

UNIVERZITA KOMENSKÉHO V BRATISLAVE  
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



LINEÁRNA REDUKCIA DÁT POMOCOU  
SPEKTRÁLNEHO ROZKLADU KOVARIANČNEJ MATICE

BAKALÁRSKA PRÁCA

UNIVERZITA KOMENSKÉHO V BRATISLAVE  
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

**LINEÁRNA REDUKCIA DÁT POMOCOU  
SPEKTRÁLNEHO ROZKLADU KOVARIANČNEJ MATICE**

**BAKALÁRSKA PRÁCA**

Študijný program: Ekonomická a finančná matematika  
Študijný odbor: 1114 Aplikovaná matematika  
Školiace pracovisko: Katedra aplikovanej matematiky a štatistiky  
Vedúci práce: Mgr. Roman Kukumberg



Univerzita Komenského v Bratislave  
Fakulta matematiky, fyziky a informatiky

---

## ZADANIE ZÁVEREČNEJ PRÁCE

**Meno a priezvisko študenta:** Michaela Dobříková  
**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:** Lineárna redukcia dát pomocou spektrálneho rozkladu kovariančnej matice /  
*Linear dimensionality reduction of data using the spectral decomposition of covariance matrix*

**Cieľ:** Cieľom práce je naštudovať a spracovať štatistické metódy používané na lineárnu redukciu dimenzie dát ako analýza hlavných komponentov a kanonické korelácie, a následná aplikácia metód.

**Vedúci:** Mgr. Roman Kukumberg  
**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

**Pod'akovanie** Touto cestou sa chcem poďakovať svojmu vedúcemu bakalárskej práce Mgr. Romanovi Kukumbergovi za ochotu, pomoc, odborné rady a podnetné pripomienky, ktoré mi pomohli pri písaní tejto práce.

## Abstrakt

DOBŘÍKOVÁ, Michaela: Lineárna redukcia dát pomocou spektrálneho rozkladu kovariančnej matice [Bakalárska práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiťel: Mgr. Roman Kukumberg, Bratislava, 2014, 57s.

Práca sa zaoberá dvomi metódami na lineárnu redukciu dimenzie dát: analýzou hlavných komponentov (Principal Component Analysis, PCA) a analýzou kanonických korelácií (Canonical Correlation Analysis, CCA). Hlavnou ideou PCA je zredukovať dimenziu dát pomocou transformácie na nové premenné, tzv. hlavné komponenty a zachovať pri tom väčšinu informácie v dátach. Analýza kanonických korelácií sa zaoberá vzťahmi medzi dvomi vektorovými premennými, hľadá ich najviac korelované lineárne kombinácie. Obe tieto metódy sú vo svojej podstate problémami na vlastné čísla a vlastné vektory, preto sme najprv zhrnuli dôležité poznatky o vlastných číslach a vlastných vektoroch. Následne sme každej z metód venovali jednu kapitolu, pričom sme najprv uviedli podrobné odvodenie metód a následne sme ich aplikovali, overili ich fungovanie v praxi.

**Kľúčové slová:** analýza hlavných komponentov, kanonické korelácie, vlastné čísla, vlastné vektory, redukcia dimenzie

# Abstract

DOBŘÍKOVÁ, Michaela: Linear Dimensionality Reduction of Data Using the Spectral Decomposition of Covariance Matrix [Bachelor thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Roman Kukumberg, Bratislava, 2014, 57p.

This thesis deals with two methods for linear dimensionality reduction of data: Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA). The main idea of PCA is to reduce data dimension by transforming them to new variables called principal components, so that they retain most of the variation in the data. CCA studies relationships between two vector variables, it searches for their most correlated linear combinations. As these methods are based on eigenvalues and eigenvectors, in the first chapter we summarize important definitions and theorems about eigenvalues and eigenvectors. The next two chapters describe the methods, at first we introduce derivation of each method which is followed by its applications.

**Key words:** Principal component analysis, canonical correlation, eigenvalues, eigenvectors, dimensionality reduction

# Obsah

|                                                                       |           |
|-----------------------------------------------------------------------|-----------|
| <b>Úvod</b>                                                           | <b>8</b>  |
| <b>1 Vlastné čísla a vlastné vektory</b>                              | <b>10</b> |
| <b>2 Analýza hlavných komponentov</b>                                 | <b>15</b> |
| 2.1 Odvodenie metódy a definícia . . . . .                            | 15        |
| 2.1.1 Odvodenie metódy . . . . .                                      | 16        |
| 2.1.2 Definícia hlavných komponentov a výberové hlavné komponenty     | 19        |
| 2.1.3 PCA a korelačná matica . . . . .                                | 21        |
| 2.1.4 Koľko hlavných komponentov ponechať? . . . . .                  | 21        |
| 2.1.5 Príklad použitia PCA . . . . .                                  | 23        |
| 2.2 Aplikácie PCA . . . . .                                           | 25        |
| 2.2.1 Spracovanie obrazu . . . . .                                    | 25        |
| 2.2.2 Ceny akcií . . . . .                                            | 28        |
| <b>3 Kanonické korelácie</b>                                          | <b>31</b> |
| 3.1 Odvodenie metódy a definícia . . . . .                            | 31        |
| 3.1.1 Odvodenie metódy . . . . .                                      | 32        |
| 3.1.2 Definícia kanonických premenných a výberové kanonické korelácie | 38        |
| 3.1.3 Príklad použitia CCA . . . . .                                  | 40        |
| 3.2 Aplikácie CCA . . . . .                                           | 43        |
| 3.2.1 Koeficient mnohonásobnej korelácie . . . . .                    | 43        |
| 3.2.2 CCA a NHL . . . . .                                             | 45        |
| <b>Záver</b>                                                          | <b>49</b> |
| <b>Zoznam použitej literatúry</b>                                     | <b>51</b> |
| <b>Príloha A</b>                                                      | <b>53</b> |

## Úvod

V štatistike sa často riešia problémy, ktoré si vyžadujú sledovanie veľkého počtu premenných. S takýmto množstvom dát sa však veľmi ťažko pracuje, preto môže byť užitočné tieto dáta nejakým spôsobom transformovať a sledovať len zredukovaný počet premenných. Jednou z možností ako to dosiahnuť je nájsť menej rozmerné projekcie vstupných dát, ktoré majú akési optimálne vlastnosti. Medzi metódy, ktoré niečo takéto umožňujú patria aj Analýza hlavných komponentov (Principal Component Analysis, PCA) a Analýza kanonických korelácií (Canonical Correlation Analysis, CCA).

Obe tieto metódy sú úzko spojené s menom Harold Hotteling a sú postavené na teórii vlastných čísel a vlastných vektorov [7]. Analýza hlavných komponentov patrí medzi najstaršie metódy v mnohorozmernej analýze. Napriek tomu, že v minulosti nebola veľmi používaná, vďaka rozvoju počítačovej techniky je dnes súčasťou mnohých štatistických softwareov [8]. Možno ju použiť napríklad pri spracovaní obrazu [11], ale aj vo financiách [8], či pri rozoznávaní tvárí [7]. V PCA sa tak, ako je uvedené v [8], súbor dát transformuje na nový súbor náhodných premenných, nazvaných hlavné komponenty, ktoré sú navzájom nekorelované tak, aby si uchovali väčšinu informácie pôvodných dát. Analýza kanonických korelácií bola ešte donedávna relatívne neznáma štatistická metóda, no podobne ako analýza hlavných komponentov s vývojom rôznych počítačových programov sa dnes využíva čoraz častejšie [3]. V [9] využili autori CCA na skúmanie vzťahov medzi dopravou a ekonomickým vývojom, v [18] autori využili túto metódu pri skúmaní reči. Táto metóda je založená na skúmaní lineárnych vzťahov medzi dvomi vektorovými premennými, pričom jedna z nich môže, ale aj nemusí byť závislá premenná.

Cieľom bakalárskej práce bolo naštudovať a spracovať analýzu hlavných komponentov a analýzu kanonických korelácií ako príklad metód na lineárnu redukciu dimenzie dát. Následne sme sa snažili poukázať na široké využitie týchto metód naštudovaním a spracovaním ich aplikácií.

Naša práca je rozdelená na tri hlavné kapitoly. Obe nami skúmané metódy využívajú spektrálny rozklad matice, sú vo svojej podstate problémami na vlastné čísla a vlastné vektory. Z toho dôvodu je prvá kapitola venovaná niektorým poznatkom o vlastných číslach a vlastných vektoroch, ktoré môžu byť užitočné pre lepšie pochopenie spome-

nutých metód.

Druhá kapitola je venovaná analýze hlavných komponentov. Najprv uvidíme odvodenie metódy a definíciu hlavných komponentov, následne ukážeme možné aplikácie tejto metódy pri spracovaní obrazu a vo financiách.

Tretia kapitola je venovaná analýze kanonických korelácií. Opäť najprv uvidíme odvodenie metódy, ktoré bude nasledované aplikáciami tejto metódy. Tentokrát uvidíme jednu aplikáciu vo financiách a následne vlastnú aplikáciu na dáta spojené s národnou hokejovou ligou (NHL).

## 1 Vlastné čísla a vlastné vektory

Keďže obe metódy na redukciiu dimezie dát, ktorým sa venujeme v tejto práci, sú vo svojej podstate problémami na vlastné čísla a vlastné vektory, táto kapitola je venovaná práve vlastným vektorom a vlastným číslam. Zhrnieme základné definície, vzťahy a vety, ktoré môžu byť užitočné pre pochopenie týchto metód. Na spracovanie tejto kapitoly boli použité zdroje [16] a [20].

Teória vlastných čísel a vlastných vektorov je postavená na riešení rovnice

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}. \quad (1)$$

Rovnica (1) je nelineárna, keďže má dve neznáme: číslo  $\lambda$  a vektor  $\mathbf{x}$ . Túto rovnicu možno ekvivalentne prepísať na tvar

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}.$$

Ak by sme poznali neznámu  $\lambda$ , získali by sme lineárnu rovnicu v  $\mathbf{x}$ , pričom vidíme, že vektor  $\mathbf{x}$  patrí do nulového priestoru matice  $\mathbf{A} - \lambda\mathbf{I}$ . Je zrejmé, že takáto rovnica má triviálne riešenie  $\mathbf{x} = \mathbf{0}$ , nás však zaujímajú netriviálne riešenia tejto rovnice. Preto hľadáme také  $\lambda$ , pre ktoré je matica  $\mathbf{A} - \lambda\mathbf{I}$  singulárna a teda má netriviálny nulový priestor. Ak je matica singulárna, znamená to, že jej determinant je rovný nule.

**Definícia 1.1.** Číslo  $\lambda$  sa nazýva vlastným číslom štvorcovej matice  $\mathbf{A}$  práve vtedy, keď

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0. \quad (2)$$

Rovnicu (2) nazývame charakteristickou rovnicou matice  $\mathbf{A}$ . Pre každé riešenie  $\lambda$  tejto rovnice bude existovať nenulový vektor spĺňajúci

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}.$$

Takýto vektor nazveme vlastným vektorom matice  $\mathbf{A}$ .

Vlastné čísla sú teda koreňmi tzv. charakteristického polynómu  $\det(\mathbf{A} - \lambda\mathbf{I})$  štvorcovej matice  $\mathbf{A}$ . Ide o polynóm stupňa  $n$  s koeficientom  $(-1)^n$  pri najvyššej mocnine  $\lambda^n$ .

Charakteristický polynóm matice je polynóm stupňa  $n$ , ktorý môže teoreticky mať aj viacnásobné korene a teda matica  $\mathbf{A}$  môže mať viacnásobné vlastné čísla. Avšak v realite pri riešení príkladov založených na skutočných dátach očakávame, že získame rôzne vlastné čísla. Pre rôzne vlastné čísla platí nasledujúce tvrdenie:

**Tvrdenie 1.1.** *Nech  $\lambda_1, \lambda_2, \dots, \lambda_k$  sú navzájom rôzne vlastné hodnoty  $n \times n$  matice  $\mathbf{A}$ . Potom k nim prislúchajúce vlastné vektory  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$  sú lineárne nezávislé.*

Vlastné čísla a vlastné vektory za určitých podmienok dovoľujú upraviť maticu  $\mathbf{A}$  na diagonálny tvar, hovoríme o diagonalizácii matice  $\mathbf{A}$ :

**Tvrdenie 1.2.** *Predpokladajme, že  $n \times n$  matica  $\mathbf{A}$  má  $n$  lineárne nezávislých vektorov. Potom ak tieto vektory zvolíme za stĺpce matice  $\mathbf{S}$*

$$\mathbf{S} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \end{bmatrix}, \quad (3)$$

*matica  $\mathbf{S}^{-1}\mathbf{A}\mathbf{S}$  je diagonálna matica  $\mathbf{\Lambda}$  s vlastnými číslami matice  $\mathbf{A}$  na diagonále*

$$\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \dots & \\ & & & \lambda_n \end{bmatrix}. \quad (4)$$

*Poznámka:* Ak má matica  $\mathbf{A}$   $n$  rôznych vlastných čísel, má podľa Tvrdenia 1.1  $n$  lineárne nezávislých vektorov a teda môže byť diagonalizovateľná.

*Poznámka:* Matica  $\mathbf{S}$  v diagonalizácii nie je určená jednoznačne. Ak je vektor  $\mathbf{x}$  vlastným vektorom matice  $\mathbf{A}$  prislúchajúci vlastnému číslu  $\lambda$ , potom ním je aj jeho nenulový násobok. Viacnásobné vlastné čísla nám tiež dovoľujú vymeniť poradie prislúchajúcich vektorov v matici  $\mathbf{S}$ .

Obe metódy na redukcii dát, ktorým sa budeme v tejto práci venovať, sú založené na počítaní vlastných vektorov a vlastných čísel kovariančnej matice, ktorá je symetrická. Pre symetrické matice platí:

**Veta 1.1.** *Nech  $\mathbf{A} = \mathbf{A}^T$  je reálna symetrická matica. Potom každé vlastné číslo matice  $\mathbf{A}$  je reálne číslo a navyše vlastné vektory matice  $\mathbf{A}$ , prislúchajúce rôznym vlastným číslam, sú navzájom ortogonálne.*

*Dôkaz:*

Pri dokazovaní vety sa využíva skutočnosť, že reálna symetrická matica je zároveň aj hermitovská, tzn.  $\mathbf{A}^H = \overline{\mathbf{A}}^T = \mathbf{A}$ . Najprv ukážeme, že pre hermitovské matice je výraz  $\mathbf{x}^H \mathbf{A} \mathbf{x}$  reálne číslo. Pre ľubovoľný komplexný vektor  $\mathbf{x}$  je tento výraz vo všeobecnosti komplexné číslo. Keďže  $\mathbf{A}$  je hermitovská platí nasledujúca rovnosť

$$(\mathbf{x}^H \mathbf{A} \mathbf{x})^H = \mathbf{x}^H \mathbf{A} \mathbf{x}, \quad (5)$$

aby sa komplexné číslo rovnalo k nemu komplexne združenému číslu, jeho imaginárna časť musí byť nulová, ide teda o reálna číslo.

Využitím, že výraz  $\mathbf{x}^H \mathbf{A} \mathbf{x}$  je reálne číslo možno ukázať reálnosť vlastných čísel hermitovskej a teda aj symetrickej matice. Stačí ak rovnicu  $\mathbf{A} \mathbf{x} = \lambda \mathbf{x}$  prenásobíme vektorom  $\mathbf{x}^H$  zľava a z poznatku, že  $\mathbf{x}^H \mathbf{x} = \|\mathbf{x}\|^2$  je reálne kladné číslo, vieme  $\lambda$  vyjadriť ako podiel dvoch reálnych čísel  $\lambda = \frac{\mathbf{x}^H \mathbf{A} \mathbf{x}}{\mathbf{x}^H \mathbf{x}}$ , čiže aj  $\lambda$  je reálne číslo.

Nech  $\mathbf{x}_1, \mathbf{x}_2$  sú vlastné vektory matice  $\mathbf{A}$  zodpovedajúce rôznym vlastným číslam  $\lambda_1$  a  $\lambda_2$ , platí teda  $\mathbf{A} \mathbf{x}_1 = \lambda_1 \mathbf{x}_1$  a  $\mathbf{A} \mathbf{x}_2 = \lambda_2 \mathbf{x}_2$ . Prenásobením prvej z rovníc vektorom  $\mathbf{x}_2^H$  dostaneme nasledujúcu postupnosť rovností:

$$\lambda_1 \mathbf{x}_2^H \mathbf{x}_1 = \mathbf{x}_2^H \mathbf{A} \mathbf{x}_1 = \mathbf{x}_2^H \mathbf{A}^H \mathbf{x}_1 = (\mathbf{A} \mathbf{x}_2)^H \mathbf{x}_1 = \lambda_2 \mathbf{x}_2^H \mathbf{x}_1. \quad (6)$$

Keďže  $\lambda_1 \neq \lambda_2$ , potom aby bola splnená rovnosť (6) musí platiť  $\mathbf{x}_1^H \mathbf{x}_2 = 0$  a teda vektory  $\mathbf{x}_1, \mathbf{x}_2$  sú navzájom ortogonálne.  $\square$

Ak vektor  $\mathbf{x}_i$  pre  $i = 1, \dots, n$  je vlastným vektorom matice  $\mathbf{A}$  zodpovedajúci vlastnému číslu  $\lambda_i$ , potom je ním aj jeho nenulový násobok, čiže aj vektor  $\mathbf{x}_i / \|\mathbf{x}_i\|$ . V prípade reálnej symetrickej matice  $\mathbf{A} \in \mathbf{R}^{n \times n}$  teda možno zvoliť ortonormálne vlastné vektory. Matica  $\mathbf{S}$  má potom ortonormálne stĺpce, ide teda o ortogonálnu maticu  $\mathbf{Q}$ , pre ktorú platí  $\mathbf{Q}^{-1} = \mathbf{Q}^T$ .

Pre symetrickú maticu  $\mathbf{A}$  (nie nutne s rôznymi vlastnými číslami) platí nasledujúca veta (tzv. veta o hlavných osiach alebo spektrálna veta):

**Veta 1.2.** *Reálna symetrická matica  $\mathbf{A}$  môže byť rozložená do tvaru  $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ , kde  $\mathbf{Q}$  je ortogonálna matica a  $\mathbf{\Lambda}$  je matica s vlastnými číslami na diagonále.*

Vzťah  $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$  sa dá ďalej rozpísať nasledujúcim spôsobom

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T = \lambda_1 \mathbf{x}_1 \mathbf{x}_1^T + \lambda_2 \mathbf{x}_2 \mathbf{x}_2^T + \dots + \lambda_n \mathbf{x}_n \mathbf{x}_n^T, \quad (7)$$

kde  $\mathbf{x}_1, \dots, \mathbf{x}_n$  sú ortonormálne vlastné vektory.  $\mathbf{A}$  je teda súčet  $n$  matic hodnosti jedna.

**Príklad 1:** Pre symetrickú maticu

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

dostaneme charakteristický polynóm

$$\det(\mathbf{A} - \lambda \mathbf{I}) = \lambda^2 - 4\lambda + 3.$$

Jeho korene  $\lambda_1 = 3$  a  $\lambda_2 = 1$  sú vlastnými číslami matice  $\mathbf{A}$ . Vlastný vektor  $\mathbf{x}_1$  prislúchajúci vlastnému číslu  $\lambda_1 = 3$  patrí do nulového priestoru matice

$$\mathbf{A} - 3\mathbf{I} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$$

z čoho vidíme, že  $\mathbf{x}_1 = (1, 1)^T$ . Podobne vlastný vektor  $\mathbf{x}_2$  prislúchajúci vlastnému číslu  $\lambda_2 = 1$  patrí do nulového priestoru matice

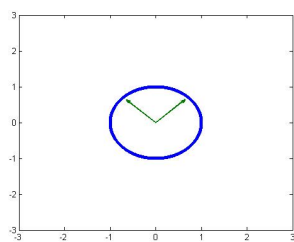
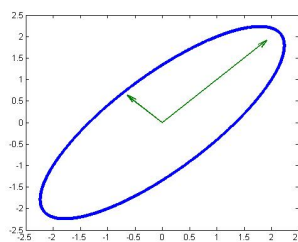
$$\mathbf{A} - \mathbf{I} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

a teda  $\mathbf{x}_2 = (1, -1)^T$ . Vektory  $\mathbf{x}_1$  a  $\mathbf{x}_2$  sú navzájom ortogonálne, ak ich zložky navyše predelíme  $\sqrt{2}$ , získame vektory ortonormálne. Matice  $\mathbf{Q}$  a  $\mathbf{\Lambda}$  vo Vete 1.2 majú teda tvar

$$\mathbf{Q} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}$$

a

$$\mathbf{\Lambda} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}.$$

**Obr. 1:** Kružnica**Obr. 2:** Elipsa

Násobenie vektora  $\mathbf{x}$  maticou  $\mathbf{A}$  predstavuje lineárne zobrazenie vektora  $\mathbf{x}$  do stĺpcového priestoru matice  $\mathbf{A}$ . Vlastný vektor  $\mathbf{x}$ , spĺňajúci rovnicu (1), sa zobrazí na svoj násobok a patrí teda do stĺpcového priestoru matice  $\mathbf{A}$ . Ak zobrazíme jednotkovú kružnicu v Obr.1 pomocou matice  $\mathbf{A}$  z predchádzajúceho príkladu, získame elipsu (Obr. 2), ktorej osami sú vektory  $\lambda_1 \mathbf{x}_1$  a  $\lambda_2 \mathbf{x}_2$ .

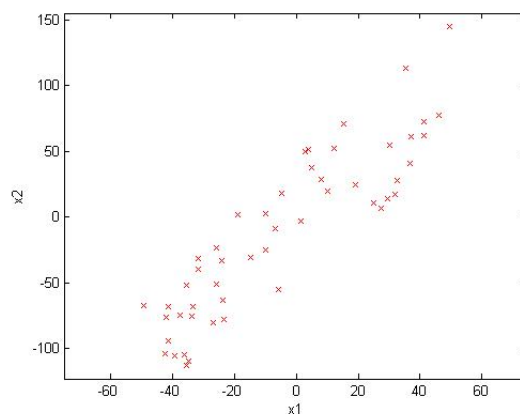
## 2 Analýza hlavných komponentov

Nasledujúca kapitola je celá venovaná analýze hlavných komponentov. V prvej časti tejto kapitoly je uvedené odvodenie metódy a niektoré užitočné teoretické poznatky. Na spracovanie tejto časti boli použité najmä zdroje [7], [4] a [8], ale aj [14]. V druhej časti tejto kapitoly uvádzame dve aplikácie analýzy hlavných komponentov týkajúce sa spracovania obrazu a cien akcií na trhu. Prvá aplikácia je založená na zdroji [11], druhá na zdroji [8].

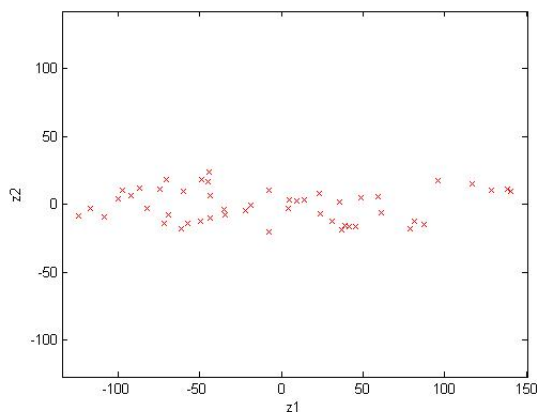
### 2.1 Odvodenie metódy a definícia

Hlavnou myšlienkou analýzy hlavných komponentov (Principal Component Analysis, PCA) je zmenšiť dimenziu súboru dát, ktorý pozostáva z veľkého počtu navzájom súvisiacich korelovaných premenných tak, aby sa zachovalo čo najviac informácie obsiahnutej v týchto dátach. Táto metóda vytvorí nový súbor premenných, tzv. hlavné komponenty, ktoré sú navzájom nekorelované a zoradené zostupne podľa variance, čiže stupňa informácie, ktorú obsahujú.

Nech  $\mathbf{X}$  je vektor  $r$  náhodných premenných a zaujímajú nás variance medzi týmito premennými. Pozrieť sa na všetkých  $\frac{1}{2}r(r-1)$  kovariancií medzi danými premennými, hlavne pre veľké  $r$ , nie je veľmi efektívne. Cieľom PCA je nájsť  $m \ll r$  odvodených premenných, ktoré si zachovávajú čo najviac informácie, ktorá je obsiahnutá v disperziách a kovarianciách.



**Obr. 3:** Grafické znázornenie 50 pozorovaní dvoch premenných  $x_1, x_2$



**Obr. 4:** Grafické znázornenie 50 pozorovaní vzhľadom na hlavné komponenty  $z_1, z_2$

Uvažujme na chvíľu nie veľmi reálny, avšak jednoduchý prípad  $r = 2$ . Prvý graf Obr.3 zobrazuje 50 pozorovaní dvoch náhodných korelovaných premenných  $x_1$  a  $x_2$ . Možno si všimnúť, že obe zložky náhodného vektora majú nemalú disperziu. Tieto zložky sme v grafe Obr.4 zobrazili v súradnicovej sústave danej vektormi hlavných komponentov  $z_1$  a  $z_2$ . Ako vidno v tomto prípade premenná  $z_1$  má oveľa väčšiu disperziu ako premenná  $z_2$  a teda premenná  $z_1$  už pomerne dobre reprezentuje pôvodné dáta. Vo všeobecnosti pre prípad  $r > 2$  niekoľko prvých hlavných komponentov vysvetľuje väčšinu variancie pôvodných premenných. Naopak posledných pár hlavných komponentov určuje smery, v ktorých je variancia veľmi malá.

### 2.1.1 Odvodenie metódy

Uvažujme teda náhodný vektor  $\mathbf{X}$  so strednou hodnotou  $\mu$  a kovariančnou maticou  $\Sigma$ . V prvom kroku hľadáme lineárnu kombináciu zložiek vektora  $\mathbf{X}$

$$\alpha_1^T \mathbf{X} = \alpha_{11}x_1 + \alpha_{12}x_2 + \dots + \alpha_{1r}x_r$$

tak, aby mala maximálnu varianciu. Následne sa hľadá lineárna kombinácia  $\alpha_2^T \mathbf{X}$ , ktorá má maximálnu varianciu a zároveň je nekorelovaná s  $\alpha_1^T \mathbf{X}$ . V k-tom kroku potom hľadáme lineárnu kombináciu  $\alpha_k^T \mathbf{X}$ , ktorá má maximálnu varianciu a zároveň je nekorelovaná s  $\alpha_1^T \mathbf{X}, \alpha_2^T \mathbf{X}, \dots, \alpha_{k-1}^T \mathbf{X}$ .

Nech  $\mathbf{X} = (x_1, x_2, \dots, x_r)^T$  je náhodný vektor, ktorý má  $(r \times r)$  kovariančnú maticu  $\Sigma$ . Táto matica, ak je známa, má na pozícii  $(i, j)$  kovarianciu medzi i-tou a j-tou zložkou

vektora  $\mathbf{X}$  ak  $i \neq j$  a disperziu  $j$ -tej zložky vektora  $\mathbf{X}$  ak  $i = j$ .

Hľadaný vektor  $\alpha_1$  je taký, ktorý maximalizuje varianciu  $\text{var}(\alpha_1^T \mathbf{X}) = \alpha_1^T \Sigma \alpha_1$ . Možno si všimnúť, že toto maximum nie je dosiahnuté pre konečný vektor  $\alpha_1$ , a preto treba použiť normalizačné ohraničenie  $\alpha_1^T \alpha_1 = 1$ , budeme teda hľadať normalizovanú lineárnu kombináciu zložiek vektora  $\mathbf{X}$ . Výsledná úloha je preto problémom na viazaný extrém, ktorý možno riešiť metódou Lagrangeových multiplikátorov. Budeme maximalizovať Lagrangeovu funkciu tvaru

$$f(\alpha_1) = \alpha_1^T \Sigma \alpha_1 - \lambda(\alpha_1^T \alpha_1 - 1),$$

kde  $\lambda$  je multiplikátor. Z nutnej podmienky pre viazaný extrém dostávame rovnicu

$$\frac{\partial f(\alpha_1)}{\partial \alpha_1} = 2\Sigma \alpha_1 - 2\lambda \alpha_1 = 0,$$

ktorú možno upraviť nasledovne

$$\Sigma \alpha_1 = \lambda \alpha_1,$$

čo je podľa Definície 1.1 problém na vlastné čísla a vlastné vektory.  $\lambda$  je teda vlastné číslo matice  $\Sigma$  a  $\alpha_1$  je k nemu prislúchajúci ortonormálny vlastný vektor. Z vlastných vektorov matice  $\Sigma$  potrebujeme vybrať ten, pre ktorý sa maximalizuje variancia

$$\alpha_1^T \Sigma \alpha_1 = \alpha_1^T \lambda \alpha_1 = \lambda \alpha_1^T \alpha_1 = \lambda,$$

a preto hodnota  $\lambda$  bude rovná najväčšiemu vlastnému číslu  $\lambda_1$  matice  $\Sigma$  a  $\alpha_1$  je k nemu prislúchajúci vlastný vektor.

V druhom kroku opäť hľadáme lineárnu kombináciu zložiek pôvodného vektora  $\mathbf{X}$ ,  $\alpha_2^T \mathbf{X}$  tak, aby sme maximalizovali varianciu  $\alpha_2^T \Sigma \alpha_2$  a zároveň bola táto lineárna kombinácia nekorelovaná s  $\alpha_1^T \mathbf{X}$ , t.j.  $\text{cov}(\alpha_1^T \mathbf{X}, \alpha_2^T \mathbf{X}) = 0$ . Túto kovarianciu možno prepísať nasledovne:

$$\text{cov}(\alpha_1^T \mathbf{X}, \alpha_2^T \mathbf{X}) = \alpha_1^T \Sigma \alpha_2 = \alpha_2^T \Sigma \alpha_1 = \alpha_2^T \lambda_1 \alpha_1 = \lambda_1 \alpha_1^T \alpha_2 = 0. \quad (8)$$

Z predošlého vzťahu vidíme, že na dosiahnutie nekorelovanosti hlavných komponentov je potrebné, aby vektory lineárnych kombinácií boli ortogonálne. Analogicky ako v prvom prípade dostávame úlohu na viazaný extrém s dodatočnou väzbou zabezpečujúcou nekorelovanosť náhodných premenných. Na nájdenie vektora  $\alpha_2$  opäť možno

použiť metódu Lagrangeových multiplikátorov, pričom tentokrát má maximalizovaná funkcia tvar

$$f(\alpha_2) = \alpha_2^T \Sigma \alpha_2 - \lambda(\alpha_2^T \alpha_2 - 1) - \phi \alpha_2^T \alpha_1,$$

Z nutnej podmienky pre viazaný extrém má platiť

$$\frac{\partial f(\alpha_2)}{\partial \alpha_2} = 2\Sigma \alpha_2 - 2\lambda \alpha_2 - \phi \alpha_1 = 0, \quad (9)$$

ak túto rovnicu pre násobíme vektorom  $\alpha_1^T$  zľava dostávame

$$2\alpha_1^T \Sigma \alpha_2 - 2\lambda \alpha_1^T \alpha_2 - \phi \alpha_1^T \alpha_1 = 0.$$

Podľa vzťahu (8) sú prvé dva členy predošlej rovnice nulové a taktiež z konštrukcie  $\alpha_1$  platí  $\alpha_1^T \alpha_1 = 1$ , z čoho vyplýva  $\phi = 0$ . Rovnica (9) sa teda zjednoduší na tvar:

$$\Sigma \alpha_2 = \lambda \alpha_2,$$

čiže  $\lambda$  je opäť vlastné číslo matice  $\Sigma$  a  $\alpha_2$  je k nemu prislúchajúci vlastný vektor, pričom zo symetrie kovariančnej matice  $\Sigma$  a za predpokladu, že má  $\Sigma$  rôzne vlastné čísla, vieme, že vektory  $\alpha_1$  a  $\alpha_2$  sú navzájom ortogonálne.

Požadujeme, aby  $\lambda = \alpha_2^T \Sigma \alpha_2$  bolo čo najväčšie. Za predpokladu, že matica  $\Sigma$  má rôzne vlastné čísla,  $\lambda$  bude rovná druhému najväčšiemu vlastnému číslu  $\lambda_2$  tejto matice.

Podobne sa dá ukázať, že pre  $k$ -ty vlastný vektor  $\alpha_k$  matice  $\Sigma$  prislúchajúci  $k$ -temu najväčšiemu vlastnému číslu platí

$$\text{var}(\alpha_k^T \mathbf{X}) = \lambda_k \quad \text{pre} \quad k = 1, 2, \dots, r$$

pričom vektory  $\alpha_1, \dots, \alpha_k$  sú navzájom ortogonálne, preto budú náhodné premenné  $\alpha_1^T \mathbf{X}, \alpha_2^T \mathbf{X}, \dots, \alpha_k^T \mathbf{X}$  navzájom nekorelované.

## 2.1.2 Definícia hlavných komponentov a výberové hlavné komponenty

**Definícia 2.1.** *Nech  $\mathbf{X}$  ( $r \times 1$ ) je náhodný vektor so strednou hodnotou  $\mu$  a  $r \times r$  kovariančnou maticou  $\Sigma$ . Nech spektrálny rozklad matice  $\Sigma$  je*

$$\Sigma = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T, \quad (10)$$

*kde  $\mathbf{A}$  je ortogonálna matica vlastných vektorov a  $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_r)$  je diagonálna matica vlastných čísel, pre ktoré platí  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r$ . Potom náhodný vektor  $\mathbf{Y} = \mathbf{A}^T(\mathbf{X} - \mu)$  nazývame vektorom hlavných komponentov náhodného vektora  $\mathbf{X}$ . Pre  $i = 1, 2, \dots, r$  náhodnú premennú  $Y_i = \alpha_i^T(\mathbf{X} - \mu)$  nazývame  $i$ -tým hlavným komponentom náhodného vektora  $\mathbf{X}$ .*

Pre takto definovaný náhodný vektor  $\mathbf{Y}$  platí nasledujúca veta:

**Veta 2.1.** *Náhodný vektor  $\mathbf{Y} = \mathbf{A}^T(\mathbf{X} - \mu)$  má nulovú strednú hodnotu a kovariančnú maticu  $D(\mathbf{Y}) = \text{diag}(\lambda_1, \dots, \lambda_r)$ , čiže zložky vektora  $\mathbf{Y}$  sú nekorelované a ich disperzie sú  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r$ .*

*Poznámka:* Pri odvodzovaní metódy sme ukázali, že prvý hlavný komponent  $Y_1$  vektora  $\mathbf{X}$  má najväčšiu disperziu spomedzi všetkých normovaných lineárnych kombinácií zložiek vektora  $\mathbf{X}$ . Pre  $k \geq 2$  má  $k$ -ty hlavný komponent náhodného vektora  $\mathbf{X}$  najväčšiu disperziu spomedzi normovaných lineárnych kombinácií zložiek vektora  $\mathbf{X}$ , ktoré sú zároveň nekorelované s hlavnými komponentami  $Y_1, \dots, Y_{k-1}$ .

Vieme teda ako vyzerajú hlavné komponenty náhodného vektora so známym vektorom stredných hodnôt a známou kovariančnou maticou. V skutočnosti však tento vektor a túto maticu väčšinou nepoznáme. V takom prípade ich odhadneme výberovým priemerom a výberovou kovariančnou maticou.

Nech  $\mathbf{X}_i, i = 1, \dots, n$ , je  $n$  nezávislých pozorovaní vektora  $\mathbf{X}$ . Potom vektor stredných hodnôt  $\mu$  odhadneme pomocou výberového priemeru

$$\bar{\mathbf{X}} = n^{-1} \sum_{i=1}^n \mathbf{X}_i.$$

Označme

$$\mathbf{X}_{ci} = \mathbf{X}_i - \bar{\mathbf{X}}$$

$i = 1, 2, \dots, n$  a  $\mathbf{X}_c = (\mathbf{X}_{c1}, \dots, \mathbf{X}_{cn})$  je  $r \times n$  matica. Potom kovariačnú maticu  $\Sigma$  odhadneme výberovou kovariančnou maticou

$$\hat{\Sigma} = n^{-1} \mathbf{X}_c \mathbf{X}_c^T. \quad (11)$$

Vlastné čísla  $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_r \geq 0$  matice  $\hat{\Sigma}$  budeme nazývať výberovými disperziami hlavných komponentov a k nim prislúchajúce vlastné vektory  $\mathbf{a}_1, \dots, \mathbf{a}_r$  budeme nazývať vektormi výberových hlavných komponentov. Náhodné vektory  $\mathbf{Y}_j = (\mathbf{a}_1, \dots, \mathbf{a}_r)^T (\mathbf{X}_j - \bar{\mathbf{X}})$ ,  $j = 1, \dots, n$  charakterizujú vektory  $\mathbf{X}_1, \dots, \mathbf{X}_n$  v súradnicovej sústave určenej výberovými vlastnými vektormi a realizácie týchto náhodných vektorov sa nazývajú vektory skóre.

Ak máme objekt reprezentovaný veľkým súborom dát, ktorý obsahuje viac korelovaných premenných, môže nás po vykonaní analýzy zaujímať akým spôsobom možno objekt aproximovať/zrekonštruovať pomocou získaných výberových vektorov hlavných komponentov a dosiahnutých skóre. Najlepšia rekonštrukcia rádu  $t$  vektora  $\mathbf{X}$  je daná ako

$$\hat{\mathbf{X}}^{(t)} = \bar{\mathbf{X}} + \sum_{j=1}^t \mathbf{a}_j \mathbf{a}_j^T (\mathbf{X} - \bar{\mathbf{X}}). \quad (12)$$

Všimnime si, že pri aproximácii využívame iba súradnice v smeroch s najväčšou informáciou o dátach, ktorými sú vektory výberových hlavných komponentov. Ak by sme využili všetkých  $r$  vektorov hlavných komponentov dostaneme presne pôvodný vektor  $\mathbf{X}$ .

### 2.1.3 PCA a korelačná matica

V praxi sa často stáva, že jednotlivé zložky vektora  $\mathbf{X}$  predstavujú rôzne typy meraní. Niektoré môžu predstavovať hmotnosť, iné teplotu, dĺžku a podobne. V takom prípade jednotlivé hlavné komponenty závisia od voľby jednotiek. Veľkou nevýhodou hlavných komponentov založených na kovariančnej matici, hlavne v prípade rôznych typov dát, je práve ich citlivosť na zmenu jednotiek, t.j metóda hlavných komponentov nie je škálovo invariantná.

Preto sa na odhad hlavných komponentov častokrát namiesto výberovej kovariančnej matice používa výberová korelačná matica. Pre korelačnú maticu sú jej zložky bezrozmerné, výberové hlavné komponenty sú teda škálovo invariantné. Poznamenejme, že použitie korelačnej matice je ekvivalentné štandardizácii premenných pred ich použitím.

Označme  $\mathbf{X}^*$  vektor, ktorého zložky sú  $\frac{x_j - \mu_j}{\sigma_{jj}^{1/2}}, j = 1, \dots, r$ . Hlavné komponenty sú potom definované ako  $\mathbf{Z} = \mathbf{A}^T \mathbf{X}^*$ , kde stĺpce matice  $\mathbf{A}$  sú vlastné vektory korelačnej matice.

Keďže transformácia z  $\mathbf{X}$  do  $\mathbf{X}^*$  nie je ortogonálna, tak hlavné komponenty odvodené z korelačnej matice nepodávajú rovnakú informáciu ako hlavné komponenty odvodené z kovariančnej matice a nemožno jedny z druhých jednoducho priamo odvodiť.

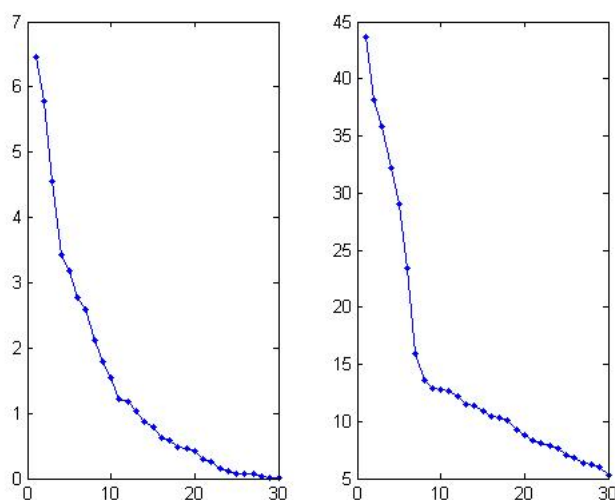
### 2.1.4 Koľko hlavných komponentov ponechať?

Pri použití analýzy hlavných komponentov ako metódy na redukcii dát treba riešiť ešte jednu dôležitú otázku: koľko hlavných komponentov ponechať? Množstvo informácie v  $j$ -tom hlavnom komponente je dané jeho disperziou, ktorá je vyjadrená  $j$ -tym najväčším vlastným číslom kovariančnej, resp. korelačnej matice. Ako veľké musí teda vlastné číslo byť, aby sme k nemu prislúchajúci hlavný komponent považovali za dôležitý?

Jednou z možností je použiť lakťový diagram, čo je lomená čiara spájajúca vlastné čísla zobrazené oproti ich poradovému číslu. Ak niekoľko najväčších vlastných čísel dominuje svojou veľkosťou nad ostatnými, v lakťovom diagrame budú tieto čísla oddelené, bude v ňom viditeľné miesto najvýraznejšieho zlomu, tzv. lakeť. Vo všeobecnosti sa potom odporúča použiť tie hlavné komponenty, ktoré zodpovedajú vlastným číslam pred spomínaným lakťom a naviac ešte prvý, ktorý je za ním. V prípade, že v tomto grafe sa

žiadny lakeť neobjaví, obľúbeným pravidlom je použiť najväčšie hlavné komponenty, ktoré dokopy zodpovedajú 90% celkovej disperzie.

Nasledujúce dva grafy zobrazujú vlastné čísla kovariančnej matice Gaussovských dát. V prvom prípade sme vygenerovali  $n = 30$  pozorovaní náhodného vektora dĺžky  $r = 30$ , v druhom sme pre náhodný vektor rovnakej dĺžky vygenerovali až  $n = 300$  pozorovaní. Možno si všimnúť, že pri väčšom počte pozorovaní je pozorovateľný výrazný zlom me-



**Obr. 5:** Laktový diagram

dzi 7 a 8 vlastným číslom, čo sa však nedá povedať o grafe pre menší počet pozorovaní. Pre PCA pomocou korelačnej matice existuje Kaiserovo pravidlo, podľa ktorého je dobré ponechať tie hlavné komponenty, ktorých zodpovedajúce vlastné číslo je väčšie ako jedna. Toto pravidlo je založené na úvahe, že ak celková disperzia  $r$  normalizovaných premenných je  $r$ , tak by hlavný komponent mal zodpovedať aspoň priemernej disperzii jednej štandardizovanej premennej. Toto pravidlo je populárne, no hodnota jedna je dosť vysoká a preto sa používa aj upravené pravidlo s hodnotou 0.9, prípadne 0.7.

### 2.1.5 Príklad použitia PCA

Pre lepšiu predstavu o fungovaní metódy uvádzame aj jednu jednoduchú aplikáciu PCA. Použili sme dáta crabs, ktoré sú dostupné v jednom z balíkov programu R. Tieto dáta opisujú veľkosť krabov pomocou piatich rôznych rozmerov: veľkosti predného lalku (frontal lobe size, FL), šírky zadnej časti tela (rear width, RW), dĺžky panciera (carapace length, CL), šírky panciera (carapace width, CW) a hrúbky tela (body depth, BD). Sledovať všetkých päť rozmerov naraz je celkom náročné, preto by sme radi každého kraba reprezentovali menším počtom premenných. Keďže boli tieto dáta namerané na 200 jedincoch, máme k dispozícii 200 pozorovaní náhodného vektora  $\mathbf{X} = [\text{FL}, \text{RW}, \text{CL}, \text{CW}, \text{BD}]$ . Na tieto dáta sme aplikovali PCA. Vzhľadom na to, že všetky dáta boli v milimetroch a teda v rovnakých jednotkách, pri výpočte sme použili kovariančnú maticu. Získané výberové variancie sú uvedené v tabuľke 1.

|                             | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_5$ |
|-----------------------------|-------------|-------------|-------------|-------------|-------------|
| Výberová variácia           | 140.7075    | 1.2968      | 1.0003      | 0.1353      | 0.0779      |
| Percento celkovej variancie | 98.25%      | 0.91 %      | 0.7%        | 0.09%       | 0.05%       |

**Tabuľka 1:** Výberová variácia a percento celkovej variancie

Možno si všimnúť, že prvá výberová variácia je mnohonásobne väčšia ako ostatné štyri a zodpovedá až 98.25% celkovej variancie v dátach.

V tabuľke 2 sú uvedené získané vektory výberových hlavných komponentov. Keďže

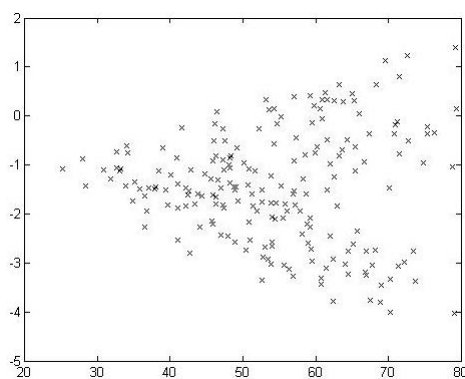
|    | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ |
|----|------------|------------|------------|------------|------------|
| FL | 0.2890     | -0.3233    | 0.5072     | 0.7949     | -0.1249    |
| RW | 0.1973     | -0.8647    | -0.4141    | -0.1483    | 0.1409     |
| CL | 0.5994     | 0.1982     | 0.1753     | -0.1436    | 0.7417     |
| CW | 0.6617     | 0.2880     | -0.4914    | 0.1256     | -0.4712    |
| BD | 0.2837     | -0.1598    | 0.5469     | -0.6344    | -0.4387    |

**Tabuľka 2:** Vektory výberových hlavných komponentov

prvý hlavný komponent má v sebe väčšinu informácie obsiahnutej v našich dátach, je

postačujúce sústrediť sa pri analýze len na prvý hlavný komponent. Podarilo sa nám teda zredukovať 5 náhodných premenných do jednej. Zložky prvého výberového komponentu sú všetky kladné, možno ho teda interpretovať ako akúsi celkovú veľkosť kraba. V tomto prípade však možno interpretovať aj druhý hlavný komponent. Vzhľadom na znamienka koeficientov ho možno chápať ako porovnanie veľkosti tela a veľkosti panciera.

V grafe Obr. 6 je zobrazených všetkých 200 pozorovaní vzhľadom na výberové vektory hlavných komponentov  $\alpha_1$  a  $\alpha_2$ . V tomto obrázku sú teda zobrazené 5-rozmerné dáta v rovine, pričom sa nám podarilo zachovať väčšinu informácie v dátach. Možno si všimnúť, že horizontálna os má oveľa väčší rozsah ako vertikálna os. Je to spôsobené tým, že prvý hlavný komponent má mnohonásobne väčšiu variáciu ako druhý hlavný komponent.



**Obr. 6:** Grafické znázornenie pozorovaní vzhľadom na vektory výberových hlavných komponentov  $\alpha_1$  a  $\alpha_2$

## 2.2 Aplikácie PCA

Analýza hlavných komponentov sa používa hlavne na redukciiu dimenzie dát, často sa však používa aj pri hľadaní dôležitých vlastností dát. Využíva sa napríklad vo financiách, ekonometrii či pri spracovaní obrazu [11]. V [7] je uvedené použitie PCA pri rozoznávaní tvárí, ale aj pri rozoznávaní písma.

### 2.2.1 Spracovanie obrazu

Jednou z mnohých aplikácií PCA je aj jej využitie pri spracovaní obrazov. Keďže takúto aplikáciu možno aj zobrazit', rozhodli sme sa, inšpirovaní [11], takúto aplikáciu vyskúšať.

Farebný obrázok veľkosti  $m \times n$  sa skladá z pixelov, ktorých farba závisí od intenzity troch farebných zložiek: červenej (R), zelenej (G) a modrej (B). Možno ho teda reprezentovať 3D maticou s rozmermi  $m \times n \times 3$ , pričom každá z troch nadrovín hovorí o intenzite danej farebnej zložky konkrétneho pixelu.

Pre nami vybraný obrázok Obr.7 sme teda získali takúto maticu, ktorú sme trans-



**Obr. 7:** Pôvodný obrázok

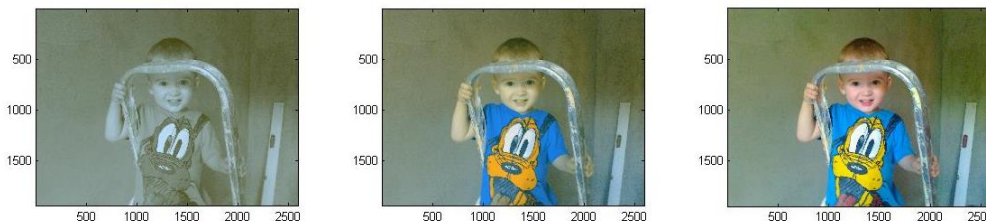
formovali na 3 vektory dĺžky  $m \times n$ . Pre náhodný vektor  $\mathbf{X} = [R, G, B]^T$  sme tak získali  $m \times n$  pozorovaní. Na tieto dáta sme aplikovali PCA. Získané výberové disperzie hlavných komponentov sú v tabuľke 3 spolu s kumulatívnym percentom vysvetlenej variancie.

Pri PCA sme taktiež vypočítali vektory hlavných komponentov, pomocou ktorých sme sa následne pokúsili o rekonštrukciu pôvodného obrázku pomocou vzťahu (12). Ob-

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ |
|-------------|-------------|-------------|
| 4071,7      | 2457,0      | 179,3       |
| 60,7%       | 97,33%      | 100%        |

**Tabuľka 3:** Výberové disperzie hlavných komponentov a kumulatívna variancia

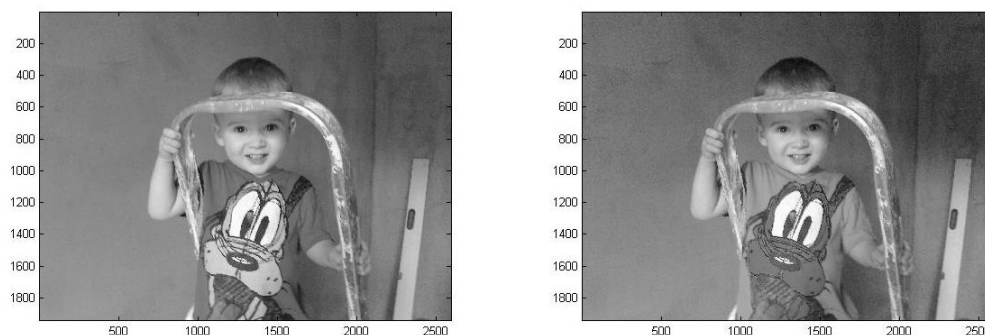
rázok sme postupne vykreslili pre  $t = 1, 2, 3$ . Poznamenajme, že sme rekonštrukciu obrázka zopakovali pre každý pixel obrázka a teda pre každé pozorovanie náhodného vektora  $\mathbf{X}$ .



**Obr. 8:** Rekonštrukcia obrázka pomocou hlavných komponentov

Možno si všimnúť, že prvý obrázok, hoci je zatiaľ takmer jednofarebný, už celkom dobre popisuje pôvodný obrázok. Tento fakt nie až taký prekvapivý, ak vezmeme do úvahy, že prvý hlavný komponent zodpovedá 60,7% celkovej variancie obsiahnutej v našich dátach. Spolu s druhým komponentom už zodpovedajú 97,33% celkovej variancie, pridaním druhého člena súčtu vo vzorci (12) sme tak získali už pomerne dobrý farebný obrázok. Poznamenajme, že z našich pokusov vyplynulo, že prvé dva komponenty veľmi dobre popisujú väčšinu obrázkov a tvoria viac ako 90 % celkovej variancie. V prípade niektorých obrázkov bol takto zrekonštruovaný obrázok od pôvodného takmer nerozoznateľný. Pridaním posledného člena sme obrázok úplne zrekonštruovali a dostali sme teda pôvodný obrázok.

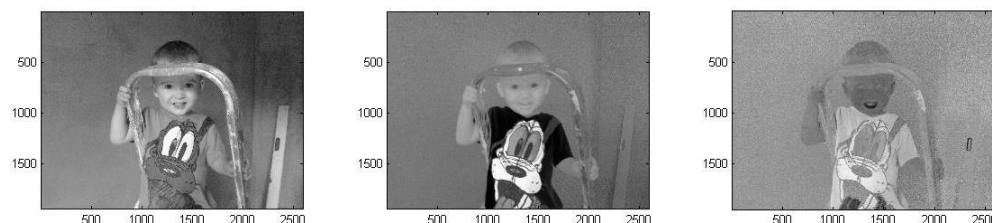
V článku [11] autori používajú PCA na kompresiu obrázkov, transformáciu obrázka z farebného obrázka do stupňov šedej. Tri farebné zložky sú redukované do jednej tak, aby bolo zachované čo najviac informácie. S každým pixelom obrázka sa spája jedno pozorovanie náhodného vektora  $\mathbf{X} = [R, G, B]^T$  a teda pre každý takýto vektor možno spočítať vektory skóre. Pre porovnanie uvádzame aj čiernobielu verziu pôvodného obrázka vykreslenú spolu s prvými zložkami vektorov skóre.



**Obr. 9:** Čiernobiely pôvodný obrázok a prvé zložky vektorov skóre

Vidno, že druhý obrázok, čiže rekonštrukcia pomocou prvých zložiek vektorov skóre, celkom dobre vystihuje pôvodný obrázok. Nemôžno sa však čudovať, že sa tieto obrázky odlišujú, keďže čiernobiela verzia pôvodného obrázku má v sebe plnú informáciu o variancii, zatiaľ čo rekonštrukcia obrázku má v sebe len časť tejto informácie.

Iný spôsob akým možno vyzdvihnúť rozdiel vo vysvetlených varianciách je vykresliť vektory skóre. Obrázky v Obr. 10 postupne vyobrazujú prvé, druhé a tretie zložky takýchto



**Obr. 10:** Zložky vektorov skóre

vektorov. Z obrázkov vidno, že najčitateľnejší je prvý z nich, čo je zrejme spôsobené tým, že prvý hlavný komponent má najväčšiu varianciu. Na druhej strane variancia tretieho hlavného komponentu je len na úrovni 4,4% variancie prvého komponentu, preto je obrázok najmenej čitateľný. Variancia totiž hovorí o tom, ako veľmi sa jednotlivé pozorovania odlišujú od ich priemeru. Čím je teda variancia menšia, tým sú si zložky podobnejšie a teda sa jednotlivé pixely na seba viac podobajú.

### 2.2.2 Ceny akcií

Pôvodným cieľom PCA bolo znížiť veľký počet premenných  $r$  na čo najmenšie číslo  $m$  a zachovať pritom čo najviac informácie z pôvodných  $r$  premenných. I keď PCA má množstvo ďalších aplikácií, zrejme najpoužívanejšou je naďalej spomínaná aplikácia zameraná na redukciu dimenzie dát. Pri takomto použití PCA je občas možné niekoľko prvých hlavných komponentov intuitívne interpretovať. V kapitole 4 v [8] je uvedených niekoľko takýchto príkladov. Jedným z nich je aj aplikácia PCA na ceny akcií obchodovateľných na burze. Podobnú aplikáciu sme spravili aj my, pričom nás zaujímalo, či budeme schopní výsledné dáta nejako interpretovať.

V našom prípade sme sa zamerali na akcie šiestich rôznych spoločností, ktoré sú obchodovateľné na burze v New Yorku. V tabuľke 4 sú uvedené nami vybrané spoločnosti spolu so skratkami, ktorými ich budeme označovať.

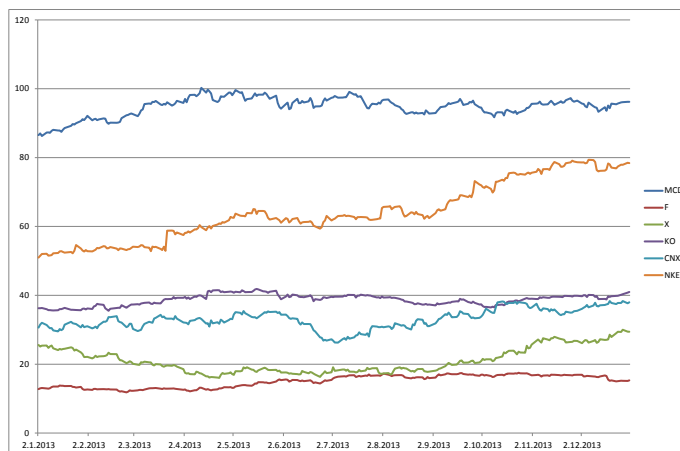
|                           |     |
|---------------------------|-----|
| McDonald's Corp.          | MCD |
| Ford Motor Co.            | F   |
| United States Steel Corp. | X   |
| The Coca-Cola Com.        | KO  |
| CONSOL Energy Inc.        | CNX |
| Nike, Inc.                | NKE |

**Tabuľka 4:** Zoznam spoločností

Na stránke [19] sme našli potrebné dáta - denné ceny jednotlivých akcií za obdobie od 2.1.2013 po 31.12.2013. Získali sme tak 255 pozorovaní náhodného vektora, ktorého zložkami sú ceny akcií v daný deň. Tieto sú zobrazené v Obr. 11.

Pre tieto dáta sme vypočítali koeficienty vektorov výberových hlavných komponentov najprv použitím kovariančnej matice. V tabuľke 5 sú uvedené koeficienty prvých dvoch takýchto vektorov spolu s percentom vysvetlenej variancie.

Z dlhodobého hľadiska ceny všetkých šiestich akcií narástli, preto prvý z vektorov, ktorého všetky koeficienty sú kladné, možno považovať za akýsi komponent veľkosti. Z grafu v Obr. 11 si možno všimnúť, že najväčšia zmena cien za toto obdobie nastala u akcií spoločnosti Nike. Preto nás neprekvapuje skutočnosť, že práve pre túto spoločnosť



Obr. 11: Ceny akcií

| Spoločnosť                  | 1      | 2       |
|-----------------------------|--------|---------|
| MCD                         | 0.0991 | 0.6065  |
| F                           | 0.1470 | 0.0729  |
| X                           | 0.2243 | -0.7045 |
| KO                          | 0.662  | 0.2845  |
| CNX                         | 0.2273 | -0.1877 |
| NKE                         | 0.9285 | 0.1196  |
| Percento celkovej variancie | 76.94% | 15.31%  |

Tabuľka 5: Vektory výberových hlavných komponentov použitím kovariančnej matice

je koeficient najväčší.

Koeficienty druhého vektora výberových hlavných komponentov sa v znamienkach odlišujú. Opačné znamienko oproti zvyšným spoločnostiam majú United States Steel Corp., ktorá vyrába produkty z ocele a CONSOL Energy Inc., ktorá produkuje uhlie a zemný plyn. Druhý hlavný komponent teda možno interpretovať ako akýsi kontrast medzi spoločnosťami, ktoré svoje produkty predávajú ďalším spoločnostiam a zvyšnými štyrmi spoločnosťami Nike, Inc., The Coca-Cola Com., Ford Motor Co. a McDonald's Corp., ktoré svoje produkty predávajú priamo zákazníkom.

Pre porovnanie sme sa pozreli aj na vektory výberových hlavných komponentov vy-

počítaných z korelačnej matice. Tieto možno ekvivalentne získať aj štandardizáciou dát. V tabuľke 6 sú uvedené koeficienty prvých dvoch vektorov výberových hlavných komponentov pre štandardizované dáta.

| Spoločnosť                  | 1      | 2       |
|-----------------------------|--------|---------|
| MCD                         | 0.3006 | -0.5757 |
| F                           | 0.4434 | 0.0608  |
| X                           | 0.2519 | 0.5556  |
| KO                          | 0.3376 | -0.5192 |
| CNX                         | 0.4369 | 0.2751  |
| NKE                         | 0.5871 | 0.1044  |
| Percento celkovej variancie | 45.15% | 33.86%  |

**Tabuľka 6:** Vektory výberových hlavných komponentov použitím korelačnej matice

Prvý komponent opäť možno interpretovať ako komponent veľkosti, všetky koeficienty v ňom majú rovnaké znamienko. Tentokrát však vysvetľuje len 45.15% celkovej variancie obsiahnutej v našich dátach. Avšak druhý vektor nemožno interpretovať tak ako v predošlom prípade. Napriek tomu aj v tomto prípade vieme nájsť intuitívnu interpretáciu: ide o kontrast medzi spoločnosťami The Coca-Cola Com. a McDonald's Corp. pôsobiacimi v potravinárskom priemysle a spoločnosťami pôsobiacimi v iných odvetviach.

### 3 Kanonické korelácie

Tretia kapitola je celá venovaná analýze kanonických korelácií, ďalšej metóde na lineárnu redukciu dát. Opäť je rozdelená na dve časti. V prvej časti je uvedené odvodenie metódy spolu s definíciou kanonických korelácií a jednoduchým príkladom, pričom boli využité najmä zdroje [7], [4] a [18], ale aj [14] a [3]. V druhej časti sú uvedené dve aplikácie tejto metódy využívajúce poznatky zo zdrojov [4] a [3].

#### 3.1 Odvodenie metódy a definícia

Analýza kanonických korelácií (Canonical Correlation Analysis, CCA) predstavuje zrejme najpoužívanjšiu metódu pre štúdium lineárnych vzťahov medzi dvomi vektorovými premennými. Napriek tomu, že táto metóda pristupuje k obojm vektorom rovnako, často sa používa v prípade, keď zložky jedného vektora sú nezávislé premenné, zatiaľ čo zložky druhého vektora sú premennými závislými. CCA sa tiež používa vtedy, ak potrebujeme nájsť vzťah medzi dvomi rôznymi reprezentáciami toho istého predmetu. Výsledky získané aplikáciou CCA by mali pomôcť zodpovedať otázky o tom, aké vzťahy existujú medzi dvomi viacrozmernými vektorovými premennými a aké silné tieto vzťahy sú.

Nech vektor

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$$

je súbor  $r + s$  náhodných premenných, pričom  $\mathbf{X} = (X_1, \dots, X_r)^T$  a  $\mathbf{Y} = (Y_1, \dots, Y_s)^T$  sú náhodné vektory, dva podsúbory náhodných premenných, ktorých vektor stredných hodnôt je daný:

$$E \left\{ \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \right\} = \begin{pmatrix} \mu_{\mathbf{X}} \\ \mu_{\mathbf{Y}} \end{pmatrix}$$

a kovariančná matica má tvar:

$$E \left\{ \begin{pmatrix} \mathbf{X} - \mu_{\mathbf{X}} \\ \mathbf{Y} - \mu_{\mathbf{Y}} \end{pmatrix} \begin{pmatrix} \mathbf{X} - \mu_{\mathbf{X}} \\ \mathbf{Y} - \mu_{\mathbf{Y}} \end{pmatrix}^T \right\} = \begin{pmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{YY}} \end{pmatrix},$$

pričom predpokladáme, že matice  $\Sigma_{\mathbf{XX}}$  a  $\Sigma_{\mathbf{YY}}$  sú regulárne a keďže ide o kovariančné matice, sú aj pozitívne definitné.

Cieľom metódy CCA je nahradiť tieto dva súbory korelovaných premenných  $t$  dvojicami nových premenných

$$(\xi_j, \omega_j), \quad j = 1, 2, \dots, t, \quad t \leq \min(r, s),$$

kde

$$\xi_j = \mathbf{g}_j^T \mathbf{X} = g_{1j}X_1 + g_{2j}X_2 + \dots + g_{rj}X_r$$

$$\omega_j = \mathbf{h}_j^T \mathbf{Y} = h_{1j}Y_1 + h_{2j}Y_2 + \dots + h_{sj}Y_s.$$

Premenné  $\xi_j, \omega_j$ , ktoré budeme nazývať kanonické premenné, sú teda lineárne kombinácie zložiek vektorov  $\mathbf{X}$  a  $\mathbf{Y}$ . Vektory koeficientov  $\mathbf{g}_j$  a  $\mathbf{h}_j$ , nazývané aj vektory kanonických korelácií, sú vyberané tak, aby:

- dvojice  $(\xi_j, \omega_j)$  boli zoradené zostupne podľa ich korelácie, ktorá je daná ako:

$$\rho_j = \frac{\text{cov}(\mathbf{g}_j^T \mathbf{X}, \mathbf{h}_j^T \mathbf{Y})}{\sqrt{\mathbf{D}(\mathbf{g}_j^T \mathbf{X})} \sqrt{\mathbf{D}(\mathbf{h}_j^T \mathbf{Y})}} = \frac{\mathbf{g}_j^T \Sigma_{\mathbf{XY}} \mathbf{h}_j}{(\mathbf{g}_j^T \Sigma_{\mathbf{XX}} \mathbf{g}_j)^{1/2} (\mathbf{h}_j^T \Sigma_{\mathbf{YY}} \mathbf{h}_j)^{1/2}}, \quad (13)$$

- $\xi_j$  je nekorelované s  $\xi_k$  pre  $k = 1, \dots, j - 1$ :  $\text{cov}(\xi_j, \xi_k) = 0$ ,
- $\omega_j$  je nekorelované s  $\omega_k$  pre  $k = 1, \dots, j - 1$ :  $\text{cov}(\omega_j, \omega_k) = 0$ .

Je dôležité uvedomiť si, že koeficient korelácie hovorí len o lineárnej závislosti premenných. Ak sú teda dve premenné nekorelované, znamená to, že neexistuje lineárna funkcia, ktorá popisuje ich vzťah. Môže však existovať kvadratická alebo iná nelineárna funkcia, ktorá ho popisuje.

### 3.1.1 Odvodenie metódy

Nech premenné  $\xi = \mathbf{g}^T \mathbf{X}$  a  $\omega = \mathbf{h}^T \mathbf{Y}$  sú ľubovoľné lineárne kombinácie zložiek vektorov  $\mathbf{X}$  a  $\mathbf{Y}$ . Bez ujmy na všeobecnosti možno predpokladať, že  $E(\mathbf{X}) = \mu_{\mathbf{X}} = \mathbf{0}$  a  $E(\mathbf{Y}) = \mu_{\mathbf{Y}} = \mathbf{0}$  a teda aj  $\xi, \omega$  majú nulové stredné hodnoty.

Ak sa pozrieme na korelačný koeficient v (13), vidíme, že je invariantný vzľadom na preškáľovanie vektorov  $\mathbf{g}$  a  $\mathbf{h}$ , a teda ich možno násobiť ľubovoľným koeficientom  $k$ . Môžeme preto zaviesť dodatočný predpoklad, že  $\xi$  a  $\omega$  majú jednotkové variancie:  $\mathbf{g}^T \Sigma_{\mathbf{XX}} \mathbf{g} = 1$  a  $\mathbf{h}^T \Sigma_{\mathbf{YY}} \mathbf{h} = 1$ .

Najprv treba nájsť vektory  $\mathbf{g}, \mathbf{h}$  také, aby korelácia medzi náhodnými premennými  $\xi,$

$\omega$ ,  $\text{corr}(\xi, \omega) = \mathbf{g}^T \Sigma_{\mathbf{XY}} \mathbf{h}$ , bola maximálna zo všetkých možných lineárnych kombinácií vektorov  $\mathbf{X}$  a  $\mathbf{Y}$ , ktoré spĺňajú podmienky  $\mathbf{g}^T \Sigma_{\mathbf{XX}} \mathbf{g} = 1$  a  $\mathbf{h}^T \Sigma_{\mathbf{YY}} \mathbf{h} = 1$ . Potrebujeme teda riešiť úlohu na viazaný extrém.

Na nájdenie riešenia možno použiť metódu Lagrangeových multiplikátorov, pričom budeme pracovať s funkciou

$$f(\mathbf{g}, \mathbf{h}) = \mathbf{g}^T \Sigma_{\mathbf{XY}} \mathbf{h} - \frac{1}{2} \lambda (\mathbf{g}^T \Sigma_{\mathbf{XX}} \mathbf{g} - 1) - \frac{1}{2} \mu (\mathbf{h}^T \Sigma_{\mathbf{YY}} \mathbf{h} - 1), \quad (14)$$

kde  $\lambda$  a  $\mu$  sú Lagrangeove multiplikátory.

Rovnicu (14) postupne parciálne zderivujeme podľa  $\mathbf{g}$  a  $\mathbf{h}$  a z nutnej podmienky pre viazaný extrém dostávame

$$\frac{\partial f}{\partial \mathbf{g}} = \Sigma_{\mathbf{XY}} \mathbf{h} - \lambda \Sigma_{\mathbf{XX}} \mathbf{g} = \mathbf{0}, \quad (15)$$

$$\frac{\partial f}{\partial \mathbf{h}} = \Sigma_{\mathbf{YX}} \mathbf{g} - \mu \Sigma_{\mathbf{YY}} \mathbf{h} = \mathbf{0}. \quad (16)$$

Prenásobením rovnice (15) vektorom  $\mathbf{g}^T$  zľava a rovnice (16) vektorom  $\mathbf{h}^T$  zľava dostaneme

$$\mathbf{g}^T \Sigma_{\mathbf{XY}} \mathbf{h} - \lambda \mathbf{g}^T \Sigma_{\mathbf{XX}} \mathbf{g} = 0, \quad (17)$$

$$\mathbf{h}^T \Sigma_{\mathbf{YX}} \mathbf{g} - \mu \mathbf{h}^T \Sigma_{\mathbf{YY}} \mathbf{h} = 0. \quad (18)$$

Využitím predpokladov  $\mathbf{g}^T \Sigma_{\mathbf{XX}} \mathbf{g} = 1$  a  $\mathbf{h}^T \Sigma_{\mathbf{YY}} \mathbf{h} = 1$  potom získame vzťah

$$\lambda = \mu = \mathbf{g}^T \Sigma_{\mathbf{XY}} \mathbf{h},$$

Lagrangeove multiplikátory sú teda rovné maximalizovanému korelačnému koeficientu.

V rovnici (16) možno za  $\mu$  dosadiť  $\lambda$ , zároveň prenasobíme rovnicu (15) maticou  $\Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1}$ :

$$-\lambda \Sigma_{\mathbf{YX}} \mathbf{g} + \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} \mathbf{h} = \mathbf{0}, \quad (19)$$

$$\Sigma_{\mathbf{YX}} \mathbf{g} - \lambda \Sigma_{\mathbf{YY}} \mathbf{h} = \mathbf{0}. \quad (20)$$

Po následnom dosadení vzťahu (20) do rovnice (19) dostaneme

$$(\Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} - \lambda^2 \Sigma_{\mathbf{YY}}) \mathbf{h} = \mathbf{0}.$$

Analogicky, ak v rovnici (16) za  $\mu$  dosadíme  $\lambda$  a prenasobíme ju zľava maticou  $\Sigma_{\mathbf{XY}} \Sigma_{\mathbf{YY}}^{-1}$  dostaneme

$$\Sigma_{\mathbf{XY}} \Sigma_{\mathbf{YY}}^{-1} \Sigma_{\mathbf{YX}} \mathbf{g} - \lambda \Sigma_{\mathbf{XX}} \mathbf{g} = \mathbf{0}. \quad (21)$$

Po dosadení vzťahu  $\Sigma_{\mathbf{XY}}\mathbf{h} = \lambda\Sigma_{\mathbf{XX}}\mathbf{g}$  získaného z rovnice (15) do rovnice (21), získame rovnicu

$$(\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}} - \lambda^2\Sigma_{\mathbf{XX}})\mathbf{g} = \mathbf{0}.$$

Na získanie hľadaných vektorov  $\mathbf{g}$  a  $\mathbf{h}$  je teda potrebné vyriešiť rovnice

$$(\Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XY}} - \lambda^2\Sigma_{\mathbf{YY}})\mathbf{h} = \mathbf{0}, \quad (22)$$

$$(\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}} - \lambda^2\Sigma_{\mathbf{XX}})\mathbf{g} = \mathbf{0}. \quad (23)$$

ktoré sú zovšeobecnenými problémami na vlastné čísla.

Nasledujúcou jednoduchou úpravou rovníc možno ukázať, že vlastné čísla sú pre obe rovnice zhodné a medzi vlastnými vektormi existuje jednoduchý vzťah.

Najprv rovnicu (22) pre násobíme maticou  $\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}$

$$(\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XY}} - \lambda^2\Sigma_{\mathbf{XY}})\mathbf{h} = \mathbf{0},$$

vložíme maticu  $\Sigma_{\mathbf{XX}}\Sigma_{\mathbf{XX}}^{-1} = \mathbf{I}$

$$(\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XY}} - \lambda^2\Sigma_{\mathbf{XX}}\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XY}})\mathbf{h} = \mathbf{0},$$

ak sprava vyjmeme spoločný člen  $\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XY}}$  dostávame

$$(\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}} - \lambda^2\Sigma_{\mathbf{XX}})\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XY}}\mathbf{h} = \mathbf{0}. \quad (24)$$

Porovnaním rovníc (23) a (24) si možno všimnúť, že ich riešením sú rovnaké nenulové vlastné čísla a pre vlastné vektory platí  $\mathbf{g} = c\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XY}}\mathbf{h}$ , pričom  $c$  je taká konštanta, pre ktorú  $\mathbf{g}^T\Sigma_{\mathbf{XX}}\mathbf{g} = 1$ . Stačí teda vyriešiť rovnicu (22), aby sme získali hľadané riešenia.

Aby mala táto rovnica netriviálne riešenie, musí byť determinant

$$\det(\Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XY}} - \lambda^2\Sigma_{\mathbf{YY}})$$

rovný nule. Keďže matica  $\Sigma_{\mathbf{YY}}$  je podľa predpokladu regulárna, tak platí

$$0 \neq \det(\Sigma_{\mathbf{YY}}^{-1}) = \det(\Sigma_{\mathbf{YY}}^{-1/2}) \det(\Sigma_{\mathbf{YY}}^{-1/2}).$$

Ekvivalentne teda možno riešiť rovnicu

$$\begin{aligned} \det(\Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2}) \det(\Sigma_{\mathbf{Y}\mathbf{X}} \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \Sigma_{\mathbf{X}\mathbf{Y}} - \lambda^2 \Sigma_{\mathbf{Y}\mathbf{Y}}) \det(\Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2}) = \\ = \det(\Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \Sigma_{\mathbf{Y}\mathbf{X}} \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \Sigma_{\mathbf{X}\mathbf{Y}} \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} - \lambda^2 \mathbf{I}_s) = 0 \end{aligned}$$

Označme  $\mathbf{R}$  maticu  $\Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \Sigma_{\mathbf{Y}\mathbf{X}} \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \Sigma_{\mathbf{X}\mathbf{Y}} \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2}$ . Z predchádzajúceho vidno, že jej vlastné čísla sú taktiež riešeniami rovnice (22). Charakteristický polýnom tejto matice je polýnom v  $\lambda^2$  stupňa  $s$ , ktorý, keďže ide o kladne semidefinitnú maticu, má  $s$  reálnych koreňov  $\lambda_1^2 \geq \lambda_2^2 \geq \dots \geq \lambda_s^2$ , ku ktorým prislúcha ortonormálny systém vlastných vektorov  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_s$ . Tieto vlastné vektory spolu s vlastnými číslami teda riešia rovnicu

$$\Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \Sigma_{\mathbf{Y}\mathbf{X}} \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \Sigma_{\mathbf{X}\mathbf{Y}} \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \mathbf{v} = \lambda^2 \mathbf{v}.$$

Ak túto rovnicu pre násobíme maticou  $\Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2}$ , dostaneme

$$\Sigma_{\mathbf{Y}\mathbf{Y}}^{-1} \Sigma_{\mathbf{Y}\mathbf{X}} \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \Sigma_{\mathbf{X}\mathbf{Y}} \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \mathbf{v} = \lambda^2 \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \mathbf{v}. \quad (25)$$

Porovnaním rovnice (25) s rovnicou (22) vidno, že medzi vektormi  $\mathbf{v}$  a  $\mathbf{h}$  existuje vzťah  $\mathbf{h} = \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \mathbf{v}$ .

Keďže úlohou bolo maximalizovať koreláciu medzi  $\xi$  a  $\omega$   $\text{corr}(\xi, \omega) = \lambda = \mathbf{g}^T \Sigma_{\mathbf{X}\mathbf{Y}} \mathbf{h}$ , zvolíme teda  $\lambda = \lambda_1$ , čo je odmocnina najväčšieho vlastného čísla matice  $\mathbf{R}$ . Vektory kanonických korelácií  $\mathbf{g}$  a  $\mathbf{h}$  sú potom dané ako

$$\mathbf{g}_1 = c_1 \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \Sigma_{\mathbf{X}\mathbf{Y}} \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \mathbf{v}_1,$$

$$\mathbf{h}_1 = \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \mathbf{v}_1,$$

kde  $\mathbf{v}_1$  je ortonormálny vlastný vektor matice  $\mathbf{R}$  prislúchajúci vlastnému číslu  $\lambda_1^2$ .

Konštanta  $c_1$  je taká, aby  $\mathbf{g}^T \Sigma_{\mathbf{X}\mathbf{X}} \mathbf{g} = 1$  čiže aby platilo

$$c_1^2 \mathbf{v}_1^T \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \Sigma_{\mathbf{Y}\mathbf{X}} \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \Sigma_{\mathbf{X}\mathbf{X}} \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \Sigma_{\mathbf{X}\mathbf{Y}} \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \mathbf{v}_1 = 1.$$

Predošlú rovnicu môžno prepísať s použitím matice  $\mathbf{R}$

$$c_1^2 \mathbf{v}_1^T \mathbf{R} \mathbf{v}_1 = 1. \quad (26)$$

Keďže  $\mathbf{v}_1$  je vlastný vektor matice  $\mathbf{R}$  zodpovedajúci vlastnému číslu  $\lambda_1^2$  rovnicu (26) možno upraviť

$$c_1^2 \lambda_1^2 \mathbf{v}_1^T \mathbf{v}_1 = 1. \quad (27)$$

Vzhľadom na to, že  $\mathbf{v}_1$  je ortonormálny vektor a využitím predošlej rovnice nakoniec dostávame

$$c_1 = \frac{1}{\lambda_1}. \quad (28)$$

Poznajúc prvú dvojicu kanonických premenných  $(\xi_1, \omega_1)$ , hľadáme ďalšiu dvojicu lineárnych kombinácií  $\xi = \mathbf{g}^T \mathbf{X}$  a  $\omega = \mathbf{h}^T \mathbf{Y}$ , ktorých koreláciu  $\mathbf{g}^T \Sigma_{\mathbf{XY}} \mathbf{h}$  maximalizujeme. Okrem podmienok  $\mathbf{g}^T \Sigma_{\mathbf{XX}} \mathbf{g} = 1$  a  $\mathbf{h}^T \Sigma_{\mathbf{YY}} \mathbf{h} = 1$  navyše vyžadujeme, aby  $\xi$  bolo nekorelované s  $\xi_1$  a aby  $\omega$  bolo nekorelované s  $\omega_1$ , t.j. aby platilo  $\mathbf{g}^T \Sigma_{\mathbf{XX}} \mathbf{g}_1 = \mathbf{h}^T \Sigma_{\mathbf{YY}} \mathbf{h}_1 = 0$ . Navyše pre  $\mathbf{g}_1$ ,  $\mathbf{h}_1$  a  $\lambda_1$  z rovnice (15) platí:

$$\Sigma_{\mathbf{XY}} \mathbf{h}_1 = \lambda_1 \Sigma_{\mathbf{XX}} \mathbf{g}_1,$$

z čoho spolu s podmienkou  $\mathbf{g}^T \Sigma_{\mathbf{XX}} \mathbf{g}_1 = 0$  vyplýva

$$\text{corr}(\xi, \omega_1) = \mathbf{g}^T \Sigma_{\mathbf{XY}} \mathbf{h}_1 = \lambda_1 \mathbf{g}^T \Sigma_{\mathbf{XX}} \mathbf{g}_1 = 0. \quad (29)$$

Podobne z rovnice (16) dostávame

$$\Sigma_{\mathbf{YX}} \mathbf{g}_1 = \lambda_1 \Sigma_{\mathbf{YY}} \mathbf{h}_1$$

a využitím predpokladu  $\mathbf{h}^T \Sigma_{\mathbf{YY}} \mathbf{h}_1 = 0$  nakoniec máme

$$\text{corr}(\omega, \xi_1) = \mathbf{h}^T \Sigma_{\mathbf{YX}} \mathbf{g}_1 = \lambda_1 \mathbf{h}^T \Sigma_{\mathbf{YY}} \mathbf{h}_1 = 0. \quad (30)$$

Na nájdenie druhého páru kanonických premenných teda hľadáme vektory  $\mathbf{g}$ ,  $\mathbf{h}$  také, ktoré maximalizujú funkciu

$$f(\mathbf{g}, \mathbf{h}) = \mathbf{g}^T \Sigma_{\mathbf{XY}} \mathbf{h} - \frac{1}{2} \lambda (\mathbf{g}^T \Sigma_{\mathbf{XX}} \mathbf{g} - 1) - \frac{1}{2} \mu (\mathbf{h}^T \Sigma_{\mathbf{YY}} \mathbf{h} - 1) + \eta \mathbf{g}^T \Sigma_{\mathbf{XX}} \mathbf{g}_1 + \nu \mathbf{h}^T \Sigma_{\mathbf{YY}} \mathbf{h}_1,$$

kde  $\lambda, \mu, \nu, \eta$  sú Lagrangeove multiplikátory.

Z nutných podmienok pre viazaný extrém vyplýva

$$\frac{\partial f}{\partial \mathbf{g}} = \Sigma_{\mathbf{XY}} \mathbf{h} - \lambda \Sigma_{\mathbf{XX}} \mathbf{g} + \eta \Sigma_{\mathbf{XX}} \mathbf{g}_1 = \mathbf{0}, \quad (31)$$

$$\frac{\partial f}{\partial \mathbf{h}} = \Sigma_{\mathbf{YX}} \mathbf{g} - \mu \Sigma_{\mathbf{YY}} \mathbf{h} + \nu \Sigma_{\mathbf{YY}} \mathbf{h}_1 = \mathbf{0}. \quad (32)$$

Po pre násobení prvej rovnice vektorom  $\mathbf{g}_1^T$  zľava a druhej rovnice vektorom  $\mathbf{h}_1^T$  taktiež zľava dostaneme:

$$\mathbf{g}_1^T \Sigma_{\mathbf{XY}} \mathbf{h} - \lambda \mathbf{g}_1^T \Sigma_{\mathbf{XX}} \mathbf{g} + \eta \mathbf{g}_1^T \Sigma_{\mathbf{XX}} \mathbf{g}_1 = 0,$$

$$\mathbf{h}_1^T \Sigma_{YX} \mathbf{g} - \mu \mathbf{h}_1^T \Sigma_{YY} \mathbf{h} + \nu \mathbf{h}_1^T \Sigma_{YY} \mathbf{h}_1 = 0.$$

Keďže  $\mathbf{g}_1^T \Sigma_{XX} \mathbf{g}_1 = 1$ ,  $\mathbf{h}_1^T \Sigma_{YY} \mathbf{h}_1 = 1$  a  $\mathbf{g}^T \Sigma_{XX} \mathbf{g}_1 = \mathbf{h}^T \Sigma_{YY} \mathbf{h}_1 = \mathbf{0}$  a navyše platia vzťahy (29) a (30), tak  $\eta = 0$  a  $\nu = 0$ . Rovnice (31) a (32) preto možno zjednodušiť na nasledujúci tvar

$$\Sigma_{XY} \mathbf{h} - \lambda \Sigma_{XX} \mathbf{g} = \mathbf{0}, \quad (33)$$

$$\Sigma_{YX} \mathbf{g} - \mu \Sigma_{YY} \mathbf{h} = \mathbf{0}, \quad (34)$$

pričom tieto vzťahy sú zhodné s rovnicami (15) a (16). Korelačný koeficient medzi  $\xi_2$  a  $\omega_2$  je teda rovný odmocnine druhého najväčšieho vlastného čísla matice  $\mathbf{R}$  a vektory  $\mathbf{g}_2$  a  $\mathbf{h}_2$  sú dané ako

$$\mathbf{g}_2 = c_2 \Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-1/2} \mathbf{v}_2, \quad \mathbf{h}_2 = \Sigma_{YY}^{-1/2} \mathbf{v}_2, \quad (35)$$

kde  $\mathbf{v}_2$  je ortonormálny vlastný vektor matice  $\mathbf{R}$  zodpovedajúci vlastnému číslu  $\lambda_2^2$  a konštanta  $c_2 = \frac{1}{\lambda_2}$ , aby  $\mathbf{g}_2^T \Sigma_{XX} \mathbf{g}_2 = \mathbf{1}$ . Podobným spôsobom možno nájsť  $k = \min(r, s)$  dvojíc premenných  $(\xi_1, \omega_1), \dots, (\xi_k, \omega_k)$ , pričom  $j$ -ta dvojica je tvorená vektormi  $\mathbf{g}_j = c_j \Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-1/2} \mathbf{v}_j$ ,  $\mathbf{h}_j = \Sigma_{YY}^{-1/2} \mathbf{v}_j$ , kde  $\mathbf{v}_j$  je vlastný vektor zodpovedajúci vlastnému číslu  $\lambda_j^2$  a  $c_j = \frac{1}{\lambda_j}$ .

### 3.1.2 Definícia kanonických premenných a výberové kanonické korelácie

Uvažujme náhodné vektory  $\mathbf{X} = (X_1, \dots, X_r)^T$  a  $\mathbf{Y} = (Y_1, \dots, Y_s)^T$ , ktorých kovariančná matica má tvar

$$\Sigma = \begin{pmatrix} \Sigma_{\mathbf{X}\mathbf{X}} & \Sigma_{\mathbf{X}\mathbf{Y}} \\ \Sigma_{\mathbf{Y}\mathbf{X}} & \Sigma_{\mathbf{Y}\mathbf{Y}} \end{pmatrix},$$

pričom predkladáme, že matice  $\Sigma_{\mathbf{X}\mathbf{X}}$  a  $\Sigma_{\mathbf{Y}\mathbf{Y}}$  sú regulárne. Označme

$$\mathbf{B} = \Sigma_{\mathbf{X}\mathbf{X}}^{-1/2} \Sigma_{\mathbf{X}\mathbf{Y}} \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2}$$

$$\mathbf{R} = \mathbf{B}^T \mathbf{B} = \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \Sigma_{\mathbf{Y}\mathbf{X}} \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \Sigma_{\mathbf{X}\mathbf{Y}} \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2}$$

Matica  $\mathbf{R}$  je kladne semidefinitná a nech jej hodnosť je  $m$ . Keďže je kladne semidefinitná a teda aj symetrická, k jej vlastným číslam vieme podľa spektrálnej vety vybrať ortonormálny systém vlastných vektorov.

**Definícia 3.1.** *Nech  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$  je ortonormálny systém vlastných vektorov matice  $\mathbf{R}$ , ktorým zodpovedajú nenulové vlastné čísla  $\lambda_1^2 \geq \dots \geq \lambda_m^2 > 0$ . Pre  $i = 1, \dots, m$  nech*

$$\mathbf{g}_i = \frac{\Sigma_{\mathbf{X}\mathbf{X}}^{-1} \Sigma_{\mathbf{X}\mathbf{Y}} \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \mathbf{v}_i}{\lambda_i}, \quad \mathbf{h}_i = \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \mathbf{v}_i,$$

$$\xi_i = \mathbf{g}_i^T \mathbf{X}, \quad \omega_i = \mathbf{h}_i^T \mathbf{Y},$$

$$\rho_i = \lambda_i$$

Potom vektory  $\mathbf{g}_1, \dots, \mathbf{g}_m, \mathbf{h}_1, \dots, \mathbf{h}_m$  nazveme vektory kanonických korelácií, náhodné premenné  $\xi_1, \dots, \xi_m, \omega_1, \dots, \omega_m$  nazveme kanonické premenné a čísla  $\rho_1, \dots, \rho_m$  nazveme kanonické korelácie.

Pre takto definované kanonické premenné platí nasledujúca veta:

**Veta 3.1.** *Pre všetky  $i, j \in \{1, \dots, m\}$  platí*

$$\text{cov}(\xi_i, \xi_j) = \text{cov}(\omega_i, \omega_j) = \begin{cases} 1 & \text{ak } i = j \\ 0 & \text{ak } i \neq j \end{cases} \quad (36)$$

$$\text{cov}(\xi_i, \omega_j) = \text{corr}(\xi_i, \omega_j) = \begin{cases} \rho_i & \text{ak } i = j \\ 0 & \text{ak } i \neq j \end{cases} \quad (37)$$

Dôkaz: Najprv dokážeme rovnosť (36), pričom využijeme ortonormalitu vektorov  $\mathbf{v}_i, \mathbf{v}_j$

$$\begin{aligned} \text{cov}(\xi_i, \xi_j) &= \mathbf{g}_i^T \Sigma_{\mathbf{X}\mathbf{X}} \mathbf{g}_j = \frac{\mathbf{v}_i^T \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \Sigma_{\mathbf{Y}\mathbf{X}} \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \Sigma_{\mathbf{X}\mathbf{X}} \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \Sigma_{\mathbf{X}\mathbf{Y}} \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \mathbf{v}_j}{\lambda_i \lambda_j} \\ &= \frac{\mathbf{v}_i^T \mathbf{R} \mathbf{v}_j}{\lambda_i \lambda_j} = \begin{cases} \frac{\lambda_j^2}{\lambda_j^2} = 1 & \text{ak } i = j \\ \frac{0}{\lambda_i \lambda_j} = 0 & \text{ak } i \neq j \end{cases} \end{aligned}$$

Podobne

$$\text{cov}(\omega_i, \omega_j) = \mathbf{h}_i^T \Sigma_{\mathbf{Y}\mathbf{Y}} \mathbf{h}_j = \mathbf{v}_i^T \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \Sigma_{\mathbf{Y}\mathbf{Y}} \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \mathbf{v}_j = \mathbf{v}_i^T \mathbf{v}_j = \begin{cases} 1 & \text{ak } i = j \\ 0 & \text{ak } i \neq j \end{cases}$$

Pričom posledná rovnosť opäť vyplýva z ortonormality vektorov  $\mathbf{v}_i$  a  $\mathbf{v}_j$ . Túto vlastnosť využijeme aj na dôkaz rovnosti (37)

$$\begin{aligned} \text{cov}(\xi_i, \omega_j) &= \text{corr}(\xi_i, \omega_j) = \mathbf{g}_i^T \Sigma_{\mathbf{X}\mathbf{Y}} \mathbf{h}_j = \frac{\mathbf{v}_i^T \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \Sigma_{\mathbf{Y}\mathbf{X}} \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \Sigma_{\mathbf{X}\mathbf{Y}} \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1/2} \mathbf{v}_j}{\lambda_i} = \\ &= \frac{\mathbf{v}_i^T \mathbf{R} \mathbf{v}_j}{\lambda_i} = \begin{cases} \frac{\lambda_j^2}{\lambda_i} = \rho_i & \text{ak } i = j \\ 0 & \text{ak } i \neq j \end{cases} \end{aligned}$$

□

*Poznámka:* Pri odvodzovaní metódy sme ukázali, že kanonické premenné  $\xi_1 = \mathbf{g}_1^T \mathbf{X}$  a  $\omega_1 = \mathbf{h}_1^T \mathbf{Y}$  majú najväčší korelačný koeficient spomedzi všetkých dvojíc netriviálnych lineárnych kombinácií  $\mathbf{g}^T \mathbf{X}$  a  $\mathbf{h}^T \mathbf{Y}$ . Pre  $k \geq 2$  majú kanonické premenné  $\xi_k = \mathbf{g}_k^T \mathbf{X}$  a  $\omega_k = \mathbf{h}_k^T \mathbf{Y}$  najväčší korelačný koeficient spomedzi všetkých dvojíc netriviálnych lineárnych kombinácií  $\mathbf{g}^T \mathbf{X}$  a  $\mathbf{h}^T \mathbf{Y}$ , ktoré sú nekorelované s  $\mathbf{g}_1^T \mathbf{X}, \dots, \mathbf{g}_{k-1}^T \mathbf{X}$  a  $\mathbf{h}_1^T \mathbf{Y}, \dots, \mathbf{h}_{k-1}^T \mathbf{Y}$ .

Keďže v skutočnosti kovariančnú maticu vektorov  $\mathbf{X}$  a  $\mathbf{Y}$  nepoznáme, používame pre výpočty výberovú kovariančnú maticu. Nech  $\mathbf{X}_1, \dots, \mathbf{X}_n$  je  $n$  nezávislých pozorovaní vektora  $\mathbf{X}$  a  $\mathbf{Y}_1, \dots, \mathbf{Y}_n$  je  $n$  nezávislých pozorovaní vektora  $\mathbf{Y}$ . Potom výberová kovariančná matica má tvar

$$\hat{\Sigma} = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T & \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T \\ \sum_{i=1}^n (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{X}_i - \bar{\mathbf{X}})^T & \sum_{i=1}^n (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T \end{pmatrix} \quad (38)$$

V tomto prípade nezáleží na tom, či použijeme maticu  $\hat{\Sigma}$  alebo maticu  $\mathbf{S} = \frac{n}{n-1} \hat{\Sigma}$ , keďže počas výpočtov sa deliteľ  $n$  a  $n-1$  vykrátia.

Navyše, na rozdiel od PCA, majú kanonické korelácie veľkú výhodu v tom, že sú invariantné vzhľadom na regulárne lineárne transformácie vektorov  $\mathbf{X}$  a  $\mathbf{Y}$ . Ak teda vezmeme  $r \times r$  maticu  $\mathbf{D}$  a  $s \times s$  maticu  $\mathbf{F}$ , potom vektory  $\mathbf{DX}$  a  $\mathbf{FY}$  majú rovnaké kanonické korelácie ako vektory  $\mathbf{X}$  a  $\mathbf{Y}$ . Dôsledkom toho nezáleží, či na výpočet kanonických korelácií použijeme maticu kovariančnú alebo korelačnú, dostane totiž rovnaký výsledok.

### 3.1.3 Príklad použitia CCA

Tak ako pri analýze hlavných komponentov aj pre analýzu kanonických korelácií sme sa rozhodli uviesť jednoduchý príklad na lepšie pochopenie metódy. Tentokrát sme použili dáta detroit, ktoré sú dostupné na stránke [15]. V týchto dátach je viacero premenných, my sme sa však rozhodli vybrať z nich len niektoré. Za vektor nezávislých premenných sme zvolili  $\mathbf{X} = [\text{FPT}, \text{UEMP}, \text{MAN}, \text{WE}]^T$ , kde premenná FPT hovorí o množstve policajtov pracujúcich na plný úväzok na 100 000 obyvateľov, premenná UEMP je nezamestnanosť, MAN počet obyvateľov, ktorý pracujú v priemyselnej výrobe v tisícoch a WE je priemerná týždenná mzda. Za vektor závislých premenných sme vzali vektor  $\mathbf{Y} = [\text{HOM}, \text{ACC}, \text{ASR}]^T$ , kde HOM je počet vrážd na 100 000 obyvateľov, ACC meria úmrtnosť pri nehodách na 100 000 obyvateľov a ASR je počet prepadnutí opäť na sto tisíc obyvateľov. Vektor  $\mathbf{X}$  teda meria sociálne podmienky v Detrote, zatiaľ čo vektor  $\mathbf{Y}$  hovorí o množstve kriminálnych činov v tomto meste. Keďže si myslíme, že sociálne podmienky ovplyvňujú kriminalitu, aplikovali sme na tieto dáta CCA a pokúsili sme sa nájsť lineárne závislosti medzi kanonickými premennými.

V tabuľke 7 sú v stĺpcoch uvedené vektory kanonických korelácií pre vektor  $\mathbf{X}$  a v tabuľke 8 sú uvedené vektory kanonických korelácií pre vektor  $\mathbf{Y}$ .

Po aplikovaní metódy sa treba rozhodnúť, ktoré kanonické premenné obsahujú dôležitú informáciu. V [3] autori odporúčajú použiť kombináciu troch kritérií. V prvom rade sa pozrieme na veľkosť získaných kanonických korelácií. Chceme, aby tento koeficient bol čo najbližšie k číslu 1, avšak neexistuje presné pravidlo, ktoré by nám povedalo, kedy je tento koeficient dostatočne veľký. Autori tiež odporúčajú pozrieť sa na hladinu významnosti jednotlivých kanonických korelácií. Funkcia `canoncorr(X,Y)` v matlabe

|      | $\mathbf{g}_1$ | $\mathbf{g}_2$ | $\mathbf{g}_3$ |
|------|----------------|----------------|----------------|
| FTP  | -0.0161        | -0.0129        | -0.0185        |
| UEMP | -0.1312        | -0.1073        | -0.4764        |
| MAN  | -0.0094        | 0.0163         | -0.0299        |
| WE   | 0.0017         | 0.0052         | 0.0496         |

**Tabuľka 7:** Koeficienty výberových vektorov kanonických korelácií pre nezávislé premenné

|     | $\mathbf{h}_1$ | $\mathbf{h}_2$ | $\mathbf{h}_3$ |
|-----|----------------|----------------|----------------|
| HOM | -0,0575        | -0,0176        | -0,1408        |
| ACC | -0.0043        | 0.2575         | 0.2445         |
| ASS | -0.0011        | 0.0093         | 0.0420         |

**Tabuľka 8:** Koeficienty výberových vektorov kanonických korelácií pre závislé premenné

podáva informáciu, ktorá sa týka testovania hypotéz  $H_0^{(k)}$  :  $k + 1$ -vá až  $m$ -tá korelácia sú všetky nulové pre  $k = 0, \dots, m - 1$ . Okrem toho nám pri rozhodovaní môže pomôcť aj index redundancie, ktorý meria schopnosť nezávislých premenných vysvetliť variáciu jednotlivých závislých premenných. Jeho výpočet možno rozdeliť do troch krokov. V prvom kroku sa vypočíta množstvo spoločnej variácie množiny závislých premenných zahrnutéj v závislej kanonickej premennej. V druhom kroku sa vypočíta množstvo variácie v závislej kanonickej premennej, ktorá môže byť vysvetlená pomocou nezávislej kanonickej premennej. Nakoniec spočítame index redundancie ako súčin týchto dvoch čísel. Chceme aby tento index bol čo najbližšie k číslu 1, avšak tak ako pri kanonických koreláciách ani pre tento index neexistuje presné pravidlo, ktoré by hovorilo, kedy je dostatočne veľký.

V tabuľke 9 sú uvedené všetky tri kritériá, ktoré sme získali pri riešení nášho problému. Keďže kanonické korelácie sú pre prvé dva páry kanonických premenných blízko k číslu 1 a boli zamietnuté hypotézy o ich nulovosti, usúdili sme, že by mohli obsahovať pre nás zaujímavé informácie.

Možno si všimnúť, že pre prvú dvojicu kanonických premenných  $(\xi_1, \omega_1)$  je kanonická korelácia  $\rho_1$  rovná takmer 1 a teda by medzi týmito premennými mala existovať takmer lineárna závislosť. V premennej  $\xi_1$  má celkom veľkú váhu veľkosť nezamestnanosti,

|                            | $(\xi_1, \omega_1)$     | $(\xi_2, \omega_2)$ | $(\xi_3, \omega_3)$ |
|----------------------------|-------------------------|---------------------|---------------------|
| kanonické korelácie        | 0.9884                  | 0.9549              | 0.4291              |
| p hodnota pre F štatistiku | $2.8399 \times 10^{-6}$ | 0.0021              | 0.4431              |
| index redundancie          | 0.5681                  | 0.3222              | 0.0120              |

**Tabuľka 9:** Hodnoty kritérií

ktorá má rovnaké znamienko ako koeficienty vektora  $h_1$ . S rastom nezamestnanosti teda očakávame aj rast úrovne kriminality. Aj pre druhý pár kanonických korelácií  $(\xi_2, \omega_2)$  sme získali vysokú hodnotu kanonickej korelácie  $\rho_2$ . Vidíme, že pre kanonickú premennú  $\xi_2$  má opäť veľkú váhu miera nezamestnanosti, pre premenú  $\omega_2$  má teraz najväčšiu váhu premenná ACC, čiže počet nehôd.

## 3.2 Aplikácie CCA

Analýza kanonických korelácií sa používa na riešenie problémov v ekonometrii, financiách, geografii či ekológii. V [9] je táto metóda použitá na skúmanie vzťahov medzi ekonomickým vývojom a dopravou, v [18] sú kanonické korelácie použité pri analýze reči. Medzi spôsoby využitia CCA patrí taktiež dôležitý špeciálny prípad - koeficient mnohonásobnej korelácie [4], ktorý predstavíme v nasledujúcej podkapitole. Pri aplikácii sme sa tentokrát inšpirovali [10].

### 3.2.1 Koeficient mnohonásobnej korelácie

Uvažujme špeciálny prípad analýzy kanonických korelácií, kedy  $s$ , čiže rozmer vektora  $\mathbf{Y}$ , je rovný 1. V matici  $\Sigma$  je potom  $\Sigma_{\mathbf{Y}\mathbf{Y}}$  reálne číslo, ktoré pre jednoduchosť označíme  $\sigma_Y^2$  a  $\Sigma_{\mathbf{X}\mathbf{Y}}$  je vektor s rozmermi  $r \times 1$ , ktorý označíme  $\mathbf{a}$ . Matica  $\mathbf{R}$  v definícii 3.1 má v takomto prípade rozmery  $1 \times 1$ , čiže sa jedná o reálne číslo

$$\mathbf{R} = \frac{\mathbf{a}^T \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{a}}{\sigma_Y^2},$$

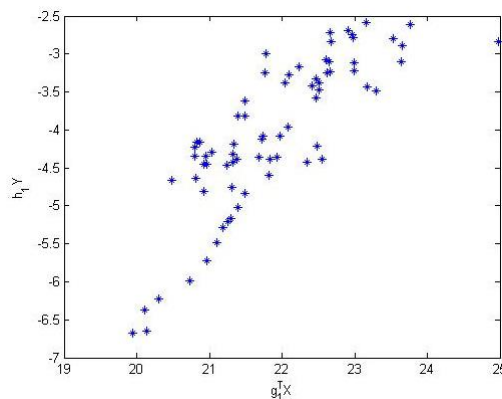
a za predpokladu  $\Sigma_{\mathbf{X}\mathbf{X}}$  je kladne definitná, hodnosť  $\mathbf{R}$  je jedna. Existuje preto len jeden pár kanonických premenných  $(\xi_1, \omega_1)$ , pričom vektory kanonických korelácií majú tvar  $\mathbf{g}_1 = \pm \frac{\Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{a}}{\sigma_Y \lambda_1}$  a  $\mathbf{h}_1 = \pm \frac{1}{\sigma_Y}$ . Jedinú kanonickú koreláciu  $\rho_1$  v takomto prípade nazveme koeficient mnohonásobnej korelácie a vypočítame ju pomocou vzťahu

$$\rho_1 = \frac{\sqrt{\mathbf{a}^T \Sigma_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{a}}}{\sigma_Y}.$$

Koeficient mnohonásobnej korelácie je úzko spätý s lineárnou regresiou  $y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_r x_{ir} + \epsilon_i$ , kde  $\epsilon_i$  sú nekorelované chyby s rovnakou disperziou. Ak vezmeme vektory vysvetľujúcich premenných za realizácie náhodného vektora  $\mathbf{X}$ , tak druhá mocnina výberového koeficientu mnohonásobnej korelácie je  $R^2$ , čiže koeficientu determinácie.

Pre ilustráciu uvažujme náhodnú premennú  $Y$  je cena akcií spoločnosti Vodafone na londýnskej burze a náhodný vektor  $\mathbf{X} = [\text{inflácia}, \text{úrok}, \text{nezamestnanosť}, \text{výmenný kurz}, \text{cena ropy}]$ , kde inflácia hovorí o výške inflácie vo Veľkej Británii, úrok je výška úroku na peňažnom trhu, nezamestnanosť je percento nezamestnaných taktiež vo Veľkej Británii a výmenný kurz je výška výmenného kurzu medzi librou a eurom. Zaujímá nás teda, či existuje lineárna závislosť medzi makroekonomickými premennými a cenami akcií

danej spoločnosti. Potrebné dáta sme našli na stránkach: [19], [5], [17] a [6]. Pre túto aplikáciu sme využili mesačné dáta za obdobie od 1. januára 2008 po 1. január 2014. Pre tieto dáta sme po aplikovaní analýzy kanonických korelácií, získali hodnotu koeficientu mnohonásobnej korelácie rovnú 0.8235, pričom bola zamietnutá hypotéza o nulovosti tohto koeficientu. Ďalšími výstupmi získanými z analýzy sú vektor  $\mathbf{g}_1 = [-23.6068, 153.4489, 179.6780, 11.9092, -0.0305]^T$  a číslo  $\mathbf{h}_1 = -0.0127$ . Možno si všimnúť, že cena akcií spoločnosti Vodafone rastie spolu s infláciou a cenami ropy a klesá s nezamestnanosťou, rastom úroku na peňažnom trhu a výškou výmenného kurzu. Takmer lineárnu závislosť možno pozorovať aj v grafe Obr.12, ktorý zobrazuje name-



**Obr. 12:** Namerané dáta vzhľadom na prvý pár výberových kanonických premenných

rané dáta vzhľadom na prvý pár výberových kanonických premenných  $(\xi_1, \omega_1)$ .

### 3.2.2 CCA a NHL

V tejto práci sme sa rozhodli vyskúšať aj vlastnú aplikáciu CCA. Pokúsili sme sa zistiť, čo ovplyvňuje defenzívne a ofenzívne schopnosti hráčov kanadsko-americkej národnej hokejovej ligy (NHL).

V prvom rade sme potrebovali zvoliť vektory premenných. Za vektor nezávislých premenných sme tentokrát zvolili vektor  $\mathbf{X} = [\text{salary}, \text{age}, \text{height}, \text{weight}]^T$ , pričom salary je mzda hráča v miliónoch dolárov v sezóne 2013/2014, age je jeho vek, height je výška hráča v centimetroch a weight je jeho váha v kilogramoch. Defenzívne schopnosti hráča sme merali počtom hitov (HITS) a počtom blokováných striel (BS). Za meradlo ofenzívnych schopností hráčov sme vzali počet strelených gólov (G) a počet asistencií (A). Keďže hráči z rôznych dôvodov odohrali rozdielne počty zápasov, považovali sme za vhodné predeliť tieto štatistiky počtom odohraných zápasov (GP). Zvolili sme teda  $\mathbf{Y} = [\text{HITS/GP}, \text{BS/GP}, \text{G/GP}, \text{A/GP}]^T$ . Aby sme získali pozorovania vektorov  $\mathbf{X}$  a  $\mathbf{Y}$ , pozreli sme sa na štatistiky hráčov dvoch tímov NHL: Pittsburghu Penguins a Edmontonu Oilers. Potrebné dáta pochádzajú z oficiálnych stránok týchto klubov ([1], [13]), ale tiež zo stránok [12] a [2].

Vďaka vlastnosti škálovej invariantnosti sme mohli CCA aplikovať priamo na dáta, rozdielne jednotky nami zvolených premenných nám preto neprekážajú. V tabuľke 10 sú uvedené vektory kanonických korelácií  $\mathbf{g}_i, i = 1, \dots, 4$  pre nezávislé premenné a tiež vektory kanonických korelácií  $\mathbf{h}_i, i = 1, \dots, 4$  pre závislé premenné. Vzhľadom na hodnoty kritérií v tabuľke 11, len pri prvom páre kanonických premenných je kanonická korelácia blízka číslu 1, pričom sme tiež zamietli hypotézu o jej nulovosti. Preto sme usúdili, že pre nás podstatnú informáciu v sebe nesie len prvý pár kanonických premenných  $(\xi_1, \omega_1)$ . Kanonická korelácia medzi týmito premennými je rovná 0.8889, čo je podľa nás dostatočne veľké číslo a teda medzi týmito premennými by mohol existovať lineárny vzťah.

Pri interpretácii vektorov kanonických premenných sa pozrieme na absolútnu hodnotu váh a tiež porovnáme znamienka jednotlivých koeficientov. Pri pohľade na vektor  $\mathbf{g}_1$  si možno všimnúť, že najväčšiu váhu má mzda hráčov. Tento fakt nie je až taký prekvapivý, keďže vysoký plat si musia hráči zaslúžiť, závisí od ich výkonov v predchádzajúcich sezónach a zároveň je motiváciou pre hráčov podávať dobré výkony. Pre závislé

|         | $\mathbf{g}_1$ | $\mathbf{g}_2$ | $\mathbf{g}_3$ | $\mathbf{g}_4$ |
|---------|----------------|----------------|----------------|----------------|
| wage    | 0.3890         | 0.0815         | -0.1814        | -0.0628        |
| age     | -0.0621        | 0.0190         | -0.1580        | 0.1450         |
| height  | -0.0019        | 0.2318         | 0.0021         | 0.0744         |
| weight  | -0.0437        | -0.0619        | -0.0599        | -0.1753        |
|         | $\mathbf{h}_1$ | $\mathbf{h}_2$ | $\mathbf{h}_3$ | $\mathbf{h}_4$ |
| HITS/GP | -0.1198        | -0.4623        | -0.4502        | -0.3689        |
| BS/GP   | 0.3434         | 1.5783         | -0.3852        | 0.5556         |
| G/GP    | 2.0150         | -1.0482        | -5.9621        | 11.4952        |
| A/GP    | 3.8649         | 0.8514         | 1.2829         | -6.9039        |

**Tabuľka 10:** Koefficienty vektorov kanonických korelácií

|                            | $(\xi_1, \omega_1)$     | $(\xi_2, \omega_2)$ | $(\xi_3, \omega_3)$ | $(\xi_4, \omega_4)$ |
|----------------------------|-------------------------|---------------------|---------------------|---------------------|
| kanonické korelácie        | 0.8889                  | 0.2960              | 0.2211              | 0.0187              |
| p hodnota pre F štatistiku | $1.2585 \times 10^{-7}$ | 0.8049              | 0.7557              | 0.9089              |
| redundancy index           | 0.3692                  | 0.0158              | 0.0130              | 0.0000              |

**Tabuľka 11:** Hodnoty kritérií

premenné majú väčšiu váhu góly na zápas a asistencie na zápas, ich váhy majú rovnaké znamienko ako váha priradená mzde, preto možno očakávať, že s rastom mzdy narastá aj počet strelených gólov a počet asistencií. V nami vybranej vzorke hráčov sa nachádzajú niektorí z najlepších útočníkov v NHL, akými sú Crosby, Malkin, Neal či Hall, ktorých platy sa pohybujú na úrovni niekoľkých miliónov dolárov, preto nie je prekvapivé, že existuje závislosť medzi mzdou a ofenzívnymi schopnosťami hráčov. Naopak malé váhy majú výška a váha hráča. To možno vysvetliť tým, že na veľkosti pri strielaní gólov a nahrávaní na góly nezáleží. Medzi hráčmi s vyšším počtom bodov sú napríklad Malkin, ktorý je vysoký a Neal, ktorý je na svoju výšku celkom ťažký, ale aj Eberle, ktorý patrí medzi hráčov nižšieho veku.

Keďže koefficienty vektorov kanonických korelácií, nazývané aj kanonické váhy, sú nestabilné a menia sa s rôznymi vzorkami dát, na interpretáciu kanonických korelácií sa používajú aj záťaže kanonických premenných (koefficienty kanonickej štruktúry, cano-

nical loadings) a krížové kanonické záťaže (canonical cross-loadings). Záťaže kanonických premenných odzrkadľujú varianciu, ktorú pozorovaná závislá premenná zdieľa so závislou kanonickou premennou a podobne aj pre nezávislé premenné. Pri krížových kanonických záťažiach sa hľadá korelácia jednotlivých závislých premenných s nezávislou kanonickou premennou a naopak.

Pre naše dáta sme získali vektor kanonických záťaží pre nezávislé premenné

$$\mathbf{L}_\xi = [0.8970, -0.3036, -0.2494, -0.3980]^T$$

a vektor loadings pre závislé premenné

$$\mathbf{L}_\omega = [-0.3833, -0.2870, 0.8308, 0.9744]^T.$$

Tieto naznačujú, že tvar prvej kanonickej premennej  $\xi_1$  je ovplyvnený najmä premennou wage, zatiaľ čo na tvar premennej  $\omega_1$  majú vplyv premenné G/GP a A/GP. Podobne sme spočítali aj krížové kanonické záťaže, pričom pre nezávislé premenné a kanonickú premennú  $\omega_1$  sme dostali vektor

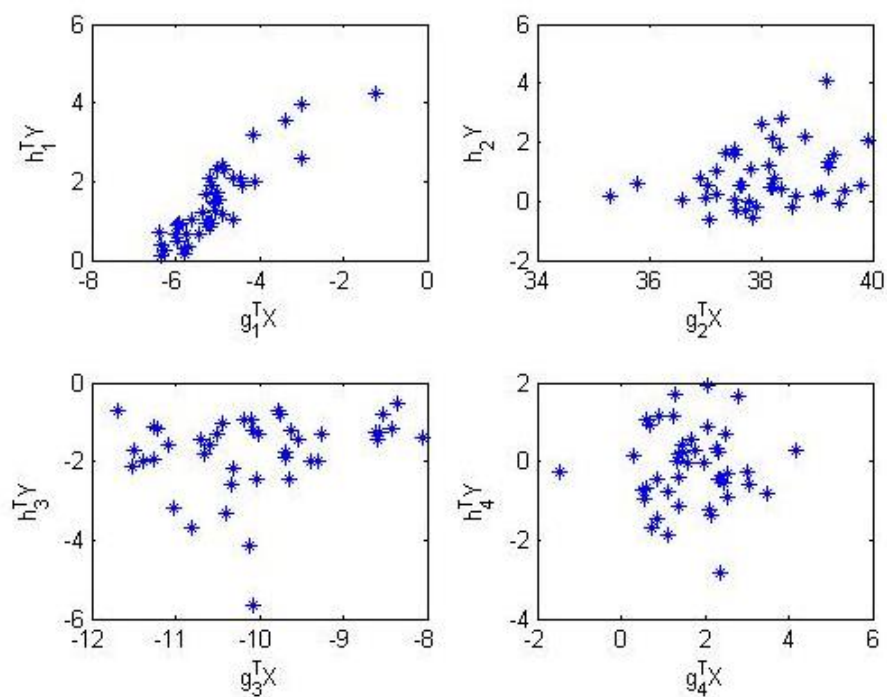
$$\mathbf{CL}_\omega = [0.7973, -0.2699, -0.2217, -0.3538],$$

zatiaľ čo krížové kanonické záťaže pre závislé premenné a kanonickú premennú  $\xi_1$  majú tvar

$$\mathbf{CL}_\xi = [-0.3407, -0.2552, 0.7085, 0.8661].$$

Tieto vektory len potvrdzujú naše tušenie, že medzi nezávislou premennou wage a kanonickou premennou  $\omega_1$  existuje závislosť. Taktiež kanonická premenná  $\xi_1$  je spätá so závislými premennými G/GP a A/GP.

V našom prípade teda kanonické váhy viedli k rovnakým záverom ako kanonické záťaže. Samozrejme možno očakávať, že pre inú vzorku hráčov by sme získali iné výsledky, nami použité dáta nás však priviedli k záveru, že medzi mzdami hráčov a počtom nimi strelených gólov a počtom prihrávk by mohol existovať lineárny vzťah. V obrázku 13 sú postupne vyobrazené namerané dáta vzhľadom na páry kanonických premenných  $(\xi_i, \omega_i)$  pre  $i = 1, \dots, 4$ . Z týchto vidno, že medzi prvým párom kanonických premenných by mohla existovať lineárna závislosť, no tak ako už naznačili hodnoty výberových kanonických premenných, zvyšné tri pári kanonických premenných sú zrejme veľmi slabo korelované.



**Obr. 13:** Zobrazenie dát vzhľadom na kanonické premenné

## Záver

Cieľom bakalárskej práce bolo naštudovať a spracovať metódy na lineárnu redukciu dát, konkrétne analýzu hlavných komponentov a analýzu kanonických korelácií a následne tieto metódy aplikovať. Prácu sme rozdelili na tri hlavné kapitoly. V prvej kapitole sme zhrnuli dôležité poznatky lineárnej algebry týkajúce sa teórie vlastných čísel a vlastných vektorov, vrátane spektrálnej vety (veta o hlavných osiach) (Veta 1.2).

Analýze hlavných komponentov bola venovaná druhá kapitola. V tejto kapitole sme najprv opísali podrobné odvodenie PCA, pričom sme využívali najmä zdroje [8] a [7], a uviedli sme definíciu hlavných komponentov (Definícia 2.1). Pre lepšie pochopenie metódy sme uviedli jednoduchý príklad. V druhej časti tejto kapitoly bolo naším cieľom poukázať na možnosti aplikácií PCA. Najprv sme ukázali, inšpirovaní [11], ako možno PCA využiť pri spracovaní obrazu. Vybraný obrázok sme aproximovali využitím vektorov skóre, pričom sme získali celkom dobrú aproximáciu čierneho bielej verzie obrázka. Tiež sme ukázali ako možno obrázok využitím hlavných komponentov a vzťahu (12) postupne zrekonštruovať. Druhá aplikácia PCA je spojená s cenami akcií na trhu. Vytvorili sme jednoduché portfólio skladajúce sa z akcií šiestich spoločností, ktoré sú obchodovateľné na burze v New Yorku. Keďže sledovať ceny všetkých týchto akcií je celkom náročné, aplikovali sme na dáta PCA, aby sme zmenšili počet premenných. V prípade použitia kovariančnej matice prvé dva hlavné komponenty zodpovedali viac ako 92% celkovej variancie v dátach a preto ich už pomerne dobre opisujú. Tento príklad sme však zvolili hlavne preto, aby sme ukázali, že v niektorých prípadoch možno niekoľko prvých hlavných komponentov aj interpretovať. Nám sa to podarilo pre prvé dva hlavné komponenty.

Tretia kapitola sa venovala analýze kanonických korelácií. Opäť sme najprv opísali podrobné odvodenie metódy, pričom sme využili najmä zdroje [7] a [18] a následne sme uviedli definíciu kanonických premenných (Definícia 3.1). Tak ako pri PCA, aj pri tejto metóde sme najskôr uviedli jednoduchý príklad a následne dve aplikácie. Prvou z nich je špeciálny prípad kanonických korelácií, tzv. koeficient mnohonásobnej korelácie. V tomto prípade sme hľadali lineárnu závislosť medzi vektorovou premennou, hodnotami rôznych makroekonomických ukazovateľov, a číselnou premennou - cenami akcií spoločnosti Vodafone. Zaujímalo nás, ako makroekonomické premenné vo Veľkej

Británii ovplyvňujú ceny akcií tejto spoločnosti na burze v Londýne. Druhá aplikácia CCA je naším vlastným nápadom. Snažili sme sa zistiť čo ovplyvňuje defenzívne a ofenzívne schopnosti hráčov v kanadsko-americkej národnej hokejovej lige. Pri tejto aplikácii sme využili dáta spojené s dvomi hokejovými klubmi Edmontonom Oilers a Pittsburghom Penguins. Naše výsledky naznačujú, že existuje vzťah najmä medzi mzdou hráča a počtom gólov a asistencií. Treba však vziať do úvahy fakt, že výsledky sú značne ovplyvnené výberom dát a pre iné tímy tejto ligy by sme mohli získať iné výsledky.

Táto bakalárska práca umožnila autorovi rozšíriť si vedomosti v oblasti štatistických metód, priniesla tiež možnosť dané metódy aplikovať a zamýšľať sa nad interpretáciou výsledkov. Prínosom práce pre čitateľa je najmä podrobné spracovanie daných metód od ich základu v lineárnej algebre až po ich použitie v rôznych oblastiach praxe. Práca oboznamuje čitateľa s možnosťami lineárnej redukcie dát, ako aj s možnosťou využitia metód v reálnom živote.

## Zoznam použitej literatúry

- [1] Edmonton Oilers, oficiálna stránka klubu, dostupné na internete (23.5.2014):  
<http://oilers.nhl.com/>
- [2] ESPN: The Worldwide Leader In Sports, dostupné na internete (23.5.2014):  
<http://espn.go.com/nhl/>
- [3] Hair J. F. a kol.: Multivariate Data Analysis, 5th edition, Prentice Hall, 1998, dostupné na internete (22.5.2014):  
[http://www.mvstats.com/Downloads/Supplements/Canonical\\_Correlation\\_6e.pdf](http://www.mvstats.com/Downloads/Supplements/Canonical_Correlation_6e.pdf)
- [4] Harman, R.: Mnohorozmerné štatistické analýzy, učebné texty, KAMŠ, FMFI UK, Bratislava, 2014
- [5] Inflation Rate and CPI, dostupné na internete (23.5.2014):  
<http://www.rateinflation.com/>
- [6] Investing.com: a definitive source for tools and information relating to financial markets, dostupné na internete (23.5.2014): <http://www.investing.com/>
- [7] Izenman, A. J.: Modern Multivariate Statistical Techniques, Springer Science + Business Media, New York, 2008
- [8] Jolliffe, I. T.: Principal Component Analysis, Springer-Verlag, New York, 2002
- [9] Loo, B.P.Y., An application of canonical correlation analysis in regional science: The interrelationships between transport and development in China's Zhujiang delta, Journal of regional science 40 (2000), 143-171, dostupné na internete (23.5.2014): <http://onlinelibrary.wiley.com/doi/10.1111/0022-4146.00168/pdf>
- [10] Mazuruse, P.: Canonical Correlation analysis: Macroeconomic variables versus stock returns, Journal of Financial Economic Policy 6 (2014), dostupné na internete: <http://www.emeraldinsight.com/journals.htm?issn=1757-6385&volume=6&issue=2&articleid=17109469&show=abstract>
- [11] Mudrová, M., Procházka A.: Principal component analysis in image processing, Proc. of Int. Conf. Technical Computing 2005, 15.11.2005, Praha,

- str. MUD/1-MUD/4, Humusoft Praha, 2005, dostupné na internete (22.5.2014):  
[http://dsp.vscht.cz/konference\\_matlab/matlab05/prispevky/mudrova/mudrova.pdf](http://dsp.vscht.cz/konference_matlab/matlab05/prispevky/mudrova/mudrova.pdf)
- [12] NHLNumbers.com: National Hockey League Salaries and Stats, dostupné na internete (23.5.2014): <http://nhlnumbers.com/>
- [13] Pittsburgh Penguins, oficiálna stránka klubu, dostupné na internete (23.5.2014):  
<http://penguins.nhl.com/>
- [14] Seber, G. A. F.: Multivariate Observation, John Wiley and Sons, Hoboken, 2004
- [15] StatLib—Datasets Archive, dostupné na internete (23.5.2014):  
<http://lib.stat.cmu.edu/datasets/>
- [16] Strang, G.: Linear Algebra and its Applications, Thompson Learning, USA, 1988
- [17] Štatistická databáza Eurostatu, dostupné na internete (23.5.2014):  
[http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search\\_database](http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database)
- [18] Weenink, D.,: Canonical correlation analysis, Proceedings 25 (2003), 81-99, dostupné na internete (22.5.2014):  
[http://www.fon.hum.uva.nl/Proceedings/Proceedings25/cca\\_ifaproc2003.pdf](http://www.fon.hum.uva.nl/Proceedings/Proceedings25/cca_ifaproc2003.pdf)
- [19] Yahoo! finance- Business Finance, Stock Market, Quotes, News:  
<http://finance.yahoo.com/>
- [20] Zlatoš, P.: Lineárna algebra a geometria, Marenčin PT, Bratislava, 2011

## Príloha A - zdrojové kódy programov

### Zdrojový kód - spracovanie obrazu

```
function [Vuspo, l2] = obr(a)
O = imread(a, 'jpg');
t = size(O);
R = O(:, :, 1);
r = zeros(t(1) * t(2), 1);
for j = 1 : t(2)
    for i = 1 : t(1)
        r(t(1) * (j - 1) + i) = R(i, j);
    end
end

G = O(:, :, 2);
g = zeros(t(1) * t(2), 1);
for j = 1 : t(2)
    for i = 1 : t(1)
        g(t(1) * (j - 1) + i) = G(i, j);
    end
end

B = O(:, :, 3);
b = zeros(t(1) * t(2), 1);
for j = 1 : t(2)
    for i = 1 : t(1)
        b(t(1) * (j - 1) + i) = B(i, j);
    end
end

X = [r, g, b];
C = cov(X);
```

---

```

rp = mean(r);
gp = mean(g);
bp = mean(b);
rm = (r - rp);
gm = (g - gp);
bm = (b - bp);

[V, L] = eig(C);
l = diag(L);
[l2, in] = sort(-l);
l2 = -l2;
kumul = l2/sum(l2)
for k = 1 : length(in)
    Vuspo(:, k) = V(:, in(k));
end

Y = Vuspo' * [rm'; gm'; bm'];
Xpruh = diag([rp, gp, bp]) * ones(3, length(r));

o1 = Xpruh + Vuspo(:, 1) * Y(1, :);
rr = reshape(o1(1, :), t(1), t(2));
gg = reshape(o1(2, :), t(1), t(2));
bb = reshape(o1(3, :), t(1), t(2));
O1(:, :, 1) = rr;
O1(:, :, 2) = gg;
O1(:, :, 3) = bb;
o2 = o1 + Vuspo(:, 2) * Y(2, :);
rr = reshape(o2(1, :), t(1), t(2));
gg = reshape(o2(2, :), t(1), t(2));
bb = reshape(o2(3, :), t(1), t(2));
O2(:, :, 1) = rr;

```

```

O2(:, :, 2) = gg;
O2(:, :, 3) = bb;
o3 = o2 + Vuspo(:, 3) * Y(3, :);
rr = reshape(o3(1, :), t(1), t(2));
gg = reshape(o3(2, :), t(1), t(2));
bb = reshape(o3(3, :), t(1), t(2));
O3(:, :, 1) = rr;
O3(:, :, 2) = gg;
O3(:, :, 3) = bb;
sed1 = reshape(Y(1, :), t(1), t(2));
sed2 = reshape(Y(2, :), t(1), t(2));
sed3 = reshape(Y(3, :), t(1), t(2));
subplot(1, 3, 1), subimage(preprocet(O1))
subplot(1, 3, 2), subimage(preprocet(O2))
subplot(1, 3, 3), subimage(preprocet(O3))
%subplot(1, 2, 2), subimage(rgb2gray(preprocet(O1)));
%subplot(1, 2, 1), subimage(rgb2gray(O));
%subplot(1, 2, 2), subimage(preprocet(sed1));
%subplot(1, 3, 2), subimage(preprocet(sed2));
%subplot(1, 3, 3), subimage(preprocet(sed3));

function [x] = preprocet(x)
x = (x - min(min(min(x))))/(max(max(max(x))) - min(min(min(x))));

```

### Zdrojový kód - kanonické korelácie

```

function[g, h, kor, stats, redIndex, loadingsNEZavPre,
loadingsZavPre, CloadingsNEZavPre, CloadingsZavPre] = KanKorelacie2(X, Y)
[g, h, kor, U, V, stats] = canoncorr(X, Y);
p = size(X);
q = size(Y);
n = p(1);

```

---

```

roz = [p(2), q(2)];
for j = 1 : p(1)
    xi(:, j) = g' * (X(j, :))';
    omega(:, j) = h' * (Y(j, :))';
end
for l = 1 : min(roz)
    xi(l, :) = g(:, l)' * X';
end
xi = xi';
for l = 1 : min(roz)
    omega(l, :) = h(:, l)' * Y';
end
omega = omega';
for i = 1 : min(roz)
    for j = 1 : q(2)
        maticaKorel2(j, i) = corr(omega(:, i), Y(:, j))^2;
    end
end
sharedVar = mean(maticaKorel2);
explainedVar = kor.^2;
redIndex = sharedVar. * explainedVar;
for i = 1 : min(roz)
    for j = 1 : q(2)
        loadingsZavPre(j, i) = corr(omega(:, i), Y(:, j));
    end
end
for i = 1 : min(roz)
    for j = 1 : p(2)
        loadingsNEZavPre(j, i) = corr(xi(:, i), X(:, j));
    end
end
end

```

```
for i = 1 : min(roz)
    for j = 1 : p(2)
        CloadingsNEZavPre(j,i) = corr(omega(:,i), X(:,j));
    end
end
for i = 1 : min(roz)
    for j = 1 : q(2)
        CloadingsZavPre(j,i) = corr(xi(:,i), Y(:,j));
    end
end
```