UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Alternativní přístupy ke statistické analýze kompozičních dat

Katedra matematické analýzy a aplikací matematiky
Vedoucí diplomové práce: **doc. RNDr. Karel Hron, Ph.D.**
Vypracoval: **Bc. Jiří Válek**
Studijní program: N1103 Aplikovaná matematika
Studijní obor Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2017

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Jiří Válek

Název práce: Alternativní přístupy ke 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: 2017

Abstrakt: Přestože pro statistickou analýzu kompozičních dat se v současné době užívá především tzv. logpodílová metodika (založená na logkonstrastech složek kompozic), paralelně se rozvíjí též další přístupy. Tato diplomová práce se zabývá alternativními přístupy založenými na hypersférickém vyjádření kompozic a odmocninové transformaci. Detailněji se pak práce zaměřuje na popis několika metod, které pracují s takto transformovanými daty. Následně jsou tyto metody porovnány se známými metodami využívajícími logpodílové transformace.

Klíčová slova: kompoziční data, logpodílové souřadnice, odmocninová transformace, hypersférické vyjádření kompozic, Kentův regresní model, robustní PCA v tangentovém prostoru

Počet stran: 73

Počet příloh: 5

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Jiří Válek

Title: Alternative approaches to 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: 2017

Abstract: Although for statistical analysis of compositional data now mainly so-called logratio methodology (based on logcontrasts between components of compositions) is used, another approaches are developed in parallel. This thesis deals in particular with alternative approaches based on square root transformation and expressing compositions in hyperspherical coordinates. The thesis is focused in a more detail on description of several methods, which are based on data after the mentioned transformations. Finally, these methods are compared with well-known methods based on the logratio methodology.

Key words: compositional data, logratio coordinates, square root transformation, hyperspherical expression of compositions, Kent regression model, robust PCA in tangent space

Number of pages: 73

Number of appendices: 5

Language: Czech

Prohlášení

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

V Olomouci dne ....

podpis

Obsah

Úvod **8**

1 Kompoziční data a jejich vývoj **10**
1.1 Úvod do historie analýzy kompozičních dat 10
1.2 Logpodílový přístup k analýze kompozičních dat 12
1.2.1 Aitchisonova geometrie na simplexu 13
1.2.2 Souřadnicový systém, logpodílové souřadnice 14
1.3 Přehled alternativních přístupů ke statistické analýze kompozičních dat 17
1.4 Sférický přístup ke statistické analýze kompozičních dat 19
1.4.1 Hypersférické vyjádření kompozic 19
1.4.2 Odmocninová transformace 23
1.5 Principy statistické analýzy kompozičních dat - jejich vznik a vliv na alternativní přístupy 24

2 Regresní model založený na odmocninové transformaci **30**
2.1 Obecný Kentův regresní model 30
2.2 Multiplikativní a aditivní Kentův regresní model 32
2.3 Asymptotické chování Kentova modelu a odhad jeho parametrů . 35
2.4 Vlastnosti Kentova regresního modelu a jejich srovnání s logpodílovou metodikou 40

3 Robustní metoda hlavních komponent v tangentovém prostoru **43**
3.1 Základní pojmy z diferenciální geometrie 44
3.1.1 Topologický prostor, topologická a diferenciální varieta . . 44
3.1.2 Tangentové vektory, tangentový prostor 45
3.2 Robustní PCA v tangentovém prostoru 48
3.2.1 Popis metody 48
3.2.2 Odhady parametrů 52
3.3 Porovnání metody s dalšími známými přístupy 54
3.3.1 Vztah k logpodílovému přístupu 54
3.3.2 Vztah k PCA na hyperkouli 55
3.4 Příklad: příčiny úmrtí v evropských zemích 56

Závěr **64**

A Rekurzivní vztahy pro výpočet úhlového rozptylu a kovariance **65**

B Příklad: příčiny úmrtí - datový soubor **67**

C Příklad: příčiny úmrtí - proporciální datový soubor **68**

D Příklad: příčiny úmrtí - datový soubor po logaritmické transformaci **69**

E Příklad: příčiny úmrtí - projekce na tangentový prostor **70**

Literatura **71**

Poděkování

Rád bych tímto poděkoval vedoucímu diplomové práce panu doc. RNDr. Karlu Hronovi, Ph.D. za cenné rady, věcné připomínky, ochotu a čas, který mi v průběhu zpracování diplomové práce věnoval.

Úvod

Tématem této práce je zevrubně popsat vybrané alternativní metodiky ke statistické analýze kompozičních dat. V současnosti jsou k tomuto účelu používány zejména metody založené na transformacích kompozic do eukleidovského prostoru, které využívají poměrů mezi složkami kompozic. Těmto transformacím se říká logpodílové transformace a výsledky analýzy transformovaných dat v eukleidovském prostoru pak můžeme zpětně přenést na poznatky o vlastnostech původních kompozic.

Úkolem této práce je představit další možné způsoby analýzy kompozičních dat a porovnat jejich vlastnosti s logpodílovou metodikou. K tomu bude nutné nejprve stručně popsat logpodílový přístup k analýze kompozičních dat. Ten přiblížíme na začátku první kapitoly, a to včetně náhledu do historie kompozičních dat. Poté uvedeme přehled uvažovaných alternativních přístupů a zaměříme se detailněji na jeden z nich, který se v současnosti nejvíce rozvíjí. Označuje se jako tzv. sférický přístup k analýze kompozičních dat, konkrétně se u něj ještě rozlišují dvě skupiny metod. První z nich se zaměřuje na vyjádření kompozic pomocí hypersférických geografických souřadnic a určení tzv. úhlů absolutní reference. Druhá skupina využívá odmocninovou transformaci zobrazující kompoziční data na povrch jednotkové hyperkoule. Hyperkoulí rozumíme vícerozměrné zobecnění koule.

V dalších kapitolách se budeme věnovat metodám, které pracují s výše zmíněnou odmocninovou transformací. V druhé kapitole představíme model používaný k regresní analýze takto transformovaných dat. Třetí kapitola se pak věnuje robustní metodě hlavních komponent v tangentovém prostoru. Ta se používá

především při analýze kompozičních dat, jejichž složky nabývají malých hodnot a jsou tak značně variabilní. Využívá se speciální relativní mocninné transformace k zobrazení kompozic na tzv. variety, které mohou být poměrně komplikovanými plochami. Dále se pokračuje projekcí na eukleidovský tangentový prostor, na němž se provádí samotná metoda hlavních komponent. Třetí kapitola je zakončena aplikací této robustní metody hlavních komponent na reálná data.

Téma alternativních přístupů k analýze kompozičních dat jsem jsem si zvolil proto, neboť se domnívám, že je to oblast zajímavá, ale přesto se o ní v současnosti příliš nepíše. Je tak dle mého názoru vhodná k dalšímu bádání. Statistika mi byla při studiu blízká a proto jsem vnímal toto téma jako výzvu naučit se něco nového, co se stále aktivně rozvíjí.

Kapitola 1

Kompoziční data a jejich vývoj

Cílem této kapitoly je seznámit čtenáře s vývojem pohledu na kompoziční data napříč zlomovými okamžiky v historii kompozic. V této kapitole jsou dále uvedeny základní pojmy týkající se kompozičních dat a práce s nimi. Nejprve se zaměříme na popis vývoje kompozičních dat, poté stručně popíšeme tzv. logpodílový přístup ke statistické analýze kompozičních dat, který je v současné době nejrozšířenějším přístupem k práci s těmito daty. Následně jsou ve stručnosti uvedeny a popsány jednotlivé alternativní přístupy. Zvýšená pozornost je věnována tzv. sférickému přístupu.

1.1. Úvod do historie analýzy kompozičních dat

Vývoj analýzy kompozičních dat do podoby, v jaké ji dnes uvažujeme, lze rozdělit zhruba do čtyř základních fází. Při jejich popisu budeme vycházet z literatury [7], [13], [15].

První fází nazýváme období do roku 1960. V této době se ke statistické analýze kompozičních dat používaly výhradně nástroje mnohorozměrné statistické analýzy. Vědci používali ke statistické analýze kompozičních vektorů, jejichž složky jsou proporcemi nějakého celku, všechny metody mnohorozměrné statistické analýzy, a to zejména korelační analýzu. Důležitým mezníkem této fáze je článek [15] Karla Pearsona z roku 1897, který se zabývá úskalím interpretace výsledků takovéto analýzy kompozic. Pearson v něm považuje korelaci mezi

složkami kompozičního vektoru za tzv. falešnou v případě, že složky vektoru před jeho uzavřením na konstantní součet byly nekorelované. Naznačme nyní hlavní myšlenku příkladu, který uvedl Pearson. Uvažujeme-li p-složkovou kompozici jako vektor $\mathbf{u} = (u_1, u_2, \ldots, u_p)^T$ s konstantním součtem $\sum_{i=1}^{p} u_i = 1$, pak platí-li

$$\mathrm{cov}(u_1, u_1 + u_2 + \cdots + u_p) = 0,$$

dostaneme po úpravě

$$\mathrm{cov}(u_1, u_2) + \mathrm{cov}(u_1, u_3) + \cdots + \mathrm{cov}(u_1, u_p) = -\mathrm{var}(u_1).$$

Je tedy chybné interpretovat korelaci mezi složkami vektoru fixovaného na konstantní součet jako korelaci mezi složkami neuzavřeného vektoru. I přes toto zjištění ovšem nebyl tento přístup až do roku 1960 odsouzen.

V další fázi již došlo k hromadné kritice aplikace standardní mnohorozměrné statistické analýzy na kompoziční data, a to v čele s geologem Felixem Chayesem v [7]. Jeho kritika mířila především proti již zmíněné interpretaci výběrové korelace, konkrétně u geochemické kompozice. Chayes považoval negativní bias, tedy uvedený zvýšený výskyt negativní korelace u proporciálních dat, jako problém jakékoliv smysluplné interpretace výsledků mnohorozměrné analýzy aplikované na kompoziční data. Druhá fáze se vyznačovala spíše patologickým přístupem, než-li snahou nalézt vhodnou metodiku.

Zlomovým okamžikem v historii kompozičních dat je označován revoluční pohled na kompozice od roku 1980. Skotský statistik John Aitchison nahlížel na kompoziční data jako na vektory nesoucí výhradně relativní informaci. Pak mohl každé tvrzení o těchto datech vyjádřit z hlediska poměrů mezi složkami kompozic. Aitchison v této třetí fázi zavedl celou metodiku založenou na logpodílových transformacích do reálného prostoru. Na transformovaná data lze použít nástroje standardní mnohorozměrné statistické analýzy a následně můžeme takto získané poznatky o vlastnostech reálných vektorů zpětně převést na poznatky o vlastnostech kompozic.

Čtvrtá fáze pramení ze zjištění, že vnitřní a vnější operace na simplexu, tzv.

perturbace a mocninná transformace, spolu se skalárním součinem tvoří eukleidovský vektorový prostor. Kromě skutečnosti, že je možné kompozice analyzovat přímo v rámci této algebraicko-geometrické struktury, je výhodné kompoziční data reprezentovat pomocí reálných souřadnic v eukleidovském vektorovém prostoru. Zde jsou kompozice statisticky analyzovány a prostřednictvím této analýzy jsou interpretovány jejich vztahy na simplexu. Jako výběrový prostor kompozic je v této fázi chápán výhradně simplex, na kterém je definována již zmíněná algebraicko-geometrická struktura, tzv. Aitchisonova geometrie. Zejména této čtvrté fázi bude věnována následující podkapitola.

1.2. Logpodílový přístup k analýze kompozičních dat

Tato podkapitola se blíže zaměřuje na čtvrtou fázi uvedenou výše, tedy na přístup k modelování kompozičních dat, který uvedl J. Aitchison v 80. letech 20. století. Aitchison popisuje kompoziční data jako vícerozměrné statistické soubory, jejichž složky představují kvantitativně vyjádřené příspěvky částí na celku. Uveďme nyní definici p-složkové kompozice. Dále definujme také výběrový prostor kompozičních dat a uzávěr libovolného vektoru. V této podkapitole vycházíme zejména z literatury [4], [13], [18].

Definice 1.2.1 *Sloupcový vektor* $\mathbf{u} = (u_1, u_2, \ldots, u_p)^T$ *nazýváme p-složkovou kompozicí, jsou-li všechny jeho složky kladná reálná čísla nesoucí pouze relativní informaci.*

Definice 1.2.2 *Výběrovým prostorem kompozičních dat je simplex* Δ^p *dimenze* $p-1$ *ve tvaru*

$$\Delta^p = \Big\{ \mathbf{u} = (u_1, u_2, \ldots, u_p)^T \mid u_i > 0, i = 1, 2, \ldots, p, \sum_{i=1}^{p} u_i = k \Big\},$$

kde k je libovolná pevně zvolená konstanta představující součet složek kompozic.

Definice 1.2.3 *Uzávěrem libovolného vektoru* $\mathbf{b} = (b_1, b_2, \ldots, b_p)^T \in \mathbb{R}_+^p$, $b_i > 0, i = 1, 2, \ldots, p$ *nazýváme kompozici*

$$\mathcal{C}(\mathbf{b}) = \left(\frac{kb_1}{\sum_{i=1}^p b_i}, \frac{kb_2}{\sum_{i=1}^p b_i}, \ldots, \frac{kb_p}{\sum_{i=1}^p b_i} \right)^T .$$

1.2.1. Aitchisonova geometrie na simplexu

Na začátku této podkapitoly uveďme motivační příklad. Mějme čtyři kompozice $\mathbf{u} = (5, 70, 25)^T, \mathbf{v} = (15, 60, 25)^T, \widetilde{\mathbf{u}} = (40, 35, 25)^T$ a $\widetilde{\mathbf{v}} = (50, 25, 25)^T$. Rozdíl mezi kompozicemi $\mathbf{u}$ a $\mathbf{v}$ není stejný jako rozdíl mezi $\widetilde{\mathbf{u}}$ a $\widetilde{\mathbf{v}}$. V obou případech je sice jejich eukleidovská vzdálenost rovna $10\sqrt{2}$ a mezi první a druhou složkou je rozdíl 10 jednotek, nicméně relativní nárůst je rozdílný. U první kompozice je to 300% a u druhé pouze 25%. Tato skutečnost nás vede ke zjištění, že bude nutné pracovat s citlivější geometrií, než je eukleidovská.

Nejprve definujeme dvě operace - perturbaci a mocninnou transformaci. Tyto operace jsou analogií operací sčítání vektorů a násobení vektorů skalárem. Poté zavedeme Aitchisonův skalární součin, normu a vzdálenost.

Definice 1.2.4 *Perturbace kompozice* $\mathbf{u} \in \Delta^p$ *kompozicí* $\mathbf{v} \in \Delta^p$ *je kompozice*

$$\mathbf{u} \oplus \mathbf{v} = \mathcal{C}(u_1 v_1, u_2 v_2, \ldots, u_p v_p)^T.$$

Definice 1.2.5 *Mocninná transformace kompozice* $\mathbf{u} \in \Delta^p$ *konstantou* $\alpha \in \mathbb{R}$ *je kompozice*

$$\alpha \odot \mathbf{u} = \mathcal{C}(u_1^\alpha, u_2^\alpha, \ldots, u_p^\alpha)^T.$$

Poznámka 1.2.1 *Trojice* $(\Delta^p, \oplus, \odot)$ *je vektorovým prostorem nad* Δ^p.

Definice 1.2.6 *Aitchisonův skalární součin kompozic* $\mathbf{u}, \mathbf{v} \in \Delta^p$ *definujeme jako*

$$\langle \mathbf{u}, \mathbf{v} \rangle_a = \frac{1}{2p} \sum_{i=1}^p \sum_{j=1}^p \ln \frac{u_i}{u_j} \ln \frac{v_i}{v_j} = \sum_{i=1}^p \ln \frac{u_i}{g(\mathbf{u})} \ln \frac{v_i}{g(\mathbf{v})},$$

kde

$$g(\mathbf{u}) = \sqrt[p]{\prod_{i=1}^p u_i} \ , \quad g(\mathbf{v}) = \sqrt[p]{\prod_{i=1}^p v_i} \ .$$

Definice 1.2.7 *Aitchisonova norma kompozice* $\mathbf{u} \in \Delta^p$ *je definována jako*

$$\|\mathbf{u}\|_a = \sqrt{\frac{1}{2p}\sum_{i=1}^{p}\sum_{j=1}^{p}\left(\ln\frac{u_i}{u_j}\right)^2} = \langle \mathbf{u}, \mathbf{u}\rangle_a.$$

Definice 1.2.8 *Aitchisonova vzdálenost mezi kompozicemi* $\mathbf{u}, \mathbf{v} \in \Delta^p$ *je definována jako*

$$d_a(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} \ominus \mathbf{v}\|_a = \sqrt{\frac{1}{2p}\sum_{i=1}^{p}\sum_{j=1}^{p}\left(\ln\frac{u_i}{u_j} - \ln\frac{v_i}{v_j}\right)^2}, \tag{1.1}$$

kde $\mathbf{u} \ominus \mathbf{v} = \mathbf{u} \oplus [(-1) \odot \mathbf{v}]$.

Vektorový prostor $(\Delta^p, \oplus, \odot)$ spolu s operací Aitchisonova skalárního součinu, normy a vzdálenosti tvoří eukleidovský vektorový prostor na simplexu Δ^p zvaný Aitchisonova geometrie.

1.2.2. Souřadnicový systém, logpodílové souřadnice

Vzhledem ke skutečnosti, že absolutní hodnoty nejsou při práci s kompozičními daty relevantní, zavedl Aitchison transformace kompozic ze simplexu do reálného prostoru založené na poměrech. V tomto kontextu budeme místo o *transformacích* hovořit spíše o *souřadnicích*, což lépe odpovídá geometrickému pohledu na kompoziční data v rámci logpodílové metodiky.

Před popisem jednotlivých souřadnicových systémů nejdříve naznačíme, jakým způsobem lze vytvořit bázi vzhledem k Aitchisonově geometrii na simplexu Δ^p. Generující systém této báze můžeme získat jako systém kompozic $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p$, pro které platí $\mathbf{w}_i = \mathcal{C}\,(1, 1, \ldots, e, \ldots, 1)^T$, kde Eulerovo číslo e je na pozici i-té složky daného vektoru, $i = 1, 2, \ldots, p$. Kompozici $\mathbf{u} \in \Delta^p$ pak lze psát jako

$$\mathbf{u} = \ln u_1 \odot (e, 1, \ldots, 1)^T \oplus \ln u_2 \odot (1, e, \ldots, 1)^T \oplus \cdots \oplus \ln u_p \odot (1, 1, \ldots, e)^T =$$

$$= \bigoplus_{i=1}^{p} \ln u_i \odot \mathbf{w}_i. \tag{1.2}$$

Je známo, že toto vyjádření není jednoznačné, a proto lze použít i následující ekvivalentní vyjádření

$$\mathbf{u} = \bigoplus_{i=1}^{p} \ln \frac{u_i}{g(\mathbf{u})} \odot \mathbf{w}_i.$$

Pak tedy můžeme definovat zobrazení $clr : \Delta^p \to \mathbb{R}^p$, které kompozici $\mathbf{u}$ přiřadí reálný vektor $clr(\mathbf{u})$ ve tvaru

$$clr(\mathbf{u}) = \left(\ln \frac{u_1}{g(\mathbf{u})}, \ln \frac{u_2}{g(\mathbf{u})}, \ldots, \ln \frac{u_p}{g(\mathbf{u})} \right)^T. \tag{1.3}$$

Vektor $clr(\mathbf{u})$ nazýváme *clr* souřadnice (z angl. centered logratio) kompozice $\mathbf{u}$. Zmíněný vektor pak představuje souřadnice kompozice $\mathbf{u}$ vzhledem ke zvolenému generujícímu systému $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p\}$.

Ve vztahu pro *clr* transformaci (1.3) můžeme $g(\mathbf{u})$ nahradit jednou ze složek $u_1, u_2, \ldots, u_p$. Zvolíme-li například u_p, pak v tomto případě je koeficient odpovídající p-té složce roven nule. Generující systém $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{p-1}\}$ potom tvoří bázi a kompozici $\mathbf{u} \in \Delta^p$ lze psát jako

$$\mathbf{u} = \bigoplus_{i=1}^{p-1} \ln \frac{u_i}{u_p} \odot \mathbf{w}_i. \tag{1.4}$$

Můžeme tedy definovat zobrazení *alr*: $\Delta^p \to \mathbb{R}^{p-1}$ takové, že kompozici $\mathbf{u}$ přiřadí reálný vektor

$$alr(\mathbf{u}) = \left(\ln \frac{u_1}{u_p}, \ln \frac{u_2}{u_p}, \ldots, \ln \frac{u_{p-1}}{u_p} \right)^T.$$

Pak reálný vektor $alr(\mathbf{u})$ nazýváme *alr* souřadnice (z angl. additive logratio) kompozice $\mathbf{u}$. Tento vektor udává souřadnice vzhledem k bázi $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{p-1}\}$.

Z generujícího systému (1.2) lze snadno získat bázi zvolením $p-1$ kompozic ze systému kompozic $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p$. Pak pro $\mathbf{u} \in \Delta^p$ platí například vztah (1.4), kde báze $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{p-1}\}$ ovšem není ortonormální. Ortonormální bázi lze najít například pomocí Gram-Schmidtovy metody. Další možnou alternativou je využít tzv. metodu postupného binárního dělení. Způsob, jakým je vybrána

ortonormální báze je důležitý pro následnou interpretaci souřadnic. Metoda postupného binárního dělení je podrobně popsána v literatuře [13].

Pro vybranou ortonormální bázi $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{p-1}\}$ a libovolnou kompozici $\mathbf{u} \in \Delta^p$ nám zobrazení *ilr*: $\Delta^p \to \mathbb{R}^{p-1}$ dané vztahem

$$ilr(\mathbf{u}) = (\langle \mathbf{u}, \mathbf{e}_1 \rangle_a, \langle \mathbf{u}, \mathbf{e}_2 \rangle_a, \ldots, \langle \mathbf{u}, \mathbf{e}_{p-1} \rangle_a)^T$$

přiřadí kompozici $\mathbf{u}$ reálný vektor $ilr(\mathbf{u})$, který nazýváme *ilr* souřadnice (z angl. isometric logratio) kompozice $\mathbf{u}$. Tento vektor tedy vyjadřuje souřadnice kompozice $\mathbf{u}$ vzhledem ke zvolené ortonormální bázi. Navíc se jedná o izomorfismus vektorových prostorů.

Porovnejme nyní ve stručnosti vlastnosti uvedených transformací. Nevýhodou *alr* transformace je skutečnost, že není izometrická a navíc není symetrická. *Clr* transformace sice izometrická je, vede ovšem k singulární varianční matici. Zřejmě *ilr* transformace bude z tohoto pohledu nejlepší, protože zachovává všechny metrické vlastnosti. Navíc při vhodné volbě ortonormální báze jsou souřadnice velmi dobře interpretovatelné. Dále lze ukázat (viz [18]), že všechny výše uvedené transformace jsou lineární logpodílové transformace (tzv. logkontrasty), tj. každou výslednou souřadnici lze zapsat jako

$$a_1 \ln u_1 + a_2 \ln u_2 + \cdots + a_p \ln u_p, \text{ kde } a_1 + a_2 + \cdots + a_p = 0.$$

Nakonec uveďme pro úplnost poslední transformaci, která tuto výše uvedenou vlastnost nemá. Definujme zobrazení *mlr*: $\Delta^p \to \mathbb{R}^{p-1}$ vztahem

$$mlr(\mathbf{u}) = \left(\ln \frac{u_1}{\sum_{j=2}^{p} u_j}, \ln \frac{u_2}{\sum_{j=3}^{p} u_j}, \ldots, \ln \frac{u_{p-1}}{u_p} \right)^T.$$

Pak reálný vektor $mlr(\mathbf{u})$ nazýváme *mlr* souřadnicemi (z angl. multiplicative logratio) kompozice $\mathbf{u}$.

Podkapitolu ukončíme definováním dimenzi redukujících operací na kompozicích, které uvedl Aitchison v [4]. Níže uvedené operace budeme využívat zejména v závěru celé kapitoly.

Podmnožinou kompozice se rozumí vektor (nikoliv kompozice) $\mathcal{S}(\mathbf{u}) = \mathbf{Su}$, kde $\mathbf{S}$ je matice nul typu $s \times p, 1 \leq s \leq p$, taková, že se v každém řádku vyskytuje prvek 1 právě jednou a v každém sloupci nejvýše jednou. Celkem je matice $\mathbf{S}$ tvořená právě s jedničkami.

Amalgamací složek kompozice se rozumí kompozice $\mathcal{A}(\mathbf{u}) = \mathbf{Au}$, kde $\mathbf{A}$ je matice nul typu $a \times p, a \leq p$ taková, že se v každém sloupci nachází prvek 1 právě jednou a v každém řádku alespoň jednou (jedná se o sloučení složek $\mathbf{u}$).

Podkompozicí kompozice $\mathbf{u}$ se rozumí kompozice

$$\mathcal{C}(\mathcal{S}(\mathbf{u})) = \mathcal{C}(\mathbf{Su}) = \frac{\mathbf{Su}}{\mathbf{u}^T\mathbf{S}^T\mathbf{1}_s},$$

která představuje kompozici určenou podvektorem $\mathbf{u}$.

1.3. Přehled alternativních přístupů ke statistické analýze kompozičních dat

V [1] Aitchison uvedl pod názvem *„Pockets of resistance and confusion“* přehled jednotlivých alternativních přístupů. V následujícím textu jsou uvedeny základní alternativní přístupy tak, jak je klasifikoval Aitchison. Jejich jednotlivá pojmenování jsou v textu ponechána bez překladu. Poznamenejme, že tento Aitchisonův souhrn působí v originálním znění přinejmenším kriticky. Vybrané významné argumenty, na nichž je založen jeho kritický postoj, jsou součástí popisu jednolivých přístupů. V souvislosti s dnešním pohledem na definici kompozičních dat v rámci logpodílové metodiky podotkněme, že konstantní součet složek představuje pouze vhodnou reprezentaci informace obsažené v podílech mezi složkami kompozic. Na druhou stranu, jak uvidíme později (viz podkapitola 1.4), v případě alternativních metod se jedná vskutku o součást definice kompozic.

The Wishful thinkers: ignorují omezení na konstantní součet nebo vynechávají jednu složku kompozice. Počítají a interpretují korelace mezi složkami neuzavřených vektorů.

The Describers: argumentují tím, že v popisné statistice je možné ignorovat omezení na konstantní součet. Popisují kompoziční soubor pomocí aritmetických průměrů, kovariančních matic původních složek a používají lineární metody jako je metoda hlavních komponent. Lze tedy říct, že tento přístup se vyznačuje spíše popisem, než-li analýzou.

The Openers: hledají vhodný způsob jak „otevřít“ kompozice a provádět analýzu takto vzniklého statistického souboru. Aitchison tento přístup odsuzuje zvláště kriticky, a to jak z hlediska geometrického, tak především z hlediska požadavku invariance na změnu měřítka. Pokud bychom totiž provedli libovolnou operaci invariantní na změnu měřítka na „otevřenou“ kompozici, musel by být efekt stejný jako při jejím provedení na dané kompozici. To hodnotí Aitchison doslova jako pošetilost.

The Null correlationists: připouštějí, že kvůli negativnímu biasu v korelacích mezi kompozicemi nulová korelace „nevede k nezávislosti“. Jejich cílem je tedy stanovit takovou korelaci, která by „vedla k nezávislosti“. Aitchison komentuje tím, že celý uvažovaný koncept nulové korelace je neopodstatněný a opravdu zbytečný. Všechny smysluplné koncepty kompoziční závislosti a nezávislosti mohou být pozorovány na simplexu a v souvislosti s logpodílovou varianční strukturou, která je již vymezena.

The Pathologists: spíše popisují problémy, než-li přinášejí řešení.

The Non-transformationists: nedůvěřují transformacím a interpretaci transformovaných dat a tak raději pracují na netransformovaných datech.

The Dirichlet extenders: snaží se zobecnit Dirichletovo rozdělení na simplexu. Tento přístup také podlehl značné kritice, neboť i přes elegantní matematické vlastnosti není Dirichletovo rozdělení plně kompatibilní s Aitchisonovou geometrií.

The Sphericists: transformují kompozice na body na povrchu hyperkoule a využívají směrových kosinů. Ani tato skupina nebyla ušetřena značné kritiky. Blíže specifikujeme rozepři v podkapitole 1.5, konkrétní kritice se budeme věnovat v podkapitole 2.4.

1.4. Sférický přístup ke statistické analýze kompozičních dat

Při tvorbě této podkapitoly jsme čerpali především ze zdroje [23], dále jsme použili [16], [21], [26].

Jako nejvýznamněji rozvíjejícím se alternativním přístupem ke statistické analýze kompozičních dat se jeví přístup skupiny *The Sphericists*. V dalším textu již budeme Aitchisonovo pojmenování této skupiny překládat do češtiny. Sféricisty jsou, v publikacích zabývajících se kompozičními daty, označovány zejména dvě skupiny vědců. Se jmény D. F. Watson, G. M. Philip a C. Stanley se pojí alternativní přístup založený na využití tzv. směrových kosinů a úhlů absolutní reference. Se jmény M. A. Stephens, J. L. Scealy a A. H. Welsh souvisí použití tzv. odmocninové transformace, která zobrazuje kompoziční data na data směrová. Zdůrazněme hned na začátku podkapitoly, že v dalším textu této podkapitoly budeme kompozici definovat jinak, než tomu bylo v definici 1.2.1.

Definice 1.4.1 *Sloupcový vektor* $\mathbf{u} = (u_1, u_2, \ldots, u_p)^T$ *nazýváme p-složkovou kompozicí, jestliže platí*

$$u_i \geq 0,\ i = 1, 2, \ldots, p \quad a \quad \sum_{i=1}^{p} u_i = 1.$$

Takto definované kompozice nabývají svých hodnot v $(p-1)$-rozměrném simplexu

$$\Delta^p = \Big\{ \mathbf{u} = (u_1, u_2, \ldots, u_p)^T \mid u_i \geq 0,\ i = 1, 2, \ldots, p,\ \sum_{i=1}^{p} u_i = 1 \Big\}.$$

1.4.1. Hypersférické vyjádření kompozic

Hlavní myšlenka přístupu analýzy kompozic podle Watsona a Philipa v [16], [26] spočívá v transformování kompozic na tzv. úhly absolutní reference. Nejprve se provádí projekce kompozic z jednotkové nadroviny (jak se v tomto kontextu simplex také nazývá) na jednotkovou hyperkouli. Tuto projekci získáme

vydělením každé ze složek kompozice odmocninou ze součtu čtverců všech složek kompozice, což můžeme zapsat následovně

$$\mathbf{u}_s = (u_{s1}, u_{s2}, \ldots, u_{sp})^T = \left(\frac{u_1}{\sqrt{\sum_{j=1}^{p} u_j^2}}, \frac{u_2}{\sqrt{\sum_{j=1}^{p} u_j^2}}, \ldots, \frac{u_p}{\sqrt{\sum_{j=1}^{p} u_j^2}} \right)^T ,$$

kde u_{si} jsou transformované hodnoty, u_i jsou složky kompozice $\mathbf{u}$, $i = 1, 2, \ldots, p$ a $p = m + 1$ je počet složek kompozice, kde m je počet nezávislých proměnných. Tímto se omezení na konstantní součet ve tvaru

$$\sum_{i=1}^{p} u_i = 1$$

mění na omezení ve tvaru

$$\sum_{i=1}^{p} u_{si}^2 = 1.$$

Proměnné u_{si} se nazývají směrové kosiny úhlů definujících vektor, který začíná v počátku soustavy souřadnic a končí v bodě daného prostoru, kde se nachází kompozice. Použitím směrových kosinů lze vypočítat úhlovou vzdálenost mezi dvěma kompozicemi $\mathbf{v}$ a $\mathbf{w}$ vztahem

$$\phi_d = \arccos \left(\sum_{i=1}^{p} u_{si\mathbf{v}} u_{si\mathbf{w}} \right),$$

kde $u_{si\mathbf{v}}, u_{si\mathbf{w}}$ značí hodnoty směrových kosinů kompozic $\mathbf{v}, \mathbf{w}$. Pomocí směrových kosinů můžeme určit střední hodnotu a rozptyl kompozic výpočtem součtu těchto vektorů. Součet vektorů se používá obvykle v prostorech nižší dimenze (2 či 3) k popisu orientací, lze jej ovšem také zobecnit součtem příslušných směrových kosinů a následným vydělením počtem pozorování n. Pak tedy k určení p-rozměrného průměrného směru výsledného vektoru nejprve vypočteme součty směrových kosinů t_{si} vztahem

$$\mathbf{t}_s = (t_{s1}, t_{s2}, \ldots, t_{sp})^T = \left(\sum_{j=1}^{n} u_{s1j}, \sum_{j=1}^{n} u_{s2j}, \ldots, \sum_{j=1}^{n} u_{spj} \right)^T ,$$

kde u_{sij} jsou hodnoty směrových kosinů. Dále tyto součty zobrazíme na jednotkovou hyperkouli a tím získáme směrové kosiny průměrného směru

$$\overline{\mathbf{u}}_s = (\overline{u}_{s1}, \overline{u}_{s2}, \ldots, \overline{u}_{sp})^T = \left(\frac{t_{s1}}{\sqrt{\sum_{j=1}^{p} t_{sj}^2}}, \frac{t_{s2}}{\sqrt{\sum_{j=1}^{p} t_{sj}^2}}, \ldots, \frac{t_{sp}}{\sqrt{\sum_{j=1}^{p} t_{sj}^2}} \right)^T .$$

Použitím vztahu

$$\overline{\mathbf{u}}_r = (\overline{u}_{r1}, \overline{u}_{r2}, \ldots, \overline{u}_{rp})^T = \left(\frac{t_{s1}}{\sum_{j=1}^{p} t_{sj}}, \frac{t_{s2}}{\sum_{j=1}^{p} t_{sj}}, \ldots, \frac{t_{sp}}{\sum_{j=1}^{p} t_{sj}} \right)^T$$

můžeme tyto součty zobrazit zpět na jednotkovou nadrovinu a tím získáme průměrný směr výslednice součtu vektorů $\overline{\mathbf{u}}_r$. Pro délku této výslednice d platí

$$d = \sqrt{\sum_{i=1}^{p} \left(\frac{\sum_{j=1}^{n} u_{sij}}{n} \right)^2}$$

a $d \in (0, 1)$, výslednice tedy leží uvnitř jednotkové hyperkoule.

Navzdory tomu, že míru centrality a rozptylu můžeme vypočítat pomocí směrových kosinů, lze data také zkoumat a popsat ve smyslu jejich úhlů absolutní reference pomocí vztahů, které popisují závislost mezi směrovými kosiny a jejich příslušnými geograficky definovanými úhly. Výše popisovaný vztah je ve tvaru

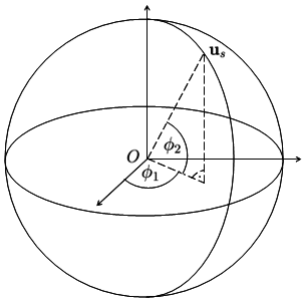
$$u_{sh} = \sin \phi_{h-1} \prod_{i=h}^{m} \cos \phi_i, \tag{1.5}$$

kde $\sin \phi_0 = 1$, tj. $\phi_0 = \pi/2$, $h = 1, 2, \ldots, p$ a $m = p - 1$ je počet nezávislých proměnných.

Jako příklad uveďme sférické geografické souřadnice, které dostaneme volbou $p = 3$ v (1.5). Tyto souřadnice jsou ve tvaru

$$\mathbf{u}_s = (u_{s1}, u_{s2}, u_{s3})^T = (\cos \phi_1 \cos \phi_2, \sin \phi_1 \cos \phi_2, \sin \phi_2)^T.$$

Úhel ϕ_1 v tomto případě nazýváme zeměpisnou délkou a ϕ_2 zeměpisnou šířkou. Situace je znázorněna na obrázku 1.1.



Obrázek 1.1: Sférické geografické souřadnice

Dále můžeme uvést do souvislosti geograficky definované úhly a jejich příslušné hypersféricky zobrazené kompozice vztahem

$$\phi_g = \arcsin\left(\frac{u_{s(g+1)}}{\prod_{i=g+1}^{p} cos\phi_i}\right), \tag{1.6}$$

kde $\cos\phi_p = 1$, tj. $\phi_p = 0$ a $g = m, m-1, \ldots, 1$.

Pomocí vztahů (1.5) a (1.6) je možné každý kompoziční soubor transformovat do úhlů hypersférického geografického souřadnicového systému a zpět. Počet těchto úhlů je určen počtem nezávislých proměnných (tedy jich je o jednu méně než složek kompozice) a tyto úhly jsou navzájem nezávislé.

Úhly dané vztahem (1.6) můžeme použít ke statistické analýze úhlových vzdáleností mezi kompozicemi v různých dimenzích. Vzhledem k absenci vhodného mnohorozměrného rozdělení, které by bylo možné ke statistické analýze těchto úhlů použít, postupujeme čistě popisně. Vzniklé úhlové statistiky je možné posléze transformovat zpět do prostoru kompozičních dat za účelem popisu středních hodnot a variability dat ve smyslu složek kompozice. Střední hodnoty složek kompozice lze transformovat z průměrných úhlů použitím (1.5), rozptyl a kovarianci kompozic musíme odvodit z úhlového rozptylu a kovariance použitím Taylorova rozvoje

$$\sigma_q^2 = \sum_{j=1}^{m}\left(\frac{\partial q}{\partial \theta_j}\right)^2 \sigma_j^2 + 2\sum_{j=1}^{m}\left[\sum_{\substack{k=1 \\ k\neq j}}^{m}\left(\frac{\partial q}{\partial \theta_j}\right)\left(\frac{\partial q}{\partial \theta_k}\right)\sigma_{jk}\right] \tag{1.7}$$

a

$$\sigma_{qr} = \sum_{j=1}^{m} \left(\frac{\partial q}{\partial \theta_j}\right)\left(\frac{\partial r}{\partial \theta_j}\right)\sigma_j^2 + \sum_{j=1}^{m}\left[\sum_{\substack{k=1\\k\neq j}}^{m}\left(\frac{\partial q}{\partial \theta_j}\right)\left(\frac{\partial r}{\partial \theta_k}\right) + \left(\frac{\partial r}{\partial \theta_k}\right)\left(\frac{\partial q}{\partial \theta_j}\right)\sigma_{jk}\right], \tag{1.8}$$

kde q a r uvažujeme jako funkce parametrů θ_j a θ_k. V našem případě budou tyto funkce nabývat svých hodnot v intervalu $\langle -1, 1\rangle$. Využitím vztahu (1.5), tj. dosazením za q a r do (1.7), resp. (1.8) lze odvodit vztahy pro rozptyl $\sigma^2_{u_{sh}}$, resp. kovarianci $\sigma_{u_{sg}u_{sh}}$, $g = 1, 2, \ldots, h-1$ a $h = 1, 2, \ldots, m+1$. Jejich tvar je značně rozsáhlý a proto jsou tyto vztahy uvedeny odděleně v příloze A. Na základě těchto vztahů z přílohy A a vztahu (1.6) je možné odvodit také rozptyl a kovarianci pomocí úhlových vzdáleností

$$\sigma^2_{\phi_g} = \frac{\sigma^2_{u_{s(g+1)}}}{\sum_{i=1}^{g} u^2_{si}} \quad \text{a} \quad \sigma_{\phi_g\phi_h} = \frac{\sigma_{u_{s(g+1)}u_{s(h+1)}}}{\sqrt{\sum_{i=1}^{g} u^2_{si}\sum_{j=1}^{h} u^2_{sj}}}.$$

1.4.2. Odmocninová transformace

Hlavní myšlenkou přístupu, který byl nejprve spojován se jménem M. A. Stephense, je založen na transformaci kompozičních dat na data směrová tzv. odmocninovou transformací ve tvaru

$$\mathbf{y} = \sqrt{\mathbf{u}} = (\sqrt{u_1}, \sqrt{u_2}, \ldots, \sqrt{u_p})^T, \quad \sum_{i=1}^{p} u_i = 1. \tag{1.9}$$

Tato transformace zobrazuje kompoziční data (včetně nul) na povrch $(p-1)$-rozměrné jednotkové hyperkoule a poté data modelujeme vhodným rozdělením na hyperkouli. Scealy a Welsh v [19], [20], [21] využívají k modelování takto transformovaných dat tzv. Kentovo rozdělení pro směrová data, neboť má stejný počet parametrů jako $(p-1)$-rozměrné normální rozdělení a tyto parametry lze smysluplně interpretovat. Dalšími důvody jsou již zmíněná přirozená práce s nulami a také dostatečně obecná varianční struktura. Tomuto přístupu, konkrétně

regresním modelům s využitím uvedené odmocninové transformace, je věnována celá kapitola 2.

1.5. Principy statistické analýzy kompozičních dat - jejich vznik a vliv na alternativní přístupy

Následující podkapitola byla vytvořena ze zdrojů [2], [3], [4], [5], [14], [18], [23], [24], [25], [26].

Analýza kompozičních dat se značně rozvinula na základě výměny názorů skrze vědecký časopis *Mathematical Geology* v letech 1989 až 1992, a to především mezi J. Aitchisonem, D. F. Watsonem spolu s G. M. Philipem a C. Stanleym. Jak jsme naznačili v podkapitole 1.2, tak základní Aitchisonův přístup v [4], [5] spočívá v náhledu na kompoziční data jako na množinu vektorů na simplexu. Tyto vektory následně transformuje do eukleidovského prostoru prostřednictvím transformací, které jsou založeny na logaritmech podílů jednotlivých složek kompozic. U takto transformovaných dat poté předpokládá vlastnosti standardních mnohorozměrných.

Watson a Philip v [26] měli odlišný pohled, než jaký měl Aitchison. Používali směrové kosiny k zobrazení kompozic, na které nahlíželi jako na radiální vektory, do kladného ortantu hyperkoule. Tvrdili, že rozdíly mezi kompozicemi jsou rozdíly radiálních směrů, a tedy je nutné použití měr, které zahrnují jejich úhlové vztahy. Vhodné míry by tedy měly být rotačně invariantní. Dospěli k názoru, že logpodílové metody k těmto vhodným mírám nevedou a simplex je tak pro ně nevhodným prostorem pro analýzu kompozičních dat.

Stanley v [23] vnesl do této problematiky nový pohled. Tvrdil, že předmětem analýzy by měl být neuzavřený vektor $\mathbf{b}$ kompozice $\mathbf{u} = \mathcal{C}(\mathbf{b})$, neboť zachování vlastností $\mathbf{b}$ při analýze založené na $\mathbf{u}$ je důležité. Dále tvrdil, že stejně jako poměry, tak i úhly mohou být použity k porovnávání kompozic, protože jsou stejné pro $\mathbf{u}$ i $\mathbf{b}$. Následně uvedl příklad, kde u čtyř dvousložkových kompozic $\mathbf{a} = (0.1, 0.9)^T, \mathbf{b} = (0.2, 0.8)^T, \mathbf{c} = (0.5, 0.5)^T, \mathbf{d} = (0.6, 0.4)^T$ sledoval rozdíly v poměrech a úhlech mezi dvojicemi kompozic $\mathbf{a}, \mathbf{b}$ a $\mathbf{c}, \mathbf{d}$. Na základě stejných

rozdílů mezi kompozicemi v absolutním měřítku a značně odlišných hodnot úhlů a poměrů mezi těmito dvojicemi usoudil, že jsou rozdíly v absolutním měřítku nesprávné, tedy se ke kompozicím nelze chovat jako k bodům ve standardní eukleidovské geometrii.

Aitchison v [2] komentuje Watsonovy a Philipovy argumenty v [26] odkazem na tzv. *invarianci na změnu měřítka*, pro kterou míra variability musí splňovat vztah

$$\rho(a_1u_1, a_2u_2, \ldots, a_nu_n) = \rho(u_1, u_2, \ldots, u_n),\ a_i > 0, i = 1, 2, \ldots, n.$$

Poukázal na to, že ρ má tuto vlastnost jen tehdy, lze-li ji vyjádřit ve smyslu poměrů složek v kompozici. Dále uvedl, že eukleidovská vzdálenost mezi kompozicemi zvaná Aitchisonova vzdálenost, která je definována pro dvě kompozice $\mathbf{u}, \mathbf{v} \in \Delta^p$ dle (1.1), tuto invarianční vlastnost má. Dále konstatoval, že ani míry založené na směrových kosinech tuto vlastnost nepostrádají, nicméně tímto motivoval k použití logpodílového přístupu.

Dále už komunikace probíhala pouze mezi Aitchisonem a Watsonem. Watson v [24] vysvětlil, že invariance na změnu měřítka není to, co měli s Philipem na mysli. Jako příklad uvedl čtyři různé kompozice, mezi kterými naměřil stejné úhly a dospěl k závěru, že Aitchisonova vzdálenost dává desetkrát horší výsledek než eukleidovská vzdálenost, která je sama o sobě nesprávná. Interpretoval to tak, že by kompoziční data neměla být analyzována na simplexu, protože eukleidovská vzdálenost není invariantní (na rotaci) a transformace (jiné než úhlové) vytvářejí zkreslení. Tvrdil, že jediná správná a jednoznačná míra vzdálenosti mezi dvěma p-složkovými kompozicemi je p-rozměrný prostorový úhel, který radiálně navzájem odděluje dvě kompozice.

Aitchison v [3] odůvodněně odpověděl, že úhly jsou opravdu invariantní mírou variability mezi kompozicemi, nicméně nejsou jedinou takovou mírou a konstatoval, že Watsonovy argumenty pramení z chybného předpokladu jedinečnosti úhlů jakožto míry variability mezi kompozicemi. Dále ukázal na příkladu, že vhodná míra vzdálenosti závisí na kontextu.

Watson v [25] nebyl s Aitchisonovým závěrem spokojen, ale přesto pone-

chal Aitchisonovi poslední slovo. Na základě toho uvedl Aitchison jeho pohled na invarianci na změnu měřítka jako na hlavní předpoklad relevantní analýzy kompozičních dat a navíc k ní dodal další kritéria, která dohromady tvoří tzv. principy statistické analýzy kompozičních dat (z angl. The Principles of Compositional Data Analysis). Konkrétně se jedná o tyto předpoklady:

- *Invariance vůči perturbaci*: $\mathrm{d}(\mathbf{v} \oplus \mathbf{u}_1, \mathbf{v} \oplus \mathbf{u}_2) = \mathrm{d}(\mathbf{u}_1, \mathbf{u}_2)$ musí platit pro libovolnou perturbaci $\mathbf{v}$, kde $\mathbf{v} \oplus \mathbf{u} = \mathcal{C}(v_1u_1, v_2u_2, \ldots, v_pu_p)$ pro každý kladný vektor $\mathbf{v}$.

- *Invariance vůči permutaci*: $\mathrm{d}(\mathbf{P}\mathbf{u}_1, \mathbf{P}\mathbf{u}_2) = \mathrm{d}(\mathbf{u}_1, \mathbf{u}_2)$ musí platit pro libovolnou permutační matici $\mathbf{P}$.

- *Podkompoziční dominance*: Vzdálenost mezi dvěma podkompozicemi nesmí být větší než vzdálenost mezi dvěma kompozicemi. Musí tedy platit nerovnost $\mathrm{d}(\mathbf{u}_1, \mathbf{u}_2) \geq \mathrm{d}(\mathcal{C}(\mathbf{S}\mathbf{u}_1), \mathcal{C}(\mathbf{S}\mathbf{u}_2))$.

- *Podkompoziční soudržnost*: Výsledky analýzy podkompozic musí být v souladu s výsledky analýzy celých kompozic. Jinak řečeno bychom měli získat stejnou informaci o složkách podkompozic a vztazích mezi nimi, ať už analyzujeme celý kompoziční soubor, nebo jen jeho redukovaný tvar.

Dále Aitchison uvedl příklad, aby poukázal na skutečnost, že úhly nesplňují výše uvedenou podkompoziční dominanci. Tím obhájil používání logpodílových transformací a logpodílovou analýzu tak povýšil nad všechny známé alternativní přístupy.

Po přiblížení vzniku principů statistické analýzy kompozičních dat se nyní zaměříme především na odlišné názory v chápání těchto principů a také na to, jak ovlivnily zastánce sférického přístupu. Na závěr celé kapitoly bude uveden příklad, jehož cílem je zpochybnit platnost principů statistické analýzy kompozičních dat jakožto nutné podmínky pro relevantní statistickou analýzu těchto dat.

Invariance na změnu měřítka a podkompoziční soudržnost jsou silně založeny na pohledu, že předmětem analýzy jsou jednotlivé složky kompozice (samozřejmě

ve vztahu ke složkám ostatním) a jediným způsobem, jak je vyšetřovat jako relativní hodnoty, je užití poměrů. Alternativním pohledem, který je vlastní transformacím do kladného ortantu hyperkoule, je analyzovat nikoliv jednotlivé složky, ale analyzovat kompozice jako takové. Pak by tedy kompozice měly být vybrány zejména na základě kontextu a předmětem zájmu není jejich konkrétní podkompozice. V tomto případě není důvod předpokládat, že objekty různých dimenzí mají stejné vlastnosti a stejně tak není důvod předpokládat odvozování vlastností z těchto navzájem různých objektů. Budeme-li tedy pracovat přímo s kompozicemi, nevzniká nutnost předpokladu podkompoziční soudržnosti.

Za zmínku stojí také pohled na podkompoziční soudržnost v literatuře [14], kde autoři tvrdí, že podkompozice by se měly chovat jako ortogonální projekce ve standardní reálné analýze. Podkompoziční soudržnost je dle nich možné prakticky vyjádřit jako „shrnutí“ podkompoziční dominance a invariance na změnu měřítka. To se ovšem zdá být dle autorů v [18] slabší předpoklad, než jak jej mínil Aitchison.

Závěrem kapitoly naznačíme hlavní myšlenku příkladu, který je podrobně popsán ve zdroji [18].

Příklad 1.5.1 *Nechť* $\mathbf{q} = (q_1, q_2, q_3)^T$ *je normálně rozdělený náhodný vektor, dále nechť*

$$r_j = \frac{e^{q_j}}{1 + e^{q_j}}, \quad j = 1, 2, 3, 4.$$

Nechť je dána kompozice $\mathbf{u} = (u_1, u_2, u_3, u_4)^T$, *kde*

$$u_1 = r_1 r_2, \quad u_2 = (1 - r_1) r_3, \quad u_3 = (1 - r_1)(1 - r_3), \quad u_4 = r_1 (1 - r_2).$$

Standardní logpodílové souřadnice jsou ve tvaru

$$alr(\mathbf{u}) = \begin{pmatrix} q_2 \\ -q_1 + q_3 + \ln\left(1 + e^{q_2}\right) - \ln\left(1 + e^{q_3}\right) \\ -q_1 + \ln\left(1 + e^{q_2}\right) - \ln\left(1 + e^{q_3}\right) \end{pmatrix},$$

$$clr(\mathbf{u}) = 1/4 \begin{pmatrix} 2q_1 + 3q_2 - q_3 - 2\ln\left(1 + e^{q_2}\right) + 2\ln\left(1 + e^{q_3}\right) \\ -2q_1 - q_2 + 3q_3 + 2\ln\left(1 + e^{q_2}\right) - 2\ln\left(1 + e^{q_3}\right) \\ -2q_1 - q_2 - q_3 + 2\ln\left(1 + e^{q_2}\right) - 2\ln\left(1 + e^{q_3}\right) \\ 2q_1 - q_2 - q_3 - 2\ln\left(1 + e^{q_2}\right) + 2\ln\left(1 + e^{q_3}\right) \end{pmatrix},$$

$$ilr(\mathbf{u}) = \begin{pmatrix} 1/\sqrt{2}\,(q_1 + q_2 - q_3 - \ln(1+e^{q_2}) + \ln(1+e^{q_3})) \\ 1/\sqrt{6}\,(q_1 + q_2 + q_3 - \ln(1+e^{q_2}) + \ln(1+e^{q_3})) \\ 1/\sqrt{12}\,(-2q_1 + q_2 + q_3 + 2\ln(1+e^{q_2}) - 2\ln(1+e^{q_3})) \end{pmatrix},$$

$$mlr(\mathbf{u}) = \begin{pmatrix} q_1 + q_2 - \ln(1+e^{q_2}) - \ln(1 + e^{q_1}/(1+e^{q_2})) \\ q_3 - \ln(1+e^{q_3}) - \ln(1/(1+e^{q_3}) + e^{q_1}/(1+e^{q_2})) \\ -q_1 + \ln(1+e^{q_2}) - \ln(1+e^{q_3}) \end{pmatrix}.$$

Je zřejmé, že marginální rozdělení uvedených souřadnic, mimo souřadnici $alr(\mathbf{u})_1 = q_2$*, jsou asymetrická. Tedy žádná ze standardních logpodílových transformací nemá mnohorozměrné normální rozdělení pravděpodobnosti. Budeme-li uvažovat podkompozici* $\mathbf{u}'$ *založenou na druhé a třetí složce kompozice* $\mathbf{u}$ *ve tvaru* $\mathbf{u}' = (r_3, 1-r_3)^T$*, pak jsou její logpodílové souřadnice ve tvaru*

$$alr(\mathbf{u}') = q_3, \quad clr(\mathbf{u}') = (1/2\ q_3, -1/2\ q_3)^T, \quad ilr(\mathbf{u}') = 1/\sqrt{2}q_3.$$

Jedná se tedy o lineární logpodílové souřadnice. Vzhledem k asymetrii se používají jako odhady mediány, které jsou nelineárními odhady polohy. Medián uvedené lineární transformace není (ani asymptoticky) lineární transformací mediánů původního rozdělení. U multiplikativní logpodílové transformace dokonce ani nelze $mlr(\mathbf{u}')_2 = q_3$ *vyjádřit z mlr souřadnic původní kompozice* $\mathbf{u}$*. V tomto případě tedy standardní logpodílová analýza na celé kompozici není v souladu s analýzou na podkompozicích, tedy nesplňuje podkompoziční soudržnost.*

Na příkladu autoři v [18] demonstrovali svůj pohled založený na tom, že i logpodílová metodika má v určitých případech problémy s konzistencí. Konstatují, že podkompoziční soudržnost tedy v některých případech nesplňuje ani logpodílová metodika, která byla na základě této vlastnosti povýšena před ostatní známé alternativní přístupy. Autoři navrhují zamítnutí podkompoziční soudržnosti jakožto předpokladu relevantní statistické analýzy kompozičních dat, nikoliv však zamítnutí zároveň logpodílové metodiky, kterou označují jako vhodnou pro určitý typ problémů.

Lze si povšimnout, že ztěžejní rozdíl mezi logpodílovým přístupem a alternativními přístupy v čele se sféricisty je ve skutečnosti, že logpodílová metodika se na konstantní součet nejen že nefixuje, ale je pro ní dokonce irelevantní. To je

zřejmě i příčinou zavádějícího myšlenkového postupu v příkladu 1.5.1. Vzhledem k faktu, že se kompoziční data neomezují na proporciální reprezentaci, zdá se být logpodílový přístup v porovnání s ostatními nejobecnějším.

Kapitola 2

Regresní model založený na odmocninové transformaci

Následující podkapitola se zaměřuje na specifický regresní model, který je navržen pro data, která získáme použitím odmocninové transformace definované vztahem (1.9). Nejprve představíme model a seznámíme se s vhodným rozdělením pravděpodobnosti pro modelování těchto dat. Dále se zaměříme na jeho asymptotické chování a odhad parametrů. Nakonec poukážeme na jeho přednosti a nedostatky v porovnání s logpodílovou metodikou.

Hlavním zdrojem informací pro tvorbu celé kapitoly byly publikace [20], [21]. Dále bylo čerpáno z [12], [19].

2.1. Obecný Kentův regresní model

Jak již bylo zmíněno v podkapitole 1.4.2, odmocninová transformace zobrazuje kompoziční data (a to včetně nul) na povrch $(p-1)$-rozměrné hyperkoule. Nechť $S^{p-1} = \{\mathbf{y} \in \mathbb{R}^p \mid \|\mathbf{y}\| = 1\}$, kde $\|.\|$ značí eukleidovskou normu, je p-rozměrná hyperkoule s jednotkovým poloměrem. Nechť $\{\mathbf{y}_i \in S^{p-1} \mid i = 1, 2, \ldots, n\}$ jsou vysvětlované proměnné a $\{\mathbf{x}_i \in \mathbb{R}^q \mid i = 1, 2, \ldots, n\}$ jsou vysvětlující proměnné. Dále nechť jsou vysvětlované proměnné vzájemně nezávislé a mají Kentovo rozdělení dané hustotou

$$f(\mathbf{y}|\mathbf{x}) = c(\kappa, \boldsymbol{\beta})^{-1} \exp\Big[\kappa \boldsymbol{\mu}(\mathbf{x})^T \mathbf{y} + \beta_2(\boldsymbol{\gamma}_2(\mathbf{x})^T \mathbf{y})^2 + \cdots + \beta_{p-1}(\boldsymbol{\gamma}_{p-1}(\mathbf{x})^T \mathbf{y})^2$$

$$- (\beta_2 + \beta_3 + \cdots + \beta_{p-1})(\boldsymbol{\gamma}_p(\mathbf{x})^T\mathbf{y})^2\Big], \qquad \mathbf{y} \in S^{p-1}, \tag{2.1}$$

kde $\boldsymbol{\mu}(\mathbf{x})$ je průměrný směr, $\boldsymbol{\gamma}_2(\mathbf{x}), \boldsymbol{\gamma}_3(\mathbf{x}), \ldots, \boldsymbol{\gamma}_p(\mathbf{x})$ jsou p-rozměrné ortonormální vektory (osy), které jsou kolmé na $\boldsymbol{\mu}(\mathbf{x}), \kappa > 0$ *a* $\boldsymbol{\beta} = (\beta_2, \beta_3, \ldots, \beta_{p-1})^T \in \mathbb{R}^{p-2}$ jsou parametry určující tvar a splňující nerovnost

$$\frac{\kappa}{2} > \beta_2 \geq \beta_3 \geq \cdots \geq \beta_{p-1} \geq -(\beta_2 + \beta_3 + \cdots + \beta_{p-1})$$

a $c(\kappa, \boldsymbol{\beta})$ je normalizační konstanta.

Je známo, že v regresních modelech můžeme zvolit, zda jsou na proměnné $\mathbf{x}$ závislé všechny parametry, nebo jen jejich část. Tento model předpokládá, že $\boldsymbol{\mu}$ je funkcí proměnné $\mathbf{x}$ (tj. $\boldsymbol{\mu}$=$\boldsymbol{\mu}(\mathbf{x})$) a ostatní parametry κ, $\boldsymbol{\beta}$ a $\mathbf{K}^*$(uvedené níže) jsou konstanty nezávislé na $\mathbf{x}$.

Je zvykem sloučit $\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\gamma}_2(\mathbf{x}), \ldots, \boldsymbol{\gamma}_p(\mathbf{x})$ do $(p \times p)$-rozměrné matice

$$\boldsymbol{\Gamma}(\mathbf{x}) = (\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\gamma}_2(\mathbf{x}), \ldots, \boldsymbol{\gamma}_p(\mathbf{x}))$$

a ortogonální matici $\boldsymbol{\Gamma}(\mathbf{x})$ modelujeme jako

$$\boldsymbol{\Gamma}(\mathbf{x}) = \mathbf{H}(\mathbf{x})\mathbf{K} = (\boldsymbol{\mu}(\mathbf{x}), \mathbf{H}^*(\mathbf{x})\mathbf{K}^*),$$

kde

$$\mathbf{H}(\mathbf{x}) = (\boldsymbol{\mu}(\mathbf{x}), \mathbf{H}^*(\mathbf{x})) \text{ a } \mathbf{K} = \begin{pmatrix} 1 & \mathbf{0}_{p-1}^T \\ \mathbf{0}_{p-1} & \mathbf{K}^* \end{pmatrix}$$

jsou $(p \times p)$-rozměrné ortogonální matice a $\mathbf{0}_{p-1}$ je $(p-1)$-rozměrný vektor nul. Dále se definuje $\mathbf{D}_c = \mathrm{diag}(0, \beta_2, \beta_3, \ldots, \beta_{p-1}, -\sum_{m=2}^{p-1} \beta_m)$ a hustotu danou vztahem (2.1) lze psát v úspornějším maticovém tvaru

$$f(\mathbf{y}|\mathbf{x}) = c(\kappa, \boldsymbol{\beta})^{-1} \exp\Big(\kappa\boldsymbol{\mu}(\mathbf{x})^T\mathbf{y} + \mathbf{y}^T\boldsymbol{\Gamma}(\mathbf{x})\mathbf{D}_c\boldsymbol{\Gamma}(\mathbf{x})^T\mathbf{y}\Big)$$

a pro normalizační konstantu Kentova rozdělení pak můžeme psát vztah

$$c(\kappa, \boldsymbol{\beta}) = \int_{\mathbf{y} \in S^{p-1}} \exp\Big(\kappa\boldsymbol{\mu}(\mathbf{x})^T\mathbf{y} + \mathbf{y}^T\boldsymbol{\Gamma}(\mathbf{x})\mathbf{D}_c\boldsymbol{\Gamma}(\mathbf{x})^T\mathbf{y}\Big)\mathrm{d}\mathbf{y}. \tag{2.2}$$

Než uvedeme způsoby, jak zvolit $\mathbf{H}^*(\mathbf{x})$ a $\mathbf{K}^*$, zmíníme ještě lemma, na jehož základě definujeme v textu dále používanou transformaci. Důkaz tohoto lemmatu lze najít v literatuře [21].

Lemma 2.1.1 *Nechť* $\mathbf{y}|\mathbf{x}$ *má Kentovo rozdělení s parametry* $\kappa, \boldsymbol{\beta}$ *a* $\boldsymbol{\Gamma}(\mathbf{x})$. *Nechť* $\mathbf{C}$ *je* $(p \times p)$*-rozměrná ortogonální matice. Pak transformace* $\mathbf{y}^{(C)} = \mathbf{C}\mathbf{y}$ *má opět Kentovo rozdělení se stejnými parametry* $\kappa, \boldsymbol{\beta}$ *avšak s ortogonální maticí rovnou* $\mathbf{C}\boldsymbol{\Gamma}(\mathbf{x})$.

Transformace

$$\mathbf{y}^* = \boldsymbol{\Gamma}(\mathbf{x})^T\mathbf{y} = (y_1^*, y_2^*, \ldots, y_p^*) \in S^{p-1} \tag{2.3}$$

se nazývá *populační standardizační transformace*, neboť její hustota je funkcí pouze parametrů určujících tvar a tudíž $\mathbf{y}^*$ lze považovat za standardizovanou formu Kentova rozdělení s ortogonální maticí rovnou jednotkové matici.

2.2. Multiplikativní a aditivní Kentův regresní model

Následující text se bude zabývat volbou $\mathbf{K}^*, \mathbf{H}^*(\mathbf{x})$ a $\boldsymbol{\mu}(\mathbf{x})$ v obecném modelu z podkapitoly 2.1 a na základě jejich tvaru pak uvedeme multiplikativní a aditivní Kentův regresní model.

Jednou z možných parametrizací matice $\mathbf{K}^*$ podle [12] je určit ji jako výsledek $(p-1)(p-2)/2$ rotací v rovině určených $(p-1) \times (p-1)$-rozměrnými maticemi $\mathbf{R}_{ij}^*(\psi)$, které jsou funkcemi $p-2$ rovníkových úhlů (tj. úhlů udávajích zeměpisnou délku) $\psi_1, \psi_2, \ldots, \psi_{p-2} \in \langle -\pi, \pi)$ a $(p-2)(p-3)/2$ poledníkových úhlů (tj. úhlů udávajících zeměpisnou šířku) $\nu_1, \nu_2, ..., \nu_{(p-2)(p-3)/2} \in \langle 0, \pi \rangle$. Matice $\mathbf{R}_{ij}^*(\psi)$ jsou jednotkové a na pozicích s příslušnými indexy $(i,i), (i,j), (j,i)$ a (j,j) jsou prvky po řadě nahrazeny funkcemi $\cos(\psi), -\sin(\psi), \sin(\psi)$ a $\cos(\psi)$. Například pro $p = 4$ jsou matice $\mathbf{R}_{ij}^*(\psi)$ ve tvaru

$$\mathbf{R}_{12}^*(\psi) = \begin{pmatrix} \cos(\psi) & -\sin(\psi) & 0 \\ \sin(\psi) & \cos(\psi) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

a

$$\mathbf{R}_{23}^*(\psi) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\psi) & -\sin(\psi) \\ 0 & \sin(\psi) & \cos(\psi) \end{pmatrix}.$$

Pro $p = 3$ je $\mathbf{K}^* = \mathbf{R}^*_{12}(\psi_1)$ a pro $p \geq 4$

$$\mathbf{K}^* = \left\{ \prod_{m=1}^{p-3} \left[\mathbf{R}^*_{12}(\psi_{p-m-1}) \prod_{j=1}^{p-m-2} \left(\mathbf{R}^*_{j+1,j+2} \left(\nu_{1-j+(p-m-1)(p-m-2)/2} \right) \right) \right] \right\} \mathbf{R}^*_{12}(\psi_1). \tag{2.4}$$

Způsob, jak parametrizovat průměrný směr $\boldsymbol{\mu}(\mathbf{x})$ spočívá v jeho vyjádření pomocí hypersférických polárních souřadnic a poté ve vyjádření těchto souřadnic lineární funkcí proměnné $\mathbf{x}$, kterou nazýváme tzv. spojovací funkcí proměnné $\mathbf{x}$. Průměrný směr $\boldsymbol{\mu}(\mathbf{x})$ uvažujeme ve tvaru

$$\mu_i(\mathbf{x}) = \begin{cases} \cos(\alpha_1(\mathbf{x})) & i = 1, \\ \left[\prod_{j=1}^{i-1} \sin(\alpha_j(\mathbf{x})) \right] \cos(\alpha_i(\mathbf{x})) & i \neq 1, i \neq p-1, i \neq p, \\ \left[\prod_{j=1}^{p-2} \sin(\alpha_j(\mathbf{x})) \right] \cos(\alpha_{p-1}(\mathbf{x})) & i = p-1, \\ \left[\prod_{j=1}^{p-2} \sin(\alpha_j(\mathbf{x})) \right] \sin(\alpha_{p-1}(\mathbf{x})) & i = p, \end{cases}$$

kde $\mu_i(\mathbf{x})$ je i-tá složka $\boldsymbol{\mu}(\mathbf{x})$ a restrikce $\alpha_1(\mathbf{x}), \alpha_2(\mathbf{x}), \ldots, \alpha_{p-1}(\mathbf{x}) \in (0, \pi/2)$ zajistí, že $\boldsymbol{\mu}(\mathbf{x})$ leží v kladném ortantu S^{p-1}. Pak můžeme psát

$$g_1(\alpha_1(\mathbf{x})) = \mathbf{a}_1^T \mathbf{x}, \;\; g_2(\alpha_2(\mathbf{x})) = \mathbf{a}_2^T \mathbf{x}, \;\; \ldots, \;\; g_{p-1}(\alpha_{p-1}(\mathbf{x})) = \mathbf{a}_{p-1}^T \mathbf{x},$$

kde g_j, $j = 1, 2, \ldots, p-1$ jsou spojovací funkce, $\mathbf{A} = (\mathbf{a}_1^T, \mathbf{a}_2^T, \ldots, \mathbf{a}_{p-1}^T) \in \mathbb{R}^{q \times (p-1)}$ je matice regresních parametrů. Je vhodné poznamenat, že nemusíme volit stejné funkce pro každé j. Jedním z možných tvarů spojovací funkce je log-tangentová funkce ve tvaru

$$g_j(\psi) = -2 \log(\mathrm{tg}(\psi)).$$

Použitím log-tangentové funkce pro $j = 1, 2, \ldots, p-1$ dostáváme

$$\mu_i(\mathbf{x}) = \begin{cases} \frac{\exp(\mathbf{a}_i^T \mathbf{x}/2)}{\sqrt{\prod_{j=1}^{i} (1 + \exp(\mathbf{a}_j^T \mathbf{x}))}} & i = 1, 2, \ldots, p-1, \\ \frac{1}{\sqrt{\prod_{j=1}^{p-1} (1 + \exp(\mathbf{a}_j^T \mathbf{x}))}} & i = p. \end{cases}$$

Další možný způsob, jak definovat vztah mezi $\boldsymbol{\mu}(\mathbf{x})$ a lineárními funkcemi vysvětlujících proměnných je

$$\mu_i(\mathbf{x}) = \begin{cases} \frac{1}{\sqrt{1+\sum_{j=1}^{p-1}\exp(\mathbf{a}_j^T\mathbf{x})}} & i = 1, \\ \frac{\exp(\mathbf{a}_{i-1}^T\mathbf{x}/2)}{\sqrt{1+\sum_{j=1}^{p-1}\exp(\mathbf{a}_j^T\mathbf{x})}} & i = 2, 3, \ldots, p. \end{cases} \tag{2.5}$$

Obecný Kentův regresní model z podkapitoly 2.1 vyžaduje výběr ortogonální matice $\mathbf{H}(\mathbf{x})$. Tu je možné vyjádřit pomocí matic rotací v rovině, a to jako

$$\mathbf{H}(\mathbf{x}) = \mathbf{R}_{(p-1)p}(\alpha_{p-1}(\mathbf{x}))\mathbf{R}_{(p-2)(p-1)}(\alpha_{p-2}(\mathbf{x}))\ldots\mathbf{R}_{12}(\alpha_1(\mathbf{x})). \tag{2.6}$$

Definice 2.2.1 *Multiplikativní Kentův regresní model je model definovaný v podkapitole 2.1 se spojovací funkcí* $\alpha_j(\mathbf{x}) = \operatorname{cotg}(\exp(-\mathbf{a}_j^T\mathbf{x}/2))$, $j = 1, 2, \ldots, p-1$, *kde je kotangentová funkce definována na* $(-\pi/2, \pi/2)$ *a s maticí* $\mathbf{H}(\mathbf{x})$ *ve tvaru (2.6).*

Multiplikativní Kentův model není invariantní vůči permutaci. Není-li předem rozmyšleno konkrétní uspořádání složek, doporučuje se uspořádat je tak, že

$$\overline{y}_1 \leq \overline{y}_2 \leq \cdots \leq \overline{y}_p, \quad \text{kde } \overline{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i = (\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_p)^T.$$

Důvodem, proč postrádá permutační invarianci je mimo jiné i skutečnost, že matice $\mathbf{H}(\mathbf{x})$ v (2.6) není symetrická. Její symetrická varianta je ve tvaru

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} \mu_1 & \boldsymbol{\mu}_L(\mathbf{x})^T \\ \boldsymbol{\mu}_L(\mathbf{x}) & \mathbf{H}_L^*(\mathbf{x}) \end{pmatrix}, \tag{2.7}$$

kde

$$\mathbf{H}_L^*(\mathbf{x}) = \frac{1}{1+\mu_1(\mathbf{x})}\,\boldsymbol{\mu}_L(\mathbf{x})\boldsymbol{\mu}_L(\mathbf{x})^T - \mathbf{I}_{p-1}, \quad \boldsymbol{\mu}_L(\mathbf{x}) = (\mu_2(\mathbf{x}), \mu_3(\mathbf{x}), \ldots, \mu_p(\mathbf{x}))^T.$$

Za předpokladu volby symetrické matice $\mathbf{H}(\mathbf{x})$ dle (2.7) a $\boldsymbol{\mu}(\mathbf{x})$ ve tvaru (2.5) je možné získat permutační ekvivarianci.

Definice 2.2.2 *Aditivní Kentův regresní model je model definovaný v podkapitole 2.1 s maticí* $\mathbf{H}(\mathbf{x})$ *ve tvaru (2.7) a s* $\boldsymbol{\mu}(\mathbf{x})$ *ve tvaru (2.5).*

Aditivní Kentův model je tedy permutačně ekvivariantní, platí následující věta, jenž je dokázána v [21].

Věta 2.2.1 *Předpokládejme, že* $\mathbf{y}|\mathbf{x}$ *se řídí aditivním Kentových regresním modelem a nechť* $\mathbf{A} = (\mathbf{a}_1^T, \mathbf{a}_2^T, \ldots, \mathbf{a}_{p-1}^T)$. *Dále nechť* $\mathbf{y}^{(C)} = \mathbf{C}\mathbf{y}$ *je permutací* $\mathbf{y}$ *takovou, že pozice* y_1 *zůstává nezměněna. Pak se* $\mathbf{y}^{(C)}$ *řídí aditivním Kentovým regresním modelem s permutovanými maticemi* $\mathbf{A}$ *a* $\mathbf{K}^*$ *a nezměněným parametrem* $\boldsymbol{\beta}$.

2.3. Asymptotické chování Kentova modelu a odhad jeho parametrů

Nejprve se zaměříme na momenty Kentova rozdělení. Pro první dva momenty platí vztahy ve tvaru

$$\mathsf{E}(\mathbf{y}|\mathbf{x}) = \boldsymbol{\mu}(\mathbf{x})\mathsf{E}(y_1^*) \tag{2.8}$$

$$\mathsf{E}(\mathbf{y}\mathbf{y}^T|\mathbf{x}) = \boldsymbol{\Gamma}(\mathbf{x})\ \mathrm{diag}(\mathsf{E}(y_1^{*2}), \mathsf{E}(y_2^{*2}), \ldots, \mathsf{E}(y_p^{*2}))\boldsymbol{\Gamma}(\mathbf{x})^T \tag{2.9}$$

Při analyzování kompozičních dat se zajímáme zejména o podmíněnou střední hodnotu $\mathsf{E}(\mathbf{u}|\mathbf{x})$, kde $\mathbf{u} = \mathbf{y}^2 = (y_1^2, y_2^2, \ldots, y_p^2)^T$. Úpravou vztahu (2.9) získáme pro $\mathsf{E}(\mathbf{u}|\mathbf{x})$ vztah

$$\mathsf{E}(\mathbf{u}|\mathbf{x}) = \boldsymbol{\Gamma}(\mathbf{x})^2\mathsf{E}(\mathbf{y}^{*2}).$$

Očekávánou hodnotu $\mathsf{E}(\mathbf{y}^{*2})$ nelze algebraicky zjednodušit, neboť normalizační konstanta $c(\kappa, \boldsymbol{\beta})$ ve tvaru (2.2) neexistuje v uzavřeném tvaru (jedná se o $(p-1)$-rozměrný integrál). Je však možné stanovit její aproximaci za předpokladu, že $\kappa \to \infty$. K tomuto účelu se využívá následující populační standardizační transformace $\mathbf{z}^* = (z_2^*, z_3^*, \ldots, z_p^*)^T$ taková, že platí

$$y_1^* = 1 - \frac{\|\mathbf{z}^*\|^2}{2} \text{ a } y_m^* = \sqrt{1 - \frac{\|\mathbf{z}^*\|^2}{4}} z_m^*, \text{ kde } \|\mathbf{z}^*\| \leq 2,\ m = 2, 3, \ldots, p.$$

Dále zmíníme větu o chování $\mathbf{z}^*$ při $\kappa \to \infty$, která je dokázana v [21].

Věta 2.3.1 *Předpokládejme, že* $\kappa \to \infty$ *a* $\beta_m/\kappa \to d_m$, $m = 2, 3, \ldots, p-1$ *a platí*

$$\frac{1}{2} > d_2 \geq d_3 \geq \cdots \geq d_{p-1} \geq -(d_2 + d_3 + \cdots + d_{p-1}).$$

Pak $z_2^*, z_3^*, \ldots, z_p^*$ *jsou asymptoticky nezávislé s rozdělením*

$$\sqrt{\kappa} z_m^* \sim \mathcal{N}\left(0, \frac{1}{1-2d_m}\right) \ , \ \ m = 2, 3, \ldots, p-1$$

a

$$\sqrt{\kappa} z_p^* \sim \mathcal{N}\left(0, \frac{1}{1+2\sum_{m=2}^{p-1} d_m}\right).$$

Navíc také platí

$$|\sqrt{\kappa}(z_2^*, z_3^*, \ldots, z_p^*)^T - \sqrt{\kappa}(y_2^*, y_3^*, \ldots, y_p^*)^T| \xrightarrow{P} \mathbf{0}_{p-1}^T.$$

Z věty 2.3.1 vyplývá, že $\mathbf{y}^* = (1, z_2^*, z_3^*, \ldots, z_p^*)^T + \mathbf{O}_P(1/\kappa)$ a asymptoticky je $\mathbf{y}^*$ tvořeno $p-1$ nezávislými přibližně normálně rozdělenými složkami. Symbol $\mathbf{O}_P(a)$, kde $a \in \mathbb{R}_+$ značí tvz. O–notaci v pravděpodobnosti, která označuje stochastickou ohraničenost posloupnosti náhodných vektorů. Pro posloupnost náhodných veličin se tato notace značí $O_P(a)$, kde $a \in \mathbb{R}_+$. Nechť $\{X_n\}_{n\in\mathbb{N}}$ je posloupnost náhodných veličin, pak $X_n = O_P(a)$ právě tehdy, když platí $\forall \epsilon > 0 \ \exists c_\epsilon \in \mathbb{R}, n_\epsilon \in \mathbb{N} : \mathsf{P}(|X_n| \leq c_\epsilon a) > 1 - \epsilon, \forall n \geq n_\epsilon$. Říkáme, že posloupnost náhodných veličin $\{X_n/a\}_{n\in\mathbb{N}}$ je stochasticky shora omezená. Potom pro posloupnost náhodných vektorů $\{\mathbf{X}_n\}_{n\in\mathbb{N}}$ platí, že $\mathbf{X}_n = \mathbf{O}_P(a)$ právě tehdy, když pro jednotlivé složky náhodného vektoru $\mathbf{X}_n = (X_n^1, X_n^2, \ldots, X_n^p)^T$ platí $X_n^k = O_P(a),\ k = 1, 2, \ldots, p$. Uvedené informace o O-notaci byly čerpány z [8].

Jelikož z (2.3) po úpravě plyne, že

$$\mathbf{y} = \mathbf{H}(\mathbf{x})\mathbf{K}\mathbf{y}^* = (\boldsymbol{\mu}(\mathbf{x}), \mathbf{H}^*(\mathbf{x})\mathbf{K}^*)\mathbf{y}^*,$$

pak

$$\mathbf{y} = \boldsymbol{\mu}(\mathbf{x}) + \mathbf{H}^*(\mathbf{x})\mathbf{K}^*\mathbf{z}^* + \mathbf{O}_P\left(1/\kappa\right)$$

a pro $\mathbf{y}^2 = (y_1^2, y_2^2, \ldots, y_p^2)^T$ platí

$$\mathbf{y}^2 = \boldsymbol{\mu}(\mathbf{x})^2 + 2\mathrm{diag}(\boldsymbol{\mu}(\mathbf{x}))\mathbf{H}^*(\mathbf{x})\mathbf{K}^*\mathbf{z}^* + \mathbf{O}_P\left(1/\kappa\right).$$

Tedy pro κ velké a $\boldsymbol{\mu}(\mathbf{x})$ nenulové má $\mathbf{u} = \mathbf{y}^2$ přibližně mnohorozměrné normální rozdělení a platí

$$\mathbf{u}|\mathbf{x} \sim \mathcal{N}_{p-1}\Big[\boldsymbol{\mu}(\mathbf{x})^2, 4\mathrm{diag}(\boldsymbol{\mu}(\mathbf{x}))\mathbf{H}^*(\mathbf{x})\mathbf{K}^*\mathbf{D}_\beta\mathbf{K}^{*T}\mathbf{H}^*(\mathbf{x})^T\mathrm{diag}(\boldsymbol{\mu}(\mathbf{x}))\Big],$$

kde

$$\mathbf{D}_\beta = \mathrm{diag}\left(\frac{1}{\kappa - 2\beta_2}, \frac{1}{\kappa - 2\beta_3}, \ldots, \frac{1}{\kappa - 2\beta_{p-1}}, \frac{1}{\kappa + 2\sum_{m=2}^{p-1}\beta_m}\right).$$

Ke stanovení odhadu $\widehat{\mathbf{a}}$ parametru $\mathbf{a}$ se využívá asymptotická metoda maximální věrohodnosti. Nechť je dán nezávislý výběr o rozsahu n z modelu definovaném v podkapitole 2.1. Logaritmická věrohodnostní funkce pro $\mathbf{a}$ je pak ve tvaru

$$\kappa\sum_{i=1}^{n}\boldsymbol{\mu}(\mathbf{x}_i)^T\mathbf{y}_i + \sum_{i=1}^{n}\mathbf{y}_i^T\boldsymbol{\Gamma}(\mathbf{x}_i)\mathbf{D}_c\boldsymbol{\Gamma}(\mathbf{x}_i)^T\mathbf{y}_i.$$

Z věty 2.3.1 plyne, že pro $\kappa \to \infty$ platí

$$\kappa\boldsymbol{\mu}(\mathbf{x})^T\mathbf{y} = \kappa + O_P(1), \quad \mathbf{y}^T\boldsymbol{\Gamma}(\mathbf{x})\mathbf{D}_c\boldsymbol{\Gamma}(\mathbf{x})^T\mathbf{y} = O_P(1).$$

Vzhledem k tomu, že první ze sčítanců dominuje hodnotu věrohodnostní funkce, potom odhad $\widehat{\mathbf{a}}$ hledáme jako řešení soustavy rovnic

$$\sum_{i=1}^{n}\frac{\partial\boldsymbol{\mu}(\mathbf{x}_i)}{\partial\mathbf{a}}\mathbf{y}_i = \mathbf{0}.$$

Odhad $\widehat{\mathbf{a}}$ bude blízký maximálně věrohodnému odhadu, je-li κ velké, nebo je-li β_2/κ malé. Lze ukázat, že uvedený odhad je nestranný (viz [21]).

Pro odhad $\mathbf{K}$ nejprve definujeme populační transformaci ve tvaru

$$\mathbf{y}_i^{**} = \mathbf{H}(\mathbf{x}_i)^T\mathbf{y}_i = \mathbf{H}(\mathbf{x}_i)^T\boldsymbol{\Gamma}(\mathbf{x}_i)\mathbf{y}_i^* = \mathbf{K}\mathbf{y}_i^* = (y_{1,i}^{**}, y_{2,i}^{**}, \ldots, y_{p,i}^{**})^T, \quad i = 1, 2, \ldots, n, \tag{2.10}$$

kde $\mathbf{y}_i^*$ je populační standardizační transformace z (2.3), platí $\mathbf{y}_i^* = \boldsymbol{\Gamma}(\mathbf{x}_i)^T\mathbf{y}_i = (y_{1,i}^*, y_{2,i}^*, \ldots, y_{p,i}^*)^T$, $i = 1, 2, \ldots, n$. Z lemmatu 2.1.1 vyplývá, že tato populační standardizační transformace má Kentovo rozdělení s ortogonální maticí ve tvaru $\mathbf{H}(\mathbf{x}_i)^T\boldsymbol{\Gamma}(\mathbf{x}_i) = \mathbf{K}$. Z (2.8) a (2.9) tedy po dosazení vyplývá, že

$$\mathsf{E}(\mathbf{y}_i^{**}) = \boldsymbol{\mu}^*\mathsf{E}(y_{1,i}^*) \quad a \quad \mathsf{E}(\mathbf{y}_i^{**}(\mathbf{y}_i^{**})^T) = \mathbf{K}\ \mathrm{diag}\ (\mathsf{E}(y_{1,i}^{*2}), \mathsf{E}(y_{2,i}^{*2}), \ldots, \mathsf{E}(y_{p,i}^{*2}))\mathbf{K}^T,$$

kde $\boldsymbol{\mu}^* = (1, 0, \ldots, 0)^T$. Výběrová verze vztahu (2.10) je tak ve tvaru

$$\widetilde{\mathbf{y}}_i^{**} = \widehat{\mathbf{H}}(\mathbf{x}_i)^T\mathbf{y}_i = (\widetilde{y}_{1,i}^{**}, \widetilde{y}_{2,i}^{**}, \ldots, \widetilde{y}_{p,i}^{**})^T, \qquad i = 1, 2, \ldots, n,$$

kde matice $\widehat{\mathbf{H}}(\mathbf{x}_i)$ je dána vztahem (2.6) nebo (2.7) dle vybraného modelu a odhad $\widehat{\mathbf{a}}$ nahrazuje $\mathbf{a}$ ve vztazích pro výpočet složek $\boldsymbol{\mu}(\mathbf{x})$. Nechť $\mathbf{S}^{**} = 1/n \sum_{i=1}^n \widetilde{\mathbf{y}}_i^{**}(\widetilde{\mathbf{y}}_i^{**})^T$ a dále nechť $\mathbf{S}_L^{**}$ je $(p-1) \times (p-1)$-rozměrná matice, která vznikne odstraněním prvního řádku a sloupce matice $\mathbf{S}^{**}$. Momentový odhad matice $\mathbf{K}^*$ lze získat spektrálním rozkladem

$$\mathbf{S}_L^{**} = \mathbf{K}^*\mathbf{D}_\lambda\mathbf{K}^{*T}, \tag{2.11}$$

kde $\mathbf{D}_\lambda = \mathrm{diag}(\lambda_2, \lambda_3, \ldots, \lambda_p)$ a výsledné odhady vlastních čísel označíme tak, že $\widehat{\lambda}_2 \geq \widehat{\lambda}_3 \geq \cdots \geq \widehat{\lambda}_p$. Pro určení jednoznačného tvaru $\widehat{\mathbf{K}}^*$ v rozkladu (2.11) je nutná restrikce rovníkových úhlů $\psi_1, \psi_2, \ldots, \psi_{p-2}$ pouze na $\langle 0, \pi\rangle$ a navíc musí být $\det(\mathbf{K}^*) = 1$, protože se jedná o matici rotace.

Dále tedy budeme pokračovat stanovením odhadů orientačních úhlů. Jak je stanovit ukážeme na případu, kdy $p = 4$. Použitím (2.4) dostáváme

$$\mathbf{K}^* = \mathbf{R}_{12}^*(\psi_2)\mathbf{R}_{23}^*(\nu_1)\mathbf{R}_{12}^*(\psi_1) =$$

$$= \begin{pmatrix} \cos(\psi_2) & -\sin(\psi_2) & 0 \\ \sin(\psi_2) & \cos(\psi_2) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\nu_1) & -\sin(\nu_1) \\ 0 & \sin(\nu_1) & \cos(\nu_1) \end{pmatrix} \begin{pmatrix} \cos(\psi_1) & -\sin(\psi_1) & 0 \\ \sin(\psi_1) & \cos(\psi_1) & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.12}$$

Nechť $\mathbf{F}$ je matice typu 3×3 tvořena vlastními vektory matice $\mathbf{S}_L^{**}$ dané vztahem (2.11), které jsou ve sloupcích seřazeny dle velikosti jejich odpovídajích vlastních čísel. Označme $\mathbf{F}_i$ sloupce matice $\mathbf{F}$, tj. $\mathbf{F} = (\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3)$. Vytvoříme nyní matici

$\mathbf{F}^*$ tak, že

$$\mathbf{F}_1^* = \begin{cases} -\mathbf{F}_1 & \text{je-li } F_{31} < 0, \\ \mathbf{F}_1 & \text{jinak,} \end{cases} \qquad \mathbf{F}_3^* = \begin{cases} -\mathbf{F}_3 & \text{je-li } F_{13} < 0, \\ \mathbf{F}_3 & \text{jinak} \end{cases}$$

a položíme $\mathbf{F}^* = (\mathbf{F}_1^*, \mathbf{F}_2, \mathbf{F}_3^*)$. Dále nechť

$$\mathbf{F}_2^* = \begin{cases} -\mathbf{F}_2 & \text{je-li } \det(\mathbf{F}^*) = -1, \\ \mathbf{F}_2 & \text{jinak,} \end{cases}$$

dále položíme $\mathbf{F}^{**} = (\mathbf{F}_1^*, \mathbf{F}_2^*, \mathbf{F}_3^*)$. Momentovým odhadem matice $\mathbf{K}^*$ je $\widehat{\mathbf{K}}^* = \mathbf{F}^{**}$ a momentové odhady orientačních úhlů ν_1, ψ_1, ψ_2 je možné jednoduše vyjádřit výpočtem součinu matic v (2.12)(stačí určit pouze členy K_{33}^*, K_{32}^* a K_{23}^*) a z rovnosti $\widehat{\mathbf{K}}^* = \mathbf{F}^{**}$ dostáváme

$$\widehat{\nu}_1 = \arccos(F_{33}^*), \qquad \widehat{\psi}_1 = \arccos\left(\frac{F_{32}^*}{\sin(\widehat{\nu}_1)}\right), \qquad \widehat{\psi}_2 = \arccos\left(\frac{-F_{23}^*}{\sin(\widehat{\nu}_1)}\right).$$

Nakonec ještě naznačíme, jak stanovit maximálně věrohodné odhady parametrů κ a $\boldsymbol{\beta}$. V případě, že by $\mathbf{K}$ a každé $\mathbf{H}(\mathbf{x}_i)$ byly známé, můžeme pomocí (2.3) vyčíslit hodnoty populační standardizační transformace

$$\mathbf{y}_i^* = \boldsymbol{\Gamma}(\mathbf{x}_i)^T \mathbf{y}_i = \mathbf{K}^T \mathbf{H}(\mathbf{x}_i)^T \mathbf{y}_i = (y_{1,i}^*, y_{2,i}^*, \ldots, y_{p,i}^*)^T, \tag{2.13}$$

kde $i = 1, 2, \ldots, n$. Logaritmická věrohodnostní funkce pro κ a $\boldsymbol{\beta}$ je

$$-n(\log(c) - \kappa) + \sum_{i=1}^{n} \Big[-\frac{\kappa}{2}(y_{1,i}^* - 1)^2 - \frac{1}{2}\mathbf{y}_{L,i}^{*T} \mathbf{D}_\beta^{-1} \mathbf{y}_{L,i}^* \Big],$$

kde $\mathbf{y}_{L,i}^* = (y_{2,i}^*, y_{3,i}^*, \ldots, y_{p,i}^*)^T$. S použitím věty 2.3.1 dostáváme $\mathbf{y}_{L,i}^* = \mathbf{O}(1/\sqrt{\kappa})$, $y_{1,i}^* = 1 + O_P(1/\kappa)$ a tedy výraz $-\kappa/2(y_{1,i}^* - 1)^2$ lze zanedbat. Pak pro maximálně věrohodné odhady platí vztahy

$$\frac{1}{n}\sum_{i=1}^{n} y_{m,i}^{*2} = \frac{1}{\kappa - 2\beta_m}, \ m = 2, 3, \ldots, p-1 \qquad \text{a} \qquad \frac{1}{n}\sum_{i=1}^{n} y_{p,i}^{*2} = \frac{1}{\kappa + 2\sum_{j=2}^{p-1}\beta_j}.$$

Vzhledem k tomu, že $\mathbf{K}$ a $\mathbf{H}(\mathbf{x}_i)$ jsou zpravidla neznámé, využijeme odhadů $\widehat{\mathbf{K}}$ a $\widehat{\mathbf{H}}(\mathbf{x}_i)$. Nejprve definujeme následující výběrovou standardizační transformaci,

která vznikne užitím odhadů $\widehat{\mathbf{K}}$ a $\widehat{\mathbf{H}}(\mathbf{x}_i)$ v (2.13), tj.

$$\widetilde{\mathbf{y}}_i^* = \widehat{\mathbf{K}}^T \widehat{\mathbf{H}}(\mathbf{x}_i)^T \mathbf{y}_i = (\widetilde{y}_{1,i}^*, \widetilde{y}_{2,i}^*, \ldots, \widetilde{y}_{p,i}^*)^T, \quad i = 1, 2, \ldots, n.$$

Lze ukázat (viz [21]), že $\widehat{\lambda}_m = 1/n \sum_{i=1}^n \widetilde{y}_{m,i}^{*2}$ pro $m = 2, 3, \ldots, p$, a uvedené odhady pak splňují následující vztahy

$$\widehat{\lambda}_m = \frac{1}{\kappa - 2\beta_m}, \quad m = 2, 3, \ldots, p-1 \qquad \text{a} \qquad \widehat{\lambda}_p = \frac{1}{\kappa + 2\sum_{j=2}^{p-1} \beta_j},$$

což vede k maximálně věrohodným odhadům

$$\widehat{\kappa}^{ML} = \frac{1}{p-1} \sum_{j=2}^{p} \frac{1}{\widehat{\lambda}_j},$$

$$\widehat{\beta}_m^{ML} = \frac{1}{2} \left(\frac{1}{p-1} \sum_{j=2}^{p} \frac{1}{\widehat{\lambda}_j} - \frac{1}{\widehat{\lambda}_m} \right), \quad m = 2, \ldots, p-1.$$

Tyto asymptotické maximálně věrohodné odhady tvarových parametrů jsou nevychýlené, jsou-li κ a n velké. Pro n malé jsou odhady značně vychýlené, a to i v případě, že κ je velké.

2.4. Vlastnosti Kentova regresního modelu a jejich srovnání s logpodílovou metodikou

Oba uvedené Kentovy modely, aditivní i multiplikativní, stejně jako ostatní možné varianty předpokládají, že složky $\boldsymbol{\mu}(\mathbf{x})$ jsou nenulové, jinak by některé z parametrů $\mathbf{a}_i$, $i = 1, 2, \ldots, p-1$ mohly nabývat nekonečných hodnot, nebo by nebyly definovány vůbec. Toto je zásadní rozdíl oproti Aitchisonově přístupu použití logistického normálního rozdělení (viz [4]), kde se předpokladá, že složky kompozic jsou nenulové, tedy se jedná o předpoklad značně striktnější.

S použitím Kentových modelů se pojí také jisté problémy, na které poukázali Scealy a Welsh zejména v [20]. Jedná se konkrétně o skutečnost, že

(i) normalizační konstanta $c(\kappa, \boldsymbol{\beta})$ neexistuje v uzavřené formě a zahrnuje mnohonásobné integrály, jejichž řešení je náročné v případě, že p je velké.

(ii) odmocninová transformace zobrazuje kompoziční data pouze přímo do kladného ortantu hyperkoule. V případě, že se složky vektoru $\mathbf{u}$ nacházejí blízko hranic simplexu, Kentův model může expandovat mimo kladný ortant hyperkoule.

Abychom se vypořádali s problémem (ii), lze uvažovat tzv. „složené“ Kentovo rozdělení. Předpokládejme, že $\mathbf{y}$ má Kentovo rozdělení, pak s využitím absolutních hodnot má $|\mathbf{y}| = (|y_1|, |y_2|, \ldots, |y_p|)^T$ tzv. složené Kentovo rozdělení. Pracovat se složeným Kentovým rozdělením je poměrně náročné (viz [19]) a proto se mu v této práci blíže věnovat nebudeme. Nicméně při analyzování mnoha datových souborů můžeme předpokládat, že většina složek $\mathbf{u}$ je rozdělena mimo hranice simplexu, pak tedy parametr koncentrace transformovaných dat κ bude značně vysoký a to nám dovolí využít aproximace uvedené v podkapitole 2.3. Tímto způsobem snížíme negativní vliv problémů (i) a (ii) a pro velké κ můžeme toto složení ignorovat.

Nyní se ještě vraťme k tématu podkapitoly 1.3 o kritice (nejen) sférických přístupů ze strany Aitchisona. Jedná se konkrétně o tyto skutečnosti

(a) rozdělení na hyperkouli nepracují přímo v kladném ortantu,

(b) perturbace na simplexu nelze vyjádřit na hyperkouli, protože rotace nevykazuje žádný vztah k perturbaci,

(c) není splněna podkompoziční soudržnost jakožto vztah mezi úhly ve vyšších a nižších dimenzích.

Scealy a Welsh v [18] oponují, že každý pohyb bodu na simplexu je možné reprezentovat jednoznačně rotací na hyperkouli a tedy je rotace na hyperkouli vhodnou alternativou k perturbaci na simplexu. Zamítají tedy problém (b). Problém (a) je totožný s (ii) a je diskutovaný výše. Kritika (c) je oprávněná, neboť podkompoziční soudržnost mezi úhly různých dimenzí opravdu není splněna. Nicméně se

splněním tohoto předpokladu mají v některých případech potíže všechny známé přístupy, z pohledu sféricistů i včetně samotné logpodílové metodiky (viz příklad 1.5.1).

Druhá kapitola demonstrovala principiální rozdíl mezi pohledem logpodílové metodiky a sféricistů v případě regresní analýzy. Zatímco v logpodílovém případě bychom postupovali vyjádřením vysvětlovaných kompozic v logpodílových souřadnicích a dále užitím standardních nástrojů vícenásobné regrese (více o interpretaci výsledků přímo v souřadnicích nebo zpět na simplexu najdeme v [13]), sféricisté jdou cestou tvorby komplexních statistických modelů šitých na míru konkrétnímu zadání. I to může být z praktického hlediska jednou z příčin, proč se v průběhu uplynulých desetiletí stala logpodílová metodika dominantní volbou, nejen při regresní analýze kompozic.

Kapitola 3

Robustní metoda hlavních komponent v tangentovém prostoru

Následující podkapitola bude pojednávat o nové robustní metodě hlavních komponent (dále jen PCA), která byla vyvinuta k analýze geochemických dat. Využívá se při ní tzv. relativní mocninné transformace s cílem analyzovat kompoziční data, jejichž některé složky jsou malé a relativně hodně variabilní. Kompozice s takovými složkami mají rozdělení blízko hranic simplexu. V souvislosti s touto metodou budeme v dalším textu pracovat s pojmy jako jsou tzv. variety, tangentové vektory a tangentové prostory, a tak se na začátku této kapitoly budeme nejprve věnovat stručnému seznámení se s těmito pojmy z diferenciální geometrie. Poté metodu popíšeme, odhadneme parametry a v neposlední řadě ji srovnáme s PCA založenou na logpodílových souřadnicích a s PCA na hyperkouli.

V tomto textu budeme považovat standardní PCA za známou. Zmíníme pouze to, že se jedná o metodu pro redukci dimenze mnohorozměrných dat s cílem vysvětlit co nejvíce variability obsažené v datech. Hlavní komponenty jsou vlastně lineárními kombinacemi původních složek náhodného vektoru, z geometrického hlediska se jedná o osy nového souřadnicového systému. Bázové vektory tohoto souřadnicového systému nazýváme zátěže. Souřadnice vzhledem k těmto bázovým vektorům nazýváme skóry. Výsledky analýzy hlavních komponent zobrazujeme pomocí speciálního grafického nástroje, tzv. biplotu, který je dvoudimenzionálním

zobrazením objektů (skórů) a proměnných (zátěží) prvních dvou hlavních komponent v jednom grafu. Pro případné detaily se odkazujeme na literaturu [9].

3.1. Základní pojmy z diferenciální geometrie

Vzhledem ke skutečnosti, že budeme v další podkapitole pracovat s prostory obecnějšími než je hyperkoule, přiblížíme zejména tzv. varietu, k tomu bude zapotřebí nejprve definovat tzv. topologický prostor a některé další pojmy. Při tvorbě celé této podkapitoly byla využita literatura [22].

3.1.1. Topologický prostor, topologická a diferenciální varieta

Definice 3.1.1 *Množina* $\mathbf{N}$ *spolu se systémem* $\mathcal{T} = \{T\}$ *podmnožin množiny* $\mathbf{N}$ *takových, že platí:*

1. $\emptyset \in \mathcal{T}, \mathbf{N} \in \mathcal{T}$,
2. $T_i \in \mathcal{T}$, $i \in I$: $\bigcup_{i \in I} T_i \in \mathcal{T}$,
3. $T_i \in \mathcal{T}$, $i \in I$, I *je konečná množina:* $\bigcap_{i \in I} T_i \in \mathcal{T}$,

se nazývá topologickým prostorem na $\mathbf{N}$. *Systém podmnožin* $\mathcal{T} = \{T\}$ *se nazývá topologie.*

Definice 3.1.2 *Množiny z* $\mathbf{N}$ *nazýváme otevřené množiny. Podmnožiny* $\mathbf{N}$ *se nazývají uzavřené, je-li jejich doplňkem otevřená množina.*

Definice 3.1.3 *Zobrazení* $f : T \to R$ *se nazývá homeomorfismus, je-li zobrazení* f *bijektivní a obě zobrazení* f, f^{-1} *jsou spojitá.*

Definice 3.1.4 *Zobrazení* $f : T \to R$ *se nazývá* C^r*-difeomorfismus, jsou-li* f, f^{-1} *spojitě diferencovatelné funkce až do řádu* r *včetně.*

Nechť $\mathbf{N}^p$ je p-rozměrný topologický prostor s otevřenými podmnožinami $\{U_i | i \in I\}$ takovými, že $\bigcup_{i\in I} U_i = \mathbf{N}^p$ a s funkcemi $c_i : U_i \to \mathbb{R}^p$, které jsou homeomorfismy na otevřené podmnožiny $f(U_i) \in \mathbb{R}^p$. Funkce $c_i, i \in I$ nazýváme mapami na $\mathbf{N}^p$ za předpokladu, že

$$c_j \circ c_i^{-1} : c_i(U_i \cap U_j) \to c_j(U_i \cap U_j) \tag{3.1}$$

je homeomorfismus $\forall i, j \in I$. Mapy lze interpretovat jako lokální souřadnicový systém na $\mathbf{N}^p$.

Definice 3.1.5 *Podmnožiny $\{U_i\}_{i\in I}$ spolu s mapami $\{c_i\}_{i\in I}$ tvoří atlas na $\mathbf{N}^p$. Množina $\mathbf{N}^p$ spolu s atlasem $\{(U_i, c_i) | i \in I\}$ se nazývá topologickou varietou dimenze p.*

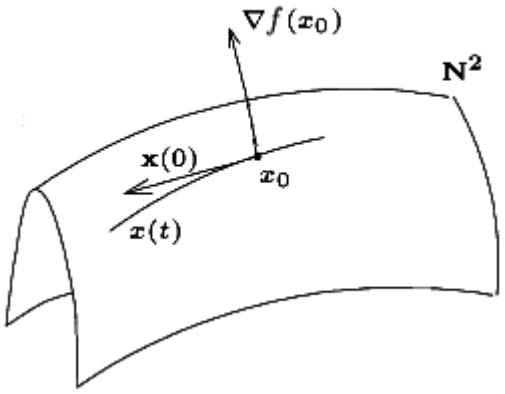
Definice 3.1.6 *Množina $\mathbf{N}^p$ spolu s atlasem $\{(U_i, c_i) | i \in I\}$ se nazývá C^r-diferenciální varietou dimenze p, je-li zobrazení $c_j \circ c_i^{-1}$ v (3.1) C^r-difeomorfismus.*

3.1.2. Tangentové vektory, tangentový prostor

Nejprve se zaměříme na intuitivní představu variety. Varietu chápeme jako zobecnění vnímání zakřivené plochy v 3-rozměrném prostoru. Zakřivenou plochu si obvykle představujeme jako podmnožinu v $\mathbb{R}^3$, která zachovává geometrické vlastnosti struktury eukleidovského prostoru, kterému náleží.

Diferenciální varietu v $\mathbb{R}^3$ lze uvažovat jako řešení rovnice $f(x_1, x_2, x_3) = 0$, kde f je reálná funkce definovaná na $\mathbb{R}^3$. Tuto plochu označíme $\mathbf{N}^2$, bod této plochy můžeme označit $x = (x_1, x_2, x_3)^T$. Je-li gradient $\nabla f(x) \neq 0$, potom odpovídá normálovému vektoru k ploše $\mathbf{N}^2$. Tangentové vektory v bodě x budou tedy vektory v $\mathbb{R}^3$ takové, které jsou na tento normálový vektor kolmé. Množina všech tangentových vektorů vzhledem k dané ploše v bodě x se nazývá tangentovým prostorem plochy $\mathbf{N}^2$ v bodě x. Pak tedy vektor $\mathbf{u} = (u_1, u_2, u_3)^T$ je tangentovým vektorem k ploše $\mathbf{N}^2$ v bodě x právě tehdy, když

$$\mathbf{u} \cdot \nabla f(x) = \sum_{j=1}^{3} u_j \frac{\partial f}{\partial x_j} = 0.$$



Obrázek 3.1: Tangentový a normálový vektor k diferenciální varietě.

Pro přechod k obecným diferenciálním varietám tato konstrukce nelze zobecnit a přistupuje se tak k abstraktnějšímu vyjádření. Tangentové vektory lze totiž uvést do souvislosti s ekvivalentními třídami tzv. cest procházejících bodem x.

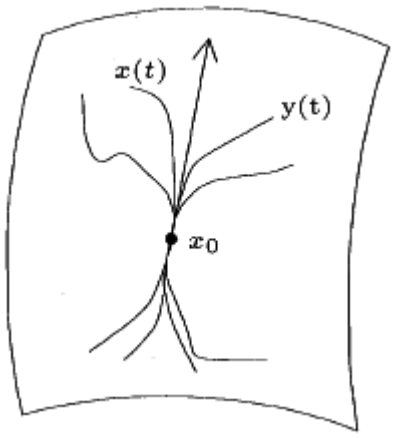
Nechť x_0 je bod na ploše $\mathbf{N}^2$ a dále nechť $x(t) = (x_1(t), x_2(t), x_3(t))^T$ je cestou na ploše procházející bodem x_0 v $t = 0$ a definovanou pro $t \in (-\eta, \eta)$. Dále $\forall t$ nechť je definován vektor

$$\mathbf{x}(t) = \left(\frac{dx_1(t)}{dt}, \frac{dx_2(t)}{dt}, \frac{dx_3(t)}{dt} \right)^T .$$

Lze tedy pozorovat, že $\mathbf{x}(0)$ je tangentový vektor k ploše $\mathbf{N}^2$ v bodě x_0, tedy každá hladká cesta procházející bodem x_0 definuje tangentový vektor v bodě x_0. Tento tangentový vektor není jednoznačný k cestě, neboť existuje mnoho cest skrze x_0, které mají v daném bodě stejný tangentový vektor. Nicméně, všechny cesty skrze bod x_0 se stejným tangentovým vektorem tvoří třídu ekvivalence cest v daném bodě. Situaci znázorňují obrázky 3.1 a 3.2.

Nechť $\mathbf{N}^p$ je diferenciální varieta. Dále nechť $x(t)$ a $y(t)$ jsou dvě hladké cesty v $\mathbf{N}^p$ procházející společným bodem x_0 v $t = 0$ a souřadnicový systém je dán mapou okolo x_0 tak, že uvedené cesty mají souřadnice

$$x(t) = (x_1(t), x_2(t), \ldots, x_p(t))^T, \quad y(t) = (y_1(t), y_2(t), \ldots, y_p(t))^T,$$



Obrázek 3.2: Tangentový vektor jako třída ekvivalence cest procházejících skrze bod na varietě.

$$\text{a} \quad x_0 = (x_{01}, x_{02}, \ldots, x_{0p})^T.$$

Cesty $x(t)$ a $y(t)$ jsou hladké, jsou-li jejich souřadnice diferencovatelné funkce proměnné t. Cesty $x(t)$ a $y(t)$ nazýváme tangentové v x_0 za předpokladu, že

$$\frac{dx_j(0)}{dt} = \frac{dy_j(0)}{dt}, \ j = 1, 2, \ldots, p.$$

Zde si můžeme povšimnout, že vlastnost „tangentnosti“ je nezávislá na výběru souřadnicového systému.

Definice 3.1.7 *Tangentový vektor* $\mathbf{x}$ *k cestě* $x(t)$ *v bodě* $x_0 = x(0)$ *je definován jako třída ekvivalence všech cest* $y(t)$ *takových, že* $y(0) = x_0$ *a cesta* $y(t)$ *je tangentová v* $t = 0$.

Definice 3.1.8 *Vektorový prostor všech tangentových vektorů vzhledem k varietě* $\mathbf{N}^p$ *v daném bodě* $x \in \mathbf{N}^p$ *nazýváme tangentovým prostorem v bodě* x *a značíme jej* $T_x(\mathbf{N}^p)$.

Lze také ukázat, že $T_x(\mathbf{N}^p)$ je stejné dimenze jako varieta $\mathbf{N}^p$. Navíc platí, že $T_x(\mathbf{N}^p)$ je izomorfní k eukleidovskému prostoru $\mathbb{R}^p$.

3.2. Robustní PCA v tangentovém prostoru

Nyní popíšeme robustní PCA užitím transformace z variety na eukleidovský tangentový prostor. Jak již bylo zmíněno, tato robustní metoda byla navržena pro geochemická data, která mají často některé složky malé a relativně hodně variabilní. Takové složky mají „asymetrické“ rozdělení blízko hranic simplexu. Cílem metody uvedené v následujícím textu je pomocí tzv. relativní mocninné transformace data přesunout a symetrizovat. Ještě před samotnou transformací se data relativně centrují vůči tzv. centrovacímu parametru, který je odhadován z celého kompozičního datového souboru. Zde lze pozorovat zásadní rozdíl oproti tradičním metodám, které ke změně měřítka používají poměry složek z každé kompozice odděleně. Aplikováním výše zmíněné transformace se data zobrazují na varietu, která může být značně komplikovanou plochou. Data jsou transformována z variety na eukleidovský tangentový prostor a v tomto prostoru se pak s nimi dále pracuje.

Tato podkapitola byla vytvořena zejména na základě publikace [17].

3.2.1. Popis metody

Pro libovolné $\alpha \in \mathbb{R}$ nejprve definujeme α-tou mocninu p-rozměrného vektoru $\mathbf{u}$ vztahem

$$\mathbf{u}^{\alpha} = \left(u_1^{\alpha}, u_2^{\alpha}, \ldots, u_p^{\alpha}\right)^T.$$

Dále použijeme tzv. Box-Coxovu mocninnou transformaci

$$T_{\alpha}(\mathbf{u}) = \begin{cases} \dfrac{\mathbf{u}^{\alpha} - \mathbf{1}_p}{\alpha} & \alpha \neq 0, \\ \ln(\mathbf{u}) & \alpha = 0. \end{cases}$$

V souladu se značením v definici 1.4.1 budeme p-rozměrné kompozice nabývající hodnot v $(p-1)$-rozměrném simplexu Δ^p značit $\mathbf{u} = (u_1, u_2, \ldots, u_p)^T \in \Delta^p$. Množinou $S^{p-1} = \{\mathbf{y} \in \mathbb{R}^p \mid \|\mathbf{y}\| = 1\}$ budeme rozumět $(p-1)$-rozměrnou jednotkovou hyperkouli. Pak pro libovolný vektor $\mathbf{y} = (y_1, y_2, \ldots, y_p)^T$, který leží

v kladném ortantu hyperkoule $S_+^{p-1} = \{\mathbf{y} \in S^{p-1} \mid y_j \geq 0,\ j = 1, 2, \ldots, p\}$ platí, že $\mathbf{y}^2 \in \Delta^p$. Naopak pro každou kompozici $\mathbf{u} \in \Delta^p$ platí, že $\sqrt{\mathbf{u}} \in S_+^{p-1}$.

Pro $\boldsymbol{\mu} \in S_+^{p-1}$ s kladnými složkami $\mu_j > 0$, $j = 1, 2, \ldots, p$ definujeme relativní mocninnou transformaci, která zobrazuje kompozice $\mathbf{u} \in \Delta^p$ na vektory $\mathbf{z} = (z_1, z_2, \ldots, z_p)^T$ vztahem

$$\mathbf{z} = \frac{1}{2}\mathrm{diag}(\boldsymbol{\mu}) T_\alpha \left(\mathrm{diag}\left(\frac{1}{\boldsymbol{\mu}^2} \right) \mathbf{u} \right) =$$

$$= \begin{cases} \frac{1}{2\alpha}\mathrm{diag}(\boldsymbol{\mu}) \left[\left(\mathrm{diag}\left(\frac{1}{\boldsymbol{\mu}^2} \right) \mathbf{u} \right)^\alpha - \mathbf{1}_p \right] & \alpha \neq 0, \\ \frac{1}{2}\mathrm{diag}(\boldsymbol{\mu}) \ln \left[\mathrm{diag}\left(\frac{1}{\boldsymbol{\mu}^2} \right) \mathbf{u} \right] & \alpha = 0. \end{cases} \tag{3.2}$$

Invertováním (3.2) dostáváme

$$\begin{aligned} \mathbf{u}^\alpha &= \boldsymbol{\mu}^{2\alpha} + 2\alpha \mathrm{diag}\left(\boldsymbol{\mu}^{2\alpha - 1} \right) \mathbf{z} && \alpha \neq 0, \\ \ln(\mathbf{u}) &= \ln(\boldsymbol{\mu}^2) + 2\mathrm{diag}\left(\frac{1}{\boldsymbol{\mu}} \right) \mathbf{z} && \alpha = 0. \end{aligned} \tag{3.3}$$

Transformace (3.2) zobrazuje kompozice ze simplexu Δ^p na varietu

$$\mathbf{N}_\alpha^{p-1} = \begin{cases} \left\{ \mathbf{z} \in \mathbb{R}^p : z_j \geq -\frac{\mu_j}{2\alpha}, f(\mathbf{z}) = 1 \right\} & \alpha > 0, \\ \left\{ \mathbf{z} \in \mathbb{R}^p : f(\mathbf{z}) = 1 \right\} & \alpha = 0, \\ \left\{ \mathbf{z} \in \mathbb{R}^p : z_j \leq -\frac{\mu_j}{2\alpha}, f(\mathbf{z}) = 1 \right\} & \alpha < 0, \end{cases}$$

kde

$$f(\mathbf{z}) = \begin{cases} \mathbf{1}_p^T \mathrm{diag}(\boldsymbol{\mu}^2) \left(\mathbf{1}_p + 2\alpha \mathrm{diag}\left(\frac{1}{\boldsymbol{\mu}} \right) \mathbf{z} \right)^{1/\alpha} & \alpha \neq 0, \\ \mathbf{1}_p^T \mathrm{diag}(\boldsymbol{\mu}^2) \exp\left(2\mathrm{diag}\left(\frac{1}{\boldsymbol{\mu}} \right) \mathbf{z} \right) & \alpha = 0. \end{cases} \tag{3.4}$$

Všimněme si, že dosazením $\mathbf{u} = \boldsymbol{\mu}^2$ do (3.2) se $\boldsymbol{\mu}^2$ zobrazí $\forall \alpha$ na $\mathbf{z} = \mathbf{0}_p$. Pro gradient funkce $f(\mathbf{z})$ platí

$$\nabla f(\mathbf{z}) = 2\mathrm{diag}(\boldsymbol{\mu}) \left(\mathbf{1}_p + 2\alpha \mathrm{diag}\left(\frac{1}{\boldsymbol{\mu}} \right) \mathbf{z} \right)^{1/\alpha - 1}.$$

Na varietě je $\mathbf{z} = \mathbf{0}_p$ definován dle (3.4), tj. $\forall \alpha$ je $f(\mathbf{0}_p) = 1$ a dále platí $\nabla f(\mathbf{0}_p) = 2\boldsymbol{\mu}$. To znamená, že $\boldsymbol{\mu}$ je normálou k varietě $\mathbf{N}_\alpha^{p-1}$ v $\mathbf{z} = \mathbf{0}_p$. Tangentové vektory k varietě $\mathbf{N}_\alpha^{p-1}$ v $\mathbf{z} = \mathbf{0}_p$ jsou tedy vektory v $\mathbb{R}^p$ ortogonální k $\boldsymbol{\mu}$.

Tyto tangentové vektory tvoří tangentový prostor $T_{\mathbf{0}_p}(\mathbf{N}_\alpha^{p-1})$. K promítnutí $\mathbf{z}$ na tangentový prostor, který je kolmý k $\boldsymbol{\mu}$, se používá projekce ve tvaru

$$\boldsymbol{\xi} = \mathbf{z} - \boldsymbol{\mu}\boldsymbol{\mu}^T\mathbf{z} = (\mathbf{I}_p - \boldsymbol{\mu}\boldsymbol{\mu}^T)\mathbf{z}, \tag{3.5}$$

kde $\mathbf{I}_p$ je $(p \times p)$-rozměrná jednotková matice.

Vektor $\mathbf{z}$ můžeme (přičtením $\boldsymbol{\mu}\boldsymbol{\mu}^T\mathbf{z}$ k oběma stranám rovnice a úpravou) zapsat pomocí (3.5) v souřadnicích tangentového prostoru jako

$$\mathbf{z} = (\boldsymbol{\mu}\boldsymbol{\mu}^T + \mathbf{I}_p - \boldsymbol{\mu}\boldsymbol{\mu}^T)\mathbf{z} = \boldsymbol{\mu}\delta + \boldsymbol{\xi}, \tag{3.6}$$

kde $\delta = \boldsymbol{\mu}^T\mathbf{z}$ a po zpětné substituci do (3.3) dostáváme $\mathbf{u}^\alpha$ v souřadnicích tangentového prostoru

$$\begin{array}{ll} \mathbf{u}^\alpha = \boldsymbol{\mu}^{2\alpha} + \boldsymbol{\mu}^{2\alpha}2\alpha\delta + 2\alpha\mathrm{diag}(\boldsymbol{\mu}^{2\alpha-1})\boldsymbol{\xi} & \alpha \neq 0, \\ \ln(\mathbf{u}) = \ln(\boldsymbol{\mu}^2) + \mathbf{1}_p 2\delta + 2\mathrm{diag}\left(\frac{1}{\boldsymbol{\mu}}\right)\boldsymbol{\xi} & \alpha = 0. \end{array} \tag{3.7}$$

PCA v tangentovém prostoru $T_{\mathbf{0}_p}(\mathbf{N}_\alpha^{p-1})$ je založena na analýze spektrálního rozkladu odhadu matice $\mathrm{var}(\boldsymbol{\xi})$. Aby bylo možné rozklad provést, je potřeba stanovit určité předpoklady o $\boldsymbol{\xi}$ a poté je prostřednictvím (3.3) a (3.6) přenést zpět přes $\mathbf{z}$ k $\mathbf{u}^\alpha$ (resp. $\ln(\mathbf{u})$). Relativní mocninnou transformací (3.2) chceme přesunout data dále od hranic simplexu a transformovat kompozice blízké nule za účelem dosažení aproximativně symetrického rozdělení. Předpokládáme, že
(P) *transformovaný náhodný vektor $\boldsymbol{\xi}$ má elipticky symetrické rozdělení s nulovou střední hodnotou a konečným rozptylem.*
Tento předpoklad je užitečný, ale v praxi většinou nelze očekávat jeho platnost. Právě z tohoto důvodu se používá robustních odhadů střední hodnoty a rozptylu, které dovolí ignorovat mírné odchylky od předpokladu (P).

První vlastní číslo $\mathrm{var}(\boldsymbol{\xi})$ je nula, protože jedna dimenze byla ztracena projekcí na tangentový prostor. Nechť $\lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p \geq 0$ jsou zbývající vlastní čísla, dále nechť $\mathbf{D}_\lambda = \mathrm{diag}(\lambda_2, \lambda_3, \ldots, \lambda_p)$ a $\boldsymbol{\Gamma}^* = (\boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \ldots, \boldsymbol{\gamma}_p)$ je $p \times (p-1)$-rozměrná matice příslušných vlastních vektorů. Pak lze psát $\mathrm{var}(\boldsymbol{\xi}) = \boldsymbol{\Gamma}^*\mathbf{D}_\lambda\boldsymbol{\Gamma}^{*T}$. Komponenta $\boldsymbol{\mu}^T\mathbf{z}$ ve směru normálového vektoru $\boldsymbol{\mu}$ je ortogonální k tangen-

tovému prostoru, tj. $\boldsymbol{\Gamma} = (\boldsymbol{\mu}, \boldsymbol{\Gamma}^*) = (\boldsymbol{\mu}, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \ldots, \boldsymbol{\gamma}_p)$ je $(p \times p)$-rozměrná ortogonální matice

$$\boldsymbol{\Gamma}^T\mathbf{z} = \begin{pmatrix} \boldsymbol{\mu}^T\mathbf{z} \\ \boldsymbol{\Gamma}^{*T}\mathbf{z} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}^T\mathbf{z} \\ \boldsymbol{\Gamma}^{*T}(\boldsymbol{\mu}\delta + \boldsymbol{\xi}) \end{pmatrix} = \begin{pmatrix} \delta \\ \boldsymbol{\Gamma}^{*T}\boldsymbol{\xi} \end{pmatrix}, \tag{3.8}$$

neboť $\boldsymbol{\Gamma}^{*T}\boldsymbol{\mu} = \mathbf{0}_{p-1}$. Z toho dále plyne, že

$$\mathsf{E}\left(\boldsymbol{\Gamma}^T\mathbf{z}\right) = \begin{pmatrix} m^* \\ \mathbf{0}_{p-1}^T \end{pmatrix} \text{ a } \operatorname{var}(\boldsymbol{\Gamma}^T\mathbf{z}) = \begin{pmatrix} v^* & \mathbf{c}_v^{*T} \\ \mathbf{c}_v^* & \mathbf{D}_\lambda \end{pmatrix}, \tag{3.9}$$

kde $m^* = \mathsf{E}(\delta)$, $v^* = \operatorname{var}(\delta)$ jsou čísla a $\mathbf{c}_v^* = \operatorname{cov}(\boldsymbol{\Gamma}^{*T}\boldsymbol{\xi}, \delta)$ je $(p-1)$-rozměrný vektor.

Lze dokázat, že platí vztahy

$$\mathsf{E}(\mathbf{z}) = \boldsymbol{\mu}m^* \text{ a } \operatorname{var}(\mathbf{z}) = v^*\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Gamma}^*\mathbf{c}_v^*\boldsymbol{\mu}^T + \boldsymbol{\mu}\mathbf{c}_v^{*T}\boldsymbol{\Gamma}^{*T} + \boldsymbol{\Gamma}^*\mathbf{D}_\lambda\boldsymbol{\Gamma}^{*T},$$

z nichž s využitím (3.3) přejdeme k $\mathbf{u}^\alpha$, resp. $\ln(\mathbf{u})$. Platí

$$\mathsf{E}(\mathbf{u}^\alpha) = \boldsymbol{\mu}^{2\alpha}(1+2\alpha m^*) \text{ a } \operatorname{var}(\mathbf{u}^\alpha) = 4\alpha^2\operatorname{diag}(\boldsymbol{\mu}^{2\alpha-1})\operatorname{var}(\mathbf{z})\operatorname{diag}(\boldsymbol{\mu}^{2\alpha-1}). \tag{3.10}$$

Pro $\alpha = 0$ jsou momenty v (3.10) ve tvaru

$$\mathsf{E}(\ln(\mathbf{u})) = \ln(\boldsymbol{\mu}^2) + \mathbf{1}_p 2m^* \text{ a } \operatorname{var}(\ln(\mathbf{u})) = 4\operatorname{diag}\left(\frac{1}{\boldsymbol{\mu}}\right)\operatorname{var}(\mathbf{z})\operatorname{diag}\left(\frac{1}{\boldsymbol{\mu}}\right).$$

Poměr vysvětlené variability v tangentovém prostoru ve směru j-tého vlastního vektoru matice $\operatorname{var}(\boldsymbol{\xi})$ je

$$\frac{\lambda_j}{\sum_{k=2}^p \lambda_k}, \quad 2 \le j \le p. \tag{3.11}$$

Dále nás bude zajímat, jak tangentový prostor $T_{\mathbf{0}_p}(\mathbf{N}_\alpha^{p-1})$ aproximuje varietu $\mathbf{N}_\alpha^{p-1}$. Určíme tedy celkovou variabilitu $\mathbf{z}$. Platí, že $\operatorname{tr}(\operatorname{var}(\mathbf{z})) = v^* + \sum_{k=2}^p \lambda_k$. Poměr celkové variability ve směru j-tého vlastního vektoru je pak

$$\frac{\lambda_j}{v^* + \sum_{k=2}^p \lambda_k}, \quad 2 \le j \le p. \tag{3.12}$$

3.2.2. Odhady parametrů

Nechť jsou data tvořena p-rozměrnými kompozicemi $\mathbf{u}_1, \ldots, \mathbf{u}_n$. Použijeme-li PCA jako metodu pro redukci dimenze, není nutné předpokládat žádné jejich explicitní rozdělení.

Nejprve odhadneme parametr α, aby bylo možné pro složky rozdělené blízko nuly vytvořit rozdělení transformovaných složek

$$\mathbf{y}_i = \mathbf{u}_i^{\alpha} \text{ pro } \alpha \neq 0,\ \mathbf{y}_i = \ln(\mathbf{u}_i) \text{ pro } \alpha = 0,\ i = 1, 2, \ldots, n,$$

tak symetrické, jak je to jen možné. Tento odhad můžeme provést graficky.

Pro zvolenou hodnotu α odhadneme $\boldsymbol{\mu}$. Při odhadu $\boldsymbol{\mu}$ se použije odhad $\mathbf{m}$ střední hodnoty $\mathsf{E}(\mathbf{u}^{\alpha})$. Poté z (3.10) plyne, že $\sqrt[2\alpha]{\mathbf{m}}$ je odhadem $\boldsymbol{\mu}\sqrt[2\alpha]{1 + 2\alpha m^*}$. Normováním dostáváme

$$\frac{\sqrt[2\alpha]{\mathbf{m}}}{\|\sqrt[2\alpha]{\mathbf{m}}\|} \text{ pro } \alpha \neq 0 \text{ , resp. } \frac{\sqrt{e^{\mathbf{m}}}}{\|\sqrt{e^{\mathbf{m}}}\|} \text{ pro } \alpha = 0.$$

Z toho vyplývá, že $\boldsymbol{\mu}$ můžeme odhadnout jako

$$\widehat{\boldsymbol{\mu}} = \frac{\sqrt[2\alpha]{\overline{\mathbf{y}}}}{\|\sqrt[2\alpha]{\overline{\mathbf{y}}}\|} \text{ pro } \alpha \neq 0 \text{ , resp. } \widehat{\boldsymbol{\mu}} = \frac{\sqrt{e^{\overline{\mathbf{y}}}}}{\|\sqrt{e^{\overline{\mathbf{y}}}}\|} \text{ pro } \alpha = 0 \text{ , kde } \overline{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i. \quad (3.13)$$

Bohužel data blízká nule nejsou po transformaci úplně symetrická a jsou zde obvykle odlehlá pozorování. Z tohoto důvodu jsou preferovány robustní odhady. Robustní odhad můžeme získat tak, že nahradíme $\overline{\mathbf{y}}$ v (3.13) prostorovým mediánem $\widetilde{\mathbf{y}}$ za účelem minimalizovat výraz

$$\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \widetilde{\mathbf{y}}\|.$$

Je-li vybráno α a stanoveno $\widehat{\boldsymbol{\mu}}$, je možné vypočítat odhady $\widehat{\boldsymbol{\xi}}_i = (\mathbf{I}_p - \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T)\widehat{\mathbf{z}}_i$, kde $\widehat{\mathbf{z}}_i$ se získá substitucí $\widehat{\boldsymbol{\mu}}$ za $\boldsymbol{\mu}$ v (3.2).

Dále potřebujeme odhadnout $\boldsymbol{\Gamma}^*$ a $\mathbf{D}_{\lambda}$. Můžeme je získat z vlastních vektorů a vlastních čísel z

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{\xi}}_i \widehat{\boldsymbol{\xi}}_i^T. \quad (3.14)$$

Odhady z (3.14) jsou ovšem nerobustní. Robustní odhady obdržíme z vlastních vektorů a čísel z

$$\frac{1}{n}\sum_{i=1}^{n} I(\widehat{\boldsymbol{\xi}}_i \neq \mathbf{0})\frac{\widehat{\boldsymbol{\xi}}_i\widehat{\boldsymbol{\xi}}_i^T}{\|\widehat{\boldsymbol{\xi}}_i\|^2}, \tag{3.15}$$

kde $I(\cdot)$ je indikátorová funkce. Tento odhad oslabí vliv tangentových vektorů, které jsou daleko od nuly, a tak budou mít menší vliv na výsledný odhad. Robustní PCA v tangentovém prostoru lze použít i na data kolineární a velkých rozměrů. Metoda má vysokou robustnost a je eficientní i pro rozdělení s těžkými chvosty. Lze ukázat, že

$$\mathsf{E}\left(\frac{\boldsymbol{\xi}\boldsymbol{\xi}^T}{\|\boldsymbol{\xi}\|^2}\right) = \boldsymbol{\Gamma}^*\mathsf{E}\left(\frac{\boldsymbol{\Gamma}^{*T}\mathbf{z}\mathbf{z}^T\boldsymbol{\Gamma}^*}{\|\boldsymbol{\Gamma}^{*T}\mathbf{z}\|^2}\right)\boldsymbol{\Gamma}^{*T} = \boldsymbol{\Gamma}^*\boldsymbol{\Psi}\boldsymbol{\Gamma}^{*T},$$

kde $\boldsymbol{\Psi}$ je diagonální matice. Jak už bylo uvedeno, první vlastní číslo je rovno nule. Tedy nechť $\widehat{\psi}_2 \geq \widehat{\psi}_3 \geq \cdots \geq \widehat{\psi}_p \geq 0$ jsou odhady zbylých vlastních čísel a nechť $\widehat{\boldsymbol{\Gamma}}^* = (\widehat{\boldsymbol{\gamma}}_2, \widehat{\boldsymbol{\gamma}}_3, \ldots, \widehat{\boldsymbol{\gamma}}_p)$ je odhad matice vlastních vektorů, které přísluší výše uvedeným odhadům vlastních čísel.

K odhadnutí $\mathbf{D}_\lambda$ se používá speciální odhad uvedený v [11]

$$\widehat{\lambda}_j = \left[\underset{1\leq i\leq n}{\text{median}}\left(|\widehat{\boldsymbol{\gamma}}_j^T\widehat{\mathbf{z}}_i - \underset{1\leq i'\leq n}{\text{median}}(\widehat{\boldsymbol{\gamma}}_j^T\widehat{\mathbf{z}}_{i'})|\right)\right]^2 \quad \text{pro} \quad j = 2, 3, \ldots, p, \tag{3.16}$$

$$\widehat{\mathbf{D}}_\lambda = \text{diag}(\widehat{\lambda}_2, \widehat{\lambda}_3, \ldots, \widehat{\lambda}_p). \tag{3.17}$$

Protože $\delta = \boldsymbol{\mu}^T\mathbf{z}$, můžeme odhadnout $m^* = \mathsf{E}(\delta)$ a použít odhad pro zpětnou transformaci k $v^* = \text{var}(\delta)$. K robustnímu odhadu těchto hodnot je nutno nejdříve vypočítat odhadnuté vzdálenosti

$$\widehat{d}_i = \widehat{\mathbf{z}}_i^T\widehat{\boldsymbol{\Gamma}}^*\widehat{\mathbf{D}}_\lambda^{-1}\widehat{\boldsymbol{\Gamma}}^{*T}\widehat{\mathbf{z}}_i \tag{3.18}$$

a definovat výběrové váhy $I(\widehat{d}_i \leq q), 1 \leq i \leq n$, kde $P(\chi^2_{p-1} \leq q) = 0.9$. Robustní odhady m^* a v^* jsou pak ve tvaru

$$\widehat{m}^* = \frac{\sum_{i=1}^n I(\widehat{d}_i \leq q)\widehat{\boldsymbol{\mu}}^T\widehat{\mathbf{z}}_i}{\sum_{i=1}^n I(\widehat{d}_i \leq q)} \text{ a } \widehat{v}^* = \frac{\sum_{i=1}^n I(\widehat{d}_i \leq q)(\widehat{\boldsymbol{\mu}}^T\widehat{\mathbf{z}}_i - \widehat{m}^*)^2}{\sum_{i=1}^n I(\widehat{d}_i \leq q)}. \tag{3.19}$$

3.3. Porovnání metody s dalšími známými přístupy

3.3.1. Vztah k logpodílovému přístupu

V podkapitole 1.2 byly představeny logpodílové souřadnice. Mezi základní logpodílové souřadnice patří *alr*, *clr* a *ilr* souřadnice. Tyto transformace lze vyjádřit také maticově ve tvaru

$$alr(\mathbf{u}) = \mathbf{F}_p \ln(\mathbf{u}) \ , \ \ clr(\mathbf{u}) = \mathbf{G}_p \ln(\mathbf{u}) \text{ a } \ ilr(\mathbf{u}) = \mathbf{H}_p \ln(\mathbf{u}), \tag{3.20}$$

kde $\mathbf{F}_p = (\mathbf{I}_{p-1}, -\mathbf{1}_{p-1})$, $\mathbf{G}_p = \mathbf{I}_p - 1/p\mathbf{1}_p\mathbf{1}_p^T$ a $\mathbf{H}_p$ je $(p-1) \times p$-rozměrná matice, jejíž řádky jsou ortogonální k $\mathbf{1}_p$. Logpodílové souřadnice $alr(\mathbf{u}), clr(\mathbf{u}), ilr(\mathbf{u})$ je možné vyjádřit v souřadnicích tangentového prostoru po řadě jako $\mathbf{F}_p$, $\mathbf{G}_p$ a $\mathbf{H}_p$ vynásobené výrazem $\ln(\mathbf{u})$ dle (3.7), tj. výrazem

$$\ln(\boldsymbol{\mu}^2) + 2\text{diag}\left(\frac{1}{\boldsymbol{\mu}}\right)\boldsymbol{\xi},$$

kde $\boldsymbol{\xi}$ je s využitím vztahů (3.2) a (3.5) možné možné psát ve tvaru

$$\boldsymbol{\xi} = (\mathbf{I}_p - \boldsymbol{\mu}\boldsymbol{\mu}^T)\frac{1}{2}\text{diag}(\boldsymbol{\mu}) \ln\left(\text{diag}\left(\frac{1}{\boldsymbol{\mu}^2}\right)\mathbf{u}\right).$$

Logpodílové transformace jsou tedy lineární v $\boldsymbol{\xi}$ a nezávisí na parametru křivosti δ. Co se týče vhodnosti těchto transformací pro PCA, tak *alr* transformace se nedoporučuje, protože výsledky nejsou invariantní na pořadí složek kompozice $\mathbf{u}$. *Ilr* transformaci zase nelze aplikovat na kolineární data. Nejvhodnější bude tedy porovnat PCA v tangentovém protoru a centrovanou logpodílovou PCA, která je pro PCA v rámci logpodílové metodiky využívána nejčastěji. Hlavní rozdíl mezi těmito dvěma metodami spočívá v tom, že PCA v tangentovém prostoru začíná s relativní transformací centrovanou v $\boldsymbol{\mu}$, kdežto centrovaná logpodílová transformace nestandardizuje data přímo, ale místo toho logaritmuje a poté centruje každou transformovanou kompozici zvlášť.

Abychom mohli prakticky porovnat výše zmíněné metody, bude nejprve nutné

zavést tzv. symetrickou verzi mocninné tranformace a dále její izometrickou podobu. Jedná se o transformace

$$\mathbf{u}^*_{c(\alpha)} = \frac{p}{\alpha \mathbf{1}_p^T \mathbf{u}^\alpha} \mathbf{G}_p \mathbf{u}^\alpha \quad \text{a} \quad \mathbf{u}^*_{i(\alpha)} = \frac{p}{\alpha \mathbf{1}_p^T \mathbf{u}^\alpha} \mathbf{H}_p \mathbf{u}^\alpha.$$

Pro $\alpha \to 0$ platí $\mathbf{u}^*_{c(\alpha)} \to clr(\mathbf{u})$, $\mathbf{u}^*_{i(\alpha)} \to ilr(\mathbf{u})$. Nyní pokračujeme známým postupem při robustní PCA založené na *ilr* transformaci. Vzhledem k tomu, že pro $\alpha = 0$ platí $ilr(\mathbf{u}) = \mathbf{H}_p clr(\mathbf{u})$, odhadneme nejprve $\mathrm{var}(ilr(\mathbf{u}))$ a zpětnou transformací získáme odhad $\mathrm{var}(clr(\mathbf{u})) = \mathbf{H}_p^T \mathrm{var}(ilr(\mathbf{u}))\mathbf{H}_p$, ze kterého obdržíme zátěže. Příslušné skóry můžeme vzít přímo z *ilr* transformovaných dat. Při volbě $\alpha \neq 0$ postupujeme stejně, využijeme ovšem symetrickou variantu mocninné transformace a zátěže získáme z odhadu $\mathrm{var}(\mathbf{u}^*_{c(\alpha)}) = \mathbf{H}_p^T \mathrm{var}(\mathbf{u}^*_{i(\alpha)})\mathbf{H}_p$.

3.3.2. Vztah k PCA na hyperkouli

Jak již bylo zmíněno v druhé kapitole, odmocninová transformace zobrazuje kompoziční data do kladného ortantu hyperkoule. Dále za předpokladu, že data nejsou příliš blízko k hranicím simplexu a jsou dostatečně málo rozptýlena, lze modelovat tato kompoziční data Kentovým rozdělením.

Momenty v (3.9), odvozené pro $\mathbf{\Gamma}^T\mathbf{z}$ za předpokladu (P), platí pro Kentovo rozdělení s volbou $\mathbf{c}^*_v = \mathbf{0}_{p-1}$. Tedy odhadování parametrů Kentova rozdělení je samo o sobě PCA. Podstatou analýzy kompozic na hyperkouli je absolutní odmocninová transformace, narozdíl od relativní odmocninové transformace při PCA v tangentovém prostoru. Navzdory tomuto rozdílu, volbou $\alpha = 1/2$ u PCA v tangentovém prostoru dávají obě analýzy stejné výsledky.

Hlavní výhoda PCA v tangentovém prostoru spočívá v tom, že stačí odhadnout pouze první dva momenty, které jsou pro analýzu v tangentovém prostoru postačující. Další parametry můžeme ignorovat. Toto zjednodušení umožňuje pracovat ve vyšších dimenzích a stanovit robustní odhady. Takovým odhadem je například normalizovaný prostorový medián zmíněný v podkapitole 3.2.2, který je standardním robustním odhadem středního směru.

3.4. Příklad: příčiny úmrtí v evropských zemích

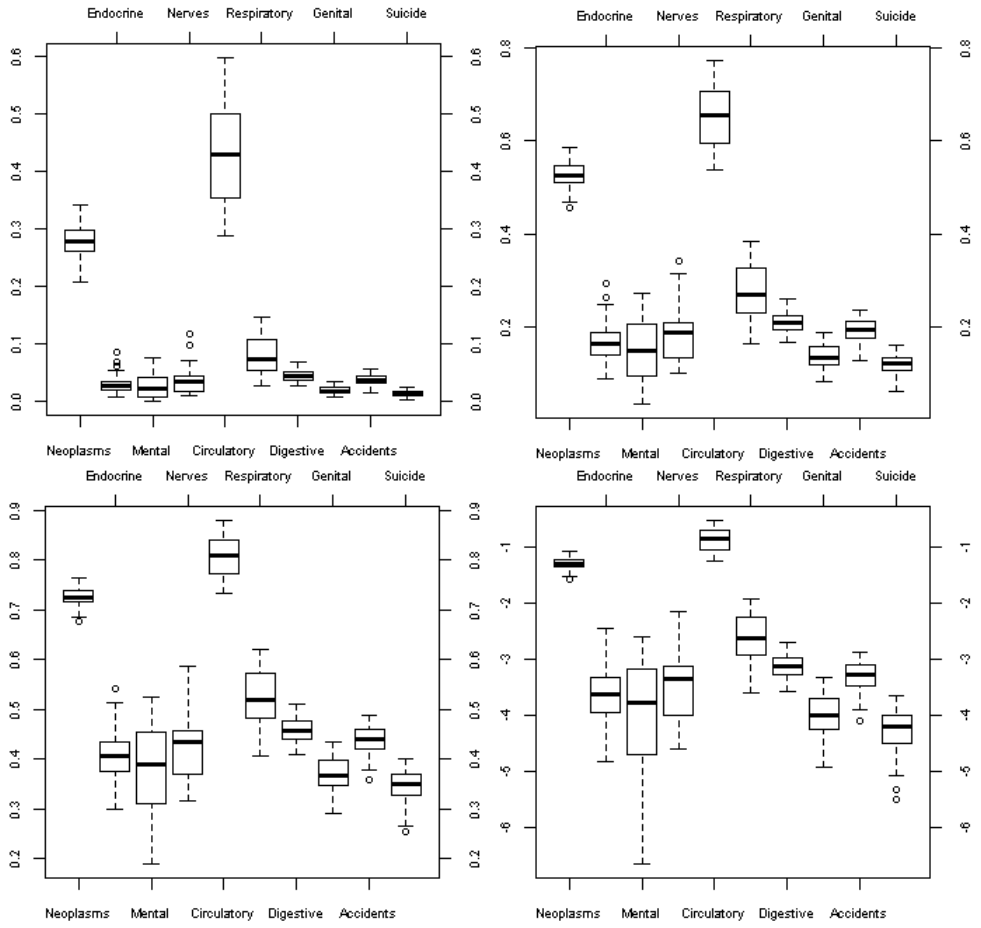
Vycházíme z datového souboru z přílohy B. Tato data byla převzata z databáze Eurostatu z webové stránky [6]. Byla zvolena data pro rok 2011 a vybráno 31 evropských zemí. Uvažujeme dohromady muže i ženy všech věkových kategorií. Databáze Eurostatu je značně rozsáhlá a proto se omezíme jen na následujících 10 nejčastěji se vyskytujících příčin smrti. Mezi ně patří *malignant neoplasms* - zhoubné novotvary, *endocrine, nutritional and metabolic diseases* - endokrinní, vyživovací a metabolické choroby, *mental and behavioural disorders* - duševní poruchy a poruchy chování, *diseases of the nervous system and the sense organs* - choroby nervového systému a smyslových orgánů, *diseases of the circulatory system* - nemoci oběhové soustavy, *diseases of the respiratory system* - nemoci dýchací soustavy, *diseases of the digestive system* - nemoci trávicí soustavy, *diseases of the genitourinary system* - nemoci pohlavního a močového ústrojí, *accidents* - nehody, *intentional self-harm* - sebevraždy.

I když byla původní data obdržena v absolutních počtech úmrtí na jednotlivé nemoci, zjevně nás zajímá spíše jejich relativní struktura, proto se vskutku jedná o kompoziční data. Pro aplikaci PCA v kontextu podkapitoly 3.2 tyto kompozice vyjádříme v proporciální reprezentaci, viz příloha C.

Cílem analýzy těchto dat bude tedy najít vzájemné vztahy mezi jednotlivými státy a příčinami smrti. Pro přehlednost biplotů (grafických výstupů PCA) při interpretaci výsledků byly zvoleny (v souladu s pořadím výše) následující zkratky příčin smrti: Neoplasms, Endocrine, Mental, Nerves, Circulatory, Respiratory, Digestive, Genital, Accidents, Suicide. Jednotlivá pozorování (státy) byla také označena zkratkami (viz přílohy B - E).

Při tvorbě tohoto příkladu jsem využil programy, které mi zprostředkoval vedoucí diplomové práce od spoluautorky [17], Dr. J. L. Scealy, Ph.D. Tyto programy v softwaru R jsem uzpůsobil pro analýzu zvolených dat a přiložil je na CD.

Prvním krokem robustní PCA v tangentovém prostoru je provést odhad parametru α. Ten provedeme graficky pro několik zvolených hodnot, například



Obrázek 3.3: Boxploty transformovaných dat $\mathbf{y} = \mathbf{u}^{\alpha}$. Vlevo nahoře $\mathbf{y} = \mathbf{u}$, vpravo nahoře $\mathbf{y} = \sqrt{\mathbf{u}}$, vlevo dole $\mathbf{y} = \sqrt[4]{\mathbf{u}}$, vpravo dole $\mathbf{y} = \ln(\mathbf{u})$.

pro $\alpha = 1, 1/2, 1/4, 0$. Pro konkrétní volbu α vypočteme relativní mocninnou transformaci $\mathbf{z}$ podle (3.2) a poté pomocí (3.3) vyčíslíme $\mathbf{y} = \mathbf{u}^{\alpha}$. Výsledná data zobrazíme pomocí boxplotů a rozhodneme. Z obrázku 3.3 vidíme, že žádné α neodstraní všechno zakřivení, neboť data jsou značně variabilní a již pouhým pohledem na proporciální datový soubor v příloze C můžeme vidět, že je zde několik odlehlých hodnot. Nejsymetričtější volbou se zdá být $\alpha = 0$. Další analýzu tedy budeme provádět na $\ln(\mathbf{u})$. Z důvodu úspory místa uvádíme datový soubor po logaritmické transformaci v příloze D.

Další krok je odhadnutí středního směru $\boldsymbol{\mu}$ a hlavních směrů $\boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \ldots, \boldsymbol{\gamma}_p$. Pro výpočet $\widehat{\boldsymbol{\mu}}$ použijeme vztah (3.13), kde místo $\overline{\mathbf{y}}$ uvažujeme prostorový medián

$$\widetilde{\mathbf{y}} = (-1.266, -3.574, -3.781, -3.397, -0.891,$$
$$-2.626, -3.118, -3.983, -3.299, -4.261)^T$$

Funkci pro jeho výpočet lze najít v R (funkce spatial.median, knihovna ICSNP). Pak tedy

$$\widehat{\boldsymbol{\mu}} = (0.541, 0.171, 0.154, 0.187, 0.653, 0.274, 0.214, 0.139, 0.196, 0.121)^T.$$

Máme-li zvoleno α a stanoveno $\widehat{\boldsymbol{\mu}}$, můžeme pokračovat projekcí na tangentový prostor

$$\widehat{\boldsymbol{\xi}}_i = (\mathbf{I}_p - \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^T)\widehat{\mathbf{z}}_i, \ i = 1, 2, \ldots, 31.$$

Výše uvedené vektory jsou vyčísleny v podobě datového souboru po provedení projekce na tangentový prostor v příloze E, pomocí nich následně odhadneme $\boldsymbol{\Gamma}^*$. K určení odhadů hlavních směrů $\boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \ldots, \boldsymbol{\gamma}_p$ budeme potřebovat určit vlastní čísla a vlastní vektory matice (3.15). První vlastní číslo je rovno nule, neboť jsme ztratili jednu dimenzi projekcí na tangentový prostor. Další vlastní čísla jsou (zapíšeme je pro úsporu místa pomocí matice $\widehat{\boldsymbol{\Psi}}$)

$$\widehat{\boldsymbol{\Psi}} = \mathrm{diag}(\widehat{\psi}_2, \widehat{\psi}_3, \ldots, \widehat{\psi}_{10}) =$$
$$= \mathrm{diag}(0.514, 0.163, 0.120, 0.090, 0.043, 0.022, 0.019, 0.017, 0.012).$$

Matici odhadů vlastních vektorů pak uvažujeme ve tvaru $\widehat{\boldsymbol{\Gamma}}^* = (\widehat{\boldsymbol{\gamma}}_2, \widehat{\boldsymbol{\gamma}}_3, \ldots, \widehat{\boldsymbol{\gamma}}_{10})$ a výpočtem dostáváme

$$\widehat{\boldsymbol{\Gamma}}^* = \begin{pmatrix} -0.178 & 0.144 & 0.218 & 0.118 & -0.166 & 0.105 & -0.241 & 0.689 & 0.147 \\ -0.053 & 0.266 & -0.506 & -0.540 & -0.541 & -0.135 & -0.028 & -0.105 & 0.164 \\ -0.560 & -0.679 & -0.302 & 0.232 & -0.157 & -0.136 & -0.009 & -0.070 & -0.086 \\ -0.379 & -0.112 & 0.150 & -0.605 & 0.464 & 0.412 & 0.044 & -0.128 & 0.130 \\ 0.587 & -0.259 & -0.238 & 0.039 & 0.262 & -0.050 & 0.057 & -0.160 & 0.058 \\ -0.383 & 0.532 & -0.040 & 0.381 & 0.251 & -0.245 & 0.027 & -0.418 & 0.222 \\ 0.020 & 0.087 & 0.157 & 0.196 & -0.413 & 0.634 & -0.095 & -0.418 & -0.356 \\ -0.123 & 0.243 & -0.232 & -0.056 & 0.167 & -0.043 & 0.497 & 0.263 & -0.710 \\ -0.017 & -0.066 & 0.517 & -0.294 & -0.084 & -0.555 & -0.294 & -0.211 & -0.397 \\ 0.002 & -0.117 & 0.423 & -0.013 & -0.319 & -0.090 & 0.770 & -0.057 & 0.294 \end{pmatrix}.$$

Využitím těchto vektorů a použitím speciálního odhadu (3.16) můžeme odhadnout $\mathbf{D}_\lambda$. Platí

$$\widehat{\mathbf{D}}_\lambda = \mathrm{diag}(\widehat{\lambda}_2, \widehat{\lambda}_3, \ldots, \widehat{\lambda}_{10})$$

a výpočtem dle (3.16) získáme

$$\widehat{\mathbf{D}}_\lambda = \mathrm{diag}(0.0241, 0.0023, 0.0017, 0.0012, 0.0006, 0.0002, 0.0002, 0.0005, 0.0001).$$

Nakonec ještě odhadneme skaláry m^* a v^*. Využijeme k tomu vztahů (3.18), (3.19) a výpočtem získáme

$$\widehat{m}^* = 0.9834, \qquad \widehat{v}^* = 0.00009.$$

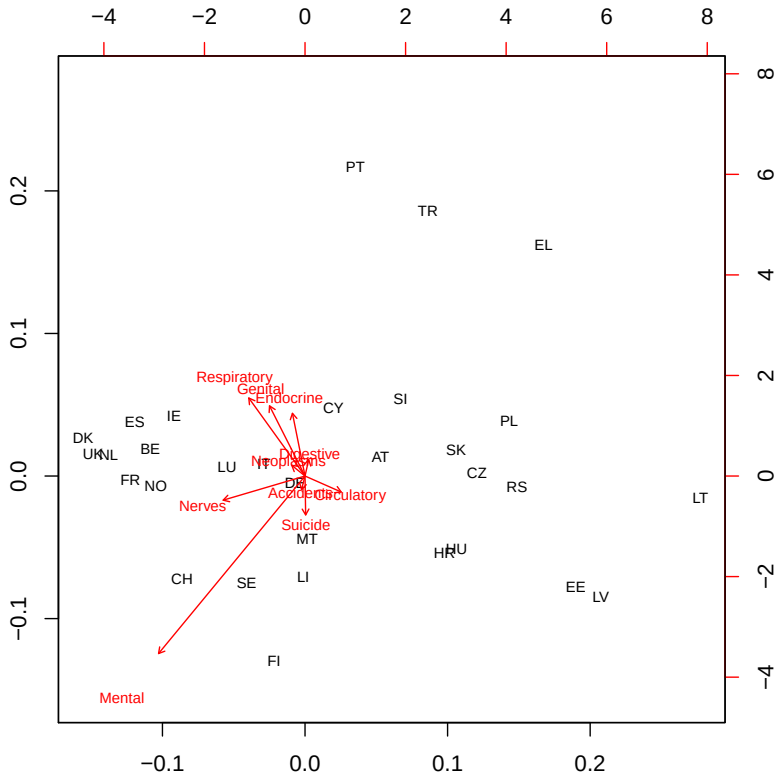
Konečně můžeme vyjádřit podíly vysvětlené variability, tedy

$$\frac{100\widehat{\lambda}_2}{\widehat{v}^* + \sum_{k=2}^{10} \widehat{\lambda}_k} = 77.33\% \ , \quad \frac{100\widehat{\lambda}_3}{\widehat{v}^* + \sum_{k=2}^{10} \widehat{\lambda}_k} = 7.55\%,$$

$$\frac{100\widehat{v}^*}{\widehat{v}^* + \sum_{k=2}^{10} \widehat{\lambda}_k} = 0.29\%.$$

Vidíme tedy, že první dvě komponenty vysvětlují dostatečně dobře variabilitu v datech, neboť součet prvních dvou komponent je přibližně 85%, což bude velmi dobře odrážet strukturu datového souboru.

Nyní ukážeme zkonstruovaný biplot na základě této robustní PCA a interpretujeme jej. Při tvorbě biplotů jsme v souladu s literaturou [17] na x-ovou osu umístili komponentu, která vysvětluje méně variability než druhá, což se může zdát nezvyklé vzhledem ke konvenci ve značení os. Na interpretaci výsledků tato skutečnost ovšem nemá žádný vliv. Při samotné interpretaci jsme využili poznatky o biplotech z literatury [10]. Na obrázku 3.4 si můžeme všimnout, že nedaleko středu leží Německo (DE). Je to dáno tím, že má u všech nemocí hodnoty blízké průměrným. Vzdálenost mezi státy můžeme interpretovat tím způsobem, že čím jsou si státy blíže, tím jsou si podobnější. Jedná se totiž o aproximovanou Mahalanobisovu vzdálenost. Můžeme si tedy všímat tvořících se shluků států. Například z geografického hlediska jsou si blízké státy Česká republika (CZ), Slovensko (SK), Polsko (PL) a dále pobaltské země Estonsko (EE) a Litva (LV). Také Dánsko (DK), Spojené království (UK) a Nizozemsko (NL) jsou si blízké. O těchto skupinách lze říct, že je zde životní styl, lékařská péče a nehodovost na srovnatelné úrovni. Zajímavá jsou také odlehlá pozorování Portugalsko (PT), Turecko (TR) a Řecko (EL), ve kterých je výskyt mentálních poruch výrazně nižší

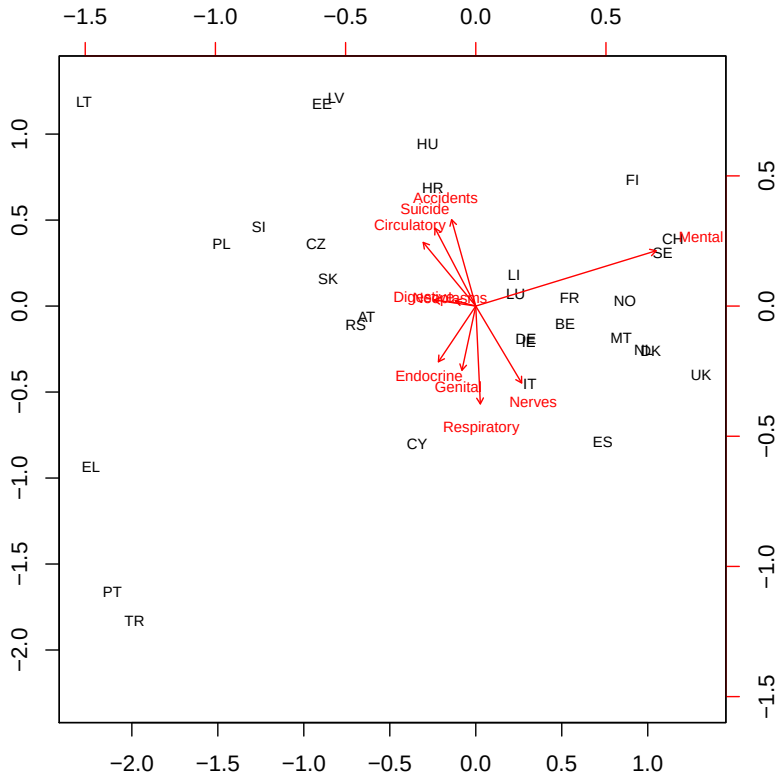


Obrázek 3.4: Biplot $\mathrm{diag}(\boldsymbol{\mu}^{-1})\boldsymbol{\gamma}_3$ proti $\mathrm{diag}(\boldsymbol{\mu}^{-1})\boldsymbol{\gamma}_2$

než průměr. Naopak Finsko (FI), Švédsko (SE) a Švýcarsko (CH) se nacházejí na opačné straně biplotu a výskyt těchto onemocnění je zde vyšší než je průměr. Může to být dáno například nadprůměrnou lékařskou péčí a proto se lidé dožívají vyššího věku, ve kterém potom na tyto nemoci často trpí. Kvůli velikosti paprsku (tedy směrodatné odchylce) právě mentální choroby význačně ovlivňují redukci dimenze celého datového souboru. Nejkratší paprsek mají zhoubné nádory, nemoci trávicí soustavy a smrt následkem nehody. O nich můžeme tvrdit, že jejich poměry jsou stabilní.

Podíváme-li se na vertikální uspořádání států, tak postup zdola nahoru ovlivňuje celá řada příčin. V tomto směru postupně klesá výskyt chorob oběhové soustavy, frekvence nehod a sebevražd a naopak vzrůstá výskyt endokrinních, pohlavních a dýchacích chorob. Co se týče horizontálního uspořádání příčin smrti, nejvíce je postup zleva doprava ovlivňován mentálními nemocemi a chorobami nervové soustavy a smyslových orgánů. V tomto směru postupně výskyt obou zmíněných příčin smrti klesá.

Na závěr ještě porovnáme výsledky naší analýzy s analýzou, která je založena



Obrázek 3.5: Biplot PC2 proti PC1

na logpodílovém přístupu. Konkrétně budeme porovnávat se známou robustní PCA, která je založena na tzv. MCD odhadu. Její princip je popsán v podkapitole 3.3.1. Připomeňme, že se začíná *ilr* transformací dat, poté odhadneme $\mathrm{var}(ilr(\mathbf{u}))$ a zpětně transformujeme na odhad $\mathrm{var}(clr(\mathbf{u})) = \mathbf{H}_p^T \mathrm{var}(ilr(\mathbf{u}))\mathbf{H}_p$, ze kterého obdržíme zátěže. Příslušné skóry vezmeme přímo z *ilr* transformovaných dat. Aplikací tohoto přístupu zjistíme, že první komponenta vysvětluje 59,32% a druhá 21,31% celkové variability. Dohromady tedy první dvě komponenty vysvětlují 80,62% celkové variability v datech.

Podívejme se ještě, v čem se liší biploty z obrázku 3.4 a 3.5. Na první pohled si můžeme všimnout, že jsou si biploty velmi podobné (jsou pouze oba jinak orientované, což na interpretaci nemá vliv). Vidíme také, že Německo (DE) se vzdálilo od průměru celého kompozičního souboru. Shluky států zůstaly zachovány, jen Polsko (PL) se lehce vzdálilo. Všimneme-li si paprsků, tak můžeme pozorovat, že smrt následkem nervových poruch se značně odklonila od mentálních poruch. Co se týče horizontálního uspořádání, tak zhruba souhlasí s vertikálním uspořádáním z minulého grafu. Postupem zprava doleva postupně klesá výskyt mentálních po-

ruch. Vertikální uspořádání směrem zdola nahoru nejvíce ovlivňuje skupina nervových, respiračních, pohlavních, endokrinních chorob v sestupném charakteru a další skupina nehod, sebevražd a nemocí oběhové soustavy v charakteru vzestupném.

Závěr

Oblast analýzy kompozičních dat pro mě byla zpočátku úplně novým pojmem. Když jsem začal psát tuto diplomovou práci, obával jsem se, že nebudu mít k dispozici dostatek zdrojů, neboť je toto téma poměrně nové a stále se rozvíjející. Seznámení se s logpodílovým přístupem analýzy kompozičních dat mi netrvalo dlouho. Logpodílový přístup je již dobře zpracován v mnoha publikacích v angličtině, dokonce existují i zdroje v češtině mezi kvalifikačními pracemi na naší univerzitě. Musím ovšem přiznat, že poslední dvě kapitoly pro mě byly z teoretického hlediska poměrně náročné. Nastudovat a následně zprostředkovat tyto teoretické poznatky čtenářům se mi zpočátku zdálo složité.

Přesto jsem rád, že jsem si zvolil právě toto téma a nahlédnul tak mimo jiné i do historie kompozičních dat. Zajímavé pro mě bylo zejména to, jak probíhala komunikace mezi J. Aitchisonem a dalšími statistiky, ze které posléze vznikly diskutované principy statistické analýzy kompozičních dat. Je zajímavé, jak vývoj analýzy kompozičních dat ovlivnil požadavek podkompoziční soudržnosti jako podmínky relevantní statistické analýzy kompozičních dat. Tento požadavek ve své podstatě zamítal alternativní přístupy i přesto, že i samotná logpodílová metodika má se splněním tohoto požadavku v některých případech (při fixaci na konstantní součet složek kompozice) potíže. Její síla totiž spočívá v požadavku důsledného respektování invariance na změnu měřítka, kde ostatní přístupy nutně selhávají. Dále jsem se dozvěděl, že regresní analýza kompozičních dat založená na odmocninové transformaci je poměrně náročnou metodou ve srovnání s regresní analýzou založenou na logpodílových souřadnicích. Zajímavou metodou pro mě byla zejména robustní PCA v tangentovém prostoru. Zpočátku se mi metoda zdála značně složitá, především kvůli transformování dat na těžko představitelné a komplikované plochy, ale při tvorbě příkladu se mi její princip dostatečně ozřejmil. Potěšilo mě, že výsledky analýzy reálných dat založené na aplikaci této robustní PCA byly smysluplné, dobře interpretovatelné a porovnatelné s robustní

PCA využívající logpodílové souřadnice. Přesto, stejně jako u regresní analýzy, je žrejmé, že logpodílová metodika umožňuje dosažení relevantních výsledků daleko jednodušší cestou.

V závěru si tedy myslím, že cíl práce - seznámit sebe i čtenáře s alternativními přístupy ke statistické analýze kompozičních dat - se mi povedlo splnit. Samozřejmě je zde určitě prostor pro další zkoumání. Například zaměřit se detailněji na ostatní alternativní přístupy, nejen na přístup sférický, či blíže popsat problematiku složeného Kentova rozdělení. To již ale rád přenechám svým následovníkům.

Příloha A

Rekurzivní vztahy pro výpočet úhlového rozptylu a kovariance

Následující vztahy byly převzaty z literatury [23]. Platí

$$\begin{aligned}
\sigma^2_{u_{sh}} = & \prod_{i=h-1}^{m} (\cos^2\phi_i)\ \sigma^2_{\phi_{h-1}} + \sum_{i=h}^{m} \left(\sin^2\phi_{h-1} \sin^2\phi_i \prod_{\substack{j=h \\ j\neq i}}^{m} (\cos^2\phi_j)\ \sigma^2_{\phi_i} \right) \\
& - \sum_{i=h}^{m} \left(2\cos\phi_{h-1}\sin\phi_{h-1}\cos\phi_i\sin\phi_i \prod_{\substack{j=h \\ j\neq i}}^{m} (\cos^2\phi_j)\ \sigma_{\phi_{h-1}\phi_i} \right) \\
& + \sum_{i=h}^{m-1} \sum_{j=i+1}^{m} \left(2\sin^2\phi_{h-1}\cos\phi_i\sin\phi_i\cos\phi_j\sin\phi_j \prod_{\substack{k=h \\ k\neq i,\ k\neq j}}^{m} (\cos^2\phi_k)\ \sigma_{\phi_i\phi_j} \right), \\
\sigma_{u_{sg}u_{sh}} = & -\sin\phi_{g-1} \prod_{i=g}^{h-2} (\cos\phi_i) \cos\phi_{h-1}\sin\phi_{h-1} \prod_{i=h}^{m} (\cos^2\phi_i)\ \sigma^2_{\phi_{h-1}} \\
& + \sum_{i=h}^{m} \left(\sin\phi_{g-1} \prod_{j=g}^{h-2} (\cos\phi_j) \cos\phi_{h-1}\sin\phi_{h-1}\sin^2\phi_i \prod_{\substack{j=h \\ j\neq i}}^{m} (\cos^2\phi_j)\ \sigma^2_{\phi_i} \right) \\
& + \prod_{i=g-1}^{h-2} (\cos\phi_i) \prod_{i=h-1}^{m} (\cos^2\phi_i) \sigma_{\phi_{g-1}\phi_{h-1}} - \sum_{i=h}^{m} \left(\prod_{j=g-1}^{h-2} (\cos\phi_j) \times \right. \\
& \left. \times \cos\phi_{h-1}\sin\phi_{h-1}\cos\phi_i\sin\phi_i \prod_{\substack{j=h \\ j\neq i}}^{m} (\cos^2\phi_j)\ \sigma_{\phi_{g-1}\phi_i} \right) \\
& - \sum_{i=g}^{h-2} \left(\sin\phi_{g-1}\sin\phi_i \prod_{\substack{j=g \\ j\neq i}}^{h-2} (\cos\phi_j) \prod_{j=h-1}^{m} (\cos^2\phi_j)\ \sigma_{\phi_i\phi_{h-1}} \right) \\
& + \sum_{i=g}^{h-2} \sum_{j=h}^{m} \left(\sin\phi_{g-1}\sin\phi_i \prod_{\substack{k=g \\ k\neq i}}^{h-2} (\cos\phi_k) \times \right. \\
& \left. \times \cos\phi_{h-1}\sin\phi_{h-1}\cos\phi_j\sin\phi_j \prod_{\substack{k=h \\ k\neq j}}^{m} (\cos^2\phi_k)\ \sigma_{\phi_i\phi_j} \right)
\end{aligned}$$

$$+\sum_{i=h}^{m}\Bigg(\sin\phi_{g-1}\prod_{j=g}^{h-2}(\cos\phi_j)(\sin^2\phi_{h-1}-\cos^2\phi_{h-1})\times$$
$$\times\cos\phi_i\sin\phi_i\prod_{\substack{j=h\\ j\neq i}}^{m}(\cos^2\phi_j)\ \sigma_{\phi_{h-1}\phi_i}\Bigg)$$
$$+\sum_{i=h}^{m-1}\sum_{j=i+1}^{m}\Bigg(2\sin\phi_{g-1}\prod_{k=g}^{h-2}(\cos\phi_k)\cos\phi_{h-1}\sin\phi_{h-1}\cos\phi_i\sin\phi_i\times$$
$$\times\cos\phi_j\sin\phi_j\prod_{\substack{k=h\\ k\neq i,\ k\neq j}}^{m}(\cos^2\phi_k)\ \sigma_{\phi_i\phi_j}\Bigg),$$

kde $\sin\phi_0=1,\sigma^2_{\phi_0}=0,\sigma_{\phi_0\phi_g}=0,g=1,2,\ldots,h-1$ a $h=1,2,\ldots,m+1$, kde m je počet nezávislých proměnných, $m+1$ je počet složek kompozice.

Příloha B

Příklad: příčiny úmrtí - datový soubor

		Malignant neoplasms	Endocrine, nutritional and metabolic diseases	Mental and behavioural disorders	Diseases of the nervous system and the sense organs	Diseases of the circulatory system
BE	Belgium	27639	2693	3903	4757	31382
CZ	Czech Republic	27244	2801	939	2145	52963
DK	Denmark	15481	1692	3052	1868	13459
DE	Germany	221693	29835	27123	22741	342373
EE	Estonia	3629	191	122	266	8173
IE	Ireland	8667	634	808	1113	9236
EL	Greece	27341	1513	126	1517	47741
ES	Spain	105432	12789	15276	20254	118326
FR	France	155160	19263	19474	33769	141355
HR	Croatia	13701	1241	1048	877	25109
IT	Italy	167883	26801	15970	23257	223107
CY	Cyprus	1151	426	81	164	2119
LV	Latvia	5890	519	319	267	15703
LT	Lithuania	8070	313	77	551	23021
LU	Luxembourg	1060	86	99	162	1292
HU	Hungary	32670	3131	2764	1693	64249
MT	Malta	875	55	138	96	1491
NL	Netherlands	43017	3581	7600	5078	38460
AT	Austria	20104	3861	946	2473	32720
PL	Poland	92270	7137	1800	5400	170184
PT	Portugal	25593	5520	181	3107	31669
SI	Slovenia	5865	350	131	337	7298
SK	Slovakia	12917	941	445	991	23777
FI	Finland	11580	595	2071	5772	20226
SE	Sweden	21747	2355	5446	3564	34921
UK	United Kingdom	159674	7704	35378	21338	159256
LI	Liechtenstein	36	3	3	13	55
NO	Norway	10839	1017	2045	1649	12961
CH	Switzerland	16736	1569	4379	3307	21279
RS	Serbia	21007	3270	1120	1518	55513
TR	Turkey	65997	19634	512	11621	121143

		Diseasesof the respiratory system	Diseases of the digestive system	Diseases of the genito-urinary system	Intentional self-harm	Accidents
BE	Belgium	10580	4680	2561	4230	2118
CZ	Czech Republic	5726	4562	1218	4046	1622
DK	Denmark	5839	2470	795	1465	598
DE	Germany	60047	40536	19670	19254	10166
EE	Estonia	418	557	105	776	218
IE	Ireland	3438	1048	689	1011	554
EL	Greece	10335	2772	1782	2774	477
ES	Spain	42243	19577	11011	10253	3180
FR	France	34621	22971	8416	25486	10574
HR	Croatia	2062	2325	807	2048	712
IT	Italy	40559	22924	11103	18634	4156
CY	Cyprus	366	196	174	212	31
LV	Latvia	728	987	321	1145	441
LT	Lithuania	1241	2049	286	2097	1009
LU	Luxembourg	311	171	64	171	61
HU	Hungary	6594	7307	910	3939	2422
MT	Malta	234	104	81	71	22
NL	Netherlands	13343	5165	3510	4057	1709
AT	Austria	4076	3003	1320	2655	1308
PL	Poland	20006	16458	5226	14557	6120
PT	Portugal	11930	4556	2812	1823	1018
SI	Slovenia	1201	1113	322	971	443
SK	Slovakia	3072	3272	655	1637	536
FI	Finland	2027	2437	407	2755	910
SE	Sweden	5703	2658	1085	2992	1114
UK	United Kingdom	76070	27996	10881	13472	4157
LI	Lichtenstein	5	4	3	6	3
NO	Norway	4026	1283	860	1890	598
CH	Switzerland	3773	2485	931	2580	1067
RS	Serbia	5032	3513	2208	1591	1256
TR	Turkey	31384	7916	10094	10610	1148

Příloha C

Příklad: příčiny úmrtí - proporciální datový soubor

		Malignant neoplasms	Endocrine, nutritional and metabolic diseases	Mental and behavioural disorders	Diseases of the nervous system and the sense organs	Diseases of the circulatory system
BE	Belgium	0.292343	0.028484	0.041283	0.050316	0.331934
CZ	Czech Republic	0.263824	0.027124	0.009093	0.020772	0.512879
DK	Denmark	0.331364	0.036217	0.065327	0.039984	0.288084
DE	Germany	0.279408	0.037602	0.034184	0.028661	0.431506
EE	Estonia	0.251055	0.013213	0.008440	0.018402	0.565410
IE	Ireland	0.318663	0.023311	0.029708	0.040922	0.339584
EL	Greece	0.283685	0.015699	0.001307	0.015740	0.495352
ES	Spain	0.294223	0.035689	0.042630	0.056522	0.330205
FR	France	0.329365	0.040890	0.041338	0.071683	0.300060
HR	Croatia	0.274404	0.024855	0.020989	0.017565	0.502884
IT	Italy	0.302823	0.048343	0.028806	0.041950	0.402434
CY	Cyprus	0.233943	0.086585	0.016463	0.033333	0.430691
LV	Latvia	0.223784	0.019719	0.012120	0.010144	0.596619
LT	Lithuania	0.208452	0.008085	0.001989	0.014233	0.594643
LU	Luxembourg	0.304861	0.024734	0.028473	0.046592	0.371585
HU	Hungary	0.259948	0.024913	0.021993	0.013471	0.511215
MT	Malta	0.276287	0.017367	0.043574	0.030313	0.470793
NL	Netherlands	0.342710	0.028529	0.060548	0.040456	0.306405
AT	Austria	0.277427	0.053280	0.013054	0.034126	0.451522
PL	Poland	0.272056	0.021043	0.005307	0.015922	0.501784
PT	Portugal	0.290140	0.062579	0.002052	0.035223	0.359022
SI	Slovenia	0.325273	0.019411	0.007265	0.018690	0.404747
SK	Slovakia	0.267749	0.019505	0.009224	0.020542	0.492859
FI	Finland	0.237392	0.012198	0.042456	0.118327	0.414637
SE	Sweden	0.266556	0.028866	0.066752	0.043685	0.428032
UK	United Kingdom	0.309490	0.014932	0.068572	0.041359	0.308680
LI	Lichtenstein	0.274809	0.022901	0.022901	0.099237	0.419847
NO	Norway	0.291622	0.027362	0.055020	0.044366	0.348714
CH	Switzerland	0.288025	0.027002	0.075362	0.056913	0.366210
RS	Serbia	0.218759	0.034053	0.011663	0.015808	0.578092
TR	Turkey	0.235654	0.070107	0.001828	0.041495	0.432562

		Diseasesof the respiratory system	Diseases of the digestive system	Diseases of the genito-urinary system	Intentional self-harm	Accidents
BE	Belgium	0.111907	0.049501	0.027088	0.044742	0.022403
CZ	Czech Republic	0.055449	0.044177	0.011795	0.039180	0.015707
DK	Denmark	0.124981	0.052869	0.017017	0.031358	0.012800
DE	Germany	0.075680	0.051089	0.024791	0.024267	0.012813
EE	Estonia	0.028917	0.038533	0.007264	0.053684	0.015081
IE	Ireland	0.126406	0.038532	0.025333	0.037172	0.020369
EL	Greece	0.107234	0.028762	0.018490	0.028783	0.004949
ES	Spain	0.117885	0.054632	0.030728	0.028612	0.008874
FR	France	0.073491	0.048761	0.017865	0.054100	0.022446
HR	Croatia	0.041298	0.046565	0.016163	0.041017	0.014260
IT	Italy	0.073159	0.041350	0.020027	0.033611	0.007496
CY	Cyprus	0.074390	0.039837	0.035366	0.043089	0.006301
LV	Latvia	0.027660	0.037500	0.012196	0.043503	0.016755
LT	Lithuania	0.032056	0.052927	0.007388	0.054166	0.026063
LU	Luxembourg	0.089445	0.049180	0.018407	0.049180	0.017544
HU	Hungary	0.052467	0.058140	0.007241	0.031342	0.019271
MT	Malta	0.073887	0.032839	0.025576	0.022419	0.006947
NL	Netherlands	0.106302	0.041149	0.027964	0.032322	0.013615
AT	Austria	0.056247	0.041440	0.018215	0.036638	0.018050
PL	Poland	0.058987	0.048526	0.015409	0.042921	0.018045
PT	Portugal	0.135247	0.051650	0.031879	0.020667	0.011541
SI	Slovenia	0.066608	0.061727	0.017858	0.053852	0.024569
SK	Slovakia	0.063678	0.067823	0.013577	0.033932	0.011110
FI	Finland	0.041554	0.049959	0.008344	0.056478	0.018655
SE	Sweden	0.069903	0.032580	0.013299	0.036673	0.013654
UK	United Kingdom	0.147444	0.054264	0.021090	0.026112	0.008057
LI	Lichtenstein	0.038168	0.030534	0.022901	0.045802	0.022901
NO	Norway	0.108319	0.034519	0.023138	0.050850	0.016089
CH	Switzerland	0.064933	0.042767	0.016022	0.044402	0.018363
RS	Serbia	0.052401	0.036583	0.022993	0.016568	0.013080
TR	Turkey	0.112062	0.028265	0.036042	0.037885	0.004099

Příloha D

Příklad: příčiny úmrtí - datový soubor po logaritmické transformaci

		Malignant neoplasms	Endocrine, nutritional and metabolic diseases	Mental and behavioural disorders	Diseases of the nervous system and the sense organs	Diseases of the circulatory system
BE	Belgium	-1.229827	-3.558399	-3.187309	-2.989438	-1.102820
CZ	Czech Republic	-1.332475	-3.607332	-4.700248	-3.874169	-0.667715
DK	Denmark	-1.104537	-3.318240	-2.728354	-3.219283	-1.244503
DE	Germany	-1.275082	-3.280693	-3.375993	-3.552206	-0.840475
EE	Estonia	-1.382083	-4.326522	-4.774775	-3.995299	-0.570204
IE	Ireland	-1.143621	-3.758850	-3.516337	-3.196084	-1.080035
EL	Greece	-1.259891	-4.154184	-6.639751	-4.151543	-0.702487
ES	Spain	-1.223419	-3.332900	-3.155202	-2.873133	-1.108042
FR	France	-1.110590	-3.196861	-3.185967	-2.635504	-1.203773
HR	Croatia	-1.293153	-3.694705	-3.863738	-4.041870	-0.687396
IT	Italy	-1.194608	-3.029436	-3.547164	-3.171269	-0.910224
CY	Cyprus	-1.452677	-2.446624	-4.106615	-3.401197	-0.842364
LV	Latvia	-1.497073	-3.926181	-4.412893	-4.590836	-0.516477
LT	Lithuania	-1.568048	-4.817753	-6.220151	-4.252222	-0.519794
LU	Luxembourg	-1.187901	-3.699578	-3.558805	-3.066329	-0.989978
HU	Hungary	-1.347274	-3.692379	-3.817052	-4.307229	-0.670965
MT	Malta	-1.286316	-4.053207	-3.133286	-3.496192	-0.753338
NL	Netherlands	-1.070870	-3.556823	-2.804317	-3.207548	-1.182846
AT	Austria	-1.282199	-2.932191	-4.338630	-3.377685	-0.795131
PL	Poland	-1.301747	-3.861174	-5.238679	-4.140067	-0.689586
PT	Portugal	-1.237390	-2.771331	-6.188967	-3.346051	-1.024371
SI	Slovenia	-1.123090	-3.941915	-4.924650	-3.979765	-0.904492
SK	Slovakia	-1.317706	-3.937063	-4.685932	-3.885291	-0.707532
FI	Finland	-1.438041	-4.406514	-3.159289	-2.134302	-0.880351
SE	Sweden	-1.322170	-3.545105	-2.706764	-3.130762	-0.848557
UK	United Kingdom	-1.172829	-4.204224	-2.679873	-3.185474	-1.175450
LI	Lichtenstein	-1.291678	-3.776585	-3.776585	-2.310248	-0.867864
NO	Norway	-1.232297	-3.598591	-2.900050	-3.115279	-1.053503
CH	Switzerland	-1.244707	-3.611830	-2.585449	-2.866227	-1.004548
RS	Serbia	-1.519784	-3.379850	-4.451311	-4.147246	-0.548023
TR	Turkey	-1.445391	-2.657738	-6.304431	-3.182186	-0.838029

		Diseasesof the respiratory system	Diseases of the digestive system	Diseases of the genito-urinary system	Intentional self-harm	Accidents
BE	Belgium	-2.190089	-3.005757	-3.608657	-3.106853	-3.798583
CZ	Czech Republic	-2.892291	-3.119547	-4.440098	-3.239579	-4.153648
DK	Denmark	-2.079591	-2.939933	-4.073564	-3.462296	-4.358315
DE	Germany	-2.581248	-2.974185	-3.697281	-3.718657	-4.357327
EE	Estonia	-3.543314	-3.256230	-4.924835	-2.924643	-4.194301
IE	Ireland	-2.068254	-3.256260	-3.675657	-3.292204	-3.893734
EL	Greece	-2.232742	-3.548709	-3.990542	-3.547988	-5.308517
ES	Spain	-2.138046	-2.907130	-3.482590	-3.553915	-4.724604
FR	France	-2.610587	-3.020814	-4.024912	-2.916918	-3.796649
HR	Croatia	-3.186946	-3.066902	-4.125054	-3.193758	-4.250299
IT	Italy	-2.615118	-3.185691	-3.910660	-3.392888	-4.893323
CY	Cyprus	-2.598430	-3.222949	-3.342009	-3.144478	-5.067077
LV	Latvia	-3.587783	-3.283414	-4.406643	-3.134924	-4.089040
LT	Lithuania	-3.440284	-2.938849	-4.907965	-2.915694	-3.647242
LU	Luxembourg	-2.414132	-3.012262	-3.995042	-3.012262	-4.043051
HU	Hungary	-2.947571	-2.844898	-4.928042	-3.462804	-3.949137
MT	Malta	-2.605219	-3.416149	-3.666091	-3.797860	-4.969498
NL	Netherlands	-2.241473	-3.190560	-3.576849	-3.432021	-4.296557
AT	Austria	-2.878001	-3.183506	-4.005486	-3.306673	-4.014618
PL	Poland	-2.830434	-3.025654	-4.172820	-3.148394	-4.014904
PT	Portugal	-2.000653	-2.963264	-3.445813	-3.879226	-4.461869
SI	Slovenia	-2.708938	-2.785033	-4.025296	-2.921521	-3.706278
SK	Slovakia	-2.753922	-2.690849	-4.299371	-3.383385	-4.499872
FI	Finland	-3.180764	-2.996553	-4.786262	-2.873903	-3.981631
SE	Sweden	-2.660653	-3.424071	-4.320065	-3.305703	-4.293688
UK	United Kingdom	-1.914309	-2.913902	-3.858945	-3.645350	-4.821170
LI	Lichtenstein	-3.265759	-3.488903	-3.776585	-3.083438	-3.776585
NO	Norway	-2.222675	-3.366247	-3.766271	-2.978871	-4.129613
CH	Switzerland	-2.734398	-3.151996	-4.133765	-3.114480	-3.997418
RS	Serbia	-2.948822	-3.308169	-3.772553	-4.100277	-4.336708
TR	Turkey	-2.188702	-3.566114	-3.323059	-3.273203	-5.496979

Příloha E

Příklad: příčiny úmrtí - projekce na tangentový prostor

		Malignant neoplasms	Endocrine, nutritional and metabolic diseases	Mental and behavioural disorders	Diseases of the nervous system and the sense organs	Diseases of the circulatory system
BE	Belgium	0.063968	0.005305	0.381653	0.315436	-0.609089
CZ	Czech Republic	-0.171419	-0.030973	-0.584508	-0.371672	0.564758
DK	Denmark	-0.171419	-0.030973	-0.584508	-0.371672	0.564758
DE	Germany	0.297980	0.141506	0.489591	0.111494	-0.633191
EE	Estonia	-0.160124	0.344913	0.442187	-0.264315	0.104936
IE	Ireland	-0.093351	-0.281827	-0.340661	-0.241238	0.550060
EL	Greece	0.267066	-0.146569	0.172029	0.156291	-0.572696
ES	Spain	0.087221	0.154216	0.360220	0.365506	-0.528543
FR	France	0.297423	0.219556	0.309203	0.477913	-0.656792
HR	Croatia	-0.089459	-0.094836	-0.061480	-0.511821	0.520779
IT	Italy	0.169454	0.638064	0.228649	0.266406	-0.223824
CY	Cyprus	-0.371871	0.732267	-0.187080	0.000580	0.133295
LV	Latvia	-0.208416	-0.110724	-0.193786	-0.462847	0.609334
LT	Lithuania	-0.179643	-0.322395	-0.595156	-0.232557	0.509985
LU	Luxembourg	0.198514	-0.194377	0.219882	0.416302	-0.615200
HU	Hungary	-0.157237	-0.071587	-0.020798	-0.591690	0.493998
MT	Malta	-0.087440	-0.385358	0.444548	-0.097549	0.365643
NL	Netherlands	0.376194	0.015346	0.515947	0.125872	-0.627516
AT	Austria	-0.125507	0.622907	-0.527790	-0.004463	0.280140
PL	Poland	-0.064531	-0.159088	-0.726323	-0.448900	0.422729
PT	Portugal	0.081920	0.312761	-0.791438	0.037174	-0.130836
SI	Slovenia	0.277523	-0.245881	-0.678733	-0.422098	-0.055706
SK	Slovakia	-0.139730	-0.262403	-0.579731	-0.382898	0.462254
FI	Finland	-0.189657	-0.365430	0.277118	0.659961	0.094564
SE	Sweden	-0.192705	0.013780	0.838119	0.242033	0.098029
UK	United Kingdom	0.194719	-0.284911	0.490398	0.128881	-0.456613
LI	Lichtenstein	-0.026170	-0.111366	0.008113	0.699396	0.077262
NO	Norway	0.060944	-0.024060	0.584531	0.222072	-0.483865
CH	Switzerland	0.045872	-0.029351	0.797898	0.428331	-0.327045
RS	Serbia	-0.336723	0.102044	-0.270622	-0.368340	0.654870
TR	Turkey	-0.145498	0.335987	-0.778373	0.099889	0.134243

		Diseasesof the respiratory system	Diseases of the digestive system	Diseases of the genito-urinary system	Intentional self-harm	Accidents
BE	Belgium	0.496699	0.094228	0.215943	0.152485	0.232937
CZ	Czech Republic	-0.310702	-0.011376	-0.265578	0.038388	0.047367
DK	Denmark	0.462420	0.128392	-0.026849	-0.080078	-0.025940
DE	Germany	0.031301	0.187253	0.273188	-0.674264	-0.116982
EE	Estonia	-0.558017	-0.048158	-0.290819	0.189320	0.030465
IE	Ireland	0.660445	-0.140444	0.182002	-0.003154	0.190119
EL	Greece	0.228464	-0.158138	0.009743	-0.076566	-0.231653
ES	Spain	0.500552	0.169259	0.260643	-0.186443	-0.209768
FR	France	0.022856	0.076154	-0.014672	0.255027	0.190724
HR	Croatia	-0.656595	0.034331	-0.089589	0.075693	-0.001015
IT	Italy	-0.034874	-0.149171	0.044693	-0.173265	-0.579071
CY	Cyprus	0.033789	-0.081181	0.340606	0.118409	-0.367375
LV	Latvia	-0.542056	-0.052052	-0.111832	0.092995	0.059549
LT	Lithuania	-0.323414	0.099321	-0.189393	0.156865	0.143091
LU	Luxembourg	0.371150	0.122020	-0.041488	0.373679	0.169366
HU	Hungary	-0.309004	0.201598	-0.457935	-0.113212	0.130158
MT	Malta	0.006785	-0.307060	0.191932	-0.459976	-0.400340
NL	Netherlands	0.366977	-0.046386	0.196616	-0.082585	-0.010944
AT	Austria	-0.445527	-0.112442	-0.037266	-0.035786	0.159272
PL	Poland	-0.182539	0.063042	-0.085797	0.094537	0.095796
PT	Portugal	0.396704	0.091057	0.174746	-0.229400	-0.042080
SI	Slovenia	-0.096551	0.265564	-0.027092	0.275983	0.252644
SK	Slovakia	-0.157030	0.366273	-0.187076	-0.077016	-0.124498
FI	Finland	-0.380876	0.095310	-0.286728	0.248348	0.105666
SE	Sweden	-0.067794	-0.351741	-0.249988	-0.020189	-0.028675
UK	United Kingdom	0.574190	0.143695	0.062169	-0.171100	-0.178412
LI	Lichtenstein	-0.588693	-0.263390	0.103625	0.151777	0.204975
NO	Norway	0.471572	-0.239187	0.126561	0.266009	0.064966
CH	Switzerland	-0.131615	-0.033582	-0.092060	0.155309	0.137554
RS	Serbia	-0.222414	-0.096420	0.089400	-0.414009	-0.016692
TR	Turkey	0.271429	-0.175427	0.201065	0.029334	-0.293843

Literatura

[1] Aitchison, J.: *A concise guide to compositional data analysis*, 2nd Compositional Data Analysis Workshop, Girona, 2003.

[2] Aitchison, J.: *Letter to the Editor. Comment on: „Measures of variability for geological data“* by D. F. Watson and G. M. Philip. Mathematical Geology **22** (1990), s. 223–226.

[3] Aitchison, J.: *Letter to the Editor. Delusions of uniqueness and ineluctability.* Mathematical Geology **23** (1991), s. 275–277.

[4] Aitchison, J.: *The Statistical Analysis of Compositional Data*. Chapman and Hall, London, 1986.

[5] Aitchison, J.: *The Statistical Analysis of Compositional Data*. Journal of the Royal Statistical Society: Series B (Methodological) **44** (1982), s. 139–177.

[6] *Eurostat Database - Causes of death - Deaths by country of residence and occurrence* [online]. [cit. 2016-02-18]. Dostupné z: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=hlth_cd_aro&lang=en.

[7] Chayes, F.: *On correlation between variables of constant sum.* Journal of Geophysical Research **65** (1960), s. 4185–4193.

[8] Janson, S.: *Probability asymptotics: Notes on notation.* [online]. 2009, [cit. 2017-02-20]. Dostupné z: http://www2.math.uu.se/ svante/papers/sjN6.pdf.

[9] Johnson, R. A., Wichern, D. W.: *Applied Multivariate Statistical Analysis.* Prentice Hall, Englewood Cliffs, 1992.

[10] Kalivodová, A.: *Diplomová práce: Kompoziční biplot.* Univerzita Palackého, Olomouc, 2012.

[11] Maronna, R., Martin, D., Yohai, V.: *Robust Statistics: Theory and Methods.* John Wiley & Sons, Chichester, 2006.

[12] Murnaghan, F. D.: *The element of volume of the rotation group.* Proceedings of the National Academy of Sciences of the United States of America **36** (1950), s. 670–672.

[13] Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R.: *Lecture Notes on Compositional Data Analysis.* [online]. 2007, [cit. 2016-01-25]. Dostupné z: http://www.sediment.uni-goettingen.de/staff/tolosana/extra/CoDa.pdf.

[14] Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R.: *Modeling and Analysis of Compositional Data.* John Wiley & Sons, Chichester, 2015.

[15] Pearson, K.: *Mathematical contributions on the theory of evolution - On a form of spurious correlation which may arise when indices are used in the measurement of organs.* Proceedings of the Royal Society of London **60** (1896), s. 489–498.

[16] Philip, G. M., Watson, D. F.: *Angles Measure Compositional Differences.* Geology **16** (1988), s. 976–979.

[17] Scealy, J. L., De Caritat, P., Grunsky, E. C., Tsagris, M. T., Welsh, A. H.: *Robust principal component analysis for power transformed compositional data.* Journal of the American Statistical Association **110** (2015), s. 136–148.

[18] Scealy, J. L., Welsh, A. H.: *Colours and cocktails: compositional data analysis.* Australian & New Zealand Journal of Statistics **56** (2014), s. 145–169.

[19] Scealy, J. L., Welsh, A. H.: *Fitting Kent models to compositional data with small concentration.* Statistics and Computing **24** (2014), s. 165–179.

[20] Scealy, J. L., Welsh, A. H.: *Properties of a square root transformation regression model.* Proceedings of the 4th International Workshop on Compositional Data Analysis, Girona, 2011.

[21] Scealy, J. L., Welsh, A. H.: *Regression for compositional data by using distributions defined on the hypersphere.* Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73** (2011), s. 351–375.

[22] Small, Ch. G.: *The Statistical Theory of Shape (Springer Series in Statistics).* Springer-Verlag, New York, 1996, s. 43–47.

[23] Stanley, C. R.: *Descriptive statistics for N-dimensional closed arrays: A spherical coordinate approach.* Mathematical Geology **22** (1990), s. 933–956.

[24] Watson, D. F.: *Letter to the Editor. Reply to comment on „Measures of variability for geological data“* by D.F. Watson and G.M. Philip. Journal of Mathematical Geology **22** (1990), s. 227–231.

[25] Watson, D. F.: *Letter to the Editor. Reply to „Delusions of uniqueness and ineluctability“* by J. Aitchison. Mathematical Geology **23** (1991), s. 279.

[26] Watson, D. F., Philip, G. M.: *Measures of variability for geological data.* Mathematical Geology **21** (1989), s. 233–254.