UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

VIACROZMERNÉ NEPARAMETRICKÉ TESTY POLOHY

DIPLOMOVÁ PRÁCA

2023 VILIAM ŽIGO

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

VIACROZMERNÉ NEPARAMETRICKÉ TESTY POLOHY

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 2023 **VILIAM ŽIGO**

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Viliam Žigo
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: Viacrozmerné neparametrické testy polohy
Multivariate nonparametric location tests

Anotácia: V prípade viacrozmerných dát vidno už niekoľko desaťročí snahy budovať testy o ich polohe, ktoré fungujú bez ohľadu na rozdelenie, z ktorého dáta pochádzajú (t. j. neparametrickosť). Tieto testy bývajú zvyčajne oveľa odolnejšie voči odľahlým dátam (t. j. robustnejšie) než klasické testy pre tento problém založené na predpoklade normality dát.

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

Dátum zadania: 11.01.2022

Dátum schválenia: 14.01.2022

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

.....
študent

.....
vedúci práce

Abstrakt

Táto práca sa zaoberá oblasťou viacrozmerných neparametrických testov polohy. Primárne sa zameriava na testy polohy viac ako dvoch súborov. Práca ponúka prehľad jednorozmerných testov spolu so základnými princípmi, z ktorých v ďalších kapitolách pri návrhu testov vychádzame. Jej hlavným cieľom je preskúmať spomenutú oblasť a pokúsiť sa navrhnúť spoľahlivé a silné zovšeobecnenie jednorozmerného Kruskalovho-Wallisovho testu pre viacrozmerné pozorovania. Práca poskytuje niekoľko našich návrhov zovšeobecnení spolu s analýzou ich chybovosti prvého druhu a kritikou ich slabých stránok. V práci taktiež uvádzame simulačnú štúdiu, ktorá sa venuje simuláciám sily navrhovaných testov pre rôzne scenáre a porovnaniu s existujúcimi testami. Táto štúdia naznačuje, že niektoré z našich návrhov sú z hľadiska sily v uvedených prípadoch aspoň tak dobré ako zaužívaná metóda od uznávaných štatistikov v oblasti neparametrických testov H. Oju a R. H. Randlesa.

Kľúčové slová: Viacrozmerné testy polohy, neparametrické testy, viacsúborové testy, zovšeobecnenia jednorozmerných testov, Kruskalov-Wallisov test.

Abstract

This work investigates the area of multivariate non-parametric location tests. It primarily focuses on location tests of more than two samples. The work offers an overview of one-dimensional tests together with the basic principles on which we base the design of the tests in the following chapters. Its main goal is to investigate the mentioned area and to try to propose a reliable and robust generalization of the univariate Kruskal-Wallis test for multivariate observations. We propose generalizations along with an analysis of their type I error rate and a critique of their weaknesses. In the paper, we also present a simulation study of power of the proposed tests under various scenarios and a comparison with existing tests. This study suggests that in terms of power some of our proposals are at least as good as the established method by respected nonparametricians H. Oja and R. H. Randles.

Keywords: Multivariate location tests, non-parametric tests, multi-sample tests, generalizations of univariate tests, Kruskal-Wallis test.

Predhovor

Analýza dát je v dnešnej dobe neoddeliteľnou súčasťou takmer každého odvetvia. Pre kvalitnú analýzu však potrebujeme spoľahlivé štatistické nástroje, prostredníctvom ktorých vieme dospieť k zmysluplným a robustným výsledkom. Mojim osobným cieľom v tejto práci bolo rozšíriť si vedomosti v oblasti viacrozmerných testov polohy, preskúmať existujúce a potenciálne nástroje a čo najviac sa z hľadiska kvality priblížiť svojimi nápadmi k zaužívaným testom, čo sa mi, verím, podarilo. Teší ma, že mi táto práca dala príležitosť vyskúšať si prácu takých štatistikov, akými sú napríklad H. B. Mann, D. R. Whitney, F. Wilcoxon, W. H. Kruskal, W. A. Wallis a mnohí ďalší.

Touto cestou sa chcem srdečne poďakovať vedúcemu diplomovej práce, Mgr. Jánovi Somorčíkovi, PhD. za ochotu, čas, konzultácie, odborné rady a podnetné pripomienky, ktoré mi pomohli pri písaní tejto práce. Ďakujem aj svojej rodine a priateľke za podporu počas celého štúdia.

Obsah

Úvod

Analýza a správna interpretácia dát je nevyhnutnou súčasťou napredovania empirického poznania. Súčasným trendom v mnohých odvetviach je zbierať dáta, ktoré je následne potrebné analyzovať a vyvodiť z nich racionálne závery. Dalo by sa povedať, že jedným z nástrojov, ktoré spájajú filozofie empirizmu a racionalizmu sú štatistické testy, pretože vychádzajúc z logiky a rácia nastavujú empirickým hypotézam hranicu medzi platnosťou a neplatnosťou. Štatistické testy sa využívajú napríklad v oblasti matematického modelovania, medicíny, farmakológie a v ďalších kľúčových odvetviach. Z tohto dôvodu je táto oblasť veľmi populárna, a keďže v mnohých situáciách zatiaľ neboli objavené tie najlepšie a najsilnejšie testy, tak ide stále o veľmi živú oblasť. Svedčia o tom aj viaceré publikácie z nedávnej minulosti využívajúce rôzne metódy ako napríklad priestorové znamienka a usporiadania [17] (2004), redukciu dimenzie [27] (2004), tzv. ε-štatistiku [22] (2004), Bayesovskú štatistiku [3] (2022) a tak ďalej.

V práci sa špecializujeme na podoblasť neparametrických testov polohy viacerých súborov. Kapitola 1 obsahuje teoretické východiská a označenia s predpokladom znalosti elementárnych poznatkov z oblasti štatistiky. V kapitole 2 sa venujeme skúmaniu možností rozšírenia Kruskalovho-Wallisovho testu, jednorozmerného neparametrického testu polohy pre viac ako dva súbory, pre použitie vo viacrozmernom priestore. Najskôr predstavíme konkrétnu problematiku, s ktorou sa v práci stretávame a v podkapitole 2.2 postupne prezentujeme naše návrhy rozšírení, ktoré zdieľajú charakteristiku redukcie dimenzie pozorovaní do jednorozmerného priestoru a následnú aplikáciu zaužívaných jednorozmerných testov o parametroch polohy. Rozšírenia v tejto podkapitole sme ďalej rozdelili do troch podkapitol podľa princípu, z ktorého vychádzajú. Ako neskôr uvádzame, rozšírenia v podkapitole 2.2.1 vychádzajú z výskumu v diplomovej práci P. Somogyiho [24], ktorý sa venoval neparametrickým testom polohy pre dva súbory. Podkapitoly 2.2.2 a 2.2.3 aplikujú v navrhnutých testoch matematické nástroje, ktorých využitie v štatistických testoch nie je bežné. V podkapitole 2.3 sa venujeme ešte iných možnostiam rozšírenia Kruskalovej-Wallisovej testovej štatistiky pre viacrozmerné pozorovania. Pokiaľ nie je uvedené inak, tak najmä z dôvodu prehľadnejšej vizualizácie demonštrujeme prístupy na dvojrozmerných dátach pochádzajúcich z troch nezávislých súborov. Dáta boli vygenerované v soft-

véri R [20] pre účely demonštrácie kombináciou vlastných zdrojových kódov uvedených v prílohe A a knižnice *mvtnorm* [8]. Testy však sformulujeme všeobecne pre ľubovoľnú dimenziu dát a počet súborov väčší ako jeden. Kapitola 3 obsahuje simulačnú štúdiu navrhnutých testov, ktorá ich porovnáva s niektorými existujúcimi alternatívami.

1 Teoretické východiská

V tejto kapitole prezentujeme základné označenia, princípy a teoretické východiská, ktoré sme využívali pri diplomovom výskume. Predstavíme najznámejšie jednorozmerné štatistické testy polohy, na ktoré v ďalších kapitolách nadviažeme, uvedieme definície pravdepodobnostných rozdelení, ktoré naprieč výskumom využívame a taktiež ďalšie využité poznatky.

1.1 Rozdelenia pravdepodobnosti dát

V nasledujúcich definíciách predstavujeme viacrozmerné spojité rozdelenia pravdepodobnosti dát, ktoré sme pri diplomovom výskume využili. Definície 1, 4 a 5 sme čerpali z prednášok [7], definíciu 2 z práce [24] a definíciu 3 z prednášok [9].

Definícia 1. *Nech $\mu \in \Re^p$ a Σ je regulárna $p \times p$ matica, pričom $p \geq 1$. Hovoríme, že náhodný vektor $X = (X_1, X_2, \ldots, X_p)^T$ má p-rozmerné normálne rozdelenie s vektorom strednej hodnoty μ a kovariančnou maticoou Σ, označujeme ako $X \sim \mathcal{N}_p(\mu, \Sigma)$, ak sa jeho hustota dá zapísať v tvare*

$$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{|\Sigma|}} \times \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}.$$

Definícia 2. *Nech $X \sim \mathcal{N}_p(0_p, \Sigma)$, $X = (X_1, X_2, \ldots, X_p)^T$, má p-rozmerné normálne rozdelenie s nulovou strednou hodnotou a kovariančnou maticou Σ. Nech $Y \sim \chi^2_n$ a nech X, Y sú nezávislé. Potom hovoríme, že*

$$Z = \frac{1}{\sqrt{\frac{Y}{n}}} \times X + \mu$$

má p-rozmerné Studentovo t-rozdelenie s n stupňami voľnosti a parametrami μ a Σ, označujeme ako $Z \sim t_n(\mu, \Sigma)$.

Definícia 3. *Nech $U, X_1, X_2, \ldots, X_p$ sú nezávislé náhodné premenné, pričom $U \sim \mathcal{R}(0,1)$, $X = (X_1, X_2, \ldots, X_p)^T$ a $X \sim \mathcal{N}_p(0, I_p)$. Potom*

$$Y = \frac{U^{\frac{1}{p}}}{\|X\|} \times X$$

má rovnomerné rozdelenie v p-rozmernej jednotkovej guli so stredom v bode $0 \in \Re^p$, označujeme ako $Y \sim \mathcal{B}_p$.

Definícia 4. *Nech $X_1, X_2, \ldots, X_n$ je náhodný výber z rozdelenia $\mathcal{N}_p(0, \Sigma)$. Nech $X^T = (X_1, X_2, \ldots, X_n)$ je matica rozmeru $p \times n$. Rozdelenia matice $M = X^T X$ nazývame Wishartovým rozdelením s n stupňmi voľnosti a používame označenie $M \sim \mathcal{W}_p(\Sigma, n)$.*

Definícia 5. *Nech $A \sim \mathcal{W}_p(I_p, m)$ a $B \sim \mathcal{W}_p(I_p, n)$ sú navzájom nezávislé, $m \geq p$ a I_p značí p-rozmernú identickú maticu. Potom hovoríme, že náhodná premenná*

$$L = \frac{\det(A)}{\det(A + B)}$$

má Wilksovo Lambda-rozdelenie s parametrami p, m a n, označujeme ako $L \sim \Lambda(p, m, n)$.

Pravdepodobnostné rozdelenia z definícií 1, 2 a 3 patria do triedy tzv. elipticky symetrických rozdelení, čo je v nasledujúcich kapitolách dôležité. Definíciu 6 elipticky symetrického rozdelenia sme čerpali z publikácie [3].

Definícia 6. *Pravdepodobnostné rozdelenie v p-rozmernom priestore s maticou rozptylu Σ sa nazýva elipticky symetrické so stredom symetrie θ, ak jeho hustota $f(\cdot)$ má tvar*

$$f(x - \theta) = \det(\Sigma)^{-1/2} g((x - \theta)^T \Sigma^{-1} (x - \theta)),$$

kde jednorozmerná nezáporná funkcia $g(\cdot)$ je taká, že

$$\int_{\Re^p} g(u^T u)\, du = 1.$$

1.2 Štatistické testy polohy

Test polohy (angl. *location test*) je štatistický test o parametri polohy (napr. o strednej hodnote) štatistického súboru. Takýchto testov existuje množstvo a podľa možnosti ich využitia sú zatriedené do rôznych kategórií. Jeden z možných spôsobov kategorizácie štatistických testov je rozdeliť ich podľa toho, z akého rozdelenia môžu pochádzať dáta analyzované daným testom. *Parametrické testy* vyžadujú od používateľa poznať pravdepodobnostné rozdelenie dát. Nevýhodou parametrických testov je práve táto naviazanosť na rozdelenie, pretože nie vždy ho poznáme alebo ho vieme spoľahlivo určiť. Z tohto dôvodu boli v minulosti vyvinuté aj *neparametrické varianty testov*. V tabuľke 1 uvádzame niekoľko známych jednorozmerných neparametrických testov polohy a ich využitie.

Tabuľka 1: Príklady jednorozmerných neparametrických testov polohy

Názov testu	**Využitie**
Wilcoxonov test	Porovnanie parametrov polohy **dvoch nezávislých** alebo aj **závislých** dátových súborov
Mannov-Whitneyho U-test	Porovnanie parametrov polohy **dvoch nezávislých** dátových súborov
Friedmanov test	Porovnanie parametrov polohy **dvoch a viac závislých** dátových súborov
Kruskalov-Wallisov test	Porovnanie parametrov polohy **dvoch a viac nezávislých** dátových súborov

Zdroj: [1]

Majme I nezávislých súborov p-rozmerných nezávislých rovnako rozdelených pozorovaní, pričom $I \geq 2$ a $p \geq 2$. Pre každé $i = 1, 2, \ldots, I$ uvažujme nasledovné nezávislé p-rozmerné náhodné vektory X_{ij} pre $j = 1, 2, \ldots, n_i$ pochádzajúce z pravdepodobnostného rozdelenia s distribučnou funkciou $F_i(\cdot)$:

$$X_{11}, X_{12}, \ldots, X_{1n_1} \sim F_1(\cdot),$$

$$X_{21}, X_{22}, \ldots, X_{2n_2} \sim F_2(\cdot),$$

$$\vdots$$

$$X_{I1}, X_{I2}, \ldots, X_{In_I} \sim F_I(\cdot),$$

kde n_i je počet náhodných vektorov v i-tom súbore. Potom môžeme testovať napríklad platnosť rovnosti stredných hodnôt naprieč týmito súbormi. Nulová hypotéza by v takom prípade mala tvar

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I \, (= \mu),$$

kde $\mu_i \in \Re^p$ je stredná hodnota rozdelenia, z ktorého pochádzajú pozorovania v i-tom súbore. Všimnime si, že obsahuje $p \cdot (I - 1)$ jednorozmerných rovností, z ktorých nesmieme ani jednu zamietnuť, aby sme nezamietli nulovú hypotézu. V štatistike sa taktiež využíva silnejšia formulácia takejto hypotézy prostredníctvom distribučných funkcií, keď testujeme

rovnosť nie len stredných hodnôt, ale celkovo pravdepodobnostných rozdelení. Nulovú hypotézu pre rovnosť pravdepodobnostných rozdelení potom vieme zapísať v tvare:

$$H_0' : F_1(\cdot) = F_2(\cdot) = \cdots = F_I(\cdot)\ (= F(\cdot)),$$

pričom pre zamietnutie nulovej hypotézy stačí zamietnuť jednu rovnosť. Ak by sme chceli testovať jednotlivé rovnosti z uvedenej hypotézy oddelene na hladine významnosti α, ale výsledky vyhodnotiť ako celok, tak by sme narazili na problém − simultánna pravdepodobnosť chyby prvého druhu (angl. *family-wise error rate*) by v takom prípade bola väčšia ako stanovená hladina významnosti α. Existujú však metódy, ktoré zabezpečia, aby simultánna pravdepodobnosť chyby prvého druhu neprekročila stanovenú hladinu významnosti α. V našej práci využívame nasledovné tvrdenia, ktoré sme čerpali z publikácie [10].

Tvrdenie 1. *Uvažujme testy m nulových hypotéz $H_1, H_2, \ldots, H_m$ a k nim zodpovedajúce p-hodnoty $p_1, p_2, \ldots, p_m$. Nech α označuje hladinu významnosti. Potom tzv. Bonferroniho korekcia zamieta hypotézu H_i, ak platí*

$$p_i \leq \frac{\alpha}{m},$$

čo zabezpečí, že simultánna pravdepodobnosť chyby prvého druhu neprekročí hladinu významnosti α.

Tvrdenie 2. *Uvažujme testy m nulových hypotéz $H_1, H_2, \ldots, H_m$ a k nim zodpovedajúce p-hodnoty $p_1, p_2, \ldots, p_m$. Nech α označuje hladinu významnosti. Nech $p^{(1)} \leq p^{(2)} \leq \cdots \leq p^{(m)}$ sú zoradené p-hodnoty podľa veľkosti a $H^{(i)}$ k nim prislúchajúce hypotézy. Potom tzv. Holmova-Bonferroniho sekvenčná metóda zamieta hypotézu $H^{(i)}$, ak platí*

$$p^{(i)} \leq \frac{\alpha}{m - i + 1},$$

čo zabezpečí, že simultánna pravdepodobnosť chyby prvého druhu neprekročí hladinu významnosti α.

1.2.1 Mannov-Whitneyho U-test

Ako sme už uviedli v tabuľke 1, Mannov-Whitneyho U-test je neparametrický test, ktorý nachádza využitie pri testovaní hypotéz o parametroch polohy dvoch nezávislých

jednorozmerných dátových súborov. Tento test navrhli páni H. B. Mann a D. R. Whitney v publikácii [15] ako rozšírenie testu pána F. Wilcoxona, ktorý vo svojej publikácii [28] uvažoval iba prípad, keď sú obe sady vyvážené. Keďže Wilcoxonov test sa dá jednoduchou transformáciou previesť na Mannov-Whitneyho U-test, tak uvádzame algoritmus Wilcoxonovho testu, ktorý sme čerpali z publikácie [1]. Uvažujme nasledovné jednorozmerné nezávislé náhodné výbery:

$$X_1, X_2, \ldots, X_n \sim F_x(\cdot),$$

$$Y_1, Y_2, \ldots, Y_m \sim F_y(\cdot),$$

kde $F_x(\cdot)$ je distribučná funkcia pravdepodobnostného rozdelenia, z ktorého pochádzajú pozorovania X_i a $F_y(\cdot)$ analogicky pre Y_i, pričom $F_x(\cdot)$ a $F_y(\cdot)$ sa môžu líšiť jedine posunom. Všetkých $n + m$ pozorovaní, ktoré dokopy tvoria tzv. *združený výber*, zoradíme vzostupne podľa veľkosti a priradíme im tzv. „ranky“ (číselne vyjadrené pozície v nadobudnutom poradí). Toto poradie je unikátne s pravdepodobnosťou 1, ak pravdepodobnostné rozdelenia, z ktorých pochádzajú náhodné premenné sú spojité. Označme T_x súčet rankov hodnôt $X_1, X_2, \ldots, X_n$ v zoradenom združenom výbere a T_y analogicky pre hodnoty $Y_1, Y_2, \ldots, Y_m$. Je jasné, že pre T_x a T_y platí vzťah

$$T_x + T_y = \frac{(n+m)(n+m+1)}{2},$$

a teda je postačujúce poznať iba jednu hodnotu, bez ujmy na všeobecnosti $T \triangleq T_x$. Testová štatistika Mannov-Whitneyho U-testu, tzv. *U-štatistika*, je potom definovaná ako

$$U_{MW} = \frac{U - E[U]}{\sqrt{D[U]}},$$

kde $U = nm + \frac{n(n+1)}{2} - T$. Náhodnú premennú U je možné ekvivalentne definovať ako $U = \#\{X_i < Y_j\}$. V publikácii [1] autor uvádza aj s dôkazmi štatistické vlastnosti oboch testov ako napríklad, že U_{MW} má asymptoticky normálne rozdelenie.

1.2.2 Kruskalov-Wallisov test

Kruskalov-Wallisov test je zovšeobecnením Mannov-Whitneyho U-testu pre viac ako dva štatistické súbory. Test bol navrhnutý pánmi W. H. Kruskalom a W. A. Wallisom v publikácii [14]. Všeobecným cieľom tohto neparametrického testu je rozhodnúť,

či viaceré súbory pochádzajú z rovnakého rozdelenia alebo nie. Algoritmus testu sme opäť čerpali z publikácie [1]. Uvažujme nasledovné jednorozmerné nezávislé náhodné výbery $X_{i1}, X_{i2}, \ldots, X_{in_i} \sim F_i(\cdot)$, $i = 1, 2, \ldots, I$. Podobne ako v Mannovom-Whitneyho U-teste vytvoríme združený výber, zoradíme pozorovania podľa veľkosti, priradíme ranky a pre každý výber vypočítame súčet rankov T_i. Nech $N = \sum_{i=1}^{I} n_i$, kde n_i je počet pozorovaní v i-tom súbore. Testová štatistika, ako ju navrhli Kruskal a Wallis, má následne tvar

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{I} n_i (T_i - \overline{T})^2. \tag{1}$$

Dá sa dokázať, že H má za platnosti H_0 asymptoticky χ^2_{I-1} rozdelenie. Opäť základné vlastnosti aj s dôkazmi je možné nájsť v publikácii [1] a komplexnejší pohľad na Kruskalov-Wallisov test je možné nájsť v originálnej publikácii Kruskala a Wallisa [14].

1.2.3 Viacrozmerná analýza rozptylu

Viacrozmerná analýza rozptylu, z anglického *multivariate analysis of variance*, skrátene MANOVA, skúma rozdiely v parametroch polohy viacerých spojitých pravdepodobnostných rozdelení dané nezávislou skupinovou premennou, tzv. *kvalitatívnym faktorom*. Uvažujme I súborov p-rozmerných pozorovaní X_{ij} pochádzajúcich z rozdelenia $F_i(\cdot) = \mathcal{N}_p(\mu_i, \Sigma)$, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, I$. Skúmané pozorovania však musia spĺňať niekoľko predpokladov. Ako vidíme, náhodné výbery pochádzajú z viacrozmerných normálnych rozdelení s rovnakou kovariančnou maticou. Predpoklad normálneho rozdelenia výrazne obmedzuje využitie tohto parametrického prístupu, v ktorom testujeme hypotézu:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I \ (= \mu).$$

Pre MANOVU existuje viacero testových štatistík. Najznámejšie testové štatistiky sú založené na výberovej kovariančnej matici modelu $\Sigma_{model} = \sum_{i=1}^{I} n_i (\overline{X_i} - \overline{X})(\overline{X_i} - \overline{X})^T$ a výberovej kovariančnej matici rezíduí $\Sigma_{rez} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (X_{ij} - \overline{X_i})(X_{ij} - \overline{X_i})^T$, kde $\overline{X_i}$ je p-rozmerný aritmetický priemer pozorovaní v i-tom súbore, n_i je počet pozorovaní v i-tom súbore a $\overline{X}$ je p-rozmerný aritmetický priemer všetkých pozorovaní. Medzi takéto testové štatistiky patria napríklad:

1. Wilksovo Λ: $\Lambda_{Wilks} = \det((\Sigma_{rez} + \Sigma_{model})^{-1} \Sigma_{rez})$,

2. Hotellingovo-Lawleyho stopové kritérium: $U_{HL} = tr(\Sigma_{rez}^{-1}\Sigma_{model})$,

3. Pillaiovo stopové kritérium: $V_{Pillai} = tr((\Sigma_{rez} + \Sigma_{model})^{-1}\Sigma_{model})$.

Žiadna z uvedených štatistík nie je rovnomerne najlepšia, avšak pre počet dát blížiaci sa nekonečnu sú všetky ekvivalentné. Toto tvrdenie ako aj ďalšie poznatky ohľadom viacrozmernej analýzy rozptylu sme čerpali z prednášok [7]. V prípade testovej štatistiky Λ_{Wilks}, je dokázané, že sa za platnosti H_0 riadi Wilksovým Lambda-rozdelením s parametrami p, $\nu_{rez} := I(N-1)$ a $\nu_{model} := I-1$, pričom p je dimenzia pozorovaní a $N := \sum_{i=1}^{I} n_i$ je celkový počet pozorovaní (pozri definíciu 5). Hoci testové štatistiky U_{HL} a V_{Pillai} majú tiež odvodené svoje rozdelenie pravdepodobnosti, tak zvyčajne sa využívajú ich aproximácie založené na Fisherovom-Schnedeckorovom rozdelení. Ďalšie podrobnosti o aproximáciách a ich použiteľnosti obsahuje napríklad diplomová práca B. Svetlíkovej [25].

1.3 Princíp permutačných testov

V oblasti štatistických testov sa často stretávame so situáciami, keď nepoznáme platnosť niektorého z predpokladov o dátach. V prípade porušenia predpokladov často dochádza k deformácii rozdelenia testovej štatistiky za platnosti nulovej hypotézy a vyhodnocovanie testu je v takom prípade nespoľahlivé. K podobným problémom dochádza aj v prípade, že máme k dispozícii malý počet pozorovaní, čo pri testoch využívajúcich asymptotické rozdelenie testovej štatistiky opäť zapríčiní nespoľahlivosť. Tieto problémy sa zväčša dajú vyriešiť aplikovaním tzv. *resamplingových techník* na tieto podkladové testy. Medzi resamplingové techniky patrí aj permutovanie, čím vznikajú tzv. *permutačné testy*.

Nulovou hypotézou pri permutačných testoch je hypotéza o totožnosti pravdepodobnostných rozdelení všetkých skúmaných súborov, ktorú sme v podkapitole 1.2 označili ako H_0'. Za predpokladu platnosti H_0' sa následne akoby dôkazom sporom dospeje k vyhodnoteniu permutačného testu. Tento „dôkaz sporom“ funguje nasledovne. Majme I súborov, každý s n_i realizáciami X_{ij} $(j = 1, 2, \ldots, n_i)$ z rozdelenia s distribučnou funkciou $F_i(\cdot)$ pre $i = 1, 2, \ldots, I$, ktoré určujú hodnotu testovej štatistiky T. Ak predpokladáme platnosť H_0', tak všetky realizácie sú z toho istého rozdelenia s distribučnou funkciou $F(\cdot) \triangleq F_1(\cdot) = F_2(\cdot) = \cdots = F_I(\cdot)$. To znamená, že aj ľubovolná permutácia realizácií

naprieč súbormi naďalej pochádza z rovnakého pravdepodobnostného rozdelenia. Keďže testová štatistika je funkciou realizácií, tak každá z B permutácií pôvodných realizácií X_{ij} určuje hodnotu testovej štatistiky T_b^*. Získaných B testových štatistík opisuje rozdelenie testovej štatistiky T podkladového testu za platnosti nulovej hypotézy. To umožňuje porovnať pôvodnú hodnotu testovej štatistiky T s kvantilom nasimulovaného rozdelenia na hladine významnosti α. Ak sa ukáže, že za predpokladu platnosti H_0' je pôvodná hodnota testovej štatistiky podkladového testu príliš extrémna, tak možno s pravdepodobnosťou chyby prvého druhu na hladine významnosti α usúdiť, že H_0' zrejme neplatí.

Celkový počet permutácií $\kappa_I(n_1, n_2, \ldots, n_I)$ pre I súborov, každý s n_i pozorovaniami pre $i = 1, 2, \ldots, I$, je rovný

$$\kappa_I(n_1, n_2, \ldots, n_I) = \prod_{i=1}^{I} \binom{\sum_{k=i}^{I} n_k}{n_i}.$$

$\kappa_I(n_1, n_2, \ldots, n_I)$ je však vo všetkých argumentoch rýchlo rastúca funkcia, napríklad už $\kappa_3(5, 5, 5) = 756756$. Využiť pri aplikácii všetky permutácie by bolo časovo veľmi náročné. Z tohto dôvodu v práci na aproximáciu rozdelenia testovej štatistiky využívame Monte Carlo verziu permutačných testov s $B = 1000$ náhodnými permutáciami. V práci taktiež využívame slovné spojenie „permutačná verzia testu X“ alebo „permutačný test X“, ktorým myslíme, že sa jedná o Monte Carlo verziu permutačného testu, v ktorom sa využíva testová štatistika podkladového testu X.

2 Zovšeobecnenia jednorozmerných testov polohy viacerých súborov

Vrámci prvej kapitoly sme v časti 1.2 predstavili najznámejšie jednorozmerné neparametrické testy polohy. V praxi však nastávajú aj situácie, keď je potrebné analyzovať viacrozmerné dáta, na ktoré by sme chceli mať takéto nástroje. V tejto kapitole prezentujeme vybrané prístupy, akými sa podľa nás dajú zovšeobecniť jednorozmerné testy polohy. Konkrétne sa venujeme problematike zovšeobecnenia Kruskalovho-Wallisovho testu. Vo všetkých testoch uvažujme p-rozmerné pozorovania pochádzajúce z I nezávislých súborov, pričom každé pozorovanie $X_{ij} = (X_{ij}^{(1)}, X_{ij}^{(2)}, \ldots, X_{ij}^{(p)})^T$ pochádza z neznámeho elipticky symetrického rozdelenia s distribučnou funkciou $F_i(\cdot)$, pre $i = 1, 2, \ldots, I$, pre $j = 1, 2, \ldots, n_i$, kde n_i je počet pozorovaní v i-tom súbore a $N = \sum_{i=1}^{I} n_i$ je celkový počet pozorovaní. Vo všetkých testoch predpokladáme, že neznáme distribúcie $F_1(\cdot), F_2(\cdot), \ldots, F_I(\cdot)$ sú navzájom totožné až na možné posuny δ_i definované implicitne

$$F_i(\cdot) := F(\cdot - \delta_i),$$

kde $F(\cdot)$ predstavuje nejakú distribučnú funkciu.

2.1 Problematika priameho rozšírenia

Jedným z prístupov, akým môžeme viacrozmerné testy získať, je pokúsiť sa zovšeobecniť jednorozmerné testy. To však nie je vždy možné vykonať priamou zámenou jednorozmerných premenných za viacrozmerné. Na tento problém narážame aj pri pokuse zovšeobecniť testy uvedené v podkapitole 1.2. Problém pri priamom zovšeobecnení je napríklad v myšlienke zoradenia premenných v združenom výbere, pretože nie je jasné, ako zoradiť vektorové premenné. Jeden z neparametrických a robustných prístupov uvádzajú H. Oja a R. H. Randles v publikácii [17] z roku 2004. Tento momentálne populárny prístup využíva koncept priestorového znamienka (angl. *spatial sign*) a priestorového usporiadania (angl. *spatial rank*). Definície týchto pojmov ako aj samotný algoritmus testu je možné nájsť v [17]. Naše návrhy ako je možné takpovediac priamo zovšeobecniť jednorozmerný test uvádzame v podkapitole 2.3.

Iný prístup, akým sa dá vysporiadať s uvedeným problémom, je redukcia dimenzie, resp. transformácia dát do jednorozmerného priestoru, kde už vieme aplikovať jednorozmerné testy. V podkapitole 2.2 sa sústredíme práve na tento druhý prístup. Otázkou však zostáva, akým spôsobom vieme čo najlepšie zachytiť informácie ukryté vo viacrozmerných dátach tak, aby sme na základe ich jednorozmerných projekcií vedeli spoľahlivo vyhodnotiť testovanú hypotézu. P. Somogyi sa vo svojej diplomovej práci [24] zaoberal problematikou dvojsúborových testov, kde zovšeobecnil a modifikoval Senov-Mathurov test a dospel k záveru, že viacrozmerné údaje je vhodné transformovať ortogonálnou projekciou na priamku, ktorá prechádza odhadnutými stredmi súborov a následne využiť permutačnú verziu Mannovho-Whitneyho U-testu. K rovnakému postupu, avšak vychádzajúc z odlišnej myšlienky, dospel už pred P. Somogyim aj R. R. Wilcox vo svojom článku [27]. Ak by sme však tento postup chceli zovšeobecniť pre viacero súborov, tak narážame na prekážky. Napríklad, ak by sme mali tri nezávislé súbory, t. j. aj tri stredy, tak až na nepravdepodobný špeciálny prípad, keď ležia všetky na jednej priamke, nevieme cez všetky preložiť priamku. V nasledujúcej kapitole preto navrhujeme viaceré riešenia.

2.2 Navrhované rozšírenia využívajúce projekciu na priamku

V tejto podkapitole postupne predstavíme sadu navrhovaných testov, ktoré zdieľajú myšlienku projekcie do jednorozmerného priestoru a následnú aplikáciu jednorozmerných testov. Najskôr sa venujeme testom, ktoré využívajú projekciu pozorovaní na priamku určenú dvojkombináciou niektorých z I odhadov stredov súborov. V týchto testoch sa postupne prechádza všetkých $\binom{I}{2}$ priamok. Ďalej predstavíme testy využívajúce algoritmus lineárnej regresie a testy využívajúce výberové hlavné komponenty.

2.2.1 Projekcia na priamku určenú dvojkombináciou stredov súborov

Vychádzajúc z prác P. Somogyiho a R. R. Wilcoxa je prirodzeným rozšírením pre viac súborov využiť ich permutačný test pre dva súbory pre každú dvojkombináciu súborov a zistiť, či aspoň jeden z $\binom{I}{2}$ testov zamieta nulovú hypotézu o rovnosti parametrov polohy dvoch súborov. Takýto test by sme však nemohli vyhodnotiť priamočiarym zamietnutím nulovej hypotézy o rovnosti I súborov v prípade, že niektorá z $\binom{I}{2}$ p-hodnôt dvojsúboro-

vých testov je menšia ako stanovená hladina významnosti α. Totiž simultánna pravdepodobnosť chyby prvého druhu by v takom prípade bola väčšia ako hladina významnosti α. Z toho dôvodu sme pri tomto prístupe využili tzv. Bonferroniho korekciu (pozri tvrdenie 1 v podkapitole 1.2), ktorá zaručí, že hladina významnosti α nebude prekročená. Avšak Bonferroniho korekcia je s rastúcim počtom upravovaných p-hodnôt čoraz konzervatívnejšia pri zamietaní nulovej hypotézy. V snahe zmierniť konzervatívnosť sme sa namiesto všetkých $\binom{I}{2}$ dvojkombinácií súborov rozhodli preskúmať aj variantu, kde využijeme permutačný test pre dva súbory iba pre kombináciu jedného súboru s ostatnými, čo zredukuje počet kombinácií na $(I-1)$. Pre ďalšie odvolávanie sa na tieto testy, označme ich skratkami BK-VP (*test s Bonferroniho Korekciou - Všetky Páry*) a BK-JO (*test s Bonferroniho Korekciou - Jeden vs. Ostatné*). Výsledky simulácií pravdepodobnosti chyby prvého druhu uvádzame v tabuľke 2. Simulácie nenaznačili výrazné odchýlenie od zvolenej hladiny významnosti $\alpha = 5\%$, takže môžeme predpokladať, že navrhnuté testy sú vhodné na štatistické testovanie. Ich silu skúmame v ďalších kapitolách.

Tabuľka 2: Odhad pravdepodobnosti chyby prvého druhu testov s Bonferroniho korekciou na hladine významnosti $\alpha = 5\%$ prostredníctvom 10000 simulácií pre $I = 3$ nezávislé vyvážené súbory, každý s $n_i = 30$ dátami rozmeru $p = 2$ z príslušného rozdelenia.

	$\mathcal{N}_2(0_2, I_2)$	$\mathcal{B}_2$	$t_1(0_2, I_2)$	$t_3(0_2, I_2)$
BK-VP	0,043	0,045	0,044	0,043
BK-JO	0,047	0,041	0,050	0,046

Zdroj: vlastné spracovanie

Konzervatívnosť Bonferroniho korekcie nás viedla k hľadaniu ďalších prístupov. Nasledujúce dva prístupy hľadajú optimálnu priamku, na ktorú sa následne vykoná ortogonálna projekcia všetkých N dát zo všetkých I súborov. Myšlienka voľby jednej priamky spočíva v tom, že ak nulová hypotéza o rovnosti všetkých I posunov neplatí, tak to bude vidno aj z jednorozmerných dát vzniknutých projekciou na vhodnú priamku. Pre jednoduchosť nech je optimálna priamka v oboch prístupoch daná stredmi a-teho a b-teho súboru, označíme S_a, S_b, a je definovaná nasledovne

$$\hat{q} : S_b = S_a + k \times \overrightarrow{S_b S_a}, \text{ pre } k \in \Re.$$

Každá projekcia pozorovania X_η pre $\eta = 1, 2, \ldots, N$ je jednoznačne určená svojim parametrom k, ozn. k_η. Jednorozmerné hodnoty k_η sa následne vyhodnotia prostredníctvom Kruskalovho-Wallisovho testu opísaného v podkapitole 1.2.2. V čom sa však tieto prístupy líšia, je kritérium pre voľbu optimálnej priamky.

Pri prvom spôsobe sa pre každú dvojkombináciu (a, b) (kde $a, b = 1, 2, \ldots, I$ a $a < b$) vykoná dvojsúborový test porovnávajúci a-ty a b-ty súbor, ktorý opísali P. Somogyi [24] a R. R. Wilcox [27] vo svojich publikáciach. Za optimálnu priamku sa následne zoberie tá, ktorá mala v dvojsúborovom teste najmenšiu p-hodnotu, a teda má najväčší potenciál ju mať podobne extrémnu aj v Kruskalovom-Wallisovom teste. Teda nech p_{ab} je p-hodnota spomenutého testu pre dva súbory pre stredy $a, b = 1, 2, \ldots, I$, pričom $a < b$, potom optimálna priamka je daná dvojicou stredov (a^*, b^*), kde

$$(a^*, b^*) = \underset{\substack{a,b=1,2,\ldots,I \\ a<b}}{\arg\min}\, p_{ab}.$$

Pre ďalšie odvolávanie sa na tento test, označme ho skratkou MPH (*test s Minimálnou P-Hodnotou*).

Druhý spôsob volí optimálnu priamku z $\binom{I}{2}$ priamok tak, aby projektované dáta boli čo najviac rozptýlené v zmysle vzdialenosti na optimálnej priamke, pretože sa dá očakávať, že to zvyšuje šance na odhalenie neplatnosti nulovej hypotézy. Pri tomto spôsobe sa teda rieši maximalizačná úloha

$$(a^*, b^*) = \underset{\substack{a,b=1,2,\ldots,I \\ a<b}}{\arg\max} \left[\max_{i,j=1,2,\ldots,N} \| l_i^{(a,b)} - l_j^{(a,b)} \| \right], \tag{2}$$

kde $\| \cdot \|$ značí euklidovskú normu a $l_\eta^{(a,b)}$ je ortogonálna projekcia bodu X_η zo združeného súboru na priamku určenú kombináciou stredov (a, b). Pre ďalšie odvolávanie sa na tento test, označme ho skratkou MR (*test s Maximálnym Rozptylom*).

Pri vyhodnocovaní primeranosti prístupov MPH a MR však dochádza k porušeniu predpokladu Kruskalovho-Wallisovho testu definovaného v podkapitole 1.2.2, pretože dáta po projekcii na priamku, ktorá je nimi daná, už nespĺňajú predpoklad nezávislosti. To spôsobuje, že pri vyhodnocovaní výsledkov testov na hladine významnosti α pravdepodobnosť chyby prvého druhu prekračuje túto hladinu významnosti, čo nie je prekvapujúce, keďže χ^2-aproximácia testovej štatistiky nie je v prípade závislých pozorovaní teóriou za-

ručená. Núkajúcou sa praktikou je využiť Monte Carlo permutačnú verziu Kruskalovho-Wallisovho testu. Pri vyššie uvedených dvoch prístupoch je to nutné, čo sa ukázalo prostredníctvom simulácií, ktorých výsledky uvádzame v tabuľke 3. Pre permutačné verzie oboch našich testov sme tiež simulovali odhad pravdepodobnosti chyby prvého druhu, aby sme overili ich kvalitu. Pre oba testy sme v permutačnej verzii využili 1000 permutácií. Zatiaľ čo v prístupe MR sa ukázal byť tento počet dostatočný, tak pri prístupe MPH sa ukázal byť nedostatočný, pretože odhad pravdepodobnosti chyby prvého druhu bol napríklad pre normálne rozdelenie uvedené v tabuľke 3 na úrovni až 0,068, čo je naďalej výrazne nad zvolenou hladinou významnosti $\alpha = 5\%$. Tento problém pravdepodobne nastáva z dôvodu, že 1000 permutácií nedostatočne presne odhadlo skutočné rozdelenie testovej štatistiky, čo je možné vyriešiť zvýšením počtu permutácií. Avšak už 1000 permutácií a pri 10000 simuláciách je pri našich výpočtových možnostiach veľmi časovo náročné a iba zdvojnásobenie počtu permutácií by tento čas rovnako zdvojnásobilo, pričom nemáme istotu, že by dvojnásobný počet už bol dostatočný. Zníženie počtu simulácií z dôvodu skrátenia doby výpočtu by odhad pre pravdepodobnosť chyby prvého druhu spravilo menej dôveryhodný, a keďže v nasledujúcich kapitolách plánujeme porovnávať navrhnuté prístupy aj z hľadiska sily pre rôzne scenáre, tak by to taktiež zapríčinilo inkonzistenciu pri porovnávaní. Z týchto dôvodov by bolo zahrnutie prístupu MPH za hranicou našich výpočtových a časových možností, a preto sme prístup MPH ďalej v práci neanalyzovali.

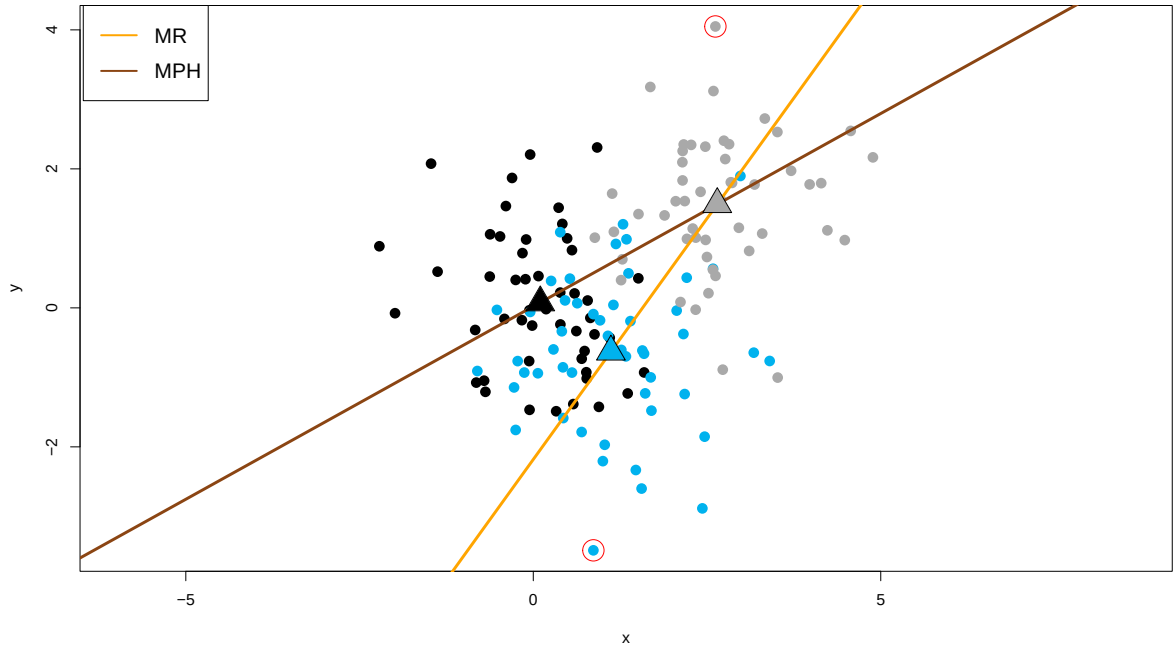
Tabuľka 3: Odhad pravdepodobnosti chyby prvého druhu na hladine významnosti $\alpha = 5\%$ testov MPH, MR a ich permutačných verzií prostredníctvom 10000 simulácií pre $I = 3$ nezávislé vyvážené súbory, každý s $n_i = 30$ dátami rozmeru $p = 2$ z príslušného rozdelenia. V každom permutačnom teste bolo vykonaných 1000 permutácií.

	$\mathcal{N}_2(0_2, I_2)$	$\mathcal{B}_2$	$t_1(0_2, I_2)$	$t_3(0_2, I_2)$
MPH	0,143	0,151	0,094	0,131
MR	0,097	0,105	0,054	0,082
Permutačný MPH	0,068	0,068	0,067	0,069
Permutačný MR	0,048	0,050	0,049	0,050

Zdroj: vlastné spracovanie

Tiež by sme chceli upozorniť na to, že prístupy MPH a MR nevolia vždy iden-

tickú projekčnú priamku za optimálnu. Na obrázku 1 uvádzame príklad, keď prístupy MPH a MR zvolili navzájom odlišné projekčné priamky. Prístup MPH zvolil ako optimálnu priamku spájajúcu stredy čierneho a sivého súboru (hnedá priamka) a prístup MR priamku spájajúcu stredy čierneho a modrého súboru (oranžová priamka). Táto odlišnosť bola zapríčinená pozorovaniami označenými červeným kruhom, ktoré by sme mohli nazvať odľahlými a práve tieto pozorovania boli od seba po projekcii najvzdialenejšie.



Obr. 1: Príklad rozdielnych optimálnych projekčných priamok prístupov MPH a MR. Trojuholníky označujú stredy súborov. Všetky tri súbory (čierny, modrý a sivý) obsahujú 50 pozorovaní, ktoré pochádzajú z dvojrozmerných normálnych rozdelení s identickou kovariančnou maticou a so strednými hodnotami $\mu_1 = (0; 0)^T$, $\mu_2 = (1; -0,6)^T$ a $\mu_3 = (2,8; 1,4)^T$. (*Zdroj: vlastné spracovanie*)

Nevýhodou spomenutých prístupov je kvadratický rast výpočtovej náročnosti s rastúcim počtom súborov I, keďže sa v týchto testoch musí analyzovať všetkých $\binom{I}{2}$ priamok. To nás viedlo k alternatívnym prístupom predstaveným v nasledujúcich podkapitolách.

2.2.2 Projekcia na priamku určenú lineárnou regresiou

Uvažujme p-rozmerné pozorovania X_η zo združeného súboru a lineárny regresný model

$$Z = Y\beta + \varepsilon, \tag{3}$$

kde $\beta \in \Re^p$ je neznámy parameter, ε je vektor chýb a Z, Y vysvetlíme neskôr. Myšlienka testu spočíva vo využití $(p-1)$-rozmernej regresnej nadroviny odhadnutej z pozorovaní metódou najmenších štvorcov, t. j. $\hat{\beta} = (Y^T Y)^{-1} Y^T Z$. V odhadnutej nadrovine následne zostrojíme priamku s predpisom

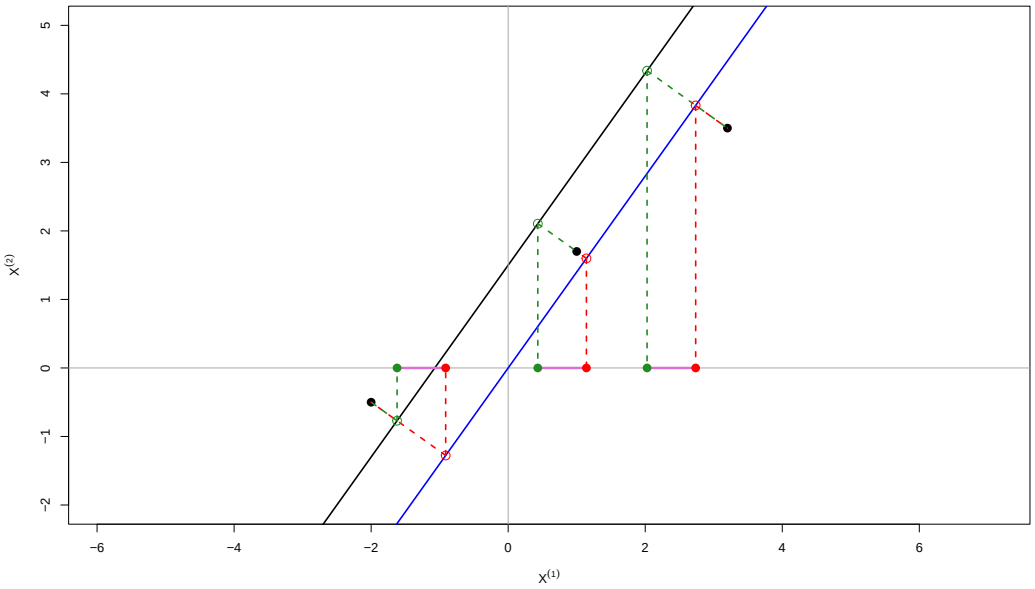
$$\hat{q} : \begin{bmatrix} \hat{X}^{(1)} \\ \vdots \\ \hat{X}^{(p-1)} \\ \hat{X}^{(p)} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \hat{\beta}_0 \end{bmatrix} + k \times \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \sum_{i=1}^{p-1} \hat{\beta}_i \end{bmatrix}, \text{ pre } k \in \Re,$$

na ktorú následne ortogonálne projektujeme pozorovania X_η. Tieto projekcie na odhadnutej priamke $\hat{q}$ sa naďalej nachádzajú v p-rozmernom priestore, no my potrebujeme ich jednorozmernú reprezentáciu. Jedným z riešení, ktoré využívame pri testoch BK-VP, BK-JO a MR, je využitie faktu, že pre $\forall \eta = 1, 2, \ldots, N$ je projekcia X_η jednoznačne určená hodnotou parametra k vystupujúcim v predpise priamky $\hat{q}$. Pre redukciu dimenzie sme sa však v predstavených lineárnych modeloch rozhodli využiť iný prístup. Nech $P_{X_\eta} = (P_{X_\eta}^{(1)}, P_{X_\eta}^{(2)}, \ldots, P_{X_\eta}^{(p)})^T$ je ortogonálna projekcia pozorovania X_η na priamku $\hat{q}$. Ako naše jednorozmerné pozorovanie navrhujeme využiť prvú zložku vektora P_{X_η}, čo nám dokopy pre $\forall \eta$ dá vektor jednorozmerných hodnôt $P_X^{(1)} = (P_{X_1}^{(1)}, P_{X_2}^{(1)}, \ldots, P_{X_N}^{(1)})^T$, ktoré sú jednoznačne určené pôvodnými pozorovaniami. Na tieto jednorozmerné projekcie následne aplikujeme Kruskalov-Wallisov test.

Všimnime si, že vzájomná poloha takto získaných jednorozmerných pozorovaní je nezávislá od $(0, \ldots, 0, \hat{\beta}_0)^T$, respektíve od interceptu lineárneho modelu $\hat{\beta}_0$. Pre lepšiu predstavu na obrázku 2 uvádzame príklad pre $p = 2$. Čierna priamka predstavuje odhadnutú priamku z regresného modelu a čierne body predstavujú vybrané pozorovania pre túto vizualizáciu. Modrá priamka je rovnobežná s čiernou a s interceptom $\hat{\beta}_0 = 0$. Prázdne zelené body vznikli projekciou na odhadnutú priamku a plné zelené predstavujú ich prvé súradnice vykreslené v zredukovanom priestore (os $X^{(1)}$). Analogicky to platí

pre červené body a modrú priamku. Ako je možné pozorovať na obrázku 2, tak zelené a červené body sa po redukcii dimenzie líšia len o konštantný posun Δ, znázornený ružovou farbou, ktorý neovplyvní test o rovnosti parametrov polohy I súborov. Dá sa ukázať, že pre $p = 2$ platí:

$$\Delta = \frac{|\hat{\beta}_0 \hat{\beta}_1|}{1 + \hat{\beta}_1^2}.$$



Obr. 2: Vizualizácia zanedbateľnosti interceptu $\hat{\beta}_0$ lineárneho modelu pre $p = 2$. *(Zdroj: vlastné spracovanie)*

Bez ujmy na všeobecnosti teda môžeme intercept zanedbať a pôvodné pozorovania projektovať na priamku danú predpisom

$$\hat{q} : \begin{bmatrix} \hat{X}^{(1)} \\ \vdots \\ \hat{X}^{(p-1)} \\ \hat{X}^{(p)} \end{bmatrix} = k \times \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \sum_{i=1}^{p-1} \hat{\beta}_i \end{bmatrix} \triangleq k \times u, \text{ pre } k \in \Re.$$

Nech $\mathcal{U} = u(u^T u)^{-1} u^T$. Potom každú projekciu P_{X_η} pre $\forall \eta$ je možné získať nasledovne:

$$P_{X_\eta} = \mathcal{U} X_\eta. \tag{4}$$

Poznamenávame, že matica $\mathcal{U}$ je špeciálny prípad projekčnej matice, tiež známej ako „*hat matrix*“[1], pre projekciu do jednorozmerného priestoru s bázovým vektorom u. Nech

[1]Podrobnejšie informácie o „*hat matrix*“ je možné nájsť napríklad v knihe [13] v kapitole 5.2.2.

$e_1 = (1, 0, \ldots, 0)^T \in \Re^p$. Je zrejmé, že

$$P_{X_\eta}^{(1)} = e_1^T \mathcal{U} X_\eta = e_1^T u (u^T u)^{-1} u^T X_\eta = \frac{1}{u^T u} u^T X_\eta = \frac{1}{u^T u} X_\eta^T u,$$

respektíve, že jednorozmernú reprezentáciu dát je možné vyjadriť maticovo ako

$$P_X^{(1)} = \frac{1}{u^T u} X u.$$

V tejto práci sa zaoberáme dvomi postupmi pre získanie odhadu lineárneho modelu. Pripomíname, že hlavným problémom pri viac-súborovom zovšeobecnení dvojsúborového testu bolo, že nevieme preložiť priamku prechádzajúcu viac ako dvomi bodmi, čo nás priviedlo k myšlienke nájsť regresnú priamku danú odhadmi stredov I súborov. Konkrétne premenné Z a Y v regresnom modeli (3) zvolíme nasledovne:

$$\begin{bmatrix} | & | \\ Y_{I\times p} & Z_{I\times 1} \\ | & | \end{bmatrix}_{I\times(p+1)} \equiv \begin{bmatrix} 1 & - & \overline{X}_1^T & - \\ 1 & - & \overline{X}_2^T & - \\ \vdots & & \vdots & \\ 1 & - & \overline{X}_I^T & - \end{bmatrix}_{I\times(p+1)},$$

kde $\overline{X}_i$ je p-rozmerný aritmetický priemer pozorovaní z i-teho súboru. Tento prístup označíme skratkou LMS (*test s Lineárnym Modelom daným Stredmi*). Z dôvodu jednoznačnosti odhadu $\hat{\beta}$ je podmienkou použiteľnosti prístupu LMS, že počet súborov musí byť väčší ako ich dimenzia, teda $I > p$.

Druhým návrhom je využiť v modeli pre odhad regresnej priamky všetkých N pozorovaní X_η zo združeného výberu ($\eta = 1, 2, \ldots, N$). To znamená, že Z a Y v regresnom modeli (3) v tomto prípade zvolíme nasledovne:

$$\begin{bmatrix} | & | \\ Y_{N\times p} & Z_{N\times 1} \\ | & | \end{bmatrix} \equiv \begin{bmatrix} 1 & - & X_1^T & - \\ 1 & - & X_2^T & - \\ \vdots & & \vdots & \\ 1 & - & X_N^T & - \end{bmatrix}_{N\times(p+1)}.$$

Tento druhý prístup ďalej v práci označujeme LMP (*test s Lineárnym Modelom daným Pozorovaniami*).

Nedostatkom testu LMS je, že, opäť z rovnakého dôvodu ako pre testy MPH a MR, nie je splnený predpoklad nezávislosti dát pre Kruskalov-Wallisov test. Z toho dôvodu je

nutné pre vyhodnotenie hypotézy pomocou tohto testu využiť Monte Carlo permutačnú verziu. Pri teste LMP sa pri simuláciách pravdepodobnosti chyby prvého druhu testu, ktorý využíva χ^2-aproximáciu v Kruskalovom-Wallisovom teste, nepreukázalo porušenie predpokladov, pretože odhad pravdepodobnosti chyby prvého druhu sa veľmi nevychýlil od zvolenej hladiny významnosti 5%. Výsledky simulácií uvádzame v tabuľke 4.

Tabuľka 4: Odhad pravdepodobnosti chyby prvého druhu na hladine významnosti $\alpha = 5\%$ prostredníctvom 10000 simulácií pre $I = 3$ nezávislé vyvážené súbory, každý s $n_i = 30$ dátami rozmeru $p = 2$ z príslušného rozdelenia. V permutačnom teste bolo vykonaných 1000 permutácií.

	$\mathcal{N}_2(0_2, I_2)$	$\mathcal{B}_2$	$t_1(0_2, I_2)$	$t_3(0_2, I_2)$
LMP	0,051	0,053	0,045	0,050
LMS	0,127	0,135	0,063	0,112
Permutačný LMS	0,051	0,049	0,049	0,050

Zdroj: vlastné spracovanie

Neočakávané nepreukázanie porušenia predpokladov Kruskalovho-Wallisovho testu v teste LMP nás viedlo k hlbšiemu skúmaniu tohto výsledku. Dospeli sme k záveru, že pri hľadaní regresnej nadroviny prostredníctvom všetkých pozorovaní je kolísavosť tejto nadroviny značne obmedzená a nie je na dátach tak závislá, aby sa to prenieslo aj do pravdepodobnosti chyby prvého druhu. Myslíme tým, že napríklad pri parametroch uvedených v popise tabuľky 4 a pozorovaniach z normálneho rozdelenia (prvý stĺpec tabuľky 4) sa sklon priamky (t. j. $\hat{\beta}_1$) v prístupe LMP pohyboval približne v intervale $[-0,601; 0,402]$, zatiaľ čo pri využití LMS to bolo $[-70,678; 356,778]$.

Nevýhodou navrhnutých prístupov LMS a LMP pre $p > 2$ je tiež výrazne subjektívny a v podstate neodôvodniteľný spôsob zostrojenia priamky $\hat{q}$, pretože sa dá jednoducho nájsť príklad, kde síce $(p-1)$-rozmerná regresná rovina dobre zachytí štruktúru dát, no priamka $\hat{q}$ neodhalí rozdiely medzi súbormi. Navrhujeme preto rekurentný algoritmus, ktorý si vie poradiť aj v situáciach, keď LMS a LMP zlyhajú. Myšlienka sa dá aplikovať v oboch prístupoch LMS a LMP, no my skúmame iba aplikáciu pre LMS, ktorú označíme RLMS (*test s Rekurentným Lineárnym Modelom daným Stredmi*). Uvažujme N p-rozmerných pozorovaní X_η z I súborov. Nech $X = (X_1, X_2, \ldots, X_N)^T$ a nech

v značení ${}_jX_\eta$ hodnota η značí poradie v združenom súbore, j označuje j-tu iteráciu algoritmu a ${}_jX_\eta^{(b)}$ značí b-tu zložku ${}_jX_\eta$. Uvažujme lineárny regresný model (3). Nech ${}_j\hat{\beta} = ({}_j\hat{\beta}_0, {}_j\hat{\beta}_1, \ldots, {}_j\hat{\beta}_{p-j-1})^T = ({}_jY^T\,{}_jY)^{-1}\,{}_jY^T\,{}_jZ$ je odhad metódou najmenších štvorcov v j-tej iterácii pri označeniach:

$$
{}_jY = \begin{bmatrix} 1 & | & \ldots & | \\ \vdots & {}_j\overline{X}^{(1)} & \ldots & {}_j\overline{X}^{(p-j-1)} \\ 1 & | & \ldots & | \end{bmatrix},
$$

$$
{}_jZ \equiv {}_j\overline{X}^{(p-j)},
$$

kde ${}_j\overline{X}^{(b)}$ označuje b-ty stĺpec v ${}_j\overline{X}$, čo je $I \times (p-j)$ rozmerná matica odhadov stredov I súborov pozorovaní ${}_jX$ v j-tej iterácii. Teda v uvedenom lineárnom modeli (3) je prvých $(p-j-1)$ zložiek matice ${}_j\overline{X}$ tzv. regresormi a $(p-j)$-ta zložka matice ${}_j\overline{X}$ modelovaná premenná. V navrhovanom rekurentnom algoritme v každej iterácii postupne redukujeme dimenziu dát o jeden takým spôsobom, že v $(p-j)$-rozmernom priestore prostredníctvom lineárneho modelu odhadneme $(p-j-1)$-rozmernú nadrovinu

$${}_jh' : y = {}_j\hat{\beta}_0 + {}_j\hat{\beta}_1 x_1 + \cdots + {}_j\hat{\beta}_{p-j-1} x_{p-j-1},$$

na ktorú následne ortogonálne projektujeme $(p-j)$-rozmerné pozorovania. Opäť je možné bez ujmy na všeobecnosti položiť intercept ${}_j\hat{\beta}_0$ nadroviny ${}_jh'$ rovný nule, pretože vzájomná poloha ortogonálnych projekcií sa ani v tomto prípade nemení v závislosti od interceptu. Teda nech

$${}_jh : y = {}_j\hat{\beta}_1 x_1 + \cdots + {}_j\hat{\beta}_{p-j-1} x_{p-j-1},$$

v ďalších značeniach značí nadrovinu pre ${}_j\hat{\beta}_0 = 0$. Vektory ${}_ju_1 := (1, 0, 0, \ldots, 0, 0, {}_j\hat{\beta}_1)^T$, ${}_ju_2 := (0, 1, 0, \ldots, 0, 0, {}_j\hat{\beta}_2)^T$ až ${}_ju_{p-j-1} := (0, 0, 0, \ldots, 0, 1, {}_j\hat{\beta}_{p-j-1})^T$ zjavne ležia v nadrovine ${}_jh$. Rekurentný vzťah teda definujeme pre $j = 0, 1, 2, \ldots, (p-2)$ a začiatočný stav ${}_0X \equiv X$ nasledovne:

$${}_{j+1}X = {}_jX\,{}_j\mathcal{U}\,{}_jV,$$

kde matica ${}_j\mathcal{U} := {}_jU({}_jU^T\,{}_jU)^{-1}\,{}_jU^T$ je „*hat matrix*“ (pozri poznámku pod čiarou naviazanú na vetu pod vzťahom (4)) a $(p-j) \times (p-j-1)$ rozmerné matice ${}_jU$ a ${}_jV$ sú

definované nasledovne:

$$ {}_jU = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 \\ {}_j\hat{\beta}_1 & {}_j\hat{\beta}_2 & \dots & {}_j\hat{\beta}_{p-j-1} \end{bmatrix}, \qquad {}_jV = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 \\ 0 & \dots & 0 & 0 \end{bmatrix}. $$

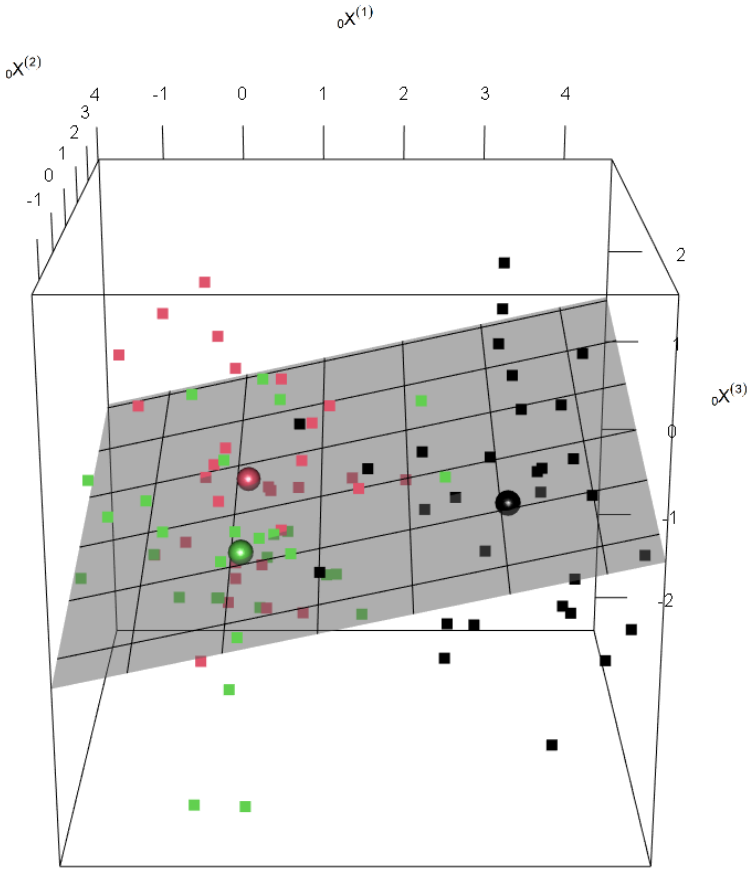
Využitím takéhoto rekurentného algoritmu získame po $(p-1)$ krokoch jednorozmernú projekciu viacrozmerných pozorovaní X, na ktorú môžeme rovnako ako v predchádzajúcich návrhoch testov aplikovať Kruskalov-Wallisov test opísaný v podkapitole 1.2.2. Pre overenie kvality RLMS sme opäť simulačne overili pravdepodobnosť chyby prvého druhu a výsledky, ktoré uvádzame v hornej časti tabuľky 5, sú totožné ako pri prístupe LMS, keďže pre $p = 2$ sú tieto testy totožné. Rozdiely sa ukázali až pri vyšších dimenziách, kde sa tieto prístupy správajú odlišne. Túto skutočnosť možno pozorovať aj v treťom a štvrtom riadku tabuľky 5. Poznamenávame, že rovnosť nasimulovaných pravdepodobností chyby prvého druhu v posledných dvoch stĺpcoch druhej časti tabuľky 5 je čisto náhodná a nie je to pravidlo.

Tabuľka 5: Odhad pravdepodobnosti chyby prvého druhu testu RLMS na hladine významnosti $\alpha = 5\%$ prostredníctvom 10000 simulácií pre I nezávislých vyvážených súborov, každý s n_i p-dimenzionálnymi realizáciami z príslušného rozdelenia.

	I	n_i	p	$\mathcal{N}_p(0_p, I_p)$	$\mathcal{B}_p$	$t_1(0_p, I_p)$	$t_3(0_p, I_p)$
RLMS	3	30	2	0,127	0,135	0,063	0,112
Permutačný RLMS	3	30	2	0,051	0,049	0,049	0,050
Permutačný LMS	4	30	3	0,048	0,048	0,050	0,051
Permutačný RLMS	4	30	3	0,049	0,047	0,050	0,051

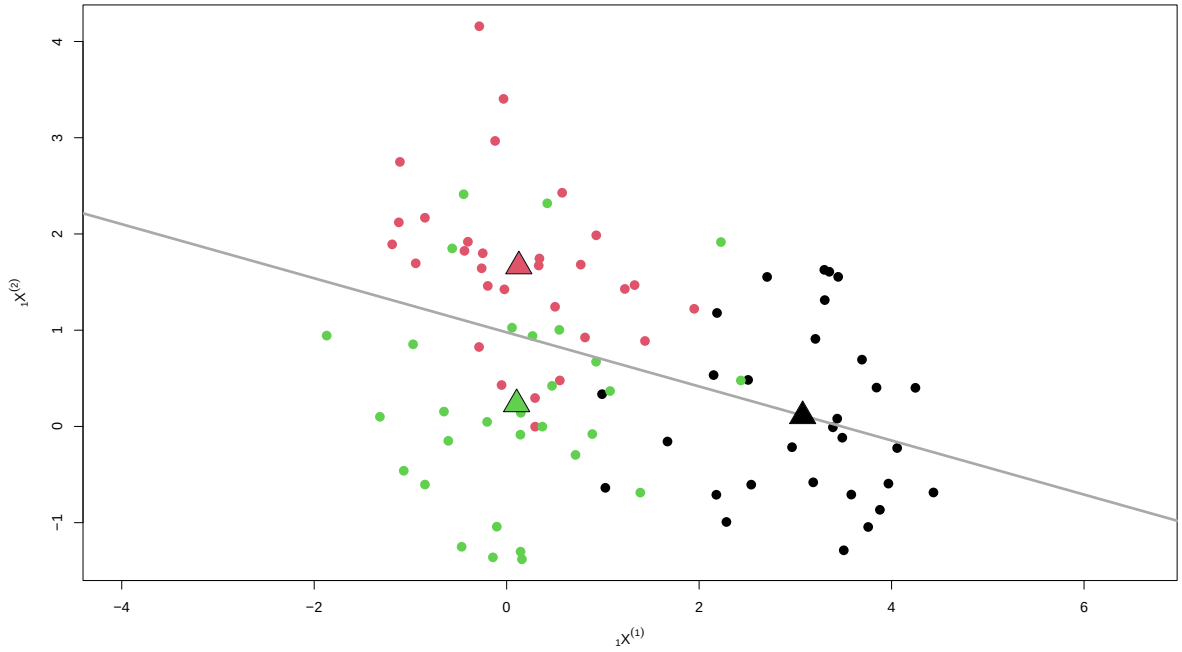
Zdroj: vlastné spracovanie

Pre lepšiu predstavu o fungovaní testu RLMS uvádzame na obrázkoch 3 až 5 ukážku krokov pre trojrozmerné pozorovania ($p = 3$) z normálneho rozdelenia so strednými hodnotami $\mu_1 = (3; 0; 0, 5)^T$, $\mu_2 = (0; 2; 0, 5)^T$, $\mu_3 = (0; 0; 0)^T$ a identickou kovariančnou maticou. Každý z $I = 3$ súborov obsahuje $n_i = 30$ pozorovaní.

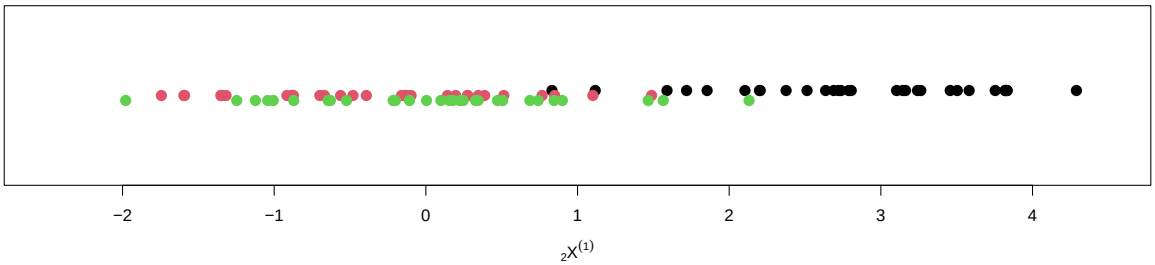


Obr. 3: Ukážka testu RLMS časť 1 - odhad roviny daný odhadmi stredov súborov pôvodných pozorovaní. *(Zdroj: vlastné spracovanie)*

Niekto by mohol namietať, že využitie lineárnej regresie nie je vhodné, pretože lineárna regresia je určená na modelovanie lineárnej závislosti, pričom nepredpokladáme, že by pozorovania boli lineárne závislé. V našom prípade to však nepredstavuje logický problém, pretože hľadáme najlepšiu možnú lineárnu reprezentáciu len za účelom redukcie dimenzie s cieľom aplikovať jednorozmerný test.



Obr. 4: Ukážka testu RLMS časť 2 - odhad priamky daný odhadmi stredov súborov ortogonálnych projekcií z obrázku 3. *(Zdroj: vlastné spracovanie)*



Obr. 5: Ukážka testu RLMS časť 3 - ortogonálne projekcie pozorovaní z obrázku 4. Tieto jednorozmerné pozorovania vstupujú do Kruskalovho-Wallisovho testu. Z dôvodu lepšej vizualizácie sú tieto jednorozmerné dáta zámerne vertikálne vychýlené. *(Zdroj: vlastné spracovanie)*

2.2.3 Projekcia na priamku určenú hlavnými smermi

Analýza hlavných komponentov (angl. *principal component analysis, PCA*), ktorú predstavil H. Hotelling v publikácii [11] je metóda na odvodenie množiny ortogonálnych lineárnych projekcií korelovaných premenných, pričom projektované premenné už nie sú korelované a hlavné komponenty sú zoradené zostupne podľa disperzie na nich projektovaných pozorovaní. Formálne definujeme hlavné komponenty nasledovne.

Definícia 7. *Nech X je p-rozmerný náhodný vektor so strednou hodnotou μ a kovariančnou maticou Σ so spektrálnym rozkladom $\Sigma = U\Delta U^T$. Potom náhodný vektor*

$$Y = U^T(X - \mu)$$

nazývame vektor hlavných komponentov náhodného vektora X. Nech $p \times p$ matica vlastných vektorov U má stĺpce (vlastné vektory) $u_1, u_2, \ldots, u_p$, potom $Y_i = u_i^T(X - \mu)$ je i-ty hlavný komponent náhodného vektora X, pričom $Var(Y) = \Delta = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ a $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Vlastný vektor u_i nazývame i-ty hlavný smer rozptylu.

V klasickom prístupe sa hlavné komponenty počítajú postupne. Prvý hlavný komponent zodpovedá hlavnému smeru, v ktorom majú projektované pozorovania najväčší rozptyl. Druhý hlavný smer je potom kolmý k prvému hlavnému smeru a opäť maximalizuje rozptyl. Pokračujúc týmto spôsobom sa vypočítajú všetky hlavné smery a hlavné komponenty. Formálne:

$$Var(Y_1) = \max\{Var(a^T X) \mid a \in \Re^p, \|a\| = 1\}, \tag{5}$$

a pre $k \geq 2$:

$$Var(Y_k) = \max\{Var(a^T X) \mid a \in \Re^p, \|a\| = 1, a \perp u_1, a \perp u_2, \ldots, a \perp u_{k-1}\}.$$

Vlastné hodnoty λ_i kovariančnej matice Δ sa interpretujú ako miera variability zachytená v tom ktorom hlavnom komponente. Keďže pracujeme s realizáciami z neznámych rozdelení, tak kovariančnú maticu musíme empiricky odhadnúť. Ďalšie poznatky na tému analýzy hlavných komponentov je možné nájsť napríklad v publikácii [13], z ktorej sme aj my vychádzali.

K myšlienke využiť hlavné komponenty nás priviedol spôsob MR, ktorý volí z $\binom{I}{2}$ priamok takú, aby projektované pozorovania boli čo najviac rozptýlené. Analýza hlavných komponentov z definície hľadá ortogonálne smery, ktoré sú zároveň zoradené zostupne podľa rozptylu v tom ktorom smere. Kľúčovým bodom v tejto myšlienke je, že pri prvom hlavnom komponente sa iba spojito hľadá najviac rozptýlený pohľad na pozorovania, pretože ešte nie je viazaný žiadnym iným, na ktorý by mal byť kolmý (pozri rovnicu (5)). A teda, keďže priamku, na ktorú projektuje pozorovania hľadá z nekonečne veľa možností, tak aspoň teoreticky by sme mali dostávať lepšie výsledky ako v prípade testu MR, ktorý hľadá tú optimálnu len z menšej množiny $\binom{I}{2}$ priamok. Druhým dôležitým rozdielom medzi MR a týmto prístupom je odlišný spôsob merania variability. Zatiaľ čo v prístupe MR sa hľadá najväčšia vzdialenosť sprojektovaných bodov (pozri rovnicu (2)), tak pri tomto prístupe sa maximalizuje celková variancia (pozri rovnicu (5)). Tiež by sme chceli poznamenať, že v dôsledku maximalizácie rozptylu v prvom hlavnom smere nadobudne prvý hlavný komponent ďalšiu vlastnosť, ktorú by sme od optimálnej priamky mohli vyžadovať, a to, že stredná hodnota μ a prvý hlavný smer u_1 udávajú takú priamku, pre ktorú sú kolmé vzdialenosti pozorovaní od tejto priamky v súčte minimálne. To okrem iného prináša aj druhý spôsob, akým je možné vypočítať hlavné komponenty. Postup výpočtu hlavných komponentov a odvodenie optimálnosti z hľadiska minimalizácie súčtu kolmých vzdialeností pozorovaní je možné nájsť napríklad v spomenutej publikácii [13] v podkapitole 7.2.3.

Hlavné komponenty sú však citlivé na škálovanie pozorovaní a z toho dôvodu sa vo väčšine prípadov odporúča pozorovania pred aplikáciou PCA štandardizovať. Keďže to nie je všeobecné pravidlo, štandardizáciu sme do algoritmu testu nezahrnuli a ponechávame na uvážení užívateľa, či pozorovania pred aplikáciou testu štandardizuje. Vo viacrozmernom teste polohy viacerých súborov teda navrhujeme nasledovný prístup, ktorý sme pre ďalšie referencie označili HK (*test s Hlavnými Komponentami*).

Navrhovaný algoritmus HK:

1. Výpočet prvého hlavného smeru a prvých hlavných komponentov všetkých pozorovaní.

2. Aplikácia Kruskalovho-Wallisovho testu na prvé hlavné komponenty.

Opäť však narážame na otázku, či sú splnené predpoklady Kruskalovho-Wallisovho testu z dôvodu spomenutého pri predchádzajúcich návrhoch. Konkrétne, pozorovania projektované na priamku danú prvým hlavným smerom už pravdepodobne nebudú nezávislé, pretože odhad hlavného smeru udávajú samotné pozorovania. Tento fakt môže opäť spôsobiť, že pravdepodobnosť chyby prvého druhu bude vychýlená k vyšším hodnotám ako bude stanovená hladina významnosti α. My sme však aspoň prostredníctvom simulácií overili, či nám odhad pravdepodobnosti chyby prvého druhu odhalí porušenie niektorého z predpokladov a dospeli sme k výsledkom uvedeným v tabuľke 6. Po tom, ako sa pri simulácii pravdepodobnosti chyby prvého druhu pre $n = 30$ dvoj-dimenzionálnych dát ako pri predchádzajúcich testoch neukázalo významné prekročenie zvolenej hladiny významnosti α, ktoré by nebolo možné považovať za dielo náhody, sme simulovali pravdepodobnosti aj pre menší počet dát $n = 10$ a taktiež aj pre väčšie dimenzie a viac súborov. Tieto výsledky uvádzame v tabuľke 6.

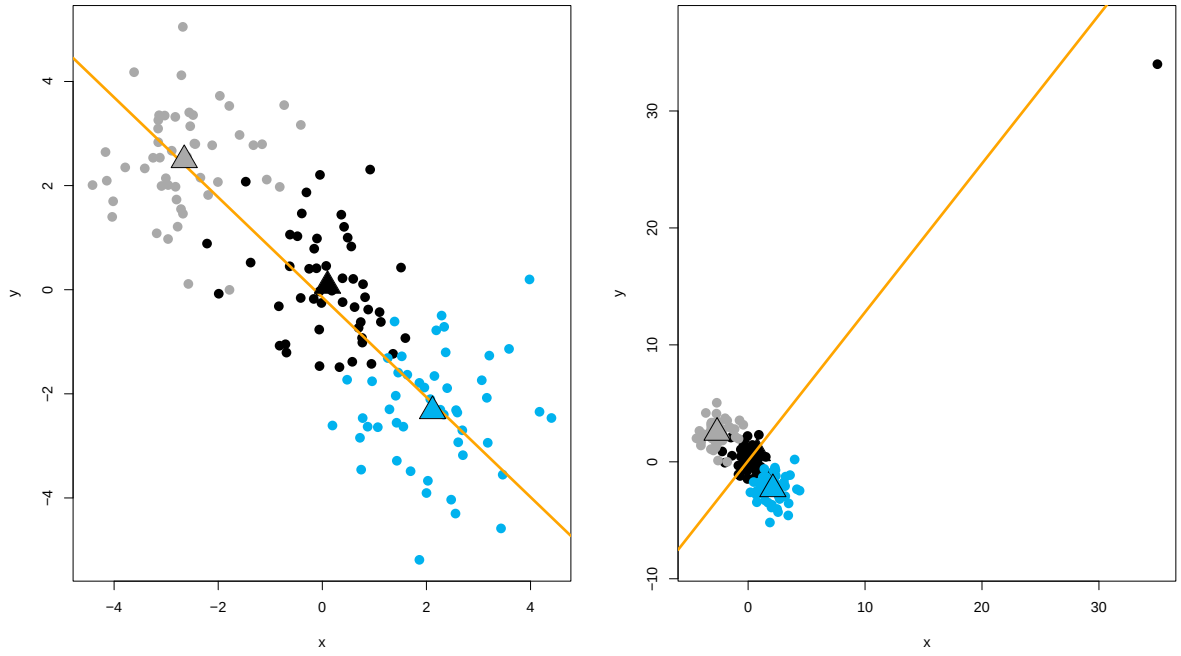
Tabuľka 6: Odhad pravdepodobnosti chyby prvého druhu testu HK na hladine významnosti $\alpha = 5\%$ prostredníctvom 10000 simulácií pre I nezávislých vyvážených súborov, každý s n_i p-dimenzionálnymi realizáciami z príslušného rozdelenia.

	I	n_i	p	$\mathcal{N}_p(0_p, I_p)$	$\mathcal{B}_p$	$t_1(0_p, I_p)$	$t_3(0_p, I_p)$
HK	3	30	2	0,047	0,052	0,047	0,049
	3	10	2	0,047	0,046	0,047	0,049
	3	30	5	0,048	0,051	0,047	0,049
	5	15	10	0,048	0,045	0,045	0,048

Zdroj: vlastné spracovanie

Avšak empirický odhad kovariančnej matice a rozptyl, ktorý sa maximalizuje, sú veľmi citlivé na odľahlé pozorovania (angl. *outliers*). V dôsledku toho sú často prvé hlavné smery priťahované k týmto anomálnym pozorovaniam. To môže negatívne ovplyvniť celú analýzu. Hlavne v navrhovanom prístupe testovania polohy, kde sa využíva len prvý hlavný komponent. Tento nedostatok demonštrujeme na obr. 6. Vľavo na obr. 6 sú znázornené

3 súbory (čierny, modrý a sivý), ktoré obsahujú po 30 realizácií z dvojrozmerného normálneho rozdelenia s identickou kovariančnou maticou a s odlišnými strednými hodnotami $\mu_1 = (0; 0)^T$, $\mu_2 = (2; -2, 3)^T$ a $\mu_3 = (-2, 5; 2, 4)^T$ znázornenými trojuholníkmi. Vpravo sú tie isté pozorovania s jedným odľahlým pozorovaním s hodnotami $(x, y)^T = (35; 34)^T$, ktoré sme doplnili do čierneho súboru. Oranžová priamka v oboch prípadoch znázorňuje smer prvého hlavného smeru.



Obr. 6: Demonštrácia citlivosti testu HK na anomálne body. *(Zdroj: vlastné spracovanie)*

Pozorujeme, že dokonca aj jedno odľahlé pozorovanie dokáže spôsobiť, že prvý a druhý hlavný smer sú určené naopak. To má v takomto prípade negatívny vplyv na výsledok testu HK. Zatiaľ čo v prípade vľavo sme vypočítali hodnotu Kruskalovej-Wallisovej χ^2 štatistiky približne 131,45 a zodpovedajúcu p-hodnotu menšiu ako $2, 2 \cdot 10^{-16}$, tak v prípade vpravo sme vypočítali hodnotu testovej štatistiky približne 4,043 a zodpovedajúcu p-hodnotu približne 0,1325. To znamená, že v prvom prípade hypotézu o rovnosti stredných hodnôt celkom isto správne zamietame a v druhom prípade túto hypotézu celkom isto nezamietame. Tento rozdielny výsledok je spôsobený práve orientáciou prvého hlavného smeru. V prípade vľavo prvý hlavný smer dostatočne dobre odhadol približne lineárnu štruktúru polohy súborov a jednorozmerné projekcie súborov bolo možné dostatočne

dobre rozlíšiť a zamietnuť nulovú hypotézu. V prípade vpravo bol smer hlavného smeru ortogonálny na túto lineárnu štruktúru, čo spôsobilo, že rozdielnosť μ_1, μ_2, μ_3 po projekcií pozorovaní nebolo možné zistiť.

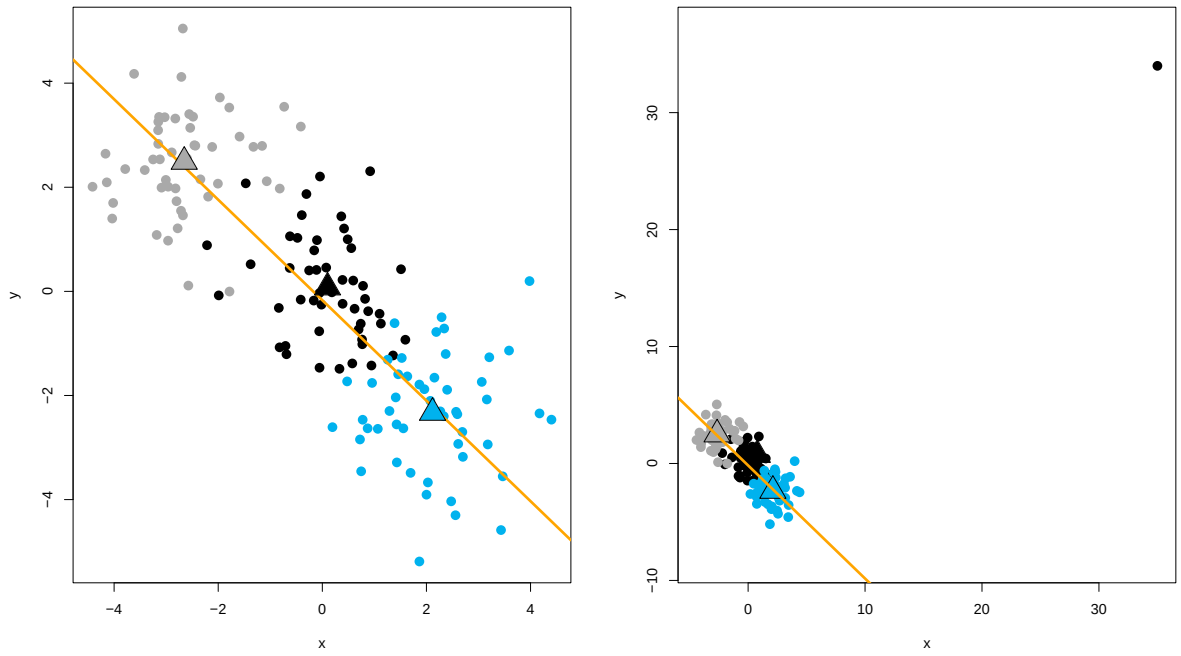
Tieto problémy nás priviedli k skúmaniu robustných verzií PCA, ktorých cieľom je získať hlavné smery, ktoré nie sú príliš ovplyvnené odľahlými pozorovaniami. Takýchto verzií bolo v minulosti predstavených niekoľko. Jedna skupina metód sa snaží namiesto klasickej kovariančnej matice využiť jej rôzne robustné odhady. Druhá skupina metód využíva projekčné sledovanie (angl. *projection pursuit*) s cieľom maximalizovať robustnú mieru rozptylu. Jedna z robustných verzií, ktorá využíva kombináciu robustného odhadu kovariančnej matice a projekčného sledovania bola predstavená v publikácii [12] autormi M. Hubert, P. J. Rousseeuw a K. V. Branden v roku 2005. Táto publikácia okrem návrhu robustného algoritmu *ROBPCA* taktiež obsahuje stručný prehľad alternatívnych metód spoločne so simulačnou štúdiou, ktorá porovnáva navrhnutý algoritmus s alternatívami. Po preštudovaní úvah, záverov a simulačnej štúdie sme sa pre našu aplikáciu rozhodli využiť práve algoritmus *ROBPCA* navrhnutý v tejto publikácií. Pri výpočtoch sme využili implementáciu v jazyku R [20] z balíka *rospca* [21]. Robustnú verziu navrhnutého testu HK, teda verziu, v ktorej namiesto klasického prístupu k výpočtu PCA využívame spomenutý robustný algoritmus, ďalej v tejto práci označujeme ako RHK (Robustný HK). Uvažujme príklad, ktorý sme vyššie opisovali na obrázku 6. Na totožné pozorovania sme aplikovali aj robustnú verziu HK, ktorej výsledok uvádzame na obrázku 7.

Tabuľka 7: Odhad pravdepodobnosti chyby prvého druhu testu RHK na hladine významnosti $\alpha = 5\%$ prostredníctvom 10000 simulácií pre I nezávislých vyvážených súborov, každý s n_i p-dimenzionálnymi realizáciami z príslušného rozdelenia.

	I	n_i	p	$\mathcal{N}_p(0_p, I_p)$	$\mathcal{B}_p$	$t_1(0_p, I_p)$	$t_3(0_p, I_p)$
RHK	3	30	2	0,046	0,051	0,047	0,049
	3	10	2	0,048	0,046	0,047	0,046
	3	30	5	0,048	0,049	0,046	0,045
	5	15	10	0,048	0,045	0,046	0,047

Zdroj: vlastné spracovanie

Tento algoritmus sa už nenechal pomýliť jedným pozorovaním a v oboch prípadoch



Obr. 7: Demonštrácia robustnosti RHK na anomálne body. Vľavo - testová štatistika Kruskalovho-Wallisovho testu 131,45 a zodpovedajúca p-hodnota menej ako $2,2 \cdot 10^{-16}$. Vpravo - testová štatistika Kruskalovho-Wallisovho testu 132,33 a zodpovedajúca p-hodnota menej ako $2,2 \cdot 10^{-16}$. *(Zdroj: vlastné spracovanie)*

hypotézu o rovnosti stredných hodnôt test správne s určitosťou zamietol. Pre takto upravený algoritmus sme opäť simulovali odhady pravdepodobnosti chyby prvého druhu, ktoré uvádzame v tabuľke 7. Z výsledkov simulácií sa domnievame, že úprava algoritmu na robustný nenarušila skutočnú pravdepodobnosť chyby prvého druhu. Pri väčších dimenziách však môžu bežne nastať situácie, kedy prostredníctvom prvého hlavného komponentu nebude možné zachytiť potrebné množstvo variancie a sila testu RHK by v takých prípadoch bola nízka. Preto sme sa tento prístup rozhodli upraviť na nasledujúci algoritmus, ktorý zohľadňuje mieru variancie zachytenú aj v ďalších hlavných komponentoch. Pre ďalšie referencie sme tento algoritmus označili skratkou KRHK (*Korigovaný Robustný test s Hlavnými Komponentami*).

Navrhovaný algoritmus KRHK:

1. Výpočet všetkých robustných hlavných smerov podľa algoritmus *ROBPCA* [12].

2. Voľba prvých k hlavných smerov, tak aby odhad podielu vysvetlenej variancie bol aspoň 80%, t. j.
$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq 80\%, \tag{6}$$
alebo aby
$$\frac{\lambda_k}{\lambda_1} \leq 10^{-3}$$
kde λ_i, je vlastná hodnota kovariančnej matice, pričom vlastné hodnoty sú usporiadané zostupne.

3. Aplikácia Kruskalovho-Wallisovho testu pre projektované pozorovania v smere každého z k zvolených hlavných smerov.

4. Vyhodnotenie nadobudnutých p-hodnôt prostredníctvom Holmovej-Bonferroniho sekvenčnej metódy (pozri tvrdenie 2).

5. Vyhodnotenie - zamietnutie rovnosti stredných hodnôt, ak aspoň jeden z k Kruskalových-Wallisových testov zamietol jemu príslušnú nulovú hypotézu.

Rovnako ako pre všetky ostatné návrhy sme opäť simulovali odhady pravdepodobnosti chyby prvého druhu. Výsledky simulácií uvádzame v tabuľke 8. Pozorujeme, že všetky odhady sú menšie ako zvolená hladina významnosti $\alpha = 5\%$, takže môžeme z tohto pohľadu prehlásiť KRHK za spoľahlivý. Taktiež pozorujeme, že KRHK má s klesajúcim počtom pozorovaní a rastúcim počtom súborov I tendenciu mať menšiu pravdepodobnosť chyby prvého druhu (pozri prvý, druhý a štvrtý riadok), no iba zvýšenie dimenzie nemalo na pravdepodobnosť chyby prvého druhu vplyv (pozri prvý a tretí riadok).

Hoci sme KRHK navrhli s cieľom zachytiť potrebné množstvo variancie vo vyšších dimenziách, tak dokonca aj pre nižšie dimenzie existujú špeciálne prípady, keď HK a RHK zlyhajú a KRHK správne odhalí neplatnosť nulovej hypotézy o rovnosti parametrov polohy. Jeden z takýchto špeciálnych prípadov demonštrujeme na obrázku 8. Prístupy HK

Tabuľka 8: Odhad pravdepodobnosti chyby prvého druhu testu KRHK na hladine významnosti $\alpha = 5\%$ prostredníctvom 10000 simulácií pre I nezávislých vyvážených súborov, každý s n_i p-dimenzionálnymi realizáciami z príslušného rozdelenia.

	I	n_i	p	$\mathcal{N}_p(0_p, I_p)$	$\mathcal{B}_p$	$t_1(0_p, I_p)$	$t_3(0_p, I_p)$
KRHK	3	30	2	0,045	0,044	0,047	0,048
	3	10	2	0,041	0,039	0,044	0,041
	3	30	5	0,047	0,049	0,047	0,049
	5	15	10	0,034	0,035	0,036	0,038

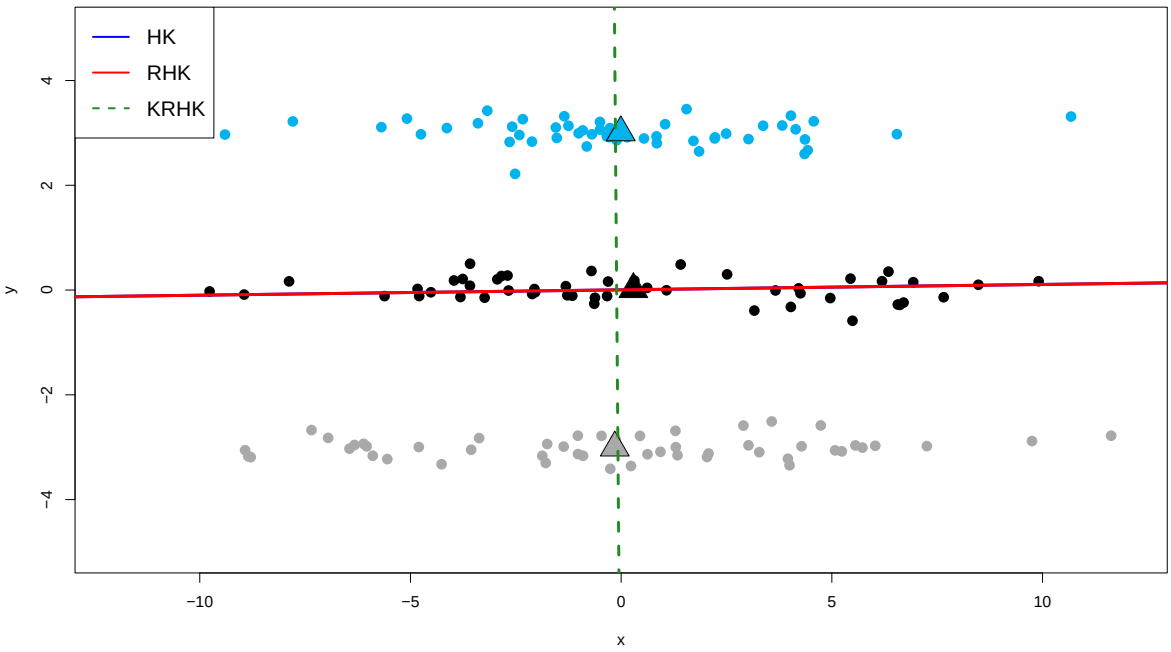
Zdroj: vlastné spracovanie

a RHK správne odhalili smer s najväčšou varianciou. Avšak nie vždy to je ten správny smer pre odhalenie neplatnosti nulovej hypotézy, čo sa v tomto prípade potvrdilo. Oba prístupy, HK a RHK, s určitosťou nezamietli hypotézu o rovnosti parametrov polohy s p-hodnotou rovnou $0,9609$. Prístup KRHK v prípade kovariančnej matice $\Sigma = \text{diag}(20; 0,05)$ dvojrozmerného normálneho rozdelenia, uvažoval aj druhý hlavný smer znázornený zelenou prerušovanou priamkou, pri ktorom je jasne vidieť rozdielnosti v parametroch polohy súborov, pretože prvý hlavný komponent ($k = 1$) nespĺňal podmienku (6) v druhom kroku definície algoritmu KRHK, pretože pri odhadnutých hodnotách $\lambda_1 \doteq 19,808$ a $\lambda_2 \doteq 6,495$

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} = \frac{19,808}{19,808 + 6,495} = 0,753 \leq 0,8.$$

V príklade na obrázku 8 teda prístup KRHK správne zamietol nulovú hypotézu. Ak by však kovariančná matica bola $\Sigma = \text{diag}(2000; 0,05)$, tak by algoritmus KRHK v podmienke v druhom kroku vyhodnotil, že je dostatočný iba prvý hlavný komponent a rovnako ako prístupy HK a RHK by neodhalil nerovnosť parametrov polohy. Ako sme však uviedli pri definícii algoritmu HK, hlavné smery sú citlivé na škálovanie pozorovaní, čo sa presne ukázalo byť v tomto príklade problémom. Ak však testy vykonáme na štandardizovaných pozorovaniach, tak všetky tri testy správne odhalia neplatnosť nulovej hypotézy.

Tento príklad nás priviedol k ďalšej úprave testovacieho algoritmu, ktorá využíva koncept priestorového znamienka. Priestorové znamienko sme už spomenuli v podkapitole 2.1 v súvislosti s testom od H. Oju a R. H. Randlesa z publikácie [17], kde je definované



Obr. 8: Demonštrácia špeciálneho prípadu slabosti HK a RHK. Trojuholníky označujú stredy súborov. Všetky tri súbory (čierny, modrý a sivý) obsahujú po 50 pozorovaní, ktoré pochádzajú z dvojrozmerných normálnych rozdelení s kovariančnou maticou $\Sigma = \text{diag}(20; 0,05)$ a so strednými hodnotami $\mu_1 = (0;0)^T$, $\mu_2 = (0;3)^T$ a $\mu_3 = (0;-3)^T$. *(Zdroj: vlastné spracovanie)*

ako funkcia

$$S(x) = \begin{cases} \|x\|^{-1}x, & x \neq 0, \\ 0, & x = 0, \end{cases}$$

kde $\|\cdot\|$ značí Euklidovskú normu. Geometricky funkcia $S(x)$ predstavuje projekciu bodu $x \in \Re^p$ na povrch p-rozmernej jednotkovej sféry so stredom v bode 0. Uvažujme opäť I súborov, každý s n_i pozorovaniami $X_{ij} \in \Re^p$, $i = 1, 2, \ldots, I$ a $j = 1, 2, \ldots, n_i$. Navrhujeme vypočítať body S_{ij} využijúc funkciu $S(x)$ nasledujúcim spôsobom

$$S_{ij} := \overline{X}_i + S(X_{ij} - \overline{X}_i), \tag{7}$$

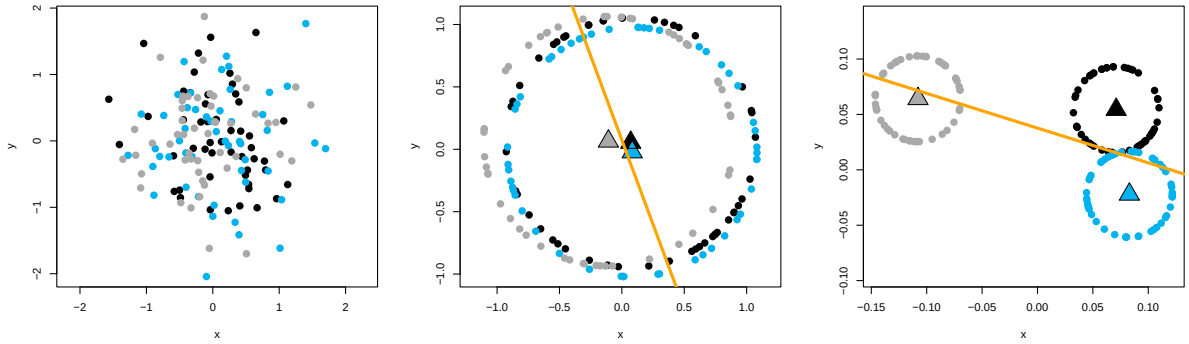
kde $\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$ a následne na transformované body S_{ij} aplikovať algoritmus hlavných komponentov z cieľom nájsť optimálny smer pre zamietnutie nulovej hypotézy. Získaný prvý hlavný smer a celkový priemer pozorovaní nám udajú priamku, na ktorú následne projektujeme pôvodné pozorovania X_{ij} a aplikujeme Kruskalov-Wallisov test pre jednorozmerné pozorovania. Poznamenávame, že v tomto prípade nie je nutné využiť

robustnú verziu hlavných komponentov, pretože pozorovania S_{ij} neobsahujú „outliery“, keďže všetky ležia na jednotkovej sfére so stredom v $\overline{X}_i$. Cieľom tohto postupu je využiť myšlienku, že ak existuje rozdiel v polohe súborov, tak sa tento rozdiel prejaví aj v rozptyle v smere tohto rozdielu, avšak pri hľadaní optimálnej priamky zároveň odfiltrovať rozptyl daný pravdepodobnostným rozdelením, z ktorého pozorovania pochádzajú. Pri simuláciách pravdepodobnosti chyby prvého druhu sa v tomto prípade ukázal problém s prekročením stanovenej hladiny významnosti $\alpha = 5\%$, takže je opäť nutné využiť permutačnú verziu. Problémy s nezanedbateľným prekročením hladiny významnosti α sa však nepodarilo odstrániť ani prostredníctvom 1000 permutácií ako v predchádzajúcich prípadoch. I keď by sme pravdepodobne dokázali nájsť počet permutácií, pre ktorý by už tento postup nevykazoval problémy s pravdepodobnosťou chyby prvého druhu, tak radšej sme sa pozreli hlbšie na príčinu problému. Pri detailnejšej analýze sme objavili, že aj keď je tento postup nenechá pomýliť situáciou, ktorú prezentujeme na obrázku 8, kde sú vzdialenosti medzi strednými hodnotami výrazne nad 1, tak pre malé až žiadne rozdiely transformované pozorovania S_{ij} ležia približne na rovnakej kružnici. Na obrázku 9 uvádzame príklad takejto situácie. Vidno, že v takýchto prípadoch sa „zašumí“ veľká časť informácie o smere najväčšieho rozptylu a v dôsledku toho sa odhadnutá priamka veľmi kolíše aj pri malých zmenách v pozorovaniach. To spôsobí, že priamka je tak veľmi závislá od pozorovaní, že ani permutačný test nepomôže pri stabilizácii pravdepodobnosti chyby prvého druhu. Ak by však transformované pozorovania S_{ij} neležali na jednotkovej sfére, ale na sfére s dostatočne malým polomerom r, tak by sa jednotlivé sféry pre ľubovoľne malé rozdiely v stredných hodnotách nepretínali. Pripomíname, že transformované pozorovania S_{ij} sa využijú len na odhad optimálnej priamky prostredníctvom metódy hlavných komponentov, na ktorú následne projektujeme pôvodné pozorovania X_{ij}. Všimnime si, že polomer r istým spôsobom udáva pri hľadaní optimálnej priamky pomer váh medzi odhadom strednej hodnoty v i-tom súbore a disperziou v i-tom súbore, ktoré vstupujú do metódy hlavných komponentov. Zdá sa nám, že polomer r je vhodné nastaviť tak, aby sa najbližšie sféry dotýkali, pretože nedochádza k prekrytiu sfér. V zmysle pomeru váh je to tiež rozumné, pretože čím sú súbory bližšie, tým menej nás zaujíma ich disperzia a záleží nám hlavne na odhade strednej hodnoty. Transformované pozorovania S_{ij} teda

navrhujeme vypočítať nasledovne.

$$S_{ij} := \overline{X}_i + r \cdot S(X_{ij} - \overline{X}_i), \text{ kde } r := \min_{\substack{i,j=1,2,\ldots,I \\ i<j}} \frac{\|\overline{X}_i - \overline{X}_j\|_2}{2}. \tag{8}$$

Pre lepšie porozumenie uvedenej úpravy uvádzame na obrázku 9 príklad, kde je vidieť vplyv úpravy výpočtu S_{ij} zo vzťahu (7) na vzťah (8).



Obr. 9: Porovnanie odhadu optimálnej priamky pre výpočet S_{ij}. Vľavo sú zobrazené pôvodné pozorovania X_{ij}, v strede sú S_{ij} počítané prostredníctvom vzťahu (7) a vpravo prostredníctvom vzťahu (8). *(Zdroj: vlastné spracovanie)*

Pre testovanie hypotézy o rovnosti parametrov polohy viacerých súborov navrhujeme nasledujúci postup, ktorý ďalej v práci označujeme ako HSPZ (*test s Hlavným Smerom Priestorových Znamienok*).

Navrhovaný algoritmus HSPZ:

1. Výpočet transformovaných pozorovaní S_{ij} prostredníctvom vzťahu (8).
2. Výpočet prvého hlavného smeru pre S_{ij}.
3. Projekcia pôvodných pozorovaní X_{ij} na priamku danú vypočítaným prvým hlavným smerom a celkovým priemerom z X_{ij}.
4. Aplikácia Kruskalovho-Wallisovho testu pre sprojektované X_{ij}.

Výsledky simulácií pravdepodobnosti prvého druhu HSPZ uvádzame v tabuľke 9. Keďže aj v tomto prípade bola pravdepodobnosť základného HSPZ nad zvolenou hladinou významnosti $\alpha = 5\%$, tak je potrebné využiť jeho permutačnú verziu. Pri 1000 permutáciách sa už neukázalo prekročenie hladiny významnosti α, ktoré by nebolo možné pokladať za dielo náhody.

Tabuľka 9: Odhad pravdepodobnosti chyby prvého druhu na hladine významnosti $\alpha = 5\%$ prostredníctvom 10000 simulácií pre $I = 3$ nezávislé vyvážené súbory, každý s $n_i = 30$ dátami rozmeru $p = 2$ z príslušného rozdelenia. V permutačnom teste bolo vykonaných 1000 permutácií.

	$\mathcal{N}_2(0_2, I_2)$	$\mathcal{B}_2$	$t_1(0_p, I_p)$	$t_3(0_p, I_p)$
HSPZ	0,097	0,100	0,058	0,085
Permutačný HSPZ	0,050	0,047	0,049	0,050

Zdroj: vlastné spracovanie

2.3 Navrhované rozšírenia vychádzajúce z testovej štatistiky Kruskala a Wallisa

V tejto podkapitole postupne predstavíme sadu navrhovaných testov, ktoré zdieľajú myšlienku viacrozmerného zovšeobecnenia testovej štatistiky Kruskala a Wallisa. Ako píšeme neskôr v tejto kapitole, táto testová štatistika sa dá jednoduchou úpravou prepísať na súčet kvadratických foriem. Otázkou však zostáva, ako zmysluplne zovšeobecniť jednotlivé premenné, ktoré v testovej štatistike vystupujú. Odpovede na túto otázku prezentujeme v nasledujúcich podkapitolách. V istom zmysle podobne ako my, no nie úplne rovnako, pristupovali k zovšeobecneniu Kruskalovho-Wallisovho testu aj Bhapkar, Chatterjee, Choi, Marden, Puri, Sen a ďalší vo svojich publikáciach. Pre prehľad uvádzame výber z týchto publikácií [2, 4, 5, 19].

Vychádzajúc z pôvodnej publikácie Kruskala a Wallisa [14] je možné testovú štatistiku testu opísaného v podkapitole 1.2.2 zapísať rôznymi spôsobmi. Prvý spôsob, ktorý sa zvykne používať, sme uviedli aj my pri opise testu (pozri vzťah (1) v podkapitole 1.2.2). Druhý spôsob autori uvádzajú v publikácií [14] s ich označením (1.5), konkrétne

$$H = \frac{N-1}{N} \sum_{i=1}^{I} \frac{n_i[\overline{R}_i - \frac{1}{2}(N+1)]^2}{(N^2-1)/12}, \tag{9}$$

kde $\overline{R}_i$ je priemer n_i rankov $X_{ij}^{(k)}$ v i-tom súbore získaných zo vzostupného usporiadania hodnôt v združenom súbore, $\frac{1}{2}(N+1)$ je stredná hodnota a $(N^2-1)/12$ variancia rovnomerného rozdelenia na prvých N prirodzených číslach[2]. Poznamenávame, že za platnosti hypotézy H_0 sa každý rank riadi takýmto rovnomerným rozdelením. Všimnime si, že testovú štatistiku je možné zovšeobecniť do tvaru

$$H_{MKW} = \frac{N-1}{N} \sum_{i=1}^{I} n_i(\overline{R}_i - E_R)^T \Sigma_R^{-1}(\overline{R}_i - E_R), \tag{10}$$

kde v jednorozmernom prípade je $E_R := \frac{1}{2}(N+1)$ a $\Sigma_R := (N^2-1)/12$. Tento tvar je však už formálne použiteľný aj pre $\overline{R}_i = (\overline{R}_i^{(1)}, \overline{R}_i^{(2)}, \ldots, \overline{R}_i^{(p)})^T$, p-rozmernú strednú hodnotu E_R a $p \times p$-rozmernú kovariančnú maticu Σ_R. V tomto momente narážame na problém priameho rozšírenia spomenutý v podkapitole 2.1, pretože nie je jasné ako vhodne zoradiť vektory a aký tvar by mali mať E_R a Σ_R. Prirodzene sa ponúka možnosť zoradiť pozorovania $X_{ij} \in \Re^p$ zo združeného výberu v každej z p zložiek samostatne a tak vytvoriť p-rozmerné vektory priemerných rankov $\overline{R}_i$, t. j. nech $R_{ij}^{(k)}$ je rank $X_{ij}^{(k)}$ vo vzostupnom usporiadaní hodnôt množiny združeného výberu $\{X_\eta^{(k)}, \eta = 1, 2, \ldots, N\}$, a $\overline{R}_i^{(k)} = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij}^{(k)}$ pre $k = 1, 2, \ldots, p$. Strednú hodnotu v takom prípade vieme jednoducho zovšeobecniť na

$$E_R = \frac{N+1}{2} \cdot 1_p,$$

kde 1_p je p-rozmerný vektor jednotiek. Jednoduchým, avšak naivným, spôsobom, ako zovšeobecniť kovariančnú maticu, by bolo ju zvoliť nasledovne

$$\Sigma_R = \frac{N^2-1}{12} \cdot I_p, \tag{11}$$

kde I_p značí p-rozmernú maticu identity. Takýto test by mal šancu fungovať iba ak by sme si boli istí, že je v pozorovaniach nie je medzi jednotlivými zložkami závislosť. Ak by tento predpoklad nebol splnený, tak by sa závislosť preniesla z pozorovaní aj do rankov

[2]Keďže predpokladáme, že rozdelenia, z ktorých pozorovania pochádzajú sú spojité, tak pravdepodobnosť rovnosti dvoch pozorovaní je nula. Preto modifikáciu pre tento problém v práci nespomíname.

a matica Σ_R by bola nesprávna, čo by spôsobilo nespoľahlivosť výsledkov testu. V nasledujúcich podkapitolách rozoberieme niektoré z možných odhadov kovariančnej matice Σ_R. Existujú aspoň dva prístupy akými je možné sa vysporiadať s týmto problémom. Buď odhadneme maticu Σ_R tak, aby zachytávala všetky závislosti, alebo sa pokúsime závislosti eliminovať. Po eliminácii závislostí by matica Σ_R s určitosťou mala uvedený tvar (11). Výhodou eliminačného prístupu je diagonálny tvar tejto matice, pretože tvar testovej štatistiky (10) sa jednoduchými úpravami prevedie na súčet p jednorozmerných H štatistík nasledovne

$$H_{MKW}\Big|_{\Sigma_R=\frac{N^2-1}{12}\cdot I_p} = \sum_{k=1}^{p} \frac{N-1}{N} \sum_{i=1}^{I} \frac{n_i[\overline{R}_i^{(k)} - \frac{1}{2}(N+1)]^2}{(N^2-1)/12} \triangleq \sum_{k=1}^{p} H^{(k)}. \tag{12}$$

Vďaka tomu, že za platnosti hypotézy H_0 sa jednorozmerná štatistika H asymptoticky riadi χ^2 rozdelením s $I-1$ stupňami voľnosti a vďaka aditívnej vlastnosti χ^2 rozdelenia vieme, že za platnosti hypotézy H_0 sa testová štatistika H_{MKW} asymptoticky riadi χ^2 rozdelením s $p(I-1)$ stupňami voľnosti.

2.3.1 Dekorelovaný viacrozmerný Kruskalov-Wallisov test

V snahe dosiahnuť asymptoticky χ^2 rozdelenie testovej štatistiky využívajúcej naivný fakt (11) navrhujeme z pozorovaní odstrániť aspoň korelácie, t. j. lineárne závislosti, zložiek. To je možné dosiahnuť aplikáciou metódy hlavných komponentov, ktorú sme opísali v podkapitole 2.2.3. Pre *dekorelovaný viacrozmerný Kruskalov-Wallisov test*, ktorý ďalej označujeme ako „DecorMKW“, navrhujeme nasledujúci algoritmus.

Navrhovaný algoritmus DecorMKW:

1. Dekorelácia pôvodných pozorovaní $X_\eta \in \Re^p$ zo združeného súboru transformáciou na hlavné komponenty $Y_\eta \in \Re^p$ pre $\eta = 1, 2, \ldots, N$.

2. Určenie hodnôt rankov $R_\eta^{(k)}$ pozorovaní $Y_\eta^{(k)}$ pre $k = 1, 2, \ldots, p$ a zostrojenie vektorov $\overline{R}_i$.

3. Výpočet testovej štatistiky (12) a porovnanie s príslušným kvantilom $\chi^2_{p(I-1)}$ rozde-

lenia.

Aby sme potvrdili vhodnosť tohto testu, tak sme opäť simulovali odhad pravdepodobnosti chyby prvého druhu pre rôzne scenáre. Výsledky simulácií uvádzame v tabuľke 10. V žiadnom z uvedených scenárov nedošlo k prekročeniu stanovenej hladiny významnosti $\alpha = 5\%$. V tabuľke 10 možno pozorovať aj asymptotické správanie testovej štatistiky, pretože pre malý počet dát pozorujeme istú konzervatívnosť testu pri zamietaní nulovej hypotézy. Ako môžeme vidieť v poslednom riadku tabuľky 10, tento efekt sa ešte umocňuje zvýšením dimenzie.

Tabuľka 10: Odhad pravdepodobnosti chyby prvého druhu testu DecorMKW na hladine významnosti $\alpha = 5\%$ prostredníctvom 10000 simulácií pre I nezávislých vyvážených súborov, každý s n_i d-dimenzionálnymi realizáciami z príslušného rozdelenia.

	I	n_i	p	$\mathcal{N}_p(0_p, I_p)$	$\mathcal{B}_p$	$t_1(0_p, I_p)$	$t_3(0_p, I_p)$
DecorMKW	3	30	2	0,050	0,050	0,048	0,048
	3	10	2	0,040	0,046	0,042	0,044
	3	30	5	0,040	0,043	0,044	0,047
	5	15	10	0,035	0,035	0,040	0,038

Zdroj: vlastné spracovanie

I keď sa nám dekoreláciou podarilo rozšíriť aplikovateľnosť testu voči naivnému prístupu aj na pozorovania pochádzajúce z elipticky symetrického rozdelenia, čo demonštrujeme v tabuľke 11, tak sme si vedomí, že nie je vhodný na situácie, keď sa medzi zložkami pozorovaní vyskytujú aj nelineárne závislosti, čiže napríklad pozorovania nepochádzajú z elipticky symetrického rozdelenia. V takých prípadoch nie je jasné, ako vo všeobecnosti zabezpečiť nekorelovanosť rankov a aký vplyv by to malo na silu testu.

2.3.2 Viacrozmerný Kruskalov-Wallisov test s Ojovými rankami

K nasledujúcemu návrhu nás priviedla snaha priblížiť sa testu H. Oju a R. H. Randlesa z publikácie [17], ktorý využíva koncept priestorového ranku. Keďže takýchto konceptov existuje viacero, priestorový rank z publikácie [17] nazývame Ojovým. Po bližšej

Tabuľka 11: Odhad pravdepodobnosti chyby prvého druhu na hladine významnosti $\alpha = 5\%$ prostredníctvom 10000 simulácií pre $I = 3$ nezávislé vyvážené súbory, každý s $n_i = 30$ dátami rozmeru $p = 2$ z príslušného rozdelenia s kovariančnou maticou $\Sigma = \begin{bmatrix} 10 & 3 \\ 3 & 1 \end{bmatrix}$

	$\mathcal{N}_2(0_2, \Sigma)$	$t_1(0_2, \Sigma)$	$t_3(0_2, \Sigma)$
MKW	0,089	0,083	0,089
DecorMKW	0,049	0,056	0,048

Zdroj: vlastné spracovanie

analýze nie je náročné zistiť, že Ojove ranky sú v podstate netriviálne transformované pozorovania do jednotkovej p-rozmernej sféry. Prípadné rozdiely v parametre polohy však zostanú zachované aj po tejto transformácii a zároveň za platnosti nulovej hypotézy sú prípadné korelácie eliminované, pretože podľa tvrdenia Oju a Randlesa z publikácie [17] na strane 600, kovariančná matica Ojových rankov je vďaka nimi navrhnutej špeciálnej transformácii identická matica vynásobená skalárom c_X^2/p, kde c_X^2 je konštanta bližšie popísaná v spomenutej publikácii. Ich výhodou je tiež ich afinná invariantnosť. To nás priviedlo k myšlienke dodatočne priradiť Ojovým rankom pozložkové ranky. Z uvedeného vyplýva, že Ojove ranky sa za platnosti nulovej hypotézy správajú ako keby pochádzali zo sféricky symetrického rozdelenia na jednotkovej p-rozmernej sfére. To zaručí, že je možné kovariančnú maticu Σ_R odhadnúť pomocou vzťahu (11). Poznamenávame, že nové ranky sa nezhodujú s Ojovými rankami. Navrhovaný *viacrozmerný Kruskalov-Wallisov test s Ojovými Rankami* ďalej v práci označujeme ako „MKWOR“.

Navrhovaný algoritmus MKWOR:

1. Výpočet vektorov priestorových rankov Y_η pre $\eta = 1, 2, \ldots, N$ podľa kapitoly 2.3 publikácie [17][3].

2. Určenie hodnôt pozložkových rankov $R_\eta^{(k)}$ pozorovaní $Y_\eta^{(k)}$ pre $k = 1, 2, \ldots, p$ a zostrojenie vektorov $\overline{R}_i$.

[3] V publikácii [17] sú Y_η v kapitole 2.3 značené ako R_i.

3. Výpočet testovej štatistiky (12) a porovnanie s príslušným kvantilom $\chi^2_{p(I-1)}$ rozdelenia.

Odhad pravdepodobnosti chyby prvého druhu testu MKWOR pre rôzne scenáre uvádzame v tabuľke 12. Test nevykazuje výrazné prekročenie hladiny významnosti $\alpha = 5\%$. V niektorých prípadoch by sa dalo povedať, že test je až príliš konzervatívny.

Tabuľka 12: Odhad pravdepodobnosti chyby prvého druhu testu MKWOR na hladine významnosti $\alpha = 5\%$ prostredníctvom 10000 simulácií pre I nezávislých vyvážených súborov, každý s n_i d-dimenzionálnymi realizáciami z príslušného rozdelenia.

	I	n_i	p	$\mathcal{N}_p(0_p, I_p)$	$\mathcal{B}_p$	$t_1(0_p, I_p)$	$t_3(0_p, I_p)$
MKWOR	3	30	2	0,048	0,051	0,046	0,048
	3	10	2	0,042	0,046	0,040	0,042
	3	30	5	0,042	0,043	0,043	0,047
	5	15	10	0,035	0,034	0,036	0,037

Zdroj: vlastné spracovanie

2.3.3 Viacrozmerný Kruskalov-Wallisov test s priamym odhadom kovariancie

Prístup eliminácie závislostí má svoje limitácie, a preto sme skúmali prístupy pre odhad kovariančnej matice Σ_R, ktoré by bola schopná zachytiť závislosti medzi rankami. V takom prípade už nie je možné upraviť testovú štatistiku H_{MKW} do tvaru (12). To spôsobí, že štatistika H_{MKW} sa už nutne nemusí asymptoticky riadiť $\chi^2_{p(I-1)}$ rozdelením. Predpokladáme však, že pokiaľ je matica Σ_R odhadnutá tak, že správne zachytáva všetky závislosti, tak sa predsa len asymptotické rozdelenie štatistiky nezmení[4]. V nasledujúcom odseku sme predstavili algoritmus, ktorým je možné odhadnúť kovariančnú maticu rankov.

Uvažujme pozorovania X, tak ako sú definované na začiatku kapitoly 2. Spolu N pozorovaní. Nech $R_\eta = (R_\eta^{(1)}, R_\eta^{(2)}, \ldots, R_\eta^{(p)})^T$ je pozložkový priestorový rank pre nejaké

[4]Predpoklad sa neopiera o žiadnu vetu či tvrdenie.

pozorovanie η. Nech rank $R_\eta^{(k)} \in 1, 2, \ldots, N$ značí pozíciu vo vzostupnom usporiadaní všetkých pozorovaní v k-tej zložke, potom ho je možné zapísať v tvare

$$R_\eta^{(k)} = \frac{1}{2} \sum_{i=1}^{N} \operatorname{sgn}(X_\eta^{(k)} - X_i^{(k)}) + \frac{N+1}{2},$$

kde $\operatorname{sgn}(\cdot)$ označuje funkciu signum definovanú pre $z \in \Re$ ako

$$\operatorname{sgn}(z) = \begin{cases} 1, & \text{ak } z > 0, \\ 0, & \text{ak } z = 0, \\ -1, & \text{ak } z < 0. \end{cases}$$

Označme kovariančnú maticu rankov

$$\Sigma_R = \begin{bmatrix} \Sigma_{R11} & \Sigma_{R12} & \ldots & \Sigma_{R1p} \\ \Sigma_{R21} & \Sigma_{R22} & \ldots & \Sigma_{R2p} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{Rp1} & \Sigma_{Rp2} & \ldots & \Sigma_{Rpp} \end{bmatrix}.$$

Diagonálne zložky Σ_{Rkk} predstavujú varianciu rankov k-tej zložky. Ako sme už vyššie uviedli, za platnosti nulovej hypotézy ranky R_η pochádzajú z rovnomerného rozdelenia na prvých N prirodzených číslach, z čoho vyplýva, že $\forall k = 1, 2, \ldots, p : \Sigma_{Rkk} = (N^2 - 1)/12$. Keďže kovariančná matica je symetrická, tak je potrebné odhadnúť iba polovicu matice. Pripomíname, že stredná hodnota $R_\eta^{(k)}$ je rovná $\mathrm{E}[R_\eta^{(k)}] = (N+1)/2$. Počítajme kovarianciu pre $k \neq l$:

$$\begin{aligned}
\Sigma_{Rkl} &= \operatorname{Cov}(R_\eta^{(k)}, R_\eta^{(l)}) \\
&= \mathrm{E}[R_\eta^{(k)} R_\eta^{(l)}] - \mathrm{E}[R_\eta^{(k)}]\,\mathrm{E}[R_\eta^{(l)}] \\
&= \mathrm{E}[R_\eta^{(k)} R_\eta^{(l)}] - \left(\frac{N+1}{2}\right)^2 \\
&= \mathrm{E}\left[\left[\frac{1}{2}\sum_{i=1}^{N} \operatorname{sgn}(X_\eta^{(k)} - X_i^{(k)}) + \frac{N+1}{2}\right]\left[\frac{1}{2}\sum_{j=1}^{N} \operatorname{sgn}(X_\eta^{(l)} - X_j^{(l)}) + \frac{N+1}{2}\right]\right] - \\
&\quad - \left(\frac{N+1}{2}\right)^2 \\
&= \frac{1}{4}\sum_{i=1}^{N}\sum_{j=1}^{N} \mathrm{E}\left[\operatorname{sgn}(X_\eta^{(k)} - X_i^{(k)}) \operatorname{sgn}(X_\eta^{(l)} - X_j^{(l)})\right] + \\
&\quad + \frac{N+1}{4}\left[\sum_{i=1}^{N} \mathrm{E}\left[\operatorname{sgn}(X_\eta^{(k)} - X_i^{(k)})\right] + \sum_{j=1}^{N} \mathrm{E}\left[\operatorname{sgn}(X_\eta^{(k)} - X_j^{(k)})\right]\right].
\end{aligned}$$

Všimnime si, že stredné hodnoty $\mathrm{E}\Big[\operatorname{sgn}(X_\eta^{(k)} - X_i^{(k)})\Big]$ a $\mathrm{E}\Big[\operatorname{sgn}(X_\eta^{(k)} - X_j^{(k)})\Big]$ sú nulové, pretože predpokladáme, že za platnosti nulovej hypotézy sú pozorovania X_η nezávislé a rovnako rozdelené, čo znamená, že $\Pr[X_\eta^{(k)} < X_i^{(k)}] = \Pr[X_\eta^{(k)} > X_i^{(k)}] = \frac{1}{2}$. Bez ujmy na všeobecnosti môžeme tiež ignorovať nulové sčítance, keď $\eta = i$ alebo $\eta = j$. Dospeli sme teda k výsledku

$$\operatorname{Cov}(R_\eta^{(k)}, R_\eta^{(l)}) = \frac{1}{4} \sum_{\substack{i=1 \\ i \neq \eta}}^{N} \sum_{\substack{j=1 \\ j \neq \eta}}^{N} \mathrm{E}\Big[\operatorname{sgn}(X_\eta^{(k)} - X_i^{(k)}) \operatorname{sgn}(X_\eta^{(l)} - X_j^{(l)}) \Big], \tag{13}$$

ktorý už nevieme presne vypočítať a musíme ho odhadnúť. Suma sa v princípe skladá z dvoch typov sčítancov:

$$c_1(k,l) = \mathrm{E}\Big[\operatorname{sgn}(X_\eta^{(k)} - X_i^{(k)}) \operatorname{sgn}(X_\eta^{(l)} - X_i^{(l)}) \Big], \text{ kde } \eta \neq i (= j),$$

$$c_2(k,l) = \mathrm{E}\Big[\operatorname{sgn}(X_\eta^{(k)} - X_i^{(k)}) \operatorname{sgn}(X_\eta^{(l)} - X_j^{(l)}) \Big], \text{ kde } \eta \neq i \neq j.$$

Oba sčítance $c_1(k,l)$ a $c_2(k,l)$ vieme pre fixné k, l odhadnúť prostredníctvom priemeru a kombinácií všetkých N pozorovaní X_η. Konkrétne pre odhad $c_1(k,l)$ je k dispozícii $N(N-1)/2$ kombinácií a $N(N-1)(N-2)/6$ kombinácií pre $c_2(k,l)$, teda:

$$\hat{c_1}(k,l) = \frac{2}{N(N-1)} \sum_{i<j}^{N} \operatorname{sgn}(X_i^{(k)} - X_j^{(k)}) \operatorname{sgn}(X_i^{(l)} - X_j^{(l)}),$$

$$\hat{c_2}(k,l) = \frac{6}{N(N-1)(N-2)} \sum_{i<j<z}^{N} \operatorname{sgn}(X_i^{(k)} - X_j^{(k)}) \operatorname{sgn}(X_i^{(l)} - X_z^{(l)}),$$

Výhodou je, že už aj pre relatívne malý počet dát N je k dispozícii nemalé množstvo kombinácií, čo by mohlo znamenať, že odhady prostredníctvom priemeru budú stabilné. Všimnime si, že $c_1(k,l)$ je v takom prípade korelačný koeficient známy ako Kendallovo τ medzi zložkami náhodného vektora $(X_\eta^{(k)}, X_\eta^{(l)})^T$. Odhad kovariancie zo vzťahu (13) vieme pri daných označeniach zapísať ako lineárnu kombináciu $c_1(k,l)$ a $c_2(k,l)$:

$$\hat{\Sigma}_{Rkl} = \widehat{\operatorname{Cov}(R_\eta^{(k)}, R_\eta^{(l)})} = \frac{N-1}{4} \hat{c_1}(k,l) + \frac{(N-1)(N-2)}{4} \hat{c_2}(k,l). \tag{14}$$

Takýmto spôsobom je možné odhadnúť celú kovariančnú maticu.

Rovnako ako vo všetkých prípadoch je žiaduce overiť pravdepodobnosť chyby prvého druhu takto vzniknutého testu, ktorý budeme pre ďalšie referencie označovať „MKWCE“, *viacrozmerný Kruskalov-Wallisov test s priamym Odhadom Kovariancie*. Tú sme odhadli prostredníctvom 10000 simulácií, ktorých výsledky uvádzame v tabuľke 13. Konštatujeme, že získané odhady nevykazujú nezanedbateľné prekročenie zvolenej hladiny významnosti $\alpha = 5\%$. Z druhého riadku tabuľky 13 je možné dospieť k záveru, že pre malý počet dát je test MKWCE konzervatívny pri zamietaní nulovej hypotézy, čo nie je nezvyčajné. Bohužiaľ, podobný efekt je prítomný aj pri zvýšenej dimenzii a v kombinácií s vyšším počtom súborov sa odhad pravdepodobnosti chyby prvého druhu blíži až k úrovni 3%. Okrem rastúcej konzervatívnosti je nevýhodou tohto testu v porovnaní s ostatnými uvedenými návrhmi taktiež rýchlejší nárast výpočtovej náročnosti odhadu kovariančnej matice, konkrétne kvadratický v dimenzii (v kovariančnej matici Σ_R je potrebné odhadnúť $p(p-1)/2$ zložiek) a kubický v počte dát (pri každom odhade Σ_{Rkl} sa využije rádovo $N(N-1)(N-2)$ kombinácií pozorovaní).

Tabuľka 13: Odhad pravdepodobnosti chyby prvého druhu testu MKWCE na hladine významnosti $\alpha = 5\%$ prostredníctvom 10000 simulácií pre I nezávislých vyvážených súborov, každý s n_i d-dimenzionálnymi realizáciami z príslušného rozdelenia.

	I	n_i	p	$\mathcal{N}_p(0_p, I_p)$	$\mathcal{B}_p$	$t_1(0_p, I_p)$	$t_3(0_p, I_p)$
MKWCE	3	30	2	0,049	0,050	0,046	0,049
	3	10	2	0,040	0,044	0,038	0,042
	3	30	5	0,041	0,042	0,042	0,046
	5	15	10	0,035	0,032	0,031	0,034

Zdroj: vlastné spracovanie

2.3.4 Viacrozmerný Kruskalov-Wallisov test s metódou bootstrap

V podkapitole 2.3.3 sme predstavili spôsob, akým je možné odhadnúť kovariančnú maticu po zložkách. Alternatívne je možné pre odhad kovariančnej matice Σ_R využiť napríklad štatistickú metódu *bootstrap* v jej klasickej forme uvedenej v publikácii B. Efrona [6]. Bootstrap sa využíva pri odhadovaní disperzie $D(T)$ odhadu T nejakého neznámeho parametra θ z pozorovaní $X_1, X_2, \ldots, X_N$ v prípadoch, keď známy odhad pre $D(T)$,

nie je dôveryhodný (napr. je dostupných málo pozorovaní, no odhad platí iba asymptoticky) alebo vzorec pre disperziu $D(T)$ je neznámy. Hlavná myšlienka tejto metódy spočíva v generovaní B sád zložených z N pozorovaní získaných z pôvodnej sady prostredníctvom rovnomerne náhodných výberov s vrátením a vypočítavaní odhadu T pre každú takúto sadu (získame tak $T_1, T_2, \ldots, T_B$). Potom odhad pre disperziu $D(T)$ bude výberová variancia z vypočítaných T_b, t. j.

$$D(T) = \frac{1}{B-1}\sum_{b=1}^{B}(T_b - \overline{T})^2, \text{ kde } \overline{T} = \frac{1}{B}\sum_{b=1}^{B} T_b.$$

Efron ukázal, že pokiaľ je počet sád B dostatočne veľký, tak odhad $D(T)$ je dôveryhodný. Bootstrap metóda je jednoducho rozšíriteľná pre viacrozmerné odhady T a odhad kovariančnej matice pre T.

V našej aplikácii využijeme bootstrap práve vo viacrozmernej forme. Označme X_η^b η-té pozorovanie a R_η^b pozložkový rank pre η-té pozorovanie v b-tom náhodnom výbere. Na základe $B = 1000$ náhodných výberov s vrátením N pozorovaní $X_\eta \in \Re^p$ zo združeného výberu navrhujeme odhadnúť kovariančnú maticu Σ_R. Navrhujeme, aby rolu T_b zastával rank R_η^b pre nejaké fixné η. Potom odhad kovariančnej matice rozdelenia vektora R_η^b pre nejaké fixné η vieme vypočítať vzťahom

$$\hat{\Sigma}_{R_\eta} = \frac{1}{B-1}\sum_{b=1}^{B}(R_\eta^b - \overline{R^b})(R_\eta^b - \overline{R^b})^T.$$

Poukazujeme na fakt, že dva odhady pre hocijaké η a η^* nie sú nutne rovnaké, ale sú medzi sebou rovnocenné v zmysle, že neexistuje η^*, pre ktoré by bol odhad $\hat{\Sigma}_{R_{\eta^*}}$ matice Σ_{R_η} v nejakom zmysle lepší ako $\hat{\Sigma}_{R_\eta}$. Navrhujeme preto vypočítať všetky odhady kovariančných matíc $\hat{\Sigma}_{R_\eta}$ pre $\eta = 1, 2, \ldots, N$ a výsledný odhad kovariančnej matice rankov vypočítať ako priemer z týchto matíc

$$\begin{aligned}\hat{\Sigma}_R &= \frac{1}{N}\sum_{\eta=1}^{N}\hat{\Sigma}_{R_\eta} \\ &= \frac{1}{N(B-1)}\sum_{\eta=1}^{N}\sum_{b=1}^{B}(R_\eta^b - \overline{R_\eta})(R_\eta^b - \overline{R_\eta})^T. \qquad (15)\end{aligned}$$

Test využívajúci takýto odhad kovariančnej matice Σ_R pre ďalšie referencie označujeme skratkou „MKWB“, *viacrozmerný Kruskalov-Wallisov test s Bootstrapom.*

Navrhovaný algoritmus MKWB:

1. Určenie pozložkových rankov $R_\eta^{(k)}$ pozorovaní $X_\eta^{(k)}$ pre $\eta = 1, 2, \ldots, N$.
2. Odhad kovariančnej matice Σ_R prostredníctvom metódy bootstrap a vzťahu (15).
3. Výpočet testovej štatistiky (10) a porovnanie s príslušným kvantilom $\chi^2_{p(I-1)}$ rozdelenia.

Pre overenie vhodnosti testu a predpokladu neporušenia chí-kvadrátovosti testovej štatistiky pri takto odhadnutej kovariančnej matici sme simulovali odhad pravdepodobnosti chyby prvého druhu. Výsledky uvádzame v tabuľke 14. Opäť pozorujeme, že odhad pravdepodobnosti chyby prvého druhu neprekračuje zvolenú hladinu významnosti $\alpha = 5\%$ a pri zmenšovaní počtu pozorovaní alebo pri zvyšujúcej sa dimenzii opäť dochádza k stúpaniu konzervatívnosti testu. Každopádne aj tento návrh testu je vhodný pre testovanie nulovej hypotézy za platnosti spomenutých predpokladov.

Tabuľka 14: Odhad pravdepodobnosti chyby prvého druhu testu MKWB na hladine významnosti $\alpha = 5\%$ prostredníctvom 10000 simulácií pre I nezávislých vyvážených súborov, každý s n_i d-dimenzionálnymi realizáciami z príslušného rozdelenia.

	I	n_i	p	$\mathcal{N}_p(0_p, I_p)$	$\mathcal{B}_p$	$t_1(0_p, I_p)$	$t_3(0_p, I_p)$
MKWB	3	30	2	0,049	0,049	0,044	0,050
	3	10	2	0,044	0,041	0,042	0,042
	3	30	5	0,044	0,046	0,045	0,040
	5	15	10	0,035	0,035	0,030	0,031

Zdroj: vlastné spracovanie

Na záver kapitoly ešte overíme, či aj testy MKWOR, MKWCE a MKWB rovnako ako DecorMKW sú z hľadiska pravdepodobnosti chyby prvého druhu vhodné aj pre eliptícky symetrické pozorovania, kde naivné zovšeobecnenie označené ako MKW zlyhá. V tabuľke 15, ktorá je rozšírením tabuľky 11, uvádzame výsledky simulácií pravdepodobnosti

chyby prvého druhu pre všetky testy uvedené v podkapitole 2.3 pre homoskedastické pozorovania z príslušného elipticky symetrického rozdelenia. Pozorujeme, že všetky navrhnuté prístupy okrem naivného MKW nevykazujú pri takýchto dátach problémy.

Tabuľka 15: Odhad pravdepodobnosti chyby prvého druhu na hladine významnosti $\alpha = 5\%$ prostredníctvom 10000 simulácií pre $I = 3$ nezávislé vyvážené súbory, každý s $n_i = 30$ dátami rozmeru $p = 2$ z príslušného rozdelenia s kovariančnou maticou $\Sigma = \begin{bmatrix} 10 & 3 \\ 3 & 1 \end{bmatrix}$

	$\mathcal{N}_2(0_2, \Sigma)$	$t_1(0_2, \Sigma)$	$t_3(0_2, \Sigma)$
MKW	0,089	0,083	0,089
DecorMKW	0,049	0,056	0,048
MKWOR	0,048	0,045	0,050
MKWCE	0,047	0,041	0,047
MKWB	0,041	0,038	0,043

Zdroj: vlastné spracovanie

3 Porovnanie navrhovaných metód

V kapitole 2 sme predstavili niekoľko možných rozšírení jednorozmerných štatistických testov polohy viac ako dvoch súborov pre viacrozmerné pozorovania. Keďže sa v tejto kapitole často odvolávame na testy prostredníctvom ich skratiek, tak v tabuľke 16 uvádzame pre prehľad zoznam všetkých našich navrhnutých testov aj s celým názvom.

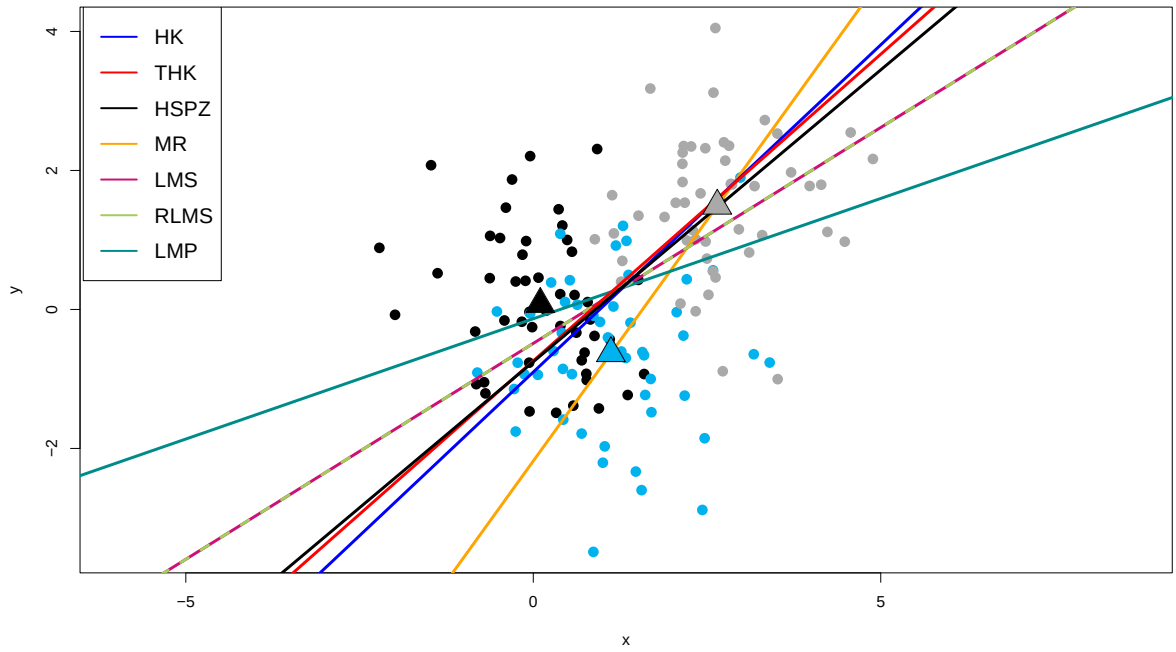
Tabuľka 16: Prehľad skratiek navrhnutých testov.

Skratka	Názov
BK-VP	test s Bonferroniho Korekciou - Všetky Páry
BK-JO	test s Bonferroniho Korekciou - Jeden vs. Ostatné
MPH	test s Minimálnou P-Hodnotou
MR	test s Maximálnym Rozptylom
LMS	test s Lineárnym Modelom daným Stredmi
LMP	test s Lineárnym Modelom daným Pozorovaniami
RLMS	test s Rekurentným Lineárnym Modelom daným Stredmi
HK	test s Hlavnými Komponentami
RHK	Robustný test s Hlavnými Komponentami
KRHK	Korigovaný Robustný test s Hlavnými Komponentami
HSPZ	test s Hlavným Smerom Priestorových Znamienok
DecorMKW	dekorelovaný viacrozmerný Kruskalov-Wallisov test
MKWOR	viacrozmerný Kruskalov-Wallisov test s Ojovými Rankami
MKWCE	viacrozmerný Kruskalov-Wallisov test s priamym Odhadom Kovariancie
MKWB	viacrozmerný Kruskalov-Wallisov test s Bootstrapom

Zdroj: vlastné spracovanie

Na obrázku 10 uvádzame príklad vizuálneho porovnania priamok, na ktoré sme v jednotlivých prístupoch projektovali pozorovania. Prístupy, kde sa využívalo viacero priamok, sme na obrázku 10 neuvádzali. Pozorujeme mierne odlišnosti naprieč prístupmi. Iba v prípade LMS a RLMS sú priamky identické, čo je v $\Re^2$ nastáva vždy, pretože sa v RLMS vykoná iba jedna iterácia, ktorá je identická s prístupom LMS. Obrázok 10

má však čisto ilustratívny charakter. Porovnanie kvality metód z hľadiska simulačného odhadu ich sily, ktorý ponúka lepšiu predstavu o kvalitách testov, uvádzame v nasledujúcej podkapitole.



Obr. 10: Príklad vizuálneho porovnania prístupov HK, RHK, HSPZ, MR, LMS, RLMS a LMP. Trojuholníky označujú stredy $I = 3$ súborov. Každý súbor obsahuje 50 pozorovaní, ktoré pochádzajú z dvojrozmerného normálneho rozdelenie s identickou kovariančnou maticou a so strednými hodnotami $\mu_1 = (0; 0)^T$, $\mu_2 = (1; -0, 6)^T$ a $\mu_3 = (2, 8; 1, 4)^T$. (*Zdroj: vlastné spracovanie*)

3.1 Simulačné odhady sily

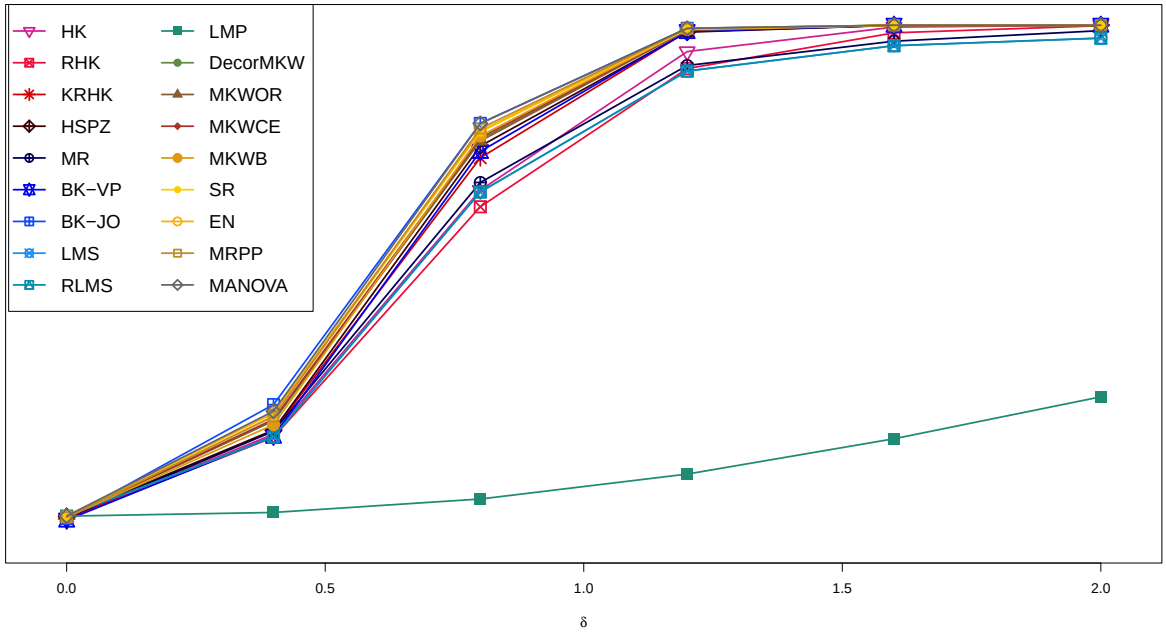
Sila testu, je definovaná ako pravdepodobnosť, že test hypotézu H_0 zamietne, ak táto hypotéza neplatí. V našom prípade to znamená, že aspoň jeden súbor je vychýlený a test to odhalí. Na obrázku 11 uvádzame porovnanie navrhovaných metód pre $I = 3$ nezávislé vyvážené súbory. Každý súbor obsahuje $n_i = 30$ pozorovaní z dvojrozmerného (čiže $p = 2$) normálneho rozdelenia s identickou kovariančnou maticou. Na vodorovnej osi na obrázku 11 je znázornené číslo δ určujúce vychýlenie prvého z troch súborov v smere

osi druhej dimenzie, t. j. vychýlenia dát v prvom súbore postupne o

$$\begin{bmatrix} 0 \\ \delta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0,4 \end{bmatrix}, \begin{bmatrix} 0 \\ 0,8 \end{bmatrix}, \begin{bmatrix} 0 \\ 1,2 \end{bmatrix}, \begin{bmatrix} 0 \\ 1,6 \end{bmatrix}, \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

Odhad sily sme pre každý z testov vo všetkých prípadoch vychýlenia prvého zo súborov simulovali pomocou 10000 simulácií. Výsledky simulácií uvádzame v tabuľke 17. Pri testoch MR, HSPZ, RLMS a LMS sme využili ich permutačné varianty s 1000 permutáciami. Taktiež sme využili aj parametrický prístup MANOVA, ktorý sa špecializuje práve na pozorovania z normálneho rozdelenia, aby sme mali predstavu, ako si viedli naše návrhy oproti tejto špecializovanej metóde. Do porovnania sme doplnili aj test využívajúci koncept priestorového usporiadania z kapitoly 2.3 publikácie [17] s testovou štatistikou označenou ako vzťah (3) na strane 600 a jeho implementáciu v jazyku R [20] v knižnici *SpatialNP* [23]. Tento test sme v nasledujúcich častiach označili skratkou SR. V porovnaní taktiež uvádzame neparametrický test od autorov G. Székely a M. Rizzo z publikácie [22] využívajúc jeho implementáciu v knižnici *energy* [26], ktorý sme označili skratkou EN. Tento test vo svojom algoritme využíva neparametrický bootstrap, pre ktorý sme nastavili počet sád na 1000. Do porovnania sme tiež pridali neparametrický test nazývaný „Multiple Response Permutation Procedure“, skrátene MRPP, prostredníctvom jeho implementácie v knižnici *vegan* [18]. Procedúra testu MRPP je bližšie popísaná napríklad v knihe [16]. Oba testy MRPP a EN testujú okrem rozdielov v polohe súborov aj rozdiely v rozptýlení. V uvedenom porovnaní sa súbory nelíšia v rozptýlení, takže aj tieto testy sa v takom prípade stávajú, rovnako ako ostatné uvedené, testami o parametre polohy.

Na obrázku 11 pozorujeme, že pre pozorovania pochádzajúce z viacrozmerného normálneho rozdelenia sú odhadnuté silofunkcie, až na metódu LMP, blízko seba a rozdiely je v jednotlivých testoch náročné vizuálne rozoznať. Metóda LMP výrazne zaostáva za ostatnými, čo je spôsobené slabou kolísavosťou jej „optimálnej“ priamky pre projekciu. Je však nutné uviesť, že sme simulovali silu pre taký prípad, že rozdiel v polohe súborov bol v smere premennej modelovanej regresným modelom v algoritme metódy, čo je prípad, ktorý nevie lineárny model namodelovať, keďže β z lineárneho modelu by musela v takomto prípade byť ∞. Ostatné prístupy si s takouto situáciou poradili lepšie. Zvyšné testy by sme mohli rozdeliť do troch skupín. Najhoršie si viedla trojica založená na rozptyle v jednom smere HK, RHK a MR, čo sa môže zdať prekvapivé, keďže rozdiel v polohe



Obr. 11: Vizuálne porovnanie odhadnutých silofunkcií pre $I = 3$ nezávislé vyvážené súbory, každý s $n_i = 30$ pozorovaniami z $\mathcal{N}_2(\mu_i, I_2)$, kde $\mu_1 = (0; \delta)^T$ a $\mu_2 \equiv \mu_3 = (0; 0)^T$. (*Zdroj: vlastné spracovanie*)

súborov bol práve v jednom smere. Dalo by sa povedať, že robustná verzia HK bola z tejto trojice najslabšia, no jej sila tkvie hlavne pri dátach s outliermi, ktorých pri normálnom rozdelení nie je veľa. Do druhej skupiny by sa dali zaradiť prístupy KRHK, HSPZ a BK-VP. Pozorujeme, že úprava v algoritme RHK s potenciálnym využitím viacerých hlavných komponentov a následná korekcia nezanedbateľne zvýšili silu. Prístup HSPZ, ku ktorému sme dospeli s cieľom pokryť slabiny testov HK, RHK a KRHK, si pri pozorovaniach z normálneho rozdelenia viedol približne rovnako dobre ako KRHK. Prístup BK-VP sa ukázal byť horší ako prístup BK-JO. Výhoda BK-JO bola spôsobená primárne návrhom experimentu pre meranie sily, pretože rozdiel v polohe sme vykonali v prvom súbore a prístup BK-JO skúma rozdiely práve prvého súboru s ostatnými. Teda BK-JO vyhodnocoval test na základe výsledkov dvojsúborových testov medzi súbormi $\{1, 2\}$ a $\{1, 3\}$, kde bol najväčší potenciál zamietnuť nulovú hypotézu, zatiaľ čo prístup BK-VP bral do úvahy aj kombináciu $\{2, 3\}$, kde rozdiel v polohe nebol, čo negatívne ovplyvnilo silu tohto testu. V najsilnejšej skupine sa ukázali byť testy MANOVA, MRPP, EN, SR, BK-JO a testy vychádzajúce z viacrozmerného zovšeobecnenia Kruskalovej-Wallisovej štatistiky uvedené

Tabuľka 17: Simulačný odhad sily prístupov pre pozorovania z dvojrozmerného normálneho rozdelenia.

δ	0	0,4	0,8	1,2	1,6	2
MR	0,048	0,215	0,696	0,922	0,969	0,990
HK	0,047	0,209	0,681	0,949	0,996	1
RHK	0,046	0,203	0,649	0,916	0,985	0,999
KRHK	0,045	0,216	0,744	0,988	1	1
BK-JO	0,047	0,266	0,810	0,994	1	1
BK-VP	0,043	0,205	0,756	0,988	1	1
LMS	0,051	0,204	0,678	0,911	0,960	0,975
RLMS	0,051	0,204	0,678	0,911	0,960	0,975
LMP	0,051	0,058	0,084	0,132	0,200	0,281
HSPZ	0,050	0,217	0,766	0,987	1	1
DecorMKW	0,050	0,235	0,779	0,991	1	1
MKWB	0,049	0,228	0,786	0,992	1	1
MKWCE	0,049	0,238	0,782	0,991	1	1
MKWOR	0,048	0,236	0,776	0,990	1	1
SR	0,050	0,244	0,794	0,992	1	1
EN	0,052	0,250	0,796	0,993	1	1
MRPP	0,048	0,244	0,800	0,994	1	1
MANOVA	0,051	0,254	0,810	0,994	1	1

Zdroj: vlastné spracovanie

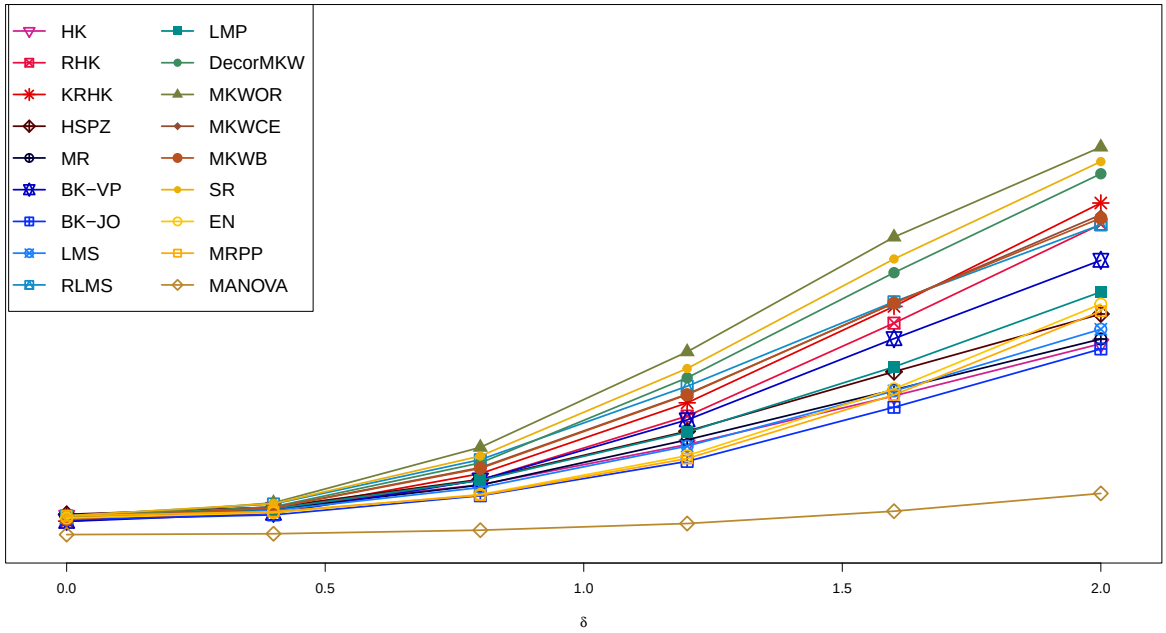
v podkapitole 2.3. Podľa očakávaní MANOVA ako špecializovaný test na normálne dáta bol medzi najsilnejšími. Ako sme už spomenuli, navrhnutý test BK-JO bol v takomto nastavení vo výhode, no ukázalo sa, že za takýchto podmienok vie byť ohľadom sily na úrovni špecializovaného prístupu MANOVA. Najsilnejšiu trojicu doplnil test SR, ktorý sa tiež priblížil k úrovni sily prístupu MANOVA. Tesne za testom SR, boli približne na jednej úrovni testy DecorMKW, MKWB, MKWCE a MKWOR. Môžeme vyhodnotiť, že pre pozorovania pochádzajúce z viacrozmerného normálneho rozdelenia je MANOVA najlepší prístup, no jeho dominancia je obmedzená práve na toto rozdelenie. Preto sme

odhad sily simulovali aj pre viacrozmerné Studentovo $t_1(0_3, I_3)$ rozdelenie. Keďže marginálne rozdelenia tohto rozdelenia majú nekonečnú disperziu, očakávali sme väčší počet „outlierov“. Cieľom bolo zistiť, akú silu majú navrhované prístupy pre takéto rozdelenie. V nasledujúcich simuláciách sme odhadovali silu testov pre $I = 4$ nezávislé vyvážené súbory, každý súbor s $n_i = 20$ pozorovaniami rozmeru $p = 3$. Aby sme eliminovali výhodu prístupu BK-JO, menili sme v tomto prípade polohu druhého súboru a tiež z dôvodu eliminácie spomenutých problémov v prístupe LMP sme vychyľovali pozorovania v prvej zložke druhého súboru, postupne o

$$\begin{bmatrix} \delta \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0,4 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0,8 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1,2 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1,6 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}.$$

Odhad sily sme pre každý z testov vo všetkých prípadoch vychýlenia druhého zo súborov simulovali pomocou 10000 simulácií. Presné výsledky simulácií uvádzame v tabuľke 18 a porovnanie odhadnutých silofunkcií spoločne s relatívnym porovnaním medzi skúmanými testami uvádzame na obrázku 12.

V tomto prípade pozorujeme väčšie rozdiely medzi jednotlivými prístupmi. Môžeme zhodnotiť, že MANOVA je podľa očakávaní pre pozorovania zo Studentovho t_1 rozdelenia nepoužiteľná. Pozorujeme, že ostatné prístupy by sa dali v tomto prípade rozdeliť opäť do troch skupín. Najhoršie dopadli prístupy v poradí BK-JO, HK, MR, LMS, HSPZ, MRPP, EN, LMP a BK-VP. V tomto príklade sa preukázala slabosť BK-JO, ktorá spočíva v usporiadaní súborov, čo malo veľký dopad na silu tohto testu. Tieto prístupy taktiež ovplyvnil veľký počet outlierov. Tie v prístupoch BK-JO, MR, LMS, RLMS, HSPZ a BK-VP negatívne vplývajú okrem iného aj na priemery, ktoré sa v týchto prístupoch využívajú na odhad stredov, pretože priemer nie je robustný odhad. V ďalšej časti uvádzame, ako by si niektoré z týchto prístupov viedli, ak by sme pre odhad stredov súborov využili priestorový medián. Aj pri testoch EN a MRPP pozorujeme značnú konzervatívnosť, ktorú sme pri týchto neparametrických testoch neočakávali. Do druhej skupiny by sme mohli zaradiť testy RLMS, RHK, MKWB, MKWCE a KRHK. Konštatujeme, že prístup RLMS si poradil najlepšie z prístupov, ktoré využívajú vo svojich algoritmoch odhad stredov súborov a navrhnutý prístup postupného znižovania dimenzie ukázal svoje výhody oproti testu LMS. Oba prístupy RHK a KRHK robustne odhadli svoje optimálne priamky a vý-



Obr. 12: Vizuálne porovnanie odhadnutých silofunkcií pre $I = 4$ nezávislé vyvážené súbory, každý s $n_i = 20$ pozorovaniami z $t_1(\mu_i, I_3)$, kde $\mu_2 = (\delta; 0; 0)^T$ a $\mu_1 \equiv \mu_3 = (0; 0; 0)^T$. (*Zdroj: vlastné spracovanie*)

hodu, ktorú im to prinieslo v porovnaní s HK, nie je možné zanedbať. Prístupu KRHK sa v porovnaní s RHK podarilo z ďalších dimenzií „vydolovať“ dodatočné informácie, čo sa prejavilo väčšou silou. Prístupy MKWB a MKWCE, ktoré odhadujú kovariančnú maticu rankov, si viedli v tomto prípade horšie ako prístupy DecorMKW a MKWOR. Do najlepšej skupiny by sme mohli zaradiť prístupy DecorMKW, SR a MKWOR. Konštatujeme, že prístupy využívajúce Ojove ranky sú očakávane najlepšie, no teší nás, že metóda DecorMKW veľmi nezaostáva za našim „benchmarkom“ akým je test SR od popredných štatistikov v odvetví viacrozmerných testov polohy viacerých súborov. Najviac nás však teší test MKWOR, ktorý v takejto situácii test SR nezanedbateľne porazil.

Tabuľka 18: Simulačný odhad sily pre pozorovania z trojrozmerného Studentovho $t_1(0_3, I_3)$ rozdelenia.

δ	0	0,4	0,8	1,2	1,6	2
MR	0,049	0,064	0,111	0,198	0,295	0,393
HK	0,053	0,068	0,111	0,189	0,283	0,384
RHK	0,047	0,062	0,121	0,244	0,424	0,615
KRHK	0,041	0,060	0,133	0,270	0,456	0,656
BK-JO	0,044	0,053	0,090	0,157	0,261	0,374
BK-VP	0,041	0,056	0,121	0,237	0,394	0,545
LMS	0,049	0,062	0,106	0,186	0,293	0,412
RLMS	0,051	0,074	0,160	0,302	0,465	0,614
LMP	0,048	0,065	0,120	0,213	0,340	0,484
HSPZ	0,054	0,068	0,122	0,215	0,33	0,442
DecorMKW	0,048	0,069	0,154	0,318	0,522	0,713
MKWB	0,044	0,067	0,143	0,286	0,463	0,627
MKWCE	0,045	0,067	0,145	0,287	0,463	0,633
MKWOR	0,045	0,077	0,184	0,368	0,590	0,764
SR	0,051	0,075	0,167	0,336	0,548	0,736
EN	0,047	0,057	0,092	0,168	0,297	0,461
MRPP	0,051	0,059	0,091	0,162	0,285	0,447
MANOVA	0,015	0,017	0,024	0,036	0,060	0,095

Zdroj: vlastné spracovanie

3.2 Analýza vplyvu priestorového mediánu na silu vybraných testov

Ako sme spomenuli v podkapitole 3.1, pri simulácií sily viaceré prístupy trpeli pri pozorovaniach z viacrozmerného Studentovho t_1 rozdelenia z dôvodu nerobustnosti priemerov v prítomnosti outlierov. Pri definíciách algoritmov testov BK-JO, MR, LMS, RLMS, HSPZ a BK-VP spomíname, že na odhad stredov využívame priemer, no vo všetkých prípadoch je bez ujmy na všeobecnosti možné ako odhad stredu súboru využiť namiesto priemeru ľubovoľný odhad stredu ako napríklad priestorový medián. Rozhodli sme sa preto zanalyzovať, aký vplyv má na silu testu zmena z priestorového priemeru na priestorový medián. V práci sme využili jeho implementáciu v jazyku R [20] v knižnici *SpatialNP* [23]. Rovnako ako v podkapitole 3.1 sme simulovali silu pre pozorovania z trojrozmerného Studentovho $t_1(0_3, I_3)$ rozdelenia, $I = 4$ nezávislé vyvážené súbory, každý súbor s $n_i = 20$ realizáciami a rovnako sme postupne vychyľovali druhý súbor o $(\delta; 0; 0)^T$ pre $\delta = 0; 0,4; 0,8; 1,2; 1,6; 2$. Výsledky simulácií sily vybraných testov uvádzame v tabuľke 19. Pre krátkosť času sme analyzovali iba prístupy MR a HSPZ. V oboch prípadoch sme pozorovali nezanedbateľný nárast sily testov v porovnaní s ich variantami využívajúcimi priemer. Mediánový prístup HSPZ je dokonca v tomto prípade na úrovni sily Ojovho a Randlesovho testu SR a nášho MKWOR. Na základe výsledkov simulácií konštatujeme, že vplyv mediánu výrazne zvýšil robustnosť prístupov, dôsledkom čoho bolo zvýšenie sily.

Tabuľka 19: Simulačný odhad sily vybraných navrhovaných prístupov s využitím priestorového mediánu pre pozorovania z trojrozmerného Studentovho $t_1(0_3, I_3)$ rozdelenia.

δ	0	0,4	0,8	1,2	1,6	2
Mediánový MR	0,054	0,071	0,135	0,245	0,371	0,478
Mediánový HSPZ	0,049	0,074	0,167	0,351	0,567	0,739

Zdroj: vlastné spracovanie

Záver

Oblasť viacrozmerných neparametrických testov polohy je v posledných desaťročiach aktívna, keďže, ako sme v úvode uviedli, v mnohých situáciách zatiaľ neboli objavené tie najlepšie a najsilnejšie testy. Cieľ, ktorý sme sa v práci pokúsili splniť, bol preskúmať oblasť viacrozmerných testov polohy viacerých súborov a navrhnúť testy, ktoré by sa svojou kvalitou priblížili známym testom v spomínanej oblasti.

V prvej kapitole sme definovali pravdepodobnostné rozdelenia, ktoré sme naprieč prácou využívali. Taktiež sme stručne spomenuli jednorozmerné testy polohy, z ktorých sme v ďalších kapitolách vychádzali a vysvetlili sme princíp permutačných testov. Po predstavení problematiky viacrozmerných testov polohy viacerých súborov sme v podkapitolách 2.2 a 2.3 predstavili naše návrhy testov. V podkapitole 2.2 sme k problematike zaujali postoj redukcie dimenzie dát na priamku a následné využitie jednorozmerného Kruskalovho-Wallisovho testu. V podkapitole 2.2.1 sme vychádzajúc z prác [24, 27] navrhli štyri testy (BK-VP, BK-JO, MR a MPH) využívajúce pre odhad optimálnej priamky kombinácie stredov súborov. Test MPH sa však pri simuláciách pravdepodobnosti chyby prvého druhu ukázal byť nevyhovujúci, no nepodarilo sa nám odhaliť pôvod jeho problémov. Ďalej v podkapitole 2.2.2 sme predstavili tri návrhy testov (LMP, LMS a RLMS), ktoré využívali pre odhad optimálnej priamky metódu lineárnej regresie. Ukázali sme, že test LMP je robustnejší z pohľadu sklonu priamky a pre jeho využitie nebolo potrebné aplikovať permutačný test. Avšak v istých prípadoch sa ukázala byť táto vlastnosť obmedzujúca, čo následne malo vplyv aj na kvalitu testu. Pre testy LMP a LMS taktiež existujú prípady, pre ktoré sú tieto testy takpovediac „slepé“ a vôbec neodhalia neplatnosť nulovej hypotézy. V snahe vyriešiť tento problém sme navrhli test RLMS, ktorý redukuje dimenziu dát prostredníctvom lineárneho modelu postupne, až kým doiteruje k jednorozmernej reprezentácii pozorovaní. Konštatujeme, že aj naprieč snahe vylepšiť testy uvedené v podkapitole 2.2.2 zostávajú prípady, pre ktoré nie sú tieto testy vhodné. V podkapitole 2.2.3 sme v navrhnutých testoch pre odhad optimálnej jednorozmernej reprezentácie dáta využili metódu hlavných komponentov, ako aj jednu z jej robustných verzií. Taktiež sme v prístupe HSPZ preskúmali kombináciu nástrojov akými sú priestorové znamienko a analýza hlavných komponentov. Táto podkapitola obsahuje dokopy štyri

návrhy testov (HK, RHK, KRHK a HSPZ). V podkapitole 2.3 sme sa pokúsili priamo zovšeobecniť testovú štatistiku jednorozmerného Kruskalovho-Wallisovho testu, z čoho vyplynul problém odhadu kovariancie rankov. Postupne sme v tejto podkapitole predstavili štyri naše návrhy testov (DecorMKW, MKWOR, MKWCE a MKWB). Pokus zovšeobecniť Kruskalovu-Wallisovu testovú štatistiku sa ukázal byť menej originálny ako projekcia do jednorozmerného priestoru, keďže takýchto pokusov sme našli v dostupnej literatúre niekoľko, no nepodarilo sa nám dopátrať k testu, ktorý by postupoval identicky ako hociktorý z našich návrhov.

Tretiu kapitolu sme venovali porovnaniu sily navrhnutých testov s niektorými alternatívami, ako napríklad MANOVA, test založený na ε-štatistike, test využívajúci MRPP procedúru, ale aj test s priestorovým usporiadaním od popredných štatistikov v skúmanej oblasti, H. Oju a R. H. Randlesa. Z výsledkov simulácií uvedených v podkapitole 3.1 sa dá konštatovať, že testy využívajúce projekciu na priamku navrhnuté v podkapitole 2.2 sú menej kvalitné ako testy, ktoré zovšeobecňujú Kruskalovu-Wallisovu testovú štatistiku. Rozdiel sa ukázal byť hlavne pri pozorovaniach z viacrozmerného Studentovho t_1 rozdelenia, čo bolo sčasti zapríčinené nerobustnosťou odhadov stredov v jednotlivých testoch. Testy však boli navrhnuté tak, aby sme bez ujmy na všeobecnosti vedeli odhad stredu vymeniť za robustný. Vplyv takejto zámeny sme skúmali v podkapitole 3.2, kde sme ukázali, že táto zámena nezanedbateľne zvýšila silu. V prípade testu HSPZ až na úroveň najsilnejších testov v celej štúdii.

Práca dokopy poskytuje 15 našich návrhov viacrozmerných testov polohy viacerých súborov, pričom môžeme konštatovať, že testy MKWOR a mediánový HSPZ sa z pohľadu sily dokázali vyrovnať testu označenému ako SR od Oju a Randlesa, ba dokonca ho v uvedenom prípade aj prekonať, čím považujeme ciele diplomového výskumu za splnené.

Zoznam použitej literatúry

[1] Anděl, J.: *Základy matematické statistiky.* Univerzita Karlova v Praze, Matematicko-fyzikální fakulta, 2002

[2] Bhapkar, V. P.: *Some non-parametric test for the multivariate several sample location problem.* Proc. International Symp. Multivariate Analysis. Dayton, Ohio, 1965

[3] Bhattacharya, I., Ghosal, S.: *Bayesian nonparametric tests for multivariate locations.* Journal of Statistical Planning and Inference, Vol. 219, 2022, 1-12

[4] Chatterjee, S. K., Sen, P. K.: *A multisample nonparametric scale test based on U-statistics.* Calcutta Stat. Assoc. Bull., Vol. 15, 1966, 109-120

[5] Choi, K., Marden, J.: *An Approach to Multivariate Rank Tests in Multivariate Analysis of Variance.* Journal of the American Statistical Association, Vol. 92(440), 1997, 1581–1590

[6] Efron, B.: *Bootstrap methods: Another look at the jackknife.* The Annals of Statistics, Vol. 7(1), 1979, 1–26

[7] Filová, L., Szűcs, G.: *Viacrozmerné štatistické analýzy: poznámky k prednáškam.* Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, 2021

[8] Genz, A., et al.: *mvtnorm: Multivariate Normal and t Distributions.* R package version 1.1-3, 2021, `https://CRAN.R-project.org/package=mvtnorm`

[9] Harman, R., Rosa, S.: *Stochastické simulačné metódy: poznámky k prednáškam.* Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, 2019

[10] Holm, S.: *A Simple Sequentially Rejective Multiple Test Procedure.* Scandinavian Journal of Statistics, Vol. 6, 1979, 65-70

[11] Hotelling, H.: *Analysis of a complex of statistical variables into principal components.* Journal of Educational Psychology, Vol. 24, 1933, 417-441

[12] Hubert, M., Rousseeuw, P. J., Branden, K. V.: *ROBPCA: A New Approach to Robust Principal Component Analysis.* Technometrics, Vol. 41, 2005, 64-79

[13] Izenman, A. J.: *Modern Multivariate Statistical Techniques.* Springer, 2008, 195-215

[14] Kruskal, W, H., Wallis, W. A.: *Use of Ranks in One-Criterion Variance Analysis.* Journal of the American Statistical Association, Vol. 47, 1952, 583–621

[15] Mann, H. B., Whitney, D. R.: *On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other.* The Annals of Mathematical Statistics, Ann. Math. Statist., Vol. 18, 1947, 50-60

[16] Mielke, P. W., Berry, K. J.: *Permutation Methods: A Distance Function Approach.* Springer Series in Statistics, Springer, 2001

[17] Oja, H., Randles, R. H.: *Multivariate Nonparametric Tests.* Statistical Science, Vol. 19, 2004, 598-605

[18] Oksanen, J., et al.: *vegan: Community Ecology Package.* R package version 2.6-4, 2022, `https://CRAN.R-project.org/package=vegan`

[19] Puri, M. L., Sen, P. K.: *On a Class of Multivariate Multisample Rank-Order Tests.* Sankhyā: The Indian Journal of Statistics, Series A, Vol. 28, 1966, 353-376

[20] R Core Team: *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, 2021, `http://www.R-project.org/`

[21] Reynkens, T.: *rospca: Robust Sparse PCA using the ROSPCA Algorithm.* R package version 1.0.4, 2018, `https://CRAN.R-project.org/package=rospca`

[22] Székely, G., Rizzo, M.: *Testing for equal distributions in high dimension.* InterStat, 2004

[23] Sirkia, S., et al. : *SpatialNP: Multivariate Nonparametric Methods Based on Spatial Signs and Ranks.* R package version 1.1-5, 2021, `https://CRAN.R-project.org/package=SpatialNP`

[24] Somogyi, P.: *Viacrozmerné neparametrické testy polohy dvoch súborov.* Diplomová práca, Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, 2021

[25] Svetlíková, B.: *Rôzne spôsoby testovania v MANOVA.* Diplomová práca, Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, 2015

[26] Rizzo, M., Szekely, G.: *energy: E-Statistics: Multivariate Inference via the Energy of Data.* R package version 1.7-11, 2022, `https://CRAN.R-project.org/package=energy`

[27] Wilcox, R. R.: *A Multivariate Projection-type Analogue of the Wilcoxon-Mann-Whitney Test.* The British Journal of Mathematical and Statistical Psychology, Vol. 57, 2004, 205-213

[28] Wilcoxon, F.: *Individual Comparisons by Ranking Methods.* Biometrics Bulletin, Vol. 1, 1945, 80-83

Prílohy

A Kód v jazyku R

A.1 Pomocné funkcie

```
################################################################################
# R version: 4.1                                                               #
# Libraries: mvtnorm, rospca, SpatialNP, energy, vegan                         #
################################################################################

### groupMeans #################################################################

#' @title groupMeans
#' @description Vypocíta priemery po skupinách
#'
#' @param data_matrix Matica dát, kazdý riadok predstavuje jeden bod.
#' @param group Vektor oznacujúci skupinu pre kazdý riadok matice
#'        \code{data_matrix}.
#' @return Matica priemerov \code{data_matrix} po skupinách. Kazdý riadok
#'         predstavuje viacrozmerný priemer.
groupMeans <- function(data_matrix, group) {
  unique_groups <- unique(group)
  g_means <-
    matrix(0,
           ncol = ncol(data_matrix),
           nrow = length(unique_groups))
  for (g in seq_along(unique_groups)) {
    g_means[g, ] <- colMeans(data_matrix[group == unique_groups[g], ])
  }
  return(g_means)
}
```

```
### groupMedians ###################################################################

#' @title groupMedians
#' @description Vypocíta mediany po skupinách
#'
#' @param data_matrix Matica dát, kazdý riadok predstavuje jeden bod.
#' @param group Vektor oznacujúci skupinu pre kazdý riadok matice
#'        \code{data_matrix}.
#' @return Matica medianov \code{data_matrix} po skupinách. Kazdý riadok
#'         predstavuje viacrozmerný median.
groupMedians <- function(data_matrix, group) {
  unique_groups <- unique(group)
  g_means <-
    matrix(0,
           ncol = ncol(data_matrix),
           nrow = length(unique_groups))
  for (g in seq_along(unique_groups)) {
    g_means[g, ] <-
      as.vector(
        SpatialNP::ae.spatial.median(
          data_matrix[group == unique_groups[g], ],
          shape = FALSE
        )
      )
  }
  return(g_means)
}
```

```
### perm_test ########################################################################

#' @title perm_test
#' @description Vykoná permutacnú verziu testu \code{call}.
#'
#' @param call Funkcia testu.
#' @param data_matrix Matica dát, kazdý riadok predstavuje jeden bod.
#' @param group Vektor oznacujúci skupinu pre kazdý riadok matice
#'        \code{data_matrix}.
#' @param perm_statistic_type Spôsob vyhodnotenia permutacnej statistiky
#'        ("one-sided" / "two-sided"). (default: "one-sided")
#' @param n_perm Pocet permutácií. (default: 1000)
#' @return Objekt triedy htest obsahujúci p-hodnotu, statistiku, permutacné
#'         statistiky a dodatocné informácie.
perm_test <- function(call,
                      data_matrix,
                      group,
                      perm_statistic_type = "one-sided",
                      n_perm = 1000) {
  STATISTIC_PERM <- numeric(n_perm)
  group_length <- length(group)
  result <- call(data_matrix, group)
  STATISTIC <- result$statistic

  for (p in seq_len(n_perm)) {
    result <- call(data_matrix,
                   group[sample(group_length, replace = FALSE)])
    STATISTIC_PERM[p] <- result$statistic
  }

  if (perm_statistic_type == "one-sided") {
    p_value <- (1 + sum(STATISTIC <= STATISTIC_PERM)) / (1 + n_perm)
  } else {
```

```
    all_mean <- mean(c(STATISTIC_PERM, STATISTIC))
    centered_statistic <- STATISTIC - all_mean
    if (centered_statistic >= 0) {
      p_value <- (
        1 +
          sum(centered_statistic <= STATISTIC_PERM - all_mean) +
          sum(-centered_statistic >= STATISTIC_PERM - all_mean)
      ) /
        (1 + n_perm)
    } else {
      p_value <- (
        1 +
          sum(-centered_statistic <= STATISTIC_PERM - all_mean) +
          sum(centered_statistic >= STATISTIC_PERM - all_mean)
      ) /
        (1 + n_perm)
    }
  }

  names(STATISTIC) <- "Statistic"
  rval <- list(
    p.value = p_value,
    statistic = STATISTIC,
    statistic_perm = STATISTIC_PERM,
    method = paste("Permutacna verzia", substitute(call)),
    data.name = deparse(substitute(data_matrix))
  )
  class(rval) <- "htest"
  return(rval)
}
```

```
### projection ########################################################################

#' @title projection
#' @description Projekcia dát \code{data_matrix} v smere \code{v}.
#'
#' @param data_matrix Matica dát, kazdý riadok predstavuje jeden bod.
#' @param v Smerový vektor.
#' @return Projekcia dát \code{data_matrix} v smere \code{v}.
projection <- function(data_matrix, v) {
  return(tcrossprod(data_matrix,
                    tcrossprod(v) / drop(crossprod(v))))
}

### matrix_projection ##################################################################

#' @title matrix_projection
#' @description Projekcia dát \code{data_matrix} do priestoru \code{A}.
#'
#' @param data_matrix Matica dát, kazdý riadok predstavuje jeden bod.
#' @param A Matica bazovych vektorov priestoru.
#' @return Projekcia dát \code{data_matrix} do priestoru \code{A}.
matrix_projection <- function(data_matrix, A) {
  return(crossprod(t(data_matrix), A) %*% solve(crossprod(A)) %*% t(A))
}
```

```
### generate_data_matrix ###########################################################

#' @title generate_data_matrix
#' @description Vygeneruje viacrozmerné dáta zo zadanej distribúcie.
#'              Vyuzíva kniznicu \code{mvtnorm}.
#' @param distribution Distribúcia ("norm", "ball", "t_(\\d+)").
#' @param rows Pocet dát.
#' @param cols Dimenzia dát.
#' @param sigma Variancná matica pre "norm" a "t_(\\d+)".
#'        (default: identictá matica)
#' @return Matica realizácií.
generate_data_matrix <- function(distribution,
                                 rows,
                                 cols,
                                 sigma = diag(cols)) {
  if (distribution == "norm") {
    data_matrix <- mvtnorm::rmvnorm(n = rows,
                                    sigma = sigma,
                                    pre0.9_9994 = TRUE)
  } else if (distribution == "ball") {
    data_matrix <- matrix(ncol = cols, nrow = rows)
    for (i in 1:rows) {
      X <- rnorm(cols)
      data_matrix[i, ] <- runif(1) ^ (1 / cols) * X / sqrt(sum(X ^ 2))
    }
  } else if (grepl("t_(\\d+)", distribution)) {
    df <- as.numeric(sub("t_", "", distribution))
    data_matrix <- mvtnorm::rmvt(rows, sigma, df = df)
  } else {
    stop("Unknown distribution")
  }
  return(data_matrix)
}
```

```
### simulate_power ###################################################################

#' @title simulate_power
#' @description Vykoná simuláciu sily testu \code{call}.
#'
#' @param call Funkcia testu.
#' @param group_counts Vektor pocetností dát v skupinách.
#' @param delta Matica stredných hodnôt súborov, i-ty riadok je stredná hodnota
#'        i-teho súboru.
#' @param n_sim Pocet simulácií (default: 10000)
#' @param dimension Dimenzia dát. (default: 2)
#' @param seed Seed. (default: 1)
#' @param alpha Hladina významnosti. (default: 0.05)
#' @param fwer_correction Typ korekcie, pre testy:
#'        "tbk_vp",
#'        "tbk_jo",
#'        "krthk". (default: NA)
#' @param perm_version Boolean, vykonat permutacnú verziu? (default: FALSE)
#' @param n_perm Pocet permutácií. (default: 1000)
#' @param sigma Variancná matica pre "norm" a "t_(\\d+)".
#'        (default: identictá matica)
#' @param perm_statistic_type Spôsob vyhodnotenia permutacnej statistiky
#'        ("one-sided" / "two-sided"). (default: "one-sided")
#' @param distribution Distribúcia ("norm", "ball", "t_(\\d+)").
#'        (default: "norm")
#'
#' @return Nasimulovaný odhad sily testu \code{call}.
simulate_power <- function(call,
                           group_counts,
                           delta,
                           n_sim = 10000,
                           dimension = 2,
                           seed = 1,
```

```
                          alpha = 0.05,
                          fwer_correction = NA,
                          perm_version = FALSE,
                          n_perm = 1000,
                          sigma = diag(dimension),
                          perm_statistic_type = "one-sided",
                          distribution = "norm") {
set.seed(seed)
p_values <- numeric(n_sim)
group <- as.factor(rep(seq_along(group_counts), group_counts))
row_length <- sum(group_counts)

if (substitute(call) == "tbk_vp" || substitute(call) == "tbk_jo"
    || substitute(call) == "krthk") {
  for (sim in seq_len(n_sim)) {
    if (sim %% 100 == 1)
      print(sim)
    data_matrix <-
      generate_data_matrix(distribution, row_length, dimension, sigma)
    p_values[sim] <- !call(data_matrix + delta[group, ],
                           group,
                           method = fwer_correction,
                           alpha = alpha)$results
  }
} else if (perm_version) {
  for (sim in seq_len(n_sim)) {
    if (sim %% 100 == 1)
      print(sim)
    data_matrix <-
      generate_data_matrix(distribution, row_length, dimension, sigma)
    p_values[sim] <-
      perm_test(
        call,
```

```
          data_matrix + delta[group, ],
          group,
          perm_statistic_type = perm_statistic_type,
          n_perm = n_perm
        )$p.value
    }
  } else {
    for (sim in seq_len(n_sim)) {
      if (sim %% 100 == 1)
        print(sim)

      data_matrix <-
        generate_data_matrix(distribution, row_length, dimension, sigma)

      p_values[sim] <-
        call(data_matrix + delta[group, ], group)$p.value
    }
  }

  return(list(
    "power" = sum(p_values < alpha) / n_sim,
    "p_values" = p_values
  ))
}
```

```
### componentwise_rank ##########################################################

#' @title componentwise_rank
#' @description Zostrojí rank_matrix z pozícií v usporiadanom zduzenom subore
#'              v kazdej dimenzii.
#'
#' @param data_matrix Matica dát, kazdý riadok predstavuje jeden bod.
#' @return Matica rankov.
componentwise_rank <- function(data_matrix) {
  rank_matrix <- matrix(NA, ncol = 0, nrow = nrow(data_matrix))
  for (p_i in seq_len(ncol(data_matrix))) {
    rank_matrix <- cbind(rank_matrix,
                         rank(data_matrix[, p_i]))
  }
  return(rank_matrix)
}

### matrix_sqrt #################################################################

#' @title matrix_sqrt
#' @description Odmocnina z matice prostredníctvom vlastných hodnôt.
#'
#' @param A Matica pre odmocnenie.
#' @return Odmocnina z matice.
matrix_sqrt <- function(A) {
  eig <- eigen(A)
  eig$vectors %*% (diag(eig$values))^(1 / 2) %*% t(eig$vectors)
}
```

```
### cov_estimate ###################################################################

#' @title cov_estimate
#' @description Odhad kovariancie medzi v1 a v2.
#'
#' @param v1 Vektor 1.
#' @param v2 Vektor 2.
#' @return Odhad kovariancie medzi v1 a v2.
cov_estimate <- function(v1, v2) {
  n <- length(v1)
  combinations <- subset(expand.grid(rep(list(seq_len(n)), 2)), Var1 != Var2)
  n_c1 <- n - 1

  c1 <- mean(
    sign(v1[combinations[, 1]] - v1[combinations[, 2]]) *
      sign(v2[combinations[, 1]] - v2[combinations[, 2]])
  )

  combinations <-
    subset(expand.grid(rep(list(seq_len(
      n
    )), 3)), Var1 != Var2 & Var1 != Var3 & Var2 != Var3)
  n_c2 <- (n - 1) * (n - 2)

  c2 <- mean(
    sign(v1[combinations[, 1]] - v1[combinations[, 2]]) *
      sign(v2[combinations[, 1]] - v2[combinations[, 3]])
  )

  return((n_c1 * c1 + n_c2 * c2) / 4)
}
```

A.2 Testy

```
################################################################################
# Tests                                                                        #
# Viliam Zigo                                                                  #
################################################################################
# R version: 4.1                                                               #
# Libraries: mvtnorm, rospca, SpatialNP, energy, vegan                         #
################################################################################

### test_two_sample ############################################################

test_two_sample <- function(data_matrix, group) {
  g_means <- groupMeans(data_matrix, group)
  v <- g_means[2, ] - g_means[1, ]
  proj_data <- projection(data_matrix, v)
  k <- (proj_data[, 1] - g_means[2, 1]) / (v[1])

  result <- wilcox.test(k ~ group)
  rval <- list(
    p.value = result$p.value,
    statistic = result$statistic,
    v = v,
    proj_data = proj_data,
    method = "Test pre dva subory od SOMOGYI",
    data.name = deparse(substitute(data_matrix))
  )
  class(rval) <- "htest"
  return(rval)
}
```

```
### tmph ##############################################################################

tmph <- function(data_matrix, group) {
  combinations <- combn(length(unique(group)), 2)
  combinations_length <- ncol(combinations)
  p_values <- numeric(combinations_length)
  g_means <- groupMeans(data_matrix, group)
  for (com in seq_len(combinations_length)) {
    r <- as.numeric(group) %in% combinations[, com]
    p_values[com] <-
      test_two_sample(data_matrix[r, ], group[r])$p.value
  }

  com <-  which.min(p_values)
  v <-
    g_means[combinations[2, com], ] - g_means[combinations[1, com], ]
  proj_data <- projection(data_matrix, v)
  k <- (proj_data[, 1] - g_means[combinations[2, com], 1]) / (v[1])
  result <- kruskal.test(k ~ group)

  rval <- list(
    p.value = result$p.value,
    statistic = result$statistic,
    proj_data = proj_data,
    method = "Test s Minimálnou P-Hodnotou",
    data.name = deparse(substitute(data_matrix)),
    comb = combinations[, com],
    p_values = p_values,
    k = k
  )
  class(rval) <- "htest"
  return(rval)
}
```

```
### tlms ##############################################################################

tlms <- function(data_matrix, group) {
  unique_group_length <- length(unique(group))
  g_means <- groupMeans(data_matrix, group)
  dimension <- ncol(g_means)
  if (dimension > unique_group_length)
    stop("Cannot process data with higher dimension than number of groups!")

  df_means <- as.data.frame(g_means)
  df_colnames <- colnames(df_means)
  form <- paste(df_colnames[dimension],
                " ~ ",
                paste0(df_colnames[-dimension], collapse = "+"))
  model <- lm(form, df_means)
  coeff <- model$coefficients
  v <- c(rep(1, dimension - 1), sum(coeff[-1]))
  proj_data <- projection(data_matrix, v)
  x <- proj_data[, 1]
  result <- kruskal.test(x ~ group)

  rval <- list(
    p.value = result$p.value,
    statistic = result$statistic,
    method = "Test s Lineárnym Modelom daným Stredmi",
    data.name = deparse(substitute(data_matrix)),
    x = x,
    line_coeff = coeff
  )
  class(rval) <- "htest"
  return(rval)
}
```

```
### tlmp ##############################################################################

tlmp <- function(data_matrix, group) {
  dimension <- ncol(data_matrix)
  if (dimension > nrow(data_matrix))
    stop(
      "Cannot process data with higher dimension than number of observations!"
    )

  df_data <- as.data.frame(data_matrix)
  df_colnames <- colnames(df_data)
  form <- paste(df_colnames[dimension],
                " ~ ",
                paste0(df_colnames[-dimension], collapse = "+"))
  coeff <- lm(form, df_data)$coefficients
  v <- c(rep(1, dimension - 1), sum(coeff[-1]))
  proj_data <- projection(data_matrix, v)
  x <- proj_data[, 1]
  result <- kruskal.test(x ~ group)

  rval <- list(
    p.value = result$p.value,
    statistic = result$statistic,
    method = "Test s~Lineárnym Modelom daným Pozorovaniami",
    data.name = deparse(substitute(data_matrix)),
    x = x,
    line_coeff = coeff
  )
  class(rval) <- "htest"
  return(rval)
}
```

```
### tmr ########################################################################

tmr <- function(data_matrix, group) {
  groups <- length(levels(group))
  combinations <- combn(groups, 2)
  combinations_length <- ncol(combinations)
  distances <- numeric(combinations_length)
  g_means <- groupMeans(data_matrix, group)
  for (com in seq_len(combinations_length)) {
    direction <-
      g_means[combinations[2, com],] - g_means[combinations[1, com],]
    proj_data <- projection(data_matrix, direction)
    distances[com] <- max(dist(proj_data))
  }

  com <- which.max(distances)
  v <-
    g_means[combinations[2, com], ] - g_means[combinations[1, com], ]
  proj_data <- projection(data_matrix, v)
  k <- (proj_data[, 1] - g_means[combinations[2, com], 1]) / (v[1])
  result <- kruskal.test(k ~ group)
  rval <- list(
    p.value = result$p.value,
    statistic = result$statistic,
    method = "Test s Maximálnym Rozptylom",
    data.name = deparse(substitute(data_matrix)),
    comb = combinations[, com],
    distances = distances,
    k = k
  )
  class(rval) <- "htest"
  return(rval)
}
```

```
### tbk_vp ###########################################################################

tbk_vp <-
  function(data_matrix,
           group,
           alpha = 0.05,
           method = "bonferroni") {
    groups <- length(unique(group))
    combinations <- combn(groups, 2)
    combinations_length <- ncol(combinations)
    p_values <- numeric(combinations_length)

    for (com in seq_len(combinations_length)) {
      r <- as.numeric(group) %in% combinations[, com]
      p_values[com] <-
        perm_test(test_two_sample,
                  data_matrix[r, ],
                  group[r],
                  perm_statistic_type = "two-sided")$p.value
    }

    adjusted_p_values <- p.adjust(p_values, method)
    results <- sum(adjusted_p_values <= alpha) > 0

    return(
      list(
        "results" = results,
        "p_values" = p_values,
        "adjusted_p_values" = adjusted_p_values
      )
    )
  }
```

```
### tbk_jo ###########################################################################

tbk_jo <-
  function(data_matrix,
           group,
           alpha = 0.05,
           method = "bonferroni") {
    groups <- length(unique(group))
    combinations <- rbind(1, seq_len(groups - 1) + 1)
    combinations_length <- ncol(combinations)
    p_values <- numeric(combinations_length)

    for (com in seq_len(combinations_length)) {
      r <- as.numeric(group) %in% combinations[, com]
      p_values[com] <-
        perm_test(test_two_sample,
                  data_matrix[r, ],
                  group[r],
                  perm_statistic_type = "two-sided")$p.value
    }

    adjusted_p_values <- p.adjust(p_values, method)
    results <- sum(adjusted_p_values <= alpha) > 0

    return(
      list(
        "results" = results,
        "p_values" = p_values,
        "adjusted_p_values" = adjusted_p_values
      )
    )
  }
```

```
### thk ##########################################################################

thk <- function(data_matrix, group) {
  pca <- prcomp(data_matrix, scale = FALSE)
  x <- pca$x[, 1]

  if (length(unique(group)) == 2) {
    result <- wilcox.test(x ~ group)
  } else {
    result <- kruskal.test(x ~ group)
  }

  rval <- list(
    p.value = result$p.value,
    statistic = result$statistic,
    method = "Test s Hlavnými Komponentami",
    data.name = deparse(substitute(data_matrix)),
    pca = pca,
    x = x
  )
  class(rval) <- "htest"
  return(rval)
}
```

```
### rthk ##############################################################################

rthk <- function(data_matrix, group) {
  rpca <- rospca::robpca(data_matrix, k = 1)
  x <- rpca$scores

  if (length(unique(group)) == 2) {
    result <- wilcox.test(x ~ group)
  } else {
    result <- kruskal.test(x ~ group)
  }

  rval <- list(
    p.value = result$p.value,
    statistic = result$statistic,
    method = "Robustný Test s Hlavnými Komponentami",
    data.name = deparse(substitute(data_matrix)),
    rpca = rpca
  )
  class(rval) <- "htest"
  return(rval)
}
```

```
### krthk ##############################################################################

krthk <-
  function(data_matrix,
           group,
           alpha = 0.05,
           method = "holm") {
    rpca <- rospca::robpca(data_matrix, kmax = ncol(data_matrix))

    p_values <- numeric(rpca$k)
    for (i in seq_len(rpca$k)) {
      x <- rpca$scores[, i]

      if (length(unique(group)) == 2) {
        p_values[i] <- wilcox.test(x ~ group)$p.value
      } else {
        p_values[i] <- kruskal.test(x ~ group)$p.value
      }
    }

    adjusted_p_values <- p.adjust(p_values, method)
    results <- sum(adjusted_p_values <= alpha) > 0
    rval <- list(
      results = results,
      p_values = p_values,
      adjusted_p_values = adjusted_p_values,
      method = "Korigovaný Robustný Test s~Hlavnými Komponentami",
      data.name = deparse(substitute(data_matrix)),
      rpca = rpca
    )
    class(rval) <- "htest"
    return(rval)
  }
```

```
### thspz ###########################################################################

thspz <- function(data_matrix, group) {
  g_means <- groupMeans(data_matrix, group)
  dimension <- ncol(data_matrix)
  lambda <- min(dist(g_means)) / 2
  transformed <- matrix(0, nrow = 0, ncol = dimension)

  for (g in seq_len(nrow(g_means))) {
    transformed <- rbind(
      transformed,
      matrix(
        g_means[g, ],
        nrow = sum(group == g),
        ncol = dimension,
        byrow = TRUE
      ) + lambda * SpatialNP::spatial.signs(data_matrix[group == g, ],
                                            center = g_means[g, ],
                                            shape = FALSE)
    )
  }

  pca <- prcomp(transformed, scale = FALSE)
  x <- data_matrix %*% pca$rotation[1, ]

  if (length(unique(group)) == 2) {
    result <- wilcox.test(x ~ group)
  } else {
    result <- kruskal.test(x ~ group)
  }

  rval <- list(
    p.value = result$p.value,
```

```
    statistic = result$statistic,
    method = "Test v Hlavnom Smere Priestorového Znamienka",
    data.name = deparse(substitute(data_matrix)),
    x = x,
    transformed = transformed,
    pca = pca
  )
  class(rval) <- "htest"
  return(rval)
}

### test_manova ######################################################################

test_manova <- function(data_matrix, group) {
  fit <- manova(data_matrix ~ group)
  result <- summary(fit, test = "Wilks", tol = 0)
  rval <- list(
    p.value = result$stats[1, "Pr(>F)"],
    method = "Test s vyuzitim manova",
    data.name = deparse(substitute(data_matrix)),
    fit = fit
  )
  class(rval) <- "htest"
  return(rval)
}
```

```
### trlms ##############################################################################

trlms <- function(data_matrix, group) {
  unique_group_length <- length(unique(group))
  dimension <- ncol(data_matrix)
  data_length <- nrow(data_matrix)

  if (dimension > unique_group_length)
    stop("Cannot process data with higher dimension than number of groups!")

  data_matrix_d <- data_matrix
  dimension_d <- dimension
  for (d in seq_len(dimension - 1)) {
    g_means_d <- groupMeans(data_matrix_d, group)
    df_means <- as.data.frame(g_means_d)
    df_colnames <- colnames(df_means)
    form <- paste(df_colnames[dimension_d],
                  " ~ ",
                  paste0(df_colnames[-dimension_d], collapse = "+"))
    coeff <- lm(form, df_means)$coefficients

    dimension_d <- dimension - d
    A <- rbind(diag(dimension_d), coeff[-1])
    data_matrix_d <- matrix_projection(data_matrix_d, A)[, seq_len(dimension_d)]
  }

  result <- kruskal.test(data_matrix_d ~ group)
  rval <- list(
    p.value = result$p.value,
    statistic = result$statistic,
    method = "Test s Rekurentným Lineárnym Modelom daným Stredmi",
    data.name = deparse(substitute(data_matrix)),
    x = data_matrix_d,
```

```
    line_coeff = coeff
  )
  class(rval) <- "htest"
  return(rval)
}

### mrpp_test ##################################################################

mrpp_test <-  function(data_matrix, group) {
  # Je to obalene vo vlasntej funkcii iba pre konzistenciu v simulate_power

  result <- vegan::mrpp(data_matrix, group)

  rval <- list(
    p.value = result$Pvalue,
    method = "MRPP",
    data.name = deparse(substitute(data_matrix))
  )
  class(rval) <- "htest"
  return(rval)
}
```

```
### decor_mkw ##########################################################################

decor_mkw <- function(data_matrix, group) {
  group_counts <- as.vector(table(group))
  n <- nrow(data_matrix)
  p <- ncol(data_matrix)
  groups <- length(unique(group))
  rank_matrix <- componentwise_rank(prcomp(data_matrix)$x)
  mean_rank_matrix <- groupMeans(rank_matrix, group)
  mean_rank <- (n + 1) / 2 * rep(1, p)
  centered_mean_rank_matrix <-
    mean_rank_matrix - matrix(rep(mean_rank, groups),
                              ncol = p,
                              byrow = TRUE)


  STATISTIC <- ((n - 1) / n) * 12 / (n ^ 2 - 1) * sum(
    group_counts * diag(centered_mean_rank_matrix %*% t(centered_mean_rank_matrix))
  )
  DF <- p * max(1, groups - 1)
  PVAL <- 1 - pchisq(STATISTIC, DF)
  names(STATISTIC) <- "MH"
  names(DF) <- "df"
  rval <- list(
    p.value = PVAL,
    statistic = STATISTIC,
    method = "DecorMKW",
    parameter = DF,
    data.name = deparse(substitute(data_matrix))
  )
  class(rval) <- "htest"
  return(rval)
}
```

```
### mkwor ##############################################################################

mkwor <- function(data_matrix, group) {
  group_counts <- as.vector(table(group))
  n <- nrow(data_matrix)
  p <- ncol(data_matrix)
  groups <- length(unique(group))

  V <- SpatialNP::rank.shape(data_matrix)
  spatial_rank_matrix <- SpatialNP::spatial.rank(
    data_matrix %*% solve(matrix_sqrt(V)),
    shape = FALSE
  )

  rank_matrix <- componentwise_rank(spatial_rank_matrix)
  mean_rank_matrix <- groupMeans(rank_matrix, group)
  mean_rank <- (n + 1) / 2 * rep(1, p)
  centered_mean_rank_matrix <- mean_rank_matrix - matrix(
    rep(mean_rank, groups),
    ncol = p,
    byrow = TRUE
  )

  STATISTIC <- ((n - 1) / n) * 12 / (n ^ 2 - 1) * sum(
    group_counts * diag(centered_mean_rank_matrix %*% t(centered_mean_rank_matrix))
  )
  DF <- p * max(1, groups - 1)
  PVAL <- 1 - pchisq(STATISTIC, DF)
  names(STATISTIC) <- "MH"
  names(DF) <- "df"
  rval <- list(
    p.value = PVAL,
    statistic = STATISTIC,
```

```
    method = "MKWOR",
    parameter = DF,
    data.name = deparse(substitute(data_matrix))
  )
  class(rval) <- "htest"
  return(rval)
}

### sr_test ###########################################################################

sr_test <- function(data_matrix, group) {
  # Je to obalene vo vlasntej funkcii iba pre konzistenciu v simulate_power,
  # pretoze sr.loc.test nema defaultne druhy argument skupinu a argument
  # skupiny je nazvany g a nie group ako v nasich testoch.

  return(SpatialNP::sr.loc.test(data_matrix, g = group, score = "rank"))
}

### energy_test #######################################################################

energy_test <-  function(data_matrix, group) {
  # Je to obalene vo vlasntej funkcii iba pre konzistenciu v simulate_power
  group_counts <- as.vector(table(group))
  return(energy::eqdist.etest(data_matrix, group_counts, R = 1000))
}
```

```
### mkwce ########################################################################

mkwce <- function(data_matrix, group) {
  group_counts <- as.vector(table(group))
  n <- nrow(data_matrix)
  p <- ncol(data_matrix)
  groups <- length(unique(group))
  mean_rank_matrix <-
    groupMeans(componentwise_rank(data_matrix), group)
  mean_rank <- (n + 1) / 2 * rep(1, p)

  centered_mean_rank_matrix <-
    mean_rank_matrix - matrix(rep(mean_rank, groups),
                              ncol = p,
                              byrow = TRUE)

  sigma_rank <- (n^2 - 1)/12 * diag(p)

  combinations <- combn(p, 2)
  combinations_length <- ncol(combinations)
  for (c in seq_len(combinations_length)) {
    row <- combinations[1, c]
    col <- combinations[2, c]

    sigma_rank[row, col] <-
      sigma_rank[col, row] <-
      cov_estimate(data_matrix[, row], data_matrix[, col])
  }

  sigma_rank_inv <- solve(sigma_rank)

  rank_sum_items <- numeric(groups)
  for (g in seq_len(groups)) {
```

```
    rank_sum_items[g] <-
      group_counts[g] * t(centered_mean_rank_matrix[g, ]) %*%
      sigma_rank_inv %*% centered_mean_rank_matrix[g, ]
  }

  STATISTIC <- ((n - 1) / n) * sum(rank_sum_items)
  DF <- p * max(1, groups - 1)
  PVAL <- 1 - pchisq(STATISTIC, DF)

  names(STATISTIC) <- "MH"
  names(DF) <- "df"
  rval <- list(
    p.value = PVAL,
    statistic = STATISTIC,
    method = "MKWCE",
    parameter = DF,
    data.name = deparse(substitute(data_matrix))
  )
  class(rval) <- "htest"
  return(rval)
}
```

```
### mkwb ########################################################################

mkwb <- function(data_matrix, group) {
  B <- 1000
  group_counts <- as.vector(table(group))
  n <- nrow(data_matrix)
  p <- ncol(data_matrix)
  groups <- length(unique(group))
  mean_rank_matrix <-
    groupMeans(componentwise_rank(data_matrix), group)
  mean_rank <- (n + 1) / 2 * rep(1, p)

  centered_mean_rank_matrix <-
    mean_rank_matrix - matrix(rep(mean_rank, groups),
                              ncol = p,
                              byrow = TRUE)


  x <- array(rep(NA, B*n*p), dim=c(B, n, p))

  for (b in seq_len(B)) {
    indexes <- sample(n, replace = TRUE)
    x[b, , ] <- componentwise_rank(data_matrix[indexes, ])
  }

  sigma_rank <- matrix(0, ncol = p, nrow = p)
  for (i in seq_len(n)) {
    sigma_rank <- sigma_rank + cov(x[, i, ])
  }

  sigma_rank <- sigma_rank / n
  sigma_rank_inv <- solve(sigma_rank)
```

```
  rank_sum_items <- numeric(groups)
  for (g in seq_len(groups)) {
    rank_sum_items[g] <-
      group_counts[g] * t(centered_mean_rank_matrix[g, ]) %*%
      sigma_rank_inv %*% centered_mean_rank_matrix[g, ]
  }

  STATISTIC <- ((n - 1) / n) * sum(rank_sum_items)
  DF <- p * max(1, groups - 1)
  PVAL <- 1 - pchisq(STATISTIC, DF)

  names(STATISTIC) <- "MH"
  names(DF) <- "df"
  rval <- list(
    p.value = PVAL,
    statistic = STATISTIC,
    method = "MKWB",
    parameter = DF,
    data.name = deparse(substitute(data_matrix))
  )
  class(rval) <- "htest"
  return(rval)
}
```