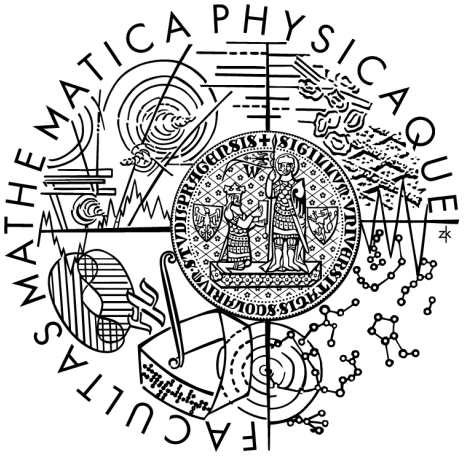
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Mária Dubová

Metoda hlavních komponent a její aplikace

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	Mgr. Radek Hendrych
Studijní program:	Matematika
Studijní obor:	Finanční matematika

Praha 2011

Rada by som poďakovala vedúcemu mojej bakalárskej práce Mgr. Radkovi Hendrychovi za pripomienky, návrhy, cenné rady a odborné vedenie pri písaní tejto práce. Ďalej chcem poďakovať tým, ktorí mi pomáhali s korektúrami textu.

Prehlasujem, že som túto bakalársku prácu vypracovala samostatne a výhradne s použitím citovaných prameňov, literatúry a ďalších odborných zdrojov.

Beriem na vedomie, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., autorského zákona v platnom znení, najmä skutočnosť, že Univerzita Karlova v Prahe má právo na uzavretie licenčnej zmluvy o užití tejto práce ako školného diela podľa §60 odst. 1 autorského zákona.

V Prahe dňa 9.12.2011 Mária Dubová

Názov práce: Metoda hlavních komponent a její aplikace
Autor: Mária Dubová, DubovaMaria@seznam.cz
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedúci bakalárskej práce: Mgr. Radek Hendrych, hendrych@karlin.mff.cuni.cz, Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V predloženej práci sa zaoberáme metódou hlavných komponentov. V prvej časti textu študujeme hlavné komponenty z rôznych aspektov, ako napríklad ich odvodenie pre viacrozmerný náhodný vektor z obecného rozdelenia alebo rozlíšime ich výpočet na základe kovariančnej či korelačnej matice. Dôležitý je taktiež správny výber počtu hlavných komponentov, čím efektívne znížime počet dimenzií dát pri snahe zachovať čo najväčšie množstvo informácie. Teoretické znalosti podkladáme ilustračnými príkladmi. V druhej časti sa zameriavame na hodnotu v riziku. Tento pojem je v práci definovaný spolu so vzťahmi na jej výpočet. Ďalej venujeme pozornosť praktickej aplikácií tohoto konceptu a metódy hlavných komponentov v prípade úrokových mier s rôznou dobou splatnosti, čo následne využijeme k výpočtu hodnoty v riziku pre rozličné portfólia.

Kľúčové slová: metóda hlavných komponentov, hlavné komponenty, variancia, hodnota v riziku

Title: Principal components analysis and its applications
Author: Mária Dubová, DubovaMaria@seznam.cz
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Radek Hendrych, hendrych@karlin.mff.cuni.cz, Department of Probability and Mathematical Statistics

Abstract: In the present thesis, we deal with the principal components analysis. In the first of this text, we study different aspects of principals components, for instance, their derivation for a multidimensional random vector from general distribution or their calculation based on a covariance or correlation matrix. It is also important to choose the proper number of principal components for reducing the dimensionality of data in order to preserve most of information. Theoretical knowledge are illustrated with several examples. In the second part of the thesis, we focus on the value at risk. This term is defined in the text also with several usual formulas to calculate it. Then, we deal with a practical application of this concept and the principal component analysis. Concretely, we analyse the portfolio of some different interest rates to obtain the value at risk in some cases.

Keywords: principal component analysis, principal components, variance, value at risk

Obsah

Úvod

Metóda hlavných komponentov je pravdepodobne najstaršou a najznámejšou technikou využívanou v mnohorozmernej analýze. Základnou myšlienkou metódy je zredukovať dimenziu množiny dát, v ktorej sa nachádza množstvo vzájomne súvisiacich premenných, pri zachovaní maximálnej možnej variancií pôvodných dát. To sa dosiahne transformovaním množiny na nové premenné nazývané hlavné komponenty, ktoré sú vzájomne nekorelované a sú zoradené tak, že prvých niekoľko komponentov vyjadruje väčšinu variancie.

Cieľom tejto práce je popísať základné princípy a vlastnosti metódy hlavných komponentov a následne aplikovať jej použitie na reálnych ekonomických dátach.

Prvá kapitola predstavuje teoretickú časť práce s ilustračnými príkladmi. Najskôr popíšeme celkové odvodenie hlavných komponentov spolu so špecifikáciou ich rozptylov a potom sa dotkneme dvoch komplikácií, ktoré pri tom môžu nastať. Ďalej sa zaoberáme centrovanými hlavnými komponentami a ich vlastnostiam, ktoré ukážeme aj na príklade. Následne formulujeme výberové hlavné komponenty za pomoci výberovej kovariančnej matice a výberového priemeru, taktiež spomenieme asymptotické vlastnosti hlavných komponentov. Na záver prvej kapitoly sa venujeme vybraným problémom hlavných komponentov, s ktorými sa môžeme stretnúť počas analýzy dát. Pozrieme sa na to, ako zistíme pomer variancie pôvodných dát zachytený komponentami. Rozlíšime medzi metódou hlavných komponentov aplikovanou na kovariančnú a korelačnú maticu. Priblížime problém mierky použitej na meranie vstupných dát. Ďalej si ukážeme tri najpoužívanejšie metódy na zistenie vhodného počtu hlavných komponentov pre zachytenie určitej miery variancie dát. Nezabudneme ani na geometrickú interpretáciu tejto metódy a na záver si povieme niečo málo o súvislosti s faktorovou analýzou.

Druhá kapitola je zameraná na hodnotu v riziku a s tým súvisiacu praktickú aplikáciu metódy hlavných komponentov. V úvode túto štatistickú mieru rizika definujeme a poskytneme niekoľko štandardných vzorcov na priamy výpočet. Ďalšia časť je najmä praktického charakteru. Na súbor úrokových mier aplikujeme metódu hlavných komponentov a za pomoci jej výsledkov prezentujeme výpočet hodnoty v riziku pre rôzne portfólia.

Väčšina značenia je prebraná z knihy Dupač & Hušková (1999), prípadné výnimky sú v texte explicitne vysvetlené. Všetky obrázky, grafy a analýzy sú vytvorené v programe Wolfram Mathematica 8.0. V práci budeme značiť symbolom $\triangle$ koniec príkladu a symbolom $\square$ koniec dôkazu.

1. Metóda hlavných komponentov

Metóda hlavných komponentov (Principal Component Analysis, PCA) je zrejme najstaršou a najznámejšou metódou používanou v mnohorozmernej analýze. Umožňuje efektívne transformovať dáta z viacrozmerného vstupného priestoru do priestoru s nižšou dimenziou, čím sa dáta premietnu do najdôležitejších lineárnych smerov a tie najmenej podstatné sa zanedbajú.

Redukcia dimenzie dát má z pohľadu štatistického rozpoznávania veľký význam. Umožňuje nám znížiť počet súradníc potrebných na efektívny popis dát. Takto môžeme ignorovať smery, v ktorých majú dáta len malú varianciu (rozlíšiteľnosť) a všímať si len tie s veľkou varianciou. V tomto prípade je teda nositeľom informácie rozptyl.

Metóda hlavných komponentov je teda metóda, ktorá umožňuje vytvárať nové premenné, ktoré sú lineárnou kombináciou premenných z pôvodných dát, pri čo najmenšej strate informácií. Pôvodných p vzájomne korelovaných (pozorovaných) premenných je nahradených novými q ($q << p$) vzájomne nekorelovanými (ortogonálnymi) nemerateľnými syntetickými premennými tak, že prvá nová súradnicová os (prvý hlavný komponent) je vedená v smere maximálnej variancie medzi objektmi. Druhá os (druhý hlavný komponent) je kolmá na prvú os a je vedená v smere druhej najväčšej variancie medzi objektmi, atď.

Relatívna pozícia objektov v pôvodnom priestore a v novom priestore (danom hlavnými komponentami) je rovnaká. To znamená, že pôvodný súradnicový systém sa natáča do smeru maximálnej variancie medzi objektmi, pričom euklidovské vzdialenosti medzi objektmi sa zachovávajú.

Metódu hlavných komponentov pôvodne navrhol už v roku 1901 Karl Pearson ako opisnú štatistickú metódu na redukciu viacrozmerných údajov. Harold Hotelling v roku 1933 zovšeobecnil postup metódy hlavných komponentov na náhodné vektory a navrhol použitie tejto metódy na rozbor kovariančnej štruktúry premenných.

1.1 Odvodenie hlavných komponentov

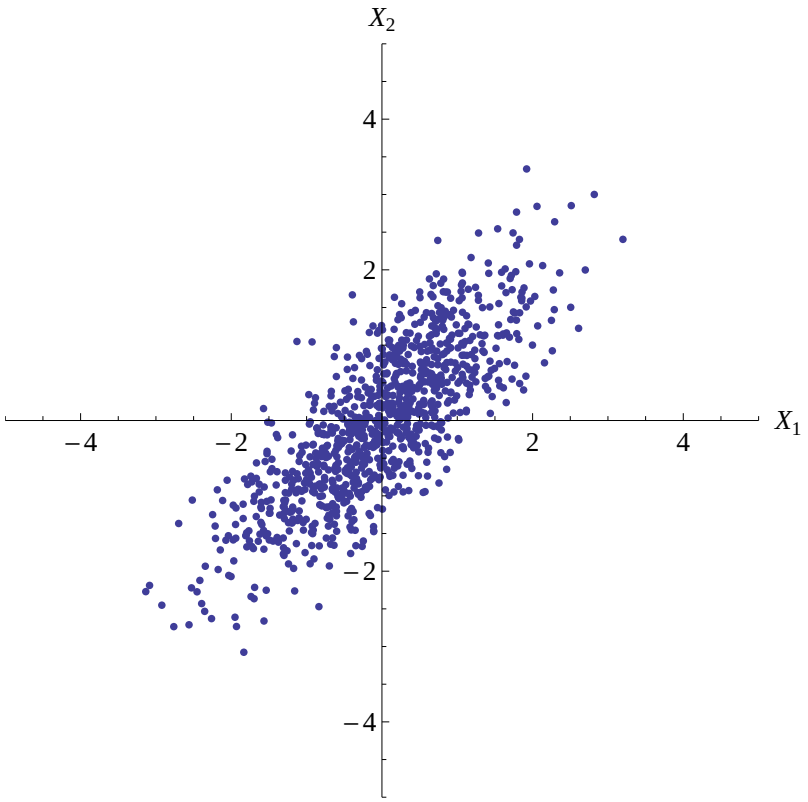
Vezmime reálny p-rozmerný náhodný vektor $\mathbf{X}$, tj. $\mathbf{X} = (X_1, X_2, ..., X_p)^T$. Nech $\mathbf{X}$ má strednú hodnotu $\boldsymbol{\mu}$ a kovariančnú maticu $\boldsymbol{\Sigma}$, značíme $\mathbf{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Našim cieľom je nájsť nový p-rozmerný náhodný vektor, ktorý budeme značiť $\mathbf{Y} = (Y_1, Y_2, ..., Y_p)^T$ a ktorého zložky sú nekorelované a ich variancia postupne klesá, tj. $Var(Y_1) \geq Var(Y_2) \geq ... \geq Var(Y_p)$. Každú zložku Y_j vyjadríme prostredníctvom lineárnej kombinácie prvkov vektora $\mathbf{X}$, teda

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{pj}X_p = \mathbf{a}_j^T\mathbf{X},$$

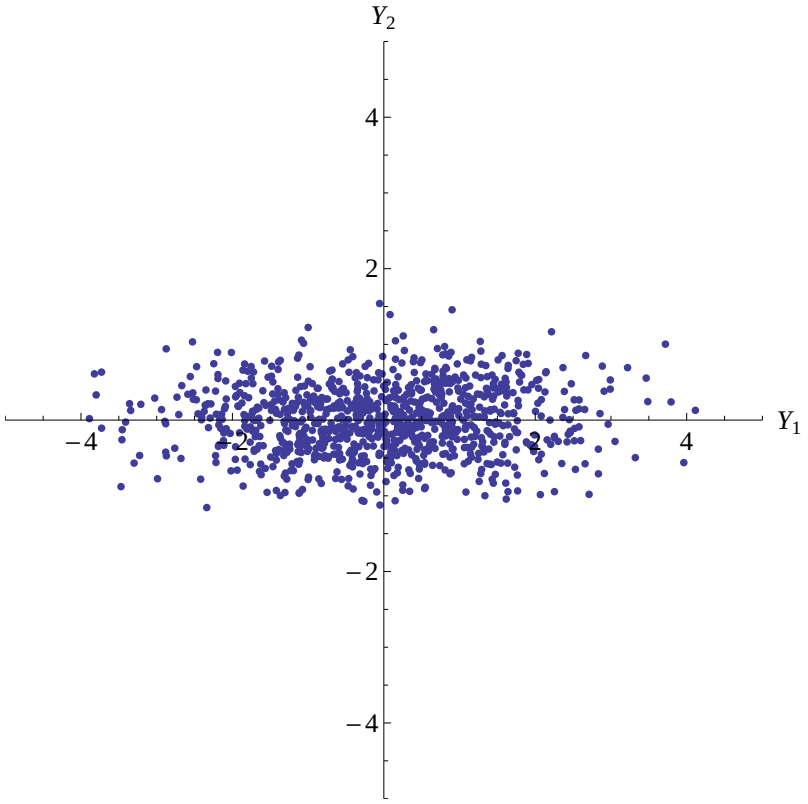
kde $\mathbf{a}_j = (a_{1j}, a_{2j}, ..., a_{pj})^T$ je vektor konštánt. Navyše kladieme podmienku, že

$$\mathbf{a}_j^T\mathbf{a}_j = \sum_{k=1}^{p} a_{kj}^2 = 1.$$

Tento konkrétny normalizačný postup zaisťuje, že celková transformácia je ortonormálna, takže vzdialenosti v p-rozmernom euklidovskom priestore sú zachované.



Obr. 1.1: 1000 pozorovaní náhodného vektora (X_1, X_2)



Obr. 1.2: 1000 pozorovaní z obrázku 1.1 zobrazených vzhľadom k ich hlavným komponentom (Y_1, Y_2)

Jednoduchý prípad si môžeme ukázať v dvojrozmernom euklidovskom priestore. Na obrázku 1.1 vidíme graf 1000 pozorovaní dvoch silno korelovaných náhodných veličín X_1 a X_2. Pozorovania boli náhodne vygenerované z dvojrozmerného normálneho rozdelenia s nulovou strednou hodnotou a kovariančnou maticou

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \frac{4}{5} \\ \frac{4}{5} & 1 \end{pmatrix}.$$

Metódou hlavných komponentov transformujeme veličiny na hlavné komponenty (Y_1, Y_2) a dostaneme usporiadanie na grafe obrázku 1.2. Vidíme, že najviac variancie je zachytenej hlavným komponentom Y_1, naopak Y_2 zachytáva málo variancie.

Teraz si vyložíme postup na odvodenie hlavných komponentov, inšpirujeme sa pri tom publikáciami Jolliffe (2002) a Chatfield & Collins (2000). Prvý hlavný komponent Y_1 nájdeme tak, že $\mathbf{a}_1$ položíme také, aby Y_1 malo čo najväčší možný rozptyl. Inými slovami, vyberieme $\mathbf{a}_1$ so snahou maximalizovať varianciu $\mathbf{a}_1^T\mathbf{X}$ za podmienky, že $\mathbf{a}_1^T\mathbf{a}_1 = 1$. Druhý hlavný komponent určíme výberom $\mathbf{a}_2$ tak, aby Y_2 malo druhú najväčšiu možnú varianciu a aby bolo nekorelované s Y_1. Podobným spôsobom odvodíme $Y_3, Y_4, ..., Y_p$ so snahou zachovať vzájomnú nekorelovanosť a aby mali postupne klesajúci rozptyl. Dúfame však, že väčšina variancie je vyjadrená v prvých q hlavných komponentoch, kde $q << p$, $p, q \in \mathbb{N}$.

Najprv uvažujme $Y_1 = \mathbf{a}_1^T\mathbf{X}$. Hľadajme vektor $\mathbf{a}_1$ maximalizujúci $Var(\mathbf{a}_1^T\mathbf{X}) = \mathbf{a}_1^T\mathbf{\Sigma}\mathbf{a}_1$ za platnosti normalizačnej podmienky $\mathbf{a}_1^T\mathbf{a}_1 = 1$. Štandardným postupom pri maximalizácií funkcie o niekoľkých premenných, za daných podmienok, je použitie metódy Lagrangeových multiplikátorov. Pre $\mathbf{a}_1$ si teda definujme Lagrangeovu funkciu

$$L(\mathbf{a}_1) = \mathbf{a}_1^T\mathbf{\Sigma}\mathbf{a}_1 - \lambda(\mathbf{a}_1^T\mathbf{a}_1 - 1).$$

Funkciu zderivujeme podľa vektora $\mathbf{a}_1$:

$$\frac{\partial L}{\partial \mathbf{a}_1} = 2\mathbf{\Sigma}\mathbf{a}_1 - 2\lambda\mathbf{a}_1.$$

Výsledok derivácie položíme rovno nulovému vektoru a dostaneme

$$(\mathbf{\Sigma} - \lambda\mathbf{I}_p)\mathbf{a}_1 = \mathbf{0}, \tag{1.1}$$

kde $\mathbf{I}_p$ je jednotková matica $(p \times p)$. Aby táto rovnica mala nenulové riešenie pre vektor $\mathbf{a}_1$, musí byť $(\mathbf{\Sigma} - \lambda\mathbf{I}_p)$ singulárna matica. Takže λ zvolíme tak, aby

$$|\mathbf{\Sigma} - \lambda\mathbf{I}_p| = 0.$$

Odtiaľ plynie, že nenulové riešenie rovnice (1.1) existuje práve vtedy, keď λ je vlastné číslo matice $\mathbf{\Sigma}$. Ale $\mathbf{\Sigma}$ má p vlastných čísel, ktoré musia byť všetky nezáporné, pretože $\mathbf{\Sigma}$ je pozitívne semidefinitná (jedná sa o kovariančnú maticu). Označme si vlastné čísla $\lambda_1, \lambda_2, ..., \lambda_p$ a uvažujme ich postupné usporiadanie $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Z rovnice (1.1) vidíme, že

$$\mathbf{a}_1^T\mathbf{\Sigma}\mathbf{a}_1 = \mathbf{a}_1^T\lambda\mathbf{I}_p\mathbf{a}_1 = \lambda\mathbf{a}_1^T\mathbf{a}_1 = \lambda,$$

takže za vlastné číslo, ktoré dáva $\mathbf{a}_1^T\mathbf{X}$ s najväčšou varianciou vezmeme λ_1. Potom podľa (1.1) musí byť $\mathbf{a}_1$ vlastný vektor matice $\mathbf{\Sigma}$ prislúchajúci najväčšiemu vlastnému číslu.

Druhý hlavný komponent $Y_2 = \mathbf{a}_2^T\mathbf{X}$ získame podobným spôsobom. Okrem prvej podmienky, $\mathbf{a}_2^T\mathbf{a}_2 = 1$, musí Y_2 spĺňať ešte druhú a to je nekorelovanosť s Y_1. Potrebujeme teda, aby kovariancia medzi hlavnými komponentami Y_2 a Y_1 bola nulová, tj.

$$Cov(Y_2, Y_1) = Cov(\mathbf{a}_2^T\mathbf{X}, \mathbf{a}_1^T\mathbf{X}) = E[\mathbf{a}_2^T(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\mathbf{a}_1] = \mathbf{a}_2^T\mathbf{\Sigma}\mathbf{a}_1 = 0. \quad (1.2)$$

Z vlastností vlastných čísel a vektorov matice vieme, že $\mathbf{\Sigma}\mathbf{a}_1 = \lambda_1\mathbf{a}_1$, takže po zjednodušení výrazu (1.2) dostávame podmienku, že $\mathbf{a}_2^T\mathbf{a}_1 = 0$. Použijeme teda dva Lagrangeove multiplikátory, ktoré označíme λ a δ a definujeme funkciu

$$L(\mathbf{a}_2) = \mathbf{a}_2^T\mathbf{\Sigma}\mathbf{a}_2 - \lambda(\mathbf{a}_2^T\mathbf{a}_2 - 1) - \delta\mathbf{a}_2^T\mathbf{a}_1.$$

V stacionárnom bode musí platiť

$$\frac{\partial L}{\partial \mathbf{a}_2} = 2(\mathbf{\Sigma} - \lambda\mathbf{I}_p)\mathbf{a}_2 - \delta\mathbf{a}_1 = \mathbf{0}. \quad (1.3)$$

Prenásobením rovnice prvkom $\mathbf{a}_1^T$ dostaneme

$$2\mathbf{a}_1^T\mathbf{\Sigma}\mathbf{a}_2 - \delta = 0,$$

pretože $\mathbf{a}_1^T\mathbf{a}_2 = 0$ a $\mathbf{a}_1^T\mathbf{a}_1 = 1$. Z rovnice (1.2) vieme, že $\mathbf{a}_1^T\mathbf{\Sigma}\mathbf{a}_2$ má byť nula, takže δ je v stacionárnom bode rovno 0. Ďalej z rovnice (1.3) plynie, že

$$(\mathbf{\Sigma} - \lambda\mathbf{I}_p)\mathbf{a}_2 = \mathbf{0}.$$

V tomto prípade vyberáme za λ druhé najväčšie vlastné číslo matice $\mathbf{\Sigma}$ a $\mathbf{a}_2$ je príslušný vlastný vektor. λ sa už nemôže rovnať λ_1, pretože ak by sa tak stalo, dostali by sme rovnosť $\mathbf{a}_1 = \mathbf{a}_2$ a v tom prípade sme nesplnili podmienku $\mathbf{a}_1^T\mathbf{a}_2 = 0$.

Takýmto spôsobom sa dá ukázať, že pre všetky ostatné hlavné komponenty sú $\mathbf{a}_3, \mathbf{a}_4, ..., \mathbf{a}_p$ vlastné vektory matice $\mathbf{\Sigma}$ príslušné vlastným číslam $\lambda_3, \lambda_4, ..., \lambda_p$.

Označme si maticu vlastných vektorov $(p\times p)$ symbolom $\mathbf{A}$. Keďže $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_p$ sú normované a ortogonálne vektory, potom matica $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_p)$ je ortonormálna. Pre $\mathbf{A}$ tak platí, že $\mathbf{A}^T = \mathbf{A}^{-1}$, teda $\mathbf{A}^T\mathbf{A} = \mathbf{I}_p$. Ďalej označme vektor hlavných komponentov $(p \times 1)$ symbolom $\mathbf{Y}$, tj.

$$\mathbf{Y} = \mathbf{A}^T\mathbf{X}. \quad (1.4)$$

Kovariančná matica vektora $\mathbf{Y}$ je taktiež rozmerov $(p \times p)$ a budeme ju značiť $\mathbf{\Lambda}$. Hlavné komponenty sme volili tak, aby boli vzájomne nekorelované a vzhľadom ku konštrukcii teda platí $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, ..., \lambda_p)$. Strednú hodnotu a rozptyl vektora $\mathbf{Y}$ je možné vyjadriť v maticovom tvare:

$$\begin{aligned} E(\mathbf{Y}) &= E(\mathbf{A}^T\mathbf{X}) = \mathbf{A}^T\boldsymbol{\mu} \\ Var(\mathbf{Y}) &= Var(\mathbf{A}^T\mathbf{X}) = \mathbf{A}^T\mathbf{\Sigma}\mathbf{A} = \mathbf{\Lambda}. \end{aligned}$$

Poslednú rovnosť vo vyjadrení rozptylu $\mathbf{Y}$ môžeme vďaka vlastnostiam matice $\mathbf{A}$ prepísať následovne

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T. \tag{1.5}$$

V tomto prípade hovoríme o spektrálnom rozklade kovariančnej matice $\mathbf{\Sigma}$.

Táto problematika je sformulovaná tiež v knihe Härdle & Simar (2003) nasledujúcou vetou:

Veta 1.1.1. Ak $\mathbf{C}$ a $\mathbf{D}$ sú symetrické matice $(p \times p)$, $p \in \mathbb{N}$, $\mathbf{D}$ je pozitívne definitná a $\boldsymbol{x} \in \mathbb{R}^p$, potom maximum $\boldsymbol{x}^T\mathbf{C}\boldsymbol{x}$ za podmienky $\boldsymbol{x}^T\mathbf{D}\boldsymbol{x} = 1$ je dané najväčším vlastným číslom maticového súčinu $\mathbf{D}^{-1}\mathbf{C}$. Obecne:

$$\max_{\{\boldsymbol{x}:\ \boldsymbol{x}^T\mathbf{D}\boldsymbol{x}=1\}} \boldsymbol{x}^T\mathbf{C}\boldsymbol{x} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p = \min_{\{\boldsymbol{x}:\ \boldsymbol{x}^T\mathbf{D}\boldsymbol{x}=1\}} \boldsymbol{x}^T\mathbf{C}\boldsymbol{x},$$

kde $\lambda_1, \lambda_2, ..., \lambda_p$ sú vlastné čísla matice $\mathbf{D}^{-1}\mathbf{C}$. Vektor maximalizujúci (minimalizujúci) $\boldsymbol{x}^T\mathbf{C}\boldsymbol{x}$ za podmienky $\boldsymbol{x}^T\mathbf{D}\boldsymbol{x} = 1$ je vlastný vektor súčinu matíc $\mathbf{D}^{-1}\mathbf{C}$, ktorý prislúcha najväčšiemu (najmenšiemu) vlastnému číslu súčinu $\mathbf{D}^{-1}\mathbf{C}$.

Dôkaz vety 1.1.1 môžeme nájsť v spomínanej knihe. Avšak, ak si vhodne priradíme označenie z vety 1.1.1 k označeniu, ktoré sme použili v predošlom texte zistíme, že dôkaz vety plynie z predchádzajúceho odvodenia hlavných komponentov. Za maticu $\mathbf{C}$ zvolíme kovariančnú maticu $\mathbf{\Sigma}$, ktorá spĺňa predpoklad symetrickosti. Potom teda symbolu $\boldsymbol{x}$ priradíme vektor $\mathbf{a}_j$. Nakoniec za maticu $\mathbf{D}$ zvolíme jednotkovú maticu $\mathbf{I}_p$ rozmerov $(p \times p)$, ktorá taktiež splňuje požadované predpoklady, jednotková matica je symetrická a pozitívne definitná.

Hlavné komponenty teda získame lineárnou kombináciou pôvodných premenných. Hlavné komponenty sú vzájomne nekorelované a ich rozptyly vyjadrujú množstvo informácie, ktoré v sebe nesie vstupný súbor premenných. Pre efektívny popis týchto dát postačuje niekoľko prvých hlavných komponentov s najväčšou varianciou.

Poznámka 1.1.1. Pri metóde hlavných komponentov sa môžeme stretnúť s dvoma problémami, ktoré sa vyskytujú skôr v teoretickej rovine ale v praxi sú zväčša neobvyklé. Tieto situácie sú taktiež popísané v knihe Jolliffe (2002).

Komplikácia môže nastať, keď sa niektoré vlastné čísla rovnajú nule. Ak teda pre nejaké $k \in \{1, 2, ..., p\}$, $p \in \mathbb{N}$, platí $\lambda_k = Var(Y_k) = 0$, potom $Y_k = konst$, s.u. Dimenzia euklidovského priestoru obsahujúceho pozorovania je rovnaká ako hodnosť matice $\mathbf{\Sigma}$ a tá je daná hodnotou $(p - q)$, kde q je počet nulových vlastných čísel, $q << p,\ q \in \mathbb{N}$. Každý hlavný komponent s nulovým rozptylom teda definuje konštantný lineárny vzťah prvkov vektora $\mathbf{X}$. Jedna premenná je nepotrebná pre každý taký vzťah, pretože jej hodnota môže byť určená z hodnôt ostatných premenných uvedených vo vzťahu. Preto môžeme znížiť počet premenných z p na $(p - q)$ bez straty akejkoľvek informácie. V ideálnom prípade, by mala byť lineárna závislosť spozorovaná pred použitím metódy hlavných komponentov a počet premenných odpovedajúcim spôsobom znížiť.

Iný problém nastane, keď vlastné čísla a teda rozptyly niektorých hlavných komponentov budú rovnaké. Ak

$$\lambda_{q+1} = ... = \lambda_{q+k},$$

potom $\lambda = \lambda_{q+1}$ nazývame k-násobným koreňom. Príslušné vlastné vektory tvoria k-dimenzionálny priestor, v ktorom sú ortonormálne. Hlavný komponent s varianciou λ nie je jednoznačne určený. Dôležitým špeciálnym prípadom je, keď posledných k vlastných čísel je rovnakých. V takom prípade je posledných k hlavných komponentov považovaných za merania nešpecifikovanej variability a charakteristiky $\mathbf{X}$ sú reprezentované prvými $(p-k)$ hlavnými komponentami.

1.2 Vlastnosti centrovaných hlavných komponentov

Opäť predpokladajme, že pracujeme s p-rozmerným reálnym náhodným vektorom $\mathbf{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. V predchádzajúcej časti sme odvodili metódu hlavných komponentov. Teraz bližšie nahliadneme na centrované hlavné komponenty:

$$\mathbf{Y} = \mathbf{A}^T(\mathbf{X} - \boldsymbol{\mu}),$$

kde $\mathbf{A}$ značí ortonormálnu maticu vlastných vektorov kovariančnej matice $\boldsymbol{\Sigma}$. Je to rovnaké ako rovnica (1.4) ale okrem ortogonálnej rotácie navyše posunieme aj počiatok súradnicovej sústavy tak, aby $\mathbf{Y}$ mal nulový vektor stredných hodnôt.

Teraz si ukážeme niektoré vlastnosti centrovaných hlavných komponentov, ktoré sú taktiež uvedené napríklad v Härdle & Simar (2003). Pre daný p-rozmerný reálny vektor $\mathbf{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ majme vektor hlavných komponentov $\mathbf{Y} = \mathbf{A}^T(\mathbf{X} - \boldsymbol{\mu})$. Potom

$$\begin{aligned}
EY_j &= 0, \quad j = 1, 2, ..., p && (1.6)\\
Var(Y_j) &= \lambda_j, \quad j = 1, 2, ..., p && (1.7)\\
Cov(Y_i, Y_j) &= 0, \quad i \neq j && (1.8)\\
Cov(\mathbf{X}, \mathbf{Y}) &= \mathbf{A}\boldsymbol{\Lambda} && (1.9)\\
\sum_{j=1}^{p} Var(Y_j) &= tr(\boldsymbol{\Sigma}) && (1.10)\\
\rho_{X_i Y_j} &= a_{ij} \left(\frac{\lambda_j}{Var(X_i)} \right)^{1/2} && (1.11)
\end{aligned}$$

Postupne si tieto vzťahy dokážeme. Začneme prvou vlastnosťou (1.6):

$$E(\mathbf{Y}) = E(\mathbf{A}^T(\mathbf{X} - \boldsymbol{\mu})) = \mathbf{A}^T(E\mathbf{X} - \boldsymbol{\mu}) = \mathbf{0},$$

takže pre každé $j = 1, 2, ..., p$ platí $EY_j = 0$.

Rovnosť (1.7) obdržíme ako

$$\begin{aligned}
Var(\mathbf{Y}) &= Var(\mathbf{A}^T(\mathbf{X} - \boldsymbol{\mu})) = \mathbf{A}^T Var(\mathbf{X} - \boldsymbol{\mu})\mathbf{A}\\
&= \mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A} = \mathbf{A}^T\mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^T\mathbf{A} = \boldsymbol{\Lambda}.
\end{aligned}$$

Predposledná rovnosť plynie z rovnice (1.5) a posledná rovnosť plynie z ortonormality matice $\mathbf{A}$. Takže pre každé $j = 1, 2, ..., p$ platí $Var(Y_j) = \lambda_j$.

Rovnicu (1.8) dokážeme nasledujúcim spôsobom:

$$\begin{aligned} Cov(Y_i, Y_j) &= E(Y_iY_j) + E(Y_i)E(Y_j) = E(Y_iY_j) = E(\mathbf{a}_i^T(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\mathbf{a}_j) \\ &= \mathbf{a}_i^T\boldsymbol{\Sigma}\mathbf{a}_j = \mathbf{a}_i^T\lambda_j\mathbf{a}_j = \lambda_j\mathbf{a}_i^T\mathbf{a}_j = 0,\ i \neq j. \end{aligned}$$

Druhá rovnosť plynie z vlastnosti (1.6), piata rovnosť plynie z vlastností vlastných čísel a vlastných vektorov matice $\boldsymbol{\Sigma}$ a posledná rovnosť plynie z ortogonality vektorov $\mathbf{a}_i$ a $\mathbf{a}_j$.

Na vzťah (1.9) nahliadneme takto:

$$\begin{aligned} Cov(\mathbf{X}, \mathbf{Y}) &= Cov(\mathbf{X}, \mathbf{A}^T(\mathbf{X}-\boldsymbol{\mu})) = Cov(\mathbf{X}, \mathbf{X})\mathbf{A} \\ &= \boldsymbol{\Sigma}\mathbf{A} = \mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^T\mathbf{A} = \mathbf{A}\boldsymbol{\Lambda}. \end{aligned}$$

Predposledná a posledná rovnosť plynie opäť z vlastnosti (1.6) a ortonormality matice $\mathbf{A}$.

Rovnicu (1.10) dostaneme:

$$\sum_{j=1}^{p} Var(Y_j) = \sum_{j=1}^{p} \lambda_j = tr(\boldsymbol{\Lambda}) = tr(\mathbf{A}^T\mathbf{A}\boldsymbol{\Lambda}) = tr(\mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^T) = tr(\boldsymbol{\Sigma}).$$

Výpočet plynie zo základných vlastností stopy matice.

Poslednú rovnosť (1.11) získame ako

$$\rho_{X_iY_j} = \frac{Cov(X_i, Y_j)}{\sqrt{Var(X_i)}\sqrt{Var(Y_j)}} = \frac{a_{ij}\lambda_j}{\sqrt{Var(X_i)}\sqrt{\lambda_j}} = a_{ij}\left(\frac{\lambda_j}{Var(X_i)}\right)^{1/2}.$$

Druhá rovnosť plynie z vlastnosti (1.9).

Korelácia (1.11) vhodne popisuje vzťah medzi hlavnými komponentami a pôvodnými premennými. Všimnime si, že $\sum_{j=1}^{p} \lambda_j a_{ij}^2 = \mathbf{a}_i^T\boldsymbol{\Lambda}\mathbf{a}_i$ je (i,i)-tý element matice $\mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^T = \boldsymbol{\Sigma}$ (viď rovnica (1.5)), takže

$$\sum_{j=1}^{p} \rho_{X_iY_j}^2 = \sum_{j=1}^{p} a_{ij}^2 \frac{\lambda_j}{Var(X_i)} = \frac{1}{Var(X_i)} Var(X_i) = 1.$$

Ako uvádza Härdle & Hlávka (2007), koreláciu $\rho_{X_iY_j}^2$ môžeme chápať ako podiel variancie i-tej premennej X_i zachytený j-tým hlavným komponentom Y_j. Podiel rozptylu X_i zachytený prvými q hlavnými komponentami je $\sum_{j=1}^{q} \rho_{X_iY_j}^2 \leq 1$. Všetky body sú teda vo vnútri jednotkovej guli. Vzdialenosť bodu so súradnicami $[\rho_{X_iY_1}, \rho_{X_iY_2}, ..., \rho_{X_iY_q}]$ od plochy tejto jednotkovej gule v q-rozmernom euklidovskom priestore môže byť použitá na meranie zachyteného rozptylu premennej X_i.

Uvažujme dvojrozmerný kruh korelácie s polomerom 1, ktorého osi popisujú veľkosť korelácie medzi pôvodnými premennými a hlavnými komponentami Y_1 a Y_2, teda graf zobrazujúci body so súradnicami $(\rho_{X_iY_1}, \rho_{X_iY_2})$ pre $i = 1, 2, ..., p$. Originálne premenné sú silno korelované s Y_1 a Y_2, ak sú ich korelačné koeficienty umiestené blízko kružnice. Dva body, ktoré ležia blízko seba pri kružnici sú vysoko pozitívne korelované. Ak sú oba body blízko kružnice ale vzájomne vzdialené, potom sú vysoko negatívne korelované. Dva body ležiace pri kružnici, ktorých vektory v počiatku zvierajú pravý uhol, sú nekorelované.

Poznámka 1.2.1. Ak poznáme rozdelenie $\mathbf{X}$, môžeme zistiť aj rozdelenie $\mathbf{Y}$. Nech $\mathbf{X} \sim \mathsf{N_p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Každý prvok vektora $\mathbf{Y}$ je teda lineárnou kombináciou normálne rozdelenej náhodnej veličiny, takže každý hlavný komponent $Y_j,\ j = 1, 2, ..., p$, bude mať normálne rozdelenie. $\mathbf{Y}$ má teda mnohorozmerné normálne rozdelenie so strednou hodnotou $\mathbf{0}$ a kovariančnou maticou $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, ..., \lambda_p)$. Píšeme $\mathbf{Y} \sim \mathsf{N_p}(\mathbf{0}, \boldsymbol{\Lambda})$. Bližšie sa touto problematikou zaoberá Chatfield (2000).

Príklad 1.2.1. Uvažujme dvojrozmerné normálne rozdelenie vektora $\mathbf{X} \sim \mathsf{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$, kde $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$ a $0 < \rho < 1$. Vlastné čísla matice $\boldsymbol{\Sigma}$ sú $\lambda_1 = 1 + \rho$ a $\lambda_2 = 1 - \rho$ s príslušnými normovanými vlastnými vektormi $\mathbf{a}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ a $\mathbf{a}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Potom sú hlavné komponenty

$$\mathbf{Y} = \mathbf{A}^T(\mathbf{X} - \boldsymbol{\mu}) = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\mathbf{X}$$

alebo

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} X_1 + X_2 \\ X_1 - X_2 \end{pmatrix}.$$

Takže prvý hlavný komponent je

$$Y_1 = \frac{1}{\sqrt{2}}(X_1 + X_2)$$

a druhý hlavný komponent je

$$Y_2 = \frac{1}{\sqrt{2}}(X_1 - X_2).$$

Teraz overíme, či vypočítané hlavné komponenty spĺňajú niektoré z vyššie uvedených vlastností.

$$EY_1 = E(\frac{1}{\sqrt{2}}(X_1 + X_2)) = \frac{1}{\sqrt{2}}(EX_1 + EX_2) = 0.$$

Analogicky

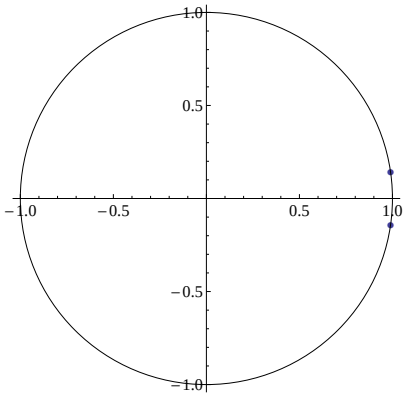
$$EY_2 = 0.$$

$$\begin{aligned} Var(Y_1) &= Var(\frac{1}{\sqrt{2}}(X_1 + X_2)) = \frac{1}{2}Var(X_1 + X_2) \\ &= \frac{1}{2}(Var(X_1) + Var(X_2) + 2Cov(X_1, X_2)) \\ &= \frac{1}{2}(1 + 1 + 2\rho) = 1 + \rho = \lambda_1 \end{aligned}$$

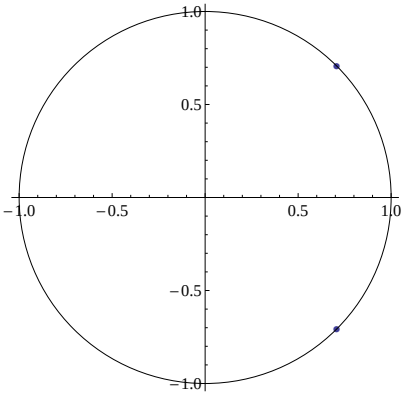
Analogicky

$$Var(Y_2) = 1 - \rho = \lambda_2.$$

$$\begin{aligned} Cov(Y_1, Y_2) &= E(Y_1Y_2) - E(Y_1)E(Y_2) = E(Y_1Y_2) \\ &= E(\frac{1}{2}(X_1 + X_2)(X_1 - X_2)) = \frac{1}{2}E(X_1^2 - X_2^2) = 0 \end{aligned}$$



Obr. 1.3: Kruh korelácie s ρ blízke jednej ($\rho = 0.96$)



Obr. 1.4: Kruh korelácie s ρ blízke nule ($\rho = 0.04$)

Teraz si vypočítame kovarianciu vektorov $\mathbf{X}$ a $\mathbf{Y}$, ktorú použijeme pri výpočte korelácie:

$$Cov(\mathbf{X}, \mathbf{Y}) = \mathbf{A\Lambda} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1+\rho & 1-\rho \\ 1+\rho & -1+\rho \end{pmatrix}$$

Potom korelácie medzi pôvodnými premennými a jednotlivými komponentami sú:

$$\rho_{X_1Y_1} = \frac{1}{\sqrt{2}} \frac{1+\rho}{\sqrt{1+\rho}} = \sqrt{\frac{1+\rho}{2}},$$

potom podobne tiež

$$\rho_{X_1Y_2} = \sqrt{\frac{1-\rho}{2}}, \qquad \rho_{X_2Y_1} = \sqrt{\frac{1+\rho}{2}}, \qquad \rho_{X_2Y_2} = -\sqrt{\frac{1-\rho}{2}}.$$

Body so súradnicami $[\rho_{X_1Y_1}, \rho_{X_1Y_2}]$ a $[\rho_{X_2Y_1}, \rho_{X_2Y_2}]$ si zobrazíme na korelačnom kruhu. Na obrázku 1.3 môžeme vidieť body, kde sme za ρ dosadili číslo blízke jednej a na obrázku 1.4 máme znázornené body, kde sme za ρ dosadili číslo blízke nule. Keďže pracujeme s dvojrozmernými dátami a používame obidva hlavné komponenty, body ležia priamo na kružnici. V prvom prípade vidíme, že body sú blízko seba. To znamená, že sú vysoko pozitívne korelované. Na druhom grafe sú body od seba ďalej a ich vektory v počiatku zvierajú pravý uhol. Takže tieto body sú nekorelované. △

Ilustračný príklad z podkapitoly 1.1 je možné priamo použiť v príklade 1.2.1, kde $\rho = \frac{4}{5}$ a vlastné čísla sú $\lambda_1 = \frac{9}{5}$ a $\lambda_2 = \frac{1}{5}$.

1.3 Výberové hlavné komponenty

V predchádzajúcich častiach textu sme pre výpočet hlavných komponentov požadovali znalosť niektorých charakteristík náhodného vektora $\mathbf{X}$, tj. najmä strednú hodnotu $\boldsymbol{\mu}$ a rozptylovú maticu $\boldsymbol{\Sigma}$. V praxi však obvykle tieto charakteristiky nepoznáme. Obyčajne mávame k dispozícii n nezávislých náhodných vektorov - pozorovaní - reprezentujúcich rozdelenie náhodného vektora $\mathbf{X}$,

tj. $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$. Zaveďme $\boldsymbol{\mathcal{X}} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)^T$ maticu pozorovaní rozmerov $(n \times p)$. Strednú hodnotu $\boldsymbol{\mu}$ potom môžeme nahradiť odhadom (výberovým priemerom)

$$\bar{\boldsymbol{x}} = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_p)^T, \text{ kde } \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij},\ j = 1, 2, ..., p.$$

Kovariančnú maticu $\boldsymbol{\Sigma}$ potom nahradíme výberovou kovariančnou maticou $\mathbf{S}$, ktorej prvky sú

$$s_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j),\ i, j = 1, 2, ..., p.$$

Takže

$$\mathbf{S} = \frac{1}{n}(\boldsymbol{\mathcal{X}}^T \boldsymbol{\mathcal{H}} \boldsymbol{\mathcal{X}}),$$

kde $\boldsymbol{\mathcal{H}} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$, $\boldsymbol{\mathcal{H}}$ je symetrická a idempotentá matica (tj. $\boldsymbol{\mathcal{H}}^2 = \boldsymbol{\mathcal{H}}$) a $\mathbf{1}_n = (1, 1, ..., 1)^T$. Výberovými hlavnými komponentami rozumieme

$$\boldsymbol{\mathcal{Y}} = (\boldsymbol{\mathcal{X}} - \mathbf{1}_n\bar{\boldsymbol{x}}^T)\boldsymbol{\mathcal{A}},$$

kde $\mathbf{S} = \boldsymbol{\mathcal{A}}\boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{A}}^T$ je spektrálny rozklad matice $\mathbf{S}$. Výberová kovariančná matica náhodnej matice $\boldsymbol{\mathcal{Y}}$ sa dá vyjadriť ako

$$\begin{aligned}\mathbf{S}_{\boldsymbol{\mathcal{Y}}} &= \frac{1}{n}\boldsymbol{\mathcal{Y}}^T\boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{Y}} = \frac{1}{n}\boldsymbol{\mathcal{A}}^T(\boldsymbol{\mathcal{X}} - \mathbf{1}_n\bar{\boldsymbol{x}}^T)^T\boldsymbol{\mathcal{H}}(\boldsymbol{\mathcal{X}} - \mathbf{1}_n\bar{\boldsymbol{x}}^T)\boldsymbol{\mathcal{A}} \\ &= \frac{1}{n}\boldsymbol{\mathcal{A}}^T\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{A}}^T\mathbf{S}\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{L}},\end{aligned}$$

kde $\boldsymbol{\mathcal{L}} = diag(l_1, l_2, ..., l_p)$ je matica vlastných čísel výberovej kovariančnej matice $\mathbf{S}$. Tretia rovnosť platí, pretože

$$\boldsymbol{\mathcal{H}}\mathbf{1}_n\bar{\boldsymbol{x}}^T = (\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{1}_n\bar{\boldsymbol{x}}^T = \mathbf{I}_n\mathbf{1}_n\bar{\boldsymbol{x}}^T - \frac{1}{n}\mathbf{1}_n n\bar{\boldsymbol{x}}^T = \mathbf{0}.$$

Označme si vlastné čísla matice $\mathbf{S}$ v klesajúcom poradí $l_1, l_2, ..., l_p$ a príslušné vlastné vektory $\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_p$. Ak sme nepoužili všetky pozorovania v populácií ale len ich časť, potom postupnosti $\{l_i\}$ a $\{\boldsymbol{a}_i\}$ môžu byť považované za odhady vlastných čísel a vlastných vektorov matice $\boldsymbol{\Sigma}$.

Následujúca veta, čerpaná z Härdle & Simar (2003), popisuje asymptotické vlastnosti hlavných komponentov vypočítaných z výberovej kovariančnej matice.

V znení vety budeme používať *Wishartovo rozdelenie*: Nech je daná dátová matica $\boldsymbol{\mathcal{X}}$ rozmerov $(n \times p)$. Je to matica n vzájomne nezávislých náhodných vektorov $\mathbf{X}_1 \sim \mathsf{N_p}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \mathbf{X}_2 \sim \mathsf{N_p}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}), ..., \mathbf{X}_n \sim \mathsf{N_p}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma})$. Teda $\boldsymbol{\mathcal{X}} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)^T$. Ďalej označme $\mathbf{M} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_n)^T$ maticu rozmerov $(n \times p)$ tvorenú strednými hodnotami náhodných vektorov a $\boldsymbol{\Sigma}$ je kovariančná matica rozmerov $(p \times p)$. Združené rozdelenie prvkov matice $\mathbf{W} = \boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}} = \sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T$ nazveme *p-rozmerným Wishartovým rozdelením o n stupňoch voľnosti s parametrami* $\boldsymbol{\Sigma}$ *a* $\mathbf{M}$. Značíme $\mathbf{W} \sim \mathsf{W_p}(n, \boldsymbol{\Sigma}, \mathbf{M})$. Ak $\mathbf{M} = \mathbf{0}$, hovoríme o *centrálnom Wishartovom rozdelení* a značíme $\mathbf{W} \sim \mathsf{W_p}(n, \boldsymbol{\Sigma})$.

Veta 1.3.1. Nech je $\mathbf{\Sigma}$ pozitívne definitná matica a má jednoznačne určené vlastné čísla a nech $\boldsymbol{\mathcal{U}} \sim \frac{1}{m}\mathsf{W_p}(m, \mathbf{\Sigma})$ so spektrálnymi rozkladmi $\mathbf{\Sigma} = \mathbf{A\Lambda A}^T$ a $\boldsymbol{\mathcal{U}} = \boldsymbol{\mathcal{G}\mathcal{L}\mathcal{G}}^T$, kde $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_p)^T$ a $\boldsymbol{\mathcal{G}} = (\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_p)^T$. Potom

a) $\sqrt{m}(\boldsymbol{l} - \boldsymbol{\lambda}) \xrightarrow{D} \mathsf{N_p}(\mathbf{0}, 2\mathbf{\Lambda}^2)$
kde $\boldsymbol{l} = (l_1, l_2, ..., l_p)^T$ a $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_p)^T$ sú diagonály matíc $\boldsymbol{\mathcal{L}}$ a $\mathbf{\Lambda}$ a symbol D značí konvergenciu v distribúcií.

b) $\sqrt{m}(\mathbf{g}_j - \mathbf{a}_j) \xrightarrow{D} \mathsf{N_p}(\mathbf{0}, \boldsymbol{\mathcal{V}}_j)$,
kde $\boldsymbol{\mathcal{V}}_j = \lambda_j \sum_{k \neq j} \frac{\lambda_k}{(\lambda_k - \lambda_j)^2} \mathbf{a}_k \mathbf{a}_k^T$,

c) $Cov(\mathbf{g}_j, \mathbf{g}_k) = \boldsymbol{\mathcal{V}}_{jk}$,
kde (r, s)-tý prvok matice $\boldsymbol{\mathcal{V}}_{jk}(p \times p)$ je $-\frac{\lambda_j \lambda_k a_{rk} a_{sj}}{m(\lambda_j - \lambda_k)^2}$,

d) prvky vektora $\boldsymbol{l}$ sú asymptoticky nezávislé na prvkoch matice $\boldsymbol{\mathcal{G}}$.

Poznámka 1.3.1. Veta 1.3.1 sa používa najmä k testovaniu rôznych hypotéz.

1.4 Vybrané problémy metódy hlavných komponentov

1.4.1 Pomer informácie zachytenej hlavnými komponentami

Keďže $\sum_{i=1}^{p} Var(X_i) = tr(\mathbf{\Sigma})$, z rovnosti (1.10) vidíme, že súčet rozptylov všetkých pôvodných premenných a súčet rozptylov hlavných komponentov je rovnaký. Pomerom $\lambda_i / \sum_{j=1}^{p} \lambda_j$ zistíme, koľko percent informácie z originálnych premenných je zachytenej v i-tom hlavnom komponente. Prvých q hlavných komponentov teda vyjadruje varianciu

$$\psi_q = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_q}{\sum_{j=1}^{p} \lambda_j}, \; q = 1, 2, ..., p. \tag{1.12}$$

1.4.2 Kovariančná vs. korelačná matica

V praxi sa mnohokrát používa na výpočet hlavných komponentov korelačná matica. Hlavné komponenty vzniknuté na základe použitia korelačnej matice si označíme $\mathbf{Z}$.

$$\mathbf{Z} = \mathbf{A}^T \mathbf{X}^*,$$

kde A obsahuje v stĺpčekoch vlastné vektory korelačnej matice a $\mathbf{X}^*$ obsahuje štandardizované premenné. Ako je uvedené v knihe Jolliffe (2002) pre $\mathbf{X}^*$ platí: j-tý prvok má tvar $X_j / \sqrt{\sigma_{jj}}$, $j = 1, 2, ..., p$, X_j je j-tý element $\mathbf{X}$ a σ_{jj} je rozptyl X_j. Potom kovariančná matica pre vektor $\mathbf{X}^*$ je korelačná matica pre vektor $\mathbf{X}$.

Môže sa zdať, že hlavné komponenty pre korelačnú maticu by mohli byť získané pomerne ľahko z hlavných komponentov pre zodpovedajúcu kovariančnú

maticu, pretože $\mathbf{X}^*$ je spojená s $\mathbf{X}$ veľmi jednoduchou transformáciou. Avšak, vlastné čísla a vektory korelačnej matice nemajú jednoduchý vzťah s kovariančnou maticou. Transformácia $\mathbf{X}$ do $\mathbf{X}^*$ nie je ortogonálna. Hlavné komponenty pre korelačnú a kovariančnú maticu neposkytujú rovnocenné informácie a nemôžu byť odvodené priamo od seba navzájom.

Hlavným argumentom pre použitie korelačnej matice na odvodenie hlavných komponentov je, že výsledky analýz pre rôzne sady náhodných veličín sú ľahšie porovnateľné, než na základe analýzy kovariančnej matice. Veľkou nevýhodou metódy hlavných komponentov za použitia kovariančnej matice je citlivosť hlavných komponentov na jednotky používané k meraniu každého prvku vektora $\mathbf{X}$.

Ak existujú veľké rozdiely medzi rozptylmi prvkov vektora $\mathbf{X}$, potom tie premenné, ktorých rozptyl je najväčší, obsadia pozície niekoľkých prvých hlavných komponentov. To je v poriadku, ak sú všetky prvky vektora $\mathbf{X}$ namerané v rovnakých jednotkách. Vtedy môžeme bez problémov použiť kovariančnú maticu. V praxi sa však často stáva, že rôzne prvky vektora $\mathbf{X}$ sú úplne odlišné typy merania. Niektoré môžu byť dĺžky, niektoré váhy, iné teploty... V takom prípade štruktúra hlavných komponentov závisí na voľbe jednotiek merania. Pri ich zmene sa taktiež mení podiel variancie vysvetlený jednotlivými hlavnými komponentami.

V korelačnej matici sú na diagonále samé jednotky. Preto súčet prvkov na diagonále alebo súčet rozptylov štandardizovaných premenných sa rovná p. Teda súčet vlastných čísel korelačnej matice je tiež rovný p, takže podiel variability, ktorá je vyjadrená j-tým hlavným komponentom je λ_j/p.

1.4.3 Problém mierky

Je dôležité si uvedomiť, že hlavné komponenty množiny premenných závisia na mierke použitej na meranie premenných. Túto problematiku tiež rozvíja Chatfield & Collins (2000).

Napríklad si zoberme množinu n jedincov a pri každom zmeráme váhu v kilogramoch, výšku v centimetroch a vek v rokoch, čo nám dá vektor $\mathbf{X}$. Jeho kovariančnú maticu označíme $\mathbf{S}_X$. Pre zmenu jednotiek, v ktorých boli premenné namerané, si vytvoríme vektor $\tilde{\mathbf{X}}^T =$ (váha v gramoch, výška v metroch, vek v mesiacoch). Potom $\tilde{\mathbf{X}} = \mathbf{K}\mathbf{X}$, kde $\mathbf{K} = diag(1000, 1/100, 12)$. Kovariančná matica nových premenných je potom $\mathbf{S}_{\tilde{\mathbf{X}}} = \mathbf{K}\mathbf{S}_X\mathbf{K}$, keďže $\mathbf{K}^T = \mathbf{K}$. Vo všeobecnosti sú vlastné čísla a vlastné vektory matice $\mathbf{S}_{\tilde{\mathbf{X}}}$ odlišné od tých z matice $\mathbf{S}_X$. Ak ich aj pretransformujeme na pôvodné premenné, hlavné komponenty budú poväčšine odlišné.

Hlavné komponenty zostanú rovnaké ak:

- Všetky prvky na diagonále matice $\mathbf{K}$ sú rovnaké, takže $\mathbf{K} = c\mathbf{I}$, kde c je skalár.

- Premenné korešpondujúce s nerovnakými prvkami na diagonále $\mathbf{K}$ sú nekorelované. Takže ak všetky prvky $\mathbf{K}$ sú odlišné, potom $\mathbf{S}_X$ musí byť diagonálna matica. V takom prípade je zbytočné používať metódu hlavných komponentov, keďže pôvodné premenné sú už nekorelované.

Ukážeme si príklad, v ktorom zmeníme mierku len jednej veličiny:

Príklad 1.4.1. Použitie metódy hlavných komponentov v prípade $\mathbf{X} \sim \mathsf{N}_2(\mathbf{0}, \mathbf{\Sigma})$, kde $\mathbf{\Sigma} = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$ a $0 < \rho < 1$ sme už predviedli v príklade 1.2.1. Spektrálna dekompozícia kovariančnej matice $\mathbf{\Sigma}$ je teda

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Vynásobme teraz X_1 konštantou c predstavujúcou zmenu mierky, $c > 0$. Potom teda $\mathbf{K} = diag(c, 0)$ a $\tilde{\mathbf{X}} = \mathbf{K}\mathbf{X}$. Keďže

$$Var(cX_1) = c^2 Var(X_1)$$

a

$$Cov(cX_1, X_2) = c\, Cov(X_1, X_2),$$

potom vektor $\tilde{\mathbf{X}} = (cX_1, X_2)^T$ má opäť normálne rozdelenie $\mathsf{N}_2(\mathbf{0}, \tilde{\mathbf{\Sigma}})$, kde

$$\tilde{\mathbf{\Sigma}} = \begin{pmatrix} c^2 & c\rho \\ c\rho & 1 \end{pmatrix}.$$

Vyšetríme vlastné čísla matice $\tilde{\mathbf{\Sigma}}$:

$$\begin{vmatrix} c^2 - \lambda & c\rho \\ c\rho & 1 - \lambda \end{vmatrix} = 0.$$

Riešením príslušnej kvadratickej rovnice

$$\lambda^2 - \lambda(c^2 + 1) + c^2(1 - \rho^2) = 0$$

dostaneme vlastné čísla

$$\lambda_{1,2} = \frac{1}{2}(1 + c^2 \pm \sqrt{(c^2-1)^2 + 4c^2\rho^2}).$$

Vlastný vektor príslušný vlastnému číslu λ_1 vypočítame za pomoci sústavy

$$\begin{pmatrix} c^2 & c\rho \\ c\rho & 1 \end{pmatrix} \begin{pmatrix} z_1^{\lambda_1} \\ z_2^{\lambda_1} \end{pmatrix} = \lambda_1 \begin{pmatrix} z_1^{\lambda_1} \\ z_2^{\lambda_1} \end{pmatrix}.$$

Z tej plynie, že

$$z_1^{\lambda_1} = \frac{z_2^{\lambda_1}(\lambda_1 - 1)}{c\rho}.$$

Vlastný vektor príslušný vlastnému číslu λ_2 si označme $(z_1^{\lambda_2}, z_2^{\lambda_2})^T$, získali by sme ho analogicky. Potom pre hlavné komponenty platí

$$\tilde{\mathbf{Y}} = \mathbf{A}^T\tilde{\mathbf{X}} = \begin{pmatrix} z_1^{\lambda_1} & z_2^{\lambda_1} \\ z_1^{\lambda_2} & z_2^{\lambda_2} \end{pmatrix} \begin{pmatrix} cX_1 \\ X_2 \end{pmatrix},$$

naviac požadujeme splnenie normalizačných podmienok:

$$\sqrt{(z_1^{\lambda_1})^2 + (z_2^{\lambda_1})^2} = 1 \quad \text{a} \quad \sqrt{(z_1^{\lambda_2})^2 + (z_2^{\lambda_2})^2} = 1.$$

Takže

$$\begin{aligned}\tilde{Y_1} &= z_1^{\lambda_1} cX_1 + z_2^{\lambda_1} X_2 = \frac{z_2^{\lambda_1}(\lambda_1 - 1)}{c\rho} cX_1 + z_2^{\lambda_1} X_2 \\ &= z_2^{\lambda_1} \left(\frac{cX_1(\lambda_1 - 1)}{c\rho} + X_2 \right).\end{aligned}$$

Pre $\lambda_1 > 1$ je teda funkcia λ_1/c rastúca v závislosti na $c > 0$. Potom $z_1^{\lambda_1} > z_2^{\lambda_1}$ a zlomok $z_1^{\lambda_1}/z_2^{\lambda_1}$ je rastúcou funkciou konštanty c. S rastúcim c rastie aj λ_1 a cX_1 nadobúda väčšiu váhu v prvom hlavnom komponente. △

Výber mierky môže mať vplyv na výsledky hlavných komponentov. Pri rozličných mierkach sa odporúča využiť centrované hlavné komponenty (viď podkapitola 1.2).

1.4.4 Počet hlavných komponentov

Vo všeobecnosti predpokladáme, že len niekoľko prvých hlavných komponentov stačí na dostatočné zachytenie celkového rozptylu pôvodných premenných. Na určenie optimálneho počtu komponentov $1 \leq q \leq p$ sa používa niekoľko pravidiel:

- Snáď najjednoduchším kritériom pre voľbu q je určiť percentuálny podiel na celkovej variancií, ktorá je zachytená hlavnými komponentami (viď kapitola 1.4.1.). Ten sa zväčša udáva 80% alebo 90%. Požadovaný počet hlavných komponentov je potom najmenšia hodnota q, pre ktorú je táto úroveň prekročená. Niekedy môže byť veľkosť podielu vyššia alebo nižšia. To záleží konkrétne od každej množiny dát. Napríklad je vhodné zvoliť hodnotu vyššiu než 90%, pokiaľ je väčšina variancie vysvetlená jedným alebo dvoma komponentami. Naopak, ak je treba veľký počet komponentov na vysvetlenie aspoň 80% variancie, v takom prípade budeme uvažovať o znížení hodnoty tohoto percentuálneho podielu.

- Pravidlo opísané v tejto časti je konštruované špeciálne pre použitie korelačných matíc, hoci môže byť prispôsobené aj pre niektoré druhy kovariančných matíc. Myšlienka pravidla je, že ak sú všetky prvky $\mathbf{X}$ nezávislé, potom hlavné komponenty sú rovnaké ako pôvodné premenné a všetky majú jednotkovú varianciu v prípade korelačnej matice. Teda akýkoľvek komponent s varianciou menej než 1 obsahuje menej informácie ako jedna z pôvodných premenných, a tak ho môžeme zanedbať. Toto pravidlo sa nazýva *Kaiserovo pravidlo* a ponecháva len tie hlavné komponenty, ktorých rozptyly presiahnu hodnotu 1. Na základe simulačných modelov sa zistilo, že za hranicu by sa malo brať číslo, ešte o niečo nižšie a to konkrétne 0.7. Pri kovariančnej matici namiesto hodnoty jedna použijeme aritmetický priemer vlastných čísel matice (označme $\bar{\lambda}$) alebo dokonca hodnotu $0.7\bar{\lambda}$.

- Rozhodnúť sa taktiež môžeme na základe grafického zobrazenia vysvetleného rozptylu hlavnými komponentami. Takýto graf sa nazýva *scree plot*. Na osi x sa nachádza poradie hlavných komponentov a os y zobrazuje percento vysvetlenej variancie hlavnými komponentami. V tomto grafe treba nájsť zlom, kde na ľavo od neho sú úsečky spájajúce jednotlivé body strmšie, než

úsečky na pravej strane. Potom použijeme len hlavné komponenty, ktoré sú vľavo od tohoto zlomového bodu.

Tieto pravidlá sú intuitívne a v praxi bežne využívané. Ďalšie metódy pre určenie vhodného počtu hlavných komponentov môžeme nájsť v Jolliffe (2002). Napríklad pravidlá založené na testovaní hypotéz, tie však zväčša odporúčajú viac komponentov než je nutné, ďalej štatistické pravidlá a iné.

V nasledujúcom príklade si ukážeme použitie vymenovaných metód na určenie počtu hlavných komponentov.

Príklad 1.4.2. Uvažujme mnohorozmerné normálne rozdelenie reálneho vektora $\mathbf{X} \sim \mathsf{N}_7(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

$$\text{kde} \quad \boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{pmatrix} \quad \text{a} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \frac{1}{2} & -\frac{1}{3} & \frac{1}{2} & \frac{2}{3} & \frac{1}{2} & \frac{2}{3} \\ \frac{1}{2} & 1 & \frac{2}{3} & \frac{1}{3} & \frac{2}{3} & \frac{1}{3} & \frac{1}{2} \\ -\frac{1}{3} & \frac{2}{3} & 2 & \frac{1}{2} & \frac{1}{2} & -\frac{1}{3} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & 1 & \frac{2}{3} & \frac{1}{2} & \frac{1}{3} \\ \frac{2}{3} & \frac{2}{3} & \frac{1}{2} & \frac{2}{3} & 1 & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{3} & -\frac{1}{3} & \frac{1}{2} & \frac{1}{3} & 2 & \frac{1}{3} \\ \frac{2}{3} & \frac{1}{2} & \frac{1}{2} & \frac{1}{3} & \frac{2}{3} & \frac{1}{3} & 2 \end{pmatrix}.$$

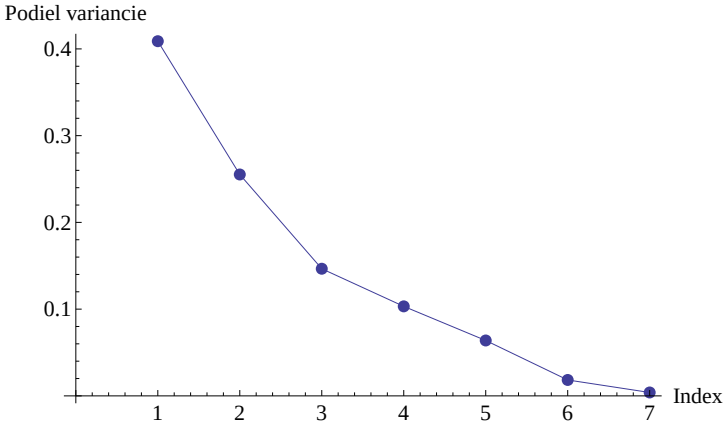
Matica $\boldsymbol{\Sigma}$ je volená ako pozitívne definitná. Vektor vlastných čísel kovariančnej matice je

$$\boldsymbol{\lambda} = (4.08885, 2.55225, 1.46566, 1.032, 0.63861, 0.18341, 0.03922)^T.$$

Percentuálny podiel variancie vysvetlený hlavnými komponentami vypočítaný podľa vzorca (1.12) je

$$\boldsymbol{\psi} = (0.408885, 0.66411, 0.810676, 0.913876, 0.977737, 0.996078, 1)^T.$$

Ak si teda zvolíme za hranicu 80%, tak nám stačia prvé tri hlavné komponenty. Pokiaľ by sme zvolili za hranicu 90%, v takom prípade musíme uvažovať prvé štyri hlavné komponenty.



Obr. 1.5: Scree plot vlastných čísel kovariančnej matice

Teraz nájdeme priemer vlastných čísel:

$$\bar{\lambda} = \frac{1}{7}\sum_{i=1}^{7} \lambda_i = 1.42857.$$

Podľa Kaiserovho pravidla teda budeme uvažovať hlavné komponenty, ktoré majú väčší rozptyl, než je hodnota 1.42857 alebo hodnota $0.7\bar{\lambda} \doteq 1$. V prvom prípade si teda všímame prvé tri hlavné komponenty a v druhom až prvé štyri komponenty. Na obrázku 1.5 vidíme znázornené podiely variancie hlavných komponentov kovariančnej matice $\mathbf{\Sigma}$. Na grafe môžeme postrehnúť mierny zlom v bode 3. Takže aj táto metóda nás utvrdzuje v tom, že vhodný počet hlavných komponentov na zachytenie väčšinového podielu variancie pôvodných dát je tri.

Z kovariančnej matice $\mathbf{\Sigma}$ si odvodíme korelačnú maticu $\mathbf{R}$ vzťahom $\mathbf{R} = \mathbf{M}^{-1}\mathbf{\Sigma}\mathbf{M}^{-1}$, kde $\mathbf{M} = diag(\sqrt{\lambda_1}, \sqrt{\lambda_2}, ..., \sqrt{\lambda_7})$. $\mathbf{M}$ je teda regulárna a diagonálna matica, ktorej prvky sú odmocniny vlastných čísel kovariančnej matice. Korelačná matica je

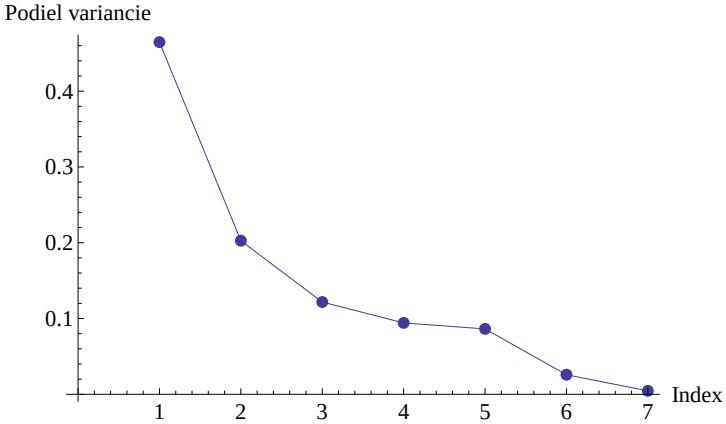
$$\mathbf{R} = \begin{pmatrix} 1 & \frac{1}{2} & -\frac{1}{3\sqrt{2}} & \frac{1}{2} & \frac{2}{3} & \frac{1}{2\sqrt{2}} & \frac{\sqrt{2}}{3} \\ \frac{1}{2} & 1 & \frac{\sqrt{2}}{3} & \frac{1}{3} & \frac{2}{3} & \frac{1}{3\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ -\frac{1}{3\sqrt{2}} & \frac{\sqrt{2}}{3} & 1 & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{6} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{2\sqrt{2}} & 1 & \frac{2}{3} & \frac{1}{2\sqrt{2}} & \frac{1}{3\sqrt{2}} \\ \frac{2}{3} & \frac{2}{3} & \frac{1}{2\sqrt{2}} & \frac{2}{3} & 1 & \frac{1}{3\sqrt{2}} & \frac{\sqrt{2}}{3} \\ \frac{1}{2\sqrt{2}} & \frac{1}{3\sqrt{2}} & -\frac{1}{6} & \frac{1}{2\sqrt{2}} & \frac{1}{3\sqrt{2}} & 1 & \frac{1}{6} \\ \frac{\sqrt{2}}{3} & \frac{1}{2\sqrt{2}} & \frac{1}{4} & \frac{1}{3\sqrt{2}} & \frac{\sqrt{2}}{3} & \frac{1}{6} & 1 \end{pmatrix}.$$

Vektor vlastných čísel korelačnej matice je

$$\boldsymbol{l} = (3.25276, 1.41861, 0.85249, 0.6594, 0.60381, 0.18096, 0.03198)^T.$$

Percentuálny podiel variancie vysvetlený hlavnými komponentami vypočítaný podľa vzorca (1.12) je

$$\boldsymbol{\varphi} = (0.46468, 0.667339, 0.789123, 0.883323, 0.969581, 0.995433, 1)^T.$$



Obr. 1.6: Scree plot vlastných čísel korelačnej matice

Podľa Kaiserovho pravidla pre korelačnú maticu berieme v úvahu prvé dva hlavné komponenty. Pokiaľ budeme uvažovať hodnotu 0.7 namiesto jednotky, v takom prípade nám pribudne o jeden komponent viac.
Na obrázku 1.6 vidíme znázornené podiely variancie hlavných komponentov korelačnej matice. Významnejšie zlomy na grafe si môžeme všimnúť v bode 2 a 3.

Vyšetrenie korelačnej matice podľa dvoch kritérií ukazuje, že zrejme prvé dva či tri hlavné komponenty zachytávajú dostatočné množstvo variancie. Keď sa však pozrieme na $\boldsymbol{\varphi}$, dvom komponentom zodpovedá len 66.73% a trom komponentom zodpovedá 78.91% variancie. To sú v porovnaní s analýzou kovariančnej matice horšie výsledky. Prvé tri hlavné komponenty kovariančnej matice totiž zachytávajú až 81.07% variancie pôvodných dát, preto je lepšie použiť práve tie. △

1.4.5 Geometrická interpretácia hlavných komponentov

Veta 1.4.1. Nech $\mathbf{X}$ je p-rozmerný reálny náhodný vektor s nulovou strednou hodnotou $\boldsymbol{\mu}$ a pozitívne definitnou kovariančnou maticou $\boldsymbol{\Sigma}$. Uvažujme množinu p-rozmerných elipsoidov

$$\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} = d, \tag{1.13}$$

kde d je konštanta, $d \geq 0$. Potom jednotlivé hlavné komponenty definujú osi týchto elipsoidov.

Dôkaz: Hlavné komponenty sú definované ako $\mathbf{Y} = \mathbf{A}^T\mathbf{X}$ (1.4). Keďže je $\mathbf{A}$ ortogonálna, existuje inverzná transformácia $\mathbf{X} = \mathbf{AY}$. Po substitúcií do vzorca (1.13) dostaneme

$$(\mathbf{AY})^T\boldsymbol{\Sigma}^{-1}(\mathbf{AY}) = \mathbf{Y}^T\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{AY} = d.$$

Matice $\boldsymbol{\Sigma}^{-1}$ a $\boldsymbol{\Sigma}$ majú rovnaké vlastné vektory a vlastné čísla matice $\boldsymbol{\Sigma}^{-1}$ sú inverzie k vlastným číslam matice $\boldsymbol{\Sigma}$ (vlastné čísla sú nenulové a pozitívne). Ďalej z rovnice (1.5) vidíme, že $\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A} = \boldsymbol{\Lambda}^{-1}$ a opäť po dosadení dostaneme

$$\mathbf{Y}^T\boldsymbol{\Lambda}^{-1}\mathbf{Y} = d.$$

Poslednú rovnosť môžeme prepísať ako

$$\sum_{k=1}^{p}\frac{Y_k^2}{\lambda_k} = d.$$

To je rovnica elipsoidu s dĺžkami poloos $\sqrt{d\lambda_1}, \sqrt{d\lambda_2}, ..., \sqrt{d\lambda_p}$, kde $\lambda_i, i = 1, 2, ..., p$ sú vlastné čísla kovariančnej matice $\boldsymbol{\Sigma}$ a hlavné osi elipsoidu majú smer vlastných vektorov kovariančnej matice $\boldsymbol{\Sigma}$. □

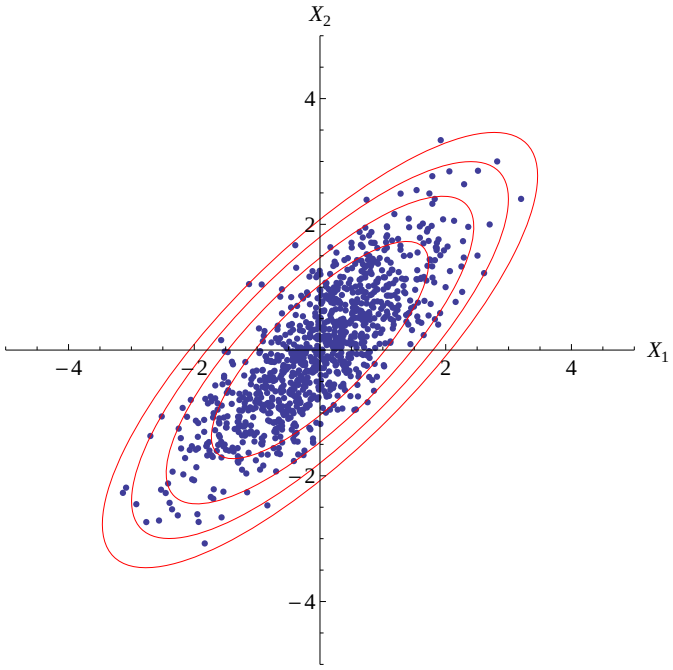
Poznámka 1.4.1. Vo vete 1.4.1 sme použili p-rozmerný reálny náhodný vektor s nulovou strednou hodnotou. Ak je $\boldsymbol{\mu} \neq \mathbf{0}$, dostaneme elipsoid so stredom v bode $\boldsymbol{\mu}$.

V prípade, že náhodný vektor $\mathbf{X}$ má mnohorozmerné normálne rozdelenie, elipsoidy z rovnice (1.13) popisujú konštantnú pravdepodobnosť rozdelenia vektora $\mathbf{X}$. Najdlhšia hlavná os elipsoidov definuje smer, v ktorom je štatistická

variancia najväčšia a má teda smer prvého vlastného vektora kovariančnej matice. Druhá hlavná os popisujúca druhý vlastný vektor je ortogonálna na prvú os. Podobne pre ostatné osi elipsoidov.

Na ilustračnom príklade z podkapitoly 1.1 si predvedieme geometrickú interpretáciu. Pozorovania boli generované z dvojrozmerného normálneho rozdelenia. Na obrázku 1.7 vidíme teda elipsy, ktorých osi majú smer vlastných vektorov $\mathbf{\Sigma}$.

Podrobnejšie objasnenie geometrickej interpretácie hlavných komponentov a ďalšie geometrické vlastnosti uvádza kniha Jolliffe (2002).



Obr. 1.7: Elipsy tvorené hlavnými komponentami súboru dvojrozmerných pozorovaní

1.4.6 Faktorová analýza

Niektorí autori uvádzajú metódu hlavných komponentov a *faktorovú analýzu* za príbuzné postupy, iní zase za samostatné metódy. Obe slúžia na analýzu vzťahov medzi premennými. Metóda hlavných komponentov aj faktorová analýza sa snažia zredukovať rozmernosť skupiny údajov. Hlavný rozdiel medzi týmito metódami je ten, že metóda hlavných komponentov vysvetľuje všetku varianciu medzi originálnymi premennými a faktorová analýza iba varianciu, ktorú majú premenné spoločnú. Cieľom metódy hlavných komponentov je teda odvodenie malého množstva lineárnych kombinácií (hlavných komponentov) z množiny premenných pri zachovaní čo najväčšieho množstva informácií obsiahnutých v pôvodných premenných. Cieľom faktorovej analýzy je vysvetliť korelácie alebo kovariancie medzi premennými pomocou malého množstva nepozorovateľných, latentných premenných (faktorov). Latentné premenné nemožno všeobecne vypočítať ako lineárnu kombináciu originálnych premenných.

2. Aplikácia metódy hlavných komponentov

Výpočet hlavných komponentov je priamočiary, ich využitie je však rôznorodé. Metóda hlavných komponentov sa používa v mnohých odvetviach, my sa zameriame na oblasť financií. V tomto smere je dôležitým aspektom odhad rizika, teda špeciálne množstvo peňazí, o ktoré môžeme prísť. Týmto problémom sa zaoberá *hodnota v riziku* a za pomoci metódy hlavných komponentov ju budeme odhadovať.

2.1 Value-at-Risk

Hodnota v riziku (Value-at-Risk, VaR) je jedným zo štandardných nástrojov na meranie trhového rizika. VaR je štatistická miera rizika, ktorá sa dá ľahko interpretovať. Je to číslo zahrňujúce všetko riziko portfólia finančných aktív. VaR je založená na odhade najhoršej straty, ku ktorej môže dôjsť s vopred určenou pravdepodobnosťou (spoľahlivosťou) v stanovenom budúcom období.

Metodiku VaR teda špecifikujú dva parametre:

- *časový horizont*: je to obdobie, počas ktorého sa možná strata uvažuje; hovoríme potom o dennej hodnote v riziku cez jeden obchodný deň alebo o desaťdennej hodnote v riziku cez dva kalendárne týždne s desiatimi obchodnými dňami (podľa doporučenia Bazilejského výboru pre bankový dohľad);
- *spoľahlivosť*: je to pravdepodobnosť, s ktorou skutočná strata neprevýši hodnotu v riziku (behom príslušného časového horizontu); pracuje sa napríklad s 95% spoľahlivosťou, alebo 99% (podľa doporučenia Bazilejského výboru pre bankový dohľad).

Pre nasledujúce výpočty zaveďme označenia:

X	náhodná veličina so spojitým rozdelením predstavujúca zisk (ak X je kladná) alebo stratu (ak X je záporná) vzniknutú počas príslušného časového horizontu (napríklad v posudzovanom investičnom portfóliu)
α	požadovaná spoľahlivosť (napríklad $\alpha = 0.95, 0.99$)
P	cena (napríklad posudzovaného investičného portfólia)
r	miera zisku (napríklad v posudzovanom investičnom portfóliu) počas príslušného časového horizontu uvažovaná ako náhodná veličina so strednou hodnotou μ
$r_{1-\alpha}$	$(1-\alpha)$-kvantil náhodnej veličiny r
$z_{1-\alpha}$	$(1-\alpha)$-kvantil normovaného normálneho rozdelenia $\mathsf{N}(0,1)$

Pri takomto značení definujeme:

- *Absolútna hodnota v riziku* VaR^{abs}: hodnota $-VaR^{abs}$ je $(1-\alpha)$-kvantil $x_{1-\alpha}$ náhodnej veličiny X, tj.

$$P(-X > VaR^{abs}) = P(X < -VaR^{abs}) = 1 - \alpha$$

- *Relatívna hodnota v riziku* VaR^{rel}: vzťahuje sa k strednej hodnote $\mathrm{E}(X)$ náhodnej veličiny X ako vzdialenosť absolútnej hodnoty v riziku od stredného zisku; v praxi sa nazýva *hodnota v riziku (VaR)*

$$VaR = VaR^{real} = VaR^{abs} + E(X)$$

- Absolútna hodnota v riziku vyjadrená pomocou miery zisku:

$$VaR^{abs} = -P \cdot r_{1-\alpha}$$

- Hodnota v riziku vyjadrená pomocou miery zisku:

$$VaR = VaR^{rel} = -P \cdot (r_{1-\alpha} - \mu)$$

Ako je uvedené v knihe Dupačová & Hurt & Štěpán (2002), je možný parametrický výpočet hodnoty v riziku, ak sa dá popísať rozdelenie miery zisku počas príslušného časového horizontu pomocou nejakého parametrického rozdelenia s odhadnuteľnými parametrami. Predpokladajme, že náhodná veličina X má rozdelenie z triedy rozdelení zohľadňujúcich polohu μ a merítko σ. Nech $G(x)$ je obecná distribučná funkcia a predpokladajme, že pre distribučnú funkciu $F_{\mu,\sigma}(x)$ zisku X platí

$$F_{\mu,\sigma}(x) = G\left(\frac{x-\mu}{\sigma}\right),$$

kde μ je reálne číslo predstavujúce strednú hodnotu a $\sigma > 0$ je smerodajná odchýlka. Potom

$$\begin{aligned} P(X < -VaR^{abs}) &= P\left(\frac{X-\mu}{\sigma} < \frac{-VaR^{abs}-\mu}{\sigma}\right) && (2.1)\\ &= G\left(\frac{-Var^{abs}-\mu}{\sigma}\right) = 1-\alpha. && (2.2) \end{aligned}$$

Ak $u_{1-\alpha}$ predstavuje $(1-\alpha)$-kvantil distribučnej funkcie $G(x)$, potom z rovnosti (2.2) plynie

$$VaR^{abs} = -\mu - \sigma \cdot u_{1-\alpha}.$$

Zväčša sa predpokladá, že G je distribučná funkcia normovaného normálneho rozdelenia Φ, aj keď vo finančnej praxi sa ukázalo, že toto rozdelenie nie je až také frekventované. Ďalej, ak za zisk X v rovnosti (2.1) dosadíme mieru zisku r, tak po zahrnutí počiatočnej ceny portfólia P vyzerajú vzorce na výpočet VaR následovne:

- Absolútna hodnota v riziku vyjadrená pomocou miery zisku s normálnym rozdelením $r \sim \mathsf{N}(\mu, \sigma^2)$:

$$VaR^{abs} = -P \cdot (\mu + \sigma \cdot z_{1-\alpha})$$

- Hodnota v riziku vyjadrená pomocou miery zisku s normálnym rozdelením $r \sim \mathsf{N}(\mu, \sigma^2)$:

$$VaR = -P \cdot \sigma \cdot z_{1-\alpha}$$

- Hodnota v riziku vyjadrená pomocou miery zisku s normálnym rozdelením $r \sim \mathsf{N}(\mu, \sigma^2)$ počas časového horizontu Δt (pritom náhodná miera zisku r sa vzťahuje na časovú jednotku, takže napríklad pre ročnú mieru zisku r a denný časový horizont pri 252 obchodných dňoch v roku je $\Delta t = 1/252$):

$$VaR = -P \cdot \sigma \cdot z_{1-\alpha} \cdot \sqrt{\Delta t}$$

- Hodnota v riziku vyjadrená pomocou miery zisku s normálnym rozdelením $r \sim \mathsf{N}(\mu, \sigma^2)$ počas časového horizontu Δt a so spoľahlivosťou 95%:

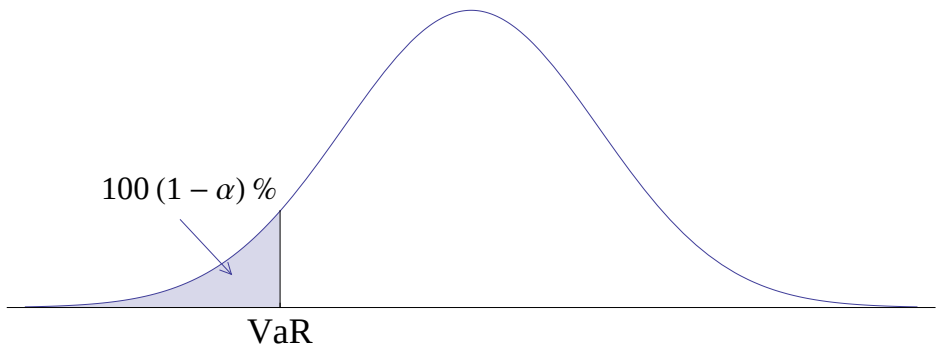
$$VaR = 1.65 \cdot P \cdot \sigma \cdot \sqrt{\Delta t}$$

- Hodnota v riziku vyjadrená pomocou miery zisku s normálnym rozdelením $r \sim \mathsf{N}(\mu, \sigma^2)$ počas časového horizontu Δt a so spoľahlivosťou 99%:

$$VaR = 2.33 \cdot P \cdot \sigma \cdot \sqrt{\Delta t}$$

Ďalšie špecifické vzorce na výpočet VaR je možné nájsť v knihe Cipra (2006).

Rozdelenie X sa zväčša predpokladá normálne, v praxi však často vadí nedostatočne ťažký záporný koniec na strane strát (pri nevhodnom použití aproximácie normálnym rozdelením teda môže vychádzať hodnota v riziku menšia, než by mala správne byť). Preto boli navrhnuté rôzne alternatívy, napr. t-rozdelenie, zmes normálnych rozdelení, atď.



Obr. 2.1: VaR počítaný z normálneho rozdelenia zmeny hodnoty portfólia s úrovňou spoľahlivosti $100\alpha\%$

Vo všeobecnosti, ak je Δt časový horizont a α úroveň spoľahlivosti, tak VaR je strata odpovedajúca $100(1-\alpha)$ percentilu rozdelenia zmeny hodnoty portfólia počas Δt dní. Na obrázku 2.1 je ilustrovaný VaR na normálnom rozdelení zmeny hodnoty portfólia.

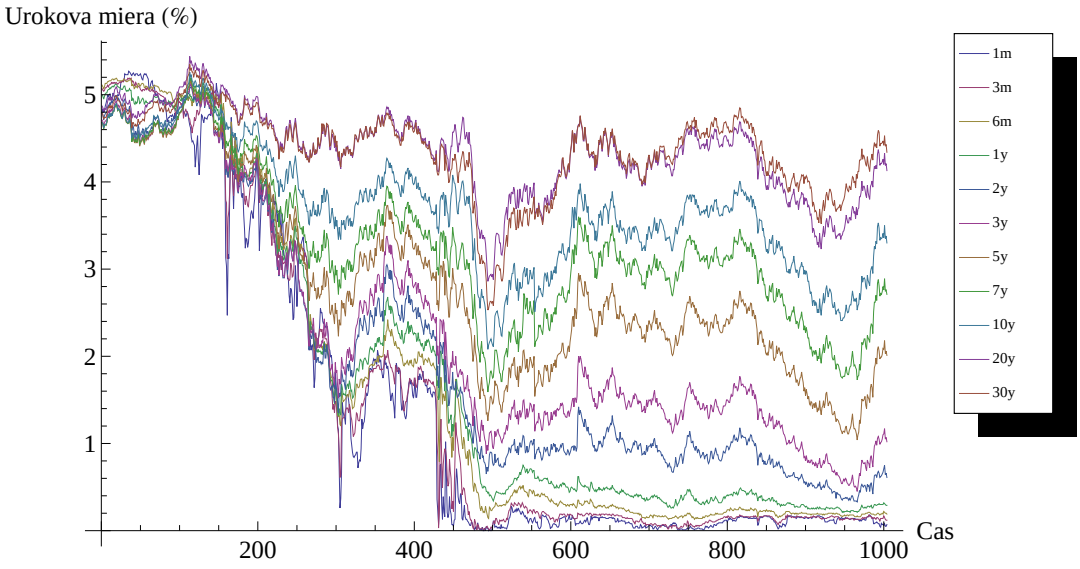
V praxi sa často počíta najprv s $\Delta t = 1$ deň a až následne sa dopočíta VaR pre $\Delta t = 10$ dní. Dôvodom je nedostatok dát na odhad správania sa premenných po dobu dlhšiu než jeden deň. Vzorec na výpočet platí, ak sú zmeny hodnôt portfólia v následujúcich dňoch nezávislé a rovnako rozdelené z normálneho rozdelenia s nulovou strednou hodnotou. Inak ide o aproximáciu

$$\Delta t\text{-denný VaR} = \text{1-denný VaR} \cdot \sqrt{\Delta t}.$$

2.2 Metóda hlavných komponentov použitá na výpočet VaR

Princíp určenia hodnoty v riziku v tejto časti práce spočíva stručne povedané v tom, že vezmeme zmeny dát trhových premenných, z ktorých určíme hlavné komponenty vystihujúce vývoj premenných, čím znížime počet premenných vstupujúcich do výpočtu VaR.

Postup si ukážeme na príklade. Za trhové premenné použijeme Daily Treasury Yield Curve Rates so splatnosťami 1, 3 a 6 mesiacov a 1, 2, 3, 5, 7, 10, 20, 30 rokov získané zo stránky www.treasury.gov. Údaje pochádzajú z rokov 2007 až 2010 a ich počet je 1004. Nedostupné dáta sú nahradené jednoduchým kĺzavým priemerom šiestich okolitých hodnôt. Vývoj sadzieb vidíme na obrázku 2.2.



Obr. 2.2: Vývoj úrokových sadzieb v priebehu štyroch rokov

Podobnou problematikou sa zaoberá Hull (2006). Analýzu budeme prevádzať na prvých diferenciách úrokových sadzieb. To znamená, že pre každú dobu splatnosti zistíme medzidenné rozdiely sadzieb. Takže počet týchto dát je 1003. Budeme sa zaoberať centrovanými hlavnými komponentami, viď podkapitoly 1.2 a 1.3.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
1m	-0.30	0.70	-0.59	-0.26	0.11	0.03	-0.04	0.00	0.00	0.00	0.00
3m	-0.24	0.46	0.32	0.50	-0.60	-0.12	0.07	-0.02	0.00	-0.01	-0.01
6m	-0.21	0.23	0.44	0.22	0.55	0.31	-0.49	0.16	0.04	0.01	-0.01
1r	-0.23	0.13	0.33	-0.07	0.43	-0.14	0.74	-0.23	-0.05	-0.01	0.03
2r	-0.31	-0.07	0.24	-0.44	-0.10	-0.44	-0.06	0.65	-0.01	0.07	-0.01
3r	-0.35	-0.11	0.18	-0.35	-0.13	-0.16	-0.38	-0.67	-0.22	0.15	-0.05
5r	-0.38	-0.19	0.03	-0.17	-0.16	0.31	0.04	-0.06	0.47	-0.66	0.08
7r	-0.37	-0.22	-0.08	0.00	-0.14	0.48	0.19	0.09	0.22	0.68	-0.05
10r	-0.33	-0.22	-0.17	0.16	-0.03	0.24	0.08	0.17	-0.78	-0.22	0.19
20r	-0.28	-0.21	-0.25	0.34	0.15	-0.27	-0.02	0.00	0.05	-0.10	-0.77
30r	-0.27	-0.20	-0.26	0.38	0.19	-0.43	-0.11	-0.07	0.26	0.10	0.60

Tabuľka 2.1: Hlavné komponenty

Počas celej analýzy budeme pracovať s kovariančnou maticou, nakoľko všetky premenné pôvodného súboru dát sú rovnakého charakteru (úrokové sadzby), tj. nemáme problém s rôznorodosťou mierky použitej na meranie jednotlivých premenných (viď podkapitola 1.4.2).

V tabuľke 2.1 sú v prvom stĺpci doby splatnosti a ostatné stĺpce sú hlavné komponenty popisujúce vývoj úrokových mier. Ak máme jednu jednotku prvého hlavného komponentu, jednomesačná úroková miera klesne o 0.30 bazického bodu, trojmesačná úroková miera klesne o 0.24 bazického bodu, atď.

Keďže máme 11 úrokových mier a 11 hlavných komponentov, zmeny úrokových mier pozorovaných v ľubovolný deň môžeme vždy vyjadriť ako lineárnu kombináciu hlavných komponentov prostredníctvom sústavy jedenástich lineárnych rovníc. Množstvo určitého komponentu použitého pri zmenách úrokových mier za konkrétny deň nazývame *skóre* daného hlavného komponentu na daný deň.

V tabuľke 2.2 vidíme v druhom stĺpci vlastné čísla kovariančnej matice vstupných dát. Dôležitosť hlavného komponentu je určená smerodajnou odchýlkou jeho faktorového skóre (odmocnina príslušného vlastného čísla), ktorú vidíme v treťom stĺpci tabuľky. Hodnoty sú namerané v bázických bodoch. Množstvo (dôle-

PC	vlastné čísla	smerodajná odchýlka	podiel variancie	súčet podielov
PC1	0.03996	0.19990	0.5873	0.5873
PC2	0.01845	0.13582	0.2712	0.8585
PC3	0.00466	0.06827	0.0685	0.9270
PC4	0.00264	0.05140	0.0388	0.9658
PC5	0.00113	0.03365	0.0166	0.9824
PC6	0.00050	0.02234	0.0073	0.9897
PC7	0.00025	0.01580	0.0037	0.9934
PC8	0.00017	0.01286	0.0025	0.9959
PC9	0.00012	0.01093	0.0018	0.9977
PC10	0.00010	0.00989	0.0015	0.9992
PC11	0.00006	0.00781	0.0008	1.0000

Tabuľka 2.2: Vlastnosti hlavných komponentov

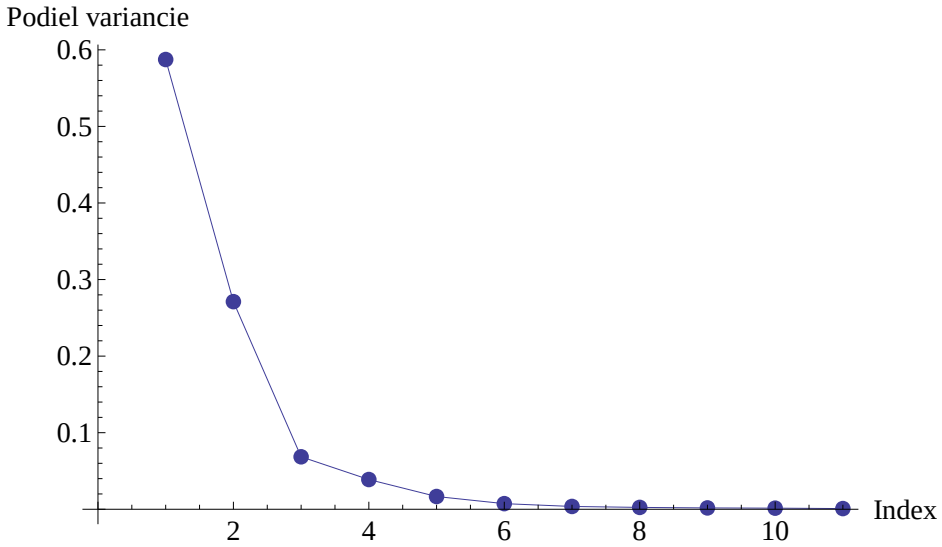
žitosť) prvého hlavného komponentu sa teda rovná smerodajnej odchýlke, preto jednomesačná úroková miera klesne o $0.30 \times 0.1999 = 0.05997$ bázického bodu, trojmesačná úroková miera klesne o $0.24 \times 0.1999 = 0.047976$ bázického bodu, atď. Celková variancia vstupných dát je súčet vlastných čísel, čo je 0.06804.

Podľa kapitoly 1.4.1 sú vypočítané zvyšné dva stĺpce tabuľky 2.2. Vo štvrtom je teda podiel variancie zachytenej príslušným hlavným komponentom a v poslednom stĺpci sú sčítané podiely variancie zachytené jednotlivými podmnožinami hlavných komponentov (viď vzorec (1.12)).

Teraz vyšetríme, s akým počtom hlavných komponentov budeme pracovať. Za požadovaný podiel variancie zachytený hlavnými komponentami si zvolme hranicu 90%.

Podľa Kaiserovho pravidla sa majú vybrať tie komponenty, ktorých príslušné vlastné čísla sú väčšie než samotný aritmetický priemer vlastných čísel alebo hodnota o niečo nižšia (viď kapitola 1.4.4.). Keďže súčet vlastných čísel je 0.06804, tak ich priemer je $0.06804/11 = 0.0062$. Pre menšiu hodnotu počítame: $0.0062 \times 0.7 = 0.0043$. Z tabuľky 2.2 teda vyplýva, že by sme v prípade prísnejšieho kritéria mali zvoliť prvé tri hlavné komponenty.

Ďalej si vykreslíme scree plot. V tomto grafe na obrázku 2.3 sú použité hodnoty zo štvrtého stĺpca tabuľky 2.2. Zlomový bod krivky vidíme v hodnote 3 na osi x. V poslednom stĺpci tabuľky 2.2 vidíme, že prvé tri komponenty zachytávajú až 92.7%, čo prevyšuje nami určenú hranicu.



Obr. 2.3: Scree plot vlastných čísel kovariančnej matice

Teraz si vyšetríme vypovedaciu hodnotu (*explanatory power*) jednotlivých hlavných komponentov v závislosti na dobe splatnosti podľa Malava (2006). Popíšeme postup: Majme p-rozmerný náhodný vektor $\mathbf{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Bez ujmy na všeobecnosti položme $\boldsymbol{\mu} = \mathbf{0}$. Vektor hlavných komponentov je daný vzťahom $\mathbf{Y} = \mathbf{A}^T\mathbf{X}$. Vzhľadom k tomu, že $\mathbf{A}$ je ortogonálna, platí

$$\mathbf{X} = \mathbf{A}\mathbf{Y}.$$

Jednotlivé zložky náhodného vektora $\mathbf{X}$ možno vyjadriť

$$X_i = \sum_{k=1}^{p} a_{ik} Y_k.$$

Ak vyjadríme X_i iba s použitím vybranej q-tej komponenty, dostaneme

$$\hat{X}_i^q = a_{iq} Y_q.$$

Platí

$$Var(X_i) = \sum_{k=1}^{p} a_{ik}^2 Var(Y_k) = \sum_{k=1}^{p} a_{ik}^2 \lambda_k,$$

ďalej

$$Var(\hat{X}_i^q) = a_{iq}^2 \lambda_q.$$

Podiel vysvetleného rozptylu q-tým hlavným komponentom je v prípade náhodnej veličiny X_i rovný podielu

$$\frac{Var(\hat{X}_i^q)}{VarX_i} = \frac{a_{iq}^2 \lambda_q}{\sum_{k=1}^{p} a_{ik}^2 \lambda_k}.$$

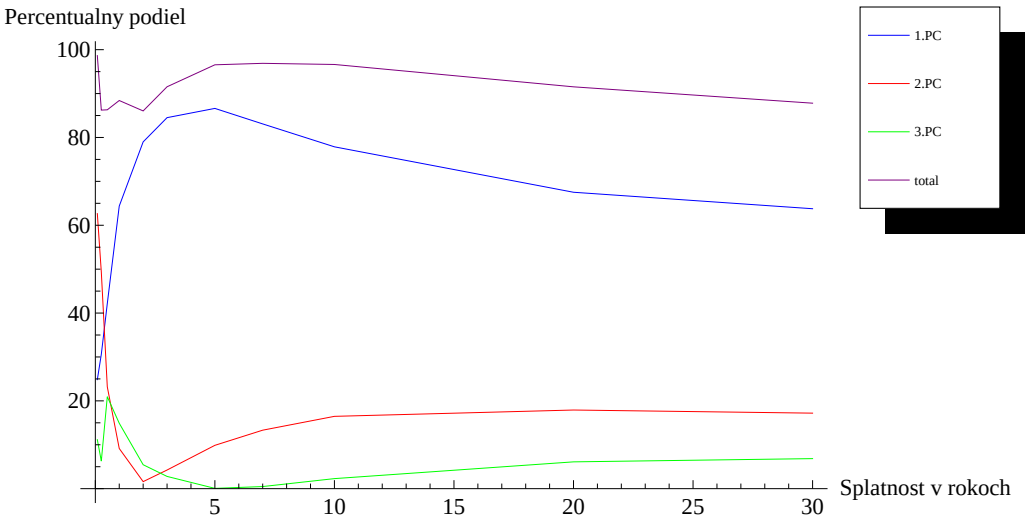
Následne pre každé $i = 1, 2, ..., p$ hľadáme $Q \in \mathbb{N}$ tak, aby

$$\sum_{q=1}^{Q} \frac{Var(\hat{X}_i^q)}{VarX_i} \geq f,$$

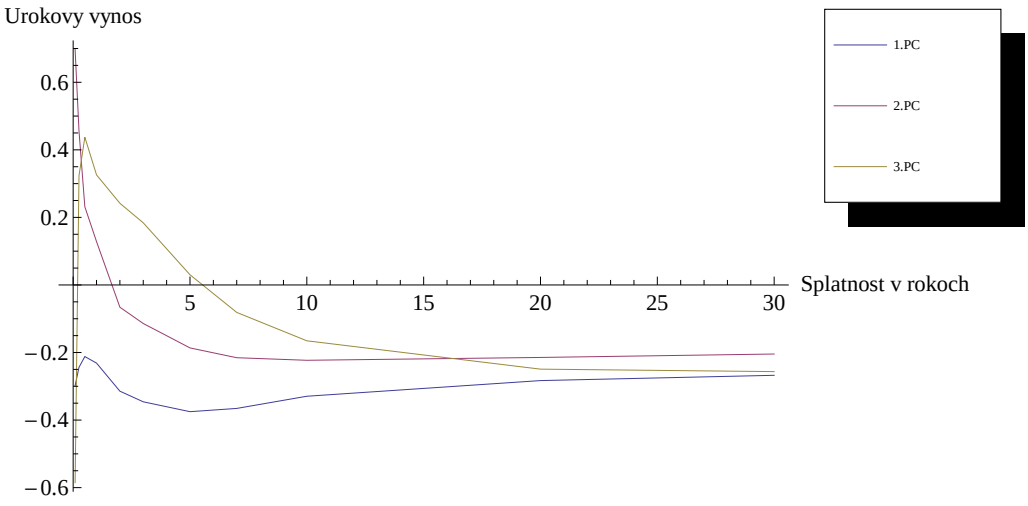
kde $0 \leq f \leq 1$ je zvolený pomer variancie, ktorý má byť vysvetlený pre každú dobu splatnosti. Poznamenajme, že $\sum_{q=1}^{Q} Var(\hat{X}_i^q)$ určuje rozptyl X_i vysvetlený prostredníctvom prvých Q hlavných komponentov (my používame výberové proťajšky). Pokiaľ uvedenú konštrukciu zrovnáme s komentárom týkajúcim sa korelácie hlavných komponentov a pôvodných premenných (viď strana 9), vidíme, že sú oba postupy zhodné.

Na obrázku 2.4 je znázornená vypovedacia hodnota prvých troch hlavných komponentov. Povšimnime si, že v prípade úrokovej miery s trojmesačnou, šesťmesačnou a dvojročnou dobou splatnosti vysvetľujú tieto komponenty len niečo cez 86% variancie. Aj keď celkovo je teda zachytených až 92.7% variancie, pokiaľ sa zameriame na jednotlivé doby splatnosti tieto pomery sú rôzne. Prvý hlavný komponent vystihuje len veľmi málo z úrokových mier peňažného trhu, najviac pre doby splatnosti medzi druhým až siedmym rokom a ešte podstatnú mieru u dlhodobých úrokových sadzbách. Druhý hlavný komponent má opačné správanie. Najviac vystihuje krátkodobé a dlhodobé úrokové miery. Tretí hlavný komponent vysvetľuje ešte varianciu hlavne v sadzbách peňažného trhu a menej u dlhodobých.

Na obrázku 2.5 sú zobrazené úrokové výnosy prvých troch hlavných komponentov vzhľadom k dobe splatnosti. Vyjadrujú, aký veľký efekt má každý hlavný komponent na jednotlivé doby splatnosti bez posudzovania ich celkovej miery efektívnosti.



Obr. 2.4: Explanatory power prvých troch komponentov



Obr. 2.5: Výnosové krivky tvorené hodnotami hlavných komponentov zobrazených v stĺpcoch tabuľky 2.1

splatnosť	1m	3m	6m	1r	2r	3r	5r	7r	10r	20r	30r
1. portfólio	2	-4	-5	-5	-5	1	3	-1	0	-2	-5
2. portfólio	0	5	4	5	0	1	4	-5	-5	1	-1
3. portfólio	1	0	0	2	3	-5	-3	-2	3	0	-4
4. portfólio	1	-2	-4	-1	-5	-5	-3	-3	5	-2	5
5. portfólio	1	1	3	-3	0	0	2	0	-5	1	-5

Tabuľka 2.3: Citlivosť hodnoty portfólia na zmenu úrokovej miery o jeden bázický bod (v miliónoch dolárov)

	1.PC	2.PC	3.PC
1. portfólio	4.90	-0.93	-5.36
2. portfólio	-1.57	5.19	6.57
3. portfólio	2.03	2.45	0.49
4. portfólio	4.36	-0.82	-6.92
5. portfólio	1.48	2.97	2.02

Tabuľka 2.4: Expozície portfólií pre jednotlivé komponenty pri zmene úrokových mier o jeden bázický bod (v miliónoch dolárov)

Pre ilustráciu výpočtu VaR za pomoci hlavných komponentov predpokladajme, že máme portfólio s expozíciami na zmenu úrokovej sadzby zobrazenými v tabuľke 2.3. Tieto hodnoty sú náhodne vygenerované v rozmedzí -5 až 5 miliónov dolárov. Pre porovnanie vypočítame viacero hodnôt VaR pre rôzne portfólia.

V prvom portfóliu zmena o jeden bázický bod v jednomesačnej úrokovej miere spôsobí nárast hodnoty portfólia o 2 milióny dolárov, zmena o jeden bázický bod v trojmesačnej úrokovej miere spôsobí pokles hodnoty portfólia o 4 milióny dolárov, atď. Na simulovanie pohybu sadzieb využijeme vyššie určené prvé tri hlavné komponenty. Užitím hodnôt z tabuľky 2.1 je expozícia pre prvý hlavný komponent prvého portfólia

$$\begin{gathered} 2\cdot(-0.30)-4\cdot(-0.24)-5\cdot(-0.21)-5\cdot(-0.23) \\ -5\cdot(-0.31)+1\cdot(-0.35)+3\cdot(-0.38)-1\cdot(-0.37) \\ +0\cdot(-0.33)-2\cdot(-0.28)-5\cdot(-0.27)=4.90, \end{gathered}$$

expozícia pre druhý hlavný komponent prvého portfólia je

$$\begin{gathered} 2\cdot(0.70)-4\cdot(0.46)-5\cdot(0.23)-5\cdot(0.13) \\ -5\cdot(-0.07)+1\cdot(-0.11)+3\cdot(-0.19)-1\cdot(-0.22) \\ +0\cdot(-0.22)-2\cdot(-0.21)-5\cdot(-0.20)=-0.93. \end{gathered}$$

Ďalšie výsledky výpočtov pre päť rôznych portfólií sú zachytené v tabuľke 2.4.

Predpokladajme, že f_1, f_2 a f_3 sú skóre prvých troch hlavných komponentov (merané v bázických bodoch). Zmeny hodnoty portfólií $P_1, P_2, ..., P_5$ sú odhadnuté

	1. portfólio	2. portfólio	3. portfólio	4. portfólio	5. portfólio
σ_i	1.053220	0.892534	0.525867	0.997587	0.518930
$VaR\ P_i$	2.45401	2.07960	1.22527	2.32438	1.20911

Tabuľka 2.5: Smerodajné odchýlky $\sigma_i,\ i = 1, 2, ..., 5$ a hodnoty jednodenných 99% VaR pre portfólia $P_i,\ i = 1, 2, ..., 5$ (v miliónoch dolárov)

funkciami

$$\begin{aligned}
\Delta P_1 &= 4.90 f_1 - 0.93 f_2 - 5.36 f_3, \\
\Delta P_2 &= -1.57 f_1 + 5.19 f_2 + 6.57 f_3, \\
\Delta P_3 &= 2.03 f_1 + 2.45 f_2 + 0.49 f_3, \\
\Delta P_4 &= 4.36 f_1 - 0.82 f_2 - 6.92 f_3, \\
\Delta P_5 &= 1.48 f_1 + 2.97 f_2 + 2.02 f_3.
\end{aligned}$$

Skóre hlavných komponentov sú nekorelované a ich smerodajné odchýlky vidíme v tabuľke 2.2. Pri výpočte však použijeme priamo rozptyly pre menšie zaokrúhľovacie chyby. Smerodajná odchýlka funkcie ΔP_1 je teda

$$\sigma_1 = \sqrt{4.90^2 \cdot \lambda_1 + 0.93^2 \cdot \lambda_2 + 5.36^2 \cdot \lambda_3}.$$

Numerickú hodnotu σ_1 nájdeme v tabuľke 2.5, ako aj ostatné výsledky výpočtov pre zvyšné portfólia. Potom jednodenný 99% VaR pre prvé portfólio je

$$\sigma_1 \cdot z_{0.01} \cdot \sqrt{\Delta t} = 1.05322 \cdot 2.33 \cdot 1 = 2.45401,$$

kde $z_{0.01}$ predstavuje 99% kvantil normovaného normálneho rozdelenia $\mathsf{N}(0, 1)$. To znamená, že máme 99% istotu, že v následujúcom dni nestratíme viac než 2.45401 milióna dolárov. VaR všetkých portfólií je zachytená v tabuľke 2.5.

Z piatich expozíc z tabuľky 2.3 nám podľa výsledkov v tabuľke 2.5 vychádza, že najvýhodnejšie rozloženie expozíc je piate, pretože hodnota VaR v tomto prípade vyšla najnižšia. Pri piatom portfóliu máme teda 99% istotu, že v následujúcom dni nestratíme viac než 1.20911 milióna dolárov.

Na výpočet VaR sme aplikovali metódu hlavných komponentov, ktorá transformovala vstupný súbor údajov a tým znížila dimenziu dátovej štruktúry. V analýze počtu hlavných komponentov sme sa rozhodli pre prvé tri komponenty, čím sa rozmernosť dát podstatne zredukovala. Ak by sme na výpočet hodnoty v riziku použili napríklad len prvé dva komponenty, VaR by pri každom portfóliu klesla v priemere až o 7%. Preto je dôležité správne analyzovať potrebný počet hlavných komponentov. Inak by VaR vyšla príliš mierna.

Záver

V tejto práci sme sa pokúsili zachytiť všetky dôležité aspekty metódy hlavných komponentov. Väčšinu teoretických základov popísaných v prvej kapitole sme overili na ilustračných príkladoch. Metóda hlavných komponentov teda poskytuje postup na zjednodušenie vstupného súboru dát, pričom zachováva pomerné množstvo informácií. Tým je možné ľahšie uskutočniť následnú interpretáciu.

Prvá časť práce poskytuje teoretické základy pre výpočet hlavných komponentov. Na začiatku sme sa zamerali na odvodenie hlavných komponentov a ich vlastnosti. Ďalej sme spomenuli problematiku výberových hlavných komponentov. Nakoniec sme sa venovali situáciám, ktoré sa v priebehu analýzy môžu vyskytnúť, ako napríklad výber počtu relevantných hlavných komponentov alebo použitie kovariančnej či korelačnej matice.

V druhej časti sme tieto poznatky použili na konkrétnych reálnych dátach. Najprv sme však definovali hodnotu v riziku, ktorej výpočet je v závere tejto práce. Hodnota v riziku zahrňuje všetky druhy tržných rizík a odhaduje maximálnu stratu spôsobenú nepriaznivými zmenami tržných sadzieb. Dôležité parametre pre výpočet hodnoty VaR sú časový horizont, počas ktorého sa odhadovaná strata uvažuje a spoľahlivosť, s ktorou skutočná strata neprevýši hodnotu v riziku. Pri konkrétnych vzorcoch na výpočet hodnoty v riziku sme sa zamerali aj na výpočet, kde zisk (miera zisku) má normálne rozdelenie, ktorý sme potom využili na konci kapitoly pre výpočet hodnoty VaR konkrétnych portfólií. Za trhové premenné sme zvolili úrokové sadzby pre jedenásť rôznych dôb splatnosti. Keďže nás zaujímalo ako sa menia úrokové miery v čase, pracovali sme so súborom prvých diferencií sadzieb dvoch nasledujúcich dní. Vďaka metóde hlavných komponentov sme určili, že postačuje ak sa zaoberáme prvými troma hlavnými komponentami. Tým sme zredukovali dimenziu dát, čo nám výpočet hodnoty v riziku zjednodušilo.

Bibliografia

Cipra T.: *Finančnı́ a pojistné vzorce*, Grada, Praha, 2006

Dupačová J., Hurt J., Štěpán J.: *Stochastic modeling in economics and Finance*, Kluwer Academic Publishers, Dordrecht, 2002

Dupač V., Hušková M.: *Pravděpodobnost a matematická statistika*, Karolinum, Praha, 1999

Härdle W., Hlávka Z.: *Multivariate Statistics: Exercises and Solutions*, Springer, New York, 2007

Härdle W., Simar L.: *Applied Multivariate Statistical Analysis*, Springer, Berlin, 2003

Hull J.C.: *Options, Futures and Other Derivatives (6th Edition)*, Pearson/Prentice Hall, Upper Saddle River, 2006

Chatfield C., Collins A.J.: *Introduction to multivariate analysis*, Chapman & Hall, Boca Raton, 2000

Jolliffe I.T.: *Principal Component Analysis, Second Edition*, Springer, New York, 2002

Malava A.: *Principal Component Analysis on Term Structure of Interest Rates*, Helsinki University of Technology, Helsinky, 2006

Stankovičová I., Vojtková M.: *Viacrozmerné štatistické metódy s aplikáciami*, IURA EDITION, spol. s. r. o., Bratislava, 2007

http://www.treasury.gov/resource-center/data-chart-center/interest-rates