UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

BAKALÁŘSKÁ PRÁCE

Interpretace odlehlých hodnot v kompozičních datech

Katedra matematické analýzy a aplikací matematiky
Vedoucí bakalářské práce: **doc. RNDr. Karel Hron, Ph.D.**
Vypracoval: **Patrik Vidlář**
Studijní program: B1103 Aplikovaná matematika
Studijní obor : Aplikovaná statistika
Forma studia: prezenční
Rok odevzdání: 2015

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Patrik Vidlář

Název práce: Interpretace odlehlých hodnot v kompozičních datech

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

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

Rok obhajoby práce: 2015

Abstrakt: Odlehlá pozorování se vyskytují téměř ve všech typech datových souborů a kompoziční data nejsou výjimkou. Tato práce se zabývá právě takovýmto typem odlehlých pozorování a jejich interpretací prostřednictvím různých vizualizačních technik, které jsou součástí knihovny mvoutlier statistického softwaru R. V práci je uvedena teorie nutná pro pochopení problematiky kompozičních dat včetně jejich vlastností. Dále jsou zde uvedeny rozdíly mezi jednorozměrnými a mnohorozměrnými odlehlými pozorováními a také metody jejich detekce. V praktické části je hlavním úkolem zkoumat výskyt odlehlých hodnot v oblastech školství a zaměstnanosti v rámci regionů Spolkové republiky Německo. Na konkrétních datech jsou zobrazeny různé typy grafů doplněné o odpovídající komentář a také příkazy, které jsou při jejich tvorbě použity.

Klíčová slova: Kompoziční data, odlehlá pozorování, Mahalanobisova vzálenost, logratio souřadnice.

Počet stran: 41

Počet příloh: 2

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Patrik Vidlář

Title: Interpreation of multivariate outliers in compositional data

Type of thesis: Bachelor's

Department: Department of Mathematical Analysis and Application of Mathematics

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

The year of presentation: 2015

Abstract: Outliers can be found in almost every type of data files. Compositional data are no exception. The bachelor thesis deals with this particular type of extrem values and their interpretation using different types of visualization techniques. These techniques appear to be a part of the mvoutlier library of the statistical software R. The theoretical part of the bachelor thesis is devoted to understand the compositional data issues as well as their attributes. The thesis then continues to deal with differences between univariate and multivariate outliers and their subsequent detection. The aim of the practical part is to explore outliers in educational system and employment structure of German regions. Specific data are used to display different types of graphs as well as commands that are used for their visualization.

Key words: Compositional data, outliers, Mahalanobis distance, logratio coordinates.

Number of pages: 41

Number of appendices: 2

Language: Czech

Prohlášení

Prohlašuji, že jsem vytvořil tuto bakalářskou práci samostatně za vedení doc. RNDr. Karla Hrona, Ph.D. a že jsem v seznamu použité literatury uvedl všechny zdroje použité při zpracování práce.

V Olomouci dne 30. dubna 2015

Poděkování

Rád bych na tomto místě poděkoval vedoucímu diplomové práce doc. RNDr. Karlovi Hronovi, Ph.D. za obětavou spolupráci, trpělivost i za čas, který mi věnoval při konzultacích. Za korekturu českého jazyka bych chtěl poděkovat Mgr. Ivě Petříkové. Dále bych chtěl poděkovat mé rodině, přítelkyni a přátelům, kteří mě po celou dobu studia podporovali.

Obsah

Úvod

Jako téma své bakalářské práce jsem si zvolil interpretaci odlehlých hodnot v kompozičních datech, mnohorozměrných pozorováních nesoucích relativní informaci (speciálně procenta či proporce), se kterými se můžeme setkat ve velkém množství oborů, například geologie, chemie, ekonomie atd. Kompoziční data jsou v rámci statistiky stále relativně novou disciplínou, a jejich specifické vlastnosti je třeba při statistickém zpracování zohlednit. Standardní metody, aplikované přímo na původní kompoziční data, totiž udávájí nepřesné, či zcela chybné výsledky.

Odlehlá pozorování se vyskytují téměř v každém datovém souboru, ale ne vždy jsme schopni přesně určit, jak vznikají. V některých případech se tato pozorování z datového souboru vyloučí úplně. Tímto se však můžeme připravit o důležité informace o jevech, jež se vyskytují s nízkou pravděpodobností. Na místo toho je vhodné odlehlá pozorování detekovat a následně vhodně interpretovat v rámci mnohorozměrné struktury datového souboru.

V teoretické části této bakalářské práce nejdříve stručně představím kompoziční data a jejich výběrový prostor, Aitchisonovu geometrii, dále roli odlehlých pozorování, metody jejich detekce, nástroje pro vizualizaci odlehlých hodnot v kompozičních datech, a knihovny `mvoutlier` a `robCompositions` statistického softwaru R.

V praktické části se již budu zabývat konkrétními datovými soubory, s primárním úkolem nalézt a vhodně interpretovat odlehlá pozorování. První datový soubor se týká struktury vzdělání, druhý struktury zaměstnanosti na území

Německa. V práci se zaměřím především na to, zda se projeví nějakým způsobem dřívější rozdělení Německa na Německou demokratickou republiku a Německou spolkovou republiku. Na oba datové soubory budu aplikovat metody pro detekci odlehlých hodnot pomocí knihovny `mvoutlier` a `robCompositions` statistického softwaru R.

1. Kompoziční data

Při psaní této kapitoly jsem vycházel zejména ze zdrojů [1], [5].

Kompoziční data řadíme mezi mnohorozměrná pozorování. Na rozdíl od standardních mnohorozměrných dat nenesou absolutní informaci, nesou však informaci relativní. Populárně jsou kompoziční data reprezentována s konstantním součtem složek. Pod tím si můžeme představit například procentuální podíly (což jsou data s konstantním součtem složek 100) nebo proporce (jejichž součet složek je roven 1). Máme-li kompoziční data reprezentována takto, není možné interpretovat jejich jednotlivé složky zvlášť, ale pouze jako součást celku.

Standardní metody v případě použití na původní kompoziční data selhávají z několika důvodů. Jedním z nich je fakt, že tyto postupy jsou sestrojeny pro data, která mohou nabývat libovolných hodnot na reálné ose. Pro tento typ dat nám běžné metody poskytují širokou paletu možností jejich statistické analýzy. Oproti tomu kompoziční data mohou vždy nabývat pouze kladných hodnot a v případě procent pouze hodnot mezi 0 a 100. Pokud hodnota jedné složky kompozice vzroste, hodnota alespoň jedné libovolné jiné složky musí poklesnout, protože se jedná o data s konstantním součtem. Toto úzce souvisí s výběrovým prostorem kompozičních dat; D-složkovou kompozicí budeme rozumět sloupcový vektor $\mathbf{x} = (x_1, \ldots, x_D)^T$, jestliže jsou všechny jeho složky kladná čísla, a jenž je možné beze ztráty informace reprezentovat s libovolným součtem složek.

Definice 1.1. *Výběrový prostor D-složkových kompozičních dat* $\mathbf{x} = (x_1, \ldots, x_D)^T$ *je tzv. simplex, definovaný jako*

$$S^D = \left\{ \mathbf{x} = (x_1, x_2, \ldots, x_D)^T, x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = \kappa \right\}.$$

Dalším důvodem proti statistickému zpracování původních kompozičních dat je skutečnost, že se kompoziční data neřídí standardní euklidovskou geometrií, ale tzv. Aitchisonovou geometrií na simplexu. Abychom mohli aplikovat standardní metody na kompoziční data, musíme nejdříve kompozice vhodně převést pomocí logratio souřadnic z D-složkového simplexu do reálného prostoru $\mathbb{R}^{D-1}$, resp. $\mathbb{R}^D$.

1.1. Vlastnosti kompozičních dat

Při psaní této kapitoly jsem vycházel zejména ze zdrojů [3] [5].

Jakákoliv statistická metoda použitá pro analýzu kompozičních dat by měla mít tyto tři vlastnosti: invariantnost na změnu škály, invariantnost na permutaci složek kompozice a podkompoziční soudržnost.

Jak již bylo uvedeno, nejdůležitější vlastností kompozičních dat je, že nesou pouze relativní informaci. Toto si nejlépe ukážeme na příkladu. Mějme třísložkové kompozice s hodnotami složek $[140, 40, 20]$ a $[35, 10, 5]$. Je vidět, že se absolutní hodnoty liší, nicméně pokud se podíváme na relativní zastoupení složek, zjistíme, že dostaneme v obou případech stejné procentuální zastoupení, a to $[70, 20, 10]$. Invariantnost na změnu škály pak říká, že libovolná reprezentace kompozic nemá

mít vliv na výsledek jejich statistické analýzy.

Invariance na permutaci složek znamená, že při statistické analýze nezáleží na pořadí jednotlivých složek kompozice. I při různých permutacích dostaneme stejné výsledky.

Podkompozice získáme jako podvektory z celkových kompozic, které případně reprezentujeme se zvoleným konstantním součtem složek. Podstatou podkompoziční soudržnosti je požadavek, aby se podkompozice chovaly jako ortogonální projekce ve standardním euklidovském prostoru. Speciálně vzdálenost mezi kompozicemi musí být větší nebo rovna vzdálenosti mezi dvěma odpovídajícími podkompozicemi.

Jak již bylo zmíněno, uvedené geometrické vlastnosti kompozičních dat jsou charakterizovány pomocí tzv. Aitchisonovy geometrie, která též respektuje jejich relativní měřítko. Pro představu si uvedeme jednoduchý příklad. Mějme třísložkové kompozice $(5, 65, 30)$, $(10, 60, 30)$, $(50, 20, 30)$ a $(55, 15, 30)$. Na první pohled můžeme intuitivně vycítit, že rozdíl mezi kompozicemi $(5, 65, 30)$ a $(10, 60, 30)$ není stejný jako u kompozic $(50, 20, 30)$ a $(55, 15, 30)$, přestože euklidovská vzdálenost mezi nimi je naprosto totožná. V obou případech je rozdíl mezi hodnotami první složky u první a druhé kompozice přesně pět jednotek. Přitom ovšem u první dvojice kompozic se jedná o zdvojnásobení původní hodnoty, kdežto u druhé dvojice je tento nárůst jen 10 %. Právě tento relativní rozdíl se zdá být vhodnější pro popis variability kompozičních dat, a je tak dalším důvodem, proč euklidovská geometrie není pro kompoziční data vhodná.

Při konstrukci Aitchisonovy geometrie nejprve zavedeme dvě operace, díky

nimž na simplexu získáme vektorový prostor. Jedná se o operace perturbace a mocninná trasformace. V případě perturbace jde o analogii k vektorovému sčítání v reálném prostoru, v případě mocninné transformace se jedná o obdobu násobení vektoru skalárem. Operaci uzávěru (značíme C) užíváme pro případné vyjádření kompozice s předepsaným součtem složek κ.

Definice 1.2. *Perturbace kompozic* $\mathbf{x} \in S^D$ a $\mathbf{y} \in S^D$ *je kompozice*

$$\mathbf{x} \oplus \mathbf{y} = C(x_1 y_1, x_2 y_2, \ldots, x_D y_D)^T.$$

Definice 1.3. *Mocninná transformace kompozice* $\mathbf{x} \in S^D$ *reálným číslem* $\alpha \in \mathbb{R}$ *je kompozice*

$$\alpha \odot \mathbf{x} = C(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha)^T.$$

Takto vzniklý vektorový prostor budeme značit $(S^D, \oplus, \odot)$. Kromě právě zmíněných operací jsme na simplexu schopni definovat skalární součin, k němu příslušnou normu a vzdálenost.

Definice 1.4. *Aitchisonův skalární součin kompozic* $\mathbf{x}, \mathbf{y} \in S^D$ *je definován jako*

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{i=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}.$$

Definice 1.5. *Aitchisonova norma je zavedena jako norma kompozice* $\mathbf{x} \in S^D$

$$\|\mathbf{x}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} \right)^2}.$$

Definice 1.6. *Aitchisonova vzdálenost je zavedena jako vzdálenost mezi kompozicemi* $\mathbf{x}, \mathbf{y} \in S^D$

$$d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\| = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}.$$

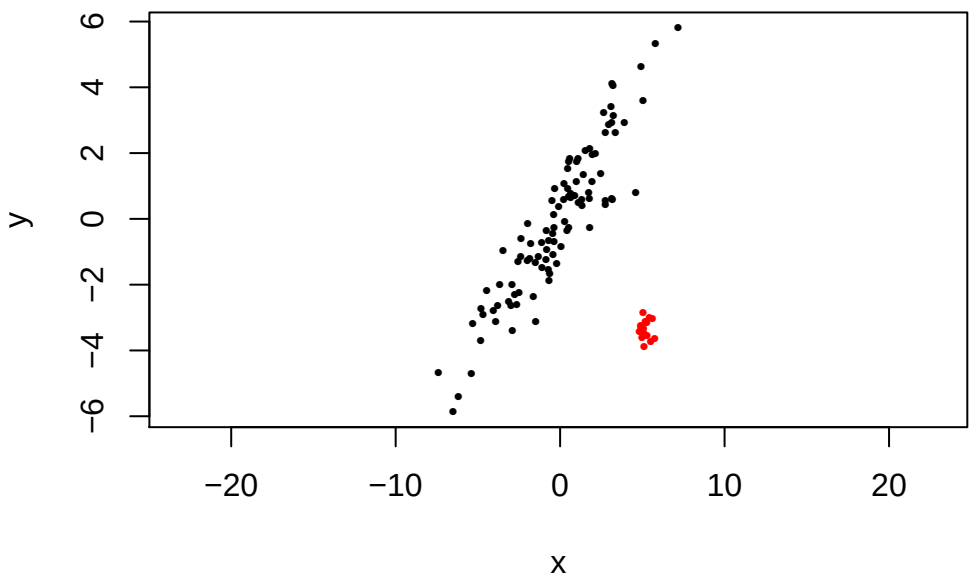
Vektorový prostor $(S^D, \oplus, \odot)$, a na něm definovaný skalární součin, norma a vzdálenost se souhrnně nazývá Aitchisonova geometrie na simplexu a má stejné vlastnosti jako euklidovská geometrie v reálném prostoru.

1.2. Odlehlá pozorování

Odlehlá pozorování jsou v případě jednorozměrných dat hodnoty, které se výrazně liší od zbylého datového souboru. Není vždy jednoduché určit, jak tato pozorování vznikají. Může se jednat o selhání měřicího zařízení nebo pouhý překlep, pokud data vypisuje člověk. Nicméně odlehlá pozorování mohou vzniknout i pouhým vlivem náhody, kdy se nemusí jednat o žádnou systematickou chybu.

V mnohorozměrné situaci je mnohem složitější rozhodnout, jaká pozorování jsou odlehlá, protože pozorování např. nemusí být odlehlé v žádné z jednotlivých proměnných, ale jen v jejich určité kombinaci. Při práci s odlehlými pozorováními můžeme narazit na některé efekty, které nám jejich detekci ještě zkomplikují. Může se jednat o tzv. maskovací efekt, který spočívá ve skrytí jednoho odlehlého pozorování nějakým jiným. Při analýze (podmnožiny) statistických znaků se pozorování nejeví jako odlehlé. Odlehlost závisí na celkové struktuře našich po-

zorování. V důsledku shluk odlehlých pozorování zkresluje výběrový průměr a výběrovou varianční matici jako odhady příslušných teoretických charakteristik. Tomuto problému se snažíme vyhnout použitím robustních metod. V praktické



Obrázek 1: Shluk odlehlých pozorování (červeně) v rámci datového souboru

části si ukážeme použití standardních i robustních metod a budeme zjišťovat, zda a jak se mění počet nalezených odlehlých pozorování. Dalším z těchto efektů může být tzv. swamping efekt. Ten naopak říká, že pozorování je odlehlé, ale pouze v přítomnosti jiného odlehlého pozorování. Pokud bychom první pozorování odstranili, z druhého, které bylo původně považováno za odlehlé, se stane pozorování, které již za odlehlé nepovažujeme.

1.3. Detekce odlehlých hodnot

Při tvorbě této kapitoly jsem vycházel zejména z [1], [2].

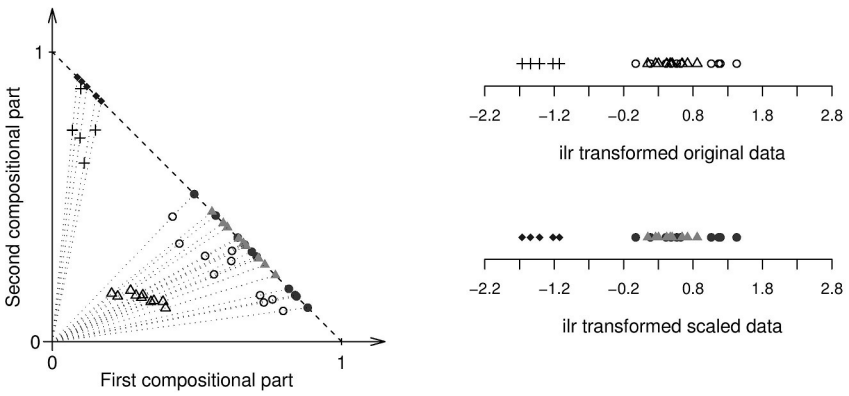
Pro detekci mnohorozměrných odlehlých pozorování se nejčastěji používá jedna, ze dvou následujících metod. První z nich je tzv. projekce, založená na převedení dat do jednorozměrného prostoru. To je výhodné z důvodu snadnější detekce odlehlých pozorování v případě jednorozměrných dat. Tato metoda je vhodná pro data s velkým počtem proměnných a malým rozsahem výběru. Druhá metoda je založena na odhadu číselných charakteristik polohy a variability neboli určení vzdálenosti pozorování od středu souboru pozorování vzhledem k jeho varianční struktuře. Stále je nutno mít na paměti, že výpočty s kompozičními daty se neprovádí pomocí standardní euklidovské geometrie, ale v rámci Aitchisonovy geometrie.

Z tohoto důvodu je nedříve nutno zvolit vhodné vyjádření kompozic v logratio souřadnicích, pomocí nichž kompoziční data vyjádříme v reálném prostoru s euklidovskou geometrií a umožníme jejich další statistické zpracování. V úvahu přicházejí tři typy logratio souřadnic. Jsou to alr (aditivní logratio), clr (centrované logratio) a ilr (izometrické logratio) souřadnice. Pro řešení problému detekce odlehlých pozorování jsou nejvhodnější izometrické logratio (ilr) souřadnice, protože s jejich pomocí získáme ideální vlastnosti dat. Takto vzniklá $D-1$ rozměrná data jsou souřadnice vzhledem k ortonormální bázi na simplexu. To se ukazuje jako klíčové pro statistickou analýzu a její interpretaci, jelikož ortonormální souřadnice se řídí standardními pravidly euklidovské geometrie. Mějme $D-$složkovou kompozici $\mathbf{x} = (x_1, \ldots, x_D)^T$ a její ortonormální (ilr) souřadnice

$\mathbf{z} = (z_1, \ldots, z_{D-1})^T$. Konkrétní volbou pak dostaneme $D-1$ rozměrné reálné vektory $\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})^T, l = 1, \ldots, D$, kde jednotlivé souřadnice jsou ve tvaru

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

přitom $(x_1^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})$ jsou takovými permutacemi složek $(x_1, \ldots, x_D)$, že l-tá kompoziční složka je vždy na první pozici $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)$. Odtud následně vidíme, že první ilr proměnná $z_1^{(l)}$ vysvětluje veškerou relativní informaci o složce x_l (resp. příslušné logaritmy podílů s touto složkou) a souřadnice $z_2^{(l)}, \ldots, z_{D-1}^{(l)}$ reprezentují podkompozici $(x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})^T$. Souřadnice $z_1^{(l)}$ pak interpretujeme ve smyslu dominance složky $x_i^{(l)}$ vzhledem k průměru (průměrnému chování) ostatních složek v kompozici.



Obrázek 2: Příklad transformovaných kompozičních dat

Vzhledem ke geometrickým vlastnostem kompozic umíme počítat též se specifiky týkajícími se odlehlých pozorování. Z obrázku je patrné, že nezáleží na tom, v jaké vzdálenosti od počátku pozorování na paprsku leží, protože pozo-

rování ležící na témže paprsku nesou stejnou relativní informaci (podíly mezi jejich složkami se nemění). Místo toho nás bude zajímat, jaký úhel mezi sebou svírají, což je rozhodující pro posouzení toho, zda pozorování je, či není odlehlé (a v našem případě se zřejmě jedná o pozorování označené křížkem).

Po vyjádření kompozic v ilr souřadnicích lze použít metodu detekce odlehlých pozorovnání na základě Mahalanobisovy vzdálenosti. Ta se typicky používá k počítání vzdálenosti pozorování od středu datového souboru vzhledem k jeho varianční struktuře. Mějme např. výběr $\mathbf{z}_1, \ldots, \mathbf{z}_n$ $D-1$ rozměrných pozorování vzniklý vyjádřením kompozic v ilr souřadnicích. Mahalanobisovu vzdálenost spočítáme jako

$$MD\left(\mathbf{z}_i\right) = \left[\left(\mathbf{z}_i - T\right)' C^{-1} \left(\mathbf{z}_i - T\right]\right)^{1/2}, i = 1, \ldots, n,$$

kde $T = T\left(\mathbf{z}_1, \ldots, \mathbf{z}_n\right)$ je odhad polohy (střední hodnoty) a $C = C\left(\mathbf{z}_1, \ldots, \mathbf{z}_n\right)$ jsou odhady variability (varianční matice). Pro detekci odlehlých pozorování je kvalita těchto odhadů zásadní.

Jako odhady těchto charakteristik totiž v praxi není většinou vhodné volit klasický aritmetický průměr a výběrovou varianční matici, protože tyto odhady mohou být samy ovlivněny odlehlými pozorováními. Z tohoto důvodu se pro výpočet odhadů uvedených charakteristik aplikují robustní metody, které vliv těchto odlehlých hodnot eliminují. Často se používá tzv. MCD (Minimum Covariance Determinant) odhad, jenž je právě jedním z nejpopulárnějších robustních odhadů v mnohorozměrné statistice. Při nalezení MCD odhadů postupujeme tak, že hledáme podmnožinu h z celkového počtu n pozorování, jejichž výběrová varianční matice má nejmenší determinant. Robustním odhadem po-

lohy je aritmetický průměr těchto pozorování, robustním odhadem variability pak jejich výběrová varianční matice (vynásobená faktorem pro konzistenci při normálním rozdělení). Kromě vlastnosti robusnosti je důležitou vlastností MCD odhadů T a C tzv. afinní ekvivariance. Tu máme zajištěnu, pokud pro libovolnou regulární matici $\mathbf{A}$ o rozměru $(D-1) \times (D-1)$ a libovolný vektor $\mathbf{b} \in \mathbf{R}^{D-1}$ jsou splněny podmínky:

$$T(\mathbf{A}\mathbf{z}_1 + \mathbf{b}, \ldots, \mathbf{A}\mathbf{z}_n + \mathbf{b}) = \mathbf{A}T(\mathbf{z}_1, \ldots, \mathbf{z}_n) + \mathbf{b},$$

$$C(\mathbf{A}\mathbf{z}_1 + \mathbf{b}, \ldots, \mathbf{A}\mathbf{z}_n + \mathbf{b}) = \mathbf{A}C(\mathbf{z}_1, \ldots, \mathbf{z}_n)\mathbf{A}^T.$$

Vzhledem k existenci ortogonální transformace mezi různými volbami ilr souřadnic tak máme díky afinní ekvivarianci zaručeno, že jsou nalezené odlehlé hodnoty invariantní vůči konkrétní volbě ortonormálních souřadnic. Pokud u daného výběru budeme předpokládat (většinově) mnohorozměrné normální rozdělení pak se druhá mocnina Mahalanobisovy vzdálenosti řídí χ^2 rozdělením o $D-1$ stupních volnosti. Toho se dá využít při hledání odlehlých pozorování, pokud za kritickou hodnotu položíme např. 0.975 kvantil uvedeného rozdělení. Následně všechna pozorování, která tuto kritickou hodnotu překročí, považujeme za odlehlá.

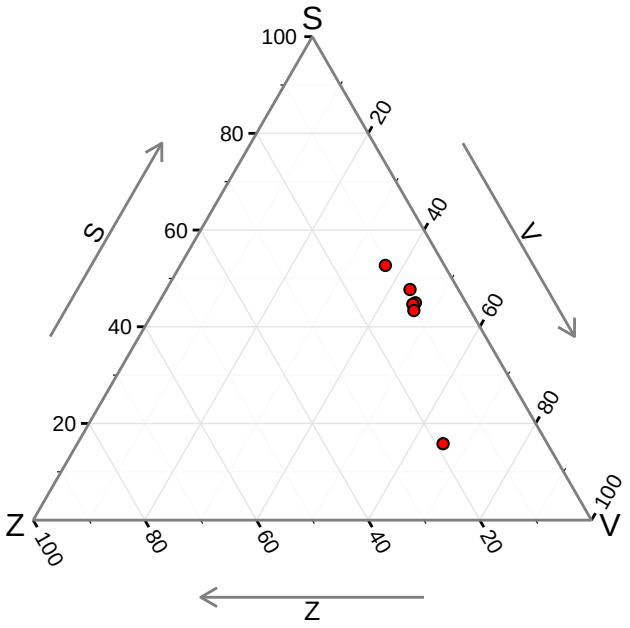
2. Zpracování ve statistickém softwaru

Ke zpracování kompozičních dat a nalezení odlehlých pozorování budeme používat statistický software R. Tento software je volně šiřitelný a uživatelé jej dále doplňují pomocí knihoven, které sami vytvářejí. Kromě toho, že nám tento software umožňuje provést samotnou statistickou analýzu, můžeme s jeho pomocí výsledky interpretovat i graficky. Pro detekci odlehlých hodnot v kompozičních datech budeme používat knihovnu `mvoutlier`, ale ukážeme si i využití knihovny `robCompositions`. Knihovna `mvoutlier` jako vstup potřebuje neupravená kompoziční data minimálně se třemi složkami. Po aplikaci funkce `mvoutlier.CoDa()` na tato data dostaneme výstup v celkem sedmi proměnných (položkách seznamu). První z nich je matice izometrických logratio souřadnic. Dalším výstupem je vektor logických hodnot `TRUE` , nebo `FALSE` podle toho, zda je konkrétní pozorování odlehlé či nikoliv (na základě Mahalanobisovy vzdálenosti, jak bylo popsáno výše). Tento vektor bude pro práci jedním z nejdůležitějších, protože právě na jeho základě budeme výsledná odlehlá pozorování interpretovat. Knihovna `mvoutlier` používá pro výpočty a hledání odlehlých pozorování pomocí Mahalanobisovy vzdálenosti robustní odhady číselných charakteristik. Knihovnou `robCompositions` si budeme demonstrovat použití klasických odhadů a bude nás zajímat, jak a zda se budou nalezená odlehlá pozorování lišit.

2.1. Grafické zobrazení

Při tvorbě této kapitoly bylo použito zejména zdroje [1].

Zjištění, která pozorování jsou odlehlá, nám nedá ucelenou představu o datové struktuře, pokud výsledky nebudeme schopni reprezentovat graficky. Protože kompoziční data reprezentujeme nejčastěji jako data s konstantním součtem, nelze na ně použít ani v tomto případě standardních metod pro zobrazení mnohorozměrných dat. Pokud máme třísložková kompoziční data, jsme schopni je zobrazit v tzv. ternárním diagramu. V něm zobrazujeme podíly mezi třemi složkami jako souřadnice v rovnostranném trojúhelníku. Hodnoty složek se rovnají vzdálenostem odpovídajícího bodu trojúhelníku od strany protilehlé k vrcholu reprezentujícího danou složku.



Obrázek 3: Ternární diagram s několika pozorováními z datového souboru školství

Jednou z nejpoužívanějších metod pro zobrazování kompozičních dat je bi-

plot. Biplot se konstruuje pomocí metody nejmenších komponent (PCA; Principal Component Analysis), která je jednou z nejpopulárnějších metod standardní mnohorozměrné statistiky. Princip metody hlavních komponent spočívá ve snížení dimenze dat a vyjádření co největší variability s pomocí několika hlavních komponent. V biplotu jsou zobrazeny skóry a zátěže prvních dvou hlavních komponent. Skóry jsou zobrazeny jako body a zátěže jako šipky. Díky tomu, že biplot tradičně konstruujeme v clr souřadnicích (vzhledem ke generujícímu systému), daných pro kompozici $\mathbf{x} = (x_1, \ldots, x_D)^T$ jako

$$clr(\mathbf{x}) = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)^T,$$

jež jsou $\sqrt{\frac{D-1}{D}}$ násobky $z_1^{(1)}, \ldots, z_1^{(D)}$, můžeme délky paprsků v biplotu vnímat ve smyslu relativních příspěvků původních jednotlivých kompozičních složek zobrazených hlavních komponent. Skóry jsou tak vlastně souřadnice pro každý objekt datového souboru na osách hlavních komponent. V případě kompozičních dat nás u biplotu nejvíce zajímá vzdálenost mezi paprsky, která aproximuje rozptyl logaritmu podílu příslušné dvojice složek. Čím je tato vzdálenost menší, tím je uvedený podíl stabilnější a vztah mezi složkami (ve smyslu přímé úměrnosti) silnější. Poznamenejme, že přestože jsou pro konstrukci kompozičního biplotu preferované clr souřadnice pro robustní detekci odlehlých hodnot se nehodí, protože MCD odhady vyžadují pozitivně definitní varianční matici.

Grafické nástroje nám poskytuje již zmíněná knihovna `mvoutlier`, s jejíž pomocí jsme schopni vykreslit nejenom biplot, ale také tzv. bodový graf (scatter

plot) či spojnicový graf (parallel plot) pro jednotlivé souřadnice $z_1^{(1)}, \ldots, z_1^{(D)}$. Všechny tyto grafy si ukážeme v průběhu samotného zpracování reálných datových souborů.

2.2. Představení datových souborů

V další části se již budeme zabývat nasbíranými daty. První datový soubor je tvořen strukturou zaměstnanosti v Německu, druhý se zabývá strukturou vzdělání obyvatel v téže zemi. U obou datových souborů se jedná o celkem 37 oblastí či měst. Školství je rozděleno na čtyři podskupiny, a to podle stupně vzdělání. Jmenovitě se jedná o stupeň základní, střední, odborného vyučení a vysokoškolský. V případě zaměstnanosti se jedná o pět podskupin. Jsou to: lesnictví a rybářství, výroba, obchod, finančnictví, a nakonec služby. Analyzovaná data jsou z roku 2011; zdrojem je Spolkový statistický úřad [6]. Bude nás například zajímat, zda uvidíme rozdíly pro oblasti bývalé Německé demokratické republiky a Německé spolkové republiky ve struktuře vzdělání a zaměstnanosti, potažmo která bude generovat více odlehlých hodnot, nebo zda-li se žádný rozdíl neprojeví. Ve struktuře vzdělání můžeme například očekávat vyšší podíl vysokoškolsky vzdělaných občanů v oblastech velkých měst v důsledku vyššího výskytu univerzit a vzdělávacích institucí. U zaměstnanosti pak očekáváme větší podíl obyvatel ve službách a finančnictví v oblastech bývalé Nemecké spolkové republiky. Zda se tato očekávání potvrdí, nebo ne, zjistíme až po analýze, která bude následovat.

2.3. Analýza školství

V následujících dvou kapitolách jsem čerpal zejména z [1], [4].

Nyní přistoupíme k samotné analýze datového souboru školství pomocí softwaru R. Jak už bylo zmíněno, jedná se o 37 oblastí Německa, přičemž každá z nich má čtyři vzdělanostní kategorie (viz příloha A). Kromě samotných výstupů a grafů si zde uvedeme i používané příkazy.

Ze všeho nejdříve si v R-ku nastavíme pracovní adresář příkazem **setwd()** a poté aktivujeme knihovnu na detekci odlehlých pozorování příkazem

```
> library(mvoutlier)
```

Nyní můžeme načíst data z datového souboru `školství.cvs` do matice `y`

```
> y = read.csv2("školství3.csv", header = TRUE)
```

Pomocí parametru `header = TRUE` nastavíme načtení dat s hlavičkou. V tomto bodě si můžeme ukázat několik řádků výsledné datové matice, abychom viděli, jak naše data vypadají v R-ku.

```
> y
```

	Zakladni	Stredni	Odborne	Vysokoskolske
1	5.4	26.8	40.4	27.4
2	5.8	26.8	40.0	27.4
3	5.2	29.5	40.0	25.3
4	5.1	28.1	41.1	25.7
5	5.8	24.8	42.8	26.6
⋮	⋮	⋮	⋮	⋮

Na takto připravená kompoziční data již budeme aplikovat funkci `mvoutlier.CoDa()` a výstup si uložíme do nové datové matice `x`.

```
> x = mvoutlier.CoDa(y)
```

Nyní se podíváme, jak vypadají naše data v izometrických logratio souřadnicích.

```
> x
```

	Zakladni	Stredni	Odborne	Vysokoskolske
[1,]	-1.5122472	0.33758649	0.8115078	0.36315286
[2,]	-1.4474895	0.31983048	0.7822621	0.34539685
[3,]	-1.5467497	0.45750053	0.8090944	0.28015484
[4,]	-1.5618904	0.40863959	0.8477012	0.30554956
[5,]	-1.4360777	0.24168567	0.8717993	0.32259272
⋮	⋮	⋮	⋮	⋮

V dalším kroku zjistíme, jaká pozorování jsou odlehlá, pomocí vektoru logických hodnot `TRUE` nebo `FALSE`. Tento vektor jsme také získali jako výstup funkce `mvoutlier.CoDa()` jako druhou položku seznamu.

1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
10	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
19	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
28	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE
37	FALSE								

Odtud vidíme, že v našem datovém souboru je celkem osm odlehlých pozorování. Jedná se o oblasti Berlín, Brandenburg, Hamburg, Mecklenburg-Vorpommern, Saarland, Chemnitz, Dresden a Leipzig.

Těchto osm odlehlých pozorování bylo detekováno využitím robustních odhadů polohy a varianční matice při počítání Mahalanobisovy vzdálenosti. Nyní se podíváme, jak by náš vektor hodnot `TRUE` nebo `FALSE` vypadal při použití standardních odhadů. K tomuto budeme využívat knihovnu `robCompositions` a funkci `outCoDa`. Nejprve si tuto knihovnu zavoláme příkazem

```
> library(robCompositions).
```

Zjištění, která pozorování jsou odlehlá při použití standardních odhadů místo robustních, provedeme pomocí příkazu

```
> outStandard = outCoDa(y, quantile = 0.975, method = "standard")
```

Výstupem funkce `outCoDa` je

```
"2 out of 37 observations are detected as outliers."
```

Vidíme, že při nepoužití robustních odhadů se nám počet nalezených odlehlých pozorování zmenšil z osmi na dva. Nyní se můžeme podívat na přesné hodnoty Mahalanobisovy vzdálenosti pro každé pozorování.

Ty si vypíšeme použitím příkazu

```
> outStandard$mahalDist

 1   1.0637446  0.9226057  1.2945767  1.2584830  0.7392713  1.7885657
 7   1.3022072  1.0277624  1.5942759  1.2210205  1.9039396  2.2264222
13   2.6641335  0.9898690  2.4401218  0.1817467  0.9342221  1.1885695
19   3.2599770  1.6483279  1.4443019  1.4892381  1.5513049  0.5690782
25   1.9447069  1.1703291  1.3950050  0.1498318  1.6949591  0.6648334
31   0.9468964  3.1642256  2.8529223  2.6725256  2.1818394  1.6935159
37   1.6318092
```

Z tabulky můžeme určit dvě nejvyšší hodnoty Mahalanobisovy vzdálenosti a takto identifikovat odlehlá pozorování (poznamenejme, že v tomto případě je kritickou hodnotou $\sqrt{\chi^2_{3,0.975}} = 3.058$). Jedná se o pozorování číslo 19 a 32. Vektor logických hodnot `TRUE` a `FALSE` si vypíšeme pomocí příkazu

```
> outStandard$outlierIndex

 1  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE
10  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE
19   TRUE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE
28  FALSE  FALSE  FALSE  FALSE   TRUE  FALSE  FALSE  FALSE  FALSE
37  FALSE
```

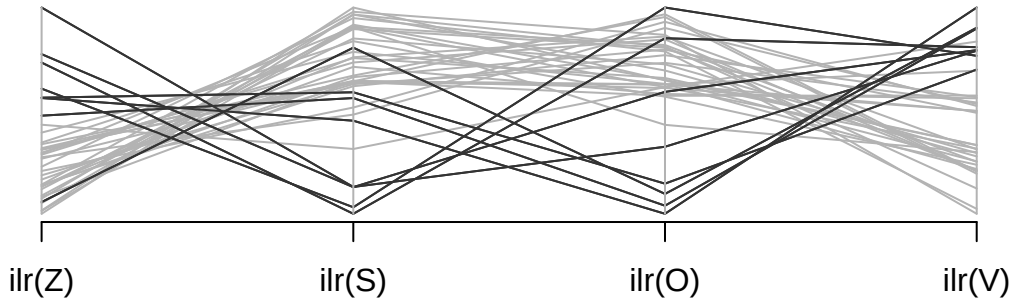
Z tohoto vektoru bychom identifikovali jako odlehlá pozorování 19 a 32, což se shoduje s předchozím určením pomoci Mahalanobisovy vzdálenosti. V původním datovém souboru se jedná o oblasti Mecklenburg-Vorpommern a Saarland. Lze očekávat, že tato pozorování budou extrémní i v rámci odlehlých pozorování, které nalezneme pomocí MCD odhadu. Pro srovnání si zde uvedeme i Mahalanobisovy vzdálenosti při použití robustních metod. Ty získáme příkazem

```
> outRobust$mahalDist

 1   1.1367811  0.9615621  1.4237895  1.3789833  0.5040644  1.5099191
 7   1.1809595  0.9127919  1.6184366  1.2983081  1.4333558  4.0039476
13   4.8293950  0.8127244  4.3687689  0.9955032  0.8647634  1.1331930
19   5.6132265  1.6392514  1.5640287  1.6882259  1.8189548  1.1325073
25   1.7310130  0.9429459  1.1246369  0.9553196  1.6928298  0.7526766
31   1.1185812  4.1842207  3.8012098  4.0524519  3.9623943  2.0279298
37   2.2806777
```

V dalším textu se již budeme zabývat pouze odlehlými pozorováními, které nalezneme pomocí knihovny `mvoutlier` s využitím robustních metod. Pro přehlednost si přejmenujeme naše vzdělanostní proměnné (kompoziční složky) pouze s využitím jejich počátečního písmena. V grafech tedy bude proměnná základní reprezentována písmenem Z atp.

V následujícím kroku zjistíme, jak vypadá celková struktura vzdělanosti ve všech 37 oblastech dohromady. Strukturu budeme zkoumat graficky, a to nejdříve pomocí spojnicového grafu, díky němuž získáme ucelenou představu o datech v ilr souřadnicích, kdy $\text{ilr}(\mathbf{x})$ vrací první souřadnici takového systému, v němž složka X obsadila první pozici ve vstupní kompozici. Postupně tatko obdržíme hodnoty $z_1^1, \ldots, z_1^4$. Spojnicový graf si vykreslíme příkazem

```
> plot(x, which = "parallel", bw = TRUE, onlyout = FALSE, symb =
FALSE, symbtxt = FALSE)
```

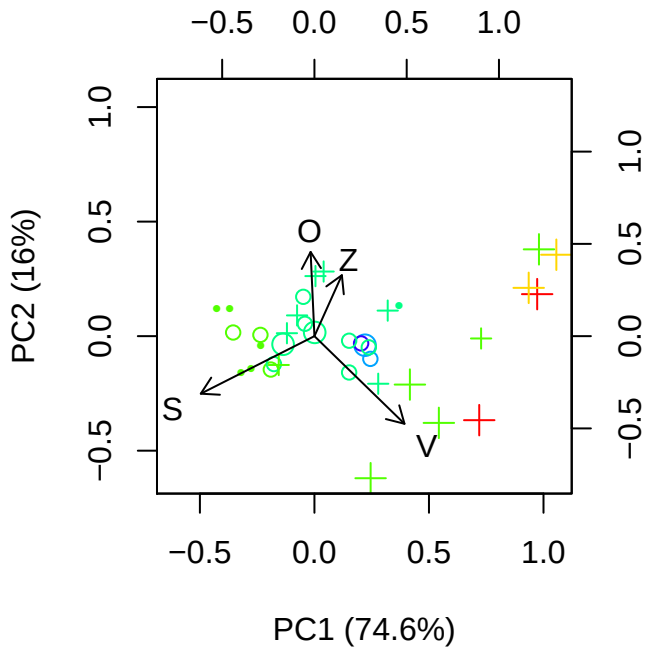


Obrázek 4: Celková struktura vzdělanosti

Nyní máme zobrazenou mnohorozměrnou strukturu kompozičních dat pomocí spojitých čar. Z grafu lze vyčíst, že hlavní část dat má vyšší zastoupení u středního a odborného vzdělání. Také si můžeme povšimnout rozdílu v odstínu, kdy standardní pozorování je zobrazeno světlým odstínem a odlehlé pozorování odstínem tmavým. Lze tak vidět rozdíl mezi mnohorozměrnou strukturou odlehlých pozorování a pozorování standardních. Pro hlubší analýzu se v dalších krocích zaměříme právě na odlehlá pozorování.

Za tímto účelem vytvoříme kompoziční biplot, který se konstruuje pomocí již zmíněné metody hlavních komponent (PCA), v němž jsou použity robustní MCD odhady polohy a variability, ovšem spočítané v ilr souřadnicích a následně vyjádřené v clr proměnných v nichž je PCA aplikována. Pro vytvoření kompozičního biplotu použijeme příkaz

```
> plot(x, which = "biplot", onlyout = FALSE, symb = TRUE, symbtxt
= TRUE)
```



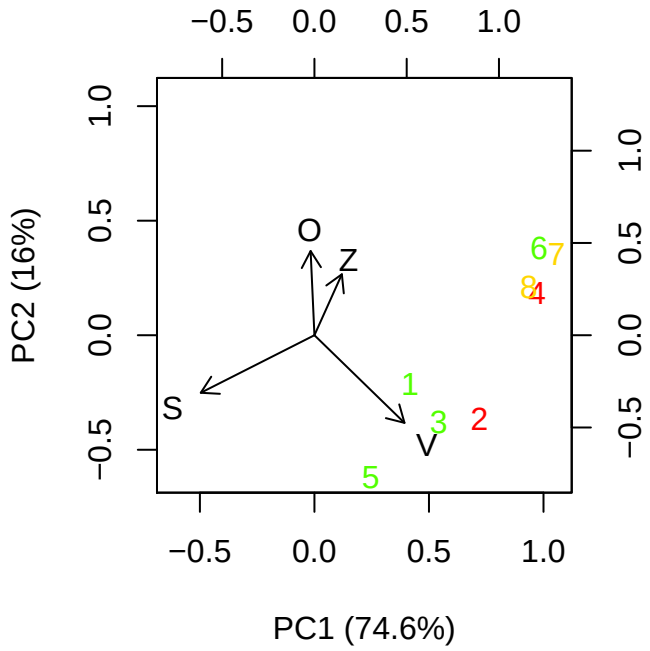
Obrázek 5: Kompoziční biplot pro data školství

V tuto chvíli máme všechna pozorování zobrazena v biplotu, v němž je pomocí prvních dvou hlavních komponent zachyceno téměř 91 % celkové variability v datech. Z biplotu jsme schopni vyčíst nejsilnější vztah mezi základním a odborným vzděláním, vzhledem k nejmenší vzdálenosti mezi odpovídajícími šipkami. Oproti tomu vidíme velmi slabý proporciální vztah mezi středním a vysokoškolským vzděláním. Barva je přiřazena na základě výsledných ilr souřadnic, pro každou z nich je počítána její vzdálenost od příslušného mediánu. Barva je pak přiřazována pomocí mediánu všech těchto vzdáleností, červená (tmavá) pro vysoké hodnoty a modrá (světlá) pro hodnoty nízké.

Pro větší přehlednost si vykreslíme stejný biplot s tím rozdílem, že si v něm

necháme pouze identifikovaná odlehlá pozorování. Toho docílíme úpravou příkazu na

```
> plot(x, which = "biplot", onlyout = TRUE, symb = TRUE, symbtxt
= TRUE)
```



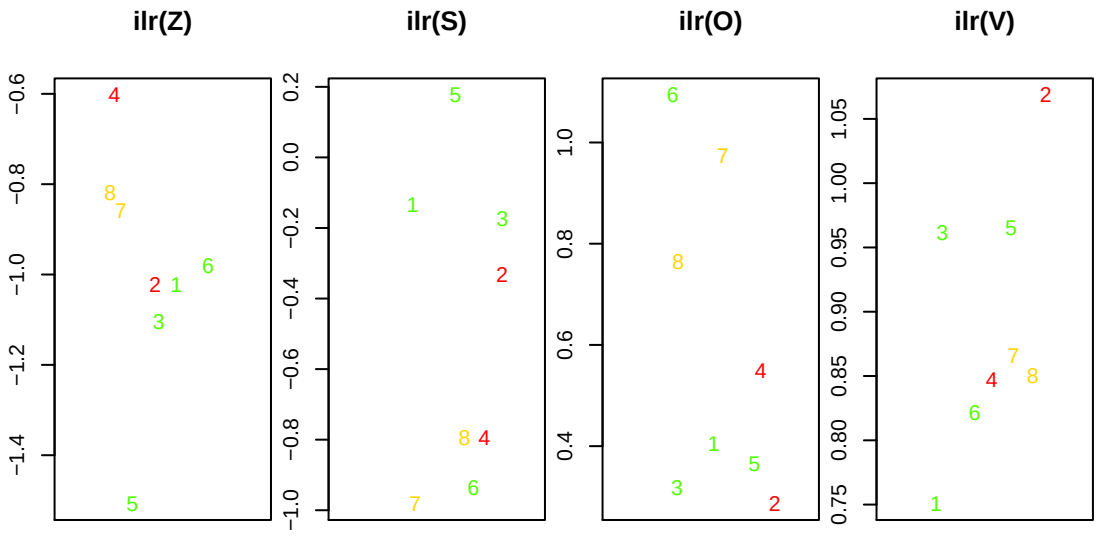
Obrázek 6: Odlehlá pozorování datového souboru školství

Z obrázku je na první pohled vidět výskyt odlehlých pozorování u paprsku reprezentujícího vysokoškolské vzdělání. V původním datovém souboru se jedná o pozorování 12, 13, 15 a 32. Odpovídající oblasti jsou Berlin, Brandenburg, Hamburg a Saarland. Berlin a Hamburg jsou dvě největší města v Německu, dá se tedy očekávat vyšší počet univerzit a vysokých škol a s tím spojený vyšší výskyt vysokoškolsky vzdělaných lidí. V případě oblasti Brandenburg je hodnota u vysokoškolského vzdělání nejvyšší v celém datovém souboru. Tato oblast se geograficky nachází v okolí hlavního města Berlina. Saarland je pak nejmenší

spolkovou zemí, pomineme-li městské spolkové země Berlin, Hamburg a Bremen. Při pohledu na pozorování mající v původním datovém souboru číslo 19 naopak vidíme odlehlost ve směru základního vzdělání. Jedná se oblast Mecklenburg-Vorpommern a v tomto případě je hodnota u základního vzdělání nejvyšší v celém datovém souboru. Jedním z důvodů může být fakt, že se jedná o oblast s nejnižší hustotou zalidnění v celém Německu a také slabší ekonomikou. Toto pozorování bylo detekováno jako jedno ze dvou odlehlých i při použití standardních odhadů pro výpočet Mahalanabisovy vzdálenosti.

Nyní si ukážeme rozložení pozorování v jednorozměrných bodových grafech, vytvořených pro jednotlivé souřadnice $z_1^{(l)}$ odpovídající původním kompozičním složkám. Necháme si rovnou vykreslit pouze odlehlá pozorování, což provedeme příkazem

```
> plot(x, which = "uni", onlyout = TRUE, symb = TRUE, symbtxt = TRUE)
```



Obrázek 7: Odlehlá pozorování pro jednotlivé složky

Povšimněme si pozorování čtyři. V původním datasetu se jedná o číslo 19, které již bylo zmíněno v případě kompozičního biplotu. Opět zde vidíme vysokou hodnotu u základního vzdělání a nižší hodnoty u ostatních typů vzdělání.

Z celkem osmi nalezených odlehlých pozorování se šest nachází na území bývalé Německé demokratické republiky a pouze dvě na území bývalé Německé spolkové republiky. Ve druhém případě se jedná o oblasti Hamburg a Saarland. V původním datovém souboru jsou označené čísly 15 a 32. Většina detekovaných odlehlých pozorování tak byla extrémní v souřadnici odpovídající vysokoškolskému vzdělání, ať už se jedná o města, jako jsou Hamburg či Berlin, nebo oblasti s těmito městy sousedícími, např. oblast Brandenburg.

2.4. Analýza zaměstnanosti

Analýzu týkající se zaměstnanosti budeme provádět obdobně jako v předchozím případě. Jedná se o stejných 37 oblastí jako v minulém příkladu, každá tato oblast je rozdělena na pět zaměstnaneckých kategorií (viz příloha B). Konkrétně se jedná o kategorie lesnictví a rybářství, výroba, obchod, finančnictví a služby. K analýze budeme opět používat statistický software R. V tomto příkladu si již nebudeme vypisovat všechny používané příkazy. V prvním kroku si opět nastavíme pracovní adresář a načteme data do datové matice `y` pomocí příkazu

```
> y = read.csv2("zaměstnanost.csv", header = TRUE)
```

Ukážeme si několik prvních řádků datové matice, abychom měli představu o tom, jak naše data vypadají.

```
> y
```

	Lesnictvni a rybarstvi	Vyroba	Obchod	Financnictvi	Sluzby
1	1,6	33,6	21,7	17,9	25,2
2	1,1	28,4	23,1	18,9	28,5
3	2,5	31,4	24,4	12,3	29,4
4	2,5	33,7	22,2	12,5	29,1
5	2,1	23,2	24,6	21,3	28,8
⋮	⋮	⋮	⋮	⋮	⋮

V tomto bodě aktivujeme knihovnu `mvoutlier` a pomocí funkce `mvoutlier.CoDa()` získáme ortonormální souřadnice, zjistíme, která pozorování jsou odlehlá, atp. Obdržené souřadnice si uvádět nebudeme, nicméně uvedeme vektor logických hodnot `TRUE` nebo `FALSE` a na jeho základě rozhodneme, jaká pozorování jsou odlehlá.

```
1   FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE
10  FALSE  FALSE   TRUE  FALSE   TRUE   TRUE   TRUE  FALSE  FALSE
19  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE
28  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE
37  FALSE
```

Z tohoto výstupu jsou jasně vidět čtyři odlehlá pozorování (12, 14, 15 a 16). V našem původním datovém souboru se jedná o oblasti Berlin, Bremen, Hamburg a Darmstadt. I v tomto příkladu si ukážeme použití knihovny `robCompositions` a zjistíme, zda bychom nalezli odlehlá pozorování využitím standardních odhadů pro výpočet Mahalanobisovy vzdálenosti. Samotné tyto vzdálenosti si opět vypíšeme pomocí

```
> outStandard$mahalDist

 1   2.4077624  1.6662864  1.2547527  1.5080543  2.1477962  1.7836693
 7   1.3227512  1.2448559  1.6612056  1.0699665  1.5882499  4.2731671
13   1.6752055  3.6542198  3.2066997  3.1316164  1.4218284  0.9991804
19   2.5566885  1.4178291  1.0640925  2.5390344  1.6804150  1.5969496
25   1.3817390  0.7055658  1.0648947  1.3656016  1.3556859  2.3796075
31   1.2035049  2.1698295  1.3632780  1.7794139  1.9900709  2.1164318
37   1.5811082
```

Nejvyšší hodnota se nachází u pozorování číslo 12. Nicméně zatím nevíme, kolik odehlých pozorování jsme zatím nalezli. Toto rozhodnutí uděláme pomocí vektoru logických hodnot vytvořeného srovnáním získaných Mahalanobisových vzdáleností s odmocninou kvantilu $\chi^2_{4,0.975}$, tedy hodnotou 3.338. Vypíšeme jej pomocí příkazů

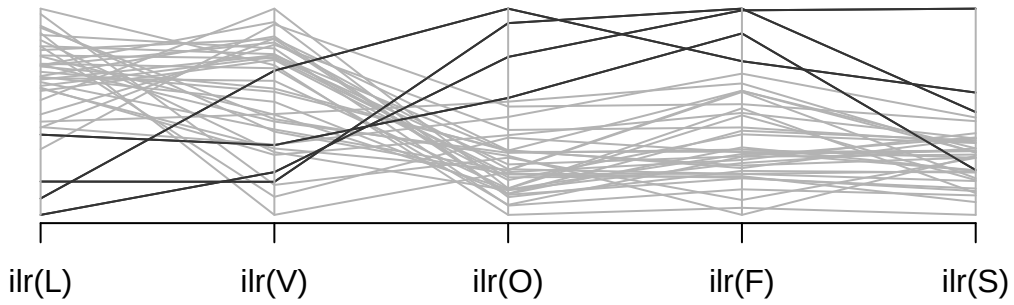
```
> outStandard = outCoDa(y, quantile = 0.975, method = "standard")

> outStandard$outlierIndex
```

Výsledný vektor vypadá takto

```
 1  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE
10  FALSE  FALSE   TRUE  FALSE   TRUE  FALSE  FALSE  FALSE  FALSE
19  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE
28  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE
37  FALSE
```

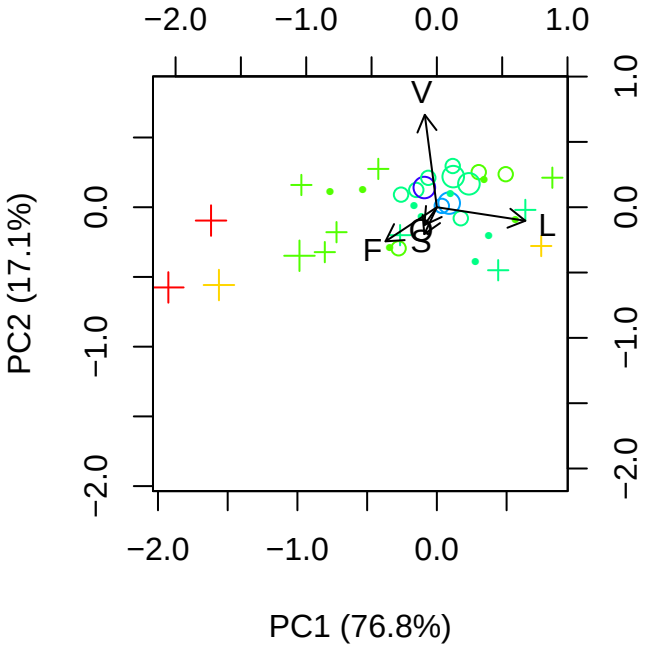
Podařilo se nám tedy takto nalézt pouze dvě odlehlá pozorování, jedná se o oblasti Berlin a Bremen. Při následné hlubší analýze budeme zkoumat, čím jsou tato pozorování extrémní. Nyní se již ovšem vrátíme k použití knihovny `mvoutlier`. Pro přehlednost opět proměnné přejmenujeme pomocí jejich počátečních písmen.

Pomocí příkazu `plot()` si vykreslíme spojnicový graf a zobrazíme tím celkovou strukturu našich pozorování.



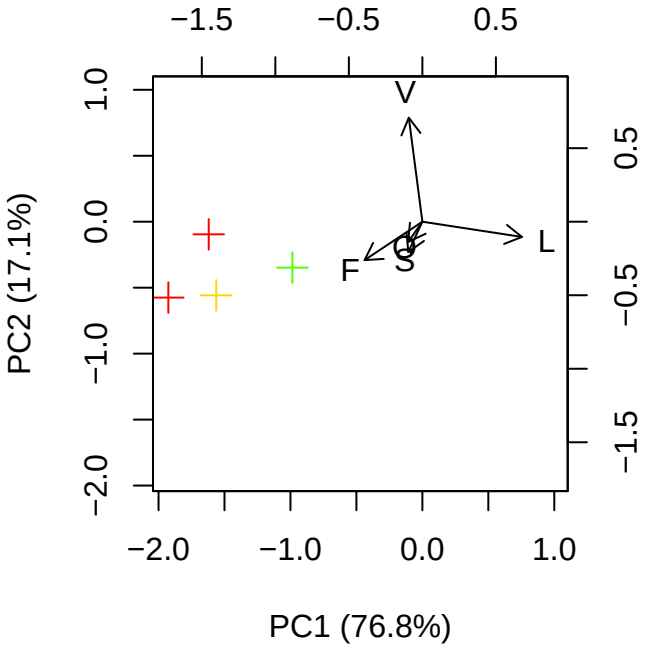
Obrázek 8: Celková struktura zaměstnanosti

Odlehlá pozorování jsou zobrazena tmavě a lze si povšimnout jejich odlišné struktury. Například v souřadnici odpovídající finančnictví se jedná o čtyři nejvyšší hodnoty v celém datovém souboru. Nyní si vykreslíme biplot a budeme zkoumat, zda-li je mezi proměnnými či skupinou proměnných nějaký vztah.



Obrázek 9: Biplot

Z grafu, ve kterém máme vysvětleno téměř 94 % variability v datech, lze vyčíst závislost mezi skupinou proměnných obchod, služby a finančnictví. Vztah mezi nimi je tak těsný, že se paprsky v našem grafu téměř překrývají. Také si lze povšimnout skupinky odlehlých pozorování v oblasti finančnictví a jednoho pozorování lišícího se vyšší hodnotou u lesnictví. Úpravou parametru `onlyout = TRUE` si zobrazíme pouze odlehlá pozorování, která jsou včechna ve směru finančnictví. Jak už bylo zmíněno výše, jedná se o oblasti Berlin, Bremen, Hamburg a Darmstadt. Ve všech případech se jedná o města, nikoliv větší územní celky. Dá se usuzovat, že v těchto městech sídlí velké finanční společnosti, jako jsou banky, pojišťovny atp. Můžeme si také povšimout, že nám zmizelo pozorování, které se mohlo zdát odlehlé ve směru lesnictví. V původním datovém souboru se jedná o oblast Luneburg. Hodnota tohoto pozorování je v proměnné lesnictví



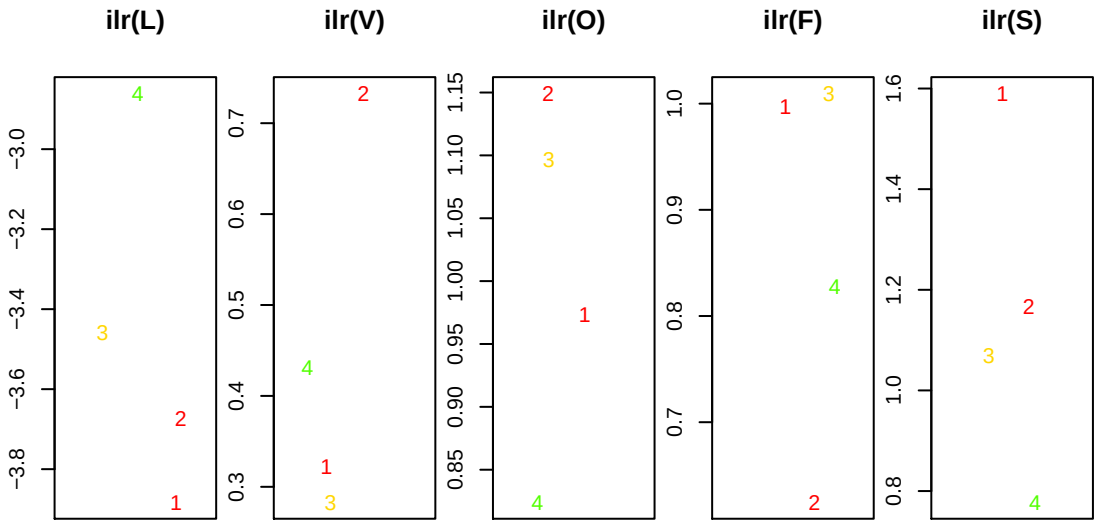
Obrázek 10: Odlehlá pozorování

druhá nejvyšší v celém datovém souboru, nicméně svou celkovou strukturou se neliší natolik, abychom jej detekovali jako odlehlé.

V posledním kroku si zobrazíme odlehlá pozorování v bodovém grafu jednotlivých souřadnic a detailněji se podíváme na jejich strukturu v rámci kompozic.

Povšimněme si zde pozorování číslo čtyři, které reprezentuje město Darmstadt. To se liší od ostatních vyšší hodnotou u proměnné lesnictví, nicméně toto platí, pouze pokud uvažujeme naše nalezená odlehlá pozorování. Pokud bychom uvažovali celý datový soubor, zjistili bychom, že se určitě v této proměnné nejedná o extrém. To, co dělá pozorování odlehlým, je opět celková mnohorozměrná struktura kompozic.

V našem datovém souboru o relativní struktuře zaměstnanosti se nachází celkem čtyři odlehlá pozorování, z nichž tři jsou na území bývalé Německé spol-



kové republiky. Tímto se nám alespoň částečně potvrdilo naše očekávání z úvodu praktické části o vyšším zastoupení obyvatel ve finančnictví na území bývalého kapitalistického „západního Německa".

Závěr

V této práci jsem se nejprve věnoval nezbytným teoretickým základem pro pochopení rozdílu mezi standardními a kompozičními daty, včetně uvedení některých jejich specifických vlastností. Představil jsem zde izomerické logratio souřadnice, které jsem s pomocí statistického softwaru R následně využil pro detekci odlehlých pozorování pomocí Mahalanobisovy vzdálenosti. Na tomto místě bych zdůraznil důležitost použití robustních odhadů při jejich výpočtu, jelikož i Mahalanobisovy vzdálenosti samotné mohou být při použití klasických odhadů (výběrový průměr a výběrová varianční matice) odlehlými pozorováními ovlivněny.

V praktické části jsem již pracoval s knihovnou `mvoutlier` statistického softwaru R a různými typy grafů, které poskytuje pro analýzu kompozičních dat. Výsledky jsem se snažil vhodně a objektivně interpretovat. Pomocí provedené analýzy jsem byl schopen potvrdit některá očekávání, jež jsem měl před jejím samotným začátkem.

V průběhu psaní jsem nabyl nové znalosti nejen o problematice kompozičních dat a odlehlých pozorování, ale také o práci se statistickým softwarem R a typografickým systémem LaTeX. Právě práce s těmito programy byla pro mne asi největší výzva. Doufám, že tato práce bude nápomocna všem, kdo se zajímají o problematiku kompozičních dat či odlehlých pozorování.

Literatura

[1] Filzmoser, P., Hron, K., Reimann, C., *Interpretation of multivariate outliers for compositional data*, Computer & Geosciences 39, 77-85, (2012)

[2] Filzmoser, P., Hron, K., *Outlier detection for compositional data using robust methods*, Mathematical Geosciences 40, 233-248, (2008)

[3] Pawlowsky-Glahn V., Egozcue, J.J., Tolosana-Delgado, R., *Lecture Notes on Compositional Data Analysis*, [online], dostupné z: http://dugi-doc.udg.edu//handle/10256/297

[4] Pawlowsky-Glahn V., Buccianti A., *Compositional data analysis: Theory and applications*, 1. vydání. Chichester: Wiley, 2011

[5] Pawlowsky-Glahn V., Egoycue, J.J., *Compositional data and their analysis: an introduction*, Geological Society, Londýn, 1-10, Special publication 2006

[6] *Ausgewählte Regionaldaten für Deutschland*, Statistische Ämter des Bundes und der Länder, 2011, Hannover 2012

Příloha A

Data týkající se struktury školství na území Německa. Data pocházejí z [6].

	Základní	Střední	Odborná	Vysoká
Stuttgart	5,4	26,8	40,4	27,4
Karlsruhe	5,8	26,8	40,0	27,4
Freiburg	5,2	29,5	40,0	25,3
Tubingen	5,1	28,1	41,1	25,7
Oberbayern	5,8	24,8	42,8	26,6
Niederbayern	5,7	28,4	46,1	19,8
Oberpfalz	5,4	28	44,4	22,2
Oberfranken	5,5	25,8	44,5	24,2
Mittelfranken	7,3	25,7	40,4	26,6
Unterfranekn	5,1	25,8	44,3	24,8
Schwaben	6,3	28,1	44,8	20,8
Berlin	8,9	19,2	30,6	41,3
Brandenburg	8,3	15,1	25,8	50,8
Bremen	6,0	18,6	42,5	32,9
Hamburg	7,9	17,7	27,1	47,3
Darmstadt	6,9	19,7	40,5	32,9
Geiben	5,7	21,7	46,1	26,5
Kassel	6,7	21,6	46,9	24,8
Macklenburg-Vorpommern	12,4	10,5	33,6	43,5
Braunschweig	5,9	15,6	47,2	31,3
Hannover	6,3	15,1	46,7	31,9
Lüneburg	6,7	19,6	49,7	24 ,0
Weser-Ems	6,8	18,9	50,1	24,2
Düsseldorf	6,6	19,4	39,1	34,9
Köln	5,4	18,6	38,8	37,2
Münster	5,7	19,2	44,0	31,1
Detmold	5,7	18,1	43,0	33,2
Arnsberg	6,9	19,4	40,9	32,8
Koblenz	7,6	24,1	42,2	26,1
Trier	6,8	22,7	42,1	28,4
Rheinhessen-Pfaltz	6,0	21	38,9	34,1
Saarland	5,3	22,8	26,8	45,1
Chemnitz	7,8	8,1	47,0	37,1
Dresden	8,9	8,0	43,5	39,6
Leipzig	9,8	10	38,6	41,6
Schleswig - Holstein	7,0	28,0	36,8	28,2
Thuringen	7,2	12,3	38,3	42,2

Příloha B

Data týkající se zaměstnanosti na území Německa. Data pocházejí z [6].

	Lesnictví	Výroba	Obchod	Finančnictví	Služby
Stuttgart	1,6	33,6	21,7	17,9	25,2
Karlsruhe	1,1	28,4	23,1	18,9	28,5
Freiburg	2,5	31,4	24,4	12,3	29,4
Tubingen	2,5	33,7	22,2	12,5	29,1
Oberbayern	2,1	23,2	24,6	21,3	28,8
Niederbayern	5,4	32,6	23,7	11,1	27,2
Oberpfalz	3,8	32,9	23,2	11,9	28,2
Oberfranken	3,1	33,0	23,3	12,4	28,2
Mittelfranken	2,2	27,7	23,2	18,3	28,6
Unterfranekn	2,9	30,6	24,8	12,7	29,0
Schwaben	3,3	31,8	25,1	12,8	27,0
Berlin	0,3	12,9	23,1	23,6	40,1
Brandenburg	3,6	22,7	24,5	14,2	35
Bremen	0,4	20,6	29,9	18,7	30,4
Hamburg	0,5	14,2	29,4	27,2	28,7
Darmstadt	1,0	19,0	27,0	27,1	25,9
Geiben	1,9	28,5	23,7	13,2	32,7
Kassel	2,5	26,8	26,1	13,1	31,5
Macklenburg-Vorpommern	3,9	18,1	26	14,5	37,5
Braunschweig	1,8	28,3	22,6	15,3	32
Hannover	1,9	20,4	26,6	17,9	33,2
Lüneburg	5,0	20,8	28,1	11,5	34,6
Weser-Ems	4,4	25,4	26,8	13,4	30,0
Düsseldorf	1,2	22,0	27,6	20,2	29,0
Köln	1,1	19,2	25,6	21,0	33,1
Münster	2,6	24,2	26,4	14,8	32
Detmold	2,1	29,5	25,4	14,0	29,0
Arnsberg	1,3	28,2	25,1	15,1	30,3
Koblenz	2,3	25,9	25,6	12,6	33,6
Trier	4,1	25,3	27,1	10,0	33,5
Rheinhessen-Pfaltz	2,9	25,2	23,8	14,6	33,5
Saarland	0,8	27,8	24,8	16,1	30,5
Chemnitz	2,2	31,8	22,3	14,5	29,2
Dresden	2,2	25,3	22,5	16,3	33,7
Leipzig	1,9	21,1	24,0	20,9	32,1
Schleswig-Holstein	3,3	19,0	28,6	15,0	34,1
Thuringen	2,6	29,1	22,1	14,0	32,2