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

DIPLOMOVÁ PRÁCE

Kompoziční biplot

Vedoucí diplomové práce:
RNDr. Karel Hron, Ph.D.
Rok odevzdání: 2012

Vypracovala:
Bc. Alžběta Kalivodová
AME, II. ročník

Prohlášení

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

V Olomouci dne 22. března 2012

Poděkování

Ráda bych na tomto místě poděkovala vedoucímu diplomové práce RNDr. Karlu Hronovi, Ph.D. za obětavou spolupráci i za čas, který mi věnoval při konzultacích. Dále bych ráda poděkovala všem svým blízkým, že se mnou měli trpělivost, a také svému počítači.

Obsah

Úvod

Tématem této práce je konstrukce a aplikace kompozičního biplotu. Výběr námětu byl pro mě mnohem jednodušší než dříve, při výběru tématu bakalářského. Důvodem je hlavně skutečnost, že zde navazuji na již zmíněnou práci bakalářskou. V té jsem uvedla základy problematiky týkající se biplotu - teorie, ze kterých vychází (singulární rozklad matice, metoda hlavních komponent), principy konstrukce a v neposlední řadě také praktické příklady.

Úkolem této práce je ukázat širší využití biplotu. Ten se totiž často užívá při statistické analýze speciálního typu dat, tzv. kompozičních dat, nesoucích pouze relativní informaci.

V první kapitole se seznámíme s kompozičními daty a jejich základními charakteristikami. Představíme si geometrii užívanou pro tato data a také tři základní transformace, které lze na kompozice aplikovat. Dále naváže kratší kapitola se shrnutím metody hlavních komponent, která je i v případě kompozičních dat důležitá pro vytvoření biplotu. V úvodu následující kapitoly si zopakujeme, jak vypadá standardní biplot. Následně si ukážeme biplot kompoziční, který, jak zjistíme, lze rozlišit na tzv. kovarianční a vytvořený, logratio a ilr biplot. V závěru této kapitoly bude shrnuta grafická interpretace. V poslední teoretické kapitole se budeme zabývat případem odlehlých hodnot. Představíme si robustní metodu užívanou pro kompoziční data. Závěr celé práce bude tvořit praktická část - uvedeme si zde dva příklady s reálnými daty.

Důvodů, proč jsem se rozhodla v tématu biplotu pokračovat, je hned několik. Samotnou mě toto téma „chytlo za srdce“. Oblast statistiky mi byla vždy blízká a kombinace kompozičních dat a biplotu mi přijde jako vhodná k probádání. Také jsem se přesvědčila, že je tato problematika opravdu zajímavá i pro člověka, který o tomto slyší poprvé. Při několika prezentacích mé bakalářské práce jsem mohla sledovat, jak přítomní zaujatě poslouchají výklad, zejména pak popisy grafických výstupů praktických příkladů. Co mě ještě více potěšilo, byl fakt, že jevili zájem o tuto tématiku a kladli mi dotazy. To mě utvrdilo v myšlence, že si kompoziční biplot zaslouží, aby se o něm dozvědělo co nejvíce lidí.

1 Kompoziční data

1.1 Základní pojmy

V této kapitole si přestavíme kompoziční data a jejich základní charakteristiky. S tímto typem dat budeme pracovat v průběhu celé práce. Při tvorbě kapitoly byly informace čerpány zejména z [1], [2], [15], [16], [19].

Mějme dán vícerozměrný statistický soubor, jehož složky představují kvantitativně vyjádřené části nějakého celku. Tato data potom označujeme jako kompoziční (nebo též zkráceně kompozice). Data mohou mít podobu procent (pak je součet jejich složek roven 100) nebo jsou vyjádřena jako podíly na celku velikosti k (nejčastěji je přitom k rovno právě 100 nebo 1, v daném případě se jedná o proporce). Častými příklady souborů kompozičních dat jsou koncentrace chemických sloučenin v horninách, procentuální zastoupení několika živočišných či rostlinných druhů v určité lokalitě nebo výdaje jednotlivce na různé potřeby do domácnosti. Ve všech těchto případech nás nezajímají přímo absolutní hodnoty složek, ale již zmíněné podíly na celku.

První komplexní přístup ke zpracování kompozičních dat zavedl v osmdesátých letech minulého století skotský statistik John Aitchison. Své příznivce nalezla jeho teorie zejména v oblasti geologie.

Nyní si zavedeme definice pojmů D-složková kompozice, simplex, uzávěr a podkompozice, které jsou pro další teorii velmi důležité.

Definice 1.1.1. *Sloupcový vektor* $\mathbf{x} = (x_1, x_2, \ldots, x_D)^T$ *se nazývá* D*-složková kompozice, jestliže jsou všechny jeho prvky kladná reálná čísla nesoucí pouze relativní informaci.*

Definice 1.1.2. *Výběrový prostor kompozičních dat je* D*-složkový simplex o dimenzi* $D - 1$*,*

$$S^D = \left\{ \mathbf{x} = (x_1, x_2, \ldots, x_D)^T | x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = k \right\}, \tag{1.1}$$

kde k je libovolná pevně zvolená konstanta.

Definice 1.1.3. *Pro libovolný vektor o D složkách, které obsahují kladná reálná čísla $\mathbf{x} = (x_1, x_2, \ldots, x_D)^T \in \mathbb{R}_+^D, \quad x_i > 0 \quad pro \quad i = 1, 2, \ldots, D$, je uzávěr definován takto*

$$C(\mathbf{x}) = \left(\frac{k \cdot x_1}{\sum_{i=1}^{D} x_i}, \frac{k \cdot x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{k \cdot x_D}{\sum_{i=1}^{D} x_i} \right)^T. \tag{1.2}$$

Výsledkem uzávěru je vlastně reprezentace kompozice jako kladného vektoru se součtem složek k. Volba k přitom závisí na jednotkách měřítka. Nejčastějšími hodnotami jsou $k = 1, 100, 10^6, 10^9$.

Definice 1.1.4. *Mějme dánu kompozici $\mathbf{x}$. Její podkompozici $\mathbf{x}_l$ o l prvcích ($l < D$) získáme použitím operace uzávěru na podvektor $(x_{i_1}, x_{i_2}, \ldots, x_{i_l})^T$ vektoru $\mathbf{x}$. Indexy $i_1, i_2, \ldots, i_l$ nám určují, které složky z $\mathbf{x}$ byly vybrány.*

Soubory kompozičních dat často obsahují data s velkým množstvím složek. Proto se v literatuře (zejména v oblasti geologie) objevuje kvůli snazší interpretaci omezení na trojrozměrné (pod)kompozice. V tomto případě je pak simplexem rozuměn rovnostranný trojúhelník, jehož vrcholy jsou v bodech $A = [k, 0, 0,]$, $B = [0, k, 0]$, $C = [0, 0, k]$. Data se ovšem častěji zobrazují do tzv. ternárního diagramu. Je to rovinný diagram tvořený opět rovnostranným trojúhelníkem, ve kterém pro kompozici $\mathbf{p} = (p_a, p_b, p_c)^T$ platí, že p_a (resp. p_b, p_c) označuje vzdálenost příslušné strany od protějšího vrcholu A (resp. B,C).

Podkompozice mohou být vyjádřeny také jiným způsobem. Mějme dánu matici $\mathbf{X}$ kompozičních dat o rozměru $n \times D$. Její typický řádek má tvar vektoru

$(x_1, x_2, \ldots, x_D)^T$, kde $x_i,\ i = 1, \ldots, D$ jsou kladná čísla s konstantním součtem $x_1 + x_2 + \ldots + x_D = k$. Podkompozice o $l < D$ složkách (bez újmy na obecnosti zvolíme prvních l složek) má tvar

$$C(\mathbf{s}) = \left(\frac{x_1}{\sum_{i=1}^{l} x_i}, \ldots, \frac{x_l}{\sum_{i=1}^{l} x_i} \right)^T \qquad pro \quad \mathbf{s} = (s_1, \ldots, s_l)^T. \tag{1.3}$$

Z tohoto vztahu také vyplývá, že podíly jsou invariantní vůči tvorbě podkompozic: $s_j/s_{j'} = x_j/x_{j'}, j, j' = 1, \ldots, l$. Toto jsou podíly složek uvnitř kompozičního datového souboru. Pokud je používáme k počítání skalárního součinu a kovariancí pro biplot (viz dále) a porovnáváme data mezi sebou, musíme brát v úvahu, z jaké škály data pocházejí. Tato informace byla převzata z [19]. Proto je vhodné vypočítat logaritmy těchto podílů (logratio) a uvažovat rozdíly mezi nimi.

Jak již bylo řečeno v úvodu, kompoziční data nesou pouze relativní informaci. Díky tomu můžeme vyvozovat vlastnost, že změna měřítka informaci, obsaženou v kompozici, nezmění. Tedy platí následující definice.

Definice 1.1.5. *Mějme dány dva D-složkové vektory kladných reálných složek $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D$ $(x_i, y_i > 0$ pro $i = 1, 2, \ldots, D)$, kde $\mathbf{y} = \lambda \cdot \mathbf{x}$ pro $\lambda \in \mathbb{R}^+$. Pak $\mathbf{x}$ a $\mathbf{y}$ jsou kompozičně ekvivalentní. Platí také $C(\mathbf{x}) = C(\mathbf{y})$.*

Soubor kompozičních dat je dále invariantní vůči permutaci. Tedy pokud změníme pořadí složek vektoru, podíly mezi nimi se nemění.

Poslední vlastností kompozičních dat je podkompoziční soudružnost. Ta nám říká, že podíly mezi složkami v podkompozici jsou vždy stejné jako podíly v rámci celé kompozice a obdobně by se měl chovat i libovolný rozumný přístup ke statistické analýze kompozičního datového souboru. Uveďme si příklad, který je převzat z [15]. Představme si dva vědce A a B, kteří zkoumají stejný vzorek horniny. Vědec A si vyjádří podíly všech prvků - organických i anorganických. Vědec B bere v úvahu pouze prvky organické (má tedy podkompozici) a podíly počítá pouze v rámci nich. Výsledky obou vědců týkající se organických prvků by

měly být stejné bez ohledu na to, že chemik B používá redukovaný soubor. Tedy z celého datového souboru (tj. uvažujeme-li původní kompozici) bychom měli obdržet stejné informace o prvcích nějaké podkompozice jako při analýze pouze této podkompozice. Ovšem standardní statistické metody, jako např. analýza hlavních komponent založená na kovariancích původních složek kompozičních dat, tuto vlastnost nemají.

1.2 Aitchisonova geometrie

Informace k této kapitole jsou čerpány zejména z [19].

Práce s kompozičními daty je mnohem složitější, než jak jsme zvyklí z klasického datového souboru využívajícího principy euklidovské geometrie. V něm umíme sčítat vektory či je násobit skalárem, počítat vzdálenosti mezi vektory nebo zjišťovat jejich vzájemnou ortogonalitu. Odpovídající geometrie pro kompoziční data se chová analogicky, její interpretace je ovšem poněkud složitější. Říkáme jí souhrnným názvem Aitchisonova geometrie.

Ukažme si následující příklad. Rozdíl mezi kompozicemi $(5, 80, 15)^T$ a $(10, 75, 15)^T$ není zdaleka stejný jako mezi kompozicemi $(50, 35, 15)^T$ a $(55, 30, 15)^T$. Vidíme, že rozdíl mezi hodnotami první a druhé složky je v obou případech pět, tedy je stejná i euklidovská vzdálenost $\left(5\sqrt{2}\right)$. Ovšem když se podíváme na vzájemné podíly složek, rozdíl je podstatný. V prvním případě jde u první složky o nárůst z 5 na 10, tedy 100 % nárůst. V druhém případě máme ale nárůst pouze 10 % (z 50 na 55).

Tato skutečnost ovšem není jediným důvodem, proč nepoužíváme pro statistickou analýzu kompozic euklidovskou geometrii. Můžeme například také narazit na problém, kdy výsledek statistické analýzy vyjde mimo výběrový prostor nebo není splněna podmínka podkompoziční soudružnosti.

Aitchisonova geometrie je tvořena celou řadou operací. První dvě, perturbace a mocninná transformace, tvoří vektorový prostor na simplexu a vyžadují operaci uzávěru, která je zavedena v Definici 1.1.3. Perturbace je analogií pro sčítání

v reálném prostoru. Mocninná transformace nahrazuje násobení skalárem. Následně přidáme do tohoto vektorového prostoru skalární součin, normu a vzdálenost mezi kompozicemi, abychom obdrželi euklidovský vektorový prostor.

Definice 1.2.1. *Perturbace kompozice* $\mathbf{x} \in S^D$ *kompozicí* $\mathbf{y} \in S^D$ *je kompozice*

$$\mathbf{x} \oplus \mathbf{y} = C(x_1 y_1, x_2 y_2, \ldots, x_D y_D)^T. \tag{1.4}$$

Definice 1.2.2. *Mocninná transformace kompozice* $\mathbf{x} \in S^D$ *konstantou* $\alpha \in \mathbb{R}$ *je dána jako*

$$\alpha \odot \mathbf{x} = C(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha)^T. \tag{1.5}$$

Nyní máme nadefinován vektorový prostor, který je tvořen simplexem $(S^D, \oplus, \odot)$ a má následující vlastnosti:

Vlastnost 1.2.1. $(S^D, \oplus)$ *je komutativní grupa, tedy pro* $\mathbf{x}, \mathbf{y}, \mathbf{z} \in S^D$ *platí*

- *Komutativita:* $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$;
- *Asociativita:* $(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} = \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})$;
- *Existence neutrálního prvku:* $\mathbf{n} = C(1, 1, \ldots, 1)^T = \left(\frac{1}{D}, \frac{1}{D}, \ldots, \frac{1}{D}\right)^T$;
- *Existence inverzního prvku:* $\mathbf{x}^{-1} = C\left(x_1^{-1}, x_2^{-1}, \ldots, x_D^{-1}\right)^T$, *pro který platí* $\mathbf{x} \oplus \mathbf{x}^{-1} = \mathbf{n}$.

  Jako analogii se standardními operacemi v reálném prostoru budeme používat $\mathbf{x} \oplus \mathbf{y}^{-1} = \mathbf{x} \ominus \mathbf{y}$.

Všimněme si, že operace uzávěru ruší veškeré konstanty, a proto není z matematického hlediska sama o sobě důležitá. Tento poznatek nám umožňuje vynechat operaci uzávěru jako mezikrok při složitějších výpočtech. Tato skutečnost má také

důležité důsledky pro praktické použití, například při analýze hlavních komponent, kterou si ukážeme dále. Uvedenou vlastnost můžeme vyjádřit pro $\mathbf{z} \in \mathbb{R}_+^D$ a $\mathbf{x} \in S^D$ takto

$$\mathbf{x} \oplus (\alpha \odot \mathbf{z}) = \mathbf{x} \oplus (\alpha \odot C(\mathbf{z})). \tag{1.6}$$

Vlastnost 1.2.2. *V případě mocninné transformace pro* $\mathbf{x}, \mathbf{y} \in S^D$ *a* $\alpha, \beta \in \mathbb{R}$ *platí*

- *Asociativita:* $(\alpha \cdot \beta) \odot \mathbf{x} = \alpha \odot (\beta \odot \mathbf{x})$;
- *Distributivita:* $\alpha \odot (\mathbf{x} \oplus \mathbf{y}) = (\alpha \odot \mathbf{x}) \oplus (\alpha \odot \mathbf{y})$, $(\alpha + \beta) \odot \mathbf{x} = (\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{x})$;
- *Existence neutrálního prvku:* $1 \odot \mathbf{x} = \mathbf{x}$.

Nyní si nadefinujeme výše zmíněný skalární součin, normu a vzdálenost mezi kompozicemi.

Definice 1.2.3. *Aitchisonův skalární součin kompozic* $\mathbf{x}, \mathbf{y} \in S^D$ *je definován jako*

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}. \tag{1.7}$$

Z vlastností logaritmů podílu může být Aitchisonův skalární součin zapisován také zjednodušeně,

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}. \tag{1.8}$$

Definice 1.2.4. *Aitchisonova norma kompozice* $\mathbf{x} \in S^D$ *je definována jako*

$$||\mathbf{x}||_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} \right)^2} = \langle \mathbf{x}, \mathbf{x} \rangle_a. \tag{1.9}$$

Definice 1.2.5. *Aitchisonova vzdálenost mezi* $\mathbf{x}$ *a* $\mathbf{y} \in S^D$ *je dána jako*

$$d_a(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} \ominus \mathbf{y}||_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}. \tag{1.10}$$

1.3 Souřadnicový systém

Při zpracování tohoto tématu bylo čerpáno zejména z [5], [10], [11], [16], [19].

Jak již bylo řečeno v předchozí kapitole, ve standardním reálném euklidovském prostoru nelze s kompozičními daty díky jejich odlišné geometrii pracovat. Proto byly sestrojeny tzv. logratio transformace ze simplexu do reálného prostoru. John Aitchison využil faktu, že velikost složek je pro kompoziční data nepodstatná a v osmdesátých letech minulého století zavedl dva druhy těchto transformací - aditivní logratio (alr) a centrovanou logratio (clr) transformaci. První z nich (alr) nezachovává vzdálenosti mezi daty, druhá (clr) tuto vlastnost (izometrii) sice neztrácí, ale vede k singulární varianční matici. Tato omezení a také fakt, že obě transformace není možné ztotožnit s ortogonálním souřadnicovým systémem na simplexu, vedli k myšlence definovat třetí transformaci [5]. Ta se jmenuje izometrická (ilr) logratio transformace a je izometrií mezi S^D a $\mathbb{R}^{D-1}$.

1.3.1 Clr transformace

Kompozice jsou v simplexu S^D formulovány pomocí kanonické báze $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_D\}$ v $\mathbb{R}^D$. Tedy každý vektor může být vyjádřen jako

$$\mathbf{x} = x_1(1, 0, \ldots, 0)^T + x_2(0, 1, \ldots, 0)^T + \cdots + x_D(0, 0, \ldots, 1)^T = \sum_{i=1}^{D} x_i \cdot \mathbf{e}_i. \tag{1.11}$$

V tomto vyjádření ovšem nastává problém. Kanonická báze zmíněná výše není generujícím systémem a dokonce ani bází vzhledem k Aitchisonově geometrii

na simplexu S^D. Proto musíme nějakou takovou bázi vytvořit. Využijeme k tomu generující systém

$$\mathbf{w}_i = C(\exp(\mathbf{e}_i))^T = C(1, 1, \ldots, e, \ldots, 1)^T,\ i = 1, 2, \ldots, D. \tag{1.12}$$

Kompozici $\mathbf{x} \in S^D$ můžeme vyjádřit následujícím způsobem

$$\mathbf{x} = \bigoplus_{i=1}^{D} \ln \frac{x_i}{g(\mathbf{x})} \odot \mathbf{w}_i = \ln \frac{x_1}{g(\mathbf{x})} \odot (e, 1, \ldots, 1)^T \oplus \ldots \oplus \ln \frac{x_D}{g(\mathbf{x})} \odot (1, 1, \ldots, e)^T, \tag{1.13}$$

kde $g(\mathbf{x}) = \sqrt[D]{\prod_{i=1}^{D} x_i} = \exp\left(\frac{1}{D}\sum_{i=1}^{D} \ln x_i\right)$ značí geometrický průměr složek kompozice $\mathbf{x}$. Do jmenovatele přitom můžeme obecně dosadit libovolnou konstantu. Pravá strana této rovnosti označuje centrovanou logratio (clr) transformaci. Tuto si můžeme vyjádřit také po složkách jako

$$\mathrm{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \ln \frac{x_2}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})}\right)^T = \boldsymbol{\xi}. \tag{1.14}$$

Inverze, která nám dává koeficienty kanonické báze reálného prostoru, pak vypadá následovně

$$\mathrm{clr}^{-1}(\boldsymbol{\xi}) = C(\exp(\xi_1), \exp(\xi_2), \ldots, \exp(\xi_D))^T = \mathbf{x}. \tag{1.15}$$

Jak již bylo řečeno dříve, varianční matice $\boldsymbol{\xi}$ je singulární, tedy její determinant je nulový. Také součet složek v clr transformaci je nula.

Formální definice clr transformace je následující.

Definice 1.3.1. *Mějme dánu kompozici $\mathbf{x} \in S^D$. Clr koeficienty jsou složky vektoru $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_D)^T = \mathrm{clr}(\mathbf{x})$ jediného, který splňuje*

$$\mathbf{x} = \mathrm{clr}^{-1}(\boldsymbol{\xi}) = C(\exp(\boldsymbol{\xi}))^T,\ \sum_{i=1}^{D} \xi_i = 0; \tag{1.16}$$

i-tý clr koeficient má tvar

$$\xi_i = \ln \frac{x_i}{g(\mathbf{x})}. \tag{1.17}$$

Dále bychom měli věnovat pozornost převodu operací a metriky ze simplexu do reálného prostoru. Než vyjádříme větu, která tyto vlastnosti vyslovuje, musíme si označit euklidovskou vzdálenost, normu a skalární součin v $\mathbb{R}^D$ pomocí $d(\cdot,\cdot)_e$, $||\cdot||_e$ a $\langle\cdot,\cdot\rangle_e$.

Věta 1.3.1. *Mějme dány vektory $\mathbf{x}_1, \mathbf{x}_2 \in S^D$ a reálné konstanty α, β, pak platí*

$$\begin{gathered} \operatorname{clr}(\alpha \odot \mathbf{x}_1 \oplus \beta \odot \mathbf{x}_2) = \alpha \cdot \operatorname{clr}(\mathbf{x}_1) + \beta \cdot \operatorname{clr}(\mathbf{x}_2); \\ \langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = \langle \operatorname{clr}(\mathbf{x}_1), \operatorname{clr}(\mathbf{x}_2) \rangle_e; \\ ||\mathbf{x}_1||_a = ||\operatorname{clr}(\mathbf{x}_1)||_e \; ; \quad d_a(\mathbf{x}_1, \mathbf{x}_2) = d(\operatorname{clr}(\mathbf{x}_1), \operatorname{clr}(\mathbf{x}_2))_e. \end{gathered} \tag{1.18}$$

1.3.2 Ilr transformace

Transformace do reálných souřadnic zachovávající všechny metrické vlastnosti kompozic se nazývá izometrická logratio (ilr) transformace.

Mějme dánu ortonormální bázi $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ na simplexu S^D vzhledem k Aitchisonově geometrii. Ilr transformace je dána tak, že platí $\operatorname{ilr}(\mathbf{e}_i) = \vec{e}_i$, kde $\vec{e}_i$ značí i-tý vektor kanonické báze v $\mathbb{R}^{D-1}$. Přesná definice je následující.

Definice 1.3.2. *Mějme dánu libovolnou kompozici $\mathbf{x} \in S^D$. Izometrická logratio (ilr) transformace spojená s ortonormální bází v S^D, $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$, je transformace z S^D do $\mathbb{R}^{D-1}$ a má tvar*

$$\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_{D-1}^*)^T = \operatorname{ilr}(\mathbf{x}) = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \langle \mathbf{x}, \mathbf{e}_2 \rangle_a, \ldots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a)^T. \tag{1.19}$$

Reálný vektor $\operatorname{ilr}(\mathbf{x}) \in \mathbb{R}^{D-1}$ vyjadřuje souřadnice kompozice $\mathbf{x} \in S^D$ vzhledem k bázi $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$. Ekvivalentně jej můžeme ztotožnit se souřadnicemi vzhledem ke kanonické bázi v $\mathbb{R}^{D-1}$, což vede k vyjádření $\operatorname{ilr}(\mathbf{x}) = \sum_i \langle \mathbf{x}, \mathbf{e}_i \rangle_a \vec{e}_i$.

Potom ze vztahu (1.19) a z ortonormality báze na S^D získáme požadovanou vlastnost $\mathrm{ilr}(\mathbf{e}_i) = \vec{e}_i$ pro $i = 1, 2, \ldots, D-1$. Inverzní ilr transformace odpovídá vyjádření $\mathbf{x}$ v odpovídající bázi S^D,

$$\mathbf{x} = \mathrm{ilr}^{-1}(\mathbf{x}^*) = \bigoplus_{i=1}^{D-1} (\langle \mathbf{x}^*, \vec{e}_i \rangle_e \odot \mathbf{e}_i), \tag{1.20}$$

kde $\langle \mathbf{x}^*, \vec{e}_i \rangle_e = \langle \mathbf{x}, \mathbf{e}_i \rangle_a = x_i^*$.

Každou kompozici $\mathbf{x} \in S^D$ lze tedy zapsat také jako

$$\mathbf{x} = \langle \mathbf{x}, \mathbf{e}_1 \rangle_a \odot \mathbf{e}_1 \oplus \ldots \oplus \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a \odot \mathbf{e}_{D-1}. \tag{1.21}$$

Již ze samotné definice ilr transformace je zřejmé, že se jedná o izomorfismus mezi Aitchisonovou geometrií na simplexu a euklidovskou geometrií v reálném prostoru. Pro ilr transformace tak platí analogické vlastnosti, jako je uvedeno ve Větě 1.3.1 pro clr transformaci. Jinak řečeno, aplikací ilr transformace je Aitchisonova geometrie nahrazena euklidovskou. Pro reálná čísla α, β můžeme psát

$$\mathrm{ilr}(\alpha \odot \mathbf{x} \oplus \beta \odot \mathbf{y}) = \alpha \mathrm{ilr}(\mathbf{x}) + \beta \mathrm{ilr}(\mathbf{y}) = \alpha \mathbf{x}^* + \beta \mathbf{y}^*, \quad \langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \mathbf{x}^*, \mathbf{y}^* \rangle_e, \tag{1.22}$$

kde $\langle ., . \rangle_e$ značí (stejně jako dříve) euklidovský skalární součin.

Ilr transformace jsou preferovány zejména v případech, kde je interpretace výsledku statistické analýzy zaměřena spíše na objekty (pozorování) než na jednotlivé složky kompozice.

Pro interpretaci souřadnic je důležité, jakým způsobem vybereme ortonormální bázi na S^D. Nejčastěji k tomuto účelu volíme proces zvaný postupné binární dělení (PBD), který funguje následujícím způsobem. Nejprve rozdělíme složky kompozic do dvou skupin označených $+1$ a -1. Toto nám vymezí první souřadnici vyjadřující rovnováhu mezi složkami $+1$ a -1 a vlastně i podíly mezi jednotlivými

skupinami. V následující krocích rozdělujeme jednu z těchto dvou skupin na další dvě atd. Poté samozřejmě rozdělíme i druhou část. Nezahrnuté prvky v daném kroku dělení označujeme 0. Počet kroků potřebných k dosažení rozkladu celé kompozice je $D - 1$.

Označme si r_i počet $+1$ a s_i počet -1 v souboru v i-tém kroku. Nové souřadnice budou mít tvar:

$$x_i^* = \sqrt{\frac{r_i s_i}{r_i + s_i}} \ln \frac{\left(\prod_+ x_j\right)^{1/r_i}}{\left(\prod_- x_k\right)^{1/s_i}}, \quad i = 1, 2, \ldots, D-1. \tag{1.23}$$

Celý proces je většinou zapisován do tabulky. Například pro pětisložkovou kompozici ($D = 5$) může vypadat následovně (x_i jsou označeny jednotlivé složky kompozice):

souřadnice	krok	x_1	x_2	x_3	x_4	x_5	r	s
x_1^*	1	$+1$	-1	-1	-1	-1	1	4
x_2^*	2	0	$+1$	$+1$	-1	-1	2	2
x_3^*	3	0	$+1$	-1	0	0	1	1
x_4^*	4	0	0	0	$+1$	-1	1	1

Nově vzniklé souřadnice tvoříme dle vzorce (1.23):

x_1^*	x_2^*	x_3^*	x_4^*
$\sqrt{\frac{4}{5}} \ln \frac{x_1}{(x_2 x_3 x_4 x_5)^{1/4}}$	$\ln \frac{(x_2 x_3)^{1/2}}{(x_4 x_5)^{1/2}}$	$\frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3}$	$\frac{1}{\sqrt{2}} \ln \frac{x_4}{x_5}$

Podrobnější popis celého procesu bude vysvětlen u praktického příkladu.

1.3.3 Vztah clr a ilr transformací

V této krátké kapitole si uvedeme několik poznatků ohledně vztahu mezi clr a ilr transformacemi.

Vztah (1.14) můžeme maticově zapsat také jako

$$\boldsymbol{\xi} = \mathbf{F} \ln(\mathbf{x}), \tag{1.24}$$

kde $\mathbf{F} = \mathbf{I}_D - \frac{1}{D}\mathbf{J}_D$, $\mathbf{I}_D$ je jednotková matice řádu D a $\mathbf{J}_D$ je čtvercová matice jedniček řádu D.

Clr transformace umožňuje interpretovat jednotlivé koeficienty jako proměnné vysvětlující všechnu relativní informaci o příslušné složce kompozice, pohybující se v čitateli daného koeficientu. Proto je v popisu výstupů zvolené statistické analýzy možné použít jména pro jednotlivé proměnné odpovídající názvům původních složek kompozice. Tato transformace je také využívána ke konstrukci biplotu, jak se zmíníme později. Nevýhodou této transformace je, že výsledné proměnné jsou kolineární, protože $\sum_{i=1}^{D} \xi_i = 0$. Nemůžeme tedy clr transformované kompozice zpracovávat metodami, které závisí na plné hodnosti datové matice.

Tento problém řeší ilr transformace. Zvolme ortonormální bázi na nadrovině $H : \xi_1 + \ldots + \xi_D = 0$ v $\mathbb{R}^D$, která je generována clr transformovanými daty, tedy například

$$\mathbf{v}_i = \sqrt{\frac{i}{i+1}}\left(\underbrace{\frac{1}{i}, \ldots, \frac{1}{i}}_{\#i}, -1, 0, \ldots, 0\right)^T, \; i = 1, 2, \ldots, D-1. \tag{1.25}$$

Poznamenejme přitom, že příslušné ilr souřadnice jsou v tomto konkrétním případě ve tvaru

$$x_i^* = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}}, \; i = 1, 2, \ldots, D-1. \tag{1.26}$$

Pak vztah mezi clr i ilr transformacemi je maticově vyjádřen jako

$$\boldsymbol{\xi} = \mathbf{V}\mathbf{x}^*, \tag{1.27}$$

kde $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{D-1})$ je matice rozměru $D \times (D-1)$ s ortonormálními vektory ze vztahu (1.25) na nadrovině H. Platí, že $\mathbf{V}^T\mathbf{V} = \mathbf{I}_{D-1}$.

1.3.4 Alr transformace

Nyní se znovu podívejme na vztah pro vyjádření clr transformace (1.14). Řekli jsme si, že do jmenovatelů v tomto vztahu můžeme dosadit libovolnou hodnotu. Pokud tak učiníme a použijeme jednu ze složek $x_1, \ldots, x_D$, pak dostaneme koeficient odpovídající dané složce roven nule. Proto můžeme z původního generujícího systému vyjmout vektor, kterým je při volbě x_D kompozice $\mathbf{w}_D$ (obecně nemusí jít nutně o poslední vektor). Potom dostaneme bázi $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{D-1}\}$. V takovém případě můžeme libovolnou kompozici $\mathbf{x} \in S^D$ psát jako

$$\mathbf{x} = \bigoplus_{i=1}^{D-1} \ln \frac{x_i}{x_D} \odot \mathbf{w}_i = \ln \frac{x_1}{x_D} \odot (e, 1, \ldots, 1, 1)^T \oplus \ldots \oplus \ln \frac{x_{D-1}}{x_D} \odot (1, 1, \ldots, e, 1)^T. \tag{1.28}$$

Takovouto úpravou dostáváme aditivní logratio (alr) transformaci, která má následující tvar

$$\mathrm{alr}(\mathbf{x}) = \left(\ln \frac{x_1}{x_D}, \ln \frac{x_2}{x_D}, \ldots, \frac{x_{D-1}}{x_D} \right)^T = \mathbf{y}. \tag{1.29}$$

Alr transformace má ovšem dva nedostatky. Prvním z nich je nesymetrie (jedna ze složek, x_D, hraje ve vztahu (1.29) zcela jinou roli než složky ostatní). Mnohem větší potíže ale nastávají díky tomu, že tato transformace není izometrická.

1.4 Popisná statistika kompozic

Podklady k tvorbě této kapitoly tvořily zejména [16], [19].

V úvodu je třeba říci, že v celé této kapitole budeme pracovat s datovou maticí kompozic $\mathbf{X}$ o n řádcích a D sloupcích se složkami x_{ik}. Než ovšem začneme datový soubor analyzovat, musíme jej zkontrolovat.

Nejprve zjistíme, zda v datech nejsou chyby, jestli mají logický význam a není mezi nimi žádné pozorování, které by svými hodnotami značně vybočovalo vzhled-

em k ostatním. Dále se podíváme na odlehlé hodnoty - kolik jich je, jak jsou vzdálené od ostatních dat v souboru. Problematice odlehlých hodnot se budeme podrobně věnovat ve čtvrté kapitole o robustních metodách. Nakonec zkontrolujeme, zda soubor neobsahuje nuly.

Standardní popisná statistika využívaná pro data s euklidovskou metrikou není pro kompozice příliš vhodná. Například rozptyl nebo směrodatná odchylka nedává v Aitchisonově geometrii smysl. Proto byly zavedeny číselné charakteristiky centrum, variační matice a celkový rozptyl.

Definice 1.4.1. *Charakteristika polohy pro kompoziční data se nazývá centrum. Pro datový soubor o rozsahu n je definováno jako*

$$\mathbf{g} = C(g_1, g_2, \ldots, g_D)^T, \quad kde \quad g_k = \left(\prod_{i=1}^{n} x_{ik} \right)^{\frac{1}{n}}, \ k = 1, 2, \ldots, D. \tag{1.30}$$

Definice 1.4.2. *Rozptyl souboru tvořeného kompozičními daty je vyjádřen pomocí variační matice, matice rozptylů logaritmů podílů (resp. ortonormálních souřadnic) jednotlivých podkompozic* (x_{ik}, x_{jk}), $i, j = 1, 2, \ldots, D; k = 1, 2, \ldots, n$,

$$\mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & \ldots & t_{1D} \\ t_{21} & t_{22} & \ldots & t_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ t_{D1} & t_{D2} & \ldots & t_{DD} \end{pmatrix}, \quad kde\ t_{ij}\ je\ rozptyl\ souboru \left\{ \ln \frac{x_{ik}}{x_{jk}}, k = 1, \ldots, n \right\}, \tag{1.31}$$

resp. pomocí normované variační matice

$$\mathbf{T}^* = \begin{pmatrix} t^*_{11} & t^*_{12} & \ldots & t^*_{1D} \\ t^*_{21} & t^*_{22} & \ldots & t^*_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ t^*_{D1} & t^*_{D2} & \ldots & t^*_{DD} \end{pmatrix}, \quad kde\ t^*_{ij}\ je\ rozptyl\ souboru \left\{ \frac{1}{\sqrt{2}} \ln \frac{x_{ik}}{x_{jk}}, k = 1, \ldots, n \right\}. \tag{1.32}$$

Pro t_{ij} a t^*_{ij} platí: $t^*_{ij} = \mathrm{var}\left(\frac{1}{\sqrt{2}} \ln \frac{x_{ik}}{x_{jk}}\right) = \frac{1}{2} t_{ij}$, tedy $\mathbf{T}^* = \frac{1}{2}\mathbf{T}$. Tyto matice jsou symetrické a jejich hlavní diagonály jsou tvořeny nulami.

Definice 1.4.3. *Měřítkem celkové variability kompozičního datového souboru je celkový rozptyl,*

$$\mathrm{totvar}(\mathbf{X}) = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} t_{ij} = \frac{1}{D} \sum_{i=1}^{D} \sum_{j=1}^{D} t^*_{ij}. \tag{1.33}$$

Variační matice nám říká, jak je celkový rozptyl rozdělen mezi jednotlivé složky kompozice. Prvky matice $\mathbf{T}^*$ jsou interpretovány následujícím způsobem: čím blíž je hodnota t^*_{ij} k nule, tím jsou hodnoty podílů mezi i-tou a j-tou složkou kompozic stabilnější. Navíc celková variance nezávisí na konstantě k (součtu složek kompozice), tudíž případné přeškálování dat nemá na tuto hodnotu žádný vliv.

V praxi se využívá centrování a škálování dat. Centrování není nic jiného, než aplikace operace perturbace. Používáme jej následujícím způsobem. Nejprve sestrojíme centrum $\mathbf{g}$ našeho výběru. Pak na data aplikujeme perturbaci inverzním prvkem $\mathbf{g}^{-1}$. Toto nám umožňuje posunout centrum kompozic do těžiště simplexu. Pokud chceme datový soubor $\mathbf{X}$ přeškálovat, umocníme jej hodnotou $1/\sqrt{\mathrm{totvar}(\mathbf{X})}$. Tímto způsobem získáme kompozice s jednotkovým celkovým rozptylem.

1.5 Reprezentace kompozičních datových souborů

V úvodu této kapitoly si krátce představíme singulární rozklad matice, který je důležitý pro správné pochopení dalších postupů. Informace o něm byly čerpány z [17].

Mějme dánu reálnou matici $\mathbf{X}$ o rozměrech $n \times D$, zapisujeme $\mathbf{X}_{n,D}$, a dále $\mathbf{U}_{n,n}$, $\mathbf{D}_{n,D}$ a $\mathbf{V}_{D,D}$. Matici $\mathbf{X}$ lze rozložit na součin

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{1.34}$$

kde **U** a **V** jsou ortogonální matice. Sloupce matice **U** se nazývají skóry (scores), odpovídající sloupce matice **V** se nazývají zátěže (loadings). Dále je zde matice **D** nezáporných (tzv. singulárních) hodnot nacházejících se na hlavní diagonále. Jsou uspořádány sestupně. Tedy

$$d_{11} \geq d_{22} \geq \cdots \geq d_{kk} \geq 0 \quad \text{kde} \quad k = \min(n, D); \tag{1.35}$$

přitom prvky mimo hlavní diagonálu jsou rovny nule. Singulární hodnoty mají vypovídající funkci o vztahu **U** a **V**. Čím blíže k nule je číslo $d_{ii}, i = 1, 2, \ldots, k$ tím mají odpovídající skóry a zátěže v celkovém rozkladu menší vliv.

A nyní se můžeme vrátit k reprezentaci kompozičních datových souborů. Hlavní zdroj informací tvořil článek [2]. V něm jsou shrnuty tři ekvivalentní způsoby, jak vyjádřit podíly mezi složkami kompozičních dat:

1. $\frac{1}{2}D(D-1)$ podílů $\frac{x_j}{x_{j'}}$ mezi páry složek, při výběru páru předpokládáme $j < j', j = 1, 2, \ldots, D-1, j' = 1, 2, \ldots, D$;

2. $D-1$ podílů $\frac{x_j}{x_D}$ mezi prvními $D-1$ složkami a poslední složkou (vede k alr transformaci);

3. D podílů $\frac{x_j}{g(\mathbf{x})}$ mezi složkami a jejich geometrickým průměrem

   $g(\mathbf{x}) = (x_1 \cdot x_2 \cdots x_D)^{1/D}$ (vede k clr transformaci).

Na logaritmické škále tak uvažujeme rozdíly $\ln(x_j) - \ln(x_{j'}), \ln(x_j) - \ln(x_D)$ a odchylku od průměru $\ln(x_j) - 1/D \sum_j \ln(x_j)$. Druhá z možností není symetrická vzhledem ke všem složkám, a proto se jí již dále nebudeme zabývat. Nejvíce nás bude zajímat vztah mezi logaritmy podílů (logratios) $\ln(x_j/x_{j'}) = \ln(x_j) - \ln(x_{j'})$.

Označme logaritmus $\ln(x_{ik}), i = 1, 2, \ldots, n,\ k = 1, 2, \ldots, D$, prvků kompoziční datové matice pomocí j_{ik}, které tvoří matici **J** o rozměru $n \times D$. Dále zavedeme symboly $j_{i.}, j_{.k}$ a $j_{..}$ pro průměry přes odpovídající indexy. Nakonec

definujme $\pi_{i,kk'} = j_{ik} - j_{ik'}$ jako prvky matice $\mathbf{E}$ rozměru $n \times \frac{1}{2}D(D-1)$, která je tvořena pro každou z kompozic ve výběru všemi logaritmy podílů, kde $k < k'$. Budeme se snažit ukázat, že je možné získat všechny potřebné informace vedoucí k matici $\mathbf{E}$ při použití menší matice tvořené pouze D sloupci založenými na clr proměnných (viz bod 3. výše).

Matici $\mathbf{E}$, můžeme centrovat vzhledem ke sloupcovým průměrům $\pi_{.,kk'} = j_{.k} - j_{.k'}$. Takto získáme matici $\mathbf{Y}$ o prvcích $y_{i,kk'} = \pi_{i,kk'} - \pi_{.,kk'}$. Singulární rozklad matice $\mathbf{Y}$ má tvar $\mathbf{Y} = \mathbf{A\Psi B}$, kde $\mathbf{B}$ má $\frac{1}{2}D(D-1)$ řádků reprezentujících každý z logaritmů podílů (kk'). Matice $\mathbf{E}$ má $\frac{1}{2}D(D-1)$ sloupců, ale její hodnost je rovna číslu $D-1$. Tedy zde máme $\frac{1}{2}(D-1)(D-2)$ nadbytečných sloupců.

Nyní si zavedeme matici $\mathbf{Z}$ o rozměrech $n \times D$, která je tvořena clr transformovanými kompozicemi $j_{ik} - j_{i.}$, centrovanými vzhledem ke sloupcovým průměrům $z_{.k} = j_{ik} - j_{..}$. Tedy $\mathbf{Z}$ je matice dvakrát centrovaných prvků matice $\mathbf{J}$: $z_{ik} = j_{ik} - j_{i.} - j_{.k} + j_{..}$. $\mathbf{Z}$ má singulární rozklad tvaru $\mathbf{Z} = \mathbf{UDV}^T$. Singulární vektory v maticích $\mathbf{U}$ a $\mathbf{V}$ jsou centrované a matice $\mathbf{Z}$ má hodnost rovnu $D-1$.

Singulární rozklady matic $\mathbf{Y}$ a $\mathbf{Z}$ spolu souvisí následujícím způsobem:

1. Singulární hodnoty mají vztah: $\mathbf{\Psi} = \sqrt{D}\mathbf{D}$;
2. Levé singulární vektory jsou shodné: $\mathbf{A} = \mathbf{U}$;
3. Prvky matice $\mathbf{B}$ souvisí s maticí $\mathbf{V}$ následovně: $b_{kk',l} = \frac{(v_{kl} - v_{k'l})}{\sqrt{D}}$.

Předchozí odstavce nám tedy říkají, že k analýze datové matice $\mathbf{Y}$ (centrovaných) logaritmů podílů nám stačí znát menší matici $\mathbf{Z}$. Toho využijeme později při konstrukci biplotu.

2 Metoda hlavních komponent

V následující kapitole se seznámíme s metodou hlavních komponent. Ta je klíčová pro pochopení a sestavení biplotu. Informace jsou čerpány z [4], [11], [12], [19]. Detailně je metoda hlavních komponent popsána v bakalářské práci [17].

Analýza hlavních komponent je jednou z nejdůležitějších statistických metod. Jejím cílem je zredukovat dimenzi mnohorozměrných dat tak, aby se stala jednoduchými a dobře čitelnými, ale abychom touto redukcí ztratili co nejméně informace. Vychází z vlastností singulárního rozkladu matice uvedeného v kapitole 1.5 (podrobně viz [17]).

Mějme dánu datovou matici $\mathbf{X}$ o rozměrech $n \times D$, kde ve většině případů n řádků představuje pozorování a D sloupců proměnné (v našem případě složky kompozice). Obvykle, v závislosti na typu dat, je $\mathbf{X}$ transformována na matici $\mathbf{Z}$. Nejobvyklejšími transformacemi jsou centrování proměnných nebo jejich normování, v případě kompozičních dat je vhodná logratio transformace. Níže popíšeme nejprve standardní přístup.

Následující text je zkrácený a převzatý z [17].

Nechť je dán výběrový průměr $\bar{\mathbf{X}}$ a výběrová varianční matice $\mathbf{S}$. Jejich definice zde nebudeme uvádět, pro zájemce jsou k dispozici v [17] (str. 10, Definice 2.4 a 2.5).

Výběrové hlavní komponenty, resp. jednotlivé výběrové komponenty - sloupce matice $\mathbf{W}_{n,D}$, počítáme ze vztahů

$$\mathbf{W} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{X}}^T)\mathbf{G}, \tag{2.1}$$

resp.

$$\mathbf{W}_j = (\mathbf{X} - \mathbf{1}\bar{\mathbf{X}}^T)\mathbf{g}_j, \quad j = 1, \ldots, D. \tag{2.2}$$

Přitom $\mathbf{1}$ je vektor n jedniček a $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_D)$ je matice, jejíž sloupce jsou

jednotlivé vlastní vektory matice $\mathbf{S}$.

Dále zde platí

$$\mathbf{S} = \mathbf{G}\mathbf{D}\mathbf{G}^T, \qquad \mathbf{G}^T\mathbf{S}\mathbf{G} = \mathbf{D}, \tag{2.3}$$

kde $\mathbf{D} = \mathrm{Diag}(g_1, \ldots, g_D)$ je diagonální matice vlastních čísel (seřazených sestupně) matice $\mathbf{S}$.

Poznamenejme přitom, že matice $\mathbf{W}$ má v tomto případě stejné rozměry jako datová matice $\mathbf{X}$ a její prvky $W_{ij}, i = 1, \ldots, n;\ j = 1, \ldots, D$, můžeme označit jako skóry. Sloupce matice $\mathbf{G}$ se nazývají zátěže j-té hlavní komponenty, $j = 1, 2, \ldots, D$, a vyjadřují vliv původních proměnných na nové hlavní komponenty.

Libovolné dvě hlavní komponenty jsou nekorelované, protože jsou příslušné vlastní vektory $\mathbf{g}_j, j = 1, 2, \ldots, D$, ortonormální. Takto lze tedy vytvořit D komponent. Z hlediska zmenšování dimenze dat (která je naším hlavním cílem) je ale lepší mít komponent méně.

Poznamenejme přitom, že metoda hlavních komponent je velmi citlivá na měřítko dat. To znamená, že pokud jsou hodnoty jedné proměnné výrazně větší než hodnoty ostatních (například v jiných jednotkách), bude tato proměnná vzhledem k celkovému rozptylu dominovat též při konstrukci hlavních komponent. Právě především z tohoto důvodu data normujeme (a tedy mimo nulových sloupcových průměrů budou rozptyly sloupců datové matice rovny jedné). Tento postup ovšem není možné použít vždy, jak uvidíme v kapitole 3.2 o konstrukci kompozičního biplotu.

3 Konstrukce biplotu

V této kapitole se již začneme věnovat hlavnímu tématu této práce - biplotu. Nejdříve si v krátkosti shrneme jeho klasickou podobu, potom se podíváme, k jakým změnám dochází v případě kompozičního biplotu, logratio a ilr biplotu. Nakonec si shrneme grafickou interpretaci. Hlavním zdrojem pro tvorbu této kapitoly byly články a knihy [2], [3], [4], [14], [17], [19].

Konstrukci biplotu popsal roku 1971 matematik K. R. Gabriel [13]. Biplot se dá zjednodušeně vysvětlit jako dvojdimenzionální zobrazení objektů a proměnných v jednom grafu. V roce 2002 přizpůsobil J. Aitchison biplot pro kompoziční data [2] a dokázal, že je to užitečný nástroj jak ke zkoumání datové struktury, tak i k vysvětlování vztahů mezi proměnnými.

Než přejdeme k dalším kapitolám, uvedeme si v krátkosti rozdíly v názvosloví jednotlivých biplotů. Nejprve můžeme rozdělit biploty na standardní a kompoziční. První z nich jsou tvořeny standardními datovými soubory, druhé jsou konstruovány pro kompoziční data. Dále můžeme dělit biploty na klasické a robustní. Klasický biplot je konstruován pro homogenní datový soubor, respektive soubor bez odlehlých hodnot. Robustní biplot použijeme v případě, kdy datový soubor (standardní nebo kompoziční) obsahuje odlehlé hodnoty či jiné „špatné“ hodnoty. Tento biplot je totiž vůči nim invariantní. Všechny dále uváděné biploty jsou klasické, o robustních se zmíníme v kapitole 4.

3.1 Standardní biplot

Text v následující kapitole je citován z [17]. Další zdroje byly [2], [14].

Mějme dánu datovou matici $\mathbf{X}$ o n řádcích a D sloupcích, kde u každého z n objektů bylo provedeno D měření. Přesněji řečeno, matice $\mathbf{X}$ obsahuje statistický soubor o rozsahu n odpovídající D statistickým znakům $X_1, \ldots, X_D$.

Pro konstrukci biplotu je důležitý singulární rozklad matice $\mathbf{X}$ pomocí matic $\mathbf{U}$, $\mathbf{D}$ a $\mathbf{V}$ (viz (1.34)). Princip konstrukce je založen na nahrazení matice $\mathbf{X}$ pomocí její aproximace $\mathbf{X}_{(2)}$ s hodností rovnou dvěma. Ta se jeví optimální z hlediska minimalizace součtu čtverců odchylek jejích prvků od příslušných prvků

matice $\mathbf{X}$. Ve vyjádření matice $\mathbf{X}_{(2)}$ přitom použijeme pouze první dva sloupce matice $\mathbf{U}$ a první dva sloupce matice $\mathbf{V}$ ze singulárního rozkladu. Maticově lze tuto skutečnost zapsat jako

$$\mathbf{X} \approx \mathbf{X}_{(2)} = \mathbf{U}\mathbf{D}\mathbf{V}^T = (\mathbf{u}_1, \mathbf{u}_2) \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix} \begin{pmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{pmatrix}. \tag{3.1}$$

Je zřejmé, že $\mathbf{X}_{(2)}$ je opět rozměrů $n \times D$. Můžeme ji rozdělit takto:

$$\mathbf{X}_{(2)} = \mathbf{G}\mathbf{H}^T, \tag{3.2}$$

kde

$$\mathbf{G} = (\mathbf{u}_1, \mathbf{u}_2) \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}^{1-c}, \tag{3.3}$$

$$\mathbf{H} = (\mathbf{v}_1, \mathbf{v}_2) \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}^{c} \tag{3.4}$$

pro $0 \leq c \leq 1$. Řádky matice $\mathbf{G}$ pak využijeme v biplotu k zobrazení objektů a řádky matice $\mathbf{H}$ k zobrazení proměnných.

Nyní se podíváme na vztah (3.1) z trochu jiného pohledu. Matice $\mathbf{X}_{(2)}$ minimalizuje tzv. Frobeniovu maticovou normu vyjádřenou jako $\left\|\mathbf{X}_{(2)} - \mathbf{X}\right\|^2 = \sum_i \sum_j \left(x_{(2)ij} - x_{ij}\right)^2$, pro libovolnou matici $\mathbf{X}$ hodnosti 2. Vypovídající schopnost tohoto biplotu potom bude souviset s kvalitou aproximace $\mathbf{X}$ pomocí $\mathbf{X}_{(2)}$.

Mějme dánu datovou matici $\mathbf{X}$. Tuto můžeme centrovat pomocí sloupcových průměrů $\bar{x}_j,\ j = 1, 2, \ldots, D$. Také může být požadována normalizace proměnných. V tomto případě využijeme odhad směrodatné odchylky j-té proměnné s_j. Prvky transformované matice $\mathbf{Z}$ mají tvar

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, D. \tag{3.5}$$

Nyní se podíváme na vztahy (3.3) a (3.4) podrobněji. Nejdříve za parametr c dosadíme $c = 1$. Tento zápis nám vyjadřuje tzv. kovarianční biplot (covariance biplot), který je vhodný zejména k zobrazení proměnných. Z pohledu matic $\mathbf{G}$ a $\mathbf{H}$ tvoříme kovarianční (standardní) biplot tak, že matice $\mathbf{G}$ a $\mathbf{H}$ vyjádříme jako $\mathbf{G} = \mathbf{U}$, $\mathbf{H} = \mathbf{VD}$, kde matice $\mathbf{U}, \mathbf{D}$ a $\mathbf{V}$ jsou matice singulárního rozkladu původní datové matice $\mathbf{X}$. Název je dán díky vlastnosti, že výraz $\mathbf{HH}^T = \mathbf{VD}^2\mathbf{V}^T$ vydělený výrazem $(n-1)$ aproximuje příslušnou varianční matici. Zde, až na konstantu $\sqrt{n-1}$, aproximují délky paprsků směrodatné odchylky proměnných a kosinus úhlu mezi paprsky vyjadřuje jejich korelaci.

Naopak, pokud dosadíme $c = 0$, dostaneme tzv. vytvořený biplot (form biplot). Tento je vhodný k přednostnímu zobrazení jednotlivých pozorování. Maticově pro biplot vytvořený platí $\mathbf{G} = \mathbf{UD}, \mathbf{H} = \mathbf{V}$. Matice skalárních součinů řádků datové matice, $\mathbf{XX}^T$, je v něm aproximována pomocí $\mathbf{GG}^T$. Biplot vytvořený se v praxi neužívá v takové míře jako kovarianční podoba.

V obou případech v biplotu zobrazujeme pozorování pomocí bodů a proměnné pomocí paprsků vycházejících z počátku. Délky paprsků i jejich úhly jsou důležité pro další interpretaci. Toto je patrné také z grafů v Příloze A. Grafy jsou převzaty a přeloženy z [2]. Zde například vidíme, že v kovariančním biplotu aproximuje délka paprsku (vektoru) směrodatnou odchylku proměnných, kosinus úhlu aproximuje korelaci proměnných nebo vzdálenost mezi body aproximuje Mahalanobisovu vzdálenost mezi pozorováními. Ta je užívána především ve své výběrové podobě a má pro i-tý prvek výběru z rozdělení náhodného vektoru $\mathbf{X} = (X_1, \ldots, X_D)^T$, příslušný výběrový průměr $\bar{\mathbf{X}}$ a výběrovou varianční matici $\mathbf{S}$ tvar

$$MD(\mathbf{X}_i) = \sqrt{(\mathbf{X}_i - \bar{\mathbf{X}})^T \mathbf{S}^{-1} (\mathbf{X}_i - \bar{\mathbf{X}})}. \tag{3.6}$$

Mahalanobisova vzdálenost vyjadřuje vzdálenost pozorování $\mathbf{X}_i$ od centra distribuce datového souboru, vyjádřeného pomocí výběrového průměru $\bar{\mathbf{X}}$, vzhledem ke kovarianční struktuře, dané výběrovou varianční maticí $\mathbf{S}$.

Další vlastnosti biplotu si uvedeme v následujících kapitolách.

3.2 Kompoziční biplot

V úvodu bych chtěla upřesnit, že pokud mluvíme ve vztahu ke kompoziční datové matici o proměnných, myslíme tím jednotlivé složky kompozice.

Datovou matici $\mathbf{X}$ jsme si již nadefinovali v předchozí kapitole. Tuto matici budeme centrovat dle kapitoly 1.4 (normování se v tomto případě neužívá). Dále vytvoříme matici $\mathbf{Z}$ pomocí clr koeficientů uvedených v (1.14). Všimněme si, že matice $\mathbf{X}$ a $\mathbf{Z}$ mají stejné rozměry, tedy n řádků a D sloupců, $n \geq D$. Také si připomeňme, že clr transformace zachovává vzdálenosti mezi objekty. Proto můžeme použít na matici $\mathbf{Z}$ standardní postupy. Zejména pak skutečnost, že nejlepší nahrazení matice $\mathbf{Z}$, z pohledu minimalizace součtu čtverců odchylek jejích prvků od příslušných prvků aproximované matice $\mathbf{Y}$, je při použití singulárního rozkladu matice $\mathbf{Z}$ (konkrétně viz (3.8)). V následujícím textu systematizujeme poznatky zavedené v předchozích kapitolách.

Singulární rozklad $\mathbf{Z}$ tedy získáme pomocí matice vlastních vektorů $\mathbf{Z}\mathbf{Z}^T$ s označením $\mathbf{L}$, matice vlastních vektorů $\mathbf{Z}^T\mathbf{Z}$ označené $\mathbf{M}$ a diagonální matice čtvercových odchylek s kladných vlastních čísel $(\lambda_1, \lambda_2, \ldots, \lambda_s)$ obou matic $\mathbf{Z}\mathbf{Z}^T$ i $\mathbf{Z}^T\mathbf{Z}$ s označením $\mathbf{K}$ ($k_i = \sqrt{\lambda_i}, i = 1, 2, \ldots, s$). Maticově potom

$$\mathbf{Z} = \mathbf{L} \begin{pmatrix} k_1 & 0 & \ldots & 0 \\ 0 & k_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & k_s \end{pmatrix} \mathbf{M}^T, \tag{3.7}$$

kde s je hodnost matice $\mathbf{Z}$ a čísla $k_1, k_2, \ldots, k_s$ jsou seřazena sestupně. Nejčastěji je $s = D - 1$. Tento zápis je obdobou (3.1), pouze jsme předefinovali matice $\mathbf{U}$, $\mathbf{D}$ a $\mathbf{V}$ na $\mathbf{L}$, $\mathbf{K}$ a $\mathbf{M}$. K tomuto přeznačení matic dochází také v případě vyjádření aproximované matice $\mathbf{Y}$ (opět s užitím matic hodnosti 2)

$$\mathbf{Z} \approx \mathbf{Y} = \mathbf{L}\mathbf{K}\mathbf{M}^T = (\mathbf{l}_1, \mathbf{l}_2) \begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix} \begin{pmatrix} \mathbf{m}_1^T \\ \mathbf{m}_2^T \end{pmatrix}. \tag{3.8}$$

Podíl variability zachovaný pomocí této aproximace je $\frac{\lambda_1+\lambda_2}{\sum_{i=1}^{s}\lambda_i}$.

Také vyjádření matice $\mathbf{Y}$ pomocí $\mathbf{G}$ a $\mathbf{H}$, jak je popsáno v (3.2), probíhá stejným způsobem, tedy

$$\mathbf{Y} = \mathbf{G}\mathbf{H}^T, \tag{3.9}$$

kde pro $0 \leq c \leq 1$

$$\mathbf{G} = (\mathbf{l}_1, \mathbf{l}_2)\begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix}^{1-c}, \qquad \mathbf{H} = (\mathbf{m}_1, \mathbf{m}_2)\begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix}^{c}. \tag{3.10}$$

Matice $\mathbf{G}$ je typu $n \times 2$, matice $\mathbf{H}$ je typu $D \times 2$ a pro volbu $c = 1$ nám vychází

$$\begin{aligned} \mathbf{Y} &= \begin{pmatrix} \sqrt{n-1}l_{11} & \sqrt{n-1}l_{21} \\ \sqrt{n-1}l_{12} & \sqrt{n-1}l_{22} \\ \vdots & \vdots \\ \sqrt{n-1}l_{1n} & \sqrt{n-1}l_{2n} \end{pmatrix} \begin{pmatrix} \frac{k_1 m_{11}}{\sqrt{n-1}} & \frac{k_1 m_{12}}{\sqrt{n-1}} & \dots & \frac{k_1 m_{1D}}{\sqrt{n-1}} \\ \frac{k_2 m_{21}}{\sqrt{n-1}} & \frac{k_2 m_{22}}{\sqrt{n-1}} & \dots & \frac{k_2 m_{2D}}{\sqrt{n-1}} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{g}_1^T \\ \mathbf{g}_2^T \\ \vdots \\ \mathbf{g}_n^T \end{pmatrix} (\mathbf{h}_1\ \mathbf{h}_2\ \dots \mathbf{h}_D). \end{aligned} \tag{3.11}$$

Vektory $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_n$ nazýváme skóry, vektory $\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_D$ jsou zátěže.

Grafická interpretace kompozičního biplotu vychází z interpretace biplotu standardního, kterou jsme zmínili v kapitole 3.1. Řídíme se tedy obdobnými pravidly, jako v případě popisu grafů v Příloze A. Jsou zde ale jisté odlišnosti, které jsou způsobeny faktem, že v kompozičním biplotu nezobrazujeme samotné proměnné, ale jejich transformace (nejčastěji clr). Proto například nemůžeme interpretovat úhly mezi šipkami jako ukazatel korelace mezi původními složkami kompozice.

Slovní popis obsažený v následujícím odstavci odpovídá grafu v Příloze B, který je převzatý z [2]. Při interpretaci biplotu je potřeba dobře sledovat paprsky

(šipky). Šikmá osa vedoucí skrze paprsky se nazývá osa biplotu. Hodnota každého prvku (původní) datové matice je aproximována skalárním součinem mezi příslušnými řádky matic **G** a **H** biplotu (mezi pozorováním a proměnnou) a je shodná s projekcí pozorování na osu biplotu násobenou délkou paprsku.

V následujícím odstavci budeme opět popisovat graf z Přílohy B. Libovolná kombinace paprsků (rays) v biplotu nám dává vektor, který reprezentuje odpovídající lineární kombinaci proměnných. Hodnota rozdílu mezi dvěma proměnnými může být graficky zřetelná z přímky, která spojuje dva vrcholy různých paprsků a je nazývána spojnice (link). V obrázku je rozdílnost (resp. podobnost) proměnných A a B vyjádřena pomocí čárkované spojnice. Tuto skutečnost můžeme chápat také tak, že čím mají dvě proměnné kratší spojnici, tím jsou si podobnější v rámci všech hodnot pozorování, a naopak s rostoucí délkou spojnice se jejich hodnoty skrz celý soubor objektů různí. Podrobněji si postup interpretace uvedeme v kapitole 3.5.

Pro kompoziční biplot můžeme také vytvořit jeho kovarianční a vytvořenou podobu. Postup tvorby bude analogický (až na označení a následnou interpretaci) jako v kapitole 3.1.

3.3 Logratio biplot

Výše uvedený kompoziční biplot je v některých zdrojích označován jako tzv. clr biplot. Například v [3] je uvedena také druhá podoba kompozičního biplotu - logratio biplot (biplot logaritmů podílů). V této kapitole si jej krátce představíme.

Předpokládejme, že určíme logaritmy $\ln(x_{ij})$ kompoziční datové matice **X**. Tyto logaritmy poté uspořádáme do matice **J** o rozměru $n \times D$. Matice všech logaritmů podílů $\ln(x_{ij}/x_{ij'})$ je rovna $\mathbf{JB}_D$, kde $\mathbf{B}_D$ je matice o rozměru $D \times D(D-1)$ obsahující čísla 1 a -1 v každém sloupci na pozici j a j' a číslo 0 ve zbylých pozicích sloupce, tedy

$$\mathbf{B}_D = \begin{pmatrix} 1 & 1 \ldots & 1 & -1 & 0 \ldots & 0 \ldots & -1 \ldots & 0 \\ -1 & 0 \ldots & 0 & 1 & 1 \ldots & 1 \ldots & \ldots \ldots \ldots & \ldots \\ 0 & -1 \ldots \ldots & 0 & -1 & \ldots \ldots \ldots & \ldots \ldots \ldots & \ldots \\ \ldots & \ldots \ldots \ldots & \ldots & \ldots & \ldots \ldots & 0 \ldots & 0 \ldots & -1 \\ 0 & 0 \ldots & -1 & 0 & 0 \ldots & -1 \ldots & 1 \ldots & 1 \end{pmatrix}. \tag{3.12}$$

V tomto případě jsme brali v úvahu všechny dvojice logaritmů podílů, nikoliv jen ty splňující $j < j'$, jak jsme uvažovali v kapitole 1.5.

Matice logaritmů podílů je dále centrována tak, že ji zleva vynásobíme maticí $\mathbf{C}_n$. Tedy $\mathbf{Y} = \mathbf{C}_n\mathbf{J}\mathbf{B}_D$, kde $\mathbf{C}_n$ je idempotentní centrovací matice řádu n. Pro jednotkovou matici $\mathbf{I}_n$ a vektor n jedniček $\mathbf{1}$ platí

$$\mathbf{C}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T = \begin{pmatrix} \frac{n-1}{n} & \frac{-1}{n} & \ldots & \frac{-1}{n} \\ \frac{-1}{n} & \frac{n-1}{n} & \ldots & \frac{-1}{n} \\ \ldots & \ldots & \ldots & \ldots \\ \frac{-1}{n} & \frac{-1}{n} & \ldots & \frac{n-1}{n} \end{pmatrix}. \tag{3.13}$$

Předpokládejme, že singulární rozklad matice $\mathbf{Y}$ má tvar $\mathbf{Y} = \mathbf{P}\Phi\mathbf{N}^T$. Tento rozklad je propojený s clr vektory $\mathbf{Z} = \mathbf{L}\mathbf{K}\mathbf{M}^T$, kde $\mathbf{Z} = \mathbf{C}_n\mathbf{J}\mathbf{C}_D$ následujícím způsobem:

$$\begin{aligned} \mathbf{Y}\mathbf{Y}^T &= \mathbf{C}_n\mathbf{J}\mathbf{B}_D\mathbf{B}_D^T\mathbf{J}^T\mathbf{C}_n = \mathbf{C}_n\mathbf{J}2D\mathbf{C}_D\mathbf{J}^T\mathbf{C}_n = 2D\mathbf{C}_n\mathbf{J}\mathbf{C}_D\mathbf{J}^T\mathbf{C}_n \\ &= 2D\mathbf{L}\mathbf{K}^2\mathbf{L}^T = \mathbf{L}(2D\mathbf{K}^2)\mathbf{L}^T, \end{aligned}$$

protože $\mathbf{B}_D\mathbf{B}_D^T = 2D\mathbf{C}_D$. Vztah mezi clr a logratio biplotem je takový, že platí $\mathbf{\Phi} = \sqrt{2D}\mathbf{K}$. Levé singulární vektory jsou identické, $\mathbf{P} = \mathbf{L}$.

Pro tvorbu matic $\mathbf{G}_{lr}$ a $\mathbf{H}_{lr}$ v kovariančním logratio biplotu platí

$$\begin{aligned} \mathbf{G}_{lr} &= \mathbf{P} = \mathbf{L}, \\ \mathbf{H}_{lr} &= \mathbf{N}\Phi = \tfrac{\mathbf{B}_D^T\mathbf{M}}{\sqrt{2D}}\sqrt{2D}\mathbf{K} = \mathbf{B}_D^T\mathbf{M}\mathbf{K}. \end{aligned} \tag{3.14}$$

3.4 Ilr biplot

Posledním z řady kompozičních biplotů, který si zde velmi v krátkosti představíme, je ilr biplot. Hlavním zdrojem informací o této problematice byla kniha [19].

Pro získání ilr biplotu musíme nejdříve na kompoziční data aplikovat ilr transformaci. K tomu využijeme postupné binární dělení, kterému jsme se věnovali v kapitole 1.3.2. Provedeme tedy transformaci původních kompozic na ilr data s využitím tabulky tvořené jednotlivými kroky PBD. Tuto novou datovou matici potom centrujeme a dále s ní pracujeme jako v případě kovariančního biplotu - tedy ji rozkládáme pomocí singulárního rozkladu a tvoříme z něj další matice **G** a **H**.

Interpretace toho biplotu je stejná, jako v případě standardního biplotu [17]. Tedy druhé mocniny délek paprsků nám aproximují rozptyl jednotlivých složek kompozice, kosinus úhlu mezi šipkami aproximuje korelaci mezi dvěma proměnnými. Jednotlivá pozorování jsou reprezentována body, u kterých sledujeme jejich vzájemnou polohu. Jediným problémem v případě ilr biplotu bývá interpretace samotných ilr proměnných, vytvořených pomocí PBD. Podrobnější postup konstrukce a interpretace ilr biplotu uvedeme u praktického příkladu věnovanému této problematice.

3.5 Grafická interpretace clr biplotu

Metodika popisu kompozičního biplotu byla v [2] vysvětlena na příkladu 22 obrazů, které jsou složeny vždy ze 6 barev. Protože nám jde o relativní zastoupení barev na obrazech, jedná se zřejmě o kompoziční data. Je zde ukázáno, jak můžeme pomocí singulárního rozkladu clr transformované datové matice získat příslušný biplot a jak může být pomocí něj odhalena možná souvislost mezi jednotlivými barvami. Než si analýzu uvedeného datového souboru rozebereme v této práci, podívejme se podrobněji na interpretaci kompozičního (clr) biplotu.

Grafická interpretace kompozičního biplotu vychází z rozšíření interpretace

biplotu standardního. Kompoziční biplot se skládá ze tří částí:

1. Počátek O, který vyjadřuje centrum kompozičního datového souboru;
2. Vrchol na pozici $\mathbf{h}_j$ pro každou z D proměnných (složek kompozice);
3. Ukazatel na pozici $\mathbf{g}_i$ pro každý z n objektů.

Jak jsme se zmínili již dříve, spojení počátku O a vrcholu $\mathbf{h}_j$ nazýváme paprsek $\overline{O\mathbf{h}_j}$ a propojení dvou vrcholů a $\mathbf{h}_j$ a $\mathbf{h}_k$ označujeme jako spojnici $\overline{\mathbf{h}_j\mathbf{h}_k}$. Tyto funkce tvoří základní charakteristiky biplotu s následujícími hlavními vlastnostmi pro interpretaci kompoziční variability.

1. Spojnice a paprsky poskytují informace o relativní variabilitě v kompozičních datech vyjádřené jako

$$|\overline{\mathbf{h}_j\mathbf{h}_k}|^2 \approx \mathrm{var}\left(\ln\frac{x_j}{x_k}\right), \qquad |\overline{O\mathbf{h}_j}|^2 \approx \mathrm{var}\left(\ln\frac{x_j}{g(\mathbf{x})}\right). \tag{3.15}$$

   Při interpretaci paprsků ve smyslu jednotlivých složek kompozice ovšem musíme být opatrní, protože uvažovaná clr proměnná souvisí s celou kompozicí skrze $g(\mathbf{x})$ a zmenší se, pokud pracujeme pouze s podkompozicí.

2. Informace o korelaci podkompozic nám dávají spojnice. Když se spojnice $\overline{\mathbf{h}_j\mathbf{h}_k}$ a $\overline{\mathbf{h}_i\mathbf{h}_l}$ protínají v M, pak platí

$$\cos(\mathbf{h}_j M \mathbf{h}_i) \approx \mathrm{corr}\left(\ln\frac{x_j}{x_k}, \ln\frac{x_i}{x_l}\right). \tag{3.16}$$

   Jestliže dvě spojnice svírají pravý úhel, pak můžeme očekávat nekorelovanost jejich logaritmů podílů. Tato vlastnost je užitečná při vyšetřování možné nezávislosti podkompozic.

3. Analýza podkompozic: Počátek O je těžištěm D vrcholů. Při vytvoření libovolné podkompozice mezi zbývajícími složkami jsou poměry zachovány.

Z toho vyplývá, že biplot pro jakoukoliv podkompozici je vytvářen na základě výběru vrcholů odpovídajících částem podkompozice a centrum vzniklého biplotu je těžiště těchto vrcholů.

4. Shodné vrcholy: Jestliže vrcholy j a k splývají, je $\mathrm{var}(\ln(x_j/x_k))$ přibližně roven nule, tedy podíl x_j/x_k je konstantní a x_j a x_k můžeme považovat za vzájemně nahraditelné složky.

5. Kolineární vrcholy: Jestliže je podmnožina vrcholů kolineární, může to znamenat, že příslušná podkompozice má jednodimenzionální biplot, tedy že výběrové podkompozice leží (po clr transformaci) na jedné přímce.

Abychom lépe pochopili celou problematiku, převezmeme příklad použitý ve zdrojové literatuře [2]. Nyní se podívejme na Přílohu C. Zde je vidět tabulka proporcí jednotlivých barev na 22 obrazech.

Rozdíl v zastoupení jednotlivých barev se značně liší obraz od obrazu. Hlavním úkolem je zjistit, jak budou vypadat podkompozice například v případě, kdy chceme znát vztah pouze mezi třemi barvami.

Podívejme se na grafy v Příloze D. Oba znázorňují výsledky zpracování dat z Přílohy C (vztahy jednotlivých barev na uměleckých plátnech) a byly převzaty z [2]. První z nich vyjadřuje vytvořený biplot, druhý biplot je kovarianční. V druhém případě jsme použili clr transformované kompozice, které byly dále škálovány vydělením konstantou $\sqrt{n-1} = \sqrt{21}$.

Než přejdeme k popisu obou grafů, uvedeme si několik základních označení:

- i : i-tý řádkový bod (pozorování) ve standardních souřadnicích,
- I : i-té pozorování v hlavních souřadnicích,
- j : j-tý sloupcový bod (proměnná) ve standardních souřadnicích,
- J : j-tá proměnná v hlavních souřadnicích.

Na základě tohoto značení můžeme definovat, že vytvořený biplot zobrazuje I a j a kovarianční biplot naopak i a J. Paprsky a spojnice jsou rozeznávány na základě koncových bodů. Například OJ v kovariančním biplotu značí paprsek k vrcholu j-té složky, JJ' je spojnice vrcholů J a J' reprezentující rozdíl $J' - J$. Vzdálenosti mezi body budeme označovat $|OI|, |JJ'|, |ii'|$.

V tomto odstavci si uvedeme vztahy, které nám pomohou osvětlit spojitosti v rámci popisu grafu, které budeme následně používat. Mějme dány složky A, B, C, D. Vzdálenost mezi body A a B označíme λ, vzdálenost mezi B a C μ. Konstantní kontrast logaritmů má tvar

$$\mu \ln A + \lambda \ln C - (\lambda + \mu) \ln B = \mu \ln \frac{A}{B} - \lambda \ln \frac{B}{C} = konstanta, \tag{3.17}$$

tento vztah nám vyjadřuje proporcionalitu

$$\left(\frac{A}{B}\right)^{\mu} \propto \left(\frac{B}{C}\right)^{\lambda}. \tag{3.18}$$

V případě kontrastu logaritmů v této podobě

$$\ln A - \ln B + \ln C - \ln D = \ln \frac{A}{B} - \ln \frac{D}{C} = \ln \frac{A}{D} - \ln \frac{B}{C} = konstanta, \tag{3.19}$$

je proporcionalitu možné vyjádřit dvěma způsoby

$$\left(\frac{A}{B}\right) \propto \left(\frac{D}{C}\right) \quad , \quad \left(\frac{A}{D}\right) \propto \left(\frac{B}{C}\right). \tag{3.20}$$

Nyní přejdeme k samotné interpretaci a popisu grafů z Přílohy D. Datová matice je centrována do středu O. Data jsou centrována skrze řádky i sloupce, tedy můžeme říci, že jsme použili dvojité centrování. To znamená, že průměr hodnot pozorování (obrazů) i proměnných (barev) je znázorněn ve středu grafu.

Jednotlivá pozorování mají v obou grafech stejné prostorové upořádání. Liší se pouze měřítkem. Rozdílným prvkem jsou zde paprsky. Ty jdou stejnými směry, liší se ale jejich délky. To znamená, že jednotlivé proměnné mají v každém grafu trochu jiný vliv na celkové uspořádání.

Vzdálenosti $|II'|$ ve vytvořeném biplotu aproximují vzdálenosti mezi pozorováními. Tuto hodnotu počítáme buď z matice clr transformovaných kompozic nebo z matice párových logaritmů podílů.

Vzdálenosti $|JJ'|$ v kovariančním biplotu reprezentují směrodatné odchylky příslušných logaritmů podílů. Když se podíváme na paprsek černé barvy a šipku ostatní vidíme, že jejich vrcholy jsou velmi blízko u sebe, tedy jejich spojnice je velmi krátká. To značí, že tyto dvě proměnné mají téměř konstantní podíl pro všechna pozorování. Naopak nejdelší spojnici mají paprsky červené a modré. Tyto dvě proměnné mají tedy velmi proměnlivé hodnoty podílů v rámci všech obrazů.

Kosinus úhlu mezi spojnicemi v kovariančním biplotu vyjadřuje korelace příslušných logaritmů podílů. V grafu vidíme, že spojnice paprsků modré, červené a žluté jsou téměř kolmé na spojnice bílé, ostatní a černé. To znamená, že logaritmy podílů mezi těmito dvěma skupinami jsou téměř nekorelované. Naopak v rámci těchto dvou skupin je korelace silná. Tato vysoká korelace je také zřetelná, když porovnáme délky jednotlivých spojnic. Délka spojnice červená (dále označíme jako C) - žlutá (Z) je zhruba 2.5 násobkem délky spojnice žlutá - modrá (M). Obě spojnice mohou být přeneseny do počátku, tedy symbolicky

$$\ln\frac{C}{Z} - \text{ave}\left(\ln\frac{C}{Z}\right) = 2.5\left[\ln\frac{Z}{M} - \text{ave}\left(\ln\frac{Z}{M}\right)\right], \tag{3.21}$$

kde ave(...) označuje průměr příslušných logaritmů podílů. Jak jsme si uvedli výše, tento vztah můžeme dále převést na tvar v podobě kontrastu logaritmů

$$2.5\ln M + \ln C - 3.5\ln Z = konstanta. \tag{3.22}$$

Proporcionální vztah mezi příslušnými barvami může být také zapsán jako

$$\frac{C}{Z} \propto \left(\frac{Z}{M}\right)^{2.5}. \tag{3.23}$$

Dále se podívejme na paprsky barev černá (Ce), modrá (M), bílá (B) a červená (C). Spojnice barev černá - červená a modrá - bílá můžeme přenést do středu. Tímto krokem získáme vztah

$$\ln\frac{Ce}{C} - \mathrm{ave}\left(\ln\frac{Ce}{C}\right) = \ln\frac{M}{B} - \mathrm{ave}\left(\ln\frac{M}{B}\right), \tag{3.24}$$

odkud dostaneme kontrast logaritmů

$$\ln Ce - \ln C + \ln B - \ln M = konstanta \tag{3.25}$$

a odtud proporcionální vztahy

$$\frac{Ce}{C} \propto \frac{M}{B} \quad nebo \quad \frac{Ce}{M} \propto \frac{C}{B}. \tag{3.26}$$

Mějme dánu podmnožinu I pozorování a J proměnných. Pokud tato pozorování, resp. proměnné, leží vždy v jedné přímce tak, že jsou tyto dvě přímky ortogonální, potom má podmatice tvořená příslušnými I řádky a J sloupci (přibližně) konstantní logaritmy podílů pro všechny komponenty. Tedy dvojitě centrovaná submatice logaritmů má složky blízké nule. Tuto vlastnost můžeme v obou biplotech pozorovat například v případě maleb s čísly 9, 15 a 21 a spojnice barev bílá - ostatní - černá.

Při interpretaci biplotu musíme být velmi obezřetní. Nemůžeme pouze slepě sledovat pravidla. Musíme si všímat, jaká data máme, jestli v nich není mnoho odlehlých hodnot, zda nám výsledky vycházejí logické. Další úskalí a jejich řešení si ukážeme v kapitole s praktickými příklady. Alternativní metodu konstrukce clr biplotu v případě nehomogenních dat či dat s odlehlými hodnotami si ukážeme v další kapitole.

4 Robustní přístup ke konstrukci clr biplotu

V následující kapitole nastíníme konstrukci clr biplotu pomocí metod robustní statistiky. Zdroje tvořily články [9], [10], [11].

Nejideálnější datový soubor pro statistickou analýzu je ten, který splňuje předpoklad normálního rozdělení, nevyskytují se v něm žádná odlehlá pozorování ani rozdílná měřítka proměnných. Takový soubor ovšem v reálném životě najdeme jen stěží. Proto se jako vhodnější často jeví užití tzv. robustních statistických metod, které nám pomohou vliv odlehlých hodnot a různost měřítek potlačit. Dále se zaměříme na robustní zpracování metody hlavních komponent v případě kompozičních dat.

Z interpretačních důvodů obvykle aplikujeme metodu hlavních komponent pro clr transformované kompozice. Jak víme, hlavní komponenty bývají využívány k zjednodušení mnohorozměrné datové struktury a mohou nám pomoci redukovat dimenzi úlohy. Pokusy o zachování co největšího dílu informace o souboru mohou být ovšem ovlivněny existencí odlehlých hodnot. V klasické metodě hlavních komponent je ve výpočtech využíván výběrový průměr a výběrová varianční matice. V případě robustní metody hlavních komponent je nahrazujeme robustními odhady těchto charakteristik. V takovém případě již k výpočtům skórů a zátěží nemůžeme použít clr transformaci. Ta je nahrazena ilr transformací a teprve výsledné hlavní komponenty jsou převedeny do prostoru clr transformovaných kompozic.

Připomeňme si v tomto ohledu vztah (2.1). Zde je pro výpočet použit klasický výběrový průměr $\bar{\mathbf{X}}$ a výběrová varianční matice $\mathbf{S}$. Tento přístup je plně akceptovatelný pro normálně rozložená data. Pokud ovšem máme v datech nějaké významné odlehlé hodnoty, je lepší tyto vztahy upravit, protože jak výběrový průměr, tak výběrová varianční matice jsou citlivé právě na tato odlehlá pozorování. Proto je lepší místo obou použít jejich robustnější protějšky. Označme si odhad polohy $T(\mathbf{X})$ a odhad variability $C(\mathbf{X})$. Potom bude mít vztah (2.1) podobu

$$\mathbf{W} = (\mathbf{X} - \mathbf{1}T(\mathbf{X})^T)\mathbf{G}_\mathbf{x}, \tag{4.1}$$

kde $\mathbf{G}_\mathbf{x}$ je matice, jejíž sloupce jsou jednotlivé vlastní vektory matice $C(\mathbf{X})$. Singulárním rozkladem totiž dostaneme $C(\mathbf{X}) = \mathbf{G}_\mathbf{x}\mathbf{L}\mathbf{G}_\mathbf{x}{}^T$, $\mathbf{L}$ je diagonální matice vlastních čísel matice $C(\mathbf{X})$.

Vedle požadavku robustnosti odhadů $T(\mathbf{X})$ a $C(\mathbf{X})$ je důležitá také jejich afinní ekvivariance. Ta platí v případě, jestliže pro výběr $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ (řádky matice $\mathbf{X}$), pro každou regulární matici $\mathbf{A}$ řádu D a vektor $\mathbf{a} \in \mathbb{R}^D$ platí

$$\begin{aligned} T(\mathbf{A}\mathbf{x}_1 + \mathbf{a}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{a}) &= \mathbf{A}T(\mathbf{x}_1, \ldots, \mathbf{x}_n) + \mathbf{a}, \\ C(\mathbf{A}\mathbf{x}_1 + \mathbf{a}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{a}) &= \mathbf{A}C(\mathbf{x}_1, \ldots, \mathbf{x}_n)\mathbf{A}^T, \end{aligned} \tag{4.2}$$

tedy že odhady $T(\mathbf{X})$ a $C(\mathbf{X})$ se při lineární transformaci chovají obdobně jako výběrový průměr a výběrová varianční matice.

Nejznámějším robustním odhadem je minimální kovarianční determinant (minimum covariance determinant - MCD) [10]. Metoda jeho výpočtu je značně rozsáhlá a již přesahuje rozsah této práce. Proto si ji zde představíme velmi zjednodušeně. Konkrétně, odhady polohy a variability dostaneme vypočtením výběrového průměru a výběrové varianční matice (přenásobené ještě určitou konstantou tzv. konzistenčním faktorem) pro takovou podmnožinu h pozorování ($h \leq n$), které odpovídá nejmenšímu determinantu výběrové kovarianční matice. Výběr čísla h určuje jak robustnost, tak eficienci odhadu. Hodnota h musí být rovna alespoň polovině rozsahu celého výběru n, což nám v tomto případě dává robustní výsledek. Příslušný odhad je ale málo eficientní. Druhá možnost je, že h je velké, blízké n. Zde je efekt opačný, máme tedy málo robustní, zato poměrně efektivní odhad. Kompromisem mezi těmito dvěma krajními situacemi je výběr o velikosti $h = (3/4)n$. Toto číslo nám říká, že vzniklý odhad je schopný „zvládnout“ zhruba $(n - h)/n = 1/4$ odlehlých pozorování z celkového počtu dat ve výběru. Musíme si ovšem dát pozor, aby ve výběru bylo opravdu přibližně $1/4$ odlehlých hodnot.

Pokud by jich totiž bylo méně, byl by takto vzniklý odhad zbytečně zkreslený.

Metoda MCD je ovšem určena pouze pro regulární datové matice s hodností rovnou počtu složek pozorování, nelze ji tedy aplikovat pro clr transformované kompozice. Místo toho je potřeba použít ilr transformovaná data.

Než přejdeme k odvozování příslušných vztahů, připomeňme si značení z kapitoly 1.3.3 o vztahu clr a ilr transformací. Mějme dánu datovou matici $\mathbf{X}$ o rozměru $n \times D$ kompozičních dat. Úpravou vztahu (1.24) dostaneme ve výběrovém případě

$$\mathbf{Y} = \ln(\mathbf{X})\mathbf{F}^T. \tag{4.3}$$

Dále platí vztah obdobný s (1.27), pouze v opačném směru

$$\mathbf{Z} = \mathbf{Y}\mathbf{V}, \tag{4.4}$$

kde matice $\mathbf{V}$ je tvořena např. stejnými vektory jako ve vztahu (1.25). $\mathbf{Z}$ je ilr transformovaná datová matice typu $n \times (D-1)$.

Použitím odhadů $T(\mathbf{X})$ a $C(\mathbf{X})$ zmíněných výše můžeme vytvořit aplikací vztahu (4.1) hlavní komponenty pro data ve tvaru ilr transformace

$$\mathbf{Z}^* = (\mathbf{Z} - \mathbf{1}T(\mathbf{Z})^T)\mathbf{G_z}. \tag{4.5}$$

Matice $\mathbf{G_z}$ je rozměru $(D-1) \times (D-1)$ a vznikne singulárním rozkladem MCD odhadu varianční matice

$$C(\mathbf{Z}) = \mathbf{G_z}\mathbf{L_z}\mathbf{G_z}^T. \tag{4.6}$$

Pokud má původní kompoziční datová matice $\mathbf{X}$, reprezentovaná pomocí proporcionálních dat, plnou sloupcovou hodnost $D-1$, pak má matice $\mathbf{Z}$ také plnou hodnost $D-1$. Na $\mathbf{Z}$ tedy můžeme MCD odhad aplikovat a dostaneme odhady $T(\mathbf{Z})$ a $C(\mathbf{Z})$. V takovém případě můžeme označit $\mathbf{Z}^*$ jako robustní skóry a $\mathbf{G_z}$ jako robustní zátěže. Protože v případě ilr transformací je interpretace biplotu

přece jen poněkud komplikovanější, převedeme výsledky zpět do clr. Použitím (4.4) dostaneme robustní složky v clr prostoru

$$\mathbf{Y}^* = \mathbf{Z}^*\mathbf{V}^T. \tag{4.7}$$

Dále platí

$$C(\mathbf{Y}) = C(\mathbf{Z}\mathbf{V}^T) = \mathbf{V}C(\mathbf{Z})\mathbf{V}^T = \mathbf{V}\mathbf{G}_\mathbf{z}\mathbf{L}_\mathbf{z}{\mathbf{G}_\mathbf{z}}^T\mathbf{V}^T, \tag{4.8}$$

protože MCD odhad rozptylu má vlastnost afinní ekvivalence, a tedy pro zátěže platí

$$\mathbf{G}_\mathbf{y} = \mathbf{V}\mathbf{G}_\mathbf{z}. \tag{4.9}$$

5 Příklady

V následující kapitole si ukážeme aplikace výše zmíněných teoretických postupů na dvou příkladech. Seznámíme se s tím, jak interpretovat konkrétní kovariančí a vytvořený biplot, dále si představíme ilr biplot a logratio biplot. Nakonec bude uvedena také jedna z robustních variant biplotu. K výpočtům a grafickému vyjádření použijeme statistický software R (`www.r-project.org`)[21].

Data použitá v následujících příkladech jsou čerpána z internetových stránek Eurostatu [6]. Na těchto stránkách je k dispozici celá řada tabulek s informacemi z mnoha oblastí o většině evropských států za roky 1997 - 2009 (mírně se liší u jednotlivých datových souborů). Pro účely této práce byla vybrána data týkající se kriminality a příčin smrti v důsledku nemoci. Data jsou vztažena k roku 2008 proto, aby mohly být jednotlivé příklady porovnávány i mezi sebou. Z tohoto důvodu bylo také vybráno konkrétních 24 zemí, které měly pro jednotlivá témata dostačující dostupné údaje. Konkrétní zdroje a podrobnosti budou uvedeny v příslušných kapitolách.

Jak víme, jednotlivé státy Evropy mají různé rozlohy a počty obyvatel. V případě standardního (nekompozičního) biplotu by mohla tato skutečnost vést ke zkreslujícím výsledkům. Protože se ale nyní pohybujeme v oblasti kompozičních dat, zajímají nás proporce a nikoliv absolutní hodnoty. Navíc, protože mají jednotlivé státy dostatečně velké počty obyvatel, proporce z nich vytvořené mají dostatečnou vypovídající hodnotu. Relativní informace obsažená v datech byla reprezentována prostřednictvím procentuálního zastoupení složek v rámci jednotlivých států. Tímto jsme ve vyjádření dat potlačili vliv počtu obyvatel jednotlivých států.

Dalšími zdroji pro tvorbu této kapitoly, zejména z oblasti softwaru, byly [12], [20].

5.1 Kriminalita

První datový soubor, použitý k ukázce práce s kompozičními biploty, nese název Kriminalita. Data jsou převzata z databáze Eurostatu [8]. Jak již bylo

řečeno v úvodu, konkrétní data byla zvolena pro rok 2008 a bylo vybráno 24 zemí, jejichž mezinárodní zkratky používané v příkladu jsou uvedeny v Příloze E.

Popisy jednotlivých kriminálních činů jsou v původním jazyce (angličtině) a jsou následující: homicide = vražda, violent crime = násilný čin, robbery = loupež, domestic burglary = vloupání do domácnosti, motor vehicle theft = krádež auta/vykradení auta, drug trafficking = distribuce drog. Úkolem je zjistit, jaký je vzájemný vztah mezi jednotlivými státy a také mezi jednotlivými kriminálními činy.

5.1.1 Kovarianční biplot

Nejdříve si ukážeme postup vytvoření kovariančního biplotu. Jak již bylo řečeno dříve, k výpočtům a grafickým zobrazením je používán statistický software R. Než začneme pracovat, nastavíme příslušnou zdrojovou knihovnu v daném počítači. K tomu použijeme příkaz:

`>setwd("Cesta k příslušné knihovně").`

Pro práci s kompozičními daty musíme otevřít doplňkové knihovny:

`>library (compositions)`, `>library(robCompositions).`

První z nich slouží ke statistické analýze kompozičních dat, druhá obsahuje (nejen) robustní statistické metody pro kompoziční data. Dále načteme zdrojová data, která jsou uložena v textovém souboru s názvem `Krimi.txt`, jako matici o 24 řádcích (státy) a 6 sloupcích (kriminální činy):

`>X=matrix(scan("Krimi.txt"),nrow=24,byrow=T).`

Označíme jednotlivé sloupce a řádky jmény:

```
>colnames(X)=c("Homicide","ViolentCrime","Robbery","Burglary",
"VehicleTheft","Drugs")
>rownames(X)=c("BG","CZ","DK","DE","EE","IE","GR","ES","FR","IT","LV",
"LT","LU","MG","AT","PL","PT","RO","SL","SK","FI","SE","GB","NO")
```

Nyní můžeme data reprezentovat prostřednictvím procentuálních podílů. Využijeme k tomu funkci:

```
>X=acomp(X,total=100).
```

Tímto příkazem jsme dostali kompoziční datový soubor, který vyjadřuje podíly jednotlivých přečinů na celkové kriminalitě (všechny kriminální činy v daném státě). Tento datový soubor můžeme nalézt v Příloze F. Protože ovšem výše zmíněný příkaz vrací kromě upravené datové matice ještě další informace, musíme pomocí

```
>X=X[1:24,]
```

„vytáhnout“ pouze tuto matici. Následně provedeme clr transformaci

```
>Y=clr(X)[1:24,]
```

a centrování transformované datové matice

```
>Z=scale(Y,center=TRUE,scale=FALSE).
```

Příkaz

```
>pdf("Krimi_cov.pdf")
```

nám vytvoří grafický výstup ve formátu pdf. Dále vytvoříme matice $\mathbf{L}$, $\mathbf{K}$ a $\mathbf{M}$ dle vztahu (3.7):

```
>L=svd(Z)$u[,c(1,2)]
>K=diag(svd(Z)$d[1:2])
>M=svd(Z)$v[,c(1,2)]
```

a matice $\mathbf{G}$ a $\mathbf{H}$ dle (3.8), respektive dle rozdělení pro kovarianční biplot, jak je uvedeno v kapitole 3.5:

```
>G=(sqrt(nrow(X)-1))*L
>H=(1/sqrt(nrow(X)-1))*M%*%K
>rownames(H)=colnames(X)
>rownames(G)=rownames(X)
```

Poslední dva řádky převádí označení řádků a sloupců původní datové matice na nově vytvořené matice. Následuje příkaz

`>summary(princomp(Z))`,

který ukazuje, kolik procent celkové variability statistického souboru (po vynásobení 100) je vysvětleno pomocí hlavních komponent, nově vytvořených statistických znaků. Pro účely biplotu nás zajímají první dva číselné sloupce, které odpovídají prvním dvěma (z hlediska variability nejvýznamnějším) hlavním komponentám. V tomto případě máme tabulku:

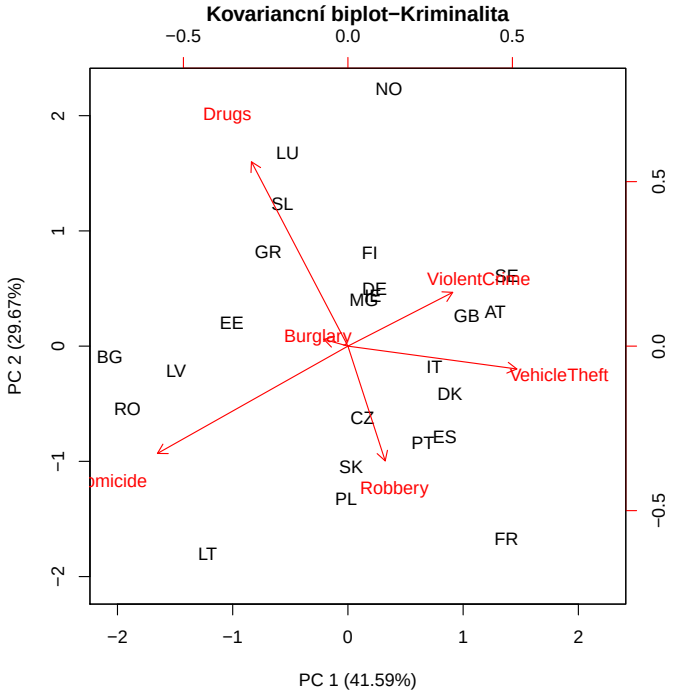
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Standard deviation	1.0975731	0.9269876	0.6161308	0.5862384	0.33028911	3.893152e-09
Proportion of Variance	0.4159249	0.2966852	0.1310671	0.1186579	0.03766488	5.232999e-18
Cumulative Proportion	0.4159249	0.7126101	0.8436772	0.9623351	1.00000000	1.00000000

První řádek tabulky nám ukazuje směrodatné odchylky jednotlivých hlavních komponent. Druhý řádek vyjadřuje, že první hlavní komponenta vysvětluje 41.59% a druhá hlavní komponenta dalších 29.67 % celkové variability, což dá v součtu 71.26 %, a tedy dobrý výsledek. Tento součet je vidět v poli třetího řádku, sloupce s nadpisem `Comp.2` (číslo 0.7126101). Biplot tedy bude dobře odrážet skutečnou strukturu mnohorozměrného datového souboru. Podrobnější vysvětlení problematiky celkové vysvětlené variability souboru naleznete v bakalářské práci [17] na straně 20.

Poslední příkaz, který byl použit,

```
>biplot(x=G,y=H,xlab="PC 1 (41.59%)",ylab="PC 2 (29.67%)",main=
"Kovarianční biplot-Kriminalita")
>dev.off()
```

vykreslí clr (kompoziční) biplot pro naše data do výše uvedeného pdf souboru.



Kovarianční biplot se užívá zejména pro zobrazení vztahů mezi jednotlivými složkami kompozice. Jeho obecná interpretace je uvedena v grafu v Příloze A.

Datová matice je centrována, tedy průměr hodnot pozorování (států) i složek kompozice (kriminálních činů) je znázorněn ve středu grafu.

Vzdálenosti mezi jednotlivými body nám udávají aproximovanou Mahalanobisovu vzdálenost mezi jednotlivými pozorováními. Díky této vlastnosti můžeme tedy říci, že čím jsou jednotlivé státy v grafu blíže u sebe, tím jsou si podobnější. Zhruba v dolní polovině grafu můžeme například vidět shluk středoevropských států Česká republika (CZ), Slovensko (SK) a Polsko (PL). Navíc jsou u vrcholu paprsku loupež (roberry). Tedy můžeme říci, že tento druh kriminality v daných zemích převažuje. Blízko vrcholu této šipky jsou také státy Španělsko (ES) a Portugalsko (PT). Pokud bychom si kompoziční datový soubor seřadili právě podle loupeží, tyto dva státy by měly největší procentuální zastoupení ze všech. Další zajímavou skupinu tvoří Finsko (FI), Maďarsko (MG) a Německo (DE). Tyto státy spolu geograficky příliš nesouvisí, ovšem podíly drogové kriminality (drugs) a násilných činů (violent crime) (a v případě MG a DE i vloupání do domácnosti

- burglary) jsou zde velmi podobné. Zajímavá jsou také „osamocená“ pozorování Norsko (NO) a Francie (FR); u Norska převládá drogová kriminalita (je to patrné také z dat), ve Francii mají zase problémy s krádežemi aut (vehicle theft) a loupežnými přepadeními (robbery). Relativní zastoupení ostatní kriminality je podprůměrné.

Další prvek, kterého si na grafu všímáme, jsou délky jednotlivých paprsků. Ty aproximují směrodatnou odchylku proměnných (clr transformovaných složek kompozice). Vidíme, že nejdelší jsou paprsky drog a vražd (homicide). Pokud se podíváme na datový soubor, vidíme opravdu u těchto dvou složek značný „rozptyl“ hodnot vzhledem k jejich velikostem. Znatelně nejkratší paprsek, a tedy i nejmenší odchylku, má prvek vloupání do domácnosti. Ostatní tři proměnné jsou na tom z tohoto hlediska téměř stejně.

Poslední část rozboru tohoto grafu se bude týkat problematiky naznačené v Příloze B, tedy paprsků a spojnic. Zde vidíme, že spojnice mezi dvojicemi drogy - loupež a vražda - násilný čin jsou vzájemně ortogonální, tedy jsou téměř nekorelované. Navíc jsou tyto spojnice velmi dlouhé. To znamená, že v rámci těchto dvojic jsou proporce kriminálních činů velmi proměnlivé skrze všechny státy. Z hlediska délek spojnic k sobě mají vzájemně nejblíže proměnné odpovídající násilným činům a krádežím vozidel, tedy i odpovídající podíly mezi nimi v jednotlivých státech budou docela stabilní.

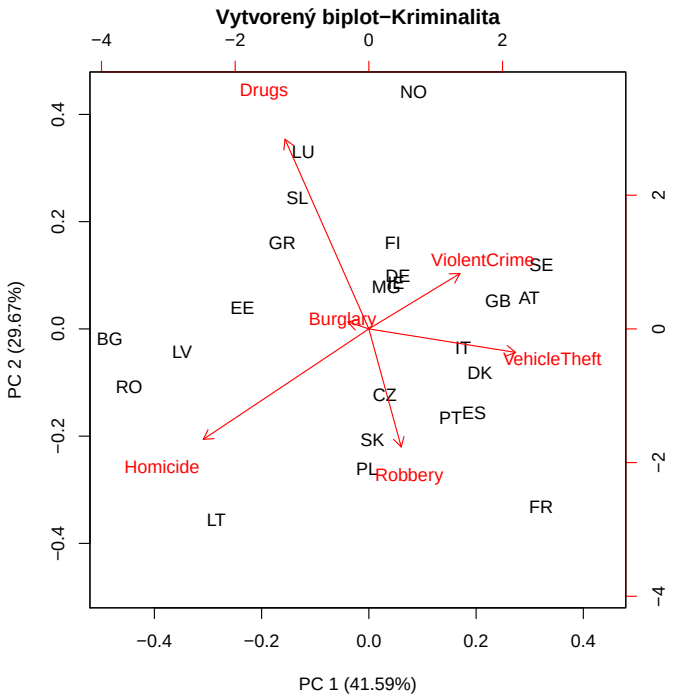
5.1.2 Vytvořený biplot

Nyní se podíváme na vytvořený biplot. V tomto případě dojde ve výše popsaném postupu pouze k jedné změně. Matice **G** a **H** jsou vytvořeny dle příslušného postupu v kapitole 3.5:

```
>G=(1/sqrt(nrow(X)-1))*L%*%K
>H=(sqrt(nrow(X)-1))*M
>rownames(H)=colnames(X)
>rownames(G)=rownames(X)
```

Ostatní části algoritmu zůstávají stejné. Tabulka vysvětlené variability sou-

boru je také stejná. Je to dáno tím, že zkoumáme variabilitu původní matice **Z** před tím, než ji rozdělíme pomocí singulárního rozkladu, a dále na matice **G** a **H**, kterými se kovarianční a vytvořený biplot liší. Výsledný biplot potom vypadá následovně.



Graf pro vytvořený biplot je v podstatě stejný jako ten pro biplot kovarianční. Došlo zde pouze k posunu měřítek jednotlivých os. Přesto se zde slovní popis řídí trochu jinými pravidly - uvedeno v Příloze A. Tento biplot je vhodný k zobrazení jednotlivých pozorování (států).

Výše zmíněný posun měřítek nám říká, že jednotlivé kriminální činy mají v každém grafu trochu jiný vliv na celkové uspořádání. Vzdálenosti mezi jednotlivými státy se nezměnily, došlo pouze k jejich celkovému posunu vzhledem k paprskům.

V případě vytvořeného biplotu aproximuje vzdálenost mezi pozorováními v grafu euklidovskou vzdálenost mezi jednotlivými státy. Interpretace je proto analogická jako v úvodu kovariančního biplotu.

Délky paprsků nám v tomto případě udávají kvalitu zobrazení jednotlivých složek kompozice. Nejkratší šipku má vloupání do domácnosti. Můžeme tedy říci, že ta je v rámci celého souboru nejhůře zobrazena. Paprsky ostatních proměnných jsou dostatečně dlouhé, a proto můžeme říci, že ostatní kriminální činy jsou zobrazeny kvalitně.

Protože se tento biplot zaměřuje spíše na pozorování, podíváme se ještě, jaký význam má jejich vertikální a horizontální uspořádání. Uspořádání států zleva doprava je dáno ekonomickou vyspělostí daných států od méně vyspělých (např. Rumunsko - RO, Bulharsko - BG) až po vyspělé ekonomiky jako je Velká Británie (GB) nebo Rakousko (AT). To souvisí také s faktem, že postupně tímto směrem přecházíme od těžkých kriminálních zločinů, jako je např. vražda, po ty méně závažné, třeba v podobě krádeže auta. Pokud si státy v tabulce procentuálního zastoupení jednotlivých kriminálních činů (Příloha F) seřadíme právě podle proporcí vražd, nejvyšší zastoupení mají zmíněné země, kdežto druhé vyjmenované jsou na posledním a předposledním místě. Uspořádání shora dolů je dáno na základě proporcí distribuce drog, kdy nahoře jsou státy tímto neduhem nejvíce postižené (Norsko - NO, Lucembursko - LU) a směrem dolů tento druh kriminality postupně klesá. Naopak uspořádání zdola nahoru ovlivňují loupeže od nejčetnějšího výskytu, např. v Polsku (PL), k menšímu.

5.1.3 Ilr biplot

Dalším druhem biplotu, který si na datovém souboru Kriminalita ukážeme, je ilr biplot. Je to z hlediska interpretace standardní biplot vytvořený pro ilr transformovaná data.

V úvodu tvorby tohoto biplotu v softwaru R postupujeme stejně jako u předchozích případů. Nastavíme si tedy příslušné knihovny, načteme a pojmenujeme datovou matici $\mathbf{X}$ a převedeme ji na kompoziční tvar dle postupu u kovariančního biplotu.

Nyní musíme určit postup, jakým budeme složky kompozice rozdělovat do

skupin dle postupného binárního dělení. Postupný průběh dělení je naznačen v níže uvedené tabulce a řídí se následujícími pravidly. První bilance x_1^* byla vytvořena na základě rozdělení dle závažnosti jednotlivých kriminálních činů podle nového Trestního zákoníku ([22], par. 14). V první skupině jsou uvedeny zločiny (ozn. $+1$), jejichž trestní sazba se nachází nad hranicí 5 let, druhou skupinu tvoří přečiny (-1) s trestní sazbou nižší než 5 roků. Bilance x_2^* rozděluje zločiny podle toho, zda jsou vztaženy vůči jiné osobě ($+1$) či vůči jejímu majetku (-1). x_3^* je logickým rozdělením zbývajících dvou prvků dané skupiny. Bilance x_4^* je rozdělena stejně jako x_2^*, tedy dělí přečiny s ohledem na osobu ($+1$) či majetek (-1). Poslední bilance x_5^* je opět pouze doplňujícím oddělením posledních dvou složek kompozice.

Bilance	Homicide	Violent crime	Robbery	Domestic burglary	Motor vehicle theft	Drug trafficking
x_1^*	$+1$	$+1$	$+1$	-1	-1	-1
x_2^*	$+1$	$+1$	-1	0	0	0
x_3^*	$+1$	-1	0	0	0	0
x_4^*	0	0	0	-1	-1	$+1$
x_5^*	0	0	0	$+1$	-1	0

V softwaru R tuto tabulku vyrobíme jako

```
>bin=cbind(c(1,1,1,-1,-1,-1),c(1,1,-1,0,0,0),c(1,-1,0,0,0,0),
c(0,0,0,-1,-1,1),c(0,0,0,1,-1,0)).
```

Nyní vytvoříme ilr transformovaná data dle výše uvedeného binárního dělení:

```
>base=gsi.buildilrBase(bin)
>Y=ilr(X,V=base)[1:24,].
```

Další postup je opět stejný jako u kovariančního biplotu - centrujeme tedy nově vytvořenou datovou matici $\mathbf{Y}$, provedeme singulární rozklad a následné rozdělení na matice $\mathbf{G}$ a $\mathbf{H}$. Jediným rozdílem je pojmenování řádků matice $\mathbf{H}$,

```
>rownames(H)=c("xs1","xs2","xs3","xs4","xs5"),
```

protože ilr souřadnice již nemůžeme (i nepřímo) ztotožnit s původními kompozičními složkami.

Podíl vysvětlené variability souboru je následující (a totožný se situací u prvních pěti hlavních komponent clr transformovaných dat):

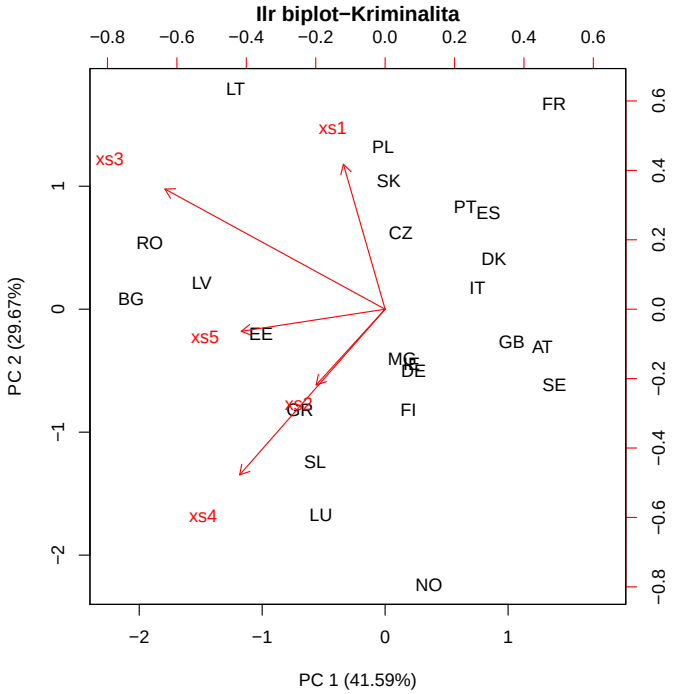
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1.0975731	0.9269876	0.6161308	0.5862384	0.33028911
Proportion of Variance	0.4159249	0.2966852	0.1310671	0.1186579	0.03766488
Cumulative Proportion	0.4159249	0.7126101	0.8436772	0.9623351	1.00000000

První komponenta tedy znovu vysvětluje 41.59 % a druhá 29.67 % celkové variability, jejich součet je tedy 71.26 % .

Příkazem

```
>biplot(x=G,y=H,xlab="PC 1 (41.59%)",ylab="PC 2 (29.67%)",
main="Ilr biplot-Kriminalita")
```

opět vykreslíme biplot.



S interpretací ilr biplotu bývají občas problémy. Velmi totiž závisí na tom, jakým způsobem jsme vytvořili příslušné bilance. Definovat postupné binární dělení tak, aby měl graf nějakou výpovědní hodnotu, je někdy obtížné. Pokud se nám to podaří, je interpretace pak již jednoduchá. Nevýhodou tohoto přístupu ke konstrukci kompozičního biplotu je fakt, že interpretace ilr souřadnic není vždy jednoznačná.

V našem biplotu můžeme vidět dva splývající paprsky x_2^* a x_4^*. Tuto skutečnost můžeme interpretovat tak, že rozdělení zločinů i přečinů na vztažené vůči osobě či majetku spolu úzce souvisí. Ovšem šipka x_2^* je méně než poloviční oproti x_4^*. To znamená, že na dané rozdělení má x_4^* větší vliv. Dále zde můžeme pozorovat přibližnou ortogonalitu paprsků x_2^* a x_3^*, x_3^* a x_4^*, x_1^* a x_5^*, která značí nízkou korelaci mezi příslušnými bilancemi. Naopak uvnitř dvojic paprsků x_1^* a x_2^*, x_1^* a x_4^* dochází k negativní korelaci. Tedy rozdělení kriminálních činů na zločiny a přečiny je negativně korelované s následným rozdělením jak zločinů, tak i přečinů, na vztažené vůči osobě a vůči majetku. Tuto skutečnost můžeme interpretovat slovy, že čím větší je v kompozici relativní zastoupení zločinů vůči přečinům, tím budou tyto události spíše vztaženy k majetku. Naopak, čím více jsou relativně zastoupeny přečiny na úkor zločinů, tím spíše budou delikty vztaženy k lidem.

Pokud se podíváme na uspořádání jednotlivých států v biplotu a porovnáme jej s kovariančním biplotem, vidíme, že tyto dva grafy mají úplně stejné rozložení. Liší se pouze otočením. Tedy státy, které byly dříve nahoře, jdou dole, a naopak. Interpretace tohoto postavení tedy vychází z dříve zmíněného v případě clr biplotu.

5.1.4 Logratio biplot

Poslední biplot, který bude představen u tohoto příkladu, nese označení logratio biplot. Z jeho názvu je patrné, že bude zobrazovat vztahy mezi logaritmy podílů jednotlivých kriminálních činů.

Nastavení knihoven, načtení a ostatní práce s původní datovou maticí je opět stejná jako u příkladů výše. Další průběh odpovídá postupu uvedenému v kapitole

3.3. Označme si

```
>n=24
```

počet pozorování,

```
>D=6
```

počet složek kompozice. Dále vytvoříme matice $\mathbf{J}, \mathbf{B}, \mathbf{C}_n$ a $\mathbf{C}_D$, zmíněné v kapitole 3.3:

```
>J=log(X)
```

```
>b1=matrix(c(1,1,1,1,1,-1,0,0,0,0,0,-1,0,0,0,0,0,-1,0,0,
0,0,0,-1,0,0,0,0,0,-1),D,D-1,TRUE)
>b2=matrix(c(-1,0,0,0,0,1,1,1,1,1,0,-1,0,0,0,0,0,-1,0,0,
0,0,0,-1,0,0,0,0,0,-1),D,D-1,TRUE)
>b3=matrix(c(-1,0,0,0,0,0,-1,0,0,0,1,1,1,1,1,0,0,-1,0,0,
0,0,0,-1,0,0,0,0,0,-1),D,D-1,TRUE)
>b4=matrix(c(-1,0,0,0,0,0,-1,0,0,0,0,0,-1,0,0,1,1,1,1,1,
0,0,0,-1,0,0,0,0,0,-1),D,D-1,TRUE)
>b5=matrix(c(-1,0,0,0,0,0,-1,0,0,0,0,0,-1,0,0,0,0,0,-1,0,
1,1,1,1,1,0,0,0,0,-1),D,D-1,TRUE)
>b6=matrix(c(-1,0,0,0,0,0,-1,0,0,0,0,0,-1,0,0,0,0,0,-1,0,
0,0,0,0,-1,1,1,1,1,1),D,D-1,TRUE)
>B=cbind(b1,b2,b3,b4,b5,b6)
```

```
>cn1=diag(n)
>cn2=matrix(1, nr = n, nc = 1)
>Cn=cn1-1/n*cn2%*%t(cn2)
```

```
>cD1=diag(D)
>cD2=matrix(1, nr = D, nc = 1)
>CD=cD1-1/D*cD2%*%t(cD2).
```

Dále vytvoříme matici $\mathbf{Z}$ (vlastně se jedná o matici clr transformovaných kompozic) a její singulární rozklad

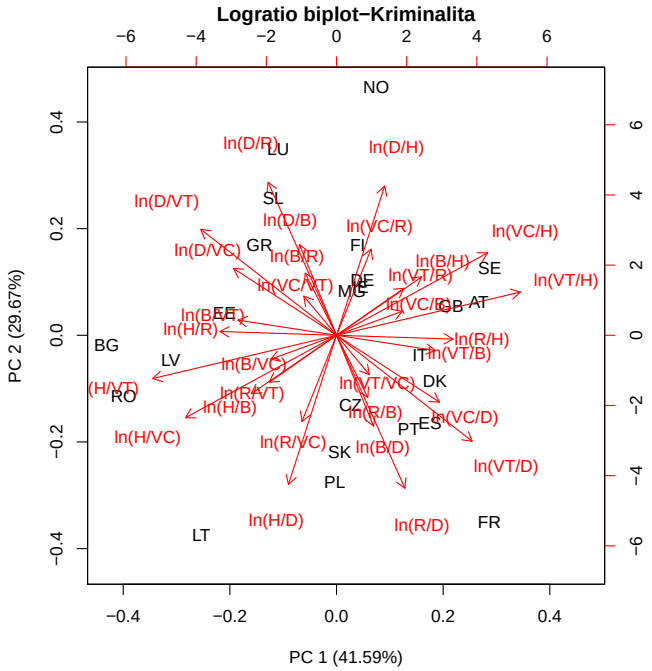
```
>Z=Cn%*%J%*%CD
>L=svd(Z)$u[,c(1,2)]
>K=diag(svd(Z)$d[1:2])
>M=svd(Z)$v[,c(1,2)]
```

Posledním krokem je tvorba matic **G** a **H** dle (3.14)

```
>G=L, >H=t(B)%*%M%*%K
>rownames(H)=c("ln(H/VC)","ln(H/R)","ln(H/B)","ln(H/VT)","ln(H/D)",
"ln(VC/H)","ln(VC/R)","ln(VC/B)","ln(VC/VT)","ln(VC/D)","ln(R/H)",
"ln(R/VC)","ln(R/B)","ln(R/VT)","ln(R/D)","ln(B/H)","ln(B/VC)",
"ln(B/R)","ln(B/VT)","ln(B/D)","ln(VT/H)","ln(VT/VC)","ln(VT/R)",
"ln(VT/B)","ln(VT/D)","ln(D/H)","ln(D/VC)","ln(D/R)","ln(D/B)",
"ln(D/VT)")
>rownames(G)=rownames(X)
```

Podíl vysvětlené variability souboru pomocí jednotlivých hlavních komponent je totožný jako v případě kovariančního clr biplotu.

Nakonec vykreslíme biplot.



Grafické znázornění logratio biplotu se nám může na první pohled zdát trochu chaotické. Je to dáno zejména velkým počtem paprsků. Interpretace tohoto grafu je obtížná.

První skutečností, které si zde můžeme všimnout, je silně negativně korelovaná odpovídající si dvojice paprsků logaritmů inverzních podílů (tedy např. $\ln\left(\frac{x_1}{x_2}\right)$ a $\ln\left(\frac{x_2}{x_1}\right)$). Je to dáno způsobem konstrukce tohoto biplotu.

Dále si můžeme všimnout, že v případě proměnné drogy jdou všechny paprsky s touto proměnnou v čitateli stejným směrem jako původní paprsek drog (D). Je to způsobeno faktem, že je tento paprsek v kovariančním clr biplotu nejdelší, a proto „přetahuje“ ostatní na svou stranu. Naopak paprsky s proměnnou vloupání do domácnosti (B) v čitateli se nechají „táhnout“ silnějšími proměnnými. Například šipka $\ln\left(\frac{B}{H}\right)$ je úplně mimo oblast, kde se vyskytoval původní paprsek, na místě, které určuje proměnná vražda (H).

Nejdelší paprsky mají ty logaritmy podílů, ve kterých se vyskytuje jedna z proměnných s původně nejdelšími šipkami - drogy a vražda.

Nakonec se můžeme podívat, že je v grafu několik paprsků, jejichž vrcholy jsou velmi blízko u sebe (např. $\ln\left(\frac{R}{H}\right)$ a $\ln\left(\frac{VT}{B}\right)$) a tedy jejich spojnice je velmi krátká. To značí, silný vztah mezi těmito dvěma logaritmy podílů.

Uspořádání jednotlivých států je opět stejné (až na rotaci) jako v prvním příkladu.

5.2 Příčiny úmrtí v důsledku nemoci

Druhý datový soubor nese název Příčiny úmrtí v důsledku nemoci. Data jsou převzata z databáze Eurostatu [7]. Jak již bylo řečeno u předchozího příkladu, data byla zvolena pro rok 2008 a bylo vybráno 24 zemí, jejichž zkratky používané v příkladu jsou uvedeny v Příloze G. Dále je zde uvedeno i číselné označení jednotlivých zemí.

Původní databáze byla značně rozsáhlá a byl zde uveden velký počet nemocí, z nichž některé měly pouze malé podíly na celkové nemocnosti v dané zemi. Proto bylo vybráno 10 nejčetnějších chorob. Jejich popisy v grafu jsou opět v anglič-

tině a jsou následující: malignant neoplasms = zhoubné novotvary, endocrine, nutritional and metabolic diseases = endokrinní, vyživovací a metabolické choroby, diabetes mellitus = cukrovka, mental and behavioural disorders = duševní poruchy a poruchy chování, diseases of the nervous system = poruchy nervového systému, diseases of the circulatory system = poruchy oběhového systému, ischaemic heart diseases = ischemická choroba srdeční, cerebrovascular diseases = cerebrovaskulární nemoci (týkající se mozkových cév), diseases of the respiratory system = nemoci dýchací soustavy, pneumonia = zápal plic. Jako zdroj pro porozumění názvům určitých nemocí lze doporučit literaturu [23].

Úkolem je zjistit, jaký je vzájemný vztah mezi jednotlivými státy a také mezi jednotlivými nemocemi.

Nejdříve si data zobrazíme pomocí kovariančního (clr) biplotu. Postup je stejný jako u příkladu s názvem Kriminalita, a proto jej zde nebudeme celý znovu uvádět. Liší se pouze názvy zdrojových souborů a pojmenování jednotlivých složek kompozice

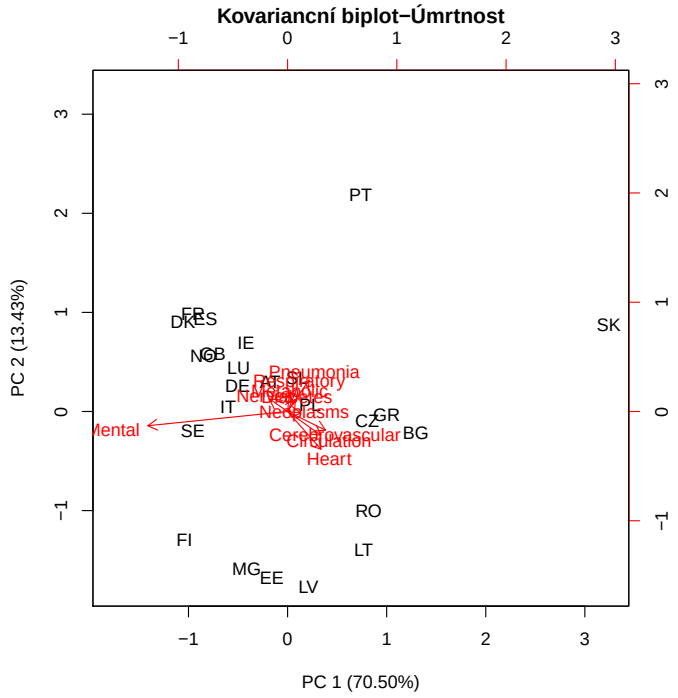
```
>colnames(X)=c("Neoplasms","Metabolic","Diabetes","Mental","Nerves",
"Circulation","Heart","Cerebrovascular","Respiratory","Pneumonia").
```

Datový soubor převedeme na procentuální reprezentaci stejným způsobem jako výše. Takto vytvořený soubor je uveden v Příloze H.

Tabulka charakteristik jednotlivých hlavních komponent vypadá následovně:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1.7466664	0.7624151	0.58819899	0.41434979	0.31737665
Proportion of Variance	0.7049725	0.1343183	0.07994675	0.03967222	0.02327567
Cumulative Proportion	0.7049725	0.8392908	0.91923756	0.95890978	0.98218545
	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
Standard deviation	0.22730454	0.134745644	0.072965511	0.0441207850	4.482615e-09
Proportion of Variance	0.01193901	0.004195481	0.001230233	0.0004498199	4.643175e-18
Cumulative Proportion	0.99412447	0.998319947	0.999550180	1.0000000000	1.0000000000

Zde vidíme, že první hlavní komponenta vysvětluje 70.50 % a druhá dalších 13.43 % celkové variability, což dá v součtu 83.93 %. Biplot tedy bude velmi dobře odrážet skutečnou strukturu mnohorozměrného datového souboru.



Výsledný biplot je z hlediska interpretace paprsků trochu nepřehledný. Později si ukážeme, jak tento problém řešit. Nejdříve si ale popíšeme alespoň ty vlastnosti, které jsou zde viditelné.

Z pohledu vzdáleností jednotlivých pozorování můžeme opět říct, že čím jsou jednotlivé státy v grafu blíže u sebe, tím jsou si podobnější. Můžeme zde například vidět blízký shluk států Dánsko (DK), Francie (FR) a Španělsko (ES). Tyto státy jsou si velmi podobné ve výskytu všech uvedených nemocí. Může to být dáno jejich vyspělostí a odtud plynoucí podobnou zdravotní péčí. Další takovou skupinu (odlišnou od předchozí) tvoří Česká republika (CZ), Řecko (GR) a Bulharsko (BG). Ty mají všechny velmi malý výskyt mentálních poruch. Další viditelnou skupinu tvoří například ugrofinské národy z Maďarska (MG), Estonska (EE) a Lotyšska (LV). Pokud se podíváme na graf velmi dobře, uvidíme, že téměř

uprostřed grafu se nachází Polsko (PL). To má při porovnání s původními daty opravdu u všech nemocí (kromě nemocí oběhové soustavy) průměrné hodnoty.

Další výraznou charakteristikou tohoto biplotu jsou odlehlé hodnoty Slovensko (SK) a Portugalsko (PT). Díky těmto pozorováním se nám v grafu vytvořil nepřehledný shluk paprsků. Portugalsko má nejvyšší výskyt cukrovky a druhý nejvyšší u metabolických poruch a zápalu plic. Na druhou stranu má nejnižší přítomnost ischemické choroby srdeční. Tyto extrémy zřejmě způsobily jeho oddělení se od ostatních. Vychýlení Slovenska může být způsobeno velmi nízkou proporcí u mentálních poruch. Hodnota složky kompozice u tohoto státu je 46-krát nižší než u druhého v pořadí (Bulharsko) a 3938-krát nižší než maximum (Dánsko). Tato skutečnost se také projevuje tak, že SK je velmi daleko od vrcholu paprsku reprezentujícího danou chorobu.

Horizontální uspořádání států ovlivňuje celá řada nemocí. Při postupování zdola nahoru postupně roste výskyt zhoubných novotvarů, metabolických chorob, cukrovky, poruch nervového systému, nemocí dýchací soustavy a zápalu plic. Při tomto postupu naopak klesá zastoupení poruch oběhového systému a ischemické choroby srdeční. Při přechodu shora dolů je postup opačný. Vertikální rozložení je dáno výskytem mentálních poruch. Při průběhu zleva, doprava hodnoty klesají. Vlevo se nachází státy jako je Dánsko, Francie či Finsko (FI), které bychom mohli označit za vyspělé v oblasti zdravotnictví. V pravé části biplotu se naopak nachází např. Slovensko, Řecko nebo Bulharsko, tedy státy s horšími podmínkami v této oblasti. Skutečnost, že ve státech vyskytujících se v levé části grafu umírá větší podíl pacientů s mentálními chorobami, by mohla být proto matoucí. Zmíněný fakt může být způsoben také tím, že ve státech v levé části grafu je obecně lepší zdravotní péče, což značně přispívá k tomu, že se obyvatelé dožívají vyššího věku (viz zdroj [18]), ve kterém jsou pak náchylnější k duševním nemocem, v důsledku kterých umírají (např. Alzheimerova choroba).

Zkoumání paprsků není v tomto grafu jednoduché. Můžeme pouze s jistotou říci, že nejdelší šipku mají mentální poruchy. Ty tedy značně ovlivňují celý soubor. Potom zde můžeme pozorovat rozdělení zbylých nemocí do dvou skupin

s navzájem stabilními podíly v rámci datového souboru. Například tu menší z nich tvoří poruchy oběhového systému, ischemická choroba srdeční a cerebrovaskulární nemoci.

Jak již bylo řečeno v popisu biplotu, v tomto grafu se nachází viditelné odlehlé hodnoty (státy SK a PT), které ovlivňují podobu celého zobrazení. Proto zkusíme na data aplikovat robustní podobu biplotu, která vliv těchto pozorování potlačí.

Nastavení knihoven, načtení, pojmenování dat a převod na kompoziční tvar je stejný jako výše. Další příkazy vypadají následovně:

```
>resRob=pcaCoDa(as.data.frame(X),method ="robust")
```

nám provede převod původní datové matice na robustní. Předpis

```
>resRob
```

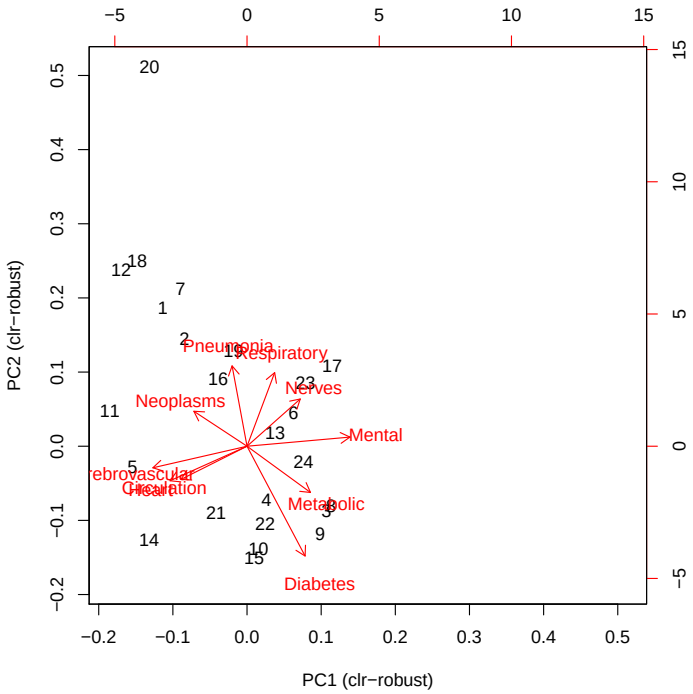
nám poté vypíše procenta vysvětlené variability pomocí klasických komponent pro data po clr transformaci:

```
                         Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
Proportion of Variance 0.6264544 0.1911580 0.1126522 0.0333043 0.0223270
Cumulative Proportion  0.6264544 0.8176124 0.9302646 0.9635689 0.9858959
                         Comp.6     Comp.7     Comp.8     Comp.9
Proportion of Variance 0.0077756 0.0041890 0.0018681 0.0002714
Cumulative Proportion  0.9936715 0.9978605 0.9997286 1.0000000
```

Vzhledem k robustním odhadům polohy a variability vyjdou vždy procenta vysvětlené variability trochu jiná než v případě klasického (nerobustního) biplotu. To také odpovídá poněkud odlišné konfiguraci pozorování a proměnných ve výsledném biplotu.
Poslední příkaz

```
>plot(resRob)
```

vykreslí příslušný biplot.



V příkazu `pcaCoDa()` je uložen již kompletně vytvořený algoritmus pro převod na robustní datový soubor a tvorbu příslušného biplotu. Tento algoritmus ovšem neumí převádět názvy pozorování, proto se budeme muset při interpretaci biplotu řídit číselným označením jednotlivých států z Přílohy H.

Jak vidíme, v tomto grafu je již zobrazení paprsků mnohem zřetelnější. Paprsek mentálních poruch již nepřevyšuje délky ostatních paprsků. Všechny paprsky jsou přibližně stejně dlouhé, tedy i stejně významné pro uspořádání grafu. Jsou zde zřetelné tři skupiny nemocí. První z nich je stejná jako v předchozím případě. Zbylé nemoci z druhé skupiny se nyní rozdělily do dvou dalších. Menší z nich tvoří metabolické poruchy a cukrovka, větší je zastoupena zhoubnými novotvary, zápalem plic, poruchami nervové a dýchací soustavy. Mentální choroby zůstávají opět samotné.

Zkrácení paprsku mentálních poruch bylo způsobeno potlačením vlivu odlehlých hodnot (např. Portugalsko(17)). Díky této skutečnosti zde již není tak výrazný vliv duševních chorob, jak bylo popsáno u předchozího biplotu. Rozdíl mezi výše zmíněnými skupinami je zde ovšem pořád patrný. Skupina vyspělých států,

dříve umístěných v levé části grafu - Dánsko (3), Francie (9) a Finsko (21), je zde zachována. Protože ale nyní míří šipka mentálních poruch zleva doprava, nachází se tato skupina vpravo dole (u vrcholu této šipky - zejména Dánsko a Francie). Naopak státy Řecko (7) a Bulharsko (1) jsou opět dále od vrcholu příslušné šipky - vlevo nahoře. Slovensko (20) má v tomto grafu opět postavení odlehlého pozorování, pořád je ovšem nejdále od vrcholu paprsku mentálních poruch. Interpretace týkající se duševních chorob z předchozího grafu by proto mohla platit i zde.

Nejkratší spojnici, a tedy i největší stabilitu odpovídajícího podílu, mají srdeční a oběhové choroby. Tento fakt vychází také z biologického propojení obou systémů.

Dříve odlehlé pozorování Portugalsko (17) se posunulo blíže středu grafu. Shluk států Dánsko (3), Francie (9) a Španělsko (8) zůstal zachován. To stejné platí i pro trojici Česká republika (2), Řecko (7) a Bulharsko (1). U poslední výše zmíněné skupiny, Maďarska (14), Estonska (5) a Lotyšska (11), došlo k jejich vzájemnému odstupu, ale pořád jsou si vzájemně nejblíže oproti ostatním státům. Nejblíže středu je nyní Lucembursko (13), původní středové Polsko (16) se posunulo blíže k paprsku zápalu plic.

Horizontální uspořádání států nyní zhruba koresponduje s vertikálním uspořádáním předchozího grafu. Tentokrát jsou nemocemi, které nejvíce ovlivňují postup zdola nahoru, cukrovka a metabolické poruchy (od největšího k nejmenšímu). Vertikální seřazení je ovlivněno dříve zmíněnou dvojicí chorob srdce a oběhové soustavy. Jejich (relativní) výskyt postupem zleva doprava klesá.

V případech, jako byl tento, kdy mohou odlehlé hodnoty značně zkreslit interpretaci klasického biplotu, se zdá být robustní biplot velmi dobrým nástrojem.

Závěr

Když jsem začínala psát tuto práci, myslela jsem si, že nebudu mít žádné nové informace, které by se lišily od těch, uvedených v mé bakalářské práci. To jsem se ale velmi zmýlila.

Již v úvodu, při poznávání kompozičních dat, jsem byla překvapená, jaké skýtají možnosti. S daty ve tvaru kompozic - nejčastěji procentuálního zastoupení - se člověk setkává poměrně často. Až nyní mě ovšem napadlo o nich přemýšlet v takovýchto širších souvislostech. Také mě překvapilo, jak je teorie spojená s kompozičními daty celkem mladá.

Dalším překvapením pro mě bylo zjištění, kolik druhů biplotů pro kompoziční data existuje. Zejména při řešení praktických příkladů jsem si uvědomila, jak lze jeden datový soubor znázornit a popsat tolika různými způsoby. Každý biplot je něčím zajímavý. Zajímavá pro mě byla i interpretace datových souborů. Protože jsou to reálná data, doufám, že mé výsledky mohou vypovídat něco málo i o stavu evropské společnosti.

Když porovnám výše uvedené grafické výstupy, dle mého názoru je z kompozičních biplotů nejlepší kovarianční clr biplot. Myslím si, že z něj můžeme vyčíst nejvíce informací. Je totiž dostatečně přehledný, ale zase ne moc stručný. Navíc pravidla pro jeho výklad jsou nejpodrobnější. Tímto tvrzením samozřejmě neshazuji ostatní biploty. Myslím, že každý vypovídá „svou část příběhu“.

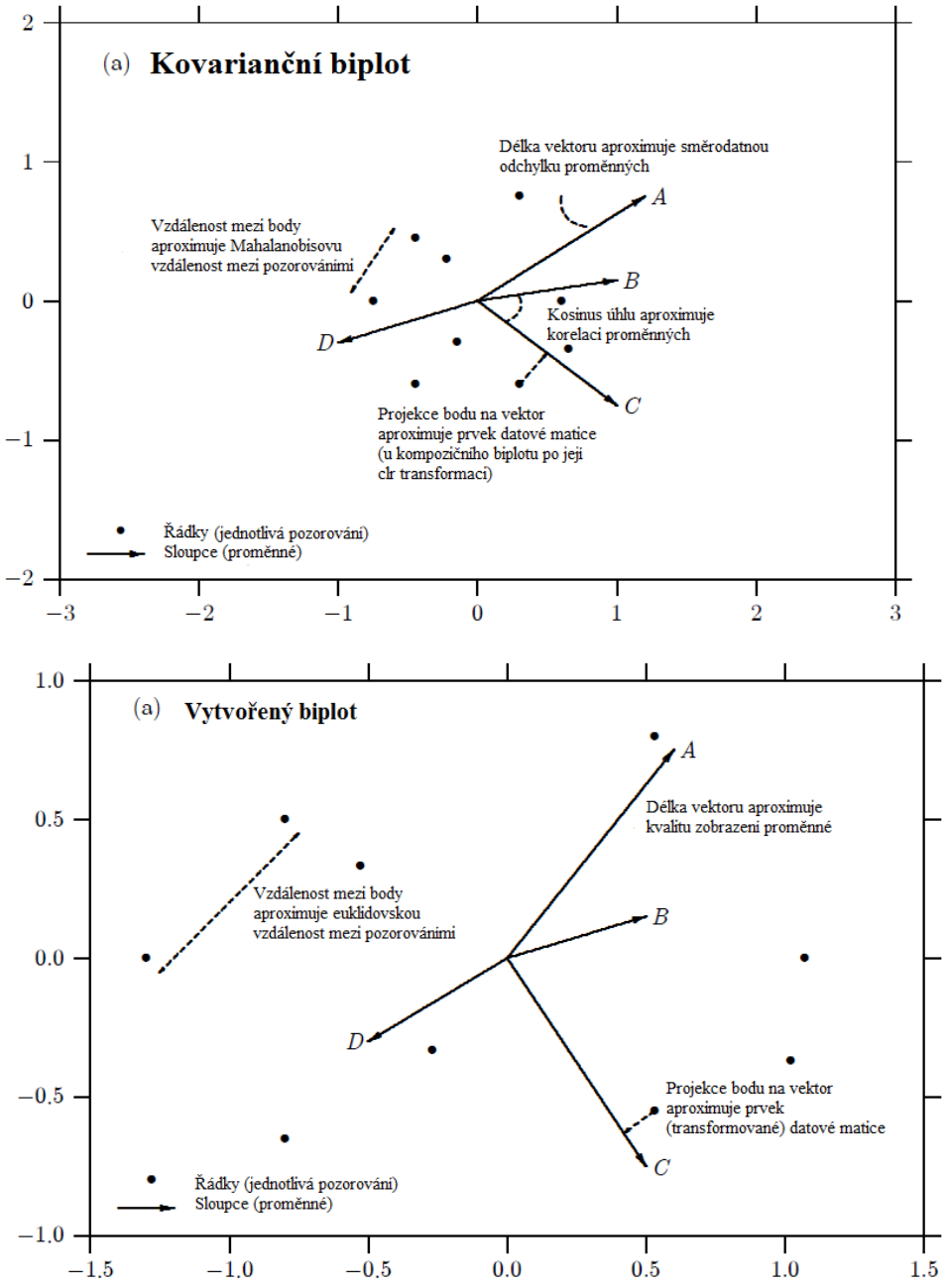
Myslím, že cíl práce - seznámit sebe i čtenáře této práce s kompozičními daty a jejich biploty - byl splněn. Určitě je ale ještě velký prostor pro doplňování informací a pátrání po podrobnostech. Například by mohlo být dále rozvinuto praktické využití robustních biplotů. Také by mohlo v budoucnu proběhnout porovnání přínosnosti jednotlivých druhů biplotů.

Těší mě, že jsem zvolila právě toto téma. Jak jsem již napsala u standardního biplotu v mé bakalářské práci, i tento grafický rozbor dat mi přijde jako velmi praktický a použitelný v mnoha oblastech. Sama se začínám přesvědčovat, jak lze tyto postupy efektivně využít například v chemii nebo lékařství. Určitě se tomuto tématu budu ráda věnovat i v budoucnu, protože si myslím, že právě tam patří.

Přílohy

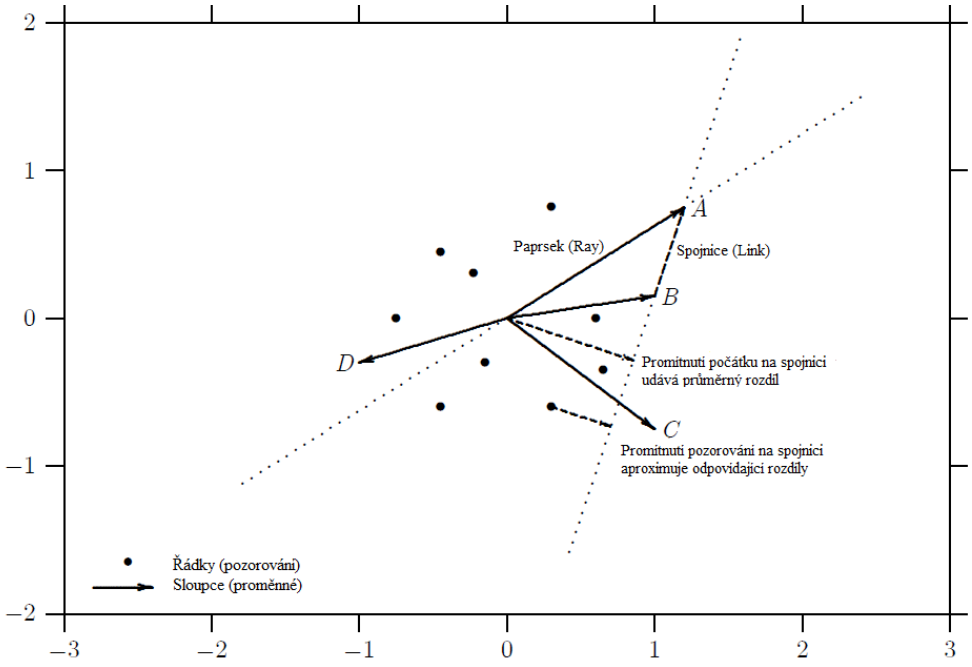
Příloha A

Grafické vyjádření kovariančního a vytvořeného biplotu (převzato z [2]).



Příloha B

Osy biplotu skrze paprsky a spojnice (převzato z [2]).



Příloha C

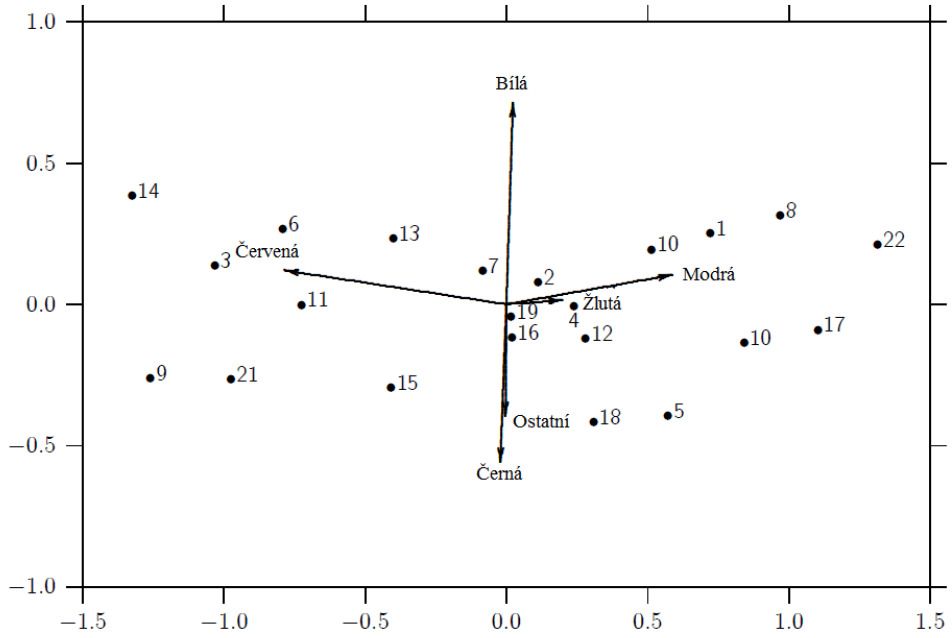
Data - proporcionální zastoupení 6 barev na 22 obrazech

Malba	Černá	Bílá	Modrá	Červená	Žlutá	Ostatní
1	0.125	0.243	0.153	0.031	0.181	0.266
2	0.143	0.224	0.111	0.051	0.159	0.313
3	0.147	0.231	0.058	0.129	0.133	0.303
4	0.164	0.209	0.120	0.047	0.178	0.282
5	0.197	0.151	0.132	0.033	0.188	0.299
6	0.157	0.256	0.072	0.116	0.153	0.246
7	0.153	0.232	0.101	0.062	0.170	0.282
8	0.115	0.249	0.176	0.025	0.176	0.259
9	0.178	0.167	0.048	0.143	0.118	0.347
10	0.164	0.183	0.158	0.027	0.186	0.281
11	0.175	0.211	0.070	0.104	0.157	0.283
12	0.168	0.192	0.120	0.044	0.171	0.305
13	0.155	0.251	0.091	0.085	0.161	0.257
14	0.126	0.273	0.045	0.156	0.131	0.269
15	0.199	0.170	0.080	0.076	0.158	0.318
16	0.163	0.196	0.107	0.054	0.144	0.335
17	0.136	0.185	0.162	0.020	0.193	0.304
18	0.184	0.152	0.110	0.039	0.165	0.350
19	0.169	0.207	0.111	0.057	0.156	0.300
20	0.146	0.240	0.141	0.038	0.184	0.250
21	0.200	0.172	0.059	0.120	0.136	0.313
22	0.135	0.225	0.217	0.019	0.187	0.217

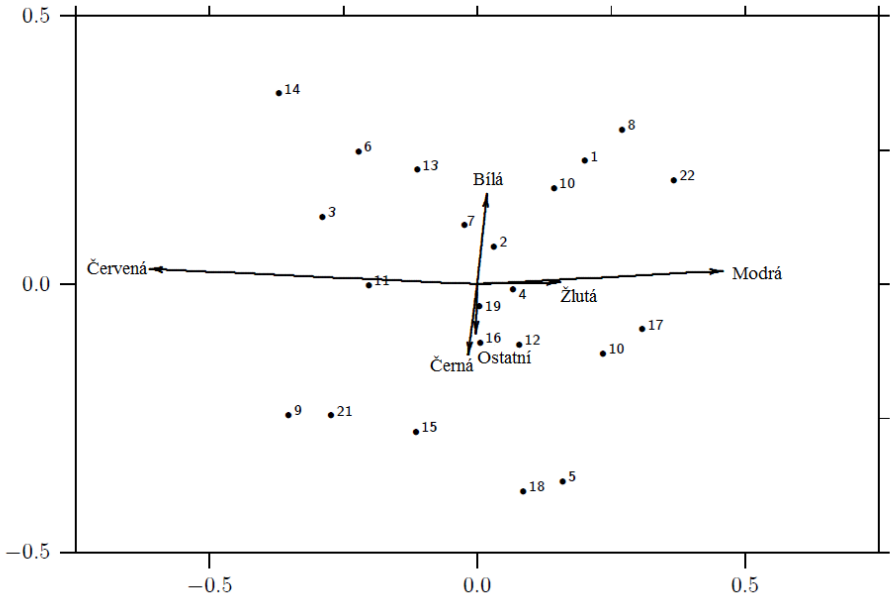
Příloha D

Grafická interpretace vzorového příkladu (převzato z [2]).

Vytvořený biplot - uchovává vzdálenosti mezi řádky (pozorováními - obrazy).



Kovarianční biplot - zachovává kovarianční strukturu clr transformovaných dat.



Příloha E

Kriminalita - původní datový soubor

		Homicide	Violent crime	Robbery	Domestic burglary	Motor vehicle theft	Drug trafficking
Bulgaria	BG	172	8 538	2 868	19 980	430	2 857
Czech Republic	CZ	202	18 187	4 641	9 111	18 893	2 812
Denmark	DK	79	24 928	10 747	43 974	26 804	3 237
Germany	DE	656	210 885	49 913	108 284	89 036	55 905
Estonia	EE	84	9 082	909	3 321	1 035	1 558
Ireland	IE	89	19 152	2 299	24 864	14 307	4 028
Greece	GR	118	11 220	3 097	44 150	7 834	9 852
Spain	ES	408	116 567	93 186	82 135	96 314	14 574
France	FR	839	331 778	106 633	166 250	211 484	6 128
Italy	IT	654	146 598	64 535	153 080	229 961	34 082
Latvia	LV	119	1 928	1 441	3 538	1 868	2 512
Lithuania	LT	304	4 372	3 452	6 076	2 553	793
Luxembourg	LU	7	3 197	260	1 731	343	1 343
Hungary	MG	147	33 035	3 128	19 239	16 539	5 464
Austria	AT	46	129 613	4 786	18 648	9 049	1 980
Poland	PL	460	52 122	21 085	31 481	17 669	3 317
Portugal	PT	124	24 516	20 856	29 663	25 274	3 730
Romania	RO	493	6 842	2 464	10 285	2 355	3 621
Slovenia	SL	11	2 638	386	2 282	582	1 434
Slovakia	SK	94	9 030	1 371	2 118	4 135	524
Finland	FI	132	42 035	1 696	5 978	13 804	5 659
Sweden	SE	82	108 448	8 909	18 176	44 717	7 797
Great Britain	GB	785	1 094 066	84 362	309 001	161 847	40 763
Norway	NO	34	23 848	1 598	8 125	11 901	17 547

Příloha F

Kriminalita - proporcionální datový soubor

		Homicide	Violent crime	Robbery	Domestic burglary	Motor vehicle theft	Drug trafficking
Bulgaria	BG	0.49361458	24.50280	8.230736	57.339647	1.234036	8.1991677
Czech Republic	CZ	0.37514393	33.77595	8.619025	16.920477	35.087100	5.2223006
Denmark	DK	0.07196932	22.70951	9.790560	40.060491	24.418552	2.9489200
Germany	DE	0.12745809	40.97408	9.697889	21.039133	17.299326	10.8621102
Estonia	EE	0.52536119	56.80155	5.685159	20.770530	6.473200	9.7441991
Ireland	IE	0.13747509	29.58340	3.551182	38.406525	22.099507	6.2219064
Greece	GR	0.15471149	14.71070	4.060521	57.885697	10.271270	12.9170982
Spain	ES	0.10119449	28.91161	23.112524	20.371592	23.888349	3.6147268
France	FR	0.10193024	40.30776	12.954859	20.197737	25.693223	0.7444916
Italy	IT	0.10398944	23.30985	10.261405	24.340526	36.565009	5.4192174
Latvia	LV	1.04331054	16.90338	12.633702	31.018762	16.377345	22.0234964
Lithuania	LT	1.73219373	24.91168	19.669516	34.621083	14.547009	4.5185185
Luxembourg	LU	0.10172940	46.46127	3.778521	25.156227	4.984741	19.5175120
Hungary	MG	0.18955024	42.59723	4.033423	24.807871	21.326336	7.0455952
Austria	AT	0.02802793	78.97357	2.916123	11.362279	5.513581	1.2064196
Poland	PL	0.36469152	41.32272	16.716349	24.958378	14.008118	2.6297430
Portugal	PT	0.11904419	23.53619	20.022465	28.477482	24.263894	3.5809260
Romania	RO	1.89178818	26.25480	9.455104	39.466616	9.036838	13.8948580
Slovenia	SL	0.15000682	35.97436	5.263876	31.119596	7.936724	19.5554343
Slovakia	SK	0.54423344	52.28115	7.937703	12.262622	23.940482	3.0338119
Finland	FI	0.19046520	60.65306	2.447189	8.625765	19.918042	8.1654739
Sweden	SE	0.04358711	57.64555	4.735580	9.661456	23.769328	4.1444966
Great Britain	GB	0.04642707	64.70608	4.989402	18.275172	9.572078	2.4108364
Norway	NO	0.05392289	37.82215	2.534376	12.885985	18.874598	27.8289693

Příloha G

Příčiny úmrtí v důsledku nemoci - původní datový soubor

			Malignant neoplasms	Endocrine, nutritional and metabolic diseases	Diabetes mellitus	Mental, behavioural disorders	Diseases of the nervous system
1	Bulgaria	BG	18 017	2 181	2 120	90	1 027
2	Czech Republic	CZ	27 571	2 211	1 979	235	1 279
3	Denmark	DK	15 145	1 743	1 281	3 061	1 629
4	Germany	DE	216 002	27 331	22 330	18 850	19 841
5	Estonia	EE	3 543	262	224	192	234
6	Ireland	IE	8 203	640	468	564	887
7	Greece	GR	27 379	1 591	1 242	119	1 371
8	Spain	ES	100 675	12 337	10 153	13 040	17 496
9	France	FR	153 975	20 178	11 713	17 680	31 071
10	Italy	IT	164 847	24 999	20 233	13 915	21 502
11	Latvia	LV	5 833	445	402	205	333
12	Lithuania	LT	8 266	316	285	98	608
13	Luxembourg	LU	932	76	51	64	127
14	Hungary	MG	32 111	3 253	2 865	2 510	1 753
15	Austria	AT	19 780	4 398	3 385	825	2 347
16	Poland	PL	93 060	7 027	6 599	2 077	4 955
17	Portugal	PT	24 033	5 126	4 278	209	2 693
18	Romania	RO	46 086	2 278	2 206	593	2 034
19	Slovenia	SL	5 679	316	274	129	276
20	Slovakia	SK	11 891	708	625	1	722
21	Finland	FI	10 961	667	540	2 369	4 082
22	Sweden	SE	21 751	2 496	1 999	4 961	3 205
23	Great Britain	GB	157 032	8 695	6 450	22 374	19 789
24	Norway	NO	10 632	1 005	713	1 749	1 509

			Diseases of the circulatory system	Ischaemic heart diseases	Cerebrovascular diseases	Diseases of the respiratory system	Pneumonia
1	Bulgaria	BG	71 492	14 425	22 440	4 466	1 768
2	Czech Republic	CZ	52 280	25 844	11 685	5 736	2 859
3	Denmark	DK	15 031	5 570	3 912	5 622	1 923
4	Germany	DE	356 729	134 822	63 127	59 049	21 839
5	Estonia	EE	9 074	4 588	1 570	489	143
6	Ireland	IE	9 883	5 188	2 116	3 435	1 303
7	Greece	GR	49 214	11 624	16 064	10 239	935
8	Spain	ES	122 793	35 928	31 833	44 200	9 191
9	France	FR	149 541	38 306	33 162	33 259	10 808
10	Italy	IT	225 588	75 514	63 617	37 771	6 905
11	Latvia	LV	16 516	8 638	4 894	725	335
12	Lithuania	LT	23 623	14 633	5 843	1 684	729
13	Luxembourg	LU	1 307	378	303	272	85
14	Hungary	MG	64 749	32 828	13 996	6 231	693
15	Austria	AT	32 294	14 453	5 358	4 130	1 181
16	Poland	PL	172 943	48 909	37 248	19 297	9 128
17	Portugal	PT	33 811	7 784	14 583	11 580	5 145
18	Romania	RO	153 137	52 534	48 582	12 310	5 314
19	Slovenia	SL	7 225	2 004	1 943	1 142	550
20	Slovakia	SK	28 502	17 226	6 170	2 981	1 838
21	Finland	FI	20 281	11 761	4 246	1 980	457
22	Sweden	SE	37 466	16 488	8 364	5 848	2 257
23	Great Britain	GB	190 856	88 227	53 143	81 322	32 308
24	Norway	NO	14 135	5 636	3 570	4 118	1 641

Příloha H

Příčiny úmrtí v důsledku nemoci - proporcionální datový soubor

			Malignant neoplasms	Endocrine, nutritional and metabolic diseases	Diabetes mellitus	Mental, behavioural disorders	Diseases of the nervous system
1	Bulgaria	BG	13.05334	1.5801371	1.5359425	0.065205106	0.7440627
2	Czech Republic	CZ	20.93804	1.6790832	1.5028972	0.178464296	0.9713014
3	Denmark	DK	27.57798	3.1738806	2.3326110	5.573866016	2.9662946
4	Germany	DE	22.98089	2.9078007	2.3757341	2.005489829	2.1109243
5	Estonia	EE	17.43688	1.2894335	1.1024165	0.944928392	1.1516315
6	Ireland	IE	25.09560	1.9579649	1.4317619	1.725456604	2.7136170
7	Greece	GR	22.85812	1.3282907	1.0369183	0.099350465	1.1446175
8	Spain	ES	25.31774	3.1025083	2.5532760	3.279298673	4.3998934
9	France	FR	30.81392	4.0380794	2.3440392	3.538172438	6.2180179
10	Italy	IT	25.17167	3.8172765	3.0895218	2.124781070	3.2832945
11	Latvia	LV	15.21943	1.1610917	1.0488963	0.534884935	0.8688619
12	Lithuania	LT	14.73834	0.5634305	0.5081573	0.174734778	1.0840688
13	Luxembourg	LU	25.92490	2.1140473	1.4186370	1.780250348	3.5326843
14	Hungary	MG	19.94608	2.0206350	1.7796247	1.559112734	1.0888943
15	Austria	AT	22.43877	4.9891663	3.8400018	0.935894091	2.6624769
16	Poland	PL	23.19293	1.7513078	1.6446393	0.517641429	1.2349125
17	Portugal	PT	21.99978	4.6923345	3.9160762	0.191318357	2.4651691
18	Romania	RO	14.17708	0.7007635	0.6786147	0.182420003	0.6257037
19	Slovenia	SL	29.06643	1.6173610	1.4023953	0.660251817	1.4126318
20	Slovakia	SK	16.82752	1.0019246	0.8844673	0.001415148	1.0217367
21	Finland	FI	19.11447	1.1631557	0.9416853	4.131208147	7.1184431
22	Sweden	SE	20.74784	2.3808842	1.9068059	4.732198216	3.0571851
23	Great Britain	GB	23.78566	1.3170331	0.9769826	3.388993572	2.9974432
24	Norway	NO	23.78098	2.2479198	1.5947929	3.912051534	3.3752349

			Diseases of the circulatory system	Ischaemic heart diseases	Cerebrovascular diseases	Diseases of the respiratory system	Pneumonia
1	Bulgaria	BG	51.79604	10.450930	16.257807	3.235622	1.2809181
2	Czech Republic	CZ	39.70261	19.626516	8.873852	4.356048	2.1711890
3	Denmark	DK	27.37040	10.142579	7.123477	10.237267	3.5016479
4	Germany	DE	37.95312	14.343987	6.716210	6.282343	2.3234956
5	Estonia	EE	44.65771	22.579851	7.726758	2.406614	0.7037748
6	Ireland	IE	30.23526	15.871753	6.473522	10.508765	3.9862942
7	Greece	GR	41.08768	9.704620	13.411478	8.548314	0.7806108
8	Spain	ES	30.87998	9.035172	8.005362	11.115414	2.3113523
9	France	FR	29.92657	7.665907	6.636475	6.655887	2.1629280
10	Italy	IT	34.44665	11.530774	9.714136	5.767525	1.0543739
11	Latvia	LV	43.09346	22.538225	12.769399	1.891666	0.8740803
12	Lithuania	LT	42.12000	26.090755	10.418115	3.002585	1.2998128
13	Luxembourg	LU	36.35605	10.514604	8.428373	7.566064	2.3643950
14	Hungary	MG	40.21952	20.391455	8.693762	3.870451	0.4304642
15	Austria	AT	36.63487	16.395730	6.078207	4.685143	1.3397466
16	Poland	PL	43.10181	12.189372	9.283153	4.809305	2.2749307
17	Portugal	PT	30.95055	7.125465	13.349261	10.600319	4.7097270
18	Romania	RO	47.10835	16.160628	14.944905	3.786830	1.6347047
19	Slovenia	SL	36.97922	10.256935	9.944723	5.845020	2.8150271
20	Slovakia	SK	40.33454	24.377335	8.731462	4.218555	2.6010415
21	Finland	FI	35.36726	20.509556	7.404436	3.452846	0.7969448
22	Sweden	SE	35.73806	15.727572	7.978252	5.578290	2.1529069
23	Great Britain	GB	28.90899	13.363759	8.049579	12.317857	4.8936982
24	Norway	NO	31.61627	12.606245	7.985148	9.210879	3.6704840

Literatura

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

[2] Aitchison, J., Greenacre, M., *Biplots of compositional data*, Journal of the Royal Statistical Society 51 (4), 375–392, (2002).

[3] Daunis-i-Estadella, J., Thió-Henestrosa, S., Mateu-Figueras, G., *Including supplementary elements in a compositional biplot*, Journal Computers & Geosciences 37 (5), 696-701, (2011).

[4] Daunis-i-Estadella, J., Thió-Henestrosa, S., Mateu-Figueras, G., *Two more things about compositional biplots: Quality of projection and inclusion of supplementary elements*, [online], dostupné z: http://congreso.cimne.com/codawork11/Admin/Files/FilePaper/p18.pdf [citováno 22. 8. 2011]

[5] Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., *Isometric logratio transformations for compositional data analysis*, Mathematical Geology 35, 279-300, (2003).

[6] *Eurostat Database* [online], dostupné z: http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database [citováno 6. 2. 2012].

[7] *Eurostat Database - Causes of death - Absolute number (Annual data)* [online], dostupné z: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=hlth_cd_anr&lang=en [citováno 6. 2. 2012].

[8] *Eurostat Database - Crimes recorded by the police* [online], dostupné z: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=crim_gen&lang=en [citováno 6. 2. 2012].

[9] Filzmoser, P., Hron, K., *Outlier detection for compositional data using robust methods*, Mathematical Geosciences 40 (3), 233-248, (2008).

[10] Filzmoser, P., Hron, K., *Robust methods for compositional data*, [online], dostupné z: http://www.statistik.tuwien.ac.at/public/filz/papers/2010Compstat.pdf [citováno 29. 10. 2011].

[11] Filzmoser, P., Hron, K., Reimann, C., *Principal component analysis for compositional data with outliers*, Environmetrics 20 (6), 621–632, (2009).

[12] Filzmoser, P., Hron, K., Templ, M., *robCompositions: An R-package for robust statistical analysis of compositional data*, In: Pawlowsky-Glahn, V., Buccianti, A. Compositional data analysis: Theory and applications. Wiley, New York, 2011, 341-355.

[13] Gabriel, K. R., *The biplot graphic display of matrices with application to principal component analysis*, Biometrika 58 (3), 453-467, (1971).

[14] Greenaracre, M.J., Lewi, P., *Weighted logratio biplots, correspondence analysis and spectral maps*, [online], dostupné z: http://hdl.handle.net/10256/664 [citováno 6. 11. 2011].

[15] Greenaracre, M.J., Lewi, P., *Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements*, [online], dostupné z: http://www.econ.upf.edu/∼michael/work/GreenacreLewi5.pdf [citováno 2. 10. 2011].

[16] Hron, K., *Elementy statistické analýzy kompozičních dat*, Informační bulletin České statistické společnosti 21, 41 - 48, (2010).

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

[18] *List of countries by life expectancy* [online], dostupné z: http://en.wikipedia.org/wiki/List_of_countries_by_life_expectancy [citováno 10. 3. 2012]

[19] Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R., *Lecture notes on compositional data analysis*, [online], dostupné z: http://hdl.handle.net/10256/297, [citováno 31. 7. 2011].

[20] Šebková, T., *Bakalářská práce: Kompoziční data a statistický software R*, Olomouc: UPOL, 2009.

[21] *The R Project for Statistical Computing* [online], dostupné z: http://www.r-project.org/ [citováno 6. 2. 2012].

[22] *Trestní zákoník, Zákon č. 40/2009 Sb.* [online], dostupné z: http://business.center.cz/business/pravo/zakony/trestni-zakonik/cast1h2d1.aspx [citováno 13. 2. 2012].

[23] *Velký lékařský slovník* [online], dostupné z: http://lekarske.slovniky.cz/ [citováno 13. 2. 2012].