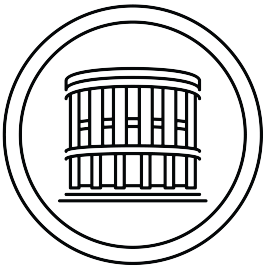
UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



NEPARAMETRICKÉ TESTY POLOHY VO VIACROZMERNOM SVETE

DIPLOMOVÁ PRÁCA

2025 Bc. Lucia TÓTHOVÁ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

NEPARAMETRICKÉ TESTY POLOHY VO VIACROZMERNOM SVETE

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	Matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Mgr. Ján Somorčík, PhD.

Bratislava 2025 **Bc. Lucia TÓTHOVÁ**





Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Lucia Tóthová
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Neparametrické testy polohy vo viacrozmernom svete
Non-parametric location tests in multidimension

Anotácia: Popri klasickom Hotellingovom teste odvodenom za predpokladu normality dát existujú v literatúre mnohé viac či menej úspešné postupy, ako testovať hypotézy o polohe jedného či viacerých súborov viacrozmerných dát. Ich výhodou býva najmä odolnosť voči outlierom (t. j. robustnosť). Pridanou hodnotou je aj samotný prívlastok neparametrické, ktorý znamená, že netreba poznať rozdelenie vstupných dát.

Vedúci: Mgr. Ján Somorčík, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: doc. Mgr. Igor Melicherčík, PhD.

Dátum zadania: 12.01.2024

Dátum schválenia: 25.01.2024

prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Rada by som vyslovila veľkú vďaku svojmu vedúcemu diplomovej práce Mgr. Jánovi Somorčíkovi, PhD. za čas, ochotu, podnetné pripomienky a motiváciu pri písaní tejto práce. Ďakujem tiež všetkým ľuďom, ktorí ma podporovali počas ostatných niekoľkých rokov, či už akademicky alebo emočne. Išli sme cestou síce dlhšou ale o to náročnejšou, nezabúdajúc, že pomaly ďalej zájdeš ale rýchlejšie tam budeme skôr.

Abstrakt

TÓTHOVÁ Lucia: Neparametrické testy polohy vo viacrozmernom svete [Diplomová práca] Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Ján Somorčík, PhD., Bratislava, 2025, 74 s.

V tejto práci sa zaoberáme neparametrickými testami polohy vo viacrozmernom prostredí. V úvode popíšeme základnú teóriu potrebnú na štúdium tejto problematiky. Prvá časť práce je venovaná Mathurovmu testu, ktorý testuje hypotézu o polohe jedného viacrozmerného dátového súboru. Tento test vychádza z Blumenovho neparametrického testu polohy, ktorý využíva štandardizáciu dát na jednotkovú kružnicu. Pokúsime sa reprodukovať výsledky uvedené v Mathurovom článku a vytvoríme pri tom osem variantov testovej štatistiky. V druhej časti sa venujeme dvojsúborovému testu Mathura a Smithovej, ktorý porovnáva priemernú vzdialenosť medzi dvoma dátovými súbormi s priemernou vzdialenosťou dátových bodov v rámci jednotlivých súborov. Ukazujeme, že testová štatistika konverguje k U-štatistike, pričom opravujeme niektoré nepresnosti z pôvodného článku a rozvíjame teóriu o hľadaní asymptotického rozdelenia príslušnej U-štatistiky. Ďalej predstavujeme vlastnú upravenú verziu testu a porovnávame jej výkonnosť s viacerými inými viacrozmernými neparametrickými testami polohy.

Kľúčové slová: Neparametrické testy polohy, Blumenov test, U-štatistika, dvojsúborové testy polohy

Abstract

TÓTHOVÁ Lucia: Non-parametric location tests in multidimension [Diploma Thesis] Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Ján Somorčík, PhD., Bratislava, 2025, 74 p.

In this thesis, we study nonparametric location tests in a multivariate setting. In the introduction, we outline the basic theory necessary for understanding this topic. The first part of the thesis focuses on Mathur's test, which evaluates the hypothesis concerning the location of a single multivariate data sample. This test is based on Blumen's nonparametric location test, which uses standardization of data onto the unit sphere. We attempt to reproduce the results presented in Mathur's paper and, in doing so, construct eight variants of the test statistic. In the second part, we focus on the two-sample test of Mathur and Smith, which compares the average distance between two data samples with the average distance of points within separate samples. We show that the test statistic converges to a U-statistic, correct some inaccuracies in the original paper, and further develop the theory related to the asymptotic distribution of the corresponding U-statistic. Finally, we propose our own modified version of the test and compare its performance with several other multivariate nonparametric location tests.

Keywords: Nonparametric location tests, Blumen's test, U-statistic, two-sample location tests

Obsah

Úvod **9**

1 Základné pojmy **11**

1.1 Blumenov test 11

1.2 Dvojvýberový Hottelingov T^2-test 13

1.3 Teória U-štatistík 14

2 Mathurov S-test **17**

2.1 Mathurov navrhovaný test 17

2.2 Čomu udeliť rank? 20

2.3 Ako udeliť rank? A ako zoradiť dáta? 20

2.4 Kritické hodnoty 22

2.5 Aplikácia Mathurovho S-testu 23

2.6 Hľadanie postupu 24

2.7 Overovanie sily Mathurovych S-testov 28

3 Mathur-Smithovej T štatistika **32**

3.1 Asymptotické rozdelenie 43

3.2 Ilustrácia asymptotického rozdelenia 46

3.2.1 Asymptotické rozdelenie a neplatnosť H_0 50

3.3 Aplikácia T-štatistiky 53

3.4 Naša verzia Mathur-Smithovej testu 59

3.5 Simulačné porovnanie našej verzie Mathur-Smithovej testu a iných testov . 63

Záver **71**

Zoznam použitej literatúry **73**

Príloha A **75**

A.1 Ilustrácia Bluemnovej metódy **76**

A.2 Ilustrácia S-testov s rankovaním 1 **77**

A.2.1 S^1_α 77
A.2.2 S^1_γ 78

A.3 Ilustrácia S-testov s rankovaním 2 **79**
A.3.1 S^2_α 79
A.3.2 S^2_β 80
A.3.3 S^2_γ 81
A.3.4 S^3_α 82
A.3.5 S^3_β 83
A.3.6 S^3_γ 84

Úvod

Jedným zo základných a najdôležitejších krokov skúmania dát je otázka ich polohy, teda toho, kde v priestore sa nachádzajú. Poloha rozdelenia predstavuje kľúčovú charakteristiku, ktorá má praktický význam v mnohých oblastiach: od biologických meraní, cez obrazové dáta až po analýzu finančných trhov. Testy polohy nám v prípade jedného dátového súboru umožňujú formálne testovať hypotézy o tom, či dáta majú konkrétnu strednú hodnotu, respektíve stred symetrie, v prípade viacerých dátových súborov zas testujú hypotézy o tom, či dáta majú rovnakú strednú hodnotu, alebo naopak, medzi nimi existujú systematické rozdiely.

Práve mnohé snahy popísať svet okolo nás a jeho zákonitosti a štruktúry nám ukazujú, že na popísanie komplikovaných systémov a princípov nám nestačí skúmať každú premennú samostatne. Jednorozmerné testy môžu viesť k neúplným či skresleným záverom, ktoré neberú do úvahy závislosti medzi rôznymi premennými. Poznáme aj spôsoby, ako testovať jednorozmernými testami hypotézy o viacrozmernej strednej hodnote súčasne, no žiaden z nich nerieši problém straty informácie o korelácii medzi premennými. Sila týchto testov preto často býva nedostatočná. Viacrozmerné testy polohy tieto nedostatky prekonávajú tým, že pracujú s dátovým súborom ako s celkom a nestrácajú tým informácie o tom, aké sú vzťahy medzi jednotlivými premennými.

Najznámejším viacrozmerným testom polohy dát je Hotellingov T^2-test, respektíve jeho dvojsúborová verzia. Tento test pracuje s predpokladom normality dát. V praxi je však často nemožné či nežiaducce predpokladať, že dáta pochádzajú z normálneho rozdelenia. Dát môžeme mať k dispozícii málo, prípadne môžu pochádzať z iného ako normálneho rozdenia. Práve v takýchto situáciách môžeme siahnuť po neparametrických testoch, ktoré na naše dáta nekladú takéto nároky. Ďalšou výhodou neparametrických testov je ich robustnosť, teda odolnosť voči odľahlým pozorovaniam.

Problematike viacrozmerných neparametrických testov polohy sa už venoval nejeden autor pred nami a v literatúre možno nájsť mnoho testov, ktorými získame viac či menej spoľahlivé odpovede na naše otázky. Uvedieme a popíšeme niektoré známe testy a pozrieme sa bližšie na niektoré menej známe. Cieľom našej bude pochopiť a prípadne upraviť niektoré už existujúce testy. Špeciálnu pozornosť budeme venovať článku od Sunila K. Mathura [3] a článku od Sunila K. Mathura a Pamely F. Smithovej [4]. V prípade

článku [3] podrobne rozoberieme navrhnutú testovú štatistiku a pokúsime sa zistiť, ako sa má tento test používať v praxi. Pri práci s článkom [4] napravíme niektoré v ňom uvedené nepresnosti, rozvinieme ďalej časť teórie o asymptotických rozdeleniach a navrhneme vlastnú verziu tohto testu, ktorej silu otestujeme pomocou simulačných štúdií, kde ju porovnáme s inými známymi testami.

1 Základné pojmy

V tejto kapitole popíšeme teóriu a známe testy, s ktorými budeme v našej práci pracovať.

1.1 Blumenov test

Ako sme uviedli v úvode, často sa stáva, že pracujeme s dátami, pre ktoré predpoklad normality nie je splnený, prípadne máme dáta s odľahlými pozorovaniami. V týchto prípadoch prichádzajú na rad neparametrické testy.

V roku 1965 predstavil Isadore Blumen v článku [2] svoj dvojrozmerný neparametrický test polohy, ktorý sa neskôr stal akýmsi štandardom kvality pre tento typ testov.

Autor v článku pracuje s dvojrozmernými dátami a skúma hypotézu o tom, že dáta pochádzajú z rozdelenia s nulovou strednou hodnotou. Označme dátové body:

$$X_i = (x_i, y_i)^T,\ i = 1, 2, \ldots, n.$$

Ako robustný odhad strednej hodnoty si Blumen vybral medián a testoval hypotézu

$$H_0 :\ median(X) = (0,0)^\top \quad \text{oproti} \quad H_1 : median(X) \neq (0,0)^\top.$$

Pokiaľ nulová hypotéza platí, približne rovnako veľa dát je „napravo“ od nuly, ako ich je „naľavo.“ Ako ale niečo takéto merať v dvojrozmernom svete? A čo korelácia medzi rozmerni?

Definícia 1.1. *Pre dátový bod X_i definujeme dátovú priamku ako priamku so sklonom*

$$m_i = \frac{y_i}{x_i},$$

ktorá prechádza bodom $(0,0)^\top$ a bodom X_i.

Každý dátový bod teda vieme jednoznačne určiť sklonom m_i, jeho vzdialenosťou od bodu $(0,0)^\top$ a znamienkom niektorej súradnice. Ako znamienko použijeme $sign(y_i)$. Každý dátový bod presunieme po jeho dátovej priamke tak, aby ležal na priesečníku dátovej priamky a jednotkovej kružnice. Dátové body so zápornou y-ovou zložkou navyše stredovo symetricky preklopíme cez bod $(0,0)^\top$, teda budú ležať na „opačnom“ priesečníku svojej dátovej priamky a jednotkovej kružnice.

Definícia 1.2. *Zoradenie dát $X_1, X_2, \ldots, X_n$ je vytvorenie novej postupnosti*

$$\tilde{X_1}, \tilde{X_2}, \ldots, \tilde{X_n},$$

ktorá je permutáciou pôvodných dátových bodov $X_1, X_2, \ldots, X_n$ spĺňajúca určitú vopred zvolenú podmienku zoradenia.

V Bluemenovom teste dáta zoradíme na základe m_i. Ako $\tilde{X_1}$ označíme ten dátový bod, ktorého dátová priamka zviera s kladnou časťou osi x najmenší uhol, ako $\tilde{X_2}$ ten bod, ktorého dátová priamka zviera druhý najmenší uhol s kladnou časťou osi x, a tak ďalej, až dátová priamka posledného dátového bodu $\tilde{X_n}$ bude s touto polpriamkou zvierať najväčší uhol (pričom sú všetky tieto uhly v intervale $[0, \pi]$ rad). Z hľadiska hodnôt m_i teda bude mať prvý dátový bod najmenšiu kladnú hodnotu, pokračujeme po kružnici až po najväčšiu kladnú hodnotu m_i zo vzorky, potom pokračujeme najmenšou (najzápornejšou) hodnotou m_i, až posledný bude ten dátový bod, ktorý má hodnotu m_i najbližšie 0 zľava.

Ďalším krokom je odstránenie korelácie. Dátové priamky upravíme tak, aby bola uhlová vzdialenosť susediacich priamok vždy rovnaká, $\frac{\pi}{n}$ rad. Pre prehľadnosť upravíme dátovú priamku, ktorá zviera najmenší uhol s kladnou x-ovou osou tak, že jej udelíme práve sklon $0rad$. Zoradeným dátam sme teda priradili uhly

$$\tau_1 = 0,\ \tau_2 = \frac{1\pi}{n}, \ldots, \tau_n = \frac{(n-1)\pi}{n}.$$

Vďaka tejto štandardizácii bude náš test nezávislý na jednotkách, v ktorých sú dáta merané, aj na korelácií medzi premennými. Dá sa dokázať, že tento test je afínne invariantný. Nakoniec dáta stredovo symetricky preklopíme cez súradnicový počiatok, ak to-ktoré dáto malo pôvodnú y-ovú súradnicu zápornú.

Za platnosti nulovej hypotézy sú dáta približne rovnomerne rozmiestnené okolo počiatku súradnicovej sústavy. Pre každú priamku by teda malo byť náhodné, či savýsledný dátový bod nachádza nad x-ovou osou, alebo pod ňou a tieto udalosti by mali nastávať s rovnakou pravdepodobnosťou. Inak napísané, ak platí H_0, potom

$$P(sign(y_i) = 1) = P(sign(y_i) = -1) = \frac{1}{2}.$$

Takto sme získali n dátových bodov:

$$\tilde{X_i} = a_i \begin{pmatrix} \cos\left(\frac{\pi(i-1)}{n}\right) \\ \sin\left(\frac{\pi(i-1)}{n}\right) \end{pmatrix},$$

kde

$$a_i = \begin{cases} -1 & \text{ak } y_i < 0, \\ 1 & \text{inak .} \end{cases}$$

Všetky tieto body sú vzdialené 1 od bodu $(0,0)^\top$. Ak každému tomuto bodu priradíme „hmotnosť“ 1, získame n-uholník, ktorého ťažisko by malo byť za platnosti nulovej hypotézy blízko bodu $(0,0)^\top$. Jednotlivé súradnice $\frac{v_1}{n}$ a $\frac{v_2}{n}$ tohto ťažiska môžeme získať pomocou nasledujúcich premenných:

$$\frac{v_1}{n} = \frac{1}{n}\sum_{j=0}^{n-1} a_j cos\left(\frac{\pi j}{n}\right) \quad \text{a} \tag{1}$$

$$\frac{v_2}{n} = \frac{1}{n}\sum_{j=0}^{n-1} a_j sin\left(\frac{\pi j}{n}\right), \tag{2}$$

kde

$$a_j = \begin{cases} -1 & \text{ak } y_{j+1} < 0, \\ 1 & \text{inak .} \end{cases}$$

Definície v_1, v_2 a a_j sme teraz uviedli konzistentne s pôvodným článkom [2]. Na prvý pohľad pôsobí mierne nejasne zisťovanie hodnoty a_j na základe y_{j+1}, teda na základe dátového bodu s iným indexom, no tento posun je spôsobený tým, že sumy v rovniciach (1) a (2) sú definované pre $j = 0, 1, \ldots, (n-1)$.

Dá sa ukázať, že za platnosti nulovej hypotézy sú v_1 a v_2 približne nezávislé normálne premenné so strednou hodnotou 0 a varianciou $\frac{n}{2}$. Pokiaľ nulová hypotéza neplatí, stredná hodnota jednej alebo oboch týchto premenných bude nenulová. Môžeme teda zaviesť testovú štatistiku

$$v^2 = \frac{2(v_1^2 + v_2^2)}{n},$$

ktorá sa za platnosti H_0 správa podľa χ^2 rozdelenia s 2 stupňami voľnosti.

1.2 Dvojvýberový Hottelingov T^2-test

Pri skúmaní polohy viacrozmerných dát je Hottelingov T^2-test asi najznámejšou metódou, ktorého dvojvýberovú verziu popíšeme konzistentne s [1]. Pri aplikácii tohto testu využívame predpoklad, že dáta pochádzajú z p-rozmerných normálnych rozdelení s rovnakou kova-

riančnou maticou, teda

$$X_{11}, X_{12}, \ldots, X_{1n} \sim \mathcal{N}_p(\mu_1, \Sigma),$$
$$X_{21}, X_{22}, \ldots, X_{2m} \sim \mathcal{N}_p(\mu_2, \Sigma).$$

Testujeme hypotézu o tom, že stredná hodnota normálnych rozdelení je rovnaká

$$H_0 : \ \mu_1 = \mu_2 \quad \text{oproti} \quad H_1 : \ \mu_1 \neq \mu_2.$$

Pre vyhodnotenie testu potrebujeme odhadnúť strednú hodnotu a kovariančnú maticu oboch rozdelení, ktoré získame ako výberový priemer a výberovú kovariančnú maticu oboch súborov. Tie sa z realizácií $X_1, X_2, \ldots X_n$ počítajú podľa nasledujúcich vzorcov:

$$\overline{X} = \sum_{i=1}^{n} X_i$$
$$S = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})^T.$$

Z prvého a druhého dátového súboru spočítame výberové priemery $\overline{X}_1, \overline{X}_2$ a výberové kovariančné matice S_1, S_2. Pretože predpokladáme, že obe dátové sady sú realizáciou náhodných rozdelení s rovnakou kovariančnou maticou, odhadneme ju váženým priemerom výberových kovariančných matíc

$$S = \frac{(n-1)S_1 + (m-1)S_2}{n+m-2}.$$

Potom môžeme vypočítať testovú štatistiku

$$T = \frac{nm(n+m-2)}{(n+m)^2} (\overline{X}_1 - \overline{X}_2)^\top S^{-1} (\overline{X}_1 - \overline{X}_2).$$

pričom platí, že testová štatistika T má Hottelingovo T^2 rozdelnie s parametrami p a $n+m-2$. Nulovú hypotézu zamietame, ak je hodnota T vyššia ako kritická hodnota Hotellingovho T^2 rozdelenia s parametrami p a $n+m-2$ na hladine významnosti α. Vidno, že tento test je akýmsi zovšeobecnením dvojsúborového Studentovho t- testu do viacrozmerného sveta.

1.3 Teória U-štatistík

V tejto podkapitole popíšeme U-štatistiky a niektoré ich vlastnosti. Predstavíme pojem U štatistiky s degenerovaným kernelom a uvedieme jej asymptotické rozdelenie. Špeciálne

sa budeme venovať U-štatistikám druhého rádu a naším hlavným zdrojom bude kniha [9].

Predpokladajme, že máme náhodný výber zo spojitého rozdelenia s distribučnou funkciou F:

$$X_1, X_2, \ldots, X_n.$$

Pomocou tohto výberu (našej dátovej sady) by sme chceli odhadnúť nejaký parameter θ rozdelenia F. V našom prípade

$$\theta = E_F(h(X_1, X_2, \ldots, X_r))$$

pre nejakú kernelovú funkciu (skrátene kernel) h, ktorá má $r \leq n$ argumentov. Predpokladá sa, že táto kernelová funkcia je permutačne symetrická vo svojich argumentoch.

Definícia 1.3. *Symetrickej kernelovej funkcii $h = h(x_1, x_2, \ldots, x_r)$ prislúcha U-štatistika daná predpisom*

$$U_n = \frac{1}{\binom{n}{r}} \sum_c h(X_{i_1}, X_{i_2}, \ldots, X_{i_r}),$$

pričom $\sum_c$ značí sumáciu cez $\binom{n}{r}$ odlišných kombinácií prvkov $\{i_1, i_2, \ldots, i_r\}$ z $\{1, \ldots, n\}$.

Z definícií θ a U_n ľahko vidno, že U_n je skutočne nevychýlený odhad pre θ. V našej diplomovej práci sa budeme zaoberať U-štatistikami s kernelmi, ktoré vyberajú práve 2 prvky zo sady, teda $r = 2$. V takomto špeciálnom prípade môžeme zapísať

$$U_n = \frac{1}{\binom{n}{2}} \sum_{1 \leq i <} \sum_{j \leq n} h(X_i, X_j).$$

Aby sme mohli pracovať s U-štatistikou a skúmať jej vlastnosti, zadefinujeme *pridružené funkcie*

$$h_1(x_1) = E_F(h(x_1, X_2))$$

a

$$\tilde{h}_1 = h_1 - \theta$$

Funkcia $\tilde{h}_1$ centruje funkciu h_1 na nulovú strednú hodnotu, lebo sa dá dokázať, že $E_F(h_1(X_1)) = \theta$.

Definícia 1.4. [9, str. 177] *Definujme* $\zeta_0 = 0$ *a*

$$\zeta_1 = Var_F(h_1(X_1)) = E_F(\tilde{h}_1^2(X_1)).$$

Rovnosť v tejto definícii platí práve vďaka spomínanému $E_F(h_1(X_1)) = \theta$. Dá sa tiež dokázať, že platí $0 = \zeta_0 \leq \zeta_1 \leq Var_F(h(X_1, X_2))$.

Definícia 1.5. *Ak pre U-štatistiku* U_n *platí* $\zeta_1 = 0$, *hovoríme, že* U_n *je U-štatistika s degenerovanou kernelovou funkciou.*

Pretože budeme chcieť pomocou U-štatistík U_n vykonávať testy, budeme potrebovať poznať ich asymptotické rozdelenia. Tým sa budú venovať nasledujúce dve Vety, ktoré sú špeciálnymi prípadmi výsledkov zo Serflingovej knihy [9] pre $r = 2$.

Veta 1.6. *Ak platí* $Var_F(h(X_1, X_2)) < \infty$ *a* $\zeta_1 > 0$, *potom*

$$n^{1/2}(U_n - \theta) \xrightarrow{d} N(0, r^2\zeta_1), \text{ kde } r = 2.$$

U_n *je teda asymptoticky normálna náhodná premenná. Symbol* $\xrightarrow{d}$ *značí konvergenciu podľa distribučnej funkcie pre* $n \to \infty$.

Veta 1.7. *Ak platí* $0 < Var_F(h(X_1, X_2)) < \infty$ *a* $\zeta_1 = 0$, *potom asymptotické rozdelenie* U_n *nie je normálne ale označuje sa Gausovský chaos. Platí*

$$n(U_n - \theta) \xrightarrow{d} \frac{r(r-1)}{2} Y,$$

kde $r = 2$ *a*

$$Y = \sum_{j=1}^{\infty} \lambda(\chi_{1j}^2 - 1),$$

kde $\chi_{11}^2, \chi_{12}^2, \ldots$ *sú nezávislé* χ_1^2 *premenné a* $\lambda_1, \lambda_2, \ldots$ *sú vlastné hodnoty prislúchajúce rôznym vlastným funkciám Fredholmovej funkcionálnej rovnice*

$$\lambda\varphi(x) = \int h(x, y)\varphi(y)dF(y).$$

Hoci je vo všeobecnosti Y dané sumou obsahujúcou nekonečno členov, v praxi môže nastať prípad, kedy má Fredholmova rovnica konečný počet vlastných hodnôt. Potom toto limitné rozdelenie možno ľahko odhadnúť pomocou počítačových simulácií.

2 Mathurov S-test

Sunil K. Mathur je autorom a spoluautorom viacerých článkov, ktoré navrhujú rôzne neparametrické testy polohy pre viacrozmerné dáta. V tejto časti sa podrobne pozrieme na článok [3] z roku 2006, v ktorom autor navrhuje neparametrický test polohy pre dvojrozmerné dáta. Tento test budeme označovať Mathurov S-test.

Ako Mathur uvádza vo svojom článku, tento test je inšpirovaný Blumenovym testom (ktorý sme popísali v kapitole 1.1) so snahou vylepšiť ho a lepšie využiť informácie o dátach. S týmto zámerom autor uvádza rankovú verziu tohto testu.

Intuícia by nám mohla napovedať, že ak by sme chceli dať dátam rôznu váhu a nejako určiť ich rank, mohli by sme zohľadniť napr. vzdialenosť dát od počiatku súradnicovej sústavy. Za zmienku stojí, že takýto test, ktorý sa zároveň zaoberá hypotézami vo viac ako dvojrozmernom priestore, publikovali napríklad D. Peters a R. H. Randles v roku 1990 v článku [5].

Mathur vo svojom článku [3] popisuje test, ktorý dátam udeľuje rank podľa veľkosti uhla, ktorý zviera jeho dátová priamka s kladnou x-ovou poloosou. Začnime však od začiatku.

2.1 Mathurov navrhovaný test

Nech sú $(X_{1i}, Y_{1i})^\top,\ i = 1, 2, \ldots, n$ náhodným výberom n nezávislých dát z dvojrozmernej populácie s hladkou distribučnou funkciou (cdf) $F(x + \delta_1, y + \delta_2)$. Nech je populácia symetrická okolo mediánu $\delta = (\delta_1, \delta_2)^\top$. Potom môžeme testovať hypotézu:

$$H_0 :\ (\delta_1, \delta_2)^\top = (\delta_{11}, \delta_{22})^\top \quad \text{oproti} \quad H_1 : (\delta_1, \delta_2)^\top \neq (\delta_{11}, \delta_{22})^\top$$

pre konkrétne hodnoty $(\delta_{11}, \delta_{22})^\top$.

Označme $X_i := X_{1i} - \delta_{11},\ Y_i := Y_{1i} - \delta_{22}$. Podobne ako v Blumenovom teste definujme dátové priamky $m_i = Y_i/X_i$ a označme $\theta_i = arctg(m_i)$. Samotné dátové body teda môžeme definovať ako kombináciu dátovej priamky, vzdialenosti od bodu $(0, 0)$ a znamienka $sign(Y_i)$. O znamienkach môžeme urobiť podobnú úvahu, ako pri Blumenovom teste. Pokiaľ je medián dát skutočne $(\delta_{11}, \delta_{22})^\top$, ako predpokladá nulová hypotéza, pre každú priamku by mala byť rovnaká pravdepodobnosť toho, že sa na nej ležiaci dátový

bod nachádza nad x-ovou osou, ako že sa nachádza pod ňou. Za platnosti H_0 potom platí $P(Y_i > 0) = P(Y_i < 0) = \frac{1}{2}$.

Ďalší postup autor popisuje nasledovne: *„Now we arrange the angles in ascending order with respect to positive horizontal axis. To make test statistic independent of correlation between two variates, we change slope of adjacent vectors so that angle between vectors represented by adjacent slopes is 180/n degrees.“* „Teraz zoradíme uhly vzostupne vzhľadom na kladnú x-ovú os. Aby bola naša testová štatistika nezávislá od korelácii dvoch premenných, zmeníme sklon susedných vektorov tak, že uhol daný susednými sklonmi je $180/n$ stupňov.“ V texte bohužiaľ nie je uvedené presnejšie, ako má prebehnúť takéto zoradenie dát. Na rozdiel od pôvodného Blumenovho testu Mathur nespomína žiadne preklápanie bodov nad x-ovú os. Presným nejasnostiam a ich následkom sa budeme venovať neskôr.

Mathur ďalej postupuje úvahou podobne ako Blumen, teda nájde priesečníky dátových priamok a jednotkovej kružnice a tie považuje za nové dátové body. Za platnosti H_0 by ťažisko takto získaného útvaru malo byť blízko bodu $(0, 0)$.

V článku je ďalej popísaný Blumenov test (kap. 1.1) s dovetkom „*In the construction of proposed test statistic, we will try to use ranks of m_i's as well as signs of Y_i's in an ordered m_i's. We believe that inclusion of ranks of θ_is will provide more information about population and make the test statistic more powerful than its competitors.*“ „Pri tvorbe navrhovanej testovej štatistiky skúsime využiť ranky m_i, rovnako ako znamienka Y_i pre zoradené m_i. Veríme, že zahrnutie rankov θ_i nám dodá viac informácií o populácii a naša testová štatistika bude mať väčšiu silu než jej súperi.“ Tento text napovedá, že Mathur zoradil dáta rovnakým spôsobom, ako Blumen. Bohužiaľ, ako ukážeme neskôr, skutočné použité zoradenie nie je jednoduché určiť.

Pri popísaní logiky využitia rankov sa autor odkazuje na svoj starší článok, ktorý uverejnil s K. Senom v roku 1997 [6]. Tomuto článku a nezrovnalostiam v ňom sa vo svojej diplomovej práci venoval P. Somogyi [7].

Testové štatistiky sú ďalej definované ako

$$S_1 = \frac{1}{n} \sum_{i=1}^{n} A_i \cos(\pi \frac{i}{n}) Rank(\theta_i) \tag{3}$$

a

$$S_2 = \frac{1}{n}\sum_{i=1}^{n} A_i \sin(\pi\frac{i}{n}) Rank(\theta_i) \tag{4}$$

kde

$$A_i = \begin{cases} 1 & \text{ak } Y_i > 0 \text{ v } (i+1)\text{-vom zoradenom sklone,} \\ -1 & \text{ak } Y_i < 0 \text{ v } (i+1)\text{-vom zoradenom sklone.} \end{cases}$$

Podobne ako v kapitole 1.1 sa môžeme pozastaviť nad tým, prečo by sme používali znamienka prislúchajúce nie tomu-ktorému dátovému bodu, ale „susednému.“ Domnievame sa, že ide o chybu vzniknutú pri prepise definície z článku [2], kde boli testové štatistiky definované sumou pre $i = 0, 1, \ldots, (n-1)$. V tomto prípade by bola pre $i = n$ premenná A_i nedefinovaná.

Za platnosti nulovej hypotézy platí

$$P(A_i = 1) = \frac{1}{2} = P(A_i = -1).$$

Navrhovaná testová štatistika je

$$S = S_1^2 + S_2^2.$$

Autor ďalej odvádza, že pokiaľ by sme testové štatistiky S_1 a S_2 normalizovali, potom by sa testová štatistika

$$Z^2 = \left(\frac{S_1 - E(S_1)}{\sqrt{Var(S_1)}}\right)^2 + \left(\frac{S_2 - E(S_2)}{\sqrt{Var(S_2)}}\right)^2$$

riadila χ^2 rozdelením s dvomi stupňami voľnosti. Postup normalizácie však má určité technické nedostatky, napríklad disperzia testovej štatistiky S_1 je vypočítaná ako

$$Var(S_1) = \frac{1}{2n^2}\sum_{i=1}^{n} \cos\left(\frac{360i}{n}\right) \mathrm{Rank}^2(\theta_i) + \frac{n}{6}\left(1+\frac{1}{n}\right)\left(1+\frac{1}{2n}\right),$$

čo zjavne nie je korektné tvrdenie, keďže premenná $Rank(\theta_i)$ je vo všeobecnosti náhodná premenná. Toto tvrdenie sa nám bohužiaľ nepodarilo uviesť na pravú mieru.

Možno aj preto autor textu neodporúča vyhodnocovať tento test na hladine významnosti α porovnaním hodnoty Z^2 s $(1-\alpha)$ kvantilom rozdelenia χ^2_2. Miesto toho však v článku [3] uvádza tabuľku kritických hodnôt, s ktorými máme pri použití testu porovnať testovú štatistiku S. Túto tabuľku uvedieme v kapitole 2.4 ako Tabuľku 1 a vyjadríme sa k nej bližšie.

Posledná časť článku uvádza konkrétny príklad použitia tohto testu, bohužiaľ však autor uvádza iba vstupné dáta a hodnoty testových štatistík, zostáva tak veľa priestoru pre odhadovanie, ako sa má tento test teda používať.

V tomto texte sme našli viacero nezrovnalostí a nejasností, postupne ich teda rozoberieme a uvedieme naše závery. Pri hľadaní odpovedí na niektoré otázky sme sa pokúsili využiť jeden príklad aplikácie tejto testovej štatistiky, ktorý Mathur uviedol vo svojom článku [3]. Tejto aplikácii sa budeme podrobnejšie venovať v kapitole 2.5.

2.2 Čomu udeliť rank?

Autor v texte nie je konzistentný v tom, či ranky udeľujeme sklonom m_i alebo uhlom $\theta_i = arctg(m_i)$, vidno však, že funkcia $arctg(x)$ je neklesajúca. Teda ak udelíme ranky sklonom m_i vzostupne podľa ich veľkosti, nebudú sa líšiť od toho, keby sme vzostupne udelili ranky uhlom θ_i. Táto nejasnosť teda neovplyvňuje samotný test a získané testové štatistiky. Ďalej v tejto práci teda budeme pracovať s $Rank(\theta_i)$.

2.3 Ako udeliť rank? A ako zoradiť dáta?

Kľúčovými aspektami pre výpočet testovej štatistiky S sú spôsob rankovania a spôsob zoradenia dát. V článku [3] Mathur hovorí iba o tom, že využíva znamienka $sign(Y_i)$ a ranky $Rank(\theta_i)$ pre zoradené m_i.

Samotná myšlienka udeľovania ranku dátam je založená na tom, že uhly θ_i nejako zoradíme od najmenšieho po najväčší a tak im udelíme rank. Ak uvažujeme definíciu $\theta_i = arctg(m_i)$, parameter θ_i nadobúda hodnoty z intervalu $[-\pi/2, \pi/2]$. Ak berieme do úvahy iba tento parameter, nevieme rozlíšiť z ktorého kvadrantu to-ktoré dáto pochádza. V prípade, že je hodnota θ_i kladná, dáto môže pochádzať z 1. alebo 3. kvadrantu. Ak je hodnota θ_i záporná, dáto pochádza z 2. alebo 4. kvadrantu. Ak by sme sa na túto problematiku pozreli „blumenovským“ spôsobom, dáta by sme premietli do 1. a 2. kvadrantu. V takomto prípade by sme ranky udelili najprv dátovým bodom s kladnými hodnotami θ_i a potom dátovým bodom so zápornými hodnotami tohto parametra, v oboch prípadoch od najmenšej po najväčšiu. Takéto rankovanie budeme označovať **rankovanie 1.**

Ďalší spôsob rankovania, ktorým sa zaoberáme, je rankovanie konzistentné s pôvodnými hodnotami θ_i, teda rank určíme vzostupne pre $\theta_i \in [-\pi/2, \pi/2]$ bez ďalšej normalizácie

alebo dodatočného zoradenia. To je ekvivalentné tomu, keby sme dáta premietli do 1. a 4. kvadrantu. Keďže autor bližšie nepopisuje spôsob rankovania, takýto priamočiary prístup sa zdá byť vhodný. Toto rankovanie budeme označovať ako **rankovanie 2**. Pôvodne sme pracovali iba s týmto prístupom, pri výpočtoch sme však zistili, že naše výsledky sú veľmi odlišné od tých, ktoré získal Mathur vo svojom článku a to nás podnietilo v ďalšom bádaní v tomto smere.

Tretí prístup k udeľovaniu ranku je podľa skutočnej polohy dát. Uvažujme z každej dátovej priamky iba tú polpriamku, na ktorej leží dátový bod $(X_i, Y_i)^\top$. Začnime priamo na kladnej x-ovej polosi a postupujme proti smeru hodinových ručičiek. Bodu, ktorého dátovú polpriamku stretneme ako prvú, udelíme rank 1. Bodu, ktorého dátová polpriamka má najmenšiu uhlovú vzdialenosť (meranú proti smeru hodinových ručičiek) od bodu s rankom 1, udelíme rank 2. Takto postupne proti smeru hodinových ručičiek udelíme ranky všetkým dátovým bodom, až po rank n. Takéto udeľovanie ranku budeme označovať **rankovanie 3**. Rankovanie 3 je podobné ako rankovanie 1, no berie do úvahy pôvodnú polohu dát, nie len ich verzie premietnuté do jednej polroviny.

Po popísaní troch rôznych spôsobov rankovania popíšeme rôzne zoradenia. Zoradenie určí v rovnostiach (3) a (4) znamienko A_i, teda či budeme uvažovať ten-ktorý štandardizovaný dátový bod nad x-ovou osou alebo pod ňou, rovnako ako hodnotu $Rank($ $theta_i)$, teda váhu, ktorú udelíme danému štandardizovanému bodu. Je dôležité podotknúť, že z ilustračného príkladu použitia S-testu v článku [3] poznáme iba vstupné dáta a hodnotu testovej štatistiky S, správny spôsob rankovania a správny spôsob zoradenia teda hľadáme súčasne a nemáme spôsob, ako najprv nájsť spôsob rankovania a potom spôsob zoradenia, prípadne naopak. Môžeme sa však inšpirovať Mathurovymi ďalšími článkami (napr. [4] alebo [6]) či diplomovou prácou [7].

Podobne ako v prípade rankovania, zoradeniu dát, resp. im prináležiacich sklonov, sa autor tohto testu špecificky nevenuje. Môžeme uvažovať, že priamo prevzal zoradenie od Blumena z článku [2]. Toto zoradenie podrobne popisujeme v kapitole 1.1. Takéto zoradenie dáva zmysel najmä z geometrického uhla pohľadu. Blumenov test je založený na hľadaní ťažiska n-uholníka daného normalizovanými bodmi na jednotkovej kružnici. Vzorcami (1) a (2) z kapitoly 1.1 teda zistíme x-ovú a y-ovú súradnicu normalizovaného bodu iba na základe poradia bodov a znamienok Y_i. Toto zoradenie budeme označovať

zoradenie α.

Zoradeniu dát sa v skutočnosti venuje aj celá problematika udeľovania ranku. Rôzne spôsoby rankovania sú rôzne spôsoby, ako zoradiť dáta, resp. uhly dátových priamok. Môžeme teda skúmať aj také zoradenie, ktoré je totožné s rankom. Takéto zoradenie budeme nazývať **zoradenie β.** Je nutné dodať, že v takomto prípade platí $Rank(\theta_i) = i$ a ak by bolo možné urobiť takúto úpravu v pôvodných Mathurovych úvahách, bolo by mierne nečakané ak by túto úpravu sám neurobil.

Posledné zoradenie dát, ktorému sa venujeme, je také, v akom poradí sa dáta nachádzajú v pôvodnej sade dát. Nejde teda o zoradenie v takom zmysle, ako popisoval Blumen vo svojom článku [2]. Hodnotám S_1 a S_2 získaným týmto postupom len ťažko hľadať geometrickú interpretáciu, no podobný prístup Mathur využíva v článku [6]. Toto zoradenie označíme ako **zoradenie γ.** Pôvodne sme túto možnosť nebrali do úvahy, no pri hľadaní podobného výsledku, aký našiel Mathur v ilustračnom príklade, sa ukázal byť prekvapivo užitočný. Tieto jednotlivé prístupy graficky znázorníme v prílohe A.

2.4 Kritické hodnoty

Štatistické testy typicky vyhodnocujume tak, že hodnotu testovej štatistiky porovnáme s nejakou kritickou hodnotou. Výška tejto kritickej hodnoty záleží od hladiny významnosti α na ktorej test vyhodnocujeme a predovšetkým od samotného rozdelenia pravdepodobnosti testovej štatistiky za platnosti H_0. Mnoho kritických hodnôt je tabelovaných, najmä ak ide o známe rozdelenie. Existujú však aj testy, ktorých testová štatistika sa takto priateľsky nespráva. V takom prípade zvyknú autori vytvoriť tabuľku kritických hodnôt napr. permutačným prístupom. Toto urobil aj Mathur v tomto článku a uviedol tabuľku kritických hodnôt, ktorú uvádzame v Tabuľke 1.

Pozastavili sme sa nad tým, že v tejto tabuľke sú uvedené ľavostranné aj pravostranné kritické hodnoty.

Základnou myšlienkou Blumenovho testu bolo hľadanie ťažiska istého n-uholníka. Pokiaľ by toto ťažisko bolo príliš ďaleko od bodu $(0, 0)$, nulovú hypotézu by sme zamietli. Následne spočítame Z^2, čiže štvorec normalizovanej vzdialenosti ťažiska od $(0, 0)^\top$, a ak je príliš veľký, nulovú hypotézu zamietame. Tým pádom v tomto teste uvažujeme pravostrannú kritickú hodnotu danú 95%-ným kvantilom χ^2 rozdelenia s dvomi stupňami

Size	C.V. at 5%	
n	**Left**	**Right**
05	0.04	5.29
07	0.08	8.19
10	0.10	12.17
12	0.11	14.28
15	0.21	18.08
18	0.29	22.40
20	0.30	24.16
25	0.30	29.83

Tabuľka 1: Kritické hodnoty Mathurovho S-testu na 5% hladine vyznamnosti.
Zdroj: [3, str. 89]

voľnosti. Ľavostrannú kritickú hodnotu nepotrebujeme. Ak je hodnota testovej štatistiky malá, znamená to, že ťažisko n-uholníka je veľmi blízko bodu $(0, 0)^\top$. Zmietať nulovú hypotézu na základe malej hodnoty testovej štatistiky by teda bolo rovnaké, ako tvrdiť, že nulová hypotéza vyzerá až príliš vierohodne.

O hodnotách v Tabuľke 1 autor tvrdí, že ich získal pomocou permutačnej teórie, no viac detailov neuvádza. Predpokladáme, že Mathur vygeneroval sady dát, pre ktoré platila nulová hypotéza, vyhodnotil testovú štatistiku S a následne z týchto hodnôt vybral 2.5%-ný kvantil a 97.5%-ný kvantil ako odhady kritických hodnôt rozdelenia jeho testovej štatistiky. Bohužiaľ nevieme, z akých dát

2.5 Aplikácia Mathurovho S-testu

Ako sme už spomenuli, v poslednej časti článku Mathur uvádza príklad aplikácie, ktorý má demonštrovať aplikáciu tohto testu na dáta. Dáta v tomto prípade pochádzajú z [8] a obsahujú merania systolického a diastolického tlaku Peržanov, ktorí žili väčšinu svojho života vo vysokých nadmorských výškach. Dáta uvádzame v tabuľke 2.

Testujeme hypotézu o tom, či títo ľudia majú rovnaký krvný tlak ako zdraví muži

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X_{1i}	170	125	148	140	106	108	124	134	116	114	118	138	134	124	114
Y_{1i}	76	75	120	78	72	62	70	64	76	74	68	78	86	64	66

Tabuľka 2: Dáta o krvnom tlaku ľudí žijúcich vo vysokých nadmorských výškach.

Zdroj: [8]

starší ako 21 rokov žijúci v Spojených štátoch, teda:

$$H_0 : \ (\delta_1, \delta_2) = (120, 80) \quad \text{oproti} \quad H_1 : (\delta_1, \delta_2) \neq (120, 80).$$

Bohužiaľ článok [3] neobsahuje viac informácií o postupe výpočtu testových štatistík, získame z neho iba informácie o ich hodnotách, konkrétne $S_1 = -4.33277$, $S_2 = -0.12824$, $S = 18.78937$.

Spolu so simulovanými kritickými hodnotami môžeme tento výsledok použiť ako akúsi nápovedu pri hľadaní postupu, ktorý by sme podľa autora mali využiť pri vykonávaní testu.

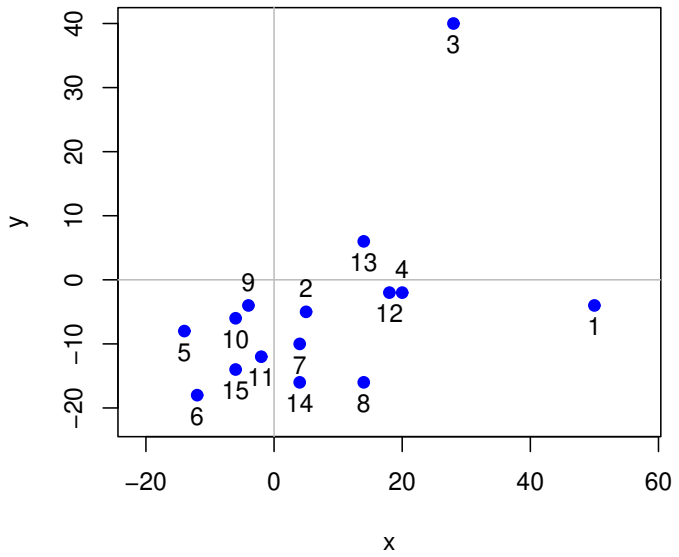
2.6 Hľadanie postupu

Ďalším krokom v bádaní nie je nič iné, ako vyhodnotiť jednotlivé možné postupy a zistiť, ktorý mal autor skutočne na mysli. Použijeme teda rovnaké dáta, ako Sunil K. Mathur v článku [3]. Od týchto dát odčítame vektor $(120, 80)$, čím získame nové súradnice, ktoré budeme označovať $(X_i, Y_i)^\top$ a budeme testovať hypotézu o tom, že stredná hodnota dát je $(0, 0)^\top$. Upravené dáta uvádzame v Tabuľke 3, vykresľujeme na Obr. 1.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X_i	50	5	28	20	-14	-12	4	14	-4	-6	-2	18	14	4	-6
Y_i	-4	-5	40	-2.0	-8	-18	-10	-16	-4	-6	-12	-2	6	-16	-14

Tabuľka 3: Vycentrované dáta o krvnom tlaku ľudí žijúcich vo vysokých nadmorských výškach.

V Tabuľke 4 vidíme, že postup, ktorého výsledky sa najviac podobajú na tie z článku [3], je rankovanie 1 a zoradenie γ. Ranky teda udeľujeme tak, že všetky body preklopíme cez bod $(0, 0)^\top$ do polroviny danej 1. a 2. kvadrantom a najmenší rank udelíme tomu



Obr. 1: Vykreslené dáta z Tabuľky 3.

	rankovanie 1			rankovanie 2			rankovanie 3		
zoradenie	α	β	γ	α	β	γ	α	β	γ
S_1	4.01	4.01	-0.0013	0.68	2.14	1.14	3.76	3.92	-0.14
S_2	-4.15	-4.15	-4.34	-3.83	-2.76	-3.71	-4.44	-4.62	-4.64
S	33.36	33.36	18.87	15.11	12.19	15.09	33.82	36.76	21.58

Tabuľka 4: Testové štatistiky pre dáta z Tabuľky 1 pri rôznych postupoch vyhodnocovania testu.

dátovému bodu, ktorého dátová priamka zviera najmenší uhol s kladnou x-ovou polosou, a najväčší rank tomu, ktorého dátová priamka zviera najväčší uhol s touto polosou. Následne každému bodu udelíme nový uhol dátovej priamky. Ten udelíme iba na základe toho, v akom poradí boli dáta v pôvodnom dátovom súbore, teda i-temu bodu v dátovom súbore udelíme uhol $(\frac{i\pi}{n})$ radiánov. Následne dátové body, ktoré mali pôvodne zápornú y-ovú súradnicu, preklopíme okolo počiatku súradnicovej sústavy a hľadáme ťažisko takto vzniknutého n-uholníka.

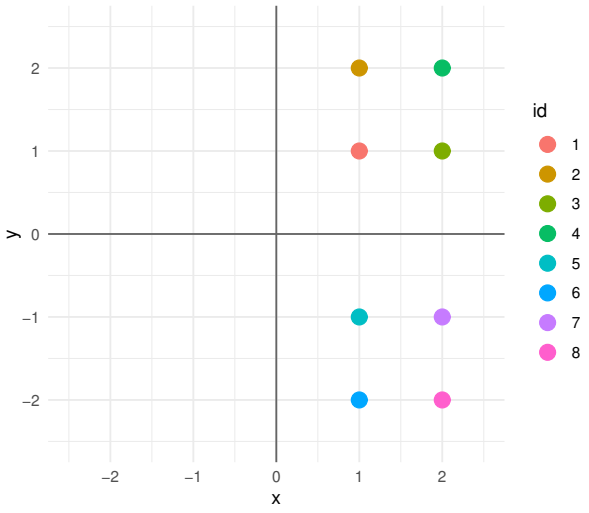
Problémom tohto prístupu je, že pri zoradení bodov neberieme nijako do úvahy skutočnú polohu dát. Následkom toho je jediná spoľahlivá informácia, ktorú máme o štandardizovaných

bodoch na jednotkovej kružnici tá, koľko bodov bolo pôvodne nad x-ovou osou a koľko pod ňou a akú váhu danému bodu priradiť. Takto vypočítané súradnice ťažiska (najmä jeho x-ová súradnica) iba veľmi malú výpovednú hodnotu a robiť na ich základe závery o dátach môže viesť k zavádzajúcim výsledkom. Tento jav vykreľujeme na obrázkoch nižšie.

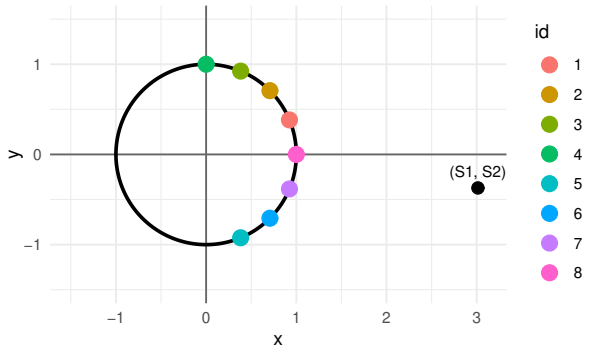
Uvažujme 8 dátových bodov uvedených v Tabuľke 5. Tieto dátové body vykresľujeme na Obr. 2 a štandardizovaný n-uholník, ktorý z týchto bodov vznikne na Obr. 3.

i	1	2	3	4	5	6	7	8
X_i	1	1	2	2	1	1	2	2
Y_i	1	2	1	2	-1	-2	-1	-2

Tabuľka 5: Vygenerované dáta, ktoré slúžia ako ukážka nespoľahlivosti testovej štatistiky S_1^γ.



Obr. 2: Body z Tabuľky 5



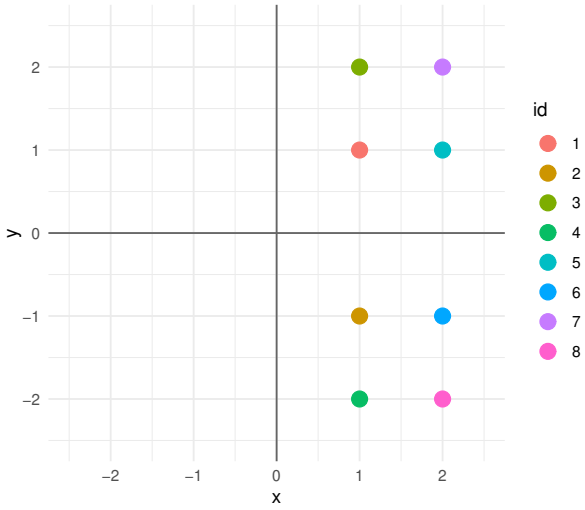
Obr. 3: Štandardizovaný n-uholník z dát v Tabuľke 5 pri použití testovej štatistiky S_1^γ a jeho rankované ťažisko $(S_1, S_2)^\top = (3.014, -0.371)^\top$. $S_1^\gamma = 9.22$

Zmeňme teraz poradie dát z Tabuľky 5:

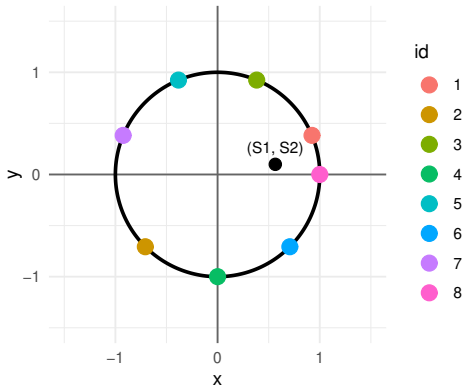
i	1	2	3	4	5	6	7	8
X_i	1	1	1	1	2	2	2	2
Y_i	1	-1	2	-2	1	-1	2	-2

Tabuľka 6: Vygenerované dáta, ktoré slúžia ako ukážka nespoľahlivosti testovej štatistiky S_1^γ v inom poradí.

K týmto dátam prislúchajúci štandardizovaný n-uholník uvádzame na Obr.



Obr. 4: Body z Tabuľky 6



Obr. 5: Štandardizovaný n-uholník z dát v Tabuľke 6 pri použití testovej štatistiky S_1^γ a jeho rankované ťažisko $(S_1, S_2)^\top = (0.565, 0.01)^\top$. $S_1^\gamma = 0.323$.

Vidno, že pre rovnaké dáta, ktoré sú iba inak usporiadané v pôvodnom dátovom súbore, sme získali výrazne odlišné hodnoty testovej štatistiky.

Okrem hodnoty testovej štatistiky v uvedenom príklade sa môžeme pokúsiť nájsť postup tým, že sa pokúsime zrekonštruovať tabuľku kritických hodnôt (tabuľka 1).

Kritické hodnoty sme odhadovali nasledovne:

1. Vygenerovali sme n realizácií $X_i \sim N_2((0,0), I_2)$.
2. Pre takto získaný dataset sme spočítali testovú štatistiku S_{cal}.
3. Kroky 1. a 2. sme zopakovali 10^6 krát.
4. Kritickými hodnotami sú 2.5%-ný a 97.5%-ný kvantil S_{cal}.

Odhady kritických hodnôt uvádzame v Tabuľkách 7 až 9.

Možno pozorovať, že odhady kritických hodnôt sú pre rankovania 1 a 2 podobné, kým pri rankovaní 3 sú výrazne vyššie. Najviac sa odlišujú hodnoty pre rankovanie 3 a zoradenie β. Bohužiaľ, ani jedna získaná tabuľka kritických hodnôt sa nezhoduje s Tabuľkou 1 dostatočne na to, aby sme mohli konštatovať, že sa nám podarilo nájsť práve ten postup, ktorý popisoval Mathur v článku [3]. Je môžné, že pre odhad kritických hodnôt generoval realizácie z dvojrozmerného normálneho rozdelenia s inou variačno-kovariančnou maticou, takéto detaily však neuvádza.

Zoradenie	α		β		γ	
n	**Left**	**Right**	**Left**	**Right**	**Left**	**Right**
5	0.28	5.20	0.28	5.20	0.15	4.90
7	0.08	9.03	0.08	9.03	0.11	8.16
10	0.09	14.19	0.09	14.19	0.12	12.29
12	0.12	16.61	0.12	16.61	0.13	14.83
15	0.14	20.52	0.14	20.52	0.16	18.56
18	0.16	24.79	0.16	24.79	0.19	22.25
20	0.18	27.36	0.18	27.36	0.20	24.65
25	0.22	34.31	0.22	34.31	0.25	30.95

Tabuľka 7: Obojstranné kritické hodnoty S-testov pri **rankovaní 1.**

Zoradenie	α		β		γ	
n	**Left**	**Right**	**Left**	**Right**	**Left**	**Right**
5	0.28	5.20	0.28	5.20	0.15	4.89
7	0.08	9.03	0.08	9.03	0.11	8.17
10	0.09	14.20	0.09	14.21	0.12	12.27
12	0.12	16.57	0.12	16.54	0.13	14.79
15	0.14	20.48	0.14	20.54	0.16	18.56
18	0.16	24.64	0.16	24.78	0.19	22.25
20	0.18	27.47	0.18	27.42	0.20	24.72
25	0.22	34.21	0.22	34.34	0.25	30.90

Tabuľka 8: Obojstranné kritické hodnoty S-testov pri **rankovaní 2.**

2.7 Overovanie sily Mathurovych S-testov

Hoci sa naši kandidáti na S-test jednoznačne nezhodujú s tým, ktorý popísal S. K. Mathur v článku [3], rozhodli sme sa simulačne porovnať ich silu s Blumenovym testom a Hottelingovym jednosúborovým T^2-testom. Ako sme uviedli v kapitole 2.4, vyhodnocovanie testu porovnaním vypočítanej hodnoty testovej štatistiky s ľavostrannou aj pravostrannou kritickou hodnotou nepovažujeme za vhodné. Malá hodnota testovej štatistiky nie je dôvodom na zamietnutie testovej hypotézy, preto sme pre účely testovania sily S-testov odhadli nové kritické hodnoty. Tie sme hľadali rovnako ako v predošlej kapitole, ale miesto

Zoradenie	α		β		γ	
n	**Left**	**Right**	**Left**	**Right**	**Left**	**Right**
5	0.10	5.20	2.57	5.20	0.10	5.29
7	0.19	8.57	4.66	9.16	0.18	8.54
10	0.26	14.20	8.13	17.52	0.22	13.79
12	0.30	18.82	11.35	24.52	0.29	17.78
15	0.48	25.86	17.36	37.13	0.45	24.55
18	0.69	33.95	24.75	52.31	0.62	32.11
20	0.87	39.63	31.42	63.45	0.81	37.97
25	1.51	55.60	50.11	96.49	1.33	53.19

Tabuľka 9: Obojstranné kritické hodnoty S-testov pri **rankovaní 3.**

2.5%-ného a 97.5%-ného kvantilu S_{cal} sme určili jedinú kritickú hodnotu ako 95%-ný kvantil S_{cal}. Testy zamietnu nulovú hypotézu pokiaľ bude hodnota testovej štatistiky vyššia ako táto hodnota. Tieto kritické hodnoty uvádzame v Tabuľke 10.

V označení S-testov predstavuje horný index spôsob rankovania a dolný index spôsob zoradenia dát. V porovnaní neuvádzame samostatne testov S^1_α a S^1_β, pretože ide o identické testy z definície.

Test	S^1_α	S^1_γ	S^2_α	S^2_β	S^2_γ	S^3_α	S^3_β	S^3_γ
krit. hodnota	17.05	15.57	17.03	16.93	15.67	23.10	37.13	21.89

Tabuľka 10: Kritické hodnoty S-testov pre $n = 15$ určené ako 95%-ný kvantil S_{cal}.

Dáta sú generované z normálneho rozdelenia so strednou hodnotou $\mu = (\delta_1, \delta_2)^\top$ a testujeme hypotézu o nulovosti strednej hodnoty. Simulačné dátové súbory obsahujú $n = 15$ pozorovaní a pre každú uvedenú dvojicu $(\delta_1, \delta_2)^\top$ vygenerujeme 15000 súborov. V tabuľkách nižšie uvádzame, pre akú časť vygenerovaných súborov test zamietol nulovú hypotézu.

V Tabuľkách 11 a 13 vidno, že tasty využívajúce rankovanie 1 a rankovanie 2 majú veľmi podobné výsledky, pričom spoľahlivo najslabšie sú testy využívajúce rankovanie γ, čo spĺňa naše očakávania vzhľadom na jeho nezmyselnosť, ktorú sme ilustrovali príkladom v kapitole 2.6.

δ_1, δ_2	S^1_α	S^1_γ	S^2_α	S^2_β	S^2_γ	S^3_α	S^3_β	S^3_γ	BL	T^2
0.00, 0.00	0.051	0.050	0.051	0.054	0.050	0.051	0.055	0.050	0.046	0.051
0.01, 0.00	0.051	0.050	0.052	0.051	0.051	0.051	0.056	0.050	0.046	0.048
0.03, 0.02	0.054	0.050	0.053	0.051	0.053	0.045	0.049	0.045	0.046	0.053
0.05, 0.03	0.050	0.051	0.051	0.052	0.049	0.043	0.046	0.044	0.049	0.053
0.10, 0.08	0.049	0.052	0.060	0.059	0.055	0.033	0.033	0.036	0.060	0.066
0.20, 0.10	0.054	0.053	0.076	0.068	0.056	0.035	0.029	0.033	0.087	0.096
0.40, 0.30	0.054	0.059	0.211	0.160	0.089	0.026	0.006	0.016	0.253	0.316
0.60, 0.50	0.075	0.069	0.478	0.363	0.172	0.038	0.001	0.017	0.555	0.665
0.80, 0.60	0.106	0.077	0.677	0.552	0.254	0.063	0.000	0.024	0.764	0.874
1.40, 1.00	0.305	0.184	0.962	0.947	0.680	0.272	0.000	0.115	0.993	1.000
2.60, 2.20	0.982	0.867	1.000	1.000	0.989	0.982	0.000	0.763	1.000	1.000
3.60, 3.10	1.000	0.983	1.000	1.000	0.992	1.000	0.000	0.891	1.000	1.000

Tabuľka 11: Porovnanie sily S-testov, Blumenovho testu a Hottelingovho T^2-testu. Testy boli porovnané na dátach generovaných z rozdelenia $N_2((\delta_1, \delta_2)^\top, I_2)$.

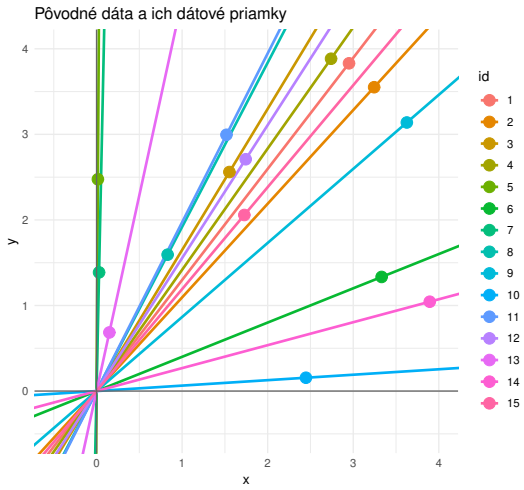
δ_1, δ_2	S^1_α	S^1_γ	S^2_α	S^2_β	S^2_γ	S^3_α	S^3_β	S^3_γ	BL	T^2
0.00, 0.00	0.047	0.052	0.051	0.053	0.054	0.051	0.055	0.050	0.045	0.050
0.01, 0.00	0.051	0.052	0.052	0.053	0.050	0.050	0.055	0.050	0.046	0.051
0.03, 0.02	0.050	0.051	0.051	0.050	0.050	0.045	0.049	0.046	0.046	0.049
0.05, 0.03	0.053	0.051	0.052	0.050	0.048	0.043	0.046	0.044	0.046	0.051
0.10, 0.08	0.053	0.051	0.053	0.055	0.052	0.033	0.033	0.036	0.052	0.056
0.20, 0.10	0.055	0.053	0.063	0.058	0.052	0.033	0.029	0.032	0.064	0.073
0.40, 0.30	0.050	0.050	0.114	0.089	0.061	0.019	0.006	0.016	0.130	0.153
0.60, 0.50	0.056	0.058	0.215	0.160	0.089	0.027	0.001	0.017	0.265	0.338
0.80, 0.60	0.072	0.057	0.347	0.253	0.111	0.044	0.000	0.023	0.422	0.522
1.40, 1.00	0.137	0.070	0.769	0.645	0.289	0.240	0.000	0.115	0.867	0.950
2.60, 2.20	0.541	0.340	0.992	0.993	0.875	0.978	0.000	0.762	1.000	1.000
3.60, 3.10	0.896	0.683	0.999	1.000	0.978	1.000	0.000	0.892	1.000	1.000

Tabuľka 12: Porovnanie sily S-testov, Blumenovho testu a Hottelingovho T^2-testu. Testy boli porovnané na dátach generovaných z rozdelenia $N((\delta_1, \delta_2)^\top, \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right))$.

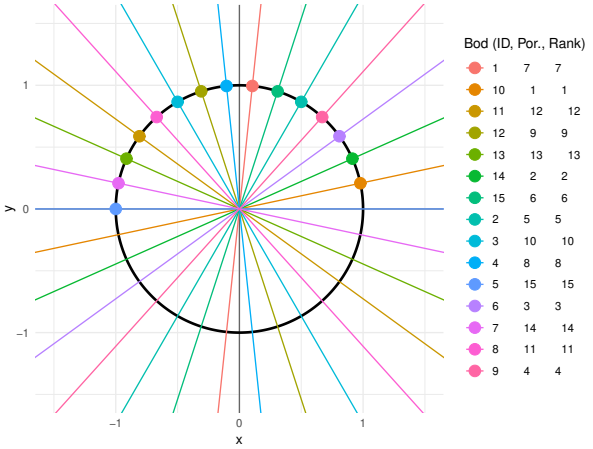
δ_1, δ_2	S^1_α	S^1_γ	S^2_α	S^2_β	S^2_γ	S^3_α	S^3_β	S^3_γ	BL	T^2
0.00, 0.00	0.050	0.051	0.051	0.050	0.053	0.051	0.056	0.050	0.043	0.049
0.01, 0.00	0.051	0.053	0.051	0.051	0.051	0.050	0.056	0.050	0.047	0.050
0.03, 0.02	0.052	0.051	0.050	0.050	0.050	0.047	0.051	0.048	0.045	0.050
0.05, 0.03	0.052	0.051	0.054	0.053	0.050	0.046	0.050	0.047	0.048	0.051
0.10, 0.08	0.051	0.051	0.056	0.054	0.055	0.040	0.042	0.042	0.051	0.057
0.20, 0.10	0.048	0.050	0.057	0.054	0.053	0.043	0.039	0.039	0.061	0.071
0.40, 0.30	0.051	0.051	0.084	0.081	0.065	0.039	0.017	0.024	0.112	0.136
0.60, 0.50	0.049	0.057	0.150	0.138	0.084	0.046	0.007	0.016	0.206	0.258
0.80, 0.60	0.060	0.055	0.223	0.193	0.105	0.074	0.004	0.015	0.331	0.417
1.40, 1.00	0.109	0.077	0.574	0.518	0.265	0.150	0.000	0.022	0.772	0.882
2.60, 2.20	0.542	0.351	0.970	0.975	0.840	0.537	0.000	0.248	0.998	1.000
3.60, 3.10	0.894	0.684	0.998	0.998	0.969	0.897	0.000	0.575	1.000	1.000

Tabuľka 13: Porovnanie sily S-testov, Blumenovho testu a Hottelingovho T^2-testu. Testy boli porovnané na dátach generovaných z rozdelenia $N((\delta_1, \delta_2)^\top, \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 4 \end{smallmatrix}\right))$.

Najnižšiu silu majú testy využívajúce rankovanie 3, ktorých sila najprv klesá pod 5%. V prípade testov S^3_α a S^3_γ sa s narastajúcou vzdialenosťou strednej hodnoty $(X_i, Y_i)^\top$ od bodu $(0, 0)^\top$ táto sila opäť zvýši. Test S^3_β celkom prestane zamietať nulovú hypotézu v prípade vysokých parametrov $(\delta_1, \delta_2)^\top$. Toto správanie je spôsobené tým, že ak sú všetky dáta nad x-ovou osou, štandardizovaný n-uholník, ktorého ťažisko počítame, sa skladá iba z bodov nad x-ovou osou, ktorých ranky sú navyše dané jednoznačne. V takom prípade bude hodnota testovej štatistiky vždy rovnaká, $S^3_\beta = 35.23$, čo je hodnota nižšia ako nami odhadnutá kritická hodnota. Takúto situáciu ilustrujeme na obrázkoch 6 a 7.



Obr. 6: Dáta s kladnými hodnotami X_i aj Y_i kol z kladných dát pri použití testovej štatistiky S_3^{β}.



Obr. 7: Štandardizovaný n-uholník, ktorý vzni-

3 Mathur-Smithovej T štatistika

V tejto kapitole sa budeme venovať testovej štatistike, ktorú navrhli a popísali autori K. S. Mathur a P. Smithová v článku [4], ktorý bol publikovaný v roku 2008 v časopise Statistical Methodology. Na rozdiel od testu, ktorému sme sa venovali v predošlej podkapitole, tento test sa zaoberá riešením dvojsúborového bivariančného problému polohy. Tento problém sa zaoberá situáciou, kedy máme dva súbory dát a chceli by sme zistiť, či je ich poloha rovnaká.

Predpokladajme, že $(X_i, Y_i)^\top$, $i = 1, 2, \ldots, n$ sú merania prináležiace n subjektom z populácie s distribučnou funkciou $F(X, Y)$ a $(U_j, V_j)^\top$, $j = 1, 2, \ldots, m$ sú merania prináležiace m subjektom z rozdelenia s distribučnou funkciou $G(X, Y)$, pričom F a G sú spojité.

Chceme otestovať, či majú obe populácie rovnakú polohu. Autori uvádzajú, že chcú testovať hypotézu ohľadom parametra polohy $\theta = (\theta_1, \theta_2)^\top$ dvoch neznámych rozdelení:

$$H_0 : \ \theta_F = \theta_G \quad \text{oproti} \quad H_1 : \theta_F = \theta_G + (\delta_1, \delta_2)^\top, \ \text{kde } (\delta_1, \delta_2)^\top \neq (0, 0)^\top. \tag{5}$$

Predpokladáme, že parametre θ_F a θ_G sú stredné hodnoty príslušného rozdelenia, no článok [4] to explicitne nikde neuvádza. Ekvivalentne môžeme označiť $\theta = (\delta_1, \delta_2)^\top =$

$\theta_F - \theta_G$ a testovať hypotézu

$$H_0 : \ \theta = (0,0)^\top \quad \text{oproti} \quad H_1 : \theta \neq (0,0)^\top. \tag{6}$$

Navrhovaná testová štatistika je založená na porovnaní vzdialeností dát z dvoch rôznych súborov a vzdialeností dát v rámci každého súboru jednotlivo. Testová šatistika je daná:

$$T = T_{AB} - \frac{T_A + T_B}{2}, \tag{7}$$

kde

$$\begin{aligned}
T_{AB} &= \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} [(X_i - U_j)^2 + (Y_i - V_j)^2], \\
T_A &= \frac{1}{\binom{n}{2}} \sum_{1 \leq i < \ j \leq n} [(X_i - X_j)^2 + (Y_i - Y_j)^2], \\
T_B &= \frac{1}{\binom{m}{2}} \sum_{1 \leq j < \ k \leq m} [(U_j - U_k)^2 + (V_j - V_k)^2].
\end{aligned}$$

Ďalej sa autori v texte [4] venujú výlučne špeciálnemu prípadu $n = m$. V tomto prípade platí

$$\begin{aligned}
T_{AB} &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[(X_i - U_j)^2 + (Y_i - V_j)^2\right] \\
&= \frac{1}{n} \sum_{i=1}^{n} X_i^2 + \frac{1}{n} \sum_{j=1}^{n} U_j^2 + \frac{1}{n} \sum_{i=1}^{n} Y_i^2 + \frac{1}{n} \sum_{j=1}^{n} V_j^2 - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i U_j + Y_i V_j) \\
\frac{T_A}{2} &= \frac{1}{n(n-1)} \sum_{1 \leq i < \ j \leq n} [(X_i - X_j)^2 + (Y_i - Y_j)^2] \\
&= \frac{1}{n} \sum_{i=1}^{n} X_i^2 + \frac{1}{n} \sum_{i=1}^{n} Y_i^2 - \binom{n}{2}^{-1} \sum_{1 \leq i < \ j \leq n} (X_i X_j + Y_i Y_j) \\
\frac{T_B}{2} &= \frac{1}{n(n-1)} \sum_{1 \leq j < \ k \leq m} [(U_j - U_k)^2 + (V_j - V_k)^2] \\
&= \frac{1}{n} \sum_{i=1}^{n} U_i^2 + \frac{1}{n} \sum_{i=1}^{n} V_i^2 - \binom{n}{2}^{-1} \sum_{1 \leq i < \ j \leq n} (U_i U_j + V_i V_j)
\end{aligned}$$

Členy so štvorcami sa vzájomne odčítajú a zostáva

$$T = \binom{n}{2}^{-1} \sum_{1 \leq i < \ j \leq n} (X_i X_j + Y_i Y_j + U_i U_j + V_i V_j) - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i U_j + Y_i V_j). \tag{8}$$

Mathur a Smithová tvrdia, že rovnicu (8) môžeme prepísať do tvaru

$$T = \binom{n}{2}^{-1} \sum_{1\le i<} \sum_{j\le n} (X_iX_j + Y_iY_j + U_iU_j + V_iV_j - X_iU_j - U_iX_j - Y_iV_j - V_iY_j)$$
$$- \frac{2}{n^2} \sum_{i=1}^{n} (X_iU_i + Y_iV_i) - \frac{2}{n^2(n-1)} \sum_{1\le i<} \sum_{j\le n} (X_iU_j + U_iX_j + Y_iV_j + V_iY_j). \qquad (9)$$

Úpravami nižšie však ukážeme, že znamienko pred posledným členom na pravej strane tejto rovnice by malo byť kladné (teda + a nie −). Upravme posledný člen rovnice (8):

$$\frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_iU_j + Y_iV_j) =$$
$$= \frac{2(n-1)}{n^2(n-1)} \sum_{1\le i<} \sum_{j\le n} (X_iU_j + Y_iV_j + U_iX_j + V_iY_j) + \frac{2}{n^2} \sum_{i=1}^{n} (X_iU_i + Y_iV_i)$$
$$= \frac{2n}{n^2(n-1)} \sum_{1\le i<} \sum_{j\le n} (X_iU_j + X_jU_i + Y_iV_j + Y_jV_i) -$$
$$- \frac{2}{n^2(n-1)} \sum_{1\le i<} \sum_{j\le n} (X_iU_j + X_jU_i + Y_iV_j + Y_jV_i) + \frac{2}{n^2} \sum_{i=1}^{n} (X_iU_i + Y_iV_i).$$

Dosaďme do (8):

$$T = \binom{n}{2}^{-1} \sum_{1\le i<} \sum_{j\le n} (X_iX_j + Y_iY_j + U_iU_j + V_iV_j)$$
$$- \binom{n}{2}^{-1} \sum_{1\le i<} \sum_{j\le n} (X_iU_j + X_jU_i + Y_iV_j + Y_jV_i)$$
$$+ \frac{2}{n^2(n-1)} \sum_{1\le i<} \sum_{j\le n} (X_iU_j + X_jU_i + Y_iV_j + Y_jV_i) - \frac{2}{n^2} \sum_{i=1}^{n} (X_iU_i + Y_iV_i)$$
$$= \binom{n}{2}^{-1} \sum_{1\le i<} \sum_{j\le n} (X_iX_j + Y_iY_j + U_iU_j + V_iV_j - X_iU_j - U_iX_j - Y_iV_j - V_iY_j) \qquad (10)$$
$$- \frac{2}{n^2} \sum_{i=1}^{n} (X_iU_i + Y_iV_i) + \frac{2}{n^2(n-1)} \sum_{1\le i<} \sum_{j\le n} (X_iU_j + U_iX_j + Y_iV_j + V_iY_j).$$

Rozdiel medzi (9) a (10) je práve v znamienku pred posledným členom na pravej strane.

Za predpokladu $E(XU) < \infty$ a $E(YV) < \infty$ sa dá ukázať, že pre veľké n oba posledné členy rovnice (9) resp. (10) konvergujú podľa pravdepodobnosti k 0. Mathur a Smithová toto odvodenie urobili ledabolo, nižšie uvádzame takéto odvodenie s väčším dôrazom na detail a opravujeme niektoré nekorektné tvrdenia. Využijeme pri tom slabý zákon veľkých čísel a Slutského vetu a budeme potrebovať definíciu označení rádu konvergencie.

Veta 3.1 (Čebyševov slabý zákon veľkých čísel)**.** *Nech $X_1, \ldots, X_n$ sú po dvoch nezávislé, náhodné premenné z rovnakého rozdelenia so strednou hodnotou a konečnou disperziou, teda $E(X_i) = \mu, Var(X_i) < 0$. Potom*

$$\frac{1}{n}(X_1 + X_2 + \cdots + X_n) \xrightarrow{P} \mu,$$

kde $\xrightarrow{P}$ značí konvergenciu podľa pravdepodobnosti pre $n \to \infty$.

Veta 3.2 (Špeciálny prípad Slutského vety)**.** *Nech $X_n \xrightarrow{P} c$ a $Y_n \xrightarrow{P} d$, kde X_n, Y_n sú postupnosti náhodných premenných a c, d sú konštanty. Potom*

a) $X_n + Y_n \xrightarrow{P} c + d$;

b) $X_n Y_n \xrightarrow{P} cd$.

Definícia 3.3. *Definujme označenia rádu konvergencie:*

a) označenenie $X_n = o_P(n^\alpha)$ znamená, že rad $n^{-\alpha} X_n$ konverguje k 0 podľa pravdepodobnosti pre $n \to \infty$. Teda platí

$$\lim_{n\to\infty} P\left(|n^{-\alpha} X_n| > \epsilon\right) = 0 \text{ pre každé } \epsilon > 0.$$

b) označenenie $X_n = O_P(n^\alpha)$ znamená, že rad $n^{-\alpha} X_n$ je stochasticky ohraničený, teda pre každé $\epsilon > 0$ existuje $M > 0$ také, že

$$P\left(|n^{-\alpha} X_n| > M\right) < \epsilon.$$

Venujme pozornosť najprv prostrednému členu (10). Zo slabého zákona veľkých čísel a z nezávislosti vektorov (X, Y) a (U, V) vyplýva

$$\frac{1}{n}\sum_{i=1}^{n} X_i U_i \xrightarrow{P} E(XU) = E(X)E(U)$$
$$\frac{1}{n}\sum_{i=1}^{n} Y_i V_i \xrightarrow{P} E(YV) = E(Y)E(V).$$

Aplikáciou Vety 3.2 *a*) dostaneme

$$\frac{1}{n}\sum_{i=1}(X_i U_i + Y_i V_i) = \frac{1}{n}\sum_{i=1}^{n}(X_i U_i) + \frac{1}{n}\sum_{i=1}^{n}(Y_i V_i) \xrightarrow{P} E(X)E(U) + E(Y)E(V).$$

a nakoniec z Vety 3.2 *b*)

$$\frac{2}{n^2}\sum_{i=1}^{n}(X_iU_i + Y_iV_i) = \frac{2}{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_iU_i) + \frac{1}{n}\sum_{i=1}^{n}(Y_iV_i)\right) \xrightarrow{P} 0. \tag{11}$$

Mathur a Smithová tvrdia, že

$$\frac{2}{n^2}\sum_{i=1}^{n}(X_iU_i + Y_iV_i) = o_P\left(\frac{1}{n}\right),$$

no toto tvrdenie nie je pravdivé, pretože platí

$$n\left(\frac{2}{n^2}\sum_{i=1}^{n}(X_iU_i + Y_iV_i)\right) = \frac{2}{n}\sum_{i=1}^{n}(X_iU_i + Y_iV_i) \xrightarrow{P} 2E(X)E(U) + 2E(Y)E(V) \not\equiv 0. \tag{12}$$

Opravená verzia tohto tvrdenia je teda

$$\frac{2}{n^2}\sum_{i=1}^{n}(X_iU_i + Y_iV_i) = o_P(1).$$

S posledným členom rovnice (10) sa vysporiadme iba o trochu komplikovanejšie. Uvedomme si, že platí

$$\sum_{1\le i<}\sum_{j\le n}(X_iU_j + U_iX_j) = \sum_{i\neq j}\sum(X_iU_j) = \sum_{i=1}^{n}\sum_{j=1}^{n}(X_iU_j) - \sum_{i=1}^{n}(X_iU_i).$$

Pôvodná dvojitá suma skutočne prechádza cez všetky dvojice X_i, U_j okrem prípadov, kedy $i = j$. Môžeme teda zapísať

$$\frac{1}{n^2}\sum_{1\le i<}\sum_{j\le n}(X_iU_j + U_iX_j) = \left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)\left(\frac{1}{n}\sum_{j=1}^{n}U_j\right) - \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_iU_i)\right). \tag{13}$$

Vďaka slabému zákonu veľkých čísel vieme, že

$$\frac{1}{n}\sum_{i=1}^{n}X_i \xrightarrow{P} E(X),$$
$$\frac{1}{n}\sum_{j=1}^{n}U_j \xrightarrow{P} E(U),$$
$$\frac{1}{n}\sum_{i=1}^{n}(X_iU_i) \xrightarrow{P} E(X)E(U)$$

a teda podľa Vety 3.2 *b*) máme

$$\frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_iU_i)\right) \xrightarrow{P} 0.$$

Ak využijeme Vety 3.2a) aj 3.2b), získame

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} U_j\right) - \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_iU_i)\right) \xrightarrow{P} E(X)E(U) - 0 \cdot E(X)E(U). \quad (14)$$

Z (13) a (14) máme

$$\frac{1}{n^2}\sum_{1\le i<}\sum_{j\le n}(X_iU_j + U_iX_j) \xrightarrow{P} E(X)E(U).$$

Ekvivalentnú úvahu možno urobiť pre výraz

$$\frac{1}{n^2}\sum_{1\le i<}\sum_{j\le n}(Y_iV_j + V_iY_j)$$

a vďaka Vete 3.2 a) môžeme napísať

$$\frac{1}{n^2}\sum_{1\le i<}\sum_{j\le n}(X_iU_j + Y_iV_j + U_iX_j + V_iY_j) \xrightarrow{P} E(X)E(U) + E(Y)E(V). \quad (15)$$

Tento medzivýsledok je odlišný od tvrdenia v článku [4], podľa ktorého postupnosť v (15) konverguje k $2E(X)E(U)+2E(Y)E(V)$. Domnievame sa, že v tomto prípade ide o chybu v článku.

Z výrazu (15) vyplýva, že tretí člen rovnice (10) tiež konverguje k 0 podľa pravdepodobnosti,

$$\frac{2}{n^2(n-1)}\sum_{1\le i<}\sum_{j\le n}(X_iU_j + U_iX_j + Y_iV_j + V_iY_j) \xrightarrow{P} 0, \quad (16)$$

a platí

$$\frac{2}{n^2(n-1)}\sum_{1\le i<}\sum_{j\le n}(X_iU_j + U_iX_j + Y_iV_j + V_iY_j) = o_P(1).$$

Aj v tomto prípade Mathur a Smithová v článku uvádzajú, že platí

$$\frac{2}{n^2(n-1)}\sum_{1\le i<}\sum_{j\le n}(X_iU_j + U_iX_j + Y_iV_j + V_iY_j) = o_P\left(\frac{1}{n}\right),$$

čo nie je korektné tvrdenie, pretože podľa (15)

$$n\left(\frac{2}{n^2(n-1)}\right)\sum_{1\le i<}\sum_{j\le n}(X_iU_j+U_iX_j+Y_iV_j+V_iY_j) \tag{17}$$
$$=2\left(\frac{n-1+1}{n^2(n-1)}\right)\sum_{1\le i<}\sum_{j\le n}(X_iU_j+U_iX_j+Y_iV_j+V_iY_j)$$
$$=2\left(\frac{1}{n^2}\sum_{1\le i<}\sum_{j\le n}(X_iU_j+U_iX_j+Y_iV_j+V_iY_j)\right)$$
$$+\frac{2}{n-1}\left(\frac{1}{n^2}\sum_{1\le i<}\sum_{j\le n}(X_iU_j+U_iX_j+Y_iV_j+V_iY_j)\right)$$
$$\xrightarrow{P}2(E(X)E(U)+(Y)E(V))+0\cdot 2(E(X)E(U)+(Y)E(V))\not\equiv 0.$$

Označme

$$U_n=\binom{n}{2}^{-1}\sum_{1\le i}\sum_{<j\le n}(X_iX_j+Y_iY_j+U_iU_j+V_iV_j-X_iU_j-X_jU_i-Y_iV_j-Y_jV_i). \tag{18}$$

Podľa (10), (11) a (16) dostávame, že výraz T sa pre veľké n blíži k U_n, teda

$$T-U_n\xrightarrow{P}0. \tag{19}$$

Autori článku [4] zvolili mierne odlišný prístup a tvrdia nasledovné: „*Under the null hypothesis, we can center* (X,Y) *and* (U,V) *at their respective mean, then without loss of generality, we assume that the liming expectations* $E(X)E(U)+E(Y)E(V)$ *and* $2E(X)E(U)+2E(Y)E(V)$ *are zero under the null hypothesis.*“ „Za platnosti nulovej hypotézy môžeme $(X,Y)^\top$ a $(U,V)^\top$ vycentrovať na ich príslučnú strednú hodnotu. Potom bez ujmy na všeobecnosti predpokladáme, že limitné stredné hodnoty $E(X)E(U)+E(Y)E(V)$ a $2E(X)E(U)+2E(Y)E(V)$ sú za platnosti nulovej hypotézy nulové.“ Následne dodávajú, že rovnicu T možno zapísať ako U_n. Centrovanie dát ani platnosť H_0 nie sú potrebné, pretože vyššie popísané konvergencie platia bez ohľadu na konkrétnu hodnotu $E(X)E(U)+E(Y)E(V)$. V kapitole 3.3 navyše vidno, že autori pri aplikácii tohto testu dáta necentrovali. Záver o tom, že rovnicu T možno zapísať ako U_n je doslova vzaté nepravdivý, tieto dva výrazy sa k sebe podľa (19) skutočne blížia pre veľké n, no nie sú ekvivalentné.

Označme niektoré premenné: $\omega_i=(X_i,Y_i,U_i,V_i)^\top$, $E(X)=\mu_1$, $E(Y)=\mu_2$, $E(U)=m_1$, $E(V)=m_2$. Autori v článku [4] ďalej dokazujú, že U_n je U-štatistika, čo zjavne plynie z tvaru U_n a z Definície 1.3. Kernelová funkcia tejto U-štatistiky má tvar $h(s,t)=x_1x_2+y_1y_2+u_1u_2+v_1v_2-x_1u_2-x_2u_1-y_1v_2-y_2v_1$, kde $s=(x_1,y_1,u_1,v_1)$ a $t=(x_2,y_2,u_2,v_2)$.

Za platnosti hypotézy H_0 platí pre strednú hodnotu U_n

$$\begin{aligned}\theta &= E[U_n(h)] = E[h(\omega_1, \omega_2)] \\ &= \mu_1^2 + \mu_2^2 + m_1^2 + m_2^2 - 2\mu_1 m_1 - 2\mu_2 m_2 \\ &= (\mu_1 - m_1)^2 + (\mu_2 - m_2)^2 = 0.\end{aligned} \tag{20}$$

Aby sme mohli lepšie pracovať s U-štatistikou U_n, pomohlo by nám poznať jej asymptotické rozdelenie za platnosti H_0. Mathur a Smithová v článku [4] tvrdia, že ide o štatistiku s degenerovaným kernelom. Hľadajme $\zeta_1 = E(\hat{h}_1^2)$ z Definície 1.5.

$$\begin{aligned}\hat{h}_1(s) &= E[h(\omega_1, \omega_2)|\omega_1 = s] - \theta \\ &= E(x_1X_2 + y_1Y_2 + u_1U_2 + v_1V_2 - x_1U_2 - y_1V_2 - u_1X_2 - v_1Y_2) - 0 \\ &= x_1\mu_1 + y_1\mu_2 + u_1m_1 + v_1m_2 - x_1m_1 - y_1m_2 - u_1\mu_1 - v_1\mu_2 \\ &= x_1(\mu_1 - m_1) + y_1(\mu_2 - m_2) + u_1(m_1 - \mu_1) + v_1(m_2 - \mu_2) \\ &= 0,\end{aligned}$$

pretože nulová hypotéza platí, čiže $\theta = 0$, $\mu_1 = m_1$ a $\mu_2 = m_2$. Z toho vyplýva, že aj $E(\hat{h}_1^2) = 0$ a podľa Definície 1.5 je (18) U-štatistika s degenerovanou kernelovou funkciou. Podľa Vety 1.7 pre túto štatistiku platí

$$n(U - \theta) \xrightarrow{d} A, \tag{21}$$

kde

$$A = \sum_{j=1}^{\infty} \lambda_j(\chi_{1j}^2 - 1),$$

kde $\chi_{11}^2, \chi_{12}^2, \ldots$ sú nezávislé χ_1^2 premenné a $\lambda_1, \lambda_2, \ldots$ sú vlastné hodnoty prislúchajúce rôznym vlastným funkciám Fredholmovej funkcionálnej rovnice

$$\lambda\varphi(s) = \int h(s,t)\varphi(t)dH(t). \tag{22}$$

Autori v článku [4] uvádzajú, že štatistika (18) by mala konvergovať podľa distribúcie k náhodnej premennej tvaru $\sum_{k=1}^{\infty} \lambda_k(Z_k^2 - 1)$, kde $Z_1, Z_2, \ldots$ sú nezávislé náhodné premenné z rozdelenia $N(0,1)$ a $\lambda_1, \ldots, \lambda_4$ sú konečné vlastné hodnoty rovnice (22). V tejto časti majú autori pravdepodobne preklep a nekonečný mal byť nielen počet Z_k, ale aj λ_k.

Ukážeme, že vďaka elegantnému tvaru kernelovej funkcie $h(s,t)$ je každá vlastná funkcia Fredholmovej rovnice $\lambda\varphi(s) = \int h(s,t)\varphi(t)dH(t)$ lineárna, čiže $\varphi(s) = ax_1 + by_1 + cu_1 + dv_1$, kde a, b, c, d sú konštanty. Ak $s = (x_1, y_1, u_1, v_1)^\top$ a $t = (x_2, y_2, u_2, v_2)^\top$, potom musí platiť

$$\begin{aligned}
\lambda\varphi(s) &= \int (x_1x_2 + y_1y_2 + u_1u_2 + v_1v_2 - x_1u_2 - x_2u_1 - y_1v_2 - y_2v_1)\varphi(t)dH(t) \\
&= \int \big(x_1(x_2 - u_2) + y_1(y_2 - v_2) + u_1(u_2 - x_2) + v_1(v_2 - y_2)\big)\varphi(t)dH(t) \\
&= x_1 \int (x_2 - u_2)\varphi(t)dH(t) + y_1 \int (y_2 - v_2)\varphi(t)dH(t) \\
&\quad + u_1 \int (u_2 - x_2)\varphi(t)dH(t) + v_1 \int (v_2 - y_2)\varphi(t)dH(t) \\
&= x_1a' + y_1b' + u_1c' + y_1d',
\end{aligned} \tag{23}$$

čiže

$$\varphi(s) = \frac{a'}{\lambda}x_1 + \frac{b'}{\lambda}y_1 + \frac{c'}{\lambda}u_1 + \frac{d'}{\lambda}v_1 = ax_1 + by_1 + cu_1 + dv_1.$$

Dosadením $\varphi(s) = ax_1 + by_1 + cu_1 + dv_1$ aj $\varphi(t) = ax_2 + by_2 + cu_2 + dv_2$ do (23) dostaneme

$$\begin{aligned}
\lambda(ax_1 + by_1 + cu_1 + dv_1) = x_1 &\int (x_2 - u_2)(ax_2 + by_2 + cu_2 + dv_2)dH(t) \\
&+ y_1 \int (y_2 - v_2)(ax_2 + by_2 + cu_2 + dv_2)dH(t) \\
&+ u_1 \int (u_2 - x_2)(ax_2 + by_2 + cu_2 + dv_2)dH(t) \\
&+ v_1 \int (v_2 - y_2)(ax_2 + by_2 + cu_2 + dv_2)dH(t).
\end{aligned}$$

Porovnaním koeficientov pri x_1, x_2, x_3 a x_4 získame tvar

$$\begin{aligned}
\lambda a &= \int (x_2 - u_2)(ax_2 + by_2 + cu_2 + dv_2) dH(t) \\
&= \int \left(a(x_2^2 - x_2u_2) + b(x_2y_2 - y_2u_2) + c(x_2u_2 - u_2^2) + d(x_2v_2 - u_2v_2)\right) dH(t) \\
\lambda b &= \int (y_2 - v_2)(ax_2 + by_2 + cu_2 + dv_2) dH(t) \\
&= \int \left(a(y_2x_2 - v_2x_2) + b(y_2^2 - y_2v_2) + c(y_2u_2 - v_2u_2) + d(y_2v_2 - v_2^2)\right) dH(t) \\
\lambda c &= (u_2 - x_2)(ax_2 + by_2 + cu_2 + dv_2) dH(t) \\
&= \int \left(a(u_2x_2 - x_2^2) + b(u_2y_2 - x_2y_2) + c(u_2^2 - x_2u_2) + d(u_2v_2 - x_2v_2)\right) dH(t) \\
\lambda d &= \int (v_2 - y_2)(ax_2 + by_2 + cu_2 + dv_2) dH(t) \\
&= \int \left(a(v_2x_2 - y_2x_2) + b(v_2y_2 - y_2^2) + c(v_2u_2 - y_2u_2) + d(v_2^2 - y_2v_2)\right) dH(t).
\end{aligned}$$

Hľadané hodnoty λ by sme teda mali získať riešením nasledujúcej sústavy štyroch rovníc o piatich neznámych:

$$\begin{aligned}
\lambda a &= [E(X^2) - E(UX)]a + [E(XY) - E(UY)]b \\
&\quad + [E(XU) - E(U^2)]c + [E(XV) - E(UV)]d, \\
\lambda b &= [E(YX) - E(VX)]a + [E(Y^2) - E(VY)]b \\
&\quad + [E(YU) - E(VU)]c + [E(YV) - E(V^2)]d, \\
\lambda c &= [E(UX) - E(X^2)]a + [E(UY) - E(XY)]b \\
&\quad + [E(U^2) - E(XU)]c + [E(UV) - E(XV)]d, \\
\lambda d &= [E(VX) - E(YX)]a + [E(VY) - E(Y^2)]b \\
&\quad + [E(VU) - E(YU)]c + [E(V^2) - E(YV)]d.
\end{aligned} \tag{24}$$

Na ilustráciu určovania asymptotického rozdelenia zvolili Mathur a Smithová rozdelenia $(X, Y)^\top \sim N_2((\mu_1, \mu_2)^\top, \Sigma^2_{X,Y})$ a $(U, V)^\top \sim N_2((m_1, m_2)^\top, \Sigma^2_{U,V})$, pričom

$$\begin{aligned}
\mu_1 = 36, \quad \mu_2 = 29, \quad \sigma_X^2 = \sigma_Y^2 = 1, \quad \rho_{X,Y} = 0.5, \\
m_1 = 35, \quad m_2 = 30, \quad \sigma_U^2 = \sigma_V^2 = 1, \quad \rho_{U,V} = 0.5.
\end{aligned}$$

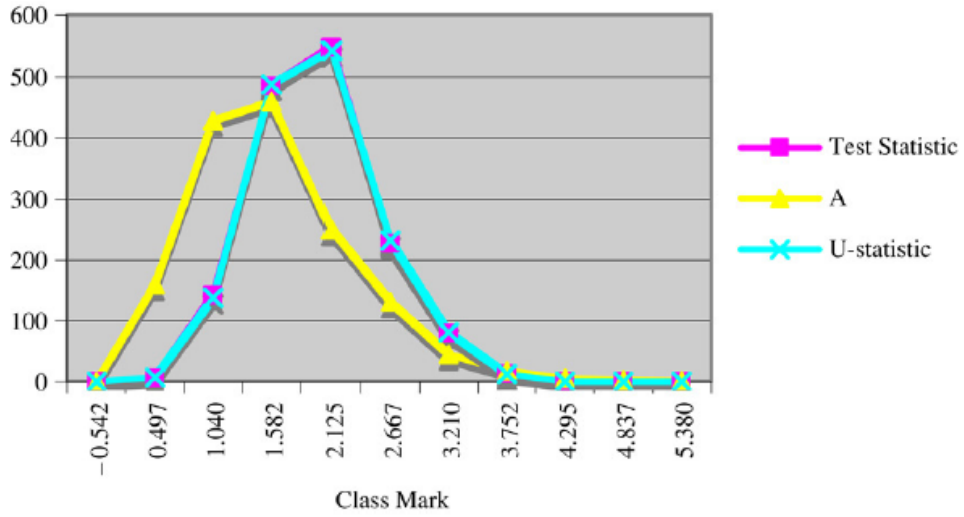
Po dosadení premenných do vyššie uvedených rovníc (24), získali autori článku riešenie systému rovníc $\lambda = 3$ a $\lambda = 3$. Autori bohužiaľ neuvádzajú koľkonásobný koreň λ uvažujú

a hoci v texte spomínajú $\lambda_1, \lambda_2, \lambda_3$ a λ_4, nie je jasné, či je náhodná premenná A daná jedným členom, dvomi alebo štyrmi.

Na Obr. 8 je kópia obrázka z článku, kde porovnávajú na základe 1500 simulácií pôvodnú testovú štatistiku T (7), štatistiku U_n (18) a náhodnú premennú

$$A = \sum_{k=1}^{\infty} \lambda_k (Z_k^2 - 1),$$

pričom dátové sady generovali z vyššie uvedených rozdelení.



Obr. 8: Pôvodné porovnanie testovej štatistiky T, U-štatistiky U_n a asymptotického rozdelenia A s parametrom $\lambda = 3$. Hodnoty štatistík boli získané pomocou 1500 simulácií dát z rozdelení $(X, Y)^\top \sim N_2((36, 29)^\top, \left(\begin{smallmatrix} 1 & 0{,}5 \\ 0.5 & 1 \end{smallmatrix}\right))$ a $(U, V)^\top \sim N_2((35, 30)^\top, \left(\begin{smallmatrix} 1 & 0{,}5 \\ 0.5 & 1 \end{smallmatrix}\right))$.
Zdroj : [4, *str.*150]

K správnosti tohto ilustračného príkladu sa vyjadríme v kapitole 3.2.

3.1 Asymptotické rozdelenie

Autori článku [4] uvádzajú, že pre nájdenie asymptotického rozdelenia štatistiky U_n musíme vyriešiť sústavu štyroch rovníc o piatich neznámych (24). Riešením tohto systému získame parametre $\lambda_k,\ k = 1, 2, \ldots$ rozdelenia daného $\sum_{k=1}^{\infty} \lambda_k (Z_k^2 - 1)$, kde Z_k sú nezávislé náhodné premenné z rozdelenia $N(0, 1)$. Takto nájdené rozdelenie je asymptotické rozdelenie U-štatistiky U_n.

V tejto časti ukážeme, že sústavu rovníc (24) je možné eliminovať na sústavu dvoch rovníc o troch neznámych. Ďalej sa budeme zaoberať existenciou parametrov λ_k a ich vlastnosťami. Napokon uvedieme úvahu nad tým, či sa rozdelenie U-štatistiky U_n (18) skutočne môže asymptoticky blížiť k takto nájdenému rozdeleniu.

Vidno, že ak v sústave rovníc (24) sčítame 1. a 3. rovnicu a 2. a 4. rovnicu, získame rovnice $\lambda a + \lambda c = 0$ a $\lambda b + \lambda d = 0$, teda $a = -c$ a $b = -d$. Systém teda vieme prepísať do tvaru

$$\lambda a = [E(X^2) - 2E(UX) + E(U^2)]a + [E(XY) - E(UY) - E(XV) + E(UV)]b, \quad (25)$$

$$\lambda b = [E(YX) - E(VX) - E(YU) + E(VU)]a + [E(Y^2) - 2E(VY) + E(V^2)]b. \quad (26)$$

Ide o systém dvoch rovníc o troch neznámych λ, a, b a posledné dve rovnice zo systému (24) boli redundantné. Keď sme náš problém zjednodušili, pusťme sa do hľadania riešení.

Hľadanie riešenia systému rovníc (25) a (26) je v skutočnosti ekvivalentné hľadaniu vlastných hodnôt matice.

Definícia 3.4. *Hovoríme, že skalár $\lambda \in R$ je vlastná hodnota lineárneho operátora $\phi : V \to V$, ak existuje vektor $0 \neq v \in V$, pre ktorý platí $\phi(v) = \lambda v$.*

V tomto prípade je matica

$$\phi = \begin{pmatrix} E(X^2) - 2E(UX) + E(U^2) & E(XY) - E(UY) - E(XV) + E(UV) \\ E(YX) - E(VX) - E(YU) + E(VU) & E(Y^2) - 2E(VY) + E(V^2) \end{pmatrix}$$

a vektor $v = \begin{pmatrix} a \\ b \end{pmatrix}$.

Vďaka nezávislosti vektorov (X, Y) a (U, V) a platnosti hypotézy H_0 o tom, že $E(X) =$

$E(U)$ a $E(Y) = E(V)$ môžeme jednotlivé výrazy v rovnici (25) upraviť nasledovne:

$$\begin{aligned} E(X^2) - 2E(UX) + E(U^2) &= E(X^2) - 2E(X)E(U) + E(U^2) \\ &= E(X^2) - E(X)^2 + E(U^2) - E(U)^2 = Var(X) + Var(U) \end{aligned}$$

a

$$\begin{aligned} E(XY) - E(UY) - E(XV) + E(UV) &= E(XY) - E(X)E(Y) + E(UV) - E(U)(V) \\ &= Cov(X,Y) + Cov(U,V). \end{aligned}$$

analogicky môžeme upraviť výrazy v rovnici (26) a získame nový systém rovníc

$$\begin{aligned} \lambda a &= (Var(X) + Var(U))a + (Cov(X,Y) + Cov(U,V))b \\ \lambda b &= (Cov(X,Y) + Cov(U,V))a + (Var(Y) + Var(V))b. \end{aligned}$$

Aby sme sprehľadnili nasledujúce úpravy, zaveďme nasledovné označenia:

$$\alpha := Var(X) + Var(U),\ \beta := Var(Y) + Var(V),\ \gamma := Cov(X,Y) + Cov(U,V).$$

Zjavne platí, že $\alpha, \beta > 0$. Rovnice (3.1) a (3.1) v tomto tvare prepíšeme ako

$$\begin{aligned} \lambda a &= \alpha a + \gamma b \\ \lambda b &= \gamma a + \beta b. \end{aligned}$$

Za predpokladu platnosti hypotézy H_0 teda hľadáme vlastné hodnoty matice

$$\phi_{H_0} = \begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix}.$$

Vlastné hodnoty tejto matice sú také λ, pre ktoré platí

$$det\left(\begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) = 0. \tag{27}$$

Vieme, že platí

$$det\left(\begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) = det\begin{pmatrix} \alpha - \lambda & \gamma \\ \gamma & \beta - \lambda \end{pmatrix} = (\alpha - \lambda)(\beta - \lambda) - \gamma^2. \tag{28}$$

Spojením (27) a (28) získame kvadratickú rovnicu

$$\lambda^2 - (\alpha + \beta)\lambda + (\alpha\beta - \gamma^2) = 0. \tag{29}$$

Prvým krokom v skúmaní vlastností koreňov kvadratickej rovnice je nájdenie jej diskriminantu a overenie jeho vlastností. V tomto prípade ukážeme, že diskriminant sa dá zapísať ako suma dvoch štvorcov, teda je nezáporný a rovnica (29) má vždy reálne riešenie:

$$\begin{aligned} D &= (\alpha+\beta)^2 - 4(\alpha\beta - \gamma^2) \\ &= \alpha^2 + \beta^2 - 2\alpha\beta + 4\gamma^2 \\ &= (\alpha-\beta)^2 + 4\gamma^2 \geq 0. \end{aligned}$$

Nulovosť diskriminantu nastáva v prípade, že platí $Var(X)+Var(U) = Var(Y)+Var(V)$ a zároveň $Cov(X,Y) = Cov(U,V) = 0$. V takomto prípade bude koreň dvojnásobný.

Keďže sme ukázali, že parametre λ sa vždy dajú nájsť, môžeme skúmať ich vlastnosti. Ukážeme, že korene rovnice (29) sú nezáporné. Napíšme tieto korene pomocou štandardného vzorčeka:

$$\lambda_{1,2} = \frac{(\alpha+\beta) \pm \sqrt{(\alpha+\beta)^2 - 4(\alpha\beta-\gamma^2)}}{2} \geq 0.$$

Nezápornosť koreňa vo vetve

$$\lambda_1 = \frac{(\alpha+\beta) + \sqrt{(\alpha+\beta)^2 - 4(\alpha\beta-\gamma^2)}}{2}$$

vidno jednoducho, pretože $\alpha, \beta > 0$. Zamerajme sa na druhú vetvu a robme ekvivalentné úpravy:

$$\begin{aligned} \lambda_2 = \frac{(\alpha+\beta) - \sqrt{(\alpha+\beta)^2 - 4(\alpha\beta-\gamma^2)}}{2} &\geq 0 \\ (\alpha+\beta) &\geq \sqrt{(\alpha-\beta)^2 + 4\gamma^2} \\ (\alpha+\beta)^2 - (\alpha-\beta)^2 &\geq 4\gamma^2 \\ 4\alpha\beta &\geq 4\gamma^2 \\ \alpha\beta - \gamma^2 &\geq 0. \end{aligned}$$

Dosaďme pôvodné výrazy za α, β, γ :

$$(Var(X)+Var(U))(Var(Y)+Var(V)) - (Cov(X,Y)+Cov(U,V))^2 \geq 0. \qquad (30)$$

Platnosť tejto nerovnosti ukážeme pomocou lineárnej algebry. Uvažujme náhodný vektor $(P,Q)^\top = (X+U, Y+V)^\top$, pričom $(X,Y)^\top$ a $(U,V)^\top$ sú nezávislé náhodné vektory. Variančno-kovariančná matica tohto vektora je

$$\Sigma_{P,Q} = \begin{pmatrix} Var(X) + Var(U) & Cov(X,Y) + Cov(U,V) \\ +2Cov(X,U) & +Cov(X,V) + Cov(U,Y) \\ Cov(X,Y) + Cov(U,V) & Var(Y) + Var(V) \\ +Cov(Y,U) + Cov(V,X) & +2Cov(Y,V) \end{pmatrix}.$$

Pretože vektory $(X,Y)^\top$ a $(U,V)^\top$ sú nezávislé, platí

$$Cov(X,U) = Cov(X,V) = Cov(Y,U) = Cov(Y,V) = 0,$$

a môžeme zapísať

$$\Sigma_{P,Q} = \begin{pmatrix} Var(X) + Var(U) & Cov(X,Y) + Cov(U,V) \\ Cov(X,Y) + Cov(U,V) & Var(Y) + Var(Y) \end{pmatrix}.$$

Vieme, že variančno-kovariančné matice sú vždy kladne semidefinitné, teda ich determinant musí byť nezáporný. Tým získame nerovnosť

$$(Var(X) + Var(U))(Var(Y) + Var(V)) - (Cov(X,Y) + Cov(U,V))^2 \geq 0,$$

čo je doteraz hľadaná nerovnosť (30) a teda platí aj pôvodná nerovnosť, z ktorej sme vyšli

$$\lambda_2 \geq 0.$$

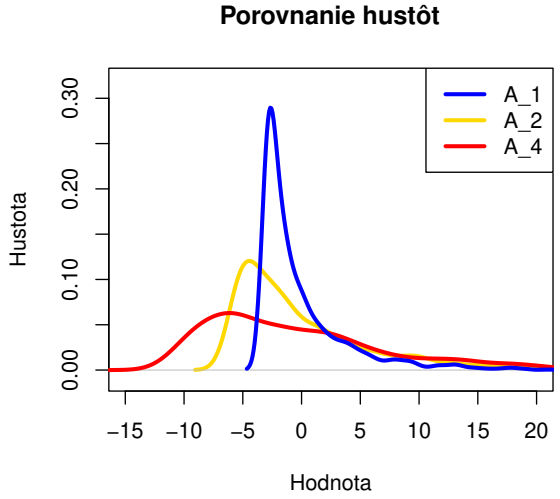
Ukázali sme, že ak platí nulová hypotéza, Fredholmova funkcionálna rovnica (22) má vždy dve reálne vlastné hodnoty (nie nutne rôzne). Obe vlastné hodnoty sú nezáporné a je možné získať ich riešením kvadratickej rovnice (29).

3.2 Ilustrácia asymptotického rozdelenia

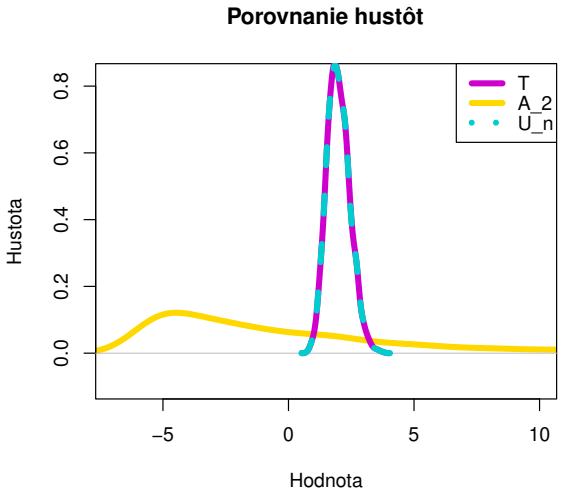
Mathur a Smithová v článku [4] ilustrujú asymptotické rozdelenie štatistík T aj U_n na príklade, v ktorom je porušená nulová hypotéza. V tejto časti práce sa vyjadríme k tomu, prečo je na Obr. 8 viditeľný signifikantný rozdiel medzi údajným asymptotickým rozdelením a nasimulovaným rozdelením štatistík T a U_n. Pokúsime sa tento obrázok zreprodukovať a následne nájdeme skutočné asymptotické rozdelenie testových štatistík v prípade, ktorý je použitý na ilustráciu.

Na Obr. 8 môžeme pozorovať, že testové štatistiky T aj U_n zrejme pochádzajú z podobného rozdelenia, náhodná premenná A je však posunutá doľava. Z pôvodného článku [4] nie je jasné, či autori parameter $\lambda = 3$ považujú za jednonáosbný, dvojnásobný alebo viacnásobný koreň Fredholmovej rovnice (22). Rozhodli sme sa teda nasimulovať rozdelenia $A_k = \sum_{i=1}^{k} 3(Z_i^2 - 1),\ k = 1, 2, 4$.

Na Obr. 9 vykresľujeme simulovanú hustotu náhodných premenných A_k. Vidno, že s rastúcim počtom členov sa zvyšuje disperzia rozdelenia. Hustotu vykresľujeme iba po hodnotu 20, pretože štatistiky T a U_n dosiahli najvyššiu hodnotu 4.21, ako vidno na Obr. 10. Na Obr. 10 vykresľujeme hustoty testových štatistík T a U_n, ktoré sme rovnako ako v prípade 8 vyhodnocovali na základe simulovaných dát $(X, Y)^\top \sim N_2\big((36, 29)^\top, \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)\big)$ a $(U, V)^\top \sim N_2\big((35, 30)^\top, \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)\big)$ s počtom dátových bodov $n = 40$. Dátové sady sme simulovali 1500 krát. Testové štatistiky navyše porovnávame s náhodnou premennou A_2, keďže (ako sme ukázali v kapitole 3.1) náležitá Fredholmova rovnica (22) má vždy práve dve vlastné hodnoty.



Obr. 9: Simulovaná hustota $A_k = \sum_{i=1}^{k} 3(Z_i^2 - 1)$, kde $Z_i \sim N(0, 1)$ pre $k = 1, 2, 4$

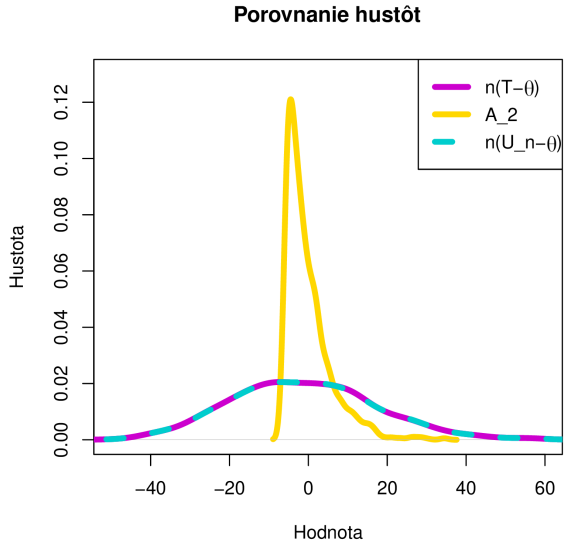


Obr. 10: Simulovaná hustota štatistiky T (7) a štatistiky U_n (18) pre dáta simulované z rozdelení $(X, Y)^\top \sim N_2\left((36, 29)^\top, \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)\right)$ a $(U, V)^\top \sim N_2\left((35, 30)^\top, \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)\right)$.

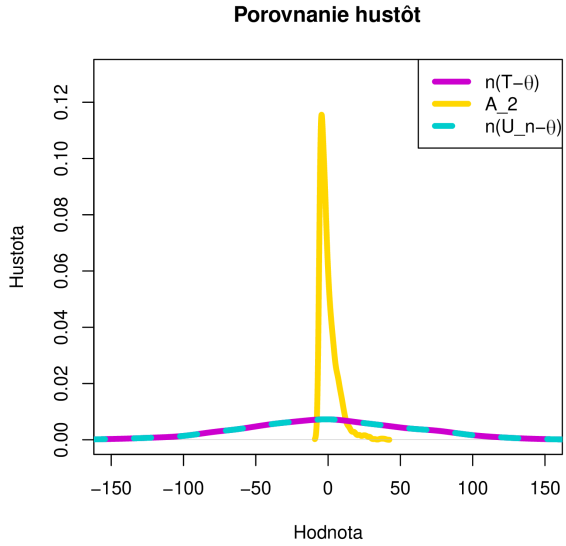
Pri pohľade na Obr. 10 vidno, že testové štatistiky T a U_n nadobúdajú výrazne iné hodnoty ako náhodná premenná A_2. Ide o náš najlepší pokus o reprodukciu Obr. 8 a na základe tohto grafu nekonštatujeme, že sa rozdelenie štatistík T a U_n limitne blíži k

rozdeleniu A_2.

Spomeňme si však na Vetu 1.7, podľa ktorej $n(U_n - \theta) \xrightarrow{d} A$. Mohol by teda znieť nádejne nápad neporovnávať A, U_n a T, miesto toho však porovnať A, $n(U_n - \theta)$ a $n(T-\theta)$. Hodnotu parametra θ určíme pomocou rovnice (20): $\theta = (\mu_1 - m_1)^2 + (\mu_2 - m_2)^2 = 1 + 1 = 2$. Z Obr. 10 vidno, že posun testových štatistík doľava by mohol vycentrovať zobrazené krivky na spoločnú strednú hodnotu. Podľa Vety 1.7 nastáva konvergencia podľa distribučnej funkcie. Je teda možné, že $n = 40$ nie je dostatočný počet dátových bodov na to, aby bola táto konvergencia zjavne viditeľná na obrázkoch a zvýšenie n by mohlo viesť k zlepšeniu tohto fitu. Okrem posunutia pomocou θ teda navyše zvýšime počet simulovaných dátových bodov na $n = 400$.



Obr. 11: Simulovaná hustota štatistiky $n(T - 2)$ a štatistiky $n(U_n - 2)$ pre dáta simulované z rozdelení $(X, Y)^\top \sim N_2\left((36, 29)^\top, \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)\right)$ a $(U, V)^\top \sim N_2((35, 30)^\top, \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right))$ a $n = 40$.



Obr. 12: Simulovaná hustota štatistiky $n(T - 2)$ a štatistiky $n(U_n - 2)$ pre dáta simulované z rozdelení $(X, Y)^\top \sim N_2\left((36, 29)^\top, \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)\right)$ a $(U, V)^\top \sim N_2((35, 30)^\top, \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right))$ a $n = 400$.

Na Obr. 11 vidíme, že nasimulované A rozdelenie má príliš malú disperziu v porovnaní s nasimulovanými $n(T - \theta)$ a $n(U_n - \theta)$. V prípade vyššieho n je táto disperzia ešte väčšia a testové štatistiky sa na údajné asymptotické rozdelenie podobajú ešte menej. Aby sme vysvetlili tento nesúlad, musíme sa vrátiť k základnej myšlienke nášho testovania a k nulovej hypotéze. Asymptotické rozdelenie U_n bolo odvodené za platnosti nulovej hypotézy o tom, že poloha dvoch dátových súborov je rovnaká, teda $\theta = 0$. Dáta použité

na ilustráciu asymptotického rozdelenia však tento predpoklad porušujú. Aký vplyv bude mať neplatnosť nulovej hypotézy na rozdelenie pravdepodobnosti testovej štatistiky?

3.2.1 Asymptotické rozdelenie a neplatnosť H_0

Spomeňme si, ako sú dokázané vlastnosti štatistiky U_n v rovniciach (20) a (3). Za neplatnosti nulovej hypotézy získame

$$\begin{aligned}\theta &= E[U_n(h)] = E[h(\zeta_1, \zeta_2)] \\ &= \mu_1^2 + \mu_2^2 + m_1^2 + m_2^2 - 2\mu_1 m_1 - 2\mu_2 m_2 \\ &= (\mu_1 - m_1)^2 + (\mu_2 - m_2)^2 > 0\end{aligned} \tag{31}$$

a

$$\begin{aligned}\hat{h}_{1H_1}(s) &= E_{H_1}[h(\zeta_1, \zeta_2)|\zeta_1 = s] - \mu_{H_1}(h) \\ &= E(x_1X_2 + y_1Y_2 + u_1U_2 + v_1V_2 - x_1U_2 - y_1V_2 - u_1X_2 - v_1Y_2) - \theta \\ &= x_1\mu_1 + y_1\mu_2 + u_1m_1 + v_1m_2 - x_1m_1 - y_1m_2 - u_1\mu_1 - v_1\mu_2 \\ &\quad - (\mu_1 - m_1)^2 - (\mu_2 - m_2)^2 \\ &= x_1(\mu_1 - m_1) + y_1(\mu_2 - m_2) + u_1(m_1 - \mu_1) + v_1(m_2 - \mu_2) \\ &\quad - (\mu_1 - m_1)^2 - (\mu_2 - m_2)^2 \\ &\not\equiv 0.\end{aligned}$$

To znamená, že $E(\hat{h}_{1H_1}^2) = Var(h_1) \not\equiv 0$. Vyjadrime konkrétne $Var(h_1)$:

$$\begin{aligned}Var(h_1) &= Var(E_{H_1}[h(\zeta_1, \zeta_2)|\zeta_1 = s]) \\ &= Var\left(X_1(\mu_1 - m_1) + Y_1(\mu_2 - m_2) + U_1(m_1 - \mu_1) + V_1(m_2 - \mu_2)\right) \\ &= Var\left(X_1(\mu_1 - m_1) + Y_1(\mu_2 - m_2)\right) + Var\left(U_1(m_1 - \mu_1) + V_1(m_2 - \mu_2)\right) \\ &\quad + 2Cov\left(\left(X_1(\mu_1 - m_1) + Y_1(\mu_2 - m_2)\right), \left(U_1(m_1 - \mu_1) + V_1(m_2 - \mu_2)\right)\right) \\ &= Var\left(X_1(\mu_1 - m_1) + Y_1(\mu_2 - m_2)\right) + Var\left(U_1(m_1 - \mu_1) + V_1(m_2 - \mu_2)\right) + 0\end{aligned} \tag{32}$$

Oba členy posledného riadku tejto úpravy sú jasne nezáporné, keďže ide o varianciu náhodných premenných. Ukážeme, aspoň jeden z týchto členov je kladný. Prvý člen môžeme zapísať ako

$$\begin{aligned}&Var\left(X_1(\mu_1 - m_1) + Y_1(\mu_2 - m_2)\right) = \\ &\qquad \begin{pmatrix}(\mu_1 - m_1), & (\mu_2 - m_2)\end{pmatrix} \begin{pmatrix}Var(X_1) & Cov(X_1, Y_2) \\ Cov(X_1, Y_2) & Var(Y_1)\end{pmatrix} \begin{pmatrix}(\mu_1 - m_1) \\ (\mu_2 - m_2)\end{pmatrix},\end{aligned}$$

teda ako kvadratickú formu variančnej matice vektora $(X, Y)^\top$. Ak je variančná matica kladne definitná, jej kvadratická forma je nulová iba ak $\big((\mu_1 - m_1), (\mu_2 - m_2)\big)^\top = \big(0, 0\big)^\top$, čo v prípade nesplnenej nulovej hypotézy nie je pravda. Podobne môžeme druhý člen zapísať ako kvadratickú formu vektora $(U, V)^\top$.

Variančná matica nie je kladne definitná (a je iba kladne semidefinitná) v prípade, že náhodné premenné sú kolineárne, teda je medzi nimi jednoznačná lineárna závislosť. Ak by boli oba členy (32) nulové, obe variančné matice by museli byť kladne semidefinitné. Avšak to by znamenalo, že by boli prvky X_i a Y_i kolineárne a súčasne by boli kolineárne aj U_i a V_i. Testovací problém by bolo možné previesť do jednorozmerného sveta a nebolo by potrebné použiť viacrozmerný test. Môžeme teda predpokladať, že variančná matica vektora $(X_1, Y_1)^\top$ alebo variančná matica vektora $(U_1, V_1)^\top$ je kladne definitná. Potom platí

$$Var\left(X_1(\mu_1 - m_1) + Y_1(\mu_2 - m_2)\right) > 0$$

alebo

$$Var\left(U_1(m_1 - \mu_1) + V_1(m_2 - \mu_2)\right) > 0$$

a ukázali sme, že platí

$$\begin{aligned} Var(h_1) &= Var\left(X_1(\mu_1 - m_1) + Y_1(\mu_2 - m_2)\right) + Var\left(U_1(m_1 - \mu_1) + V_1(m_2 - \mu_2)\right) \\ &> 0. \end{aligned}$$

Testová štatistika U_n počítaná na dátach pri ktorých bola porušená nulová hypotéza teda nemá degenerovanú kernelovú funkciu. Podľa Vety 1.6 platí

$$n^{1/2}(U_n - \theta) \xrightarrow{d} N(0, 4Var(h_1)). \tag{33}$$

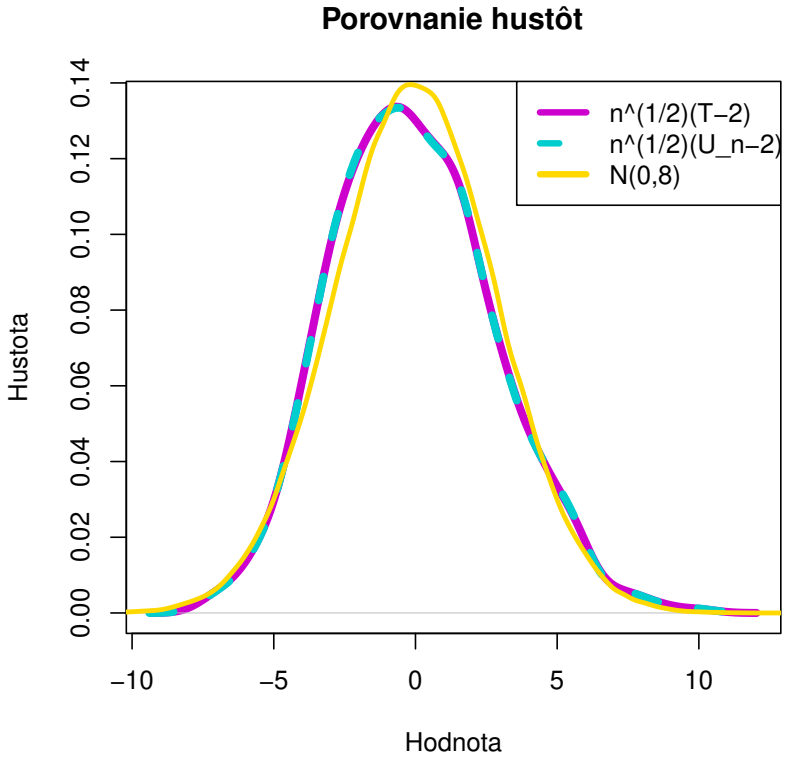
V ilustračnom príklade v článku [4] platí konkrétne

$$\begin{aligned} Var(h_1) &= Var(E_{H_1}[h(\zeta_1, \zeta_2)|\zeta_1 = s]) \\ &= Var\left(X_1(36 - 35) + Y_1(29 - 30) + U_1(35 - 36) + V_1(30 - 29)\right) \\ &= Var(X_1 - Y_1 - U_1 + V_1) \\ &= Var(X_1 - Y_1) + Var(U_1 - V_1) - 2Cov((X_1, Y_1), (U_1, V_1)) \\ &= Var(X_1) + Var(Y_1) - 2Cov(X_1, Y_1) + Var(U_1) + Var(V_1) - 2Cov(U_1, V_1) \\ &= 1 + 1 - 2(0.5) + 1 + 1 - 2(0.5) = 2. \end{aligned}$$

Navyše podľa (31) $E(U_n) = \theta = 2$, a (33) má tvar

$$n^{1/2}(U_n - 2) \xrightarrow{d} N(0, 8).$$

Porovnanie testových štatistík a náležitého normálneho rozdelenia uvádzame na Obr. 13. Vidno, že na rozdiel od Obr. 8 majú všetky tri simulované rozdelenia rovnakú strednú hodnotu a ich priebeh je veľmi podobný.

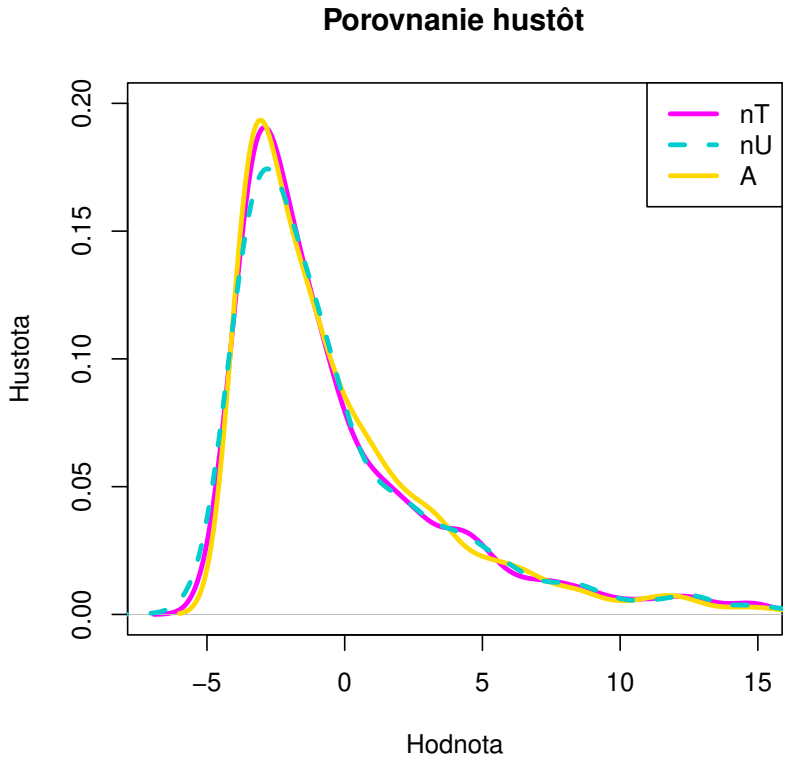


Obr. 13: Porovnanie $n(T-2)$, $n(U_n-2)$ a $N(0,8)$ pre $n = 40$ a rozdelenie dát podľa Example v článku [4, str. 148,149]

Pôvodný ilustračný príklad sme navyše upravili tak, aby skutočne ilustroval fakt, že sa testové štatistiky T a U_n blížia náhodnej premennej $A = \lambda_1(Z_1^2 - 1) + \lambda_2(Z_2^2 - 1)$. Pôvodné stredné hodnoty náhodného vektora $(U, V)^\top$ sme zmenili tak, aby bola splnená nulová hypotéza (6), teda $E(X, Y)^\top = E(U, V)^\top = (36, 29)^\top$. Variančno-kovariančné matice nezmeníme a stále predpokladáme

$$\Sigma_{X,Y} = \Sigma_{U,V} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

Zmena parametrov rozdelení, z ktorých simulujeme dáta, spôsobí, že aj vlastné hodnoty príslušnej Fredholmovej rovnice sa zmenia, v tomto prípade platí $\lambda_1 = 3$ a $\lambda_2 = 1$. Na Obr. 14 možno pozorovať, že simuláciou získané hodnoty T a U_n majú podobné rozdelenie ako simulované hodnoty náhodnej premennej $A = 3(Z_1^2 - 1) + (Z_2^2 - 1)$.



Obr. 14: Porovnanie $n(T - \theta)$, $n(u - \theta)$ a $A = 3(Z_1^2 - 1) + (Z_2^2 - 1)$ pre dáta $(X, Y)^\top \sim N_2((36, 29)^\top, \Sigma)$, $(U, V)^\top \sim N_2(36, 29)^\top, \Sigma), \Sigma = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$.

3.3 Aplikácia T-štatistiky

Podobne ako v článku [3], Mathur a Smithová v článku [4] dedikovali jednu jeho časť názornej aplikácii navrhovaného testu. Aplikáciu vykonali na dvoch rôznych dátových sadách, pričom v prípade jednej z nich sme získali rovnaký výsledok, aký je uvedený v článku [4]. V nasledujúcej kapitole popíšeme, akým spôsobom títo autori aplikovali navrhovanú T-štatistiku a uvedieme úvahu nad možnými nedostatkami prezentovaného prístupu.

V prípade oboch dátových sád bol demonštrovaný veľmi podobný test. Nulová hy-

potéza bola definovaná ako v (6). Pre tú-ktorú sadu dát vypočítali Mathur a Smithová testovú štatistiku T danú vzťahom (7) a hodnotu tejto štatistiky porovnali s kritickou hodnotou, o ktorej tvrdili, že zodpovedá hladine významnosti $\alpha = 0.05$. Prekvapilo nás však, keď sme zistili, ako autori článku [4] získavali kritické hodnoty.

Po siahodlhom odvodzovaní a dokazovaní vlastností príslušnej U-štatistiky a jej asymptotického rozdelenia sme predpokladali, že kľúčom k úspechu bude odhad asymptotického rozdelenia príslušnej U-štatistiky a kritických hodnôt tohto rozdelenia, nasledovaný porovnaním týchto kritických hodnôt a vypočítanej T-štatistiky. Mathur a Smithová sa však rozhodli pre priamočiarejší postup a hranicu významnosti získali permutačným prístupom. Ten pozostával z deviatich krokov:

1. Combine the x-coordinates into one set of size $N = m + n$.
2. Combine the y-coordinates into one set of size $N = m + n$.
3. Draw n x-coordinates and assign to sample 1.
4. Repeat step three to select the y-coordinates of sample 1.
5. The remaining values are assigned to sample 2.
6. Calculate the test statistic T_{cal}.
7. Repeat steps three through six 10,000 times.
8. Arrange the T_{cal} values in ascending order.
9. Identify the $(1 - \alpha)$ value for the stated level of significance, α.

Tieto kroky sme preložili nasledovne:

1. Spojenie dát v rámci 1. súradnice do jednej sady veľkosti $N = m + n$.
2. Spojenie dát v rámci 2. súradnice do jednej sady veľkosti $N = m + n$.
3. Náhodný výber n dátových bodov z 1. súradnice, ktoré budú priradené do vzorky 1.
4. Náhodný výber n dátových bodov z 2. súradnice, ktoré budú priradené do vzorky 1.
5. Priradenie zvyšných dátových bodov do vzorky 2.

6. Výpočet testovej štatistiky T_{cal} pomocou vzorca (7).

7. Zopakovanie krokov 3 až 6 10 000 krát.

8. Zoradenie hodnôt T_{cal} vzostupne.

9. Určenie $(1 - \alpha)$ hodnoty pre danú hladinu signifikantnosti α zo získaných hodnôt T_{cal}.

Tento postup vyzerá takmer ako štandardný permutačný prístup k odhadovaniu kritickej hodnoty. Prekvapil nás však krok 3, podľa ktorého vytvárame celkom nové dvojrozmerné dátové body a nezachovávame pôvodné dvojice dát. Tým strácame informáciu o závislosti v rámci párov a môže sa stať, že vygenerujeme také body, ktoré majú iba veľmi malú pravdepodobnosť existencie a takto získaný kvantil nezodpovedá skutočnému kvantilu rozdelenia, ktorým sa riadi T.

Pre účely implementácie tiež nie je jasné, ako presne vytvoriť vzorku 2 v bode 5. Sú všetky zvyšné prvé a druhé súradnice postupne popárované a tak zahrnuté do sady 2? Alebo majú byť aj tieto súradnice premiešané? Rozhodli sme sa pre náhodné premiešanie aj týchto súradníc.

Ďalším prvkom nejasnosti v tomto postupe je „určenie $(1 - \alpha)$ hodnoty pre danú hladinu signifikantnosti α.“ Typicky sa takáto hodnota určuje buď ako 5%-ný výberový kvantil (v takom prípade je nulová hypotéza zamietnutá, ak je testová štatistika **menšia** ako táto hodnota), 95%-ný výberový kvantil (v tomto prípade je nulová hypotéza zamietnutá, ak je testová štatistika **väčšia** ako táto hodnota), alebo sa určujú dve hodnoty: 2.5%-ný a 97.5%-ný výberový kvantil (v tomto prípade je nulová hypotéza zamietnutá, ak je testová štatistika **menšia** ako 2.5%-ným výberový kvantil, **alebo** ak je testová štatistika **väčšia** 97.5%-ný výberový kvantil). Text v článku [4] nás nabáda k tomu, že $(1 - \alpha)$ hodnotu volia ako výberový $(1 - \alpha)$ kvantil. Popis aplikácie však túto našu teóriu vyvracia. Pozrime sa podrobne na príklad použitia testovej štatistiky uvedený v tomto článku a osvetlime tieto nejasnosti.

Prvá sada dát obsahuje umelo vygenerované údaje „získané“ zo strojov počas dvoch smien v továrni. Tieto dáta uvádzame v Tabuľke 14.

Bohužiaľ nevieme z akého rozdelenia boli tieto dáta generované a či nulová hypotéza skutočne platila alebo nie, autori v článku [4] uvádzajú iba informáciu o homoskedasticite

M	1	2	3	4	5	6	7	8
X_i	24.3	29.2	18.7	19.2	22.4	18.0	16.8	19.7
Y_i	21.3	16.8	19.9	20.1	19.5	20.1	19.3	16.1
U_i	31.3	26.2	25.7	26.3	29.5	25.0	23.8	26.7
V_i	28.3	23.8	26.9	27.1	26.4	27.1	26.2	23.1
M	9	10	11	12	13	14	15	
X_i	18.8	17.3	20.0	15.0	20.6	21.4	20.1	
Y_i	21.9	24.4	21.9	22.2	18.9	21.8	18.4	
U_i	25.7	24.2	27.0	22.0	27.1	28.3	27.1	
V_i	28.8	31.4	28.8	29.1	25.5	28.8	25.4	

Tabuľka 14: Dáta zo smeny A $(X_{1i}, Y_{1i})^\top$ a smeny B $(X_{2j}, Y_{2j})^\top$ zo stroja M.

týchto dát. Každý pár údajov predstavuje meranie kvality a kvantity. V tomto prípade sme aplikáciou vzorca (7) získali hodnotu $T = 86$, čo je hodnota výrazne odlišná od $T = 286$ uvedenej v článku. Permutačným prístupom sme simulovali hodnoty T_{cal} a získané hodnoty výberových kvantilov uvádzame v Tabuľke 15. Autori článku uvádzajú, že permutačnou metódou odhadli, že kritická hodnota je 108, teda na základe hodnoty testovej štatistiky $T = 284$ zamietajú nulovú hypotézu.

kvantil	2.5%	5%	95%	97.5%
hodnota	-4.72	-4.56	9.97	13.05

Tabuľka 15: Hodnoty kvantilov simulovaných T_{cal} z dát v Tabuľke 14.

Nielen, že je naša hodnota testovej štatistiky signifikantne nižšia ako tá uvedená v článku [4], aj naše kritické hodnoty sú výrazne odlišné. Intuitívne dáva zmysel zamietať nulovú hypotézu ak je hodnota testovej štatistiky T príliš vysoká. Keďže je nami získaná hodnota testovej štatistiky $T = 86$ vyššia ako 95%-ný výberový kvantil = 9.97, na základe tohto testu by sme aj my zamietli nulovú hypotézu. Väčší úspech pri získavaní rovnakých testových štatistík, aké sú uvedené v článku [4], sme mali pri druhom uvedenom dátovom súbore.

Druhá dátová sada obsahuje informácie o dĺžkach a šírkach hláv bratov. Tieto dáta

získali Mathur a Smithová z knihy [10]. Použijúc vyššie popísané postupy sme v tomto prípade získali rovnakú hodnotu štatistiky T aj rovnaké kritické hodnoty, ako uvádzajú Mathur a Smithová.

F	1	2	3	4	5	6	7	8	9	10	11	12	13
X_{1i}	191	195	181	183	176	208	189	197	188	192	179	183	174
Y_{1i}	155	149	148	153	144	157	150	159	152	150	158	147	150
X_{2j}	179	201	185	188	171	192	190	189	197	187	186	174	185
Y_{2j}	145	152	149	149	142	152	149	152	159	151	148	147	152

F	14	15	16	17	18	19	20	21	22	23	24	25
X_{1i}	190	188	163	195	186	181	175	192	174	176	197	190
Y_{1i}	159	151	137	155	153	145	140	154	143	139	167	163
X_{2j}	195	187	161	183	173	182	165	185	178	176	200	187
Y_{2j}	157	158	130	158	148	146	137	152	147	143	158	150

Tabuľka 16: Dĺžky a šírky hláv starších bratov $(X_{1i}, Y_{1i})^\top$ a mladších bratov $(X_{2j}, Y_{2j})^\top$ z knihy [10]

Dáta o dĺžkach a šírkach hláv uvádzame v Tabuľke 16. Rovnako ako Mathurovi a Smithovej, aj nám vyšla hodnota testovej štatistiky $T = -4.8$. Zostala jediná otázka; s čím túto hodnotu porovnať? Hodnoty viacerých výberových kvantilov, ktoré sme ziskali pomocou simulácie uvádzame v Tabuľke 17. Autori článku [4] uvádzajú, že nulovú hypotézu o

kvantil	2.5%	5%	95%	97.5%
hodnota	-11.6	-11.4	25.31	31.1

Tabuľka 17: Hodnoty kvantilov simulovaných T_{cal} z dát v Tabuľke 16.

rozdielnej strednej hodnote nezamietajú, pretože dolná kritická hodnota na hladine signifikantnosti $\alpha = 0.05$ je -11.6 a hodnota testovej štatistiky $T = -4.8$. je vyššia. V Tabuľke 17 si môžeme všimnúť, že hodnota -11.6 zodpovedá presne nami nájdenému 2.5%-nému kvantilu simulovaných T_{cal}.

Zhrňme teda vyhodnocovanie testovej štatistiky, ako je uvedené v článku [4]. Autori v jednom prípade nulovú hypotézu zamietajú, pretože hodnota testovej štatistiky T je vyššia

ako kritická hodnota. V druhom prípade testovú štatistiku nezamietajú, pretože hodnota testovej štatistiky T je vyššia ako dolná kritická hodnota. Táto hodnota sa navyše zhoduje s nami získaným 2.5%-ným výberovým kvantilom simulovaných T_{cal}. Domnievame sa teda, že autori uvažovali dve kritické hodnoty, ktoré určili ako 2.5%-ný a 97.5%-ný výberový kvantil simulovaných T_{cal}. Nulovú hypotézu následne zamietali, ak sa hodnota testovej štatistiky T nenachádzala v intervale danom týmito dvomi hodnotami.

Takýto prístup však nepovažujeme za vhodný. Pripomeňme, ako je definovaná testová štatistika T:

$$T = T_{AB} - \frac{T_A + T_B}{2},$$

kde

$$T_{AB} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}[(X_i - U_j)^2 + (Y_i - V_j)^2] = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\|\begin{pmatrix}X_i\\Y_i\end{pmatrix} - \begin{pmatrix}U_j\\V_j\end{pmatrix}\right\|^2,$$

$$T_A = \frac{1}{\binom{n}{2}}\sum_{1\le i<}\sum_{j\le n}[(X_i - X_j)^2 + (Y_i - Y_j)^2] = \frac{1}{\binom{n}{2}}\sum_{1\le i<}\sum_{j\le n}\left\|\begin{pmatrix}X_i\\Y_i\end{pmatrix} - \begin{pmatrix}X_j\\Y_j\end{pmatrix}\right\|^2,$$

$$T_B = \frac{1}{\binom{m}{2}}\sum_{1\le j<}\sum_{k\le m}[(U_j - U_k)^2 + (V_j - V_k)^2] = \frac{1}{\binom{m}{2}}\sum_{1\le j<}\sum_{k\le m}\left\|\begin{pmatrix}U_j\\V_j\end{pmatrix} - \begin{pmatrix}U_k\\V_k\end{pmatrix}\right\|^2.$$

V člene T_{AB} teda zisťujeme priemernú „vzdialenosť“ medzi dátovými bodmi $(X_i, Y_i)^\top$ zo vzorky 1 a $(U_j, V_j)^\top$ zo vzorky 2. V člene T_A zisťujeme priemernú „vzdialenosť“ medzi jednotlivými bodmi vo vzorke 1 a podobne v člene T_B zisťujeme priemernú „vzdialenosť“ medzi bodmi vo vzorke 2. Hodnotu testovej štatistiky získame tak, že porovnáme člen T_{AB} s priemerom z T_A, T_B. Pýtame sa teda, či sú vzdialenosti medzi bodmi z rôznych sád väčšie (alebo prípadne menšie) ako vzdialenosti v rámci sady. Ak je T veľmi malé, znamená to, že vzdialenosti medzi dátami z rôznych sád sú podobné, alebo menšie ako vzdialenosti v rámci každej sady. Pokiaľ je, naopak, T veľké, naznačuje to, že vzdialenosti medzi dátami z rôznych sád sú výrazne väčšie než rozdiely medzi dátami v každej sade jednotlivo. Ak by boli stredné hodnoty jednotlivých dátových sád odlišné, očakávame, že vzdialenosti medzi dátami z rôznych vzoriek sú veľké, teda T bude veľké. Nedáva zmysel zamietať nulovú hypotézu na základe toho, že dáta z dvoch súborov sú si veľmi podobné.

Napriek nejasnostiam v spôsobe zamietnutia nulovej hypotézy a spochybniteľnému prístupu k získavaniu permutovaných dátových sád zostáva jednoznačnou nevýhodou

tohto postupu uvedeného v článku [4] čas potrebný na vyhodnotenie tohto testu a otázno sť permutačného prístupu ak sa dáta líšia aj inak, než iba posunom. Pokiaľ by sme zvolili porovnávanie hodnoty testovej štatistiky T s kritickými hodnotami získanými pomocou jej asymptotického rozdelenia, toto testovanie by bolo rýchlejšie a nemuseli by sme sa spoliehať na permutačný prístup.

3.4 Naša verzia Mathur-Smithovej testu

V predchádzajúcich kapitolách sme sa zaoberali testovou štatistikou T predstavenou a popísanou v článku [4]. V kapitole 3.3 sme sa pokúsili zistiť ako presne autori tohto článku využívali túto testovú štatistiku na testovanie hypotéz o polohe dát.

V tejto kapitole popíšeme iný spôsob použitia testovej štatistiky T, ktorý využíva limitné rozdelenie U-štatistiky U_n a nevyžaduje opakované vyhodnocovanie T pre rôzne permutácie dát. Okrem podrobného popisu postupu uvedieme aj ilustračný príklad použitia našej verzie T-testu.

Podobne ako v prípade pôvodného Mathur-Smithovej T-testu predstaveného v kapitole 3, predpokladajme, že $(X_i, Y_i)^\top,\ i = 1, 2, \ldots, n$ sú merania prináležiace n subjektom z populácie s distribučnou funkciou $F(X, Y)$ a $(U_j, V_j)^\top,\ j = 1, 2, \ldots, n$ sú merania prináležiace n subjektom z rozdelenia s distribučnou funkciou $G(X, Y)$, pričom F a G sú spojité.

Spočítajme hodnotu testovej štatistiky T podľa vzorca (7). V kapitole 3 sme ukázali, že platí

$$T - U_n \xrightarrow{P} 0.$$

O testovej štatistike U_n sme v kapitole 3.1 ukázali, že podľa vety 1.7 môžeme nájsť jej asymptotické rozdelenie ako

$$n(U_n - \theta) \xrightarrow{d} A,$$

kde

$$A = \lambda_1(Z_1^2 - 1) + \lambda_2(Z_2^2 - 1),$$

kde $Z_1, Z_2 \sim N(0, 1)$ a $\lambda_{1,2}$ sú korene kvadratickej rovnice (29).

Chceli by sme ukázať, že A je zároveň asymptotické rozdelenie $n(T-\theta)$, no nemôžeme priamočiaro spojiť (19) a (21). Problémom je rád konvergencie dvoch členov rovnosti (10). Pripomeňme túto rovnosť:

$$T = \binom{n}{2}^{-1} \sum_{1\le i<} \sum_{j\le n} (X_iX_j + Y_iY_j + U_iU_j + V_iV_j - X_iU_j - U_iX_j - Y_iV_j - V_iY_j)$$
$$- \frac{2}{n^2} \sum_{i=1}^{n} (X_iU_i + Y_iV_i) + \frac{2}{n^2(n-1)} \sum_{1\le i<} \sum_{j\le n} (X_iU_j + U_iX_j + Y_iV_j + V_iY_j).$$

Hoci autori článku [4] tvrdia, že posledné dva členy tejto rovnice konvergujú k 0 rýchlosťou $o_P\left(\frac{1}{n}\right)$, v kapitole 3 sme ukázali, že rád tejto konvergencie je iba $o_P(1)$. Avšak vďaka našej oprave znamienka v pôvodnej rovnici (9) môžeme napísať

$$n(T-U_n) = n\left(-\frac{2}{n^2}\sum_{i=1}^{n}(X_iU_i + Y_iV_i) + \frac{2}{n^2(n-1)}\sum_{1\le i<}\sum_{j\le n}(X_iU_j + U_iX_j + Y_iV_j + V_iY_j)\right)$$

$$= -n\left(\frac{2}{n^2}\sum_{i=1}^{n}(X_iU_i + Y_iV_i)\right)$$
$$+ n\left(\frac{2}{n^2(n-1)}\sum_{1\le i<}\sum_{j\le n}(X_iU_j + U_iX_j + Y_iV_j + V_iY_j)\right).$$

Aplikovaním náhodných limít (12) a (17) a Vety 3.2 a) získame

$$n(T-U_n) \xrightarrow{P} -2(E(X)E(U) + E(Y)E(V)) + 2(E(X)E(U) + E(Y)E(V)) = 0. \quad (38)$$

V pôvodnej verzii rovnice (9) z článku [4] by tento prenásobený rozdiel testových štatistík konvergoval k $4(E(X)E(U) + E(Y)E(V))$, čo by síce nemarilo náš prístup k testu Mathura a Smithovej, no komplikovalo by to situáciu, keďže $n(T-\theta)$ by konvergovalo k $4(E(X)E(U) + E(Y)E(V)) + A$.

Spojením tvrdení (38) a (21) dostaneme, že

$$n(T-\theta) \xrightarrow{d} A.$$

Pokiaľ platí nulová hypotéza o rovnakej polohe dát, navyše $\theta = 0$ a máme $nT \xrightarrow{d} A$.

Aby sme odhadli parametre λ_1 a λ_2, musíme nájsť korene upravenej kvadratickej rovnice (29). Keďže nepoznáme skutočné charakteristiky F a G, nahradíme ich našimi

výberovými odhadmi. Z dát spočítajme výberové charakteristiky F:

$$(\overline{X,Y}) = \frac{1}{n}\sum_{i=1}^{n}(X_i, Y_i),\ \widehat{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}),$$
$$\widehat{Var}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,\ \widehat{Var}(Y) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

Analogicky spočítame výberové charakteristiky rozdelenia G, teda $(\overline{U,V}), \widehat{Cov}(U,V), \widehat{Var}(U)$ a $\widehat{Var}(V)$. Budeme teda riešiť „odhadnutú“ kvadratickú rovnicu

$$\lambda^2 - (\hat{\alpha} + \hat{\beta})\lambda + (\hat{\alpha}\hat{\beta} - \hat{\gamma}^2) = 0,$$

kde

$$\hat{\alpha} := \widehat{Var}(X) + \widehat{Var}(U),\ \hat{\beta} := \widehat{Var}(Y) + \widehat{Var}(V),\ \hat{\gamma} := \widehat{Cov}(X,Y) + \widehat{Cov}(U,V)$$

a získame tak korene $\hat{\lambda}_1$ a $\hat{\lambda}_2$.

Keď máme odhad všetkých potrebných parametrov, vygenerujeme R realizácií náhodnej premennej

$$\hat{A} = \hat{\lambda}_1(Z_1^2 - 1) + \hat{\lambda}_2(Z_2^2 - 1).$$

Generovanie týchto realizácií nie je výpočtovo náročné a vo všeobecnosti platí, že väčší počet realizácií nám poskytne presnejšiu informáciu o tomto rozdelení. V našich simuláciach sme volili generovanie $R = 10^6$ realizácií.

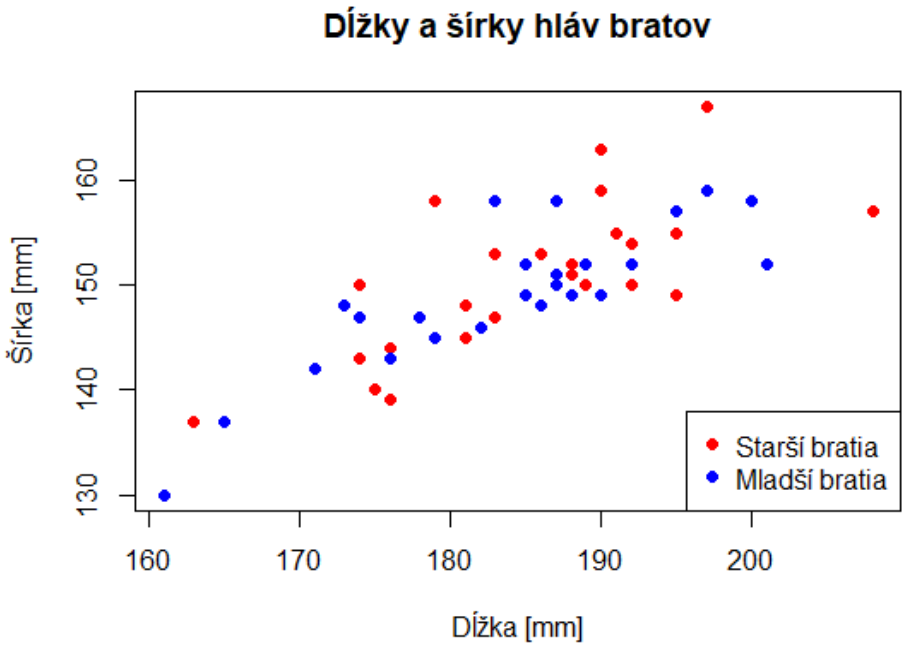
Realizácie náhodnej premennej A zoradíme vzostupne a určíme $Q := (1-\alpha)$-percentný výberový kvantil. Nakoniec porovnáme hodnotu testovej štatistiky T vynásobenú n s výberovým kvantilom Q. Ak je testová štatistika veľmi veľká, teda ak $nT > Q$, nulovú hypotézu o rovnakej polohe dát zamietame. V opačnom prípade nulovú hypotézu nezamietame.

Použitie našej verzie Mathur-Smithovej T-testu budeme ilustrovať na druhom ilustračnom príklade z kapitoly 3.3. Majme dáta o dĺžkach a šírkach hláv bratov z knihy [10] (Tabuľka 16). Pre konzistentnosť urobíme preznačenie pôvodnej tabuľky $X_i := X_{1i}, Y_i := Y_{1i}, U_j := X_{2j}, V_j := Y_{2j}$. Chceme testovať hypotézu o tom, že rozmery hlavy starších súrodencov sú v strednej hodnote rovnaké, ako rozmery hlavy mladších bratov na hladine $\alpha = 5\%$. Ako sme uviedli v predošlej kapitole, hodnota testovej štatistiky $T = -4.8$.

Výberové odhady v tomto prípade sú:

$$(\overline{X,Y}) = (185.72,\ 151.12), \quad (\overline{U,V}) = (183.84,\ 149.24)$$
$$\widehat{Cov}(X,Y) = 52.87, \quad \widehat{Cov}(U,V) = 56.54$$
$$\widehat{Var}(X) = 95.29,\ \widehat{Var}(Y) = 54.436,\ \widehat{Var}(U) = 100.81,\ \widehat{Var}(V) = 45.02,$$
$$\hat{\lambda}_1 = 28.12, \quad \hat{\lambda}_2 = 267.36.$$

Výberový kvantil nám vyšiel $Q = 757.23$. Test nakoniec vyhodnotíme tak, že porovnáme $nT = -120$ a Q. Keďže platí $nT < Q$, nulovú hypotézu nezamietame. Porovnávať hodnoty -120 a 757.23 môže na prvý pohľad pôsobiť mierne extrémne, ale ilustrácia dát na Obr. 15 napovedá, že náš test jednoznačne nemal dôvod zamietnuť nulovú hypotézu.



Obr. 15: Dáta o rozmeroch hláv starších a mladších bratov.

Pri druhej ukážke použitia nášho testu použijeme známe dataset *Iris* (Kosatec), ktorý prvýkrát predstavil v roku 1936 R. A. Fisher v článku [11]. Tento dataset obsahuje dáta o nameraných dĺžkach a šírkach listov a lupeňov kosatcov z troch rôznych druhov (setosa, versicolor a virginica). My sme sa rozhodli pracovať dĺžkami a šírkami lupeňov druhov versicolor a virginica. Náhodne sme vybrali 25 meraní druhu versicolor a 25 meraní druhu virginica. Dáta uvádzame v Tabuľke 18.

i	1	2	3	4	5	6	7	8	9	10	11	12	13
X_i	4.6	4.0	3.3	4.7	4.2	3.0	4.5	4.1	4.3	4.5	4.9	3.5	4.3
Y_i	1.5	1.3	1.0	1.4	1.3	1.1	1.5	1.0	1.3	1.5	1.5	1.0	1.3
U_i	5.7	6.6	5.0	5.4	6.1	5.8	4.5	5.3	5.6	5.3	5.5	6.7	5.4
V_i	2.5	2.1	2.0	2.1	2.3	1.6	1.7	1.9	1.8	2.3	1.8	2.0	2.3
i	14	15	16	17	18	19	20	21	22	23	24	25	
X_i	3.6	4.7	4.8	3.8	4.6	4.0	4.2	4.2	4.4	3.9	5.0	3.9	
Y_i	1.3	1.5	1.8	1.1	1.3	1.3	1.3	1.5	1.3	1.4	1.7	1.2	
U_i	6.1	5.1	6.1	5.6	5.5	5.1	6.0	5.1	6.4	4.9	5.0	5.2	
V_i	2.5	1.9	1.9	2.4	2.1	1.9	1.8	1.8	2.0	1.8	1.9	2.3	

Tabuľka 18: Dĺžka a šírka lupeňov Iris versicolor (X_i, Y_i) a Iris virginica (U_i, V_i).

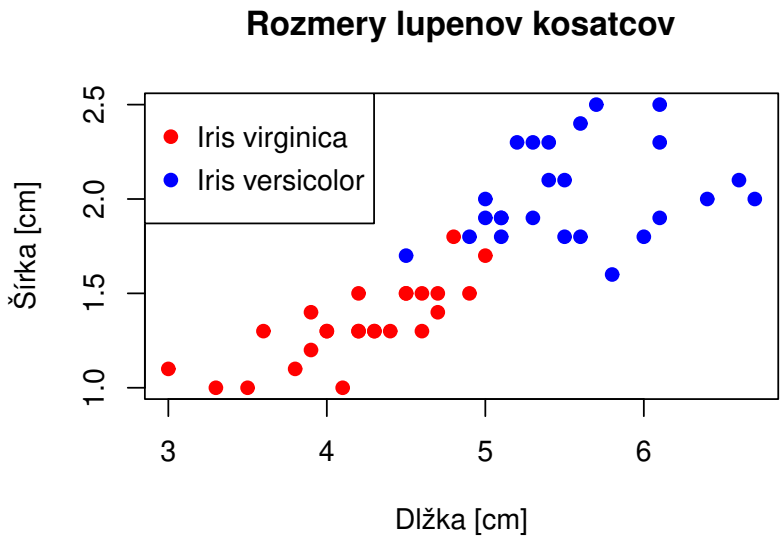
Pri týchto dátach sme vypočítali testovú štatistiku $T = 2.3$. Výberové odhady sú:

$$(\overline{X,Y})^\top = (4.2,\ 1.37)^\top, \quad (\overline{U,V})^\top = (5.56,\ 2.03)^\top$$
$$\widehat{Cov}(X,Y) = 0.081, \quad \widehat{Cov}(U,V) = 0.035$$
$$\widehat{Var}(X) = 0.25,\ \widehat{Var}(Y) = 0.41,\ \widehat{Var}(U) = 0.31,\ \widehat{Var}(V) = 0.064,$$
$$\hat{\lambda}_1 = 0.077, \quad \hat{\lambda}_2 = 0.59.$$

Výberový kvantil nám pri dátach o kosatcoch vyšiel $Q = 1.67$. Ak porovnáme $nT = 57.55$ a Q, zistíme, že $nT > Q$, a nulovú hypotézu zamietame. Vizualizáciu dát pridávame na Obr. 16.

3.5 Simulačné porovnanie našej verzie Mathur-Smithovej testu a iných testov

Aby sme overili vlasntosti našej verzie Mathur-Smithovej testu, v tejto kapitole ho porovnáma s niekoľkými inými neparametrickými viacrozmernými testami polohy. Rozhodli sme sa pre porovnanie vybrať test, ktorý popísal H. Oja v článku [12], Hotellingov dvojvýberový T^2-test a Blumenov test. Posledný v uvedených testov nie je dvojvýberový, preto pri implementácii síce vygenerujeme dve dátové sady, no tie od seba vzájomne odčítame a získame tak jednu sadu, ktorej stredná hodnota by za platnosti nulovej hy-



Obr. 16: Šírka a dĺžka lupeňov rôznych druhov kosatca.

potézy mala mať strednú hodnotu v bode $(0,0)^\top$.

Vo všetkých simuláciach sme simulovali dve skupiny dát veľkosti $n = 25$ z dvojrozmerných rozdelení. Stredná hodnota prvého dátového súboru bola vždy $(0,0)^\top$ a stredná hodnota druhého súboru bola $(\delta_1, \delta_2)^\top$. Po výpočte odhadnutých parametrov asymptotického rozdelenia A sme simulovali $R = 10^6$ realizácií tohto rozdelenia a na základe nich sme určili výberový 95%-ný kvantil Q s ktorým sme následne porovnali hodnotu testovej štatistiky. Pre každú dvojicu $(\delta_1, \delta_2)^\top$ sme sme túto simuláciu zopakovali $N = 10^5$ krát. V tabuľkách nižšie uvádzame v akej časti prípadov ten-ktorý test zamietol nulovú štatistiku o rovnakej polohe dát.

Na Obr. 17 vidno, že sila nášho T testu sa zdá byť vyššia ako sila Hotellingovho T^2 testu. To nás veľmi prekvapilo, pretože je známe, že v prípade dát pochádzajúcich z normálneho rozdelenia je práve Hotellingov T^2 test najsilnejší afínne invariantný test polohy. Ak sa však podrobnejšie pozrieme na hodnotu simulovanej pravdepodobnosti chyby 1. rádu (teda pravdepodobnosť, že test zamietne nulovú hypotézu, ktorá platí) v Tabuľkách 19 až 22, zistíme, že táto pravdepodobnosť je o čosi vyššia ako 5%. Môžeme teda pozorovať, že náš test je takpovediac „liberálny“ a celkovo má tendenciu často zamietať nulovú hypotézu.

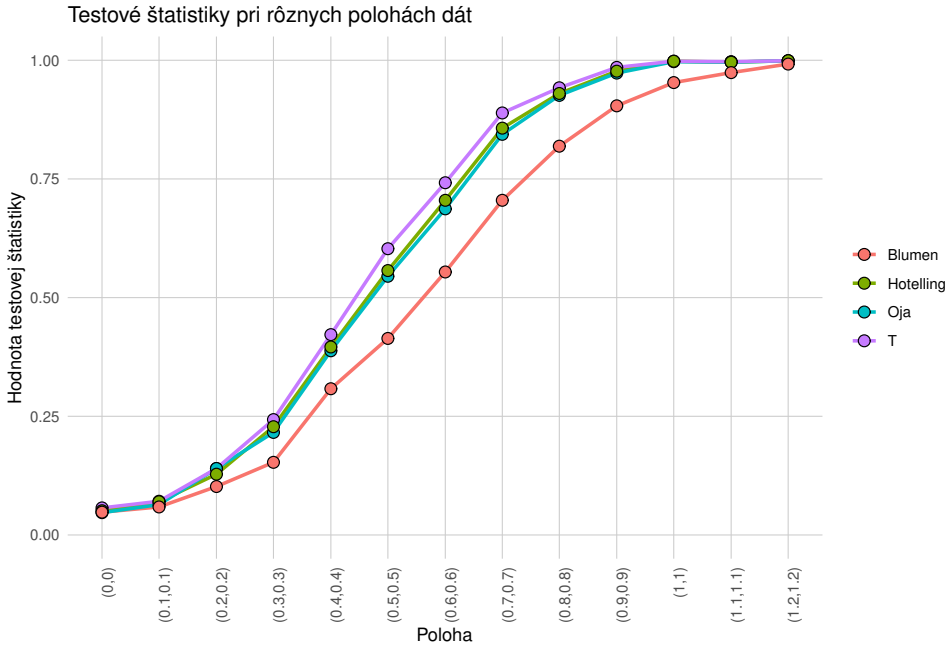
Na Obr. 18 môžeme pozorovať, že náš test veľmi spoľahlivo zamieta neplatnú nulovú

δ_1	δ_2	T	Oja	Hotelling	Blumen
0.00	0.00	0.057	0.047	0.051	0.048
0.10	0.10	0.071	0.064	0.070	0.059
0.20	0.20	0.140	0.140	0.128	0.102
0.30	0.30	0.243	0.216	0.228	0.153
0.40	0.40	0.422	0.388	0.396	0.308
0.50	0.50	0.603	0.545	0.557	0.414
0.60	0.60	0.742	0.687	0.705	0.554
0.70	0.70	0.889	0.844	0.857	0.705
0.80	0.80	0.942	0.926	0.930	0.819
0.90	0.90	0.985	0.973	0.977	0.904
1.00	1.00	0.998	0.997	0.998	0.953
1.10	1.10	0.997	0.996	0.996	0.974
1.20	1.20	0.999	0.999	0.999	0.992

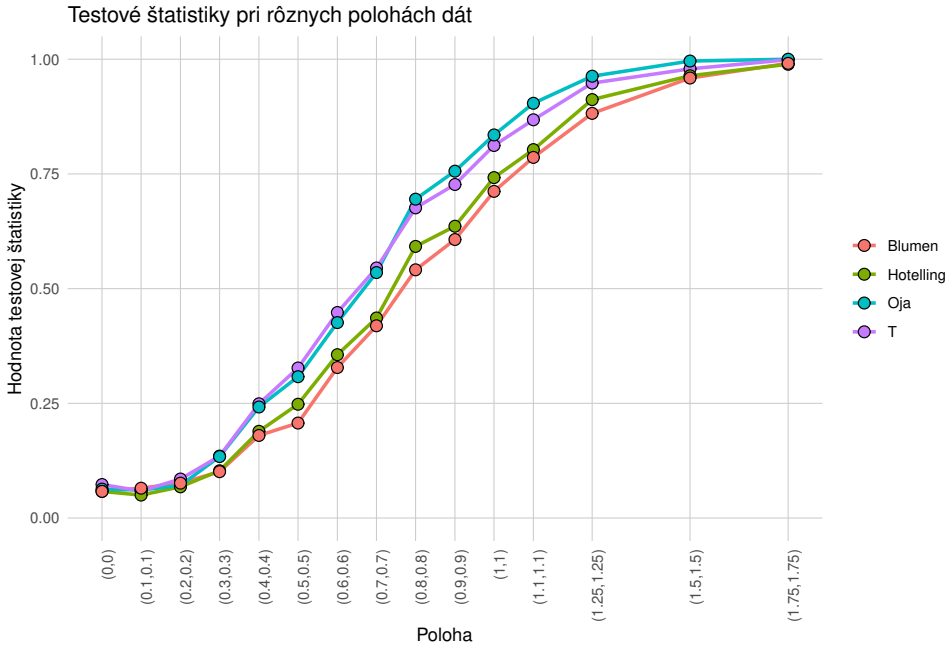
Tabuľka 19: Porovnanie sily nášho T testu, Oja, Hotelling a Blumen. Obe sady dát pochádzajú z normálneho rozdelenia: $(X,Y)^\top \sim N_2\left((0,0)^\top, \left(\begin{smallmatrix}1 & 0.5\\ 0.5 & 1\end{smallmatrix}\right)\right)$ a $(U,V)^\top \sim N_2\left((\delta_1,\delta_2)^\top, \left(\begin{smallmatrix}1 & 0.5\\ 0.5 & 1\end{smallmatrix}\right)\right)$.

hypotézu pri dátach simulovaných z rozdelenia s veľmi ťažkými chvostami a je porovnateľne silný ako test Oja. Na obrázkoch 19 a 20 však vidno, že v prípade dát pochádzajúcich z rozdelenia s tenšími chvostami, prípadne z asymetrického rozdelenia, je sila nášho testu výrazne nižšia ako sila „konkurenčných“ testov.

V článku [4] autori uvádzajú výsledky vlastných simulačných štúdií, bohužiaľ neuvádzajú konkrítne rozdelenia, z ktorých boli dátové sady simulované a nemohli sme jednoznačne porovnať naše výsledky s výsledkami uvedenými v tomto článku. Pri pokuse o reprodukciu dát uvedených v tomto článku sme nenašli také rozdelenia, pri ktorých by mal Mathur-Smithovej T test takú silu, ako je v článku uvedené.



Obr. 17: Porovnanie sily testov na dátach z dvojrozmerného normálneho rozdelenia.



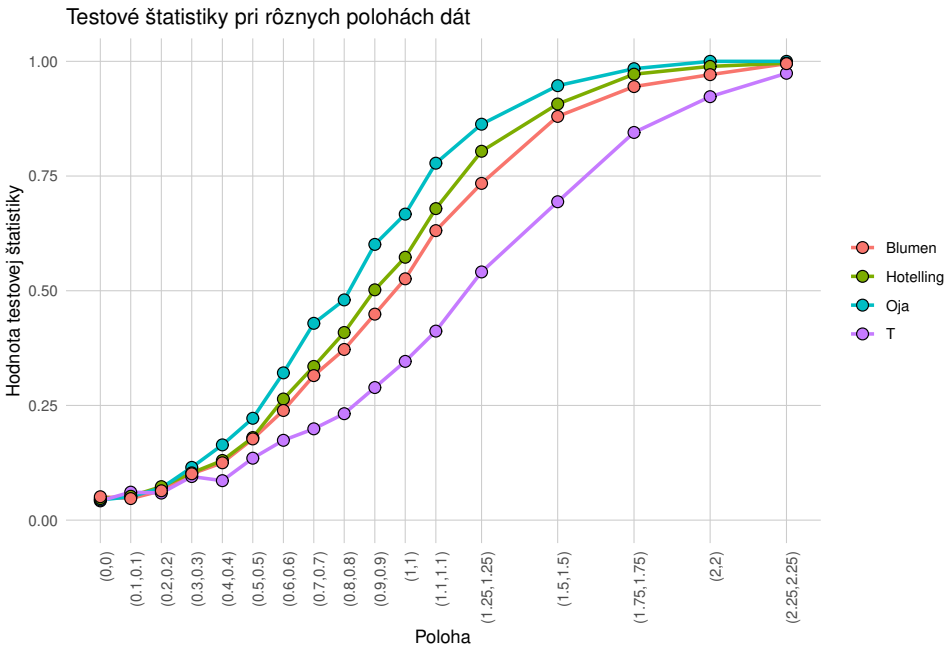
Obr. 18: Porovnanie sily testov na dátach z dvojrozmerného t-rozdelenia so 4 stupňami voľnosti.

δ_1	δ_2	T	Oja	Hotelling	Blumen
0.00	0.00	0.073	0.063	0.058	0.058
0.10	0.10	0.059	0.061	0.050	0.065
0.20	0.20	0.085	0.071	0.068	0.076
0.30	0.30	0.135	0.134	0.103	0.101
0.40	0.40	0.249	0.242	0.189	0.180
0.50	0.50	0.327	0.308	0.248	0.207
0.60	0.60	0.448	0.426	0.356	0.328
0.70	0.70	0.545	0.535	0.436	0.419
0.80	0.80	0.676	0.695	0.592	0.541
0.90	0.90	0.727	0.756	0.636	0.607
1.00	1.00	0.812	0.835	0.742	0.712
1.10	1.10	0.868	0.904	0.803	0.786
1.25	1.25	0.948	0.963	0.912	0.882
1.50	1.50	0.979	0.996	0.964	0.959
1.75	1.75	0.999	1.000	0.989	0.991

Tabuľka 20: Porovnanie sily nášho T testu, Oja, Hotelling a Blumen. Volili sme: $R = 1000000$, $n = 25, N = 10000$, $(X, Y) \sim t_2(df = 4, \rho = 0.5, \mu = (0, 0))$ a $(U, V) \sim t_2(df = 4, \rho = 0.5, m = (\delta_1, \delta_2))$.

δ_1	δ_2	T	Oja	Hotelling	Blumen
0.00	0.00	0.042	0.044	0.047	0.051
0.10	0.10	0.061	0.051	0.052	0.047
0.20	0.20	0.059	0.070	0.073	0.064
0.30	0.30	0.095	0.115	0.103	0.101
0.40	0.40	0.086	0.164	0.130	0.125
0.50	0.50	0.135	0.222	0.180	0.177
0.60	0.60	0.174	0.321	0.264	0.239
0.70	0.70	0.199	0.429	0.335	0.315
0.80	0.80	0.232	0.480	0.409	0.372
0.90	0.90	0.289	0.601	0.502	0.449
1.00	1.00	0.346	0.667	0.573	0.526
1.10	1.10	0.412	0.778	0.679	0.631
1.25	1.25	0.541	0.863	0.804	0.734
1.50	1.50	0.694	0.947	0.907	0.880
1.75	1.75	0.845	0.984	0.972	0.945
2.00	2.00	0.923	1.000	0.989	0.971
2.25	2.25	0.974	1.000	0.996	0.995

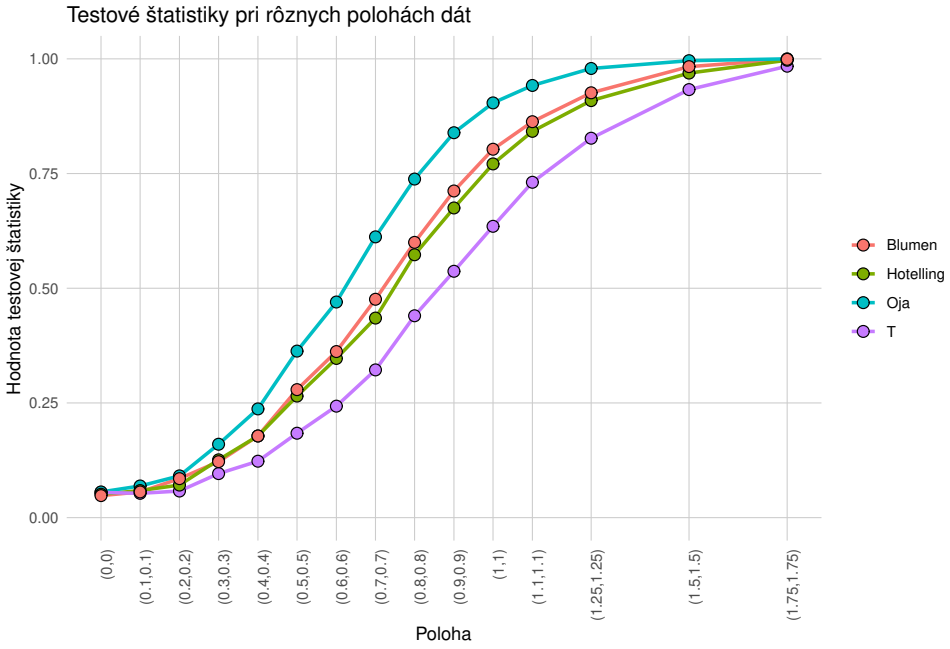
Tabuľka 21: Porovnanie sily nášho T testu, Oja, Hotelling, Blumen. Volili sme: $R = 1000000$, $n = 25, N = 10000$, rozdelenie Bivariate Gamma, $\alpha_1 = 2, \alpha_2 = 7.5$, scale $= 1$.



Obr. 19: Porovnanie sily testov na dátach z dvojrozmerného Gamma rozdelenia.

δ_1	δ_2	T	Oja	Hotelling	Blumen
0.00	0.00	0.055	0.056	0.051	0.048
0.10	0.10	0.053	0.069	0.059	0.056
0.20	0.20	0.058	0.091	0.071	0.085
0.30	0.30	0.096	0.160	0.126	0.122
0.40	0.40	0.123	0.237	0.178	0.178
0.50	0.50	0.184	0.363	0.265	0.279
0.60	0.60	0.243	0.470	0.347	0.362
0.70	0.70	0.322	0.612	0.435	0.476
0.80	0.80	0.440	0.738	0.573	0.600
0.90	0.90	0.537	0.839	0.675	0.712
1.00	1.00	0.635	0.904	0.771	0.803
1.10	1.10	0.731	0.942	0.842	0.863
1.25	1.25	0.827	0.979	0.909	0.926
1.50	1.50	0.933	0.996	0.969	0.983
1.75	1.75	0.984	1.000	0.997	0.999

Tabuľka 22: Porovnanie sily nášho T testu, Oja, Hotelling, Blumen. Volili sme: $R = 10^6$, $n = 25, N = 10^4$, rozdelenie Gumbel exponential, $\lambda_1 = 0.5, \lambda_2 = 0.5, \theta = 0.5$



Obr. 20: Porovnanie sily testov na dátach z dvojrozmerného Gumbelovho rozdelenia.

Záver

Cieľom našej diplomovej práce bolo pochopiť a prípadne upraviť niektoré už existujúce viacrozmerné neparametrické testy polohy. V prvej časti práce sme popísali niektoré známe testy a venovali sme špeciálnu pozornosť teórii U-štatistík.

V druhej kapitole sme sa pokúsili upresniť niektoré nejasnosti ohľadom testovej štatistiky v článku [3]. V snahe zreprodukovať výsledky, ktoré získal autor a publikoval vo svojom článku, sme testovali vlastnosti až deviatich testových štatistík, napokon sme však ani jednou z nich nedosiahli rovnaké výsledky ako S. K. Mathur.

Tretia kapitola bola venovaná dvojsúborovému Mathur-Smithovej testu, ktorý testuje hypotézu o tom, že dve dátové sady majú rovnakú strednú hodnotu. Hodnota testovej štatistiky je založená na porovnaní priemernej vzdialenosti bodov z dvoch rôznych sád s priemernou vzdialenosťou bodov v rámci prvej a v rámci druhej sady. Autori v pôvodnom článku ukázali, že testová štatitika sa limitne blíži k U-štatistike, no v kapitole 3 sme uviedli na pravú mieru tvrdenia o konvergenciách niektorých výrazov. V kapitole 3.1 sme rozvinuli teóriu o hľadaní limitného rozdelenia príslušnej U-štatistiky a jeho vlastnostiach. Limitné rozdelenie autori v pôvodnom článku ilustrovali obrázkom, ktorý ukazoval, že ich testová štatistika a U-štatistika k sebe síce konvergujú, no nájdené asymptotické rozdelenie bolo zjavne odlišné. Vysvetlili sme prečo došlo k rozdielu medzi testovými štatistikami a asymptotickým rozdelením, opravili sme ilustračný príklad a v kapitole 3.2.1 sme odvodili asymptotické rozdelenie platné v ilustračnom príklade, ktorý bol pôvodne použitý. V ďalšej kapitole sme sa vyjadrili k aplikácii testovej štatistiky tak, ako ju popísali Mathur a Smithová v článku [4].

Napokon sme v kapitole 3.4 uviedli upravenú verziu testu Mathura a Smithovej. Dokázali sme, že, že testová štatistika skutočne konverguje k rovnakému asymptotickému rozdeleniu, ako príslušná U-štatistika. Na rozdiel od pôvodného testu, našu verziu je možné aplikovať aj v prípadoch, kedy nie je možné využiť permutačné testy, respektíve je možnosť ich využitia otázna. Nevýhodou našej verzie je však vysoká pravdepodobnosť chyby I. druhu, teda pravdepodobnosť, že tento test zamietne hypotézu o rovnakej polohe aj v prípade, že táto hypotéza platí, je vyššia ako 5%. V poslednej kapitole tejto práce uvádzame porovnanie sily nášho testu a niektorých iných viacrozmerných neparametrických testov polohy.

Viacrozmerných neparametrických testov polohy, ktoré by sa oplatilo ďalej skúmať, je mnoho. Ďalším výskumom v oblasti našej diplomovej práce by mohlo byť preskúmanie testu Mathura a Smithovej za predpokladu, že dva súbory majú iný počet dátových bodov, prípadne jeho verzia pre viac ako dva súbory.

Zoznam použitej literatúry

[1] Filová, L., Sűcz, G.: *Viacrozmerné štatistické analýzy: poznámky k prednáškam*, FMFI UK, Bratislava, 2023

[2] Blumen, I.: *A New Bivariate Sign Test,* Journal of the American Statistical Association, vol. 53, no. 282, pp. 448–456, 1958

[3] Mathur, S. K.: *A Nonparametric Test for the One-Sample Bivariate Location Problem,* American Journal of Mathematical and Management Sciences, vol. 26(1–2), pp. 71–92, 2006

[4] Mathur, S. K., Smith, P. F.: *An efficient nonparametric test for bivariate two-sample location problem,* Statistical Methodology, vol. 5(2), pp. 142-159, 2008

[5] Peters, D., Randles, R. H.: *A Multivariate Signed-Rank Test for the One-Sample Location Problem,* Journal of the American Statistical Association, vol. 85(410), pp. 552–557, 1990

[6] Sen, K., Mathur, S. K.: *A bivariate signed rank test for two sample location problem,* Communications in Statistics - Theory and Methods, vol. 26(12), pp. 3031–3050, 1997

[7] Somogyi, P.: *Viacrozmerné neparametrické testy polohy dvoch súborov,* diplomová práca, FMFI UK, Bratislava, 2021

[8] Ryan, T. A., Joiner B.L., Ryan B. F.: *Minitab Student Handbook,* Duxbury Press, Minnesota, pp. 276, 1976

[9] Serfling, R. J.: *Approximation Theorems of Mathematical Statistics,* John Wiley, New York, 1980.

[10] Anderson, T.W.: *An Introduction to Multivariate Statistical Analysis*, John Wiley, New York, 1984.

[11] Fisher, R. A.: *The use of multiple measurements in taxonomic problems,* Annals of Eugenics, vol. 7(2), pp. 179–188, 1936

[12] Hettmansperger T.P., Oja H.: *Affine invariant multivariate multisample sign tests,* Journal of the Royal Statistical Society, Series B 56, pp. 235–249, 1994.

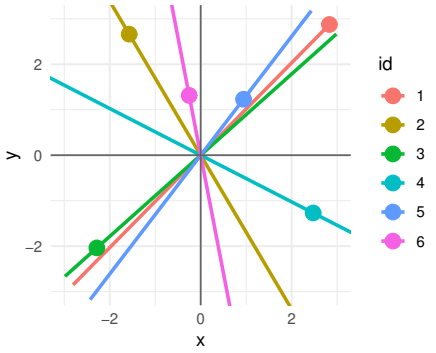
Príloha A

V tejto prílohe uvádzame ilustráciu Blumenovho testu a ilustrácie rôznych verzií Mathurovho S-testu spolu s hodnotami testových štatistík.

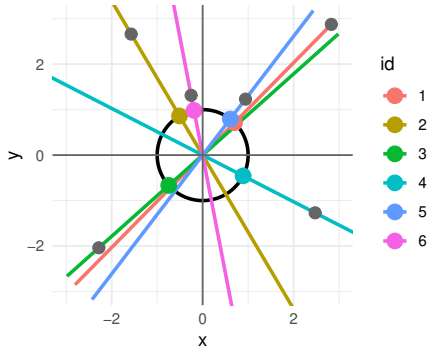
Tieto prístupy znázorníme na náležitých ilustráciach. Najprv sme vygenerovali 6 náhodných bodov (X_i, Y_i) tak, aby v každom zo 4 kvadrantov ležal aspoň 1 bod. Tieto body uvádzame v Tabuľke 23. Tieto body sme premietli na jednotkovú kružnicu. V prípade každej metódy sme postupovali inak, tieto kroky znázorníme na obrázkoch s popisom. Keďže máme iba málo dát, vplyvom náhody sa stalo, že hodnota niektorých testových štatistík vyšla rovnaká.

i	1	2	3	4	5	6
X_i	2.83	-1.57	-2.28	2.47	0.94	-0.25
Y_i	2.87	2.66	-2.04	-1.27	1.23	1.31

Tabuľka 23: Ilustračné body.

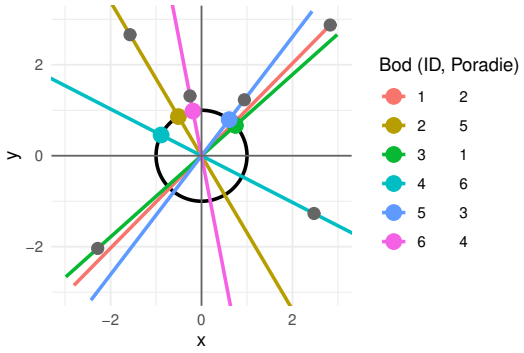


Obr. A.21: Vygenerované dáta

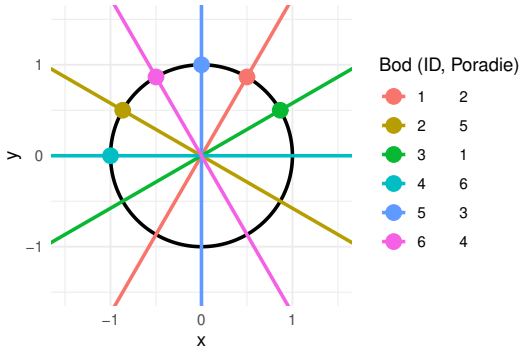


Obr. A.22: Premietnutie dátových bodov na 1-kovú kružnicu

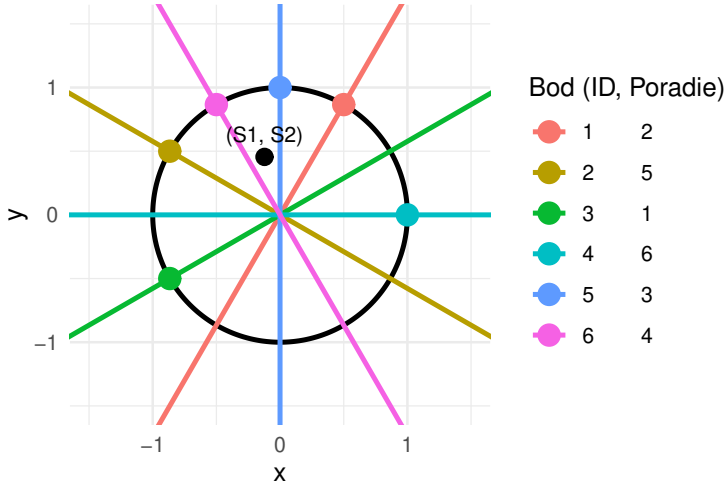
A.1 Ilustrácia Bluemnovej metódy



Obr. A.23: Premietnutie dátových bodov nad x-ovú os



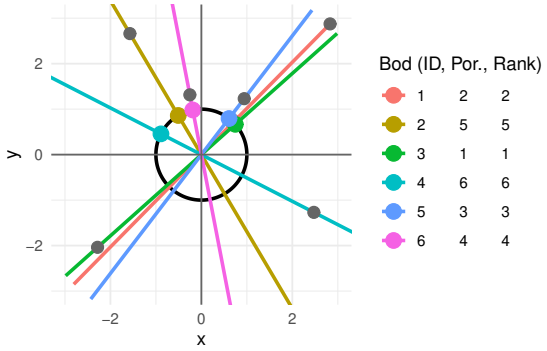
Obr. A.24: Rozostavenie do rovnomerných uhlových vzdialeností



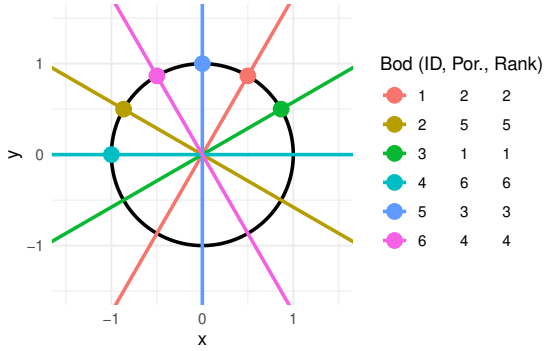
Obr. A.25: Premietnutie na pôvodnú stranu x-ovej osi. Štandardizovaný n-uholník a jeho ťažisko. $v^2 = 0.22$

A.2 Ilustrácia S-testov s rankovaním 1

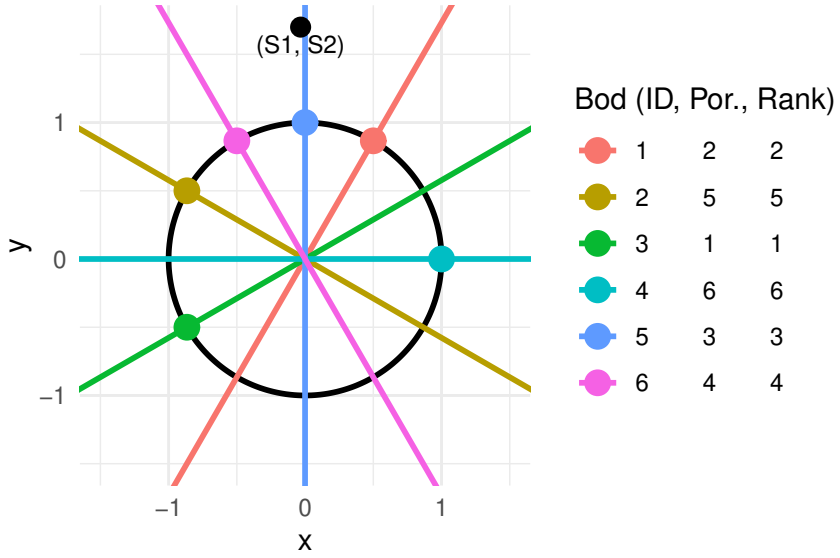
A.2.1 S_α^1



Obr. A.26: Určenie ranku a poradia dát

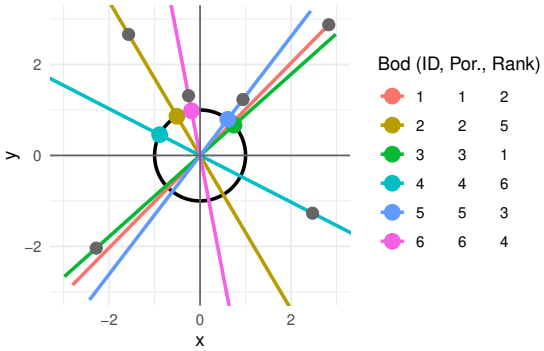


Obr. A.27: Zoradenie dát

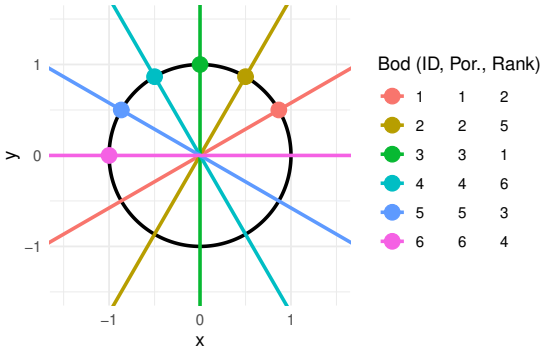


Obr. A.28: Štandardizovaný n-uholník a jeho váhované ťažisko. S_γ^1=2.88

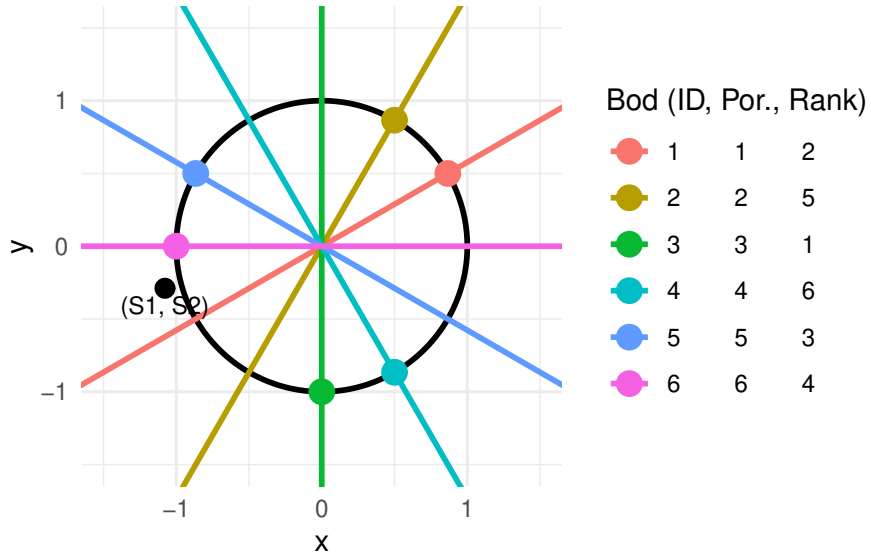
A.2.2 S_γ^1



Obr. A.29: Určenie ranku a poradia dát



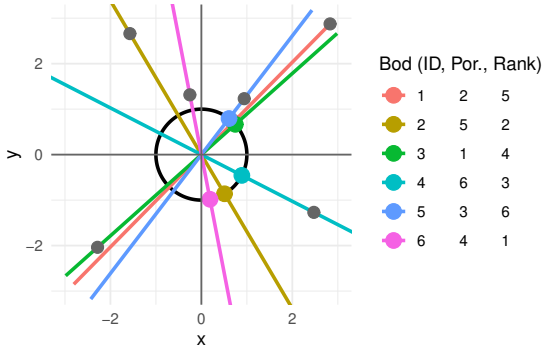
Obr. A.30: Zoradenie dát



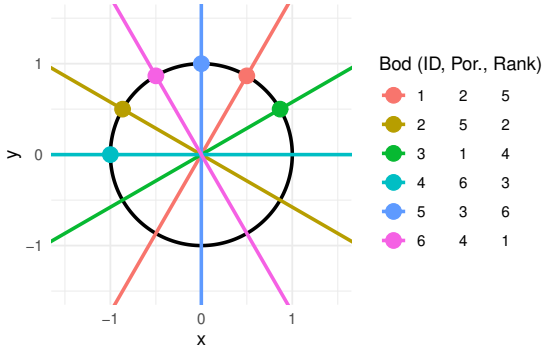
Obr. A.31: Štandardizovaný n-uholník a jeho váhované ťažisko. S_γ^1=1.24

A.3 Ilustrácia S-testov s rankovaním 2

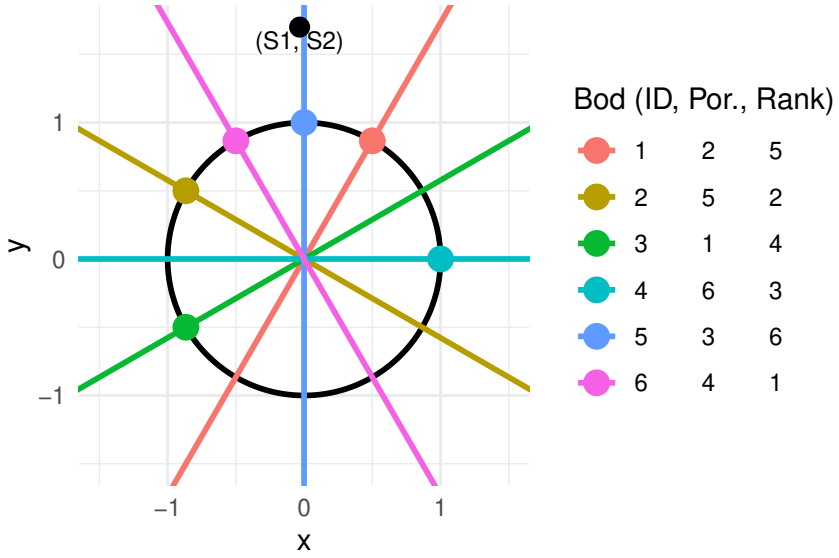
A.3.1 S_α^2



Obr. A.32: Určenie ranku a poradia dát

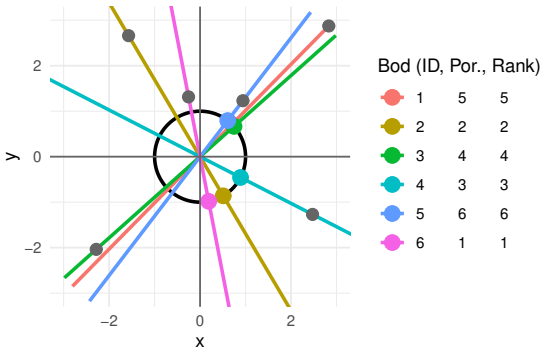


Obr. A.33: Zoradenie dát

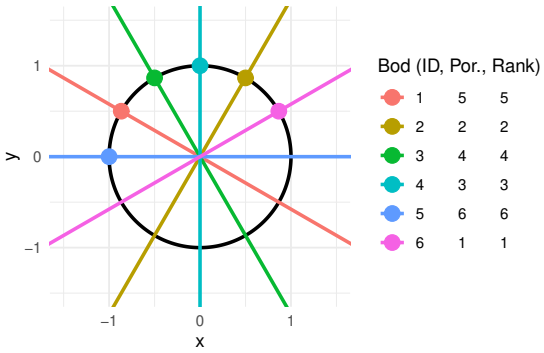


Obr. A.34: Štandardizovaný n-uholník a jeho váhované ťažisko. S_α^2=2.88

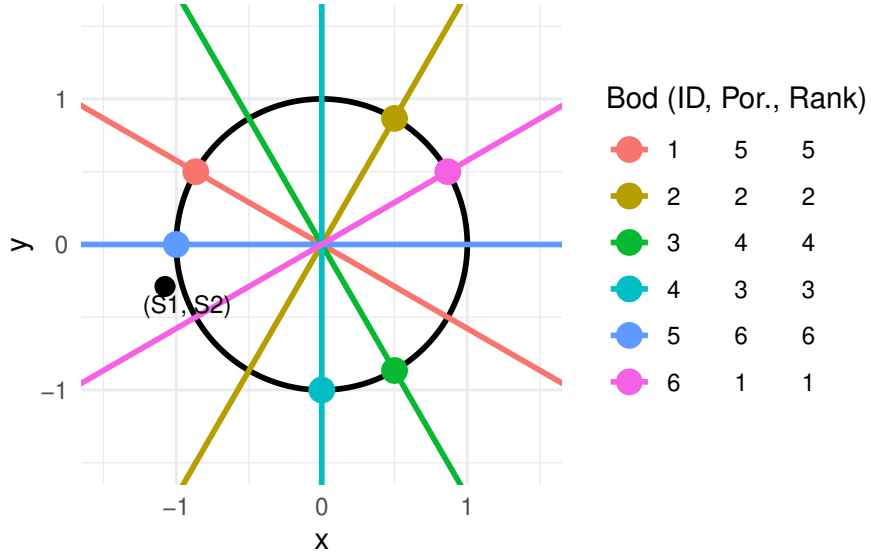
A.3.2 S_β^2



Obr. A.35: Určenie ranku a poradia dát

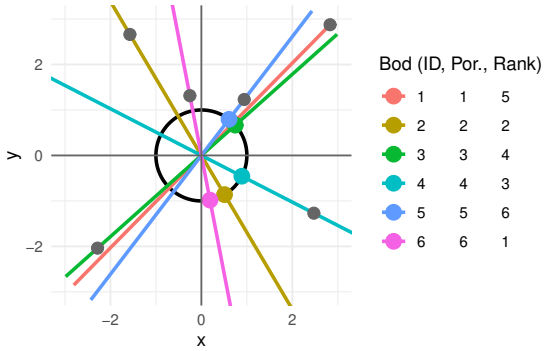


Obr. A.36: Zoradenie dát

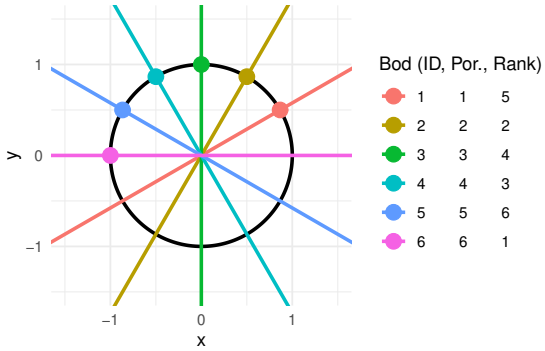


Obr. A.37: Štandardizovaný n-uholník a jeho váhované ťažisko. S_β^2=1.12

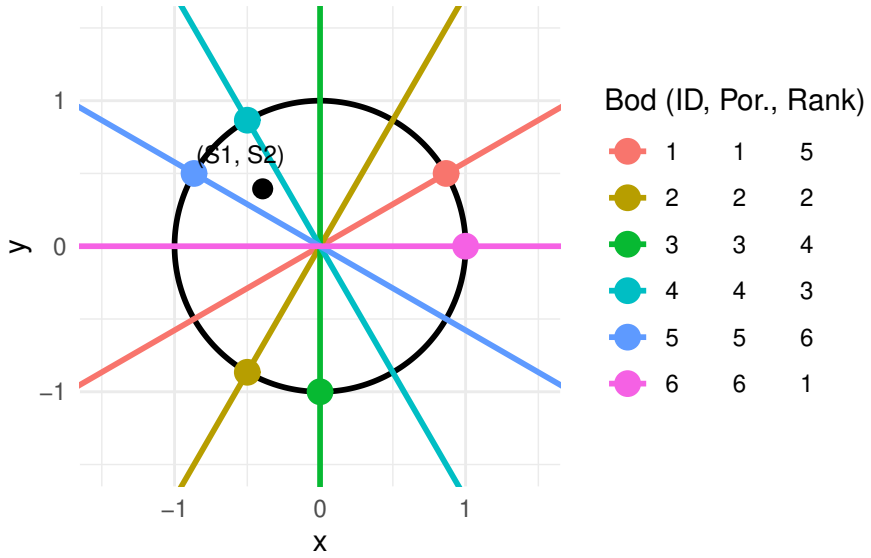
A.3.3 S_γ^2



Obr. A.38: Určenie ranku a poradia dát

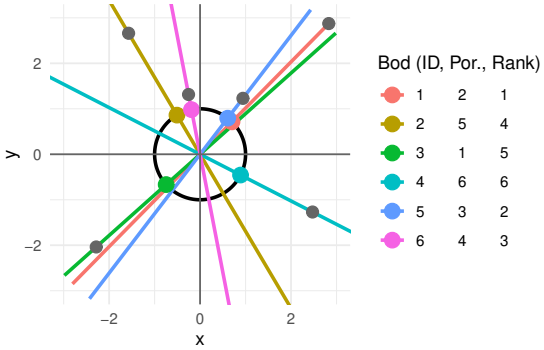


Obr. A.39: Zoradenie dát

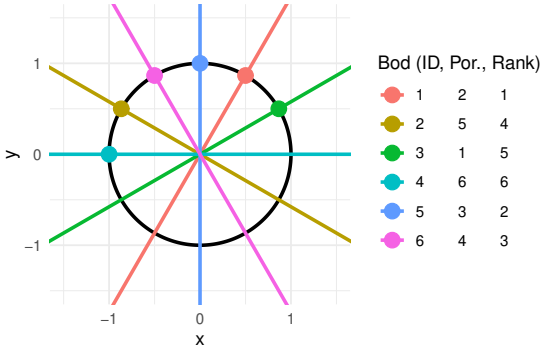


Obr. A.40: Štandardizovaný n-uholník a jeho váhované ťažisko. S_γ^2=0.31

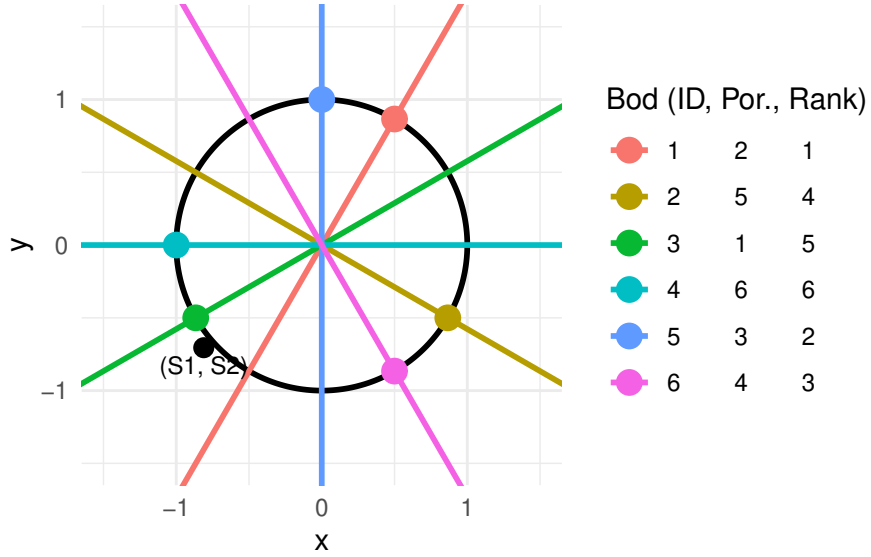
A.3.4 S_α^3



Obr. A.41: Určenie ranku a poradia dát

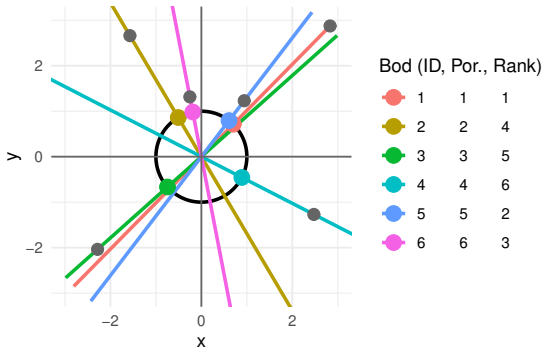


Obr. A.42: Zoradenie dát

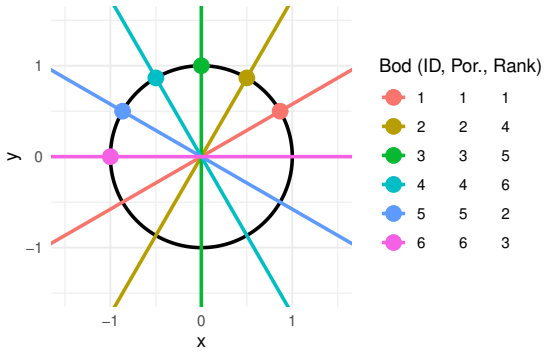


Obr. A.43: Štandardizovaný n-uholník a jeho váhované ťažisko. S_α^3=1.16

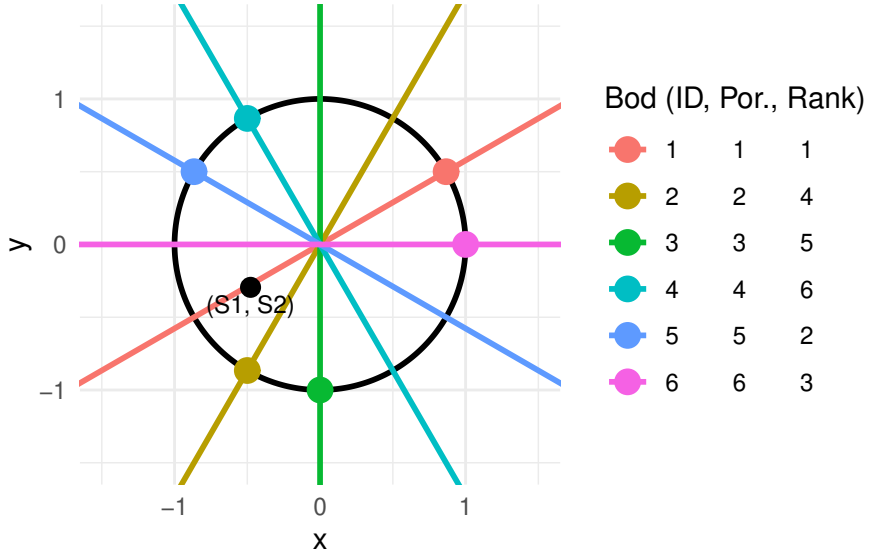
A.3.5 S_β^3



Obr. A.44: Určenie ranku a poradia dát

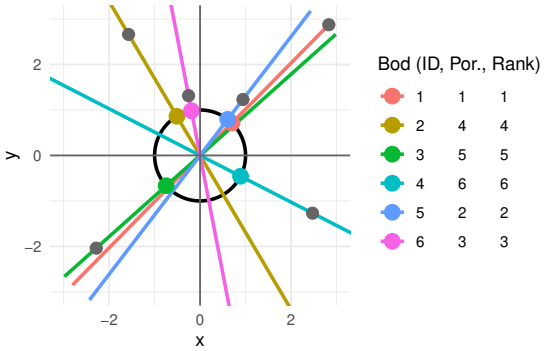


Obr. A.45: Zoradenie dát

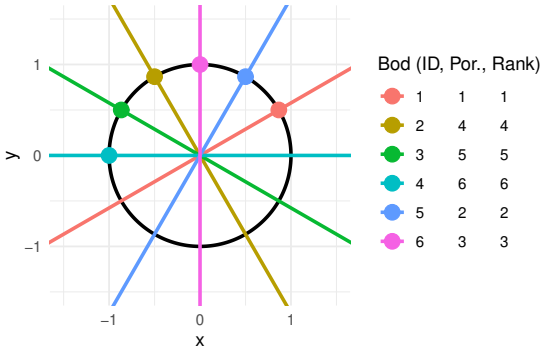


Obr. A.46: Štandardizovaný n-uholník a jeho váhované ťažisko. S_β^3=0.35

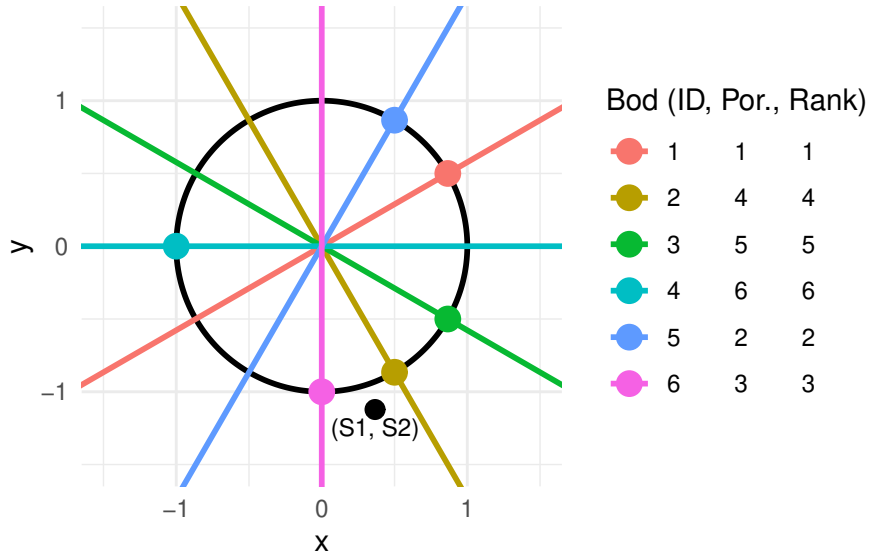
A.3.6 S_γ^3



Obr. A.47: Určenie ranku a poradia dát



Obr. A.48: Zoradenie dát



Obr. A.49: Štandardizovaný n-uholník a jeho váhované ťažisko. S_β^3=1.40