UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



VIZUALIZÁCIA KATEGORIZOVANÝCH DÁT

DIPLOMOVÁ PRÁCA

2021 Bc. Antónia IHNÁTOVÁ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

VIZUALIZÁCIA KATEGORIZOVANÝCH DÁT

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	1113 Matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	doc. Mgr. Radoslav Harman, PhD.

2021 **Bc. Antónia IHNÁTOVÁ**





Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Antónia Ihnátová
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: Vizualizácia kategorizovaných dát
Visualization of categorized data

Anotácia: Vizualizácia mnohorozmerných dát v nízkorozmernom priestore je jedným z najdôležitejších nástrojov exploratívnej analýzy. Klasické zobrazovacie metódy však berú do úvahy len 'neoznačené' dáta, čiže len vysvetľujúce premenné pozorované na objektoch. To môže mať za následok zobrazenie, ktoré je z hľadiska informovania o štruktúre dát, povedzme o ich príslušnosti k rôznym zhlukom, nevhodné.

Vedúci: doc. Mgr. Radoslav Harman, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Marek Fila, DrSc.

Dátum zadania: 17.12.2019

Dátum schválenia: 21.12.2019 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Touto cestou chcem úprimne poďakovať svojmu vedúcemu diplomovej práce doc. Mgr. Radoslavovi Harmanovi, PhD. za ochotu, ústretovosť, podnetné pripomienky a cenné rady aj do života. Rovnako veľké ďakujem patrí aj mojim rodičom a sestre s manželom, u ktorých som v ťažkých chvíľach stále našla morálnu podporu.

Abstrakt v štátnom jazyku

Bc. IHNÁTOVÁ, Antónia: Vizualizácia kategorizovaných dát [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: doc. Mgr. Radoslav Harman, PhD., Bratislava, 2021, 74 s.

Táto práca sa zaoberá vizualizáciou kategorizovaných dát, ktoré obsahujú vysvetľujúce kvantitatívne premenné a jednu závislú kategoriálnu premennú. Okrem prehľadu rôznych lineárnych metód adekvátnych na vizualizáciu kategorizovaných dát v práci predstavujeme aj vlastnú metódu, vizualizáciu kategorizovaných dát pomocou regulárneho simplexového kódovania a kanonických korelácií (catCCA). Základom catCCA je netradičná aplikácia analýzy kanonických korelácií. Pri tvorbe pomocných premenných pre kategoriálnu premennú sme čerpali inšpiráciu z vlastností regulárneho simplexu. Naša metóda je zároveň užitočným alternatívnym pohľadom na dobre známu metódu lineárnej diskriminačnej analýzy. Práca obsahuje aj praktickú časť, v ktorej demonštrujeme aplikáciu vybraných metód. Ponaučenie, ktoré plynie z tejto práce je, že cieľom nie je nájsť najlepšiu metódu vizualizácie dát, ale na vytvorenie predstavy o dátach je dobré aplikovať viacero metód a z každej vizualizácie niečo vyťažiť.

Kľúčové slová: exploratívna dátová analýza, kanonické korelácie, lineárna diskriminačná analýza, označené dáta, redukcia dimenzie.

Abstract

Bc. IHNÁTOVÁ, Antónia: Visualization of categorized data [Master Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: doc. Mgr. Radoslav Harman, PhD., Bratislava, 2021, 74 p.

This thesis is dedicated to the visualization of categorized data containing explanatory quantitative variables and one dependent categorical variable. In addition to overview of various linear methods suitable for the visualization of categorized data, the thesis also presents our own method, utilizing regular simplex coding and canonical correlations (catCCA). The basis of catCCA is an unconventional application of canonical correlation analysis. The process of creating the auxiliary variables corresponding to the categorical variable was motivated by the properties of a regular simplex. Our method is also a useful alternative view on the well-known method of the linear discriminant analysis. The thesis also contains a practical section which demonstrates the application of selected methods. The conclusion of this thesis is that the aim should not be to find a single best method of data visualization, but for the purpose of getting an insight into the data it is beneficial to apply several methods that together convey the most comprehensive information.

Keywords: exploratory data analysis, canonical correlation, linear discriminant analysis, labeled data, dimension reduction.

Obsah

Úvod **9**

1 Kategorizované dáta **11**

2 Klasifikácia metód vizualizácie dát **14**

3 Lineárne metódy vizualizácie dát bez učiteľa **16**

3.1 Analýza hlavných komponentov (*PCA*) 16

4 Lineárne metódy vizualizácie dát s učiteľom **19**

4.1 Analýza hlavných komponentov pre kategorizované dáta (catPCA) 19

4.2 Lineárna diskriminačná analýza (LDA) 22

4.3 Redukcia dimenzie riadená vnímaním využívajúca konzistenciu vzdialeností s ohľadom na hustotu (*PDD*) 26

5 Vizualizácia pomocou kanonických korelácií **30**

5.1 Analýza kanonických korelácií (CCA) 30

5.2 Vizualizácia kategorizovaných dát pomocou regulárneho simplexového kódovania a kanonických korelácií (catCCA) 33

5.2.1 Spracovanie dát 34

5.2.2 Praktický návod na aplikáciu metódy. 40

5.2.3 Vlastnosti metódy. 42

5.2.4 Súvis s inými metódami 45

6 Zhodnotenie výsledkov vizualizácie vybraných dátových súborov **50**

6.1 Umelý dátový súbor A 50

6.2 Iris 56

6.3 Pen-Based Recognition of Handwritten Digits 60

Záver **71**

Zoznam použitej literatúry **73**

Príloha **75**

Úvod

Predstavte si, že si chcete naplánovať letnú dovolenku na mieste, kde ste doteraz ešte neboli. Jednou z možnosti ako si vybrať destináciu je zavrieť oči, roztočiť glóbus a náhodne ukázať prstom. Väčšina ľudí sa však pred výberom destinácie informuje o možnostiach dopravy, kultúrnych pamiatkach v danej krajine alebo o type pláže v uvažovanej oblasti. Podobný postup je vítaný aj pri práci s dátami.

Pred aplikáciou rôznych štatistických metód je dôležité oboznámiť sa s charakterom dát a získať o nich predstavu. Štatistická oblasť, ktorá sa zaoberá prvotným skúmaním dát a hľadaním hlavných charakteristík sa nazýva exploratívna dátová analýza (angl. *explanatory data analysis*). Za zakladateľa sa považuje americký matematik John Wilder Tukey, ktorý v roku 1977 vydal knihu *Exploratory data analysis*. V tejto oblasti sú zahrnuté rôzne procesy. Za súčasť exploratívnej dátovej analýzy sa považuje aj zistenie rozmeru dát, typu premenných alebo rozsah nadobúdaných hodnôt jednotlivých premenných. Najčastejším nástrojom exploratívnej analýzy dát je vizualizácia dát. Pre voľbu adekvátnej metódy je rozhodujúci typ premenných. Pre každý typ premenných existuje široké spektrum rôznych prístupov na vizualizáciu. Následne pomocou zvolenej vizualizácie je možné získať predstavu o štruktúre dát, o trende v dátach, nájsť zaujímavé vzťahy medzi pozorovaniami, identifikovať odľahlé pozorovania (angl. *outliers*), ale aj navrhnúť hypotézy na ďalšie testovanie.

Neexistuje univerzálne najlepšia metóda vizualizácie dát, ani automat, ktorý by pre každú situáciu zvolil optimálnu metódu. Každý prístup môže prinášať užitočné pohľady a informácie. V našej práci sa zameriame na vizualizáciu mnohorozmerných kategorizovaných dát v nízkorozmernom priestore, rovine, pomocou lineárnych metód vizualizácie dát. Tieto metódy úzko súvisia s redukciou dimenzie dát. Typickou črtou je hľadanie dvoch smerov, ktoré vzniknú lineárnou kombináciou pôvodných smerov. Kategorizované dáta obsahujú nielen vysvetľujúce premenné pozorované na objektoch, ale aj kategóriu, z ktorej pozorovania pochádzajú.

„Klasické" zobrazovacie metódy berú do úvahy len „neoznačené" dáta, čiže len vysvetľujúce premenné pozorované na objektoch. Výsledné zobrazenie nemusí dobre informovať o štruktúre dát z hľadiska ich príslušnosti k rôznym kategóriam. Cieľom tejto práce je

predstaviť novú lineárnu metódu vizualizácie kategorizovaných dát, ktorá berie do úvahy aj informácie o kategóriach. Počas výskumu sa ukázalo, že naša metóda je vo výsledku formálne ekvivalentná už inej dobrej známej metóde, lineárnej diskriminačnej analýze. Naša práca teda prináša alternatívny pohľad a nové intuitívne zdôvodnenie metódy založenej na lineárnej diskriminačnej analýze.

Táto práca obsahuje 6 kapitol. V prvej kapitole uvedieme stručné členenie dát z hľadiska typu nadobúdaných hodnôt a upresníme terminológiu, ktorú používame v práci. Následne v druhej kapitole začleníme lineárne metódy vizualizácie dát do hierarchie metód vizualizácie dát spojených s redukciou dimenzie. Tretia a štvrtá kapitola je venovaná vybraným známym metódam. Presnejšie, v tretej kapitole uvedieme zástupcu lineárnych metód, ktoré nepracujú s plnou mierou informácie obsiahnutej v kategorizovaných dátach, ale môžu byť použité na ich vizualizáciu. Metódy uvedené v štvrtej kapitole využívajú plnú mieru informácie. Hlavnou časťou práce je navrhnutie vlastnej metódy vizualizácie kategorizovaných dát, ktorá využíva plnú mieru informácie. Spolu s jej vlastnosťami a dôkazom ekvivalencie s metódou lineárnej diskriminačnej analýzy ju popisujeme v piatej kapitole. Poslednú šiestu kapitolu tvorí praktická časť. V nej aplikujeme vybrané metódy vizualizácie dát spolu s nami navrhnutou metódou na umelé aj reálne dátové súbory a interpretujeme výsledky vizualizácií.

1 Kategorizované dáta

Pred samotnou analýzou dát je dôležité oboznámiť sa s typom dát a na základe neho zvoliť vhodné štatistické metódy. V súčasnosti neexistuje úplne jednotný systém klasifikácie typov dát. V jednotlivých publikáciách sa využívajú klasifikácie od rôznych autorov. V tejto kapitole uvedieme nami zvolenú klasifikáciu a definujeme typy dát, ktoré sú podstatné pre našu prácu. Pri vypracovaní tejto kapitoly sme vychádzali z [1], [8], [12] a [14].

Z hľadiska typu nadobúdaných hodnôt delíme štatistické premenné na kvantitatívne a kvalitatívne, ktoré sa označujú aj ako kategoriálne. Kategoriálna premenná (angl. *categorical variable*) je premenná, ktorá nadobúda hodnoty z diskrétnej konečnej množiny. Nemusí ísť výlučne iba o podmnožinu reálnych čísel. Množina nadobúdaných hodnôt môže byť tvorená aj textovými prvkami, intervalmi a inými symbolmi, napríklad {žena, muž}. Na základe nadobúdaných hodnôt je možné pozorovania rozdeliť do k disjunktných množín, kde k označuje počet kategórií. Kategórie spravidla vyplývajú z kvalitatívnych vlastností pozorovaní, napríklad {pšenica letná forma ozimná, pšenica tvrdá, pšenica špalda}. Kategoriálna premenná môže vzniknúť aj z kvantitatívnej premennej, a to v prípade, že nie sú dostupné presné hodnoty kvantitatívnej premennej, alebo sú informácie obsiahnuté v kvantitatívnej premennej zámerne zovšeobecnené kvôli jednoduchosti. Napríklad pre ciele analýzy môže byť užitočné pracovať s kategóriami {dieťa, dospelý, dôchodca} namiesto presnej hodnoty veku. Kategórie môžu byť reprezentované numerickými indexmi, ale vo všeobecnosti s takýmito indexmi nemôže byť zaobchádzané ako s číslami. Bližšie informácie o numerických indexoch uvedieme pri delení kategoriálnych premenných.

Kategoriálne premenné delíme z pohľadu vzťahu medzi pozorovaniami na ordinálne (angl. *ordinal variable*) a nominálne (angl. *nominal variable*). Kategórie ordinálnych premenných sa dajú prirodzene usporiadať, ale „vzdialenosť" medzi jednotlivými kategóriami nie je známa a môže sa medzi jednotlivými kategóriami líšiť. Príkladom sú dáta merajúce spokojnosť zákazníkov {veľmi nespokojný, nespokojný, spokojný, veľmi spokojný}. Ak sú kategóriam priradené numerické indexy, tak sú tiež usporiadané, napríklad $\{1, 2, 3, 4\}$. Poskytujú však len informáciu o usporiadaní. Nevyplýva z nich, že veľmi spokojný zákazník (4) je dvakrát taký spokojný ako nespokojný zákazník (2). Pri práci s ordinálnymi dátami je vhodné zvoliť metódu, ktorá využíva nielen informáciu o kategóriách, ale aj informá-

ciu o usporiadaní kategórií. Metódy by mali byť nezávislé od vzostupného, respektíve zostupného usporiadania dát. Iné preusporiadania by mali výsledok ovplyvniť.

V prípade, že medzi kategóriami neexistuje prirodzené usporiadanie, hovoríme o nominálnej premennej. Konkrétne, v prípade kategórií {muž, žena} nevieme určiť, či žena je „viac" ako muž alebo naopak. Kategórie nominálnych premenných môžu byť reprezentované numerickými indexmi. Na základe týchto indexov vieme len určiť, či pozorovania patria do rovnakej kategórie alebo do odlišných kategórií. Metódy využívané na analýzu nominálnych dát by mali byť nezávislé na poradí kategórií. Ľubovoľným preusporiadaním kategórií by sme mali získať tie isté výsledky. Platí, že metódy vhodné pre nominálne dáta môžu byť použité aj pre ordinálne dáta. Nevýhodou je neúplne využitie informácií obsiahnutých v dátach. Naopak to neplatí. Metódy navrhnuté pre ordinálne dáta nie sú vhodné pre nominálne dáta.

Špeciálnym typom kategoriálnej premennej je binárna premenná (angl. *binary variable*), ktorá má práve 2 kategórie. Zvyčajne sú kategórie reprezentované numerickými indexmi $\{0, 1\}$.

Kategoriálne dáta môžu byť reprezentované pomocnými premennými, v ktorých bude obsiahnutá informácia o kategórii. Existujú rôzne prístupy. Typickým prístupom je metóda *one hot encoding*. Pre k kategórií vytvoríme prislúchajúcich k pomocných premenných, ktoré nadobúdajú hodnotu 1, ak pozorovanie patrí do danej kategórie, inak nadobúdajú hodnotu 0. Podobným prístupom je vytvorenie *dummy* premenných. Prvým krokom je voľba referenčnej kategórie a potom pre ostatných $k - 1$ kategórií, rovnako ako v prípade *one hot encoding*, vytvoríme prislúchajúce pomocné dummy premenné. Pozorovanie z referenčnej kategórie nadobúda hodnotu 0 vo všetkých pomocných dummy premenných. Množstvo informácie obsiahnuté v $k - 1$ pomocných dummy premenných je rovné množstvu informácie obsiahnutému v k pomocných premenných z prvého uvedeného prístupu.

Kvantitatívne premenné sú premenné, ktoré nadobúdajú numerické hodnoty. Existuje poväčšine jednotka merania, v ktorej sú kvantitatívne premenné vyjadrené. Podľa počtu možných nadobúdaných hodnôt ich delíme na diskrétne, ktoré nadobúdajú konečný alebo spočítateľný počet hodnôt, a spojité premenné, ktoré nadobúdajú nespočítateľný počet hodnôt. Z pohľadu vzťahu medzi premennými rozlišujeme intervalové (angl. *interval va-*

riable) a pomerové premenné (angl. *ratio variable*). Pri obidvoch typoch má zmysel určovať rozdiely medzi hodnotami premennej, ale v prípade pomerovej premennej je zmysluplné uvažovať aj pomer medzi nimi. Často pomerové premenné nadobúdajú len nezáporné hodnoty. Na ilustráciu uvažujme intervalovú premennú zodpovedajúcu teplote vyjadrenej v stupňoch Celzia. Je zmysluplné počítať rozdiel medzi dvomi teplotami, avšak podiel dvoch teplôt nemá interpretačný význam.

V našej práci sa zaoberáme dátovými súbormi, v ktorých sú nezávislé, vysvetľujúce premenné kvantitatívne a závislá, vysvetľovaná premenná je kategoriálna. Takéto dáta budeme nazývať kategorizované. Zameriame sa na vizualizáciu dát s nominálnymi premennými, ktoré sú aplikovateľné aj na ordinálne premenné.

2 Klasifikácia metód vizualizácie dát

Vizualizácia mnohorozmerných dát v menej rozmernom priestore, typicky $\mathbb{R}^2$ alebo $\mathbb{R}^3$, úzko súvisí s metódami redukcie dimenzie dát. Na redukciu dimenzie dát sa používajú dve techniky. Prvou je selekcia premenných, ktorá sa snaží vybrať relevantnú podmnožinu pôvodných premenných. Metódy vizualizácie dát sa zvyčajne sústreďujú na druhú techniku, ktorá spočíva v extrakcii informácie zo všetkých pôvodných premenných.

Metódy vizualizácie dát môžeme členiť z rôznych hľadísk. Z hľadiska typu informácie, ktorú metódy využívajú, členíme metódy na metódy bez učiteľa (angl. *unsupervised methods*), metódy s učiteľom (angl. *supervised methods*) a metódy s čiastočným učiteľom (angl. *semi-supervised methods*). Metódy bez učiteľa pracujú len s informáciami obsiahnutými v nezávislých premenných. Metódy s učiteľom využívajú navyše aj informácie obsiahnuté v závislých premenných. Tieto metódy sú vhodné aj pre nami zadefinované kategorizované dáta, kde závislá premenná je kategoriálna. Medzník tvoria metódy s čiastočným učiteľom, ktoré sú určené pre dáta, v ktorých sú hodnoty závislej premennej známe iba pre niektoré pozorovania. Metódy s učiteľom hľadajú také nízkorozmerné reprezentácie dát, ktoré zachytávajú nielen celkovú štruktúru, ale odzrkadľujú aj vzťahy medzi rôznymi kategóriami. V porovnaní s metódami bez učiteľa spravidla lepšie zachytávajú štruktúru rôznych kategórií.

Dáta nemusia obsahovať iba jednu vysvetľovanú premennú, ako je to v našom prípade. Existujú metódy, ktoré sa zameriavajú aj na redukciu priestoru viacerých závislých premenných. Prehľad členenia metód z hľadiska priestoru, ktorý redukujú, a informácie, ktoré pri tom využívajú, je spracovaný v [13].

Ďalším zo základných členení metód je charakterizácia z pohľadu linearity na lineárne a nelineárne metódy. Ako už vyplýva z názvu, lineárne metódy zobrazujú mnohorozmerné dáta do nízkorozmerného priestoru pomocou lineárnych transformácií. Sústredia sa na vernejšiu reprezentáciu geometrie a vzájomnej polohy dát v nízkorozmernom priestore za cenu nižšej schopnosti prispôsobiť sa špecifickým nelinearitám v dátach. Výhodou lineárnych metód je ľahká interpretácia a zrozumiteľnosť, pretože každý smer v nízkorozmernom priestore, v ktorom sú dáta zobrazované, vzniká lineárnou kombináciou pôvodných smerov. Vo všeobecnosti sa tieto metódy považujú za výpočtovo efektívne. Na druhej strane

nelineárne metódy sú schopné zachytiť aj nelineárnu štruktúru v dátach. S flexibilitou nelineárnych metód je však potrebné zaobchádzať opatrne. Aplikovaním veľmi flexibilných metód na trénovacej vzorke dát môže dôjsť k „pretrénovaniu" (angl. *overfitting*). To znamená, že zdanlivo je metóda schopná dobre zachytiť štruktúru v dátach, napríklad odlíšiť pozorovania prislúchajúce rôznym kategóriam. Avšak takéto zobrazenie je zavádzajúce a pri vizualizácii dodatočných testovacích pozorovaní už metóda nebude poskytovať rovnako dobré vizualizácie. Bohatý prehľad lineárnych aj nelineárnych metód je možné nájsť v [4].

Hlavným prínosom tejto práce je predstavenie vlastnej metódy, ktorú je možné klasifikovať ako lineárnu metódu s učiteľom. Naša metóda je zároveň alternatívnym pohľadom na jednu z populárnych metód vizualizácie dát, ktorá tiež patrí medzi lineárne metódy s učiteľom. Z tohto dôvodu v nasledujúcich kapitolách predstavíme vybrané lineárne metódy s učiteľom a bez učiteľa redukujúce priestor nezávislých premenných, ktoré neskôr aplikujeme na dátové súbory a výsledky vizualizácie porovnáme.

3 Lineárne metódy vizualizácie dát bez učiteľa

Lineárne metódy vizualizácie dát bez učiteľa využívajú len informácie obsiahnuté v nezávislých kvantitatívnych premenných. Niektoré z metód majú svoje uplatnenie nielen pri vizualizácii, ale aj v iných oblastiach štatistiky, najčastejšie v metódach redukcie počtu premenných. My sa však zameriame len na formulácie, výhody a nevýhody z pohľadu vizualizácie dát. Tieto metódy môžu byť použité aj na vizualizáciu kategorizovaných dát. Možným nedostatkom je nevyužívanie informácie o kategóriách. V nasledujúcej podkapitole predstavíme vybraného zástupcu týchto metód. Neskôr v podkapitole 4.1 predstavíme aj modifikáciu tejto metódy využívajúcu aj informácie o kategóriách.

3.1 Analýza hlavných komponentov (*PCA*)

V tejto podkapitole budeme vychádzať z [5] a [9].

Analýza hlavných komponentov (angl. *principal component analysis*) je označovaná aj ako *Karhunenova-Loèveho* transformácia. Nech náhodný p-rozmerný vektor $\mathbf{X} = (X_1, \ldots, X_p)^T$ má strednú hodnotu $\boldsymbol{\mu}_1$ a kovariančnú maticu $\boldsymbol{\Sigma}_{11}$, pričom $X_1, \ldots, X_p$ sú kvantitatívne premenné. [1] Nech $\mathbf{u}_1, \ldots, \mathbf{u}_p$ je ortonormálny systém vlastných vektorov $\boldsymbol{\Sigma}_{11}$ a $\lambda_1 \geq \cdots \geq \lambda_p$ sú príslušné vlastné čísla. Pre jednoduchosť predpokladajme, že tieto vlastné čísla sú navzájom rôzne. Hlavné komponenty môžu byť definované rôznymi spôsobmi. My budeme vychádzať z idey, že hlavné komponenty sú ortogonálne projekcie dátových bodov na také smery, v ktorých je rozptyl projekcií dátových bodov maximálny. Takéto smery sú zaujímavé, lebo poskytujú veľa informácie o rozložení dát. Naopak v smeroch, v ktorých majú dáta takmer nulový rozptyl, môžu byť dáta považované za „konštantné" a pre vizualizáciu sú nezaujímavé. Hlavné komponenty sa hľadajú postupne a platí, že j-ty hľadaný hlavný komponent ψ_j musí byť nekorelovaný s predchádzajúcimi hlavnými komponentmi $cov(\psi_j, \psi_i) = 0$, $i < j$. Za predpokladu regularity kovariančnej matice je možné nájsť t hlavných komponentov, $t \leq p$. Pre účely vizualizácie dát v $\mathbb{R}^2$ je potrebné nájsť prvé dva hlavné komponenty. Na komplexnejšiu vizualizáciu je možné využiť aj väčší počet

[1] Označenie strednej hodnoty symbolom $\boldsymbol{\mu}_1$ a kovariančnej matice symbolom $\boldsymbol{\Sigma}_{11}$ pri analýze hlavných komponentov je neštandardné. My sme ho zaviedli kvôli jednoznačnosti v označeniach v ďalších častiach práce.

hlavných komponentov a zobraziť dáta pomocou viacerých dvojíc hlavných komponentov.

Prvý hlavný komponent je taká normalizovaná lineárna kombinácia premenných $X_1, \ldots, X_p$ [2]

$$\psi_1 = \boldsymbol{\alpha}_1^T \mathbf{X} = \alpha_{11} X_1 + \cdots + \alpha_{p1} X_p, \quad ||\boldsymbol{\alpha}_1|| = 1,$$

kde $\boldsymbol{\alpha}_1 = (\alpha_{11}, \ldots, \alpha_{p1})^T$, ktorá má maximálny rozptyl

$$var(\psi_1) = \max_{||\boldsymbol{\alpha}||=1} var(\boldsymbol{\alpha}^T X) = \max_{||\boldsymbol{\alpha}||=1} \left(\boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha} \right).$$

Optimalizačný problém riešime zostrojením Lagrangeovej funkcie

$$L(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha} - \lambda(\boldsymbol{\alpha}^T \boldsymbol{\alpha} - 1),$$

kde λ je Lagrangeov multiplikátor. Derivovaním a položením rovné 0 dostávame sústavu rovníc

$$\left. \frac{\partial L(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_1} = 2\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}_1 - \lambda \boldsymbol{\alpha}_1 = 2(\boldsymbol{\Sigma}_{11} - \lambda \mathbf{I}_p)\boldsymbol{\alpha}_1 = \mathbf{0}_p.$$

Ak $\boldsymbol{\alpha}_1 \neq \mathbf{0}$, potom λ musí byť vlastné číslo matice $\boldsymbol{\Sigma}_{11}$ a $\boldsymbol{\alpha}_1$ príslušný vlastný vektor. Platí

$$var(\psi_1) = \max_{||\boldsymbol{\alpha}||=1} \left(\boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha} \right) = \max_{||\boldsymbol{\alpha}||=1} \left(\boldsymbol{\alpha}^T \lambda \boldsymbol{\alpha} \right) = \max_{||\boldsymbol{\alpha}||=1} \left(\lambda \boldsymbol{\alpha}^T \boldsymbol{\alpha} \right) = \lambda.$$

Z toho vyplýva, že $\boldsymbol{\alpha}_1$ je vlastný vektor prislúchajúci najväčšej vlastnej hodnote matice $\boldsymbol{\Sigma}_{11}$, teda $\boldsymbol{\alpha}_1 = \mathbf{u}_1$ a $\lambda = \lambda_1$. Druhý hlavný komponent ψ_2 hľadáme obdobným spôsobom ako lineárnu kombináciu $\psi_2 = \boldsymbol{\alpha}_2^T \mathbf{X}$ s maximálnym rozptylom, kde $\boldsymbol{\alpha}_2 = (\alpha_{12}, \ldots, \alpha_{p2})^T$ a $||\boldsymbol{\alpha}_2|| = 1$. Pribudne len podmienka nekorelovanosti

$$0 = cov(\psi_2, \psi_1) = cov(\boldsymbol{\alpha}_2^T \mathbf{X}, \boldsymbol{\alpha}_1^T \mathbf{X}) = \boldsymbol{\alpha}_2^T \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha}_1 = \lambda_1 \boldsymbol{\alpha}_2^T \boldsymbol{\alpha}_1,$$

ktorá je ekvivalentná podmienke ortogonálnosti $\boldsymbol{\alpha}_2^T \boldsymbol{\alpha}_1 = 0$. Zostrojíme Lagrangeovu funkciu

$$L(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha} - \dot{\lambda}(\boldsymbol{\alpha}^T \boldsymbol{\alpha} - 1) - \eta(\boldsymbol{\alpha}^T \boldsymbol{\alpha}_1),$$

kde $\dot{\lambda}$ a η sú Lagrangeove multiplikátory. Derivovaním a položením rovné 0 opäť dostávame systém rovníc

$$\left. \frac{\partial L(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_2} = 2(\boldsymbol{\Sigma}_{11} - \dot{\lambda}\mathbf{I}_p)\boldsymbol{\alpha}_2 - \eta \boldsymbol{\alpha}_1 = \mathbf{0}_p.$$

[2] Hlavné komponenty sú lineárne kombinácie centralizovaných premenných $\psi_j = \boldsymbol{\alpha}_j^T(\mathbf{X} - \boldsymbol{\mu}_1)$. Bez ujmy na všeobecnosti predpokladajme $\boldsymbol{\mu}_1 = \mathbf{0}_p$.

Po prenásobení $\boldsymbol{\alpha}_1^T$ vyplýva, že $2\boldsymbol{\alpha}_1^T(\boldsymbol{\Sigma}_{11} - \dot{\lambda}\mathbf{I}_p)\boldsymbol{\alpha}_2 - \eta = 0$. Zo sústavy rovníc $(\boldsymbol{\Sigma}_{11} - \lambda_1\mathbf{I}_p)\boldsymbol{\alpha}_1 = \mathbf{0}_p$ po prenásobení $\boldsymbol{\alpha}_2^T$ vyplýva $\boldsymbol{\alpha}_2^T\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}_1 = 0$, a teda $\eta = 0$. Dôsledok vzťahu $(\boldsymbol{\Sigma}_{11} - \dot{\lambda}\mathbf{I}_p)\boldsymbol{\alpha}_2 = \mathbf{0}_p$ je, že $\dot{\lambda}$ je vlastné číslo matice $\boldsymbol{\Sigma}_{11}$ a $\boldsymbol{\alpha}_2$ príslušný vlastný vektor, presnejšie $\dot{\lambda} = \lambda_2$ a $\boldsymbol{\alpha}_2 = \mathbf{u}_2$. Aj pre ďalšie hlavné komponenty sa dá ukázať, že $\psi_j = \mathbf{u}_j^T\mathbf{X}$.

V praxi sú zvyčajne $\boldsymbol{\mu}_1$ a $\boldsymbol{\Sigma}_{11}$ neznáme a k dispozícii máme iba náhodný výber $\mathbf{X}_1, \ldots, \mathbf{X}_n$, $n \geq p$. Namiesto hlavných komponentov sa na vizualizáciu použijú výberové hlavné komponenty, ktoré vzniknú nahradením priemeru $\boldsymbol{\mu}_1$ výberovým priemerom $\overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i$ a kovariančnej matice $\boldsymbol{\Sigma}_{11}$ výberovou kovariančnou maticou $\mathbf{S}_{11} = \frac{1}{n}\sum_{i=1}^n(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T$. Na vizualizáciu dát v $\mathbb{R}^2$ využijeme prvé dva výberové hlavné komponenty, a to tak, že zobrazíme body v rovine

$$\left(\hat{\psi}_{i1}, \hat{\psi}_{i2}\right),$$

kde $\hat{\psi}_{ij} = \hat{\boldsymbol{\alpha}}_j^T\mathbf{X}_i, i \in \{1, \ldots, n\}, j \in \{1, 2\}$, sú takzvané „skóre". V prípade trojrozmerného zobrazovacieho zariadenia alebo vizualizácie pomocou rotovateľného trojrozmerného grafu zobrazíme body v priestore

$$\left(\hat{\psi}_{i1}, \hat{\psi}_{i2}, \hat{\psi}_{i3}\right),$$

kde $\hat{\psi}_{ij} = \hat{\boldsymbol{\alpha}}_j^T\mathbf{X}_i, i \in \{1, \ldots, n\}, j \in \{1, 2, 3\}$.

Nedostatkom vizualizácie ľubovoľných dát metódou hlavných komponentov je závislosť na škálovaní jednotlivých premenných. To napríklad znamená, že výsledky vizualizácie môže ovplyvniť zmena jednotiek premenných. V prípade, že premenné majú rôznu povahu, napríklad X_1 vyjadruje čas, X_2 vyjadruje dĺžku atď., je odporúčané pred výpočtom hlavných komponentov normalizovať dáta. V prípade, že všetky premenné sú vyjadrené v rovnakých jednotkách, môže byť užitočné pracovať priamo s nenormalizovanými dátami. Dodatočnou nevýhodou pre vizualizáciu kategoriálnych dát je nevyužitie plnej miery informácií obsiahnutej v dátach. Výhodou je ľahká interpretovateľnosť smerov zobrazenia, pretože vznikli iba lineárnou kombináciou pôvodných smerov.

4 Lineárne metódy vizualizácie dát s učiteľom

Metódy predstavené v tejto kapitole využívajú informácie obsiahnuté v nezávislých kvantitatívnych premenných a zároveň informácie obsiahnuté v závislej kategoriálnej premennej. Tieto metódy sa snažia vizualizovať dáta tak, aby boli odzrkadlené rozdiely medzi pozorovaniami z rôznych kategórií. Okrem toho, že pomáhajú získať predstavu o štruktúre dát, majú uplatnenie aj v klasifikačných problémoch. V nasledujúcich podkapitolách uvedieme ich formulácie, výhody a nevýhody len z pohľadu vizualizácie dát.

4.1 Analýza hlavných komponentov pre kategorizované dáta (catPCA)

V tejto podkapitole budeme vychádzať z [10].

Hlavnou myšlienkou analýzy hlavných komponentov predstavenej v podkapitole 3.1 bola projekcia dát na také smery, v ktorých je rozptyl projekcií dátových bodov maximálny. Avšak smery, v ktorých majú dáta najväčší rozptyl, nemusia zodpovedať smerom, ktoré vysvetľujú rozdiely medzi jednotlivými kategóriami. Kritérium maximalizácie rozptylu je ekvivalentné kritériu maximalizácie všetkých nevážených vzájomných vzdialeností medzi projekciami dátových bodov. Modifikácia metódy pre kategorizované dáta spočíva v maximalizácii vážených vzdialeností medzi projekciami dátových bodov. Cieľom metódy je pokúsiť sa rozlíšiť pozorovania z rôznych kategórií, preto je väčšia váha priradená vzdialenostiam medzi dátovými bodmi z rôznych kategórií.

Uvažujme náhodný vektor $\mathbf{X} = (X_1, \ldots, X_p)^T$, kde $X_1, \ldots, X_p$ sú kvantitatívne premenné. Nech $\mathbf{Y}$ je náhodná kategoriálna premenná. Hodnoty kategoriálnej premennej nemusia byť numerické. Ak aj sú, nemôžeme s nimi zaobchádzať ako s číslami, preto nech $\mathbf{Y}$ nadobúda numerické indexy kategórií $\{1, \ldots, k\}$, kde k označuje celkový počet kategórií. Aj v ďalších metódach vizualizácie dát s učiteľom budeme z rovnakých dôvodov uvažovať $\mathbf{Y}$ ako náhodný vektor, ktorý nadobúda numerické indexy kategórií. Nech $\mathbf{X}_1, \ldots, \mathbf{X}_n$ je náhodný výber $\mathbf{X}$ a $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ je príslušný náhodný výber $\mathbf{Y}$. Nech jednotlivé pozorovania tvoria stĺpce matice $\mathbf{M}$, $\mathbf{M} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$. Uvažujme predpoklady a označenia uvedené

v podkapitole 3.1. [3] Pre Euklidovskú vzdialenosť d_{ij} medzi projekciami dátových bodov $\mathbf{X}_i$ a $\mathbf{X}_j$ na smer $\hat{\boldsymbol{\alpha}}$, $||\hat{\boldsymbol{\alpha}}|| = 1$, platí

$$d_{ij}^2(\hat{\boldsymbol{\alpha}}) = \left((\mathbf{M}^T\hat{\boldsymbol{\alpha}})_i - (\mathbf{M}^T\hat{\boldsymbol{\alpha}})_j\right)^2, i, j \in \{1, \ldots, n\}.$$

Nech $\mathbf{x} \in \mathbb{R}^n$ a $\mathbf{L}$ je $n \times n$ Laplacián, teda kladne semidefinitná symetrická matica, v ktorej súčet hodnôt ľubovoľného riadku alebo stĺpca je rovný nule. Teraz odvodíme vlastnosť Laplaciánu, ktorú autori v článku uvádzajú bez odvodenia.

$$\begin{aligned}
\mathbf{x}^T\mathbf{L}\mathbf{x} &= \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j L_{ij} \\
&= \sum_{i<j} x_i x_j L_{ij} + \sum_{i>j} x_i x_j L_{ij} + \sum_{i=1}^{n} x_i^2 L_{ii} \\
&= 2\sum_{i<j} x_i x_j L_{ij} + \sum_{i=1}^{n} x_i^2 \sum_{i<j}(-L_{ij}) + \sum_{i=1}^{n} x_i^2 \sum_{i>j}(-L_{ij}) \\
&= 2\sum_{i<j} x_i x_j L_{ij} - \sum_{i<j} L_{ij} x_i^2 - \sum_{i>j} L_{ij} x_i^2 \\
&= 2\sum_{i<j} x_i x_j L_{ij} - \sum_{i<j} L_{ij} x_i^2 - \sum_{i<j} L_{ij} x_j^2 \\
&= -\sum_{i<j} L_{ij}(x_i - x_j)^2.
\end{aligned}$$

Tento vzťah využijeme v nasledujúcej časti. Uvažujme Laplacián $\mathbf{L}^u$ s prvkami $L_{ij}^u = \delta_{ij}n - 1$, kde δ_{ij} je Kroneckerovo delta, teda $\delta_{ij} = 1$ pre $i = j$, inak $\delta_{ij} = 0$. Potom pre súčet všetkých vzájomných vzdialeností medzi projekciami dátových bodov $\mathbf{X}_1, \ldots, \mathbf{X}_n$ na smer $\hat{\boldsymbol{\alpha}}$ platí

$$\begin{aligned}
\sum_{i<j} d_{ij}^2(\hat{\boldsymbol{\alpha}}) &= -\sum_{i<j} L_{ij}^u d_{ij}^2(\hat{\boldsymbol{\alpha}}) \\
&= -\sum_{i<j} L_{ij}^u \left((\mathbf{M}^T\hat{\boldsymbol{\alpha}})_i - (\mathbf{M}^T\hat{\boldsymbol{\alpha}})_j\right)^2 \\
&= (\mathbf{M}^T\hat{\boldsymbol{\alpha}})^T\mathbf{L}^u(\mathbf{M}^T\hat{\boldsymbol{\alpha}}) \\
&= \hat{\boldsymbol{\alpha}}^T\mathbf{M}\mathbf{L}^u\mathbf{M}^T\hat{\boldsymbol{\alpha}}.
\end{aligned}$$

[3] Podobne ako analýza hlavných komponentov aj analýza hlavných komponentov pre kategorizované dáta hľadá lineárnu transformáciu centralizovaných premenných. Bez ujmy na všeobecnosti predpokladajme, že súčet prvkov v každom riadku matice $\mathbf{M}$ je nulový.

Pre prvky súčinu matíc $\mathbf{M}\mathbf{L}^u\mathbf{M}^T$ platí [4]

$$\begin{aligned}(\mathbf{M}\mathbf{L}^u\mathbf{M}^T)_{kl} &= \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ki}L_{ij}M_{lj} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}(\delta_{ij}n-1)M_{ki}M_{lj} \\ &= \sum_{i=1}^{n} nM_{ki}M_{li} - \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ki}M_{lj} \\ &= n(\mathbf{M}\mathbf{M}^T)_{kl} - \sum_{i=1}^{n} M_{ki}\sum_{j=1}^{n} M_{lj} \\ &= n(\mathbf{M}\mathbf{M}^T)_{kl} \\ &= n^2(\mathbf{S}_{11})_{kl}.\end{aligned}$$

Z toho vyplýva, že kritérium maximalizácie rozptylu $\max_{||\hat{\boldsymbol{\alpha}}||=1}(\hat{\boldsymbol{\alpha}}^T\mathbf{S}_{11}\hat{\boldsymbol{\alpha}})$ je až na konštantu n^2 zhodné s kritériom maximalizácie nevážených vzájomných vzdialeností medzi projekciami dátových bodov. Konštanta však neovplyvňuje výsledok optimalizačnej úlohy ani vizualizácie.

Pri vizualizácii kategorizovaných dát touto metódou kladieme dôraz na zobrazenie odlišnosti medzi kategóriami, pritom zanedbáme vzdialenosti medzi pozorovaniami z rovnakej kategórie. Zvolíme váhy pre vzdialenosti $w_{ij} = 0$, ak $d_{ij}(\hat{\boldsymbol{\alpha}})$ je vzdialenosť medzi pozorovaniami z rovnakej kategórie, teda $\mathbf{Y}_i = \mathbf{Y}_j$, inak $w_{ij} = 1$. Riešime optimalizačnú úlohu

$$\max_{||\hat{\boldsymbol{\alpha}}||=1}\sum_{i<j} w_{ij}(d_{ij}(\hat{\boldsymbol{\alpha}}))^2,$$

ku ktorej pre hľadanie druhého, respektíve tretieho smeru pridáme podmienku kolmosti na predchádzajúce smery, podobne ako v podkapitole 3.1. Pomocou $n \times n$ Laplaciánu $\mathbf{L}^d$ s prvkami

$$L_{ij}^d = \begin{cases} -w_{ij}, \text{ak } i \neq j, \\ \sum_{l=1,l\neq i}^{n} w_{ij}, \text{ak } i = j, \end{cases}$$

upravíme účelovú funkciu do tvaru $\hat{\boldsymbol{\alpha}}\mathbf{M}\mathbf{L}^d\mathbf{M}^T\hat{\boldsymbol{\alpha}}$. Analogicky optimálne smery $\hat{\boldsymbol{\alpha}}_1$ a $\hat{\boldsymbol{\alpha}}_2$ pre vizualizáciu dát v rovine zodpovedajú vlastným vektorom matice $\mathbf{M}\mathbf{L}^d\mathbf{M}^T$ prislúchajúcim k prvým dvom najväčším vlastným hodnotám. Nech $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_p$ je ortonormálny systém

[4] Symbol $\mathbf{S}_{11}$ označuje výberovú kovariančnú maticu definovanú v podkapitole 3.1.

vlastných vektorov matice $\mathbf{M}\mathbf{L}^d\mathbf{M}^T$ a $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ sú príslušné vlastné čísla. Pre jednoduchosť predpokladajme, že sú navzájom rôzne. Potom formálne zapísané platí $\hat{\boldsymbol{\alpha}}_1 = \hat{\mathbf{u}}_1$ a $\hat{\boldsymbol{\alpha}}_2 = \hat{\mathbf{u}}_2$. Vzhľadom na výpočtovú náročnosť ide o ekvivalentný optimalizačný problém. Pri vizualizácií v rovine zobrazíme body

$$\left(\hat{\psi}_{i1}, \hat{\psi}_{i2}\right),$$

kde $\hat{\psi}_{ij} = \hat{\boldsymbol{\alpha}}_j^T \mathbf{X}_i, i \in \{1, \ldots, n\}, j \in \{1, 2\}$. V prípade trojrozmerného zobrazovacieho zariadenia alebo vizualizácie pomocou rotovateľného trojrozmerného grafu zobrazíme body v priestore

$$\left(\hat{\psi}_{i1}, \hat{\psi}_{i2}, \hat{\psi}_{i3}\right),$$

kde $\hat{\psi}_{ij} = \hat{\boldsymbol{\alpha}}_j^T \mathbf{X}_i, i \in \{1, \ldots, n\}, j \in \{1, 2, 3\}$.

Podobne ako pôvodná metóda analýzy hlavných komponentov aj jej modifikovaná verzia pre kategorizované dáta je závislá na škálovaní. Dodatočnou výhodou je využitie informácií aj z kategoriálnej premennej. Dodatočnou nevýhodou je zanedbanie štruktúry dát v rámci rovnakej kategórie.

4.2 Lineárna diskriminačná analýza (LDA)

V nasledujúcej časti budeme vychádzať z [9], pričom popis metódy doplníme o vlastné odvodenia.

Ako vyplýva z názvu, lineárna diskriminačná analýza (angl. *linear discriminant analysis*) sa snaží pomocou lineárnej transformácie zobraziť dáta v nízkorozmernom priestore tak, aby boli pozorovania z rôznych kategórií čo najviac oddelené. Smery, v ktorých sú dáta zobrazené, poskytujú informácie o rozdieloch medzi kategóriami. Princíp metódy spočíva v maximalizácii rozptylu medzi pozorovaniami z rôznych kategórií a zároveň v minimalizácii rozptylu medzi pozorovaniami z rovnakej kategórie.

Uvažujme náhodný vektor $\mathbf{X} = (X_1, \ldots, X_p)^T$, kde $X_1, \ldots, X_p$ sú kvantitatívne premenné, so strednou hodnotou $\boldsymbol{\mu}_1$ a kovariančnou maticou $\boldsymbol{\Sigma}_{11}$. [5] Taktiež uvažujme ná-

[5] Označenie strednej hodnoty symbolom $\boldsymbol{\mu}_1$ a kovariančnej matice symbolom $\boldsymbol{\Sigma}_{11}$ pri lineárnej diskriminačnej analýze je neštandardné. My sme ho zaviedli kvôli jednoznačnosti v označeniach v ďalších častiach práce.

hodnú kategoriálnu premennú $\mathbf{Y}$, ktorá nadobúda numerické indexy kategórií $\{1, \ldots, k\}$, kde k označuje počet kategórií. Kvôli jednoduchosti uvažujme hneď na začiatku náhodný výber $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1}, \ldots, \mathbf{X}_{k1}, \ldots, \mathbf{X}_{kn_k}$ vektora $\mathbf{X}$, pričom $\mathbf{X}_{ij}$ označuje j-te pozorovanie z i-tej kategórie a n_i udáva veľkosť náhodného výberu pre i-tu kategóriu. [6] Celková veľkosť náhodného výberu je $n = \sum_{i=1}^{k} n_i$. Pre náhodný výber vektora $\mathbf{Y}$, teda $\mathbf{Y}_{11}, \ldots, \mathbf{Y}_{1n_1}, \ldots, \mathbf{Y}_{k1}, \ldots, \mathbf{Y}_{kn_k}$, platí $\mathbf{Y}_{ij} = \mathbf{Y}_{il}$ pre $i \in \{1, \ldots, k\}$ a pre $j, l \in \{1, \ldots, n_i\}$. Stredné hodnoty pozorovaní z jednotlivých kategórií odhadneme výberovými priemermi

$$\overline{\mathbf{X}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{X}_{ij}, i \in \{1, \ldots, k\}.$$

Pre celkový výberový priemer a výberovú kovariančnú maticu zo všetkých pozorovaní bez ohľadu na kategóriu platí [7]

$$\overline{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \mathbf{X}_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{X}_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \overline{\mathbf{X}}_i,$$
$$\mathbf{S}_{11} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\mathbf{X}_{ij} - \overline{\mathbf{X}}\right) \left(\mathbf{X}_{ij} - \overline{\mathbf{X}}\right)^T.$$

Keďže ide o lineárnu metódu, opäť hľadáme vhodnú lineárnu kombináciu dát. Zamerajme sa najprv na hľadanie prvého smeru $\hat{\boldsymbol{\alpha}}_1$, na ktorý budeme projektovať dáta.

Čiastočným cieľom lineárnej diskriminačnej analýzy je zobraziť pozorovania v smere, v ktorom sú pozorovania z rôznych kategórií čo možno najviac odseparované. Snahou je maximalizovať „rozptyl projekcií kategórií". Rozptyl projekcií kategórií je definovaný pomocou výberových priemerov a celkového výberového priemeru

$$\hat{\boldsymbol{\alpha}}^T \mathbf{S}_B \hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}^T \left(\frac{1}{n} \sum_{i=1}^{k} n_i \left(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}\right) \left(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}\right)^T \right) \hat{\boldsymbol{\alpha}},$$

kde $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^p$ je smer projekcie.

[6] Štandardne sa pri lineárnej diskriminačnej analýze uvažuje k náhodných vektorov, pričom každý náhodný vektor prislúcha jednej kategórii. V takom prípade sa zvykne dodatočne predpokladať, že všetky náhodné vektory majú rovnakú kovariančnú maticu. My kvôli jednoduchosti v označení uvažujeme jeden náhodný vektor.

[7] Pri vizualizácii zobrazujeme centralizované premenné. Bez ujmy na všeobecnosti predpokladajme $\overline{\mathbf{X}} = \mathbf{0}_p$.

Zároveň druhým čiastočným cieľom lineárnej diskriminačnej analýzy je zobraziť pozorovania z rovnakej kategórie blízko seba. To znamená minimalizovať rozptyl v rámci jednotlivých kategórií. Pre súčet rozptylov projekcií pozorovaní jednotlivých kategórií platí

$$\hat{\boldsymbol{\alpha}}^T \mathbf{S}_W \hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}^T \left(\frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)^T \right) \hat{\boldsymbol{\alpha}}.$$

Spojením týchto čiastočných cieľov môžeme napísať výslednú formuláciu optimalizačného problému

$$\hat{\boldsymbol{\alpha}}_1 = \max_{\hat{\boldsymbol{\alpha}}^T \mathbf{S}_W \hat{\boldsymbol{\alpha}} = 1} \frac{\hat{\boldsymbol{\alpha}}^T \mathbf{S}_B \hat{\boldsymbol{\alpha}}}{\hat{\boldsymbol{\alpha}}^T \mathbf{S}_W \hat{\boldsymbol{\alpha}}}.$$

Pri optimalizácii sa snažíme maximalizovať hodnotu výrazu v čitateli a minimalizovať hodnotu výrazu v menovateli, čo je v súlade s vyššie popísanými cieľmi. Prenásobenie $\hat{\boldsymbol{\alpha}}$ ľubovoľnou nenulovou konštantou neovplyvní hodnotu účelovej funkcie, preto pridáme normalizačnú podmienku $\hat{\boldsymbol{\alpha}}^T \mathbf{S}_W \hat{\boldsymbol{\alpha}} = 1$.

Optimalizačný problém riešime zostrojením Lagrangeovej funkcie

$$L(\hat{\boldsymbol{\alpha}}) = \hat{\boldsymbol{\alpha}}^T \mathbf{S}_B \hat{\boldsymbol{\alpha}} - \lambda \left(\hat{\boldsymbol{\alpha}}^T \mathbf{S}_W \hat{\boldsymbol{\alpha}} - 1 \right),$$

kde λ je Lagrangeov multiplikátor. Derivovaním a položením rovné 0 dostávame sústavu rovníc

$$\left. \frac{\partial L(\hat{\boldsymbol{\alpha}})}{\partial \hat{\boldsymbol{\alpha}}} \right|_{\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_1} = 2\mathbf{S}_B \hat{\boldsymbol{\alpha}}_1 - 2\lambda \mathbf{S}_W \hat{\boldsymbol{\alpha}}_1 = \mathbf{0}_p$$
$$\mathbf{S}_B \hat{\boldsymbol{\alpha}}_1 = \lambda \mathbf{S}_W \hat{\boldsymbol{\alpha}}_1.$$

Predpokladajme, že matica $\mathbf{S}_W$ je regulárna a vlastné hodnoty matice $\mathbf{S}_W^{-1}\mathbf{S}_B$ sú navzájom rôzne. Potom predchádzajúcu sústavu rovníc môžeme upraviť do tvaru

$$\mathbf{S}_W^{-1} \mathbf{S}_B \hat{\boldsymbol{\alpha}}_1 = \lambda \hat{\boldsymbol{\alpha}}_1,$$

z ktorej za predpokladu $\hat{\boldsymbol{\alpha}}_1 \neq \mathbf{0}_p$ vyplýva, že optimálne riešenie je vlastným vektorom matice $\mathbf{S}_W^{-1}\mathbf{S}_B$. Platí

$$\max_{\hat{\boldsymbol{\alpha}}} \hat{\boldsymbol{\alpha}}^T \mathbf{S}_B \hat{\boldsymbol{\alpha}} = \max_{\hat{\boldsymbol{\alpha}}} \lambda \hat{\boldsymbol{\alpha}}^T \mathbf{S}_W \hat{\boldsymbol{\alpha}} = \lambda.$$

Optimálne riešenie je preto vlastný vektor matice $\mathbf{S}_W^{-1}\mathbf{S}_B$ prislúchajúci k najväčšiemu vlastnému číslu. Pri hľadaní druhého smeru $\hat{\boldsymbol{\alpha}}_2$ budeme postupovať obdobne. Pridáme

len dodatočnú podmienku $\hat{\boldsymbol{\alpha}}_1^T \mathbf{S}_W \hat{\boldsymbol{\alpha}}_2 = 0$, ktorá zaručí odlišnosť smerov. Pre $\hat{\boldsymbol{\alpha}}_2$ taktiež uvažujeme normalizačnú podmienku $\hat{\boldsymbol{\alpha}}_2^T \mathbf{S}_W \hat{\boldsymbol{\alpha}}_2 = 1$. Opäť zostrojíme Lagrangeovu funkciu

$$L\left(\hat{\boldsymbol{\alpha}}\right) = \hat{\boldsymbol{\alpha}}^T \mathbf{S}_B \hat{\boldsymbol{\alpha}} - \dot{\lambda}\left(\hat{\boldsymbol{\alpha}}^T \mathbf{S}_W \hat{\boldsymbol{\alpha}} - 1\right) - \eta\left(\hat{\boldsymbol{\alpha}}^T \mathbf{S}_W \hat{\boldsymbol{\alpha}}_1\right),$$

kde $\dot{\lambda}$ a η sú Lagrangeove multiplikátory. Derivovaním a položením rovné 0 dostávame

$$\left.\frac{\partial L\left(\hat{\boldsymbol{\alpha}}\right)}{\partial \hat{\boldsymbol{\alpha}}}\right|_{\hat{\boldsymbol{\alpha}}=\hat{\boldsymbol{\alpha}}_2} = 2\mathbf{S}_B \hat{\boldsymbol{\alpha}}_2 - 2\dot{\lambda}\mathbf{S}_W \hat{\boldsymbol{\alpha}}_2 - \eta \mathbf{S}_W \hat{\boldsymbol{\alpha}}_1 = \mathbf{0}_p.$$

Zo vzťahu $\mathbf{S}_B \hat{\boldsymbol{\alpha}}_1 = \lambda \mathbf{S}_W \hat{\boldsymbol{\alpha}}_1$ po prenásobení $\hat{\boldsymbol{\alpha}}_2^T$ sprava vyplýva $\hat{\boldsymbol{\alpha}}_2^T \mathbf{S}_B \hat{\boldsymbol{\alpha}}_1 = 0$. Tento poznatok využijeme po prenásobení zderivovanej Lagrangeovej funkcie výrazom $\hat{\boldsymbol{\alpha}}_1^T$ sprava. Z toho vyplýva $\eta = 0$ a $\mathbf{S}_B \hat{\boldsymbol{\alpha}}_2 = \dot{\lambda} \mathbf{S}_W \hat{\boldsymbol{\alpha}}_2$. Za predpokladu $\hat{\boldsymbol{\alpha}}_2 \neq \mathbf{0}_p$ je $\hat{\boldsymbol{\alpha}}_2$ tiež vlastný vektor matice $\mathbf{S}_W^{-1}\mathbf{S}_B$, ale tentokrát prináležiaci k druhému najväčšiemu vlastnému číslu. Nech $\hat{\lambda}_1 \geq \hat{\lambda}_2 > 0$ ($\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \hat{\lambda}_3 > 0$) sú prvé 2 (3) najväčšie vlastné čísla matice $\mathbf{S}_W^{-1}\mathbf{S}_B$ a $\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2$ ($\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \hat{\mathbf{u}}_3$) sú prislúchajúce vlastné vektory. Pre jednoduchosť predpokladajme, že spomínané vlastné čísla sú navzájom rôzne. Potom formálne platí $\hat{\boldsymbol{\alpha}}_1 = \hat{\mathbf{u}}_1$, $\hat{\boldsymbol{\alpha}}_2 = \hat{\mathbf{u}}_2$ ($\hat{\boldsymbol{\alpha}}_1 = \hat{\mathbf{u}}_1$, $\hat{\boldsymbol{\alpha}}_2 = \hat{\mathbf{u}}_2$, $\hat{\boldsymbol{\alpha}}_3 = \hat{\mathbf{u}}_3$).

Pri vizualizácii v rovine zobrazíme body

$$\left(\hat{\psi}_{ij1}, \hat{\psi}_{ij2}\right),$$

kde $\hat{\psi}_{ijl} = \hat{\boldsymbol{\alpha}}_l^T \mathbf{X}_{ij}, i \in \{1, \ldots, k\}, j \in \{1, \ldots, n_i\}, l \in \{1, 2\}$. V prípade trojrozmerného zobrazovacieho zariadenia alebo vizualizácie pomocou rotovateľného trojrozmerného grafu zobrazíme body v priestore

$$\left(\hat{\psi}_{ij1}, \hat{\psi}_{ij2}, \hat{\psi}_{ij3}\right),$$

kde $\hat{\psi}_{ijl} = \hat{\boldsymbol{\alpha}}_l^T \mathbf{X}_{ij}, i \in \{1, \ldots, k\}, j \in \{1, \ldots, n_i\}, l \in \{1, 2, 3\}$.

Výhodou lineárnej diskriminačnej analýzy je využívanie informácie aj z kategoriálnej premennej a nezávislosť ani na škálovaní, ani na akejkoľvek inej regulárnej lineárnej transformácii dát. Nevýhodou je zanedbanie štruktúry dát v rámci rovnakej kategórie. Smery, v ktorých zobrazujeme dátové body, nie sú ortogonálne, ale $\mathbf{S}_W$-ortogonálne. Každý ďalší smer môže v sebe zahŕňať „informácie" z predchádzajúcich smerov. Riešením je ortogonalizácia smerov. Hodnosť matice $\mathbf{S}_B$ závisí od počtu kategórií, platí $\mathbf{S}_B \leq k - 1$. V prípade dvoch kategórií nám táto metóda môže poskytnúť iba jeden smer na projekciu dát. Pre účely vizualizácie je v prípade viacerých kategórií toto obmedzenie irelevantné.

4.3 Redukcia dimenzie riadená vnímaním využívajúca konzistenciu vzdialeností s ohľadom na hustotu (*PDD*)

V tejto podkapitole budeme vychádzať z článku [15].

Metóda redukcie dimenzie riadená vnímaním využívajúca konzistenciu vzdialeností s ohľadom na hustotu (angl. *Perception-driven dimension reduction using Density-aware distance consistency*) patrí medzi metódy vychádzajúce z ľudského vnímania. Základnou ideou optimalizačných kritérií tejto triedy metód pre vizualizáciu kategorizovaných dát je vizuálne odseparovať pozorovania z rôznych kategórií. Kritérium konzistencie vzdialeností s ohľadom na hustotu zohľadňuje pre každý dátový bod dve vzdialenosti. Prvou je vzdialenosť projekcie dátového bodu od projekcie priemeru „kategórie", z ktorej dátový bod pochádza. Druhou je vzdialenosť projekcie dátového bodu od projekcie najbližšieho priemeru inej „kategórie". Optimalizáciou kritéria sa snažíme dosiahnuť, aby projekcie dátových bodov ležali blízko projekcie príslušného priemeru a čo možno najďalej od projekcie iného najbližšieho priemeru.

Uvažujme k p-rozmerných náhodných vektorov $\mathbf{X}_i$, ktoré pozostávajú z kvantitatívnych premenných, $i \in \{1, \ldots, k\}$. Nech $\mathbf{X}_i$ má strednú hodnotu $\boldsymbol{\mu}_i$ a kovariančnú maticu $\boldsymbol{\Sigma}_i$. Každý náhodný vektor prislúcha jednej kategórii. Nech $\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i}$ je náhodný výber z rozdelenia zodpovedajúceho rozdeleniu náhodného vektora $\mathbf{X}_i$, pričom n_i udáva veľkosť náhodného výberu pre i-tu kategóriu. Stredné hodnoty jednotlivých náhodných vektorov odhadneme výberovými priemermi

$$\overline{\mathbf{X}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{X}_{ij}, \quad i \in \{1, \ldots, k\}.$$

Na účely vizualizácie dát v rovine hľadáme maticu lineárnej transformácie $\hat{\mathbf{A}}_1 = (\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2)^T$, kde $\hat{\boldsymbol{\alpha}}_l = (\hat{\alpha}_{1l}, \ldots, \hat{\alpha}_{lp})^T$, $||\hat{\boldsymbol{\alpha}}_l|| = 1$, $l \in \{1, 2\}$, pomocou ktorej zobrazíme body v rovine

$$\boldsymbol{\Psi}_{ij} = \hat{\mathbf{A}}_1 \mathbf{X}_{ij}, \quad i \in \{1, \ldots, k\}, j \in \{1, \ldots, n_i\}.$$

Pri hľadaní matice $\hat{\mathbf{A}}_1$ budeme brať do úvahy vzdialenosť projekcií dátových bodov od projekcií výberových priemerov. Uvažujme Euklidovskú vzdialenosť. Nech $a(\boldsymbol{\Psi}_{ij})$ označuje vzdialenosť medzi projekciami dátového bodu $\mathbf{X}_{ij}$ a príslušného výberového priemeru $\overline{\mathbf{X}}_i$. Nech $b(\boldsymbol{\Psi}_{ij})$ označuje vzdialenosť medzi projekciami dátového bodu $\mathbf{X}_{ij}$ a iného najb-

ližšieho výberového priemeru $\overline{\mathbf{X}}_m$, $m \neq i$. Spomínané vzdialenosti formálne vyjadríme

$$a(\mathbf{\Psi}_{ij}) = ||\mathbf{\Psi}_{ij} - \hat{\mathbf{A}}_1\overline{\mathbf{X}}_i||_2,$$
$$b(\mathbf{\Psi}_{ij}) = \min_{m\in\{1,\ldots,k\},m\neq i} ||\mathbf{\Psi}_{ij} - \hat{\mathbf{A}}_1\overline{\mathbf{X}}_m||_2.$$

Kritérium konzistencie vzdialeností s ohľadom na hustotu

$$s(\mathbf{\Psi}_{ij}) = \frac{b(\mathbf{\Psi}_{ij}) - a(\mathbf{\Psi}_{ij})}{\max\{a(\mathbf{\Psi}_{ij}), b(\mathbf{\Psi}_{ij})\}}$$

berie do úvahy rozdiel spomínaných vzdialeností a normalizuje ho maximom z nich. Funkcia $s(\mathbf{\Psi}_{ij})$ nadobúda hodnoty z intervalu $[-1, 1]$. Ak $s(\mathbf{\Psi}_{ij})$ nadobúda hodnotu blízko 1, znamená to, že projekcia dátového bodu $\mathbf{\Psi}_{ij}$ je blízko projekcie výberového priemeru príslušnej kategórie. Nulová hodnota kritéria implikuje, že projekcia dátového bodu je približne rovnako vzdialená od projekcií výberového priemeru príslušnej kategórie a iného najbližšieho výberového priemeru. Hľadanú maticu $\hat{\mathbf{A}}_1$ zvolíme tak, aby maximalizovala priemernú hodnotu kritéria všetkých projekcií dátových bodov

$$\hat{\mathbf{A}}_1 = \underset{\hat{\mathbf{A}}}{\operatorname{argmax}} \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k}\sum_{j=1}^{n_i} s(\mathbf{\Psi}_{ij}).$$

Daný optimalizačný problém nebudeme riešiť analyticky, ale použijeme stochastickú optimalizačnú metódu simulované žíhanie (angl. *simulated annealing*). Nech $E(\hat{\mathbf{A}})$ označuje hodnotu optimalizačného kritéria nadobúdanú v matici $\hat{\mathbf{A}}$

$$E(\hat{\mathbf{A}}) = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k}\sum_{j=1}^{n_i} s(\mathbf{\Psi}_{ij}).$$

Schematicky sú kroky algoritmu pre tento konkrétny problém popísané nižšie.

Postup: Algoritmus simulovaného žíhania pre redukciu dimenzie riadenú vnímaním využívajúcu konzistenciu vzdialeností s ohľadom na hustotu.

Vstup: Náhodné výbery $\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i}$ náhodných vektorov $\mathbf{X}_i$, $i \in \{1, \ldots, k\}$.

1: Náhodne vygenerujeme maticu $\hat{\mathbf{A}} \in \mathbb{R}^{2\times p}$, pričom prvky matice pochádzajú z rovnomerného rozdelenia $\mathcal{R}(-1, 1)$.

2: Zvolíme hodnotu počiatočnej teploty $T = 100p$.

3: Zvolíme hodnoty parametrov, ktoré určujú prehľadávané okolie matice $\hat{\mathbf{A}}$, a to $r = 0.05$ a $m = 0.01$.

4: Zvolíme hodnotu parametra ochladzovania teploty $\alpha = 0.95$.

5: Zvolíme odchýlku teploty $\epsilon = 0.001$. Teplotu nižšiu ako zvolené ϵ budeme pokladať za dostatočne blízko nuly.

6: **while** $T > \epsilon$ **do**

7: Náhodne vygenerujeme matice $\mathbf{Q}_1$, $\mathbf{Q}_2 \in \mathbb{R}^{2\times p}$ v okolí matice $\hat{\mathbf{A}}$. Tieto matice vzniknú z matice $\hat{\mathbf{A}}$ tak, že ku každému prvku pripočítame, respektíve v prípade $\mathbf{Q}_2$ odpočítame náhodné číslo z rozdelenia $rAlt(\frac{1}{2})$.

8: Určíme kandidáta na optimálne riešenie z okolia $\hat{\mathbf{A}}$, $\mathbf{Q} = \underset{\mathbf{A}\in\{\mathbf{Q}_1,\mathbf{Q}_2\}}{\operatorname{argmax}} E(\mathbf{A})$.

9: K prvkom matice $\mathbf{Q}$, pre ktoré platí $\mathbf{Q}_{ij} = \hat{\mathbf{A}}_{ij}, i \in \{1, 2\}, j \in \{1, \ldots, p\}$, pripočítame $m(-1)^{Alt(\frac{1}{2})}$.

10: Riadky matice $\mathbf{Q}$ normalizujeme na jednotkovú dĺžku.

11: **if** $E(\mathbf{Q}) > E(\hat{\mathbf{A}})$ **then**

12: $\hat{\mathbf{A}} = \mathbf{Q}$, lebo matica transformácie $\mathbf{Q}$ dosahuje lepšiu hodnotu kritéria.

13: **else**

14: **if** $e^{\frac{E(\mathbf{Q})-E(\hat{\mathbf{A}})}{T}} > \text{rand}(0, 1)$ **then**

15: $\hat{\mathbf{A}} = \mathbf{Q}$. [8]

16: **end if**

17: **end if**

18: Zredukujeme teplotu $T = \alpha T$.

19: **end while**

20: $\hat{\mathbf{A}}_1 = \hat{\mathbf{A}}$.

[8] Táto podmienka slúži ako prevencia pred zastavením v lokálnom maxime. Má vplyv predovšetkým na začiatku algoritmu, kedy je hodnota teploty T najvyššia.

Po nájdení optimálnej matice lineárnej transformácie $\hat{\mathbf{A}}_1$, zobrazíme pri vizualizácii v rovine body

$$\boldsymbol{\psi}_{ij} = \hat{\mathbf{A}}_1 \mathbf{X}_{ij},$$

kde $i \in \{1, \ldots, k\}, j \in \{1, \ldots, n_i\}$. Ide o lineárnu metódu, ktorá je časovo menej efektívna ako ostatné spomínané metódy. Dôvodom je algoritmické riešenie optimalizačného problému namiesto analytického riešenia. Nevýhodou tejto metódy je jej závislosť na škálovaní. Táto metóda nekladie žiadnu podmienku na ortogonalitu vektorov $\hat{\boldsymbol{\alpha}}_1$ a $\hat{\boldsymbol{\alpha}}_2$. To považujeme taktiež za nevýhodu tejto metódy, lebo druhý smer $\hat{\boldsymbol{\alpha}}_2$ môže poskytovať informácie o dátach, ktoré poskytuje už prvý smer $\hat{\boldsymbol{\alpha}}_1$.

5 Vizualizácia pomocou kanonických korelácií

Táto kapitola je venovaná vizualizácii dát pomocou metódy kanonických korelácií. Prioritné uplatnenie metódy je pri hľadaní súvislosti medzi dvoma súbormi dát. Jej využitie na vizualizáciu dát je neobvyklé. Metódu kanonických korelácií by bolo možné uviesť v kapitole lineárnych metód bez učiteľa. Vzhľadom na to, že predstavíme vlastnú modifikáciu tejto metódy pre vizualizáciu kategorizovaných dát, rozhodli sme sa jej venovať v osobitnej kapitole. Najprv uvedieme základnú formuláciu metódy, z ktorej potom budeme vychádzať pri odvodení našej modifikovanej metódy. V závere kapitoly uvedieme súvis s inými metódami.

5.1 Analýza kanonických korelácií (CCA)

Analýza kanonických korelácií (angl. *canonical correlation analysis*) sa v rámci oblasti vizualizácie dát klasicky používa na vizualizáciu dvoch sád kvantitatívnych premenných. V tejto podkapitole sa budeme venovať tomuto „klasickému" použitiu. Pri vypracovaní tejto podkapitoly sme vychádzali z [6] a [9], pričom sme odvodenia doplnili aj o vlastné výpočty.

Uvažujme náhodný $(p+q)$-rozmerný vektor $\mathbf{X} = (\mathbf{U}, \mathbf{V})^T$, ktorý rozložíme na dva podvektory. Prvý podvektor $\mathbf{U} = (X_1, \ldots, X_p)^T = (U_1, \ldots, U_p)^T$ má dimenziu p a druhý podvektor $\mathbf{V} = (X_{p+1}, \ldots, X_{p+q})^T = (V_1, \ldots, V_q)^T$ má dimenziu q. Náhodný vektor $\mathbf{X}$ má strednú hodnotu

$$E\left\{\begin{pmatrix}\mathbf{U}\\ \mathbf{V}\end{pmatrix}\right\} = \begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}$$

a kovariančnú maticu

$$E\left\{\begin{pmatrix}\mathbf{U}\\ \mathbf{V}\end{pmatrix}\right\} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}.$$

Predpokladajme, že matice $\boldsymbol{\Sigma}_{11}$ a $\boldsymbol{\Sigma}_{22}$ sú regulárne. Základná myšlienka metódy je hľadanie dvojíc smerov pre podvektory $\mathbf{U}$ a $\mathbf{V}$ tak, aby sa maximalizovala korelácia medzi projekciami dát na príslušné smery. Dvojice smerov sa hľadajú postupne. V poradí j-ta dvojica (ψ_j, ω_j) sa nazýva j-ty pár kanonických premenných. Je možné nájsť t párov kanonických premenných, $t \leq \min(p, q)$. Platí, že j-ta kanonická premenná je nekorelovaná

s predchádzajúcimi kanonickými premennými prislúchajúcimi tomu istému podvektoru, teda

$$cov(\psi_j, \psi_i) = 0, i < j,$$

$$cov(\omega_j, \omega_i) = 0, i < j.$$

Na vizualizáciu dát sa využívajú celé páry kanonických premenných, preto je možné dáta vizualizovať iba v priestore $\mathbb{R}^2$. Najčastejšie sa na vizualizáciu dát používa len prvý pár kanonických premenných. Hypoteticky je možné na komplexnejšiu vizualizáciu využiť aj viacero párov a zobraziť pre každý pár príslušnú rovinnú vizualizáciu. Naša myšlienka vizualizácie kategorizovaných dát, ktorú uvedieme neskôr, je podstatne iná a umožňuje nám dáta bez problémov vizualizovať nielen formou zobrazenia v rovine, respektíve viacerých zobrazení v rovine, ale aj v jednorozmernom či trojrozmernom priestore.

Prvý pár kanonických premenných je taká lineárna kombinácia premenných

$$\psi_1 = \boldsymbol{\alpha}_1^T \mathbf{U} = \alpha_{11} U_1 + \cdots + \alpha_{p1} U_p,$$

$$\omega_1 = \boldsymbol{\beta}_1^T \mathbf{V} = \beta_{11} V_1 + \cdots + \beta_{q1} V_q,$$

kde $\boldsymbol{\alpha}_1 = (\alpha_{11}, \ldots, \alpha_{p1})^T \neq \mathbf{0}_p$ a $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{q1})^T \neq \mathbf{0}_q$, ktorá maximalizuje koreláciu[9]

$$\begin{aligned} corr(\psi_1, \omega_1) &= corr(\boldsymbol{\alpha}_1^T \mathbf{U}, \boldsymbol{\beta}_1^T \mathbf{V}) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} corr(\boldsymbol{\alpha}^T \mathbf{U}, \boldsymbol{\beta}^T \mathbf{V}) \\ &= \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \frac{\boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{12} \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha}} \sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_{22} \boldsymbol{\beta}}}. \end{aligned}$$

Prenásobenie vektorov $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ konštantou nemá vplyv na výsledok. Nech $\boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha} = 1$ a $\boldsymbol{\beta}^T \boldsymbol{\Sigma}_{22} \boldsymbol{\beta} = 1$, potom

$$corr(\psi_1, \omega_1) = \max_{\boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha} = 1, \boldsymbol{\beta}^T \boldsymbol{\Sigma}_{22} \boldsymbol{\beta} = 1} \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{12} \boldsymbol{\beta}.$$

Optimalizačný problém riešime zostrojením Lagrangeovej funkcie

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{12} \boldsymbol{\beta} - \frac{\lambda_1}{2} \left(\boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha} \right) - \frac{\lambda_2}{2} \left(\boldsymbol{\beta}^T \boldsymbol{\Sigma}_{22} \boldsymbol{\beta} \right),$$

kde λ_1 a λ_2 sú Lagrangeove multiplikátory. Hodnota derivácií funkcií v optimálnom riešení musí byť nulová. Získame systém rovníc

$$\left. \frac{\partial L(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{\alpha}} \right|_{\boldsymbol{\alpha} = \boldsymbol{\alpha}_1, \boldsymbol{\beta} = \boldsymbol{\beta}_1} = \boldsymbol{\Sigma}_{12} \boldsymbol{\beta}_1 - \lambda_1 \boldsymbol{\Sigma}_{11} \boldsymbol{\alpha}_1 = \mathbf{0}_p, \tag{1}$$

[9] Kvôli jednoduchosti a bez ujmy na všeobecnosti predpokladajme $E(\mathbf{X}) = \mathbf{0}_{p+q}$.

$$\left.\frac{\partial L(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_1,\boldsymbol{\beta}=\boldsymbol{\beta}_1} = \boldsymbol{\Sigma}_{21}\boldsymbol{\alpha}_1 - \lambda_2\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}_1 = \mathbf{0}_q. \tag{2}$$

Rovnice prenásobíme zľava $\boldsymbol{\alpha}_1^T$, respektíve $\boldsymbol{\beta}_1^T$

$$\boldsymbol{\alpha}_1^T\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}_1 = \lambda_1\boldsymbol{\alpha}_1^T\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}_1,$$

$$\boldsymbol{\beta}_1^T\boldsymbol{\Sigma}_{21}\boldsymbol{\alpha}_1 = \lambda_2\boldsymbol{\beta}_1^T\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}_1.$$

Po odčítaní rovníc z toho vyplýva, že $\lambda_1 = \lambda_2 := \lambda$. V úvode sme predpokladali regulárne matice $\boldsymbol{\Sigma}_{11}$ a $\boldsymbol{\Sigma}_{22}$, preto môžeme vyjadriť

$$\boldsymbol{\alpha}_1 = \frac{1}{\lambda}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}_1, \tag{3}$$

$$\boldsymbol{\beta}_1 = \frac{1}{\lambda}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\alpha}_1 \tag{4}$$

a dosadiť (3) do (2) a (4) do (1). Po úprave sme sa dopracovali k vzťahom

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\alpha}_1 = \lambda^2\boldsymbol{\alpha}_1,$$

$$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}_1 = \lambda^2\boldsymbol{\beta}_1.$$

Ak $\boldsymbol{\alpha}_1 \neq \mathbf{0}_p$ a $\boldsymbol{\beta}_1 \neq \mathbf{0}_q$, potom $\boldsymbol{\alpha}_1$ je vlastný vektor matice $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ prislúchajúci k vlastnému číslu λ^2 a $\boldsymbol{\beta}_1$ je vlastný vektor matice $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ tiež prislúchajúci k rovnakému vlastnému číslu. Po dosadení (3) platí

$$\begin{aligned} corr(\psi_1,\omega_1) &= \max_{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}_{11}\boldsymbol{\alpha}=1,\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}=1} \boldsymbol{\alpha}^T\boldsymbol{\Sigma}_{12}\boldsymbol{\beta} = \max_{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}=1} \frac{1}{\lambda}\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\beta} \\ &= \max_{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}=1} \frac{1}{\lambda}\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\beta} \\ &= \max_{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}=1} \frac{1}{\lambda}\lambda^2\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{22}\boldsymbol{\beta} = \lambda. \end{aligned}$$

Z toho vyplýva, že $\boldsymbol{\alpha}_1$ je vlastný vektor matice $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ prislúchajúci k najväčšiemu vlastnému číslu, $\boldsymbol{\beta}_1$ je vlastný vektor matice $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ prislúchajúci k najväčšiemu vlastnému číslu a prvá kanonická korelácia $\rho_1 = corr(\psi_1,\omega_1) = \lambda = \sqrt{\lambda^2}$ je rovná odmocnine z daného najväčšieho vlastného čísla.

Ďalšie páry kanonických premenných sa hľadajú obdobným spôsobom. Dá sa ukázať, že pre j-ty pár kanonických premenných (ψ_j,ω_j) platí $\psi_j = \boldsymbol{\alpha}_j^T\mathbf{U}$, kde $\boldsymbol{\alpha}_j$ je vlastný vektor matice $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ prislúchajúci k j-temu najväčšiemu vlastnému číslu, $\omega_j =$

$\boldsymbol{\beta}_j^T\mathbf{V}$, kde $\boldsymbol{\beta}_j$ je vlastný vektor matice $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ prislúchajúci k j-temu najväčšiemu vlastnému číslu a j-ta kanonická korelácia $\rho_j = corr(\psi_j, \omega_j)$ je rovná odmocnine z daného j-teho najväčšieho vlastného čísla.

V praxi sú zvyčajne $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{12}$, $\boldsymbol{\Sigma}_{21}$ a $\boldsymbol{\Sigma}_{22}$ neznáme a k dispozícii máme iba náhodný výber $\mathbf{X}_1, \ldots, \mathbf{X}_n$, $n \geq (p+q)$. Namiesto kanonických premenných sa na vizualizáciu použijú výberové kanonické korelácie, ktoré vzniknú nahradením kovariančnej matice s podmaticami $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{12}$, $\boldsymbol{\Sigma}_{21}$, $\boldsymbol{\Sigma}_{22}$ výberovou kovariančnou maticou

$$\mathbf{S}_n = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}.$$

Na „klasickú" vizualizáciu dát v $\mathbb{R}^2$ využijeme prvý pár výberových kanonických premenných, a to tak, že zobrazíme body v rovine

$$\left(\hat{\psi}_{i1}, \hat{\omega}_{i1}\right),$$

kde $\hat{\psi}_{i1} = \hat{\alpha}_1^T\mathbf{U}_i, \hat{\omega}_{i1} = \hat{\beta}_1^T\mathbf{V}_i, i \in \{1, \ldots, n\}$. V prípade, že chceme využiť t párov výberových kanonických premenných, zobrazíme pre každé j v samostatných rovinách body

$$\left(\hat{\psi}_{ij}, \hat{\omega}_{ij}\right),\ i \in \{1, \ldots, n\},$$

kde $\hat{\psi}_{ij} = \hat{\alpha}_j^T\mathbf{U}_i, \hat{\omega}_{ij} = \hat{\beta}_j^T\mathbf{V}_i, i \in \{1, \ldots, n\}, j \in \{1, \ldots, t\}$. Výhodou tejto metódy je jej nezávislosť na škálovaní a ľahká interpretovateľnosť smerov zobrazenia, pretože vznikli iba lineárnou kombináciou pôvodných smerov.

5.2 Vizualizácia kategorizovaných dát pomocou regulárneho simplexového kódovania a kanonických korelácií (catCCA)

Našim zámerom je modifikovať analýzu kanonických korelácií pre kategorizované dáta. Cieľom je vytvoriť metódu, ktorá bude využívať aj informácie o kategóriach. V analýze kanonických korelácií vystupujú dve sady premenných $\mathbf{U}$ a $\mathbf{V}$. Základnou myšlienkou modifikovanej metódy je vložiť do $\mathbf{U}$ informácie obsiahnuté v kvantitatívnych vysvetľujúcich premenných, do $\mathbf{V}$ vložiť informácie obsiahnuté v kategoriálnej vysvetľovanej premennej a na vizualizáciu použiť prvé dve, prípadne tri kanonické premenné prislúchajúce $\mathbf{U}$.

5.2.1 Spracovanie dát

Uvažujme náhodný vektor $\mathbf{X} = (X_1, \ldots, X_p)^T$, kde $X_1, \ldots, X_p$ sú kvantitatívne premenné, a náhodnú kategoriálnu premennú $\mathbf{Y}$, ktorá reprezentuje numerické indexy kategórií a nadobúda hodnoty $1, \ldots, k$, kde k označuje počet kategórií. Všetky kvantitatívne premenné vložíme do jednej sady premenných $\mathbf{U} = (X_1, \ldots, X_p)^T$. Pre kategoriálnu premennú vytvoríme pomocné premenné $Z_1, \ldots, Z_q$, v ktorých bude zahrnutá informácia o kategóriach, a tie vložíme do druhej sady premenných $\mathbf{V} = (Z_1, \ldots, Z_q)^T$. Určiť tvar a ani počet pomocných premenných nie je priamočiare. Je potrebné vytvoriť aspoň dve pomocné premenné, aby sme vedeli nájsť dve kanonické premenné prislúchajúce kvantitatívnym premenným.

Prvým spôsobom ako vytvoriť pomocné premenné je zakódovať informáciu o kategóriach pomocou *one hot encoding*. Pre k kategórií vytvoríme príslušných $q = k$ pomocných premenných, ktoré nadobúdajú hodnotu 1, ak pozorovanie patrí do danej kategórie, inak nadobúdajú hodnotu 0. To znamená, že pre pozorovanie z j-tej kategórie

$$\begin{aligned} &Z_j = 1, \\ &Z_i = 0,\ i \in \{1, \ldots, j-1, j+1, \ldots, q\}. \end{aligned}$$

Označme n_j počet pozorovaní v j-tej kategórii a n celkový počet pozorovaní, $n = \sum_{j=1}^{q} n_j$. Potom pre zložky kovariančnej podmatice $(\mathbf{\Sigma}_{22})_{i,j} = cov(Z_i, Z_j)$ platí

$$\begin{aligned} cov(Z_i, Z_j) &= \mathbf{E}((Z_i - \mathbf{E}(Z_i))(Z_j - \mathbf{E}(Z_j))) \\ &= \frac{n_i(1 - \frac{n_i}{n})(0 - \frac{n_j}{n}) + n_j(0 - \frac{n_i}{n})(1 - \frac{n_j}{n}) + (n - n_i - n_j)(0 - \frac{n_i}{n})(0 - \frac{n_j}{n})}{n} \\ &= \frac{-n_i n_j}{n^2},\ i \neq j, \\ cov(Z_j, Z_j) &= \mathbf{E}((Z_j - \mathbf{E}(Z_j))(Z_j - \mathbf{E}(Z_j))) \\ &= \frac{n_j(1 - \frac{n_j}{n})^2 + (n - n_j)(0 - \frac{n_j}{n})^2}{n} \\ &= \frac{n_j(n - n_j)}{n^2}. \end{aligned}$$

Kovariančnú podmaticu $\mathbf{\Sigma}_{22}$ upravíme do nasledujúceho tvaru

$$\boldsymbol{\Sigma}_{22} = \frac{1}{n^2}\begin{pmatrix} n_1(n-n_1) & -n_1n_2 & \dots & -n_1n_q \\ -n_1n_2 & n_2(n-n_2) & \dots & -n_2n_q \\ \vdots & \vdots & \ddots & \vdots \\ -n_1n_q & -n_2n_q & \dots & n_q(n-n_q) \end{pmatrix}$$
$$= \frac{1}{n}\begin{pmatrix} n_1 & 0 & \dots & 0 \\ 0 & n_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & n_q \end{pmatrix}\left[\mathbf{I}_q + \begin{pmatrix} \frac{-1}{n} \\ \frac{-1}{n} \\ \vdots \\ \frac{-1}{n} \end{pmatrix}\begin{pmatrix} n_1 & n_2 & \dots & n_q \end{pmatrix}\right]. \tag{5}$$

Nech $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, potom sa dá dokázať, že platí $\det(\mathbf{I}_n + \mathbf{u}\mathbf{v}^T) = 1 + \mathbf{u}^T\mathbf{v}$ ([7], dôsledok vety 13.3.8). Zároveň pre matice $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ a konštantu $c \in \mathbb{R}$ platí $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$, $\det(c\mathbf{A}) = c^n \det(\mathbf{A})$. Aplikovaním týchto vzťahov vypočítame $\det(\boldsymbol{\Sigma}_{22})$

$$\det(\boldsymbol{\Sigma}_{22}) = \frac{1}{n^q}\left(\prod_{i=1}^{q} n_i\right)\left(1 - \frac{1}{n}\sum_{i=1}^{q} n_i\right) = 0, \tag{6}$$

pričom sme využili $\sum_{j=1}^{q} n_j = n$. Kovariančná podmatica $\boldsymbol{\Sigma}_{22}$ je singulárna. Nemôžeme teda použiť postup uvedený v časti 5.1, lebo nie je splnený predpoklad regulárnosti matice. Tento problém by bolo možné vyriešiť perturbáciou kovariančnej matice alebo využitím pseudoinverzií, a tak zabezpečiť regulárnosť, ale týmto smerom sme ďalej neuvažovali.

Druhým spôsobom je vytvoriť pomocné premenné v tvare *dummy* premenných. Prvým krokom je voľba referenčnej kategórie a pre zvyšných $(k-1)$ kategórií vytvoríme príslušných $q = (k-1)$ pomocných premenných, ktoré nadobúdajú hodnotu 1, ak pozorovanie patrí do danej kategórie, inak nadobúdajú hodnotu 0. Nech k-ta kategória je referenčná, potom pre pozorovanie z j-tej kategórie, $j \neq k$, platí

$$\begin{aligned} &Z_j = 1, \\ &Z_i = 0,\ i \in \{1, \dots, j-1, j+1, \dots, q\}. \end{aligned} \tag{7}$$

Pozorovanie z referenčnej kategórie nadobúda hodnotu 0 vo všetkých premenných

$$Z_i = 0,\ i = 1, \dots, q.$$

Označme n_j počet pozorovaní v j-tej kategórii a n celkový počet pozorovaní $n = \sum_{j=1}^{q} n_j +$

n_k. Pre $\det(\mathbf{\Sigma}_{22})$ platí (6), ale $n \neq \sum_{j=1}^{q} n_j$, teda

$$\det(\mathbf{\Sigma}_{22}) = \frac{1}{n^q}\left(\prod_{i=1}^{q} n_i\right)\left(1 - \frac{n - n_k}{n}\right) \neq 0.$$

Aplikácia tohto prístupu odstráni problém so singulárnosťou kovariančnej podmatice $\mathbf{\Sigma}_{22}$. Problém voľby referenčnej kategórie by vo všeobecnosti mohol spôsobovať nejednoznačnosť výsledkov. Tento problém sa v prípade *one hot encoding* nevyskytol. [10]

Našou snahou je vytvoriť pomocné premenné tak, aby sme odstránili nedostatky prvého aj druhého prístupu. K vhodným pomocným premenným sa dopracujeme pomocou grafického znázornenia predchádzajúcich prístupov. Budeme vychádzať z definície simplexu.

Definícia 5.1. *(Simplex) Nech $k+1$ bodov $\mathbf{v}_0, \ldots, \mathbf{v}_k \in \mathbb{R}^n$ je afinne nezávislých. To znamená, že vektory $\mathbf{v}_1 - \mathbf{v}_0, \ldots, \mathbf{v}_k - \mathbf{v}_0$ sú lineárne nezávislé. Potom simplex definovaný týmito bodmi je*

$$C = conv\{\mathbf{v}_0, \ldots, \mathbf{v}_k\} = \{\alpha_0\mathbf{v}_0 + \cdots + \alpha_k\mathbf{v}_k | \alpha \geq 0, \mathbf{1}^T\boldsymbol{\alpha} = 1\},$$

kde conv označuje konvexný obal množiny. Afinná dimenzia simplexu je k, preto sa niekedy označuje aj ako k-rozmerný simplex v $\mathbb{R}^n$. [2]

Kvôli názornosti uvažujme kategoriálnu premennú s $k = 3$ kategóriami. V prípade *dummy* pomocných premenných sa na hodnoty $\mathbf{V}$ pre rôzne kategórie môžeme pozerať ako na vrcholy dvojrozmerného simplexu $\mathbf{v}_0 = (0,0)^T, \mathbf{v}_1 = (1,0)^T, \mathbf{v}_2 = (0,1)^T$ v $\mathbb{R}^2$. Grafické znázornenie je na Obr. 1. Vektory smerujúce zo stredu simplexu $\mathbf{c} = \frac{1}{k+1}\sum_{i=0}^{k}\mathbf{v}_i$ k jednotlivým vrcholom nemajú rovnakú dĺžku

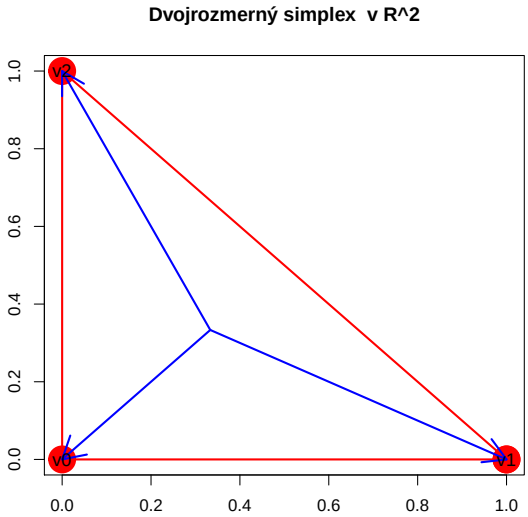
$$\exists i, j \in \{0, \ldots, k\}, i \neq j : ||\mathbf{v}_i - \mathbf{c}|| \neq ||\mathbf{v}_j - \mathbf{c}||$$

a nezvierajú rovnaký uhol s ostatnými vektormi

$$\exists i, j, l \in \{0, \ldots, k\}, i \neq j \neq l : \angle\mathbf{v}_i\mathbf{c}\mathbf{v}_j \neq \angle\mathbf{v}_i\mathbf{c}\mathbf{v}_l.$$

Jednotlivé dvojice vrcholov nemajú rovnakú vzájomnú polohu. Presnejšie, vzájomná poloha vrcholu prislúchajúcemu referenčnej kategórii a ostatných vrcholov je iná ako vzájomná poloha ostatných vrcholov navzájom. To, že kategórie nie sú rovnocenné, je príčinou prečo voľba referenčnej kategórie môže vo všeobecnosti ovplyvniť výsledok vizualizácie.

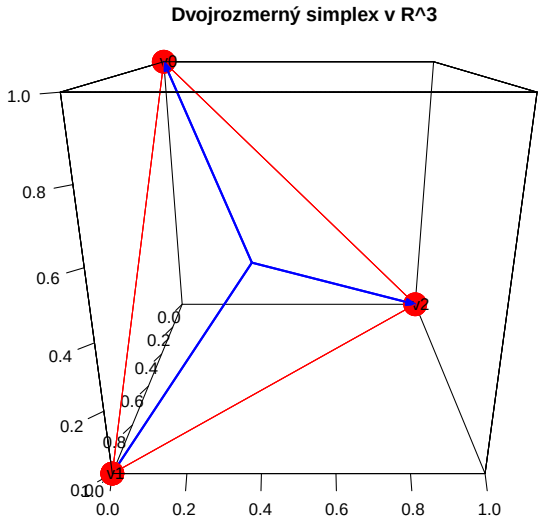
[10] V ďalšej časti práce ukážeme, že nutnosť voľby referenčnej kategórie nemá vplyv na jednoznačnosť výsledkov. Dôvodom sú vlastnosti metódy kanonických korelácií. Vo všeobecnosti to však pre ľubovoľnú metódu platiť nemusí.



Obr. 1: Dvojrozmerný simplex, ktorého vrcholy reprezentujú hodnoty $\mathbf{V}$ pre jednotlivé kategórie reprezentované *dummy* pomocnými premennými. Modré šípky znázorňujú smerové vektory zo stredu simplexu k jednotlivým vrcholom.

V prípade prístupu *one hot encoding* sa na hodnoty $\mathbf{V}$ pre rôzne kategórie môžeme pozerať ako na vrcholy dvojrozmerného simplexu $\mathbf{v}_0 = (1,0,0)^T, \mathbf{v}_1 = (0,1,0)^T, \mathbf{v}_2 = (0,0,1)^T$ v $\mathbb{R}^3$. Dimenzia priestoru je o jeden väčšia ako minimálna dimenzia priestoru potrebná pre k-rozmerný simplex, teda $n = k + 1$. Grafické znázornenie je na Obr. 2. Všetky vektory smerujúce zo stredu simplexu k jednotlivým vrcholom majú rovnakú dĺžku a zvierajú navzájom rovnaký uhol. Simplex s danou vlastnosťou nazývame regulárny simplex. Pre tieto vrcholy platí, že majú medzi sebou rovnakú vzájomnú polohu. Kategórie sú rovnocenné. To, že existuje k-rozmerný podpriestor, v ktorom ležia všetky vrcholy, v tomto prípade rovina, je zase dôvodom, prečo je pri použití *one hot encoding* podmatica $\mathbf{\Sigma}_{22}$ singulárna.

Našu metódu vytvárania pomocných premenných nazveme *regulárne simplexové kódovanie*. Pre k kategórií zostrojíme $(k-1)$-rozmerný regulárny simplex v $\mathbb{R}^{k-1}$, teda v minimálnej dimenzii pre $(k-1)$-rozmerný simplex. Smerové vektory zo stredu simplexu k jednotlivým vrcholom s rovnakou veľkosťou budeme pokladať za realizácie $\mathbf{V}$ pre jednotlivé kategórie. Pre k kategórií zostrojíme $q = k - 1$ pomocných premenných.



Obr. 2: Dvojrozmerný simplex, ktorého vrcholy reprezentujú hodnoty $\mathbf{V}$ pre jednotlivé kategórie reprezentované pomocou *one hot encoding*. Modré šípky znázorňujú smerové vektory zo stredu simplexu k jednotlivým vrcholom.

Lema 5.2. *(Vektory smerujúce k vrcholom k-rozmerného regulárneho simplexu v* $\mathbb{R}^k$ *) Nech* $\mathbf{u}_1, \ldots, \mathbf{u}_k$ *sú vlastné vektory úplne symetrickej matice*

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & \ldots & -1 \\ -1 & 1 & \ldots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \ldots & 1 \end{pmatrix}_{(k+1)\times(k+1)}$$

prislúchajúce k vlastnému číslu $\lambda = 2$ *a nech tvoria ortonormálny systém, pričom* $\mathbf{u}_i = (u_{1i}, \ldots, u_{(k+1)i})^T, i \in \{1, \ldots, k\}$. *Nech stĺpce matice* $\mathbf{W}$ *sú tvorené týmito vektormi*

$$\mathbf{W} = \begin{pmatrix} \mathbf{u}_1 & \ldots & \mathbf{u}_k \end{pmatrix}_{(k+1)\times k}.$$

Potom vektory $\mathbf{v}_i = (u_{(i+1)1}, \ldots, u_{(i+1)k})^T \in \mathbb{R}^k, i \in \{0, \ldots, k\}$, *ktoré tvoria riadky matice* $\mathbf{W}$, *smerujú k vrcholom k-rozmerného regulárneho simplexu v* $\mathbb{R}^k$.

Dôkaz. Pre vlastné čísla matice platí $\lambda(\mathbf{A} + c\mathbf{I}) = \lambda(\mathbf{A}) + c$, kde $c \in \mathbb{R}$ je konštanta. Nech $c = -2$, potom matica $\mathbf{A} - 2\mathbf{I}$ je singulárna symetrická matica s hodnosťou $h(\mathbf{A} - 2\mathbf{I}) = 1$. Pre symetrické matice platí, že počet nenulových vlastných čísel zodpovedá hodnosti matice. Z toho vyplýva, že algebraická násobnosť $\lambda = 0$ je $(k+1) - h(\mathbf{A} - 2\mathbf{I}) = k$, označme

$\lambda_1 = \cdots = \lambda_k = 0$. Vychádzajúc z vlastnosti stopy $tr(\mathbf{A} - 2\mathbf{I}) = \sum_{i=1}^{k+1} \lambda_i$ určíme poslednú neznámu hodnotu vlastného čísla $\lambda_{k+1} = -(k+1)$. Pre vlastné čísla matice $\mathbf{A}$ potom platí $\lambda_1 = \cdots = \lambda_k = 2$, $\lambda_{k+1} = 1-k$. Pre symetrické matice sa algebraická násobnosť vlastných čísel rovná geometrickej násobnosti. Existuje k vlastných vektorov $\mathbf{u}_1, \ldots, \mathbf{u}_k$ prislúchajúcich vlastnému číslu $\lambda = 2$. Z definície vlastných vektorov $(\mathbf{A} - 2\mathbf{I})\mathbf{u}_i = 0$ vyplýva $-\mathbf{1}_{k+1}^T \mathbf{u}_i = 0, i \in \{1, \ldots, k\}$, preto vektory $\mathbf{u}_i$ môžu byť ľubovoľné nenulové nezávislé vektory kolmé na $-\mathbf{1}_{k+1}$. Zvoľme tieto vektory tak, aby tvorili ortonormálny systém. Vlastný vektor jednotkovej dĺžky prislúchajúci k $\lambda = 1 - k$ má tvar $\mathbf{u}_{k+1} = \pm(\frac{1}{\sqrt{k+1}}, \ldots, \frac{1}{\sqrt{k+1}})^T$. Je známe tvrdenie, že vlastné vektory prislúchajúce k rôznym vlastným hodnotám symetrickej matice sú navzájom kolmé. Všetky nami zvolené vlastné vektory matice $\mathbf{A}$ teda tvoria ortonormálny systém. Vektory $\mathbf{v}_i = (u_{(i+1)1}, \ldots, u_{(i+1)k})^T$, $i \in \{0, \ldots, k\}$, smerujú k vrcholom k-rozmerného regulárneho simplexu práve vtedy, ak každý vektor má rovnakú dĺžku a navzájom zvierajú rovnaké uhly. Ekvivalentné podmienky sú

$$||\mathbf{v}_i|| = ||\mathbf{v}_j||, \forall i, j \in \{0, \ldots, k\} \iff \mathbf{v}_i^T \mathbf{v}_i = \mathbf{v}_j^T \mathbf{v}_j, \forall i, j \in \{0, \ldots, k\},$$
$$\text{vektory zvierajú rovnaké uhly} \iff \mathbf{v}_i^T \mathbf{v}_j = \mathbf{v}_i^T \mathbf{v}_l, \forall i, j, l \in \{0, \ldots, k\}.$$

Matica $\mathbf{W}$, ktorej stĺpce sú tvorené vektormi $\mathbf{u}_i, i = 1, \ldots, k+1$,

$$\mathbf{W} = \begin{pmatrix} u_{11} & u_{12} & \ldots & u_{1k} & u_{1(k+1)} \\ u_{21} & u_{22} & \ldots & u_{2k} & u_{2(k+1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ u_{k1} & u_{k2} & \ldots & u_{kk} & u_{k(k+1)} \\ u_{(k+1)1} & u_{(k+1)2} & \ldots & u_{(k+1)k} & u_{(k+1)(k+1)} \end{pmatrix}$$

je ortogonálna. Ortogonálne matice spĺňajú $\mathbf{I}_{k+1} = \mathbf{W}\mathbf{W}^T$. Ekvivalentný zápis je

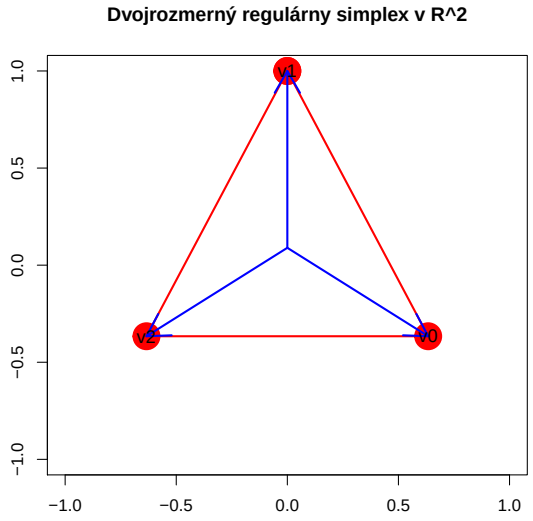
$$\mathbf{I}_{k+1} = \begin{pmatrix} \mathbf{v}_0^T & u_{1(k+1)} \\ \mathbf{v}_1^T & u_{2(k+1)} \\ \vdots & \vdots \\ \mathbf{v}_{k-1}^T & u_{k(k+1)} \\ \mathbf{v}_k^T & u_{(k+1)(k+1)} \end{pmatrix} \begin{pmatrix} \mathbf{v}_0 & \mathbf{v}_1 & \ldots & \mathbf{v}_{k-1} & \mathbf{v}_k \\ u_{1(k+1)} & u_{2(k+1)} & \ldots & u_{k(k+1)} & u_{(k+1)(k+1)} \end{pmatrix}.$$

Využijúc skôr definovaný tvar vektora $\mathbf{u}_{k+1}$ dospejeme k vlastnostiam vektorov $\mathbf{v}_i$

$$\begin{pmatrix} \mathbf{v}_0^T \\ \vdots \\ \mathbf{v}_k^T \end{pmatrix} \begin{pmatrix} \mathbf{v}_0 & \dots & \mathbf{v}_k \end{pmatrix} = \begin{pmatrix} \frac{k}{k+1} & -\frac{1}{k+1} & \dots & -\frac{1}{k+1} & \\ -\frac{1}{k+1} & \frac{k}{k+1} & \dots & -\frac{1}{k+1} & \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{1}{k+1} & -\frac{1}{k+1} & -\frac{1}{k+1} & \frac{k}{k+1} & \end{pmatrix}_{(k+1)\times(k+1)} .$$

□

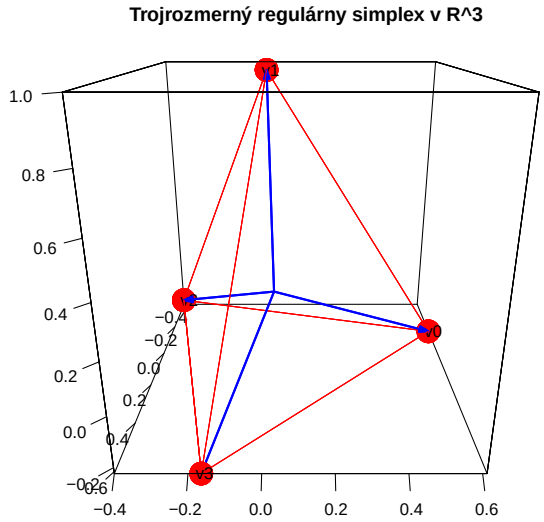
Vektory smerujúce k vrcholom dvojrozmerného regulárneho simplexu v $\mathbb{R}^2$ a trojrozmerného regulárneho simplexu v $\mathbb{R}^3$ sú zobrazené na Obr. 3 a Obr. 4.



Obr. 3: Vektory smerujúce k vrcholom dvojrozmerného regulárneho simplexu v $\mathbb{R}^2$.

5.2.2 Praktický návod na aplikáciu metódy.

V predchádzajúcej podkapitole sme plynule prešli z metódy vizualizácie kvantitatívnych dát pomocou kanonických korelácií na našu metódu vizualizácie kategorizovaných dát pomocou regulárneho simplexového kódovania a kanonických korelácií. Zamerali sme sa v nej najmä na spracovanie kategoriálnej premennej a sled myšlienok, ktoré nás doviedli k zavedeniu regulárneho simplexového kódovania. V tejto podkapitole uvedieme explicitný návod na aplikáciu našej metódy, ktorý je možné aplikovať aj bez oboznámenia sa s predchádzajúcimi kapitolami.



Obr. 4: Vektory smerujúce k vrcholom trojrozmerného regulárneho simplexu v $\mathbb{R}^3$.

Podobne ako pri predchádzajúcich metódach aj v tomto prípade platí, že v praxi sú zvyčajne matice $\mathbf{\Sigma}_{11}, \mathbf{\Sigma}_{12}, \mathbf{\Sigma}_{21}, \mathbf{\Sigma}_{22}$ neznáme a k dispozícií máme iba náhodný výber $\mathbf{X}_1, \ldots, \mathbf{X}_n$, $n \geq (p+q)$. Jediná zmena, ktorá v našej metóde nastane, je nahradenie kanonických premenných výberovými kanonickými premennými.

V praxi na hľadanie vektorov $\mathbf{u}_i$ z *Lemy* 5.2 využijeme algoritmus na hľadanie Schurovho rozkladu $\mathbf{A} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T$, pričom vektory $\mathbf{u}_i$ budú tvoriť stĺpce matice $\mathbf{W}$. Na nájdenie Schurovho rozkladu v jazyku `R` môžeme použiť funkciu `eigen(x, ...)`.

Postup: Vizualizácia kategorizovaných dát pomocou regulárneho simplexového kódovania a kanonických korelácií

Vstup: Náhodný výber $\mathbf{X}_1, \ldots, \mathbf{X}_n$ náhodného vektora $\mathbf{X} = (X_1, \ldots, X_p)^T$, $n \geq p$, ktorý obsahuje p kvantitatívnych premenných $X_1, \ldots, X_p$. Náhodný výber $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ náhodného vektora $\mathbf{Y}$, ktorý obsahuje numerické indexy kategórií, ku ktorým $\mathbf{X}_1, \ldots, \mathbf{X}_n$ patria.

1: Matica $\mathbf{A} \in \mathbb{R}^{k \times k}$: $\forall i, j \in \{1, \ldots, k\} : \mathbf{A}_{ij} = 1$, ak $i = j$, inak $\mathbf{A}_{ij} = -1$, kde k označuje počet kategórií.

2: Algoritmus na nájdenie Schurovho rozkladu $\mathbf{A} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T$ a preusporiadanie tak, aby $\mathbf{\Lambda}_{kk} = \lambda_{\min}(\mathbf{A})$.

3: Náhodný výber $\mathbf{V}_1, \ldots, \mathbf{V}_n$ náhodného vektora $\mathbf{V}$: $\mathbf{V}_r = \mathbf{W}[r, 1 : (k-1)]$, kde r označuje numerický index r-tej kategórie.

4: $\mathbf{U} = \mathbf{X}$, $\mathbf{U}_i = \mathbf{X}_i, i \in \{1, \ldots, n\}$.

5: Nech $\mathbf{S}_n$ je výberová kovariančná matica náhodných vektorov $\mathbf{U}, \mathbf{V}$. Označme príslušné blokové podmatice $\mathbf{S}_{11}, \mathbf{S}_{12}, \mathbf{S}_{21}$ a $\mathbf{S}_{22}$.

6: Pre vizualizáciu v $\mathbb{R}^2$ ($\mathbb{R}^3$) určíme vlastné vektory $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ ($\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3$) prislúchajúce k prvým 2 (3) najväčším vlastným hodnotám matice $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$ [11].

7: V $\mathbb{R}^2$ ($\mathbb{R}^3$) zobrazíme body $(\boldsymbol{\alpha}_1^T\mathbf{X}_i, \boldsymbol{\alpha}_2^T\mathbf{X}_i)$, $i \in \{1, \ldots, n\}$ $((\boldsymbol{\alpha}_1^T\mathbf{X}_i, \boldsymbol{\alpha}_2^T\mathbf{X}_i, \boldsymbol{\alpha}_3^T\mathbf{X}_i)$, $i \in \{1, \ldots, n\})$.

5.2.3 Vlastnosti metódy.

V tejto časti analyticky odvodíme niektoré vlastnosti našej metódy vizualizácie kategorizovaných dát pomocou regulárneho simplexového kódovania a kanonických korelácií.

Lema 5.3. *(Prenásobenie matice $\mathbf{V}$ regulárnou maticou nemá vplyv na výsledok vizualizácie) Nech $\mathbf{A} \in \mathbb{R}^{q\times q}$ je regulárna matica. Potom pre náhodné vektory $\mathbf{X} = (\mathbf{U}, \mathbf{V})^T$ a $\tilde{\mathbf{X}} = (\tilde{\mathbf{U}}, \tilde{\mathbf{V}}) = (\mathbf{U}, \mathbf{AV})^T$, kde $\mathbf{U} \in \mathbb{R}^p$ a $\mathbf{V} \in \mathbb{R}^q$, sa súčiny príslušných podmatíc kovariančnej matice alebo ich inverzií $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$, respektíve $\tilde{\boldsymbol{\Sigma}}_{11}^{-1}\tilde{\boldsymbol{\Sigma}}_{12}\tilde{\boldsymbol{\Sigma}}_{22}^{-1}\tilde{\boldsymbol{\Sigma}}_{21}$ zhodujú.*

Dôkaz. Nech

$$cov(\mathbf{X}, \mathbf{X}) = \begin{pmatrix} cov(\mathbf{U}, \mathbf{U}) & cov(\mathbf{U}, \mathbf{V}) \\ cov(\mathbf{V}, \mathbf{U}) & cov(\mathbf{V}, \mathbf{V}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$
$$cov(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}) = \begin{pmatrix} cov(\mathbf{U}, \mathbf{U}) & cov(\mathbf{U}, \mathbf{AV}) \\ cov(\mathbf{AV}, \mathbf{U}) & cov(\mathbf{AV}, \mathbf{AV}) \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{\Sigma}}_{11} & \tilde{\boldsymbol{\Sigma}}_{12} \\ \tilde{\boldsymbol{\Sigma}}_{21} & \tilde{\boldsymbol{\Sigma}}_{22} \end{pmatrix},$$

potom

$$cov(\mathbf{U}, \mathbf{AV}) = cov(\mathbf{U}, \mathbf{V})\mathbf{A}^T = \boldsymbol{\Sigma}_{12}\mathbf{A}^T = \tilde{\boldsymbol{\Sigma}}_{12},$$
$$cov(\mathbf{AV}, \mathbf{U}) = \mathbf{A}cov(\mathbf{U}, \mathbf{U}) = \mathbf{A}\boldsymbol{\Sigma}_{21} = \tilde{\boldsymbol{\Sigma}}_{21},$$
$$cov(\mathbf{AV}, \mathbf{AV}) = \mathbf{A}cov(\mathbf{V}, \mathbf{V})\mathbf{A}^T = \mathbf{A}\boldsymbol{\Sigma}_{22}\mathbf{A}^T = \tilde{\boldsymbol{\Sigma}}_{22}.$$

[11] Predpokladáme, že 2. (3.) najväčšia vlastná hodnota je rôzna od 3. (4.) najväčšej vlastnej hodnoty.

Pri vyjadrení $\tilde{\mathbf{\Sigma}}_{11}^{-1}\tilde{\mathbf{\Sigma}}_{12}\tilde{\mathbf{\Sigma}}_{22}^{-1}\tilde{\mathbf{\Sigma}}_{21}$ využijeme, že inverzia súčinu regulárnych matíc je súčin inverzií matíc v opačnom poradí. Platí

$$\begin{aligned}\tilde{\mathbf{\Sigma}}_{11}^{-1}\tilde{\mathbf{\Sigma}}_{12}\tilde{\mathbf{\Sigma}}_{22}^{-1}\tilde{\mathbf{\Sigma}}_{21} &= \mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{A}^T\left(\mathbf{A}\mathbf{\Sigma}_{22}\mathbf{A}^T\right)^{-1}\mathbf{A}\mathbf{\Sigma}_{21}\\ &= \mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{A}^T(\mathbf{A}^T)^{-1}\mathbf{\Sigma}_{22}^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{\Sigma}_{21}\\ &= \mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}.\end{aligned}$$

□

Dôsledok Lemy 5.2.3 je nezávislosť metódy vizualizácie kategorizovaných dát pomocou regulárneho simplexového kódovania a kanonických korelácií od ľubovoľnej regulárnej transformácie pomocných premenných pre kategoriálnu premennú. Špeciálnym prípadom regulárnych transformácií sú ortogonálne transformácie, ktoré reprezentujú rotácie a permutácie. Naša metóda preto nezávisí od poradia, v akom jednotlivým kategóriam priraďíme vektory smerujúce k vrcholom regulárneho simplexu. Avšak nezávislosť metódy od ľubovoľnej regulárnej transformácie pomocných premenných pre kategoriálnu premennú taktiež značí, že rovnaké výsledky vizualizácie by sme dostali aj keby sme zvolili ľubovoľný iný systém $(k-1)$ pomocných premenných, pre ktorý by príslušná kovariančná matica $\mathbf{\Sigma}_{22}$ bola regulárna. V podkapitole 5.2.1 sme ukázali, že ak vytvoríme pomocné premenné pre kategoriálnu premennú v tvare *dummy* premenných, tak $\det(\mathbf{\Sigma}_{22}) \neq 0$. Z toho vyplýva, že voľba referenčnej kategórie neovplyvňuje výsledok. Rovnocennosť medzi kategóriami, na ktorú sme kládli dôraz pri vytváraní našej metódy, teda nie je potrebná. Napriek tomu, považujeme predstavenie vytvárania pomocných premenných pre kategoriálnu premennú pomocou regulárneho simplexového kódovania za prínos tejto práce, pretože v metódach vizualizácie kategorizovaných dát, ktoré by nemali vlastnosť uvedenú v 5.2.3, môže byť užitočné.

Lema 5.4. *(Prenásobenie matice* $\mathbf{U}$ *regulárnou maticou nemá vplyv na výsledok vizualizácie) Nech* $\mathbf{A} \in \mathbb{R}^{p\times p}$ *je ortogonálna matica. Potom pre náhodné vektory* $\mathbf{X} = (\mathbf{U}, \mathbf{V})^T$ *a* $\tilde{\mathbf{X}} = (\tilde{\mathbf{U}}, \tilde{\mathbf{V}}) = (\mathbf{A}\mathbf{U}, \mathbf{V})^T$ *platí*

$$\boldsymbol{\alpha}_i^T\mathbf{U} = \tilde{\boldsymbol{\alpha}}_i^T\tilde{\mathbf{U}},$$

kde $i \in \{1, \ldots, p\}$, $\mathbf{U} \in \mathbb{R}^p$, $\mathbf{V} \in \mathbb{R}^q$ *a* $\boldsymbol{\alpha}_i, \tilde{\boldsymbol{\alpha}}_i$ *sú vlastné vektory prislúchajúce i-tej najväčšej vlastnej hodnote matice* $\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$ *tvorenej súčinom príslušných podmatíc kovariančnej matice alebo ich inverzií pre* $\mathbf{X}$*, respektíve* $\tilde{\mathbf{\Sigma}}_{11}^{-1}\tilde{\mathbf{\Sigma}}_{12}\tilde{\mathbf{\Sigma}}_{22}^{-1}\tilde{\mathbf{\Sigma}}_{21}$ *pre* $\tilde{\mathbf{X}}$*.*

Dôkaz. Nech

$$cov(\mathbf{X},\mathbf{X}) = \begin{pmatrix} cov(\mathbf{U},\mathbf{U}) & cov(\mathbf{U},\mathbf{V}) \\ cov(\mathbf{V},\mathbf{U}) & cov(\mathbf{V},\mathbf{V}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$
$$cov(\tilde{\mathbf{X}},\tilde{\mathbf{X}}) = \begin{pmatrix} cov(\mathbf{AU},\mathbf{AU}) & cov(\mathbf{AU},\mathbf{V}) \\ cov(\mathbf{V},\mathbf{AU}) & cov(\mathbf{V},\mathbf{V}) \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{\Sigma}}_{11} & \tilde{\boldsymbol{\Sigma}}_{12} \\ \tilde{\boldsymbol{\Sigma}}_{21} & \tilde{\boldsymbol{\Sigma}}_{22} \end{pmatrix},$$

potom

$$\begin{aligned} cov(\mathbf{AU},\mathbf{AU}) &= \mathbf{A}cov(\mathbf{U},\mathbf{U})\mathbf{A}^T = \mathbf{A}\boldsymbol{\Sigma}_{11}\mathbf{A}^T = \tilde{\boldsymbol{\Sigma}}_{11}, \\ cov(\mathbf{AU},\mathbf{V}) &= \mathbf{A}cov(\mathbf{U},\mathbf{V}) = \mathbf{A}\boldsymbol{\Sigma}_{12} = \tilde{\boldsymbol{\Sigma}}_{12}, \\ cov(\mathbf{V},\mathbf{AU}) &= cov(\mathbf{V},\mathbf{U})\mathbf{A}^T = \boldsymbol{\Sigma}_{21}\mathbf{A}^T = \tilde{\boldsymbol{\Sigma}}_{21}. \end{aligned}$$

Pri vyjadrení $\tilde{\boldsymbol{\Sigma}}_{11}^{-1}\tilde{\boldsymbol{\Sigma}}_{12}\tilde{\boldsymbol{\Sigma}}_{22}^{-1}\tilde{\boldsymbol{\Sigma}}_{21}$ opäť využijeme, že inverzia súčinu regulárnych matíc je súčin inverzií matíc v opačnom poradí. Platí

$$\begin{aligned} \tilde{\boldsymbol{\Sigma}}_{11}^{-1}\tilde{\boldsymbol{\Sigma}}_{12}\tilde{\boldsymbol{\Sigma}}_{22}^{-1}\tilde{\boldsymbol{\Sigma}}_{21} &= \left(\mathbf{A}\boldsymbol{\Sigma}_{11}\mathbf{A}^T\right)^{-1}\mathbf{A}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{A}^T \\ &= (\mathbf{A}^T)^{-1}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{A}^{-1}\mathbf{A}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{A}^T \\ &= (\mathbf{A}^T)^{-1}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{A}^T. \end{aligned}$$

Pre vlastné vektory a vlastné čísla týchto matíc platí

$$\begin{aligned} (\mathbf{A}^T)^{-1}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{A}^T\tilde{\boldsymbol{\alpha}}_i &= \lambda_i\tilde{\boldsymbol{\alpha}}_i \\ \mathbf{A}^T(\mathbf{A}^T)^{-1}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{A}^T\tilde{\boldsymbol{\alpha}}_i &= \lambda_i\mathbf{A}^T\tilde{\boldsymbol{\alpha}}_i \\ \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\alpha}_i &= \lambda_i\boldsymbol{\alpha}_i. \end{aligned}$$

Matice $\tilde{\boldsymbol{\Sigma}}_{11}^{-1}\tilde{\boldsymbol{\Sigma}}_{12}\tilde{\boldsymbol{\Sigma}}_{22}^{-1}\tilde{\boldsymbol{\Sigma}}_{21}$ a $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ majú rovnaké vlastné čísla a pre vlastné vektory platí vzťah $\boldsymbol{\alpha}_i = \mathbf{A}^T\tilde{\boldsymbol{\alpha}}_i$. Na základe týchto vzťahov potom platí

$$\boldsymbol{\alpha}_i^T\mathbf{U} = \left(\mathbf{A}^T\tilde{\boldsymbol{\alpha}}_i\right)^T\mathbf{U} = \tilde{\boldsymbol{\alpha}}_i^T\mathbf{AU} = \tilde{\boldsymbol{\alpha}}_i^T\tilde{\mathbf{U}}.$$

□

Ak v Leme 5.4 zvolíme ako regulárnu transformáciu kladne definitnú diagonálnu regulárnu maticu, tak pre našu metódu to značí nezávislosť na škálovaní jednotlivých premenných. Samozrejme, od každej premennej stále vyžadujeme, aby boli všetky pozorovania vyjadrené v rovnakých jednotkách.

5.2.4 Súvis s inými metódami

Počas výskumu sme objavili metódu vizualizácie kategorizovaných dát pomocou kanonických korelácií a *dummy* pomocných premenných [9]. Na základe Lemy 5.2.3 a jej dôsledku sú výsledky vizualizácie získané touto metódou identické s výsledkami vizualizácie pomocou našej metódy. V [9] je zároveň dokázaná ekvivalencia metódy vizualizácie kategorizovaných dát pomocou kanonických korelácií a *dummy* pomocných premenných s metódou lineárnej diskriminačnej analýzy. Náš prístup založený na regulárnom simplexovom kódovaní a kanonických koreláciách teda vedie k rovnakým výsledkom ako metóda lineárnej diskriminačnej analýzy. V dôkaze v práci [9] sa vyskytovali nepresnosti, ktoré sme opravili, a zároveň sme postup doplnili o vlastné výpočty, aby bol pre čitateľa zrozumiteľnejší. Teraz uvedieme nami rozšírený dôkaz .

Optimálnymi lineárnymi transformáciami v prípade lineárnej diskriminačnej analýzy sú vlastné vektory matice $\mathbf{S}_W^{-1}\mathbf{S}_B$ [12] prislúchajúce k prvým dvom, respektíve trom najväčším vlastným číslam. V prípade kanonických korelácií a „vhodného" systému $k-1$ pomocných premenných [13] za optimálne lineárne transformácie považujeme vlastné vektory matice $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$ prislúchajúce k prvým dvom, respektíve trom najväčším vlastným číslam. Ukážeme, že matice $\mathbf{S}_W^{-1}\mathbf{S}_B$ a $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$ sa až na prenásobenie konštantou zhodujú. Z toho potom bude vyplývať ekvivalencia metód. Predtým však odvodíme pomocné vzťahy, ktoré neskôr využijeme.

Pre náhodný výber vektora kvantitatívnych premenných budeme používať označenie z podkapitoly 4.2. To znamená, že $\mathbf{X}_{ij}$ označuje j-te pozorovanie z i-tej kategórie pre $i \in \{1,\ldots,k\}, j \in \{1,\ldots,n_i\}$. Budeme uvažovať *dummy* pomocné premenné pre kategoriálnu premennú $\mathbf{Y}$. Realizácie pomocných premenných pre náhodný výber $\mathbf{Y}$ budeme označovať $\mathbf{Z}_{ij}$, pričom platí $\mathbf{Z}_{ij} = \mathbf{Z}_{il}$ pre $i \in \{1,\ldots,k\}$ a pre $j,l \in \{1,\ldots,n_i\}$. Zároveň vektor $\mathbf{Z}_{ij} \in \mathbb{R}^{k-1}$ pozostáva zo samých núl okrem i-tej pozície, na ktorej je 1, $i \in \{1,\ldots,k-1\}$. Najprv vyjadríme matice $\mathbf{S}_{12}$ a $\mathbf{S}_{22}^{-1}$ a na základe nich aj súčin matíc $\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$. Využijúc

[12] Matice $\mathbf{S}_W$ a $\mathbf{S}_B$ sme definovali v podkapitole 4.2.

[13] Pod „vhodným" systémom $k-1$ pomocných premenných myslíme taký systém, v ktorom príslušná kovariančná matica $\mathbf{\Sigma}_{22}$ je regulárna.

rovnice z predchádzajúcej časti (5) pre maticu $\mathbf{S}_{22}$ platí

$$\mathbf{S}_{22} = \frac{1}{n} \begin{pmatrix} n_1 & 0 & \dots & 0 \\ 0 & n_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & n_{k-1} \end{pmatrix} \left[\mathbf{I}_{k-1} - \frac{1}{n} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} n_1 & n_2 & \dots & n_{k-1} \end{pmatrix} \right].$$

Inverzná matica k $\mathbf{S}_{22}$ má tvar

$$\mathbf{S}_{22}^{-1} = n \left[\mathbf{I}_{k-1} + \frac{1}{n_k} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} n_1 & n_2 & \dots & n_{k-1} \end{pmatrix} \right] \begin{pmatrix} \frac{1}{n_1} & 0 & \dots & 0 \\ 0 & \frac{1}{n_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{n_{k-1}} \end{pmatrix}.$$

Skutočne, ukážme, že platí $\mathbf{S}_{22}^{-1}\mathbf{S}_{22} = \mathbf{I}_{k-1}$.

$$\mathbf{S}_{22}^{-1}\mathbf{S}_{22} = \left[\mathbf{I}_{k-1} + \frac{1}{n_k} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} n_1 & n_2 & \dots & n_{k-1} \end{pmatrix} \right] \left[\mathbf{I}_{k-1} - \frac{1}{n} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} n_1 & n_2 & \dots & n_{k-1} \end{pmatrix} \right]$$

$$\mathbf{S}_{22}^{-1}\mathbf{S}_{22} = \mathbf{I}_{k-1} + \left(\frac{1}{n_k} - \frac{1}{n} - \frac{1}{n_k}\frac{1}{n}\sum_{i=1}^{k-1} n_i \right) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} n_1 & n_2 & \dots & n_{k-1} \end{pmatrix}$$

$$= \mathbf{I}_{k-1} + \left(\frac{1}{n_k} - \frac{1}{n} - \frac{1}{n_k}\frac{1}{n}(n - n_k) \right) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} n_1 & n_2 & \dots & n_{k-1} \end{pmatrix} = \mathbf{I}_{k-1}.$$

Upravme teraz tvar matice $\mathbf{S}_{12}$, ktorá bola definovaná nasledovne

$$\mathbf{S}_{12} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\mathbf{X}_{ij} - \overline{\mathbf{X}} \right) \left(\mathbf{Z}_{ij} - \overline{\mathbf{Z}} \right),$$

kde $\overline{\mathbf{Z}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{Z}_{ij} = \left(\frac{1}{n_1},\ldots,\frac{1}{n_{k-1}}\right)^T$. Pri ďalších úpravách využijeme, že $\mathbf{Z}_{ij} = \mathbf{Z}_{il}$.

$$\begin{aligned}
\mathbf{S}_{12} &= \frac{1}{n}\sum_{i=1}^{k} n_i\left(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}\right)\left(\mathbf{Z}_{i1} - \overline{\mathbf{Z}}\right)^T \\
&= \frac{1}{n}\sum_{i=1}^{k} n_i\left(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}\right)\mathbf{Z}_{i1}^T - \frac{1}{n}\sum_{i=1}^{k} n_i\left(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}\right)\overline{\mathbf{Z}}^T \\
&= \frac{1}{n}\left(n_1\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}\right) \quad \ldots \quad n_{k-1}\left(\overline{\mathbf{X}}_{k-1} - \overline{\mathbf{X}}\right)\right) - \left(\overline{\mathbf{X}} - \overline{\mathbf{X}}\right)\overline{\mathbf{Z}}^T \\
&= \frac{1}{n}\left(n_1\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}\right) \quad \ldots \quad n_{k-1}\left(\overline{\mathbf{X}}_{k-1} - \overline{\mathbf{X}}\right)\right).
\end{aligned}$$

Na základe týchto vyjadrení upravíme tvar súčinu matíc $\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$. Kvôli prehľadnosti rozdelíme výpočet na dva čiastočné výpočty.

$$\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21} = \mathbf{A}_1 + \mathbf{A}_2,$$

$$\begin{aligned}
\mathbf{A}_1 &= \mathbf{S}_{12}\mathrm{diag}\left(\frac{n}{n_1},\ldots,\frac{n}{n_{k-1}}\right)\mathbf{S}_{12}^T \\
&= \frac{1}{n}\sum_{i=1}^{k-1} n_i\left(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}\right)\left(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}\right)^T,
\end{aligned}$$

$$\begin{aligned}
\mathbf{A}_2 &= \mathbf{S}_{12}\frac{n}{n_k}\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}\left(n_1 \quad \ldots \quad n_{k-1}\right)\mathrm{diag}\left(\frac{1}{n_1},\ldots,\frac{1}{n_{k-1}}\right)\mathbf{S}_{12}^T \\
&= \frac{n}{n_k}\mathbf{S}_{12}\begin{pmatrix}1 & \ldots & 1\\ \vdots & & \vdots \\ 1 & \ldots & 1\end{pmatrix}\mathbf{S}_{12}^T \\
&= \frac{1}{nn_k}\left(n_k(\overline{\mathbf{X}} - \overline{\mathbf{X}}_k)\ldots n_k(\overline{\mathbf{X}} - \overline{\mathbf{X}}_k)\right)\begin{pmatrix} n_1\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}\right)^T \\ \vdots \\ n_{k-1}\left(\overline{\mathbf{X}}_{k-1} - \overline{\mathbf{X}}\right)^T\end{pmatrix} \\
&= \frac{n_k}{n}\left(\overline{\mathbf{X}}_{k-1} - \overline{\mathbf{X}}\right)\left(\overline{\mathbf{X}}_{k-1} - \overline{\mathbf{X}}\right)^T.
\end{aligned}$$

Spojením týchto čiastočných výpočtov dostávame

$$\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21} = \frac{1}{n}\sum_{i=1}^{k} n_i\left(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}\right)\left(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}\right)^T = \mathbf{S}_B. \tag{8}$$

Získali sme vzťah, pomocou ktorého sme definovali maticu $\mathbf{S}_B$ v podkapitole 4.2, teda $\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21} = \mathbf{S}_B$.

Posledným pomocným vzťahom, ktorý odvodíme, je vzťah medzi maticou $\mathbf{S}_n$ a maticami $\mathbf{S}_B$ a $\mathbf{S}_W$. Matice $\mathbf{S}_n$, $\mathbf{S}_W$ sme definovali nasledovne

$$\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}\right)\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}\right)^T, \tag{9}$$

$$\mathbf{S}_W = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)^T, \tag{10}$$

kde $n = \sum_{i=1}^{k} n_i$. Odchýlku dátového bodu od celkového priemeru môžeme vyjadriť ako súčet odchýlky dátového bodu od priemeru danej kategórie, z ktorej pozorovanie pochádza, a odchýlky priemeru danej kategórie od celkového priemeru

$$\mathbf{X}_{ij} - \overline{\mathbf{X}} = (\mathbf{X}_{ij} - \overline{\mathbf{X}}_i) + (\overline{\mathbf{X}}_i - \overline{\mathbf{X}}).$$

Prenásobíme jednotlivé strany rovnice ich transpozíciami sprava, roznásobíme a sčítame všetkých n rovníc.

$$\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}\right)\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}\right)^T = \left((\mathbf{X}_{ij} - \overline{\mathbf{X}}_i) + (\overline{\mathbf{X}}_i - \overline{\mathbf{X}})\right)\left((\mathbf{X}_{ij} - \overline{\mathbf{X}}_i) + (\overline{\mathbf{X}}_i - \overline{\mathbf{X}})\right)^T$$

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}\right)\left(\mathbf{X}_{ij} - \overline{\mathbf{X}}\right)^T &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)^T + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{\mathbf{X}}_i - \overline{\mathbf{X}})(\overline{\mathbf{X}}_i - \overline{\mathbf{X}})^T \\ &+ \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)(\overline{\mathbf{X}}_i - \overline{\mathbf{X}})^T + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{\mathbf{X}}_i - \overline{\mathbf{X}})(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)^T.\end{aligned}$$

Využijeme vzťahy (8), (9) a (10)

$$n\mathbf{S}_{11} = n\mathbf{S}_W + n\mathbf{S}_B + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)(\overline{\mathbf{X}}_i - \overline{\mathbf{X}})^T + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{\mathbf{X}}_i - \overline{\mathbf{X}})(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)^T$$

$$n\mathbf{S}_{11} = n\mathbf{S}_W + n\mathbf{S}_B + \sum_{i=1}^{k}\left(n_i\overline{\mathbf{X}}_i - n_i\overline{\mathbf{X}}_i\right)(\overline{\mathbf{X}}_i - \overline{\mathbf{X}})^T + \sum_{i=1}^{k}(\overline{\mathbf{X}}_i - \overline{\mathbf{X}})\left(n_i\overline{\mathbf{X}}_i - n_i\overline{\mathbf{X}}_i\right)^T$$

$$\mathbf{S}_{11} = \mathbf{S}_W + \mathbf{S}_B. \tag{11}$$

Už máme odvodené všetky potrebné pomocné vzťahy na to, aby sme ukázali, že vlastný vektor matice $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$ prislúchajúci k i-temu najväčšiemu vlastnému číslu je zhodný s vlastným vektorom matice $\mathbf{S}_W^{-1}\mathbf{S}_b$ tiež prislúchajúcemu k i-temu najväčšiemu vlastnému číslu. Nech $\boldsymbol{\alpha}$ je vlastný vektor matice $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$ a λ je príslušné vlastné číslo. Upra-

víme nasledujúci vzťah

$$\begin{aligned}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\boldsymbol{\alpha} &= \lambda\boldsymbol{\alpha}\\ \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\boldsymbol{\alpha} &= \lambda\mathbf{S}_{11}\boldsymbol{\alpha}.\end{aligned}$$

Využijeme pomocné vzťahy (8) a (11)

$$\begin{aligned}\mathbf{S}_B\boldsymbol{\alpha} &= \lambda(\mathbf{S}_B + \mathbf{S}_W)\boldsymbol{\alpha}\\ \mathbf{S}_B\boldsymbol{\alpha} &= \lambda\mathbf{S}_B\boldsymbol{\alpha} + \lambda\mathbf{S}_W\boldsymbol{\alpha}\\ (1-\alpha)\mathbf{S}_B\boldsymbol{\lambda} &= \lambda\mathbf{S}_W\boldsymbol{\alpha}\\ \mathbf{S}_W^{-1}\mathbf{S}_B\boldsymbol{\alpha} &= \frac{\lambda}{1-\lambda}\boldsymbol{\alpha}\\ \mathbf{S}_W^{-1}\mathbf{S}_B\boldsymbol{\alpha} &= \tilde{\lambda}\boldsymbol{\alpha},\end{aligned}$$

kde $\tilde{\lambda} = \frac{\lambda}{1-\lambda}$. Vidíme, že vlastné vektory $\boldsymbol{\alpha}$ sú zároveň vlastnými vektormi matice $\mathbf{S}_W^{-1}\mathbf{S}_B$. Ukážeme ešte, že $\lambda_1 > \lambda_2$ práve vtedy, keď $\tilde{\lambda}_1 > \tilde{\lambda}_2$. Z toho bude vyplývať, že vlastné vektory matíc $\mathbf{S}_W^{-1}\mathbf{S}_B$ a $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$ prislúchajúce k i-tym najväčším vlastným číslam sa zhodujú

$$\begin{aligned}\tilde{\lambda}_1 &> \tilde{\lambda}_2\\ \frac{\lambda_1}{1-\lambda_1} &> \frac{\lambda_2}{1-\lambda_2}\\ \lambda_1(1-\lambda_2) &> \lambda_2(1-\lambda_1)\\ \lambda_1 &> \lambda_2.\end{aligned}$$

6 Zhodnotenie výsledkov vizualizácie vybraných dátových súborov

Metódy vizualizácie dát, ktoré sme predstavili v predchádzajúcich kapitolách, teraz aplikujeme na vybrané dátové súbory obsahujúce kategorizované dáta. Metódy aplikujeme najprv na umelý dátový súbor, aby sme vizuálne demonštrovali vlastnosti metód, a neskôr aj na reálne dátové súbory. Reálne dátové súbory sme vybrali tak, aby boli zastúpené aj dátové súbory s menším počtom kvantitatívnych premenných, aj dátové súbory s väčším počtom kvantitatívnych premenných. Možným prístupom na porovnávanie metód vizualizácie kategorizovaných dát je ohodnotiť vizualizácie z pohľadu kompaktnosti a separovateľnosti kategórií. Našim cieľom však nie je určiť najlepšiu metódu. Práve naopak, chceme poukázať na užitočnosť všetkých metód a ukázať, ako porovnávaním viacerých vizualizácií môžeme získať ucelenú predstavu o dátach. Všetky výpočty a vizualizácie sme realizovali v programovacom jazyku *R* [11]. Príslušné kódy je možné nájsť v prílohe. Pri generovaní náhodných čísel sme zvolili *set.seed(110)*, aby bolo možné reprodukovať všetky výsledky.

6.1 Umelý dátový súbor *A*

Prvý uvažovaný umelý dátový súbor A obsahuje tri kvantitatívne premenné a jednu kategoriálnu premennú, ktorá nadobúda 4 hodnoty. Na ich označenie použijeme numerické indexy $1, 2, 3, 4$. Z každej kategórie sme vygenerovali 250 pozorovaní z normálneho rozdelenia. Parametre normálnych rozdelení, z ktorých sme generovali dáta, sú uvedené v Tabuľke 1.

kategória	**stredná hodnota normálneho rozdelenia**	**kovariančná matica normálneho rozdelenia**
1	$(0,0,0)^T$	$\text{diag}\left(\frac{1}{2}, \frac{5}{2}, 2\right)$
2	$(2,0,0)^T$	$\text{diag}\left(\frac{1}{2}, \frac{5}{2}, 2\right)$
3	$(4,0,0)^T$	$\text{diag}\left(\frac{1}{2}, \frac{5}{2}, 2\right)$
4	$(6,0,0)^T$	$\text{diag}\left(\frac{1}{2}, \frac{5}{2}, 5\right)$

Tabuľka 1: Parametre normálnych rozdelení, z ktorých boli vygenerované pozorovania z jednotlivých kategórií umelého dátového súboru A.

Pozorovania z jednotlivých kategórií sa líšia predovšetkým v hodnotách prvej kvantitatívnej premennej. Vizualizácie tohto dátového súboru v rovine pomocou metód PCA, catPCA, catCCA, respektíve LDA, a PDD sú na Obr. 5 - Obr. 8. V ďalších častiach práce budeme používať označenia catCCA a LDA ako synonymá. Je potrebné poznamenať, že tieto dáta boli zámerne konštruované tak, aby jednotlivé vizualizácie poskytovali rôzne pohľady na ich štruktúru. Neskôr pri vizualizácii reálnych dátových súborov budeme demonštrovať, že rôzne metódy môžu poskytovať aj veľmi podobné vizualizácie. Pred aplikáciou metód sme hodnoty pozorovaní nepreškálovali na jednotkový rozptyl. Hodnoty vektorov optimálnych lineárnych transformácií získané týmito metódami sú uvedené v Tabuľke 2. V nasledujúcej časti interpretujeme výsledky vizualizácie získané rôznymi metódami bez toho, aby sme brali do úvahy rozdelenia, z ktorých dáta pochádzajú. Napriek tomu, že dátové body by bolo možné zobraziť pomocou rotovateľného trojrozmerného grafu, nebudeme sa tomu venovať, pretože hlavným cieľom tohto modelového príkladu je interpretovať a porovnať vizualizácie dát vybranými metódami v rovine.

PCA		**catPCA**		**catCCA, LDA**		**PDD**	
$\hat{\boldsymbol{\alpha}}_1$	$\hat{\boldsymbol{\alpha}}_2$	$\hat{\boldsymbol{\alpha}}_1$	$\hat{\boldsymbol{\alpha}}_2$	$\hat{\boldsymbol{\alpha}}_1$	$\hat{\boldsymbol{\alpha}}_2$	$\hat{\boldsymbol{\alpha}}_1$	$\hat{\boldsymbol{\alpha}}_2$
−0.0282178	−0.1416224	−0.0603705	0.8890772	−0.9999987	0.0333367	0.9994024	0.9999925
−0.0297449	−0.9893554	−0.0341416	0.4524829	0.0015079	−0.4086542	0.0105375	−0.0008079
0.9991591	−0.0334527	0.9975920	0.0692894	−0.0005590	0.9120803	0.0329215	−0.0037838

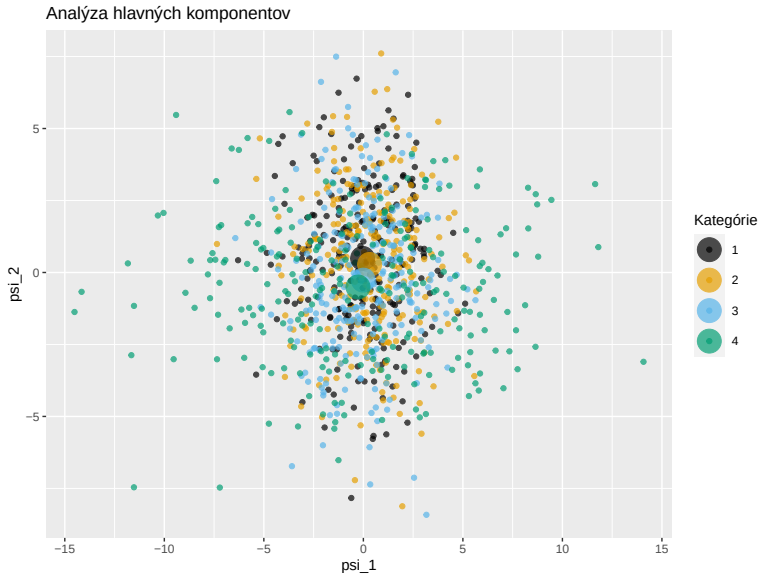
Tabuľka 2: Optimálne vektory lineárnej transformácie metód PCA, catPCA, catCCA, respektíve LDA, PDD pre umelý dátový súbor A.

Optimálne vektory lineárnej transformácie získané metódou PCA sú také smery, v ktorých majú dátové body najväčší rozptyl. Z porovnania výsledku vizualizácie metódy PCA na Obr. 5 a výsledku vizualizácie metódy catCCA na Obr. 7 vyplýva, že prvé dva smery maximálneho rozptylu sa nezhodujú so smermi, ktoré dobre vysvetľujú rozdiely medzi kategóriami. Projekcia na smery maximálneho rozptylu však odhaľuje celkovú štruktúru v dátach. Z tejto projekcie sa javí, že štruktúra dát naprieč prvou, druhou a treťou kategóriou je približne rovnaká. Zároveň sa zdá, že dátové body prislúchajúce štvrtej kategórii majú väčší rozptyl. Napriek tomu žiadne z týchto pozorovaní nepovažujeme za odľahlé (angl. *outliera*). V Tabuľke 3 uvádzame mieru rozptylu, ktorú zachytáva prvý a druhý smer lineárnej transformácie pre jednotlivé metódy. Prvý hlavný komponent vysvetľuje

44.67% celkového rozptylu v dátach. Spomedzi uvedených smerov lineárnej transformácie zachytáva najväčšiu mieru celkového rozptylu, čo je v súlade s princípom metódy PCA. Keďže hlavné komponenty sú navzájom ortogonálne, môžeme hovoriť aj o kumulatívnej miere rozptylu, ktorú zachytávajú prvé dva hlavné komponenty. Tie spolu vysvetľujú 75.52% celkového rozptylu.

	PCA	**catPCA**	**catCCA, LDA**	**PDD**
$\hat{\boldsymbol{\alpha}}_1$	44.67%	44.65%	24.62%	24.63%
$\hat{\boldsymbol{\alpha}}_2$	30.85%	26.61%	42.53%	24.63%

Tabuľka 3: Miera rozptylu, ktorú zachytáva prvý a druhý smer lineárnej transformácie získaný metódami PCA, catPCA, catCCA, respektíve LDA, PDD pre umelý dátový súbor A.



Obr. 5: Vizualizácia umelého dátového súboru A metódou analýzy hlavných komponentov (PCA). Projekcie pozorovaní z jednotlivých kategórií sú farebne odlíšené a znázornené bodkami. Väčšími kruhmi sú znázornené projekcie príslušných výberových priemerov.

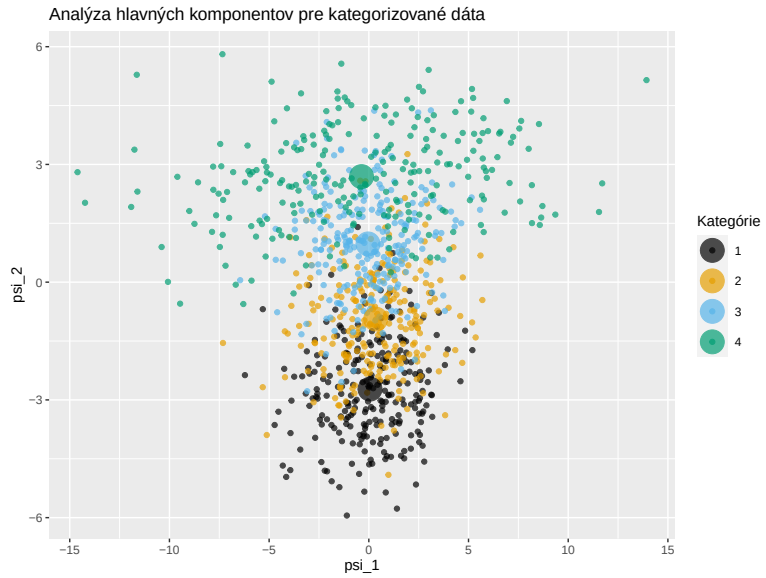
Zamerajme sa teraz na vizualizáciu pomocou metódy catPCA, ktorá je zobrazená na Obr. 6. Metóda catPCA hľadá smery, v ktorých sú vzdialenosti medzi dátovými bodmi z rôznych kategórií najväčšie. Z hodnôt vektorov uvedených v Tabuľke 2 vyplýva, že prvé smery $\hat{\boldsymbol{\alpha}}_1$ metód PCA a catPCA sú podobné a zodpovedajú predovšetkým tretiemu pôvodnému smeru, teda tretej kvantitatívnej premennej. To znamená, že smer maximálneho rozptylu približne zodpovedá smeru, ktorý najviac vysvetľuje rozptyl medzi pozorovaniami

z rôznych kategórií. Takáto situácia môže nastať v dvoch prípadoch. V prvom prípade je v zvolenom smere maximálny rozptyl aj medzi pozorovaniami z rovnakých kategórií aj medzi pozorovaniami z rôznych kategórií. V druhom prípade síce nie je v zvolenom smere maximálny rozptyl medzi pozorovaniami z rovnakých kategórií, ale je v ňom maximálny rozptyl medzi pozorovaniami z rôznych kategórií a zároveň ide o smer maximálneho rozptylu. V smere $\hat{\boldsymbol{\alpha}}_1$ na Obr. 5 sme pozorovali väčší rozptyl dátových bodov prislúchajúcich štvrtej kategórii. Rovnaký jav pozorujeme aj na Obr. 6. V druhom smere $\hat{\boldsymbol{\alpha}}_2$ sa metódy výraznejšie odlišujú. Na Obr. 6 sa „mraky" pozorovaní z jednotlivých kategórií prekrývajú. Tento prekryv by sme mohli označiť ako postupný. Neprekrývajú sa všetky „mraky" pozorovaní súčasne. Zobrazené výberové priemery naznačujú, že medzi kategóriami existuje rozdiel. Metóda catPCA čiastočne dokázala odseparovať pozorovania z rôznych kategórií. Ak by sme uvažovali iba pozorovania z kategórií jeden a štyri, boli by dané kategórie dobre odseparované.

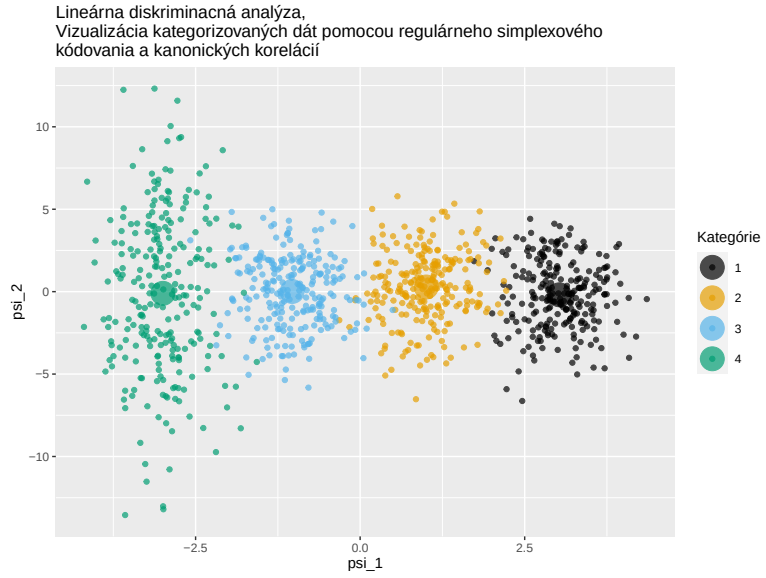
Napriek tomu, že metóda catCCA patrí do rovnakej triedy metód ako metóda catPCA, medzi lineárne metódy s učiteľom, dokázala výraznejšie lepšie odseparovať pozorovania z rôznych kategórií. Vizualizácia pomocou metódy catCCA je na Obr. 7. Dôvod pramení z formulácie metód. Obidve metódy sa istým spôsobom snažia odlíšiť projekcie pozorovaní z rôznych kategórií. Avšak z formulácie metódy LDA vieme, že zároveň kladie podmienku aj na vzťahy medzi projekciami pozorovaní z rovnakej kategórie. Snaží sa minimalizovať ich rozptyl. Metóda catPCA nekladie žiadnu podmienku na vzťahy medzi projekciami pozorovaní z rovnakej kategórie. Z hodnôt vektorov lineárnej transformácie uvedených v Tabuľke 2 vyplýva, že prvý smer $\hat{\boldsymbol{\alpha}}_1$ metódy catCCA zodpovedá predovšetkým prvej kvantitatívnej premennej. [14] Druhý smer $\hat{\boldsymbol{\alpha}}_2$ metódy catCCA zodpovedá predovšetkým tretej kvantitatívnej premennej. V tomto smere majú pozorovania zo štvrtej kategórie výrazne väčší rozptyl ako pozorovania z iných kategórií. Tento jav sme pozorovali aj pri vizualizácii pomocou metód PCA a catPCA.

Metóda PDD sa snaží vizualizovať dátové body tak, aby vzdialenosť projekcie dátového bodu od projekcie výberového priemeru príslušnej kategórie bola menšia ako vzdialenosť

[14] Poznamenajme, že dáta boli konštruované tak, aby sa „mraky" dát z jednotivých kategórií príliš neprekrývali. A zároveň tak, aby existoval 1 smer, pozdĺž ktorého sa predovšetkým líšia pozorovania z rôznych kategórií. Dá sa povedať, že metóda catCCA tento smer zvolila.

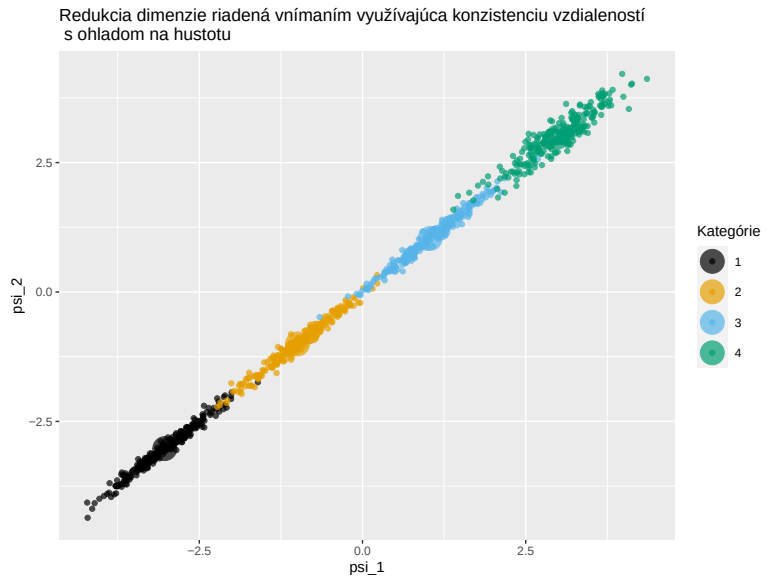


Obr. 6: Vizualizácia umelého dátového súboru A metódou analýzy hlavných komponentov pre kategorizované dáta (catPCA). Projekcie pozorovaní z jednotlivých kategórií sú farebne odlíšené a znázornené bodkami. Väčšími kruhmi sú znázornené projekcie príslušných výberových priemerov.



Obr. 7: Vizualizácia umelého dátového súboru A metódou lineárnej diskriminačnej analýzy (LDA), respektíve metódou vizualizácie kategorizovaných dát pomocou regulárneho simplexového kódovania a kanonických korelácií (catCCA). Projekcie pozorovaní z jednotlivých kategórií sú farebne odlíšené a znázornené bodkami. Väčšími kruhmi sú znázornené projekcie príslušných výberových priemerov.

projekcie dátového bodu od projekcie výberového priemeru inej kategórie. Na Obr. 8, ktorý zobrazuje vizualizáciu pomocou metódy PDD, ležia projekcie dátových bodov takmer na jednej priamke. Z toho vyplýva, že prvý a druhý smer lineárnej projekcie je približne rovnaký. Potvrdzuje to aj Tabuľka 2. Na vysvetlenie rozdielov medzi kategóriami teda postačuje jeden smer. Prvý smer lineárnej transformácie $\hat{\boldsymbol{\alpha}}_1$ je približne rovnaký pre metódy catCCA a PDD. Keďže tento smer bol postačujúci pre vysvetlenie rozdielov medzi kategóriami v metóde PDD, dá sa považovať za postačujúci aj pri metóde catCCA. Pri metóde catCCA sme nemohli dostať podobný prvý a druhý smer, pretože na smery bola kladená podmienka istého typu kolmosti. Druhý smer lineárnej transformácie $\hat{\boldsymbol{\alpha}}_2$ pre metódu catCCA teda môžeme považovať za smer, ktorý vysvetľuje štruktúru dát v rámci kategórií. Tá môže byť naprieč kategóriami iná. Túto interpretáciu potvrdzuje aj Obr. 7.



Obr. 8: Vizualizácia umelého dátového súboru A metódou redukcie dimenzie riadenou vnímaním využívajúcou konzistenciu vzdialeností s ohľadom na hustotu (PDD). Projekcie pozorovaní z jednotlivých kategórií sú farebne odlíšené a znázornené bodkami. Väčšími kruhmi sú znázornené projekcie príslušných výberových priemerov.

Napriek tomu, že v tomto modelovom príklade sme pracovali len s trojrozmernými dátami, je možné pozorovať rozdiely medzi vizualizáciami. Každá vizualizácia nám poskytla užitočný pohľad na dáta. Avšak ucelenú predstavu o dátach sme získali až analýzou všetkých vizualizácií.

6.2 Iris

Dátový súbor *Iris* sme získali z archívu [3]. Odtiaľ parafrázujeme popis dátového súboru. Tento dátový súbor je zároveň implementovaný v softvéri *R*. V jazyku *R* sme pracovali s implementovanými hodnotami.

Ide o jeden z najznámejších súborov dát. Jeho autorom je Roland Aylmer Fisher (1936). Dátový súbor obsahuje údaje o troch typoch kosatcoch, a to setosa, versicolour a virginica. Okrem tejto kategoriálnej premennej obsahuje aj 4 ďalšie kvantitatívne premenné s informáciami o dĺžke a šírke kališného lístku a okvetného lístku. Hodnoty sú vyjadrené v centimetroch. Z každého typu kosatcov je k dispozícií 50 pozorovaní. Naprieč pozorovaniami sa nevyskytujú žiadne chýbajúce hodnoty.

Na rozdiel od predchádzajúceho dátového súboru v tomto prípade by nám na vizualizáciu dát nepostačovalo ani trojrozmerné rotovateľné zariadenie. Keďže stále ide o relatívne malý počet premenných, na získanie dobrej predstavy o štruktúre dát je možné zobraziť všetky dvojice premenných (angl. *pairplot*). My sa však tomu nebudeme venovať. Zameráme sa iba na vizualizácie dátového súboru *Iris* v rovine pomocou metód PCA, catPCA, catCCA a PDD, ktoré sú na Obr. 9 - Obr. 12. Všetky kvantitatívne premenné sú vyjadrené v centimetroch, preto pred aplikovaním metód sme hodnoty pozorovaní nepreškálovali. Hodnoty vektorov optimálnych lineárnych transformácií získané týmito metódami sú uvedené v Tabuľke 4. Miery celkového rozptylu dát, ktoré jednotlivé smery lineárnych transformácií vysvetľujú, sú uvedené v Tabuľke 5.

PCA		**catPCA**		**catCCA, LDA**		**PDD**	
$\hat{\boldsymbol{\alpha}}_1$	$\hat{\boldsymbol{\alpha}}_2$	$\hat{\boldsymbol{\alpha}}_1$	$\hat{\boldsymbol{\alpha}}_2$	$\hat{\boldsymbol{\alpha}}_1$	$\hat{\boldsymbol{\alpha}}_2$	$\hat{\boldsymbol{\alpha}}_1$	$\hat{\boldsymbol{\alpha}}_2$
0.3613866	-0.6565888	0.3500640	-0.6414078	-0.2087418	0.0065320	-0.2863102	-0.3031674
-0.0845225	-0.7301614	-0.0935593	-0.7488032	-0.3862037	0.5866106	-0.2063136	-0.2213403
0.8566706	0.1733727	0.8589096	0.1606956	0.5540117	-0.2525615	0.6475004	0.6435111
0.3582892	0.0754810	0.3610906	0.0454628	0.7073504	0.7694531	0.6754290	0.6670768

Tabuľka 4: Optimálne vektory lineárnej transformácie metód PCA, catPCA, catCCA, respektíve LDA, PDD pre dátový súbor *Iris*.

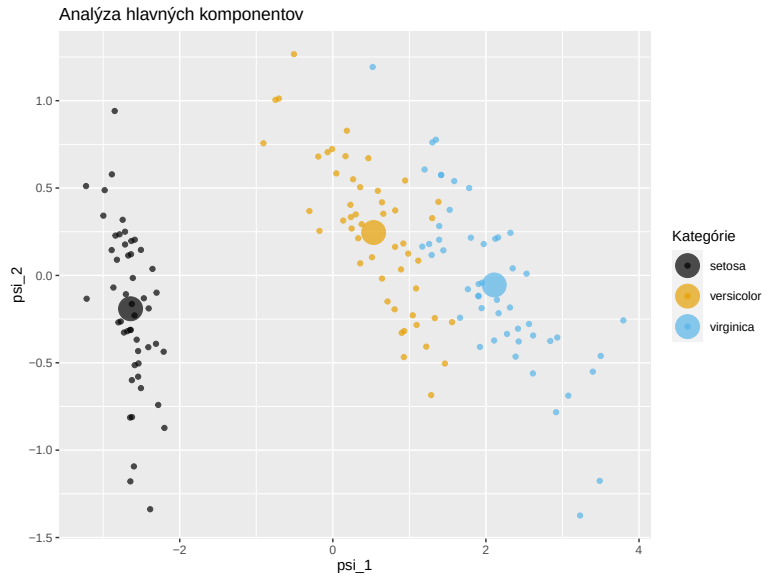
Na rozdiel od predchádzajúceho umelého dátového súboru A v tomto prípade sú oba smery $\hat{\boldsymbol{\alpha}}_1$ a $\hat{\boldsymbol{\alpha}}_2$ získané metódami PCA a catPCA podobné. Pre dátový súbor *Iris* však smery

	PCA	**catPCA**	**catCCA, LDA**	**PDD**
$\hat{\boldsymbol{\alpha}}_1$	92.46%	92.44%	45.36%	48.41%
$\hat{\boldsymbol{\alpha}}_2$	5.31%	5.32%	2.04%	47.05%

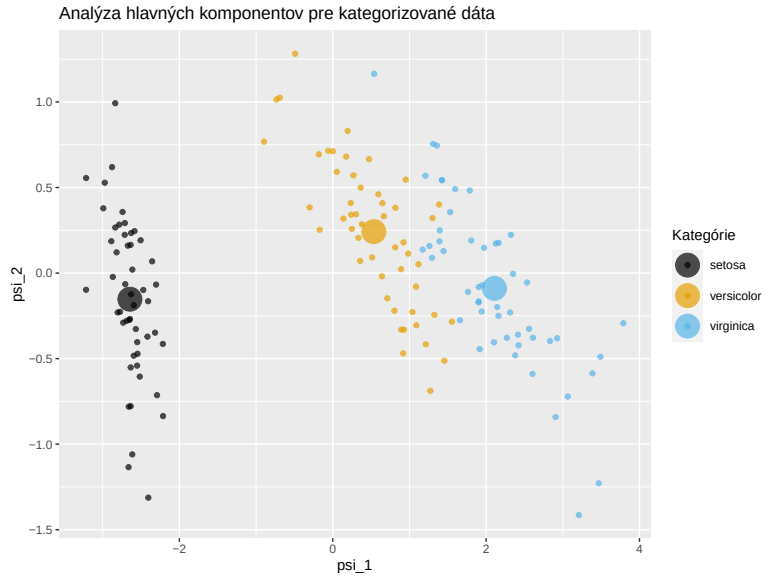
Tabuľka 5: Miera rozptylu, ktorú zachytáva prvý a druhý smer lineárnej transformácie získaný metódami PCA, catPCA, catCCA, respektíve LDA, PDD pre dátový súbor *Iris*.

maximálneho rozptylu, respektíve smery, v ktorých sú maximálne vzdialenosti medzi pozorovaniami z rôznych kategórií, dobre vysvetľujú aj rozdiely medzi kategóriami. Zdá sa, že štruktúra dát kategórií versicolor a virginica je podobná. Vidíme trend, že s rastúcou hodnotou v smere $\hat{\boldsymbol{\alpha}}_1$ pozorovania nadobúdajú nižšie hodnoty v smere $\hat{\boldsymbol{\alpha}}_2$. Pozdĺž smeru $\hat{\boldsymbol{\alpha}}_2$ majú pozorovania z kategórie setosa podobnú štruktúru ako pozorovania z ostatných kategórií. Avšak v smere $\hat{\boldsymbol{\alpha}}_1$ majú menší rozptyl. Smer $\hat{\boldsymbol{\alpha}}_1$ zodpovedá predovšetkým tretej kvantitatívnej premennej, teda dĺžke okvetného lístku. Pozorovania z kategórie setosa sú pozdĺž neho výraznejšie odseparované od pozorovaní z ostatných dvoch kategórií. Vysvetľuje viac ako 92% celkového rozptylu. Druhý smer $\hat{\boldsymbol{\alpha}}_2$ metód PCA a catPCA vysvetľuje väčšinu nevysvetleného rozptylu. Avšak v absolútnych hodnotách ide len o 5.31%, respektíve 5.32% celkového rozptylu. Kategórie versicolor a virginica nie sú na Obr. 9 a Obr. 10 jednodnoznačne odseparované. Medzi výberovými priemermi týchto kategórií existuje „oblasť prechodu", v ktorej sa vyskytujú pozorovania z obidvoch kategórií.

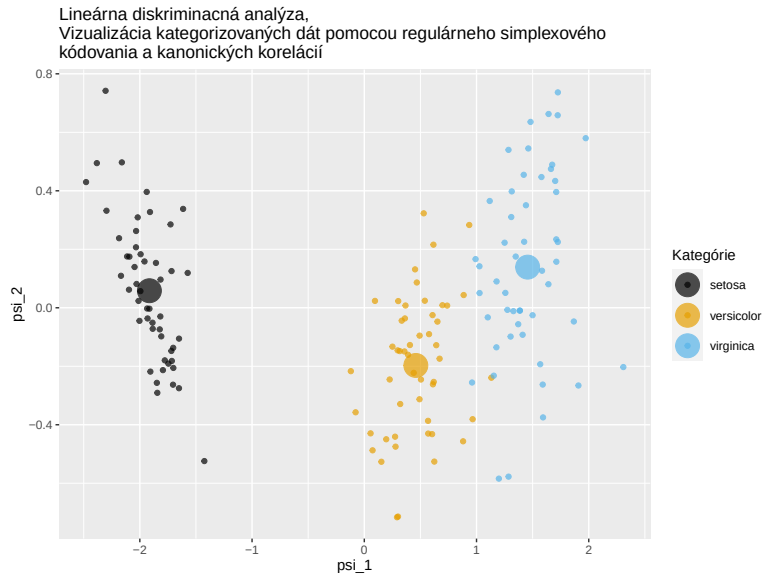
Z hodnôt uvedených v Tabuľke 4 vyplýva, že vizualizácia pomocou metódy catCCA na Obr. 11 poskytuje odlišný pohľad na dáta ako metódy PCA a catPCA. Napriek tomu aj z tohto pohľadu usudzujeme rovnaké kvalitatívne vlastnosti štruktúry dát. Pozorovania z jednotlivých kategórií sú dobre odseparované, pričom pozorovania z kategórie setosa sú výraznejšie odčlenené. V štruktúre dát z kategórií versicolor a virginica opäť pozorujeme rovnaký, tentokrát rastúci trend v smere $\hat{\boldsymbol{\alpha}}_2$ vzhľadom na smer $\hat{\boldsymbol{\alpha}}_1$. Štruktúra dát z kategórie setosa má klesajúci trend v smere $\hat{\boldsymbol{\alpha}}_2$ vzhľadom na smer $\hat{\boldsymbol{\alpha}}_1$. Na vizualizácii pomocou metódy catCCA sú pozorovania z rovnakej kategórie zobrazené kompaktnejšie. Separácia pozorovaní z kategórií versicolor a virginica je o niečo jasnejšia, avšak stále pozorujeme „oblasť prechodu" medzi kategóriami.



Obr. 9: Vizualizácia dátového súboru *Iris* metódou analýzy hlavných komponentov (PCA). Projekcie pozorovaní z jednotlivých kategórií sú farebne odlíšené a znázornené bodkami. Väčšími kruhmi sú znázornené projekcie príslušných výberových priemerov.



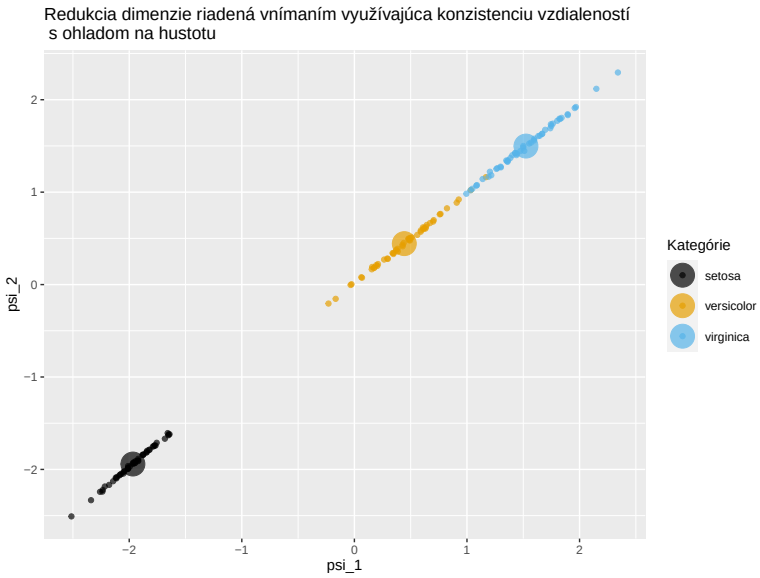
Obr. 10: Vizualizácia dátového súboru *Iris* metódou analýzy hlavných komponentov pre kategorizované dáta (catPCA). Projekcie pozorovaní z jednotlivých kategórií sú farebne odlíšené a znázornené bodkami. Väčšími kruhmi sú znázornené projekcie príslušných výberových priemerov.



Obr. 11: Vizualizácia dátového súboru *Iris* metódou lineárnej diskriminačnej analýzy (LDA), respektíve metódou vizualizácie kategorizovaných dát pomocou regulárneho simplexového kódovania a kanonických korelácií (catCCA). Projekcie pozorovaní z jednotlivých kategórií sú farebne odlíšené a znázornené bodkami. Väčšími kruhmi sú znázornené projekcie príslušných výberových priemerov.

Prvé dva smery lineárnej transformácie $\hat{\boldsymbol{\alpha}}_1$ a $\hat{\boldsymbol{\alpha}}_2$ pre metódu PDD uvedené v Tabuľke 4 sú podobné. Dátové body zobrazené metódou PDD preto ležia takmer na jednej priamke. Smery získané metódou PDD sú zároveň podobné s prvým smerom $\hat{\boldsymbol{\alpha}}_1$ získaným metódou catCCA. Konštatujeme rovnaké závery ako v prípade umelého dátového súboru A. Na vysvetlenie rozdielov medzi kategóriami postačuje jeden smer. Druhý smer $\hat{\boldsymbol{\alpha}}_2$ metódy catCCA odhaľuje štruktúru dát v rámci jednotlivých kategórií.

Podobnosť smerov $\hat{\boldsymbol{\alpha}}_1$ a $\hat{\boldsymbol{\alpha}}_2$ získaných metódou PDD, ktorú sme pozorovali aj v prípade umelého dátového súboru A, aj v prípade dátového súboru *Iris* je skôr výnimkou ako pravidlom. Vyskytuje sa v prípade, kedy postačuje jeden smer na odseparovanie pozorovaní z rôznych kategórií a ďalšie smery len vysvetľujú štruktúru jednotlivých kategórií. V prípade viacrozmerného dátového súboru ide o zriedkavý jav.



Obr. 12: Vizualizácia dátového súboru *Iris* metódou redukcie dimenzie riadenou vnímaním využívajúcou konzistenciu vzdialeností s ohľadom na hustotu (PDD). Projekcie pozorovaní z jednotlivých kategórií sú farebne odlíšené a znázornené bodkami. Väčšími kruhmi sú znázornené projekcie príslušných výberových priemerov.

6.3 Pen-Based Recognition of Handwritten Digits

Dátový súbor *Pen-Based Recognition of Handwritten Digits* sme získali z archívu [3]. Odtiaľ parafrázujeme popis dátového súboru.

Tento dátový súbor obsahuje údaje o rukou napísaných cifrách 0 - 9. Na tablet s rozlíšením 500×500 pixelov napísalo 44 ľudí po 250 cifier v náhodnom poradí. Zariadenie zaznamenalo pre každú cifru každých 100 milisekúnd hodnoty x-ovej a y-ovej súradnice a intenzitu tlaku pera. Informácie o intenzite tlaku pera autori dátového súboru nebrali do úvahy. Hodnoty x-ovej a y-ovej súradnice autori normalizovali, aby boli invariantné na posun a veľkosť. Ďalej ich upravili tak, aby nové hodnoty pochádzali z intervalu $[0, 100]$. Počet údajov o hodnotách x-ovej a y-ovej súradnice naprieč ciframi nie je jednotný, preto tieto údaje prevzorkovali. V dátovom súbore sú pre každú cifru zaznamenané hodnoty x-ovej a y-ovej súradnice pre 8 bodov. Body boli vybrané tak, aby boli rovnomerne vzdialené. Dátový súbor obsahuje 17 premenných. Okrem 16 kvantitatívnych premenných obsahuje jednu kategórialnu premennú, v ktorej je zaznamenané, o akú cifru ide. Dátový súbor je rozdelený na trénovaciu a testovaciu sadu. V trénovacej sade sú údaje o cifrách napísaných

30 ľuďmi, v testovacej sade sú zase údaje o cifrách napísaných zvyšnými 14 ľuďmi. My sme pracovali iba s trénovacou sadou. V nej sa nachádza 7494 pozorovaní, čo je o 6 menej ako by sme očakávali. V popise uvedenom v archíve [3] nie je jasne uvedené prečo tieto údaje chýbajú.

Vizualizácie dátového súboru *Pen-Based Recognition of Handwritten Digits* v rovine pomocou metód PCA, catPCA, catCCA a PDD sú na Obr. 13 - Obr. 17. Všetky kvantitatívne premenné majú rovnakú jednotku, preto sme hodnoty pozorovaní pred aplikovaním metód nepreškálovali. Hodnoty vektorov optimálnych lineárnych transformácií získané týmito metódami sú uvedené v Tabuľke 6. Miery celkového rozptylu dát, ktoré jednotlivé smery lineárnych transformácií vysvetľujú, sú uvedené v Tabuľke 7.

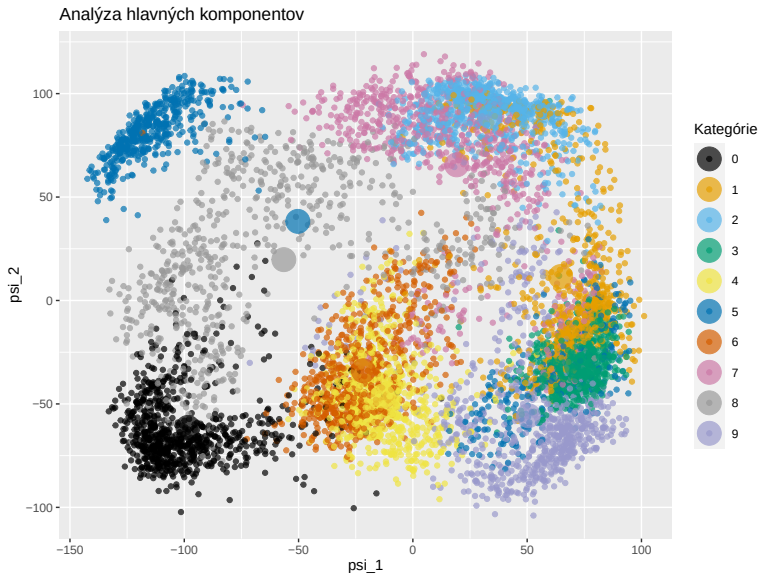
PCA		**catPCA**		**catCCA, LDA**		**PDD**	
$\hat{\boldsymbol{\alpha}}_1$	$\hat{\boldsymbol{\alpha}}_2$	$\hat{\boldsymbol{\alpha}}_1$	$\hat{\boldsymbol{\alpha}}_2$	$\hat{\boldsymbol{\alpha}}_1$	$\hat{\boldsymbol{\alpha}}_2$	$\hat{\boldsymbol{\alpha}}_1$	$\hat{\boldsymbol{\alpha}}_2$
-0.0713129	-0.1988096	0.0918037	0.1975038	-0.2214529	-0.0890543	-0.0794030	-0.2814666
-0.0721798	-0.0357250	0.0763629	0.0306358	-0.1992893	-0.3543223	-0.6028876	0.0433647
0.2017470	0.0486380	-0.2015130	-0.0309396	-0.0781603	0.0411489	0.2061285	-0.0696441
0.1531377	0.0965901	-0.1593827	-0.0833588	0.5287835	0.4195453	-0.2232802	-0.1328535
0.2704011	0.2694848	-0.2913337	-0.2502866	0.3418346	0.2471638	-0.1429922	0.1493349
0.3592598	0.1203967	-0.3668170	-0.0897505	-0.0513009	-0.0322331	0.2598766	0.2505739
0.1577870	0.0517091	-0.1630463	-0.0427347	0.1239796	-0.0638441	0.0248248	0.2408876
0.4136720	-0.0520962	-0.4082420	0.0859649	-0.2210029	0.1601587	0.2042004	-0.1400433
0.1183264	-0.3801000	-0.0884228	0.3883183	-0.0572051	0.0532427	-0.0532109	-0.1079182
0.1779246	-0.2263066	-0.1610896	0.2382830	0.1658972	0.1462787	0.2066399	-0.5235806
0.0376410	-0.5541718	0.0060337	0.5557640	-0.1618667	-0.2058978	0.0070317	-0.5635995
-0.2105912	-0.1898159	0.2221174	0.1700583	-0.0128346	-0.2242476	-0.2637535	-0.2378154
-0.0705295	-0.1092491	0.0763177	0.1038307	0.0578489	0.0578065	0.0203707	0.2060186
-0.4626658	-0.0253173	0.4632394	-0.0149680	-0.3214466	0.1277331	-0.4485269	-0.0065929
-0.0876971	0.5400302	0.0377553	-0.5421197	0.3202535	-0.2017579	0.1838891	-0.1296147
-0.4573811	0.1093699	0.4495255	-0.1482262	-0.4188425	0.6478280	-0.2388064	-0.1199909

Tabuľka 6: Optimálne vektory lineárnej transformácie metód PCA, catPCA, catCCA, respektíve LDA, PDD pre dátový súbor *Pen-Based Recognition of Handwritten Digits*.

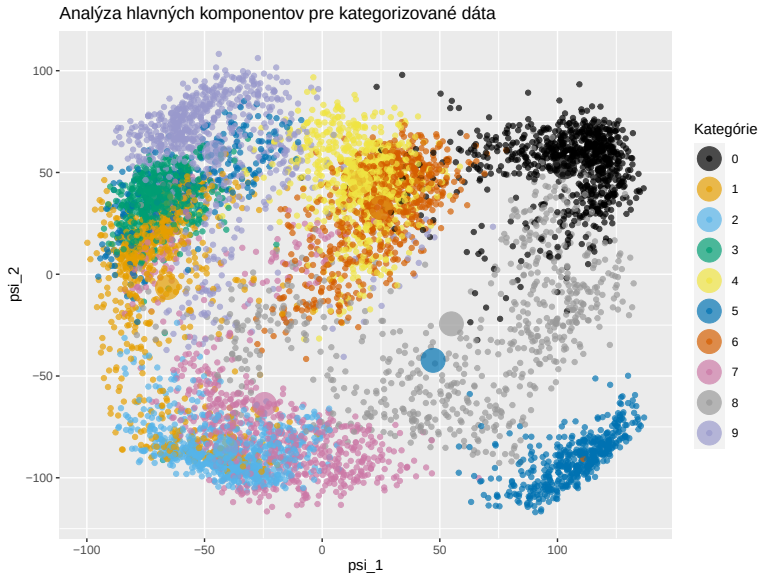
	PCA	**catPCA**	**catCCA, LDA**	**PDD**
$\hat{\boldsymbol{\alpha}}_1$	28.51%	28.48%	10.65%	11.76%
$\hat{\boldsymbol{\alpha}}_2$	24.40%	24.43%	4.05%	8.63%

Tabuľka 7: Miera rozptylu, ktorú zachytáva prvý a druhý smer lineárnej transformácie získaný metódami PCA, catPCA, catCCA, respektíve LDA, PDD pre dátový súbor *Pen-Based Recognition of Handwritten Digits*.

Ako prvé analyzujeme vizualizácie pomocou metód PCA a catPCA zobrazené na Obr. 5 a Obr. 6. Na prvý pohľad sa z rozloženia „mrakov" dát jednotlivých kategórií zdá, že ide o podobné vizualizácie, len navzájom prevrátené. Prevrátená vizualizácia vzniká použitím opačných smerov lineárnej transformácie. Nedochádza pri nej k zmene kvalitatívnych vlastností vizualizácie. Smery lineárnej transformácie $\hat{\boldsymbol{\alpha}}_1$ a $\hat{\boldsymbol{\alpha}}_2$ metód PCA a catPCA uvedené v Tabuľke 2 nie sú opačné, ale my ich budeme považovať za smery blízke opačným smerom. Rozhodli sme sa preto analyzovať iba jednu vizualizáciu, a to pomocou metódy PCA. Z podobnosti výsledkov vizualizácie metód PCA a catPCA vyplýva, že prvé dva smery maximálneho rozptylu sú podobné prvým dvom smerom, v ktorých sú pozorovania z rôznych kategórií najviac vzdialené. Dochádza k čiastočnému prekryvu „mrakov" dát prislúchajúcim jednotlivým kategóriam. Napriek tomu pozorujeme, že medzi kategóriami existujú určité rozdiely v smeroch maximálneho rozptylu. Prvý smer maximálneho rozptylu vysvetľuje iba 28.51% celkového rozptylu. Vo vektoroch $\hat{\boldsymbol{\alpha}}_1$, respektíve $\hat{\boldsymbol{\alpha}}_2$ metódy PCA sa nevyskytuje hodnota, ktorá by výrazne dominovala nad ostatnými hodnotami. Narozdiel od analýzy predchádzajúcich dátových súborov preto nevieme jednoznačne určiť, ktorým kvantitatívnym premenným zodpovedajú dané smery. Za smerodajné budeme pokladať kvantitatívne premenné, ktoré zodpovedajú prvým dvom najväčším hodnotám vektora v absolútnej hodnote. Rovnaký prístup zvolíme aj pri metóde catCCA a PDD. Na základe tohto zjednodušenia je pre smer $\hat{\boldsymbol{\alpha}}_1$ metódy PCA smerodajná 14. a 16. kvantitatívna premenná. Tieto premenné zodpovedajú y-ovej súradnici posledného a predposledného bodu. Obidve príslušné hodnoty vektora sú záporné. Pozorovania, ktoré nadobúdajú nižšie hodnoty, by mali prevažne zodpovedať cifrám, ktoré pri písaní boli zakončené hore. Naopak, pozorovania, ktoré sú zobrazené v pravej časti grafu, teda nadobúdajú vyššie hodnoty v prvom smere, by mali zodpovedať cifrám zakončeným dole. Pre druhý smer $\hat{\boldsymbol{\alpha}}_2$ metódy PCA je smerodajná 11. a 15. kvantitatívna premenná. Tieto premenné zodpove-



Obr. 13: Vizualizácia dátového súboru *Pen-Based Recognition of Handwritten Digits* metódou analýzy hlavných komponentov (PCA). Projekcie pozorovaní z jednotlivých kategórií sú farebne odlíšené a znázornené bodkami. Väčšími kruhmi sú znázornené projekcie príslušných výberových priemerov.



Obr. 14: Vizualizácia dátového súboru *Pen-Based Recognition of Handwritten Digits* metódou analýzy hlavných komponentov pre kategorizované dáta (catPCA). Projekcie pozorovaní z jednotlivých kategórií sú farebne odlíšené a znázornené bodkami. Väčšími kruhmi sú znázornené projekcie príslušných výberových priemerov.

dajú x-ovej súradnici 6. a posledného bodu. Vyššie hodnoty v smere $\hat{\boldsymbol{\alpha}}_2$ by mali nadobúdať tie pozorovania, ktoré zodpovedajú cifrám, ktoré pri písaní boli zakončené krivkou smerujúcou zľava doprava. Kvôli prehľadnosti postupne analyzujeme každú kategóriu. Z každej kategórie sme vybrali tri pozorovania, ktorých hodnoty sme graficky znázornili na Obr. 15. Prvé dve pozorovania pre každú kategóriu sme náhodne vybrali tak, aby pochádzali z rôznych časti prislúšneho „mraku" dát. Výnimkou je kategória 5, pre ktorú pozorujeme dva oddelené „mraky" dát. Približne zo stredu každého „mraku" sme vybrali jedno pozorovanie. Tretie vybrané pozorovanie zodpovedá pozorovaniu, ktoré by mohlo byť klasifikované ako odľahlé. V prípade niektorých kategórií sme ho určili jednoznačne, v prípade iných menej jednoznačne.

- **0**: Táto kategória patrí k najlepšie odseparovaným od ostatných kategórií. Zodpovedajúci „mrak" má banánovitý tvar. Pozorovania blízko výberového priemeru by mali byť nuly, ktoré sú ukončené blízko vrcholu, pretože „mrak" leží v ľavej časti grafu. Zároveň tieto pozorovania by mali byť písané v kladnom smere, pretože krivka medzi 6. a 8. bodom by mala prechádzať výrazne sprava doľava. Nami vybraté pozorovanie z tejto oblasti je na Obr. 15.1 a je v súlade s našimi očakávaniami. Horná vetva „mraku" by mala pozostávať z núl, ktoré sú tiež písané v kladnom smere, ale končia ešte pred vrcholom nuly, keďže v záverečnej časti by krivka mala mať menej výrazný posun sprava doľava. Nami zvolené pozorovanie z hornej vetvy je na Obr. 15.2 a je v súlade s očakávaniami. V dolnej vetve by sa mali nachádzať nuly, ktoré končia vľavo v strede. Pre dolnú vetvu nemáme očakávania týkajúce sa orientácie smeru. Na Obr. 15.3 je pozorovanie, ktoré sme posúdili ako odľahlé vzhľadom na ostatné pozorovania z rovnakej kategórie. Kvôli iným pozorovaniam ho na Obr. 13 nevidíme. Súradnice pozorovania preto uvádzame v popise Obr. 15.3. Aj ľudským okom by bolo náročne uhádnuť, že ide o grafické zobrazenie nuly.

- **1**: „Mrak" dát prislúchajúci kategórii 1 je umiestnený v pravej časti grafu, čo znamená, že krivky končia dole. V druhom smere je mrak natiahnutý a siaha od záporných hodnôt takmer až po maximálne kladné hodnoty. Príčinu môžeme vidieť na Obr. 15.4, ktorý reprezentuje pozorovanie z dolnej časti „mraku", a Obr. 15.5, ktorý reprezentuje pozorovanie z hornej časti „mraku". Niektorí ľudia ukončujú jed-

notku dolnou vodorovnou čiarou, ktorá znamená posunutie zľava doprava, a teda umiestnenie v hornej časti grafu. Na Obr. 15.6 je pozorovanie, ktoré sme posúdili ako odľahlé vzhľadom na ostatné pozorovania z rovnakej kategórie. Nie je zjavné, že ide o reprezentáciu jednotky.

- **2**: Pre kategóriu 2 je príslušný „mrak" kompaktný. Vybrali sme pozorovania z ľavej a pravej časti mraku, Obr. 15.7 a Obr. 15. Na príklade sú zobrazené rôzne varianty písania dvojky. Krivka obidvoch variánt však končí dole a na konci smeruje zľava doprava, preto sú aj varianty zobrazené blízko seba.

- **3**: „Mrak" dát prislúchajúci kategórii 3 je tiež kompaktný. Je pravdepodobné, že ak aj existujú rôzne varianty písania trojky, tak všetky sú písané zhora nadol a krivka na konci stále smeruje sprava doľava. Grafické zobrazenie tretieho vybraného pozorovania z danej kategórie na Obr. 15.12 pripomína skôr dvojku ako trojku. Je možné, že pozorovanie má chybne priradenú kategóriu.

- **4**: Pozorovania z kategórie 4 tvoria „mrak" okrúhleho tvaru. Natiahnutie v druhom smere zapríčiňuje sklon poslednej časti krivky. Na Obr. 15.13, na ktorom je pozorovanie z ľavej hornej časti „mraku", pozorujeme takmer zvislý koniec. Na Obr. 15.14, na ktorom je pozorovanie z pravej dolnej časti „mraku", je koniec krivky šikmý, smerujúci zľava hore vpravo dole. Domnievame sa, že príčina natiahnutia „mraku" v prvom smere spočíva v tom, či cifra bola písaná jedným ťahom. Ak áno, tak predposledný bod očakávame nižšie ako v opačnom prípade. Pozorovania písané jedným ťahom teda očakávame viac vľavo. Pozorovania z kategórie 4 sa výrazne prekrývajú s pozorovaniami z kategórie 6.

- **5**: Pozorovania z kategórie 5 tvoria výrazne oddelené dva kompaktné „mraky". Vybrané pozorovania z ľavého horného a pravého dolného mraku sú na Obr. 15.16 a Obr. 15.17. Badáme, že pravý dolný „mrak" zodpovedá tým päťkám, ktoré sú napísané jedným ťahom, a teda koniec krivky je dole a smeruje sprava doľava. Ľavý horný „mrak" tvoria tie päťky, ktoré boli napísané dvoma ťahmi. Druhý ťah tvorila horná vodorovná čiara smerujúca zľava doprava. Tretie vybrané pozorovanie považujeme za odľahlé. Leží mimo spomínaných dvoch „mrakov". Z grafického znázornenia na Obr. 15.18 sa domnievame, že ide skôr o pozorovanie z kategórie 6.

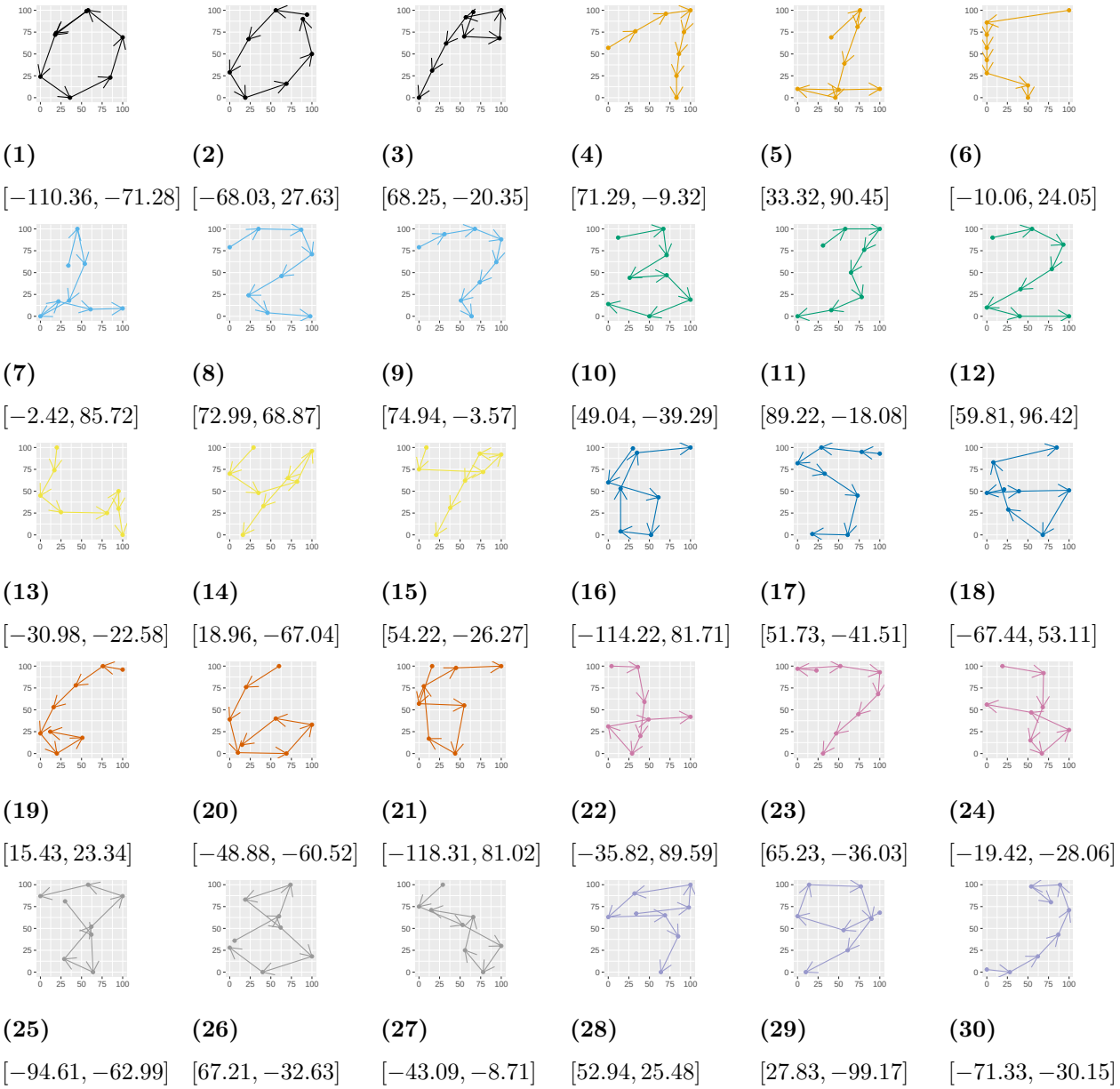
- **6**: „Mrak" prislúchajúci kategórii 6 nie je umiestnený vpravo. Z toho vyplýva, že krivky pozorovaní nie sú ukončené hore, a teda šestky sú písané zhora nadol. „Mrak" má šikmý oválny tvar. Myslíme si, že rozptyl pozorovaní v prvom a v druhom smere sa neodvíja primárne od polohy posledného bodu, ale od polohy 7., respektíve 6. bodu. Od nich závisí veľkosť dolného oblúka, teda maximálna dosahovaná výška a dĺžka medzi pravým krajným a ľavým krajným bodom oblúka. Tretie vybrané pozorovanie leží v „mraku" prislúchajúcemu kategórií 5. Aj z grafického znázornenia na Obr. 15.21 sa zdá, že skôr ide o cifru 5.

- **7**: Pre kategóriu 7 je „mrak" dát rozptýlený predovšetkým v druhom smere. Ak je sedmička písaná jedným ťahom bez strednej vodorovnej čiary, smeruje koniec krivky sprava doľava. Príkladom je pozorovanie na Obr. 15.23 pochádzajúce z dolnej časti „mraku". Ak sedmička obsahuje strednú vodorovnú čiaru, smeruje koniec krivky zľava doprava, a teda je pozorovanie umiestnené v hornej časti „mraku", viď Obr. 15.22.

- **8**: Pozorovania z kategórie 8 sú veľmi rozptýlené, pretože pre cifru 8 existuje viacero variant písania. Krivky niektorých z nich končia hore, iných dole. Taktiež pre niektoré platí, že koniec krivky smeruje sprava doľava a pre iné naopak, zľava doprava. Napriek veľkému rozptylu v rámci kategórie je väčšina pozorovaní dobre odseparovaná od porozovaní z iných kategórií.

- **9**: „Mrak" dát prislúchajúci kategórii 9 je banánovitého tvaru. Krivka pre pozorovania z pravej hornej časti „mraku" by mala končiť takmer zvislo dole bez výrazného dolného poloblúka. Grafické znázornenie reprezenta tejto oblasti je na Obr. 15.28. Krivky pozorovaní z dolnej ľavej časti „mraku" sú zakončené poloblúkom, a teda koniec krivky smeruje sprava doľava a môže byť vyššie ukončený. Ukážka pozorovania z tejto oblasti je na Obr. 15.30.

Okrem samotného „mraku" dát je zaujímavé sledovať aj polohu výberových priemerov. Tie nám naznačujú preferovaný spôsob písania cifier. Napríklad v prípade kategórie 5 je výberový priemer zobrazený bližšie k ľavému hornému „mraku" dát. Väčšina ľudí teda nenapísala päťku jedným ťahom, ale hornú vodorovnú čiaru cifry napísali nakoniec.

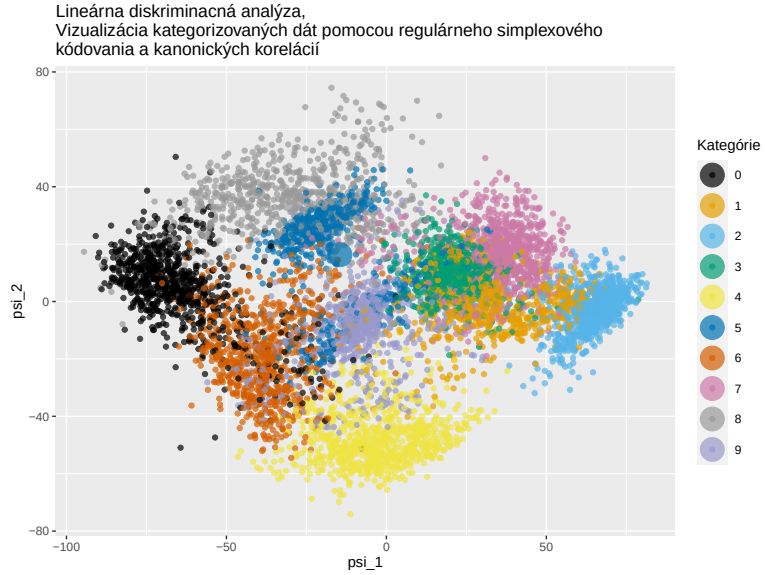


Obr. 15: Grafické znázornenie hodnôt kvantitatívnych premenných vybraných pozorovaní. Z každej kategórie sú vybrané tri pozorovania pochádzajúce z príslušného „mraku" dát získaného metódou PCA. Prvé dve pozorovania sú vybrané z rôznych častí „mraku", tretie pozorovanie sme vyberali z odľahlých pozorovaní.

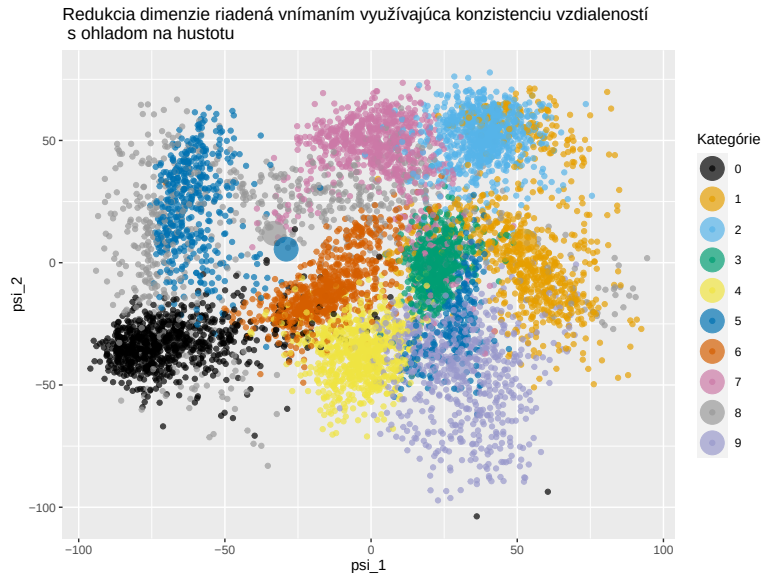
Aj vizualizácie pomocou metód catCCA a PDD je možné takto podrobne analyzovať. Avšak nechceme čitateľa zaťažiť ďalšími podobnými závermi, preto uvádzame stručnejšiu interpretáciu vizualizácií pomocou metód catCCA a PDD.

Vizualizácia pomocou metódy catCCA je na Obr. 16. Táto vizualizácia je výrazne odlišná od predchádzajúcej vizualizácie. „Mraky" dát prislúchajúce jednotlivým kategóriam sú kompaktnejšie. Zmenilo sa aj ich vzájomné rozmiestnenie. Z Tabuľky 6 vyplýva, že pre prvý smer $\hat{\boldsymbol{\alpha}}_1$ je smerodajná 4. a 16. kvantitatívna premenná. Tie zodpovedajú y-ovej súradnici 2. a 8. bodu. Hodnota pozorovania v prvom smere závisí kladne od 4. kvantitatívnej premennej a záporne od 16. kvantitatívnej premennej. Pre druhý smer $\hat{\boldsymbol{\alpha}}_2$ je opäť smerodajná 4. a 16. kvantitatívna premenná. Tentokrát je však o niečo väčší dôraz kladený na 16. kvantitatívnu premennú. Hodnota v druhom smere závisí kladne od obidvoch spomínaných premenných. V hornej časti grafu by sa mali nachádzať pozorovania, ktorých príslušná krivka začína aj končí hore. V pravej časti grafu by zase mali dominovať pozorovania, ktorých príslušná krivka začína hore a končí dole. „Mraky" dát prislúchajúce kategórii 5 sa nachádzajú podstatne bližšie pri sebe. Dátové body prislúchajúce päťkám písaným jedným ťahom sa opäť prekrývajú s pozorovaniami z kategórie 9. Niektoré kategórie sú však teraz lepšie odseparované od ostatných kategórií. Príkladom sú kategórie 4 a 6. Na vizualizácii pomocou PCA sa pozorovania z týchto kategórií prekrývali. Metóda catCCA ich dokázala relatívne dobre odseparovať. Podobný jav nastal aj pre pozorovania z kategórií 2 a 7.

Posledná vizualizácia daného dátového súboru metódou PDD je na Obr. 17. Metóda PDD nám poskytla odlišný dvojrozmerný pohľad na dáta ako predchádzajúce metódy. Líši sa vzájomné rozmiestnenie „mrakov". Napriek tomu, že na prvý a druhý smer $\hat{\boldsymbol{\alpha}}_1$ a $\hat{\boldsymbol{\alpha}}_2$ metódy PDD nie je kladená žiadna podmienka kolmosti, sa tieto smery odlišujú. Pre smer $\hat{\boldsymbol{\alpha}}_1$ je smerodajná 2. a 14. kvantitatívna premenná. Smerodajné premenné zodpovedajú y-ovej súradnici 1. a 7. bodu. Hodnota pozorovania v prvom smere závisí záporne od obidvoch týchto premenných. Pozorovania umiestnené v pravej časti grafu by mali zodpovedať pozorovaniam, ktorých krivka začína dole a tesne pred koncom je opäť dole. Pre smer $\hat{\boldsymbol{\alpha}}_2$ je smerodajná 10. kvantitatívna premenná, ktorá zodpovedá y-ovej súradnici 5. bodu, a 11. kvantitatívna premenná, ktorá zodpovedá x-ovej súradnici 6. bodu. Od obidvoch spomínaných premenných závisí hodnota v druhom smere záporne. Pre pozorovania



Obr. 16: Vizualizácia dátového súboru *Pen-Based Recognition of Handwritten Digits* metódou lineárnej diskriminačnej analýzy (LDA), respektíve metódou vizualizácie kategorizovaných dát pomocou regulárneho simplexového kódovania a kanonických korelácií (catCCA). Projekcie pozorovaní z jednotlivých kategórií sú farebne odlíšené a znázornené bodkami. Väčšími kruhmi sú znázornené projekcie príslušných výberových priemerov.



Obr. 17: Vizualizácia dátového súboru *Pen-Based Recognition of Handwritten Digits* metódou redukcie dimenzie riadenou vnímaním využívajúcou konzistenciu vzdialeností s ohľadom na hustotu (PDD). Projekcie pozorovaní z jednotlivých kategórií sú farebne odlíšené a znázornené bodkami. Väčšími kruhmi sú znázornené projekcie príslušných výberových priemerov.

umiestnené v hornej časti grafu by malo byť charakteristické, že ich 5. bod je umiestnený dole a 6. bod je umiestnený vľavo. V porovnaní s metódou catCCA dokázala metóda PDD lepšie vizuálne odseparovať pozorovania z kategórií 1, 3 a 7. „Mraky" dát sú menej kompaktné ako v prípade vizualizácie metódou catCCA. To značí, že v smerodajných súradniciach bodov sa pozorovania z rovnakej kategórie viac odlišujú.

Vizualizácie pomocou vybraných metód nám poskytli rôzne pohľady na dáta. Z každej sme dokázali vyvodiť iné užitoćné charakteristiky.

Populárny dátový súbor *MNIST* taktiež obsahuje údaje o rukou napísaných cifrách. Oproti dátovému súboru *Pen-Based Recognition of Handwritten Digits* sa odlišuje typ informácie, ktorú obsahujú kvantitatívne premenné. Ako sme už spomínali, dátový súbor *Pen-Based Recognition of Handwritten Digits* obsahuje x-ové a y-ové súradnice 8 bodov. Dátový súbor *MNIST* obsahuje pre každé pozorovanie informácie o odtieni šedej na obrázku s rozlíšením 28×28 pixelov. Okrem kategoriálnej premennej obsahuje 784 kvantitatívnych premenných. Výhodou *Pen-Based Recognition of Handwritten Digits* je, že ide o dáta menšieho rozmeru, a teda pri jeho analyzovaní sú výpočty menej náročné. Zároveň sme na vizualizáciach pozorovali, že aj tieto menej rozmerné dáta obsahujú dostatok informácie na odlíšenie pozorovaní z rôznych kategórií. Na druhej strane aj typ informácií v *MNIST* má výhodu. Reprezentácia cifier odtieňami šedej nie je závislá na smere písania cifry ani na počte potrebných ťahov.

Záver

Hlavným cieľom tejto diplomovej práce bolo predstaviť vlastnú metódu vizualizácie kategorizovaných dát. Na naplnenie tohto cieľa bolo najprv nutné charakterizovať kategorizované dáta a vymedziť miesto kvantitatívnych a kategoriálnych premenných. Potrebnej terminológii sme sa venovali v kapitole 1.

Naša metóda vizualizácie kategorizovaných dát pomocou regulárneho simplexového kódovania a kanonických korelácií patrí medzi lineárne metódy vizualizácie dát. Rozhodli sme sa uľahčiť čitateľovi orientáciu v danej oblasti a v kapitole 2 sme začlenili našu metódu do systému metód vizualizácie dát spojených s redukciou dimenzie. Ako sme v tejto kapitole uviedli, lineárne metódy majú z hľadiska flexibility a interpretácie iné črty ako nelineárne metódy. Oproti nelineárnym metódam sa sústredia na vernejšiu reprezentáciu geometrie a vzájomnej polohy dát v nízkorozmernom priestore za cenu nižšej schopnosti prispôsobiť sa špecifickým nelinearitám v dátach. Vedľajším cieľom práce bolo v praktickej časti porovnať výsledky vizualizácií našej metódy a iných vybraných metód. Chceli sme porovnávať „podobné" metódy, preto sme vyberali z lineárnych metód vizualizácie dát.

Ako zástupcov sme zvolili jednu z najpopulárnejších metód, a to analýzu hlavných komponentov (PCA), modifikovanú verziu analýzy hlavných komponentov pre kategorizované dáta (catCCA), lineárnu diskriminačnú analýzu (LDA) a redukciu dimenzie riadenú vnímaním využívajúcu konzistenciu vzdialeností s ohľadom na hustotu (PDD). Spomínané metódy boli zvolené tak, aby napriek tomu, že sú všetky lineárne, boli medzi ich ideami aj rozdiely. Pri popise metód v kapitolách 3 a 4 sme sa snažili o jednotnú štruktúru naprieč metódami. Konkrétne, v úvode sme uviedli hlavnú myšlienku metódy, potom matematickú formuláciu a prípadné odvodenie, praktický návod na vizualizáciu a v závere sme zhrnuli výhody a nevýhody. Jednotlivé kroky sme sa snažili podrobne zdôvodniť, aby boli pre čitateľa zrozumiteľné. Viaceré odvodenia sme doplnili aj o vlastné výpočty.

Vizualizácii kategorizovaných dát pomocou regulárneho simplexového kódovania a kanonických korelácií sme sa venovali v samostatnej kapitole 5. V podkapitole 5.1 sme najprv sformulovali pôvodnú metódu kanonických korelácií, z ktorej sme vychádzali pri odvodení našej myšlienky. Základnou ideou bolo vloženie pomocných premenných pre kategoriálnu premennú do druhej sady premenných. V podkapitole 5.2.3 sme ukázali, že nezáleží na

tom, aký systém pomocných premenných zvolíme, pokiaľ je príslušná kovariančná matica regulárna. Odvodenie tohto poznatku pokladáme za jeden z prínosov tejto práce. Aj keď náš prístup k tvorbe pomocných premenných pomocou regulárneho simplexového kódovania pri tejto metóde nevyužil plný potenciál, tiež ho pokladáme za prínos tejto práce. Tento typ pomocných premenných môže byť užitočný v metódach, ktoré nemajú vlastnosť uvedenú v Leme 5.3. Počas výskumu sa ukázalo, že naša metóda neposkytuje unikátne pohľady na mnohorozmerné dáta, ale je ekvivalentná s metódou lineárnej diskriminačnej analýzy. Užitočná je však alternatívna interpretácia metódy lineárnej diskriminačnej analýzy. Smery, v ktorých sa minimalizuje rozptyl pozorovaní z rovnakej kategórie a zároveň maximalizuje rozptyl priemerov rôznych kategórií, sú také, ktoré najlepšie vysvetľujú jednotlivé kategórie.

V poslednej praktickej časti, v kapitole 6, sme aplikovali vybrané metódy na umelý, menej rozmerný reálny a viacrozmerný reálny dátový súbor. Pozorovali sme, že výsledky vizualizácií daných dátových súborov sa líšili. PCA v niektorých prípadoch dokázala odkryť nielen celkovú štruktúru dát, ale aj štruktúru z hľadiska ich príslušnosti k rôznym kategóriam. Zdá sa, že jej modifikácia pre kategorizované dáta catPCA vykazuje výrazne odlišné zobrazenie ako pôvodná metóda len v špeciálnych prípadoch. Oproti metódam PDD a catCCA, respektíve LDA nekladie žiadnu podmienku na vzťah pozorovaní z rovnakej kategórie. Smery získané metódami catCCA a PDD dobre vysvetľovali rozdiely medzi kategóriami. Narozdiel od catCCA však PDD niekedy zdegenerovala a v druhom smere nepriniesla žiadnu novú informáciu.

Dôležitým posolstvom tejto diplomovej práce je, že cieľom nie je nájsť najlepšiu metódu vizualizácie dát, ale na vytvorenie predstavy o dátach je dobré aplikovať viacero metód a z každej vizualizácie niečo vyťažiť.

Zoznam použitej literatúry

[1] Agresti, A.: *An Introduction to Categorical Data Analysis*, John Wiley & Sons, New Jersey, 2007

[2] Boyd, S.,Vandenberghe, L.: *Convex Optimization*, Cambridge University Press, New York, 2004

[3] Dua, D., Graff, C.: *UCI Machine Learning repository*, University of California, School of Information and Computer Sciences, Irvine, 2019, dostupné na internete (15.4.2021): http://archive.ics.uci.edu/ml

[4] Espadoto, M. et al.: *Towards a Quantitative Survey of Dimension Reduction Techniques*, IEEE Transaction on Visualization and Computer Graphics 27 (2021), 2153-2173

[5] Gareth, J. et al.: *An Introduction to Statistical Learning with Applications in R*, Springer, New York, 2013

[6] Harman, R.: *Multivariate Statistical Analysis: Selected Lecture Notes*, učebné texty, FMFI UK, Bratislava, 2020, dostupné na internete (20.12.2020): http://www.iam.fmph.uniba.sk/ospm/Harman/VSAp.pdf

[7] Harville, D. A.: *Matrix Algebra From a Statistician's Perspective*, Springer, New York, 2008

[8] Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, 2011

[9] Izenman, A. J.: *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning*, Springer, New York, 2008

[10] Koren, Y., Carmel, L.: *Visualization of Labeled Data Using Linear Transformations*, IEEE Symposium on Information Visualization (2003), 121-128

[11] R Core Team: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020, dostupné na internete (13.4.2021): https://www.R-project.org/

[12] Rudas, T.: *Lectures on Categorical Data Analysis*, Springer, New York, 2018

[13] Siblini, W., Kuntz, P., Meyer, F.:*A Review on Dimensionality reduction for Multi-label Classification*, IEEE Transactions on Knowledge and Data Engineering 33 (2021), 839-857

[14] Upton, G.: *Categorical Data Analysis by Example*, John Wiley & Sons, New Jersey, 2017

[15] Wang, Y. et. al.: *A Perception-Driven Approach to Supervised Dimensionality Reduction for Vizualization*, IEEE Transaction on Visualization and Computer Graphics 24 (2018), 1828-1840

Príloha

```
# I. KNIZNICE
library(datasets); library(dplyr); library('rgl')
library(RColorBrewer); library(ggplot2); library(GGally)
# II. FUNKCIE
# 4.1 catPCA
catpcacomp <- function(X,catcol,scale,center){
   X.kvant <- X[,-catcol] # kvantitativne premenne
   X.cat <- X[,catcol] # kategorialna premenna
   # preskalovanie  a centrovanie
   if(scale == TRUE){
     X.kvant <- sweep(X.kvant,2,sqrt(apply(X.kvant,2,var)),FUN = "/")}
   if(center == TRUE){
     X.kvant <- sweep(X.kvant,2,apply(X.kvant,2,mean),FUN = "-")}
   # Laplacian
   L <- matrix(rep(0,length(X.cat)^2),nrow = length(X.cat))
   for (i in 1:length(X.cat)){
    for (j in 1:length(X.cat)){
     if(X.cat[i] != X.cat[j]){
           L[i,j] <- -1}}}
   L <- L - diag(apply(L,2,sum))
   H1 <- t(as.matrix(X.kvant)) %*% L %*% as.matrix(X.kvant)
   # zoradene vlastne vektory
   x <- eigen(H1)$vectors[,order(eigen(H1)$values,decreasing = TRUE)]
   print(x)
   return(as.matrix(X.kvant) %*% x)}
# 4.2 LDA,
ldacomp <- function(X,catcol,center){
    X.kvant <- X[,-catcol] # kvantitativne premenne
    X.cat <- X[,catcol] # kategorialna premenna
    # centrovanie
    if(center == TRUE){
       X.kvant <- sweep(X.kvant,2,apply(X.kvant,2,mean),FUN = "-")}
```

```
   k <- length(unique(X.cat)) # pocet kategorii
   # celkovy priemer a priemery vramci jednotlivych kategorii
   X.mean <- apply(X.kvant,2,mean)
   X.cat.mean <- matrix(rep(0,k * ncol(X.kvant)),nrow = k)
   # matice Sb a Sw
   Sb <- matrix(rep(0,ncol(X.kvant) ^ 2),nrow = ncol(X.kvant))
   Sw <- matrix(rep(0,ncol(X.kvant) ^ 2),nrow = ncol(X.kvant))
   cat.index <- 1
   for (i in unique(X.cat)){
     X.cat.mean[cat.index,] <- apply(X.kvant[X.cat == i,],2,mean)
     n_i <- length(X.cat[X.cat == i])
     Sb <- Sb + n_i * (X.cat.mean[cat.index,] - X.mean) %*%
           t(X.cat.mean[cat.index,] - X.mean)
     M <- X.kvant[X.cat == i,] - matrix(rep(X.cat.mean[cat.index,],
          length(X.cat[X.cat == i])), nrow = length(X.cat[X.cat == i]),
          byrow = TRUE)
     Sw <- Sw + t(as.matrix(M)) %*% as.matrix(M)
     cat.index <- cat.index + 1}
   Sb <- Sb / length(X.cat)
   Sw <- Sw / length(X.cat)
   H1 <- solve(Sw) %*% Sb
   # zoradene vlastne vektory
   x <- eigen(H1)$vectors[,order(eigen(H1)$values,decreasing = TRUE)]
   print(x)
   return(as.matrix(X.kvant) %*% x)}
# 5.2 catCCA
catccacomp <- function(X,catcol,center){
   X.kvant <- X[,-catcol] # kvantitativne premenne
   X.cat <- X[,catcol] # kategorialna premenna
   # centrovanie
   if(center == TRUE){
       X.kvant <- sweep(X.kvant,2,apply(X.kvant,2,mean),FUN = "-")}
   k <- length(unique(X.cat)) # pocet kategorii
```

```
  # vektory smerujuce k vrcholom pravidelneho simplexu
  W <- eigen(2 * diag(k) - 1)$vectors
  V <- W[,1:(k-1)]
  Z <- matrix(rep(0,length(X.cat) * (k - 1)),nrow = length(X.cat))
  cat.index <- 1
  for (i in unique(X.cat)){
    for (j in 1:length(X.cat)){
      if (X.cat[j] == i){
        Z[j,] <- V[cat.index,]}}
    cat.index <- cat.index + 1}
  # vyberove kovariancne matice a ich sucin, respektive ich inverzii
  S11 <- cov(X.kvant); S22 <- cov(Z)
  S12 <- cov(X.kvant,Z); S21 <- t(S12)
  H1 <- solve(S11) %*% S12 %*% solve(S22) %*% S21
  # zoradene vlastne vektory
  x <- eigen(H1)$vectors[,order(eigen(H1)$values,decreasing = TRUE)]
  print(x)
  return(as.matrix(X.kvant) %*% x)}
# 4.3 PDD
E <- function(X.kvant, X.cat.num, A){
  # pomocna funkcia, vyhodnocuje hodnotu kriteria pre projekcnu maticu A
  X.proj <- as.matrix(X.kvant) %*% A # projekcia premennych
  k <- length(unique(X.cat.num)) # pocet kategorii
  # priemery projekcii  jednotlivych kategorii
  X.proj.cat.mean <- matrix(rep(0, k * ncol(X.proj)),nrow = k)
  for (i in 1:k){
    X.proj.cat.mean[i,] <- apply(X.proj[X.cat.num == i,],2,mean)}
  # vzdialenosti projekcie datoveho bodu od projekcie vyb. priemeru, z kt.
  # pochadza, a od projekcie ineho najblizsieho vyb. priemeru
  a <- matrix(rep(0,length(X.cat.num)),nrow = length(X.cat.num))
  b <- matrix(rep(Inf,length(X.cat.num)),nrow = length(X.cat.num))
  # hodnota kriteria
  s <- matrix(rep(0,length(X.cat.num)),nrow = length(X.cat.num))
```

```
   for (j in 1:length(X.cat.num)){
     a[j] <- dist(matrix(c(X.proj[j,],X.proj.cat.mean[X.cat.num[j],]),
             nrow = 2,byrow = TRUE),method = "euclidean")
     for (i in 1:k){
       if(i != X.cat.num[j]){
         b[j] <- min(b[j],dist(matrix(c(X.proj[j,],X.proj.cat.mean[i,]),
                 nrow = 2,byrow = TRUE),method = "euclidean"))}}
     s[j] <- (b[j] - a[j]) / max(a[j],b[j])}
   return(apply(s,2,mean)) # priemerna hodnota kriteria}

pddcomp <- function(X,catcol,scale,center){
    X.kvant <- X[,-catcol] # kvantitativne premenne
    X.cat <- X[,catcol] # hodnoty kategorialna premenna
    # skalovanie a centrovanie
    if(scale == TRUE){
       X.kvant <- sweep(X.kvant,2,sqrt(apply(X.kvant,2,var)),FUN = "/")}
    if(center == TRUE){
       X.kvant <- sweep(X.kvant,2,apply(X.kvant,2,mean),FUN = "-")}
    # numericke indexy kategorii
    X.cat.num <- matrix(rep(0,length(X.cat)),nrow = length(X.cat))
    cat.index <- 1
    for (i in unique(X.cat)){
       X.cat.num[X.cat == i] <- cat.index
       cat.index <- cat.index + 1}
    # nahodne vygenerovana projekcna matica
    A <- matrix(runif(2 * ncol(X.kvant),-1,1),nrow = ncol(X.kvant))
    A <- sweep(A,2,sqrt(apply(A ^ 2,2,sum)),FUN = "/")
    T <- 100 * ncol(X.kvant) # zaciatocna teplota
    alpha <- 0.95 # parameter redukcie teploty
    # parameter okolia prvkov projekcnej matice
    r <- 0.05; m <- 0.01
    # simulovane zihanie
    while (T > 0.001){
```

```
      Q1 <- matrix(runif(2 * ncol(X.kvant)),nrow = ncol(X.kvant))
      Q1[Q1 <= 0.5] <- 0
      Q1[Q1 > 0.5] <- r
      if (E(X.kvant,X.cat.num,A + Q1) < E(X.kvant,X.cat.num,A - Q1)){
         Q1 <- -Q1}
      Q1[Q1 == 0] <- sign(runif(round(2 * ncol(X.kvant) - sum(abs(Q1))/ r),-1,1)) * m
      Q1 <- A + Q1
      Q1 <- sweep(Q1,2,sqrt(apply(Q1 ^ 2,2,sum)),FUN = "/")
      deltaE <- E(X.kvant,X.cat.num,Q1) - E(X.kvant,X.cat.num,A)
      if (deltaE > 0){
        A <- Q1}
      else if (exp(deltaE / T) > runif(1)){
        A <- Q1}
      T <- T * alpha
      print(T)
      print(E(X.kvant,X.cat.num,A))}
    return(as.matrix(X.kvant) %*% A)}
# III. DATASETY
# UMELE DATA A
set.seed(110) # zaistenie repredukcie vysledkov
X1 <- matrix(rnorm(1000,mean = 0,sd = 0.5),nrow = 1000)
X2 <- matrix(rnorm(1000,mean = 0,sd = 2.5),nrow = 1000)
X3 <- rbind(matrix(rnorm(750,mean = 0,sd = 2),nrow = 750),
matrix(rnorm(250,mean = 0,sd = 5),nrow = 250))
X4 <- c(rep(1,250),rep(2,250),rep(3,250),rep(4,250))
X <- cbind(X1,X2,X3,X4)
X[251:500,1] <- X[251:500,1] + 2
X[501:750,1] <- X[501:750,1] + 4
X[751:1000,1] <- X[751:1000,1] + 6
catcol <- 4
k <- length(unique(X[,catcol]))
# IRIS
data(iris); X <- iris; attach(X)
```

```
catcol <- 5
k <- length(unique(X[,catcol]))
# DIGITS
X <- read.table("pendigits.txt",header = FALSE,sep=",")
attach(X)
catcol <- 17
k <- length(unique(X[, catcol]))
# IV. VIZUALIZACIE
cbPalette <- c("#000000","#E69F00","#56B4E9","#009E73","#F0E442",
               "#0072B2","#D55E00","#CC79A7","#999999","#9999CC")
# 3.1 PCA
Psi.pca <- prcomp(X[,-catcol],scale = FALSE)
mean.cat <- matrix(rep(0,2 * k),ncol = 2)
for (i in 1:k){
   for (j in 1:2){
     cat <- as.character(unique(X[,catcol])[i])
     mean.cat[i,j] <- apply(Psi.pca$x[X[,catcol] == cat,],2,mean)[j]}}
(ggplot()
  + geom_point(data = data.frame(Psi.pca$x),
    aes(x = data.frame(Psi.pca$x)[,1],y = data.frame(Psi.pca$x)[,2],
    color = as.factor(X[,catcol])),shape = 19,alpha = 0.7)
  + geom_point(data = data.frame(mean.cat),
    aes(x = data.frame(mean.cat)[,1], y = data.frame(mean.cat)[,2],
    color = as.factor(unique(X[,catcol]))),size = 8,shape = 19,alpha = 0.7)
  + scale_colour_manual(values = cbPalette,name = "Kategorie")
  + ggtitle("Analyza hlavnych komponentov")
  + xlab("psi_1")
  + ylab("psi_2"))
summary(Psi.pca) # podiel vysvetleneho rozptylu
# 4.1 CATPCA
Psi.catpca <- catpcacomp(X,catcol,FALSE,TRUE)
mean.cat <- matrix(rep(0,2 * k),ncol = 2)
for (i in 1:k){
```

```
    for (j in 1:2){
     cat <- as.character(unique(X[,catcol])[i])
     mean.cat[i,j] <- apply(Psi.catpca[X[,catcol] == cat,],2,mean)[j]}}
(ggplot()
  + geom_point(data = data.frame(Psi.catpca),
    aes(x = data.frame(Psi.catpca)[,1],y = data.frame(Psi.catpca)[,2],
    color = as.factor(X[,catcol])),shape = 19,alpha = 0.7)
  + geom_point(data = data.frame(mean.cat),
    aes(x = data.frame(mean.cat)[,1],y = data.frame(mean.cat)[,2],
    color = as.factor(unique(X[,catcol]))),size = 8,shape=19,alpha = 0.7)
  + scale_colour_manual(values = cbPalette,name = "Kategorie")
  + ggtitle("Analyza hlavnych komponentov pre kategorizovane data")
  + xlab("psi_1")
  + ylab("psi_2"))
# podiel vysvetleneho rozptylu v 1. a 2. smere
cov(as.matrix(Psi.catpca[,1])) / sum(apply(X[,-catcol],2,var))
cov(as.matrix(Psi.catpca[,2])) / sum(apply(X[,-catcol],2,var))
# 4.2 LDA
Psi.lda <- ldacomp(X,catcol,TRUE)
mean.cat <- matrix(rep(0,2 * k),ncol = 2)
for (i in 1:k){
    for (j in 1:2){
     cat <- as.character(unique(X[,catcol])[i])
     mean.cat[i,j] <- apply(Psi.lda[X[,catcol] == cat,],2,mean)[j]}}
(ggplot()
  + geom_point(data = data.frame(Re(Psi.lda)),
    aes(x = data.frame(Re(Psi.lda))[,1],y = data.frame(Re(Psi.lda))[,2],
    color = as.factor(X[,catcol])),shape = 19,alpha = 0.7)
  + geom_point(data = data.frame(Re(mean.cat)),
    aes(x = data.frame(Re(mean.cat))[,1],y = data.frame(Re(mean.cat))[,2],
    color = as.factor(unique(X[,catcol]))),size = 8,shape=19,alpha = 0.7)
  + scale_colour_manual(values = cbPalette,name = "Kategorie")
  + ggtitle("Linearna diskriminacna analyza")
```

```
  + xlab("psi_1")
  + ylab("psi_2"))
# podiel vysvetleneho rozptylu v 1. a 2. smere
cov(as.matrix(Psi.lda[,1])) / sum(apply(X[,-catcol],2,var))
cov(as.matrix(Psi.lda[,2])) / sum(apply(X[,-catcol],2,var))
# 5.2 catCCA
Psi.catcca <- catccacomp(X,catcol,TRUE)
mean.cat <- matrix(rep(0,2 * k),ncol = 2)
for (i in 1:k){
    for (j in 1:2){
     cat <- as.character(unique(X[,catcol])[i])
     mean.cat[i,j] <- apply(Psi.catcca[X[,catcol] == cat,],2,mean)[j]}}
(ggplot()
  + geom_point(data = data.frame(Re(Psi.catcca)),
    aes(x = data.frame(Re(Psi.catcca))[,1],y = data.frame(Re(Psi.catcca))[,2],
    color = as.factor(X[,catcol])),shape = 19,alpha = 0.7)
  + geom_point(data = data.frame(Re(mean.cat)),
   aes(x = data.frame(Re(mean.cat))[,1],y = data.frame(Re(mean.cat))[,2],
   color = as.factor(unique(X[,catcol]))),size = 8,shape=19,alpha = 0.7)
  + scale_colour_manual(values = cbPalette,name = "Kategorie")
  + ggtitle("Vizualizacia kategorizovanych dat pomocou regularneho
    simplexoveho kodovania a kanonickych korelacii")
  + xlab("psi_1")
  + ylab("psi_2"))
# podiel vysvetleneho rozptylu v 1. a 2. smere
cov(as.matrix(Re(Psi.catcca)[,1])) / sum(apply(X[,-catcol],2,var))
cov(as.matrix(Re(Psi.catcca)[,2])) / sum(apply(X[,-catcol],2,var))
# 4.3 PDD
set.seed(110) # zaistenie repredukcie vysledkov
Psi.pdd <- pddcomp(X,catcol,FALSE,TRUE)
mean.cat <- matrix(rep(0,2 * k),ncol = 2)
for (i in 1:k){
    for (j in 1:2){
```

```
     cat <- as.character(unique(X[,catcol])[i])
     mean.cat[i,j] <- apply(Psi.pdd[X[,catcol] == cat,],2,mean)[j]}}
(ggplot()
  + geom_point(data = data.frame(Psi.pdd),
    aes(x = data.frame(Psi.pdd)[,1],y = data.frame(Psi.pdd)[,2],
    color = as.factor(X[,catcol])),shape = 19,alpha = 0.7)
  + geom_point(data = data.frame(mean.cat),
    aes(x = data.frame(mean.cat)[,1],y = data.frame(mean.cat)[,2],
    color = as.factor(unique(X[,catcol]))),size = 8,shape=19,alpha = 0.7)
  + scale_colour_manual(values = cbPalette,name = "Kategorie")
  + ggtitle("Redukcia dimenzie riadena vnimanim vyuzivajuca konzistenciu
    vzdialenosti s ohladom na hustotu")
  + xlab("psi_1")
  + ylab("psi_2"))
# podiel vysvetleneho rozptylu v 1. a 2. smere
cov(as.matrix(Psi.pdd[,1])) / sum(apply(X[,-catcol],2,var))
cov(as.matrix(Psi.pdd[,2])) / sum(apply(X[,-catcol],2,var))
```