UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

ANALÝZA VOLEBNÉHO SPRÁVANIA NA SLOVENSKU

BAKALÁRSKA PRÁCA

2014 Jana MATYAŠOVSKÁ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

ANALÝZA VOLEBNÉHO SPRÁVANIA NA SLOVENSKU

BAKALÁRSKA PRÁCA

Študijný program:	Ekonomická a finančná matematika
Študijný odbor:	1114 Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Mgr. Martin Niepel, PhD.

Bratislava 2014 **Jana MATYAŠOVSKÁ**

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Jana Matyašovská
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský

Názov: Analýza volebného správania na Slovensku / *Analysis of Electoral Behavior in Slovakia*

Cieľ: Porozumieť spektrálnym metódam lineárnej algebry, ktoré sa používajú pri štatistickej analýze dát (PCA, faktorová analýza). Získané zručnosti použiť na analýzu volebného správania na Slovensku v rokoch 1990 - 2012. Porovnanie výsledkov s relevantným politologickým výskumom.

Kľúčové slová: voľby na Slovensku, analýza hlavných komponentov, faktorová analýza

Vedúci: Mgr. Martin Niepel, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, CSc.

Dátum zadania: 18.10.2013

Dátum schválenia: 14.11.2013 doc. RNDr. Margaréta Halická, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Touto cestou sa chcem poďakovať svojmu vedúcemu bakalárskej práce Mgr. Martinovi Nieplovi, PhD. za ochotu a trpezlivosť pri konzultáciách a mnoho podnetných nápadov a pripomienok, ktoré mi pomohli pri písaní tejto práce. Taktiež by som rada poďakovala svojej rodine, priateľovi a kamarátom za podporu.

Abstrakt v štátnom jazyku

MATYAŠOVSKÁ, Jana: Analýza volebného správania na Slovensku [Bakalárska práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Martin Niepel, PhD., Bratislava, 2014, 67 s.

V našej práci sa zaoberáme dvoma viacrozmernými štatistickými metódami používanými na redukciu dimenzionality dát - analýzou hlavných komponentov a faktorovou analýzou. Venujeme sa tiež problému kompozičných dát a úpravám analýzy hlavných komponentov pre prácu s takýmito špeciálnymi dátami. Výsledky troch druhov analýzy poukazujú na to, že v prípade volebných dát nie je potrebné pristupovať k nim ako ku kompozičným. Výsledky analýzy hlavných komponentov a tiež faktorovej analýzy ďalej ukazujú, že významným faktorom v správaní voličov na Slovensku je ich národnosť.

Kľúčové slová: voľby na Slovensku, analýza hlavných komponentov, faktorová analýza, kompozičné dáta

Abstract

MATYAŠOVSKÁ, Jana: Analysis of Electoral Behavior in Slovakia [Bachelor thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Martin Niepel, PhD., Bratislava, 2014, 67p.

In our thesis we study two multivariate statistical methods which are used to reduce data dimensionality - principal component analysis and factor analysis. We also investigate the problem of compositional data and modifications of principal component analysis for this specific type of data. The results of our analysis show that it is not necessary to think of electoral data as compositional data. The results of principal component analysis and factor analysis also show that a very important feature of electoral behavior in Slovakia is the ethnicity.

Keywords: elections in Slovakia, principal component analysis, factor analysis, compositional data

Obsah

Zoznam obrázkov **9**

Zoznam tabuliek **10**

Úvod **11**

1 Analýza hlavných komponentov **13**

1.1 Definícia a odvodenie hlavných komponentov 13

1.2 Špeciálne prípady 16

1.3 PCA a singulárny rozklad 17

1.4 Kovariančná a korelačná matica 18

1.5 Počet hlavných komponentov 19

2 Kompozičné dáta a ich analýza **21**

2.1 Kompozícia, simplex ako priestor 21

2.2 Príprava kompozičných dát na analýzu 22

2.3 PCA pre kompozičné dáta 23

2.4 Výskyt núl v dátach 26

2.5 Simpliciálny singulárny rozklad 28

3 Faktorová analýza **30**

3.1 Model faktorovej analýzy 30

3.2 Riešenie faktorového modelu pomocou hlavných komponentov 32

3.3 Počet spoločných faktorov 34

3.4 Rotácia faktorov 35

3.5 Faktorové skóre 35

4 Aplikácia PCA na volebné dáta **37**

4.1 Voľby do NR SR 1992 38

4.2 Voľby do NR SR 2002 40

4.3 Voľby do NR SR 2010 41

4.4 Voľby do NR SR a národnosť 44

4.5 Prezidentské voľby 1999 45

4.6 Aproximácia pomocou singulárneho rozkladu 46
4.7 Zhrnutie výsledkov PCA na volebných dátach 49

5 Aplikácia FA na výsledky volieb **51**
5.1 Voľby do NR SR 2010 51
5.2 Voľby do NR SR 2010 a 2012 54
5.3 Zhrnutie výsledkov FA na volebných dátach 56

Záver **58**

Zoznam použitej literatúry **61**

Príloha A - Zdrojové kódy **63**

Príloha B - Zoznam kandidátov a ich volebné výsledky **65**

Zoznam obrázkov

1 Graf pre 50 pozorovaní premenných x_1 a x_2 14

2 Trojuholníkový diagram pre kompozíciu $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ 22

3 Hlasy pre maďarské a slovenské strany 44

Zoznam tabuliek

1 Hlavné komponenty pre voľby do NR SR 1992 39
2 Hlavné komponenty pre voľby do NR SR 2002 41
3 Hlavné komponenty pre voľby do NR SR 2010 za okresy 42
4 Hlavné komponenty pre voľby do NR SR 2010 za obce 43
5 Hlavné komponenty pre voľby do NR SR 2010 za obce 46
6 Chyby aproximácie 48
7 Odhad váh faktorov pre voľby do NR SR 2010 52
8 Odhad váh faktorov pre voľby do NR SR 2010 za obce 53
9 Odhad váh faktorov pre voľby do NR SR 2010 a 2012 57
10 Zoznam a výsledky politických strán kandidujúcich vo voľbách do NR SR 1992 65
11 Zoznam a výsledky politických strán kandidujúcich vo voľbách do NR SR 2002 66
12 Zoznam a výsledky politických strán kandidujúcich vo voľbách do NR SR 2010 67
13 Zoznam a výsledky kandidátov v prezidentských voľbách 1999 67

Úvod

Jedným z hlavných nástrojov demokracie sú voľby, ktorých výsledky reprezentujú názory určitej vzorky ľudí - voličov. Tieto dáta sú často predmetom záujmu štatistikov, pretože v istom zmysle predstavujú náhľad do štruktúry obyvateľstva. Pritom zaujímavé je pracovať s výsledkami na lokálnej úrovni (okresy, obce) a predmetom štúdia bývajú výsledky jednotlivých strán alebo politických skupín.

Často využívaným nástrojom na analýzu viacrozmerných dát je tzv. analýza hlavných komponentov (anglicky Principal Component Analysis - PCA), ktorá predstavuje spôsob, ako znížiť počet dimenzií analyzovaných dát. Vyšlo niekoľko štatistických článkov zaoberajúcich sa aplikovaním PCA práve na volebné dáta, napr. [12], [14].

Ako však podotýkajú autori v [14], volebné dáta sú vo svojej podstate veľmi špeciálne - viac ako absolútne čísla nás zaujímajú podiely a v takom prípade súčet zložiek každého vektora predstavujúceho volebné výsledky napríklad v konkrétnej obci je rovný jednej. Takéto súbory dát sa v angličtine nazývajú Compositional Data, v tejto práci používame názov kompozičné dáta. Pri analýze týchto špeciálnych dát musíme byť opatrní a použitie bežných metód na analýzu viacrozmerných dát tu nie je vždy vhodné a správne.

J. Aitchinson v osemdesiatych rokoch minulého storočia v [3] a neskôr spoločne s J. J. Egozcuem v [4] navrhol úpravy metódy PCA pre potreby analýzy kompozičných dát založené na vlastnostiach simplexu ako vektorového priestoru. Tieto metódy vo veľkej miere využívajú logaritmické transformácie a stále sa rozvíjajú, predovšetkým možnosti ich interpretácie sú predmetom ďalšieho výskumu J. Aitchinsona a iných štatistikov.

Okrem PCA ďalšou obľúbenou metódou na zmenšovanie dimenzionality dát je faktorová analýza, ktorá hľadá lineárny vzťah medzi skúmanými premennými a inými (nemerateľnými) premennými nazvanými faktory. Na volebné dáta sme aplikovali aj túto metódu a snažili sme sa nájdené faktory interpretovať, s pomocou dostupnej literatúry a poznatkov o správaní slovenských voličov.

Cieľom bakalárskej práce bolo naštudovať PCA a faktorovú analýzu ako metódy lineárnej algebry využívané v štatistických analýzach a tiež modifikácie PCA na kompozičné dáta navrhnuté J. Aitchinsonom. Tieto zručnosti sme neskôr aplikovali na výsledky parlamentných a prezidentských volieb na Slovensku od roku 1990. Výsledky

analýzy sme sa pokúsili interpretovať, porovnali sme ich medzi sebou a tiež s výsledkami politologických výskumov.

Túto prácu sme rozdelili na teoretickú a praktickú časť. V prvej časti práce sú postupne vysvetlené teoretické základy PCA, kompozičných dát a modifikácií PCA pre ne a napokon v tretej kapitole faktorová analýza. Druhá, praktická časť, popisuje výsledky aplikácie spomínaných metód. V štvrtej kapitole sa venujeme PCA v základnej podobe ako aj v iných podobách navrhnutých pre kompozičné dáta a ich porovnaniu na základe výsledkov ich použitia na volebné dáta. V poslednej kapitole sú následne popísané výsledky faktorovej analýzy aplikovanej na volebné dáta a ich možná interpretácia.

1 Analýza hlavných komponentov

V tejto kapitole sa venujeme často používanej metóde na znižovanie dimenzie problému, analýze hlavných komponentov (ďalej iba PCA, z angl. Principal Component Analysis), popisu jej fungovania a použitia. Kapitola je spracovaná podľa [8] a [10].

V štatistike sa často stretávame s viacrozmernými problémami. Žiaľ, schopnosť bežného človeka predstaviť si tieto problémy je spravidla ohraničená tromi rozmermi, čo viedlo k vytvoreniu rôznych metód určených na zredukovanie dimenzie priestoru dát.

Ak pracujeme s niekoľkými (navzájom korelovanými) premennými, môže byť veľmi užitočné nahradiť ich inými, medzi sebou nekorelovanými, ktoré však budú opisovať čo najväčšie množstvo informácie obsiahnutej v pôvodných premenných. Na tejto myšlienke je založená aj PCA. Zameriava sa predovšetkým na varianciu premenných, berie však do úvahy aj kovariancie a korelácie medzi premennými.

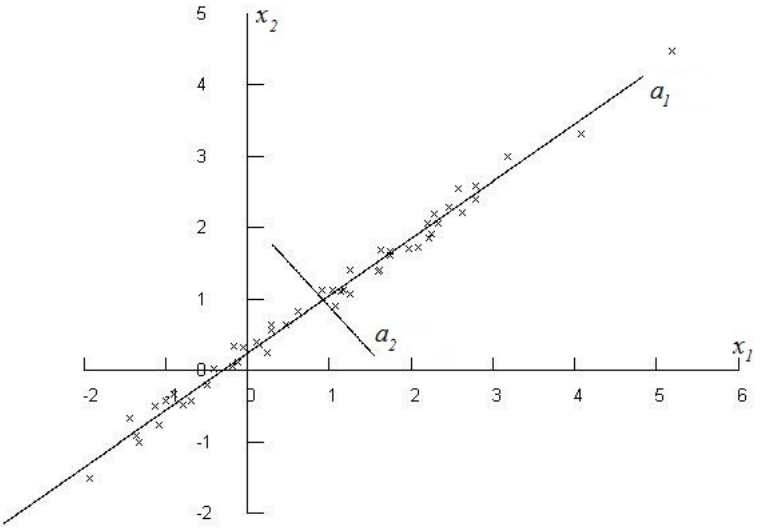
1.1 Definícia a odvodenie hlavných komponentov

Definícia 1.1. *Nech* $\mathbf{x}$ *je p-rozmerný náhodný vektor so strednou hodnotou* μ. *Nech* α_i, $i = 1, ..., p$, *sú také, že* $\alpha_i^T \alpha_i = 1$, $\alpha_i^T(\mathbf{x} - \mu)$ *je nekorelovaná so všetkými* $\alpha_j^T(\mathbf{x} - \mu)$ *pre* $j < i$, *a platí* $\mathrm{var}(\alpha_1^T(\mathbf{x} - \mu)) \geq \mathrm{var}(\alpha_2^T(\mathbf{x} - \mu)) \geq ... \geq \mathrm{var}(\alpha_p^T(\mathbf{x} - \mu))$. *Potom odvodenú náhodnú premennú* $\alpha_i^T(\mathbf{x} - \mu)$ *nazývame i-tym hlavným komponentom.*

Hlavné komponenty teda predstavujú lineárnu kombináciu pôvodných premenných očistených o strednú hodnotu μ. Prechod k novým premenným nám uľahčí analýzu dát, bez straty veľkej časti informácie. Ako mieru informácie obsiahnutej v pôvodných premenných a hlavných komponentoch používame ich varianciu. V ideálnom prípade je už v prvých k komponentoch obsiahnutá väčšina variancie pôvodných premenných, pričom k je oveľa menšie ako p.

Podotýkame, že v definícii hlavných komponentov nie je presná zhoda. Niektorí autori ako hlavné komponenty označujú vektory váh α_i z vyššie uvedenej definície. Častejšie sa však používa definícia hlavných komponentov ako kombinácie pôvodných premenných.

Hlavnú myšlienku fungovania PCA je najnázornejšie vysvetliť pomocou dvoch premenných. V takom prípade je ich analýza pomerne jednoduchá aj bez redukcie dimen-



Obr. 1: Graf pre 50 pozorovaní premenných x_1 a x_2

zie, preto sa v praxi v takomto prípade PCA nevyužíva, veľkou výhodou tohto prípadu je ale možnosť grafického znázornenia. Obr.1 znázorňuje 50 meraní dvoch premenných. Už na prvý pohľad je jasné, že sú navzájom silne korelované a väčšiu varianciu má premenná x_1. Výsledkom PCA sú osi a_1 a a_2. Je zjavné, že v smere a_1 je variancia oveľa väčšia ako v smere a_2. V skutočnosti predstavuje takmer 98 percent celkovej variancie obsiahnutej v dátach.

Podľa definície pri hľadaní hlavných komponentov začíname hľadaním lineárnej kombinácie pôvodných premenných $\alpha_1^{\mathrm{T}}(\mathbf{x} - \mu) = \alpha_{11}(\mathrm{x}_1 - \mu_1) + ... + \alpha_{1\mathrm{p}}(\mathrm{x}_\mathrm{p} - \mu_\mathrm{p})$ takej, ktorá má najvyššiu možnú varianciu a zároveň spĺňa podmienku $\alpha_1^{\mathrm{T}}\alpha_1 = 1$. Nájdená lineárna funkcia je potom prvým hlavným komponentom. Druhý komponent hľadáme podobným spôsobom, pribudne však podmienka na nekorelovanosť s tým prvým. Hľadáme tak k hlavných komponentov, pričom na určenie čísla k si môžeme zvoliť rôzne kritériá. Niektoré z nich spomenieme na konci tejto kapitoly.

Z teórie pravdepodobnosti a štatistiky vieme, že ak $\mathbf{x}$ je náhodný vektor, potom variancia náhodnej premennej $\mathbf{a}^{\mathrm{T}}\mathbf{x}$ sa dá vyjadriť ako $\mathbf{a}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{a}$, kde $\mathbf{\Sigma}$ je kovariančná matica vektora $\mathbf{x}$. Platí tiež, že posunutie o konštantu nemení variančnú maticu náhodného vektora (resp. varianciu náhodnej premennej). Hľadanie prvého hlavného komponentu

preto vedie k úlohe na viazaný extrém v tvare

$$\begin{aligned} \max var(\alpha_1^{\mathrm{T}}(\mathbf{x}-\mu)) &= \alpha_1^{\mathrm{T}}\mathbf{\Sigma}\alpha_1 \\ \alpha_1^{\mathrm{T}}\alpha_1 &= 1 \end{aligned} \tag{1}$$

Riešením tejto úlohy klasickou metódou Lagrangeových multiplikátorov získame úlohu

$$\max \alpha_1^{\mathrm{T}}\mathbf{\Sigma}\alpha_1 - \lambda(\alpha_1^{\mathrm{T}}\alpha_1 - 1) \tag{2}$$

Po zderivovaní podľa α_1 riešime rovnicu

$$2\mathbf{\Sigma}\alpha_1 - 2\lambda\alpha_1 = 0 \tag{3}$$

alebo

$$(\mathbf{\Sigma} - \lambda\mathbf{I})\alpha_1 = 0 \tag{4}$$

čo je definičná rovnica vlastnej hodnoty a vlastného vektora kovariančnej matice $\mathbf{\Sigma}$. Vďaka ohraničeniu $\alpha_1^{\mathrm{T}}\alpha_1 = 1$ jedna z dvojíc tvorených vlastnou hodnotou a príslušným vlastným vektorom bude maximalizovať Lagrangeovu funkciu, otázkou ostáva, o ktorú vlastnú hodnotu a ktorý vlastný vektor sa jedná. Maximalizovaná veličina $\alpha_1^{\mathrm{T}}\mathbf{\Sigma}\alpha_1$ sa ale dá prepísať nasledovným spôsobom: $\alpha_1^{\mathrm{T}}\mathbf{\Sigma}\alpha_1 = \alpha_1^{\mathrm{T}}\lambda\alpha_1 = \lambda\alpha_1^{\mathrm{T}}\alpha_1 = \lambda$, a teda hľadaným vektorom váh pre prvý hlavný komponent je vlastný vektor kovariančnej matice prislúchajúci najväčšej vlastnej hodnote λ_1.

Pri hľadaní druhého hlavného komponentu riešime podobnú úlohu na viazaný extrém, obsahujúcu však podmienku navyše:

$$\begin{aligned} \max \alpha_2^{\mathrm{T}}\mathbf{\Sigma}\alpha_2 \\ \alpha_2^{\mathrm{T}}\alpha_2 = 1 \\ \mathrm{cov}(\alpha_2^{\mathrm{T}}\mathbf{x}, \alpha_1^{\mathrm{T}}\mathbf{x}) = \alpha_2^{\mathrm{T}}\mathbf{\Sigma}\alpha_1 = 0 \end{aligned} \tag{5}$$

Pretože $\alpha_2^{\mathrm{T}}\mathbf{\Sigma}\alpha_1 = \alpha_2^{\mathrm{T}}\lambda_1\alpha_1 = \lambda_1\alpha_2^{\mathrm{T}}\alpha_1$ a $\lambda_1 > 0$, poslednú podmienku možno prepísať na $\alpha_2^{\mathrm{T}}\alpha_1 = 0$ a metódou Lagrangeových multiplikátorov dostávame:

$$\max \alpha_2^{\mathrm{T}}\mathbf{\Sigma}\alpha_2 - \lambda(\alpha_2^{\mathrm{T}}\alpha_2 - 1) - \theta\alpha_2^{\mathrm{T}}\alpha_1 \tag{6}$$

Derivácia podľa α_2 dáva podmienku v tvare

$$2\mathbf{\Sigma}\alpha_2 - 2\lambda\alpha_2 - \theta\alpha_1 = 0 \tag{7}$$

z čoho po prenásobení α_1^{T} zľava vyplýva

$$2\alpha_1^{\mathrm{T}}\mathbf{\Sigma}\alpha_2 - 2\lambda\alpha_1^{\mathrm{T}}\alpha_2 - \theta\alpha_1^{\mathrm{T}}\alpha_1 = 0 \tag{8}$$

Prvé dva výrazy sú podľa (5) rovné nule a podľa (1) platí $\alpha_1^{\mathrm{T}}\alpha_1 = 1$, preto dostávame $\theta = 0$. Dosadením späť do (7) vidíme, že λ je opäť vlastnou hodnotou matice $\mathbf{\Sigma}$ a α_2 je príslušným vlastným vektorom. Pretože podobne ako pri odvodení prvého hlavného komponentu maximalizujeme veličinu $\alpha_2^{\mathrm{T}}\mathbf{\Sigma}\alpha_2 = \lambda$, musí λ byť najväčšie možné. Ak predpokladáme, že matica $\mathbf{\Sigma}$ má rôzne vlastné hodnoty, musí λ byť druhou najväčšou vlastnou hodnotou λ_2, pretože keby to bola najväčšia vlastná hodnota, platilo by $\alpha_1 = \pm\alpha_2$, čo je v spore s podmienkou $\alpha_2^{\mathrm{T}}\mathbf{\Sigma}\alpha_1 = \lambda\alpha_2^{\mathrm{T}}\alpha_1 = 0$.

Takýmto spôsobom je možné po pridaní ďalších podmienok na nekorelovanosť odvodiť aj ostatné vektory váh pre hlavné komponenty α_{k} pre $k = 3, ..., p$. Na základe tohto odvodenia možno sformulovať nasledujúce tvrdenie:

Tvrdenie 1.1. *Nech* $\mathbf{x}$ *je p-rozmerný náhodný vektor so strednou hodnotou* μ *a kovariančnou maticou* $\mathbf{\Sigma}$*. Nech* $\lambda_1 > \lambda_2 > ... > \lambda_p > 0$ *sú vlastné hodnoty matice* $\mathbf{\Sigma}$*. Potom i-tym hlavným komponentom vektora* $\mathbf{x}$ *je náhodná premenná* $y_i = \alpha_{\mathrm{i}}^{\mathrm{T}}(\mathbf{x} - \mu)$*, kde* α_{i} *je vlastný vektor kovariančnej matice* $\mathbf{\Sigma}$ *prislúchajúci vlastnej hodnote* λ_i*. Platí tiež, že* $var(y_i) = \lambda_i$ *a i-ty hlavný komponent tak obsahuje* $\frac{\lambda_i}{\sum_{j=1}^{p}\lambda_j}$ *celkovej variancie.*

Poznámka: Tvrdenie je vlastne dôsledkom spektrálnej vety pre kladne definitnú (prípadne kladne semidefinitnú) symetrickú maticu $\mathbf{\Sigma}$. Vektory α_{i} predstavujú kritické body pre Rayleighov podiel.

1.2 Špeciálne prípady

V predchádzajúcej časti sme predpokladali, že vlastné hodnoty kovariančnej matice $\mathbf{\Sigma}$ sú navzájom rôzne a všetky kladné. Tento predpoklad nemusí byť vždy splnený, prípad rovnakých alebo nulových vlastných hodnôt (kovariančná matica je kladne semidefinitná, preto prípad záporných vlastných hodnôt nemôže nastať vôbec) však nastáva iba veľmi zriedka.

Ak má matica $\mathbf{\Sigma}$ rovnaké vlastné hodnoty, nie je možné jednoznačne určiť hlavné komponenty, pretože možnosť, ako určiť vlastné vektory prislúchajúce takejto viacnásobnej vlastnej hodnote, nie je len jedna.

V prípade nulovej vlastnej hodnoty matice $\mathbf{\Sigma}$ máme hlavný komponent, ktorého variancia je nulová. To hovorí o konštantnom vzťahu medzi niektorými premennými, a teda že niektorá premennná je nepotrebná, pretože jej hodnotu je možné presne dorátať z hodnôt ostatných premenných. Preto je možné niektorú premennú z ďalšej práce s dátami úplne vypustiť, bez akejkoľvek straty informácie.

1.3 PCA a singulárny rozklad

V praxi väčšinou strednú hodnotu a kovariančnú maticu vektora $\mathbf{x}$ nepoznáme. V takom prípade používame ich odhady pomocou nameraných dát. Majme teda n meraní premenných $x_1, ..., x_p$ uložených v matici $\mathbf{X}$ po riadkoch (teda matica $\mathbf{X}$ je rozmeru $n \times p$, každý stĺpec predstavuje jednu premennú). Potom odhadom strednej hodnoty je aritmetický priemer meraní, teda vektor $\bar{\mathbf{X}} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathbf{X}_{\mathrm{i}}$, kde $\mathbf{X}_{\mathrm{i}}$ je i-ty riadok matice $\mathbf{X}$. Ako odhad matice $\mathbf{\Sigma}$ použijeme výberovú kovariančnú maticu $\mathbf{S} = \frac{1}{\mathrm{n}-1}\mathbf{X}_{\mathbf{c}}^{\mathrm{T}}\mathbf{X}_{\mathbf{c}}$, kde $\mathbf{X}_{\mathbf{c}}$ je matica pozorovaní očistených od priemeru, teda od každého riadku odpočítame vektor $\bar{\mathbf{X}}$. Ako vektory váh pre hlavné komponenty potom použijeme vlastné vektory matice $\mathbf{S}$, prislúchajúce vlastným hodnotám zoradeným podľa veľkosti.

V skutočnosti na výpočet hlavných komponentov nie je potrebné odhadovať kovariančnú maticu a počítať jej vlastné hodnoty. Rovnako dobre nám poslúži aj singulárny rozklad matice pozorovaní $\mathbf{X}$, o ktorom hovorí nasledujúce tvrdenie.

Tvrdenie 1.2. *Každá $n \times p$ matica $\mathbf{X}$ sa dá napísať v tvare $\mathbf{X} = \mathbf{U}\mathbf{L}\mathbf{V}^{\mathrm{T}}$, kde $\mathbf{U}$ je rozmeru $n \times n$, $\mathbf{V}$ je rozmeru $p \times p$, obe ortogonálne, a $\mathbf{L}$ je diagonálna matica rozmeru $n \times p$. Navyše, stĺpce matice $\mathbf{U}$ tvoria vlastné vektory matice $\mathbf{X}\mathbf{X}^{\mathrm{T}}$, stĺpce matice $\mathbf{V}$ vlastné vektory matice $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ a na diagonále matice $\mathbf{L}$ sú odmocniny z vlastných čísel matice $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ zoradené od najväčšej, ktoré nazývame singulárnymi hodnotami matice $\mathbf{X}$.*

Dôkaz tohto tvrdenia je možné nájsť vo veľkom množstve literatúry, uvádza ho napríklad Strang v [15].

Vlastná hodnota λ a vlastný vektor α matice $\mathbf{S}$ spĺňajú $(\mathbf{S} - \lambda\mathbf{I})\alpha = 0$. Ako už bolo spomenuté, $\mathbf{S} = \frac{1}{\mathrm{n}-1}\mathbf{X}_{\mathbf{c}}^{\mathrm{T}}\mathbf{X}_{\mathbf{c}}$. Z toho ľahko vidieť, že vlastné vektory matice $\mathbf{S}$ a vektory tvoriace maticu $\mathbf{V}$ singulárneho rozkladu $\mathbf{X}_{\mathbf{c}}$, teda vlastné vektory matice $\mathbf{X}_{\mathbf{c}}^{\mathrm{T}}\mathbf{X}_{\mathbf{c}}$, sú zhodné. To znamená, že vektory váh hlavných komponentov tvoria stĺpce matice $\mathbf{V}$.

Platí tiež, že i-ta vlastná hodnota matice $\mathbf{X_c^T X_c}$ je $(n-1)$-násobkom i-tej vlastnej hodnoty matice $\mathbf{S}$. Preto ako mieru variancie obsiahnutej v i-tom hlavnom komponente môžeme použiť druhú mocninu i-tej singulárnej hodnoty matice $\mathbf{X_c}$ v pomere k súčtu druhých mocnín všetkých singulárnych hodnôt. Navyše, ak rovnosť $\mathbf{X_c} = \mathbf{ULV^T}$ prenásobíme sprava ortogonálnou maticou $\mathbf{V}$, získame $\mathbf{X_c V} = \mathbf{UL}$. Pretože stĺpce $\mathbf{V}$, ako sme už spomenuli, predstavujú váhy hlavných komponentov, zložky matice $\mathbf{UL}$ predstavujú po riadkoch ich hodnoty pre jednotlivé merania.

Singulárny rozklad je tiež možné použiť na aproximáciu matice. Singulárny rozklad matice $\mathbf{X_c}$ v tvare $\mathbf{X_c} = \mathbf{ULV^T}$ možno prepísať nasledujúcim spôsobom:

$$\mathbf{X_c} = \mathbf{u}_1 l_1 \mathbf{v}_1^{\mathrm{T}} + \mathbf{u}_2 l_2 \mathbf{v}_2^{\mathrm{T}} + ... + \mathbf{u}_{\mathrm{r}} l_{\mathrm{r}} \mathbf{v}_{\mathrm{r}}^{\mathrm{T}} \tag{9}$$

kde r je hodnosť matice $\mathbf{X_c}$, $\mathbf{u}_{\mathrm{i}}$ predstavuje i-ty stĺpec matice $\mathbf{U}$, $\mathbf{v}_{\mathrm{i}}$ i-ty stĺpec matice $\mathbf{V}$ a l_i i-tu singulárnu hodnotu. V prípade, že posledných niekoľko singulárnych hodnôt je veľmi malých, predovšetkým v porovnaní s prvými, nemajú už na výslednú maticu veľký vplyv, preto je možné ich zanedbať a získať tak aproximáciu matice $\mathbf{X_c}$. Potom ak $\tilde{\mathbf{X}}_\mathbf{c}$ označíme aproximáciu matice $\mathbf{X_c}$ s použitím prvých k singulárnych hodnôt a vektorov, chybu tejto aproximácie môžeme merať výrazom $\left\|\tilde{\mathbf{X}}_\mathbf{c} - \mathbf{X_c}\right\|^2$, pričom $\|.\|$ tu predstavuje Frobeniovu normu matice. Dá sa jednoducho ukázať, že táto chyba je rovná $l_{\mathrm{k+1}}^2 + l_{\mathrm{k+2}}^2 + ... + l_{\mathrm{r}}^2 = \lambda_{\mathrm{k+1}} + \lambda_{\mathrm{k+2}} + ... + \lambda_{\mathrm{r}}$, čo je nevysvetlená variancia v PCA pri použití k hlavných komponentov. Ďalšou možnosťou je posudzovať absolútnu alebo relatívnu odchýlku prvkov $\tilde{\mathbf{X}}_\mathbf{c}$ oproti $\mathbf{X_c}$.

1.4 Kovariančná a korelačná matica

Doteraz sme hovorili o hlavných komponentoch odvodených z kovariančnej matice. To je v poriadku, ak pracujeme s premennými v rovnakých jednotkách. Ak sú ale premenné rôznych typov (napríklad váha, teplota a výška), nastáva problém, pretože variancie a kovariancie nie sú invariantné na zmenu jednotiek. Jednoduchým prechodom k iným jednotkám v jednej premennej, napríklad ak sa rozhodneme výšku merať v centimetroch namiesto metrov, sa zmenia hlavné komponenty, keďže dôsledkom maximalizácie variancie je uprednostňovanie premenných s najväčšou varianciou. Preto sa v praxi často pracuje so štandardizovanými premennými, teda sa od nich odráta stredná hod-

nota (priemer) a predelia sa varianciou. Vektor takto upravených premenných označíme $\mathbf{x}^*$.

Hlavné komponenty pre vektor $\mathbf{x}^*$ získame ako skalárny súčin vlastných vektorov jeho kovariančnej matice $\mathbf{\Sigma}^*$ so samotným vektorom $\mathbf{x}^*$. V skutočnosti matica $\mathbf{\Sigma}^*$ je korelačnou maticou vektora $\mathbf{x}$. Hoci prechod od $\mathbf{x}$ k $\mathbf{x}^*$ je veľmi jednoduchou transformáciou, medzi hlavnými komponentmi získanými z kovariančej a korelačnej matice neexistuje žiadny jednoduchý vzťah. Dôvodom je, že hlavné komponenty sú invariantné na ortogonálne transformácie vektora $\mathbf{x}$, jeho štandardizácia ale nie je ortogonálnou transformáciou.

Ďalším dôvodom uprednostňovania korelačnej matice pri meraniach rôznych typov premenných je interpretácia výsledkov pri použití kovariančnej matice. Pripočítanie teploty k váhe nie je zmysluplným krokom. Štandardizáciou však prejdeme k bezrozmerným premenným, čo tento problém odstraňuje. V prípade merania premenných v rovnakých jednotkách však použitie kovariančnej matice môže mať informatívnejší charakter. Štandardizáciou premenných istú časť informácie z dát odstránime, v niektorých prípadoch je táto informácia ale dôležitá.

Záverom teda je, že rozhodnutie o použití kovariančnej alebo korelačnej matice nie je presne dané. Vždy je dôležité posúdiť situáciu, v ktorej ideme PCA robiť, a na základe charakteru dát aj požadovaných výsledkov vybrať jednu zo spomínaných matíc. Aj v konkrétnej situácii nemusí byť jasné, ktorá matica je vhodnejšia, rôzni používatelia PCA by mohli mať rôzne argumenty pre použitie jednej alebo druhej z nich.

1.5 Počet hlavných komponentov

Za predpokladu, že kovariačná, resp. korelačná matica vektora $\mathbf{x}$ má p vlastných hodnôt, má tiež p ortogonálnych vlastných vektorov. To znamená, že máme p hlavných komponentov. Hlavným cieľom PCA je ale znížiť dimenziu problému, preto nepoužívame všetky hlavné komponenty. Je teda potrebné nejakým spôsobom určiť, koľko z nich ponechať pre ďalšiu prácu. Na to existuje niekoľko spôsobov:

- veľkosť vlastnej hodnoty - vybrať tie hlavné komponenty, pre ktoré prislúchajúca vlastná hodnota je väčšia ako priemer (pri použití korelačnej matice väčšia ako 1)

- lakťový diagram (scree plot) - graf, ktorý na os x nanáša poradie komponentov a na os y príslušnú vlastnú hodnotu; hľadáme bod, kedy v grafe nastáva určitý zlom, a ten určuje počet komponentov
- percento vysvetlenej variancie - použijeme najmenší počet komponentov potrebný na vysvetlenie požadovaného percenta variancie (zvyčajne je to 80-90 %)

2 Kompozičné dáta a ich analýza

Pri niektorých dátach je podstatnou informáciou, aký podiel z celku predstavujú jednotlivé premenné. Takéto dáta, pri ktorých sú údaje nezáporné a súčet premenných je rovný konštante, napríklad 100% alebo 1 pre podiely, sa nazývajú kompozičné dáta. Z ohraničenia na súčet premenných vyplývajú špeciálne vlastnosti kompozičných dát a preto nie je vhodné pri ich analýze používať klasické multidimenzionálne metódy ako PCA. Ako jeden z prvých sa tejto problematike veľmi obšírne začal venovať J. Aitchinson (napr. [3]), počas posledných tridsiatich rokov sa výskum v tejto oblasti veľmi posunul a vyšlo niekoľko ďalších článkov a kníh ([2, 4, 6, 7, 13]).

2.1 Kompozícia, simplex ako priestor

Definícia 2.1. *Kompozícia* $\mathbf{x}$ *s* D *časťami je* D*-rozmerný vektor s nezápornými zložkami* $x_i, i = 1, ..., D$ *spĺňajúcimi* $\sum_{i=1}^{D} x_i = 1$.

Podmienka na súčet zložiek vektora je omnoho obmedzujúcejšia ako nezápornosť zložiek. Zaručuje však, že na úplné popísanie kompozície stačí jej $d = (D - 1)$ zložiek. Napríklad ak poznáme d zložiek vektora $\mathbf{x}$, poslednú je možné dopočítať z obmedzenia na súčet zložiek. Rovnako je možné kompozíciu $\mathbf{x}$ popísať vektorom podielov prvých d zložiek vektora $\mathbf{x}$ k poslednej. Ak označíme

$$r_i = \frac{x_i}{x_D}, i = 1, ..., d, \tag{10}$$

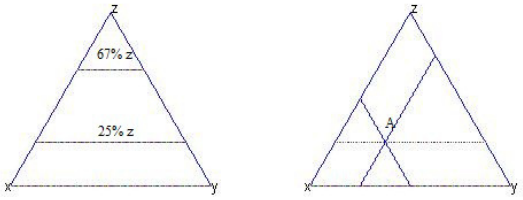
potom zložky kompozície získame nasledovne:

$$\begin{aligned} x_i &= \frac{r_i}{r_1 + ... + r_d + 1} \\ x_D &= \frac{1}{r_1 + ... + r_d + 1} \end{aligned} \tag{11}$$

Pri budovaní teórie kompozičných dát podmienku nezápornosti zložiek pre jednoduchosť ešte zosilníme a budeme predpokladať, že zložky vektora $\mathbf{x}$ sú kladné. Neskôr sa zmienime aj o dátach obsahujúcich nuly.

Pri analýze kompozičných dát hrá veľkú úlohu priestor, z ktorého pochádzajú, preto sa v tejto časti venujeme jeho opisu, nasledujúc [3] a [4].

Definícia 2.2. *Množinu* $S^d = \{(x_1, ..., x_D)^T; x_i > 0, i = 1, ..., D, \sum_{i=1}^{D} x_i = 1\}$ *nazývame d-rozmerným simplexom.*



Obr. 2: Trojuholníkový diagram pre kompozíciu $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$

Každý vektor $\mathbf{w}$ z D-rozmerného reálneho vektorového priestoru, ktorý má kladné zložky, možno jednoznačne previesť na kompozíciu a tak ho istým spôsobom sprojektovať na d-rozmerný simplex. Takýto vektor $\mathbf{w}$ voláme bázou kompozície $\mathbf{x}$ a okrem kompozície má tiež jednoznačne určenú veľkosť, $t = \sum_{i=1}^{D} w_i$. Naopak, ak máme k dispozícii kompozíciu $\mathbf{x}$ a veľkosť bázy t, dá sa spätne jednoznačne určiť báza $\mathbf{w}$.

V prípade $D = 3$ možno kompozičné dáta zakresliť do tzv. trojuholníkového diagramu, ktorého vrcholy predstavujú kompozície $(1, 0, 0), (0, 1, 0)$ a $(0, 0, 1)$. Každý bod v trojuholníku jednoznačne určuje podiel jednotlivých častí kompozície, ako to vysvetľuje Obr.1. Bod A v jeho pravej časti označuje kompozíciu $(x, y, z) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$, teda 50% bodu A tvorí prvok x.

V súvislosti so simplexom hovoríme o tzv. Aitchinsonovej geometrii, ktorá je založená na základných operáciách uzáveru, perturbácie a mocninnej transformácie.

Definícia 2.3. *Nech* $\mathbf{z}$ *je D-rozmerný vektor s kladnými zložkami,* $\mathbf{x}, \mathbf{y}$ *sú kompozície s D časťami a* α *je reálne číslo. Potom uzáverom vektora* $\mathbf{z}$ *je kompozícia* $C(\mathbf{z}) = \frac{\mathbf{z}}{\sum_{\mathrm{i}=1}^{\mathrm{D}} \mathrm{z_i}}$, *perturbáciou kompozícií* $\mathbf{x}, \mathbf{y}$ *je* $\mathbf{x} \oplus \mathbf{y} = \mathrm{C}(\mathrm{x_1 y_1}, ..., \mathrm{x_D y_D})^{\mathrm{T}}$ *a mocninnou transformáciou kompozície* $\mathbf{x}$ *je kompozícia* $\alpha \odot \mathbf{x} = \mathrm{C}(\mathrm{x_1^{\alpha}}, ..., \mathrm{x_D^{\alpha}})^{\mathrm{T}}$.

Poznámka: Operácia perturbácie je analógiou súčtu v reálnom vektorovom priestore, podobne mocninná transformácia kompozície predstavuje analógiu k násobeniu reálneho vektora skalárom.

2.2 Príprava kompozičných dát na analýzu

Z vlastností kompozičných dát plynie nevhodnosť použitia klasických štatistických metód, vedúcich v tomto prípade často k nesprávnym výsledkom. Je však možné dáta transformovať do reálneho vektorového priestoru a tam na ne použiť klasické štatis-

tické metódy. Z (10) a (11) vieme, že relevantná informácia o kompozícii je ukrytá aj v podieloch medzi jej prvkami. Tento fakt a to, že matematicky sa pracuje jednoduchšie s logaritmickými podielmi ako s obyčajnými, viedli Aitchinsona v [3] k predstaveniu transformácií kompozičných dát založených na logaritmických podieloch alr (aditive logratio - aditívne logaritmické podiely) a clr (centered logratio - centrované logaritmické podiely), ktoré umožnujú prechod od simplexu k reálnemu vektorovému priestoru:

$$alr(\mathbf{x}) = (\ln\frac{\mathrm{x_1}}{\mathrm{x_D}}, ..., \ln\frac{\mathrm{x_d}}{\mathrm{x_D}})^{\mathrm{T}} \tag{12}$$

$$\mathbf{u} = clr(\mathbf{x}) = (\ln\frac{\mathrm{x_1}}{\mathrm{g}(\mathbf{x})}, ..., \ln\frac{\mathrm{x_D}}{\mathrm{g}(\mathbf{x})})^{\mathrm{T}} \tag{13}$$

pričom menovateľ v (13) predstavuje geometrický priemer častí kompozície, teda $g(\mathbf{x}) = \sqrt[D]{\prod_{i=1}^{D} x_i}$.

Takýto prechod zo simplexu S^d do reálneho vektorového priestoru $\mathbf{R}^D$ je jednoznačný, čo zabezpečuje možnosť vrátiť sa späť do simplexu a tam interpretovať výsledky analýz. Návrat do simplexu predstavuje inverzná clr transformácia

$$clr^{-1}(\mathbf{u}) = C(e^{u_1}, ..., e^{u_D}) \tag{14}$$

Pretože prirodzene požadovanou vlastnosťou, ktorú od analýzy takýchto dát očakávame, je invariantnosť na permutácie (zmena poradia prvkov v kompozícii by nemala zmeniť výsledok analýzy vzhľadom na dané prvky), alr nie je vhodnou transformáciou, keďže jej výsledok závisí od menovateľa, ktorý sa permutáciou prvkov môže meniť. Tento problém rieši clr, kde menovateľ závisí rovnako od všetkých prvkov, bez ohľadu na ich poradie. Pre potreby ďalších metód, najmä robustných, bola v [6] predstavená aj ďalšia transformácia, ilr (isometric logratio - izometrické logaritmické podiely). Táto však nie je predmetom tejto bakalárskej práce a ďalej sa jej nebudeme venovať.

2.3 PCA pre kompozičné dáta

Analýza kompozičných dát predstavuje oblasť, v ktorej sa stále napreduje a predovšetkým štatistici navrhujú ďalšie metódy. Podnetom na rozvoj rôznych metód na analýzu

tohto špeciálneho druhu dát boli problémy nastávajúce pri používaní klasických metód na matice pozorovaní upravené operáciou uzáveru.

V prípade PCA na uzavretých dátach hovoríme o tzv. hrubej PCA (z anglického crude PCA). Ohraničenie na súčet prvkov kompozície, teda konštantný vzťah medzi premennými, sa prejaví v podobe nulovej vlastnej hodnoty výberovej kovariančnej matice, resp. nulovej singulárnej hodnoty matice uzavretých dát, ako sme už spomenuli v predchádzajúcej kapitole. Práve fakt, že pri kompozičných dátach je výberová kovariančná matica vždy singulárna, viedol k názoru, že klasické multidimenzionálne metódy na analýzu kompozičných dát nie sú vhodné.

Ako prvý priniesol novú metódu Aitchinson. Tzv. logcontrast PCA založenú na clr transformácii popisuje napríklad v [3]. V klasickej PCA hľadáme vektor α taký, že $\alpha^{\mathrm{T}}\alpha = \mathbf{1}$, ktorý maximalizuje varianciu odvodenej premennej $\alpha^{\mathrm{T}}\mathbf{x}$. Dá sa preto očakávať, že aj v tomto prípade budeme hľadať nejakú lineárnu kombináciu premenných, ktoré máme k dispozícii, teda premenných po clr transformácii. Názov metódy je potom odvodený práve od tohto výsledku.

Definícia 2.4. *Logcontrast kompozície $\boldsymbol{x}$ je akákoľvek loglineárna transformácia $\boldsymbol{a}^T ln\, \boldsymbol{x}$ spĺňajúca $\sum_{i=1}^{D} a_i = 0$. Ak navyše platí $\boldsymbol{a}^T \boldsymbol{a} = 1$ a $a_1 > 0$, hovoríme o štandardnom logcontraste.*

Podmienka $\sum_{i=1}^{D} a_i = 0$ zabezpečuje, že platí $\boldsymbol{a}^{\mathrm{T}} \ln \mathbf{x} = \boldsymbol{a}^{\mathrm{T}} \ln \frac{\mathbf{x}}{\mathrm{g}(\mathbf{x})}$, teda že logcontrast kompozície $\mathbf{x}$ je zároveň lineárnou kombináciou zložiek jej clr transformácie.

V klasickej PCA sú jednotlivé hlavné komponenty medzi sebou ortogonálne. Niečo podobné by sme očakávali aj v analýze kompozičných dát, preto pri logcontrastoch zavádza Aitchinson tiež pojem ortogonality.

Definícia 2.5. *Dva logcontrasty $\boldsymbol{a}^T ln\, \mathbf{x}$ a $\boldsymbol{b}^T ln\, \mathbf{x}$ sú ortogonálne, ak platí $\boldsymbol{a}^T \boldsymbol{b} = 0$.*

Označme $\boldsymbol{\Gamma}$ kovariančnú maticu vektora $\boldsymbol{u} = clr(\mathbf{x})$. Aitchinson ju nazýva tiež centrovanou kovariančnou maticou logaritmických podielov. Potom pre vektor $\boldsymbol{a}$ platí $var(\boldsymbol{a}^T \boldsymbol{u}) = var(\boldsymbol{a}^T ln\, \mathbf{x}) = \boldsymbol{a}^T \boldsymbol{\Gamma} \boldsymbol{a}$. Z vlastností clr transformácie tiež vyplýva zaujímavá vlastnosť matice $\boldsymbol{\Gamma}$.

Tvrdenie 2.1. *Nech $\boldsymbol{\Gamma}$= $[\gamma_{ij}]$ je kovariančná matica vektora $\boldsymbol{u} = clr(\boldsymbol{x})$, teda $\gamma_{ij} = cov(u_i, u_j)$. Potom ak označíme $\boldsymbol{j}$ vektor samých jednotiek, platí $\boldsymbol{\Gamma j} = \boldsymbol{0}$.*

Dôkaz tohto tvrdenia vyplýva priamo z definície prvkov matice $\mathbf{\Gamma}$, nulového súčtu prvkov vektora $\boldsymbol{u}$ a z toho, že $cov(\mathbf{x}, \mathbf{y} + \mathbf{z}) = cov(\mathbf{x}, \mathbf{y}) + cov(\mathbf{x}, \mathbf{z})$.

Podobne ako pri klasickej PCA hľadáme taký vektor a_1, ktorý maximalizuje varianciu príslušného logcontrastu. Budeme tiež požadovať, aby bol tento logcontrast štandardný, čo zabezpečí jednoznačnosť (až na znamienko). Budeme teda riešiť úlohu:

$$\begin{aligned} \max var(\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{u}) &= \boldsymbol{a}_1^{\mathrm{T}}\mathbf{\Gamma}\boldsymbol{a}_1 \\ \boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{a}_1 &= 1 \\ \boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{j} &= 0 \end{aligned} \tag{15}$$

kde $\boldsymbol{j}$ je vektor samých jednotiek. Použitím Lagrangeových multiplikátorov riešime úlohu na voľný extrém

$$\max \boldsymbol{a}_1^{\mathrm{T}}\mathbf{\Gamma}\boldsymbol{a}_1 - \lambda(\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{a}_1 - 1) - 2\mu\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{j} \tag{16}$$

Pretože tretia podmienka v (15) je nulová, môžeme ju prenásobiť číslom 2, čo neskôr uľahčí prácu. Zderivovaním (16) podľa $\boldsymbol{a}, \lambda, \mu$ dostávame podmienky prvého rádu:

$$2\mathbf{\Gamma}\boldsymbol{a_1} - \mathbf{2}\lambda\boldsymbol{a_1} - \mathbf{2}\mu\boldsymbol{j} = \mathbf{0} \tag{17}$$

$$\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{a}_1 = 1 \tag{18}$$

$$\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{j} = 0 \tag{19}$$

Vydelením podmienky (17) číslom 2 a jej prenásobením zľava postupne vektormi $\boldsymbol{a}_1^{\mathrm{T}}$ a $\boldsymbol{j}$ dostaneme

$$\boldsymbol{a}_1^{\mathrm{T}}\mathbf{\Gamma}\boldsymbol{a}_1 - \lambda\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{a}_1 - \mu\boldsymbol{a}_1^{\mathrm{T}}\mathbf{j} = \boldsymbol{a}_1^{\mathrm{T}}\mathbf{\Gamma}\boldsymbol{a}_1 - \lambda = 0 \tag{20}$$

$$\boldsymbol{j}^{\mathrm{T}}\mathbf{\Gamma}\boldsymbol{a}_1 - \lambda\boldsymbol{j}^{\mathrm{T}}\boldsymbol{a}_1 - \mu\boldsymbol{j}^{\mathrm{T}}\boldsymbol{j} = \boldsymbol{j}^{\mathrm{T}}\mathbf{\Gamma}\boldsymbol{a}_1 - \mathrm{D}\mu = 0 \tag{21}$$

Po dosadení (21) do (17) máme:

$$\mathbf{\Gamma}\boldsymbol{a}_1 - \lambda\boldsymbol{a}_1 - \mathrm{D}^{-1}\boldsymbol{j}\boldsymbol{j}^{\mathrm{T}}\mathbf{\Gamma}\boldsymbol{a}_1 = ((\mathbf{I} - \mathrm{D}^{-1}\mathbf{J})\mathbf{\Gamma} - \lambda\mathbf{I})\boldsymbol{a}_1 = (\mathbf{\Gamma} - \lambda\mathbf{I})\boldsymbol{a}_1 = 0 \tag{22}$$

kde $\mathbf{J}$ je matica samých jednotiek. Posledná rovnosť vyplýva z vlastnosti $\mathbf{\Gamma}\boldsymbol{j} = \mathbf{0}$. Výsledná rovnica je opäť definičnou rovnicou pre vlastnú hodnotu a vlastný vektor matice $\mathbf{\Gamma}$. Z (20) vyplýva, že maximalizovanou hodnotou je práve vlastná hodnota λ, preto hľadaným vektorom $\boldsymbol{a}_1$ je vlastný vektor matice $\mathbf{\Gamma}$ prislúchajúci jej najvyššej vlastnej hodnote λ_1. Podobným rozšírením ako v prípade klasickej PCA dostávame nasledovné tvrdenie:

Tvrdenie 2.2. *Nech $\boldsymbol{x}$ je D-prvková kompozícia, $\boldsymbol{u} = clr(\boldsymbol{x})$ a $\mathbf{\Gamma}$ je kovariančná matica vektora $\boldsymbol{u}$. Potom i-tym logcontrast hlavným komponentom je odvodená premenná $\boldsymbol{a}_i^T(ln\ \boldsymbol{x} - E(ln\ \boldsymbol{x}))$, kde $\boldsymbol{a}_i$ je vlastný vektor matice $\mathbf{\Gamma}$ prislúchajúci jej i-tej najväčšej vlastnej hodnote.*

Ukazuje sa, že logcontrast hlavné komponenty popisujú správanie sa pôvodných premenných rovnako dobre alebo lepšie ako výsledky hrubej PCA. V prípade, že v pôvodných dátach je prítomné nejaké zakrivenie, lineárna funkcia, ktorá je výsledkom klasickej PCA, toto zakrivenie popísať nedokáže. Logcontrast hlavné komponenty túto úlohu zvládajú s väčšou presnosťou.

Hoci niektoré skupiny dát poukazujú na to, že najvhodnejším spôsobom na analýzu hlavných komponentov pre kompozičné dáta je clr transformácia a následná PCA na takto transformovaných dátach, prípadne robustná PCA založená na ilr transformácii, v tejto oblasti stále prebieha výskum a väčšina štatistikov odporúča zvážiť charakter dát a na základe toho vybrať vhodný prístup, prípadne skúsiť niekoľko prístupov a rozhodnúť sa na základe toho, ako presne jednotlivé výsledky popisujú charakter dát.

2.4 Výskyt núl v dátach

V doterajšom budovaní teórie kompozičných dát sme brali do úvahy iba vektory s kladnými zložkami. To dáva zmysel predovšetkým pri práci s koncetráciami v prípade, kedy presnosť meraní dovoľuje zaznamenať aj veľmi malé množstvá blízke nule. Tento predpoklad bol dôležitý hlavne z toho dôvodu, že sme narábali s logaritmami či podielmi, matematickými operáciami, pri ktorých nastáva v prípade prítomnosti núl problém. V skutočnosti sa ale v kompozíciách z rôznych dôvodov nuly niekedy nachádzajú. Väčšinou nie je možné s istotou určiť dôvod toho, že daný prvok má v kompozícii hodnotu nula. Môže to znamenať, že v kompozícii nie je vôbec prítomný, ale tiež, že je prítomný

v príliš malom, nemerateľnom množstve, prípadne toto malé množstvo je merateľné, no vplyvom zaokrúhlenia sa v dátach javí ako nula. Jednou z možností, ako sa s problémom núl v kompozícii vyrovnať, je zvážiť predefinovanie prvkov kompozície tak, aby sme sa núl zbavili. Napríklad, ak kompozícia predstavuje výdavky domácnosti, rozdeliť ich iným spôsobom - miesto dvoch skupín výdavkov, zvlášť na oblečenie a zvlášť na obuv, brať do úvahy iba jednu takúto skupinu, ktorá bude v sebe zahŕňať výdavky aj na oblečenie, aj na obuv. Ale niekedy takéto predefinovanie z rôznych dôvodov nie je možné. V takom prípade Aitchinson v [3] navrhuje nahradiť nuly nejakou malou konštantou a zároveň od ostatných prvkov kompozície odrátať inú malú konštantu, aby bol zachovaný súčet prvkov.

Aitchinson zavádza takzvané oblasti možných nezaokrúhlených kompozícií s polomerom δ. Každá kompozícia z takejto oblasti sa zaokrúhli do jej stredu. V prípade, že takáto oblasť má svoj stred na hranici simplexu, výsledkom zaokrúhlenia bude kompozícia s niektorými prvkami nulovými.

Napríklad kompozícia $(0.00, 0.62, 0.38)$ leží na hranici simplexu a môže byť výsledkom zaokrúhlenia akejkoľvek kompozície z polovice pravidelného šesťuholníka s polomerom δ okolo tejto kompozície. Mohli by sme preto túto kompozíciu nahradiť inou - $(0.00+\epsilon, 0.62-\frac{1}{2}\epsilon, 0.38-\frac{1}{2}\epsilon)$. Jednou z možných volieb ϵ je geometrický stred príslušnej oblasti možných nezaokrúhlených kompozícií, v tomto prípade polovice pravidelného šesťuholníka, vynásobený najvyššou možnou chybou spôsobenou zaokrúhlením, čo je pre dáta zaznamenávané na dve desatinné miesta 0,005. Toto číslo predstavuje polomer oblasti možných nezaokrúhlených kompozícií δ. Takýmto uvažovaním je možné prísť k všeobecnému vzorcu na nahrádzanie núl. Ak označíme C počet nulových prvkov v kompozícii, potom pre novú kompozíciu $\mathbf{x}^*$ platí predpis:

$$x_i^* = \begin{cases} \frac{\delta(C-1)(D-C)}{D^2} & \text{ak} \quad x_i = 0 \\ \frac{\delta C(C+1)}{D^2} & \text{ak} \quad x_i \neq 0 \end{cases} \tag{23}$$

Tento spôsob je možné podľa požiadaviek a charakteru dát rôznymi spôsobmi upravovať. Napríklad ak máme dodatočné informácie o veľkosti bázy kompozície pre rôzne merania, môžeme náhradu prispôsobiť tejto veľkosti. Tak isto sa dá uvažovať aj v prípade špeciálnych znalostí o prvkoch kompozície. Príkladom môžu byť výsledky pre-

zidentských volieb 2014. Ak sa objaví v obci alebo okrese 0 na pozícii prislúchajúcej R.Ficovi či A. Kiskovi, táto informácia má iný charakter ako v prípade napríklad J. Šimka. Inú informáciu prináša tiež 0 v malej obci oproti veľkému mestu, kde je 60 tisíc voličov. V tomto prípade nemusí byť oblasť zaokrúhľovania pravidelná, v niektorých smeroch môže byť "jemnejšia".

2.5 Simpliciálny singulárny rozklad

V prvej kapitole sme spomenuli singulárny rozklad ako užitočný nástroj pri PCA. Pripomeňme si, že maticu $\mathbf{X_c}$ rozmeru $n \times p$ je možné rozložiť nasledovným spôsobom: $\mathbf{X_c} = \mathbf{ULV}^{\mathrm{T}} = \mathbf{u}_1 \mathrm{l}_1 \mathbf{v}_1^{\mathrm{T}} + \mathbf{u}_2 \mathrm{l}_2 \mathbf{v}_2^{\mathrm{T}} + ... + \mathbf{u}_{\mathrm{r}} \mathrm{l}_{\mathrm{r}} \mathbf{v}_{\mathrm{r}}^{\mathrm{T}}$. Potom i-ty riadok tejto matice sa dá rozpísať ako

$$\mathbf{X}_{\mathbf{c}\mathrm{i}\cdot} = \mathrm{u}_{\mathrm{i}1} l_1 \mathbf{v}_{\cdot 1}^{\mathrm{T}} + \mathrm{u}_{\mathrm{i}2} l_2 \mathbf{v}_{\cdot 2}^{\mathrm{T}} + ... + \mathrm{u}_{\mathrm{ir}} l_{\mathrm{r}} \mathbf{v}_{\cdot \mathrm{r}}^{\mathrm{T}} \tag{24}$$

kde $\mathbf{v}_{\cdot \mathrm{i}}$ je i-ty stĺpec matice $\mathbf{V}$ a r predstavuje menšie z čísel n a p. Ak teda matica $\mathbf{X_c}$ je maticou $\mathbf{X}$ upravenou o priemer, ako je to uvedené v prvej kapitole, potom i-ty riadok matice $\mathbf{X}$ môžeme pomocou singulárneho rozkladu matice $\mathbf{X_c}$ napísať ako

$$\mathbf{X}_{\mathrm{i}\cdot} = \bar{\mathbf{X}} + \mathrm{u}_{\mathrm{i}1} l_1 \mathbf{v}_{\cdot 1}^{\mathrm{T}} + \mathrm{u}_{\mathrm{i}2} l_2 \mathbf{v}_{\cdot 2}^{\mathrm{T}} + ... + \mathrm{u}_{\mathrm{ir}} l_{\mathrm{r}} \mathbf{v}_{\cdot \mathrm{r}}^{\mathrm{T}} \tag{25}$$

Analogický výsledok na simplexe predstavuje simpliciálny singulárny rozklad, ktorý umožňuje podobným spôsobom pracovať aj s maticami kompozičných dát. Operácie násobenia a sčítania sú tu však nahradené ich simpliciálnymi verziami. Singulárny rozklad potom bude pre i-ty riadok po riadkoch kompozičnej matice $\mathbf{X}$ o rozmeroch $n \times D$ vyzerať nasledovne:

$$\mathbf{X}_{\mathrm{i}\cdot} = \hat{\epsilon} \oplus (\mathrm{u}_{\mathrm{i}1} l_1 \odot \beta_1) \oplus (\mathrm{u}_{\mathrm{i}2} l_2 \odot \beta_2) \oplus ... \oplus (\mathrm{u}_{\mathrm{ir}} l_{\mathrm{r}} \odot \beta_{\mathrm{r}}) \tag{26}$$

kde $\hat{\epsilon}$ je uzavretý vektor geometrických priemerov stĺpcov matice $\mathbf{X}$.

Postup na získanie matíc $\mathbf{U}$ a $\mathbf{L}$ a kompozícií β_{i} uvádza Aitchinson v [1]. Prvým krokom je vytvorenie dvojnásobne centrovanej matice logaritmických podielov $\mathbf{Z} = [\mathrm{z}_{\mathrm{ri}}]$, kde

$$z_{ri} = ln \frac{x_{ri}}{g(\mathbf{X}_{\mathrm{r}})} - \frac{1}{N} \sum_{k=1}^{N} ln \frac{x_{ki}}{g(\mathbf{X}_{\mathrm{k}})} \tag{27}$$

Ak $\mathbf{Z} = \mathbf{ULV}^{\mathrm{T}}$ je singulárny rozklad matice $\mathbf{Z}$, potom matice $\mathbf{U}$ a $\mathbf{L}$ zodpovedajú tým zo simpliciálneho singulárneho rozkladu matice $\mathbf{X}$ a

$$\beta_i = C(e^{v_{1i}}, ..., e^{v_{Di}}) \tag{28}$$

Rovnako ako v reálnom vektorovom priestore, aj tu nám singulárny rozklad umožňuje aproximovať pôvodnú maticu použitím len niekoľkých prvých členov v (26). A tiež, poskytuje nám možnosť získať váhy pre logcontrast hlavné komponenty bez nutnosti odhadu matice $\mathbf{\Gamma}$. Vektorom váh $\mathbf{a}_{\mathrm{i}}$ pre i-ty logcontrast hlavný komponent je $clr(\beta_{\mathrm{i}})$. Tento prístup je v literatúre venujúcej sa kompozičným dátam nazývaný aj "staying-in-the-simplex approach", teda "prístup ostávania v simplexe".

3 Faktorová analýza

V prvej kapitole sme predstavili PCA ako jednu z metód na znižovanie dimenzie viacrozmerných štatistických problémov. Ďalšou často používanou technikou s rovnakým cieľom je faktorová analýza, na ktorú sa sústredíme v tejto kapitole. Pri jej popise vychádzame najmä z [9] a [17].

Faktorová analýza skúma, či sú pozorované premenné lineárne závislé od iných, nepozorovateľných premenných, ktoré sa nazývajú faktory. Pritom počet faktorov by mal byť výrazne menší ako počet pôvodných premenných.

3.1 Model faktorovej analýzy

Nech $X_1, ..., X_p$ sú nami pozorované náhodné premenné so strednými hodnotami $\mu_1, ..., \mu_p$. Ak očakávané faktory označíme $F_1, ..., F_k$, môžeme základnú myšlienku faktorovej analýzy zapísať nasledujúcim modelom:

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \ldots + l_{1k}F_k + e_1 \\ &\vdots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pk}F_k + e_p \end{aligned} \tag{29}$$

kde $l_{ij}, i = 1, ..., p, j = 1, ..., k$ sú váhy jednotlivých faktorov v premenných. Faktory $F_1, ..., F_k$ nazývame spoločné faktory, pretože sú rovnaké pre všetky pozorované premenné. Vektor $\mathbf{e}$ predstavuje vektor chýb, ktoré sa inak nazývajú aj špecifické faktory, keďže pre každú premennú vysvetľujú chýbajúcu informáciu, ktorú spoločné faktory nedokážu vysvetliť. Tento základný model je možné prepísať aj v maticovej forme:

$$\mathbf{X} - \mu = \mathbf{LF} + \mathbf{e} \tag{30}$$

Aby bolo možné ďalej s týmto modelom pracovať, kladieme naň dva základné predpoklady:

(i) Chyby $e_i, i = 1, ..., p$ sú navzájom nezávislé a platí pre ne $E(e_i) = 0$ a $Var(e_i) = \psi_i$.

(ii) Spoločné faktory $F_j, j = 1, ..., k$ sú nezávislé navzájom aj so špecifickými faktormi a platí pre ne $E(F_j) = 0$ a $Var(F_j) = 1$.

Predpoklad nezávislosti špecifických faktorov medzi sebou vlastne hovorí, že spoločné faktory vysvetľujú všetku možnú informáciu a teda e_i sú unikátne pre prislúchajúcu premennú X_i. Vo väčších modeloch sa často od podmienky nezávislosti spoločných faktorov upúšťa, ostáva iba podmienka na nezávislosť medzi spoločnými a špecifickými faktormi. Podmienka na strednú hodnotu a varianciu spoločných faktorov je kladená z technických dôvodov, pretože práca so štandardizovanými premennými je jednoduchšia.

Za týchto predpokladov a platnosti modelu môžeme vysloviť nasledujúce tvrdenie:

Tvrdenie 3.1. *Pre premenné $X_i, i = 1, ..., p$ riadiace sa modelom (29) platí:*

$$Var(X_i) = l_{i1}^2 + ... + l_{ik}^2 + \psi_i \tag{31}$$

$$Cov(X_i, X_j) = l_{i1}l_{j1} + ... + l_{ik}l_{jk} \tag{32}$$

Dôkaz. Po rozpísaní $Var(X_i)$ a $Cov(X_i, X_j)$ podľa modelu a použití predpokladov (i) a (ii) dostaneme:

$$\begin{aligned} Var(X_i) &= Var(l_{i1}F_1 + ... + l_{ik}F_k + e_i) = l_{i1}^2 Var(F_1) + ... + l_{ik}^2 Var(F_k) + \psi_i \\ &= l_{i1}^2 + ... + l_{ik}^2 + \psi_i \end{aligned}$$

$$\begin{aligned} Cov(X_i, X_j) &= Cov(l_{i1}F_1 + ... + l_{ik}F_k + e_i, l_{j1}F_1 + ... + l_{jk}F_k + e_j) \\ &= l_{i1}l_{j1}Var(F_1) + ... + l_{ik}l_{jk}Var(F_2) + 1.0.Var(e_1) + 0.1.Var(e_j) \\ &= l_{i1}l_{j1} + ... + l_{ik}l_{jk} \end{aligned}$$

□

Varianciu jednotlivých premenných môžeme rozdeliť na dve časti, ktoré sa nazývajú komunalita a špecifický rozptyl. i-ta komunalita je časť variancie i-tej premennej obsiahnutá spoločnými faktormi, teda $h_i^2 = l_{i1}^2 + ... + l_{ik}^2$. Špecifický rozptyl ψ_i potom predstavuje časť variancie prislúchajúcu špecifickému faktoru e_i, teda tú časť, ktorú spoločné faktory nepopisujú. Ak opäť rovnako ako pri popise PCA nazveme $\mathbf{\Sigma}$ kovariančnú maticu vektora premenných $\mathbf{X} = (X_1, ..., X_p)^T$, potom predchádzajúce tvrdenie môžeme zapísať aj v maticovej forme:

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^{\mathrm{T}} + \mathbf{\Psi} \tag{33}$$

kde $\mathbf{\Psi}$ je diagonálna matica obsahujúca na diagonále postupne variancie špecifických faktorov e_i.

V praxi väčšinou kovariančnú maticu $\mathbf{\Sigma}$ nemáme k dispozícii. Pomocou nameraných dát je ale možné ju odhadnúť. Našou snahou potom bude nájsť matice $\mathbf{L}$ a $\mathbf{\Psi}$ tak, aby $\mathbf{S} = \mathbf{L}\mathbf{L}^T + \mathbf{\Psi}$ bolo splnené čo najpresnejšie. Už na prvý pohľad je jasné, že takéto riešenie nie je jediné, pretože ak $\mathbf{L}$ spĺňa požadovanú rovnosť, bude ju spĺňať aj matica $\mathbf{L}^* = \mathbf{L}\mathbf{T}$ pre akúkoľvek ortogonálnu maticu $\mathbf{T}$. Táto vlastnosť sa ukazuje ako veľmi užitočná. Umožňuje totiž takzvanú rotáciu faktorov, ktorá môže uľahčiť ich interpretáciu. Rotáciou faktorov sa budeme zaoberať v neskoršej časti tejto kapitoly.

Dôležitou časťou faktorovej analýzy je práve jej úvodná časť, ktorou je získanie takzvaných prvých faktorových riešení. Spočíva v odhade matíc $\mathbf{L}$ a $\mathbf{\Psi}$. Na tento účel sa používajú rôzne metódy, medzi najobľúbenejšie a najčastejšie používané patria metóda hlavných komponentov, metóda maximálnej vierohodnosti či metóda využívajúca kanonické korelácie. Pretože prvá z menovaných možností je úzko spojená s PCA popisovanou v prvej kapitole, zameriame sa práve na túto metódu.

3.2 Riešenie faktorového modelu pomocou hlavných komponentov

Veľkou výhodou tejto metódy je fakt, že nekladie žiadne požiadavky na rozdelenie pozorovaných premenných, ako je to napríklad pri metóde maximálnej vierohodnosti. Je tiež pomerne jednoduchá na výpočet a preto veľmi obľúbená. Vychádza zo spektrálneho rozkladu kovariančnej alebo korelačnej matice, resp. ich výberových odhadov. Diskusia o tom, či použiť kovariančnú alebo korelačnú maticu, používa v podstate rovnaké argumenty ako v prípade PCA.

Nech teda $X_1, ..., X_p$ sú pozorované premenné, tvoriace náhodný vektor $\mathbf{X}$ s kovariančnou maticou $\mathbf{\Sigma}$. Predpokladajme ďalej, že pre vlastné hodnoty matice $\mathbf{\Sigma}$ platí $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Nech $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_\mathrm{p}$ sú príslušné vlastné vektory matice $\mathbf{\Sigma}$.

Potom je možné napísať ju ako

$$\Sigma = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^{\mathrm{T}} + ... + \lambda_{\mathrm{p}} \mathbf{v}_{\mathrm{p}} \mathbf{v}_{\mathrm{p}}^{\mathrm{T}} = \begin{pmatrix} \sqrt{\lambda_1}\mathbf{v}_1 & \dots & \sqrt{\lambda_p}\mathbf{v}_{\mathrm{p}} \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1}\mathbf{v}_1^{\mathrm{T}} \\ \vdots \\ \sqrt{\lambda_p}\mathbf{v}_{\mathrm{p}}^{\mathrm{T}} \end{pmatrix}.$$

Takýto zápis by zodpovedal modelu s toľkými faktormi, koľko je premenných, a nulovými špecifickými faktormi. Pretože ale našou snahou je popísať dáta čo najmenším počtom premenných, nie je použiteľný. Posledných niekoľko vlastných hodnôt je však často veľmi malých, a preto sa zvykne ich vplyv zanedbať. Ak počet týchto posledných vlastných hodnôt je $p - m$, dostávame teda

$$\Sigma \doteq \lambda_1 \mathbf{v}_1 \mathbf{v}_1^{\mathrm{T}} + ... + \lambda_{\mathrm{m}} \mathbf{v}_{\mathrm{m}} \mathbf{v}_{\mathrm{m}}^{\mathrm{T}} = \begin{pmatrix} \sqrt{\lambda_1}\mathbf{v}_1 & \dots & \sqrt{\lambda_m}\mathbf{v}_{\mathrm{m}} \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1}\mathbf{v}_1^{\mathrm{T}} \\ \vdots \\ \sqrt{\lambda_m}\mathbf{v}_{\mathrm{m}}^{\mathrm{T}} \end{pmatrix} = \mathbf{L}\mathbf{L}^{\mathrm{T}}.$$

Faktorový model ráta tiež so špecifickými faktormi, ktorých kovariančnou maticou je diagonálna matica $\mathbf{\Psi}$. Tú môžeme vytvoriť pomocou diagonálnych prvkov matice $\mathbf{\Sigma} - \mathbf{L}\mathbf{L}^{\mathrm{T}}$, čo umožňuje upresniť diagonálne prvky a získať tak aproximáciu

$$\mathbf{\Sigma} \doteq \begin{pmatrix} \sqrt{\lambda_1}\mathbf{v}_1 & \dots & \sqrt{\lambda_m}\mathbf{v}_{\mathrm{m}} \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1}\mathbf{v}_1 \\ \vdots \\ \sqrt{\lambda_m}\mathbf{v}_{\mathrm{m}} \end{pmatrix} + \begin{pmatrix} \psi_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \psi_p \end{pmatrix} = \mathbf{L}\mathbf{L}^{\mathrm{T}} + \mathbf{\Psi},$$

kde $\psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2$ pre $i = 1, ..., p$. V praxi sa používa odhad kovariančnej matice $\mathbf{\Sigma}$ pomocou nameraných dát, teda výberová kovariančná matica $\mathbf{S}$. Ak označíme vlastné hodnoty matice $\mathbf{S}$ $\hat{\lambda}_1 \geq ... \geq \hat{\lambda}_p \geq 0$ a príslušné vlastné vektory $\hat{\mathbf{v}}_1, ..., \hat{\mathbf{v}}_{\mathrm{p}}$, odhad matice faktorových váh $\mathbf{L}$ pre $m < p$ spoločných faktorov je potom daný ako

$$\hat{\mathbf{L}} = \begin{pmatrix} \sqrt{\hat{\lambda}_1}\hat{\mathbf{v}}_1 & \dots & \sqrt{\hat{\lambda}_m}\hat{\mathbf{v}}_{\mathrm{m}} \end{pmatrix}. \tag{34}$$

Odhadnuté variancie špecifických faktorov $\hat{\psi}_i$ pre $i = 1, ..., p$ získame ako diagonálne prvky matice $\mathbf{S} - \hat{\mathbf{L}}\hat{\mathbf{L}}^{\mathrm{T}}$.

Riešenie faktorového modelu pomocou hlavných komponentov pre korelačnú maticu získame, ak miesto výberovej kovariančnej matice $\mathbf{S}$ použijeme výberovú korelačnú maticu $\mathbf{R}$ a budeme pracovať s jej vlastnými hodnotami a vlastnými vektormi.

3.3 Počet spoločných faktorov

V niektorých prípadoch je počet spoločných faktorov m daný vopred rôznymi úvahami vyplývajúcimi väčšinou z vedeckého výskumu. Ak však takéto informácie nemáme k dispozícii, je potrebné rozhodnúť o počte použitých faktorov iným spôsobom. Nasledujúci spôsob je podobne ako pri PCA založený na veľkosti jednotlivých vlastných čísel matice $\mathbf{S}$, príp. $\mathbf{R}$.

Uvažujme maticu $\mathbf{S} - (\hat{\mathbf{L}}\hat{\mathbf{L}}^{\mathrm{T}} - \hat{\mathbf{\Psi}})$. Jej diagonálne zložky sú nulové, a ak aj ostatné zložky sú blízke nule, môžeme subjektívne prehlásiť použité m za vhodný počet faktorov. Toto je možné posúdiť pomocou posledných $p - m$ vlastných hodnôt matice $\mathbf{S}$ použitím nasledujúceho tvrdenia.

Tvrdenie 3.2. *Označme f zobrazenie priraďujúce matici $\boldsymbol{A}$ veľkosti $m \times n$ druhú mocninu jej Frobeniovej normy, teda $f(A) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2$. Potom pre maticu $\boldsymbol{S}$ s vlastnými hodnotami $\lambda_1, ..., \lambda_p$ a matice $\hat{\mathbf{L}}$, $\hat{\mathbf{\Psi}}$ získané riešením faktorového modelu pomocou hlavných komponentov z matice $\boldsymbol{S}$ platí: $f(\mathbf{S} - (\hat{\mathbf{L}}\hat{\mathbf{L}}^{\mathrm{T}} + \hat{\mathbf{\Psi}})) \leq \sum_{\mathrm{i=m+1}}^{\mathrm{p}} \lambda_{\mathrm{i}}^2$.*

Dôkaz. Pretože matica $\mathbf{S} - (\hat{\mathbf{L}}\hat{\mathbf{L}}^{\mathrm{T}} + \hat{\mathbf{\Psi}})$ má na diagonále nuly, platí

$$f(\mathbf{S} - (\hat{\mathbf{L}}\hat{\mathbf{L}}^{\mathrm{T}} + \hat{\mathbf{\Psi}})) \leq f(\mathbf{S} - \hat{\mathbf{L}}\hat{\mathbf{L}}^{\mathrm{T}}).$$

Ďalej z toho, ako bola matica $\hat{\mathbf{L}}$ vytvorená pomocou spektrálneho rozkladu matice $\mathbf{S}$, vieme, že

$$\mathbf{S} - \hat{\mathbf{L}}\hat{\mathbf{L}}^{\mathrm{T}} = \hat{\lambda}_{\mathrm{m+1}}\hat{\mathbf{v}}_{\mathrm{m+1}}\hat{\mathbf{v}}_{\mathrm{m+1}}^{\mathrm{T}} + ... + \hat{\lambda}_{\mathrm{p}}\hat{\mathbf{v}}_{\mathrm{p}}\hat{\mathbf{v}}_{\mathrm{p}}^{\mathrm{T}} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{\mathrm{T}},$$

kde $\mathbf{P} = \begin{pmatrix} \hat{\mathbf{v}}_{\mathrm{m+1}} & \dots & \hat{\mathbf{v}}_{\mathrm{p}} \end{pmatrix}$ a $\mathbf{\Lambda}$ je diagonálna matica s prvkami $\hat{\lambda}_{m+1}, ..., \hat{\lambda}_p$. Pretože zjavne $f(\mathbf{A}) = \mathbf{tr}(\mathbf{A}\mathbf{A}^{\mathrm{T}})$ a tiež vďaka tomu, že stĺpce $\mathbf{P}$ sú ortonormálne, získame požadovaný výsledok:

$$f(\mathbf{S} - (\hat{\mathbf{L}}\hat{\mathbf{L}}^{\mathrm{T}} + \hat{\mathbf{\Psi}})) \leq f(\mathbf{P}\mathbf{\Lambda}\mathbf{P}^{\mathrm{T}}) = \mathrm{tr}(\mathbf{P}\mathbf{\Lambda}\mathbf{P}^{\mathrm{T}}(\mathbf{P}\mathbf{\Lambda}\mathbf{P}^{\mathrm{T}})^{\mathrm{T}}) =$$

$$= tr(\mathbf{P}\mathbf{\Lambda}\mathbf{\Lambda}\mathbf{P}^{\mathrm{T}}) = \mathrm{tr}(\mathbf{\Lambda}\mathbf{\Lambda}) = \sum_{\mathrm{i=m+1}}^{\mathrm{p}} \lambda_{\mathrm{i}}^2.$$

□

Tento výsledok znamená, že ak súčet štvorcov posledných $p - m$ vlastných hodnôt matice $\mathbf{S}$ je malý, bude malý aj súčet štvorcov matice $\mathbf{S} - (\hat{\mathbf{L}}\hat{\mathbf{L}}^{\mathrm{T}} - \hat{\mathbf{\Psi}})$. To následne značí, že počet faktorov m je pre model dostatočný.

Používa sa tiež kritérium posudzujúce príspevok faktorov do výberových variancií premenných. Príspevok prvého spoločného faktora do prvku s_{ii} matice $\mathbf{S}$ je $\hat{l}_{i1}^2$, do celkovej výberovej variancie dát $tr(\mathbf{S})$ teda $\sum_{i=1}^{p} \hat{l}_{i1}^2 = (\sqrt{\hat{\lambda}_1}\hat{\mathbf{v}}_\mathbf{1})^{\mathrm{T}}(\sqrt{\hat{\lambda}_1}\hat{\mathbf{v}}_\mathbf{1}) = \hat{\lambda}_1$. Vo všeobecnosti j-ty faktor predstavuje $\frac{\hat{\lambda}_j}{tr(\mathbf{S})} \cdot 100\%$ celkovej výberovej variancie premenných. Počet faktorov potom určujeme tak, aby celkový príspevok spoločných faktorov predstavoval aspoň nejakú vopred stanovenú časť celkovej variancie.

3.4 Rotácia faktorov

V predchádzajúcej časti sme uviedli jednu z metód získavania začiatočných odhadov matíc $\mathbf{L}$ a $\mathbf{\Psi}$. Tieto odhady ale nemusia byť jednoducho interpretovateľné, preto je často vhodné použiť na ne rotáciu, teda prenásobiť maticu $\hat{\mathbf{L}}$ nejakou ortogonálnou maticou. Snahou je získať čo najjednoduchšiu štruktúru matice váh, najlepšie takú, kde niektoré prvky sú pomerne veľké a ostatné dosť blízke nule. Bolo navrhnutých niekoľko rôznych rotácií, medzi najobľúbenejšie patria varimax a quartimax (viď [17]). Vysvetlíme teraz používanejšiu z nich, metódu varimax, ktorá sa zameriava na maximalizáciu variancie váh faktorov. Cieľom je získať niektoré váhy v absolútnej hodnote čo najväčšie a ostatné čo najmenšie, čo často zjednoduší interpretáciu faktorov. Táto metóda umožňuje nájsť také faktory, ktoré sú spojené len s niektorými premennými. Ak označíme $\hat{k}_{ij}$ rotované odhady váh faktorov a $\tilde{k}_{ij} = \frac{\hat{k}_{ij}}{h_i}$, metóda varimax hľadá ortogonálnu transformáciu $\mathbf{T}$ takú, ktorá maximalizuje výraz

$$V = \frac{1}{p}\sum_{j=1}^{m}\left[\sum_{i=1}^{p}\tilde{k}_{ij}^4 - \frac{(\sum_{i=1}^{p}\tilde{k}_{ij}^2)^2}{p}\right]. \tag{35}$$

3.5 Faktorové skóre

Hoci zvyčajne je dôraz pri faktorovej analýze kladený na váhy jednotlivých faktorov, hlavne tých spoločných, v pôvodných premenných, často je dôležité a informatívne pozrieť sa aj na hodnoty jednotlivých faktorov pre rôzne merania, teda faktorové skóre. Preto ďalšou fázou faktorovej analýzy býva jeho odhad. Na to bolo tiež vyvinutých a predstavených niekoľko spôsobov, napríklad (vážená) metóda najmenších štvorcov alebo regresná metóda. Obe sú popísané v [9], my sa zameriame na prvú z nich, ktorá je najčastejšie používanou metódou v prípade, že prvý odhad váh sa získava metódou

hlavných komponentov.

Na začiatok predpokladajme, že vektor μ, matica $\mathbf{L}$ a špecifické variancie ψ_i modelu (30) sú známe. Ak $\mathbf{e}$ považujeme za vektor chýb, môžeme ako odhad faktorového skóre pre faktory použiť taký vektor $\hat{\mathbf{f}}$, ktorý minimalizuje výraz

$$\sum_{i=1}^{p} \frac{e_i^2}{\psi_i} = \mathbf{e}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\mathbf{e} = (\mathbf{x} - \mu - \mathbf{L}\mathbf{f})^{\mathrm{T}}\boldsymbol{\Psi}^{-1}(\mathbf{x} - \mu - \mathbf{L}\mathbf{f}). \tag{36}$$

Uvedený výraz predstavuje súčet štvorcov chýb vážených ich varianciami. Vektor $\hat{\mathbf{f}}$ potom dostaneme ako

$$\hat{\mathbf{f}} = (\mathbf{L}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}(\mathbf{x} - \mu). \tag{37}$$

V prípade, že používame namerané dáta, považujeme pri tejto metóde odhady matíc $\hat{\mathbf{L}}, \hat{\boldsymbol{\Psi}}$ a priemer meraní ako odhad strednej hodnoty za skutočné hodnoty. Faktorové skóre pre j-te meranie potom odhadneme ako

$$\hat{\mathbf{f}} = (\hat{\mathbf{L}}^{\mathrm{T}}\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}})^{-1}\hat{\mathbf{L}}^{\mathrm{T}}\hat{\boldsymbol{\Psi}}^{-1}(\mathbf{x}_{\mathrm{j}} - \bar{\mathbf{x}})^{\mathrm{T}}, \tag{38}$$

kde $\mathbf{x}_{\mathrm{j}}$ predstavuje j-te meranie pre vektor premenných $\mathbf{X}$.

4 Aplikácia PCA na volebné dáta

Častým záujmom politológov aj štatistikov sú výsledky volieb a ich analýza rôznymi metódami. Keďže ide o viacrozmerné dáta (vo voľbách do NR SR väčšinou kandiduje viac ako 20 politických strán, prezidentských kandidátov je zvyčajne aspoň 10), v ktorých by sme chceli pozorovať rôzne vzťahy, je na mieste snaha o redukciu dimenzionality týchto výsledkov aj pomocou metód popísaných v predchádzajúcich častiach tejto práce, teda pomocou PCA a FA. V tejto kapitole sa venujeme aplikácii troch rôznych druhov PCA (klasická, hrubá, logcontrast) na výsledky niektorých volieb a porovnaniu výsledkov, nasledujúca kapitola opisuje aplikáciu FA a možnú interpretáciu získaných faktorov.

Dáta použité v tejto časti práce pochádzajú z internetovej stránky Štatistického úradu SR, www.statistics.sk. V sekcii Volebná štatistika sú prístupné výsledky rôznych druhov volieb ako aj výsledky referend. Pre potreby tejto práce sme používali výsledky volieb do NR SR a prezidentských volieb, ktoré sú prístupné na rôznych úrovniach - za celú republiku, podľa krajov, podľa okresov, podľa obcí. Pri niektorých voľbách sú výsledky dostupné maximálne na úrovni krajov či okresov, no ŠÚ SR na požiadanie poskytuje aj kompletnejšie výsledky. Vo väčšine prípadov sme použili výsledky podľa okresov (SR má v súčasnosti 79 okresov, pred rokom 1996 ich bolo 42), pre porovnanie sme však v jednom prípade použili aj výsledky za obce (2926 obcí). Zoznam kandidátov vo všetkých analyzovaných voľbách spolu s ich volebnými výsledkami je uvedený v Prílohe B.

Aby sme získali komplexnejší pohľad na správanie voličov na Slovensku, zamerali sme sa pri aplikácii PCA na výsledky volieb v rôznych časových obdobiach. Vybrali sme voľby do NR SR v rokoch 1992, 2002 a 2010 a prezidentské voľby v roku 1999.

Z prvej kapitoly vyplýva, že pri použití PCA je potrebné urobiť dve rozhodnutia, a to, či pri hľadaní hlavných komponentov vychádzať z kovariančnej alebo korelačnej matice a podľa akého kritéria určiť počet hlavných komponentov. Pretože pri volebných dátach sú všetky premenné (predstavujúce volebný zisk pre rôzne politické subjekty) merané v rovnakých jednotkách a dôležité sú tiež rozdiely medzi jednotlivými premennými, ktoré by po štandardizácii premenných ustúpili, rozhodli sme sa pre použitie kovariančnej matice. Ako kritérium pre rozhodnutie o počte hlavných komponentov sme

tu používali kumulatívne percento vysvetlenej variancie, pričom hranicu sme zvolili na úrovni 85%. Výsledky každých analyzovaných volieb sú uložené v tabuľkách, ktoré obsahujú vektory váh pre hlavné komponenty, percento variancie vysvetlené daným hlavným komponentom (označené % PC var), ako aj kumulatívne percento variancie (označené % kumul).

Všetky výpočty boli robené pomocou programu MATLAB, ktorý je pre študentov FMFI UK prístupný priamo na fakulte. Výpočtová jednoduchosť PCA umožňuje získať hlavné komponenty len pomocou príkazu *cov* na výpočet výberovej kovariančnej matice $\mathbf{\Sigma}$ a príkazu *eig* na výpočet vlastných hodnôt a vlastných vektorov matice, alebo pomocou príkazu *svd*, ktorého výstupom je singulárny rozklad zadanej matice.

4.1 Voľby do NR SR 1992

Volieb do NR SR v roku 1992 sa zúčastnilo 23 politických strán, ktoré k účasti vo voľbách podnietili 84,2% oprávnených voličov. V roku 1992 bolo Slovensko rozdelené na 42 okresov, prechod na dnešný systém 79 okresov bol súčasťou reformy verejnej správy v roku 1996. Aby bolo možné výsledky analýz v jednotlivých rokoch porovnávať, z výsledkov volieb v roku 1992 podľa obcí sme získali výsledky podľa dnešných okresov. Matica $\mathbf{X}$ je tak rozmeru 79×23. Výsledky troch druhov PCA aplikovaných na volebný zisk strán v týchto voľbách sú zaznamenané v Tabuľke 1.

Ako je vidieť v Tabuľke 1, pre klasickú PCA na vysvetlenie stanovenej hranice kumulatívneho percenta variancie stačia dva hlavné komponenty, pri hrubej aj logcontrast PCA sú však už potrebné tri. Percento variancie, ktorú jednotlivé komponenty vysvetľujú, klesá v prípade logcontrast PCA najpomalšie, teda aj druhý komponent vysvetľuje ešte dosť veľký podiel variancie obsiahnutej v dátach, zatiaľ čo predovšetkým v prípade klasickej PCA už prvý komponent predstavuje veľkú časť celkovej variancie a pre ostatné tak neostáva toľko priestoru.

V prípade klasickej PCA prvý komponent hovorí o tom, že najväčším zdrojom variancie v dátach je strana HZDS, ktorá mala na úrovni celého Slovenska najvyšší zisk. Naproti tomu záporné váhy majú strany MKM-EGY a MPP-MOS, ktoré autori [11] zaradili do sústavy maďarských strán, a tiež rómska ROI, teda strany národnostných menšín.

	PCA		hrubá PCA			logcontrast PCA		
Polit. strana	PC1	PC2	PC1	PC2	PC3	PC1	PC2	PC3
HSD-SMS	0,002	0,002	0,001	0,001	0,003	0,021	0,060	0,024
HZSP-SRÚ	0,001	0,001	0,001	0,002	0,001	0,004	0,017	0,055
HZDS	**0,916**	0,161	**-0,771**	**-0,465**	-0,262	0,145	0,039	0,078
SDĽ	0,198	0,151	0,069	**0,475**	**-0,701**	0,070	-0,046	-0,006
SPI	0,004	0,008	0,015	0,044	-0,041	0,090	-0,202	-0,040
HZOS	0,006	0,000	-0,004	-0,001	0,006	0,139	0,090	0,091
SSL-SNZ	0,004	0,003	0,000	0,004	0,007	0,081	-0,028	0,003
SKDH	0,064	-0,001	-0,061	-0,001	0,198	0,168	0,078	0,167
MKM-EGY	-0,190	**0,917**	**0,584**	**-0,596**	-0,127	**-0,670**	-0,281	0,201
HSS	0,002	0,004	0,001	-0,001	0,008	**-0,346**	**0,765**	**-0,461**
SZ	0,015	0,016	0,010	0,023	0,015	0,016	0,006	0,030
KDH	0,152	-0,002	-0,060	**0,318**	**0,461**	0,166	0,033	0,022
ODÚ	0,040	0,036	0,053	0,151	0,054	0,032	-0,040	0,138
ZPR-RSČ	0,002	0,005	0,005	0,012	-0,006	0,044	-0,079	-0,082
NALI	0,002	0,000	0,000	0,002	0,002	0,071	0,029	0,139
SZS	0,038	0,017	0,001	0,043	0,020	0,063	0,023	0,031
ROI	-0,002	0,008	0,016	0,033	-0,013	0,093	**-0,450**	**-0,767**
SDSS	0,050	0,039	0,016	0,134	-0,033	0,080	-0,025	-0,011
KSS	0,003	0,017	0,008	-0,003	-0,033	0,019	-0,064	-0,078
DS-ODS	0,058	0,024	0,019	0,087	0,035	0,024	0,002	0,131
SNS	0,208	0,069	-0,094	-0,050	**0,404**	0,074	0,134	0,161
SĽS	0,006	0,000	-0,005	0,001	0,020	0,131	0,104	0,088
MPP-MOS	-0,075	**0,319**	0,196	**-0,215**	-0,021	**-0,515**	-0,164	0,084
% PC var	57,33	31,30	61,96	22,75	7,29	46,07	31,59	10,42
% kumul	57,33	88,63	61,96	84,71	92,00	46,07	77,65	88,07

Tabuľka 1: Hlavné komponenty pre voľby do NR SR 1992

Druhý komponent akoby čiastočne upravoval to, čo prvý komponent vysvetlil nesprávne. S vysokými kladnými váhami tentokrát vystupuje sústava maďarských strán. Najvyššie hodnoty teda komponent nadobúda v maďarských okresoch, nízky je predovšetkým v hlavnom meste. Ak túto metódu zhrnieme, najväčším zdrojom variancie sú tu HZDS a maďarské strany a nové premenné zložené predovšetkým z ich výsledkov vysvetľujú veľkú časť informácie ukrytej v dátach.

V prvom komponente hrubej PCA nájdeme podobne ako pri klasickej PCA v absolútnej hodnote vysokú váhu pre HZDS, s opačným znamienkom ako váhy maďarských strán. Váhy pre maďarské strany sú tu však oveľa výraznejšie, v skutočnosti predstavujú druhú a tretiu najvyššiu váhu, teda môžeme hovoriť o kontraste medzi HZDS a sústavou maďarských strán. Druhý komponent prikladá týmto trom stranám pomerne

vysoké váhy s rovnakým znamienkom, možno tu ale pozorovať protiklad s viacerými ďalšími stranami, predovšetkým SDĽ a KDH. Napokon tretí komponent stavia SDĽ proti KDH a SNS.

Už pri prvom pohľade na Tabuľku 1 je zrejmé, že logcontrast PCA opisuje vzťahy medzi premennými inak. V prvom hlavnom komponente síce aj v tomto prípade výrazne vystupujú maďarské strany, pridáva sa však aj HSS, ktorá vo voľbách získala veľmi nízky počet hlasov (0,1% všetkých hlasov) a teda výsledky volieb nijako významne neovplyvnila. Napriek tomu je výrazná už v prvom logcontrast hlavnom komponente, ktorý predstavuje takmer polovicu všetkej variancie. Druhý komponent má vysokú váhu predovšetkým pre HSS, do kontrastu s touto stranou stavia ďalšiu málo úspešnú stranu - ROI (0,6%). K výraznejšej polovici strán sa zaraďujú aj maďarské strany. Na záver tretí logcontrast hlavný komponent vysvetľujúci približne 10% celkovej variancie, stále pomerne veľkú časť, kladie silné záporné váhy na HSS a ROI, teda akoby opravoval druhý komponent, najvyššiu kladnú váhu má silná maďarská koalícia MKM-EGY.

4.2 Voľby do NR SR 2002

V roku 2002 sa volieb do NR SR zúčastnilo 25 politických strán, matica $\mathbf{X}$ je teda rozmeru 79×25. Účasť vo voľbách bola 70,06%. Výsledky troch druhov PCA aplikovaných na tieto dáta sú uložené v Tabuľke 2.

Pri analýze volieb do NR SR v roku 2002 klasickou PCA sa ako najväčší zdroj variancie javí SMK. Prvý komponent je preto najvyšší v okresoch s vysokým podielom maďarského obyvateľstva - Dunajská Streda, Komárno, Nové Zámky, Galanta či Levice. Druhý hlavný komponent potom zoskupuje strany s najvyššími predvolebnými preferenciami a najvyššími volebnými výsledkami, SDKÚ, HZDS a SMER. Ukazuje sa tu akoby vymenená tendencia oproti roku 1992, kedy najvyšším zdrojom variancie bolo HZDS a najvyššie váhy v druhom hlavnom komponente boli pri maďarských stranách.

Hlavné komponenty získané pomocou hrubej PCA tiež poukazujú na veľkú úlohu strán SMK, SDKÚ a HZDS. Prvý komponent kladie veľkú zápornú váhu na SMK. Druhý predstavuje kontrast medzi silnými stranami SDKÚ a HZDS, jeho hodnota je vysoká najmä pre okresy s vysokým podielom mestského obyvateľstva - Bratislava a Košice.

	PCA		hrubá PCA		logcontrast PCA					
Polit. strana	PC1	PC2	PC1	PC2	PC1	PC2	PC3	PC4	PC5	PC6
SZS	0,003	0,025	0,002	0,007	0,016	0,033	0,082	0,009	0,055	0,031
SDKÚ	-0,029	**0,594**	0,048	**0,769**	0,005	0,193	0,291	**0,276**	0,051	0,022
SDPO	0,000	0,003	0,000	0,006	0,067	-0,021	0,129	0,120	**0,355**	**-0,714**
SDĽ	-0,002	0,031	0,013	-0,008	-0,079	-0,003	-0,032	-0,074	0,156	0,204
SMER	-0,027	**0,404**	0,124	-0,007	-0,087	0,067	0,044	-0,07	0,072	0,016
HZDS	-0,101	**0,565**	0,332	**-0,577**	-0,187	0,035	-0,062	-0,177	-0,149	0,021
OKS	0,000	0,012	0,001	0,013	0,009	0,224	0,102	**0,215**	0,041	0,104
HZD	-0,013	0,091	0,044	-0,050	-0,152	0,113	-0,146	-0,052	-0,105	-0,254
ROMA	0,003	-0,001	-0,006	-0,004	**0,314**	**-0,396**	**-0,590**	**0,411**	0,019	-0,052
KSS	-0,016	0,112	0,064	-0,054	-0,055	-0,071	-0,035	-0,079	**0,317**	0,073
SMK-MKP	**0,992**	0,106	**-0,916**	-0,175	**0,748**	**0,492**	-0,112	**-0,290**	-0,046	0,050
KDH	-0,059	0,227	0,142	0,030	-0,187	0,028	0,229	0,115	-0,236	0,032
ĽS	0,000	0,001	0,000	0,000	-0,019	0,102	-0,136	**0,347**	**-0,529**	0,067
ZRS	-0,002	0,009	0,007	-0,013	-0,132	-0,165	-0,246	**-0,376**	0,159	-0,081
ĽB	-0,001	0,003	0,003	-0,002	-0,042	-0,110	-0,066	**0,247**	**0,304**	**0,460**
ANO	0,006	0,212	0,026	0,152	0,015	0,024	0,07	0,088	0,125	0,007
B-RRS	0,000	0,002	0,001	-0,001	-0,031	-0,105	0,062	**-0,207**	0,001	0,153
ŽAR	-0,001	0,011	0,003	0,006	-0,047	0,015	0,129	0,054	0,079	-0,035
SDA	-0,005	0,063	0,007	0,066	-0,004	0,150	0,190	0,202	0,158	0,009
SNJ	-0,001	0,003	0,002	-0,002	-0,133	0,016	-0,048	-0,091	-0,162	0,02
NOSNP	0,002	0,023	0,004	-0,010	-0,029	-0,050	-0,032	**-0,285**	0,028	0,195
SNS	-0,008	0,099	0,044	-0,055	-0,159	0,141	-0,104	-0,087	-0,14	-0,056
ROSA	0,000	0,005	0,003	-0,004	-0,046	-0,134	-0,068	-0,11	-0,001	0,016
ROISR	0,001	0,002	-0,001	0,003	**0,368**	**-0,602**	**0,502**	-0,125	**-0,289**	-0,025
P SNS	-0,030	0,141	0,054	-0,088	-0,153	0,024	-0,154	-0,063	-0,264	-0,263
% PC var	56,25	29,25	65,83	20,13	42,18	26,62	7,24	4,64	3,93	2,95
% kumul	56,25	85,5	65,83	85,96	42,18	68,8	76,04	80,68	84,6	87,55

Tabuľka 2: Hlavné komponenty pre voľby do NR SR 2002

Prvé tri logcontrast hlavné komponenty, predstavujúce spoločne 76% celkovej variancie, sa zameriavajú na národnostné menšiny. Prvý komponent kombinuje rómske strany ROMA a ROISR s maďarskou stranou SMK. Ďalej vidíme "súboj" medzi Maďarmi a rómskymi stranami a napokon tretí komponent stavia proti sebe ROISR a ROMA. Ostatné komponenty zahŕňajúce podiely variancie menšie ako 5% sa tiež zameriavajú väčšinou na menej úspešné strany.

4.3 Voľby do NR SR 2010

Volieb do NR SR v roku 2010 sa zúčastnilo 18 politických strán, volebná účasť dosiahla 58,83% oprávnených voličov. Pre tieto voľby sme sa rozhodli analýzu urobiť na dátach

za okresy a tiež za obce, porovnať výsledky a zistiť tak, aký veľký je rozdiel medzi takto získanými informáciami. Pracovali sme teda s maticami $\mathbf{X_1}$ rozmeru 79×18 a $\mathbf{X_2}$ rozmeru 2926×18. Výsledky sú zaznamenané v Tabuľke 3 a v Tabuľke 4. Pri aplikácii na výsledky podľa obcí sme sa však už neriadili percentom vysvetlenej variancie. Pre porovnanie štruktúry hlavných komponentov a vysvetľovacej sily sme uvažovali rovnaký počet hlavných komponentov ako v prípade okresov. Na porovnanie sme použili predovšetkým uhol medzi konkrétnymi dvoma vektormi váh pre hlavné komponenty.

	PCA		hrubá PCA		logcontrast PCA		
Polit. strana	PC1	PC2	PC1	PC2	PC1	PC2	PC3
EDS	0,004	0,006	0,000	0,014	0,029	0,279	0,051
Únia	0,018	0,003	-0,002	-0,015	-0,056	-0,083	-0,189
SRK	0,001	0,002	0,000	0,007	0,090	**0,730**	**-0,409**
Paliho Kapurková	0,014	0,002	-0,006	-0,005	-0,079	-0,088	-0,075
SaS	**0,321**	0,074	0,000	**-0,420**	-0,061	-0,248	-0,215
SDĽ	0,051	0,008	-0,024	-0,005	-0,085	-0,059	0,069
SMK - MKP	-0,067	**0,568**	**0,397**	**0,373**	**0,847**	-0,071	0,169
ĽS - HZDS	0,074	0,003	-0,059	0,104	-0,104	0,025	**0,328**
KSS	0,014	-0,001	-0,016	0,014	-0,112	0,029	0,266
SNS	0,134	-0,001	-0,087	0,036	-0,138	-0,158	0,292
ND	0,007	0,001	-0,001	-0,014	-0,090	-0,289	-0,250
ZRS	0,003	-0,001	-0,007	0,007	-0,156	0,163	**0,393**
KDH	0,199	-0,034	-0,141	-0,164	-0,178	-0,095	-0,297
ĽSNS	0,021	-0,002	-0,014	0,009	-0,091	0,172	0,002
SDKÚ - DS	**0,390**	0,158	0,094	**-0,634**	-0,018	-0,230	**-0,316**
AZEN	0,002	0,001	-0,001	0,002	0,003	0,172	0,061
SMER-SD	**0,820**	-0,026	**-0,696**	**0,390**	-0,142	-0,036	0,183
MOST - HíD	-0,025	**0,803**	**0,563**	**0,300**	**0,343**	-0,212	-0,062
% PC var	53,59	34,06	65,35	23,52	65,12	17,22	4,5
% kumul	53,59	87,65	65,35	88,87	65,12	82,34	86,84

Tabuľka 3: Hlavné komponenty pre voľby do NR SR 2010 za okresy

Klasická PCA aplikovaná na dáta zozbierané za okresy identifikuje ako smer najväčšej variancie kombináciu s vysokými váhami predovšetkým pre SaS, SDKÚ a SMER, čo sa ukazuje aj v prvom hlavnom komponente pre obce. Druhý hlavný komponent je v oboch prípadoch tvorený s veľkou váhou na maďarských stranách SMK a MOST, no pri analýze dát pre obce sa ako významné ukazuje aj SDKÚ, a navyše tiež SMER s opačnou váhou oproti trom menovaným. Uhol zvieraný vektormi váh pre prvý hlavný komponent je veľmi malý, iba $13,5°$, v prípade druhých hlavných komponentov už ale uhol narástol na približne $32°$. Hoci váhy pre prvý hlavný komponent sú v oboch prí-

	PCA		hrubá PCA		logcontrast PCA		
Polit. strana	PC1	PC2	PC1	PC2	PC1	PC2	PC3
EDS	0,005	0,001	0,004	0,010	0,158	**0,530**	**0,303**
Únia	0,020	0,007	-0,005	-0,006	-0,054	-0,052	0,197
SRK	0,002	0,000	0,006	-0,005	0,184	**0,583**	0,092
Paliho Kapurková	0,016	0,002	-0,006	-0,006	-0,125	**-0,318**	0,040
SaS	**0,380**	0,141	-0,054	**-0,219**	-0,052	-0,087	-0,024
SDĽ	0,053	-0,007	-0,027	-0,012	-0,102	-0,050	-0,043
SMK - MKP	0,015	**0,373**	**0,527**	**0,377**	**0,738**	**-0,208**	-0,114
ĽS - HZDS	0,075	-0,038	-0,038	0,008	-0,094	0,059	-0,022
KSS	0,016	-0,010	-0,013	0,013	-0,148	-0,027	-0,043
SNS	0,107	-0,071	-0,079	-0,013	-0,216	0,003	-0,006
ND	0,008	-0,001	-0,002	-0,007	-0,063	**-0,371**	**0,409**
ZRS	0,004	-0,004	-0,005	0,005	-0,121	0,089	-0,751
KDH	0,190	-0,042	-0,112	**-0,544**	-0,244	-0,016	0,020
ĽSNS	0,027	-0,018	-0,014	-0,021	-0,169	0,075	0,237
SDKÚ - DS	**0,510**	**0,400**	-0,028	**-0,363**	-0,011	-0,091	-0,022
AZEN	0,003	0,001	-0,001	0,001	0,016	0,087	-0,228
SMER-SD	**0,724**	**-0,448**	**-0,661**	**0,584**	-0,108	0,018	-0,044
MOST - HÍD	0,117	**0,686**	**0,510**	**0,201**	**0,410**	**-0,224**	-0,004
% PC var	89,38	6,37	65,63	12,09	26,48	9,58	7,65
% kumul	89,38	95,75	65,63	77,72	26,48	36,06	43,71

Tabuľka 4: Hlavné komponenty pre voľby do NR SR 2010 za obce

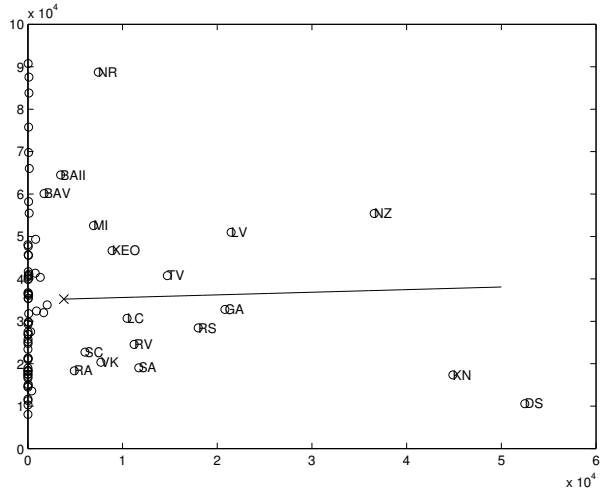
padoch veľmi podobné, vo vysvetľovacej sile prvých hlavných komponentov je celkom výrazný rozdiel. Zatiaľ čo pri okresoch vysvetľuje prvý komponent iba o niečo viac ako polovicu celkovej variancie, v prípade obcí je to takmer 90%. To môže byť spôsobené tým, že pri analyzovaní výsledkov podľa obcí je rozdiel medzi varianciou najúspešnejších strán a tých málo úspešných výraznejší, čo je dôsledkom väčších rozdielov vo veľkosti obcí.

Takáto tendencia sa už neukazuje pri použití hrubej PCA, kde sa zbavujeme priameho vplyvu veľkosti obce alebo okresu. Prvý hlavný komponent vysvetľuje v oboch prípadoch 65% celkovej variancie. Aj štruktúra prvého hlavného komponentu je veľmi podobná, v oboch prípadoch sú do kontrastu postavené maďarské strany a SMER. Dôkazom podobnosti týchto vektorov váh je aj uhol, ktorý zvierajú. Je to približne $11,5°$. Pozorovateľná je aj podobnosť medzi hlavnými komponentmi vysvetľujúcimi druhé najväčšie množstvo variancie. V oboch sú dôležité strany SMK, MOST a SMER proti dvom silným pravicovým stranám voleným najmä mestským obyvateľstvom, SaS a SDKÚ. Rovnako ako v prípade klasickej PCA zvieraný uhol stúpol, na približne $33°$.

V prípade logcontrast hlavných komponentov stále pozorujeme istú podobnosť, ktorá sa ale so stúpajúcim poradovým číslom komponentu stráca. Toto potvrdzujú aj uhly medzi vektormi váh pre hlavné komponenty, v prípade prvého komponentu necelých 15°, ďalej potom 30° a pri treťom komponente už dokonca 59°. Ako najväčší zdroj variancie vystupujú v oboch prípadoch maďarské strany. Zatiaľ čo pri analýze volebných výsledkov podľa okresov v druhom logcontrast hlavnom komponente dominuje strana SRK, výsledky podľa obcí prikladajú veľkú váhu aj niekoľkým ďalším stranám. Spoločným menovateľom tretieho komponentu je zas ZRS, ostatné strany s vysokými váhami sa však už nezhodujú.

Veľkým rozdielom v prípade logcontrast hlavných komponentov je tiež podiel vysvetlenej variancie. Vo výsledkoch podľa okresov už prvý komponent vysvetľuje približne 65% celkovej variancie, prvé tri komponenty stačia na dosiahnutie nami požadovanej 85%-nej hranice. Oproti tomu prvé tri logcontrast hlavné komponenty v prípade obcí nevysvetľujú ani polovicu variancie a pre dosiahnutie hranice 85% by bolo potrebné uvažovať až desať komponentov.

4.4 Voľby do NR SR a národnosť



Obr. 3: Hlasy pre maďarské a slovenské strany

Pri každých analyzovaných voľbách sa ako veľmi dôležité javí obyvateľstvo maďarskej národnosti, ktoré vo väčšine prípadov odovzdá hlas maďarským stranám (a tiež väčšina hlasov pre maďarské strany zrejme pochádza od obyvateľstva maďarskej ná-

rodnosti). Maďarská menšina žije pomerne koncentrovane iba na niektorých častiach územia SR, čo spôsobuje veľkú výberovú varianciu vo volebných výsledkoch maďarských strán. Preto sme sa rozhodli pre ilustračný Obrázok 3, ktorý ukazuje okresy podľa počtu hlasov odovzdaných maďarským stranám oproti počtu hlasov odovzdaným ostatným stranám (čo možno takmer rovnako presne nazvať tiež ako počet hlasov odovzdaný slovenským stranám). Pre lepšiu prehľadnosť sú však označené iba okresy, ktoré sa na maďarskej osi líšia od väčšiny (KEO predstavuje okres Košice - okolie). Obrázok v práci je založený na výsledkoch volieb do NR SR pre rok 1992, no veľmi podobné správanie je možné pozorovať aj v prípadoch ostatných volieb. Krížik predstavuje priemerný okres, teda okres s priemerným počtom Slovákov a priemerným počtom Maďarov. Takmer vodorovná čiara vychádzajúca z neho predstavuje os získanú prenásobením váh druhého hlavného komponentu volieb v roku 1992 na seba kolmými vektormi získanými tiež z váh pre druhý hlavný komponent. Prvá súradnica zodpovedá násobeniu váh vektorom s nulami na miestach zodpovedajúcich slovenským stranám a váhami komponentu na miestach pre maďarské strany, v druhej súradnici je tento vektor tvorený naopak.

4.5 Prezidentské voľby 1999

Špecifickým druhom volieb sú voľby prezidenta, v ktorých si ľud priamo volí jedného zástupcu. V prípade prezidentských volieb je oproti voľbám do NR SR zvyčajne náročnejšie urobiť nejaké zaradenie kandidátov do skupín na základe spoločných charakteristík, volebný zisk kandidáta často nezávisí len od jeho politickej príslušnosti a názorov, ale tiež od osobnosti kandidáta. Napriek tomu nám môžu metódy ako PCA priniesť cenný náhľad do vzťahov medzi volebným ziskom jednotlivých kandidátov a celkovej štruktúry výsledkov.

V tabuľke 5 sú uložené výsledky klasickej, hrubej a logcontrast PCA aplikovanej na výsledky prezidentských volieb v roku 1999, v ktorých kandidovalo 9 kandidátov a svoje volebné právo využilo 73,89% všetkých voličov.

Ako je jasne viditeľné v Tabuľke 5, pri klasickej aj hrubej PCA najvýraznejšie vystupujú v prvých hlavných komponentoch dvaja najvýraznejší kandidáti, ktorí proti sebe neskôr bojovali v druhom kole, Rudolf Schuster a Vladimír Mečiar. Už prvé dva

	PCA		hrubá PCA	logcontrast PCA		
Kandidát	PC1	PC2	PC1	PC1	PC2	PC3
Ján Demikát	0,003	0,000	0,001	0,414	0,269	0,166
Juraj Lazarčík	0,003	-0,005	-0,004	0,043	0,334	0,176
Vladimír Mečiar	0,232	-0,962	-0,663	-0,226	0,334	-0,116
Ivan Mjartan	0,029	-0,075	-0,041	-0,228	0,079	-0,104
Rudolf Schuster	0,968	0,241	0,747	0,322	-0,294	0,627
Ján Slota	0,018	-0,058	-0,033	-0,134	0,386	0,071
Juraj Švec	0,005	-0,016	-0,011	-0,280	-0,162	-0,357
Magda Vášáryová	0,083	-0,078	0,009	-0,187	-0,619	0,076
Boris Zala	0,010	-0,021	-0,008	-0,336	-0,162	0,078
Nevoliteľní kandidáti	0,004	0,002	0,003	0,613	-0,166	-0,616
% PC var	69,6	29,07	94,57	48,03	25,47	12,82
% kumul	69,6	98,67	94,57	48,03	73,47	86,59

Tabuľka 5: Hlavné komponenty pre voľby do NR SR 2010 za obce

hlavné komponenty (v prípade hrubej PCA dokonca už ten prvý) výrazne prekračujú stanovenú hranicu vysvetlenej variancie, čo hovorí o tom, že títo dvaja kandidáti najviac charakterizujú analyzovaný súbor dát. Táto informácia nie je nijako prekvapivá, keďže ich volebný zisk v oboch prípadoch prekročil milión hlasov, no z ostatných kandidátov nikto nedosiahol ani hranicu dvestotisíc.

V poslednej časti tabuľky vyzerá situácia úplne inak. Logcontrast PCA opisuje ako najväčší zdroj variancie skupinu nevoliteľných kandidátov, ktorú predstavuje pravdepodobne predovšetkým M. Kováč. Ten pôvodne tiež kandidoval na post hlavy štátu, no neskôr sa kandidatúry vzdal v prospech R. Schustera. V absolútnej hodnote vysoké sú aj váhy pri J. Demikátovi, R. Schusterovi a B. Zalovi. Druhý logcontrast hlavný komponent sa zameriava na M. Vášaryovú, ktorú spoločne s R. Schusterom stavia proti trojici J. Lazarčík, V. Mečiar a J. Slota. Zaujímavý je tretí komponent, ktorý opisuje kontrast medzi R. Schusterom (najvyšší volebný zisk) a nevoliteľnými kandidátmi (najnižší volebný zisk).

4.6 Aproximácia pomocou singulárneho rozkladu

Ako jedno z možných porovnaní prístupov k volebným dátam sme zvolili aj ich aproximáciu pomocou singulárneho rozkladu. Ako sme uviedli v prvej a druhej kapitole, singulárny rozklad a simpliciálny singulárny rozklad matice poskytujú istý spôsob od-

hadu matice. Ak $\mathbf{X_c} = \mathbf{ULV}^{\mathrm{T}}$ a z pôvodných matíc $\mathbf{U}, \mathbf{L}, \mathbf{V}$ ponecháme iba časť, teda prvých niekoľko stĺpcov, získame aproximáciu pôvodnej matice $\mathbf{X_c}$, pre ktorú platí: $\left\|\tilde{\mathbf{X}}_\mathbf{c} - \mathbf{X_c}\right\|^2 = l_{\mathrm{k}+1}^2 + l_{\mathrm{k}+2}^2 + ... + l_{\mathrm{r}}^2$. Analogicky možno po riadkoch kompozičnú maticu aproximovať aj pomocou simpliciálneho singulárneho rozkladu. Vyjadriť chybu už však nie je tak jednoduché.

Pre voľby v roku 1992 sme teda vytvorili takéto aproximácie s použitím prvých 3 stĺpcov (počet hlavných komponentov potrebných na vysvetlenie aspoň 85% celkovej výberovej variancie v dátach). Porovnanie chýb meraných tak, ako je to spomenuté v predchádzajúcom odseku, ukazuje, že v tomto prípade klasický singulárny rozklad aproximuje maticu dát lepšie ako singulárny. Frobeniova norma rozdielu matice a jej aproximácie je v prípade singulárneho rozkladu rovná približne $0,29$, zatiaľ čo pre simpliciálny singulárny rozklad je to asi $1,89$. Zaujímalo nás však, či je takýto vzťah medzi chybami aproximácie stály, teda či pre rôzne počty použitých vektorov je chyba pri simpliciálnom singulárnom rozklade vždy väčšia. Tabuľka 6 zachytáva chyby aproximácie pre rôzne počty použitých vektorov pre singulárny rozklad aj simpliciálny singulárny rozklad. Aby sme sa lepšie presvedčili o vzťahu medzi chybami, ľavá časť tabuľky zachytáva tieto chyby pre rok 1992, stredná pre rok 2010 (okresy) a pravá pre prezidentské voľby 1999.

Ako sa ukazuje, pre aproximáciu po riadkoch kompozičnej matice dát je úspešnejší obyčajný singulárny rozklad ako ten simpliciálny. Frobeniova norma rozdielu aproximovanej a pôvodnej matice je pri rovnakom počte použitých stĺpcov matíc singulárneho rozkladu menšia pre obyčajný singulárny rozklad. To môže zodpovedať aj faktu, že na vysvetlenie rovnakého percenta celkovej výberovej variancie dát je v prípade log-contrast PCA často potrebné použiť viac hlavných komponentov ako v prípade hrubej PCA.

Hoci doteraz sme s PCA spájali matice $\mathbf{V}$ a $\mathbf{L}$ zo singulárneho rozkladu matice $\mathbf{X_c}$, dôležité informácie poskytuje aj matica $\mathbf{U}$. Ako sme uviedli v prvej kapitole, $\mathbf{UL}$ je matica, ktorá obsahuje hodnoty hlavných komponentov pre jednotlivé merania. Pretože prvý komponent pre voľby v roku 1992 a v roku 2010 spolu s druhým komponentom pre voľby v roku 2002 vyšli podobne v tom zmysle, že vyzdvihujú najúspešnejšie slovenské strany, a rovnaký vzťah sa ukazuje medzi druhým komponentom pre voľby v roku 1992

Počet	Klasický	Simpliciálny
1	1,397	3,986
2	0,562	2,529
3	0,294	1,886
4	0,133	1,494
5	0,064	0,460
6	0,040	0,140
7	0,026	0,075
8	0,015	0,056
9	0,010	0,053
10	0,005	0,019
11	0,003	0,016
12	0,002	0,012
13	0,001	0,012
14	0,001	0,009
15	0,000	0,004
16	0,000	0,004
17	0,000	0,003
18	0,000	0,002
19	0,000	0,001
20	0,000	0,001
21	0,000	0,000
22	0,000	0,000
23	0,000	0,000

Počet	Klasický	Simpliciálny
1	0,967	1,395
2	0,310	0,940
3	0,146	0,644
4	0,095	0,657
5	0,051	0,483
6	0,027	0,237
7	0,014	0,166
8	0,007	0,059
9	0,004	0,016
10	0,003	0,007
11	0,002	0,007
12	0,001	0,001
13	0,001	0,001
14	0,000	0,000
15	0,000	0,000
16	0,000	0,000
17	0,000	0,000
18	0,000	0,000

Počet	Klasický	Simpliciálny
1	0,239	0,593
2	0,021	0,016
3	0,006	0,042
4	0,003	0,020
5	0,001	0,017
6	0,001	0,009
7	0,000	0,010
8	0,000	0,003
9	0,000	0,000
10	0,000	0,000

Tabuľka 6: Chyby aproximácie

a v roku 2010 a prvým komponentom pre voľby v roku 2002, ktoré kladú vysoké váhy na maďarské strany, zaujímali nás tiež hodnoty týchto komponentov v jednotlivých okresoch. Hoci na prvý pohľad by sa mohlo zdať, že štruktúra spomínaných hlavných komponentov je veľmi podobná a teda vektory ich hodnôt v jednotlivých okresoch by mali byť takmer rovnobežné, uhly medzi nimi sa pohybujú zväčša od 14° do 21°. Dokonca pri porovnaní prvého hlavného komponentu z roku 1992 a druhého hlavného komponentu z roku 2002 je to viac ako 31°.

Okrem toho, že matica **U** poskytuje informáciu o hodnotách hlavných komponentov, v špeciálnom prípade, kedy priemer každého stĺpca aj priemer každého riadku je 0, predstavuje matica **U** navyše tiež váhy hlavných komponentov pre situáciu, kedy by sme sa na maticu **X** pozerali opačne - teda keby riadky predstavovali premenné a stĺpce ich merania. Toto ale nie je prípad volebných dát, ktoré sú nezáporné.

4.7 Zhrnutie výsledkov PCA na volebných dátach

V tejto kapitole boli prezentované výsledky rôznych druhov PCA na niekoľkých súboroch dát opisujúcich výsledky volieb na Slovensku. Vo všeobecnosti možno povedať, že logcontrast PCA zvyčajne vysvetľuje v prvých komponentoch menší podiel celkovej variancie, no podiel variancie prislúchajúci jednotlivým komponentom tým pádom klesá pomalšie, teda oproti klasickej a hrubej PCA druhý či tretí komponent stále vysvetľujú vysoké percento celkovej variancie.

V prípade volieb do NR SR sú často veľkým zdrojom variancie strany menšín, predovšetkým maďarskej, čo je spôsobené geografickým rozložením obyvateľstva. Národnostné strany získavajú veľký počet aj podiel hlasov v okresoch, kde je vysoký podiel obyvateľstva danej národnostnej menšiny, no v ostatných okresoch je ich podiel zanedbateľný. Preto premenné predstavujúce volebný zisk týchto strán majú veľkú výberovú varianciu a v prvých pár hlavných komponentoch vystupujú s vysokými váhami.

Ako sa ukázalo v časti 4.3, váhy predovšetkým prvého hlavného komponentu v prípade všetkých troch použitých druhov PCA nie sú veľmi citlivé na to, či je analýza vykonávaná na výsledkoch podľa okresov alebo obcí. So stúpajúcim poradovým číslom hlavného komponentu však uhol zvieraný vektorom váh získaným z výsledkov podľa okresov a tým získaným z výsledkov za obce rastie. V prípade klasickej PCA je navyše vplyvom prítomnosti veľkosti obce alebo okresu veľký rozdiel medzi vysvetľovacou silou prvého hlavného komponentu.

Dôležitým pozorovaním je fakt, že logcontrast PCA kladie často veľký dôraz na strany alebo kandidátov, ktorí mali nízky volebný zisk. Toto je pravdepodobne dôsledkom toho, že logcontrast PCA pracuje s dátami upravenými pomocou logaritmov. Teda kým v prípade klasickej a hrubej PCA je dôležitý rozdiel medzi volebným ziskom kandidátov v rôznych okresoch alebo obciach, pri logcontrast PCA je dôležitý rozdiel zlogaritmovaných výsledkov, čo sa dá vysvetliť aj tak, že dôležitú úlohu zohráva nie rozdiel, ale podiel medzi jednotlivými výsledkami.

Kandidáti s nízkym volebným ziskom, ktorí sa javia ako dôležití v prípade logcontrast PCA, výsledok volieb výrazne neovplyvnili. Autori [4] vo svojom článku zdôrazňujú, že pri analýze dát je potrebné dôkladne zvážiť, či je analyzovaný problém iba kompozičný a či je skutočne potrebné používať špeciálne techniky, ako napríklad

logcontrast PCA. V prípade volebných výsledkov nie je dôležité iba to, aký podiel v niektorom okrese kandidát získal, pretože okresy nie sú rovnako veľké. Nerozhoduje teda iba to, v koľkých okresoch kandidát získal najvyšší počet hlasov, ale tiež veľkosť týchto okresov. Logcontrast PCA preto môže viesť k odhaleniu zaujímavých vzťahov medzi málo úspešnými kandidátmi, no ak naším hlavným záujmom je opis štruktúry výsledkov a nájdenie smerov, ktoré výsledky najviac ovplyvňujú, klasická a hrubá PCA sa javia ako presnejší nástroj.

5 Aplikácia FA na výsledky volieb

Druhou často používanou viacrozmernou štatistickou metódou, ktorú sme aj my aplikovali na volebné dáta, je faktorová analýza. Opäť je potrebné urobiť pred analýzou a počas nej niekoľko rozhodnutí. V otázke použitia kovariančnej alebo korelačnej matice ostávame pri rozhodnutí použiť maticu kovariančnú z dôvodov spomenutých už v predchádzajúcej kapitole. Pre prvú fázu faktorovej analýzy, teda prvý odhad faktorových váh, sme použili metódu hlavných komponentov.

Počet faktorov sme tentokrát vybrali priamo vzhľadom na výsledky politologických prieskumov. Podľa politológov je volebné správanie občanov SR ovplyvňované viacerými faktormi, medzi iným predovšetkým národnosťou, mestským alebo vidieckym prostredím, vzdelaním, vierovyznaním či sociálnou situáciou (napr. [11]). Pracovali sme teda so štyrmi a piatimi faktormi. Pri práci s nerotovanými faktormi získanými pomocou metódy hlavných komponentov práca s viacerými faktormi znamená len pridanie nového faktora k už získaným faktorom, ktoré sa takýmto rozšírením nezmenia. Ak však pracujeme s rotovanými faktormi, pridanie ďalšieho môže charakter predchádzajúcich faktorov výrazne zmeniť.

Na začiatok je ešte vhodné poznamenať, že významnou časťou faktorovej analýzy je aj interpretácia nájdených faktorov, ktorá je pomerne subjektívna. Rôzni ľudia môžu pri použití faktorovej analýzy dospieť k rôznym výsledkom. Ak jeden človek v nájdených faktoroch vidí nejakú interpretáciu, nemusí s tým ďalší súhlasiť. V neposlednom rade výsledné faktory môžu vyjsť každému používateľovi FA iné, vzhľadom na rôzne pravidlá výberu počtu faktorov, niekoľko možností hľadania počiatočných faktorov a tiež niekoľko možností ďalšej práce s faktormi formou rotácie.

5.1 Voľby do NR SR 2010

Voľby do NR SR v roku 2010 sme analyzovali už v predchádzajúcej kapitole pomocou analýzy hlavných komponentov, teda v tejto časti môžeme čiastočne využiť už získané výsledky. Pri analýze pomocou PCA už prvé dva hlavné komponenty vysvetľovali viac ako 87% celkovej variancie, môžeme teda bez obáv povedať, že 4 faktory budú vyčerpávať informáciu v dátach dostatočne. V skutočnosti pokrývajú viac ako 98% variability

dát, čo je už veľmi vysoký podiel. Odhad váh pre jednotlivé faktory získaný pomocou metódy hlavných komponentov je v Tabuľke 7.

Polit. strana	Faktor 1	Faktor 2	Faktor 3	Faktor 4
EDS	33,60	34,49	-22,96	41,07
Únia	142,72	20,33	71,67	37,91
SRK	8,15	9,57	-13,03	53,59
Paliho Kapurková	106,34	14,65	32,89	22,07
SaS	**2 517,77**	460,96	**1 283,08**	-253,39
SDĽ	396,97	49,60	18,27	-4,08
SMK - MKP	-526,20	**3 553,97**	-701,76	35,97
ĽS - HZDS	579,01	18,15	-243,22	-22,66
KSS	109,62	-3,23	-32,45	-9,71
SNS	**1 055,30**	-3,87	-310,57	-221,78
ND	57,95	4,69	22,54	5,14
ZRS	26,48	-4,96	-11,08	-2,90
KDH	**1 559,04**	-210,11	235,30	**1 279,39**
ĽSNS	168,07	-9,68	-19,04	37,47
SDKÚ - DS	**3 063,21**	990,69	**2 374,82**	-95,24
AZEN	14,22	6,75	3,55	-1,87
SMER-SD	**6 440,60**	-160,21	**-1 678,94**	-124,09
MOST - HíD	-198,07	**5 027,62**	-133,04	65,49
% F var	53,59	34,06	9,42	1,56
% kumul	53,59	87,65	97,07	98,63

Tabuľka 7: Odhad váh faktorov pre voľby do NR SR 2010

Pokúsime sa teraz nájdené faktory interpretovať. V tejto fáze faktorovej analýzy sme si pomohli štúdiou o profile voličov jednotlivých strán uverejnenou na stránke Inštitútu pre verejné otázky ([5]).

Prvý faktor má výrazne vysoké váhy pre všetky strany s vysokým volebným ziskom, kladné sú tieto váhy pre všetky strany okrem dvoch maďarských. Nazvime ho preto faktorom slovenského obyvateľstva.

Obrátenú situáciu môžeme pozorovať v druhom faktore. Vysoké kladné váhy majú maďarské strany SMK a Most-Híd, záporné sú tu váhy predovšetkým pri nacionálnych stranách. Všeobecne je známe, že v okresoch s vysokým podielom maďarského obyvateľstva väčšinou narastá volebný zisk maďarských strán a klesá zisk predovšetkým strán nacionálnych spolu so stranou SMER. Preto tento faktor nazveme faktorom maďarského obyvateľstva, ako dôležitá tu ale vystupuje pravdepodobne aj Bratislava.

V treťom faktore výrazne vystupujú strany SaS a SDKÚ s kladnými váhami, oproti

nim však SMER so zápornou váhou. Väčšina ostatných strán má tiež záporné váhy, teda záporne korelujú s týmto faktorom. Aj strany s kladnými váhami majú tieto váhy veľmi nízke oproti spomínaným SaS a SDKÚ, ktoré sú zvyčajne volené mestským obyvateľstvom s vyšším vzdelaním. Kladná váha pri KDH, ktoré má voličov predovšetkým v radoch vidieckeho obyvateľstva, nás napokon presvedčila k označeniu faktora ako vzdelanostného, keďže podľa [5] sa prípade KDH posilňuje segment voličov s vyšším vzdelaním.

Posledný vybraný faktor má jednoznačne najvyššiu váhu pri KDH, čo nás viedlo k tomu, aby sme ho označili ako faktor katolíckeho kresťanstva.

Podobne ako pri PCA, aj v tejto analýze sme sa rozhodli pre porovnanie využiť aj volebné dáta zozbierané podľa obcí. Výsledok faktorovej analýzy aplikovanej na túto väčšiu sadu dát sa nachádza v Tabuľke 8.

Polit. strana	Faktor 1	Faktor 2	Faktor 3	Faktor 4
EDS	5,82	0,36	-1,30	0,87
Únia	26,15	2,39	4,83	4,84
SRK	1,97	0,01	-0,93	1,15
Paliho Kapurková	20,30	0,72	1,38	2,29
SaS	**485,92**	48,24	**61,83**	-24,25
SDĽ	68,40	-2,36	-1,70	-1,77
SMK - MKP	18,97	**127,37**	**-104,97**	8,62
ĽS - HZDS	95,65	-12,81	-12,57	0,86
KSS	20,11	-3,37	-3,07	0,26
SNS	137,02	-24,19	-19,92	-19,19
ND	9,97	-0,39	0,75	0,27
ZRS	4,98	-1,27	-0,64	-0,70
KDH	**242,89**	-14,46	22,56	**90,73**
ĽSNS	33,93	-6,17	-3,73	1,89
SDKÚ - DS	**652,28**	**136,71**	**132,62**	-3,78
AZEN	3,47	0,42	0,50	-0,26
SMER-SD	**925,99**	**-153,10**	**-108,16**	-6,30
MOST - HÍD	150,02	**234,35**	**-105,27**	2,03
% F var	89,38	6,37	3,08	0,51
% kumul	89,38	95,75	98,83	99,34

Tabuľka 8: Odhad váh faktorov pre voľby do NR SR 2010 za obce

Nápadnou črtou prvého faktora je, že pre každú stranu je jeho váha kladná. Najvyššie váhy vystupujú pri stranách s najvyšším volebným ziskom. Preto tento faktor nazveme faktorom veľkosti obce.

Druhý faktor má kladné váhy hlavne pri maďarských stranách, pridáva sa tiež SDKÚ-DS. Najzápornejšiu váhu nájdeme pri strane SMER. Preto rovnako ako v prípade analýzy dát podľa okresov tento faktor označíme ako maďarský. Opäť ale možno za dôležitú považovať aj Bratislavu.

Tretí a štvrtý faktor majú podobnú štruktúru ako v prípade okresov. Najvyššie váhy tretieho faktora pri SDKÚ-DS a SaS poukazujú na dôležitosť vzdelania, štvrtý faktor je výrazný najmä pri KDH, čo naznačuje význam kresťanského obyvateľstva.

Okrem prvého faktora je teda štruktúra faktorových váh vypočítaných z volebných výsledkov podľa obcí veľmi podobná tej z dát podľa okresov. Rozdiel v prvom faktore spočíva iba v rozdielnom znamienku váh prislúchajúcich maďarským stranám. V prípade okresov teda rozhoduje predovšetkým počet Slovákov v danom okrese, pri obciach je však významný celkový počet ľudí.

5.2 Voľby do NR SR 2010 a 2012

Posledné dve voľby poslancov NR SR sa konali neobvykle krátko po sebe, delili ich iba dva roky. Obe boli spoločensky veľmi diskutované, rozhodli sme sa preto spojiť tieto dva súbory dát do jedného, nájsť faktory a pokúsiť sa ich interpretovať. Zaujímalo nás, čo najviac ovplyvnilo volebný zisk predovšetkým tých najúspešnejších strán a tiež, či a ako sa líšia medzi sebou váhy pri jednotlivých faktoroch pre strany kandidujúce v oboch rokoch. Prvý odhad faktorových váh pomocou metódy hlavných komponentov priniesol váhy, ktoré pre nás neboli jednoducho interpretovateľné, preto sme použili tiež varimax rotáciu faktorov. MATLAB umožňuje vykonať túto rotáciu príkazom *rotatefactors*. Výsledok uvádzame v Tabuľke 9. Vo vrchnej časti tabuľky sú uložené strany kandidujúce vo voľbách do NR SR v roku 2010, spodná časť tabuľky je venovaná voľbám v roku 2012. Ako je vidno v poslednom riadku tabuľky, päť faktorov vyčerpáva viac ako 99% celkovej variancie, teda v tomto smere je výsledok analýzy viac než uspokojivý.

Tieto faktory sa najprv pokúsime zinterpretovať a potom porovnáme váhy pri jednotlivých faktoroch pre strany vystupujúce v oboch voľbách.

Prvý faktor má podobne ako pri FA pre samotný rok 2010 vysoké váhy pri najúspešnejších stranách pre oba roky, a navyše všade okrem maďarských strán sú tieto váhy kladné, preto prvý faktor opäť pomenujeme faktorom slovenského obyvateľstva.

Pri druhom faktore si môžeme všimnúť vysoké váhy predovšetkým pri maďarských stranách, záporné váhy sú najvýraznejšie pri KDH a SMER-SD. Predovšetkým váhy pri maďarských stranách nás viedli k označeniu faktora ako faktora maďarského obyvateľstva.

Tretí faktor má kladné váhy pri veľkej väčšine všetkých strán v oboch rokoch. Výrazne tu vystupujú predovšetkým strany s najvyšším volebným ziskom. S ohľadom na túto skutočnosť sme ho nazvali faktorom veľkosti okresu.

Výrazne vysoké váhy pre štvrtý faktor vystupujú pri stranách KDH a SMER-SD v prípade oboch rokov, čo by mohlo navádzať na vysvetlenie faktora ako vidieckeho. Túto hypotézu potvrdzujú tiež záporné váhy SaS ako strany volenej predovšetkým mestským obyvateľstvom. Váhy pri ďalšej mestskej strane, SDKÚ-DS, sú síce kladné, no pomerne nízke. Ďalšou možnosťou na vysvetlenie tohto faktora však je vzdelanie, s vyšším faktorovým skóre v prípade nižšieho vzdelania.

Zaujímavou výzvou na interpretáciu je piaty faktor. Jeho váhy sú v oboch rokoch vysoké pre KDH ako kresťanskú stranu. V roku 2010 výrazne vystupuje tiež pri SaS, v roku 2012 pri OĽaNO. OĽaNO je strana volená tiež predovšetkým obyvateľstvom hlásiacim sa ku kresťanskej viere. V roku 2010 ešte vo voľbách do NR SR nekandidovala, no niekoľko jej predstaviteľov vystupovalo na kandidátnej listine SaS, čo aj strane SaS zabezpečilo pomerne výraznú podporu kresťanských voličov. Preto tento faktor označíme za faktor katolíckeho kresťanstva.

Keď už máme faktory interpretované, môžeme sa pozrieť na ich vplyv na volebný zisk jednotlivých strán v rôznych rokoch. Veľmi výrazný rozdiel vidno hneď v prvom faktore pri strane SDKÚ-DS. Zatiaľ čo v roku 2010 patrila medzi strany s najvyššou váhou pre tento faktor, o dva roky neskôr je táto váha menej než tretinová v porovnaní s rokom 2010. Aj v treťom faktore váha pri SDKÚ klesla, o viac než polovicu. Tieto dva faktory sme vysvetlili ako faktor slovenského obyvateľstva a veľkosti okresu. Preto výraznú zmenu váh možno vysvetliť faktom, že tejto strane v roku 2012 výrazne klesla volebná podpora. Po páde vlády v roku 2012 získala strana iba menej ako polovicu počtu hlasov z roku 2010.

Váha tretieho faktora klesla aj pri SaS. Tento pokles je zrejme dôsledkom už spomínaného faktu, že v roku 2010 figurovali na kandidátnej listine strany aj ľudia, ktorí v

roku 2012 už kandidovali za novovzniknutú stranu OĽaNO. Rozdiel v piatom faktore založený na tej istej skutočnosti už sme spomenuli pri interpretácii piateho faktora.

5.3 Zhrnutie výsledkov FA na volebných dátach

Výsledky volieb, predovšetkým do NR SR, predstavujú významný náhľad na štruktúru a názory obyvateľstva. Faktorová analýza poskytuje jeden z možných spôsobov, ako zistiť, čo správanie voličov najviac ovplyvňuje.

Ako ukazuje faktorová analýza pre samotný rok 2010 aj pre kombináciu dát z rokov 2010 a 2012, správanie slovenských voličov skutočne možno popísať predovšetkým niekoľkými faktormi, ktoré za dôležité považujú aj politológovia. Ako najvýznamnejšie sa ukazujú veľkosť okresu či obce, podiel maďarského obyvateľstva, vzdelanie, ale aj vierovyznanie. Ich významnosť pre rôzne strany sa môže meniť, predovšetkým na základe zmien volebných preferencií, ktoré často vplyvom krokov vlád nastávajú v pomerne krátkom čase a môžu sa prejaviť aj v dvoch po sebe idúcich voľbách. Tento fakt podporuje výsledok analýzy na kombinovaných dátach pre roky 2010 a 2012.

Ako bolo spomenuté v úvode tejto kapitoly, interpretácia faktorov, ktorú sme uviedli, je subjektívna, založená predovšetkým na výsledkoch prezentovaných v [5], a tiež na všeobecných vedomostiach o politických subjektoch a ich voličskej základni. Nájdené faktory môžu mať aj niekoľko rôznych interpretácií a je zložité rozhodnúť, ktorá z nich je najlepšia. Uvádzame také vysvetlenie faktorov, ktoré z nášho pohľadu najviac vyhovuje nájdenej štruktúre.

Polit. strana	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5
EDS	27,40	34,49	7,01	54,72	5,83
Únia	61,86	-3,12	141,61	51,94	27,50
SRK	0,30	10,08	-5,21	60,83	10,17
Paliho Kapurková	58,32	0,62	90,16	37,56	12,69
SaS	**1363,75**	84,98	**2490,88**	-143,73	**579,94**
SDĽ	307,92	20,85	244,45	54,02	66,47
SMK - MKP	-177,01	**3660,08**	-246,18	9,22	84,98
ĽS - HZDS	576,85	13,99	166,39	184,63	-113,45
KSS	108,61	-4,09	34,96	15,49	9,05
SNS	**1096,82**	-8,08	354,27	-70,38	83,41
ND	37,65	-1,56	46,31	-1,22	44,29
ZRS	29,38	-4,09	3,89	1,86	6,61
KDH	**861,83**	-366,54	**992,50**	**1313,71**	**746,44**
ĽSNS	135,53	-17,25	74,74	76,07	33,20
SDKÚ - DS	**1020,00**	342,71	**3842,05**	160,04	290,96
AZEN	8,99	5,05	11,97	3,12	-1,68
SMER-SD	**6005,04**	-317,35	**2449,23**	**1450,62**	-149,31
MOST - HÍD	-399,10	**4955,53**	742,22	199,08	-222,24
Zelení	23,00	12,94	29,26	17,58	7,99
KDH	**834,45**	-295,76	**1152,98**	**1232,65**	**628,22**
SDĽ	14,04	-1,84	7,17	22,84	-2,64
SNS	**986,21**	-43,71	361,75	-71,98	75,85
OĽaNO	**1124,23**	-114,31	**1425,72**	192,92	**905,10**
SaS	392,48	25,24	**1792,01**	-85,83	-12,06
PaS	48,10	6,40	59,67	22,80	15,98
NÁŠ KRAJ	-5,83	-10,50	7,85	63,74	1,67
SZ	40,83	1,28	84,70	-2,95	-0,50
ĽSNS	166,43	-26,93	122,65	70,54	69,27
SMER - SD	**7703,48**	-398,77	**3169,69**	**1962,00**	-48,83
Zmena zdola, DÚ	66,07	8,94	364,29	21,76	-19,07
NaS - ns	117,49	-17,92	53,42	4,57	13,99
KSS	92,23	-1,76	43,83	2,75	13,97
SRÚS	3,08	2,57	-4,44	32,75	-11,09
MOST - HÍD	-234,45	**3945,31**	767,35	103,37	-286,03
99%	152,26	48,11	150,25	72,44	33,05
ĽS - HZDS	108,79	7,39	19,39	60,94	-9,94
+1 HLAS	1,46	0,59	7,20	1,68	-1,10
SF	33,44	0,88	67,08	2,26	4,56
Obyčajní ľudia	20,76	-0,14	24,06	5,45	15,16
SDKÚ - DS	293,41	14,65	**1748,05**	70,93	54,28
SOSKA	11,21	-3,19	28,75	5,14	12,19
SMK - MKP	-244,04	**3797,42**	-254,41	9,14	116,41
SSS - NM	75,86	-2,62	143,21	188,01	33,71
SŽS	11,18	4,00	12,15	8,24	1,08
% F var	60,26	29,26	7,56	1,39	0,49
% kumul	60,26	89,52	97,45	98,94	99,47

Tabuľka 9: Odhad váh faktorov pre voľby do NR SR 2010 a 2012

Záver

Cieľom bakalárskej práce bolo pochopiť viacrozmerné štatistické metódy na analýzu dát založené na spektrálnych metódach lineárnej algebry, analýzu hlavných komponentov a faktorovú analýzu, a použiť ich na výsledky volieb v SR od roku 1990. Na základe istých špeciálnych vlastností volebných výsledkov (nezáporné čísla, dôležité sú aj podiely na celkovom počte hlasov) sme sa venovali aj modifikáciam PCA pre tzv. kompozičné dáta, ktoré ako jeden z prvých začal veľmi podrobne študovať J. Aitchinson v [3].

V prvých troch teoretických kapitolách práce sme popísali teoretické základy a fungovanie metód PCA a FA. Prvá kapitola spracovaná predovšetkým na základe [10] je venovaná PCA a tiež jej spojitosti so singulárnym rozkladom matice. Druhá kapitola začína budovaním teórie kompozičných dát, ďalej v náväznosti na prvú kapitolu sú popísané dve modifikácie PCA, ktoré boli J. Aitchinsonom v [3] navrhnuté na analýzu kompozičných dát - hrubá a logcontrast PCA. V tretej kapitole sme s pomocou odbornej literatúry ([9],[17]) vysvetlili základné princípy FA.

Vo veľmi obsiahlej štvrtej kapitole sme prezentovali výsledky aplikácie troch skôr opísaných druhov PCA na výsledky niekoľkých volieb na Slovensku. Porovnali sme výsledky medzi sebou a tiež so všeobecnými znalosťami o štruktúre obyvateľstva SR a politologickými výskumami. Na základe týchto výsledkov sme usúdili, že ak našou snahou je popísať štruktúru výsledkov a nájsť smery najviac ovplyvňujúce výsledky volieb, pozerať sa na volebné výsledky ako na kompozičné dáta zrejme nie je najvhodnejší spôsob. Napokon, sám Aitchinson spolu s Egozcuem v [4] zdôrazňuje, že je vždy potrebné zvážiť, či dáta, s ktorými pracujeme, sú skutočne iba kompozičné.

Pri snahe nájsť najvýraznejšie smery najuspokojivejší výsledok, zhodujúci sa s politologickými štúdiami (napr. [5]), poskytuje základná metóda PCA, no aj hrubá PCA. Naproti tomu logcontrast PCA identifikuje ako dôležité predovšetkým menej významné prvky, teda kandidátov na prezidenta či politické strany s veľmi nízkym volebným ziskom. Táto metóda je teda vhodná v prípade, kedy hľadáme neobvyklé vzťahy medzi málo úspešnými kandidátmi. Prvky s malými podielmi na celku môžu predstavovať dôležitú časť informácie napríklad v geológii, ktorá bola jedným z hlavných dôvodov pre predstavenie metódy logcontrast PCA, no v prípade volebných dát zvyknú byť na okraji záujmu.

V poslednej kapitole sa venujeme aplikácii faktorovej analýzy na volebné dáta a intepretácii faktorov. Ukazuje sa, že dôležitými faktormi sú okrem iného veľkosť okresu či obce, národnosť, vzdelanie alebo tiež vierovyznanie, čo sa zhoduje s názormi politológov.

Prínosom pre autorku bolo nadobudnutie nových poznatkov o využití metód lineárnej algebry v štatistike a tiež o kompozičných dátach a možnostiach ich analýzy. Štúdium kompozičných dát prináša aj mnoho informácií o simplexe ako vektorovom priestore a operáciách na ňom, pričom zaujímavé sú analógie medzi operáciami súčtu a násobenia skalárom v reálnom vektorovom priestore a operáciami perturbácie a mocninnej transformácie na simplexe. Definícia týchto operácií na simplexe navyše ponúka celkom prirodzený prechod od simplexu k reálnemu vektorovému priestoru pomocou logaritmu.

Metódy lineárnej algebry a tiež PCA a FA je možné na výsledky volebných dát použiť niekoľkými spôsobmi a táto práca určite neobsiahla všetky. Pomocou singulárneho rozkladu matice výsledkov volieb je napríklad možné pokúšať sa z čiastočných výsledkov zverejňovaných počas volebnej noci odhadnúť výsledný stav (viď http://thales.doa.fmph.uniba.sk/niepel/LA/volby.html). Toto by mohlo byť zaujímavé a správne predovšetkým v prípade, kedy je možné racionálne predpokladať, že štruktúra a názory obyvateľstva sa príliš nezmenili. Z dôvodu zmien množstva kandidátov je však lepšie miesto matice $\mathbf{V}$ zo singulárneho rozkladu využiť maticu $\mathbf{U}$ zodpovedajúcu okresom alebo obciam.

Ďalej by tieto metódy mohli ponúknuť spôsob, ako vypočítať presuny voličov medzi stranami. Politológovia a internetové portály po voľbách často uverejňujú tabuľky ukazujúce, koľko percent voličov nejakej strany v predchádzajúcich voľbách volilo nejakú inú stranu vo voľbách aktuálnych (napríklad http://www.sme.sk/c/7153578/ku-komu-presli-volici-knazka-a-prochazku-volebne-grafy.html popisuje presuny voličov medzi prvým a druhým kolom prezidentských volieb 2014). Takéto tabuľky sa dajú pravdepodobne získať aj pomocou lineárnej algebry, jedným z kandidátov na metódy, ktoré by sa v tomto prípade dali využiť, je faktorová analýza.

Prvé hlavné komponenty vysvetľujú najvýznamnejšie zdroje variability v dátach. Ostávajú však otvorené ďalšie otázky, ako napríklad aký je vplyv účasti na výsledky, či je možné pomocou štatistickej analýzy odhaliť nejaké trendy alebo naopak zlomy vo

volebných dátach, alebo ktorá z dvoch maďarských strán vystupujúcich na súčasnej slovenskej politickej scéne bude mať v dlhodobom horizonte viac voličov. Odpovede na niektoré z týchto otázok by mohli odhaliť aj hlavné komponenty prislúchajúce menším vlastným hodnotám výberovej kovariančnej matice.

Jednou z ďalších možností ako aplikovať použité metódy na volebné dáta je tiež spojiť niektoré strany (prípadne kandidátov na prezidenta) do jednej skupiny a skúmať takto vytvorenú novú sadu dát. Predovšetkým v prípade malých strán, ktoré majú často iba regionálnych voličov, takýto krok môže výrazne zmeniť ich variabilitu. Takáto úprava dát by však bola značne subjektívna, pretože spôsobov, ako vytvoriť nové skupiny premenných, je veľmi veľa.

V oblasti kompozičných dát stále prebieha výskum a štatistici uverejňujú nové články, postupy a možnosti interpretácie. V tomto smere je ešte veľa nepreskúmaných možností a nezodpovedaných otázok. V tejto práci popísané metódy na analýzu kompozičných dát (hrubá a predovšetkým logcontrast PCA) sú iba prvými návrhmi J. Aitchinsona, ktoré sa dodnes používajú, no odvtedy boli uverejnené ďalšie metódy predovšetkým robustných analýz a štatistici (napr. Filzmoser, Hron, Reimann) stále pracujú na nových metódach alebo na modifikáciách už skôr uverejnených postupov.

Zoznam použitej literatúry

[1] Aitchinson, J.: *A Concise Guide to Compositional Data*, elektronická publikácia, 2003, dostupné na internete (16.5.2014): http://www.leg.ufpr.br/lib/exe/fetch.php/pessoais:abtmartins:a_concise _guide_to_compositional_data_analysis.pdf

[2] Aitchinson, J.: *Conditional Compositional Biplots: Theory and Application*, prednáška workshopu CoDa Work '05, Girona, 2005, dostupné na internete (9.12.2013): http://ima.udg.edu/activitats/codawork05/CD/Session1/Aitchinson-Ng.pdf

[3] Aitchinson, J.: *The Statistical Analysis of Compositional Data*, Chapman and Hall Ltd., London, 1986

[4] Aitchinson, J., Egozcue J.J.: *Compositional Data Analysis: Where Are We and Where Should We Be Heading?*, Mathematical Geology 37 (2005), 829-850

[5] Bútorová, Z., Gyárfášová, O., Krivý, V.: *Výskum voličského správania na Slovensku*, Inštitút pre verejné otázky, Bratislava, 2010, dostupné na internete (16.5.2014): http://www.ivo.sk/buxus/docs//Parlamentne _volby_2010/tlacova_sprava_povolebny_vyskum.pdf

[6] Egozcue, J.J. , Pawlowsky-Glahn, V., Mateu-Figueraz, G., Barceló-Vidal, C.: *Isometric logratio transformations for compositional data analysis*, Mathematical Geology 37 (2003), 279-300

[7] Filzmoser, P., Hron, K., Reimann, C.: *Principal Component Analysis for Compositional Data with Outliers*

[8] Izenman, A.J.: *Modern Multivariate Statistical Techniques*, Springer Science, Business Media, New York, 2008

[9] Johnson, R.A., Wichern, D.W.: *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey, 1998

[10] Jolliffe, I. T.: *Principal Component Analysis*, Springer-Verlag, New York, 2002

[11] Krivý, V., Feglová, V., Balko, D.: *Slovensko a jeho regióny: Sociokultúrne súvislosti volebného správania*, Nadácia Médiá, Bratislava, 1996

[12] Lincoln, S. V., Piepe, A., Prior, R.: *An Application of Principal Component Analysis to Voting in Scottish Municipal Elections 1967-9*, Journal of the Royal Statistical societz, Series D (The Statistician) 20 (1971), 73-88, dostupné na internete (9.12.2013):
http://www.jstor.org/stable/2986987

[13] Pawlowsky-Glahn, V., Buccianti, A.: *Compositional Data Analysis: Theory and Applications*, John Wiley & Sons, West Sussex, 2011

[14] Rodrigues, P. C., Lima, A. T.: *Analysis of an European Union Election Using Principal Component Analysis*, Statistical Papers 50 (2009), 895-904, dostupné na internete (9.12.2013): http://link.springer.com/article/10.1007%2Fs00362-009-0264-2

[15] Strang, G.: *Linear Algebra and its Applications*, Thomson Learning, USA, 1988

[16] Štatistický úrad Slovenskej republiky, dostupné na internete (10.5.2014): http://www.statistics.sk

[17] Tryfos, P.: *Methods for Business Analysis and Forecasting: Text & Cases*, John Wiley & Sons, 1998, kapitola 14 *Factor Analysis* dostupná na internete (28.4.2014): http://www.yorku.ca/ptryfos/f1400.pdf

Príloha A - Zdrojové kódy

Získanie hlavných komponentov

```
function [klas, klasvar, klasPC, crude, crudevar, crudePC, clr, clrvar, clrPC] = analyza(data)
M1 = repmat(mean(data), size(data, 1), 1);
data2 = data − M1;
[klasPC, klasvar, klas] = svd(data2);
cdata = composition(data);
M2 = repmat(mean(cdata), size(data, 1), 1);
cdata2 = cdata − M2;
[crudePC, crudevar, crude] = svd(cdata2);
ldata = loggeo(nula(cdata));
M3 = repmat(mean(ldata), size(data, 1), 1);
ldata2 = ldata − M3;
[clrPC, clrvar, clr] = svd(ldata2);
```

Vytvorenie kompozície

```
functionA = composition(B)
A = B;
for i = 1 : length(A(:, 1))
  A(i, :) = A(i, :)/sum(B(i, :));
end
```

clr transformácia dát

```
functionB = loggeo(A)
B = A;
for i = 1 : length(A(:, 1))
  g = geomean(A(i, :));
  B(i, :) = A(i, :)/g;
  for j = 1 : length(B(1, :))
    B(i, j) = log(B(i, j));
  end
end
```

Nahradenie núl

```
functionB = nula(A)
B = A;
m = length(A(:, 1));
n = length(A(1, :));
for i = 1 : m
  pocetnul = 0;
  for j = 1 : n
    if A(i, j) == 0
      pocetnul = pocetnul + 1;
    end
  end
  eps = 0.00005 * (pocetnul + 1) * (n - pocetnul)/(n^2);
  for j = 1 : n
    if A(i, j) == 0
      B(i, j) = eps;
    else
      B(i, j) = A(i, j) - eps * pocetnul/(n - pocetnul);
    end
  end
end
```

Príloha B - Zoznam kandidátov a ich volebné výsledky

Skratka	Názov strany	Volebný výsledok
HSD-SMS	Hnutie za samosprávnu demokraciu - Spoločnosť pre Moravu a Sliezsko	0,13
HZSP-SRÚ	Hnutie za slobodu prejavu - Slovenská republikánska únia	0,07
HZDS	Hnutie za demokratické Slovensko	37,26
SDĽ	Strana demokratickej ľavice	14,70
SPI	Strana práce a istoty	0,97
HZOS	Hnutie za oslobodenie Slovenska	0,23
SSL-SNZ	Strana slobody - Strana národného zjednotenia	0,31
SKDH	Slovenské kresťansko-demokratické hnutie	3,05
MKM-EGY	Maďarské kresťanskodemokratické hnutie, Együttélés	7,42
HSS	Hnutie za sociálnu spravodlivosť	0,11
SZ	Strana zelených	1,08
KDH	Kresťanskodemokratické hnutie	8,89
ODÚ	Občianska demokratická únia	4,04
ZPR-RSČ	Združenie pre republiku - Republikánska strana Česko-Slovenska	0,33
NALI	Národní liberáli	0,08
SZS	Strana zelených na Slovensku	2,14
ROI	Rómska občianska iniciatíva	0,60
SDSS	Sociálnodemokratická strana na Slovensku	4,00
KSS	Komunistická strana Slovenska	0,76
DS-ODS	Demokratická strana - Občianska demokratická strana	3,31
SNS	Slovenská národná strana	7,93
SĽS	Slovenská ľudová strana	0,30
MPP-MOS	Magyar Polgári Párt - Maďarská občianska strana	2,29

Tabuľka 10: Zoznam a výsledky politických strán kandidujúcich vo voľbách do NR SR 1992

Skratka	Názov strany	Volebný výsledok
SZS	Strana zelených na Slovensku	0,98
SDKÚ	Slovenská kresťanská a demokratická únia	15,09
SDPO	Strana za demokratické práva občanov	0,23
SDĽ	Slovenská demokratická ľavica	1,36
SMER	SMER	13,46
HZDS	Hnutie za demokratické Slovensko	19,50
OKS	Občianska konzervatívna strana	0,32
HZD	Hnutie za demokraciu	3,28
ROMA	Politické hnutie Rómov na Slovensku - ROMA	0,21
KSS	Komunistická strana Slovenska	6,32
SMK-MKP	Strana maďarskej koalície - Magyar Koalíció Pártja	11,16
KDS	Kresťanskodemokratické hnutie	8,25
ĽS	Ľudová strana	0,02
ZRS	Združenie robotníkov Slovenska	0,54
ĽB	Ľudový blok	0,22
ANO	Aliancia nového občana	8,01
B-RRS	Béčko - Revolučná robotnícka strana	0,09
ŽAR	Žena a rodina	0,43
SDA	Sociálnodemokratická alternatíva	1,79
SNJ	Slovenská národná jednota	0,15
NONSP	Nezávislá občianska strana nezamestnaných a poškodených	0,91
SNS	Slovenská národná strana	3,32
ROSA	Robotnícka strana ROSA	0,30
ROISR	Rómska občianska iniciatíva SR	0,29
P SNS	Pravá Slovenská národná strana	3,65

Tabuľka 11: Zoznam a výsledky politických strán kandidujúcich vo voľbách do NR SR 2002

Skratka	Názov strany	Volebný výsledok
EDS	Európska demokratická strana	0,40
Únia	Únia - Strana pre Slovensko	0,70
SRK	Strana rómskej koalície	0,27
Paliho Kapurková	Paliho Kapurková	0,57
SaS	Sloboda a Solidarita	12,14
SDĽ	Strana demokratickej ľavice	2,41
SMK-MKP	Strana maďarskej koalície - Magyar Koalíció Pártja	4,33
ĽS-HZDS	Ľudová strana - Hnutie za demokratické Slovensko	4,32
KSS	Komunistická strana Slovenska	0,83
SNS	Slovenská národná strana	5,07
ND	Nová demokracia	0,31
ZRS	Združenie robotníkov Slovenska	0,24
KDH	Kresťanskodemokratické hnutie	8,52
ĽSNS	Ľudová strana Naše Slovensko	1,33
SDKÚ-DS	Slovenská demokratická a kresťanská únia - Demokratická strana	15,42
AZEN	AZEN - Aliancia za Európu národov	0,13
SMER-SD	SMER - Sociálna demokracia	34,79
MOST-HíD	MOST-HíD	8,12

Tabuľka 12: Zoznam a výsledky politických strán kandidujúcich vo voľbách do NR SR 2010

Meno kandidáta	Volebný výsledok
Ján Demikát	0,15
Juraj Lazarčík	0,52
Vladimír Mečiar	37,23
Ivan Mjartan	3,59
Rudolf Schuster	47,37
Ján Slota	2,50
Juraj Švec	0,81
Magda Vášaryová	6,60
Boris Zala	1,00
Nevoliteľní kandidáti	0,18

Tabuľka 13: Zoznam a výsledky kandidátov v prezidentských voľbách 1999