UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Vícerozměrná analýza symbolických dat

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

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Aneta Andrášiková

Název práce: Vícerozměrná analýza symbolických dat

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

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

Rok obhajoby práce: 2017

Abstrakt: Rozsáhlé datové soubory se vyskytují v mnoha vědních oborech, proto je nutné najít vhodný nástroj, jak s takovými daty zacházet. Jedním z možných přístupů je symbolické pojetí označované jako Symbolic Data Analysis. Diplomová práce se zabývá nejen symbolickými daty, ale zdůrazňuje také vztah mezi symbolickými a kompozičními daty, čímž určuje další orientaci textu. Teoretické aspekty jsou názorně vysvětleny při řešení příkladů vycházejících z různých datových souborů. K výpočtům je použit statistický software R.

Klíčová slova: symbolická data, kompoziční data, metoda hlavních komponent

Počet stran: 82

Počet příloh: 8

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Aneta Andrášiková

Title: Multivariate symbolic data analysis

Type of thesis: Master's

Department: Department of Mathematical Analysis and Application of Mathematics

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

The year of presentation: 2017

Abstract: The large data sets appear in many science fields. Therefore, it is necessary to find an appropriate tool suitable for their analyses. One of possible approaches is a symbolic concept called Symbolic Data Analysis. The master's thesis deals not only with the symbolic data, but also emphasizes the relationship between the symbolic and compositional data, thereby it determines the next orientation of the text. The theoretical aspects are illustratively explained by solving examples coming from various data sets. Statistical software R is used for calculations.

Key words: symbolic data, compositional data, principal component analysis

Number of pages: 82

Number of appendices: 8

Language: Czech

Prohlášení

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

V Olomouci dne ....

podpis

Obsah

Úvod **7**

1 Základy symbolických dat **8**
1.1 Na cestě k symbolickým datům 8
1.2 Nové typy symbolických proměnných 9

2 Modální proměnná **16**
2.1 Modální vícehodnotová proměnná 17
2.1.1 Popisná statistika pro případ jedné proměnné 17
2.1.2 Popisná statistika pro případ dvou a více proměnných . . . 21
2.2 Histogramová proměnná 26
2.2.1 Popisná statistika pro případ jedné proměnné 26
2.2.2 Popisná statistika pro případ dvou a více proměnných . . . 33

3 Spojitost symbolických a kompozičních dat **41**
3.1 Úvod do kompozičních dat 41
3.2 Provázanost symbolických a kompozičních dat 46

4 PCA tradičně i netradičně **48**
4.1 Standardní přístup 48
4.2 PCA s kompozičními vektory 49
4.3 Aplikace na Kola data 58

Závěr **72**

Literatura **73**

Přílohy **76**

A Datový soubor: Ubytování **76**
A.1 Grafické znázornění 76

B Datový soubor: Požáry **77**
B.1 Grafické znázornění 77
B.2 Výpočet symbolického výběrového průměru a rozptylu 77

C Datový soubor: Kola data **78**
C.1 Příprava dat 78
C.2 Výpočet skórů 79
C.3 Grafické znázornění 79
C.4 Postup v případě $D = 3$ 80
C.5 Ternární diagram a mapa pro druhou hlavní komponentu 81

Poděkování

Ráda bych poděkovala vedoucímu své diplomové práce doc. RNDr. Karlu Hronovi, Ph.D. za čas věnovaný konzultacím, za obětavost a inspirující nápady. Děkuji také své rodině za poskytnutou podporu nejen při psaní této práce, ale po celou dobu studia.

Úvod

Dnešní doba je již standardně spojována s rozsáhlými datovými soubory, což nás vzhledem ke stále vyšší a vyšší úrovni technického vybavení patrně nepřekvapí. Při setkání s daty tohoto typu však vyvstává otázka, jak s nimi dále zacházet. Je možné provést určitý druh sumarizace? A neztratíme tím příliš mnoho informací? Odpovědi na tyto otázky můžeme nalézt v rámci symbolického konceptu, který nese název Symbolic Data Analysis (zkráceně SDA). Tato metodika přináší návod na sumarizaci problematicky rozsáhlých datových souborů, aniž by docházelo k nepřiměřené ztrátě informací. Ona sumarizace je umožněna vznikem nových typů proměnných, které označujeme jako symbolické proměnné. Ač se nám může zdáti, že jsme se se symbolickými daty (dále SD) doposud nesetkali, tak věžme, že speciálním typem symbolických proměnných jsou kompoziční data, která jsou bezpochyby rozšířena daleko více. A pokud jsme doposud neslyšeli ani o symbolických ani o kompozičních datech, tak vzpomeňme na slavný citát ještě slavnějšího fyzika:

„*Lidské poznání má přece jen své hranice, a bude je mít vždy.*“

Albert Einstein

Cílem této práce je popis možných přístupů analýzy SD ke statistické analýze několika typů symbolických proměnných. Klíčovým aspektem, který zde bereme v potaz, je vícerozměrnost dat. Práce je složena ze čtyř kapitol. První kapitola se zabývá základy SD, popisuje různé typy symbolických proměnných. Druhá kapitola je pak věnována výhradně modálním proměnným, konkrétně charakteristikám popisné statistiky nejprve pro případ jedné proměnné, následně pro případ dvou a více modálních proměnných. Další, třetí kapitola, představuje stěžejní spojovací článek mezi první a druhou polovinou práce. Nastiňuje určitou spojitost mezi symbolickými a kompozičními daty. Čtvrtá, a tedy poslední, kapitola obsahuje nejen standardní popis metody hlavních komponent, ale také její modifikaci pro případ kompozičních vektorů. Na závěr je uveden příklad, řešený pomocí statistického softwaru R.

1. Základy symbolických dat

Tato kapitola má za cíl seznámit čtenáře se základní myšlenkou analýzy SD. Dozvíme se také, proč standardní postupy analýzy datových souborů nemusí být vždy postačující. Nakonec se seznámíme s novými typy proměnných SD. V této kapitole budeme vycházet z [2], [4], [5] a [9].

1.1. Na cestě k symbolickým datům

Představme si skupinku deseti dětí z jisté základní školy (dále ZŠ). Co nás u nich může zajímat? Může to být jejich výška, váha, kolik mají sourozenců, jaké jsou jejich oblíbené předměty apod. Z takto získaných dat není nikterak obtížné vytvořit klasické datové pole, ve kterém řádky odpovídají jednotlivým dětem a sloupce obsahují odpovědi na naše otázky, následně pak provést analýzu dat s využitím standardních metod. S takovýmto případem rozměrů dat se ovšem setkáváme velice zřídka.

V praxi obvykle uvažujeme datové soubory nesrovnatelně většího rozsahu (řádově až desetitisíce jednotek a až stovky proměnných), u nichž je vhodnější využít symbolických postupů.

Poznámka 1.1. V klasickém případě budeme uvažovat i-tého jednotlivce, přičemž $i \in \Omega = \{1, \ldots, n\}$, kde $n \in \mathbb{N}$ označuje počet jednotlivců daného datového souboru. Realizaci j-té proměnné Y_j, kde $j = 1, \ldots, p$, odpovídající i-tému jednotlivci, $i = 1, \ldots, n$, označíme jako x_{ij}, tedy $Y_j(i) = x_{ij}$. Hodnoty x_{ij} tvoří matici $\mathbf{X}$. Dále, pokud obor hodnot náhodné veličiny Y_j označíme $\mathcal{Y}_j$, $j = 1, \ldots, p$, matice $\mathbf{X}$ nabývá hodnot z $\mathcal{X} = \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_p = \times_{i=1}^{p} \mathcal{Y}_j$.

Poznámka 1.2. V rámci symbolického pojetí pak uvažujme u-tou kategorii ω_u nabývající hodnot z množiny m symbolických kategorií $E = \{\omega_1, \ldots, \omega_m\}$. Symbolickou realizaci budeme značit ξ_{uj}, tedy $Y_j(\omega_u) = \xi_{uj}$, pro $u = 1, \ldots, m$. Tyto realizace pak tvoří matici $\boldsymbol{\xi}$, nabývající hodnot z $\boldsymbol{\mathcal{Y} = \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_p}$.

Základem vhodnějšího přístupu práce s daty je shrnout datový soubor do takových rozměrů, abychom nejen zredukovali komplikovanost analýz, což je bezpochyby zásadním požadavkem, ale také např. zdůraznili zájem o celky vyššího stupně (kategorie), konkrétně v našem případě budeme v rámci měst či států patrně klást větší důraz na jednotlivé ZŠ než na konkrétní děti.

Myšlenka SD spočívá mimo jiné v tom, že ke shrnutí dat nedochází pouhým určením průměrů či modů v rámci jednotlivých proměnných, ale vytvořením nových typů proměnných, které budou brát v potaz vnitřní variabilitu jednotlivých kategorií. Nebude tak docházet k nesmyslným ztrátám hodnotných informací. Zásadním rozdílem těchto proměnných oproti proměnným klasického typu je skutečnost, že symbolické proměnné nám umožňují uvažovat mnohonásobné, resp. vážené, hodnoty. Setkáváme se zde s tím, že jednotlivá pozorování nabývají několika (různě vážených) hodnot současně.

V případě, že by některé pozorování nabývalo jediné hodnoty, uvědomíme si, že klasická realizace představuje speciální případ symbolické realizace.

1.2. Nové typy symbolických proměnných

V rámci symbolického pojetí rozlišujeme tři základní typy symbolických proměnných. Jedná se o vícehodnotové, modální a intervalové proměnné.

Vícehodnotové proměnné dále dělíme na kvalitativní a kvantitativní proměnné. Co se týče dělení modálních proměnných, setkáváme se s modálními vícehodnotovými proměnnými a histogramovými proměnnými. Přestože typů modálních proměnných existuje více, my se v další kapitole omezíme pouze na tyto dva typy. Intervalové proměnné pak představují speciální případ histogramových proměnných. Nejprve uvedeme příslušné definice. Následně předvedeme jednotlivé typy proměnných na ilustrativním příkladu. Pojem náhodné veličiny zde chápeme ve výše uvedeném rozšířeném (symbolickém) smyslu.

Definice 1.1. Vícehodnotová proměnná je taková symbolická náhodná veličina Y, která pro každou realizaci nabývá jedné nebo více hodnot ze svého oboru

hodnot $\mathcal{Y}$. Výčet možných hodnot obsažených v $\mathcal{Y}$ je konečný. V případě kvalitativní proměnné se jedná o množinu kategorií, v případě kvantitativní proměnné o množinu reálných čísel.

Definice 1.2. Nechť náhodná veličina Y nabývá hodnot z množiny $\{\eta_k,\, k \in \mathbb{N}\}$ v rámci oboru hodnot $\mathcal{Y}$. Potom se jedná o modální proměnnou, pokud konkrétní výsledná hodnota pro u-tou kategorii ω_u má tvar:

$$Y(\omega_u) = \xi_u = \{\eta_k,\, \pi_k;\, k = 1, \ldots, s_u\},$$

kde π_k je nezáporná míra spojená s η_k a kde s_u je počet hodnot skutečně vzatých z oboru hodnot $\mathcal{Y}$. Těchto možných η_k může být konečný či nekonečný počet, přitom může jít o hodnoty kvalitativní či kvantitativní.

Definice 1.3. Nechť $\mathcal{Y}_{cat}$ je oborem hodnot možných výsledků vícehodnotové proměnné Y_{cat}, kde $\mathcal{Y}_{cat} = \{\eta_1,\, \eta_2,\, \ldots\}$. Pak modální vícehodnotovou proměnnou nazveme takovou proměnnou, která nabývá hodnot patřících do podmnožiny oboru hodnot $\mathcal{Y}_{cat}$ s nezápornými mírami přidruženými ke každé hodnotě v oné podmnožině. Jednotlivé pozorování pro kategorii ω_u nabývá tvaru:

$$Y\left(\omega_u\right) = \xi_u = \left\{\eta_{u1},\, p_{u1}; \ldots;\, \eta_{us_u},\, p_{us_u}\right\},$$

kde $\{\eta_{u1},\ \ldots,\ \eta_{us_u}\} \subseteq \mathcal{Y}_{cat}$. Výsledek η_{uk} nastane s váhou p_{uk}, kde $k = 1,\ \ldots,\ s_u$ a s_u je počet hodnot skutečně vzatých z oboru hodnot $\mathcal{Y}_{cat}$. Dále platí, že $\sum_{k=1}^{s_u} p_{uk} = 1,\, \forall u = 1,\ \ldots,\ m$.

Definice 1.4. Nechť Y označuje kvantitativní náhodnou veličinu, která může nabývat hodnot na konečném počtu nepřekrývajících se intervalů $\{\langle a_k,\, b_k)\,,\, k \in \mathbb{N}\}$, kde $a_k \leq b_k$. Potom se jedná o histogramovou proměnnou, jestliže výsledná hodnota odpovídající u-té kategorii ω_u je tvaru:

$$Y\left(\omega_u\right) = \xi_u = \left\{\langle a_{uk},\, b_{uk})\,,\, p_{uk};\, k = 1,\ \ldots,\ s_u\right\},$$

přičemž $s_u < \infty$ je konečný počet intervalů vytvářejících nosič výsledné hodnoty $Y(\omega_u)$ odpovídající u-té kategorii ω_u, a p_{uk} odpovídá váze konkrétního podintervalu $\langle a_{uk},\, b_{uk})$, kde $k = 1,\ \ldots,\ s_u$. Dále platí, že $\sum_{k=1}^{s_u} p_{uk} = 1$, pro $u = 1,\ \ldots,\ m$. Intervaly s krajními body a_k, b_k mohou být otevřené nebo uzavřené.

Při standardním chápání pojmu „interval“ z uvedeného vyplývá, že histogramová proměnná je speciálním případem modální vícehodnotové proměnné, a to konkrétně pro kvantitativní případ. Dále tedy můžeme uvažovat modální vícehodnotovou proměnnou jako kvalitativní a histogramovou proměnnou jako kvantitativní verzi modální proměnné.

Definice 1.5. Intervalová proměnná je taková symbolická náhodná veličina Y, která nabývá hodnot ve tvaru intervalu, tj. $Y = \xi = \langle a,\, b\rangle \subset \mathbb{R}$, pro $a \leq b$, kde $a,\, b \in \mathbb{R}$. Interval může být tvaru: $(a,\, b)$, $\langle a,\, b\rangle$, $(a,\, b\rangle$, nebo $\langle a,\, b)$.

Nyní již přejdeme k samotnému ilustrativnímu příkladu.

Příklad 1.1. Uvažujme soubor tisíce zhruba stejně starých dětí z pěti fiktivních ZŠ. O každém z dětí si zaznamenáme hodnoty zkoumaných proměnných, a vytvoříme tak klasické datové pole, které vyobrazuje tabulka 1. Datové pole obsahuje řádky odpovídající $i = 1,\ \ldots,\ 1000$. Legendu objasňující názvy proměnných tohoto datového pole nalezneme v tabulce 2.

Tabulka 1: Klasické datové pole.

i	Y_0	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1	ZŠ Žlutá	1,42	31	2	český jazyk	kočka	1
2	ZŠ Modrá	1,37	29	1	tělesná výchova	pes	2
3	ZŠ Oranžová	1,45	37	1	matematika	žádný	2
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...

Zdůrazněme, že v našem případě je nezbytné uvádět také jméno školy, do které dítě chodí (odpovídající proměnné Y_0), neboť právě název ZŠ bude představovat hledisko, podle nějž vytvoříme kategorie. Namísto toho, abychom dále uvažovali

Tabulka 2: Legenda k tabulce 1.

Y_0	- ZŠ
Y_1	- výška [m]
Y_2	- hmotnost [kg]
Y_3	- počet sourozenců
Y_4	- oblíbený předmět
Y_5	- domácí mazlíček
Y_6	- počet kroužků

tisíc dětí, budeme pracovat již jen s pěti kategoriemi ZŠ, konkrétně ZŠ Červená, ZŠ Fialová, ZŠ Modrá, ZŠ Oranžová a ZŠ Žlutá. Jak budou vypadat jednotlivé typy proměnných SD?

Začneme vícehodnotovou proměnnou. Již dříve jsme uvedli, že vícehodnotová proměnná může být dvojího typu. Ukážeme příklad kvalitativní proměnné, resp. kvantitativní proměnné. Využijeme k tomu proměnných Y_5 (domácí mazlíček), resp. Y_6 (počet kroužků), viz tabulky 3 a 4.

Tabulka 3: Vícehodnotová kvalitativní proměnná.

škola	domácí mazlíček
ZŠ Červená	{had, kočka, křeček, pes, ryba, žádný, želva}
ZŠ Fialová	{ježek, kočka, pes, žádný}
ZŠ Modrá	{kočka, morče, pes, žádný}
ZŠ Oranžová	{kočka, pes, ryba, žádný, želva}
ZŠ Žlutá	{had, ještěrka, kočka, křeček, morče, pes, žádný}

Tabulka 4: Vícehodnotová kvantitativní proměnná.

škola	počet kroužků
ZŠ Červená	{0, 1, 2, 3}
ZŠ Fialová	{0, 1, 2}
ZŠ Modrá	{1, 2}
ZŠ Oranžová	{1, 2, 3}
ZŠ Žlutá	{0, 1, 2, 3, 4}

Jen pro úplnost dodejme, že levé sloupce tabulek 3 a 4 jsou tvořeny pěti námi určenými kategoriemi ZŠ a pravé sloupce obsahují neopakující se seznamy od-

povědí dětí z daných škol. Pokud bychom požadovali exaktní zápis, pak např. pro ZŠ Červená dostáváme:

$$Y_5(\omega_1) = Y_5\left(\text{ZŠ Červená}\right) = \xi_{15} = \{\text{had, kočka, křeček, pes, ryba, žádný, želva}\},$$
$$Y_6(\omega_1) = Y_6\left(\text{ZŠ Červená}\right) = \xi_{16} = \{0, 1, 2, 3\}.$$

Nyní budeme pokračovat ukázkou modální proměnné. Již víme, že ve spojitosti s modální proměnnou můžeme uvažovat kvalitativní i kvantitativní obdobu, což nám předvádí tabulky 5 (pro kvalitativní případ) a 7 či 8 (pro kvantitativní případ). Vycházíme zde z proměnných Y_4 (oblíbený předmět), Y_2 (hmotnost [kg]) a Y_3 (počet sourozenců). Zkratky jednotlivých předmětů uvedené v tabulce 5 pak vysvětluje tabulka 6. Dodejme, že dělení použité u proměnné Y_2 vychází z našeho rozhodnutí, neboť se jedná pouze o ilustrativní příklad.

Tabulka 5: Modální vícehodnotová proměnná.

škola	oblíbený předmět					
	aj	čj	hv	mat	tv	jiný
ZŠ Červená	0,75	0,10	0,02	0,03	0,09	0,01
ZŠ Fialová	0,55	0,10	0,02	0,03	0,30	0,00
ZŠ Modrá	0,15	0,05	0,30	0,10	0,37	0,03
ZŠ Oranžová	0,05	0,10	0,65	0,15	0,05	0,00
ZŠ Žlutá	0,30	0,25	0,15	0,12	0,17	0,01

Tabulka 6: Legenda k tabulce 5.

aj	- anglický jazyk
čj	- český jazyk
hv	- hudební výchova
mat	- matematika
tv	- tělesná výchova
jiný	- jiný předmět

K vytvoření pravých částí tabulek 5, 7 i 8 by bylo zapotřebí zjistit četnosti výskytu odpovědí jednotlivých dětí a následně tyto hodnoty přepočíst na proporce. Přesný

Tabulka 7: Histogramová proměnná Y_2.

škola	hmotnost [kg]		
	< 30	$\langle 30, 36\rangle$	> 36
ZŠ Červená	0,01	0,96	0,03
ZŠ Fialová	0,04	0,93	0,03
ZŠ Modrá	0,02	0,94	0,04
ZŠ Oranžová	0,05	0,92	0,03
ZŠ Žlutá	0,02	0,94	0,04

Tabulka 8: Histogramová proměnná Y_3.

škola	počet sourozenců			
	0	1	2	3 a více
ZŠ Červená	0,60	0,37	0,03	0,00
ZŠ Fialová	0,72	0,09	0,18	0,01
ZŠ Modrá	0,45	0,47	0,08	0,00
ZŠ Oranžová	0,12	0,76	0,10	0,02
ZŠ Žlutá	0,83	0,12	0,05	0,00

zápis by např. pro ZŠ Fialová vypadal takto:

$$
\begin{aligned}
Y_4(\omega_2) &= Y_4\left(\text{ZŠ Fialová}\right) = \xi_{24} = \\
&= \{\text{aj, 0,55; čj, 0,10; hv, 0,02; mat, 0,03; tv, 0,30; jiný, 0,00}\}, \\
Y_2(\omega_2) &= Y_2\left(\text{ZŠ Fialová}\right) = \xi_{22} = \{< 30,\ 0{,}04;\ \langle 30,\ 36\rangle,\ 0{,}93;\ > 36,\ 0{,}03\}, \\
Y_3(\omega_2) &= Y_3\left(\text{ZŠ Fialová}\right) = \xi_{23} = \{0,\ 0{,}72;\ 1,\ 0{,}09;\ 2,\ 0{,}18;\ 3 \text{ a více},\ 0{,}01\}.
\end{aligned}
$$

Zbývá nám ukázka intervalové proměnné, k jejímuž představení využijeme proměnnou Y_1 (výška [m]). Intervalovými daty a jejich možnou reprezentací (s využitím distribuční funkce, či pomocí parametrického přístupu) jsem se zabývala ve své bakalářské práci. Pro podrobnější informace odkazuji na [2].

Jak můžeme vidět z tabulky 9, intervalovou proměnnou vytvoříme jako interval z nejmenší a největší hodnoty, které máme pro danou ZŠ k dispozici. Exaktní zápis by pak např. pro ZŠ Modrá vypadal takto:

Tabulka 9: Intervalová proměnná.

škola	výška [m]
ZŠ Červená	$\langle 1{,}31;\ 1{,}49\rangle$
ZŠ Fialová	$\langle 1{,}33;\ 1{,}45\rangle$
ZŠ Modrá	$\langle 1{,}29;\ 1{,}44\rangle$
ZŠ Oranžová	$\langle 1{,}30;\ 1{,}45\rangle$
ZŠ Žlutá	$\langle 1{,}31;\ 1{,}47\rangle$

$$Y_1(\omega_3) = Y_1\left(\text{ZŠ Modrá}\right) = \xi_{31} = \langle 1{,}29;\ 1{,}44\rangle .$$

Tímto ilustrativním příkladem uzavřeme stručný úvod do SD, věnující se mimo jiné různým typům SD. Následující kapitola je pak zaměřena pouze na modální proměnnou, u které nás budou zajímat charakteristiky popisné statistiky.

2. Modální proměnná

Jak bylo uvedeno v podkapitole 1.2, rozlišujeme v rámci modálních proměnných modální vícehodnotovou proměnnou a histogramovou proměnnou. U každé z těchto proměnných se zaměříme na charakteristiky popisné statistiky, a to pro případ jedné a následně i dvou a více proměnných. Popsanou teorii pak aplikujeme při výpočtech jednotlivých charakteristik. Teoretická část této kapitoly vychází především z [4] a [5].

Pro další účely je zapotřebí zavést následující pojmy:

Definice 2.1. Nechť náhodné veličiny Y_j, $j = 1, \ldots, p$, mají obory hodnot $\mathcal{Y}_j$. Označíme-li $\mathcal{X} = \times_{j=1}^{p} \mathcal{Y}_j$, potom každý bod $\boldsymbol{x} = (x_1, \ldots, x_p) \in \mathcal{X}$ nazveme popisným vektorem.

Definice 2.2. Nechť náhodné veličiny Y_j, $j = 1, \ldots, p$, mají obory hodnot $\mathcal{Y}_j$. Označme $\mathcal{X} = \times_{j=1}^{p} \mathcal{Y}_j$. Dále nechť jsou D_j podmnožinami $\mathcal{Y}_j$, tedy $D_j \subseteq \mathcal{Y}_j$. Jestliže $D = \times_{j=1}^{p} D_j \subseteq \mathcal{X}$ je kartézským součinem množin D_j, pak D nazveme kartézskou popisnou množinou.

Definice 2.3. Nechť náhodné veličiny Y_j, $j = 1, \ldots, p$, mají obory hodnot $\mathcal{Y}_j$. Označme $\mathcal{X} = \times_{j=1}^{p} \mathcal{Y}_j$. Nechť $\mathbf{Y} = (\mathbf{Y}^1, \mathbf{Y}^2)$, kde $\mathbf{Y}^1 = (Y_1, \ldots, Y_k)$ nabývá hodnoty $d = (x_1, \ldots, x_k) \in \times_{j=1}^{k} \mathcal{Y}_j$ a $\mathbf{Y}^2 = (Y_{k+1}, \ldots, Y_p)$ nabývá hodnot z podmnožiny $D = \times_{j=k+1}^{p} D_j \subseteq \times_{j=k+1}^{p} \mathcal{Y}_j$. Potom se $\mathbf{d} = (x_1, \ldots, x_k, D_{k+1}, \ldots, D_p)$ nazývá popis d. Množinu všech možných popisů označíme jako popisný prostor $\mathcal{D}$.

Definice 2.4. Individuální popis, jenž označíme x, je takový popis, pro nějž je každé D_j jednoprvkovou množinou.

Definice 2.5. Virtuální popis $vir\,(d)$ popisného vektoru d je množinou všech individuálních popisných vektorů x, které splňují všechna pravidla (logické závislosti) v na $\mathcal{X}$[1], tj. $vir\,(d) = \{x \in D;\, v\,(x) = 1, \forall v \in V_{\mathcal{X}}\}$, kde $V_{\mathcal{X}}$ označuje množinu všech pravidel v působících na $\mathcal{X}$.

[1] Konvenční zápis pravidla $v : \langle x \in A\rangle \Rightarrow \langle x \in B\rangle$.

Ve skutečnosti slouží virtuální popis k matematickému „očištění“ dat. Dále tak pracujeme pouze s hodnotami, které jsou logicky správné. V případě malých datových souborů by bylo možné provést „očištění“ vizuálně - to ale nebude naše situace. My budeme obecně uvažovat rozsáhlé datové soubory, u kterých následně dojde k agregaci do kategorií. V souvislosti s tímto je nutné si uvědomit, že ani původně „správná“ data si nemusí svou korektnost po agregaci uchovat.

Příkladem logicky nesprávných dat mohou být například záznamy uvádějící nenulový počet potomků u malých dětí apod.

2.1. Modální vícehodnotová proměnná

2.1.1. Popisná statistika pro případ jedné proměnné

Mezi základní charakteristiky popisné statistiky pro případ jedné proměnné patří histogramy četností, výběrové průměry a v neposlední řadě také výběrové rozptyly. Vzhledem ke kvalitativní povaze modální vícehodnotové proměnné, která nás dále bude zajímat, se omezíme pouze na určení četností, a to v rámci symbolického pojetí. Připomeňme, že SD se často vyskytují v přítomnosti pravidel logické závislosti, a to z již dříve uvedeného důvodu – zajistíme tím, abychom nepracovali s logicky nesprávnými daty.

Poznámka 2.1. Pro snadnější zápis budeme dále místo u-té kategorie $\omega_u \in E$ psát $u \in E$.

Uvažujme data vícehodnotové či kategoriální proměnné Y_j, nabývající možných hodnot ξ_{jk} s vahami p_{jk}, $k = 1, \ldots, s$, přičemž $\sum_{k=1}^{s} p_{jk} = 1$. Předpokládejme, že nás zajímá konkrétní symbolická náhodná veličina $Y_j \equiv Z$ s ohledem na logickou závislost v. Potom můžeme definovat:

Definice 2.6. Napozorovaná četnost toho, že $Z = \xi_k$, $k = 1, \ldots, s$, je

$$O_Z(\xi_k) = \sum_{u \in E} \pi_Z(\xi_k; u), \tag{1}$$

kde

$$\pi_Z(\xi_k; u) = P(Z = \xi_k | x \in vir(d_u), u) = \\ = \frac{\sum_x P(x = \xi_k | x \in vir(d_u), u)}{\sum_x \sum_{k=1}^{s} P(x = \xi_k | x \in vir(d_u), u)}, \tag{2}$$

přičemž $P(x = \xi_k | x \in vir(d_u), u)$ je pro každou kategorii u pravděpodobnost toho, že individuální popis $\mathbf{x}'$ $(\equiv x_j)$ má hodnotu ξ_k a platí pravidlo logické závislosti v.

Poznámka 2.2. Pokud pro konkrétní u-tou kategorii nedisponujeme žádnými popisnými vektory, které by splňovaly pravidlo v, jinými slovy, pokud je jmenovatel ve vztahu (2) nulový, pak vynecháváme u-tý člen součtu ve vztahu (1).

Dále lze konstatovat, že

$$\sum_{k=1}^{s} O_Z(\xi_k) = m. \tag{3}$$

Kromě absolutních četností z definice 2.6 můžeme uvažovat taktéž relativní četnosti. Ty získáme vydělením napozorovaných četností počtem pozorování, resp. počtem kategorií m. Zavedené pojmy si blíže představíme na následujícím příkladu.

Příklad 2.1. Uvažujme datový soubor [18] popisující proporcionální zastoupení typu obydlí obyvatelstva 31 států pro rok 2015, tj. $m = 31$. Typem obydlí rozumíme rodinný dům, dvojdomek, byt a ostatní, tedy $k = 1, \ldots, 4$. Data jsou uvedena v tabulce 10. Protože se zabýváme jedinou proměnnou, označíme ji jako Z. Z téhož důvodu také není potřeba uvažovat žádné pravidlo logické závislosti v.

Než začneme s výpočty, je zapotřebí si uvědomit, jaké informace datový soubor obsahuje. Na první pohled je zřejmé, že se jedná o modální vícehodnotovou proměnnou. Co se týče interpretace číselných hodnot, tak například pro Českou republiku můžeme říci, že 37,1 % obyvatelstva žilo v roce 2015 v rodinném domě, 10,3 % v dvojdomku, dále 52,2 % v bytě a konečně 0,4 % jinde.

Tabulka 10: Datový soubor: Bydlení.

stát	rodinný dům	dvojdomek	byt	ostatní
Belgie	0,366	0,407	0,221	0,006
Bulharsko	0,432	0,124	0,440	0,004
Česká republika	0,371	0,103	0,522	0,004
Dánsko	0,562	0,128	0,305	0,005
Německo	0,255	0,158	0,573	0,014
Estonsko	0,321	0,047	0,627	0,005
Irsko	0,409	0,515	0,074	0,002
Řecko	0,338	0,101	0,561	0,000
Španělsko	0,127	0,210	0,658	0,005
Francie	0,447	0,237	0,315	0,001
Chorvatsko	0,734	0,079	0,187	0,000
Itálie	0,213	0,259	0,525	0,003
Kypr	0,471	0,255	0,259	0,014
Lotyšsko	0,318	0,031	0,650	0,001
Litva	0,361	0,063	0,574	0,002
Lucembursko	0,369	0,280	0,343	0,008
Maďarsko	0,622	0,048	0,325	0,005
Malta	0,051	0,402	0,545	0,002
Nizozemí	0,166	0,599	0,199	0,036
Rakousko	0,480	0,069	0,445	0,006
Polsko	0,506	0,052	0,441	0,001
Portugalsko	0,366	0,179	0,453	0,002
Rumunsko	0,602	0,019	0,379	0,000
Slovinsko	0,651	0,050	0,296	0,003
Slovensko	0,465	0,018	0,512	0,005
Finsko	0,465	0,193	0,337	0,005
Švédsko	0,495	0,091	0,402	0,012
Spojené království	0,245	0,599	0,150	0,006
Island	0,341	0,188	0,467	0,003
Norsko	0,613	0,198	0,186	0,003
Srbsko	0,660	0,104	0,235	0,001

Naším cílem je napočítat napozorované a relativní četnosti, a to s využitím vztahu (1). Napozorovanou četnost toho, že $Z = \xi_1 =$ rodinný dům, počítáme jako:

$$O_Z(\xi_1) = 0{,}366 + 0{,}432 + \ldots + 0{,}613 + 0{,}660 = 12{,}822.$$

Obdobně lze napozorované četnosti spočítat i pro ξ_2, ξ_3 a ξ_4. Dále nás zajímají relativní četnosti. Vzhledem k tomu, že máme údaje o 31 státech, podělíme absolutní četnosti právě touto hodnotou.

Výpočty lze provést například ve statistickém softwaru R. Po načtení datového souboru ve formátu `csv` a uložení do proměnné `bydleni` zadáme:

```
cetnosti = c()
rel_cetnosti = c()
for (i in 2:ncol(bydleni)){
  cetnosti[i-1] = sum(bydleni[,i])
  rel_cetnosti[i-1] = cetnosti[i-1]/nrow(bydleni)
}
```

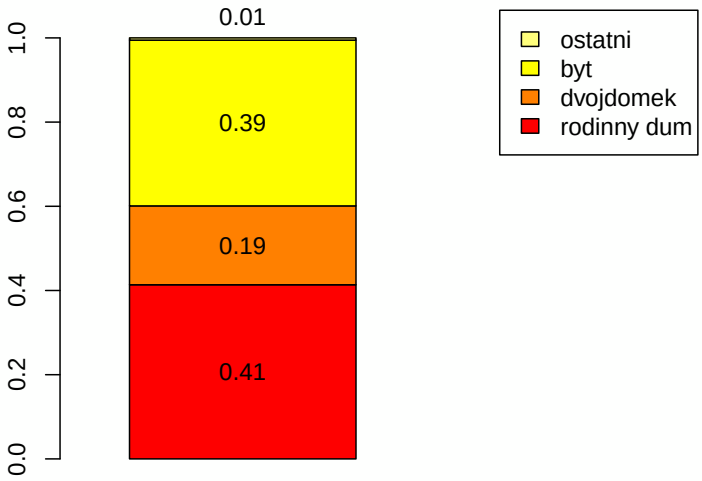
Podotkněme na tomto místě, že ač jsou hodnoty v tabulce 10 zaokrouhleny na tři desetinná místa, do softwaru načteme data nezaokrouhlená. Stejně tomu bude i v následujících příkladech 2.2, 2.3 a 2.4. Výsledné hodnoty četností sepíšeme pro přehlednost do tabulky 11.

Tabulka 11: Výsledné četnosti datového souboru Bydlení.

$Z = \xi_k,\ k = 1,\ \ldots,\ 4$	$O_Z(\xi_k)$	$O_Z(\xi_k)/m$
$Z = \xi_1$	12,822	0,414
$Z = \xi_2$	5,807	0,187
$Z = \xi_3$	12,207	0,394
$Z = \xi_4$	0,164	0,005

Z hlediska interpretace výsledků nás často zajímají právě relativní četnosti (viz obrázek 1), které uvádějí, že 41,4 % obyvatelstva uvedených států žilo v roce 2015 v rodinném domě, 18,7 % v dvojdomku, 39,4 % v bytě a 0,5 % jinde. Na závěr příkladu uveďme, že sečtením napozorovaných četností skutečně získáme hodnotu m, jak uvádí vztah (3). Součet relativních četností je pak roven jedné.

Popisnou statistiku pro případ jedné proměnné opustíme a přesuneme se k problematice charakteristik popisné statistiky pro případ dvou a více proměnných.



Obrázek 1: Proporcionální zastoupení typu obydlí evropského obyvatelstva.

2.1.2. Popisná statistika pro případ dvou a více proměnných

V této části se opět zaměříme na modální vícehodnotovou proměnnou. Nyní však situaci poněkud zkomplikujeme a naše úvahy rozšíříme na případ obecně p proměnných, kde $p > 1$. Konkrétně nás budou zajímat proměnné Z_1 a Z_2 z oboru hodnot $\mathcal{Z} = \mathcal{Z}_1 \times \mathcal{Z}_2$, tedy $p = 2$, a jim příslušející sdružené histogramy. Pro $p > 2$ bychom postupovali analogicky.

Poznámka 2.3. Modální vícehodnotové proměnné se skládají ze seznamu kategoriálních či kvalitativních proměnných s odpovídajícími pravděpodobnostmi. Následující úvahy budou vycházet z určité analogie definic pro případ vícehodnotové proměnné. Ty rozšíříme a zohledníme důležitost příslušných pravděpodobností.

Nechť modální vícehodnotová proměnná Y_j, $j = 1, \ldots, p$, nabývá pro kategorii ω_u, $u = 1, \ldots, m$, hodnot

$$Y_j(\omega_u) = \left\{ [\xi_{uj1}, p_{uj1}]; \ \ldots; \ [\xi_{ujs_j}, p_{ujs_j}] \right\},$$

kde $\sum_{k=1}^{s_j} p_{ujk} = 1$. Proměnná Y_j nabývá s_j možných hodnot $\{\xi_{jk},\ k = 1,\ \ldots,\ s_j\}$ s odpovídajícími pravděpodobnostmi $\{p_{ujk},\ k = 1,\ \ldots,\ s_j\}$ pro danou kategorii $\omega_u,\ u = 1,\ \ldots,\ m$. Dále uvažujme, že předmětem našeho zájmu bude hledání sdruženého histogramu pro konkrétní proměnné $Y_{j_1} \equiv Z_1$ a $Y_{j_2} \equiv Z_2$.

Poznámka 2.4. Pro snadnější zápis budeme dále místo u-té kategorie $\omega_u \in E$ opět psát $u \in E$.

Definice 2.7. Napozorovaná četnost toho, že $(Z_1 = \xi_{1k_1},\ Z_2 = \xi_{2k_2})$, je v případě modálních vícehodnotových proměnných $Z_j = \left\{\xi_{jk_j},\ k_j = 1,\ \ldots,\ s_j\right\},\ j = 1,\ 2,$ rovna

$$O\left(Z_1 = \xi_{1k_1},\ Z_2 = \xi_{2k_2}\right) = \sum_{u \in E} \pi_{Z_1,Z_2}\left(\xi_{1k_1},\ \xi_{2k_2};\ u\right), \tag{4}$$

kde výraz

$$\pi_{Z_1,Z_2}\left(\xi_{1k_1},\ \xi_{2k_2};\ u\right) = p_{u1k_1} \cdot p_{u2k_2} \frac{\left|\left\{x \in vir\left(d_u\right) | x_{Z_1} = \xi_{1k_1},\ x_{Z_2} = \xi_{2k_2}\right\}\right|}{\left|vir\left(d_u\right)\right|} \tag{5}$$

představuje procento jednotlivých kategoriálních dvojic $(\xi_{1k_1},\ \xi_{2k_2})$, které je obsaženo ve virtuálním popisu $vir\left(d_u\right)$.

Bylo by možné ukázat, že platí

$$\sum_{\substack{k_1 \\ k_2}} \sum_{\substack{\xi_{1k_1} \in \mathcal{Z}_1 \\ \xi_{2k_2} \in \mathcal{Z}_2}} O\left(Z_1 = \xi_{1k_1},\ Z_2 = \xi_{2k_2}\right) = m, \tag{6}$$

ovšem za předpokladu, že ve vztahu (4) sčítáme pouze přes taková pozorování (resp. kategorie), pro která $vir\left(d_u\right) \neq 0$. Z toho vyplývá následující definice.

Definice 2.8. Empirický sdružený histogram proměnných Z_1 a Z_2 je množina dvojic

$$\left\{\left[\xi_{1k_1},\ \xi_{2k_2};\ O_{Z_1Z_2}\left(\xi_{1k_1},\ \xi_{2k_2}\right)\right],\ k_j = 1,\ \ldots,\ s_j, j = 1,\ 2\right\}.$$

Poznámka 2.5. Při výpočtech se můžeme setkat také s váženými sdruženými histogramy. Potřebné váhy lze získat např. podílem počtu pozorování odpovídající dané u-té kategorii a celkového počtu pozorování n.

Příklad 2.2. Mějme k dispozici datové soubory vztahující se k návštěvnosti [19] a kapacitám hromadných ubytovacích zařízeních [20] (dále HUZ) jednotlivých krajů České republiky pro rok 2015. Z těchto datových souborů vytvoříme jeden soubor, viz tabulka 12. První proměnná, kterou dále označíme jako Z_1, popisuje proporcionální zastoupení hostů v HUZ. Tyto hosty rozlišujeme na rezidenty a nerezidenty. Druhá proměnná (Z_2) představuje proporcionální zastoupení různých typů ubytovacích zařízení z celkového počtu HUZ pro daný kraj.

V obou případech se jedná o modální vícehodnotovou proměnnou. Typem ubytovacího zařízení rozumíme hotel (zde řadíme také motel a botel), penzion, kemp (do této kategorie spadají také chatová osada a turistická ubytovna) a ostatní.

Tabulka 12: Datový soubor: Ubytování.

kraj	Hosté (Z_1)		HUZ (Z_2)			
	rezidenti	nerezidenti	hotel	penzion	kemp	ostatní
Hl. město Praha	0,135	0,865	0,670	0,127	0,087	0,117
Středočeský	0,770	0,230	0,311	0,323	0,214	0,152
Jihočeský	0,693	0,307	0,159	0,439	0,212	0,191
Plzeňský	0,639	0,361	0,223	0,399	0,202	0,177
Karlovarský	0,372	0,628	0,478	0,312	0,083	0,127
Ústecký	0,652	0,348	0,329	0,357	0,167	0,147
Liberecký	0,798	0,202	0,183	0,430	0,143	0,243
Královéhradecký	0,772	0,228	0,211	0,401	0,133	0,255
Pardubický	0,855	0,145	0,220	0,411	0,217	0,152
Vysočina	0,868	0,132	0,220	0,312	0,234	0,234
Jihomoravský	0,680	0,320	0,248	0,482	0,154	0,116
Olomoucký	0,790	0,210	0,191	0,424	0,178	0,206
Zlínský	0,844	0,156	0,297	0,360	0,157	0,187
Moravskoslezský	0,768	0,232	0,306	0,352	0,170	0,172

Pokud se budeme zabývat interpretací zadaných hodnot, zjistíme, že v Olomouckém kraji bylo v roce 2015 celkem 79 % z hostů rezidenty a 21 % z hostů nerezidenty. Dále se dozvíme, že z celkového počtu HUZ pro daný kraj se v 19,1 % jednalo o hotel (motel, botel), ve 42,4 % o penzion, v 17,8 % o kemp (chatovou osadu, turistickou ubytovnu) a ve 20,6 % o ostatní zařízení.

Obdobně jako v předchozí podkapitole nás budou zajímat napozorované četnosti. S ohledem na to, že má proměnná Z_1 dvě obměny (možné hodnoty) a proměnná Z_2 čtyři obměny, je zapotřebí spočítat napozorované četnosti celkem osmi variant. Tyto četnosti vytvoří ve spojení s obměnami proměnných osm dvojic. Množina takových dvojic pak představuje empirický sdružený histogram. S využitím vztahu (4) dostáváme pro první variantu:

$$O\left(Z_1 = 1,\, Z_2 = 1\right) = \left(0{,}135 \cdot 0{,}670\right)_{u=1} + \ldots + \left(0{,}768 \cdot 0{,}306\right)_{u=14} = 2{,}468.$$

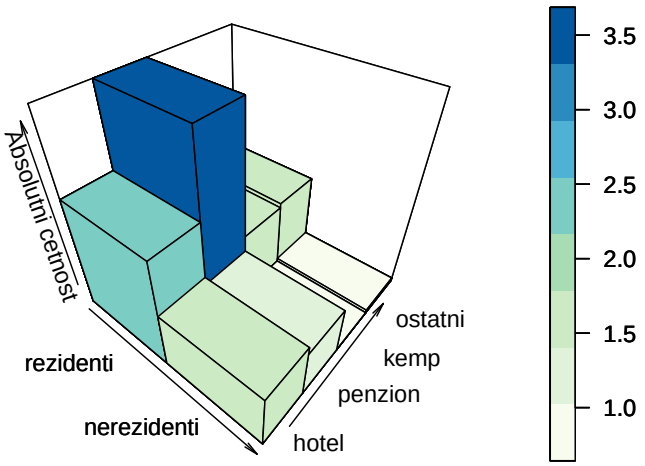
Pro ostatní varianty lze postupovat podobně. Vzhledem ke zdlouhavosti výpočtů je opět možné využít softwaru R. Datový soubor ve formátu `csv` uložíme do proměnné `ubytovani` a zadáme:

```
cetnostiA = matrix(0,2,4)
for (i in 4:7){
  cetnostiA[1,i-3] = sum(ubytovani[,2]*ubytovani[,i])
  cetnostiA[2,i-3] = sum(ubytovani[,3]*ubytovani[,i])
}
```

Výsledné absolutní četnosti je možné zakreslit do histogramu (viz obrázek 2). K tomu využijeme balíček `plot3D` s funkcí `hist3D`. Kód uvádíme v příloze A.1.

Dále je možné spočítat četnosti a relativní četnosti proměnných Z_1, případně Z_2, a to bez ohledu na druhou z uvedených. V případě relativních četností dělíme absolutní četnosti počtem kategorií. Pracujeme se 14 kraji, proto $m = 14$. Pro proměnnou Z_1 zadáme:

```
cetnosti1=c()
rel_cetnosti1=c()
```



Obrázek 2: Empirický sdružený histogram datového souboru Ubytování.

```
for (i in 1:2){
  cetnosti1[i] = sum(cetnostiA[i,])
  rel_cetnosti1[i] = cetnosti1[i]/14
}
```

Výsledné hodnoty obou proměnných uvádíme v tabulce 13.

Tabulka 13: Empirický sdružený histogram datového souboru Ubytování.

Hosté (Z_1)	HUZ (Z_2)				Četnost Z_1	Rel. četnost Z_1
	hotel	penzion	kemp	ostatní		
rezidenti	2,468	3,686	1,706	1,777	9,637	0,688
nerezidenti	1,577	1,443	0,644	0,699	4,363	0,312
Četnost Z_2	4,045	5,129	2,350	2,476	14	
Rel. četnost Z_2	0,289	0,366	0,168	0,177		1, 000

V případě proměnné Z_1 dostáváme proporcionální složení této proměnné nezávisle na proměnné Z_2. Při výpočtech četností vztahující se k proměnné Z_2 nahradíme řádkové součty sloupcovými součty. Co se týče interpretace výsledných relativních četností proměnné Z_1, uveďme, že 68,8 % z hostů bylo rezidenty a 31,2 % nerezidenty. Obdobně lze interpretovat i relativní četnosti proměnné Z_2.

Poznámka 2.6. Dodejme, že ač se jedná o případ $p = 2$, neuvažovali jsme žádné pravidlo logické závislosti v. Stejně tomu bude i v příkladu 2.4. Problém je dále možné rozšířit o pravidlo logické závislosti, příp. výpočet váženého sdruženého histogramu.

2.2. Histogramová proměnná

2.2.1. Popisná statistika pro případ jedné proměnné

Na rozdíl od modální vícehodnotové proměnné (viz definice 1.3) se setkáváme s proměnnou kvantitativního typu. Důsledkem toho můžeme uvažovat více charakteristik než v předchozím případě. Své úvahy totiž rozšíříme mimo jiné o symbolické obdoby výběrového průměru či výběrového rozptylu.

Poznámka 2.7. Zdůrazněme na tomto místě analogii úvah vztahujících se k intervalovým datům (viz [2]), neboť pro $k = 1$ dochází k degradaci histogramové proměnné na intervalovou proměnnou.

Předpokládejme, že předmětem našeho zájmu je náhodná veličina $Y_j \equiv Z$, která nabývá pro u-tou kategorii ω_u, hodnot z intervalů $\xi_{uk} = \langle a_{uk},\, b_{uk}) $ s pravděpodobnostmi p_{uk}, pro $u = 1,\ \ldots,\ m$ a pro $k = 1,\ \ldots,\ s_u$. Dále předpokládejme, že všechny individuální popisné vektory $x \in vir\,(d_u)$ mají uvnitř každého intervalu $\langle a_{uk},\, b_{uk})$ rovnoměrné rozdělení. Potom pro každé ξ_k platí, že

$$P\left\{x \leq \xi_k | x \in vir\,(d_u)\right\} = \begin{cases} 0 & \xi_k < a_{uk}, \\ \frac{\xi_k - a_{uk}}{b_{uk} - a_{uk}} & a_{uk} \leq \xi_k < b_{uk}, \\ 1 & \xi_k \geq b_{uk}. \end{cases} \tag{7}$$

Nechť $I = \langle \min_{k,u \in E} a_{ku},\, \max_{k,u \in E} b_{ku} \rangle$ reprezentuje interval, který pokrývá všechny napozorované hodnoty náhodné veličiny $Z \in \mathcal{Z}$, a nechť je takto označený interval I rozdělen do r podintervalů $I_g = \langle \zeta_{g-1},\, \zeta_g)\,,\ g = 1,\ \ldots,\ r-1$, přitom $I_r = \langle \zeta_{r-1},\, \zeta_r \rangle$.

Poznámka 2.8. I zde budeme pro snadnější zápis dále místo u-té kategorie $\omega_u \in E$ psát $u \in E$.

Definice 2.9. Pro histogramovou proměnnou Z definujeme napozorovanou četnost intervalu $I_g = \langle \xi_{g-1},\, \xi_g \rangle\,,\; g = 1,\; \ldots,\; r$ jako

$$O_Z\,(g) = \sum_{u \in E} \pi_Z\,(g;\, u)\,, \tag{8}$$

kde

$$\pi_Z\,(g; u) = \sum_{k \in Z(g)} \frac{\|Z\,(k;\, u) \cap I_g\|}{\|Z\,(k;\, u)\,\|} \cdot p_{uk}, \tag{9}$$

přičemž $Z\,(k;\, u)$ označuje interval $\langle a_{uk},\, b_{uk}\rangle$ a $Z\,(g)$ představuje množinu všech takových intervalů $Z\,(k;\, u)$, které se překrývají s intervalem I_g pro dané u. Zápisem $\|.\|$ rozumíme standardně délku daného intervalu.

Povšimněme si, že každý člen součtu ve vztahu (8) představuje takovou část intervalu $Z\,(k;\, u)$, která je překryta intervalem I_g, a tedy tu část jeho váhy p_{uk}, která se týká celého histogramového intervalu I_g. Z toho plyne, že

$$\sum_{g=1}^{r} O_Z\,(g) = m. \tag{10}$$

Definice 2.10. Relativní četnost intervalu I_g definujeme v případě histogramové proměnné jako

$$p_g = \frac{O_Z\,(g)}{m}. \tag{11}$$

Dohromady množina dvojic $\{[p_g,\, I_g]\,,\; g = 1,\; \ldots,\; r\}$ reprezentuje histogram relativních četností pro kombinovanou množinu pozorovaných histogramů, tj. histogram histogramových proměnných.

Definice 2.11. Empirickou funkci hustoty definujeme pro histogramovou proměnnou jako

$$f(\xi) = \frac{1}{m} \sum_{u \in E} \sum_{k=1}^{s_u} \frac{I_{uk}(\xi)}{\|Z(u;\, k)\|} \cdot p_{uk},\ \xi \in \mathbb{R}, \tag{12}$$

kde $I_{uk}(\cdot)$ je indikátorová funkce, tedy

$$I_{uk}(\xi) = \begin{cases} 1 & \text{pro } \xi \in Z(u;\, k) \\ 0 & \text{jinak.} \end{cases}$$

V rámci charakteristik popisné statistiky pro případ jedné proměnné uvedeme již poslední definici, a to definici vymezující pojmy symbolický výběrový průměr a symbolický výběrový rozptyl.

Definice 2.12. Symbolický výběrový průměr histogramové proměnné definujeme jako

$$\bar{Z} = \frac{1}{2m} \sum_{u \in E} \left[\sum_{k=1}^{s_u} (b_{uk} + a_{uk}) \cdot p_{uk} \right] \tag{13}$$

a symbolický výběrový rozptyl je dán vztahem

$$S^2 = \frac{1}{3m} \sum_{u \in E} \left[\sum_{k=1}^{s_u} \left(b_{uk}^2 + b_{uk} a_{uk} + a_{uk}^2 \right) \cdot p_{uk} \right] - \frac{1}{4m^2} \left[\sum_{u \in E} \sum_{k=1}^{s_u} (b_{uk} + a_{uk}) \cdot p_{uk} \right]^2. \tag{14}$$

Příklad 2.3. Uvažujme datový soubor [21] obsahující absolutní počty zraněných osob v důsledku požárů v roce 2015. Data se vztahují k jednotlivým okresům České republiky. Původní datový soubor (`pozary_puvodni.xls`) je k dispozici na přiloženém CD. Z tohoto souboru vytvoříme agregací vzhledem k příslušným krajům nový soubor, ve kterém jsou jednotlivé počty zraněných osob nahrazeny intervaly rozmezí počtů s odpovídajícími pravděpodobnostmi (viz tabulka 14), a vytvoříme tak symbolickou histogramovou proměnnou Z. Dále nás bude zajímat 13 krajů České republiky, neboť kraj „Hlavní město Praha“ nebudeme kvůli svým specifikům brát v potaz. Uvažujeme tedy $m = 13$. Z datové tabulky je možné

vyčíst, že v rámci Olomouckého kraje patřily všechny absolutní počty zraněných osob v důsledku požáru do intervalu $\langle 0,\, 20)$, zatímco například pro Moravskoslezský kraj bylo 33,3 % z absolutních počtů zraněných z intervalu $\langle 0,\, 20)$, 50 % z intervalu $\langle 20,\, 40)$ a zbývajících 16,7 % z intervalu $\langle 40,\, 60\rangle$.

Tabulka 14: Datový soubor: Požáry.

kraj	$\langle 0,\, 20)$	$\langle 20,\, 40)$	$\langle 40,\, 60\rangle$
Středočeský	0,667	0,333	0,000
Jihočeský	0,857	0,143	0,000
Plzeňský	0,857	0,143	0,000
Karlovarský	0,667	0,333	0,000
Ústecký	0,857	0,000	0,143
Liberecký	0,750	0,250	0,000
Královéhradecký	0,667	0,333	0,000
Pardubický	0,250	0,750	0,000
Vysočina	1,000	0,000	0,000
Jihomoravský	0,857	0,000	0,143
Olomoucký	1,000	0,000	0,000
Zlínský	0,750	0,000	0,250
Moravskoslezský	0,333	0,500	0,167

K výpočtům napozorovaných a relativních četností je nejprve zapotřebí vytvořit interval I, který pokryje všechny napozorované hodnoty. Získáme tak interval $I = \langle 0,\, 60\rangle$. Tento interval následně rozdělíme do r stejně dlouhých podintervalů I_g. K volbě hodnoty r můžeme využít například Sturgesova pravidla [3], podle kterého $r \doteq 5$. Naše vytvořené intervaly budou mít podobu: $I_1 = \langle 0,\, 12)$, $I_2 = \langle 12,\, 24)$, $I_3 = \langle 24,\, 36)$, $I_4 = \langle 36,\, 48)$ a $I_5 = \langle 48,\, 60\rangle$.

Nyní se budeme zabývat výpočtem požadovaných četností intervalů I_1 až I_5. Pro interval $I_1 = \langle 0,\, 12)$ můžeme dle vztahu (9) psát:

$$\pi_Z\,(1;\,1) = \frac{12-0}{20-0} \cdot 0{,}667 = 0{,}4,$$

$$\vdots$$

$$\pi_Z\,(1;\,13) = \frac{12-0}{20-0} \cdot 0{,}333 = 0{,}2.$$

K výpočtu napozorované četnosti, resp. relativní četnosti pak využijeme vztahy (8), resp. (11), a získáme tak

$$O_Z\,(g=1) = 5{,}707,$$
$$p_1 = \frac{5{,}707}{13} = 0{,}439.$$

Obdobně postupujeme i pro další intervaly I_2 až I_5. Pro zrychlení výpočtů je možné využít statistického softwaru R. Datový soubor ve formátu `csv` načteme a uložíme jej do proměnné `pozary`. Jednou z možností, jak dále postupovat, je využití logické úvahy, přičemž si uvědomíme, že např. interval $I_1 = \langle 0,\, 12)$ má nenulový průnik pouze s intervalem $\langle 0,\, 20)$, tudíž by do softwaru stačilo zadat:

```
O_1 = sum(pozary[,1]*(12/20))
p_1 = O_1/m
```

Zlomek $\frac{12}{20}$ je v kódu obsažen proto, že čitatel odpovídá délce intervalu vzniklého průnikem dvou zmíněných intervalů, tedy $12 - 0$, a jmenovatel vyjadřuje délku původního intervalu, tedy $20 - 0$. V tuto chvíli si uvědomujeme, že se jako daleko efektivnější způsob ukáže vytvoření funkce v softwaru. O nezbytnosti tohoto způsobu se navíc přesvědčíme v příkladu 2.4, kdy místo jedné proměnné budeme řešit problém se dvěma histogramovými proměnnými. Proto zadáme:

```
pozary = pozary[,2:4]
p = matrix(c(pozary[,1],pozary[,2],pozary[,3]),
    nrow = 13,ncol = 3)
b = c(0,20,40,60)
body = c(0, 12, 24, 36, 48, 60)

inters <- function(int1,int2){
  d = min(c(int1[2],int2[2])) - max(c(int1[1],int2[1]))
  if(d < 0){
    d = 0
  }
```

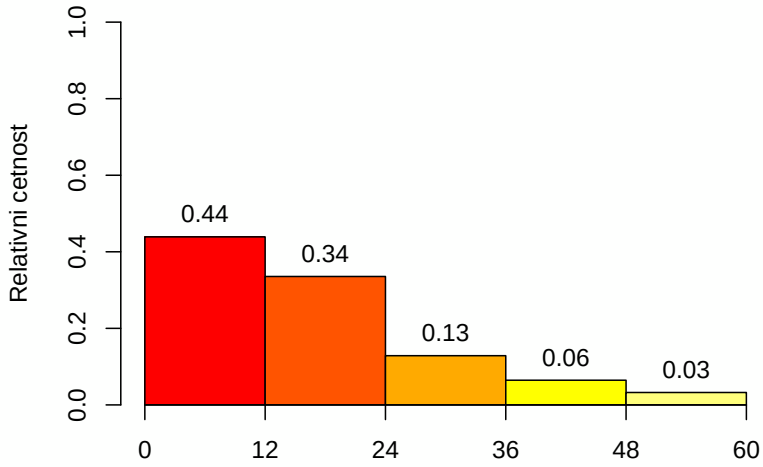
```
  return(d)
}

O = matrix(0,5,1)
for(k in 1:13){
  for(i in 1:5){
    int = body[i:(i+1)]
    for(j in 1:3){
      O[i] = O[i] + inters(int,
     b[j:(j+1)])/(b[j+1] - b[j]) * p[k,j]
    }
  }
}
m = 13
p_g = O/m
```

Výsledné hodnoty četností jsou zapsány v tabulce 15. Na základě těchto hodnot můžeme uvést, že 43,9 % absolutních počtů zraněných patřily do intervalu $\langle 0,\, 12)$, 33,6 % do intervalu $\langle 12,\, 24)$ atd. Dále si můžeme všimnout, že součet napozorovaných četností jednotlivých intervalů odpovídá hodnotě m, což uvádí vztah (10) a sečtením relativních četností obdržíme hodnotu rovnu jedné.

Tabulka 15: Výsledné četnosti histogramové proměnné z datového souboru Požáry.

g	I_g	$O_Z(g)$	p_g
1	$\langle 0,\, 12)$	5,707	0,439
2	$\langle 12,\, 24)$	4,362	0,336
3	$\langle 24,\, 36)$	1,671	0,129
4	$\langle 36,\, 48)$	0,838	0,064
5	$\langle 48,\, 60\rangle$	0,421	0,032

Než přikročíme k dalším výpočtům, znázorníme si relativní četnosti intervalů také graficky (viz obrázek 3). K vykreslení použijeme funkci `barplot`. Příslušný



Obrázek 3: Proporcionální zastoupení počtu zraněných osob v důsledku požárů.

kód uvádíme v příloze B.1.

Nyní nám zbývá provést výpočet symbolického výběrového průměru $\bar{Z}$ a symbolického výběrového rozptylu S^2 dle vztahů (13) a (14), tedy

$$\begin{aligned}
\bar{Z} &= \frac{1}{13}\left\{\left[\left(\frac{0+20}{2}\right)\cdot 0{,}667 + \ldots + \left(\frac{40+60}{2}\right)\cdot 0\right]_{u=1} + \ldots + \right. \\
&\quad \left. + \left[\left(\frac{0+20}{2}\right)\cdot 0{,}333 + \ldots + \left(\frac{40+60}{2}\right)\cdot 0{,}167\right]_{u=13}\right\} = \\
&= 16{,}447, \\
S^2 &= \frac{1}{13}\left\{\left[\left(\frac{0^2+0\cdot 20+20^2}{3}\right)\cdot 0{,}667 + \ldots + \left(\frac{40^2+40\cdot 60+60^2}{3}\right)\cdot 0\right]_{u=1} + \right. \\
&\quad + \ldots + \\
&\quad \left. + \left[\left(\frac{0^2+0\cdot 20+20^2}{3}\right)\cdot 0{,}333 + \ldots + \left(\frac{40^2+40\cdot 60+60^2}{3}\right)\cdot 0{,}167\right]_{u=13}\right\} \\
&\quad - (16{,}447)^2 = 163{,}932 \\
S &= 12{,}804
\end{aligned}$$

Pro úplnost uvádíme taktéž směrodatnou odchylku S. K samotným výpočtům jsme využili statistického softwaru R (viz příloha B.2).

Tímto jsme ukončili část věnující se popisné statistice jedné proměnné. V následující podkapitole své úvahy opět rozšíříme na případ dvou proměnných.

2.2.2. Popisná statistika pro případ dvou a více proměnných

Obdobně jako v podkapitole 2.1.2 se budeme zabývat charakteristikami popisné statistiky pro případ dvou proměnných, tedy $p = 2$, tentokrát však ve spojitosti s histogramovými proměnnými.

Uvažujme dvě konkrétní proměnné Z_1 a Z_2 a předpokládejme, že každá proměnná $Z_j(u)$ nabývá pro u-tou kategorii ω_u hodnot

$$Z_j(\omega_u) = \{\langle a_{ujk},\, b_{ujk}) ,\, p_{ujk},\, k = 1,\ \ldots,\, s_{uj}\} .$$

To znamená, že k jednotlivým podintervalům $\xi_{ujk} = \langle a_{ujk},\, b_{ujk})$ náleží váhy p_{ujk}, kde $k = 1,\ \ldots,\, s_{uj}$, $j = 1, 2$ a $u = 1,\ \ldots,\, m$. Dále platí, že $\sum_{k=1}^{s_{uj}} p_{ujk} = 1$.

Poznámka 2.9. Uvědomme si, že se v případě, kdy $s_{uj} = 1$ a $p_{ujk} = 1\ \forall j,\, k,\, u$, opět dostáváme k intervalovým proměnným. Z toho důvodu jsou následující úvahy a definice rozšiřující analogií k příslušným úvahám a definicím intervalových proměnných.

Definice 2.13. Empirickou sdruženou funkci hustoty definujeme pro dvojici histogramových proměnných $(Z_1,\, Z_2)$ v bodě $(\xi_1,\, \xi_2)$ jako

$$f(\xi_1,\, \xi_2) = \frac{1}{m} \sum_{u \in E} \left\{ \sum_{k_1=1}^{s_{u1}} \sum_{k_2=2}^{s_{u2}} \frac{p_{u1k_1} \cdot p_{u2k_2} \cdot I_{k_1k_2}(\xi_1,\, \xi_2)}{\| Z_{k_1k_2}(u) \|} \right\}, \tag{15}$$

kde $\| Z_{k_1k_2}(u) \|$ odpovídá ploše obdélníku $Z_{k_1k_2}(u) = \langle a_{u1k_1},\, b_{u1k_1}) \times \langle a_{u2k_2},\, b_{u2k_2})$ a $I_{k_1k_2}(\cdot,\, \cdot)$ označuje indikátorovou funkci, tedy

$$I_{k_1k_2}(\xi_1,\, \xi_2) = \begin{cases} 1 & \text{pro } (\xi_1,\, \xi_2) \in Z_{k_1k_2}(u) \\ 0 & \text{jinak.} \end{cases}$$

Definice 2.14. Sdružený histogram histogramových proměnných Z_1 a Z_2 získáme vykreslením množiny dvojic $\{R_{g_1g_2},\, p_{g_1g_2}\}$ nad obdélníky $R_{g_1g_2}$ vzniklé kartézským součinem dvou intervalů, tedy $R_{g_1g_2} = \langle \zeta_{1,g_1-1},\, \zeta_{1g_1}) \times \langle \zeta_{2,g_2-1},\, \zeta_{2g_2})$, pro $g_1 = 1,\ \ldots,\, r_1,\, g_2 = 1,\ \ldots,\, r_2$, kde

$$p_{g_1g_2} = \frac{f_{g_1g_2}}{m}, \tag{16}$$

přičemž

$$f_{g_1g_2} = \sum_{u \in E} \sum_{k_1 \in Z(g_1)} \sum_{k_2 \in Z(g_2)} \frac{\| Z\,(k_1,\, k_2;\, u) \cap R_{g_1g_2} \|}{\| Z\,(k_1,\, k_2;\, u)\,\|} \cdot p_{u1k_1} \cdot p_{u2k_2}. \tag{17}$$

$Z\,(g_j)$ reprezentuje všechny intervaly $Z\,(k_j;\, u) \equiv \left\langle a_{ujk_j},\, b_{ujk_j} \right\rangle$, $j = 1,\, 2$, které se překrývají s obdélníkem $R_{g_1g_2}$ pro danou kategorii u.

Poznámka 2.10. Důsledkem výše uvedeného je, že každý člen součtu zavedeného vztahem (17) odpovídá proporci pozorovaného obdélníku $Z\,(k_1,\, k_2;\, u)$, který se pro každé $(g_1,\, g_2)$ překrývá s obdélníkem $R_{g_1g_2}$.

Příklad 2.4. Zabývejme se datovým souborem [22] obsahujícím informace o průměrné výši plného starobního důchodu v jednotlivých okresech České republiky v prosinci roku 2015. Původní datový soubor (`starobni_duchod_puvodni.xls`) je na přiloženém CD. Dále uvažujme datový soubor [23] popisující absolutní počty osob pobírajících sirotčí důchod. Data se rovněž vztahují k jednotlivým okresům České republiky a odpovídají témuž časovému období. Původní datový soubor (`sirotci_duchod_puvodni.xls`) je taktéž k dispozici na přiloženém CD.

Z těchto dvou datových souborů vytvoříme agregací vzhledem ke 13 krajům České republiky jeden soubor, který bude složen ze dvou histogramových proměnných. Kraj „Hlavní město Praha“ nebudeme uvažovat ze stejného důvodu, jako v příkladu 2.3, tedy $m = 13$.

První z proměnných, která bude popisovat proporcionální zastoupení průměrné výše plného starobního důchodu, označíme dále Z_1. Druhou z proměnných,

Tabulka 16: Datový soubor: Důchody – 1. část.

kraj	Průměrná výše důchodu (Z_1)		
	⟨10 400, 10 900)	⟨10 900, 11 400)	⟨11 400, 11 900⟩
Středočeský	0,000	0,500	0,500
Jihočeský	0,000	0,857	0,143
Plzeňský	0,000	0,857	0,143
Karlovarský	0,333	0,667	0,000
Ústecký	0,143	0,571	0,286
Liberecký	0,000	1,000	0,000
Královéhradecký	0,000	0,800	0,200
Pardubický	0,000	1,000	0,000
Vysočina	0,200	0,800	0,000
Jihomoravský	0,286	0,571	0,143
Olomoucký	0,400	0,600	0,000
Zlínský	0,000	1,000	0,000
Moravskoslezský	0,167	0,333	0,500

Tabulka 17: Datový soubor: Důchody – 2. část.

kraj	Počet osob (Z_2)		
	⟨150, 650)	⟨650, 1 150)	⟨1 150, 1 650⟩
Středočeský	1,000	0,000	0,000
Jihočeský	0,857	0,143	0,000
Plzeňský	1,000	0,000	0,000
Karlovarský	1,000	0,000	0,000
Ústecký	1,000	0,000	0,000
Liberecký	0,750	0,250	0,000
Královéhradecký	1,000	0,000	0,000
Pardubický	1,000	0,000	0,000
Vysočina	1,000	0,000	0,000
Jihomoravský	0,714	0,143	0,143
Olomoucký	0,800	0,200	0,000
Zlínský	0,750	0,250	0,000
Moravskoslezský	0,167	0,500	0,333

vztahující se k proporcionálnímu zastoupení počtu osob pobírajících sirotčí důchod, označíme jako Z_2. Obě proměnné vzniknou tak, že jednotlivé hodnoty, ať už

průměrné výše důchodu či počty osob, nahradíme intervaly s příslušnými pravděpodobnostmi. Odpovídající datový soubor zobrazují tabulky 16 a 17. Z takto zadaných tabulek je možné vyčíst, že průměrná výše starobního důchodu se pro Olomoucký kraj pohybuje ve 40 % případů ze zjištěných průměrných výší v intervalu $\langle 10\ 400, 10\ 900)$ a v 60 % v intervalu $\langle 10\ 900, 11\ 400)$. Počet osob pobírajících sirotčí důchod je pak v 80 % zjištěných absolutních počtů v rozmezí $\langle 150, 650)$ a zbývajících 20 % v rozmezí $\langle 650, 1\ 150)$.

K sestrojení sdruženého histogramu pro dvojici proměnných (Z_1, Z_2) je nejprve zapotřebí určit rozmezí nabývaných hodnot obou proměnných. Průměrná výše důchodu nabývá hodnot z intervalu $\langle 10\ 400, 11\ 900\rangle$, zatímco počet osob pobírajících sirotčí důchod se pohybuje v rozmezí $\langle 150, 1\ 650\rangle$. Na základě Sturgesova pravidla rozdělíme každý z intervalů na pět podintervalů. Nyní je možné určit obdélníky $R_{g_1 g_2}$, kde $g_1 = 1, \ldots, 5$, $g_2 = 1, \ldots, 5$. Dostáváme tak 25 obdélníků $\langle 10\ 400, 10\ 700) \times \langle 150, 450)$, $\ldots$, $\langle 10\ 400, 10\ 700) \times \langle 1\ 350, 1\ 650\rangle$, $\ldots$, $\langle 11\ 600, 11\ 900\rangle \times \langle 1\ 350, 1\ 650\rangle$. Z důvodu vysokého počtu možných variant je nutností provést výpočty softwarem R, vycházíme přitom ze vztahů (16) a (17). Data ve formátu `csv` uložíme do proměnné `duchody` a zadáme:

```
duchody = duchody[,2:7]
p1 = matrix(c(duchody[,1],duchody[,2],duchody[,3]),
     nrow = 13,ncol = 3)
b1 = c(10400,10900,11400,11900)
body1 = c(10400, 10700, 11000, 11300, 11600, 11900)

p2 = matrix(c(duchody[,4],duchody[,5],duchody[,6]),
     nrow = 13,ncol = 3)
b2 = c(150, 650, 1150, 1650)
body2 = c(150, 450, 750, 1050, 1350, 1650)

inters <- function(int1,int2){
  d = min(c(int1[2],int2[2])) - max(c(int1[1],int2[1]))
```

```
  if(d < 0){
    d = 0
  }
  return(d)
}

f = matrix(0,5,5)
for(k in 1:13){
  f1 = matrix(0,5,1)
  f2 = matrix(0,5,1)
  for(i in 1:5){
    int1 = body1[i:(i+1)]
    int2 = body2[i:(i+1)]
    for(j in 1:3){
      f1[i] = f1[i] + inters(int1,
        b1[j:(j+1)])/(b1[j+1] - b1[j]) * p1[k,j]
      f2[i] = f2[i] + inters(int2,
        b2[j:(j+1)])/(b2[j+1] - b2[j]) * p2[k,j]
    }
  }

  for(i in 1:5){
    for(j in 1:5){
      f[i,j] = f[i,j] + f1[i]*f2[j]
    }
  }
}
```

Uvedený kód představuje jistou analogii ke kódu uvedeném v příkladu 2.3. Výsledné hodnoty sepíšeme do tabulky 18.

K sestrojení sdruženého histogramu je však zapotřebí určit relativní četnosti.

Tabulka 18: Četnosti datového souboru Důchody.

Z_1	Z_2				
	⟨150, 450)	⟨450, 750)	⟨750, 1050)	⟨1050, 1 350)	⟨1 350, 1 650⟩
⟨10 400, 10 700)	0,442	0,319	0,073	0,048	0,035
⟨10 700, 11 000)	1,300	0,922	0,168	0,087	0,046
⟨11 000, 11 300)	3,015	2,129	0,357	0,165	0,069
⟨11 300, 11 600)	1,350	0,963	0,189	0,108	0,068
⟨11 600, 11 900⟩	0,517	0,380	0,105	0,080	0,067

Ty získáme vydělením jednotlivých hodnot z tabulky 18 číslem m. Z těchto hodnot pak sestavíme tabulku 19 a využijeme je ke grafickému znázornění sdruženého histogramu (viz obrázek 4).

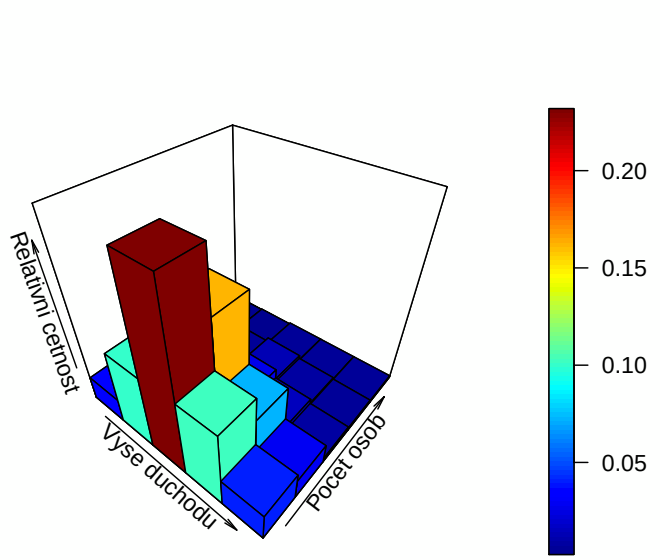
```
m = 13
r = f/m

library("plot3D")
z = r
hist3D(z=z, border="black", xlab = "Vyse duchodu",
       ylab = "Pocet osob", zlab="Relativni cetnost")
```

Tabulka 19: Relativní četnosti datového souboru Důchody.

Z_1	Z_2				
	⟨150, 450)	⟨450, 750)	⟨750, 1050)	⟨1050, 1 350)	⟨1 350, 1 650⟩
⟨10 400, 10 700)	0,034	0,025	0,006	0,004	0,003
⟨10 700, 11 000)	0,100	0,071	0,013	0,007	0,004
⟨11 000, 11 300)	0,232	0,164	0,027	0,013	0,005
⟨11 300, 11 600)	0,104	0,074	0,015	0,008	0,005
⟨11 600, 11 900⟩	0,040	0,029	0,008	0,006	0,005

Zbývá nám určit marginální a relativní četnosti. Marginální četnosti proměnné Z_1, resp. Z_2 počítáme jako řádkové, resp. sloupcové součty z tabulky 18. V softwaru zadáme:

```
radky = c()
sloupce = c()
for (i in 1:5){
```



Obrázek 4: Sdružený histogram datového souboru Důchody.

```
  radky[i] = sum(f[i,])
  sloupce[i] = sum(f[,i])
}
m = 13
rel_radky = radky/m
rel_sloupce = sloupce/m
```

Tak získáme zbývající požadované četnosti, které sepíšeme do tabulek 20 a 21.

Tabulka 20: Marginální a relativní četnosti proměnné Z_1.

Z_1	Četnost Z_1	Rel. četnost Z_1
⟨10 400, 10 700)	0,917	0,071
⟨10 700, 11 000)	2,523	0,194
⟨11 000, 11 300)	5,734	0,441
⟨11 300, 11 600)	2,677	0,206
⟨11 600, 11 900⟩	1,149	0,088
$\sum$	13	1

Tabulka 21: Marginální a relativní četnosti proměnné Z_2.

Z_2	Četnost Z_2	Rel. četnost Z_2
$\langle 150, 450)$	6,623	0,509
$\langle 450, 750)$	4,712	0,362
$\langle 750, 1\ 050)$	0,891	0,069
$\langle 1\ 050, 1\ 350)$	0,488	0,038
$\langle 1\ 350, 1\ 650\rangle$	0,286	0,022
$\sum$	13	1

Těmito tabulkami uzavřeme příklad i kapitolu věnovanou popisným charakteristikám modální proměnné.

3. Spojitost symbolických a kompozičních dat

Relativní charakter modální vícehodnotové proměnné a histogramové proměnné nás přivádí k tzv. kompozičním datům, se kterými se stručně seznámíme v rámci této kapitoly. Dále objasníme zmíněnou spojitost mezi kompozičními daty a symbolickými daty, kterými jsme se zabývali v kapitolách 1 a 2. Popsané myšlenky pak budou mít zcela zásadní význam pro další orientaci práce. Budeme zde vycházet z [12] a [16].

3.1. Úvod do kompozičních dat

Nejprve je potřeba zdůraznit, že kompoziční data popisují části určitého celku a nesou pouze relativní informaci. Přitom se může jednat o proporce, procenta, koncentrace či četnosti složek představujících podíly na celku. K základním pojmům této problematiky patří bezpochyby kompozice a kompoziční vektor, jejichž definicemi začneme.

Poznámka 3.1. Pro přehlednější zápis použijeme dva typy závorek. Závorka $[\ldots]$ označuje kompoziční data v simplexovém prostoru (viz dále), zatímco závorkou $(\ldots)$ máme na mysli vektor v prostoru reálných čísel.

Definice 3.1. Řádkový vektor $\mathbf{x} = [x_1, x_2, \ldots, x_D]$ je D-složkovou kompozicí (zkráceně kompozicí), jestliže všechny složky tohoto vektoru jsou kladná reálná čísla nesoucí pouze relativní informaci.

Definice 3.2. Sloupcový vektor $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]^T$, tvořený p D-složkovými kompozicemi, je p-složkový kompoziční vektor (zkráceně kompoziční vektor).

Definice 3.3. Množina kompozičních vektorů tvoří vícenásobnou kompoziční matici.

Nejčastějším případem, se kterým se v rámci kompozičních dat můžeme setkat, jsou tzv. uzavřená data. Jedná se o data s konstantním součtem κ. Takto obdržíme například $\kappa = 1$ (v případě proporcí) či $\kappa = 100$ (v případě procent), nicméně, konstantní součet κ může odpovídat libovolnému kladnému číslu. Samotné „uzavření“ kompoozice na předepsaný počet složek je pak formálně definováno takto:

Definice 3.4. Pro libovolný vektor o D kladných reálných složkách

$$\mathbf{x} = (x_1, x_2, \ldots, x_D) \in \mathbb{R}_+^D,\ x_i > 0,\ i = 1, 2, \ldots, D,$$

je uzávěr vektoru $\mathbf{x}$ s konstantním součtem $\kappa > 0$ definován jako

$$C(\mathbf{x}) = \left[\frac{\kappa \cdot x_1}{\sum_{i=1}^D x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^D x_i}, \ \ldots, \frac{\kappa \cdot x_D}{\sum_{i=1}^D x_i} \right].$$

Poznámka 3.2. V případě konstantního součtu $\kappa = 1$ dostáváme:

$$C(\mathbf{x}) = \left[\frac{x_1}{\sum_{i=1}^D x_i}, \frac{x_2}{\sum_{i=1}^D x_i}, \ \ldots, \frac{x_D}{\sum_{i=1}^D x_i} \right].$$

Výsledkem této operace je normování původního vektoru, a to tak, aby složky nově vzniklého vektoru dávaly v součtu κ. Vzhledem k tomu, že data mohou být díky vhodné normovací konstantě reprezentována proporcemi, budeme dále uvažovat kompoziční data bez újmy na obecnosti právě v této formě.

Ve spojení s kompozičními daty se setkáváme s jejich příslušným výběrovým prostorem, jehož definici nyní uvedeme.

Definice 3.5. Výběrovým prostorem kompozičních dat je D-rozměrný simplex (zkráceně simplex)

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

Použití statistických metod na soubory kompozic s sebou vzhledem k jejich geometrickým vlastnostem přináší jisté obtíže. Všechny tyto metody navíc musí splňovat tři základní podmínky, kterými jsou invariance na změnu měřítka, invariance na permutaci a podkompoziční soudržnost. Vzhledem k poslední podmínce je zapotřebí zavést pojem podkompozice.

Definice 3.6. Nechť je dána kompozice $\mathbf{x}$ a množina indexů $S = \{i_1, \ldots, i_s\}$. Podkompozici $\mathbf{x}_S$, obsahující s složek, získáme uplatněním operace uzávěr na podvektor $[x_{i_1}, x_{i_2}, \ldots, x_{i_s}]$ vektoru $\mathbf{x}$. Množina indexů S označuje, které složky kompozice $\mathbf{x}$ jsou vybrány, nemusí jít nutně o prvních s složek.

Výše uvedené podmínky, stručně řečeno, vypovídají o tom, že výsledky použitých statistických metod nemohou být v žádném případě ovlivněny libovolností reprezentace, či pořadím složek kompozice, a že závěry získané pro celou kompozici nemůžou být v rozporu se závěry odpovídající podkompozici. Poslední podmínka také říká, že podkompozice by se měla chovat jako ortogonální projekce ve standardním euklidovském prostoru.

Další zvláštností kompozičních dat je, že klasickou euklidovskou geometrii nahrazujeme tzv. Aitchisonovou geometrií na simplexu, a to zejména kvůli relativní povaze kompozičních dat. K tomu, aby simplex získal strukturu vektorového prostoru, je zapotřebí zavést příslušné operace, tzv. perturbaci a mocninnou transformaci. V prvním případě se jedná o analogii ke sčítání v euklidovském prostoru, v druhém pak o analogii násobení skalárem v témže prostoru. Než přejdeme k samotným definicím, uveďme, že výše zmíněná Aitchisonova geometrie na simplexu nese jméno po jedné z nejvýznamnějších osobností související s kompozičními daty. Právě J. Aitchison, a zejména jeho práce [1], inspiruje, a pro mnohé autory dokonce vytváří ideovou základnu kompozičního přístupu dodnes. Nejinak je tomu v případě publikací autorů uvedených v této práci. Dalšími zásadními podklady jsou bezesporu [10] či [11].

Definice 3.7. Perturbace kompozic $\mathbf{x} \in S^D$ a $\mathbf{y} \in S^D$ je opět kompozice

$$\mathbf{x} \oplus \mathbf{y} = C\left(x_1 \cdot y_1, x_2 \cdot y_2, \ldots, x_D \cdot y_D\right) \in S^D. \tag{18}$$

Definice 3.8. Mocninná transformace kompozice $\mathbf{x} \in S^D$ konstantou $\beta \in \mathbb{R}$ je opět kompozice

$$\beta \odot \mathbf{x} = C\left(x_1^\beta,\, x_2^\beta,\, \ldots,\, x_D^\beta\right) \in S^D. \tag{19}$$

Simplex s těmito operacemi vytváří vektorový prostor $\left(S^D,\, \oplus\,, \odot\right)$. Jinými slovy, $\left(S^D,\, \oplus\right)$ je komutativní grupa a operace $\odot$ splňuje vlastnosti vnějšího součinu.

Poznámka 3.3. Na základě vztahů (18) a (19) pak můžeme psát

$$\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus ((-1) \odot \mathbf{y}) = C\left(\frac{x_1}{y_1},\, \frac{x_2}{y_2},\, \ldots,\, \frac{x_D}{y_D}\right), \tag{20}$$

přičemž značení C ve výše uvedených vztazích (18), (19) a (20) odpovídá uzávěru z definice 3.4.

Dalším tématem, o kterém se v rámci úvodu do kompozičních dat zmíníme, bude „transformace“ kompozic. Transformaci však v tomto kontextu chápejme jako vyjádření pozorování v reálném souřadnicovém systému, což nám umožní používat dále standardní statistické metody. Ve spojitosti s kompozičními daty jsou základními typy souřadnic *alr* (additive logratio) a *ilr* (isometric logratio) souřadnice, které odpovídají vyjádření vzhledem k bázi, a dále pak *clr* (centered logratio) koeficienty, odpovídající vyjádření vzhledem ke generujícímu systému. Více například v [12]. Pro pozdější účely si uvedeme pouze vyjádření v clr koeficientech.

Definice 3.9. Nechť je dána kompozice $\mathbf{x} \in S^D$. *Clr* koeficienty kompozice jsou definovány pomocí zobrazení $clr : S^D \to \mathbb{R}^D$, přičemž

$$clr\,(\mathbf{x}) = \mathbf{z} = (z_1,\, z_2,\, \ldots,\, z_D) = \left(\ln \frac{x_1}{g\,(\mathbf{x})},\, \ln \frac{x_2}{g\,(\mathbf{x})},\, \ldots,\, \ln \frac{x_D}{g\,(\mathbf{x})}\right), \tag{21}$$

kde $g\,(\mathbf{x}) = \sqrt[D]{\prod_{i=1}^{D} x_i}$ značí geometrický průměr kompozice $\mathbf{x}$.

Poznámka 3.4. Příslušné inverzní zobrazení $clr^{-1} : \mathbb{R}^D \to S^D$ je tvaru:

$$\begin{aligned} clr^{-1}(\mathbf{z}) &= \mathbf{x} = [x_1,\, x_2,\, \ldots,\, x_D] \\ &= \left[\frac{\exp(z_1)}{\sum_{k=1}^{D} \exp(z_k)},\, \frac{\exp(z_2)}{\sum_{k=1}^{D} \exp(z_k)},\, \ldots,\, \frac{\exp(z_D)}{\sum_{k=1}^{D} \exp(z_k)} \right]. \end{aligned}$$

Se znalostí vyjádření v *clr* koeficientech se nyní můžeme věnovat definicím skalárního součinu, normy a vzdálenosti, které nám zajistí vlastnosti euklidovského vektorového prostoru pro Aitchisonovu geometrii. Dolní index S^D bude zdůrazňovat uvažovaný výběrový prostor kompozičních dat.

Definice 3.10. Aitchisonův skalární součin kompozic $\mathbf{x} \in S^D$ a $\mathbf{y} \in S^D$ je tvaru

$$\langle \mathbf{x},\, \mathbf{y} \rangle_{S^D} = \langle clr(\mathbf{x}),\, clr(\mathbf{y}) \rangle = \sum_{i=1}^{D} \ln \frac{x_i}{g(\mathbf{x})} \cdot \ln \frac{y_i}{g(\mathbf{y})}, \tag{22}$$

Aitchisonovu normu kompozice $\mathbf{x} \in S^D$ definujeme jako

$$\|\mathbf{x}\|_{S^D} = \sqrt{\langle \mathbf{x},\, \mathbf{x} \rangle_{S^D}} \tag{23}$$

a Aitchisonova vzdálenost mezi kompozicemi $\mathbf{x} \in S^D$ a $\mathbf{y} \in S^D$ je

$$d_{S^D}(\mathbf{x},\, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_{S^D}^2 = \sqrt{\sum_{i=1}^{D} \left(\ln \frac{x_i}{g(\mathbf{x})} - \ln \frac{y_i}{g(\mathbf{y})} \right)^2}, \tag{24}$$

přičemž $g(\mathbf{x})$, resp. $g(\mathbf{y})$ označuje geometrický průměr kompozice $\mathbf{x}$, resp. geometrický průměr kompozice $\mathbf{y}$.

Poznámka 3.5. Dodejme, že Aitchisonův skalární součin kompozic $\mathbf{x} \in S^D$ a $\mathbf{y} \in S^D$ je možné vyjádřit v různých, leč navzájem ekvivalentních podobách, např.:

$$\begin{aligned} \langle \mathbf{x},\mathbf{y} \rangle_{S^D} &= \frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \ln \frac{x_i}{x_j} \cdot \ln \frac{y_i}{y_j} = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \cdot \ln \frac{y_i}{y_j} \\ &= \sum_{i=1}^{D} \ln x_i \cdot \ln y_i - \frac{1}{D} \left(\sum_{j=1}^{D} \ln x_j \right) \cdot \left(\sum_{k=1}^{D} \ln y_k \right). \end{aligned}$$

Touto poznámkou jsme zakončili podkapitolu věnující se stručnému seznámení se s kompozičními daty. Z uvedených definic a poznámek vycházíme později v kapitole 4.2, která je zaměřena na popis jedné z metod vícerozměrné statistiky, a to z kompozičního hlediska.

3.2. Provázanost symbolických a kompozičních dat

Jak bylo uvedeno v úvodu této kapitoly, zaměříme se nyní na spojitost mezi symbolickými daty a kompozičními daty. Bez průtahů uveďme, že kompoziční data představují speciální případ symbolických dat. Přesněji řečeno, kompoziční data odpovídají modálním vícehodnotovým proměnným (viz definice 1.3).

V rámci symbolického přístupu jsme se setkali s tím, že je možné rozlišit modální vícehodnotovou proměnnou a histogramovou proměnnou. Použité dělení, a především následná realizace statistických analýz s proměnnými těchto dvou typů však může místy působit těžkopádně, a není ojedinělou situací, kdy je s histogramovými proměnnými zacházeno stejně jako s modálními vícehodnotovými proměnnými. Jako příklad můžeme uvést vzdálenost mezi dvěma objekty. Co se stane, když rozmezí naměřených jednotek nahradíme slovním ohodnocením? Místo intervalu $\langle 0, 1)$ odpovídající vzdálenosti v kilometrech bychom takovou vdálenost mohli nazvat „malá“. Tím z histogramové proměnné vytvoříme modální vícehodnotovou proměnnou. Podobných příkladů bychom mohli uvést celou řadu.

Na tuto záležitost proto budeme dále pohlížet z jiného úhlu a využijeme kompozičního přístupu. Tento přístup zmíněné typy proměnných nerozlišuje a v obou případech pracuje s jediným konceptem. Kompoziční metodika nám navíc umožní provést analýzy vícerozměrných dat, jejichž symbolická implementace by nebyla úplně přímočará.

Existuje veliké množství metod mnohorozměrné statistiky, kterými bychom se mohli dále zabývat. V rámci této práce si vybereme metodu hlavních komponent (viz kapitola 4), neboť patří ke zcela klíčovým mnohorozměrným metodám.

Závěrem dodejme, že ve prospěch použití kompozičního přístupu hovoří také nedostatek postačující literatury věnující se metodě hlavních komponent modifikované pro symbolická data. Návrhy takto upravené metody můžeme nalézt v [5] (pouze pro intervalová data) a v [8] (pouze pro histogramová data).

4. PCA tradičně i netradičně

V této kapitole popíšeme základní myšlenky metody hlavních komponent, kterou však častěji známe pod označením PCA (z anglického „Principal component analysis“). Pro úspornější zápis používejme dále tohoto označení.

Kapitolu rozdělíme do dvou stěžejních částí. Nejprve si ve stručnosti zopakujeme podstatu metody v rámci standardního přístupu, kdy budeme vycházet především z [7] a [16]. Následně se na metodu zaměříme z kompozičního pohledu a teoretickou část, čerpající z [12] a [16], doplníme i praktickou ukázkou použití metody.

4.1. Standardní přístup

PCA patří k velice užívaným nástrojům vícerozměrné statistiky. Svou nezastupitelnou roli hraje zejména při průzkumové analýze vícerozměrných dat a jejím cílem je redukce dimenze datových souborů. Za použití ortogonální transformace dochází k přeměně množiny původních korelovaných proměnných na množinu nekorelovaných proměnných. Takto vzniklé proměnné představují lineární kombinace původních proměnných a označujeme je jako hlavní komponenty. Je žádoucí, aby hlavní komponenty postupně vyčerpávaly největší část zbývající variability v datech, tedy aby první hlavní komponenta vysvětlila co nejvíce z celkové variability dat, a aby na poslední hlavní komponentu zůstal jen malý nevysvětlený zbytek. Celkovou variabilitou přitom rozumíme součet rozptylů všech proměnných, které zkoumáme. Obecně pak platí, že výsledný počet komponent odpovídá počtu vstupních proměnných. Z hlediska vizualizace stačí uvažovat nejvýše tři hlavní komponenty, často se však můžeme setkat s případem tří až čtyř hlavních komponent. Jejich větší počet by ovšem pro nás vzhledem k primárnímu cíli, tj. redukci dimenze, nebyl příliš přijatelný.

Hlavní komponenty představují vážený součet hodnot vstupních proměnných, ovšem s určitými omezeními. První podmínka udává, že čtverce vah jednotlivých komponent musí v součtu odpovídat jedné, druhá podmínka zaručuje vzájem-

nou nekorelovanost jednotlivých komponent. Zmíněné váhy pak označujeme jako komponentní zátěže a lze je vysvětlovat jako kovariance, příp. korelační koeficienty mezi vstupními proměnnými a nově získanými komponentami, a to v závislosti na použité datové matici. V případě nestejných jednotek či zcela rozdílných rozptylů vstupních proměnných se doporučuje, aby PCA vycházela z výběrové korelační matice. Pokud to ale není nutné, tedy v případě srovnatelných jednotek původních proměnných, je výhodnější pracovat s výběrovou varianční maticí. Zdůrazněme, že hlavní komponenty vycházející z varianční matice jsou odlišné od komponent získané z korelační matice.

Závěrem uveďme, že klíčovou úlohu v rámci využití PCA mají hodnoty označované jako komponentní skóry. Jedná se o hodnoty (souřadnice) hlavních komponent spočtené pro jednotlivá pozorování. Mohou být využity například při hledání odlehlých hodnot, nebo při ověřování předpokladů vztahujících se k vícerozměrným datům. Pro detailnější výklad odkazujeme zájemce např. na [6] nebo [7].

4.2. PCA s kompozičními vektory

Kompoziční data, která jsme si stručně představili v podkapitole 3.1, se vyskytují v mnoha oblastech. Jako příklad můžeme uvést situaci, kdy se zajímáme o koncentraci D chemických látek v půdě. V tomto kontextu uvažujme p konkrétních vrstev půdního profilu, ve kterých byla měření provedena, přičemž počet pozorování označíme standardně jako n. Našim cílem bude použití PCA k posouzení vnitřní struktury p kompozičních proměnných, nikoli však $p \cdot D$ složek.

Po uvážení okolností zjistíme, že se nacházíme ve svízelné situaci. Pokud bychom na popsaný problém použili standardní PCA, stručně naznačenou v podkapitole 4.1, dostaneme vzhledem k odlišným geometrickým vlastnostem kompozičních dat sporné výsledky. Navíc v této situaci také cítíme potřebu zabývat se PCA přizpůsobenou kompozičním vektorům (viz definice 3.2), nikoli pouze kompozicím (viz definice 3.1). Právě této problematice je věnována následující podkapitola, přičemž teoretické jádro doplníme o praktickou aplikaci na datech

zmíněných v předchozím odstavci.

Uvažujme vícenásobnou kompoziční matici (viz definice 3.3), kterou dále označíme jako $\mathbf{U}_{n\times(D\cdot p)}$. Tuto matici tvoří n kompozičních vektorů, kdy každý z nich obsahuje p D-složkových kompozic. Jednoduše řečeno, jedná se o matici, jejíž prvky jsou celé kompozice. Tuto matici můžeme zapsat následovně

$$\mathbf{U}_{n\times(D\cdot p)} = (\mathbf{U}_1,\, \mathbf{U}_2,\, \ldots,\, \mathbf{U}_p) = \begin{pmatrix} \mathbf{O}_1^T \\ \mathbf{O}_2^T \\ \vdots \\ \mathbf{O}_n^T \end{pmatrix} = \begin{pmatrix} \mathbf{u}_{11} & \mathbf{u}_{12} & \cdots & \mathbf{u}_{1p} \\ \mathbf{u}_{21} & \mathbf{u}_{22} & \cdots & \mathbf{u}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{u}_{n1} & \mathbf{u}_{n2} & \cdots & \mathbf{u}_{np} \end{pmatrix}. \tag{25}$$

Označením $\mathbf{u}_{ij} = [u_{ij1},\, u_{ij2},\, \ldots,\, u_{ijD}]$, kde $\mathbf{u}_{ij} \in S^D$, rozumíme D-složkovou kompozici, která se nachází v i-tém řádku a v j-tém sloupci matice $\mathbf{U}_{n\times(D\cdot p)}$. Konkrétní j-tá proměnná, skládající se z n D-složkových kompozic, je označena $\mathbf{U}_j = \left(\mathbf{u}_{1j}^T,\, \mathbf{u}_{2j}^T,\, \ldots,\, \mathbf{u}_{nj}^T\right)^T$ a odpovídá j-té submatici matice $\mathbf{U}_{n\times(D\cdot p)}$. Tyto submatice jsou pak seřazené za sebou ve sloupcích. Naopak i-tý kompoziční vektor, obsahující p D-složkových kompozic, značíme $\mathbf{O}_i = (\mathbf{u}_{i1},\, \mathbf{u}_{i2},\, \ldots,\, \mathbf{u}_{ip})^T$, kde $\mathbf{O}_i^T \in S^D \times S^D \times \ldots \times S^D = S^{Dp}$ je i-tý řádek téže matice.

Nyní nás budou zajímat operace, které můžeme s kompozičními vektory provést. Uvažujme kompoziční vektory $\mathbf{O}_i^T$ a $\mathbf{O}_{i'}^T$ a konstantu $\beta \in \mathbb{R}$. Zavedeme

$$\mathbf{O}_i^T \oplus \mathbf{O}_{i'}^T = (\mathbf{u}_{i1} \oplus \mathbf{u}_{i'1},\, \mathbf{u}_{i2} \oplus \mathbf{u}_{i'2},\, \ldots,\, \mathbf{u}_{ip} \oplus \mathbf{u}_{i'p}), \tag{26}$$

$$\beta \odot \mathbf{O}_i^T = (\beta \odot \mathbf{u}_{i1},\, \beta \odot \mathbf{u}_{i2},\, \ldots,\, \beta \odot \mathbf{u}_{ip}), \forall \beta \in \mathbb{R}, \tag{27}$$

$$\mathbf{O}_i^T \ominus \mathbf{O}_{i'}^T = (\mathbf{u}_{i1} \ominus \mathbf{u}_{i'1},\, \mathbf{u}_{i2} \ominus \mathbf{u}_{i'2},\, \ldots,\, \mathbf{u}_{ip} \ominus \mathbf{u}_{i'p}), \tag{28}$$

$$\left\langle \mathbf{O}_i^T,\, \mathbf{O}_{i'}^T \right\rangle_{S^{Dp}} = \sum_{j=1}^{p} \left\langle \mathbf{u}_{ij},\, \mathbf{u}_{i'j} \right\rangle_{S^D}, \tag{29}$$

$$\|\mathbf{O}_i^T\|_{S^{Dp}}^2 = \sum_{j=1}^{p} \|\mathbf{u}_{ij}\|_{S^D}^2, \tag{30}$$

$$d_{S^{Dp}}^2\left(\mathbf{O}_i^T,\, \mathbf{O}_{i'}^T\right) = \sum_{j=1}^{p} d_{S^D}^2\left(\mathbf{u}_{ij},\, \mathbf{u}_{i'j}\right). \tag{31}$$

Poznámka 4.1. Výše uvedené operace vychází ze vztahů (18)–(20) a (22)–(24). Platí, že skalární součin kompozičních vektorů (viz (29)) splňuje pozitivní definitnost, symetrii a linearitu. Pro libovolné kompoziční vektory $\mathbf{O}_i^T$, $\mathbf{O}_{i'}^T$ a $\mathbf{O}_{i''}^T \in S^{Dp}$ pak matematický zápis těchto vlastností vypadá postupně takto:

1. $\left\langle \mathbf{O}_i^T, \mathbf{O}_i^T \right\rangle_{S^{Dp}} \geq 0$, přičemž rovnost nastává pouze v případě, kdy $\mathbf{O}_i^T = \left(\left[\frac{1}{D}, \frac{1}{D}, \ldots, \frac{1}{D}\right], \left[\frac{1}{D}, \frac{1}{D}, \ldots, \frac{1}{D}\right], \ldots, \left[\frac{1}{D}, \frac{1}{D}, \ldots, \frac{1}{D}\right] \right)$,
2. $\left\langle \mathbf{O}_i^T, \mathbf{O}_{i'}^T \right\rangle_{S^{Dp}} = \left\langle \mathbf{O}_{i'}^T, \mathbf{O}_i^T \right\rangle_{S^{Dp}}$,
3. $\left\langle \mathbf{O}_i^T \oplus \mathbf{O}_{i'}^T, \mathbf{O}_{i''}^T \right\rangle_{S^{Dp}} = \left\langle \mathbf{O}_i^T, \mathbf{O}_{i''}^T \right\rangle_{S^{Dp}} + \left\langle \mathbf{O}_{i'}^T, \mathbf{O}_{i''}^T \right\rangle_{S^{Dp}}$ a $\left\langle \beta \odot \mathbf{O}_i^T, \mathbf{O}_{i'}^T \right\rangle_{S^{Dp}} = \beta \cdot \left\langle \mathbf{O}_i^T, \mathbf{O}_{i'}^T \right\rangle_{S^{Dp}}, \forall \beta \in \mathbb{R}$.

Jedná se tedy o vlastnosti identické skalárnímu součinu standardních vektorů.

Definice 4.1. Uvažujme kompoziční proměnnou $\mathbf{U}_j$ o n pozorováních. Výběrové centrum této proměnné je ve tvaru

$$E_{S^D}(\mathbf{U}_j) = \bar{\mathbf{u}}_j = C\left(g(\mathbf{L}_{j1}), g(\mathbf{L}_{j2}), \ldots, g(\mathbf{L}_{jD})\right) \tag{32}$$

a výběrový celkový rozptyl odpovídá vztahu

$$Var_{S^D}(\mathbf{U}_j) = \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{u}_{ij} \ominus \bar{\mathbf{u}}_j\|_{S^D}^2, \tag{33}$$

přičemž označením C rozumíme uzávěr z definice 3.4, $g(\cdot)$ označuje geometrický průměr a $\mathbf{L}_{jk} = (u_{1jk}, u_{2jk}, \ldots, u_{njk})^T$, kde $k = 1, \ldots, D$.

Poznámka 4.2. Uvědomme si, čemu odpovídají příslušné indexy. Označením $\mathbf{L}_{jk}$ rozumíme n-rozměrný číselný vektor vztahující se ke k-té složce D-složkové kompozice j-té proměnné.

Ve spojitosti s výběrovým celkovým rozptylem kompoziční proměnné $\mathbf{U}_j$ uveďme jeho tvar, který je možné také uvažovat:

$$totVar\left(\mathbf{U}_j\right) = \sum_{k=1}^{D} var\left[clr_k\left(\mathbf{U}_j\right)\right]. \tag{34}$$

Označení $clr_k\left(\mathbf{U}_j\right)$ odpovídá k-tému sloupci matice s transformovanými daty $clr\left(\mathbf{U}_j\right)$, kterou lze zapsat jako

$$clr\left(\mathbf{U}_j\right) = \begin{pmatrix} clr\left(\mathbf{u}_{1j}\right) \\ clr\left(\mathbf{u}_{2j}\right) \\ \vdots \\ clr\left(\mathbf{u}_{nj}\right) \end{pmatrix}. \tag{35}$$

Na základě tohoto označení pak není obtížné přesvědčit se o tom, že výběrový celkový rozptyl, určený vztahem (33), přímo odpovídá vztahu (34), neboť

$$Var_{S^D}\left(\mathbf{U}_j\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{D}\left[clr_k\left(\mathbf{u}_{ij}\right) - clr_k\left(\bar{\mathbf{u}}_j\right)\right]^2 = \sum_{k=1}^{D} var\left[clr_k\left(\mathbf{U}_j\right)\right]. \tag{36}$$

Dále má smysl zabývat se centrováním uvažované j-té proměnné $\mathbf{U}_j$. K tomu je zapotřebí zavést označení $cen\left(\mathbf{U}\right)_j = \left(\bar{\mathbf{u}}_j, \bar{\mathbf{u}}_j, , \ldots, \bar{\mathbf{u}}_j\right)^T$. Centrovaná data pak získáme provedením operace

$$\mathbf{U}_j \ominus cen\left(\mathbf{U}\right)_j. \tag{37}$$

Alternativní možností přípravy dat k dalším analýzám je standardizace, tedy v tomto případě

$$\frac{1}{\sqrt{Var_{S^D}\left(\mathbf{U}_j\right)}} \odot \left(\mathbf{U}_j \ominus cen\left(\mathbf{U}\right)_j\right). \tag{38}$$

To, kterou z variant standardizace použijeme, a jestli vůbec, záleží především na povaze vstupních dat. U kompozičních vektorů v podobě, ve které s nimi pracujeme ovšem není třeba přihlížet k jednotkám a měřítku dat.

Definice 4.2. Uvažujme kompoziční proměnné $\mathbf{U}_j$ a $\mathbf{U}_{j'}$. Výběrová kovariance těchto proměnných je definována jako

$$Cov_{S^D}\left(\mathbf{U}_j, \mathbf{U}_{j'}\right) = \frac{1}{n}\sum_{i=1}^{n}\left\langle \mathbf{u}_{ij} \ominus \bar{\mathbf{u}}_j, \mathbf{u}_{ij'} \ominus \bar{\mathbf{u}}_{j'}\right\rangle_{S^D} \tag{39}$$

a výběrový korelační koeficient je

$$r_{S^D}(\mathbf{U}_j, \mathbf{U}_{j'}) = \frac{Cov_{S^D}(\mathbf{U}_j, \mathbf{U}_{j'})}{\sqrt{Var_{S^D}(\mathbf{U}_j)} \cdot \sqrt{Var_{S^D}(\mathbf{U}_{j'})}}. \tag{40}$$

Zaměřme se nyní na právě zadefinovaný výběrový korelační koeficient, který má vlastnosti standardního korelačního koeficientu. Pokud lineární korelovanost zaznačíme jako $\mathbf{U}_j = \beta \odot \mathbf{U}_{j'} \oplus \mathbf{U}_{j''}$, kde β představuje reálné číslo a $\mathbf{U}_{j''}$ odpovídá kompozičnímu vektoru (stejných) konstant, tak můžeme tvrdit, že kompoziční proměnné $\mathbf{U}_j$ a $\mathbf{U}_{j'}$ jsou lineárně závislé právě tehdy, když pro výběrový korelační koeficient platí, že

$$r_{S^D}(\mathbf{U}_j, \mathbf{U}_{j'}) = \begin{cases} 1 & \text{pro } \beta > 0 \\ -1 & \text{pro } \beta < 0. \end{cases}$$

Další úvahy povedou k definici kompozičního podvektoru. Uvažme euklidovský prostor $\mathbb{R}^p$. Jeho podmnožinu můžeme označit jako $\mathbb{R}^q$, ovšem za předpokladu, že $q \leq p$. Obdobně i v případě podmnožiny uvažového prostoru S^{Dp}.

Definice 4.3. Nechť je dán kompoziční vektor $\mathbf{O}_i = (\mathbf{u}_{i1}, \mathbf{u}_{i2}, \ldots, \mathbf{u}_{ip})^T \in S^{Dp}$ obsahující p D-složkových kompozic $[u_{ij1}, u_{ij2}, \ldots, u_{ijD}]$. Kompoziční podvektor je vektor $\mathbf{O}_i^{p^*} = (\mathbf{u}_{i1^*}, \mathbf{u}_{i2^*}, \ldots, \mathbf{u}_{ip^*})^T \in S^{Dp^*}$ o p^* kompozicích, přičemž množina indexů $(1^*, 2^*, \ldots, p^*)$ určuje, které kompozice jsou v podvektoru vybrány, nemusí se nutně jednat o prvních p^* kompozic.

Je nutné zdůraznit, že hodnoty výběrových charakteristik se nemohou lišit v závislosti na tom, zda k výpočtům použijeme celý kompoziční vektor, nebo pouze jeho část – tedy podvektor. Uvědomme si, že definice 4.3 se soustředí na redukci dimenze p a nikoli D. Jinými slovy, nesnižujeme počet složek v rámci jednotlivých kompozic, ale v rámci jednotlivých kompozičních vektorů. Oproti tomu je třeba rozlišit pojem „podkompoziční vektor“[2], který redukuje počet složek D. V takovém případě budou výsledné hodnoty výběrových charakteristik nutně rozdílné.

[2]Rozlišujme pojmy „podkompoziční vektor“ a „podkompozice“, uvedený v definici 3.6.

Před detailním popisem samotného algoritmu PCA ještě uveďme důležitou vlastnost výběrového průměru a výběrového rozptylu, a to i s jejich důkazem.

Věta 4.1. *Pro libovolné dvě kompoziční proměnné* $\mathbf{U}_j$ *a* $\mathbf{U}_{j'}$ *platí, že*

a) $E_{S^D}(\beta \odot \mathbf{U}_j \oplus \mathbf{U}_{j'}) = \beta \odot E_{S^D}(\mathbf{U}_j) \oplus E_{S^D}(\mathbf{U}_{j'})$,

b) $Var_{S^D}(\beta \odot \mathbf{U}_j \oplus \mathbf{U}_{j'}) = \beta^2 Var_{S^D}(\mathbf{U}_j) + 2\beta Cov_{S^D}(\mathbf{U}_j, \mathbf{U}_{j'}) + Var_{S^D}(\mathbf{U}_{j'})$.

Důkaz: Nejprve dokážeme vlastnost *a*), a to podrobným rozepsáním levé strany rovnice. Využijeme přitom definic perturbace a mocniné transformace kompozic (viz definice 3.7 a 3.8) a také definici výběrového centra kompoziční proměnné (viz definice 4.1):

$$
\begin{aligned}
& E_{S^D}(\beta \odot \mathbf{U}_j \oplus \mathbf{U}_{j'}) = \\
&= C\left(\sqrt[n]{\prod_{k=1}^{n} u_{jk1}^{\beta} \cdot u_{j'k1}},\ \ldots,\ \sqrt[n]{\prod_{k=1}^{n} u_{jkD}^{\beta} \cdot u_{j'kD}}\right) = \\
&= C\left(\left(\sqrt[n]{\prod_{k=1}^{n} u_{jk1}}\right)^{\beta} \cdot \prod_{k=1}^{n} u_{j'k1},\ \ldots,\ \left(\sqrt[n]{\prod_{k=1}^{n} u_{jkD}}\right)^{\beta} \cdot \prod_{k=1}^{n} u_{j'kD}\right) = \\
&= \beta \odot E_{S^D}(\mathbf{U}_j) \oplus E_{S^D}(\mathbf{U}_{j'}).
\end{aligned}
$$

Dále dokážeme vlastnost *b*), a to s využitím vztahu (36) a definic 3.7 a 3.8. Opět vyjdeme z levé strany rovnice:

$$
\begin{aligned}
& Var_{S^D}(\beta \odot \mathbf{U}_j \oplus \mathbf{U}_{j'}) = \\
&= \sum_{k=1}^{D} var\left(clr_k(\beta \odot \mathbf{U}_j \oplus \mathbf{U}_{j'})\right) = \sum_{k=1}^{D} var\left(\beta clr_k(\mathbf{U}_j) + clr_k(\mathbf{U}_{j'})\right) = \\
&= \sum_{k=1}^{D} \left[var\left(\beta clr_k(\mathbf{U}_j)\right) + var\left(clr_k(\mathbf{U}_{j'})\right) + 2cov\left(\beta clr_k(\mathbf{U}_j),\, clr_k(\mathbf{U}_{j'})\right)\right] = \\
&= \beta^2 Var_{S^D}(\mathbf{U}_j) + Var_{S^D}(\mathbf{U}_{j'}) + 2\sum_{k=1}^{D} cov\left(\beta clr_k(\mathbf{U}_j),\, clr_k(\mathbf{U}_{j'})\right). \qquad (41)
\end{aligned}
$$

Nyní upravíme poslední člen:

$$\sum_{k=1}^{D} cov\left(\beta clr_k\left(\mathbf{U}_j\right),\, clr_k\left(\mathbf{U}_{j'}\right)\right) = \beta \sum_{k=1}^{D} cov\left(clr_k\left(\mathbf{U}_j\right),\, clr_k\left(\mathbf{U}_{j'}\right)\right) =$$

$$= \beta \sum_{k=1}^{D} \left[E\left(clr_k\left(\mathbf{U}_j\right) \cdot clr_k\left(\mathbf{U}_{j'}\right)\right) - E\left(clr_k\left(\mathbf{U}_j\right)\right) E\left(clr_k\left(\mathbf{U}_{j'}\right)\right)\right] =$$

$$= \beta \sum_{k=1}^{D} E\left[\left(clr_k\left(\mathbf{U}_j\right) - E\left(clr_k\left(\mathbf{U}_j\right)\right)\right)\left(clr_k\left(\mathbf{U}_{j'}\right) - E\left(clr_k\left(\mathbf{U}_{j'}\right)\right)\right)\right] =$$

$$= \beta \sum_{k=1}^{D} E\left[\left(clr_k\left(\mathbf{U}_j\right) - \bar{\mathbf{u}}_j\right)\left(clr_k\left(\mathbf{U}_{j'}\right) - \bar{\mathbf{u}}_{j'}\right)\right] =$$

$$= \beta \frac{1}{n} \sum_{s=1}^{n} \sum_{k=1}^{D} \left[\left(clr_k\left(\mathbf{u}_{sj}\right) - \bar{\mathbf{u}}_j\right)\left(clr_k\left(\mathbf{u}_{sj'}\right) - \bar{\mathbf{u}}_{j'}\right)\right] =$$

$$= \beta \frac{1}{n} \sum_{s=1}^{n} \left\langle \mathbf{u}_{sj} - \bar{\mathbf{u}}_j,\, \mathbf{u}_{sj'} - \bar{\mathbf{u}}_{j'} \right\rangle =$$

$$= \beta Cov_{S^D}\left(\mathbf{U}_j,\, \mathbf{U}_{j'}\right).$$

Dosazením upraveného tvaru posledního členu do vztahu (41) a přehozením pořadí posledních dvou členů v témž vztahu dostáváme tvrzení věty. □

Nyní již k samotnému algoritmu PCA. V dalším textu uvažujme n kompozičních vektorů, které jsou popsány p kompozičními proměnnými $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_p$. Předpokládejme, že se jedná o centrované proměnné. Stejně jako v klasickém případě PCA, získáme v rámci kompozičního přístupu hlavní komponenty jako lineární kombinace vstupních proměnných. Můžeme tedy uvést, že k-tá hlavní komponenta, kde $1 \leq k \leq p$, je lineární kombinací kompozičních proměnných $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_p$, což zapíšeme jako

$$\mathbf{V}_k = \bigoplus_{j=1}^{p} \left(e_{kj} \odot \mathbf{U}_j\right) = e_{k1} \odot \mathbf{U}_1 \oplus e_{k2} \odot \mathbf{U}_2 \oplus \ldots \oplus e_{kp} \odot \mathbf{U}_p. \tag{42}$$

Vektor $\mathbf{e}_k = (e_{k1}, e_{k2}, \ldots, e_{kp})^T$ musí být zvolen tak, aby rozptyl hlavní komponenty $\mathbf{V}_k$ byl co největší. Zároveň však musí být splněny podmínky $\|\mathbf{e}_k\| = 1$

a $\mathbf{e}_k^T \mathbf{e}_l = 0$, pro $l = 1, 2, \ldots, p$, $l \neq k$. Jinými slovy, u všech takových vektorů vyžadujeme ortonormalitu. Jak jsme uvedli výše, klíčovým úkolem je pro nás maximalizace rozptylu. Nejprve uveďme, jak lze zapsat rozptyl k-té hlavní komponenty.

$$\begin{aligned} Var_{S^D}(\mathbf{V}_k) &= Var_{S^D}(e_{k1} \odot \mathbf{U}_1 \oplus e_{k2} \odot \mathbf{U}_2 \oplus \ldots \oplus e_{kp} \odot \mathbf{U}_p) \\ &= \sum_{i=1}^{p} \sum_{j=1}^{p} e_{ki}\, e_{kj}\, Cov_{S^D}(\mathbf{U}_i, \mathbf{U}_j) \\ &= \mathbf{e}_k' \mathbf{W} \mathbf{e}_k, \end{aligned} \tag{43}$$

s tím, že matice $\mathbf{W}$ reprezentuje varianční matici kompozičních proměnných $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_p$.

Prvních m hlavních komponent, přičemž $m \leq p$, je určeno m ortonormálními vektory $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$. Tyto vektory maximalizují celkový rozptyl, tedy výraz $\sum_{k=1}^{m} Var_{S^D}(\mathbf{V}_k)$. Musí ovšem platit, že

$$Var_{S^D}(\mathbf{V}_1) \geq Var_{S^D}(\mathbf{V}_2) \geq \ldots \geq Var_{S^D}(\mathbf{V}_m).$$

Takto popsaný problém pak můžeme zapsat jako úlohu

$$\begin{cases} \text{maximalizovat funkci} & \sum_{k=1}^{m} \mathbf{e}_k' \mathbf{W} \mathbf{e}_k \quad \text{pro } \mathbf{e}_k \in \mathbb{R}^p,\ k = 1, 2, \ldots, m \\ \text{za podmínek} & \|\mathbf{e}_k\| = 1, \quad \text{pro } k = 1, 2, \ldots, m, \\ & \mathbf{e}_k'\, \mathbf{e}_l = 0, \quad \text{pro } k, l = 1, 2, \ldots, m,\ l \neq k, \\ & \mathbf{e}_1' \mathbf{W} \mathbf{e}_1 \geq \mathbf{e}_2' \mathbf{W} \mathbf{e}_2 \geq \ldots \geq \mathbf{e}_m' \mathbf{W} \mathbf{e}_m, \quad \text{pro } m \leq p. \end{cases}$$

Řešením dané úlohy jsou vektory $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$. Dostáváme tak m vlastních vektorů příslušných prvním m vlastním číslům varianční matice $\mathbf{W}$, označeným jako $\lambda_1, \lambda_2, \ldots, \lambda_m$. Platí, že $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m > 0$.

Poznámka 4.3. Poznamenejme, že PCA je takto i v případě kompozičních vektorů založena na spektrálním rozkladu varianční matice $\mathbf{W}$.

Ve stručnosti tedy shrňme postup PCA modifikované pro kompoziční vektory. Nejprve provedeme centrování kompozičních proměnných $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_p$

dle vztahu (37). V dalším kroku spočítáme s využitím vztahů (33) a (39) varianční matici $\mathbf{W}$. Následně provedeme její spektrální rozklad, a tak získáme m vlastních vektorů. Nakonec určíme k-tou hlavní komponentu pomocí předpisu $\mathbf{V}_k = \bigoplus_{j=1}^{p} (e_{kj} \odot \mathbf{U}_j)$, pro $k = 1, 2, \ldots, m$.

Popsanou metodou je možné snížit dimenzi prostoru S^{Dp} na S^{Dm}, pro $m \leq p$. Obdobně jako v předchozí kapitole, i nyní je nutné zvážit, zda ve výpočtech používat varianční matici, či korelační matici. Návod k rozhodnutí o volbě matice pak zůstává stejný jako v klasickém případě (viz podkapitola 4.2).

Poznámka 4.4. Uvažujme centrované kompoziční proměnné $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_p$. Z nich získáme m hlavních komponent $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_m$. Ty splňují následující vlastnosti:

$$
\begin{aligned}
E_{S^D}(\mathbf{V}_k) &= \left[\frac{1}{D}, \frac{1}{D}, \ldots, \frac{1}{D}\right], \quad 1 \leq k \leq p, \\
Var_{S^D}(\mathbf{V}_k) &= \lambda_k, \quad 1 \leq k \leq p, \\
Cov_{S^D}(\mathbf{V}_k, \mathbf{V}_l) &= 0, \quad 1 \leq k, l \leq p, l \neq k, \\
\mathbf{U}_j &= \bigoplus_{k=1}^{p} (e_{kj} \odot \mathbf{V}_k), \quad 1 \leq j \leq p, \\
\sum_{j=1}^{p} Var_{S^D}(\mathbf{U}_j) &= \sum_{k=1}^{p} Var_{S^D}(\mathbf{V}_k).
\end{aligned}
$$

Poznamenejme, že první z uvedených vztahů plyne z centrování původních proměnných, neboť i nově vzniklé komponenty musí mít tuto vlastnost. Dostáváme tak neutrální prvek na simplexu, odpovídající nulovému vektoru. Další dvě vlastnosti plynou ze samotné konstrukce komponent. Čtvrtý vztah pak z toho, jak vypadá kompoziční model (viz (42)). Poslední vlastnost souvisí s tím, že pracujeme s varianční maticí a vypovídá o tom, že součet rozptylů hlavních komponent je roven součtu rozptylů původních proměnných. Uvědomme si, že zde nepracujeme s rozptyly jednotlivých složek kompozic ale s celkovými rozptyly kompozic. Jelikož se jedná o jistou analogii, i důkazy těchto vlastností by byly obdobné jako pro případ klasické PCA.

Stejně jako v klasickém případě vypovídá hodnota $Var_{S^D}(\mathbf{V}_k)$ o rozptylu vysvětleném k-tou hlavní komponentou. Dále má smysl zabývat se ukazatelem CCR (z anglického Cumulative Contribution Rate). Příspěvek prvních m hlavních komponent k celkovému rozptylu je dán vztahem

$$CCR_m = \frac{\sum_{k=1}^{m} Var_{S^D}(\mathbf{V}_k)}{\sum_{j=1}^{p} Var_{S^D}(\mathbf{V}_j)} = \frac{\sum_{k=1}^{m} \lambda_k}{\sum_{j=1}^{p} \lambda_j}, \tag{44}$$

resp.

$$CCR_m = \frac{1}{p} \sum_{k=1}^{m} \lambda_k, \tag{45}$$

pokud místo varianční matice vycházíme z korelační matice.

Na závěr teoretické části této podkapitoly uveďme rovnici popisující vztah mezi k-tou hlavní komponentou $\mathbf{V}_k$ a j-tou kompoziční proměnnou $\mathbf{U}_j$. Vztah je dán tvarem

$$r_{S^D}(\mathbf{V}_k, \mathbf{U}_j) = \frac{\sqrt{\lambda_k}}{\sqrt{Var_{S^D}(\mathbf{U}_j)}} \cdot e_{kj}, \quad j \leq k \leq p, 1 \leq j \leq p. \tag{46}$$

Pokud by všechny původní kompoziční proměnné $\mathbf{U}_j$ byly standardizované, tj. $Var_{S^D}(\mathbf{U}_j) = 1$, vztah (46) můžeme zjednodušit a psát

$$r_{S^D}(\mathbf{V}_k, \mathbf{U}_j) = \sqrt{\lambda_k} \cdot e_{kj}.$$

4.3. Aplikace na Kola data

Příklad 4.1. Uvažujme Kola data vztahující se k měření koncentrací určitých chemických látek ve čtyřech odlišných vrstvách půdy. Jedná se o mechové patro, horizont O (nadložní organický horizont), horizont B (metamorfický horizont) a horizont C (půdotvorný substrát). Výčet vrstev uvádíme dle rostoucí hloubky, ve které je můžeme v půdě nalézt. Označení proměnných provedeme na základě abecedního uspořádání, tedy $\mathbf{U}_1 =$ horizont B, $\mathbf{U}_2 =$ horizont C,

$\mathbf{U}_3 =$ horizont O a $\mathbf{U}_4 =$ mechové patro. Více informací o půdním profilu a půdních horizontech lze najít např. ve [15] nebo [17].

Měření bylo prováděno v letech 1993 až 1998 na území Finska, Norska a Ruska. Příslušné datové soubory lze nalézt ve statistickém softwaru R v knihovně `mvoutlier` postupně pod názvy `bhorizon`, `chorizon`, `humus` a `moss`. Každý z datových souborů obsahuje kromě hodnot koncentrací také identifikátory. Jedná se o identifikační číslo daného pozorování a dále místo, kde bylo měření provedeno. Vzhledem k velkému objemu dat jsme se rozhodli postupovat následovně. Ze všech proměnných vybereme jen ty, které byly měřeny ve všech čtyřech vrstvách a zároveň spadají do kategorie geologických hlavních prvků. Vybereme tedy hliník, vápník, železo, draslík, hořčík, mangan, sodík, fosfor, síru a křemík. Při výběru odpovídajících prvků jsme vycházeli především z [13]. Ze všech pozorování nám dále budou vyhovovat jen ta, jejichž identifikátory si napříč všemi vrstvami také odpovídají. Příslušný kód je uveden v příloze C.1.

V dalším kroku si data vyjádříme v *clr* koeficientech, a to s využitím vztahu (21). K vyjádření použijeme tabulkový procesor Excel, případně odpovídající balíček a funkci v softwaru. Získáme datový soubor, který uložíme ve formátu `csv`, načteme a provedeme centrování. Datový soubor `puda_clr.csv`, ze kterého dále vycházíme, je k dispozici na přiloženém CD.

```
data = read.csv2("puda_clr.csv",header = TRUE)
for (i in 4:43){
  data[,i] = data[,i] - mean(data[,i])
}
```

Výsledný centrovaný soubor označíme jako `puda`. Pro výpočty ještě vynecháme identifikátory, tedy

```
puda =  data[,4:ncol(data)]
```

Dále pracujeme s datovým souborem, který obsahuje $n = 130$ pozorování, $p = 4$ kompozičních proměnných, přičemž každá z nich je složena z $D = 10$ složek. V tabulce 22 je zobrazeno několik prvních řádků jedné z vrstev.

Tabulka 22: Ukázka datového souboru: Kola data.

$\mathbf{U}_1$									
Al	Ca	Fe	K	Mg	Mn	Na	P	S	Si
−0,243	−0,281	0,247	1,029	−0,269	−0,409	−0,776	0,640	0,215	−0,153
−0,231	0,517	−0,588	−0,030	−0,198	0,354	0,347	0,073	−0,370	0,127
0,086	0,604	−0,074	−0,703	−0,046	−0,554	0,528	0,635	−0,140	−0,336
0,089	0,140	−0,432	−0,423	−0,399	0,217	0,324	0,519	−0,294	0,259
0,552	−0,323	−0,300	−0,275	−0,433	0,344	−0,191	0,180	0,373	0,072
0,371	0,217	−0,047	−0,589	−0,316	−0,489	0,481	0,242	0,125	0,005

Nyní spočítáme varianční matici $\mathbf{W}$, a to s využitím vztahů (36) a (39), tedy

```
W = matrix(0,4,4)
for (j in 1:4){
  for (k in (10*(j-1)+1):(10*j)){
    W[j,j] = W[j,j] + var(puda[,k])
  }
}
W = 0.5 * W
for (i in 1:4){
  for (j in (i+1):4){
    if (i<=3){
      for (k in 1:10){
        W[i,j] = W[i,j] + cov(puda[,10*(i-1)+k],puda[,10*(j-1)+k])
      }
    }
  }
}
W = W + t(W)
```

Jak je vidět z tabulky 23, která zobrazuje výslednou varianční matici $\mathbf{W}$, diagonální prvky této matice jsou stejného řádu, proto budeme dále vycházet právě z varianční matice, nikoli z korelační matice. Následně provedeme spektrální rozklad matice $\mathbf{W}$, a to příkazem

```
lambda = eigen(W)
```

Tabulka 23: Varianční matice **W**.

	$\mathbf{U}_1$	$\mathbf{U}_2$	$\mathbf{U}_3$	$\mathbf{U}_4$
$\mathbf{U}_1$	2,513	1,783	0,031	−0,156
$\mathbf{U}_2$	1,783	2,647	−0,091	−0,156
$\mathbf{U}_3$	0,031	−0,091	2,429	0,852
$\mathbf{U}_4$	−0,156	−0,156	0,852	1,836

Výstupem funkce `eigen` jsou vlastní čísla (`$values`) a vlastní vektory (`$vectors`). V našem případě dostáváme

```
$values
[1] 4.3911655 3.0169642 1.2291244 0.7879553


$vectors
            [,1]       [,2]        [,3]        [,4]
[1,] -0.68680918 -0.1045172 -0.05101154  0.71747274
[2,] -0.71528462 -0.0649468  0.11744490 -0.68582544
[3,]  0.06955316 -0.8152064 -0.56748630 -0.09252155
[4,]  0.10873559 -0.5659476  0.81336574  0.07947393
```

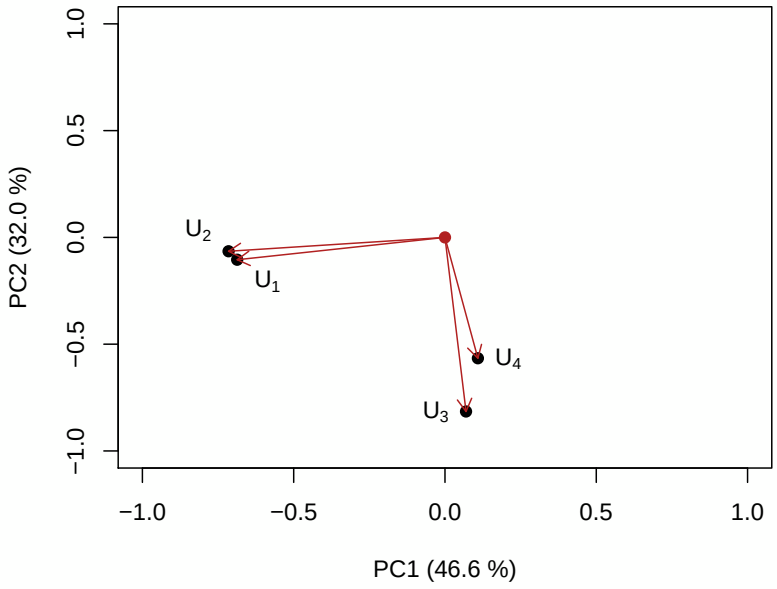
Na základě těchto hodnot je možné dopočítat procenta vysvětlené variability jednotlivých hlavních komponent, resp. hodnoty CCR dle vztahu (44).

```
vysv_var = 100*lambda$values / sum(lambda$values)
CCR = c(vysv_var[1],0,0,0)
for (i in 2:4){
  CCR[i] = CCR[i-1] + vysv_var[i]
}
```

Shrnutí výsledných hodnot nalezneme v tabulce 24. Dále určíme hodnoty skórů. Kód pro jejich výpočet uvádíme v příloze C.2. Na závěr provedeme grafické znázornění vztahu původních proměnných a prvních dvou hlavních komponent (viz obrázek 5). Příslušný kód je k dispozici v příloze C.3.

Tabulka 24: Rozptyly hlavních komponent a CCR.

PC	rozptyl PC	
	Hodnota	$CCR\,(\%)$
1	4,391	46,590
2	3,017	78,599
3	1,229	91,640
4	0,788	100,000



Obrázek 5: Biplot datového souboru: Kola data.

Interpretace uvedeného biplotu se neliší od interpretace biplotu v případě standardní PCA. Můžeme vidět silnou korelaci mezi proměnnými $\mathbf{U}_1$ a $\mathbf{U}_2$ a také mezi proměnnými $\mathbf{U}_3$ a $\mathbf{U}_4$. Zopakujme, že k analýze jsme využili vrstvy v tomto pořadí: horizont B, horizont C, horizont O a mechové patro. Tento závěr je, vzhledem k vlastnostem jednotlivých půdních horizontů a mechového patra, vcelku opodstatněný.

Nyní se zabývejme určitou modifikací tohoto příkladu, kdy nás budou zajímat

pouze tři složky ($D = 3$) tak, abychom skóry hlavních komponent mohli následně zobrazit v ternárním diagramu. Rozhodli jsme se pro volbu tří chemických látek, které se společně vyskytují v příkladech v rámci odborné literatury. Konkrétně se jedná o chemické prvky K (draslík), Mg (hořčík) a P (fosfor).

Co se týče provedení PCA, základní myšlenka postupu se oproti předchozí situaci nemění. Zdůrazněme však nutnost výběru požadovaných sloupců z původních, nikoli transformovaných dat. Data vyjádříme v *clr* koeficientech a opět provedeme centrování. Datový soubor `puda_clr3.csv`, ze kterého budeme v dalších výpočtech vycházet, je k dispozici na přiloženém CD. Do softwaru zadáme

```
data = read.csv2("puda_clr3.csv",header = TRUE)
for (i in 4:15){
  data[,i] = data[,i] - mean(data[,i])
}
puda =  data[,4:ncol(data)]
```

Rozdíl oproti předchozí situaci se projeví po výpočtu varianční matice $\mathbf{W}$, kterou získáme obdobným způsobem jako v předchozím případě (viz příloha C.4). Jak můžeme vidět v tabulce 25, v případě proměnné $\mathbf{U}_4$ se setkáváme s rozptylem jiného řádu než u ostatních proměnných.

Tabulka 25: Varianční matice $\mathbf{W}$ (pro $D = 3$).

	$\mathbf{U}_1$	$\mathbf{U}_2$	$\mathbf{U}_3$	$\mathbf{U}_4$
$\mathbf{U}_1$	0,505	0,331	0,073	0,005
$\mathbf{U}_2$	0,331	0,460	0,064	0,019
$\mathbf{U}_3$	0,073	0,064	0,194	0,060
$\mathbf{U}_4$	0,005	0,019	0,060	0,063

Tato skutečnost nás přivádí k myšlence spektrálního rozkladu korelační matice (viz tabulka 26) namísto varianční matice. Pro výpočet korelační matice zadáme

```
odchylky = sqrt(diag(W))
R = matrix(0,4,4)
```

```
for (i in 1:4){
  for (j in 1:4){
    R[i,j] = W[i,j] / (odchylky[i] * odchylky[j])
  }
}
```

Tabulka 26: Korelační matice $\mathbf{R}$ (pro $D = 3$).

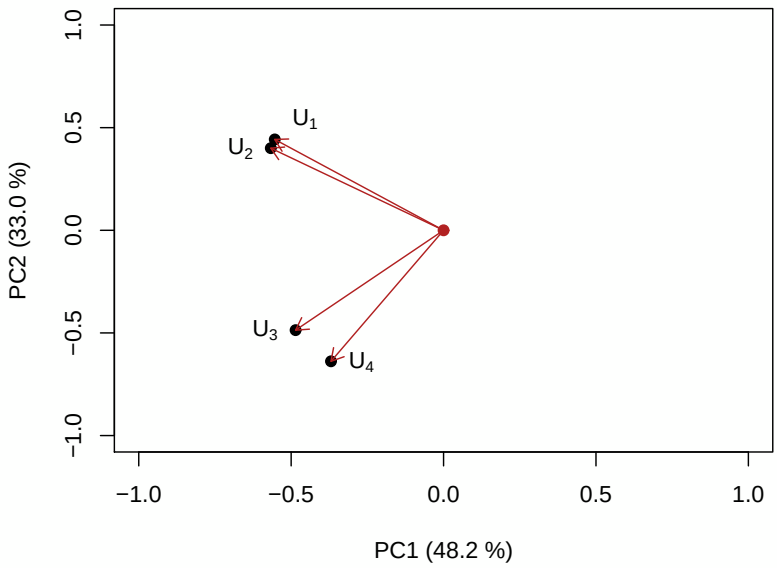
	$\mathbf{U}_1$	$\mathbf{U}_2$	$\mathbf{U}_3$	$\mathbf{U}_4$
$\mathbf{U}_1$	1,000	0,687	0,233	0,030
$\mathbf{U}_2$	0,687	1,000	0,214	0,112
$\mathbf{U}_3$	0,233	0,214	1,000	0,541
$\mathbf{U}_4$	0,030	0,112	0,541	1,000

Další postup je standardní jako v případě $D = 10$. Příslušný kód je uveden v příloze C.4. Pro zajímavost uvedeme rovněž grafické znázornění proměnných vzhledem k prvním dvěma hlavním komponentám (viz obrázek 6). To, až na mírnou rotaci zátěží, prakticky odpovídá situaci pro $D = 10$. Na závěr vykreslíme ternární diagramy pro první dvě hlavní komponenty (viz obrázky 7 a 8). K tomu využijeme napočítané hodnoty skórů, u nichž však aplikujeme inverzní *clr* transformaci. Té můžeme dosáhnout buď ručním výpočtem, případně za pomocí knihovny `robCompositions`, a to následovně

```
library("robCompositions")
V1_inv = constSum(cenLRinv(V1, useClassInfo = F),const=100)
V2_inv = constSum(cenLRinv(V2, useClassInfo = F),const=100)
```

Ke znázornění samotných ternárních diagramů pak slouží knihovna `ggtern` a její stejnojmenný příkaz. Zadáme

```
library("ggtern")
V1_inv_d = data.frame(V1_inv)
V2_inv_d = data.frame(V2_inv)
```

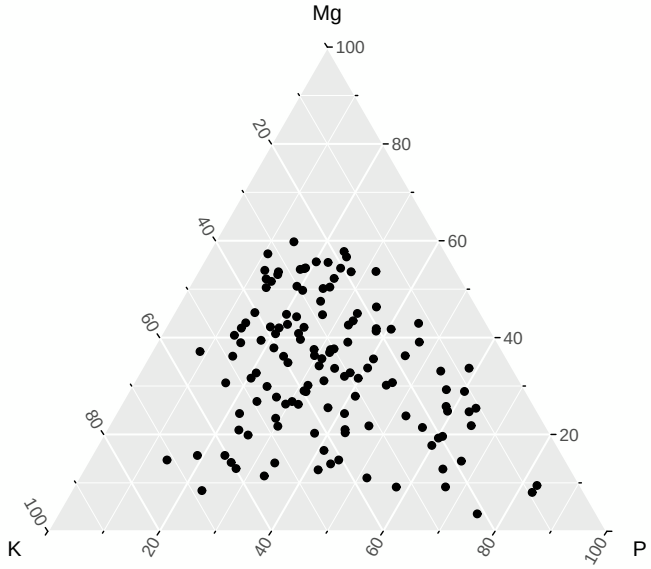


Obrázek 6: Biplot datového souboru: Kola data (pro $D = 3$).

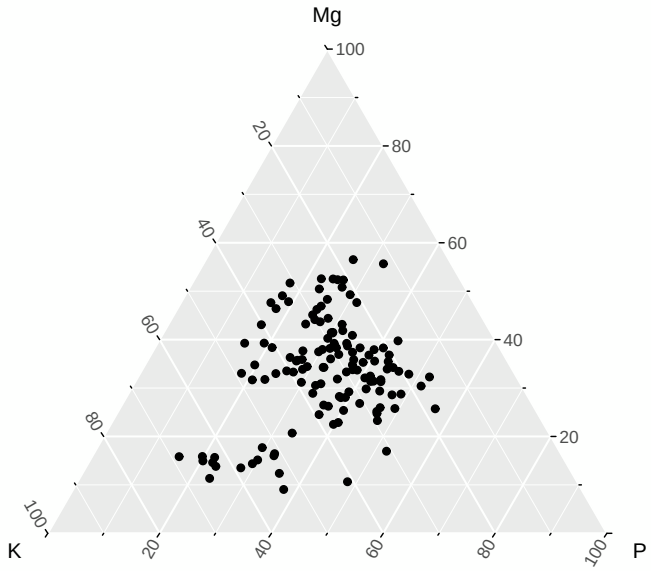
```
ggtern(data=V1_inv_d,aes(x=K,y=Mg,z=P))+geom_point(pch=19)
ggtern(data=V2_inv_d,aes(x=K,y=Mg,z=P))+geom_point(pch=19)
```

Co se týče interpretace diagramů 7 a 8, můžeme mezi sebou porovnávat jednotlivá pozorování z hlediska procentuálního zastoupení uvedených chemických prvků. Vzhledem k tomu, že každý z diagramů obsahuje 130 bodů, není efektivní zabývat se každým bodem zvlášť. Pro první hlavní komponentu uveďme, že v případě prvku K se procentuální zastoupení pohybuje v rozmezí $10 - 70$ % (s výjimkou čtyř pozorování), pro prvek Mg je rozmezí $0 - 60$ % a pro prvek P se, stejně jako v případě draslíku, jedná o rozmezí $10 - 70$ % (až na výjimku čtyř pozorování). Pro druhou hlavní komponentu je možné provést interpretaci na základě stejného klíče. Můžeme si všimnout zjevně menší variability dat, což ovšem odpovídá konstrukci hlavních komponent. Dalším postupem by mohla být identifikace odlehlých pozorování a jejich znázornění do mapy dle identifikátorů.

Příklad zakončíme právě zobrazením pozorování do mapy poloostrova Kola. Každému skóru první a následně i druhé hlavní komponenty nejprve přiřadíme

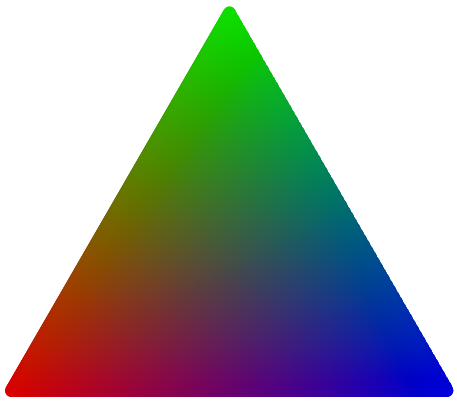


Obrázek 7: Ternární diagram pro skóry PC1.



Obrázek 8: Ternární diagram pro skóry PC2.

jinou barvu, a to na základě umístění v ternárním diagramu. Různobarevnost přitom zajistíme využitím barevného modelu RGB (viz obrázek 9). Kód grafického znázornění tohoto modelu byl převzat ze [14] a upraven.

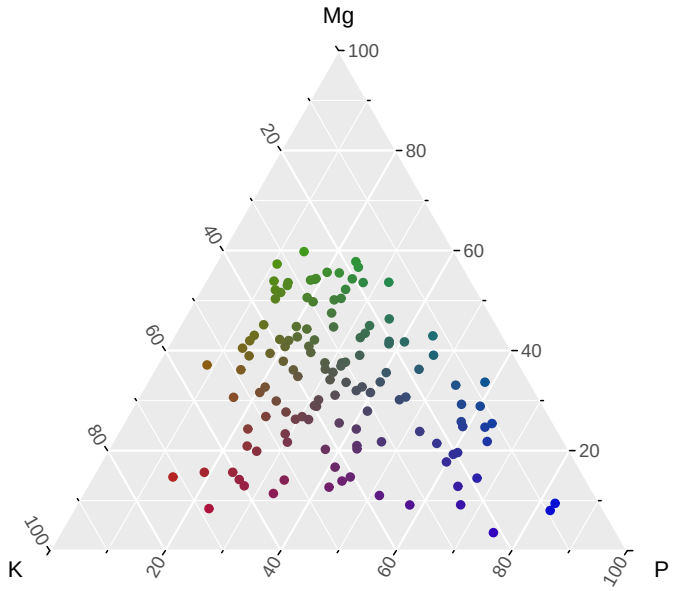


Obrázek 9: Barevný model RGB.

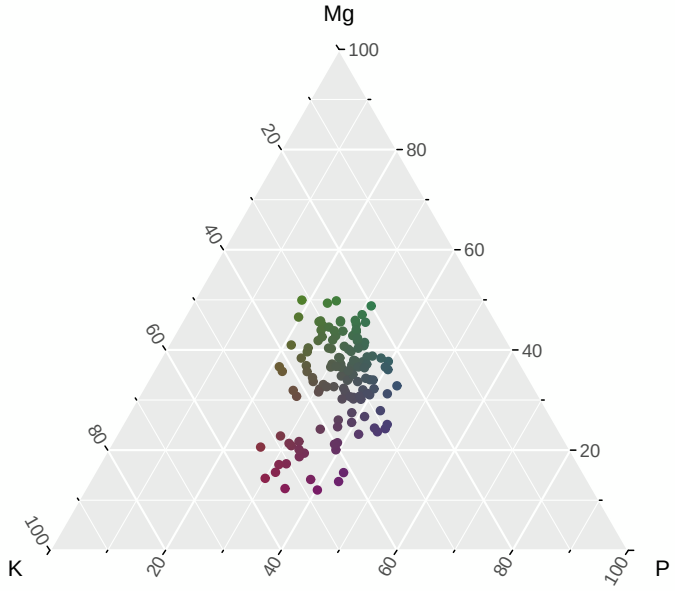
K vykreslení ternárních diagramů pro první i druhou hlavní komponentu použijeme opět příkaz `ggtern`. Uvádíme zde kód příslušící první hlavní komponentě, kód pro druhou hlavní komponentu je uveden v příloze C.5.

```
barvy = matrix(0,130,1)
for(i in 1:130){
  temp = V1_inv_d[i,]/100 *255
  barvy[i] = rgb(temp[1],temp[2],temp[3],maxColorValue=255)
}
ggtern(data=V1_inv_d,aes(x=K,y=Mg,z=P))+geom_point(pch=19,
       col=barvy)
```

Poznamenejme, že ač jsme se souřadnicemi jednotlivých pozorování prozatím nepracovali, při počátečních úpravách jsme hodnoty proměnných $XCOO$ a $YCOO$ převedli do tvaru souřadnic WGS84. V souborech `puda_clr.csv` a `puda_clr3.csv` tedy vystupuje místo proměnné $XCOO$ proměnná E (východní délka) a místo



Obrázek 10: Ternární diagram s rozlišením barev pro skóry PC1.

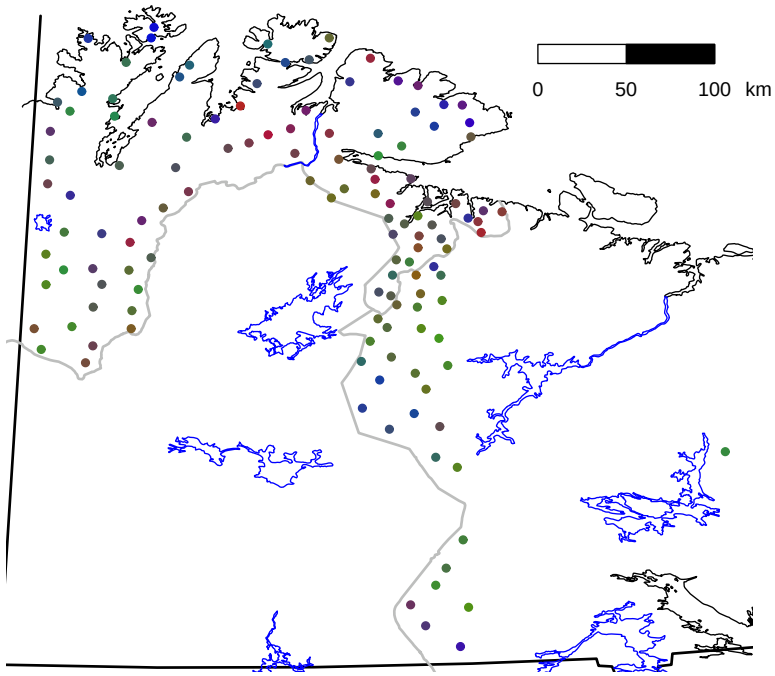


Obrázek 11: Ternární diagram s rozlišením barev pro skóry PC2.

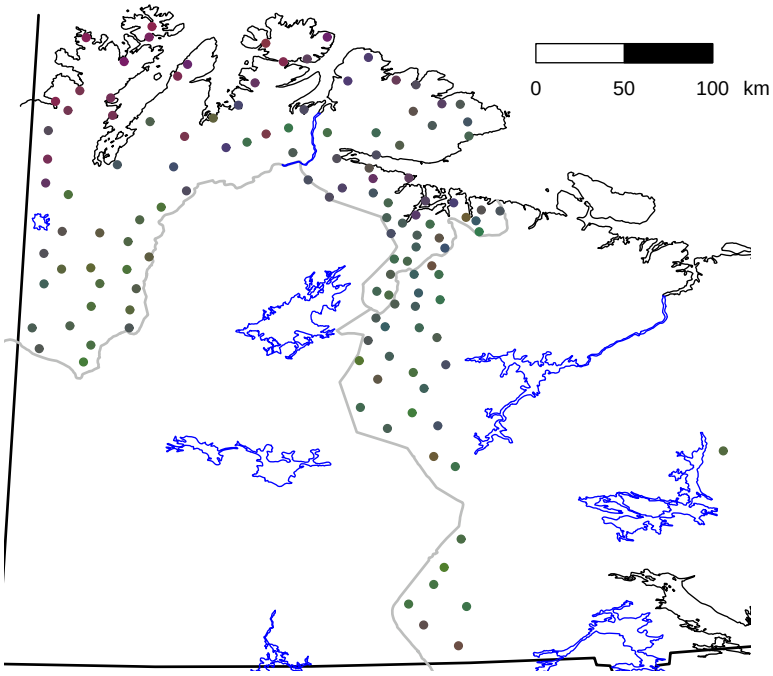
proměnné $YCOO$ uvažujeme proměnnou N (severní šířka). Vzhledem k dále po-

užitým příkazům `pkb` a `points`, kterými vykreslíme implementovanou mapu poloostrova Kola a barevně rozlišená pozorování, je nutné pracovat se souřadnicemi v nepřevedeném tvaru. Datový soubor s původními souřadnicemi (`kola.csv`) je k dispozici na přiloženém CD. Kód dokreslení měřítka mapy, stejně jako mapu v případě druhé hlavní komponenty, uvádíme v příloze C.5. Pro případ první hlavní komponenty zadáme

```
souradnice = read.csv2("kola.csv", header = TRUE)
souradnice = souradnice[,c(2,3)]
plot(souradnice,frame.plot=FALSE,xaxt="n",yaxt="n",
     xlab="",ylab="",type="n",cex.lab=1.2)
pkb(map.col=c("black","black","grey","blue"),add.plot=T)
points(souradnice,pch=20,cex=1.2,col=barvy)
```



Obrázek 12: Zobrazení první hlavní komponenty pro datový soubor Kola data.



Obrázek 13: Zobrazení druhé hlavní komponenty pro datový soubor Kola data.

Na obrázcích 12 a 13 je znázorněno 130 pozorování datového souboru Kola data v mapě stejnojmenného poloostrova. K detailní interpretaci dosažených výsledků by bylo zapotřebí se podrobně zabývat nejen členěním reliéfu poloostrova podle nadmořské výšky, ale také například tím, zda se jedná o přímořskou, či vnitrozemní oblast.

Nicméně, na obrázku 12 si na první pohled můžeme všimnout dvou jasně modrých bodů, které jsou izolovány na ostrově Magerøya v severní části mapy. Po ověření zjistíme, že se jedná o body, které nalezneme na obrázku 10 v pravém dolním rohu. Jde o pozorování s vysokým procentuálním zastoupením fosforu (v obou případech kolem 83 %). V severní části také nalezneme pozorování s nejvyšším procentuálním zastoupením draslíku (přibližně 71 %). Poznáme ho podle zářivě červené barvy (viz levý dolní roh obrázku 10). Naopak pozorování s relativně vyšším procentuálním zastoupením hořčíku (kolem 50 až 60 %) je možné

nalézt převážně ve střední či jižní části mapy, a to pod odstíny zelené barvy.

K interpretaci obrázku 13 je možné použít obdobný klíč jako u obrázku 12. Zdůrazněme však, že barevné zastoupení tentokrát není tak pestré jako v předchozím případě. Mapu by bylo možné vizuálně rozdělit na dvě části, a to na severní přímořskou oblast, která je zastoupena řadou červených a fialových bodů, a zbývající část, v níž převládá zelená a modrá barva. O tomto rozdělení se můžeme přesvědčit také na základě polohy skórů v obrázku 11, kde vidíme dvě elipsovité oblasti bodů.

Závěr

Tato práce se zabývala popisem možných přístupů analýzy SD ke statistické analýze různých typů symbolických proměnných. Jak název práce napovídá, zásadním pojmem pro analýzu byla v našem případě „vícerozměrnost“. První dvě kapitoly byly věnovány výhradně symbolickým proměnným. Uvedla jsem stručný úvod do dané problematiky, zabývala jsem se novými typy symbolických proměnných a následně jsem se zaměřila na jeden konkrétní typ symbolické proměnné. Třetí kapitola představovala myšlenkový most mezi symbolickými a kompozičními daty. Rovněž jsem se věnovala základním poznatkům o kompozičních datech. Poslední kapitola se po krátkém shrnutí PCA věnovala především modifikaci PCA pro případ kompozičních vektorů.

Celá práce je doplněna o vysvětlující příklady s komentáři. Výpočty byly ve většině případů provedeny ve statistickém softwaru R, až na drobné úpravy dat v tabulkovém procesu Excel. Odpovídající kódy jsou vloženy buď přímo v textu práce, nebo v příloze, na kterou se text odkazuje.

Největší přínos pro mě představovalo řešení uvedených příkladů, obzvláště příkladu obsaženého v poslední kapitole. Ač se jednalo o data implementovaná ve statistickém softwaru R, myslím, že jejich zpracování pěkně ilustrovalo předchozí teoretické poznatky. A doufám, že společně s ostatními částmi osloví i čtenáře této práce.

Literatura

[1] Aitchison, J., *The Statistical Analysis of Compositional Data*, London & New York: Chapman & Hall, 1986. ISBN 0-412-28060-4.

[2] Andrášiková, A., *Statistická analýza intervalových dat v symbolic data analysis*. Bakalářská práce – Univerzita Palackého v Olomouci, Přírodovědecká fakulta, 2015.

[3] Anděl, J., *Statistické metody*, 4. vydání, Praha: Matfyzpress, 2007. ISBN 80-7378-003-8.

[4] Billard, L., Diday, E., *From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis*, Journal of the American Statistical Association **98**(462), 470 – 487 (2003).

[5] Billard, L., Diday, E., *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Chichester: Wiley, 2006. ISBN 978-0-470-09016-9.

[6] Everitt, B., Hothorn, T., *An Introduction to Applied Multivariate Analysis with R*, New York: Springer, 2011. ISBN 978-1-4419-9649-7.

[7] Hebák, P., *Statistické myšlení a nástroje analýzy dat*, 2. vydání, Praha: Informatorium, 2015. ISBN 978-80-7333-118-4.

[8] Makosso-Kallyth, S., Diday, E., *Adaptation of interval PCA to symbolic histogram variables*, Advances in Data Analysis and Classification **6**(2), 147 – 159 (2012).

[9] Noirhomme-Fraiture, M., Brito, P., *Far Beyond the Classical Data Models: Symbolic Data Analysis*, Statistical Analysis and Data Mining **4**(2), 157 – 170 (2011).

[10] Pawlowsky-Glahn, V., Egozcue, J. J., *Geometric approach to statistical analysis on the simplex*, Stochastic Environmental Research and Risk Assessment **15**, 384 – 398 (2001).

[11] Pawlowsky-Glahn, V., Egozcue, J. J., *BLU Estimators and Compositional Data*, Mathematical Geology **34**(3), 259 – 274 (2002).

[12] Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R., *Modeling and Analysis of Compositional Data*, Chichester: Wiley, 2015. ISBN 978-1-118-44306-4.

[13] SGS, *Major elements*. [cit. 8. 3. 2017, 18:30]. Dostupné z: http:/www.sgs.com/en/mining/analytical-services/chemical-testing/major-elements.

[14] Stack Overflow, *Use RGB color triangle with ggtern() to spot a color.* [cit. 10. 4. 2017, 20:45]. Dostupné z: http:/stackoverflow.com/questions/42990044/use-rgb-color-triangle-with-ggtern-to-spot-a-color.

[15] Vítejte na Zemi, *Půdní horizonty – co je to půdní profil?* [cit. 7. 4. 2017, 10:15]. Dostupné z: http:/vitejtenazemi.cz/cenia/index.php?p=pudni_horizonty_co_je_to_pudni_profil&site=puda.

[16] Wang, H., Shangguan, L., Guan, R., Billard, L., *Principal component analysis for compositional data vectors*, Computational Statistics **30**(4), 1079 – 1096 (2015).

[17] Wikipedie, *Diagnostický půdní horizont.* [cit. 9. 3. 2017, 17:30]. Dostupné z: https://cs.wikipedia.org/wiki/Diagnostick%C3%BD_p%C5%AFdn%C3%AD_horizont.

Zdroje dat

[18] Eurostat, *Housing statistics.* [cit. 18. 2. 2017, 14:15]. Dostupné z: http:/ec.europa.eu/eurostat/statistics-explained/index.php/Housing_statistics.

[19] Český statistický úřad, *Návštěvnost HUZ podle kategorie - územní srovnání.* [cit. 19. 2. 2017, 18:40]. Dostupné z: https:/vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=vystup-objekt&z=T&f=TABULKA&katalog=31743&pvo=CRUD002&c=v3~8__RP2015&v=v90__KAT__154__1.

[20] Český statistický úřad, *Kapacity hromadných ubytovacích zařízení podle kategorie* - 14 tabulek dle krajů za období 2015. [cit. 19. 2. 2017, 19:00]. Dostupné z: https:/vdb.czso.cz/vdbvo2/faces/cs/index.jsf;jsessionid=24KD90Ln0hUzrfsVBaKY6M1YMkz7PhHbLZnTyGU6lwu6O3lYu-DY!-911420616?page=vystup-objekt-parametry&z=T&f=TABULKA&katalog=31742&sp=A&pvo=CRU01&u=v42__VUZEMI__100__3018&c=v3~8__RP2015&str=v42.

[21] Český statistický úřad, *Požáry - územní srovnání.* [cit. 23. 2. 2017, 21:00]. Dostupné z:

https:/vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=vystup-objekt&z=T&f=TABULKA&katalog=31008&pvo=KRI10&evo=v141_!_VUZEMI97-100-101_1&c=v3~8__RP2015#w=.

[22] Český statistický úřad, *Průměrná výše důchodu - územní srovnání*. [cit. 25. 2. 2017, 10:30]. Dostupné z: https:/vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=vystup-objekt&pvo=SZB06b&skupId=468&str=v404&evo=v260_!_VUZEMI97-100-101hal_1#w=.

[23] Český statistický úřad, *Příjemci důchodů - územní srovnání*. [cit. 25. 2. 2017, 10:45]. Dostupné z: https:/vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=vystup-objekt&z=T&f=TABULKA&skupId=468&pvo=SZB06a&str=v69&evo=v286_!_VUZEMI97-100-101hal_1&c=v287~4__RP2015MP12.

Přílohy

A. Datový soubor: Ubytování

A.1. Grafické znázornění

```
library("plot3D")
z = cetnostiA

library(RColorBrewer)
color = brewer.pal(8, "GnBu")

hist3D(z=z, border="black",zlab = "Absolutni cetnost",
       col=color, xlab="", ylab="")
text3D(x = c(-1,0), y = rep(-0.6, 2), z = rep(-0.5, 2),
       labels = c("rezidenti","nerezidenti"),
       add = TRUE, adj = 0)
text3D(x = c(1.9,2.2,2.2,2.2),   y = c(0.2,0.7,1.0,1.4),
       z = c(0.1,0.1,0.1,0.1),
       labels  = c("hotel","penzion","kemp","ostatni"),
       add = TRUE, adj = 1)
```

B. Datový soubor: Požáry

B.1. Grafické znázornění

```
vektor = c(p_g)
popis = round(p_g,2)

g = barplot(vektor,ylim = c(0,1),space = 0,col=heat.colors(5),
    ylab="Relativni cetnost")
axis(1, at = c(0:5),labels=c(0,12,24,36,48,60))
text(x = g, y = vektor, label = popis, pos = 3, col = "black")
```

B.2. Výpočet symbolického výběrového průměru a rozptylu

```
pom = c()
for (i in 1:13){
  pom[i] = ((0 + 20)/2)*pozary[i,1] +
+ ((20 + 40)/2)*pozary[i,2] +
+ ((40 + 60)/2)*pozary[i,3]
}
prumer = (sum(pom)/m)

pom2 = c()
for (i in 1:13){
  pom2[i] = ((0^2 + 0*20 + 20^2)/3)*pozary[i,1] +
+ ((20^2 + 20*40 + 40^2)/3)*pozary[i,2] +
+ ((40^2 + 40*60 + 60^2)/3)*pozary[i,3]
}
rozptyl = (sum(pom2)/m) - prumer^2
odchylka = sqrt(rozptyl)
```

C. Datový soubor: Kola data

C.1. Příprava dat

```
library("mvoutlier")
b = bhorizon
c = chorizon
o = humus
m = moss
nazvy = names(m)
indexy = matrix(0,34,4)
indexy[,4] = 1:34
for(i in 1:34){
  indexy[i,1] = max(c((1:length(b))[names(b) == nazvy[i]],0))
  indexy[i,2] = max(c((1:length(c))[names(c) == nazvy[i]],0))
  indexy[i,3] = max(c((1:length(o))[names(o) == nazvy[i]],0))
}
vyber = indexy[indexy[,1]*indexy[,2]*indexy[,3]*indexy[,4]>0,]

indexy2 = matrix(0,598,4)
indexy2[,4] = 1:598
for(i in 1:598){
  indexy2[i,1] = max(c((1:dim(b)[1])[b$XCOO == m$XCOO[i]
               & b$YCOO == m$YCOO[i]],0))
  indexy2[i,2] = max(c((1:dim(c)[1])[c$XCOO == m$XCOO[i]
               & c$YCOO == m$YCOO[i]],0))
  indexy2[i,3] = max(c((1:dim(o)[1])[o$XCOO == m$XCOO[i]
               & o$YCOO == m$YCOO[i]],0))
}
vyber2 = indexy2[indexy2[,1]*indexy2[,2]*indexy2[,3]
         *indexy2[,4]>0,]

b_horizon = b[vyber2[,1],vyber[,1]]
c_horizon = c[vyber2[,2],vyber[,2]]
o_horizon = o[vyber2[,3],vyber[,3]]
moss_layer = m[vyber2[,4],vyber[,4]]

major = c("Al","Ca","Fe","K","Mg","Mn","Na","P","S","Si")
identifikace = c("ID", "XCOO", "YCOO")
promenne = c(identifikace, major)

b_horizon = b_horizon[,promenne]
```

```
c_horizon = c_horizon[,promenne]
o_horizon = o_horizon[,promenne]
moss_layer = moss_layer[,promenne]
```

C.2. Výpočet skórů

```
V1 = lambda$vectors[1,1] * puda[,1:10] + lambda$vectors[2,1]
    * puda[,11:20] + lambda$vectors[3,1] * puda[,21:30]
    + lambda$vectors[4,1] * puda[,31:40]
V2 = lambda$vectors[1,2] * puda[,1:10] + lambda$vectors[2,2]
    * puda[,11:20] + lambda$vectors[3,2] * puda[,21:30]
    + lambda$vectors[4,2] * puda[,31:40]
V3 = lambda$vectors[1,3] * puda[,1:10] + lambda$vectors[2,3]
    * puda[,11:20] + lambda$vectors[3,3] * puda[,21:30]
    + lambda$vectors[4,3] * puda[,31:40]
V4 = lambda$vectors[1,4] * puda[,1:10] + lambda$vectors[2,4]
    * puda[,11:20] + lambda$vectors[3,4] * puda[,21:30]
    + lambda$vectors[4,4] * puda[,31:40]
```

C.3. Grafické znázornění

```
plot(c(),c(), xlim = c(-1,1), ylim = c(-1,1), xlab = "PC (46.6 %)",
     ylab = "PC2 (32.0 %)")
points(lambda$vectors[,1], lambda$vectors[,2], pch = 19)
points(0,0,pch = 19, col = "firebrick")
x = c(lambda$vectors[1:4,1])
y = c(lambda$vectors[1:4,2])

for (i in 1:4){
  arrows(0, 0, x1 = x[i], y1 = y[i], length = 0.1, angle = 30,
code = 2, col = "firebrick",lty = "solid" )
}
text(lambda$vectors[1,1]+0.1,lambda$vectors[1,2]-0.1,
     expression("U"[1]))
text(lambda$vectors[2,1]-0.1,lambda$vectors[2,2]+0.1,
     expression("U"[2]))
text(lambda$vectors[3,1]-0.1,lambda$vectors[3,2],
     expression("U"[3]))
text(lambda$vectors[4,1]+0.1,lambda$vectors[4,2],
     expression("U"[4]))
```

C.4. Postup v případě $D = 3$

```
W = matrix(0,4,4)
for (j in 1:4){
  for (k in (3*(j-1)+1):(3*j)){
    W[j,j] = W[j,j] + var(puda[,k])
  }
}
W = 0.5 * W

for (i in 1:4){
  for (j in (i+1):4){
    if (i<=3){
      for (k in 1:3){
        W[i,j] = W[i,j] + cov(puda[,3*(i-1)+k],puda[,3*(j-1)+k])
      }
    }
  }
}
W = W + t(W)

lambda = eigen(R)
vysv_var = 100*lambda$values / sum(lambda$values)
CCR = c(vysv_var[1],0,0,0)
for (i in 2:4){
  CCR[i] = CCR[i-1] + vysv_var[i]
}

plot(c(),c(), xlim = c(-1,1), ylim = c(-1,1), xlab = "PC1 (48.2 %)",
    ylab = "PC2 (33.0 %)")
points(lambda$vectors[,1], lambda$vectors[,2], pch = 19)
points(0,0,pch = 19, col = "firebrick")
x = c(lambda$vectors[1:4,1])
y = c(lambda$vectors[1:4,2])

for (i in 1:4){
  arrows(0, 0, x1 = x[i], y1 = y[i], length = 0.1, angle = 30,
code = 2, col = "firebrick",lty = "solid" )
}
text(lambda$vectors[1,1]+0.1,lambda$vectors[1,2]+0.1,
     expression("U"[1]))
text(lambda$vectors[2,1]-0.1,lambda$vectors[2,2],
     expression("U"[2]))
```

```
text(lambda$vectors[3,1]-0.1,lambda$vectors[3,2],
     expression("U"[3]))
text(lambda$vectors[4,1]+0.1,lambda$vectors[4,2],
     expression("U"[4]))

V1 = lambda$vectors[1,1] * puda[,1:3] + lambda$vectors[2,1]
   * puda[,4:6] + lambda$vectors[3,1] * puda[,7:9]
   + lambda$vectors[4,1] * puda[,10:12]
V2 = lambda$vectors[1,2] * puda[,1:3] + lambda$vectors[2,2]
   * puda[,4:6] + lambda$vectors[3,2] * puda[,7:9]
   + lambda$vectors[4,2] * puda[,10:12]
V3 = lambda$vectors[1,3] * puda[,1:3] + lambda$vectors[2,3]
   * puda[,4:6] + lambda$vectors[3,3] * puda[,7:9]
   + lambda$vectors[4,3] * puda[,10:12]
V4 = lambda$vectors[1,4] * puda[,1:3] + lambda$vectors[2,4]
   * puda[,4:6] + lambda$vectors[3,4] * puda[,7:9]
   + lambda$vectors[4,4] * puda[,10:12]
colnames(V1) = c("K","Mg","P")
colnames(V2) = c("K","Mg","P")
colnames(V3) = c("K","Mg","P")
colnames(V4) = c("K","Mg","P")
```

C.5. Ternární diagram a mapa pro druhou hlavní komponentu

```
barvy2 = matrix(0,130,1)
for(i in 1:130){
  temp = V2_inv_d[i,]/100 *255
  barvy2[i] = rgb(temp[1],temp[2],temp[3],maxColorValue=255)
}
ggtern(data=V2_inv_d,aes(x=K,y=Mg,z=P))+geom_point(pch=19,
       col=barvy2)

meritko =
  function(x_vlevo_dole,y_vlevo_dole,x_vpravo_nahore,
         y_vpravo_nahore,posun_text,posun_km,velikost)
  {
    rect(x_vlevo_dole,y_vlevo_dole,x_vpravo_nahore,y_vpravo_nahore)
    rect(x_vlevo_dole+(x_vpravo_nahore-x_vlevo_dole)/2,y_vlevo_dole,
     x_vpravo_nahore,y_vpravo_nahore,col=1)
    mtext("0",side=1,at=x_vlevo_dole,line=posun_text, cex=velikost)
```

```
    mtext("50",side=1,at=x_vlevo_dole+(x_vpravo_nahore-
      x_vlevo_dole)/2,line=posun_text, cex=velikost)
    mtext("100",side=1,at=x_vpravo_nahore,line=posun_text,
      cex=velikost)
    mtext("km",side=1,at=x_vpravo_nahore+posun_km,line=posun_text,
      cex=velikost)
    invisible()
  }

plot(souradnice,frame.plot=FALSE,xaxt="n",yaxt="n",
     xlab="",ylab="",type="n",cex.lab=1.2)
pkb(map.col=c("black","black","grey","blue"),add.plot=T)
points(souradnice,pch=20,cex=1.2,col=barvy2)
meritko(670000,7855000,770000,7870000,posun_text = -23,
        posun_km = 25000, velikost = 0.9)
```