UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

DIPLOMOVÁ PRÁCE

Metoda PARAFAC a její aplikace

Vedoucí diplomové práce:
doc. RNDr. Karel Hron, Ph.D.
Rok odevzdání: 2014

Vypracovala:
Bc. Šárka Brodinová
AME, II. ročník

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci sepsala samostatně pod vedením pana doc. RNDr. Karla Hrona, Ph.D. a že jsem v seznamu literatury uvedla všechny použité zdroje při zpracování práce.

V Olomouci dne 17. března 2014

Poděkování

Na tomto místě bych chtěla poděkovat především svému vedoucímu diplomové práce panu doc. RNDr. Karlu Hronovi, Ph.D., že měl se mnou dostatek trpělivosti, aby mi pomohl dovézt tuto práci ke zdárnému konci. Také bych ráda poděkovala své rodině, přátelům a spolužákům, kteří mě po celou dobu studia podporovali. Mé díky patří i tvůrcům typografického programu LaTeX, ve kterém je práce napsána.

Obsah

Úvod

Mnoha vědcům, ať už chemikům, biologům či ekonomům, je jasné, že bez zpracování dat (informací), které vzešly z jejich experimentů, nejsou schopni učinit žádné závěry a publikovat je. V pozadí každé publikace stojí i práce analytika (statistika), který je schopen z napozorovaných dat získat komplexnější představu o zkoumané problematice. Většinu dat pocházejících z experimentů je obtížné analyzovat kvůli jejich struktuře, proto je potřeba najít takový analytický nástroj či model, který by pomohl pochopit přirozenou strukturu těchto dat a vyvodit z nich závěry. Cílem této práce je představit jeden z možných modelů, který lze použít k tomuto účelu.

Většina experimentů je navržena tak, že jsme jejich výsledky schopni uspořádat do datové matice, např. pacienti a jejich výsledky různých vyšetření (měření tlaku, krevní obraz, atd.). Chceme-li dostat souhrnný obrázek o pacientech, můžeme se například zajímat i o průběh jednotlivých vyšetření v čase. Pro pochopení struktury takto získaných dat je vhodné použít metodu hlavních komponent, PCA (Karl Pearson, 1901), metodu Tucker3 (Kroonenberg a de Leeuw, 1980) nebo nezávisle na sobě navržené metody PARAFAC (Harshman a Berenbaum, 1981) a CANDECOMP (Carroll a Chang, 1970). Zmíněné metody rozkládají datové pole na množinu skórů a zátěží, které uchovávají co nejvíce informace z původního datového souboru a umožní tak další analýzu zkoumaných dat již v jednodušší struktuře. Tato práce se věnuje základním principům metody PCA a především teoretickým i praktickým aspektům metody PARAFAC. Dále si klade za cíl, aby čtenář pochopil rozdílné a společné vlastnosti těchto dvou metod a byl schopen s těmito metodami pracovat v prostředí statistického softwaru R.

První kapitola je věnována metodě PCA (Pricipal Component Analysis, analýza hlavních komponent); jednak jsou zde uvedeny základní pojmy, způsoby výpočtu hlavních komponent, ale i grafická interpretace výsledků v podobě biplotu. Na konec se čtenář seznámí s některými vlastnostmi této dekompoziční metody.

K pochopení základních algebraických pojmů a operací, jež jsou používány

v souvislosti s vícerozměrnými datovými poli, je určena druhá kapitola. Je zde například zmíněn Khatri-Raoův součin představující speciální případ Kroneckerova součinu. Součástí kapitoly jsou i příklady, které napomohou k lepšímu pochopení uvedených operací.

Třetí kapitola je pro tuto práci stěžejní a je věnována samotné metodě PARAFAC (PARAllel FACtor analysis), která se v současné době stává mnohem více užívanější, a to nejen díky jednoznačnému řešení, ale i uplatnění pro další zpracování. Metoda představuje zobecnění analýzy hlavních komponent a její výpočet probíhá iteračním způsobem, který je založen na metodě ALS (Alternating Least Squaers). V kapitole jsou uvedeny též praktické aspekty metody, na které by si každý analytik měl dát pozor, např. předzpracování dat či určení počtu komponent.

Praktická část je věnována aplikaci metody PARAFAC na reálných datech z různých vědních disciplín. Veškeré výpočty jsou zpracovány za pomoci statistického softwaru R a postup výpočtů je i podrobně popsán. Výstupem metody jsou zátěžové grafy, které jsou interpretovány jak zvlášť, tak i jako celek.

1 Metoda hlavních komponent (PCA)

Metoda hlavních komponent je vedle diskriminační či faktorové analýzy jednou z nejpoužívanějších vícerozměrných metod. Jelikož je metoda PARAFAC určitým zobecněním analýzy hlavních komponent, je jistě na místě se nejprve zmínit o této metodě. V kapitole si uvedeme základní principy, cíle a vlastnosti PCA (Pricipal Component Analysis), aby nám pomohly lépe pochopit fungování metody PARAFAC.

Hlavním cílem metody hlavních komponent je redukce dimenze prostoru naměřených vícerozměrných dat (znaků, proměnných), které charakterizují zkoumané objekty, jevy či procesy, s minimální ztrátou informace o prostorové struktuře dat. Metoda se jeví být výhodná také při zobrazování vysoce dimenzionálních dat, při velkém počtu proměnných, které mohou být i vzájemně závislé. Ve všech těchto případech dochází ke zjednodušení varianční struktury experimentálních dat.

Techniky k dosažení výše uvedených cílů budou zmíněny v následujících podkapitolách. Nejprve se zmíníme o lineární transformaci původních proměnných na nové latentní proměnné [2, 5, 8] nazývající se hlavní komponenty. K tomuto účelu se používá zejména spektrální rozklad varianční matice. Jiný pohled spočívá v rozkladu zdrojové matice na množinu skórů a zátěží [13, 5], k čemuž slouží singulární rozklad. A na závěr se na PCA podíváme jako na optimalizační úlohu [4], což nás přiblíží k samotné metodě PARAFAC. Čtenáři bude nastíněno, jak jednotlivé techniky fungují, pro detailnější vysvětlení je doporučena citovaná literatura.

1.1 PCA přístup I

Informace o prostorové struktuře dat jsou obsaženy v jejich variabilitě, proto je jedním z nástrojů spektrální rozklad varianční matice pro konstrukci hlavních komponent. Technika této metody spočívá v lineární transformaci původního J-složkového vektoru $\mathbf{x}$ na nový latentní (skrytý) vektor $\mathbf{a} = (A_1, \ldots, A_F)$ s menším počtem proměnných tak, aby zachoval co nejvíce informace ve smyslu za-

chování variability. Nové proměnné se nazývají hlavní komponenty a jsou oproti původním složkám vektoru $\mathbf{x}$ nekorelované.

Mějme populaci, ve které náhodné veličiny $X_1, X_2, \ldots, X_J$ tvoří náhodný vektor $\mathbf{x}$ pocházející z mnohorozměrného normálního rozdělení s varianční maticí $\boldsymbol{\Sigma}$ a bez újmy na obecnosti budeme předpokládat nulový vektor středních hodnot $\boldsymbol{\mu}$. Než přistoupíme ke konstrukci hlavních komponent, uvedeme si větu pojednávající o spektrálním rozkladu [7] čtvercové matice.

Věta 1.1. *Nechť je dána symetrická matice* $\mathbf{H}$ *řádu* J*. Pak existují ortogonální matice* $\mathbf{B}$ *a diagonální matice* $\boldsymbol{\Lambda}$ *takové, že*

$$\mathbf{H} = \mathbf{B}\boldsymbol{\Lambda}\mathbf{B}^\top = \sum_{j=1}^{J} \lambda_j \mathbf{b}_j \mathbf{b}_j^\top, \tag{1}$$

kde $\boldsymbol{\Lambda} = diag\{\lambda_1, \ldots, \lambda_J\}, \lambda_1 \geq \cdots \geq \lambda_J$ *jsou vlastní čísla* $\mathbf{H}$ *a* $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_J)$ *je matice ortonormálních* [1] *vlastních vektorů* $\mathbf{H}$*.*

Poznámka 1.1. *Je-li matice* $\mathbf{H}$ *pozitivně semidefinitní a platí-li* $r(\mathbf{H}) = F < J$*, pak prvních* F *vlastních čísel je kladných a zbytek, tj.* $J - F$*, nulových. Pro ortogonální matici* $\mathbf{B}$ *platí že,* $\mathbf{B}^\top = \mathbf{B}^{-1}, \mathbf{B}^\top\mathbf{B} = \mathbf{B}\mathbf{B}^\top = \mathbf{I}$*. Symbol* r *značí hodnost matice.*

Hlavní komponenty jsou vytvářeny postupně a seřazeny podle důležitosti. První komponenta vysvětluje nejvíce z celkové variability a každá další pak zbývající variabilitu. Poslední komponenta tudíž obsahuje nejméně informace a není důležitá. Pokud původní znak, který charakterizuje zkoumané objekty, má značně malý rozptyl, tj. nijak nepřispívá k rozlišení mezi objekty, nebude obsažen v hlavní komponentě. Při konstrukci první hlavní komponenty hledáme takový vektor $\mathbf{d} \in \mathbb{R}^F$, který splňuje podmínku $\mathbf{d}^\top\mathbf{d} = 1$ a zároveň požadujeme, aby náhodná veličina $\mathbf{d}^\top\mathbf{x}$, tj. lineární kombinace složek původního vektoru, měla co největší rozptyl. Jelikož platí $\operatorname{var}\mathbf{d}^\top\mathbf{x} = \mathbf{d}^\top\boldsymbol{\Sigma}\mathbf{d}$, budeme maximalizovat výraz na pravé

[1] Pro ortonormální vektory platí $\mathbf{b}_j^\top\mathbf{b}_k = \delta_{jk}$. Je-li $j = k$, pak hodnota výrazu je 1, jinak 0.

straně za podmínky $\mathbf{d}^\top\mathbf{d} = 1$. Podle lemmatu 1 uvedeného v [2] na str 297 a s využitím spektrálního rozkladu matice $\mathbf{\Sigma}$ vyplývá, že maximum je dosaženo pokud $\mathbf{d} = \mathbf{b}_1$ a jeho hodnota je rovna λ_1. Proto první hlavní komponenta je ve tvaru $A_1 = \mathbf{b}_1^\top\mathbf{x}$. Při konstrukci dalších komponent postupujeme dle lemmatu 3 [2], proto $A_2 = \mathbf{b}_2^\top\mathbf{x}$ a obdobně až po A_F. Obecně f-tá hlavní komponenta je ve tvaru

$$A_f = \mathbf{b}_f^\top\mathbf{x}, \qquad f = 1, \ldots, F, \tag{2}$$

a při použití spektrálního rozkladu $\mathbf{\Sigma}$ platí

$$\operatorname{var} A_f = \operatorname{var}\mathbf{b}_f^\top\mathbf{x} = \mathbf{b}_f^\top\mathbf{\Sigma}\mathbf{b}_f = \mathbf{b}_f^\top\sum_k \lambda_k\mathbf{b}_k\mathbf{b}_k^\top\mathbf{b}_f = \lambda_f, \tag{3}$$

$$\operatorname{cov}(A_f, A_k) = \operatorname{cov}(\mathbf{b}_f^\top\mathbf{x}, \mathbf{b}_k^\top\mathbf{x}) = \mathbf{b}_f^\top\mathbf{\Sigma}\mathbf{b}_k = \mathbf{b}_f^\top\sum_l \lambda_l\mathbf{b}_l\mathbf{b}_l^\top\mathbf{b}_k = \lambda_f\delta_{fk}, \tag{4}$$

kde $f, k = 1, \ldots, F$. Vztah (3) vypovídá o postupném vyčerpávání variability a vztah (4) chápeme jako splnění podmínky nekorelovanosti jednotlivých komponent. Poznamenejme, že v případě $F = J$, nutně platí $tr(\operatorname{var}\mathbf{a}) = tr(\mathbf{\Sigma})$, tedy celková variabilita nově vytvořeného vektoru odpovídá celkové variabilitě původního vektoru proměnných. Symbol tr značí stopu příslušné matice.

Pro praxi je důležité ohodnotit, jak moc veličina A_f přispívá v celkové variabilitě, k tomuto účelu slouží podíl vysvětlené variability, který je zaveden jako $\lambda_f/tr(\mathbf{\Sigma})$. Je-li hodnota blízko jedné, pak komponenta dostatečně vysvětluje variabilitu $\mathbf{x}$. Naopak hodnoty blízko nule vypovídají o nepatrném přispění k celkové variabilitě $\mathbf{x}$. Celkový počet použitých hlavních komponent je určen tak, aby součet jejich rozptylů podělený $tr(\mathbf{\Sigma})$ byl blízko jedné, a udává skutečnou dimenzi vektoru $\mathbf{x}$. Výsledky analýz ukazují, že nejčastějším případem jsou dvě až čtyři komponenty. Z grafického hlediska je pak nejvýhodnější mít nejvýše tři komponenty.

Při statistické analýze experimentálních dat máme k dispozici náhodný výběr o rozsahu I s J proměnnými, tj. $\mathbf{X} \in \mathbb{R}^{I\times J}$. Protože neznáme varianční

matici $\boldsymbol{\Sigma}$ ani vektor středních hodnot $\boldsymbol{\mu}$, provedeme jejich odhad pomocí výběrových charakteristik (výběrový průměr $\bar{\mathbf{x}}$ a výběrová varianční matice $\mathbf{S}$). Z těchto odhadů vypočítáme vlastní vektory a můžeme sestavit hlavní komponenty. V praxi se často stává, že hodnoty jednotlivých proměnných mají rozdílné jednotky. Abychom předešli zkresleným výsledkům, přistupujeme buď k centrování, ke škálování (standardizaci), nebo k logaritmické transformaci datového souboru.

Hodnota hlavní komponenty pro jednotlivé výběrové jednotky se nazývá komponentní skóre. Hodnota f-té komponenty u i-tého objektu je

$$a_{if} = \mathbf{b}_f^\top \mathbf{x}_i, \qquad i = 1, \ldots, I,\ f = 1, \ldots, F. \tag{5}$$

Uspořádání těchto hodnot do matice $\mathbf{A} \in \mathbb{R}^{I \times F}$ tvoří **matici skórů**, přičemž tato matice dostatečně vysvětluje chování experimentálních dat v matici $\mathbf{X}$.

Získané vlastní vektory $\mathbf{b}_f$ varianční matice se nazývají vektory komponentních zátěží a tvoří **matici zátěží** $\mathbf{B} \in \mathbb{R}^{F \times I}$.

Poznámka 1.2. *Na začátku kapitoly jsme předpokládali bez újmy na obecnosti nulový vektor středních hodnot, proto* $A_f = \mathbf{b}_f^\top \mathbf{x}$. *V některých literaturách a při práci s daty se používá vztah* $A_f = \mathbf{b}_f^\top (\mathbf{x} - \boldsymbol{\mu})$, *uvažujeme-li nenulový vektor středních hodnot.*

1.2 PCA přístup II

Nyní předpokládejme, že už máme k dispozici náhodný výběr tvořící zdrojovou matici $\mathbf{X} \in \mathbb{R}^{I \times J}$. Jak již bylo zmíněno, základním cílem PCA je transformovat původní počet proměnných na menší počet latentních proměnných (komponent). Jinými slovy, provádíme aproximaci objektů zobrazených v euklidovském prostoru (ortogonální systém souřadnic rozměru J) na souřadnice objektů v prostoru hlavních komponent. Rozdíl mezi těmito souřadnicemi označujeme jako chybu modelu. Dochází tedy k rozkladu původní matice $\mathbf{X}$ na skutečnou strukturu (první komponenty, které vysvětlují dostatečnou variabilitu dat) a šum

neboli chybu modelu PCA (zbývající komponenty, které popisují nejmenší proměnlivost). Situaci znázorňuje vztah

$$\mathbf{X} = \mathbf{A}\mathbf{B}^\top + \mathbf{E}. \tag{6}$$

Přitom $\mathbf{A}$ je matice skórů a prvky této matice jsou souřadnicemi objektů v prostoru hlavních komponent. Projekce i-tého objektu například na první hlavní komponentu je skóre a_{i1}. $\mathbf{B}$ je matice, jejíž prvky nám dávají informaci o vztahu mezi původními proměnnými a vytvořenými hlavními komponentami, např. b_{12} odpovídá velikosti přispění druhého znaku na první hlavní komponentu. $\mathbf{E}$ je matice reziduí a souvisí s mírou těsnosti proložení dat modelem PCA. Vlivem aproximace dochází ke ztrátě informace, velikost rezidua i-tého objektu je $\mathbf{e}_i$.

Skutečná struktura dat (rozměr) je tedy určena maticemi skórů a zátěží, ty můžeme obdržet buď pomocí spektrálního rozkladu (kapitola 1.1) výběrové variančí matice nebo singulárního rozkladu [7] zdrojové matice.

Věta 1.2. *Nechť je dána reálná matice* $\mathbf{X}$ *typu* $I \times J$*. Pak existují ortogonální matice* $\mathbf{U}$ *typu* $I \times I$ *a* $\mathbf{V}$ *typu* $J \times J$ *takové, že matici* $\mathbf{X}$ *lze zapsat ve tvaru součinu*

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top, \tag{7}$$

kde $\mathbf{D}$ *je diagonální matice rozměru* $I \times J$ *s nezápornými prvky na hlavní diagonále, tzv. singulárními hodnotami, které jsou uspořádané sestupně podle velikosti.*

Z [5] vyplývá, že při použití několika prvních, nejčastěji dvou sloupců matice $\mathbf{U}$ a $\mathbf{V}$ ze singulárního rozkladu, dostáváme součin dvou matic, který aproximuje zdrojovou matici ve smyslu metody nejmenších čtverců a odpovídá tak (2) pro $F = 2$. Model lze přepsat do tvaru

$$\mathbf{X}^{I\times J} \approx \widehat{\mathbf{X}}^{I\times J}_{(2)} = \mathbf{U}\mathbf{D}\mathbf{V}^\top = (\mathbf{u}_1\ \mathbf{u}_2)\begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}\begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{pmatrix} = \mathbf{A}\mathbf{B}^\top, \tag{8}$$

kde volíme $\mathbf{A} = \mathbf{U}\mathbf{D}$ a $\mathbf{B} = \mathbf{V}$, dolní index (2) značí počet použitých komponent. Tento přístup se jeví jako výhodný při konstrukci grafu skórů a zátěží metody

hlavních komponent, tzv. biplotu. Z [5] po vhodných úpravách vyplývá, že

$$\begin{aligned} \mathbf{a}_{i.}^{\top}\mathbf{b}_{j.} &\approx x_{ij}, \\ \mathbf{A}\mathbf{A}^{\top} &\approx \mathbf{S}. \end{aligned} \tag{9}$$

Podívejme se blíže na (9), pravou stranu lze rozepsat pomocí (1), do levé strany pouze dosadíme

$$\begin{aligned} \mathbf{S} = \hat{\mathbf{\Sigma}} &= \mathbf{B}\mathbf{\Lambda}\mathbf{B}^{\top}, \\ \mathbf{A}\mathbf{A}^{\top} &= \mathbf{U}\mathbf{D}(\mathbf{U}\mathbf{D})^{\top} = \mathbf{U}\mathbf{D}^2\mathbf{U}^{\top}, \end{aligned}$$

kde $\mathbf{D}^2$ označme pro další použití jako $\mathbf{D}^*$. Z výše rozepsaných vztahů lze vyvodit, že prvky matice $\mathbf{D}$ představují odmocněná vlastní čísla varianční matice a $\mathbf{U}$ je matice příslušných charakteristických vlastních vektorů. Proto i tento přístup pracuje s variabilitou dat a v jistém smyslu ji rozkládá, obdržíme tak stejné hodnoty skórů a zátěží jako v předchozí kapitole.

1.2.1 Biplot

Na závěr si ještě nastíníme grafické zpracování PCA (více v [5, 13]), aby pro čtenáře bylo snazší pochopit a interpretovat grafický výstup určený metodou PARAFAC. Pro zobrazení výsledků získaných metodou PCA se nejčastěji používá biplot. Tento dvojný graf umožňuje zobrazit jak komponentní skóry, tak i zátěže prvních dvou hlavních komponent v rovině (viz vztah (8)). Konstrukce biplotu vychází ze singulárního rozkladu (dekompozice) původní matice $\mathbf{X}$, jak jsme si ukázali výše a jeho interpretace je velmi intuitivní.

Řádky matice $\mathbf{A}$ jsou reprezentovány v grafu body a odpovídají zkoumaným objektům. Čím blíž jsou body u sebe, tím více se jednotlivé objekty sobě podobají a v některých případech mohou vytvářet i shluky. Body daleko od počátku poukazují na neobvyklé vlastnosti objektu.

Řádky matice $\mathbf{B}$ představují vrcholy šipek a jsou projekcí znaků. Délky šipek jsou úměrné variabilitě příslušného znaku. Kosinus úhlu, který svírají dvě šipky, představuje korelační koeficient těchto statistických znaků (sílu lineární

závislosti mezi jednotlivými znaky). Nachází-li se objekt blízko určitému znaku, znamená to, že tento znak je ve velké míře obsažen v tomto objektu. Čtenář tuto problematiku lépe pochopí z následujícího příkladu.

1.2.2 Příklad

Cílem je analyzovat emise znečišťujících látek (REZZO 1–3) na jednoho obyvatele podle krajů v roce 2002 (viz příloha A.1, rok 2002). Data byla získána ze Statistické ročenky České republiky 2002 [18]. Analýza byla provedena ve statistickém softwaru R. Měření se provádělo ve vybraných krajích, v biplotu jsou označeny zkratkami, např. Pardubický kraj (Pa). V jednotlivých krajích byly měřeny emise znečišťujících látek (emise tuhé, oxid siřičitý, oxidy dusíku a oxid uhelnatý).

Po nastavení pracovního adresáře, ve kterém je uložen datový soubor, byl k načtení dat použit příkaz:

```
>setwd("D:/UPOL/Mgr/diplomka/data/Emise")
>d1=as.matrix(read.table("02.txt", header=FALSE, sep=""))
>X=as.matrix(d1)
```

Pro přehlednost byly pojmenovány jednotlivé proměnné i objekty:

```
>colnames(X)= c("EmiseTuhe","SO2","Nox","CO")
>rownames(X)= c("JC","Pl","Ka","U","Li","Kr","Pa","Vy","JM","Ol","Z","Ms")
```

Jednotky jsou ve stejném měřítku, ale u některých krajů byly zjištěny výrazně vyšší hodnoty než u ostatních, proto byla data zlogaritmována, aby byla dobře porovnatelná:

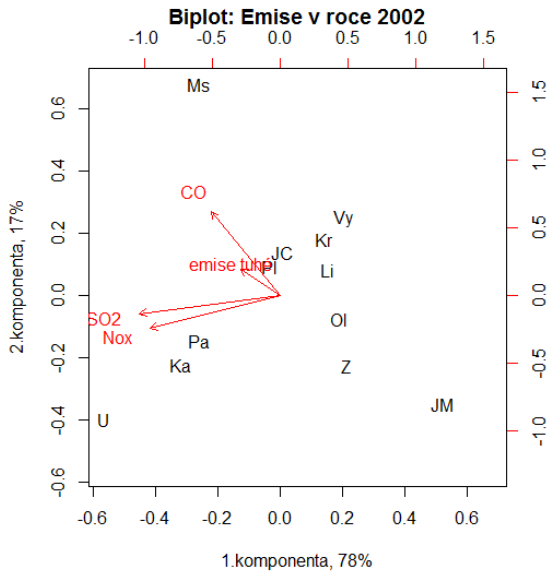
```
>Emise=log(X)
```

Na transformovaných datech byla použita PCA a zjištěny následující informace:

```
> summary(princomp(Emise))
Importance of components:
                          Comp.1    Comp.2     Comp.3      Comp.4
Standard deviation     0.5521415 0.2545473 0.13270255 0.060625290
Proportion of Variance 0.7798134 0.1657399 0.04504519 0.009401508
Cumulative Proportion  0.7798134 0.9455533 0.99059849 1.000000000
```

Z toho vyplývá, že můžeme použít dvě hlavní komponenty, jelikož součet podílů vysvětlované variability je 0.94. Můžeme tedy přikročit k použití funkce `biplot`, která slouží k zobrazení skórů a zátěží prvních dvou komponent.

Výsledky biplotu v prvé řadě ukazují, že znečištění oxidy dusíku a oxidem siřičitým je na sobě silně závislé. Prvky zleva doprava vypovídají o velikosti znečistění oxidy dusíku a oxidem siřičitým na jednoho obyvatele v jednotlivých krajích. Podíváme-li se na Ústecký kraj, jeho znečištění těmito emisemi je opravdu vysoké, o čemž vypovídají i data. V Pardubickém i Karlovarském kraji je situace podobná, ale v menším měřítku. Uspořádání prvků shora dolů je dáno znečištěním oxidem uhelnatým a tuhými emisemi, proto jsou obyvatelé v Mosteckém kraji nejvíce postiženi. Jihomoravský kraj má nejmenší míru znečištění uvedenými emisními látkami.



Obrázek 1: Zobrazení skórů a zátěží.

To vyplývá z toho, že kraje, které jsou v opačném směru než šipky, mají nejnižší znečištění na obyvatele. Další skupina krajů ve středu grafu je spíše postižena oxidem uhelnatým a tuhými emisemi. Všechny zkoumané proměnné přispívají hlavně do první komponenty.

1.3 PCA přístup III a jeho vlastnosti

Uvažujme situaci (6) s cílem najít takový model, jehož míra těsnosti aproximace zadaných dat bude co nejvyšší. Pro posouzení těsnosti aproximace se používá reziduální součet čtverců

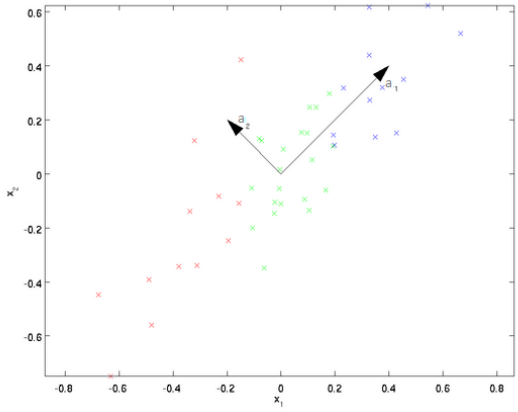
$$SSE = ||\mathbf{X} - \mathbf{A}\mathbf{B}^\top||_F^2. \tag{10}$$

Jelikož požadujeme, aby vytvořený model zachoval co nejvíce informací z původních dat, budeme hodnotu SSE minimalizovat. Touto úvahou se dostáváme k optimalizační úloze

$$\min_{\mathbf{A},\mathbf{B}} ||\mathbf{X} - \mathbf{A}\mathbf{B}^\top||_F^2 \tag{11}$$

za podmínky $\mathbf{A}^\top\mathbf{A} = \mathbf{D}^*$, $\mathbf{B}^\top\mathbf{B} = \mathbf{I}$, což nás přibližuje k základní myšlence fungování metody PARAFAC (viz kapitola 3).

Než si řekneme o vlastnostech metody PCA, je nezbytné si ještě ujasnit pojmy jako metoda, model a struktura modelu. Metoda je chápána jako určitý nástroj k tomu, abychom mohli nalézt v datech souvislosti, zákonitosti. Neboli, aplikací dané metody, ať už PCA či jiné, aproximujeme data a vytváříme model. Co se týče struktury modelu, ta je v případě použití metody PCA bilineární, jelikož dochází k lineární kombinaci dvou typů parametrů (skórů a zátěží). Jak jsme si ukázali v předchozích podkapitolách, hlavní komponenty lze konstruovat jak přes spektrální rozklad varianční matice, tak i přes singulární rozklad zdrojové matice, a proto se tato metoda řadí mezi dekompoziční metody dvourozměrných dat. K tomu, aby nám metoda nalezla správný model a parametry dávající smysl, musí být zavedena omezení (podmínky). V případě PCA jsou požadovány podmínky ortogonality a nekorelovanost umělých proměnných. Ortogonalita je určena podmínkou $\mathbf{B}^\top\mathbf{B} = \mathbf{I}$. K nalezení hlavních komponent je potřeba provést rotaci původního souřadnicového systému tak, aby nové osy procházely směry maximálního rozptylu shluku bodů a aby takto vytvořené osy byly na sebe kolmé. Takto pak můžeme chápat i grafickou interpretaci metody PCA, situaci znázorňuje obrázek 2. Je zřejmé, že splnění takového požadavku lze dosáhnout jedině použitím ortogonální rotace.



Obrázek 2: Transformace původního souřadnicového systému.

Podmínka nezápornosti dekompozice datového souboru, $\mathbf{A}^\top\mathbf{A} = \mathbf{D}^*$, zaručuje nekorelovanost hlavních komponent. U PCA nastává problém s jednoznačností výsledného modelu a to ve dvojím smyslu. Jednak jde o nejednoznačnost při orientaci rotovaných souřadnic, jelikož orientace nových souřadnic může být jak kladná, tak i záporná. A druhý typ nejednoznačnosti je, že výsledné řešení může být rotováno jakoukoliv ortogonální maticí. Uvažujme aproximaci matice $\mathbf{X}$ v podobě její dekompozice na matici skórů a zátěží (viz (6)). Hodnost aproximované matice $\hat{\mathbf{X}}$ je rovna $F \leq J$. Situaci zapíšeme následovně

$$\hat{\mathbf{X}} = \mathbf{A}\mathbf{B}^\top = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f, \tag{12}$$

kde $\circ$ je tenzorový součin, pro který platí $\mathbf{a} \circ \mathbf{b} = \mathbf{a}\mathbf{b}^\top$. Z (7) víme, že singulární rozklad matice $\mathbf{X}$ je $\mathbf{U}\mathbf{D}\mathbf{V}^\top$, a můžeme zvolit $\mathbf{A} = \mathbf{U}\mathbf{D}$ a $\mathbf{B} = \mathbf{V}$. Volbu těchto matic můžeme ovšem provést i jinak. Zvolíme-li matici skórů $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{W}$ a matici zátěží jako $\mathbf{B} = \mathbf{V}\mathbf{W}$, kde $\mathbf{W} \in \mathbb{R}^{F\times F}$ je ortogonální matice, a dosadíme do (12), získáme

$$\mathbf{A}\mathbf{B}^\top = \mathbf{U}\mathbf{D}\mathbf{W}\mathbf{W}^{-1}\mathbf{V}^\top = \mathbf{U}\mathbf{D}\mathbf{V}^\top,$$

přičemž jsme využili vlastnosti ortogonální matice, tj. $\mathbf{W}^\top = \mathbf{W}^{-1}$ a $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$. Po dosazení dostáváme stejné proložení dat jako při předchozí volbě parametrů. Jinými slovy, změníme-li parametry, změní se i podoba hlavních komponent,

ale aproximace původních dat zůstává stejná. Proto model PCA může být rotován jakoukoliv ortogonální maticí typu $F \times F$ a v takovém případě mluvíme o nejednoznačnosti řešení. Problémy s jednoznačností řešení ve smyslu rotace při použití PCA daly podmět ke vzniku metod, které komplikaci s nejednoznačností eliminují. Jedna z takových metod je právě metoda PARAFAC.

2 Základní algebraické pojmy a značení

Než přistoupíme k představení metody PARAFAC, je potřeba se seznámit se základní terminologií [10, 3], která je užívána v souvislosti s touto metodou. Jedná se zejména o pojmy jako mnohorozměrné pole (tenzor) a jeho transformace do matice. Dále si zadefinujeme termíny jako dekompozice mnohorozměrného pole či Khatri-Raoův součin matic. U některých pojmů si pro názornost uvedeme i příklady. V kapitole 2.3 si řekneme, jakým způsobem se určuje hodnost mnohorozměrného pole. Ke každému pojmu bude vždy uvedena jeho souvislost s metodou PARAFAC.

2.1 Mnohorozměrné pole

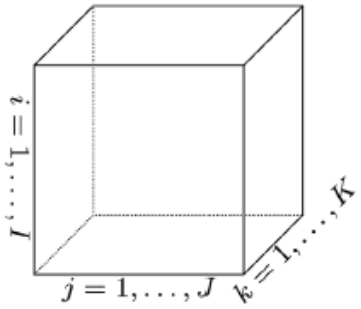
Jelikož je metoda PARAFAC určena pro analýzu vícedimenzionálních dat, které tvoří mnohorozměrné pole, definujme si tento základní pojem.

Definice 2.1. *Jakýkoliv datový soubor, jehož prvky lze uspořádat do datového pole jako*

$$x_{ijk\ldots}, \qquad i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K,\ \ldots,$$

kde počet indexů značí počet rozměrů datového souboru, se nazývá ***mnohorozměrné pole****.*

V některých literaturách jako [10] bývá mnohorozměrné pole označováno též jako tenzor. Tento termín je mnohem přirozenější, proto jej budeme používat i v následujícím textu. Jednorozměrný tenzor je vektor, dvourozměrný tenzor je matice a pro vyšší rozměry se používá pojem n-rozměrný tenzor. Obrázek 5 znázorňuje třírozměrný tenzor obsahující tři indexy (vyplývá z definice 2.1). V souvislosti s n-rozměrnými tenzory se používá pojem mód. Matice má dva módy, řádkový a sloupcový, proto třírozměrný tenzor bude mít tří módy. Dimenze označuje počet úrovní v jednom módu, čili u matice je to počet řádků, resp. počet sloupců.



Obrázek 3: Grafické znázornění třírozměrného datového souboru [10].

Značení:

$\mathbf{x} = \{x_i\}$ ... vektor

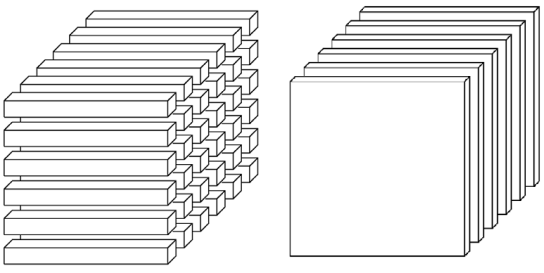
$\mathbf{X} = \{x_{ij}\}$ ... matice

$\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_n}$ ... n-rozměrný tenzor

$\underline{\mathbf{X}} = \mathcal{X} \in \mathbb{R}^{I \times J \times K} = \{x_{ijk}\}$ ... třírozměrný tenzor

a ... skalár

Fixováním indexů tenzoru získáváme jeho podmnožiny. Použití dvojtečky říká, že vybereme všechny prvky z jednoho módu. Chceme-li vyjádřit j-tý sloupec matice $\mathbf{X}$, označíme ho $\mathbf{x}_{(:j)}$, nebo jednoduše $\mathbf{x}_j$. Vektory třírozměrného pole $\underline{\mathbf{X}}$ nazýváme řádky, sloupci a tuby (viz obrázek 4a). *Vrsty* představují dvourozměrné podmnožiny tenzoru, definují se fixováním dvou indexů. Přitom u třírozměrného tenzoru rozlišujeme horizontální, vertikální a přední vrstvu (viz obrázek 4b). Například k-tou přední vrstvu značíme $\mathbf{X}_{(::k)}$ nebo jednoduše $\mathbf{X}_k$.



(a) (b)

Obrázek 4: (a) řádky: $x_{(i:k)}$, (b) přední vrstvy: $\mathbf{X}_{(::k)}$ [10].

V mnohorozměrných metodách se často používá transformace tenzoru na matici. Tento proces je velice jednoduchý, jak pochopíme z následující definice i příkladu.

Definice 2.2. ***Transformace*** *tenzoru je operace, při které dochází k přeuspořádání prvků tenzoru do matice.*

Princip transformace si ukážeme na příkladě. Mějme třírozměrný tenzor $\underline{\mathbf{X}}^{3\times3\times2}$ s předními vrstvami

$$\mathbf{X}_1 = \begin{pmatrix} 2 & 5 & 8 \\ 3 & 6 & 9 \\ 4 & 7 & 10 \end{pmatrix}, \qquad \mathbf{X}_2 = \begin{pmatrix} 11 & 12 & 13 \\ 14 & 15 & 16 \\ 17 & 18 & 19 \end{pmatrix},$$

pak jedny z možných transformací jsou

$$\mathbf{X}^{3\times6} = \begin{pmatrix} 2 & 5 & 8 & 11 & 12 & 13 \\ 3 & 6 & 9 & 14 & 15 & 16 \\ 4 & 7 & 10 & 17 & 18 & 19 \end{pmatrix},$$

$$\mathbf{X}^{3\times6} = \begin{pmatrix} 2 & 11 & 3 & 14 & 4 & 17 \\ 5 & 12 & 6 & 15 & 7 & 18 \\ 8 & 13 & 9 & 16 & 10 & 19 \end{pmatrix},$$

$$\mathbf{X}^{2\times9} = \begin{pmatrix} 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 11 & 14 & 17 & 12 & 15 & 18 & 13 & 16 & 19 \end{pmatrix}.$$

Důvody přeskupování jsou spíše praktické, v takovémto uspořádání je mnohem snazší zpracovat data pomocí softwaru. V praktické části této práce budeme používat první případ transformace.

2.2 Khatri-Raoův součin

V případě PCA jsme pracovali se zdrojovou datovou maticí a vystačili jsme si se singulárním či spektrálním rozkladem, ve kterých se používá standardní součin dvou matic. Pomocí těchto operací jsme určili hodnoty parametrů (komponentní skóry a zátěže). Díky parametrům jsme následně sestavili model, který

nám umožnil pochopit zkoumaný problém. U metody PARAFAC je situace poněkud jiná. K tomu, abychom určili parametry, používáme dekompozici tenzoru (viz kapitola 2.3) pracující s Khatri-Raovým součinem matic [3, 10]. V kapitole si uvedeme i jiné druhy maticových součinů, které je nutné znát v souvislosti s metodou PARAFAC.

Definice 2.3. ***Kroneckerův součin*** *dvou matic* $\mathbf{A} \in \mathbb{R}^{I\times J}$ *a* $\mathbf{B} \in \mathbb{R}^{K\times L}$ *je značen jako* $\mathbf{A} \otimes \mathbf{B}$. *Výsledkem tohoto součinu je matice o rozměru* $IK \times JL$, *která je definována*

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1J}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2J}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I1}\mathbf{B} & a_{I2}\mathbf{B} & \dots & a_{IJ}\mathbf{B} \end{pmatrix} = \tag{13}$$
$$= (\mathbf{a}_1 \otimes \mathbf{b}_1; \mathbf{a}_1 \otimes \mathbf{b}_2; \mathbf{a}_1 \otimes \mathbf{b}_3; \dots; \mathbf{a}_J \otimes \mathbf{b}_L),$$

kde $\mathbf{a}_j, j = 1, \dots, J$ *a* $\mathbf{b}_l, l = 1, \dots, L$ *značí postupně sloupce matice* $\mathbf{A}$ *a* $\mathbf{B}$.

Následující příklad osvětlí princip této operace. Uvažujme tedy matice

$$\mathbf{A} = \begin{pmatrix} 3 & 5 \\ 2 & 1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 8 & 4 & 7 \\ 6 & 9 & 0 \end{pmatrix},$$

potom Kroneckerův součin matice $\mathbf{A}$ a $\mathbf{B}$ je

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} 3\mathbf{B} & 5\mathbf{B} \\ 2\mathbf{B} & 1\mathbf{B} \end{pmatrix} = \begin{pmatrix} 3\cdot 8 & 3\cdot 4 & 3\cdot 7 & 5\cdot 8 & 5\cdot 4 & 5\cdot 7 \\ 3\cdot 6 & 3\cdot 9 & 3\cdot 0 & 5\cdot 6 & 5\cdot 9 & 5\cdot 0 \\ 2\cdot 8 & 2\cdot 4 & 2\cdot 7 & 1\cdot 8 & 1\cdot 4 & 1\cdot 7 \\ 2\cdot 6 & 2\cdot 9 & 2\cdot 0 & 1\cdot 6 & 1\cdot 9 & 1\cdot 0 \end{pmatrix} =$$
$$= \begin{pmatrix} 24 & 12 & 21 & 40 & 20 & 35 \\ 18 & 27 & 0 & 30 & 45 & 0 \\ 16 & 8 & 14 & 8 & 4 & 7 \\ 12 & 18 & 0 & 6 & 9 & 0 \end{pmatrix}.$$

Definice 2.4. ***Khatri-Raoův součin*** *dvou matic* $\mathbf{A} \in \mathbb{R}^{I\times J}$ *a* $\mathbf{B} \in \mathbb{R}^{K\times J}$ *je značen jako* $\mathbf{A} \odot \mathbf{B}$. *Výsledkem je matice o rozměru* $IK \times J$, *která je definována jako*

$$\mathbf{A} \odot \mathbf{B} = (\mathbf{a}_1 \otimes \mathbf{b}_1; \mathbf{a}_2 \otimes \mathbf{b}_2; \mathbf{a}_3 \otimes \mathbf{b}_3; \dots; \mathbf{a}_J \otimes \mathbf{b}_J) \tag{14}$$

Lépe tento součin pochopíme na příkladě, ve kterém uvažujeme následující dvě matice (podmínkou je pouze shodnost druhého rozměru obou matic),

$$\mathbf{A} = \begin{pmatrix} 3 & 5 \\ 2 & 1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 8 & 4 \\ 6 & 9 \\ 7 & 0 \end{pmatrix},$$

pak Khatri-Raoův součin výše uvedených matice je dán jako

$$\mathbf{A} \odot \mathbf{B} = \begin{pmatrix} 3\mathbf{b}_1 & 5\mathbf{b}_2 \\ 2\mathbf{b}_1 & 1\mathbf{b}_2 \end{pmatrix} = \begin{pmatrix} 3 \cdot 8 & 5 \cdot 4 \\ 3 \cdot 6 & 5 \cdot 9 \\ 3 \cdot 7 & 5 \cdot 0 \\ 2 \cdot 8 & 1 \cdot 4 \\ 2 \cdot 6 & 1 \cdot 9 \\ 2 \cdot 7 & 1 \cdot 0 \end{pmatrix} = \begin{pmatrix} 24 & 20 \\ 18 & 45 \\ 21 & 0 \\ 16 & 4 \\ 12 & 9 \\ 14 & 0 \end{pmatrix}.$$

Vlastnosti těchto operací jsou uvedeny v [10]. Zde si uvedeme pouze jednu z nich, kterou použijeme v algoritmu umožnující určit parametry modelu PARAFAC (kapitola 3.4). Pro Khatri-Raoův součin tří matic platí

$$\mathbf{A} \odot \mathbf{B} \odot \mathbf{C} = (\mathbf{A} \odot \mathbf{B}) \odot \mathbf{C} = \mathbf{A} \odot (\mathbf{B} \odot \mathbf{C}). \tag{15}$$

Definice 2.5. *Nechť* $\mathbf{A} = (a_{ij}), \mathbf{B} = (b_{ij}) \in \mathbb{R}^{I \times J}$ *jsou matice stejné dimenze, potom* ***Hadamardův součin*** *těchto dvou matice je definován jako*

$$\mathbf{A} \bullet \mathbf{B} = (a_{ij} b_{ij}), \tag{16}$$

pro $i = 1, \ldots, I,\ j = 1, \ldots, J,$ *a kde výsledná matice má opět rozměr* $I \times J$.

Opět ilustrujeme princip Hadamardova součinu na příkladě. Mějme dvě matice stejného typu

$$\mathbf{A} = \begin{pmatrix} 3 & 5 \\ 2 & 1 \\ 4 & 6 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 8 & 4 \\ 6 & 9 \\ 2 & 3 \end{pmatrix},$$

potom

$$\mathbf{A} \bullet \mathbf{B} = \begin{pmatrix} 3 \cdot 8 & 5 \cdot 4 \\ 2 \cdot 6 & 1 \cdot 9 \\ 4 \cdot 2 & 6 \cdot 3 \end{pmatrix} = \begin{pmatrix} 24 & 20 \\ 12 & 9 \\ 12 & 18 \end{pmatrix}.$$

Na základě Hadamardova součinu jsme v modelu PARAFAC schopni spočítat jeden parametr na základě ostatních (viz kapitola 3.4).

2.3 Dekompozice a hodnost mnohorozměrného pole

Dekompozice tenzoru [11, 10] je obdobou singulárního rozkladu pro dvoudimenzionální data (viz (12)). Oba přístupy jsou založené na stejné myšlence. Chceme vyjádřit tenzor jako součet tenzorů hodnosti jedna. Dekompozice třírozměrného tenzoru $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ tak bude

$$\mathcal{X} = (\mathbf{a}_1 \circ \mathbf{b}_1 \circ \mathbf{c}_1) + (\mathbf{a}_2 \circ \mathbf{b}_2 \circ \mathbf{c}_2) + \cdots + (\mathbf{a}_F \circ \mathbf{b}_F \circ \mathbf{c}_F) = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f, \quad (17)$$

kde F je počet použitých komponent (faktorů) a $\circ$ značí tenzorový součin. V případě metody PARAFAC se komponenty skládají z vektoru skórů a dvou vektorů zátěží, nicméně tyto termíny nejsou běžné. V praxi se pro všechny parametry používá spíše jen pojem zátěže. Na rozdíl od metody PCA, kde se komponenty skládají z jednoho vektoru skórů a jednoho vektoru zátěží. V některých aplikacích se můžeme setkat i s dekompoziční metodou nazývanou Tucker3 [11], která je obecnější než dekompoziční metoda PARAFAC.

Definice hodnosti tenzoru, $r(\mathcal{X})$, je analogická k definici hodnosti matice. Ale kvůli odlišným vlastnostem matice a tenzoru je stanovení hodnosti tenzoru poněkud složitější. Jelikož určení hodnosti tenzoru [11, 10] vychází z tenzoru hodnosti jedna, je potřeba si tento pojem zadefinovat.

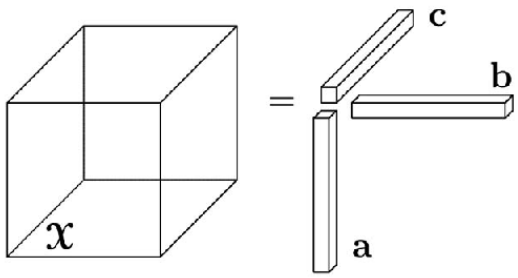
Definice 2.6. *N-rozměrný tenzor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_n}$ je hodnosti jedna, jestliže může být přepsán jako tenzorový součin n vektorů, tj.*

$$\mathcal{X} = \mathbf{a}_1 \circ \mathbf{a}_2 \circ \cdots \circ \mathbf{a}_n. \quad (18)$$

Obrázek 5 znázorňuje třírozměrný tenzor hodnosti jedna. Vidíme, že každý prvek $\mathcal{X}$ je součinem prvků vektorů $\mathbf{a}, \mathbf{b}, \mathbf{c}$.

Nyní můžeme přistoupit k samotné definici hodnosti tenzoru.

Definice 2.7. *Hodnost tenzoru $\mathcal{X}$, značíme $r(\mathcal{X})$, je definována jako nejmenší počet tenzorů hodnosti jedna, které v jejich součtu vytvoří $\mathcal{X}$.*



Obrázek 5: Třírozměrný tenzor hodnosti jedna [10].

Jinými slovy jde o nejmenší počet komponent (faktorů) nutných k popisu mnohorozměrného pole a určených metodou PARAFAC. Tato definice odpovídá vztahu (17). Stanovení hodnosti tenzoru odpovídá stanovení počtu použitých faktorů v modelu, jež je důležité pro získání jednoznačného řešení (viz kapitola 3.3). Další zvláštnost hodnosti tenzoru se týká maximální hodnosti a typické hodnosti. Zatímco maximální hodnost je definována jako největší dosažitelná hodnost, typická hodnost je jakákoliv možná hodnost, ke které může dojít s pravděpodobností větší než nula. Pro více informací může čtenář nahlédnout do [10, 11, 3], kde jsou mimo jiné uvedeny i hodnosti různě rozměrných tenzorů, např. tenzor velikosti $2 \times 2 \times 2$ je hodnosti dva nebo tři, hodnost tenzoru velikosti $9 \times 9 \times 9$ se pohybuje mezi 18 a 23.

3 Metoda PARAFAC

Metoda PARAFAC je určena pro dekompozici vícedimenzionálního pole (tenzoru) a tudíž může být chápána jako zobecnění PCA. Poprvé ji nezávisle na sobě představili Harshman v roce 1970 jako PARAFAC a dvojice Carrol a Chang v témže roce jako CANDECOMP (z anglického názvu CANonical DECOMPosition). Metoda má své původní uplatnění v psychologii, ale díky jejím vlastnostem [4, 3, 11] se využití rozšířilo do dalších oborů (přírodovědných či technických). Při měření znaků zkoumaných na objektech může dojít k tomu, že získaná data nemají dvourozměrnou strukturu, ale vícerozměrnou. Pro jednoduchost se budeme zabývat třírozměrnou strukturou dat. Například můžeme zkoumat výsledky vyšetření u pacientů v průběhu několika let nebo u sportovců můžeme zkoumat základní životní funkce při různých zátěžích. V kapitole si uvedeme výhody i nevýhody této metody. Mezi výhody patří jednak komplexní pohled na data, jednoznačnost řešení, malý počet latentních proměnných k popisu modelu. Oproti tomuto stojí početní i časová náročnost, která je způsobena použitím algoritmu založeném na ALS (z anglického názvu Alternating Least Squares), jež vyžaduje vysoký počet iterací. Jednou z dalších problematik je stanovení počtů faktorů, které popisují data a zaručují jednoznačnost výsledného modelu.

3.1 Model

Nyní si představíme strukturu modelu PARAFAC [3, 4], který zobecňuje model PCA. Obě metody jsou založené na dekompozici původního datového souboru k získání matic parametrů. Struktura modelu PCA je bilineární,

$$x_{ij} = \sum_{f=1}^{F} a_{if} b_{jf} + e_{ij}, \tag{19}$$

kde $i = 1, \ldots, I,\ j = 1, \ldots, J$, a vektorově

$$\mathbf{X} = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f^{\top} + \mathbf{E}, \tag{20}$$

přičemž a_{if} jsou prvky matice skórů, b_{jf} jsou prvky matice zátěží a e_{ij} značí rezidua. Oproti tomu model PARAFAC generuje trojici parametrů a každá komponenta se skládá z jednoho vektoru skórů a dvou vektorů zátěží

$$\mathbf{x}_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk}, \tag{21}$$

kde $i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K$. Interpretace jednotlivých parametrů je podobná jako u PCA. Hodnota parametru a_{if} koresponduje s mírou přispění i-tého objektu do f-té komponenty. Vliv j-té proměnné na f-tou komponentu představuje b_{jf} a c_{kf} ukazuje velikost přispění k-té okolnosti (fyzická zátěž nebo období) v f-té komponentě. Použitím Kroneckerova součinu obdržíme vektorovou podobu

$$\mathbf{X}^{I \times JK} = \sum_{f=1}^{F} \mathbf{a}_f (\mathbf{c}_f \otimes \mathbf{b}_f) + \mathbf{E}, \tag{22}$$

kde $\mathbf{X}^{I \times JK}$ je matice, která vznikla transformací tenzoru $\underline{\mathbf{X}}$ (viz kapitola 2.1) a $\mathbf{E}$ je matice reziduí odpovídajících rozměrů. Model můžeme také přepsat pomocí tenzorového součinu

$$\hat{\underline{\mathbf{X}}} = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f, \tag{23}$$

nebo ho vyjádřit i v maticové podobě po jednotlivých vrstvách jako

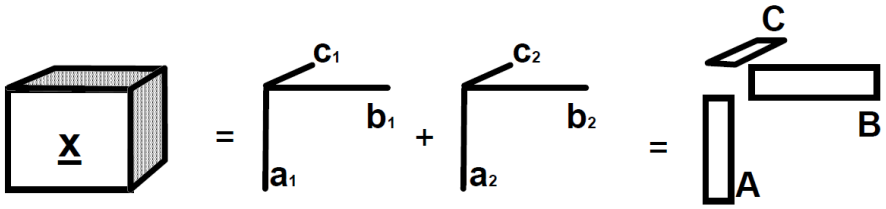
$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^\top, \qquad k = 1, \ldots, K, \tag{24}$$

kde $\mathbf{D}_k$ je diagonální matice s prvky, které odpovídají k-tému řádku matice $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_F)$, tento vztah si více rozebereme v následující kapitole. Častější zápis modelu je ovšem

$$\mathbf{X}^{I \times JK} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^\top + \mathbf{E}^{I \times JK}, \tag{25}$$

ze kterého máme jasnou představu o všech třech maticích parametrů.

Stejně jako u PCA jsou tedy vektory parametrů uspořádány do matic, tj. $\mathbf{A}, \mathbf{B}, \mathbf{C}$, které se souhrnně nazývají matice zátěží, a $\mathbf{E}$ je matice reziduí. Obrázek 6 ilustruje dvoufaktorový model PARAFAC, pro jednoduchost je vynechána



Obrázek 6: Model PARAFAC se dvěmi komponentami [10].

matice reziduí. Tenzorový součin vektorů $\mathbf{a}_1, \mathbf{b}_1$ a $\mathbf{c}_1$ určuje první komponentu (faktor), obdobně je tvořena i druhá komponenta.

K sestavení modelu bude zapotřebí odhadnout parametry, které vypočítáme pomocí optimalizační úlohy založené na ALS. Stejně jako u PCA je i zde cílem minimalizovat reziduální součet čtverců (více kapitola 3.4).

3.2 Paralelní proporcionální profily

PARAFAC je úzce provázán s principem paralelních proporcionálních profilů [3]. Princip spočívá v tom, že vektory zátěží (profily) popisující variabilitu tenzoru s odlišnými váhami, vedou k modelu nemajícího problém s rotací. Díky tomuto docílíme jednoznačné dekompozice. Uvažujme matici $\mathbf{X}$, která je popsána pomocí modelu PCA, tj. $\mathbf{X} = \mathbf{A}\mathbf{B}^\top$, se dvěmi komponentami. Model můžeme zapsat jako

$$\mathbf{X}_1 = \mathbf{a}_1\mathbf{b}_1^\top c_{11} + \mathbf{a}_2\mathbf{b}_2^\top c_{12}, \tag{26}$$

kde c_{11} a c_{12} jsou váhy, obě rovny jedné. Mějme jinou matici, která je popsána stejnými vektory skórů i zátěží, ale s odlišnými váhami, jejichž hodnoty jsou různé od těch předchozích

$$\mathbf{X}_2 = \mathbf{a}_1\mathbf{b}_1^\top c_{21} + \mathbf{a}_2\mathbf{b}_2^\top c_{22}. \tag{27}$$

Obě matice popisují jiné modely, ale skládají se ze stejných paralelních profilů (zátěží) s odlišnými váhami. Při sloučení obou modelů dostáváme

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^\top, \qquad k = 1, 2, \tag{28}$$

kde $\mathbf{D}_k$ je diagonální matice s prvky odpovídající vahám c_{kf}. Formulace (26) a (27) tak můžeme pro data s třírozměrnou strukturou vyjářit jako

$$\mathbf{X} = (\mathbf{X}_1; \mathbf{X}_2) = (\mathbf{a}_1; \mathbf{a}_2) \begin{pmatrix} \mathbf{b}_1 c_{11} & \mathbf{b}_2 c_{12} \\ \mathbf{b}_1 c_{21} & \mathbf{b}_2 c_{22} \end{pmatrix}^\top = (\mathbf{a}_1; \mathbf{a}_2)(\mathbf{b}_1 \otimes \mathbf{c}_1; \mathbf{b}_2 \otimes \mathbf{c}_2), \tag{29}$$

kde $\mathbf{c}_1 = (c_{11}; c_{21})^\top$ a $\mathbf{c}_1 = (c_{12}; c_{22})^\top$. Zobecníme-li model (29) s počtem komponent rovnu F a přepíšeme-li formulaci pomocí Khatri-Raova součinu, dostáváme

$$\mathbf{X}^{I \times JK} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^\top, \tag{30}$$

jež odpovídá třírozměrnému modelu PARAFAC s maticemi parametrů $\mathbf{A}, \mathbf{B}$ a $\mathbf{C}$ (viz (25)). Z toho plyne, že metoda PARAFAC určuje model, který nelze rotovat na rozdíl od metody PCA. Proto výsledné řešení je jednoznačné právě tehdy, když nalezneme správný počet komponent (viz kapitola 3.5.4).

3.3 Jednoznačnost

Jelikož je metoda PARAFAC známá především díky své jednoznačnosti [10, 11], věnujeme tomuto pojmu samostatnou kapitolu. Tato vlastnost spolu se snadnou interpretovatelností výsledného modelu staví tuto analýzu do popředí všech dekompozičních metod. V PCA musela být zavedena podmínka ortogonality, aby bylo dosaženo rozumného řešení, ale ani to nám nezaručuje jednoznačnost získaného modelu. Každý bilineární model může být rotován jakoukoliv ortogonální maticí typu $F \times F$, kde F značí počet komponent zahrnutých v modelu. Výhoda metody PARAFAC spočívá v nalezení jednoznačné dekompozice při správně stanoveném počtu komponent [3, 4].

Připomeňme si dekompozici třírozměrného tenzoru $\underline{\mathbf{X}}$ (23) metodou PARAFAC, je dána jako

$$\hat{\underline{\mathbf{X}}} = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f,$$

kde $\mathbf{a}_f$ je f-tý sloupec matice zátěží $\mathbf{A}$ a obdobně pro $\mathbf{b}_f$ i $\mathbf{c}_f$. Jednoznačnost znamená, že výše uvedená formulace je jedinou možnou kombinací tenzorů hod-

nosti jedna, která v součtu tvoří $\hat{\underline{\mathbf{X}}}$. Pokud bychom odhadnuté parametry rotovali, dostaneme úplně jiný model, na rozdíl od rotace parametrů u metody PCA (viz kapitola 1.3). Uvažujme maticové vyjádření podle (28), tj. $\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^\top$, rotaci modelu lze vyjádřit jako

$$\mathbf{A}\mathbf{D}_k\mathbf{B}^\top = \mathbf{A}\mathbf{T}\mathbf{T}^{-1}\mathbf{D}_k\mathbf{S}\mathbf{S}^{-1}\mathbf{B}^\top, \tag{31}$$

kde $\mathbf{AT}$ a $\mathbf{B}(\mathbf{S}^{-1})^\top$ jsou rotované matice zátěží. Ze vztahu je zřejmé, že matice zátěží $\mathbf{T}^{-1}\mathbf{D}_k\mathbf{S}$ musí být diagonální, kvůli tomuto požadavku ovšem existuje jenom málo matic $\mathbf{T}, \mathbf{S}$, které splňují tuto nutnou podmínku [6].

Harshman (1972) dokázal, že jednoznačné řešení může nastat, jsou-li vektory zátěží lineárně nezávislé ve dvou módech a ve třetím naopak lineárně závislé. Jiný přístup nabízí podmínku na počet faktorů, jejímž autorem je Kruskal (1989), a která vychází z plné hodnosti matic zátěží. Matematicky můžeme podmínku vyjádřit pomocí následující nerovnosti

$$k_A + k_B + k_C \geq 2F + 2,$$

kde k_A značí hodnost matice zátěží $\mathbf{A}$, obdobně k_B je hodnost matice $\mathbf{B}$ a k_C hodnost matice $\mathbf{C}$. Jak již bylo zmíněno F představuje možný počet komponent použitých v modelu. V některých případech se jím můžeme řídit. Uvedená pravidla nemůžeme ovšem zobecnit pro všechny případy, ale lze je považovat za přijatelná pro dosažení jednoznačného modelu.

Jestliže jsou původní data ze své podstaty trilineární, pak aplikací PARAFAC metody získáme základní informace o zkoumaných jevech. Jinými slovy získáme jednoznačnou dekompozici, čímž se dostáváme k myšlence proporcionálních profilů (viz kapitola 3.2). Zdůrazněme znovu, že jednoznačné dekompozice docílíme při správně zvoleném počtu komponent (viz kapitola 3.5.4).

3.4 Algoritmus ALS

V této kapitole si ukážeme postup, který umožní naleznout hledaný model v datech. K vyrovnání dat metodou PARAFAC můžeme použít algoritmus zalo-

žený na ALS (Alternating Least Squares) [3, 4]. Jako každý algoritmus v optimalizačních úlohách má za úkol vylepšit stávající model, tak i ALS algoritmus se snaží v každém iteračním kroku zlepšit vyrovnání modelu prostřednictvím odhadnutých parametrů. Nejprve se seznámíme s principy uvedeného algoritmu.

Koncept ALS je založen na rozdělení datového souboru do několika množin parametrů. Odhady parametrů v jednotlivých skupinách vychází z podmínky minimalizace čtvercové chyby, tj. užitím principu metody nejmenších čtverců, při zachování hodnot ostatních parametrů. Na začátku odhadujeme jednu skupinu parametrů za předpokladu, že ostatní známe. Při odhadu používáme minimalizační funkci, kterou si představíme níže. V druhém kroku odhadujeme další skupinu parametrů pomocí stejné minimalizační funkce a opět předpokládáme, že ostatní parametry známe, mezi kterými jsou ovšem již ty odhadnuté z prvního kroku. Postup opakujeme do vyrovnání modelu nebo do doby, kdy jsou změny v odhadnutých parametrech minimální.

Myšlenka algoritmu je založena na rozdělení hlavního úkolu (odhadnout všechny parametry zároveň) do dílčích, které si kladou za cíl odhadovat jednotlivé skupiny parametrů postupně.

Uvažujme nejprve pro jednoduchost bilineární model $\hat{\mathbf{X}} = \mathbf{A}\mathbf{B}^\top$, kde $\mathbf{A}$, $\mathbf{B}$ jsou parametry, které odhadneme s využitím minimalizační funkce

$$\min_{\mathbf{A},\mathbf{B}} ||\mathbf{X} - \mathbf{A}\mathbf{B}^\top||_F^2.$$

Postupně budeme jednotlivé odhady parametrů vylepšovat, a to tak, že nejprve odhadneme matici $\mathbf{B}$ jako

$$\mathbf{B} = \mathbf{X}^\top(\mathbf{A}^+)^\top.^{2} \tag{32}$$

Tu použijeme k lepšímu odhadu matice $\mathbf{A}$ dosazením do

$$\mathbf{A} = \mathbf{X}(\mathbf{B}^+)^\top, \tag{33}$$

[2]Matice $\mathbf{A}^+$ představuje Moore-Penroseovu inverzi [7] matice $\mathbf{A}$. Tato operace je zobecněním standardní inverze na obdelníkové matice typu $n \times m$ a platí pro ni vztahy $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$, $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+, (\mathbf{A}\mathbf{A}^+)^\top = \mathbf{A}\mathbf{A}^+, (\mathbf{A}^+\mathbf{A})^\top = \mathbf{A}^+\mathbf{A}$.

a následně opět použijeme ke zlepšení odhadu $\mathbf{B}$. Postupným dosazováním odhadnutých parametrů do příslušných vztahů (32) a (33) zjistíme jejich hodnoty. Iterativně pak odhadujeme jednotlivé parametry, dokud algoritmus nezačne konvergovat kolem jejich správných hodnot. Algoritmus vylepší stávající vyrovnání a nebo ho ponechá stejné, z toho důvodu rozdíl mezi hodnotou minimalizační funkce z předchozího a aktuálního kroku nikdy nenabyde záporné hodnoty. Při tomto postupu ovšem může dojít k tomu, že algoritmus bude konvergovat v případě optimalizační funkce ke svému lokálnímu minimu, nikoli ke globálnímu, které je požadováno. Proto výsledek závisí jak na typu dat, tak i na stanovení počátečních hodnot, vstupujících do algoritmu (více v kapitole 3.5).

Nyní se vrátíme k použití ALS v metodě PARAFAC, abychom odhadli parametry modelu (25), tj. matice $\mathbf{A}, \mathbf{B}, \mathbf{C}$. Poznamenejme přitom, že pro zjednodušení nebudeme pro odhady zavádět speciální omezení. Odpovídající minimalizační funkce je

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} ||\mathbf{X} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^\top||_F^2, \tag{34}$$

kterou upravíme substitucí výrazu $\mathbf{Z} = \mathbf{C} \odot \mathbf{B}$ na optimalizaci

$$\min_{\mathbf{A}} ||\mathbf{X} - \mathbf{A}\mathbf{Z}^\top||_F^2; \tag{35}$$

potom odhad $\mathbf{A}$ bude podle (33) nalezen jako

$$\mathbf{A} = \mathbf{X}(\mathbf{Z}^+)^\top = \mathbf{X}\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}, \tag{36}$$

obdobně se postupuje při odhadování $\mathbf{B}$ a $\mathbf{C}$, díky speciální struktuře modelu. Pro zajímavost si ukážeme níže, že $\mathbf{XZ}$ a $\mathbf{Z}^\top\mathbf{Z}$ lze počítat přímo z matic $\mathbf{B}$ a $\mathbf{C}$, protože

$$\mathbf{XZ} = \mathbf{X}_1\mathbf{B}\mathbf{D}_1 + \mathbf{X}_2\mathbf{B}\mathbf{D}_2 + \cdots + \mathbf{X}_k\mathbf{B}\mathbf{D}_k \tag{37}$$

a

$$\mathbf{Z}^\top\mathbf{Z} = (\mathbf{B}^\top\mathbf{B}) \bullet (\mathbf{C}^\top\mathbf{C}), \tag{38}$$

kde symbol $\bullet$ značí Hadamardův součin (viz kapitola 2.2). Shrňme si tedy dosavadní poznatky o algoritmu ALS. Na začátku nesmíme zapomenout, že se již

nepohybujeme ve dvoudimenzionálním prostoru, a proto je potřeba transformovat původní třírozměrný tenzor na matici. Volba transformace $\underline{\mathbf{X}}$ bude záležet na tom, jaký parametr se bude v danou chvíli počítat. Odhadované matice parametrů jsou pro trilineární dekompozici $\mathbf{A} \in \mathbb{R}^{I \times F}, \mathbf{B} \in \mathbb{R}^{J \times F}, \mathbf{C} \in \mathbb{R}^{K \times F}$.

Obecně lze algoritmus zapsat následovně:

0. stanovení počtu komponent F (viz kapitola 3.5.4).

1. stanovení počátečních hodnot $\mathbf{B}$ a $\mathbf{C}$ a odhadnutí matice $\mathbf{A}$:

$$\mathbf{Z} = \mathbf{C} \odot \mathbf{B}, \qquad \mathbf{A} = \mathbf{X}\mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}, \tag{39}$$

z výše uvedených rozměrů matic parametrů plyne, že matice $\mathbf{Z} \in \mathbb{R}^{F \times JK}$, a proto musíme tenzor transformovat na matici $\mathbf{X} \in \mathbb{R}^{I \times JK}$.

2. odhadnutí matice $\mathbf{B}$:

$$\mathbf{Z} = \mathbf{C} \odot \mathbf{A}, \qquad \mathbf{B} = \mathbf{X}\mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}; \tag{40}$$

obsahuje odhad $\mathbf{A}$ z předchozího kroku.

3. odhadnutí matice $\mathbf{C}$:

$$\mathbf{Z} = \mathbf{B} \odot \mathbf{A}, \qquad \mathbf{C} = \mathbf{X}\mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}, \tag{41}$$

je určeno odhadnutými maticemi $\mathbf{B}, \mathbf{A}$ z předchozích kroků.

4. opakujeme kroky 1. - 3., dokud relativní změny v odhadnutých parametrech nebudou menší než stanovená hraniční hodnota.

Jestliže ALS algoritmus konverguje ke svému globálnímu minimu, pak jsme dosáhli optimálního řešení. Obecně se ovšem v úlohách optimalizace s globální konvergencí moc často nesetkáváme, ale pouze spíše s konvergencí lokální. Z tohoto důvodu záleží na konkrétní struktuře dat a také na nastavení počátečních hodnot i počtu parametrů. Nicméně každé zlepšení stávajících odhadnutých parametrů je pro nás přínosné. Jediná nevýhoda ALS tkví v jeho časové náročnosti. V případě vysokého počtu proměnných, např. pro datové pole $50 \times 50 \times 50$, může algoritmus pracovat až hodiny. Proto je důležité jednak naprogramovat rychlý

algoritmus, zvolit vhodné počáteční hodnoty parametrů, ale mít i výkonný počítač. V praktické části této práce se ovšem setkáme s méně rozsáhlými datovými soubory, kde se s uvedenou problematikou naštěstí nesetkáme.

3.5 Praktické aspekty metody PARAFAC

Než vůbec začneme aplikovat na data metodu PARAFAC založenou na ALS algoritmu, je potřeba si uvědomit některé důležité aspekty [3, 4]. Jedná se zejména o předzpracování vstupních dat, je-li nutné použít centrování nebo jiné transformace. Pokud chceme naleznout výsledný model, musíme stanovit počáteční hodnoty vstupující do algoritmu. Všechny tyto zmíněné činnosti jsou součástí přípravné fáze. Otázkou také zůstává, kdy máme spuštěný výpočetní algoritmus zastavit, tj. jak nastavit ukončovací kritérium. A na závěr je nutné zhodnotit výsledný model, zejména posoudit zvolený počet komponent. V literatuře se v tomto smyslu často hovoří o ohodnocení modelu.

3.5.1 Předzpracování dat

Cílem předzpracování dat je použít vhodnou transformaci dat, tak abychom byli schopni nalézt vhodný model, který dostatečně popisuje zkoumaný problém. Mezi základní transformace dat patří centrování, škálování (normování) a logaritmická transformace. U každé z nich si řekneme princip a také případy, ve kterých se používají. Předzpracování vícedimenzionálních datových polí je většinou bohužel mnohem komplikovanější než je tomu u dvourozměrných matic.

Centrování má stejný význam jako ve dvourozměrných analýzách, tj. od původních hodnot se odečítá konstanta, která odpovídá nějaké vhodné charakteristice polohy (nejčastěji aritmetickému průměru). Cílem je, aby se struktura dat přiblížila struktuře modelu. Centrování prvního módu se provádí na matici typu $I \times JK$, jež vznikla transformací původního tenzoru, a je dáno jako

$$x^c_{ijk} = x_{ijk} - \overline{x_{jk}}, \qquad \overline{x_{jk}} = \frac{\sum_{i=1}^{I} x_{ijk}}{I}; \tag{42}$$

kde $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, a nazývá se též centrování přes první

mód nebo také jednoduché centrování. Tento typ transformace lze aplikovat na jakýkoliv mód v závislosti na zkoumaném problému. Existuje dvojité i trojité centrování, jehož princip spočívá v centrování přes jeden mód a následném centrování výsledných hodnot. V případě metody PARAFAC je nejvhodnější použít jednoduché centrování, které musí být prováděné přes všechny sloupce. Centrováním eliminujeme vliv $\overline{x_{jk}}$ a zachováváme relativní informaci.

Ke škálování přistupujeme v případě, že proměnné jsou měřeny v různých jednotkách. Podíváme-li se na minimalizační funkce (34) či (35), pak pokud bychom ponechali zdrojovou matici v původních jednotkách, reziduální součet čtverců by byl příliš vysoký. Škálování se provádí proto, aby data byla kompatibilní s minimalizační funkcí. Tato operace nemění strukturu dat, ale pouze velikost reziduí. Matematicky tuto transformaci lze popsat vztahem

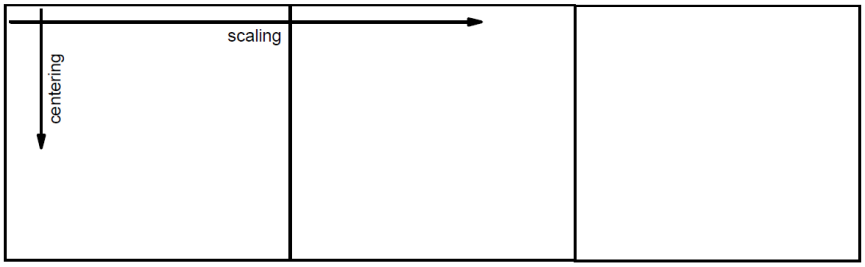
$$x^s_{ijk} = \frac{x_{ijk}}{s_i}, \qquad s_i = \sqrt{\sum_{j=1}^{J}\sum_{k=1}^{K} x^2_{ijk}}, \tag{43}$$

kde $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$. Stejně jako centrování, tak i škálování se aplikuje na transformovaný tenzor. Výše popsaný vztah odpovídá škálování prvního módu, nicméně ovlivňuje škálu též v ostatních módech. Další komplikace nastává, chceme-li škálovat i centrovat datový soubor současně. Obecně jednoduché centrování naruší škálování ve všech módech, proto se v případě metody PARAFAC doporučuje nejdříve škálovat a poté centrovat. Princip centrování a škálovaní je graficky znázorněn obrázkem 7.

Poslední možností předzpracování dat je logaritmická transformace, která se používá v případě lišících se řádů hodnot jednotlivých proměnných. Obdržíme tak hodnoty

$$x^l_{ijk} = \log\left(x_{ijk}\right), \tag{44}$$

kde $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$. Logaritmickou transformaci aplikujeme na transformovaný tenzor a to na každý prvek, jak vyplývá z výše uvedeného vztahu.



Obrázek 7: Princip centrování a škálování dat [3].

Poslední dvě transformace se provádí z důvodu lepší srovnatelnosti proměnných, v podstatě dochází ke stabilizaci rozptylu datového souboru.

3.5.2 Stanovení počátečních hodnot

Vhodné nastavení počátečních hodnot vstupujících do ALS algoritmu nám může zaručit, že algoritmus nalezne své globální minimum. Dvojice Harshman a Lundy [3] doporučují použití náhodné hodnoty a začít algoritmus z několika odlišných bodů. Takto budeme volit počáteční hodnoty i v praktické části. Jiný přístupy nabízí například [3].

3.5.3 Ukončovací kritérium

Účelem nastavení ukončovacího kritéria je, aby se ALS algoritmus v případě konvergence zastavil, až relativní změny v odhadech parametrů modelu mezi dvěma iteracemi budou pod nastavenou hranicí. Obvykle se hodnota ukončovacího kritéria volí 10^{-6}. Někdy se může stát, že i malé změny ve vstupních hodnotách jsou spojeny s velkými změnami v odhadnutých parametrech. Toto souvisí s charakterem analyzovaných dat. Abychom se ujistili, že k takovému případu nedošlo, doporučuje se algoritmus spustit dvakrát. Pokud algoritmus opravdu konverguje, obě řešení budou stejná.

3.5.4 Určení počtu komponent

V předchozích kapitolách bylo již několikrát zmíněno, že je velice důležité stanovit správný počet komponent, tj. správnou dimenzi matic zátěží, k vytvoření vhodného modelu. V případě PCA se komponenty vytváří postupně, tudíž postupně vysvětlují i variabilitu dat. Proto se občas může stát, že komponenty popisují nejen variabilitu zkoumaného systému ale i rezidua. Oproti tomu komponenty u metody PARACAF se tvoří současně a to pomocí minimalizační funkce. Použijeme-li správný počet komponent, pak tyto komponenty popisují skutečnou variabilitu dat. Pokud bychom ale k popisu modelu použili více komponent než je potřeba, velikost reziduí se bude zvyšovat a komponenty budou vzájemně korelované. Takto získaný model není samozřejmě vhodný. Proto je potřeba vždy posoudit získaný model. K tomuto účelu existuje několik nástrojů, které umožňují stanovit vhodný počet komponent. V této práci bude zmíněna „split-half analýza“ [3, 15] a diagnostika jádrové konzistence [1], pro více informací může čtenář nahlédnout do [3, 15]. Mezi další nástroje patří například bootstrap, křížová validace nebo analýza reziduí, ve které se vyšetřuje vyrovnání dat získaným modelem. Vyšetření správného počtu komponent, se provádí již na výsledném modelu. V případě špatných výsledků se jejich počet zmenší či zvětší a vytvoří se nový model, který znovu posoudíme pomocí zmíněných nástrojů.

„Split-half“ analýza

Koncept analýzy spočívá v rozdělení datového pole na dvě poloviny, na které se aplikuje metoda PARAFAC. Následně vyšetřujeme, zda se oba nalezené modely shodují. Zvolíme-li správný počet komponent, pak oba modely budou mít přibližně stejné parametry. Jestliže jsme použili špatný počet komponent, je pravděpodobné, že se modely shodovat nebudou. Doporučuje se analýzu spustit alespoň dvakrát za sebou. Je-li shoda parametrů podobná při prvním i druhém spuštění, je patrné, že nalezený model je vyhovující.

Analýza jádrové konzistence

Druhým nástrojem, který si v této kapitole popíšeme, je analýza jádrové kon-

zistence, která souvisí s dekompoziční metodou Tucker3, jež je zobecněním metody PARAFAC. Z tohoto důvodu si zde uvedeme základní strukturu modelu Tucker3, pro více informací může čtenář nahlédnout do [3], a porovnáme ji se strukturou modelu PARAFAC. Obě metody jsou určeny k získání komponent, jež jsou kombinací parametrů. Uvažujme model PARAFAC (viz kapitola 3.1) na $\underline{\mathbf{X}}$ jako

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk}, \qquad \mathbf{X} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\top}, \tag{45}$$

kde $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, a který je dán lineární kombinací parametrů, odpovídajících stejnému počtu faktorů, které přísluší jednotlivým módům. Jestliže se ukáže, že dvě komponenty popisují strukturu datového pole, z každého módu musí být určeny i dva vektorové parametry. Ve srovnání s modelem PARAFAC je model Tucker3 více flexibilnější z důvodu existence mnohorozměrného jádrového pole, jelikož umožňuje popsat model pomocí komponent, které nemusí být nutně tvořeny vektorovými parametry z každého módu. Například první komponenta může být kombinací vektorových parametrů z prvního a druhé módu a druhá komponenta může být tvořena vektorovými parametry z prvního a třetího módu. Model je dán jako

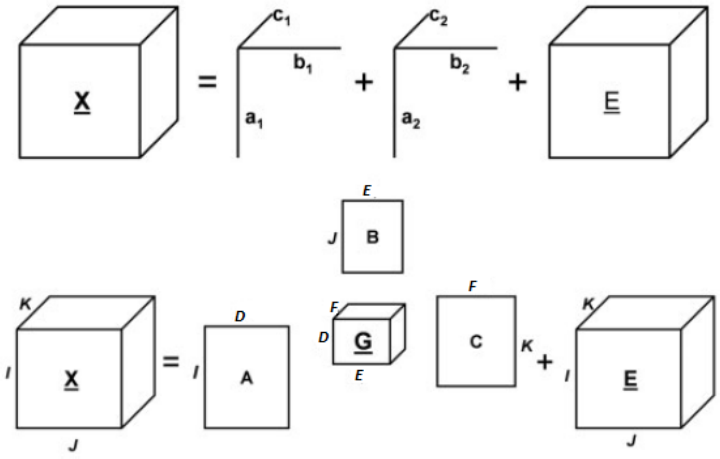
$$x_{ijk} = \sum_{d=1}^{D} \sum_{e=1}^{E} \sum_{f=1}^{F} g_{def} a_{id} b_{je} c_{kf} + e_{ijk}, \qquad \mathbf{X} = \mathbf{A}\mathbf{G}^{D \times EF}(\mathbf{C} \odot \mathbf{D})^{\top}; \tag{46}$$

značení je analogické jako u modelu PARAFAC, nicméně se nám ve vztahu objevuje navíc g_{def}, což je prvek jádrového pole $\underline{\mathbf{G}}$, transformovaného následně na matici $\mathbf{G}$. Pro srovnání můžeme oba modely graficky znázornit pomocí následujícího obrázku 8.

Model PARAFAC zapíšeme jako speciální případ modelu Tucker3 jako

$$\mathbf{X} = \mathbf{A}\mathbf{T}^{F \times FF}(\mathbf{C} \odot \mathbf{B})^{\top}, \tag{47}$$

kde $\mathbf{T}$ je transformovaná matice jádrového tenzoru obsahující nuly až na superdiagonálu, která obsahuje samé jedničky. Superdiagonálou je zde myšlena tělesová úhlopříčka třírozměrného tenzoru (krychle, viz obrázek 8). Pokud je model



Obrázek 8: Srovnání modelu PARAFAC (nahoře) s modelem Tucker3 (dole) [1].

správně navržen, pak $\underline{\mathbf{T}}$ a $\underline{\mathbf{G}}$ by měly být podobné. Stačí ověřit, zda na superdiagonále tenzoru $\underline{\mathbf{G}}$ jsou prvky blízko jedné a ostatní prvky blízko nule. V opačném případě data nejsou ze své podstaty trilineární a nebo jsme zvolili větší počet komponent. Jádrovou konzistenci lze matematicky popsat vztahem

$$\text{Jádrová konzistence} = \left(1 - \frac{\sum_{i=1}^{F}\sum_{j=1}^{F}\sum_{k=1}^{F}(g_{ijk} - t_{ijk})^2}{F^2}\right)100. \tag{48}$$

Hodnota pohybující se kolem 90% značí, že data mají trilineární strukturu a že navržený model je správně zvolený. Hodnoty blízko 50% a nebo menší indikují špatně navržený model. Jádrovou konzistenci lze zkoumat i graficky, kdy do grafu zobrazujeme prvky jádrového tenzoru a jejich hodnoty.

3.5.5 Degenerované řešení

Nyní si uvedeme některé případy degenerovaného řešení, vedoucího k nestabilitě a praktické nepoužitelnosti modelu; jiné situace jsou uvedeny v [4]. Typickým příznakem degenerovaného řešení je, když vektory zátěží stejného módu jsou silně korelované. Dalším případem je, když jsou dvě nebo žádná dvojice vektorových zátěží odpovídajících módů pozitivně korelované a současně jedna či všechny tři dvojice negativně korelované. Pro posouzení degenerovaného řešení pak slouží

ukazatel TC (z anglického Triple Cosine), který je definován pro dvojici komponent jako

$$TC_{ij} = \cos(\mathbf{a}_i, \mathbf{a}_j)\cos(\mathbf{b}_i, \mathbf{b}_j)\cos(\mathbf{c}_i, \mathbf{c}_j) = \frac{\mathbf{a}_i^\top \mathbf{a}_j}{||\mathbf{a}_i||\,||\mathbf{a}_j||}\frac{\mathbf{b}_i^\top \mathbf{b}_j}{||\mathbf{b}_i||\,||\mathbf{b}_j||}\frac{\mathbf{c}_i^\top \mathbf{c}_j}{||\mathbf{c}_i||\,||\mathbf{c}_j||}, \quad (49)$$

kde $i, j = 1, \ldots, F, i \neq j$ Hodnoty TC_{ij} pohybující se v rozmezí -1 až -0.85 poukazují na degenerované řešení.

4 Aplikace metody PARAFAC

V praktické části této práce aplikujeme metodu PARAFAC na reálné datové soubory pocházející z různých vědních oblastí. První příklad se věnuje analýze ekologických dat pocházejících z České republiky, druhý příklad zkoumá data popisující vzdělávací systém a trh práce ve vybraných státech Evropské unie na základě vybraných charakteristik a poslední model je vytvořen pro datový soubor pocházející z lékařského prostředí. U každého praktického příkladu si vždy popíšeme příslušný datový soubor, odhadneme model popisující data a tento model zhodnotíme. Stěžejním výstupem této motody jsou zátěžové grafy, stejně jako v případě PCA byl výstupem biplot. Data jsou zpracována pomocí statistického softwaru R s využitím knihovny `ThreeWay` [16]. Interpretace výsledků byla mimo jiné zpracována pomocí [14].

4.1 Příklad 1: Znečištění ovzduší na obyvatele v ČR

V tomto příkladě budeme analyzovat data získaná ze Statistické ročenky České republiky [18], jež popisují znečištění ovzduší na obyvatele během let 2002-2009 ve vybraných krajích ČR. Data jsou uspořádána do třírozměrného tenzoru, $\underline{\mathbf{X}} \in \mathbb{R}^{12\times4\times8}$ a přední vrstva má podobu následující tabulky 1.

Území	Emise tuhé	Oxid siřičitý	Oxid dusíku	Oxid uhelnatý
2002				
Jihočeský	7,7	18,3	8,3	18,8
Plzeňský	6,6	21,5	9,7	19,6
Karlovarský	5,7	56,8	26,9	18,1
Ústecký	6,2	98,4	77,1	20,0
Liberecký	4,6	11,0	6,5	15,7
Královéhradecký	5,4	13,5	5,1	17,9
Pardubický	6,3	36,2	27,5	19,2
Vysočina	6,4	8,2	5,3	18,1
Jihomoravský	2,0	3,1	4,6	5,0
Olomoucký	4,0	9,5	7,3	11,6
Zlínský	3,2	11,8	6,2	8,6
Moravskoslezský	5,9	22,8	18,8	103,0

Tabulka 1: Přední vrstva datového pole (znečištění ovzduší v ČR).

Další vrstvy pak odpovídají měření v jednotlivých letech (viz příloha A.1). Jelikož jsou data importována z textových souborů, je nezbytné nastavit cestu k pracovnímu adresáři, ve kterém jsou soubory uloženy, pomocí příkazu:

```
>setwd("D:/UPOL/Mgr/diplomka/data/Emise")
```

V úvodu jsme zmínili, že budeme pracovat s knihovnou `ThreeWay`, proto ji načteme:

```
>library(ThreeWay)
```

Nyní můžeme importovat datové tabulky a následně je převést do maticového formátu, v případě první z nich tedy zadáme:

```
> d1=as.matrix(read.table("O2.txt", header=FALSE, sep=""))
```

Takto postupujeme u načtení každého souboru, celkově pracujeme s osmi maticemi, které uspořádáme do jedné matice typu 12×32 podle definice 2.2. Jelikož se hodnoty proměnných liší jednotkami řádu, aplikujeme logaritmickou transformaci (44):

```
> data1=log(cbind(d1,d2,d3,d4,d5,d6,d7,d8))
```

Nyní máme data připravená a můžeme přistoupit k nalezení modelu pomocí ALS algoritmu. Z knihovny `ThreeWay` vybereme následující funkci a vhodně nastavíme její argumenty, především rozměr datového souboru (12, 4, 8), počet komponent (2), více v [16]:

```
> emi<- CPfuncrep(data1, 12, 4, 8, 2, 1, 1, 1, 0, 1e-6, 1000)
```

Zadáním příkazů `emi$A`, `emi$B`, `emi$C` zjistíme hodnoty parametrů:

```
> emi$A
           [,1]        [,2]
 [1,] 13.471968  0.06643899
 [2,] 13.634421  0.14901424
 [3,] 14.371087  0.67159319
 [4,] 16.034884  0.94255708
 [5,] 11.580562 -0.08500298
 [6,] 12.607683 -0.02058894
 [7,] 14.235703  0.50443471
 [8,] 12.238826 -0.18225472
 [9,]  6.917208  0.07121468
[10,] 10.740124  0.10203365
```

```
[11,]  9.566302  0.24077570
[12,] 19.279909 -0.16315026
> emi$B
           [,1]        [,2]
[1,] -0.3057162  0.06457693
[2,] -0.5352538 -1.36277831
[3,] -0.4211580 -1.55356967
[4,] -0.6300804  0.66320026
> emi$C
           [,1]      [,2]
[1,] -0.3739452 -1.163265
[2,] -0.3726193 -1.166864
[3,] -0.3723376 -1.181871
[4,] -0.3518067 -1.229719
[5,] -0.3442177 -1.287099
[6,] -0.3450902 -1.322828
[7,] -0.3333062 -1.189992
[8,] -0.3231963 -1.199020
```

Algoritmus nalezl výsledný model po 615 krocích. Tuto informaci zjistíme zadáním příkazu `emi$iter`. Ověříme procento vyrovnání dat pomocí modelu:

```
> emi$fp
[1] 99.24547
```

Vidíme, že model data vyrovnává z 99%, což je velmi dobrý výsledek. Z výsledků můžeme také zjistit minimální hodnotu TC mezi dvěma komponentami podle (49), která signalizuje degenerované řešení:

```
> emi$tripcos
Minimal triple cosine
            0.2348072
```

Hodnota se nepohybuje v rozmezí, které není žádoucí. Abychom ovšem mohli výsledný model považovat za vyhovující, aplikujeme na něj též *„split-half“* analýzu. K tomu použijeme následující funkci s vhodně zvolenými parametry, jako rozměry datového souboru, počet komponent a jiné (význam parametrů v [16]). Software nás vyzve k zadaní dvou parametrů, první volíme 1 pro nenáhodné rozdělení datového souboru a druhý volíme 0, jelikož data jsou již zlogaritmována a žádné další úpravy na nich není potřeba provádět:

```
> splitCP <- splithalfCP(data1, 12, 4, 8, 2, 0, 0, 0, 0, 1e-6, 10000, 1, 1, 1)
```

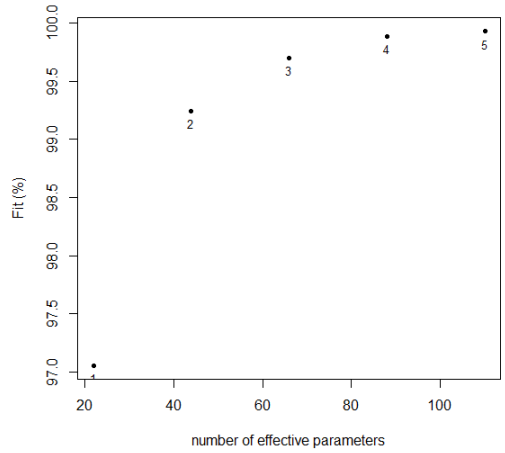
Pro hodnocení jsou stěžejní následující hodnoty:

```
Congruences for A in splits and in appropriate part of Afull
       SPL1 SPL2
Comp.1    1 1.00
Comp.2    1 0.93
Congruence values for B-mode component matrix
Comp.1 Comp.2
  1.00   0.94
Congruence values for C-mode component matrix
Comp.1 Comp.2
     1      1
```

Vidíme, že v obou souborech bylo dosaženo shody ve všech parametrech. Pokud bychom chtěli vědět, jaké je vyrovnání v případě jiného počtu komponent, zadáme následující příkaz:

```
> FitCP <- CPrunsFit(data1, 12, 4, 8, 5)
> OutCP <- FitCP[,c(1,4)]
> CPdimensionalityplot(OutCP, 12, 4, 8)
```

Výstupem této funkce je grafické zobrazení procenta vyrovnání vzhledem k počtu parametrů a počtu komponent (obrázek 9). Z grafu je zřejmé, že počet komponent



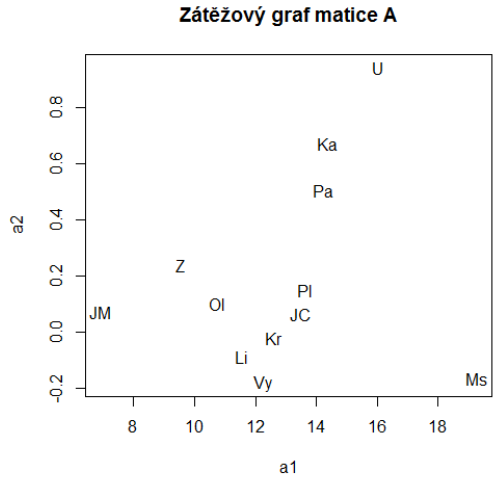
Obrázek 9: Hodnota vyrovnání modelem při různém počtu komponent (%).

byl zvolen správně. Při použití tří komponent by již nedošlo k tak razantní změně ve vyrovnání jako v případě mezi jedním a dvěmi komponenty.

Z výše uvedených výsledků lze konstatovat, že výsledný model je vyhovující a proto můžeme přistoupit ke grafické interpretaci hodnot zátěží. Pro jednoduchost si uvedeme příkaz pouze pro zobrazení hodnot matice zátěží $\mathbf{A}$, obdobně se postupuje též u zbylých dvou matic. Abychom se v grafu lépe orientovali, popíšeme si jednotlivé proměnné:

```
> ## A
> rownames(emi$A)=(paste(c("JC","Pl","Ka","U","Li","Kr","Pa","Vy","JM","Ol","Z","Ms")))
> plot(emi$A[,1],emi$A[,2],type="n",main="Zátěžový graf matice A", xlab="a1",ylab="a2")
> text(emi$A,labels=rownames(emi$A))
```

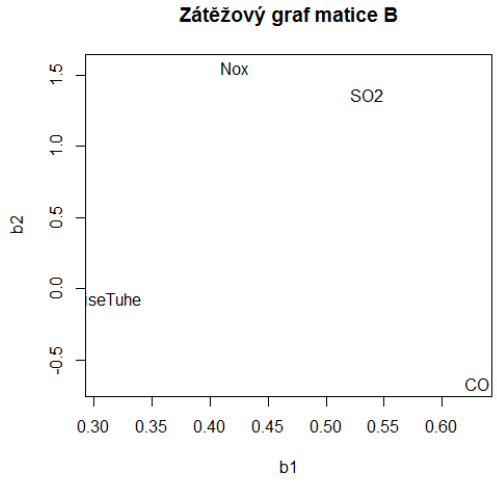
Zátěžový graf, obrázek 10, odhaluje několik skupin. Největší hodnotu podle $\mathbf{a}_1$ má Moravskoslezský kraj, naopak největší hodnotu podle $\mathbf{a}_2$ má Ústecký kraj. Kraj Pardubický a Karlovarský tvoří třetí skupinu. Jihomoravský kraj se do vek-



Obrázek 10: Zobrazení hodnot zátěží matice $\mathbf{A}$.

toru zátěží $\mathbf{a}_1$ odráží nejméně. Zbytek krajů tvoří poslední skupinu s tím, že Vysočina přispívá do zátěže $\mathbf{a}_2$ nejnižší mírou.

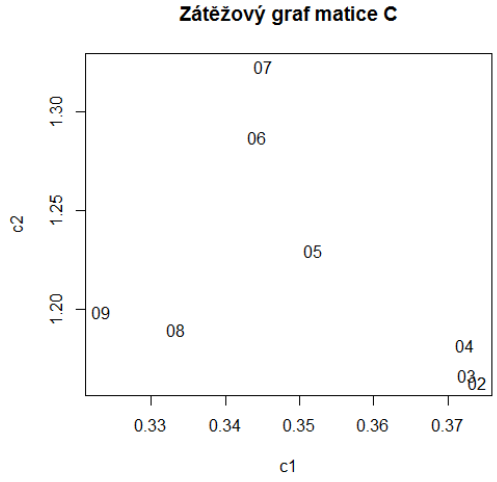
Pro mód znečišťujících složek (obrázek 11) platí, že emise tuhé nabývají nejnižší hodnoty podle vektoru zátěží $\mathbf{b}_1$. Oxid uhelnatý přispívá největší mírou do $\mathbf{b}_1$ a zároveň nejnižší mírou do $\mathbf{b}_2$. Z grafu lze vyvodit, že oxidy dusíku a oxid



Obrázek 11: Zobrazení hodnot zátěží matice **B**.

siřičitý spolu souvisí, tato souvislost byla prokázána i v případě dvourozměrné PCA v kapitole 1.2. Navíc obě složky přispívají nejvíce do $\mathbf{b}_2$ narozdíl od zbylých dvou.

Graf zobrazující vektor zátěží $\mathbf{c}_1$ proti $\mathbf{c}_2$ (obrázek 12) poukazuje na prakticky stejné znečištění v letech 2002, 2003 a 2004. Tyto roky zároveň vykazují vysokou



Obrázek 12: Zobrazení hodnot zátěží matice **C**.

míru přispění do $\mathbf{c}_1$. Druhou skupinu v tomto grafu tvoří roky 2008 a 2009, které naopak nabývají nízkých hodnot jak podle $\mathbf{c}_1$, tak i podle $\mathbf{c}_2$. Rok 2007 má

nejvyšší míru přispění do $\mathbf{c}_2$.

Kromě výše uvedených interpretací, kdy popisujeme každý mód zvlášť, se na získaný model můžeme dívat komplexně. Všechny grafy spolu totiž souvisí, konkrétně pozice jednotlivých proměnných, a každý z nich odráží vliv na ostatní. Ze zátěžových grafů lze konstatovat, že Moravskoslezský kraj má výrazné znečištění oxidem uhelnatým během let 2002-2004. V Ústeckém kraji pak můžeme pozorovat převážně znečištění oxidy dusíku a oxidem siřičitým, které se projevilo v průběhu let 2006 a 2007.

Nutno poznamenat, že v tomto příkladě došlo k nejednoznačnému řešení ve smyslu orientace os (souřadnic), jako se tomu stává při PCA. Tento problém byl vyřešen vynásobením čísla -1 vektorem zátěží druhého a třetího módu. Proto je vždy potřeba výsledný model porovat s naměřenými hodnotami.

4.2 Příklad 2: Lékařská data

Model PARAFAC aplikujeme na data z lékařského prostředí nejmenované nemocnice, viz příloha A.2. Datový soubor obsahuje výsledky měření náhodně vybraných pacientů v průběhu pěti let a tvoří třírozměrný tenzor, $\underline{\mathbf{X}} \in \mathbb{R}^{20\times10\times5}$, jehož přední vrstva má podobu tabulky 2.

Pacient	Gly	Chol	ALT	AST	Urea	Kreat	Leu	Hemo	Tromb	Ery
2009										
ID1	5	5,2	0,84	0,46	3,9	61	9,7	133	250	4,42
ID2	7,1	3,17	1,18	0,57	4,3	75	4,61	143	156	4,8
ID3	4,3	5,2	0,35	0,55	4,5	65	6,25	160	250	5,2
ID4	5,41	6,84	0,33	0,43	3,2	68	5,19	104	278	4,62
ID5	6,6	6,5	0,94	0,57	4,9	66	4,48	136	270	4,7
ID6	5,2	7,81	0,33	0,35	4,8	66	7,6	138	291	4,22
ID7	7,3	4,7	0,61	0,57	3,9	61	9,87	147	225	5,25
ID8	4,9	5,39	0,69	0,58	3,7	68	9,52	127	236	4,6
ID9	4,8	6,58	0,55	0,47	5,2	68	5,1	141	266	9,47
ID10	4,8	6,2	1	0,62	6,1	61	10,7	125	142	4,19
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Tabulka 2: Přední vrstva datového souboru lékařských dat.

Jelikož datový soubor budeme opět importovat, je potřeba nastavit odpovídající cestu k pracovnímu adresáři jako v předchozím příkladě pomocí příkazu `setwd`. Knihovnu `ThreeWay` máme již načtenou, proto můžeme přistoupit k importu dat, které rovnou převedeme do maticové podoby:

```
> p1=as.matrix(read.csv2("2009.csv", header=FALSE, sep=";", dec=","))
```

Tento příkaz provedeme u každého souboru a všech pět matic uspořádáme do jedné, v podstatě provedeme transformaci tenzoru na matici. Protože jsou jednotlivé proměnné měřeny v různých jednotkách, naměřená data normujeme přes první mód:

```
> P=cbind(p1,p2,p3,p4,p5)
> data2=norm3(P, 20, 10, 5, 1)
```

Pomocí algoritmu ALS následně hledáme vhodný model, resp. vypočítáme hodnoty parametrů. K tomu slouží následující funkce s vhodně nastavenými vstupními argumenty, např. počet komponent je nastaven na dvě:

```
> pac<- CPfuncrep(data2, 20, 10, 5, 2, 1, 1, 1, 0, 1e-6, 1000)
```

Ke zjištění hodnot matic zátěží postupujeme stejným způsobem jako u předchozího příkladu:

```
> pac$A
           [,1]        [,2]
 [1,] -6.717092 -0.32898102
 [2,] -6.800721  0.76266166
 [3,] -7.511450  0.11835886
 [4,] -6.068755 -0.57738473
 [5,] -7.315385 -0.11632492
 [6,] -7.383284 -0.18717773
 [7,] -6.935507  0.29741256
 [8,] -6.963079  0.08693758
 [9,] -7.240148 -0.08558099
[10,] -6.240860  0.46610206
[11,] -7.647392 -0.74830667
[12,] -6.826490 -0.01594428
[13,] -7.128953 -0.03066265
[14,] -6.656580  0.76258350
[15,] -7.225445  0.13658928
[16,] -7.366165 -0.02494032
[17,] -8.104946  0.11574646
[18,] -7.191147  0.18406490
[19,] -7.687320  0.61812424
[20,] -7.482944 -0.02049854
```

```
> pac$B
              [,1]         [,2]
 [1,] -0.018382376 -0.006672810
 [2,] -0.018333068  0.004634140
 [3,] -0.002003216 -0.002072476
 [4,] -0.001653640 -0.001416209
 [5,] -0.016937441 -0.009265952
 [6,] -0.237273378 -0.192174185
 [7,] -0.023828345  0.009342572
 [8,] -0.469527115 -0.159538874
 [9,] -0.852569508  1.828530860
[10,] -0.016257093 -0.004231505
> pac$C
          [,1]       [,2]
[1,] 0.4449004 -0.5642161
[2,] 0.4442561 -0.5660081
[3,] 0.4449005 -0.7179163
[4,] 0.4447688 -0.7300557
[5,] 0.4438009 -0.5736778
```

Nalezený model byl zjištěn po 170 krocích. Než přistoupíme k zobrazení hodnot zátěží, ověříme kvalitu získaného modelu. Procento vyrovnání modelem je 99% a minimální hodnota TC byla zjištěna -0.1364217. Oba výsledky vypovídají o nalezení vhodného modelu. Nicméně výsledek ohodnotíme pomocí *„split-half“* analýzy s vhodně nastavenými parametry [16]:

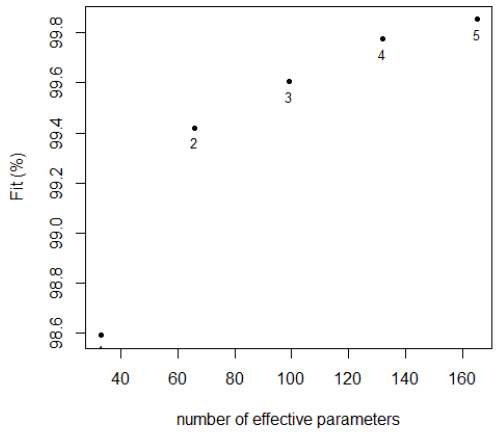
```
> splitCP <- splithalfCP(data2, 20, 10, 5, 2, 0, 0, 0, 0, 1e-6, 10000, 1, 1, 1)
```

Sledujeme hodnoty, jež vypovídají o shodě parametrů:

```
Congruences for A in splits and in appropriate part of Afull
       SPL1 SPL2
Comp.1 1.00 1.00
Comp.2 0.97 0.98
Congruence values for B-mode component matrix
Comp.1 Comp.2
  1.00   0.99
Congruence values for C-mode component matrix
Comp.1 Comp.2
  1.00   0.97
```

Vzhledem k tomu, že shoda byla dosažena ve všech parametrech, a že hodnota TC je přijatelná, považujeme model za vyhovující. Graf (obrázek 13) znázorňuje

% vyrovnání dat modelem pro různý počet komonent, kde vidíme, že počet komponent byl zvolem správně a můžeme přistoupit ke grafické interpretaci výsledků.



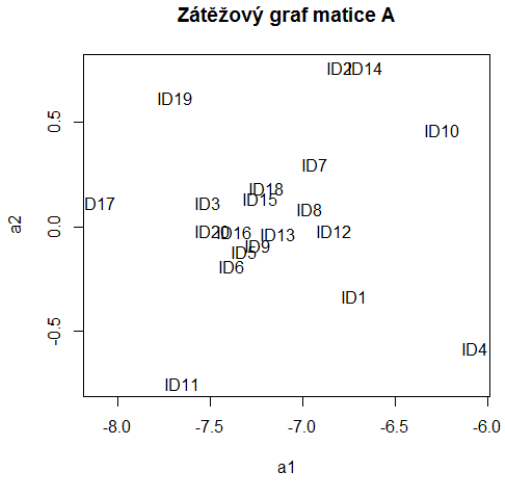
Obrázek 13: Hodnota vyrovnání modelem při různém počtu komponent (%).

Uvedeme postup pouze pro zátěžový graf prvního módu. A abychom se lépe orientovali v grafu, přiřadíme k jednotlivým proměnným popisky:

```
## A
rownames(pac$A)=(paste(c("ID1","ID2","ID3","ID4","ID5","ID6","ID7","ID8","ID9","ID10","ID11",
                         "ID12", "ID13","ID14","ID15","ID16","ID17","ID18","ID19","ID20")))
plot(pac$A[,1],pac$A[,2],type="n",main="Zátěžový graf matice A", xlab="a1",ylab="a2")
text(pac$A,labels=rownames(pac$A))
```

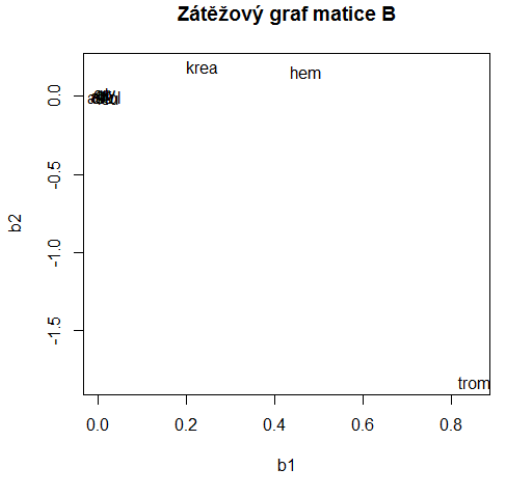
Kromě skupiny pacientů, která je situována uprostřed zátěžového grafu (obrázek 14) vidíme i odlehlá pozorování. Pacient s ID4 nejvíce přispívá do vektoru zátěží $\mathbf{a}_1$, nejnižší hodnotu podle tohoto vektoru má pacient s ID17. Co se týče míry přispění do vektoru $\mathbf{a}_2$, tak nejméně přispívá pacient s ID11 a nejvíce skupina pacientů s ID2 a ID4.

Graf (obrázek 15) mapující zátěže druhého módu ukazuje souvislosti mezi kreatinem a hemoglobinem, které mají nejvyšší hodnotu podle $\mathbf{b}_2$. Další skupina měřných látek tvoří jeden velký shluk a přispívá svoji mírou nejméně do vektoru zátěží $\mathbf{b}_1$. Trombocity se do tohoto vektoru naopak odráží nejvíce, ale svojí mírou přispívají nejméně do $\mathbf{b}_2$. Látky ležící mimo velkou skupinu mají odlišné vlastnosti



Obrázek 14: Zobrazení hodnot zátěží matice **A**.

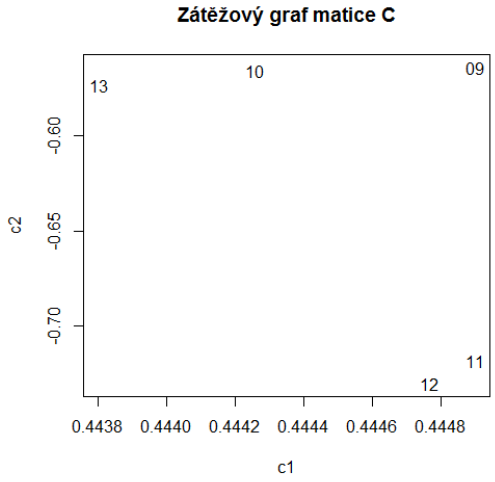
a zřejmě by stálo za to je blíže prozkoumat, zda nemají významný vliv na nějaké onemocnění.



Obrázek 15: Zobrazení hodnot zátěží matice **B**.

Zobrazení vektorů zátěží $\mathbf{c}_1$ versus $\mathbf{c}_2$ (obrázek 16) odhaluje skupinu let 2012 a 2011, které se nejvíce odráží do $\mathbf{c}_1$, stejně jako rok 2009. Nejménší míru přispění do tohoto vektoru pak májí ostatní roky, které ovšem mají vysokou hodnotu podle $\mathbf{c}_2$. Z grafu lze také vyvodit že rok 2012 se nejméně odráží do $\mathbf{c}_2$.

Na jednotlivá vyšetření během let se ovšem můžeme dívat též komplexně,



Obrázek 16: Zobrazení hodnot zátěží matice **C**.

tj. budeme sledovat pozici proměnných ve všech třech grafech současně, jako tomu bylo v předchozím příkladě. Z grafů lze vypozorovat, že oproti ostatním pacientům byly pacientovi s ID4 v letech 2011-2012 naměřeny výrazné hodnoty trombocitů. Dále můžeme konstatovat, že v letech 2009 a 2010 byla při vyšetření zjištěna u pacientů s ID2 a ID14 vyšší hladina kreatininu a hemoglobinu.

4.3 Příklad 3: Data charakterizující vzdělávací systém a trh práce

Poslední příklad je zaměřen na vybrané charakteristiky popisující vzdělávací systému a trh práce ve vybraných zemí Evropské unie, viz příloha A.3, kde jsou též popsány významy jednotlivých proměnných. Data byla získána z [17] a odráží vývoj jednotlivých ukazatelů během pěti let. Výsledky experimentu jsou opět uspořádány do třírozměrného tenzoru, $\underline{\mathbf{X}} \in \mathbb{R}^{25\times5\times5}$, kde přední vrstva má podobu tabulky 3.

Jelikož příprava dat, tzn. nastavení pracovního adresáře nebo import datových souboru, je stejná jako v předchozích příkladech, nebudeme zde tyto postupy znovu zmiňovat. Knihovna `ThreeWay` nabízí kromě již uvedené funkce `CPfuncrep` i interaktivní funkci `CP`, která v sobě zahrnuje další příkazy, například pro normo-

Země	Vzdel	Nezam	Tech	Mob	Niz
2007					
Belgie	43,6	7,5	18,4	2,6	34,8
Bulharsko	37,3	6,9	18,8	8,3	28,7
Česká republika	42	5,4	25	2,1	16,2
Dánsko	66,5	3,8	20	2,5	32,3
Německo	57,7	8,8	25,6	3,1	23,5
Estonsko	56,4	4,8	21,1	4,5	20,4
Irsko	35,4	4,6	23,7	14,2	34
Španělsko	44	8,3	26,6	1,4	50,3
Francie	56,8	8	26,7	2,5	34
Itálie	50,7	6,2	20	1,8	48,6
⋮	⋮	⋮	⋮	⋮	⋮

Tabulka 3: Přední vrstva třírozměného tenzoru (vybrané charakteristiky vzdělávacího systému a trhu práce).

vání vstupních dat, škálování komponent a jiné. Tuto funkci si tedy nyní představíme a aplikujeme na načtená data, ve scriptu uložená pod názvem `data3`. Nejprve si vytvoříme popisky pro jednotlivé země, indikátory a roky, protože patří mezi vstupní parametry interaktivní funkce:

```
> lab_zeme=paste(c("Bel", "Bul","CR","Dan", "Ne","Est","Ir","Spa", "Fra","It","Ky","Lot","Lit",
+ "Mad","Mal","Rak","Pol","Por","Rum","Slovi","Slove","Fin","Sve","VB","Sv"))
> lab_znaky=paste(c("vzdel","nezam","tech","mob","niz"))
> lab_roky=paste(c("07","08","09","10","11"))
```

Po zadání funkce s potřebnými parametry:

```
> educ <- CP(data3, lab_zeme, lab_znaky, lab_roky)
```

nás software přivítá v interaktivní analýze a vyžaduje zadat vstupní parametry. U této interaktivních funkcí si vystačíme se zadáváním číslic. Na začátku musíme zadat rozměry jednotlivých módů, tj. postupně 25, 5 a 5:

```
WELCOME to the interactive CANDECOMP/PARAFAC analysis program
Warning: If you insert an object of mode CHARACTER when not requested,
an error occurs and the program stops!

Specify the number of A-mode entities
1: 25
Read 1 item
Specify the number of B-mode entities
1: 5
Read 1 item
Specify the number of C-mode entities
1: 5
Read 1 item
```

Jelikož lze metodu PARAFAC stejně jako PCA použít jako prvotní metodu pro provedení jiných analýz, v případě PARAFAC je to především ANOVA, tak je i tato analýza zakomponována do této interaktivní funkce. Protože se práce zabývá pouze aplikací metody PARAFAC, zadáme 0:

```
To see ANOVA results, specify 1:
1: 0
Read 1 item
```

V průběhu zadávání vstupních parametrů se stane, že některé parametry nepotřebujeme specifikovat. Je-li tomu tak, zadáme 0. Dále budou zobrazeny ty části scriptu, které jsou potřebné pro naši analýzu. Dále vyžadujeme, aby byl datový soubor normován a to přes první mód, zadáme tedy 1:

```
How do you want to normalize your array?
0 = none (default)
1 = within A-mode
2 = within B-mode
3 = within C-mode
1: 1
Read 1 item
Data have been normalized within A-mode

Note: The preprocessed data are now available in Xprep.
```

Vstupní parametr, který nás zajímá v souvislosti s tímto příkladem, je počet komponent, zadáme 2:

```
How many components do you want to use?
1: 2
Read 1 item
```

Konvergenční kritérium necháme na implicitně nastavené hodnotě, proto zadáme 0:

```
Specify convergence criterion (default=1e-6)
1: 0
Read 0 items
```

Stejně tak postupujeme u maximálního počtu iterací, po zadání tohoto parametru se spustí ALS algoritmus:

```
Specify the maximum number of iterations you allow (default=10000).
1: 0
Read 0 items
```

Pro nás jsou stěžejní informace:

```
Candecomp/Parafac function value is 19.2929159110953 after 519 iterations
Fit percentage is 96.9131334542247 %
Procedure used 0.36 seconds

Candecomp/Parafac analysis with  2  components, gave a fit of  96.91 %
Simple check on degeneracy: inspect matrix of triple congruences
        Comp.1  Comp.2
Comp.1  1.0000 -0.1462
Comp.2 -0.1462  1.0000
```

Model byl nalezen po 519 krocích s téměř 97% vyrovnáním. Z výsledku můžeme vyčíst i informaci o degenerovaném řešení, kde -0.1462 odpovídá minimální hodnotě TC mezi dvěma komponentami. Automatickým výstupem funkce `CP` jsou i hodnoty parametrů, které zde uvádět nebudeme, ale zobrazíme je následně v grafech zátěží. Hodnoty zátěží jsou uloženy pod `educ$A`, `educ$B`, `educ$C`. Dále provedeme *„split-half“* analýzu, která slouží k hodnocení získaného modelu:

```
If you want to carry out a STABILITY CHECK on current or different solution, specify '1':
1: 1
Read 1 item
```

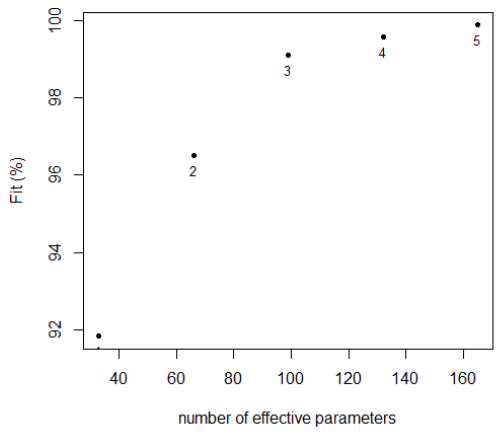
Pro nás budou stěžejní následující informace:

```
Congruences for A in splits and in appropriate part of Afull
       SPL1 SPL2
Comp.1 1.00    1
Comp.2 0.91    1
Congruence values for B-mode component matrix
Comp.1 Comp.2
  0.97   0.67
Congruence values for C-mode component matrix
Comp.1 Comp.2
     1      1
```

Vidíme, že shoda byla dosažena ve většině parametrů a s ohledem na výše uvedené můžeme model považovat za vyhovující. Ještě si ověříme, jaké má model vyrovnání:

```
If you want to carry out a FITPARTITIONING on current solution, specify '1':
1: 1
Read 1 item
Contribution to fit (in %) for all combinations of components
```

Tento příkaz vygeneruje kvalitu vyrovnání hodnot původních dat, hodnoty se vesměs pohybují okolo 97%, což je pro nás dobrý výsledek. Všechny hodnoty zde uvádět nebudeme, lze je zjistit zadáním `educ$fitA`, `educ$fitB`, `educ$fitC`. Dále si zobrazíme kvalitu vyrovnání při různém počtu komponent (postup je stejný jako v předchozích příkladech).

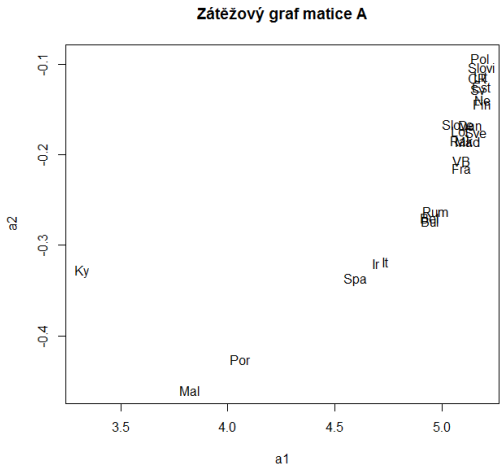


Obrázek 17: Hodnota vyrovnání model při různém počtu komponent (%).

Z grafu (obrázek 17) vidíme, že dvě komponenty jsou postačující. To plyne zejména z toho, že rozdíl mezi vyrovnáním v případě dvou a tří komponent již není tak markantní jako v případě jedné a dvou. Z výše uvedeného tedy dohromady vyplývá, že výsledný model je přijatelný, proto můžeme přistoupit k zobrazení parametrů. Stejně jako v prvním příkladě, tak i zde uvedeme pouze postup pro zátěžový graf matice **A**. Pro lepší orientaci v grafu a snazší interpretaci výsledků popíšeme proměnné:

```
## A
> rownames(educ$A)=(paste(c("Bel", "Bul","CR","Dan", "Ne","Est","Ir","Spa", "Fra","It","Ky",
+ "Lot","Lit","Mad","Mal","Rak","Pol","Por","Rum","Slovi","Slove","Fin","Sve","VB","Sv")))
> plot(educ$A[,1],educ$A[,2],type="n",main="Zátěžový graf matice A", xlab="a1",ylab="a2")
> text(educ$A,labels=rownames(educ$A))
```

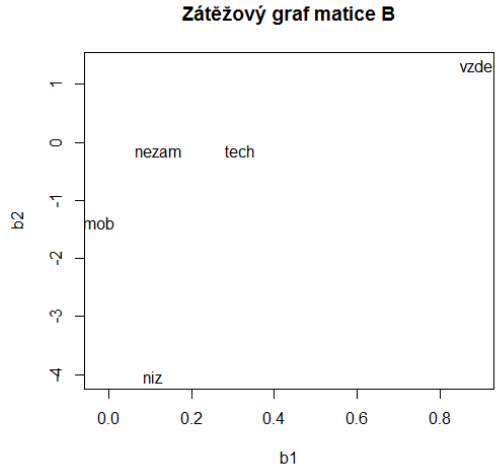
Na obrázku 18 mapujícím zátěže prvního módu vidíme několik skupin zemí. První shluk je tvořen Polskem, Slovinskem, Litvou, Českou republikou, Švýcarskem, Německem a Finskem. Tato skupina má nejvyšší míru přispění do vektoru zátěží $\mathbf{a}_1$ i $\mathbf{a}_2$. Skupina zahrnující Slovensko, Dánsko, Lotyšsko, Rakousko a Maďarsko nabývá též vysokých hodnot podle $\mathbf{a}_1$. Kypr má naopak nejnižší hodnoty podle $\mathbf{a}_1$. Do $\mathbf{a}_2$ se nejméně odráží Malta, která spolu s Portulgalskem tvoří další skupinu. Jisté podobnosti můžeme pozorovat i mezi státy jako je Španělsko, Irsko a Itálie, protože v grafu tvoří také shluk, stejně tak jako Rumunsko s Belgií a Bulharskem či Velká Británie s Francií.



Obrázek 18: Zobrazení hodnot zátěží matice $\mathbf{A}$.

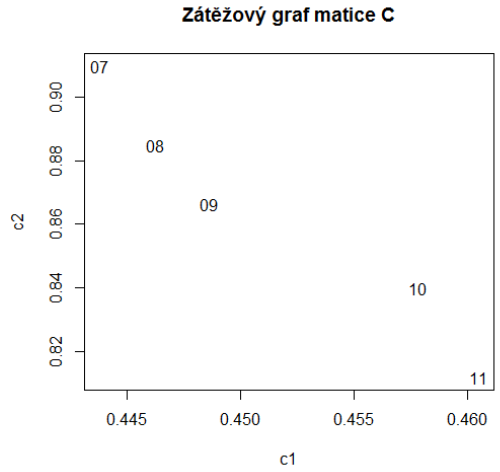
Než přistoupíme k interpretaci zátěžového grafu matice $\mathbf{B}$ (obrázek 19), nutno připomenout, že význam jednotlivých ukazatelů je uveden v příloze A.3. Z grafu vyplývá, že ukazatel „vzdel“, ležící samostatně, přispívá svojí mírou nejvíce do $\mathbf{b}_1$ i $\mathbf{b}_2$. První shluk je tvořen ukazateli „nezam“, „tech“ a „vydaje“, které nabývají vysokých hodnot podle $\mathbf{b}_1$. Ukazetel „mob“ se do vektoru $\mathbf{b}_1$ odráží nejméně, podobně jako „niz“ do $\mathbf{b}_2$. Ze zátěžového grafu můžeme vidět, že hodnoty proměnných „mob“, „niz“ a „tech“ spolu souvisí.

Poslední obrázek 20 zobrazuje zátěže třetího módu. Vidíme, že rok 2011 má největší hodnotu podle $\mathbf{c}_1$ a nejméně přispívá do $\mathbf{c}_2$. Do $\mathbf{c}_2$ nejvíce přispívá svoji mírou rok 2007, který se naopak nejméně odráží do vektoru $\mathbf{c}_1$. V tomto grafu se



Obrázek 19: Zobrazení hodnot zátěží matice **B**.

utvořila skupina roků 2008 a 2009.



Obrázek 20: Zobrazení hodnot zátěží matice **C**.

Na závěr si popíšeme souvislosti mezi jednotlivými grafy. Pozice prvních dvou skupin zemí, vzniklé dle prvního módu našeho modelu, odpovídá s umístěním ukazatele „vzdel“ a rokem 2007, jelikož tyto též nabývají nejvyšších hodnot podle druhého vektoru zátěží, potažmo druhé komponenty. Můžeme tedy konstatovat, že skupině středoevropských zemí a západoevropských zemí odpovídá vysoké procento lidí ve věku 18 až 24 let učastnících se vzdělávání. Naproti tomu pozice

Kypru se shoduje s umístěním ukazatele „mob“ i s pozicemi let 2007-2009, protože zmíněné ukazatelé přispívají nejméně do $\mathbf{b}_1$ stejně jako skupina zemí i rok. Lze tak tedy říci, že Kypr je charakterizován velkou studentskou mobilitou, která se projevila zejména během prvních let výzkumu. Obecně lze poznamenat, že se ve skupině jihoevropských států (Malta, Portugalsko) vyskytuje více obyvatel s nízkou úrovní vzdělání.

Závěr

Pro porozumění přirozené struktury vícerozměrných datových souborů se používají dekompoziční metody, které jednak naleznou skutečnou dimenzi dat, ale usnadní interpretaci výsledků a mohou sloužit jako výchozí metoda průzkumové analýzy dat, jejíž závěry jsou následně použity při dalším statistickém zpracování. V této práci jsem se zabývala vybranými dekompozičními metodami; klíčové bylo především vysvětlit princip fungování metody PARAFAC, jež je zobecněním metody PCA.

Nejprve jsem se věnovala metodě PCA, kterou lze aplikovat na dvourozměrné datové soubory. Uvedla jsem základní pojmy, propojila tři přístupy tvorby modelu a demonstrovala teoretické aspekty na příkladě, kde byly výsledné skóry a zátěže zobrazeny v podobě biplotu. Na závěr jsem zmínila i vlastnosti analýzy. Tuto část jsem zahrnula do své práce, jelikož ji považuji za stěžejní pro pochopení metody PARAFAC, která je určena pro analýzu vícedimenzionálních datových souborů. Dále jsem uvedla základní algebraické pojmy a operace, o které se opírají teoretické základy metody PARAFAC. Poslední kapitolu teoretického celku jsem věnovala samotné metodě PARAFAC, zejména pak jejím přednostem, mezi které patří jednoznačnost výsledného modelu a snadná interpretace výsledků. Podrobně jsem popsala ALS algoritmus, na základě kterého se počítají parametry. Pro sestavení správného modelu jsou zapotřebí i znalosti o předzpracování dat, určení počtu latentních proměnných (hlavních komponent) a ohodnocení výsledného modelu, proto jsem i tyto aspekty zařadila do své práce.

V praktické části jsem všechny teoretické poznatky demonstrovala na třech datových souborech, pocházející z různých vědních disciplín. V každém příkladě jsem představila příslušný datový soubor, dále pak stručně popsala postup výpočtu parametrů ve statistickém softwaru R a zhodnotila sestavený model. Byl-li získaný model vyhovující, interpretovala jsem výstupy, jež jsou v podobně zátěžových grafů.

Doufám, že moje diplomová práce přispěje k další popularizaci tohoto velmi užitečného nástroje mnohorozměrné statistické analýzy.

A Příloha: Data

A.1 Příklad 1

Emise znečišťujících látek na jednoho obyvatele podle vybraných krajů ČR [18].

Území	Emise tuhé	Oxid siřičitý	Oxid dusíku	Oxid uhelnatý
2002				
Jihočeský	7,7	18,3	8,3	18,8
Plzeňský	6,6	21,5	9,7	19,6
Karlovarský	5,7	56,8	26,9	18,1
Ústecký	6,2	98,4	77,1	20,0
Liberecký	4,6	11,0	6,5	15,7
Královéhradecký	5,4	13,5	5,1	17,9
Pardubický	6,3	36,2	27,5	19,2
Vysočina	6,4	8,2	5,3	18,1
Jihomoravský	2,0	3,1	4,6	5,0
Olomoucký	4,0	9,5	7,3	11,6
Zlínský	3,2	11,8	6,2	8,6
Moravskoslezský	5,9	22,8	18,8	103,0
2003				
Jihočeský	8,0	17,7	7,9	19,3
Plzeňský	7,1	21,3	9,4	20,8
Karlovarský	6,1	53,2	28,1	14,7
Ústecký	6,1	88,8	77,4	19,6
Liberecký	4,8	11,1	5,5	16,8
Královéhradecký	5,1	14,2	6,4	18,5
Pardubický	6,0	39,2	29,3	19,8
Vysočina	6,7	8,2	5,4	18,8
Jihomoravský	1,9	3,5	4,0	4,8
Olomoucký	4,2	10,2	6,5	11,0
Zlínský	3,4	11,0	5,0	7,9
Moravskoslezský	5,5	21,8	17,2	108,4
2004				
Jihočeský	8,0	18,6	8,2	19,0
Plzeňský	6,3	21,4	10,1	20,9
Karlovarský	4,9	56,5	27,5	14,3
Ústecký	5,4	87,3	79,3	19,7
Liberecký	4,4	9,6	5,0	16,3
Královéhradecký	4,9	16,8	5,3	18,4
Pardubický	5,7	32,8	25,5	19,9
Vysočina	6,8	7,4	5,3	17,6
Jihomoravský	2,0	3,7	4,3	5,2
Olomoucký	4,0	11,3	6,4	11,3
Zlínský	3,5	15,0	6,5	7,8
Moravskoslezský	5,6	22,8	18,9	117,9
2005				
Jihočeský	4,9	17,6	6,9	15,6
Plzeňský	4,4	21,2	9,4	16,5
Karlovarský	4,1	54,0	24,1	12,7
Ústecký	4,8	87,4	74,8	16,0
Liberecký	2,9	8,6	4,6	13,7
Královéhradecký	3,6	14,7	4,9	16,7
Pardubický	3,8	30,8	22,4	16,0
Vysočina	5,0	6,4	5,7	15,1
Jihomoravský	1,3	3,7	3,9	4,9
Olomoucký	2,7	11,1	6,3	9,3
Zlínský	1,8	12,4	5,9	7,4
Moravskoslezský	4,6	23,5	19,9	105,8

Území	Emise tuhé	Oxid siřičitý	Oxid dusíku	Oxid uhelnatý
2006				
Jihočeský	4,8	16,2	6,2	14,3
Plzeňský	4,2	19,8	9,0	14,8
Karlovarský	5,2	55,7	29,9	11,9
Ústecký	4,5	87,0	76,4	16,0
Liberecký	2,7	8,0	4,2	12,4
Královéhradecký	3,4	14,8	4,6	14,2
Pardubický	3,7	27,6	22,8	13,9
Vysočina	4,8	5,7	4,8	13,7
Jihomoravský	1,2	3,7	3,6	4,9
Olomoucký	2,5	9,3	5,8	9,9
Zlínský	1,7	11,8	5,1	6,9
Moravskoslezský	4,5	23,7	18,9	110,0
2007				
Jihočeský	5,3	15,7	6,2	13,9
Plzeňský	4,4	18,6	8,1	14,5
Karlovarský	5,1	68,9	30,5	12,4
Ústecký	4,8	92,2	75,4	17,5
Liberecký	2,9	6,9	4,1	11,8
Královéhradecký	3,6	14,1	4,3	13,2
Pardubický	4,3	27,9	26,9	13,2
Vysočina	5,4	5,2	4,7	13,9
Jihomoravský	1,5	3,7	3,3	4,6
Olomoucký	2,8	7,9	5,5	9,5
Zlínský	1,9	11,0	4,9	5,8
Moravskoslezský	5,6	24,3	19,0	131,2
2008				
Jihočeský	5,1	16,0	5,6	14,1
Plzeňský	4,3	16,4	6,9	14,8
Karlovarský	4,4	31,9	28,2	12,8
Ústecký	4,1	71,6	67,1	15,4
Liberecký	3,2	7,0	3,3	12,3
Královéhradecký	3,5	13,1	4,0	12,8
Pardubický	3,8	25,2	21,6	13,9
Vysočina	5,0	5,3	3,9	13,9
Jihomoravský	1,5	3,6	3,7	5,6
Olomoucký	2,9	6,7	5,5	9,1
Zlínský	2,1	9,2	4,9	5,8
Moravskoslezský	5,1	18,4	16,4	97,9
2009				
Jihočeský	4,5	15,4	5,4	13,6
Plzeňský	3,8	18,9	5,2	12,5
Karlovarský	4,2	29,7	26,0	11,5
Ústecký	3,7	74,5	66,3	15,5
Liberecký	2,7	6,4	2,8	11,7
Královéhradecký	3,8	10,7	3,4	12,8
Pardubický	3,5	22,6	18,6	13,0
Vysočina	4,7	5,2	3,8	13,8
Jihomoravský	1,5	3,4	3,6	5,7
Olomoucký	2,4	7,1	4,7	8,5
Zlínský	1,9	9,3	4,6	6,1
Moravskoslezský	3,6	17,6	14,9	89,1

A.2 Příklad 2

Výsledky vyšetření vybraných pacientů z nejmenované nemocnice.

Pacient	Gly	Chol	ALT	AST	Urea	Kreat	Leu	Hemo	Tromb	Ery
2009										
ID1	5	5,2	0,84	0,46	3,9	61	9,7	133	250	4,42
ID2	7,1	3,17	1,18	0,57	4,3	75	4,61	143	156	4,8
ID3	4,3	5,2	0,35	0,55	4,5	65	6,25	160	250	5,2
ID4	5,41	6,84	0,33	0,43	3,2	68	5,19	104	278	4,62
ID5	6,6	6,5	0,94	0,57	4,9	66	4,48	136	270	4,7
ID6	5,2	7,81	0,33	0,35	4,8	66	7,6	138	291	4,22
ID7	7,3	4,7	0,61	0,57	3,9	61	9,87	147	225	5,25
ID8	4,9	5,39	0,69	0,58	3,7	68	9,52	127	236	4,6
ID9	4,8	6,58	0,55	0,47	5,2	68	5,1	141	266	9,47
ID10	4,8	6,2	1	0,62	6,1	61	10,7	125	142	4,19
ID11	5,4	4,58	0,61	0,48	4	60	8,32	142	335	4,87
ID12	4,8	5,94	0,74	0,51	5,3	53	5,9	142	238	4,5
ID13	5,8	5,4	0,67	0,54	7,5	86	9,3	134	247	4,47
ID14	5,5	6,85	0,75	0,74	4,7	72	6,5	140	152	4,85
ID15	6,5	5,91	0,3	0,29	5,2	68	9,2	134	261	4,44
ID16	4,5	6,15	0,5	0,42	5,5	78	4,49	148	253	4,88
ID17	5,9	4,62	0,95	0,65	5,9	90	5,62	151	259	4,98
ID18	5	4,4	0,36	0,31	4,7	90,4	6,5	140	258	4,7
ID19	5,4	4,87	0,33	0,47	4,9	83	7,2	138	245	4,9
ID20	5,7	5,7	0,53	0,48	4,5	80	8,1	144	284	4,59
2010										
ID1	5,4	5,12	0,63	0,42	5,1	62	8,8	131	289	4,43
ID2	5,1	2,14	1,1	0,55	5,5	74	4,6	143	156	4,86
ID3	4,6	5,27	0,33	0,52	4,7	65	7,13	163	258	5,22
ID4	4,2	5,46	0,37	0,48	3,5	59	6,1	107	294	4,51
ID5	5,3	6,52	1,9	0,53	5,4	75	10,1	143	274	4,86
ID6	5,2	8,52	0,58	0,46	4,9	69	7,07	133	239	4,84
ID7	5,7	5,2	0,57	0,45	4,2	63	10,2	138	228	5,2
ID8	4,4	5,57	0,7	0,69	4,1	67	4,15	139	232	4,45
ID9	4,9	6,61	0,67	0,44	5,3	74	4,78	148	260	4,72
ID10	4,9	6,4	0,62	0,53	5,4	55	7,72	121	252	4,9
ID11	6,1	4,92	0,47	0,41	4,3	67	7,6	145	357	5,5
ID12	5,1	3,36	0,55	0,48	4,6	53	6,6	140	247	4,79
ID13	5,4	5,66	1,1	0,71	7,2	70	7,19	128	250	4,9
ID14	5,4	5,57	0,58	0,48	5,2	70	6,7	138	176	4,97
ID15	5,5	4,04	0,2	0,3	5,8	70	10,11	126	253	4,35
ID16	4,7	6,13	0,53	0,44	5,7	87	4,2	134	274	4,73
ID17	5,9	6,25	0,9	0,61	5,7	81	5,7	150	269	5,04
ID18	5,6	5,7	0,36	0,3	6,5	80,5	6,93	146	215	4,85
ID19	5,5	6,1	0,36	0,48	3,7	92	7,9	152	188	5,23
ID20	5,8	5,5	0,5	0,45	4,8	82	7,9	145	280	4,5
2011										
ID1	6,8	5,6	0,82	0,47	4,3	53	8,32	127	281	4,41
ID2	6,8	3,17	1,1	0,55	5,5	74	3,75	138	159	4,79
ID3	4,8	5,8	0,48	0,55	5,2	64	7,12	139	245	5,2
ID4	4,2	5,6	0,35	0,45	3,7	60	5,46	105	290	4,55
ID5	7	5,77	0,63	0,4	5,9	71	9,5	149	288	4,88
ID6	4,6	6,49	0,29	0,44	3,8	67	7,9	144	275	4,46
ID7	6,9	4,87	0,63	0,6	4,8	61	7,4	139	194	4,84
ID8	4,3	5,45	0,92	0,76	4,1	72	5,27	133	266	4,37
ID9	4,1	5,31	0,85	0,63	4,8	74	5,69	138	260	4,5
ID10	5,3	6,57	0,62	0,52	4,5	56	3,54	130	139	4,41
ID11	6,1	4,93	0,23	0,26	4,3	71	8,19	142	373	5,5
ID12	5,2	5,2	0,68	0,51	4,8	57	5,2	136	227	4,48
ID13	4,9	6,28	0,64	0,54	5,5	72	9	129	256	4,3
ID14	6,1	6,64	0,56	0,4	4,7	73	6,28	139	147	6,99
ID15	5,3	4,01	0,22	0,29	5,4	64	6,82	148	212	4,98

Pacient	Gly	Chol	ALT	AST	Urea	Kreat	Leu	Hemo	Tromb	Ery
ID16	4,7	5,29	0,34	0,35	5,8	72	4,1	146	262	4,91
ID17	4,8	4,45	0,62	0,52	7	106	5,5	150	275	5,2
ID18	6	7,3	0,42	0,34	7,6	79,5	8,44	140	252	4,54
ID19	6,7	5,96	0,31	0,45	4,3	101	10,28	156	215	5,21
ID20	6,02	5,17	0,78	0,72	4,7	80	8,4	150	282	4,5
2012										
ID1	6,2	5,2	0,55	0,41	5	51,1	8,13	129	253	4,41
ID2	5,7	4,71	0,99	0,57	4,8	76	5,2	145	159	4,95
ID3	5,3	6,39	0,52	0,57	5,4	70	7,1	150	256	5,1
ID4	4,6	5,64	0,37	0,42	3,9	58	5,42	105	284	4,57
ID5	7,1	6,88	0,87	0,45	7	63	8,4	150	280	4,6
ID6	5,1	6,5	0,38	0,35	5,7	73	11,83	154	302	4,59
ID7	4,3	3,46	0,65	0,63	4,4	55	8,38	146	203	5,4
ID8	4,1	5,57	0,7	0,69	4,1	67	5,11	149	210	4,78
ID9	4,5	6,3	0,63	0,5	5,9	75	5,8	144	284	4,69
ID10	4,8	6,4	0,6	0,55	4,8	61	3,13	136	166	4,47
ID11	6,9	4,93	0,33	0,26	4,9	67	8,6	143	369	4,99
ID12	4,8	4,69	0,76	0,59	5,4	51	6,16	138	251	4,38
ID13	5,4	5,49	0,85	0,67	6,2	86	9,3	134	248	4,44
ID14	5,7	6,32	0,39	0,41	3,8	74	6,95	135	136	4,96
ID15	6,2	4,26	0,35	0,29	6,2	71	6,64	147	205	4,91
ID16	4,9	4,97	0,28	0,29	4,9	54	4,91	146	270	5,2
ID17	4,8	4,56	0,72	0,54	5,7	97	5,65	152	280	4,99
ID18	5	3,69	0,35	0,42	6,7	74	7,9	137	220	4,45
ID19	8,1	6,24	0,45	0,45	5,9	101	10,57	148	196	5,15
ID20	6	4,21	1,53	1,4	3,46	50	10,3	155	279	4,63
2013										
ID1	6,2	5,36	0,5	0,35	3,7	54	7,2	127	294	4,43
ID2	5,9	4,39	0,4	0,35	6,9	78	4,61	140	156	4,8
ID3	4,9	6,19	0,91	0,82	4,6	64	7,3	154	252	5,12
ID4	4	5,5	0,35	0,4	3,7	65	4,38	115	240	4,44
ID5	8,4	5,49	0,53	0,32	5,5	72	6,65	139	239	4,75
ID6	5	5,88	0,47	0,41	4,7	67	8,44	150	295	4,67
ID7	5,3	4,54	0,49	0,49	5,1	62	9,1	144	217	5,12
ID8	4,2	6,41	0,46	0,48	3,6	66	6,65	139	239	4,75
ID9	4,2	6,1	0,6	0,52	5,5	55	6,35	137	253	4,41
ID10	4,3	6,39	0,57	0,56	4,5	60	4,3	132	157	4,49
ID11	7,1	4,77	0,42	0,34	4,2	66	7,66	144	317	5,5
ID12	5,4	4,94	0,51	0,44	4,9	59	5,8	142	250	4,4
ID13	5,5	5,32	0,41	0,38	6,5	96	8,56	126	280	4,3
ID14	6,1	5,4	0,38	0,49	4,2	72	5,5	137	155	4,94
ID15	6,8	4,55	0,23	0,31	5,7	74	8,7	138	285	4,88
ID16	5,4	5,84	0,47	0,43	4,2	69	5,4	145	255	5,1
ID17	4,8	4,1	0,84	0,64	6,7	106	5,7	150	289	5,04
ID18	5,2	4,06	0,29	0,35	6,6	73	5,4	128	227	4,34
ID19	7,1	3,36	0,77	0,56	5,7	114	7,93	142	183	4,89
ID20	6	5,79	1,28	0,94	4,5	60	7,4	152	203	4,61

Značení

Gly	Glykemie	vyšetření hladiny cukru v krvi ($mmol/l$)
Chol	Cholesterol	vyšetření hladiny cholesterolu v krvi ($mmol/l$)
ALT	Alaninaminotransferáza	vyšetření plasmy ($\mu kat/l$)
AST	Aspartátaminotransferáza	vyšetření plasmy ($\mu kat/l$)
Urea	Močovina	součástí močových testů ($mmol/l$)
Kreat	Kreatinin	součástí močových testů ($\mu mol/l$)
Leu	Leukocyty	krevní obraz ($10^9/l$)
Hemo	Hemoglobin	krevní obraz (g/l)
Tromb	Trombocyty	krevní obraz ($10^9/l$)
Ery	Erytrocyty	krevní obraz ($10^{12}/l$)

A.3 Příklad 3

Hodnoty vybraných charakteristik popisující vzdělávací systém a trh práce ve vybraných zemí Evropské unie [17].

Země	Vzdel	Nezam	Tech	Mob	Niz
2007					
Belgie	43,6	7,5	18,4	2,6	34,8
Bulharsko	37,3	6,9	18,8	8,3	28,7
Česká republika	42	5,4	25	2,1	16,2
Dánsko	66,5	3,8	20	2,5	32,3
Německo	57,7	8,8	25,6	3,1	23,5
Estonsko	56,4	4,8	21,1	4,5	20,4
Irsko	35,4	4,6	23,7	14,2	34
Španělsko	44	8,3	26,6	1,4	50,3
Francie	56,8	8	26,7	2,5	34
Itálie	50,7	6,2	20	1,8	48,6
Kypr	36,8	4	12,4	56,9	31,1
Lotyšsko	46,1	6,1	11,8	2,5	23,5
Litva	55,1	4,4	20,7	3,3	19,6
Maďarsko	50,8	7,4	13,7	1,8	26,2
Malta	31,9	6,5	15,4	9,9	63,9
Rakousko	46,8	4,5	31,7	4,7	25,2
Polsko	64,7	9,7	16,8	1,8	20,4
Portugalsko	36,9	8,5	25,2	4	71,3
Rumunsko	42,6	6,8	19,7	2,2	30,9
Slovinsko	63,1	5	17,0	2,1	22,2
Slovensko	41,3	11,2	23,4	10,2	18,4
Finsko	61,4	6,9	28,7	2,9	25,4
Švédsko	51,8	6,2	23,9	3	26,5
Velká Británie	44,7	5,4	22,6	0,6	27
Švýcarsko	62,8	3,7	21,8	7,3	21,1
2008					
Belgie	45,1	7	16,6	2,9	33,6
Bulharsko	39,1	5,7	17,9	7,9	28,3
Česká republika	44,5	4,4	26,5	2,6	15,8
Dánsko	68,5	3,5	19,4	2,4	32,9
Německo	55,2	7,6	26,4	3,5	22,3
Estonsko	54,9	5,6	20,5	4,9	20,6
Irsko	35	6,1	24,4	17,7	32,4
Španělsko	44,1	11,4	25,7	1,2	50
Francie	57	7,4	26,2	2,3	32,9
Itálie	50,7	6,8	20,4	1,8	47,8
Kypr	36,9	3,8	12,6	58,4	30,4
Lotyšsko	46,4	7,7	12,7	2,9	22,6
Litva	58,7	5,9	21	3,6	18,4
Maďarsko	50,6	7,9	13,3	1,8	25,8
Malta	37	6,1	12,9	10,9	61,7
Rakousko	47,2	3,9	28,5	4,3	24,4
Polsko	65,1	7,2	16,1	1,8	19,6
Portugalsko	39,1	8,1	27,8	3,9	70,6
Rumunsko	44,1	6,1	16,5	2	30,2
Slovinsko	63,2	4,5	17,6	2,1	21,9
Slovensko	43,5	9,5	20,8	10,7	17,6
Finsko	60,7	6,4	26,8	2,7	25,1
Švédsko	52,9	6,3	23,6	3	26,2
Velká Británie	44,6	5,7	22,9	0,6	26,9
Švýcarsko	63,5	3,4	20,6	8	20,4
2009					
Belgie	46,1	8	16,9	2,7	32,4
Bulharsko	38,1	6,9	18,8	8	27,6
Česká republika	45,2	6,8	24,8	2,7	15,2
Dánsko	71,3	6,1	19,6	2,5	32,2
Německo	57,1	7,9	24,8	3,6	22

Země	Vzdel	Nezam	Tech	Mob	Niz
Estonsko	55,9	14,1	19,4	5,2	19,4
Irsko	36,2	12,2	21,9	14,8	31,5
Španělsko	45,5	18,1	25,6	1,3	49,6
Francie	55,4	9,2	26,2	2,4	32,3
Itálie	51,2	7,9	22,2	2,1	47
Kypr	36	5,5	13,7	56,2	31,2
Lotyšsko	46,3	17,5	13,3	3,3	21,4
Litva	55,3	13,9	21	4	17,9
Maďarsko	50,5	10,1	14,8	2,1	25
Malta	33,4	7	15	11,4	59,5
Rakousko	48,5	4,9	28,7	4,5	23,6
Polsko	63,7	8,3	15,7	2	18,7
Portugalsko	43,1	10	26,6	4,4	69,1
Rumunsko	42,4	7,2	21,7	2,3	30,2
Slovinsko	65	6	17,9	2,2	20,8
Slovensko	44,3	12,1	20,6	11,4	16,5
Finsko	61	8,4	28,2	2,8	24,4
Švédsko	54,5	8,5	24,2	3,2	25,4
Velká Británie	45,7	7,7	21,9	0,6	25,7
Švýcarsko	57,6	4,2	21,7	8,3	20,5
2010					
Belgie	47,5	8,4	16,6	2,6	32,6
Bulharsko	38,2	10,3	19,8	8,1	25,9
Česká republika	47,8	7,4	24,2	2,9	14,4
Dánsko	71,5	7,6	19,3	2,5	31,8
Německo	55,9	7,2	25,7	3,9	21,4
Estonsko	60	17,3	20,5	5,6	18,4
Irsko	39,6	14,1	24	12,8	30,5
Španělsko	49	20,2	24,9	1,3	48,4
Francie	55,4	9,4	26,5	2,5	31,8
Itálie	50,7	8,5	22,7	2,4	46,2
Kypr	37	6,5	13,3	54,9	29,6
Lotyšsko	45,7	19	14,3	4,6	19,6
Litva	55,4	18,1	21,2	5	17,1
Maďarsko	50,6	11,2	15,6	2,4	24,3
Malta	36,2	7	16,3	11,1	58,7
Rakousko	50,4	4,5	29	4,3	23,1
Polsko	62,7	9,7	15,8	2	18
Portugalsko	44,3	11,4	24,9	4,7	67,1
Rumunsko	41	7,6	17,1	3	30,3
Slovinsko	69,7	7,4	21,1	2,3	20,9
Slovensko	46,9	14,4	20,8	12,5	16,3
Finsko	61,9	8,5	31,8	2,9	23,6
Švédsko	57,3	8,8	25,8	3,2	25
Velká Británie	45,4	7,9	22,6	0,7	24,1
Švýcarsko	62,9	4,7	19,9	8,1	20,8
2011					
Belgie	44,2	7,2	17,1	2,9	31,9
Bulharsko	40,8	11,4	19,1	8,6	24
Česká republika	52,4	6,8	22,7	2,9	13,9
Dánsko	71,8	7,7	20,2	2,6	30,7
Německo	57,4	6	27	3,9	18,4
Estonsko	58,9	12,8	21,1	6	17,9
Irsko	41,6	14,9	24	12,8	29,7
Španělsko	50,2	21,8	25,4	1,4	47,2
Francie	56,2	9,3	25,4	2,6	31,1
Itálie	51,1	8,5	22,3	2,6	45,4
Kypr	35,7	8,1	17,2	53,8	28,3
Lotyšsko	49,2	16,5	15,7	5,9	19,5
Litva	58	15,7	21,5	6,4	15,9
Maďarsko	49,6	11	16,5	2,5	23,8
Malta	33	6,6	13,1	11,2	57,2
Rakousko	51,1	4,2	27,3	4,5	22,9
Polsko	61,6	9,8	16,6	2	17,5

Země	Vzdel	Nezam	Tech	Mob	Niz
Portugalsko	49,5	13,4	24,6	4,9	63,8
Rumunsko	41,7	7,7	20,2	3,7	29,4
Slovinsko	73,2	8,3	23,2	2,5	19,7
Slovensko	47,6	13,7	20,3	13,8	15,7
Finsko	61,3	7,9	27,7	2,9	22,9
Švédsko	56,1	8	25,8	3,4	24,4
Velká Británie	43,5	8,2	22,5	0,8	23,8
Švýcarsko	62,5	4,1	21	10,2	20,8

Značení	
Vzdel	lidé ve věku 18 až 24 účastnící se vzdělávání (%)
Nezam	nezaměstnanost lidí s vysokým vzděláním (%)
Tech	absolventi technických oborů (% všech)
Mob	studenti studující v jiných státech EU (%)
Niz	populace s nízkým vzděláním (%)

Literatura

[1] Acar, E., Aykut-Bingol, C., Bingol, H., Bro, R., Yener, B., *Multiway analysis of epilepsy tenzor*, ISMB/ECCB, Vol. 23, Bioinformatics, 2007, i10-i18.

[2] Anděl, J., *Matematická statistika*, Praha: SNTL/ALFA, 1978.

[3] Bro, R., *Multi-Way Analysis in the Food Industry - Models, Algorithms and Applications*, PhD thesis, Universiteit van Amsterdam, The Netherlands, 1998.

[4] Bro, R., *PARAFAC. Tutorial and applications*, Chemometrics and Intelligent Laboratory Systems, 38, 1997, 149-171.

[5] Filzmoser, P., *Multivariate Statistik*, Vorlesungsskriptum, Institut für Statistik und Wahrscheinlichkeitstheorie, Wien, 2007.

[6] Harshman, R. A., *Foundations of the PARAFAC procedure: Models and conditions for an explanatory multi-modal factor analysis*, UCLA working papers in phonetics, 16, 1970, 1-84.

[7] Harvill, D. A., *Matrix algebra from a statistician's perspective*, Springer, New York, 1997.

[8] Hebák, P., a kol., *Vícerozměrné statistické metody (3)*, 2. dopl. vyd, Praha: Informatiorium, 2007, ISBN 978-80-7333-001-9.

[9] Kalivodová, B., *Biplot a jeho aplikace*, Bakalářská práce, Olomouc: UPOL, 2010.

[10] Kolda, T. G., Bader, B. W., *Tenzor decompositions and applications*, SIAM REVIEW, Vol. 51, No. 3, Society for Industrial and Applied Mathematics, 2009, 445-500.

[11] Martin, C. D., *Tenzor Decompositions Workshop Discussion Notes*, American Institute of Mathematics, Palo Alto, CA, 2004.

[12] Meloun, M., Militký J., *Interaktivní statistická analýza dat*, Vyd. 3, Praha: Karolinum, 2012, 953 s. ISBN 987-80-264-2173-9.

[13] Meloun M., Militký J., Hill M., *Statistická analýza vícerozměrných dat v příkladech*, Vyd. 2. , Praha: Academia, 2012, ISBN 978-80-200-2071-0.

[14] Pravdova, V., Boucon, C., Jong de, S., Walczak, B., Massart, D. L., *Three-way principal component analysis applied to food analysis: an example*, Analytica Chimica Acta, 462, 2000, 133-148.

[15] Rinan, A., *Application of PARAFAC on Spectral Data*, PhD thesis, Universiteit van Amsterdam, The Netherlands, 2004.

[16] *CRAN - Package ThreeWay* [online] , dostupné z: *http://cran.r-project.org/web/packages/ThreeWay/ThreeWay.pdf* [citováno 25.1.2014].

[17] *Eurostat Database* [online], dostupné z: *http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_data base* [citováno 3.2.2014].

[18] *Statistické ročenky České republiky* [online], dostupné z: *http://www.czso.cz/csu/redakce.nsf/i/statisticke_rocenky_ceske_republiky* [citováno 17.4.2013].