Ekonomická fakulta
Faculty of Economics

Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

Modifikace Randlesových nadrovin

Jana Klicnarová

KMI EF, Jihočeská univerzita v Českých Budějovicích

ROBUST 2018 - Rybník

Obsah

Klasické znaménkové a pořádkové testy jsou velmi využívané neparametrické testy v případě jednorozměrných náhodných veličin — lze tyto testy zobecnit i pro vícerozměrná rozdělení?

Uvažujeme náhodný vzorek $\mathcal{X}_n$, který obsahuje n pozorování $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^p$ sledující "symetrické" rozdělení s neznámým středem symetrie $\theta \subset \boldsymbol{R}^p$.

Symetrické rozdělení?

Sféricky symetrické rozdělení

Řekneme, že náhodný vektor **X** sleduje **sféricky symetrické rozdělení** okolo θ, jestliže rotace vektoru **X** okolo středu θ rozdělení nezmění. Přesněji:

$$\mathbf{X} - \theta \stackrel{d}{=} \mathbf{A}(\mathbf{X} - \theta)$$

pro všechny ortonormální matice **A**.

Elipticky symetrická rozdělení

Řekneme, že náhodný vektor **X** sleduje **elipticky symetrické rozdělení** s parametry θ a Σ, jestliže existuje sféricky symetricky rozdělený náhodný vektor Y takový:

$$\mathbf{X} \stackrel{d}{=} \mathbf{A}^{'}\mathbf{Y} + \theta,$$

kde matice $\mathbf{A}_{kxd}$ je taková, že $\mathbf{A}^{'}\mathbf{A} = \Sigma$.

Tři základní přístupy k vícerozměrným neparametrickým testům

1. Puri, Sen (1971);
2. prostorová znaménka, pořadí, Ojův medián – Oja (1999);
3. interdirekce (interdirections).

Interdirekce

Myšlenka přístupu pomocí interdirekcí byla představena Randlesem (1989).

Definice

Řekneme, že nadrovina je **interdirekcí mezi body** $\mathbf{y}_1, \mathbf{y}_2 \in \boldsymbol{R}^p$ ve vzorku $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, pokud tato nadrovina prochází středem rozdělení a dalšími $p-1$ body z pozorování $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ a zároveň tato nadrovina odděluje body $\mathbf{y}_1$ a $\mathbf{y}_2$.

Randles (1989): Jednovýběrový test na polohu rozdělení

Nulová hypotéza:

$$H_0 : \theta = \mathbf{0}.$$

$$V_n = \frac{p}{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \cos(\pi \hat{p}_{jk}),$$

kde

$$\begin{aligned} \hat{p}_{jk} &= (C_{jk} + d_n) / \binom{n}{p-1} \quad j \neq k \\ &= 0 \quad j = k, \end{aligned}$$

C_{jk} je počet interdirekcí mezi X_j a X_k; a $d_n = \left(\binom{n}{p-1} - \binom{n-2}{p-1} \right) / 2$.

Jednovýběrový test na polohu rozdělení

Nulová hypotéza:

$$H_0 : \theta = \mathbf{0}.$$

Randles (1989)

Jestliže pozorování pochází z elipticky symetrického rozdělení, potom za platnosti hypotézy H_0 platí: $V_n^2 \to \chi_p^2$ v rozdělení pro $n \to \infty$.

Jednovýběrový test na parametr polohy rozdělení

Později Peters and Randles (90), Jan and Randles (94) tyto výsledky vylepšili, porovnali je s jinými testy. Hallin a Paindaveine ve článku z roku 2002 ukázali nové statistiky založené na interdirekcích, ukázali jejich asymptotické vlastnosti.

Test dvou výběrů na shodu parametrů polohy

Randles and Peters (Randles and Peters (90)) uvažovali test s nulovou hypotézou

$$H_0 : \theta_1 = \theta_2$$

za podmínky elipticky symetrického rozdělení.

Vícevýběrový test na shodu parametrů polohy

Později např. Um, Randles (98) zobecnili jejich výsledky také pro situaci více než dvou výběrů.

Neparametrické testy nezávislosti

V roce 1997 Gieser a Randles představili také test nezávislosti mezi vektory také založený na Randlesových interdirekcích.

Klíčový krok (Hallin, Paindeviene (02))

Nechť $\mathbf{X}_1, \mathbf{X}_2, \ldots$ jsou i.i.d. p-rozměrné vektory se sféricky symetrickým rozdělením. Označme $\alpha(y_1, y_2) := \arccos((y_1' y_2)/(\|y_1\| \|y_2\|))$ úhel mezi y_1, y_2 (vůči středu rozdělení).
Potom

$$\binom{n}{d-1}^{-1} C^{(n)}(y_1, y_2)$$

konverguje v L_2 k $\pi^{-1}\alpha(y_1, y_2)$ pro $n \to \infty$.

Důkaz

Založen na U-statistikách...

Problém se znalostí parametrů polohy rozdělení

Výše uvedené metody předpokládají znalost středu rozdělení, k jejich realizaci je tedy zapotřebí nejprve odhadnout střed rozdělení.

Je skutečně nezbytné odhadovat střed rozdělení? Nebylo by možné najít podobnou techniku, která by nepotřebovala znalost středu rozdělení?

Lift-interdirections

Oja, Paindaveine (2005)

Oja a Paindaveine (2005) představili ve své práci tzv. lift-interdirections. **Lift-interdirections** jsou založeny na všech nadrovinách, které procházejí p body z pozorování a oddělují body y_1 a y_2.

Ukazují, že (za předpokladu elipticky symetrického rozdělení) pořadí jejich lift-interdirections se asymptoticky shoduje s pořadím v Mahalanobisově vzdálenosti. Ukázali, že lift-interdirection společně s Randlesovými interdirekcemi mohou být použity v testech, které odvodili Hallin, Paindaveine (2002).

Interdirekce hloubky δ

Nápad

Tento přístup kombinuje obě výše zmíněné metody. Interdirekce hloubky δ jsou takové nadroviny v prostoru, které procházejí p body pozorování a zároveň rozdělují pozorované hodnoty na dvě "poloviny". (Nemusíme nutně uvažovat poloviny, stačí nám, pokud povolíme, že rozdíl mezi počtem bodů v jednotlivých poloprostorech se liší maximálně o předem zvolené číslo – δ).

Nebo je možné ukázat, že některé výsledky platí i pro modifikované Randlesovy nadroviny ve smyslu, že uvažujeme všechny lift-interdirection, které procházejí dostatečně blízko ke středu rozdělení?

$\mathcal{X}_n = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ je vzorek, $\boldsymbol{q} := (q_1, q_2, \ldots q_p)$, $1 \leq q_1 < q_2 < \cdots < q_p \leq n$, nadrovina $H^{\boldsymbol{q}}$ je nadrovina procházející body $X_{q_1}, X_{q_2}, \ldots, X_{q_n}$.

δ-HLUBOKÁ NADROVINA

Je každá nadrovina $H^{\boldsymbol{q}}$: $|\sum_{k=1}^{n} S^{\boldsymbol{q}}(\boldsymbol{X}_k)| \leq \delta$, $\delta \in \mathbb{N}$, kde $S^{\boldsymbol{q}}(\boldsymbol{X}_k)$ je znaménko X_k vůči nadrovině $H^{\boldsymbol{q}}$.

Položme

$$N_n^\delta = \sum_{\boldsymbol{q}\in\mathcal{Q}} \boldsymbol{I}_{\{|\sum_{k=1}^n S^{\boldsymbol{q}}(\boldsymbol{X}_k)|\leq\delta\}} \equiv \sum_{\boldsymbol{q}\in\mathcal{Q}} w_n^\delta(\boldsymbol{q})$$

a

$$\widetilde{C}^\delta_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\mathcal{X}_n) = \sum_{\boldsymbol{q}\in\mathcal{Q}} w_n^\delta(\boldsymbol{q}) h(\mathcal{X}^{\boldsymbol{q}}),$$

kde $h(\mathcal{X}^{\boldsymbol{q}}) = (1 - S^{\boldsymbol{q}}(\boldsymbol{y}_1) S^{\boldsymbol{q}}(\boldsymbol{y}_2))\,/2$.

Chtěli bychom ukázat, že

$$\frac{C^\delta_{\mathbf{y}_1,\mathbf{y}_2}(\mathcal{X}_n)}{N_n^\delta} \to \frac{1}{\pi}\alpha(\mathbf{y}_1, \mathbf{y}_2).$$

Pokud uvažujeme

$$U_2^\varepsilon := \frac{\widetilde{B}^\varepsilon_{\mathbf{y}_1,\mathbf{y}_2}(\mathcal{X}_n)}{M_n^\varepsilon}$$

kde

$$M_n^\varepsilon = \sum_{\boldsymbol{q}\in\mathcal{Q}} \boldsymbol{I}_{\{\Delta(H^{\boldsymbol{q}},\theta)<\varepsilon\}} \equiv \sum_{\boldsymbol{q}\in\mathcal{Q}} v^\varepsilon(\boldsymbol{q})$$

a

$$\widetilde{B}^\varepsilon_{\mathbf{y}_1,\mathbf{y}_2}(\mathcal{X}_n) = \sum_{\boldsymbol{q}\in\mathcal{Q}} v^\varepsilon(\boldsymbol{q}) h(\mathcal{X}^{\boldsymbol{q}}).$$

Pro sféricky symetrická rozdělení s kladnou hustotou okolo středu rozdělení θ:

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty} \mathrm{E}\frac{\widetilde{B}^\varepsilon_{\mathbf{y}_1,\mathbf{y}_2}(\mathcal{X}_n)}{M_n^\varepsilon} = \frac{1}{\pi}\alpha(\mathbf{y}_1,\mathbf{y}_2).$$

Tvrzení

Pokud $\mathbf{X}_1, \mathbf{X}_2, \ldots$ pochází ze sféricky symmetrického spojitého rozdělení s kladnou hustotou v okolí středu symetrie θ, potom

$$\mathrm{E}\frac{C^{\delta}_{\mathbf{y}_1,\mathbf{y}_2}(\mathcal{X}_n)}{N^{\delta}_n} \to \frac{1}{\pi}\alpha(\mathbf{y}_1, \mathbf{y}_2),$$

kde $C^{\delta}_{\mathbf{y}_1,\mathbf{y}_2}(\mathcal{X}_n)$ je počet δ-hlubokých interdirekcí, které oddělují body $\mathbf{y}_1$ a $\mathbf{y}_2$ a N^{δ}_n je počet všech δ-hlubokých nadrovin, které procházejí p body z pozorování $\mathcal{X}_n$.

Simulace – Dvouvýběrový test na shodu parametrů polohy

Předpokládáme dva nezávislé p-rozměrné náhodné vzorky $\mathcal{X}_{n_A}^A$ a $\mathcal{X}_{n_B}^B$ o velikostech n_A a n_B, které pocházejí ze dvou elipticky symetrických rozdělení, která se liší pouze v parametru polohy – θ_A, θ_B.

Randles a Peters (1990)

Randles a Peters (1990) představili test založený na Randlesových interdirekcích. Jako testou statistiku položili:

$$S = \frac{p n_A n_B}{n_A + n_B} \left(\frac{1}{n_A^2} \sum_{i=1}^{n_A} \sum_{j=1}^{n_A} \cos(\beta_{ij}) + \frac{1}{n_B^2} \sum_{i=n_A+1}^{n_A+n_B} \sum_{j=n_A+1}^{n_A+n_B} \cos(\beta_{ij}) - \frac{2}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=n_A+1}^{n_A+n_B} \cos(\beta_{ij}) \right).$$

Dvouvýběrový test na shodu parametrů polohy

Randles a Peters (1990)

$$\beta_{ij} = \pi C_{\boldsymbol{x}_i - \widehat{\theta}, \boldsymbol{x}_j - \widehat{\theta}}(\mathcal{X}_n - \widehat{\boldsymbol{\theta}}) / \binom{n}{p-1}, \quad i, j = 1, \ldots, n,$$

kde $C_{\boldsymbol{x}_i}$ je počet Randlesových interdirekcí ze spojeného vzorku centrovaný odhadem parametru polohu $\widehat{\boldsymbol{\theta}}$ (za platnosti H_0).

Odhad parametru polohy

Pro odhad parametru polohy $\widehat{\boldsymbol{\theta}}$ jsme využili dvou různých metod – jednou jsme parametr polohy odhadli jako výběrový střed a podruhé jako výběrový robustní prostorový medián (Hettemanspeger a Randles (2002)) – testy označme postupně T_1 a T_2.

Dvouvýběrový test na shodu parametrů polohy

Deep Interdirections

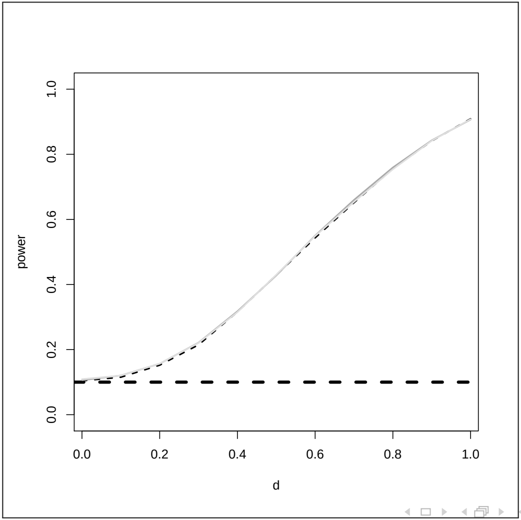
$$\beta_{ij} = \pi \widetilde{C}^{\delta}_{\mathbf{x}_i,\mathbf{x}_j}(\mathcal{X}_n)/N^{\delta}_n$$

V tomto případě jsme využili modifikovanou verzi interdirekcí, a to takových, kde nepředpokládáme znalost středu rozdělení, využíváme tzv. interdirekcí hloubky δ.

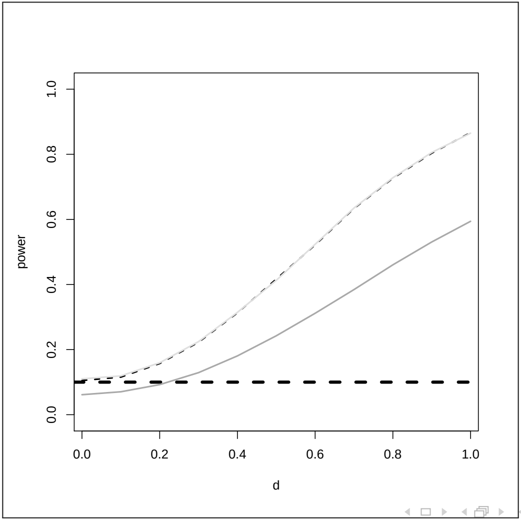
V simulaci jsme tento test označili jako T^{δ}_3.

Pro porovnání výše uvedených tří testů T_1, T_2 a T_3^δ zobrazíme empirické síly testů (založené na $N = 5\,000$ nez. experimentech) testu T_1 (světle šedá), T_2 (tmavošedá) a T_3^δ (černá čárkovaná) na hladině $\alpha = 0.1$ (horizontální černá čárkovaná) pro $\delta = 2$ (Obrázek 1) a pro $\delta = 4$ (Obrázek 2) aplikovaného na dva nezávislé vzorky o stejných velikostech $n_A = n_B = 15$ (Obrázek 1) a $n_A = n_B = 25$ (Obrázek 2) pocházející z dvourozměrných normálních rozdělení (Obrázek 1) a z dvourozměrných studentovo rozdělení $t(1)$ (Obrázek 2), jejichž parametry polohy se lišily o $(d, d)'$, $d \in \{0, 0.1, \dots, 1.0\}$. Všechny nasimulované varianty byly ohodnoceny všemi testy (za předpokladu platnosti hypotézy H_0) a postupně s využitím všem možných deseti posunů.

Porovnání klasických a modifikovaných interdirekcí



Porovnání klasických a modifikovaných interdirekcí



A co dál?

Předpokládáme, že stejně jako v dvou výběrovém testu je možné nahradit Randlesovy interdirekce interdirekcemi hloubky δ v mnoha dalších testech, které jsou založeny na interdirekcích. Např. v testech pro vícenásobný test shodnosti parametrů polohy – Um, Randles (1998), pro test nezávislosti mezi dvěma vektory – Gieser, Randles (1997), testy proporcionality rozptylu – Ghoush and Sengupta (2001), pro test nezávislosti Hallin, Paindaveine (2002), Paindaveine (2009) aj. Na druhou stranu, jsou samozřejmě testy, kde není zapotřebí původní interdirekce nahrazovat – see, e.g., Hallin, Paindaveine (2002b) and Taskinen and al. (2005).

- Gieser, P. W., Randles, R. H.: A nonparametric test of independence between two vectors. *Journal of the American Statistical Association* **92 (438)** (1997), 561-567.
- Hallin, M., Paindaveine, D.: Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. *The Annals of Statistics*, **30(4)** (2002), 1103–1133.
- Hallin, M., Paindaveine, D.: Optimal procedures based on interdirections and pseudo-Mahalanobis ranks for testing multivariate elliptic white noise against ARMA dependence. *Bernoulli* **8 (6)** (2002) 787-815.
- Oja, H.,Paindaveine, D.: Optimal signed-rank tests based on hyperplanes. *Journal of statistical planning and inference*, **135(2)** (2005), 300–323.
- Paindaveine, D.: On multivariate runs tests for randomness. *Journal of the American Statistical Association*, **104(488)** (2009), 1525-1538.

- Randles, R. H.: A distribution-free multivariate sign test based on interdirections. *Journal of the American Statistical Association*, **408(84)**, (1989), 1045–1050.
- Randles, R. H., Peters, D.: Multivariate rank tests for the two-sample location problem. *Communications in Statistics-Theory and Methods*, **19(11)** (1990), 4225-4238.
- Serfling, R. J.: Multivariate symmetry and asymmetry. *Encyclopedia of statistical sciences.* (2006).
- Taskinen, S., Oja, H., Randles, R. H.: Multivariate nonparametric tests of independence. *Journal of the American Statistical Association*, **100(471)** (2005), 916-925.
- Um, Y., Randles, R. H.: Nonparametric tests for the multivariate multi-sample location problem. *Statistica Sinica* (1998), 801-812.