UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Nuly ve statistické analýze kompozičních dat

Katedra matematické analýzy a aplikací matematiky
Vedoucí diplomové práce: **doc. RNDr. Karel Hron, Ph.D.**
Vypracovala: **Bc. Adéla Vrtková**
Studijní program: N1103 Aplikovaná matematika
Studijní obor: Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2016

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Adéla Vrtková

Název práce: Nuly ve statistické analýze kompozičních dat

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Karel Hron, Ph.D.

Rok obhajoby práce: 2016

Abstrakt: Tato práce se zabývá problematikou výskytu nul v kompozičních datech. Podstatou řešení problému je na úvod rozpoznání různých druhů nul, což následně vede k volbě vhodných metod parametrických nebo neparametrických. Tyto metody se pak snaží o eliminaci nul, popř. poskytují jiný způsob jak s nulami pracovat. Metody pro nuly vzniklé zaokrouhlením jsou založeny na nahrazení nul vhodnou malou hodnotou. Oproti tomu strukturní nuly není ze své podstaty vhodné nahrazovat a vyžadují proto jiný přístup. Závěrem jsou vybrané algoritmy demonstrovány na reálných datech a pomocí softwaru R je ukázáno jejich praktické použití.

Klíčová slova: kompoziční data, nuly vzniklé zaokrouhlením, strukturní nuly, neparametrické nahrazení, modifikovaný EM algoritmus, detekce odlehlých hodnot

Počet stran: 59

Počet příloh: 0

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Adéla Vrtková

Title: Zeros in statistical analysis of compositional data

Type of thesis: Master's thesis

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Karel Hron, Ph.D.

The year of presentation: 2016

Abstract: The thesis deals with the problem of the zero presence in compositional data. Firstly, it is essential to recognize different kinds of zeros and apply suitable parametric or nonparametric methods. Specific algorithms then either eliminate zeros or deal with zeros in a different way. Methods for rounded zeros are based on replacing zeros with suitable small values. On the other hand, it is not appropriate to replace structural zeros due to information they contain. Finally, selected methods are applied to real data sets and practical examples are shown using software R.

Key words: compositional data, rounded zeros, structural zeros, nonparametric replacement, modified EM algorithm, outlier detection

Number of pages: 59

Number of appendices: 0

Language: Czech

Prohlášení

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

V Olomouci dne 12. dubna 2016 .....

podpis

Obsah

Poděkování

Ráda bych poděkovala vedoucímu diplomové práce doc. RNDr. Karlu Hronovi, Ph.D. za spolupráci, za čas a cenné rady, které mi věnoval při konzultacích.

Úvod

Tato práce se zabývá problematikou nul ve statistické analýze kompozičních dat. Na následujících stranách budou nejprve ve zkratce uvedeny základní poznatky o analýze kompozičních dat, dále budou podrobně popsány přístupy a metody k analýze dat s výskytem nul a na závěr budou tyto metody použity na konkrétních datech. Práce je proto určena pro čtenáře, kteří již byli s teorií kompozičních dat obeznámeni, úvodní kapitola bude sloužit hlavně pro zopakování základních pojmů.

S kompozičními daty se můžeme setkat v biologii, geologii i společenských vědách, tedy všude, kde data kvantitativně popisují části jistého celku. Můžou to být např. koncentrace prvků v horninách nebo volební hlasy v jednotlivých volebních obvodech určitého státu. Jednoduše řečeno, kompoziční data nesou výhradně relativní informaci a obsahují pouze nezáporná čísla. Ukázalo se velmi výhodné pro analýzu těchto dat využít logaritmů podílů, tzv. log-ratio, a to právě díky potřebě zvýraznit relativní informaci. Prací s logaritmy podílů však vznikl problém výskytu nul v kompozičních složkách. V praxi je však téměř nemožné se jim vyhnout, a proto nastala potřeba vývoje nových metod, které by se s nulami uměly vypořádat; více k historii je možné nalézt v [12].

Cílem práce je tedy důkladná analýza problematiky nul, což znamená věnovat se i konkrétním druhům nul, protože při samotném statistickém zpracování kompozičních dat nemůže být na všechny nuly nahlíženo stejně. Následně bude využito statistického softwaru R při aplikaci vybraných metod na konkrétní data.

Práce je členěna do čtyř kapitol. Jak bylo výše naznačeno, první kapitola je věnována základním pojmům a poznatkům o kompozičních datech. Účelem této kapitoly je zavést označení, popř. upřesnit terminologii, protože teorie kompozičních dat včetně problematiky nul je k dispozici převážně v anglicky psané literatuře. V druhé a třetí kapitole je pozornost zaměřena na konkrétní druhy nul, a to na tzv. nuly vzniklé zaokrouhlením a strukturní nuly. Poslední kapitola je pak zaměřena na praktickou stránku analýzy kompozičních dat s nulami

a pomocí softwaru R jsou ukázány vybrané metody na konkrétních datech.

Domnívám se, že přínosem práce by mohl být celkový náhled na problematiku, která je stále předmětem současného výzkumu. Vědecké články, které se zabývají právě tímto tématem, se často zaměřují na určitý druh nul. Ráda bych tedy touto prací popsala současný stav poznání a využila i softwaru R pro praktickou ukázku analýzy kompozičních dat s různými druhy nul.

1 Kompoziční data

Kompoziční data, nebo také kompozice, jsou vícerozměrná data, která obsahují výhradně relativní informaci, tedy kvantitativně popisují části jistého celku. Obecněji lze říct, že jedinou podstatnou informaci obsahují podíly mezi složkami dat. V této kapitole je čerpáno hlavně z [6] a [12].

D-složková kompozice je definována jako sloupcový vektor $\mathbf{x} = (x_1, ..., x_D)'$, jehož složky jsou kladná čísla nesoucí relativní informaci. Velmi často je tato relativní informace ve formě procent nebo proporcí, ale lze se setkat i s absolutními četnostmi. Jako příklad lze uvést data s informacemi o výdajích vybraných domácností za různé komodity nebo sledování složení různých potravin.

Výběrovým prostorem kompozic je pak D-složkový simplex

$$S^D = \{\mathbf{x} = (x_1, ..., x_D)', x_i > 0, i = 1, ..., D, \sum_{i=1}^{D} x_i = \kappa\}, \tag{1.1}$$

kde κ je zvolená konstanta, kterou lze interpretovat jako součet složek kompozic. Tento součet bývá většinou 1 nebo 100, což nám dává vhodnou reprezentaci kompozic pomocí proporcí nebo procent, ale v podstatě může být jakýkoli. Díky požadavku na invarianci na změnu měřítka je totiž možné data reprezentovat s libovolným součtem složek bez ztráty informace v datech obsažené. Konkrétní reprezentaci umožní operace zvaná uzávěr, která je obecně pro $\mathbf{x} = (x_1, ..., x_D)'$, $\mathbf{x} \in \mathbb{R}_+^D$, definována jako

$$C(\mathbf{x}) = \left(\frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, ..., \frac{\kappa x_D}{\sum_{i=1}^{D} x_i} \right). \tag{1.2}$$

Dalším nutným předpokladem pro správnou analýzu kompozic je invariance na permutaci, díky kterému závěry z provedené analýzy nezávisí na pořadí složek. Například lze si dovolit seřadit složky podle abecedy, aniž by tím byly determinovány výstupy konkrétní analýzy.

Posledním nutným předpokladem je podkompoziční soudržnost, která zaručuje invarianci na změnu měřítka pro libovolnou podkompozici a nezávislost analýzy

vybraných podkompozic na podkompozicích nezahrnutých do analýzy. Podkompoziční soudržnost navíc souvisí se zachováním obecných geometrických vlastností, které charakterizují také obvykle předpokládanou eukleidovskou geometrii.

Právě podkompoziční soudržnost je narušena, pokud použijeme pro analýzu kompozičních dat eukleidovskou geometrii. Uvažujeme-li dvě kompozice a jejich eukleidovskou vzdálenost, pak její hodnota by měla být větší nebo rovna vzdálenosti podkompozic [2]. Plné kompozice totiž obsahují více informace, a proto není správné, pokud by vzdálenost podkompozic měla větší hodnotu. Tento případ však může nastat v eukleidovské geometrii, a proto je potřeba zavést takovou geometrii, která respektuje podkompoziční soudržnost a veškeré další požadavky uvedené výše.

1.1 Aitchisonova geometrie a práce v souřadnicích

Nevhodnost eukleidovské vzdálenosti a potřeba zavedení vhodné geometrie pro kompoziční data přivedla Johna Aitchisona ke koncepci, kterou uveřejnil v roce 1986 v [1] a po pozdější systematizaci byla souhrnně nazvána jako Aitchisonova geometrie na simplexu. Její součástí je zavedení několika operací, které respektují vlastnosti kompozic a vedou ke struktuře eukleidovského vektorového prostoru.

Uvažujme kompozice $\mathbf{x}, \mathbf{y} \in S^D$. Perturbace $\mathbf{x}$ s $\mathbf{y}$ je kompozice definována jako

$$\mathbf{x} \oplus \mathbf{y} = C(x_1 y_1, ..., x_D y_D). \tag{1.3}$$

Další operací je mocninná transformace kompozice $\mathbf{x}$ reálnou konstantou α, kdy obdržíme kompozici

$$\alpha \odot \mathbf{x} = C(x_1^\alpha, ..., x_D^\alpha). \tag{1.4}$$

Definujme dále Aitchisonův skalární součin kompozic $\mathbf{x}$ a $\mathbf{y}$, jehož výsledkem je reálné číslo

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{i<j} \ln\frac{x_i}{x_j} \ln\frac{y_i}{y_j}. \tag{1.5}$$

Norma kompozice $\mathbf{x}$ je pak dána vztahem

$$\|\mathbf{x}\|_a = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a} = \sqrt{\frac{1}{D} \sum_{i<j} \left(\ln \frac{x_i}{x_j} \right)^2}, \tag{1.6}$$

a Aitchisonova vzdálenost kompozic $\mathbf{x}$ a $\mathbf{y}$ je definována jako

$$\mathrm{d}_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a = \sqrt{\frac{1}{D} \sum_{i<j} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}, \tag{1.7}$$

pro $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus [(-1) \odot \mathbf{y}]$.

Jak bylo dříve naznačeno, ke statistické analýze kompozičních dat se využívá log-ratio souřadnic. Využití logaritmů podílů složek lze vidět ve vztazích (1.5)–(1.7) a bude používáno i dále při práci v souřadnicích, jejíž podstatou je zobrazení kompozic ze simplexu do reálného prostoru s eukleidovskou geometrií. Jinými slovy, kompozice vyjádříme v souřadnicích vzhledem k určité bázi. Na takto vyjádřené kompozice lze pak použít běžné statistické metody.

Jedním z možných vyjádření jsou alr souřadnice (z angl. additive-log-ratio), u kterých se jedná o vyjádření vzhledem k bázi, která není ortonormální. Alr souřadnice zobrazí kompozici ze simplexu do $\mathbb{R}^{D-1}$ a jsou pro kompozici $\mathbf{x} \in S^D$ definovány jako

$$\mathrm{alr}(\mathbf{x}) = \ln \left(\frac{x_1}{x_D}, \frac{x_2}{x_D}, ..., \frac{x_{D-1}}{x_D} \right). \tag{1.8}$$

Vyjádření v alr souřadnicích však nedodržuje invarianci na permutaci složek, protože jako jmenovatele v (1.8) lze zvolit jinou kompoziční složku. Byly proto definovány clr souřadnice (z angl. centered log-ratio), kde simplex je zobrazen do $\mathbb{R}^D$. Pro kompozici $\mathbf{x} \in S^D$ jsou clr souřadnice dány vztahem

$$\mathrm{clr}(\mathbf{x}) = \ln \left[\frac{x_1}{\mathrm{g}(\mathbf{x})}, ..., \frac{x_D}{\mathrm{g}(\mathbf{x})} \right], \ \mathrm{g}(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i \right)^{1/D}. \tag{1.9}$$

V tomto případě se ovšem jedná o souřadnice vzhledem ke generujícímu systému, což vede k singulární varianční matici.

Řešením problémů spojených s předešlými souřadnicovými systémy se ukázaly být ilr souřadnice (z angl. isometric log-ratio). Princip spočívá ve vytvoření ortonormální báze na simplexu vzhledem k Aitchisonově geometrii a následné vyjádření kompozic v této bázi. Pokud uvažujeme nějakou ortonormální bázi $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_{D-1}\}$ na simplexu S^D, pak pro kompozici $\mathbf{x} \in S^D$ jsou ilr souřadnice definovány jako

$$\text{ilr}(\mathbf{x}) = \left(\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, ..., \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a\right). \tag{1.10}$$

Takto vyjádřené kompozice se už řídí pravidly eukleidovské geometrie a lze na ně aplikovat běžné statistické metody. Výsledek statistické analýzy je možné interpretovat v souřadnicích nebo provést převedení zpět na simplex a interpretovat vzhledem k původním kompozičním složkám.

Nyní uvažujme D-složkovou kompozici $\mathbf{x} = (x_1, ..., x_D)'$ a k ní kompozici $\mathbf{x}^{(l)} = (x_1^{(l)}, ..., x_l^{(l)}, x_{l+1}^{(l)}, ...x_D^{(l)})'$, která vznikla takovou permutací složek kompozice $\mathbf{x}$, kdy l-tá složka je přesunuta na první pozici kompozice, neboli $\mathbf{x}^{(l)} = (x_l, x_1, ..., x_{l-1}, x_{l+1}, ..., x_D)'$, čerpáno z [7]. Ilr souřadnice, získané pomocí vztahu

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}}\ \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \quad i = 1, ..., D-1, \tag{1.11}$$

zobrazí kompozici $\mathbf{x}^{(l)}$ na reálný vektor $\mathbf{z}^{(l)} = (z_1^{(l)}, ..., z_{D-1}^{(l)})'$, $l = 1, ..., D$, který pak lze převést zpět na simplex podle následujících vztahů

$$\begin{aligned}
x_1^{(l)} &= \exp\left(\frac{\sqrt{D-1}}{\sqrt{D}} z_1^{(l)}\right), \\
x_i^{(l)} &= \exp\left(-\sum_{j=1}^{i-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j^{(l)} + \frac{\sqrt{D-i}}{\sqrt{D-i+1}} z_i^{(l)}\right), i = 2, ..., D-1, \\
x_D^{(l)} &= \exp\left(-\sum_{j=1}^{D-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j^{(l)}\right).
\end{aligned} \tag{1.12}$$

První souřadnice $z_1^{(l)}$ obsahuje veškerou relativní informaci, která se týká x_l ve vztahu k ostatním složkám. Zbývající ilr souřadnice $z_2^{(l)}, ..., z_{D-1}^{(l)}$ pak vysvětlují

zbytek informace, a protože nereprezentují žádnou konkrétní kompoziční složku, mohou být zvoleny jakkoli, resp. mohou být dopočteny libovolně, avšak vždy tak, aby byla dodržena ortonormalita báze. Takováto volba ilr souřadnic je vhodná, pokud je cílem analýzy predikovat x_l ze zbytku kompozice.

Lze vidět, že veškeré výše uvedené poznatky stojí na předpokladu kladnosti složek kompozic. Z většiny uvedených vztahů je zřejmé, že kvůli podílům a logaritmům je výskyt nul nežádoucí. Protože se ovšem jejich výskytu v reálných datech nelze vyhnout, je nutné nuly pochopit a rozpoznat jejich druhy, které v kompozičních datech lze najít.

1.2 Nuly v kompozičních datech

Ošetření nul v datech je nezbytnou součástí analýzy kompozičních dat. V praxi se s nulami často setkáváme a nelze je tedy přehlížet. V minulosti se objevovaly různé teorie, které se snažily problém nul vyřešit, bohužel většina těchto strategií porušovala základní předpoklady statistické analýzy kompozičních dat. Až John Aitchison dal zavedením log-ratio metodiky impuls správným směrem a problém nul se začal řešit s ohledem na povahu kompozičních dat.

1.2.1 Druhy nul

Prvním druhem nul, se kterými se lze setkat, jsou tzv. nuly vzniklé zaokrouhlením (rounded zeros). Nulová hodnota tohoto typu nemá ve skutečnosti přesně nulovou hodnotu, nýbrž zastupuje pozorovanou hodnotu, která je pod určitým detekčním limitem. To znamená, že pozorovaná hodnota je buď tak malá, že při zaokrouhlování už dostaneme nulu, nebo jsme omezeni např. citlivostí měřícího přístroje. Velmi často se s těmito nulami lze setkat při měření koncentrací látek, protože měřící přístroje mají určitou citlivost na velmi malé hodnoty koncentrací. Může se pak stát, že i když se např. daná látka ve zkoumané hornině vyskytuje, přístroj ji kvůli svému detekčnímu limitu nezachytí. Tedy skutečnou nenulovou hodnotu odpovídající složky sice neznáme, ale můžeme určit její maximální možnou hodnotu, a to právě díky zvolenému detekčnímu limitu. Intuitivně lze

předpokládat, že tento druh nul bude mít smysl nahrazovat nějakou velmi malou hodnotou, čili informace o detekčním limitu bude pro ošetření těchto nul velmi užitečná.

Dalším druhem nul jsou strukturní nuly (structural zeros). Na rozdíl od nul vzniklých zaokrouhlením nesou strukturní nuly informaci, že dané pozorování je opravdu nulové. Například data o struktuře výdajů jednotlivých rodin budou mít nulu u výdajů za alkohol, pokud se jedná o rodinu abstinentů. Tato nula nese důležitou informaci o tom, že opravdu žádné výdaje za alkohol se v dané rodině nevyskytují, a proto není vhodné takovou nulu nahrazovat.

Posledním druhem nul, který bývá rozlišován, jsou diskrétní nuly (count zeros). Jsou to nuly, které nalezneme v diskrétních kompozicích, které popisují, kolikrát daná událost nastala v rámci jednotlivých tříd. Takové kompozice si lze představit i jako realizace multinomicky rozděleného náhodného vektoru. Diskrétní nuly mohou být způsobeny buď nedostatečným rozsahem výběru, nebo omezeným časem, po který bylo prováděno pozorování. V [12] jsou diskrétní nuly rozlišovány od strukturních na základě argumentu, že s vyšším rozsahem výběru nebo prodloužením pozorovacího času by daná hodnota už nulová nebyla. Tyto nuly lze pak považovat za speciální případ nul vzniklých zaokrouhlením.

Práce se v následujících kapitolách dále zaměří na nuly vzniklé zaokrouhlením a strukturní nuly.

2 Nuly vzniklé zaokrouhlením

Metody pro nuly vzniklé zaokrouhlením se snaží o co nejšetrnější nahrazování nul, protože s každým zásahem do původních dat hrozí zkreslení informace. Pokud se tedy použije metoda pro nahrazení těchto nul nesprávně, pak hrozí narušení varianční matice, přecenění vztahu mezi složkami kompozic nebo nechtěné vytvoření odlehlých hodnot, které by mohly ovlivnit odhady statistických charakteristik. Je proto nezbytné u každé metody vědět, kdy je vhodné ji použít, a brát na vědomí všechna rizika, která jsou s imputací hodnot spojená. Jistě si lze dále povšimnout podobnosti metod s přístupy k ošetřování chybějících hodnot. Nuly vzniklé zaokrouhlením lze považovat za speciální případ tzv. NMAR hodnot (z angl. Not Missing At Random), kdy daná hodnota nemohla být pozorována, protože se nachází pod známou hodnotou detekčního limitu, více k tomuto tématu lze pak nalézt v [12].

V následujícím textu budeme dále uvažovat datovou matici $\mathbf{X} = [x_{ij}]_{n\times D}$, kde n je počet pozorování a D je počet složek kompozice (proměnných). V i-tém řádku jsou hodnoty jednotlivých proměnných pro i-té pozorování, v j-tém sloupci pak hodnoty, kterých nabývá j-tá proměnná v jednotlivých pozorováních. Navíc předpokládáme, že všechny nuly v datech vznikly zaokrouhlením. Nebude-li řečeno jinak, $\mathbf{x}_i$ bude označovat i-tý řádek, tj. i-tou kompozici matice $\mathbf{X}$, $i = 1, ..., n$.

Je třeba připomenout, že detekční limit nemusí být pouze jedna hodnota, protože každá proměnná může mít jiný s ohledem na různé způsoby pozorování a měření, s různou citlivostí a přesností. V této kapitole je čerpáno především z [7], [10], [12].

2.1 Neparametrické nahrazení

Jedna z nejjednodušších metod pro ošetření tohoto druhu nul stojí na principu nahrazení veškerých nul v datech vhodnou malou hodnotou. Výhodou této metody je právě její jednoduchost.

Uvažujme kompozici $\mathbf{x}_i = (x_{i1}, ..., x_{iD})'$ datové matice $\mathbf{X}$, která obsahuje nuly

vzniklé zaokrouhlením. Tato kompozice bude nahrazena kompozicí $\mathbf{xr}_i = (xr_{i1}, ..., xr_{iD})'$ podle vztahu

$$xr_{ij} = \begin{cases} \delta_{ij}, & x_{ij} = 0, \\ x_{ij}\left(1 - \dfrac{\sum\limits_{k|x_{ik}=0} \delta_{ik}}{\kappa_i}\right), & \text{jinak}, \end{cases} \tag{2.1}$$

kde δ_{ij} je vhodná malá hodnota, obvykle 65 % detekčního limitu. Nenulové hodnoty jsou přepočítány tak, aby byl dodržen součet složek κ_i. Kompozice $\mathbf{xr}_i$, $i = 1, ..., n$, pak vytvoří novou datovou matici s rozměry $n \times D$ s nahrazenými nulami.

Pokud však data mají více než 10 % nul, pak hrozí přecenění vztahu mezi složkami kompozic a vznik umělé korelace mezi složkami s nulami, navíc tato metoda může podcenit variabilitu v datech. Z toho důvodu je vhodná pouze pro data s menším výskytem nul. V softwaru R je tato metoda k dispozici v knihovně `zCompositions` jako funkce `multRepl`.

2.2 Modifikovaný EM algoritmus

Název metody je odvozen od EM algoritmu, který se používá pro odhad parametrů v případech, kdy data nejsou úplná. Modifikace EM algoritmu pak umožňuje generování vhodných hodnot pro nahrazení nul vzniklých zaokrouhlením, a to v závislosti na informaci, která je obsažena ve zbytku dat. Navíc nuly budou nahrazeny takovým číslem, které je menší než odpovídající detekční limit, který bude dále značen ε_{ij}.

2.2.1 Alr-EM algoritmus

Algoritmus vychází z datové matice $\mathbf{X}$, kterou je třeba nejdříve vyjádřit v alr souřadnicích. Na úvod nahradíme všechny nulové hodnoty v $\mathbf{X}$ 65 % detekčního limitu. Po tomto nahrazení už lze vyjádřit matici v alr souřadnicích, resp. dosazením jednotlivých řádků $\mathbf{x}_i$ do vztahu (1.8) dostaneme postupně alr souřadnice $\mathbf{y}_i$, které budou tvořit matici $\mathbf{Y}$.

Počáteční nahrazení nulových hodnot bude pak iteračně zlepšováno, přičemž řádek $\mathbf{y}_i$ s přepočítanými hodnotami bude značen $\mathbf{yr}_i$. Typ regrese, který je použit pro nahrazení původně nulových hodnot, je označován jako „censored regression“ a zaručuje, že imputované hodnoty jsou pod detekčním limitem. V t-tém kroku algoritmu dostáváme

$$yr_{ij}^{(t)} = \begin{cases} y_{ij}, & y_{ij} \geq \psi_{ij}, \\ \mathbf{y}'_{i,-j}\hat{\boldsymbol{\beta}}_j^{(t)} - \hat{\sigma}_j^{(t)} \dfrac{\phi\left(\dfrac{\psi_{ij} - \mathbf{y}'_{i,-j}\hat{\boldsymbol{\beta}}_j^{(t)}}{\hat{\sigma}_j}\right)}{\Phi\left(\dfrac{\psi_{ij} - \mathbf{y}'_{i,-j}\hat{\boldsymbol{\beta}}_j^{(t)}}{\hat{\sigma}_j}\right)}, & y_{ij} < \psi_{ij}, \end{cases} \tag{2.2}$$

kde $\mathbf{y}_{i,-j}$ je vektor nenulových pozorování i-té kompozice v alr souřadnicích, $\hat{\boldsymbol{\beta}}_j$ představuje vektor koeficientů lineární regrese pozorovaných hodnot j-tého sloupce na příslušné hodnoty ostatních sloupců, tj. pozorovaných hodnot $\mathbf{y}_j$ na odpovídající $\mathbf{Y}_{-j}$, $j = 1, ..., D-1$. Dále $\hat{\sigma}_j^2$ je odhad rozptylu složek $\mathbf{y}_j$, ψ_{ij} je detekční limit ε_{ij} vyjádřený v alr souřadnicích, tj. $\psi_{ij} = \ln(\varepsilon_{ij}/x_{iD})$, a ϕ a Φ jsou po řadě funkce hustoty a distribuční funkce normovaného normálního rozdělení. Nutno podotknout, že lze v algoritmu použít i robustní lineární regresi pro odhad $\hat{\boldsymbol{\beta}}_j$, která eliminuje vliv případných odlehlých hodnot. Podrobněji bude použití robustní regrese uvedeno v souvislosti s ilr-EM algoritmem v následující podkapitole.

Zobrazení $\mathbf{yr}_i$ zpět na simplex na kompozici s nahrazenými nulami provedeme podle vztahů

$$\mathbf{xr}_i = \mathrm{alr}^{-1}(\mathbf{yr}_i) = \begin{cases} xr_{ij} = \kappa \cdot \dfrac{\exp(yr_{ij})}{1 + \sum\limits_k \exp(yr_{ik})}, \\ xr_{iD} = \kappa \cdot \dfrac{1}{1 + \sum\limits_k \exp(yr_{ik})}, \end{cases} \tag{2.3}$$

přičemž $\mathbf{xr}_i$, $i = 1, ..., n$, opět vytvoří novou datovou matici s rozměry $n \times D$ s nahrazenými nulami. Jako kritérium konvergence algoritmu můžeme uvažovat

Frobeniovu normu takové matice, která vznikne rozdílem variančních matic, které vypočteme z dat vyjádřených v alr souřadnicích ve dvou po sobě jdoucích iteracích.

Pro správné fungování algoritmu je potřeba, aby alespoň jedna proměnná neobsahovala žádné nuly, což v praxi může být problém. Navíc je této metodě vytýkáno použití alr souřadnic, které nejsou izometrické a invariantní na permutaci, což vedlo ke snaze využít ilr souřadnic. V softwaru R je pak tento algoritmus dostupný v knihovně `robCompositions` jako funkce `impRZalr`.

2.2.2 Ilr-EM algoritmus

Opět uvažujeme kompoziční datovou matici $\mathbf{X}$ s nulami, které vznikly zaokrouhlením. Dále se vraťme ke značení kompozic s permutovanými složkami, které bylo použito v kapitole 1.1. Algoritmus začíná nahrazením všech nul 65 % detekčního limitu. Datovou matici upravíme tak, že seřadíme jednotlivé složky (sloupce) podle počtu nul sestupně. V prvním sloupci bude tedy proměnná, která obsahuje nejvíce nul.

V dalším kroku se upravená datová matice vyjádří v ilr souřadnicích s využitím vztahu (1.11), přičemž hodnoty původně nulové budou přepočítány pomocí

$$\psi_{i1}^{(l)} = \sqrt{\frac{D-1}{D}}\ \ln\frac{e_{i1}^{(l)}}{\sqrt[D-1]{\prod_{j=2}^{D} x_{ij}^{(l)}}}, \tag{2.4}$$

kde $e_{i1}^{(l)}$ je detekční limit l-té proměnné v původní datové matici. Takto upravenou matici označme $\mathbf{Z}^{(l)}$ a dále budeme rozumět $\mathbf{Z}^{(l)} = [\mathbf{z}_1^{(l)}, \mathbf{Z}_{-1}^{(l)}]$, kde $\mathbf{z}_1^{(l)}$ je první sloupec matice $\mathbf{Z}^{(l)}$ a $\mathbf{Z}_{-1}^{(l)}$ jsou zbylé sloupce.

Následně provedeme odhad koeficientů lineární regrese $\mathbf{z}_1^{(l)}$ na $\mathbf{Z}_{-1}^{(l)}$ pro takové řádky datové matice, kde jsou hodnoty $\mathbf{z}_1^{(l)}$ pozorovány. Vektor odhadnutých koeficientů pak označme $\hat{\boldsymbol{\beta}}^{(l)}$. Dále ho použijeme pro nahrazení původně nulových

hodnot v prvním sloupci matice $\mathbf{Z}^{(l)}$ ve vztahu

$$\hat{z}_{i1}^{(l)} = \mathbf{z}_{i,-1}^{(l)'}\hat{\boldsymbol{\beta}}^{(l)} - \hat{\sigma}^{(l)}\frac{\phi\left(\dfrac{\psi_{i1}^{(l)} - \mathbf{z}_{i,-1}^{(l)'}\hat{\boldsymbol{\beta}}^{(l)}}{\hat{\sigma}^{(l)}}\right)}{\Phi\left(\dfrac{\psi_{i1}^{(l)} - \mathbf{z}_{i,-1}^{(l)'}\hat{\boldsymbol{\beta}}^{(l)}}{\hat{\sigma}^{(l)}}\right)}, \tag{2.5}$$

kde $\hat{\sigma}^{(l)}$ je směrodatná odchylka složek $\mathbf{z}_1^{(l)}$.

Pro toto nahrazení lze použít i robustní lineární regresi, která je v praxi velice užitečná kvůli možnému výskytu odlehlých pozorování ve zkoumaných datech. Robustní odhad regresních koeficientů obdržíme prostřednictvím MM-odhadu pro regresi, který je obecně definovaný jako

$$\hat{\boldsymbol{\beta}}_{MM} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{\mathrm{r}_i(\boldsymbol{\beta})}{\hat{\sigma}}\right), \tag{2.6}$$

kde ρ je vhodná váhová funkce, $\mathrm{r}_i(\beta)$ jsou rezidua regresního modelu a $\hat{\sigma}$ je robustní odhad měřítka. Volbou vhodné váhové funkce je dosaženo snížení vlivu velkých reziduí a dále pak váhová funkce ovlivňuje vlastnosti výsledného regresního odhadu. Do obecného vztahu (2.6) pak dosadíme následovně

$$\hat{\boldsymbol{\beta}}_{MM}^{(l)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{z_{i1}^{(l)} - \mathbf{z}_{i,-1}^{(l)'}\boldsymbol{\beta}}{\hat{\sigma}^{(l)}}\right), \tag{2.7}$$

a odhady $\hat{\boldsymbol{\beta}}_{MM}^{(l)}$ a $\hat{\sigma}^{(l)}$ jsou pak použity pro nahrazení původně nulových hodnot ve vztahu (2.5).

Po ošetření nul v prvním sloupci, ať už pomocí klasické nebo robustní regrese, je matice zobrazena zpět na simplex pomocí vztahů (1.12) a přichází na řadu imputace hodnot v další proměnné. Tímto principem jsou nahrazeny všechny původně nulové hodnoty a vše se opakuje iterativně do dosažení konvergence, kdy kritérium je stejné jako pro alr-EM algoritmus. Posledním krokem celého algoritmu je zobrazení matice zpět na simplex užitím vztahů (1.12) a seřazení proměnných do původního pořadí. Veškeré nulové hodnoty v původní

datové matici tak byly nahrazeny nenulovými hodnotami, které navíc jsou pod detekčním limitem. Příslušná funkce je v softwaru R opět k dispozici v knihovně `robCompositions`, a to jako funkce `impRZilr`.

Jistě si lze všimnout podobnosti alr-EM algoritmu a ilr-EM algoritmu až na použité souřadnice. Jak už bylo naznačeno, vyjádření v alr souřadnicích má své nevýhody, a proto byl algoritmus vytvořen s ilr souřadnicemi. Lze však dokázat, že tyto přístupy jsou ekvivalentní, a tudíž nezáleží na tom, zda zvolíme variantu vycházející z alr nebo ilr souřadnic.

Nejprve bude ukázáno, že mnohonásobné lineární regrese, které jsou založeny na metodě nejmenších čtverců a aplikovány v alr souřadnicích a v ilr souřadnicích, dávají stejné hodnoty vyrovnaných hodnot, důkaz je převzat z [7].

Důkaz: Nechť $\mathbf{y}$ je n-rozměrný sloupcový vektor pozorovaných hodnot závisle proměnné. Nechť $\mathbf{X}$ je datová matice s rozměry $n \times p$, kde první sloupec je tvořen jedničkami, tj. $\mathbf{1}_n$, a zbývající sloupce odpovídají hodnotám $p-1$ proměnných $x_1, ...x_{p-1}$. Uvažujme dále mnohonásobnou lineární regresi $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, kde $\boldsymbol{\beta}$ je sloupcový vektor parametrů a $\mathbf{e}$ je sloupcový vektor náhodných chyb. Podle metody nejmenších čtverců je odhad $\boldsymbol{\beta}$ dán jako $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Vyrovnané hodnoty určíme jako $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, přičemž lze psát $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, kde $\mathbf{H}$ je projekční matice $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Nechť je dána kompozice $\mathbf{x} = (x_1, ..., x_D)'$, $\mathbf{x} \in S^D$ a $\mathbf{s}_1 = (x_1, x_D)', \mathbf{s}_2 = (x_2, x_3, ..., x_D)'$ jsou její podkompozice. Uvažujme podkompozice $\mathbf{s}_1, \mathbf{s}_2$ vyjádřené v alr souřadnicích, tj. $\text{alr}(\mathbf{s}_1) = \ln\frac{x_1}{x_D}$, $\text{alr}(\mathbf{s}_2) = \left(\ln\frac{x_2}{x_D}, ..., \ln\frac{x_{D-1}}{x_D}\right)'$. Nechť $\mathbf{C}_{2D}$ je matice o rozměrech $(D-2) \times (D-2)$ taková, že $\text{ilr}(\mathbf{s}_2) = \text{alr}(\mathbf{s}_2) \cdot \mathbf{C}_{2D}$.

Dolní indexy (a) a (i) budou dále označovat vyjádření v alr a ilr souřadnicích pro konkrétní matice a vektory. Regresní modely $\mathbf{y}_{(a)} = \mathbf{X}_{(a)}\boldsymbol{\beta}_{(a)} + \mathbf{e}_{(a)}$ a $\mathbf{y}_{(i)} = \mathbf{X}_{(i)}\boldsymbol{\beta}_{(i)} + \mathbf{e}_{(i)}$ zastupují modely mnohonásobné lineární regrese vyjádřené vzhledem k alr a ilr souřadnicím a předpokládáme, že $\mathbf{y}_{(a)} = \text{alr}(\mathbf{s}_1)$, $\mathbf{X}_{(a)} = [\mathbf{1}_n, \text{alr}(\mathbf{s}_2)]$ a $\mathbf{X}_{(i)} = [\mathbf{1}_n, \text{ilr}(\mathbf{s}_2)]$. Je nutné dodat, že přestože to není v rámci užitého označení rozlišeno, uvažujeme pro konstrukci regresního modelu práci s výběrem kom-

pozičních dat.

Definujme matici $\mathbf{M}_{2D}$ s rozměry $(D-1)\times(D-1)$, pro kterou platí $\mathbf{X}_{(i)} = \mathbf{X}_{(a)}\mathbf{M}_{2D}$, a je dána jako

$$\mathbf{M}_{2D} = \begin{pmatrix} 1\ 0\ldots 0 \\ 0 \\ \vdots\ [\mathbf{C}_{2D}] \\ 0 \end{pmatrix}.$$

Nejprve bude ukázáno, že projekční matice $\mathbf{H}_{(a)}$ a $\mathbf{H}_{(i)}$ jsou si rovny, přičemž vyjdeme z matice $\mathbf{H}_{(i)}$. Víme, že platí

$$\mathbf{H}_{(i)} = \mathbf{X}_{(i)}(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}'_{(i)},$$

a využijeme vztahu mezi maticemi $\mathbf{X}_{(i)}$ a $\mathbf{X}_{(a)}$,

$$\mathbf{H}_{(i)} = \mathbf{X}_{(a)}\mathbf{M}_{2D}(\mathbf{M}'_{2D}\mathbf{X}'_{(a)}\mathbf{X}_{(a)}\mathbf{M}_{2D})^{-1}\mathbf{M}'_{2D}\mathbf{X}'_{(a)};$$

po úpravě dostaneme

$$\mathbf{H}_{(i)} = \mathbf{X}_{(a)}(\mathbf{X}'_{(a)}\mathbf{X}_{(a)})^{-1}\mathbf{X}'_{(a)} = \mathbf{H}_{(a)}.$$

Nyní můžeme označit $\mathbf{H}_{(i)} = \mathbf{H}_{(a)} = \mathbf{H}$, a tedy $\hat{\mathbf{y}}_{(a)} = \mathbf{H}\mathbf{y}_{(a)}$ a $\hat{\mathbf{y}}_{(i)} = \mathbf{H}\mathbf{y}_{(i)}$. Projekční matice $\mathbf{H}$ je symetrická, idempotentní a platí, že $\mathbf{H}\mathbf{X}_{(a)} = \mathbf{X}_{(a)}$ a $\mathbf{H}\mathbf{X}_{(i)} = \mathbf{X}_{(i)}$, přičemž tyto vlastnosti má každá projekční matice v regresi, která je založená na metodě nejmenších čtverců.

Nechť $\mathbf{C}$ je matice, pro kterou platí $\mathrm{ilr}(\mathbf{x}) = \mathrm{alr}(\mathbf{x})\mathbf{C}$. Přidáním druhého sloupce a druhého řádku jednotkové matice do druhého sloupce a řádku matice $\mathbf{C}$ vznikne matice $\mathbf{M}$,

$$\mathbf{M} = \begin{pmatrix} c_{11} & 0 & c_{12} & \ldots & c_{1(D-1)} \\ 0 & 1 & 0 & \ldots & 0 \\ c_{21} & 0 & c_{22} & \ldots & c_{2(D-1)} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ c_{(D-1)1} & 0 & c_{(D-1)2} & \ldots & c_{(D-1)(D-1)} \end{pmatrix},$$

pro kterou dále platí $(\mathbf{y}_{(i)}, \mathbf{X}_{(i)}) = (\mathbf{y}_{(a)}, \mathbf{X}_{(a)})\mathbf{M}$.

Nyní už lze ukázat, že mnohonásobnou lineární regresí provedenou v alr a ilr souřadnicích dostaneme stejné vyrovnané hodnoty

$$(\hat{\mathbf{y}}_{(a)}, \mathbf{X}_{(a)})\mathbf{M} = (\mathbf{H}\mathbf{y}_{(a)}, \mathbf{H}\mathbf{X}_{(a)})\mathbf{M} = \mathbf{H}(\mathbf{y}_{(a)}, \mathbf{X}_{(a)})\mathbf{M} =$$
$$= \mathbf{H}(\mathbf{y}_{(i)}, \mathbf{X}_{(i)}) = (\mathbf{H}\mathbf{y}_{(i)}, \mathbf{H}\mathbf{X}_{(i)}) = (\hat{\mathbf{y}}_{(i)}, \mathbf{X}_{(i)}).$$

Bylo tedy dokázáno, že $(\hat{\mathbf{y}}_{(i)}, \mathbf{X}_{(i)}) = (\hat{\mathbf{y}}_{(a)}, \mathbf{X}_{(a)})\mathbf{M}$. Neboli při aplikaci mnohonásobné lineární regrese nezáleží na tom, zda je použita na data vyjádřená v alr nebo ilr souřadnicích. □

Díky této ekvivalenci člen $\mathbf{y}'_{i,-j}\hat{\boldsymbol{\beta}}_j^{(t)}$ ze vztahu (2.2), při aplikaci na data v alr souřadnicích, a člen $\mathbf{z}_{i,-1}^{(l)'}\hat{\boldsymbol{\beta}}^{(l)}$ z (2.5), při aplikaci na data v ilr souřadnicích, dají stejné hodnoty. Tuto ekvivalenci nyní rozšíříme na celé vztahy (2.2) a (2.5). Neboli dokážeme, že při nahrazování nul pomocí EM algoritmu nezáleží, zda použijeme alr nebo ilr souřadnice, převzato z [7].

Důkaz: Označme $\mathbf{y}$ závislou proměnnou, která obsahuje původně neznámé hodnoty, a $\mathbf{X}$ hodnoty příslušné pozorovaným proměnným. Dále se budeme držet indexování (a) a (i) z předešlého důkazu. Bez újmy na obecnosti předpokládejme, že neznámé hodnoty jsou v první složce, tj. v x_1, a pozorované proměnné ve zbytku, tj. ve složkách $x_2, ..., x_D$, který označíme $\mathbf{X}_{-1}$.

Potom $\mathbf{X}_{(a)} = [\mathbf{1}_n, \mathrm{alr}(\mathbf{X}_{-1})]$ a prvky $\mathbf{y}_{(a)}$ budou obsahovat odpovídající logaritmy podílu první a poslední složky, tj. obecně $\ln\frac{x_1}{x_D}$. Následně využijeme vztahu (1.11) pro určení $\mathbf{y}_{(i)}$ a $\mathbf{X}_{(i)}$. Každou původně neznámou hodnotu y_k proměnné $\mathbf{y}$ odhadneme ve zvolených souřadnicích pomocí vztahů

$$\hat{y}_{(a)k} = \mathbf{x}'_{(a)k}\hat{\boldsymbol{\beta}}_{(a)} - \hat{\sigma}_{(a)}\frac{\phi\left(\dfrac{\psi_{(a)k} - \mathbf{x}'_{(a)k}\hat{\boldsymbol{\beta}}_{(a)}}{\hat{\sigma}_{(a)}}\right)}{\Phi\left(\dfrac{\psi_{(a)k} - \mathbf{x}'_{(a)k}\hat{\boldsymbol{\beta}}_{(a)}}{\hat{\sigma}_{(a)}}\right)}, \tag{2.8}$$

$$\hat{y}_{(i)k} = \mathbf{x}'_{(i)k}\hat{\boldsymbol{\beta}}_{(i)} - \hat{\sigma}_{(i)}\frac{\phi\left(\dfrac{\psi_{(i)k} - \mathbf{x}'_{(i)k}\hat{\boldsymbol{\beta}}_{(i)}}{\hat{\sigma}_{(i)}}\right)}{\Phi\left(\dfrac{\psi_{(i)k} - \mathbf{x}'_{(i)k}\hat{\boldsymbol{\beta}}_{(i)}}{\hat{\sigma}_{(i)}}\right)}, \tag{2.9}$$

kde $\mathbf{x}_{(a)k}, \mathbf{x}_{(i)k}$ jsou po řadě k-té řádky matic $\mathbf{X}_{(a)}, \mathbf{X}_{(i)}$.

Bude tedy dokázáno, že získané odhady $(\hat{\mathbf{y}}_{(a)}, \mathbf{X}_{(a)})$ a $(\hat{\mathbf{y}}_{(i)}, \mathbf{X}_{(i)})$ jsou ekvivalentní, přičemž je potřeba ukázat, že platí $(\hat{y}_{(i)k}, \mathbf{x}'_{(i)k}) = (\hat{y}_{(a)k}, \mathbf{x}'_{(a)k})\mathbf{M}$, kde $\mathbf{M}$ je matice zavedená v předešlém důkazu. Protože ekvivalence $\mathbf{x}'_{(a)k}\hat{\boldsymbol{\beta}}_{(a)}$ a $\mathbf{x}'_{(i)k}\hat{\boldsymbol{\beta}}_{(i)}$ vyplývá z předešlého důkazu, bude dále rozebrána druhá část vztahů (2.8) a (2.9).

Nejprve je třeba nalézt vztah mezi $\hat{\sigma}_{(a)}$ a $\hat{\sigma}_{(i)}$. Vyjdeme z reziduálních součtů čtverců

$$(\mathbf{y}_{(a)} - \mathbf{X}_{(a)}\hat{\boldsymbol{\beta}}_{(a)})'(\mathbf{y}_{(a)} - \mathbf{X}_{(a)}\hat{\boldsymbol{\beta}}_{(a)}),$$
$$(\mathbf{y}_{(i)} - \mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)})'(\mathbf{y}_{(i)} - \mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)}),$$

kde uvažujeme $\mathbf{y}$ jako n_o-rozměrný sloupcový vektor obsahující pozorované hodnoty dané proměnné. Využijeme poznatků z předešlého důkazu, konkrétně

$$\begin{aligned}(\mathbf{y}_{(i)}, \mathbf{X}_{(i)}) &= (\mathbf{y}_{(a)}, \mathbf{X}_{(a)})\mathbf{M},\\ (\mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)}, \mathbf{X}_{(i)}) &= (\mathbf{X}_{(a)}\hat{\boldsymbol{\beta}}_{(a)}, \mathbf{X}_{(a)})\mathbf{M}.\end{aligned} \tag{2.10}$$

Rovnosti v (2.10) od sebe odečteme

$$(\mathbf{y}_{(i)} - \mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)}, \mathbf{0}) = (\mathbf{y}_{(a)} - \mathbf{X}_{(a)}\hat{\boldsymbol{\beta}}_{(a)}, \mathbf{0})\mathbf{M},$$

přičemž $\mathbf{0}$ je nulová matice s rozměry $n_o \times (D-1)$ a další úpravou dostáváme rovnost

$$(\mathbf{y}_{(i)} - \mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)}, \mathbf{0})'(\mathbf{y}_{(i)} - \mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)}, \mathbf{0}) =$$
$$= \mathbf{M}'(\mathbf{y}_{(a)} - \mathbf{X}_{(a)}\hat{\boldsymbol{\beta}}_{(a)}, \mathbf{0})'(\mathbf{y}_{(a)} - \mathbf{X}_{(a)}\hat{\boldsymbol{\beta}}_{(a)}, \mathbf{0})\mathbf{M}.$$

S poznatky o vlastnostech matice $\mathbf{M}$, které jsou podrobněji rozebrány v [7], lze odvodit vztah $\hat{\sigma}_{(i)} = m_{11}\hat{\sigma}_{(a)}$, kde m_{11} je první prvek hlavní diagonály matice $\mathbf{M}$.

Zbývá odvodit vztah mezi $\boldsymbol{\psi}_{(a)} - \mathbf{X}_{(a)}\hat{\boldsymbol{\beta}}_{(a)}$ a $\boldsymbol{\psi}_{(i)} - \mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)}$, kde $\boldsymbol{\psi}$ je n_u-rozměrný sloupcový vektor detekčních limitů příslušných n_u neznámým hodnotám, který je vyjádřen v příslušných souřadnicích. Platí rovnost $(\boldsymbol{\psi}_{(i)}, \mathbf{X}_{(i)}) = (\boldsymbol{\psi}_{(a)}, \mathbf{X}_{(a)})\mathbf{M}$, od které odečteme druhý vztah z (2.10)

$$(\boldsymbol{\psi}_{(i)} - \mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)}, \mathbf{0}) = (\boldsymbol{\psi}_{(a)} - \mathbf{X}_{(a)}\hat{\boldsymbol{\beta}}_{(a)}, \mathbf{0})\mathbf{M}.$$

Pak platí $\psi_{(i)k} - \mathbf{x}'_{(i)k}\hat{\boldsymbol{\beta}}_{(i)} = m_{11}(\psi_{(a)k} - \mathbf{x}'_{(a)k}\hat{\boldsymbol{\beta}}_{(a)})$, pro každé $k = 1, ..., n_u$, a s využitím vztahu mezi $\hat{\sigma}_{(a)}$ a $\hat{\sigma}_{(i)}$ dostaneme

$$\frac{\psi_{(i)k} - \mathbf{x}'_{(i)k}\hat{\boldsymbol{\beta}}_{(i)}}{\hat{\sigma}_{(i)}} = \frac{\psi_{(a)k} - \mathbf{x}'_{(a)k}\hat{\boldsymbol{\beta}}_{(a)}}{\hat{\sigma}_{(a)}}. \tag{2.11}$$

Dosazením pravé a levé strany z (2.11) do distribuční funkce a hustoty normovaného normálního rozdělení dostaneme tedy stejné hodnoty, přičemž pro zjednodušení v závěrečném odvození bude použit pro podíl hodnot hustoty a distribuční funkce výraz $[\phi(.)/\Phi(.)]$.

Nyní už můžeme dokázat ekvivalenci vztahů (2.8) a (2.9)

$$\begin{aligned}
(\hat{y}_{(i)k}, \mathbf{x}'_{(i)k}) &= \left(\mathbf{x}'_{(i)k}\hat{\boldsymbol{\beta}}_{(i)} - \hat{\sigma}_{(i)}[\phi(.)/\Phi(.)], \mathbf{x}'_{(i)k}\right) \\
&= \left(\mathbf{x}'_{(i)k}\hat{\boldsymbol{\beta}}_{(i)}, \mathbf{x}'_{(i)k}\right) - \left(\hat{\sigma}_{(i)}[\phi(.)/\Phi(.)], \mathbf{0}\right) \\
&= \left(\mathbf{x}'_{(a)k}\hat{\boldsymbol{\beta}}_{(a)}, \mathbf{x}'_{(a)k}\right)\mathbf{M} - \left(m_{11}\hat{\sigma}_{(a)}[\phi(.)/\Phi(.)], \mathbf{0}\right) \\
&= \left(\mathbf{x}'_{(a)k}\hat{\boldsymbol{\beta}}_{(a)}, \mathbf{x}'_{(a)k}\right)\mathbf{M} - \left(\hat{\sigma}_{(a)}[\phi(.)/\Phi(.)], \mathbf{0}\right)\mathbf{M} \\
&= \left(\mathbf{x}'_{(a)k}\hat{\boldsymbol{\beta}}_{(a)} - \hat{\sigma}_{(a)}[\phi(.)/\Phi(.)], \mathbf{x}'_{(a)k}\right)\mathbf{M} \\
&= (\hat{y}_{(a)k}, \mathbf{x}'_{(a)k})\mathbf{M}.
\end{aligned}$$

Bylo tedy dokázáno, že nezáleží na tom, zda vycházíme z alr souřadnic nebo z ilr souřadnic při nahrazování nul vzniklých zaokrouhlením. Odhady původně nulových hodnot v obou souřadnicích jsou ekvivalentní. □

Seznámili jsme se s několika způsoby, jak přistupovat k řešení výskytu nul vzniklých zaokrouhlením. Jednodušší přístupy umožňují rychlé nahrazení nul a při malém výskytu může být narušení varianční struktury minimální. Složitější iterační algoritmy jsou účinným nástrojem pro data, u kterých je podíl nul už větší. Navíc v případě podezření na výskyt odlehlých hodnot existuje možnost využití robustních metod. Obecně však neplatí, že čím složitější metoda, tím lepší výsledky. Vždy je třeba brát v potaz podíl nul v datech a celkovou povahu dat.

3 Strukturní nuly

Přístupy ke strukturním nulám se liší od metod, které jsou určené pro nuly vzniklé zaokrouhlením. Jak už bylo naznačeno, strukturní nuly není vhodné ze své podstaty nahrazovat nenulovými hodnotami, protože bychom přišli o důležitou informaci, kterou tyto nuly nesou. Tudíž není snaha o jejich nahrazení, ale spíše o přizpůsobení známých statistických metod, popř. o pochopení struktury nul. Nutno podotknout, že strukturní nuly jsou velkým problémem v kompozičních datech a neexistuje jednotně preferovaný přístup k těmto nulám. Než-li se práce zaměří na konkrétní případ detekce odlehlých hodnot v kompozičních datech se strukturními nulami, na kterém bude jeden možný přístup k tomuto typu nul demonstrován, bude uveden krátký přehled metod, které se snaží problém těchto nul vyřešit. V této kapitole bude čerpáno zejména z [15] a [16].

Nevhodnost nahrazení strukturních nul se nejvíce projeví při analýze, u které se využívá Aitchisonova vzdálenost (1.7). Pokud je nula nahrazena malým číslem, pak hodnota Aitchisonovy vzdálenosti by mohla být velmi velká, což by znamenalo nechtěný vznik odlehlé hodnoty. Jedním z nejjednodušších přístupů, jak se se strukturními nulami vypořádat bez nahrazování, je pomocí slučování složek kompozice, tzv. amalgamace. Obecně z D-složkové kompozice $\mathbf{x} = (x_1, ..., x_D)'$ vznikne r-složková kompozice, pro $r < D$, jejíž složky jsou výsledkem sloučení některých složek původní kompozice. Budeme-li uvažovat data s informacemi o výdajích domácností, pak můžeme očekávat nulovou hodnotu u výdajů za alkohol v rodinách abstinentů, popř. nulové výdaje za tabákové výrobky u nekuřáckých rodin. Sloučení těchto dvou proměnných do celkových výdajů za alkohol a tabákové výrobky by pak mohlo přispět k menšímu výskytu strukturních nul. Tato na první pohled jednoduchá a účinná operace však není lineární vzhledem k Aitchisonově geometrii a navíc nezachovává relativní informaci. Pokud se ve výjimečných případech sloučení použije, pak je nezbytností jasná interpretace výsledků analýzy vzhledem ke sloučeným složkám.

Další možností je využití modelu, který je sestaven ve dvou fázích. V první fázi proběhne identifikace nul v datech a v druhé fázi je využito vhodného rozdělení

pro modelování rozdělení nenulových hodnot, podrobněji v [4]. Tento přístup může být však výpočetně náročný a navíc pracuje s předpokladem eukleidovské geometrie, která není pro práci s kompozičními daty vhodná.

Strukturní nuly lze dále chápat jako indikátory dvou různých podskupin. Pozorování s nulou v dané proměnné mohou tvořit jednu skupinu a pozorování s nenulovou hodnotou skupinu druhou. Otázkou pak ale pořád zůstávají nuly mimo zvolenou proměnnou, která indikuje podskupiny, tudíž toto řešení nemusí vést k podstatnému zlepšení. Navíc opět hrozí narušení relativní informace, a tedy narušení celkové struktury dat, včetně nechtěného vzniku odlehlých hodnot.

3.1 Detekce odlehlých hodnot

Snaha o identifikování odlehlých pozorování vyplývá z úsilí o vhodně zvolený statistický model, nestranné odhady parametrů a celkově relevantní výsledky analýzy. Detekce odlehlých hodnot se provádí před samotnou analýzou dat a je obvykle založena na Mahalanobisově vzdálenosti, která je vypočtena z dat v ilr souřadnicích. Pro kompozice vyjádřené v ilr souřadnicích $\mathbf{z}_1, ..., \mathbf{z}_n$ je Mahalanobisova vzdálenost definována jako

$$\mathrm{MD}(\mathbf{z}_i) = [(\mathbf{z}_i - \mathbf{t})' \mathbf{C}^{-1} (\mathbf{z}_i - \mathbf{t})]^{\frac{1}{2}}, \tag{3.1}$$

kde odhad střední hodnoty $\mathbf{t}$ je $(D-1)$-rozměrný sloupcový vektor a $\mathbf{C}$ je odhad varianční matice s rozměry $(D-1) \times (D-1)$, typicky se jedná o výběrový průměr a výběrovou varianční matici. Jestliže $\mathrm{MD}^2(\mathbf{z}_i)$ překročí zvolený kvantil χ^2-rozdělení s $D-1$ stupni volnosti, tj. $\mathrm{MD}^2(\mathbf{z}_i) > \chi^2_{D-1,q}$, pak dané pozorování je označeno jako potenciální odlehlé pozorování. Samotná Mahalanobisova vzdálenost (3.1) může být ovlivněna odlehlými hodnotami, a proto se doporučuje použít robustní odhady $\mathbf{t}$ a $\mathbf{C}$, které získáme s využitím metody minimálního kovariančního determinantu (MCD).

Metoda MCD je založena na vybrání takových h pozorování ze všech n pozorování, pro která dostaneme minimální hodnotu determinantu odpovídající výběrové varianční matice (obvykle volíme $h \approx 3n/4$). Odhad střední hodnoty

$\mathbf{t}$ pak dostaneme pomocí výběrového průměru přímo z těchto h pozorování a odhad varianční matice $\mathbf{C}$ odpovídá příslušné výběrové varianční matici. Při hledání řešení se často využívá tzv. rychlá metoda MCD, která vychází z náhodného počátečního výběru h pozorování, který je dále iterativně zlepšován tak, aby determinant výběrové varianční matice klesal; více k rychlé metodě MCD k dispozici v [14].

Poměrně nenáročný přístup k detekci odlehlých hodnot v kompozičních datech však selhává, jsou-li v datech strukturní nuly, neboť taková data nelze vyjádřit v ilr souřadnicích. Otázka tedy je, zda se vyplatí jít proti povaze strukturních nul, a tedy je imputovat, a jestliže ano, pak jakým způsobem. Jednou možností je nuly jednoduše nahradit určitou hodnotou a pracovat s takto upravenou datovou maticí. Druhý komplexnější přístup je založen na nahrazení nul prostřednictvím algoritmu pro chybějící hodnoty a následné detekci odlehlých hodnot vzhledem k nenulovým pozorováním i vzhledem ke struktuře nul. Tyto dvě možnosti budou dále podrobněji rozpracovány.

3.1.1 Jednoduché nahrazení

Nenáročný algoritmus založený na nahrazení nul byl představen v [5]. Nechť m je počet nul v D-složkové kompozici. Nechť $o \subset \{1, ..., D\}$ indexuje složky s nulovou hodnotou v kompozici $\mathbf{x}$ a $v = \{1, ..., D\} \setminus o$ odpovídá indexům nenulových složek kompozice $\mathbf{x}$. Nuly budou nahrazeny podle vztahu

$$\mathbf{x}_{(o)} = \delta \frac{1}{D^2}(m+1)(D-m) \tag{3.2}$$

a nenulové hodnoty budou přepočítány pomocí

$$\mathbf{x}_{(v)} = \mathbf{x}_{(v)} - \mathbf{x}_{(v)}\delta \frac{1}{D^2} m(m+1), \tag{3.3}$$

kde δ je maximální zaokrouhlovací chyba.

Metoda sice zachovává relativní informaci obsaženou v nenulových složkách, ale je v rozporu s povahou strukturních nul. Jak bylo řečeno na začátku této kapitoly, strukturní nuly nesou důležitou informaci, že dané pozorování je skutečně

nula, tedy nahrazení malou hodnotou neodpovídá charakteru těchto nul. Na druhou stranu i přes tento zjevný rozpor může být metoda využita pro imputaci nul, aby následně bylo možné datovou matici vyjádřit v ilr souřadnicích, vypočítat Mahalanobisovy vzdálenosti a provést detekci odlehlých hodnot.

3.1.2 Dvoustupňový algoritmus pro práci se strukturními nulami

V první fázi algoritmu označíme všechny nuly za chybějící hodnoty a pomocí metod pro nahrazení chybějících hodnot je imputujeme. Toto prvotní nahrazení je použito pouze proto, aby bylo možné data vyjádřit v ilr souřadnicích. Následně se totiž využije informace pouze z původně nenulových složek, pomocí které jsou vypočteny robustní odhady $\mathbf{t}$ a $\mathbf{C}$ s využitím metody MCD.

Uvažujeme D-složkovou kompozici $\mathbf{x}_i = (x_{i1}, ..., x_{iD})'$, $i = 1, ..., n$, která obsahuje strukturní nuly. Dále předpokládejme, že kompozice $\mathbf{x}_i$ má $D - K(i)$ strukturních nul, přičemž $2 \leq K(i) \leq D - 1$. Strukturní nuly nyní dáme na začátek dané kompozice, tedy $\tilde{\mathbf{x}}_i = (0, ..., 0, x_{ij_1}, ..., x_{ij_{K(i)}})'$, kde x_{ij_k} odpovídá k-té nenulové složce kompozice $\mathbf{x}_i$, pro $k \in \{1, ..., K(i)\}$. Existuje permutační matice $\tilde{\mathbf{P}}_i$ s rozměry $D \times D$, která obsahuje pouze nuly a jedničky, a pro kterou platí $\tilde{\mathbf{x}}_i = \tilde{\mathbf{P}}_i \mathbf{x}_i$, pro $i = 1, ..., n$. Toto permutování složek kompozic umožní, po vyjádření upravených dat v ilr souřadnicích, extrahovat informaci pouze o nenulových složkách.

Následující vztahy mezi původními a permutovanými kompozicemi budou dále nezbytné pro závěrečný výpočet Mahalanobisových vzdáleností. Nyní pro přehlednost předpokládejme kompozici $\mathbf{x}$ s nahrazenými strukturními nulami a její vyjádření v ilr souřadnicích $\mathbf{z}$, které bylo provedeno pomocí (1.11). Dále uvažujme příslušné odhady $\mathbf{t}$ a $\mathbf{C}$. Permutovaná kompozice $\tilde{\mathbf{x}} = \tilde{\mathbf{P}}\mathbf{x}$ s permutační maticí $\tilde{\mathbf{P}}$ má vyjádření v ilr souřadnicích následující

$$\tilde{\mathbf{z}} = \mathbf{Q}'\mathbf{z}, \tag{3.4}$$

kde $\mathbf{Q}$ je ortogonální matice, pro kterou platí $\mathbf{Q} = \mathbf{V}'\tilde{\mathbf{V}}$. Matice $\mathbf{V}$ s rozměry $D \times (D-1)$ je dána pomocí $D-1$ sloupcových vektorů, které tvoří sloupce matice

$\mathbf{V}$, a jsou dány jako

$$\mathbf{v}_{\cdot j} = \sqrt{\frac{D-j}{D-j+1}} \left(0, ..., 0, 1, -\frac{1}{D-j}, ..., -\frac{1}{D-j}\right)', \quad j = 1, ..., D-1, \quad (3.5)$$

přičemž j-tý sloupec obsahuje $j-1$ nul, a dále platí $\tilde{\mathbf{V}} = \tilde{\mathbf{P}}\mathbf{V}$. Obdobně pak $\tilde{\mathbf{t}}$ a $\tilde{\mathbf{C}}$ jsou odhady, které obdržíme, vyjdeme-li z permutované kompozice, a pokud $\mathbf{t}$ a $\mathbf{C}$ jsou afinně ekvivariantní, tj. jejich chování při lineární transformaci náhodného vektoru odpovídá chování střední hodnoty a varianční matice při lineární transformaci, pak platí $\tilde{\mathbf{t}} = \mathbf{Q}'\mathbf{t}$ a $\tilde{\mathbf{C}} = \mathbf{Q}'\mathbf{C}\mathbf{Q}$.

Nyní už je možné extrahovat pouze informaci odpovídající nenulovým složkám kompozice. Uvažujeme opět kompozici $\mathbf{x}_i$, její permutační matici $\tilde{\mathbf{P}}_i$ a dále matici $\mathbf{Q}_i = \mathbf{V}'\tilde{\mathbf{V}}_i = \mathbf{V}'\tilde{\mathbf{P}}_i\mathbf{V}$. Využijeme (3.4) a dostaneme $\tilde{\mathbf{z}}_i = \mathbf{Q}_i'\mathbf{z}_i$, kde $\mathbf{z}_i$ je kompozice $\mathbf{x}_i$ vyjádřená v ilr souřadnicích. Posledních $K(i) - 1$ složek $\tilde{\mathbf{z}}_i$, které obsahují veškerou relativní informaci vztahující se k nenulovým složkám $\mathbf{x}_i$, označme $\tilde{\mathbf{z}}_i^*$. Analogicky máme $\tilde{\mathbf{t}}_i = \mathbf{Q}_i'\mathbf{t}$ a posledních $K(i)-1$ složek vektoru $\tilde{\mathbf{t}}_i$ označme $\tilde{\mathbf{t}}_i^*$. Nakonec pravý dolní blok matice $\tilde{\mathbf{C}}_i = \mathbf{Q}_i'\mathbf{C}\mathbf{Q}_i$, který má rozměry $(K(i)-1) \times (K(i)-1)$, označme $\tilde{\mathbf{C}}_i^*$.

Dostali jsme se tak k odhadům střední hodnoty a varianční matice pro nenulové složky kompozice $\mathbf{x}_i$, po řadě $\tilde{\mathbf{t}}_i^*$, $\tilde{\mathbf{C}}_i^*$. Robustní Mahalanobisova vzdálenost použita pro detekci odlehlých hodnot kompozice $\mathbf{x}_i$ je pak dána jako

$$\mathrm{MD}(\tilde{\mathbf{z}}_i^*) = \left[(\tilde{\mathbf{z}}_i^* - \tilde{\mathbf{t}}_i^*)'\tilde{\mathbf{C}}_i^{*-1}(\tilde{\mathbf{z}}_i^* - \tilde{\mathbf{t}}_i^*)\right]^{\frac{1}{2}}, \quad (3.6)$$

kde $i = 1, ..., n$. Jestliže $\mathrm{MD}^2(\tilde{\mathbf{z}}_i^*) > \chi^2_{K(i)-1,0,975}$, pak je kompozice $\mathbf{x}_i$ označena jako potenciální odlehlé pozorování. Tímto je tedy uzavřena první fáze, kdy probíhá identifikace potenciálních odlehlých pozorování vzhledem k nenulovým pozorováním.

Tento postup výpočtu Mahalanobisových vzdáleností lze nyní srovnat se situací, kdy jsou tyto vzdálenosti, včetně příslušných odhadů $\mathbf{t}$ a $\mathbf{C}$, vypočteny přímo z podkompozic, které odpovídají jednotlivým kombinacím nul v kompozičních složkách. Případné rozdíly pak budou signalizovat možná specifika rela-

tivní struktury takovýchto podskupin, např. struktury výdajů domácností abstinentů.

V druhé fázi se uskuteční detekce potenciálních odlehlých hodnot vzhledem ke struktuře strukturních nul, tj. mohlo by nás zajímat, které kombinace nulových hodnot signalizují odlehlá pozorování. Možných přístupů k analýze tohoto efektu je více. Zatím však nejsou dostatečně rozpracovány a u většiny se jedná o předběžnou verzi algoritmu. Z toho důvodu bude v práci uveden jen jeden z těchto předběžných přístupů, který využívá Pearsonova testu dobré shody a je popsán v [15].

Uvažujme počet všech možných konfigurací nul v D-složkové kompozici, $k = 2^D$, který odpovídá počtu všech možných tříd. Teoretické relativní četnosti jednotlivých tříd, v případě bez odlehlých pozorování, odpovídají hodnotě $p_i = \frac{1}{2^D}$, pro všechna $i = 1, ..., 2^D$. Dle nulové hypotézy očekávané teoretické relativní četnosti p_i odpovídají pozorovaným relativním četnostem $\hat{p}_i$. Pokud $\hat{p}_i$ budou svědčit ve prospěch alternativní hypotézy, pak se objeví třídy s odlehlými pozorováními, kterým budou odpovídat malé hodnoty $\hat{p}_i$. Testová statistika

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}, \tag{3.7}$$

kde X_i jsou pozorované absolutní četnosti (odpovídající pozorovaným relativním četnostem $\hat{p}_i$), má za platnosti nulové hypotézy přibližně χ^2 rozdělení s $k - 1$ stupni volnosti, a pokud $Q_{k-1} \geq \chi^2_{k-1;0,975}$, zamítáme nulovou hypotézu o shodě očekávaných a teoretických četností. Kvůli aproximaci χ^2 rozdělením je nezbytné, aby byla splněna podmínka $np_i \geq 5$, neboli $n \geq 5 \cdot 2^D$.

Pro alternativní hypotézu budou svědčit jak velké, tak malé hodnoty $\hat{p}_i$, přičemž rozepíšeme-li si součet z (3.7) na jednotlivé sčítance, pak lze poznat, které třídy přispívají k zamítnutí nulové hypotézy. Pro detekci odlehlých pozorování nás však budou zajímat pouze takové třídy, kde $p_i \geq \hat{p}_i$, $i = 1, ..., 2^D$, kdy cílem je z nich určit takové, které nejvíce svědčí pro alternativní hypotézu. Jednou z možností, jak z těch tříd, kde $p_i \geq \hat{p}_i$, vybrat ty s potenciálními odlehlými pozorováními, je seřadit odpovídající sčítance a vykreslit je pomocí scree

diagramu. Za potenciální odlehlá pozorování označí ta, která odpovídají třídám, resp. sčítancům, která jsou před „loktem“ scree diagramu. Nakonec se informace o odlehlých pozorováních obdržená z této druhé fáze spojí s informací z fáze první.

V této kapitole byly představeny dva způsoby detekce odlehlých hodnot v kompozičních datech se strukturními nulami. Oba jsou založeny na výpočtu Mahalanobisových vzdáleností, přičemž první algoritmus jako výchozí bod používá jednoduché nahrazení strukturních nul, zatímco druhý algoritmus přistupuje k detekci odlehlých pozorování poněkud komplexněji. Nejprve detekuje odlehlá pozorování vzhledem k nenulovým podkompozicím a následně provede to stejné vzhledem ke struktuře nul.

4 Praktická část

V této části bude s využitím statistického softwaru R ilustrováno použití dříve představených metod na konkrétních datech. První příklad takto ukáže použití metod pro nuly vzniklé zaokrouhlením na reálných kompozičních datech, kde se nulové hodnoty přirozeně vyskytují, navíc bude použit clr-biplot pro zobrazení dat po nahrazení nul. V druhém příkladu bude využito reálných dat bez nul, kam budou nuly imputovány simulačně a následně bude posuzována kvalita nahrazení vzhledem k původním datům. Pro třetí příklad získáme data vygenerováním hodnot z mnohorozměrného normálního rozdělení, kam imputujeme nuly simulačně, a opět nás bude zajímat kvalita nahrazení vzhledem k původním datům. Na závěr této části bude ilustrováno použití první fáze dvoustupňového algoritmu pro práci se strukturními nulami. Nejprve si ale definujme pojmy, které v následujících příkladech budou použity.

Clr-biplot, nebo také kompoziční biplot, je založen na singulárním rozkladu datové matice $\mathbf{X}$, která je nejprve vyjádřena v clr souřadnicích a centrována, čerpáno z [3], [13]. Tuto výchozí matici pro konstrukci clr-biplotu označme $\mathbf{Z}$, přičemž má stejně rozměry jako původní datová matice $\mathbf{X}$, tj. $n \times D$. Uvažujme její singulární rozklad $\mathbf{Z} = \mathbf{LKM}'$, kde $\mathbf{L}$ je matice vlastních vektorů matice $\mathbf{ZZ}'$, $\mathbf{M}$ je matice vlastních vektorů matice $\mathbf{Z}'\mathbf{Z}$, dále matice $\mathbf{K} = \mathrm{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_s})$, kde $\lambda_1, ..., \lambda_s$ jsou sestupně seřazená kladná vlastní čísla matice $\mathbf{ZZ}'$ a s je hodnost matice $\mathbf{Z}$. Dále ve smyslu výše uvedeného singulárního rozkladu aproximujeme matici $\mathbf{Z}$ pomocí matic hodnosti 2 následovně,

$$\mathbf{Z} \approx \mathbf{Y} = (\mathbf{l}_1, \mathbf{l}_2) \begin{pmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_1} \end{pmatrix} \begin{pmatrix} \mathbf{m}_1' \\ \mathbf{m}_2' \end{pmatrix}.$$

Pro $c \in \langle 0, 1 \rangle$ označme

$$\mathbf{G} = (\mathbf{l}_1, \mathbf{l}_2) \begin{pmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_1} \end{pmatrix}^{1-c}, \quad \mathbf{H} = (\mathbf{m}_1, \mathbf{m}_2) \begin{pmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_1} \end{pmatrix}^{c},$$

přičemž matice $\mathbf{G}$ má rozměry $n \times 2$, a $\mathbf{H}$ má rozměry $D \times 2$ a platí $\mathbf{Y} = \mathbf{GH}'$. Zde už se dostáváme ke konstrukci clr-biplotu, kdy položíme $c = 1$, a využijeme řádky

matice $\mathbf{G}$ pro zobrazení jednotlivých pozorování a řádky matice $\mathbf{H}$ k zobrazení proměnných.

Při grafické interpretaci clr-biplotu budeme využívat následujících poznatků. V clr-biplotu jsou vykresleny body na pozicích $\mathbf{g}_i$, $i = 1, ..., n$, přičemž každý přísluší právě jednomu z n pozorování, a lze z nich identifikovat potenciální odlehlá pozorování. Dále jsou vykresleny tzv. paprsky (šipky), kdy každý přísluší právě jedné proměnné, a vedou od počátku O k určitému vrcholu $\mathbf{h}_j$, $j = 1, ..., D$. Dále pak uvažujeme spojnice vrcholů paprsků $\mathbf{h}_j$ a $\mathbf{h}_k$. Paprsky $\overline{O\mathbf{h}_j}$ a spojnice $\overline{\mathbf{h}_j\mathbf{h}_k}$ poskytují informaci o relativní variabilitě v kompozičních datech, kterou lze vyjádřit jako

$$|\overline{O\mathbf{h}_j}|^2 \approx \mathrm{var}\left(\ln\frac{x_j}{\mathrm{g}(\mathbf{x})}\right), \quad |\overline{\mathbf{h}_j\mathbf{h}_k}|^2 \approx \mathrm{var}\left(\ln\frac{x_j}{x_k}\right),$$

kde $\mathrm{g}(\mathbf{x})$ je definováno ve vztahu (1.9). Nicméně, interpretace paprsku vzhledem ke konkrétní proměnné není jednoduchá, protože je vztažena ke všem složkám kompozice skrze $\mathrm{g}(\mathbf{x})$. Spojnice navíc poskytují informaci o korelaci mezi podkompozicemi, tj. pokud se spojnice $\overline{\mathbf{h}_j\mathbf{h}_k}$ a $\overline{\mathbf{h}_i\mathbf{h}_l}$ protínají v bodě M, pak platí

$$\cos(\mathbf{h}_j M \mathbf{h}_i) \approx \mathrm{cor}\left(\ln\frac{x_j}{x_k}, \ln\frac{x_i}{x_l}\right).$$

Pokud jsou tedy některé dvě spojnice na sebe kolmé, můžeme očekávat nulovou hodnotu korelace mezi příslušnými logaritmy podílu složek, čehož lze následně využít při zkoumání podkompozic pro případnou nezávislost. Dále čím je spojnice mezi paprsky kratší, tím je mezi příslušnými složkami kompozice silnější proporcionalita. Grafická interpretace bude později podrobněji ilustrována na konkrétním příkladu.

Dále budou zapotřebí charakteristiky, které umožňují posouzení kvality nahrazení nul vzhledem k původním datům. Relativní rozdíl variančních matic

(z angl. relative difference in covariance matrix, RDCM) je definovaný jako

$$\mathrm{RDCM} = \frac{\|\mathbf{S} - \mathbf{S}^*\|_F}{\|\mathbf{S}\|_F} = \frac{\sqrt{\sum_{i=1}^{D-1}\sum_{j=1}^{D-1}(s_{ij} - s_{ij}^*)^2}}{\sqrt{\sum_{i=1}^{D-1}\sum_{j=1}^{D-1}s_{ij}^2}}, \tag{4.1}$$

kde matice $\mathbf{S}$ je výběrová varianční matice původní datové matice $\mathbf{X}$ vyjádřené v ilr souřadnicích, a $\mathbf{S}^*$ je výběrová varianční matice dat po nahrazení imputovaných nul $\mathbf{X}^*$ vyjádřené v ilr souřadnicích. RDCM posuzuje vliv imputace na varianční strukturu dat a optimálně nabývá nulové hodnoty, čerpáno z [7].

Dále využijeme pro posouzení narušení varianční struktury kompoziční chybovou odchylku (z angl. compositional error deviation, CED), která je dána vztahem

$$\mathrm{CED} = \frac{\frac{1}{n_M}\sum_{k\in M}\mathrm{d}_a(\mathbf{x}_k, \mathbf{x}_k^*)}{\max_{\mathbf{x}_i,\mathbf{x}_j\in\mathbf{X}}\mathrm{d}_a(\mathbf{x}_i, \mathbf{x}_j)}, \tag{4.2}$$

kde n_M je počet takových pozorování $\mathbf{x}_k$, které obsahují alespoň jednu nulu vzniklou zaokrouhlením a M je množina indexů, které určují tato pozorování. V čitateli se objevuje průměrná Aitchisonova vzdálenost mezi původními pozorováními a pozorováními po nahrazení nul, přičemž se berou v úvahu pouze taková, která obsahují alespoň jednu nulu. Jmenovatel pak zaručuje zamezení možného nežádoucího efektu způsobeného variabilitou dat.

4.1 LPdata

Data obsahují informace ze stratigrafického výzkumu oblasti Cerro Pelado ve Venezuele, konkrétně údaje o složení 96 vzorků hornin, které byly odebrány z tzv. úseku La Paloma. Na každém vzorku horniny bylo měřeno zastoupení 15ti prvků v jednotkách mikrogram na gram, více k výzkumu k dispozici v [8]. Jelikož předpokládáme jistou omezenou citlivost měřících přístrojů, bude k nulám v datech přistupováno jako k nulám vzniklým zaokrouhlením; data jsou k dispo-

zici v knihovně softwaru R `zCompositions` pod názvem `LPdata`. Vybraná část datového souboru $\mathbf{X}$ pak vypadá následovně:

```
> LPdata[c(1:3,91:93),]
     Cr  B   P   V   Cu    Ti   Ni  Y  Sr La  Ce  Ba  Li     K  Rb
1  33.3 23 393  47  5.3  3715  9.1 24  38  9 180  48  76 16617  53
2  64.7 35 978  83  9.1  4215 19.8 21  93 19 259  93  95 16527  64
3  30.4 23 433  42  3.8  3305 16.6 22  59 14 240  75  80 12209  53
91 11.2 14 105  35  0.0   884  0.0  3   5  0  37  10 224 30511 167
92  4.2  4  39  21  0.0  2005  0.0 12  33 10 139  45  86 20106  83
93 22.1 11 186  59  6.4  2525 15.5 10  38  7 197  74 236 33909 195
```

Za detekční limit bude považována nejmenší nenulová hodnota dané proměnné, tj. vektor detekčních limitů pro jednotlivé proměnné je:

```
> DL
 Cr   B    P   V  Cu    Ti  Ni   Y  Sr  La   Ce   Ba   Li     K   Rb
3.6 3.0 37.0 9.0 2.0 549.0 6.3 3.0 5.0 1.0 18.0 10.0 22.0 159.0 15.0
```

Nyní budou použity algoritmy z kapitol 2.1 a 2.2 pro nahrazení nul. Výstupem funkce, která provede neparametrické nahrazení nul, je datová matice s imputovanými nulami, kterou dále budeme značit $\mathbf{X}_N^*$. Funkce `multRepl` vyžaduje zadání vektoru detekčních limitů, označení hodnot, které mají být nahrazeny, a určení proporce detekčního limitu, která bude použita pro nahrazování. Konkrétně pak nahrazení bude provedeno příkazem

```
XN <- multRepl(X,label=0,DL,delta=2/3)
```

a zobrazíme-li si stejnou část dat, pak lze vidět nahrazené nuly.

```
> round(XN[c(1:3,91:93),],2)
     Cr  B   P   V    Cu    Ti   Ni  Y  Sr    La  Ce  Ba  Li     K  Rb
1  33.3 23 393  47  5.30  3715  9.1 24  38  9.00 180  48  76 16617  53
2  64.7 35 978  83  9.10  4215 19.8 21  93 19.00 259  93  95 16527  64
3  30.4 23 433  42  3.80  3305 16.6 22  59 14.00 240  75  80 12209  53
91 11.2 14 105  35  1.33   884  4.2  3   5  0.67  37  10 224 30511 167
92  4.2  4  39  21  1.33  2005  4.2 12  33 10.00 139  45  86 20106  83
93 22.1 11 186  59  6.40  2525 15.5 10  38  7.00 197  74 236 33909 195
```

Dle očekávání byly nuly, které náležely téže proměnné nahrazeny stejnou hodnotou, která je pod detekčním limitem. Tato imputace vyplývá z principu neparametrického nahrazení, tj. ze vztahu (2.1).

Dále využijeme funkcí `impRZalr` a `impRZilr`, kterými aplikujeme EM algoritmus vycházející z příslušného vyjádření datové matice v souřadnicích. Funkce `impRZalr` vyžaduje určení proměnné bez nul, která umožní prvotní vyjádření v alr souřadnicích, dále opět vektor detekčních limitů a navíc upřesnění typu regrese, kde je možné zvolit i robustní variantu. U funkce `impRZilr` jsou parametry obdobné.

```
> fitA <- impRZalr(X, pos=3,dl=DL[c(1:2,4:length(DL))],method="lm")
> fitI <- impRZilr(as.matrix(X),dl=DL,method="lm")
```

Ve výstupu obou funkcí najdeme počet potřebných iterací, indexy nahrazovaných hodnot a samozřejmě požadované matice s imputovanými hodnotami, které po řadě označme $\mathbf{X}_A^*$, $\mathbf{X}_I^*$.

```
> XA = fitA$xImp
> XI = fitI$x
```

Oproti funkci pro ilr-EM algoritmus obdržíme pomocí funkce `impRZalr` matici, jejíž složky jsou upraveny na jednotkový součet, a je proto vhodná další úprava matice $\mathbf{X}_A^*$. Původně nenulové hodnoty budou přenásobeny příslušným řádkovým součtem výchozí datové matice $\mathbf{X}$. Původně nulové hodnoty je třeba navíc vydělit odpovídajícím řádkovým součtem matice $\mathbf{X}_A^*$, kdy součet je proveden jen z původně nenulových složek daného pozorování. V softwaru lze pak tento přepočet učinit následovně:

```
> kappa=rowSums(X)
> XA1=matrix(,nrow=nrow(XA),ncol=ncol(XA))
> for (i in 1:nrow(XA)){
+ for (j in 1:ncol(XA)){
+ if(X[i,j]!=0){
+ XA1[i,j]=kappa[i]*XA[i,j]}
+ else{
```

```
+ indexy=which(X[i,]!=0)
+ XA1[i,j]=XA[i,j]*kappa[i]/sum(XA[i,indexy])}}}
> XA=XA1
```

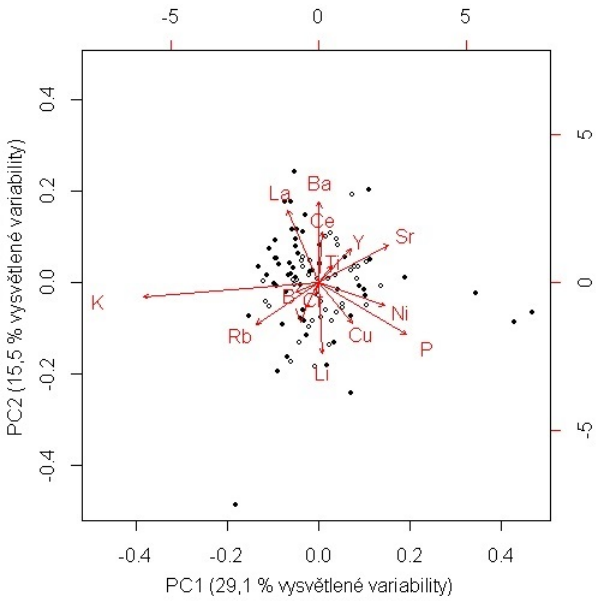
Celkově tedy dostaneme upravenou matici $\mathbf{X}_A^*$:

```
> round(XA[c(1:3,91:93),],2)
     Cr  B   P  V   Cu   Ti    Ni  Y Sr    La  Ce Ba  Li     K  Rb
1  33.3 23 393 47 5.30 3715  9.10 24 38  9.00 180 48  76 16617  53
2  64.7 35 978 83 9.10 4215 19.80 21 93 19.00 259 93  95 16527  64
3  30.4 23 433 42 3.80 3305 16.60 22 59 14.00 240 75  80 12209  53
91 11.2 14 105 35 1.40  884  2.98  3  5  0.89  37 10 224 30506 167
92  4.2  4  39 21 1.06 2005  3.10 12 33 10.00 139 45  86 20102  83
93 22.1 11 186 59 6.40 2525 15.50 10 38  7.00 197 74 236 33909 195
```

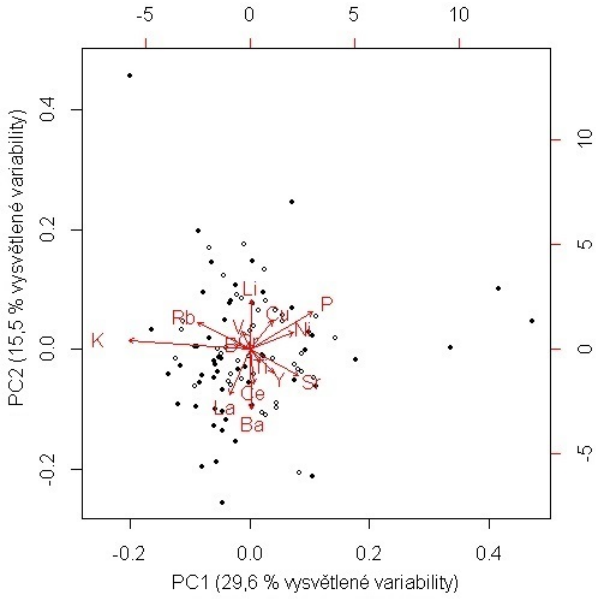
přičemž je zobrazena stejná část dat jako v předešlých případech. Z funkce pro ilr-EM algoritmus dostaneme jako přímý výstup matici $\mathbf{X}_I^*$, která nevyžaduje žádné další úpravy.

```
> round(XI[c(1:3,91:93),],2)
     Cr  B   P  V   Cu   Ti    Ni  Y Sr   La  Ce Ba  Li     K  Rb
1  33.3 23 393 47 5.30 3715  9.10 24 38  9.0 180 48  76 16617  53
2  64.7 35 978 83 9.10 4215 19.80 21 93 19.0 259 93  95 16527  64
3  30.4 23 433 42 3.80 3305 16.60 22 59 14.0 240 75  80 12209  53
91 11.2 14 105 35 1.45  884  3.29  3  5  0.9  37 10 224 30511 167
92  4.2  4  39 21 1.11 2005  3.24 12 33 10.0 139 45  86 20106  83
93 22.1 11 186 59 6.40 2525 15.50 10 38  7.0 197 74 236 33909 195
```

Na tomto místě nemá smysl porovnávat matice $\mathbf{X}_N^*$, $\mathbf{X}_A^*$ a $\mathbf{X}_I^*$ ani posuzovat kvalitu nahrazení pomocí vztahů (4.1), (4.2). Bude nás zde ale zajímat clr-biplot, pomocí kterého se podíváme na strukturu pozorování a proměnných a na polohu pozorování s nulami oproti těm bez nul. Grafy 4.1 a 4.2, zobrazují clr-biplot vycházející po řadě z matice $\mathbf{X}_N^*$ a $\mathbf{X}_I^*$, tj. z matice s nahrazenými nulami pomocí neparametrického nahrazení a ilr-EM algoritmu s klasickou regresí.



Graf 4.1: Clr-biplot matice $\mathbf{X}_N^*$



Graf 4.2: Clr-biplot matice $\mathbf{X}_I^*$

V grafu 4.1 lze vidět, že délka paprsku pro draslík (K) převyšuje délky ostatních proměnných, tedy lze se domnívat, že tato proměnná, resp. odpovídající clr

souřadnice, nejvíce ovlivňuje datový soubor. Velké hodnoty této proměnné pak značí dominanci draslíku vůči průměrné složce kompozice. Naopak nejkratší délky paprsků přísluší proměnným (prvkům) Ti, B a C. Lze si dále všímat, které paprsky mají podobnou orientaci v rámci clr-biplotu. V tomto smyslu je zřetelná skupina sestávající z prvků La, Ba a Ce nebo skupina z prvků Ti, Y a Sr. Podstatu v této podobné orientaci lze hledat skrze vlastnosti jednotlivých prvků, např. prvky La (lanthan) a Ce (cer) patří do skupiny lanthanoidů, nebo Ti (titan), Y (yttrium) a Sr (stroncium) patří mezi kovy s podobnými vlastnostmi ve smyslu počtu elektronů ve valenční vrstvě nebo protonového čísla.

V úvodu kapitoly byla pak při popisu konstrukce clr-biplotu zmíněna souvislost mezi kosinem úhlu dvou spojnic vrcholů a korelací mezi logaritmy podílu odpovídajících složek kompozice. Např. představíme-li si spojnici mezi vrcholem paprsku odpovídající proměnné K a proměnné Ni, bude tato spojnice téměř kolmá na spojnici vrcholů paprsků odpovídající proměnným Cu a Y.

Graf 4.2 je oproti grafu 4.1 opačně orientován a paprsek příslušný proměnné K dominuje svoji délkou o něco méně, nicméně podobnost v orientaci proměnných zůstává zachována. Oba grafy dále naznačují, že některá pozorování s nulami (plný kroužek), jsou zřetelně na okraji od ostatních pozorování. Při takovém výsledku lze pak uvažovat o možných odlehlých pozorováních, popř. zvážit použití robustních metod. Navíc lze při porovnání clr-biplotů vidět větší variabilitu vykreslených bodů v grafu 4.2, což je způsobeno povahou metody neparametrického nahrazení, která podceňuje variabilitu kompozičních dat.

4.2 Chorizon

Data byla získána v rámci geochemického mapování půdy na poloostrově Kola v letech 1992 až 1998. Bylo odebráno přes 600 vzorků ze čtyřech různých vrstev půdy, přičemž tato data obsahují údaje o složení vzorků z vrstvy C-horizon. Pro analýzu bude dále vybráno deset hlavních prvků, celá data jsou pak k dispozici v knihovně softwaru R `mvoutlier` pod názvem `chorizon`. Datovou matici s vybranými proměnnými dále budeme značit $\mathbf{X}$. Prvních pár řádků matice vypadá

následovně:

```
> X[1:5,]
     Al   Ca    Fe    K   Mg    Mn  Na   P  Si   Ti
1 10200 2280 15000 1300 4560 170.0 140 324 140  949
2  3540 1720  8530  600 1560  59.3 140 428 110  351
3 16100 2240 18200 3000 5780 102.0 300 631 180 1270
4 12000 3300 10500 1400 4170 102.0 510 359 190  716
5  9850 1120 20900 2800 3280 146.0  40 580 110  754
```

V tomto případě se jedná o data bez nul, které budou imputovány simulačně. Celkem budou vytvořeny čtyři scénáře se stoupajícím počtem nul. Pro každou z proměnných Mn, Na, P, Si a Ti určíme 0,1-kvantil, 0,35-kvantil, 0,5-kvantil a 0,65-kvantil, přičemž hodnoty pod příslušným kvantilem budou změněny na nulové, ostatní proměnné zůstanou bez nul. Pro každý scénář použijeme neparametrické nahrazení a ilr-EM algoritmus s klasickou regresí a následně kvalitu nahrazení porovnáme pomocí RDCM a CED.

Z prvního scénáře, který využívá 0,1-kvantilu vybraných proměnných, dostaneme data obsahující necelých 5 % nul. Vektor 0,1-kvantilů získáme jednoduše:

```
> kvantil1=sapply(X,quantile,0.1)
> names(kvantil1)=colnames(X)
> kvantil1
     Al   Ca     Fe     K     Mg    Mn   Na     P    Si    Ti
4895.0 735.0 7620.0 500.0 1770.0 68.85 60.0 189.0 100.0 382.5
```

Dále provedeme imputaci nul, tj. ve vybraných proměnných budou hodnoty, které jsou menší než příslušný 0,1-kvantil, nahrazeny nulou, takto obdrženou matici označíme $\mathbf{X}_1$.

```
> X1=X
> for(i in 1:606){
+ for(j in 6:10){
+ if (X1[i,j]<kvantil1[j]) {
+ X1[i,j]=0} else {X1[i,j]=X1[i,j]}}}
```

V prvních pár řádcích matice $\mathbf{X}_1$ pak lze vidět imputované nuly.

```
> X1[1:5,]
     Al   Ca    Fe    K   Mg  Mn  Na   P  Si   Ti
1 10200 2280 15000 1300 4560 170 140 324 140  949
2  3540 1720  8530  600 1560   0 140 428 110    0
3 16100 2240 18200 3000 5780 102 300 631 180 1270
4 12000 3300 10500 1400 4170 102 510 359 190  716
5  9850 1120 20900 2800 3280 146   0 580 110  754
```

Ještě zvolíme vektor detekčních limitů,

```
> DL1=c(sapply(X1[,1:5],min),kvantil1[6:10])
> DL1
    Al    Ca     Fe     K    Mg    Mn   Na     P    Si    Ti
1840.0 110.0 3310.0 100.0 370.0 68.85 60.0 189.0 100.0 382.5
```

tzn. pro prvních pět proměnných uvažujeme nejmenší hodnotu, kterou daná proměnná nabývá, pro druhých pět proměnných příslušný 0,1-kvantil.

Nyní už použijeme známé funkce a dostaneme se k maticím s nahrazenými nulami, přičemž označme $\mathbf{X}^*_{1_N}$ matici obdrženou prostřednictvím neparametrického nahrazení a $\mathbf{X}^*_{1_I}$ matici, kterou jsme dostali z ilr-EM algoritmu s klasickou regresí.

```
> X1N <- multRepl(X1,label=0,DL1,delta=2/3)
> fit1I <- impRZilr(as.matrix(X1),dl=DL1,method="lm")
> X1I=fit1I$x
```

Další scénáře budou vytvořeny zcela analogicky. Opět imputujeme nuly do druhé pětice proměnných a prvních pět necháváme bez nul. Uvažujeme po řadě 0,35-kvantil, 0,5-kvantil a 0,65-kvantil, díky kterým dostaneme data se 17 %, 25 % a 32 % nul, na které opět použijeme neparametrické nahrazení a ilr-EM algoritmus s klasickou regresí. Dostaneme tak pro každý ze čtyř scénářů dvě matice s imputovanými nulami.

Nyní využijeme vztahů (4.1) a (4.2) a porovnáme použité algoritmy. Pro určení hodnoty charakteristiky RDCM uvažujeme vždy původní datovou matici bez nul $\mathbf{X}$ a společně s ní do vzorce postupně vstupují matice, které v jednotlivých scénářích vznikly jako výstupy algoritmů nahrazení, tj. matice $\mathbf{X}^*_{k_N}$ a $\mathbf{X}^*_{k_I}$, $k = 1, 2, 3, 4$. Celkem tak dostaneme pro každý scénář dvě hodnoty RDCM, které

mezi sebou porovnáme a menší hodnota této charakteristiky bude znamenat větší kvalitu nahrazení, resp. menší narušení varianční struktury. Funkce RDCM není v softwaru R zabudovaná, a tak si ji lze naprogramovat třeba tímto způsobem:

```
> RDCM <- function(X,Y)
+ {norm(var(X)-var(Y),type="F")/norm(var(X),type="F")}
```

Je nutné ještě připomenout, že varianční matice se určují z matic, které jsou vyjádřeny v ilr souřadnicích, tzn. jako vstupy do funkce RDCM je nutné použít datové matice vyjádřené v ilr souřadnicích.

Naprogramování funkce CED nejprve vyžaduje výpočet jmenovatele, který je sám o sobě početně náročný, viz vztah (4.2). Naštěstí je tato hodnota stále stejná, neboť pro všechny scénáře hledáme maximální hodnotu Aitchisonovy vzdálenosti mezi pozorováními původní matice $\mathbf{X}$. Pro uvažovanou datovou matici $\mathbf{X}$ platí

$$\max_{\mathbf{x}_i,\mathbf{x}_j\in\mathbf{X}} \mathrm{d}_a(\mathbf{x}_i,\mathbf{x}_j) = 8{,}925\,012,$$

a hodnotu dále využijeme pro vlastní definování funkce CED.

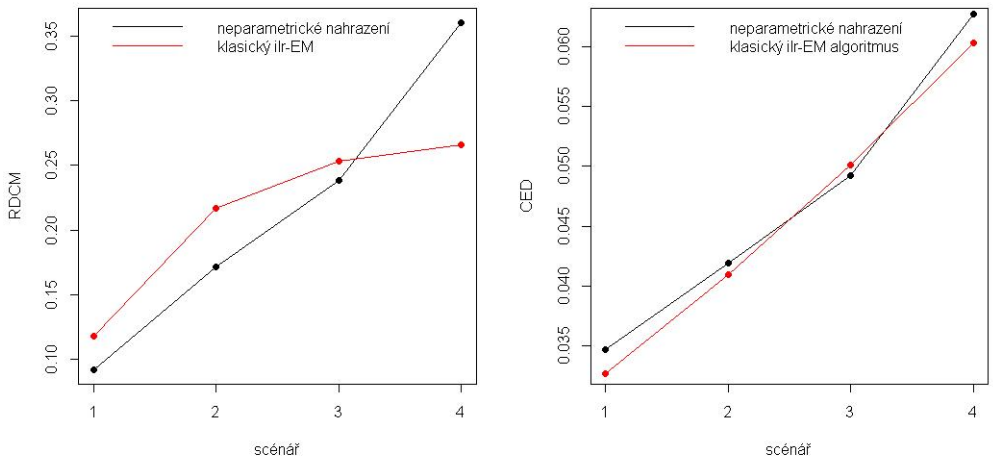
```
> CED <- function(X,Xk,XkN){
+ nM=nrow(X)-sum(0<apply(Xk,1,prod))
+ index=sort(unique(row(Xk)[which(Xk==0)]))
+ s=0
+ for (i in index){
+ a=aDist(X[i,],XkN[i,])
+ s=s+a}
+ ((1/nM)*s)/maxd}
```

Na vstupu funkce CED je potřeba zadat po řadě původní matici bez nul, matici s nulami a matici s nahrazenými nulami. Než budou uvedeny konkrétní výsledky pro všechny čtyři scénáře, je nutné dodat, že naprogramované funkce RDCM a CED jsou výhradně pro potřebu tohoto příkladu, resp. aby funkce fungovaly pro všechny možné jiné příklady, pak by bylo potřeba komplexnějšího kódu, který by zahrnoval potenciální chybová hlášení apod.

Konečně se dostáváme k porovnání charakteristik RDCM a CED pro jednotlivé scénáře a algoritmy nahrazení, které poskytuje tabulka 1 a graf 4.3.

	nep. nahrazení		klasický ilr-EM	
scénář	RDCM	CED	RDCM	CED
$\mathbf{X}_1$	0,092	0,035	0,118	0,033
$\mathbf{X}_2$	0,172	0,042	0,217	0,041
$\mathbf{X}_3$	0,238	0,049	0,253	0,050
$\mathbf{X}_4$	0,360	0,063	0,266	0,060

Tabulka 1: RDCM a CED pro jednotlivé scénáře



Graf 4.3: porovnání RDCM (vlevo) a CED (vpravo)

Co se týče charakteristiky RDCM, pak lze u neparametrického nahrazení vidět rychlejší nárůst než u klasického ilr-EM algoritmu. Neparametrické nahrazení při menším počtu nul poskytlo lepší nahrazení, resp. menší narušení varianční struktury. Na druhou stranu se zvětšujícím se podílem nul hodnota RDCM pro klasický ilr-EM algoritmus roste výrazně pomaleji, což lze vidět na grafu 4.3 (vlevo). Charakteristika CED pro oba algoritmy s přibývajícími nulami roste poměrně stejně, viz graf 4.3 (vpravo). Zároveň hodnoty CED pro neparametrické nahrazení a klasický ilr-EM algoritmus se liší v rámci scénáře velmi málo. Avšak ve třech ze čtyř scénářů je dle CED o trochu lepší klasický ilr-EM algoritmus.

4.3 Simulační studie

V tomto příkladu se zaměříme na porovnání klasického a robustního ilr-EM algoritmu. Data tentokrát získáme tak, že pomocí funkce `mvrnorm` z knihovny `MASS` vygenerujeme hodnoty z mnohorozměrného normálního rozdělení s následující střední hodnotou a varianční maticí

$$\boldsymbol{\mu} = (0,\, 0,\, 0,\, 0,\, 0)', \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & & & & \\ & 1 & & 0{,}2 & \\ & & 1 & & \\ & 0{,}2 & & 1 & \\ & & & & 1 \end{pmatrix}.$$

Vygenerujeme celkem 500 pozorování a dostaneme tak matici s rozměry 500×5. Tato matice bude představovat datovou matici vyjádřenou v ilr souřadnicích, kterou pomocí funkce `ilrInv` z knihovny `compositions` převedeme zpět na simplex, a dostaneme tak matici s rozměry 500×6 představující kompoziční datový soubor $\mathbf{X}$. Výše popsaný postup získání dat provedeme v softwaru R následujícím způsobem:

```
> Xilr=mvrnorm(500,mu,sigma)
> X=ilrInv(datailr)
```

Dále budeme postupovat analogicky jako v předešlém příkladu a nuly do dat imputujeme simulačně. Nyní vytvoříme pět scénářů, přičemž pro každý vznikne jedna datová matice s nulami, kterou označme $\mathbf{X}_k, k = 1, ..., 5$. V každém scénáři pro první tři proměnné nalezneme odpovídající kvantil a hodnoty pod příslušným kvantilem budou změněny na nulové, do druhé trojice proměnných nuly imputovány nebudou. Charakteristiky jednotlivých scénářů jsou uvedeny v tabulce 2.

scénář	kvantil	podíl nul (%)
$\mathbf{X}_1$	0,1	5
$\mathbf{X}_2$	0,2	10
$\mathbf{X}_3$	0,3	15
$\mathbf{X}_4$	0,4	20
$\mathbf{X}_5$	0,5	25

Tabulka 2: Charakteristika vytvořených scénářů

Na všechny datové matice s nulami nyní použijeme klasický i robustní ilr-EM algoritmus. Protože je zavedení vektorů detekčních limitů analogické předešlému příkladu, bude zde ukázáno pouze použití robustního ilr-EM algoritmu, který dosud nebyl demonstrován.

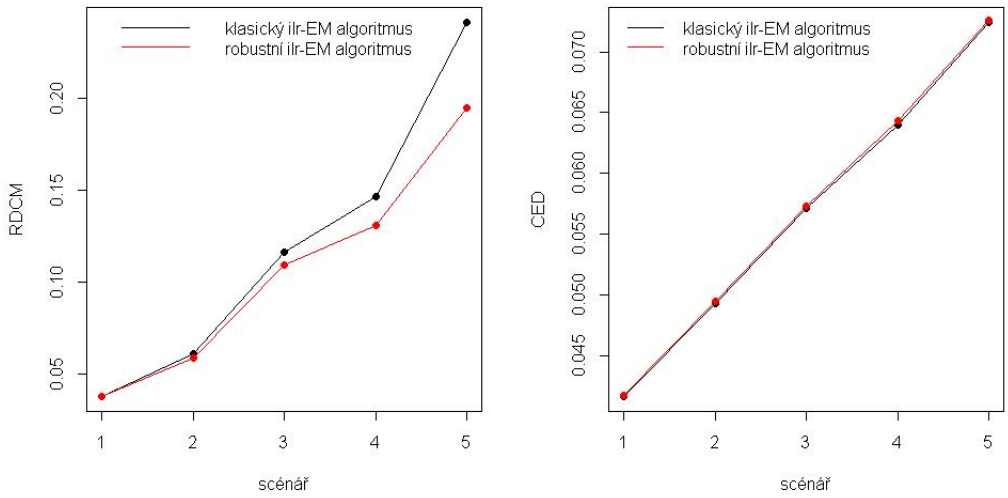
```
fitR <- impRZilr(as.matrix(X1),dl=DL1,method="MM")
```

Příkazem jsme provedli nahrazení nul v prvním scénáři, kde ve funkci `impRZilr` je potřeba nastavit odpovídající parametr pro použití robustní regrese.

Postupně opět získáme pro každý scénář dvě matice s nahrazenými nulami a přistoupíme k porovnání kvality nahrazení pomocí charakteristik RDCM a CED, viz tabulka 3 a graf 4.4.

	klasický ilr-EM		robustní ilr-EM	
scénář	RDCM	CED	RDCM	CED
$\mathbf{X}_1$	0,0379	0,0417	0,0376	0,0417
$\mathbf{X}_2$	0,0608	0,0493	0,0585	0,0494
$\mathbf{X}_3$	0,1163	0,0571	0,1092	0,0573
$\mathbf{X}_4$	0,1465	0,0639	0,1306	0,0643
$\mathbf{X}_5$	0,2410	0,0724	0,1949	0,0726

Tabulka 3: RDCM a CED pro jednotlivé scénáře

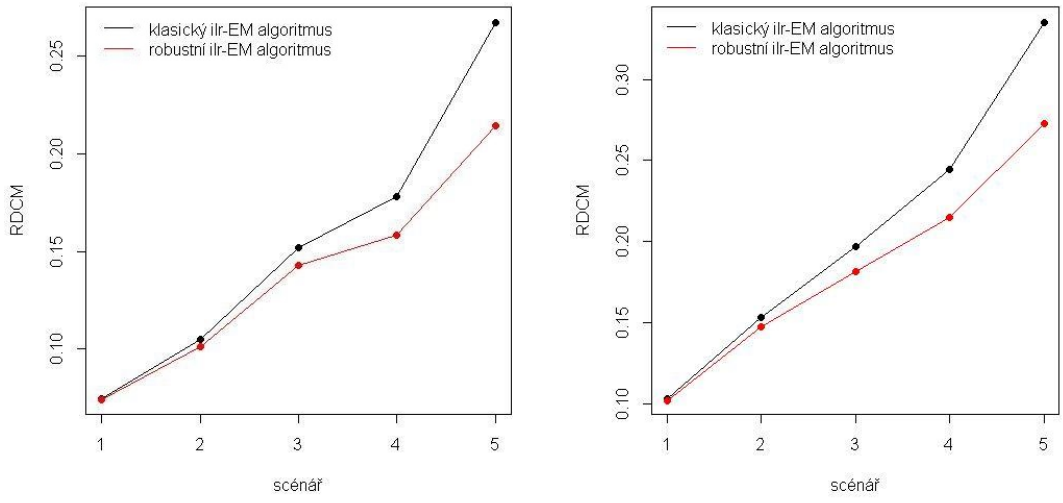


Graf 4.4: porovnání RDCM (vlevo) a CED (vpravo)

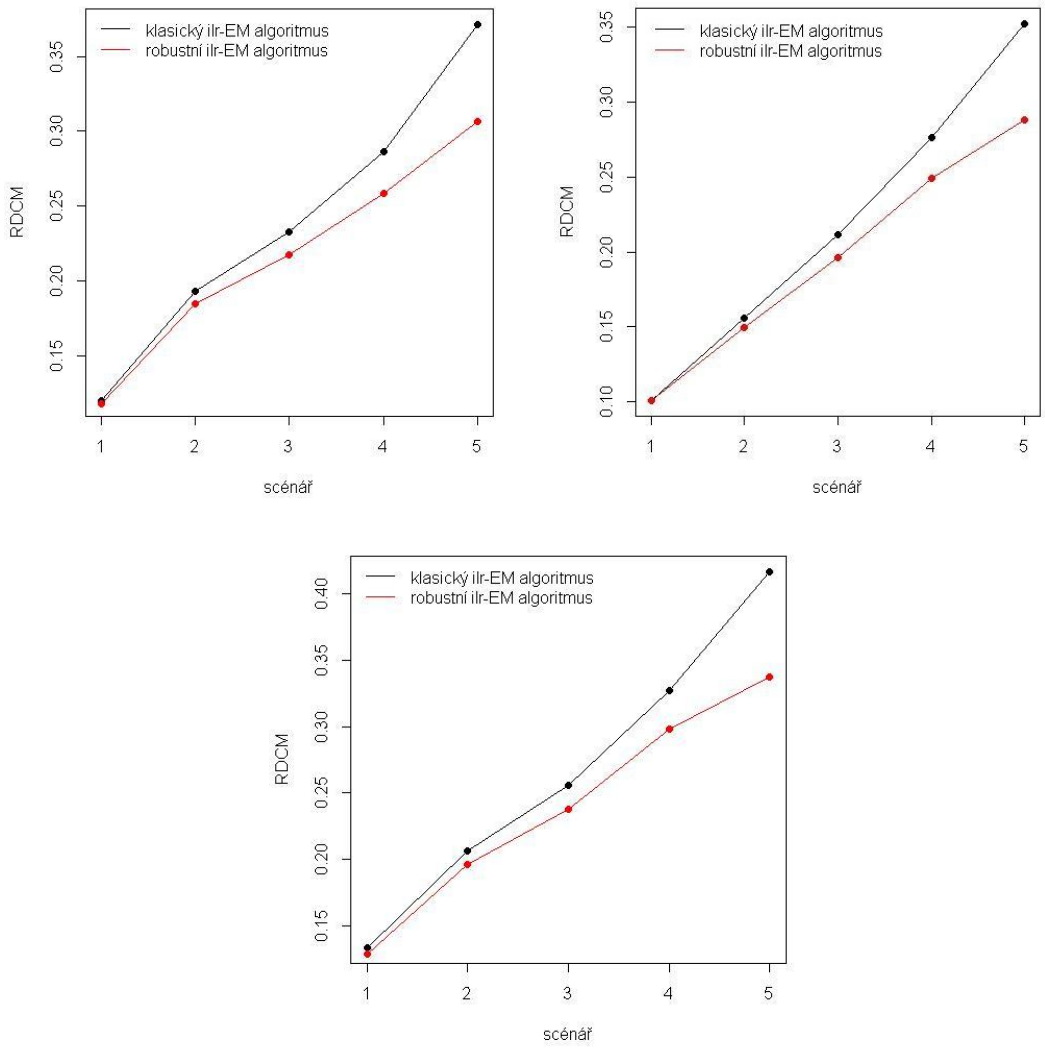
Z grafu 4.4 (vlevo) je patrné, že robustní ilr-EM algoritmus dle charakteristiky RDCM poskytuje lepší kvalitu nahrazení, resp. méně narušuje varianční strukturu dat. Navíc s přibývajícím podílem nul v datech se rozdíl mezi algoritmy zvětšuje. Oproti tomu charakteristika CED vykreslená v grafu 4.4 (vpravo) nenaznačuje výrazný rozdíl mezi robustním a klasickým ilr-EM algoritmem.

Rozdíl ve kvalitě nahrazení by se měl více projevit, pokud máme data obsahující odlehlá pozorování. U takového datového souboru lze očekávat, že robustní ilr-EM algoritmus poskytne lepší nahrazení než klasický. Proto nyní do datového souboru odlehlá pozorování přidáme, resp. budeme postupně zvyšovat jejich počet v jednotlivých scénářích, přičemž celkový počet pozorování, tj. $n = 500$, zůstane zachován. Do dat postupně přidáme 10, 20, 30, 40 a 50 odlehlých pozorování, která budou vygenerována z mnohorozměrného normálního rozdělení se střední hodnotou $\boldsymbol{\mu} = (0, 4, 0, 0, 0)'$ a se stejnou varianční maticí, která přísluší datové sadě v ilr souřadnicích. Dostaneme tak celkem pět nových datových sad s různým počtem odlehlých hodnot, přičemž pro každou datovou sadu se při tvorbě scénářů, tj. při imputaci nul, řídíme tabulkou 2.

Grafy 4.5 a 4.6 poskytují porovnání charakteristiky RDCM pro klasický a robustní ilr-EM algoritmus.



Graf 4.5: RDCM pro data s 10 (vlevo) a 20 (vpravo) odlehlými hodnotami

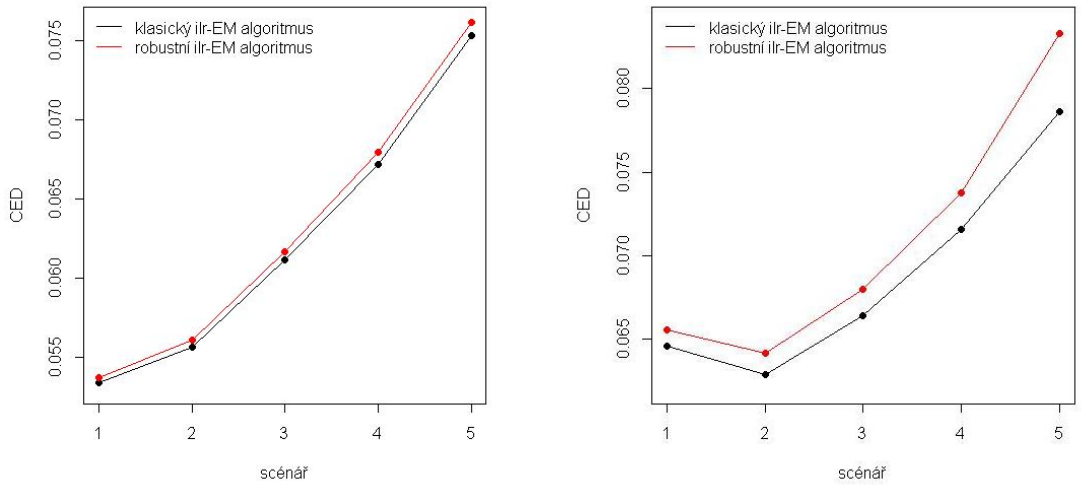


Graf 4.6: RDCM pro data s 30 (nahoře vlevo), 40 (nahoře vpravo), 50 (dole) odlehlými hodnotami

Každý graf přísluší analýze právě jedné datové sady s určitým počtem odlehlých pozorování, ve které byl postupně zvyšován podíl nul. Z grafů 4.5 a 4.6 vyplývá, že robustní ilr-EM algoritmus poskytuje podle charakteristiky RDCM lepší nahrazení. Ve všech scénářích byla hodnota RDCM pro robustní verzi algoritmu menší než pro algoritmus s klasickou regresí.

Nyní ukážeme srovnání algoritmů dle CED pouze pro data s 10 a 50 odlehlými hodnotami, protože pro ostatní datové sady dostáváme velmi podobná porovnání

algoritmů, viz graf 4.7.



Graf 4.7: CED pro data s 10 (vlevo) a 50 (vpravo) odlehlými hodnotami

Podle charakteristiky CED hodnotíme jako lepší klasický ilr-EM algoritmus, i když rozdíly jsou mnohem menší než u charakteristiky RDCM, a navíc se nezvětšují se zvyšujícím se podílem nul. Charakteristiky RDCM a CED nejsou samozřejmě srovnatelné, neboť každá se dívá na kvalitu nahrazení jinak, viz vztahy (4.1) a (4.2).

Závěrem lze tedy říci, že pokud bychom se řídili hlavně charakteristikou RDCM, která posuzuje vliv imputace na varianční strukturu dat, pak bychom vždy volili robustní ilr-EM algoritmus, i když pro malý podíl nul v datech a pro malý počet odlehlých pozorování poskytuje klasický ilr-EM algoritmus srovnatelně kvalitní nahrazení nul.

4.4 SHIW data

Data pochází z výzkumu, který byl proveden italskou bankou Banca d'Italia, a obsahuje informace o příjmech domácností v eurech z roku 2008, data jsou k dispozici v [17]. V jednotlivých domácnostech byla zjišťována např. výše mzdy, důchodů, popř. výše stipendií, výživného, příjmů z nemovitostí apod. Vzhledem k četnému výskytu nul byly proměnné sloučeny celkem do čtyř proměnných.

První proměnná YL obsahuje souhrnné informace o výši mezd, další proměnná YT udává výši důchodů, proměnná YM pak obsahuje údaje o příjmech osob samostatně výdělečně činných a proměnná YC popisuje příjmy z nemovitostí a jiných majetků. Dále z dat vyloučíme ty řádky, které obsahují záporná pozorování, a řádky s více než dvěma nulami. Dostaneme se tak celkem na 7247 pozorování. Datová sada pak obsahuje další proměnné, např. pohlaví respondenta, úroveň vzdělání apod. Prozatím si v datech necháme kvůli identifikaci pouze proměnnou s číslem dotazníku (nquest), která nám respondenta jednoznačně určuje, a pomocí které pak později dohledáme detailnější informace. Prvních pár řádků dat po úpravě vypadá následovně:

```
> head(data)
  nquest    YL    YT YM        YC
1    173     0 26000  0  8207.001
2    375     0  4860  0  7200.000
3    465 49500     0  0 10708.841
4    629 14500 17550  0  6261.051
5    632 16000 17550  0 16883.153
6    633     0 13260  0  5917.403
```

Když se zamyslíme nad významem nul v datech, pak např. nuly v první a třetí proměnné naznačují, že se nejedná o člověka v produktivním věku, ale o člověka pobírající důchod. Naopak nula v druhé a třetí proměnné napovídá o člověku, který pracuje jako zaměstnanec, ale není osobou samostatně výdělečně činnou.

V prvním kroku veškeré nuly v datech považujeme za chybějící hodnoty a pomocí funkce `impCoda` z balíčku `robCompositions` provedeme jejich nahrazení, přičemž první proměnnou vynecháváme, neboť ta slouží pouze pro identifikaci pozorování.

```
> X=data[,c(2:5)]
> X[X==0] <- NA
> fitNA <- impCoda(X,closed=TRUE)
> Ximp=fitNA$xImp
```

Dostaneme matici $\mathbf{X}_{imp}$ s imputovanými hodnotami, kterou nyní můžeme vyjádřit

v ilr souřadnicích, takovou matici označíme **Z**. Následně vypočteme robustní odhady polohy a variability **t** a **C** pomocí metody MCD.

```
> rob <- covMcd(Z)
> t=rob$center
> C=rob$cov
```

Dále budeme potřebovat matici **V**, která je definovaná v (3.5). V softwaru ji sestavíme následujícím způsobem,

```
> V=matrix(,nrow=D,ncol=(D-1))
> for(j in 1:(D-1)){
+ V[,j]=sqrt((D-j)/(D-j+1))*c(rep(0,j-1),1,rep(-1/(D-j),D-j))}
```

kde D je počet proměnných, v tomto případě $D = 4$.

Výše definované objekty budou dále výchozími pro výpočet charakteristik, které budou příslušet už konkrétnímu pozorování. Pro využití informace odpovídající nenulovým složkám pozorování je potřeba nejprve sestavit vhodnou permutaci dané kompozice. Dle kapitoly 3.1.2 ji vytvoříme tak, že strukturní nuly budou přesunuty na začátek a ostatní složky budou zařazeny za ně bez změny pořadí. K tomu je nutné si zadefinovat vlastní funkci.

```
> PresunNul<-function(x){
+ xp=rep(1,length(x))
+ index=which(x==0)
+ if(length(index)==0){xp=x}else{
+ xp[1:length(index)]=x[index]
+ xp=unlist(xp)}
+
+ index=which(xp!=0)
+ index2=which(x!=0)
+ xp[index]=x[index2]
+ if(class(xp)=="list")
+ {xp=unlist(xp)
+ xp}else{xp}}
```

Samotný algoritmus pak postupně využívá vztahů z kapitoly 3.1.2 a výstupem bude vektor s hodnotami Mahalanobisových vzdáleností pro každé pozorování, které budou vypočteny podle (3.6). Algoritmus odpovídající první fázi dvoustupňového algoritmu pro detekci odlehlých hodnot pro data se strukturními nulami lze tedy sestavit tímto způsobem:

```
> MD=rep(0,n)
> kvantily=rep(0,n)
> for(i in 1:n){
+ x=data[i,c(2:5)]
+ z=Z[i,]
+ xv=PresunNul(x)
+ xiv=c(which(x==0),which(x!=0))
+ P=as(as.integer(xiv), "pMatrix")
+ Vv=P%*%V
+ Q=t(V)%*%Vv
+ zv=t(Q)%*%z
+ tv=t(Q)%*%t
+ Cv=t(Q)%*%C%*%Q
+ K=length(which(x!=0))
+ iz=c((D-K+1):length(zv))
+ it=c((D-K+1):length(tv))
+ ic1=c((D-K+1):nrow(Cv))
+ ic2=c((D-K+1):ncol(Cv))
+ MD[i]=sqrt(t(zv[iz]-tv[it])%*%solve(Cv[ic1,ic2])%*%(zv[iz]-tv[it]))
+ kvantily[i]=qchisq(0.975,K-1)}
```

Vektor **MD** obsahuje celkem 7247 hodnot, přičemž každá odpovídá jednomu pozorování. Zároveň byl zavedený vektor s odpovídajícími kvantily χ^2 rozdělení, protože stupně volnosti jsou závislé na počtu nenulových složek v daném pozorování. Pozorování pak označíme za odlehlé, jestliže platí $\mathrm{MD}_i^2 > \chi^2_{K(i)-1,0,975}$. V tomto případě nám algoritmus označil jako odlehlé hodnoty přibližně 23 % pozorování.

Otázkou nyní je, jak výsledky interpretovat. Jedním z možných náhledů, je zkoumat označená odlehlá pozorování vzhledem k některé jiné proměnné.

Konkrétně zkusíme využít proměnnou, která popisuje dosažené vzdělání respondenta, a podíváme se, jak jsou na tom jednotlivé skupiny co se týče odlehlých pozorování. Je nutné si uvědomit, že proměnná koresponduje se vzdělávacím systémem Itálie, který ve zkratce popíšeme, čerpáno z [18]. Tzv. primární školu navštěvují děti celkem pět let do svých 11-ti let, následuje první stupeň střední školy, kam děti docházejí do svých 13-ti let. Následně se skládá státní zkouška, která je nutná pro postup do druhého stupně střední školy, kde si žáci volí buď gymnázium nebo obdobu odborných škol a učilišť. Dále se skládá obdoba maturity, která je podmínkou pro postup k univerzitnímu vzdělání. Vysokoškolské vzdělání je pak analogické s ČR.

Hodnoty uvažované proměnné mají tedy následující interpretaci: 1 – žádné vzdělání, 2 – primární vzdělání, 3 – 1. stupeň SŠ, 4 – 2. stupeň SŠ s odborným zaměřením, 5 – 2. stupeň SŠ – gymnázium, 6 – tříleté VŠ vzdělání, 7 – pětileté VŠ vzdělání, 8 – postgraduální studium.

vzdělání	1	2	3	4	5	6	7	8	celkem
počet pozorování	367	1912	2037	483	1748	50	600	50	7247
podíl odlehlých hodnot (%)	15,8	17,5	29,2	26,5	22,8	28,0	26,5	26	23,5

Tabulka 4: Výskyt odlehlých hodnot ve skupinách dle dosaženého vzdělání

Z tabulky 4 je patrné, že mezi respondenty bylo nejvíce těch, kteří dosáhli 1. stupně SŠ. V této skupině bylo rovněž procentuálně nejvíce odlehlých hodnot. Naopak poměrně stejně početná skupina respondentů s dosaženým primárním vzděláním měla podíl odlehlých hodnot jeden z nejmenších. Podobně jako v [9] by bylo možné nahlížet na podíl odlehlých hodnot jako na míru jakési konzistence, a tedy nepovažovat odlehlá pozorování výhradně za špatný signál o kvalitě dat. Například větší podíl odlehlých pozorování ve skupinách respondentů s vyšším dosaženým vzděláním by mohl naznačit výraznější rozdíly v příjmech. Naopak menší podíl by mohl napovídat o větší podobnosti příjmů respondentů v rámci skupiny. Početné množství odlehlých hodnot ve skupině by dále mohlo být známkou o přítomnosti

nějaké subpopulace. Podobně bychom dále mohli zkoumat zastoupení odlehlých hodnot vzhledem ke kterékoli jiné proměnné.

Podívejme se nyní na dosažené výsledky trochu jinak. Budeme uvažovat postupně zvlášť pozorování, která mají nulu pouze v první proměnné, zvlášť pozorování s nulou pouze v druhé proměnné atd. Pro každou konfiguraci nul pak uvažujme pouze ta pozorování, která mají zbývající složky nenulové. Tzn. mámeli pozorování, která mají nulu pouze v proměnné YL, pak uvažujeme jen hodnoty z proměnných YT, YM, YC, a z těchto hodnot vypočteme robustní Mahalanobisovy vzdálenosti. Cílem je porovnat Mahalanobisovy vzdálenosti, které jsme dostali s použitím imputace pro určení odhadů střední hodnoty a varianční matice, s Mahalanobisovými vzdálenostmi vypočtenými z jednotlivých podsouborů dle struktury nul.

Nejprve se podívejme, kolik pozorování bylo označeno za odlehlé hodnoty pro různé konfigurace nul, přičemž označení pořadí proměnné, ve které se nachází nula, respektuje pořadí YL, YT, YM, YC.

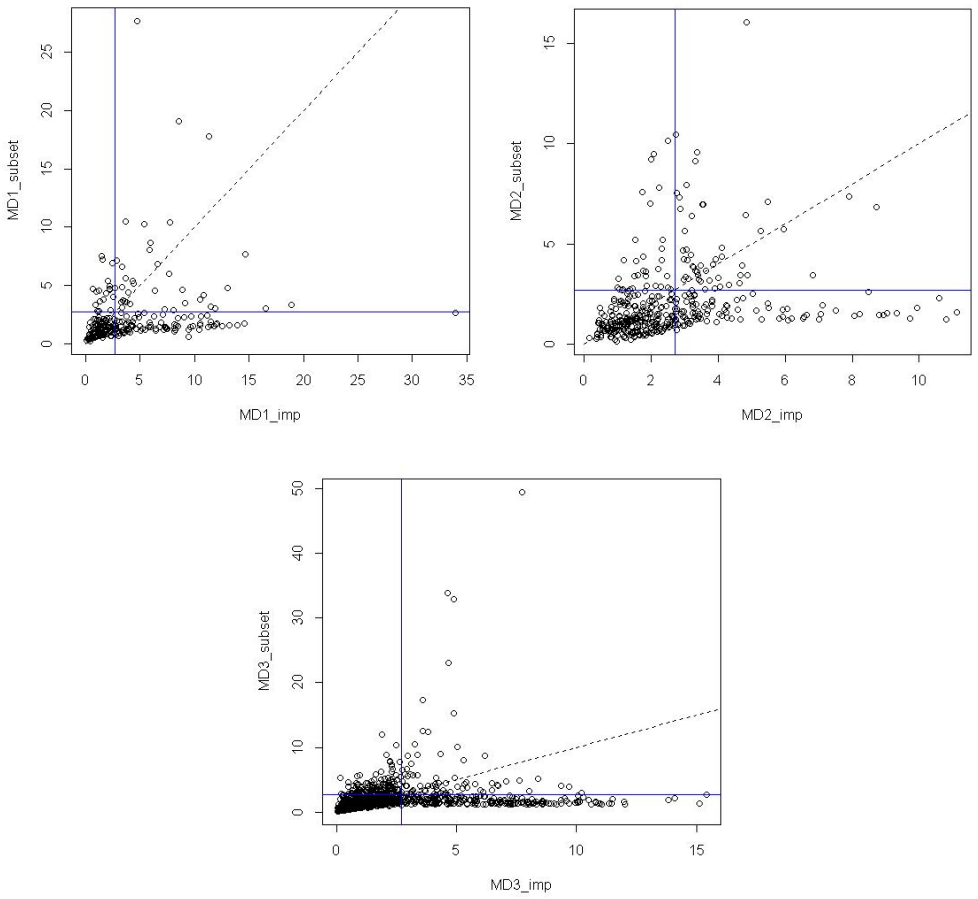
umístění nul	bez nul	1	2	3	4	1, 2	1, 3	1, 4	2, 3	2, 4	3, 4
počet pozorování	162	252	395	1109	5	344	2971	1	1941	3	64
počet odlehlých hodnot (imputace)	110	102	259	285	0	79	375	0	464	2	22
počet odlehlých hodnot (podsoubory)	38	55	82	197	NA	68	479	NA	276	NA	8

Tabulka 5: Srovnání počtů detekovaných odlehlých pozorování

V tabulce 5 lze vidět, že většinou algoritmus zahrnující imputaci strukturních nul detekoval více odlehlých hodnot. V nejpočetnější skupině respondentů s nulovými proměnnými YL a YM, tj. nedostávající příjem ze zaměstnání ani jako osoby samostatně výdělečně činné, však algoritmus odhalil méně odlehlých pozorování v porovnání s druhým přístupem. Ovšem je nutné si uvědomit, že počty odlehlých pozorování neříkají, zda oba přístupy detekovaly stejná odlehlá pozorování. Pro některé kombinace nulových hodnot bylo navíc příliš málo pozorování

v příslušných podsouborech na to, aby mohly být Mahalanobisovy vzdálenosti vypočteny.

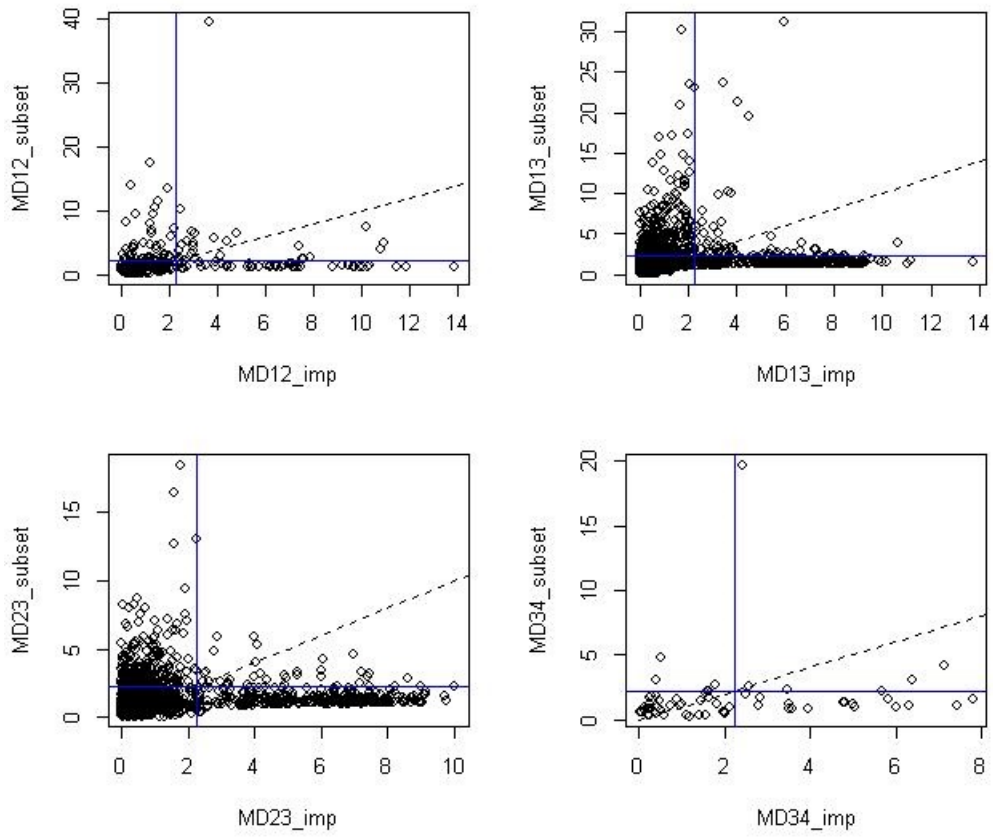
V grafu 4.8 jsou porovnány Mahalanobisovy vzdálenosti pro ta pozorování, která mají nulu pouze v jedné proměnné, a to po řadě v první (nahoře vlevo) až třetí (dole). Porovnání pro pozorování s nulami ve čtvrté proměnné není možné vzhledem k malému počtu pozorování. Na x-ové ose jsou vyneseny hodnoty Mahalanobisových vzdáleností získané algoritmem s imputací, na y-ové ose pak robustní MD získané z podsouborů po eliminaci nulové proměnné.



Graf 4.8: srovnání MD pro data s jednou nulovou proměnnou

Z grafu 4.8 je patrné, že v detekci odlehlých hodnot se přístupy příliš nepodobají. Lze vidět i rozdíly v počtech detekovaných odlehlých pozorování, které jsou uvedeny výše v tabulce 5. Navíc lze vidět, že mnoho hodnot bylo označeno

jako odlehlé pouze jedním z přístupů. Obdobně si vykreslíme Mahalanobisovy vzdálenosti pro pozorování s dvěma nulovými proměnnými.



Graf 4.9: srovnání MD pro data s dvěma nulovými proměnnými

V grafu 4.9 lze opět vidět, že jsou pozorování označena za odlehlá jedním nebo druhým přístupem, podstatná většina pozorování není označena oběma. Lze tedy konstatovat, že se rozdělení v rámci jednotlivých podskupin pozorování, určených strukturou nulových složek, liší od celkové distribuce datového souboru, kterou jsme rekonstruovali s využitím imputace nul nenulovými hodnotami. Je nutné dále připomenout, že hovoříme pouze o první fázi dvoustupňového algoritmu z kapitoly 3.1.2, a tedy detekovaná odlehlá pozorování jsou odlehlá pouze vzhledem k nenulovým pozorováním. Druhá fáze by se zabývala detekcí vzhledem ke struktuře nul, ale protože je stále ve fázi vývoje, nebude zde demonstrována.

Závěr

Práce se zaměřila na statistickou analýzu kompozičních dat s nulami. Cílem práce bylo poskytnout ucelený přehled metod a následně tyto metody použít na konkrétních datech. Na úvod byly uvedeny základní poznatky o kompozičních datech, na kterých bylo ukázáno, proč jsou nuly pro analýzu kompozičních dat nežádoucí. Dále bylo vysvětleno, že je navíc třeba rozlišovat nuly v závislosti na studovaném problému. Pro každý druh nul pak byly představeny metody a algoritmy, díky kterým je možné s kompozičními daty s nulami pracovat. Na závěr byly postupy demonstrovány na reálných datech pomocí statistického softwaru R. Práce se snažila mimo jiné poskytnout též porovnání vybraných algoritmů, a to jak teoreticky prostřednictvím důkazů, tak prakticky porovnáním konkrétních výsledků, které byly obdrženy z reálných dat.

Domnívám se, že přínosem práce je komplexnost zpracování vybrané problematiky, která je stále předmětem současného zkoumání a je k dispozici převážné v anglicky psaných vědeckých článcích. Práce poskytuje jak teoretické poznatky, tak praktické návody pro samotnou analýzu kompozičních dat s nulami. Dále přibližuje čtenáři praktické použití metod pro nahrazení nul v softwaru R, díky kterému je pak možné pokračovat v další analýze kompozičních dat.

Díky tomuto tématu jsem byla obohacena o další přístup k analýze dat. S kompozičními daty je možné se setkat poměrně často, už jen díky tomu, že data vyjadřující procentuální zastoupení nebo proporce jsou často používaná. Navíc problematika nul je velmi praktický problém, se kterým se v reálných situacích setkáváme. Tudíž toto téma bylo obohacující nejen co se týče nových teoretických poznatků, ale rovněž jsem měla možnost se zabývat praktickým problémem.

Samozřejmě veškeré otázky týkající se nul v kompozičních datech nejsou zdaleka zodpovězeny. Jedná se o oblast, která se stále vyvíjí a je předmětem současného výzkumu. Zatímco metody pro nahrazení nul vzniklých zaokrouhlením jsou v podstatě osvědčené a uznávané, tak u metod pro práci se strukturními nulami se neustále hledají nové možnosti, jak informaci v nich obsaženou využít pro relevantní statistickou analýzu kompozičního datového souboru.

Literatura

[1] AITCHISON, J. The statistical analysis of compositional data. 1986.

[2] AITCHISON, J., BARCELÓ-VIDAL, C., MARTÍN-FERNÁNDEZ, J. A, PAWLOWSKY-GLAHN, V. Logratio analysis and compositional distance. *Mathematical Geology*, 2000, 32.3: 271-275.

[3] AITCHISON, J., GREENACRE, M. Biplots of compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2002, 51.4: 375-392.

[4] AITCHISON, J., KAY, J. W. Possible solution of some essential zero problems in compositional data analysis. 2003.

[5] FRY, J. M., FRY, T. R. L., MCLAREN, K. R. Compositional data analysis and zeros in micro data. *Applied Economics*, 2000, 32.8: 953-959.

[6] HRON, K. Elementy statistické analýzy kompozičních dat. *Informační bulletin České statistické společnosti*, 2010, 21.3: 41-48.

[7] MARTÍN-FERNÁNDEZ, J. A., HRON, K., TEMPL, M., FILZMOSER, P., PALAREA-ALBALADEJO, J. Model-based replacement of rounded zeros in compositional data: classical and robust approaches. *Computational Statistics & Data Analysis*, 2012, 56.9: 2688-2704.

[8] MONTERO-SERRANO, J. C., PALAREA-ALBALADEJO, J., MARTÍN-FERNÁNDEZ, J. A., MARTÍNEZ-SANTANA, M., GUTIÉRREZ-MARTÍN, J. V. Sedimentary chemofacies characterization by means of multivariate analysis. *Sedimentary Geology*, 2010, 228.3: 218-228.

[9] MONTI, G. S., HRON, K., TEMPL, M., FILZMOSER, P. Covariance-Based Outlier Detection for Compositional Data with Structural Zeros: Application to Italian Survey of Household Income and Wealth Data. *Advances in Latent Variables-Methods, Models and Applications*. 2013.

[10] PALAREA-ALBALADEJO, J., MARTÍN-FERNÁNDEZ, J. A. Values below detection limit in compositional chemical data. *Analytica chimica acta*, 2013, 764: 32-43.

[11] PALAREA-ALBALADEJO, J., MARTÍN-FERNÁNDEZ, J. A., BUCCIANTI, A. Compositional methods for estimating elemental concentrations below the limit of detection in practice using R. *Journal of Geochemical Exploration*, 2014, 141: 71-77.

[12] PAWLOWSKY-GLAHN, V., BUCCIANTI, A. *Compositional data analysis: Theory and applications*. John Wiley & Sons, 2011.

[13] PAWLOWSKY-GLAHN, V., EGOZCUE, J. J., TOLOSANA DELGADO, R. Lecture notes on compositional data analysis. 2007.

[14] ROUSSEEUW, P. J., DRIESSEN, K. V. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 1999, 41.3: 212-223.

[15] TEMPL, M., HRON, K., FILZMOSER, P., MONTI, G. S. Outlier detection in compositional data with structural zeros. *ODAM 2013*, 2013, 61.

[16] TEMPL, M., HRON, K., FILZMOSER, P. Exploratory tools for outlier detection in compositional data with structural zeros. *Odesláno*, 2015.

[17] Bank of Italy. *Bank of Italy – Distribution of the microdata.* [online]. [2015] [cit. 2016-03-01]. Dostupné z: http://www.bancaditalia.it/statistiche/tematiche/indagini-famiglie-imprese/bilanci-famiglie/distribuzione-microdati/index.html

[18] Poznat jiné země i jejich vzdělávací systémy nikdy neuškodí. Týdeník ŠKOLSTVÍ. [online]. [2014] [cit. 2016-03-03]. Dostupné z: *http://www.tydenik-skolstvi.cz/archiv-cisel/2014/24/poznat-jine-zeme-i-jejich-vzdelavaci-systemy-nikdy-neuskodi/*