Pojem *parameter polohy* má tak znovu jasný význam. A rovnako výsledok testovania.

MONTE CARLO

Vykonali sme malú simulačnú štúdiu s $q = 3$ súbormi po $n_1 = n_2 = n_3 = 100$ dátach z $\mathbf{R}^3$. Prvý súbor sa generoval z $N_3(\mu_1, I_3)$, zvyšné dva z 3-rozmerného sféricky symetrického Cauchyho rozdelenia. Do simulácií sme zaradili aj štatistiky L_N (založená na pozložkových poradiach) a W_{ϕ_1}, W_{ϕ_2} založené na priestorových znamienkach (detaily v [2]). Z nich iba W_{ϕ_1} nevyžaduje na svoje použitie rovnaké rozdelenia súborov (až na posun). V každej z troch situácií sme vykonali 5 000 opakovaní. Ak nie je uvedené inak, parametre polohy μ_1, μ_2, μ_3 boli $(0,0,0)^T$.

Odhady pravdepodobnosti chyby 1. druhu resp. sily

	M_1	M_2	M_3	T^2	L_N	W_{ϕ_1}	W_{ϕ_2}
H_0 platí	0.060	0.064	0.063	0.029	0.049	0.049	0.057
$\mu_1 = (0.3,\ 0.3,\ 0)^T$	0.497	0.500	0.580	0.038	0.421	0.567	0.369
$\mu_2 = (0.3,\ 0.3,\ 0)^T$	0.480	0.489	0.485	0.042	0.315	0.456	0.283

- Ak je odchýlka polohy v 1. súbore (s menším rozptýlením), tak M_1 i M_2 silou zaostávajú za M_3.
- Ak je odchýlka polohy v 2. súbore (s väčším rozptýlením), tak M_1 i M_2 majú silu podobnú ako M_3.
- Cauchyho rozdelenie spôsobuje, že kvalita Lawley-Hotellingovho testu T^2 je veľmi úbohá.
- Takisto L_N a W_{ϕ_2} svojou silou výrazne zaostávajú za M_3.

Poďakovanie Autor by chcel poďakovať svojmu školiteľovi Františkovi Rublíkovi za mnohé cenné rady a diskusie. Práca bola podporená grantom VEGA 1/3016/06.

Odkazy

[1] Um, Y. & Randles, R. H. (1998). *Nonparametric tests for the multivariate multi-sample location problem.* Statistica Sinica, **8**, 801 – 812.

[2] Somorčík, J. (2006). *Tests Using Spatial Median.* Austrian Journal of Statistics, **35**, 331 – 338.

[3] Rublík, F. (2001). *Tests of some hypotheses on characteristic roots of covariance matrices not requiring normality assumptions*, Kybernetika **37**, 61 – 78.