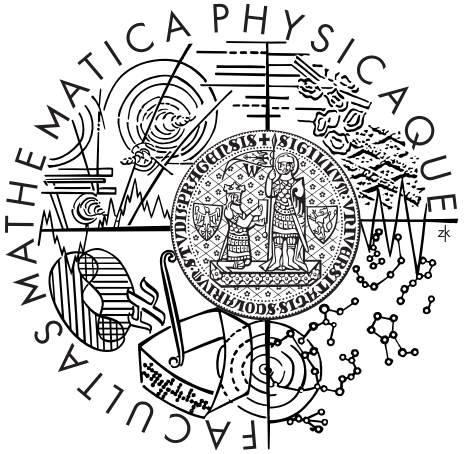
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Karolína Rezková

Archetypální analýza jako segmentační nástroj

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Marek Dvořák
Studijní program: Matematika
Studijní obor: Obecná matematika

Praha 2015

Na tomto místě bych ráda poděkovala vedoucímu práce, RNDr. Marku Dvořákovi, a to především za poskytnutí literatury, náhledu do praxe a v neposlední řadě za jeho shovívavost. Dále pak autorům vzoru pro vytvoření bakalářské práce, který jsem využila, Mgr. Martinu Marešovi, Ph.D., Doc. RNDr. Arnoštu Komárkovi, Ph.D. a Doc. Mgr. Michalu Kulichovi, Ph.D. Velké poděkování patří též rodičům a všem ostatním, kteří mě v průběhu psaní podporovali.

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne .....

Název práce: Archetypální analýza jako segmentační nástroj

Autor: Karolína Rezková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Marek Dvořák, Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Práce se zabývá metodou nazvanou archetypální analýza, jež patří do odvětví mnohorozměrné statistické analýzy. Tato moderní metoda nachází uplatnění v mnoha dalších vědních oborech. Soustředí se na hledání "čistých typů" neboli archetypů, které jsou konvexními lineárními kombinacemi prvků analyzovaných dat, přičemž se zároveň snaží stejným způsobem původní data co nejlépe aproximovat. Celý postup zpracování dat je demonstrován na reálném datovém souboru, na němž jsou též předvedeny některé vybrané vlastnosti metody. Součástí práce je také navržení několika způsobů, kterými lze na základě algoritmu archetypů rozdělit data na vhodný počet segmentů. Tyto způsoby jsou porovnány a předvedeny. Datový soubor je přiložen k práci.

Klíčová slova: Archetypální analýza, segmentace, konvexní lineární kombinace, metoda nejmenších čtverců.

Title: Archetypal Analysis as a Segmentation Tool

Author: Karolína Rezková

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Marek Dvořák, Department of Probability and Mathematical Statistics

Abstract: The thesis presents method called archetypal analysis, which belongs to the field of multivariate statistical data analysis. The method brings contribution to many different branches of science. It focuses in searching for the archetypes or so called "individuals of pure type" that are expressed as convex linear combinations of the original data. At the same time, the original data are represented as the convex linear combinations of the data minimizing the squared error in this representation. There is also a detailed example on processing of the real data. The thesis contains also a sugestion of ways how the original data can be divided into segments. The processed data set is attached.

Keywords: Archetypal analysis, segmentation, linear convex combination, least squares problem.

Obsah

Úvod

Cílem této práce představit metodu zvanou archetypální analýza, ukázat a porovnat možné způsoby segmentace dat na základě této metody a následně demonstrovat využití segmentace na reálných datech.

Nejprve je třeba přiblížit obor archetypální analýzy. Tento přístup ke zpracování dat byl představen v roce 1994 ve článku (Cutler a Breiman, 1994). Metoda se snaží popsat početnou množinu pozorování pouze několika málo *archetypy*, které ji charakterizují. Lze ji prakticky využít v různých oborech, jak ukazují zdroje (Cutler a Breiman, 1994) a (Eugster a Leisch, 2009), jmenujme kupříkladu ekonomii, statistiku, psychologii a meteorologii. Nalezené archetypy jsou směsicemi původních dat a naopak, též původní data jsou směsicemi nalezených archetypů.

Metoda byla implementována v jazyce R, který v práci využiji k demonstraci některých vlastností této metody.

Práce je tvořena třemi kapitolami, v první je přiblížena metoda archetypální analýzy, ve druhé je popsáno zpracování datasetu přiloženého k práci a ve třetí praktická ukázka zpracování dat.

Kapitola 1

Archetypální analýza

1.1 Základní poznatky

Optimalizační problém, který *archetypální analýza* řeší, lze podle (Cutler a Breiman, 1994) definovat následovně.

Definice 1. *Mějme množinu m-rozměrných vektorů $X = \{x_i, i = 1, 2, \ldots, n\}$, které reprezentují n pozorování. Zvolme pevné $p \in \mathbb{N}$ tak, aby $p \leq n$. Předmětem archetypální analýzy je najít p m-rozměrných vektorů z_j, pro něž je splněno následující:*

1. *pro všechna $j = 1, 2, \ldots, p$ je vektor z_j konvexní lineární kombinací vektorů x_i, tedy*

$$z_j = \sum_{k=1}^{n} \beta_{jk} x_k,$$

kde $\beta_{jk} \geq 0$ pro $k = 1, 2, \ldots, n$ a $\sum_{i=1}^{n} \beta_{ji} = 1$,

2. *vektory z_j minimalizují*

$$\sum_{i=1}^{n} \left\| x_i - \sum_{j=1}^{p} \alpha_{ij} z_j \right\|^2, \tag{1.1}$$

kde pro každé $i = 1, 2, \ldots, n$ posloupnost koeficientů $\{\alpha_{ij}\}_{j=1}^{p}$ splňuje podmínky $\alpha_{ij} \geq 0$, $\sum_{j=1}^{p} \alpha_{ij} = 1$ a $\left\| x_i - \sum_{k=1}^{p} \alpha_{ik} z_k \right\|^2$ je minimální.

Pak m-rozměrné vektory z_j nazveme archetypy *(neboli „čisté typy“) množiny X. Poznamenejme, že minimální hodnota reziduálního součtu čtverců* (1.1) *bývá označována jako $RSS(p)$ a koeficienty α_{ij} pro $i = 1, 2, \ldots, n$ a $j = 1, 2, \ldots, p$ bývají zmiňovány bez indexu, tedy jako koeficienty α. Symbolem $\| \cdot \|$ budeme v celé práci značit normu eukleidovskou.*

Poznámka. Metodu lze též popsat pomocí maticového zápisu. Tento způsob je zmíněn například ve článku (Cutler a Breiman, 1994).

Pro nalezení archetypů se užívá iterační *algoritmus archetypů* - nejprve se inicializuje požadovaný počet p archetypů, dále se střídavě řeší dva problémy - hledají se optimální koeficienty α pro dané archetypy z a poté se přepočítají archetypy z pro fixní koeficienty α. Algoritmus je ukončen, pokud je $RSS(p)$ dostatečně malý, nebo pokud bylo dosaženo maximálního možného počtu iterací této metody. V [1] bylo dokázáno, že tento algoritmus je konvergentní, ale v některých případech konverguje pouze k lokálnímu minimu. Proto je třeba jej několikrát opakovat s různými iniciálními hodnotami.

Jedním z klíčových problémů při užití archetypální analýzy je vhodná volba počtu archetypů p pro konkrétní data. Pro jeho řešení neexistuje žádné exaktní pravidlo, [2] však uvádí postup nazvaný *elbow criterion*, (doslovně přeloženo jako *metoda lokte*). Toto kritérium studuje závislost hodnoty $RSS(p)$ na počtu zvolených archetypů. Je zřejmé, že s rostoucím p hodnota $RSS(p)$ klesá nebo zůstává stejná. Metoda lokte volí p takové, že při jeho zvětšení již nedochází k výraznému snížení hodnoty $RSS(p)$.

1.2 Umístění archetypů

Z uvedené definice 1 archetypů plyne, že leží v konvexním obalu množiny pozorování X. Následující tvrzení ukazuje, že při volbě $p > 1$ mohou být archetypy nalezeny na hranici tohoto konvexního obalu.

Tvrzení 1 ((Cutler a Breiman, 1994) - umístění archetypů)**.** *Mějme množinu m-rozměrných vektorů $X = \{x_i, i = 1, 2, \ldots, n\}$. Nechť C je konvexní obal množiny X. Označme S množinu vektorů na hranici C, N mohutnost této množiny S a p počet hledaných archetypů množiny X.*

1. *Je-li $p = 1$, potom součet reziduálních čtverců nabývá své minimální hodnoty při volbě z_1 jako výběrového průměru z množiny X.*
2. *Je-li $1 < p < N$, potom existuje taková množina $\{z_1, z_2, \ldots, z_p\}$ minimalizující $RSS(p)$, která leží na hranici C.*
3. *Je-li $p = N$, volbou $\{z_1, z_2, \ldots, z_p\} = S$ dosáhneme nulového $RSS(p)$.*

Důkaz. Jednotlivé kroky důkazu jsou naznačeny ve článku (Cutler a Breiman, 1994), zde se pokusíme jej vysvětlit podrobněji. Pro každý z případů je třeba ověřit, že navrhované archetypy splňují definici 1

1. Výběrový průměr $\overline{X} = \sum_{i=1}^{n} \frac{x_i}{n}$ je jistě konvexní kombinací vektorů x_i, stačí volit $\beta_{1i} = \frac{1}{n}$ pro $i = 1, 2, \ldots, n$. Potom platí $\beta_{1i} \geq 0$ a $\sum_{i=1}^{n} \beta_{1i} = \sum_{i=1}^{n} \frac{1}{n} = 1$, čímž je splněna část 1 zmíněné definice.

   Podmínky pro koeficienty α při volbě $p = 1$ určují jejich hodnoty: pro $i = 1, 2, \ldots, n$ musí platit $\sum_{j=1}^{1} \alpha_{ij} = \alpha_{i1} = 1$. V důsledku jsou splněny i podmínky nezápornosti těchto koeficientů $\alpha_{i1} = 1 \geq 0$. Všimněme si, že hodnoty koeficientů α nezávisí na volbě konkrétního z_1. Zbývá ukázat, že výběrový průměr skutečně minimalizuje výraz (1.1). Upravme tedy výraz pro $p = 1$:

$$\sum_{i=1}^{n} \left\| x_i - \sum_{j=1}^{p} \alpha_{ij} z_j \right\|^2 = \sum_{i=1}^{n} \left\| x_i - z_1 \right\|^2.$$

Nyní sporem dokážeme, že jeho hodnota je skutečně minimální pro volbu $z_1 = \overline{X}$. Nechť existuje z_1' takové, že $z_1' \neq \overline{X}$ a

$$\sum_{i=1}^{n} \|x_i - z_1'\|^2 \leq \sum_{i=1}^{n} \left\|x_i - \overline{X}\right\|^2. \tag{1.2}$$

Upravujme postupně výraz:

$$\begin{aligned}
\sum_{i=1}^{n} \|x_i - z_1'\|^2 &= \sum_{i=1}^{n} \left\|x_i - \overline{X} + \overline{X} - z_1'\right\|^2 = \\
&= \sum_{i=1}^{n} \Big\langle \big(x_i - \overline{X}\big) + \big(\overline{X} - z_1'\big), \big(x_i - \overline{X}\big) + \big(\overline{X} - z_1'\big) \Big\rangle = \\
&= \sum_{i=1}^{n} \Big(\big\langle x_i - \overline{X}, x_i - \overline{X} \big\rangle + \big\langle x_i - \overline{X}, \overline{X} - z_1' \big\rangle + \\
&\quad + \big\langle \overline{X} - z_1', x_i - \overline{X} \big\rangle + \big\langle \overline{X} - z_1', \overline{X} - z_1' \big\rangle \Big) = \\
&= \sum_{i=1}^{n} \Big(\big\langle x_i - \overline{X}, x_i - \overline{X} \big\rangle + 2\big\langle x_i - \overline{X}, \overline{X} - z_1' \big\rangle + \big\langle \overline{X} - z_1', \overline{X} - z_1' \big\rangle \Big) = \\
&= \sum_{i=1}^{n} \left\|x_i - \overline{X}\right\|^2 + \sum_{i=1}^{n} 2\big\langle x_i - \overline{X}, \overline{X} - z_1' \big\rangle + \sum_{i=1}^{n} \left\|\overline{X} - z_1'\right\|^2
\end{aligned} \tag{1.3}$$

Nejprve byl k argumentu normy přičten a odečten výběrový průměr $\overline{X}$. Poté byla norma přepsána pomocí skalárního součinu, který byl následně rozepsán dle pravidel linearity pro první a druhou složku skalárního součinu a upraven dle komutativity skalárního součinu, kterou lze využít, neboť prvky skalárního součinu jsou reálná čísla. Součet byl nakonec rozdělen na tři části a pokud to bylo možné, přepsán zpět do tvaru čtverce normy. Zkoumejme nyní jednotlivé součty výsledku (1.3):

$$\sum_{i=1}^{n} \left\|\overline{X} - z_1'\right\|^2 \geq 0,$$

neboť se jedná o součet konečně mnoha nezáporných čísel. K rovnosti dochází pouze v případě, pokud $\overline{X} = z_1'$. Tuto možnost jsme však vyloučili v předpokladu tohoto důkazu.

$$\begin{aligned}
\sum_{i=1}^{n} 2\big\langle x_i - \overline{X}, \overline{X} - z_1' \big\rangle &= 2 \sum_{i=1}^{n} \big\langle x_i - \overline{X}, \overline{X} - z_1' \big\rangle = \\
&= 2\Big\langle \sum_{i=1}^{n} (x_i - \overline{X}), \overline{X} - z_1' \Big\rangle = 2\Big\langle \sum_{i=1}^{n} x_i - n\overline{X}, \overline{X} - z_1' \Big\rangle = \\
&= 2\Big\langle n \sum_{i=1}^{n} \frac{x_i}{n} - n\overline{X}, \overline{X} - z_1' \Big\rangle = 2\Big\langle n\overline{X} - n\overline{X}, \overline{X} - z_1' \Big\rangle = 0
\end{aligned}$$

Z linearity součtu byla nejprve vytknuta konstanta 2 před součet, poté bylo znovu využito linearity první složky skalárního součtu, dále proběhla úprava

této první složky, konkrétně součet n týchž hodnot byl nahrazen násobením (to bylo možné provést, neboť hodnota výběrového průměru nezávisí na i), součet byl vynásoben a vydělen nenulovou konstantou n a upraven do tvaru výběrového průměru. Tím bylo ukázáno, že první složka skalárního součinu je nulová, a tedy dle vlastností skalárního součinu je jeho celá hodnota nulová.

Vraťme se nyní k výsledku (1.3) a shrňme výsledky výpočtů výše:

$$\sum_{i=1}^{n} \|x_i - z_1'\|^2 = \sum_{i=1}^{n} \left\|x_i - \overline{X}\right\|^2 + \sum_{i=1}^{n} 2\left\langle x_i - \overline{X}, \overline{X} - z_1'\right\rangle + \sum_{i=1}^{n} \left\|\overline{X} - z_1'\right\|^2 > \\ > \sum_{i=1}^{n} \left\|x_i - \overline{X}\right\|^2. \tag{1.4}$$

Tímto jsme ale dospěli ke sporu s nerovností (1.2) a tedy dokázali, že hodnota $\overline{X}$ minimalizuje RSS pro $p = 1$.

2. Ukažme nejprve, že navrhované archetypy splňují část 1 definice 1. Jelikož se jedná o prvky C, tedy konvexního obalu množiny X, je požadované zaručeno přímo z definice konvexního obalu.

   Pro druhou část důkazu opět postupujme sporem. Z definice archetypů plyne, že archetypy vždy náleží konvexnímu obalu X. Předpokládejme tedy pro spor, že alespoň jeden z archetypů (bez újmy na obecnosti z_1) musí být vnitřním bodem C. Označme

   $$z(t) = z_j + t(z_1 - z_j), \tag{1.5}$$

   pro libovolné pevné j splňující $1 < j \leq p$, aby nedošlo k rovnosti $z_1 = z_j$, a $t \in \mathbb{R}$, $t > 1$. Jelikož C je konvexní obal konečné množiny prvků konečné dimenze, je omezený; $z(t)$ je lineární kombinací prvků z_1 a z_j. Je tedy možné volit $t = t_0$ takové, aby prvek $z(t_0)$ náležel hranici C. Porovnejme nyní konvexní obaly dvou množin, $Z = \{z_1, z_2, \ldots, z_p\}$ a $Z' = \{z(t_0), z_2, z_3, \ldots, z_p\}$, a ukažme, že $conv(Z) \subseteq conv(Z')$. Úpravou výrazu (1.5) pro $t = t_0$ získáme tvar $z_1 = \frac{1}{t_0} z(t_0) + (1 - \frac{1}{t_0}) z_j$. Jelikož $t_0 > 1$, můžeme tvrdit, že z_1 je konvexní kombinací $z(t_0)$ a z_j, a tedy

   $$conv(Z') = conv\big(\{z(t_0), z_2, z_3, \ldots, z_p\}\big) = conv\big(\{z(t_0), z_1, z_2, z_3, \ldots, z_p\}\big) \supseteq \\ \supseteq conv(\{z_1, z_2, \ldots, z_p\}) = conv(Z).$$

   Aby byly splněny podmínky části 2 definice 1 a $\left\|x_i - \sum_{k=1}^{p} \alpha_{ik} z_k\right\|^2$ byl minimální pro každé $i = 1, 2, \ldots, n$, je třeba vhodně zvolit koeficienty α. Čtverec normy nabývá své minimální hodnoty právě tehdy, když konvexní kombinace $\sum_{k=1}^{p} \alpha_{ik} z_k$ vyjadřuje takový prvek $conv(Z)$, který je pro dané i nejblíže x_i (tj. má nejmenší eukleidovskou vzdálenost od x_i). Minimalizaci výrazu (1.1) můžeme proto přepsat jako

   $$\min \Big\{ \sum_{i=1}^{n} \|x_i - u_i\|^2 \Big| \{u_1, u_2, \ldots, u_p\} \subseteq conv(Z) \Big\} \tag{1.6}$$

Uvažujme nyní jako archetypy prvky množiny Z'. Obdobně jako u množiny Z je zřejmé, že prvky množiny Z' splňují část 1 definice 1. Pro splnění části 2 zmíněné definice opět hledáme takovou konvexní kombinaci prvků Z', aby vyjadřovala prvek s nejmenší eukleidovskou vzdáleností od x_i, což lze zapsat jako

$$\min\Big\{\sum_{i=1}^{n}\|x_i-u_i\|^2\Big|\{u_1,u_2,\ldots,u_p\}\subseteq conv(Z')\Big\}.$$

Porovnejme tento výsledek s (1.6). Připomeňme, že $conv(Z)\subseteq conv(Z')$. Potom platí

$$\min\Big\{\sum_{i=1}^{n}\|x_i-u_i\|^2\Big|\{u_1,u_2,\ldots,u_p\}\subseteq conv(Z')\Big\}\leq$$
$$\leq\min\Big\{\sum_{i=1}^{n}\|x_i-u_i\|^2\Big|\{u_1,u_2,\ldots,u_p\}\subseteq conv(Z)\Big\}.$$

Ukázali jsme, že pokud zvolíme za archetypy prvky množiny Z', dosáhneme v (1.1) hodnot menších nebo rovných než kdybychom zvolili prvky Z. Nahradíme-li tedy vnitřní prvek z_1 množiny Z vhodným prvkem na hranici konvexního obalu Z, hodnota (1.1) se jistě nezvětší, čímž jsme došli ke sporu.

3. V posledním případě jsou za archetypy zvoleny přímo některé prvky původní množiny X. Pro splnění části 1 definice 1 stačí volit pro $j=1,2,\ldots,p$ a $k=1,2,\ldots,n$

$$\beta_{jk}=\begin{cases}1, & \text{je-li } z_j=x_k,\\ 0, & \text{je-li } z_j\neq x_k.\end{cases}$$

Zbývá dokázat, že při tomto výběru archetypů je $RSS=0$. Tím bude dokázán i fakt, že (1.1) je minimální, neboť se jedná o součet konečně mnoha nezáporných hodnot. Tvrzení dokažme indukcí dle m, tj.dimenze vektorů x_i. Nejprve ale připomeňme, že konvexním obalem konečně mnoha bodů o konečné dimenzi je konvexní polyedr a hranicí tohoto konvexního obalu jsou stěny polyedru, které jsou samy též konvexní polyedry, avšak jejichž dimenze je vždy ostře menší než m.

Nechť $m=1$. Pokud je $n\geq 2$, je zřejmé, že na hranici C leží pouze minimum a maximum množiny X. Všechna pozorování x_i lze přímo zapsat jako konvexní kombinaci těchto dvou hodnot, a tedy $\|x_i-\sum_{k=1}^{2}\alpha_{ik}z_k\|^2$ nabývá nuly pro každé $i=1,2,\ldots,n$, a tedy RSS $=0$. Nechť tvrzení platí pro $m-1$. Ukažme, že platí i pro m: Mějme množinu pozorování x_i, jejichž dimenze je m. Zvolíme-li pevné i, pak x_i leží buď na hranici C (je samo archetypem a triviálně platí $\|x_i-\sum_{k=1}^{p}\alpha_{ik}z_k\|^2=0$), nebo je vnitřním bodem C. Potom obdobně jako v předešlém bodu důkazu označme $z(t)=z_1+t(x_i-z_1)$. Zvolme $t_0>1$ tak, aby $z(t_0)$ leželo na hranici C, tedy na některé ze stěn polyedru. Její dimenze je ostře menší než m a (díky indukčnímu předpokladu) je $z(t_0)$ konvexní kombinací některých z archetypů, které náleží této stěně polyedru (bez újmy na obecnosti nechť jsou tyto archetypy $z_2,z_3,\ldots z_{p'}$).

Ale x_i je zřejmě konvexní kombinací z_1 a $z(t_0)$. Můžeme psát

$$x_i = \frac{1}{t_0} z(t_0) + (1 - \frac{1}{t_0}) z_1 = \frac{1}{t_0} \sum_{k=2}^{q} \alpha'_{ik} z_k + (1 - \frac{1}{t_0}) z_1,$$

ve druhé rovnosti bylo užito faktu, že $z(t_0)$ je konvexní kombinací prvků $z_2, z_3, \dots z_q$, tedy platí i $\sum_{k=2}^{q} \alpha'_{ik} = 1$

Zvolíme-li koeficienty α následovně:

$$\alpha_{ik} = \begin{cases} 1 - \frac{1}{t_0}, & \text{je-li } k = 1, \\ \frac{1}{t_0} \alpha'_{ik}, & \text{je-li } 1 < k \leq q, \\ 0, & \text{je-li } q < k \leq p, \end{cases}$$

je zřejmé, že $\alpha_{ik} \geq 0$ pro $k = 1, 2, \dots, p$ a

$$\sum_{j=1}^{p} \alpha_{ij} = 1 - \frac{1}{t_0} + \frac{1}{t_0} \sum_{k=2}^{q} \alpha'_{ik} + (p - p') \cdot 0 = 1 - \frac{1}{t_0} + \frac{1}{t_0} \cdot 1 + 0 = 1,$$

a tedy x_i je skutečně konvexní kombinací zvolených archetypů, a proto $\|x_i - \sum_{k=1}^{p} \alpha_{ik} z_k\|^2 = 0$. Tím jsme ukázali, že (1.1) při této volbě nabývá hodnoty 0 a tím i svého minima.

❑

Poznámka. Ze třetího bodu předešlé věty plyne, že volit $p > N$ nepřináší žádné vylepšení hodnoty (1.1).

Kapitola 2

Zpracování surových dat

Nedílnou součástí práce s reálnými daty je jejich úprava do podoby, ve které je možno data dále zpracovávat za pomoci výpočetní techniky. V této kapitole bude popsán způsob, jakým byl zpracován dataset, který budu později zpracovávat. Nejprve však uveďme nástroje, které byly za tímto účelem použity. Veškeré výpočty byly provedeny programem R.

2.1 Užité principy zpracování dat

2.1.1 Číselná reprezentace dat

V praxi se lze často setkat s *kvalitativními daty*. Tato data obvykle nejsou vyjádřena numerickým typem, a proto je třeba je reprezentovat vhodnými číselnými proměnnými. Někdy je však vhodné upravit i data, která již číselně vyjádřena jsou. Dle pravidel uvedených v knihách (Kaufman a Rousseeuw, 1990) a (Jain a Dubes, 1988) lze proměnné rozdělit do několika skupin, z nichž se v této práci setkáváme hlavně s následujícími třemi:

Binární proměnná může nabývat právě jedné ze dvou hodnot. Jako příklad je možno uvést odpověď na zjišťovací otázku. Odpovědi jsou v takovém případě nahrazovány hodnotami 0 a 1 vhodně přiřazenými jednotlivým možnostem (obvykle se hodnota 1 přiřazuje možnosti, kdy je splněno nějaké kritérium a hodnota 0 možnosti zbývající). Mimo pojem *binární proměnná* se v jiné literatuře lze setkat také s pojmy *proměnná dichotomická* či *alternativní*. Binární proměnná, jejíž oběma možným hodnotám je přisuzována táž váha, se obvykle nazývá *symetrická binární proměnná*. Hodnoty 0 a 1 se potom jednotlivým možnostem přiřazují náhodně.

Diskrétní nominální proměnná může nabývat několika různých hodnot, jež však vzájemně nelze porovnat ani seřadit. V tomto případě je proměnná nahrazena příslušným počtem pomocných binárních proměnných pro jednotlivé možnosti odpovědi.

Ordinální proměnná neboli *proměnná pořadová* může obdobně jako nominální proměnná nabývat více různých hodnot, ale tyto hodnoty je možné mezi sebou porovnávat a řadit. Nejsou-li proměnné numerického typu, je vhodné jednotlivé hodnoty nahradit výstižnou číselnou hodnotou, např.

$1, 2, \ldots, n$, a to tak, aby bylo zachováno jejich uspořádání. Pokud proměnná stojí za nějakou skutečnou hodnotou, ponecháme hodnotu nepozměněnou. Pokud reprezentuje nějaký interval, je vhodné použít průměrnou či jinou charakteristickou hodnotu těchto intervalů, a to obzvláště v případě, že intervaly jsou různého rozsahu. Ordinální proměnné můžeme dále rozdělit na *spojité* a *diskrétní*, mnohdy se však k diskrétním ordinálním proměnným přistupuje stejně jako ke spojitým.

2.1.2 Výběrový korelační koeficient a Spearmanův koeficient pořadové korelace

Na jednotlivé proměnné lze nahlížet jako na náhodné veličiny, na jejich hodnoty pak jako na náhodný výběr z nějakého rozdělení. Je-li třeba prozkoumat vzájemný vztah některých veličin, například za účelem snížení počtu proměnných, bývá užíváno *výběrové korelační matice* R, jako je zavedena v publikaci Anděl (2011, str.98), tedy

$$R = (r_{ij}) = \left(\frac{s_{ij}}{\sqrt{s_{ii} s_{jj}}} \right)_{i,j=1}^{m},$$

kde m značí počet zkoumaných proměnných, proměnné i a j prochází jednotlivé proměnné a r_{ij} je *výběrový korelační koeficient* mezi proměnnými určenými jeho indexy. Výběrový korelační koeficient r náhodného výběru $(X_1, Y_1), \ldots, (X_n, Y_n)$ z nějakého dvojrozměrného rozdělení je dle zdroje Anděl (2011, str.93). definován vzorcem

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}}, \tag{2.1}$$

kde S_{XY} značí statistku danou vzorcem $S_{XY} = \dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$ a $\bar{X}, \bar{Y}$ vyjadřují výběrový průměr příslušného náhodného výběru. Je však třeba poznamenat, že interpretace tohoto koeficientu předpokládá dvourozměrné normální rozdělení náhodného výběru.

V případě, že není splněna podmínka, která předpokládá normalitu náhodného výběru $(X_1, Y_1), \ldots, (X_n, Y_n)$ nebo pokud je známo pouze pořadí hodnot zkoumaných náhodných veličin, je vhodnější využít *Spearmanův koeficient pořadové korelace* r_S. Vzorec, který tento koeficient definuje, lze nalézt v knize Anděl (2011, str.256). Je obdobou vzorce 2.1 výběrového korelačního koeficientu, avšak namísto samotných hodnot náhodného výběru je do něj dosazeno pouze pořadí těchto hodnot. Označíme-li pořadí veličin $X_1, \ldots, X_n$ jako $P_1, \ldots, P_n$ a obdobně $Y_1, \ldots, Y_n$ jako $Q_1, \ldots, Q_n$ a dosadíme-li do zmíněného vzorce, postupnými úpravami a užitím definice výběrového průměru získáme přesný tvar vzorce v definici

Spearmanova koeficientu pořadové korelace ze zmíněného zdroje:

$$
\begin{aligned}
r_S &= \frac{S_{PQ}}{\sqrt{S_P^2 S_Q^2}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\big((P_i-\bar{P})(Q_i-\bar{Q})\big)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(P_i-\bar{P})^2 \cdot \frac{1}{n-1}\sum_{i=1}^{n}(Q_i-\bar{Q})^2}} = \\
&= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(P_iQ_i - P_i\bar{Q} - Q_i\bar{P} + \bar{P}\bar{Q})}{\frac{1}{n-1}\sqrt{\sum_{i=1}^{n}(P_i^2 - 2P_i\bar{P} + \bar{P}^2)\cdot\sum_{i=1}^{n}(Q_i^2 - 2Q_i\bar{Q} + \bar{Q}^2)}} = \\
&= \frac{\sum_{i=1}^{n}P_iQ_i - \bar{Q}\sum_{i=1}^{n}P_i - \bar{P}\sum_{i=1}^{n}Q_i + n\bar{P}\bar{Q}}{\sqrt{(\sum_{i=1}^{n}P_i^2 - 2\bar{P}\sum_{i=1}^{n}P_i + n\bar{P}^2)(\sum_{i=1}^{n}Q_i^2 - 2\bar{Q}\sum_{i=1}^{n}Q_i + n\bar{Q}^2)}} = \\
&= \frac{\sum_{i=1}^{n}P_iQ_i - n\bar{Q}\bar{P} - n\bar{P}\bar{Q} + n\bar{P}\bar{Q}}{\sqrt{(\sum_{i=1}^{n}P_i^2 - 2n\bar{P}^2 + n\bar{P}^2)(\sum_{i=1}^{n}Q_i^2 - 2n\bar{Q}^2 + n\bar{Q}^2)}} = \\
&= \frac{\sum_{i=1}^{n}P_iQ_i - n\bar{P}\bar{Q}}{\sqrt{(\sum_{i=1}^{n}P_i^2 - n\bar{P}^2)(\sum_{i=1}^{n}Q_i^2 - n\bar{Q}^2)}}.
\end{aligned}
$$

2.1.3 Normalizace

Hodnoty jednotlivých proměnných datasetu mohou být uvedeny v různých jednotkách, v důsledku čehož mohou mít některé proměnné při následné analýze datasetu vyšší váhu než jiné, a to může negativně ovlivnit výsledek celého procesu. Aby se tomuto problému předešlo, bývají data před začátkem provedení výpočtu upravena tak, aby bylo možno proměnné mezi sebou porovnat. Tato úprava dat se nazývá *normalizace* nebo také *standardizace*. V literatuře je uvedeno více různých způsobů, kterými lze data normalizovat. Pro účely demonstrace archetypální analýzy a následné segmentace jsem vybrala dvě metody, které budu aplikovat:

Normalizace na základě hraničních bodů oboru hodnot zaručí, že hodnoty proměnných budou zobrazeny do uzavřeného intervalu $\langle 0,1\rangle$. Označme hodnoty nějaké proměnné $X_1, \ldots, X_n$, dále $m = \min_{x\in\{X_1,\ldots,X_n\}} x$ a $M = \max_{x\in\{X_1,\ldots,X_n\}} x$. Normalizovaná data $\tilde{X}_1, \ldots, \tilde{X}_n$ dostaneme z původních dosazením do vzorce

$$\tilde{X}_i = \frac{X_i - m}{M - m} \tag{2.2}$$

pro $i = 1, \ldots, n$. Poznamenejme, že tato metoda normalizace ponechává binárním proměnným hodnoty 0 a 1.

Z-transformace bývá také označována jako *z-skóre*. Jedná se o normalizaci na základě výběrové směrodatné odchylky a výběrového průměru. Hodnoty proměnné zobrazuje do uzavřeného intervalu $\langle -1,1\rangle$, přičemž normalizované hodnoty mají nulový výběrový průměr a jednotkovou výběrovou směrodatnou odchylku. Označíme-li stejně jako v předešlém případě hodnoty nějaké proměnné $X_1, \ldots, X_n$, dále $\bar{x}$ výběrový průměr těchto hodnot a s

jejich výběrovou směrodatnou odchylku. Připomeňme, že výběrová směrodatná odchylka je dána vzorcem $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2}$. Normalizovaná data $\tilde{X}_1, \ldots, \tilde{X}_n$ dostaneme z původních lineární transformací

$$\tilde{X}_i = \frac{X_i - \bar{x}}{s} \tag{2.3}$$

pro $i = 1, \ldots, n$.

2.2 Dataset Analýza priorit spotřebitelů

Tento dataset obsahuje 233 vektorů a byl získán formou ankety umístěné na webových stránkách (Pražák, 2013). Původní dotazník se skládal ze 35 otázek. 8 z nich je zaměřených na obecnou charakteristiku dotazovaného, ostatní popisují jeho nakupovací zvyklosti. Ke každé otázce vybíral dotazovaný právě jednu z několika nabízených odpovědí, která nejlépe vystihovala skutečnost. Získaná data jsou tedy diskrétního charakteru.

2.2.1 Práce s jednotlivými proměnnými

V tomto datasetu je možné pozorovat všechny tři výše zvýšené typy proměnných. Proměnná *Sex*, která představuje pohlaví respondenta, je příkladem symetrické binární proměnné. Náhodně byla zvolena reprezentace, kdy možnost *Muž* je vyjádřena hodnotou 1 a *Žena* hodnotou 0. Proměnná *BrI* představuje odpověď na otázku, zda respondent klade při nákupu důraz na značku kupovaného produktu. Jedná se o další příklad binární proměnné, tentokrát se však nabízí vhodná reprezentace možnosti *Ano* hodnotou 1 a *Ne* hodnotou 0.

Zástupcem diskrétní nominální proměnné je proměnná *EcA* - ekonomická aktivita respondenta. Odpověď mohla být zvolena z šesti možností (*Student/ka, Nezaměstnaný/á, Podnikatel/ka,* atd.). Proměnná *EcA* tedy byla nahrazena šesti symetrickými binárními proměnnými, které byly nazvány *EA1, EA2, ..., EA6*, přičemž kupříkladu proměnná *EA1* nabývá hodnoty 1 v případě, že respondent na zadanou otázku odpověděl možností *Nezaměstnaný/á*, pokud vybral libovolnou jinou možnost, nabývá tato proměnná hodnoty 0.

Jako ukázku ordinální proměnné uveďme proměnnou *HoM* - počet členů domácnosti. Ta je vyjádřena konkrétním přirozeným číslem a její hodnoty lze nechat nepozměněny. Jinak je třeba přistupovat k další ordinální proměnné *Edu*, která vyjadřuje nejvyšší dosažené vzdělání dotazovaného. Respondenti vybírali jednu ze čtyř možností (*základní, středoškolské bez maturity/výuční list, středoškolské s maturitou, vysokoškolské*). V tomto případě byly jednotlivé možnosti jednoduše nahrazeny číselnými hodnotami tak, aby bylo zachováno jejich uspořádání. Tedy výše zmíněným možnostem byly po řadě přiřazeny hodnoty $1, 2, 3$ a 4. Poslední ordinální proměnnou, která bude zmíněna v souvislosti s datasetem *Analýza priorit spotřebitelů*, je proměnná *Age* vyjadřující věk dotazovaného. Respondent se zařadil do jedné ze čtyř věkových kategorií (*18 – 30 let, 31 – 50 let, 51 – 65 let* a *více jak 66 let*). Věk jsem pojala jako diskrétní náhodnou veličinu a pro každou ze zmíněných věkových kategorií spočetla střední hodnotu věku občanů České republiky - tou jsem se rozhodla reprezentovat jmenované možnosti. Za účelem

zmíněného výpočtu střední hodnoty jsem užila informací ze stránek (Český Statistický Úřad, 2014). Výpočet jsem pro jednotlivé intervaly prováděla dosazením do vzorce pro střední hodnotu $EX = \sum_{i=D}^{H} ip_i$, v němž D a H značí po řadě dolní a horní hranici věku pro počítaný interval, i konkrétní věk a p_i pravděpodobnost, že věk občana České republiky spadajícího do počítaného věkového intervalu je roven právě i. Hodnoty p_i lze získat jako relativní četnosti jevů vyjadřujících skutečnost, že věk občana České republiky příslušného dané věkové kategorii je roven i, tedy dle vzorce $p_i = \frac{n_i}{N}$, kde n_i značí počet občanů České republiky, jejichž věk je roven i a N počet občanů České republiky, jejichž věk spadá do počítaného intervalu, též lze zapsat jako $N = \sum_{i=D}^{H} n_i$. Namísto hodnoty EX by bylo možné použít obyčejný aritmetický průměr hodnot D a H pro první tři intervaly a hodnotu D, čili 66, pro interval poslední. Reprezentace aritmetickým průměrem je jednodušší než reprezentace užívající EX, ta však skutečnost vystihuje přesněji. U prvních tří intervalů se zmíněné reprezentace příliš neliší, ale v posledním je hodnota EX významně vyšší než zvolená hodnota 66.

Součástí dotazníku bylo též 9 otázek, které zjišťovaly, zda respondent kupuje potraviny označené některou konkrétní značkou kvality potravin. Jelikož pro některé z těchto otázek existoval pouze nízký počet respondentů, kteří odpověděli *Ano*, a tyto otázky nepovažuji v průzkumu pro své účely za klíčové, rozhodla jsem se je navzdory ztrátě informace z datasetu vyřadit a nahradit je proměnnou, která nese pouze informaci o počtu preferovaných značek kvality potravin. Tuto novou proměnnou jsem označila *Nqu*. K rozhodnutí o vyřazení těchto proměnných též přispěl fakt, že metoda archetypální analýzy je citlivá na odlehlá pozorování, neboť archetypy bývají pro $p > 1$ (jak bylo diskutováno ve tvrzení 1) umístěny na hranici konvexního obalu pozorování. Dále jsem se rozhodla vyřadit proměnnou *QuM* - informace zda respondent kupuje potraviny označené libovolnou značkou kvality, neboť tento fakt je již zahrnut ve nově zavedené proměnné *Nqu*.

Původní dotazník obsahoval ještě 9 kritérií, která měl dotazovaný seřadit sestupně podle velikosti vlivu na jeho rozhodování při nákupu. Příkladem takových kritérií je *Cena*, *Vzhled produktu* či *Reklama*. Původním záměrem bylo snížit počet těchto proměnných vytvořením nových, obecnějších. Za tímto účelem jsem za pomoci programu R vypočetla Spearmanovy koeficienty pořadové korelace těchto veličin, které by mohly odhalit souvislosti mezi těmito proměnnými. Očekávala jsem například odhalení souvislosti mezi dvojicemi kritérií s názvy *Reklama* a *Vzhled produktu* nebo *Cena* a *Zboží je ve slevě*. Avšak hodnoty Spearmanových koeficientů pořadové korelace neodhalily bližší souvislost mezi žádnými dvěma kritérii. Jelikož dotazník obsahuje další proměnné, které se dotazují na totéž jako některé z devíti kritérií, pokusila jsem se ověřit, zda existuje korelace alespoň mezi proměnnými, jejichž význam je velice blízký (např. hodnocení kritéria *Značka produktu* a binární proměnná *BrI*, která je odpovědí na otázku, zda respondent klade důraz na značku produktu při nákupu). Výběrový korelační koeficient ani Spearmanův koeficient pořadové korelace nepoukazují na souvislost mezi těmito dvěma proměnnými. Při bližším zkoumání odpovědí jednotlivých respondentů jsem došla k závěru, že někteří respondenti buď odpovídali nekonzistentně, nebo kritéria seřadili v opačném pořadí. Rozhodla jsem se proto z analýzy vyřadit i těchto 9 kritérií.

Data jsem poté normalizovala na základě hraničních bodů oboru hodnot podle vzorce 2.2, aby se zamezilo vyššímu vlivu proměnných s řádově vyššími hodno-

tami, jako například *Inc* (proměnná původně vyjadřující výši měsíčního příjmu respondenta v českých korunách) nebo *ToS* (proměnná udávající počet obyvatel obce, ve které respondent žije). Upravený dataset obsahuje 31 proměnných, z čehož je 21 binárních, zbylých 10 je ordinálních.

Kapitola 3

Metody segmentace dat

Pod pojmem *segmentace* si lze představit rozdělení celku na menší části. Exaktní definici tohoto pojmu se mi nepodařilo najít. Není tedy pevně dáno, zda segmentace musí být jednoznačná (ve smyslu že jeden konkrétní dataset vytvoří danou metodou vždy zcela shodné části), nebo třeba zda se jednotlivé části mohou překrývat, či ne. Úvodem zaveďme pro přehlednost několik pojmů.

Části, na které bude celek rozdělen, budeme nazývat *segmenty*.

Dvě množiny segmentů $M_1 = \{S_1, S_2, \ldots, S_{k_1}\}$ a $M_2 = \{\bar{S}_1, \bar{S}_2, \ldots, \bar{S}_{k_2}\}$ označíme jako *totožné* pakliže splňují následující podmínky:

1. $k_1 = k_2$.
2. Pro libovolný vektor $\boldsymbol{x}$ platí: existuje-li $k \in \{1, 2, \ldots, k_1\}$ takové, že $\boldsymbol{x} \in S_k$, potom $\boldsymbol{x} \in \bar{S}_{k_0}$ pro nějaké $k_0 \in \{1, 2, \ldots, k_2\}$.
3. pro libovolné vektory $\boldsymbol{x}, \boldsymbol{y}$ platí ekvivalence

$$\exists k \in \{1, 2, \ldots, k_1\} : (\boldsymbol{x} \in S_k \wedge \boldsymbol{y} \in S_k) \Leftrightarrow \\ \Leftrightarrow \exists k_0 \in \{1, 2, \ldots, k_2\} : (\boldsymbol{x} \in \bar{S}_{k_0} \wedge \boldsymbol{y} \in \bar{S}_{k_0}).$$

Dvojici segmentací označíme jako *ekvivalentní*, pakliže pro libovolná vstupní data vytvoří pro předem pevně zvolený počet segmentů p totožné množiny segmentů.

3.1 Zavedení konkrétních segmentačních metod

Aplikujeme-li na m-rozměrná data $X = \{\boldsymbol{x}_i, i = 1, 2, \ldots, n\}$ iterační algoritmus archetypů pro zvolený počet archetypů p, získáme p m-rozměrných vektorů $\boldsymbol{z}_j$, které jsou lineárními konvexními kombinacemi původních dat. V této kapitole se budeme věnovat různým způsobům, jakými je možné na základě provedené archetypální analýzy a nalezení archetypů rozdělit původní data do segmentů příslušných jednotlivým archetypům.

3.1.1 Segmentace podle koeficientů α

Pokusme se nejprve lépe objasnit, co zmíněné koeficienty α vyjadřují. Připomeňme, že v této práci byly zavedeny v definici 1 uvedené v kapitole 1. Dle zmíněné definice existuje pro každý vektor $\boldsymbol{x}_i$ z původních dat posloupnost $\{\alpha_{ij}\}_{j=1}^{p}$, která dle části 2 této definice pro každé $i = 1, 2, \ldots, n$ minimalizuje

$$\left\| \boldsymbol{x}_i - \sum_{k=1}^{p} \alpha_{ik} \boldsymbol{z}_k \right\|^2, \tag{3.1}$$

přičemž musí být pro každé $j = 1, 2, \ldots, p$ splněno $\alpha_{ij} \geq 0$ a $\sum_{j=1}^{p} \alpha_{ij} = 1$. Označme sumu ze vzorce (3.1) symbolem $\hat{\boldsymbol{x}}_i$, tedy $\hat{\boldsymbol{x}}_i = \sum_{k=1}^{p} \alpha_{ik} \boldsymbol{z}_k$. Pro každé $i = 1, 2, \ldots, n$ je potom vektor $\hat{\boldsymbol{x}}_i$ také m-rozměrný, speciálně se jedná o konvexní lineární kombinací nalezených archetypů, přičemž je požadováno, aby minimalizoval hodnotu danou vzorcem (3.1), tedy v novém značení $\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2$. Jde o minimalizaci složené funkce, kde vnitřní složka (eukleidovská norma) je dle definice funkce nezáporná a vnější složka, (druhá mocnina) je na intervalu $\langle 0, +\infty)$ rostoucí. Proto funkce $\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2$ nabývá svého minima v týchž bodech jako funkce $\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|$. Vektor $\hat{\boldsymbol{x}}_i$ je tedy takový prvek lineárního konvexního obalu množiny archetypů, jehož eukleidovská vzdálenost od původního vektoru $\boldsymbol{x}_i$ je nejmenší možná. Koeficienty $\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ip}$ jsou po řadě koeficienty lineární konvexní kombinace archetypů $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_p$ určující právě vektor $\hat{\boldsymbol{x}}_i$.

Samotná segmentace potom může probíhat tak, že se každý vektor $\boldsymbol{x}_i$ přiřadí segmentu danému archetypem $\boldsymbol{z}_j$ (respektive koeficientem α_{ij}), pokud pro koeficienty α platí následující:

$$\alpha_{ij} \geq \alpha_{ik} \qquad \text{pro všechna } k = 1, 2, \ldots, p. \tag{3.2}$$

Toto přiřazení však nemusí být jednoznačně určené. Může se stát, že existuje více koeficientů α s různými indexy, které splňují nerovnost (3.2). Takové koeficienty potom musí mít stejnou hodnotu. Pokud není možné jeden vektor $\boldsymbol{x}_i$ původních dat zařadit do více než jednoho segmentu, je třeba určit pravidlo, kterým se bude mezi jednotlivými indexy koeficientů α splňujících (3.2) - a tím i mezi příslušnými segmenty - rozhodovat.

Takové pravidlo může být *deterministické*, lze třeba dopředu určit *dobré uspořádání* archetypů a v případě nejednoznačnosti popsané výše vybrat pro vektor $\boldsymbol{x}_i$ segment na základě tohoto uspořádání. Vektor $\boldsymbol{x}_i$ může být kupříkladu přiřazen archetypu, jehož index má nejnižší hodnotu.

Pravidlo však může být i *stochastického* charakteru, mezi segmenty určenými indexy koeficientů α splňujících (3.2) lze rozhodnout náhodně. Tato možnost nabízí další rozšíření v podobě volby pravděpodobností, že bude vybrán ten který segment - pravděpodobnosti mohou být pro každý segment stejné (čili $\frac{1}{J}$), kde J značí počet koeficientů α splňujících (3.2)), nebo mohou být odvozeny od poměru množství vektorů $\boldsymbol{x}_i$ původních dat, které bylo možno jednoznačně přiřadit do jednotlivých segmentů. Volba segmentu pro vektor $\boldsymbol{x}_i$ na základě pravděpodobnosti s sebou však nese riziko, že dva shodné vektory $\boldsymbol{x}_{i_1}$ a $\boldsymbol{x}_{i_2}$ pro $i_1 \neq i_2$ mohou být přiřazeny různým archetypům, což by mohlo být nežádoucí.

Nespornou výhodou této segmentace je fakt, že koeficienty α není třeba po nalezení archetypů iteračním algoritmem dopočítávat, neboť se počítají v průběhu. Stačí je tedy mezi sebou porovnat, což není výpočetně složité. Za nevýhodou lze

pokládat skutečnost, že segmentace v jistých případech nemusí být jednoznačně určena, a to ani při volbě deterministického pravidla pro řešení problému popsaného výše. Nejednoznačnost vzniká v důsledku toho, že některé koeficienty α lze zvolit více způsoby, což lze demonstrovat na jednoduchém příkladě.

Příklad. Mějme následující dvojrozměrná data $\boldsymbol{x}_1 = (1{,}6), \boldsymbol{x}_2 = (2{,}1), \boldsymbol{x}_3 = (4{,}3), \boldsymbol{x}_4 = (5{,}2)$ a $\boldsymbol{x}_5 = (8{,}7)$. Aplikujme na data iterační algoritmus archetypů pro $p = 4$. Zvolme archetypy $\boldsymbol{z}_1 = \boldsymbol{x}_1, \boldsymbol{z}_2 = \boldsymbol{x}_2, \boldsymbol{z}_3 = \boldsymbol{x}_4$ a $\boldsymbol{z}_4 = \boldsymbol{x}_5$. Nejprve ověřme, že tato volba není ve sporu s definicí 1. Aby byla splněna část 1 definice 1, musí být každý archetyp lineární konvexní kombinací původních dat. Pro $i = 1, 2, 3, 4$ označme $\boldsymbol{\beta}_i$ vektor $(\beta_{i1}, \beta_{i2}, \beta_{i3}, \beta_{i4}, \beta_{i5})$. Vektory $\boldsymbol{\beta}_i$ mají potom následující tvar:

$$\begin{aligned}\boldsymbol{\beta}_1 &= (1, 0, 0, 0, 0),\\ \boldsymbol{\beta}_2 &= (0, 1, 0, 0, 0),\\ \boldsymbol{\beta}_3 &= (0, 0, 0, 1, 0),\\ \boldsymbol{\beta}_4 &= (0, 0, 0, 0, 1).\end{aligned}$$

Je zřejmé, že každý koeficient $\beta_{ij} \geq 0$ pro všechna $i = 1, 2, 3, 4$ a $j = 1, 2, 3, 4, 5$, neboť nenabývá jiných hodnot než 0 a 1. Také druhá podmínka, $\sum_{j=1}^{5} \beta_{ij} = 1$, je zřejmě splněna pro každé $i = 1, 2, 3, 4$. Zbývá ověřit část 2 definice 1. Zkoumejme tedy jaké mohou být koeficienty α příslušné jednotlivým vektorům. Pro $i = 1, 2,...,5$ označme $\boldsymbol{\alpha}_i$ vektor $(\alpha_{i1}, \alpha_{i2}, \alpha_{i3}, \alpha_{i4})$. Vektory $\boldsymbol{\alpha}_i$ pro data, která jsou zároveň archetypy, mají jasně daný tvar:

$$\begin{aligned}\boldsymbol{\alpha}_1 &= (1, 0, 0, 0),\\ \boldsymbol{\alpha}_2 &= (0, 1, 0, 0),\\ \boldsymbol{\alpha}_4 &= (0, 0, 1, 0),\\ \boldsymbol{\alpha}_5 &= (0, 0, 0, 1).\end{aligned}$$

Stejně jako u vektorů $\boldsymbol{\beta}$ je zřejmé, že splňují podmínky lineární konvexní kombinace. Vektor $\boldsymbol{\alpha}_3$ však jednoznačně určen není. Označme $\boldsymbol{v}_1 = (\frac{1}{4}, 0, \frac{3}{4}, 0)$ a $\boldsymbol{v}_2 = (0, \frac{2}{3}, 0, \frac{1}{3})$. Zvolíme-li vektor $\boldsymbol{\alpha}_3$ jako $\boldsymbol{v}_1$ nebo $\boldsymbol{v}_2$, budou podmínky z části 2 definice 1 kladené na koeficienty $\boldsymbol{\alpha}$ splněny. Suma (1.1) bude v obou případech nabývat hodnoty 0, a tedy bude nejmenší možná. Tím jsme ukázali, že zvolená čtveřice vektorů skutečně může být archetypy, ale že její koeficienty α nejsou jednoznačně určeny.

Aplikujme nyní segmentaci podle koeficientů α. Označme segmenty příslušné archetypům $\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{z}_4$ po řadě $S1, S2, S3$ a $S4$. Je zřejmé, že vektory $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_4$ a $\boldsymbol{x}_5$ náleží po řadě segmentům $S1, S2, S3$ a $S4$. Pokud nastane situace, že $\boldsymbol{\alpha}_3 = \boldsymbol{v}_1 = (\frac{1}{4}, 0, \frac{3}{4}, 0)$, bude vektor $\boldsymbol{x}_3$ patřit do segmentu $S3$. Pokud však $\boldsymbol{\alpha}_3 = \boldsymbol{v}_2 = (0, \frac{2}{3}, 0, \frac{1}{3})$, bude tento vektor zařazen do segmentu $S2$. Tím jsme ukázali, že tentýž dataset lze zmíněnou metodou korektně segmentovat více způsoby.

Poznámka. Přidáme-li do datasetu v předešlém příkladě další data, která splňují podmínku, že jsou nějakými lineárními konvexními kombinacemi dat zmíněných v příkladě, archetypy mohou zůstat stejné a vektor $\boldsymbol{x}_3$ půjde stále zařadit do zmíněných dvou segmentů.

3.1.2 Segmentace pomocí metriky

Pro přehlednost nejprve uveďme několik pojmů užívaných v této části textu.

Definice 2 (Metrika)**.** *Mějme neprázdnou množinu M. Zobrazení $d : M \times M \to \mathbb{R}$ nazveme* ***metrikou****, pokud pro $\forall x, y, z \in M$ splňuje následující podmínky:*

1. $d(x,y) \geq 0$,
2. $d(x,y) = 0 \Leftrightarrow x = y$,
3. $d(x,y) = d(y,x)$,
4. $d(x,y) + d(y,z) \geq d(x,z)$.

Uveďme několik známých metrik pro $M = \mathbb{R}^n$.

Příklad (Eukleidovská metrika)*.*

$$d_e(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Příklad (Manhattanská metrika)*.*

$$d_1(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} |x_i - y_i|$$

Příklad (Maximová metrika)*.*

$$d_\infty(\boldsymbol{x}, \boldsymbol{y}) = \max_{i=1,2,\ldots,n} |x_i - y_i|$$

Poznámka. Manhattanská metrika bývá v různé literatuře označována též jako *součtová* či *newyorská.*

Poznámka. Hodnotu $d_e(\boldsymbol{x}, \boldsymbol{y})$ (respektive $d_1(\boldsymbol{x}, \boldsymbol{y})$, $d_\infty(\boldsymbol{x}, \boldsymbol{y})$) nazveme *eukleidovskou* (respektive *manhattanskou*, *maximovou*)*vzdáleností* $\boldsymbol{x}$ a $\boldsymbol{y}$.

Segmentace na základě metriky d lze popsat tak, že vektor $\boldsymbol{x}_i$ se přiřadí segmentu danému archetypem $\boldsymbol{z}_j$, pokud platí následující:

$$d(\boldsymbol{x}_i, \boldsymbol{z}_k) \geq d(\boldsymbol{x}_i, \boldsymbol{z}_j) \qquad \text{pro všechna } k = 1, 2, \ldots, p. \tag{3.3}$$

Jinými slovy hledáme index k_0, pro který platí

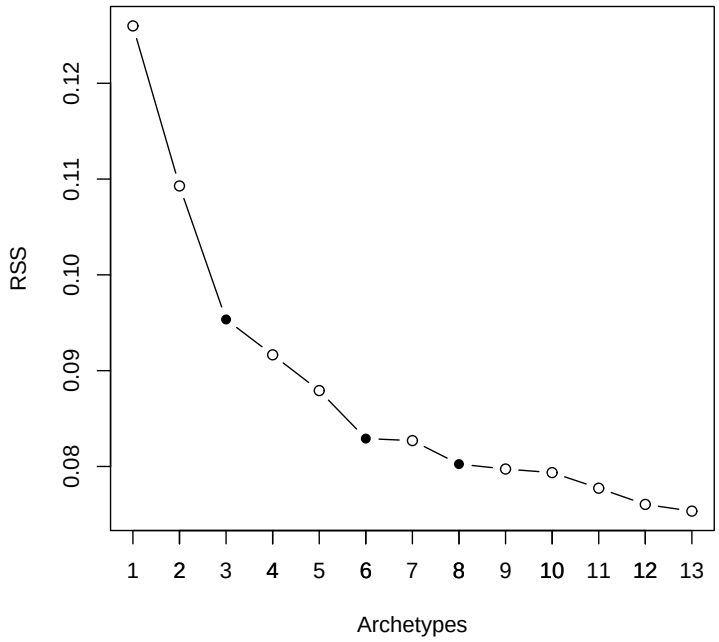
$$d(\boldsymbol{x}_i, \boldsymbol{z}_{k_0}) = \min_{k=1,2,\ldots,p} d(\boldsymbol{x}_i, \boldsymbol{z}_k).$$

Ani takto zvolený způsob segmentace není jednoznačně určen, i v tomto případě je zapotřebí dodefinovat kterému archetypu vektor $\boldsymbol{x}_i$ přiřadit, pokud existuje víc archetypů splňujících (3.3). Tento problém však lze řešit zcela stejně jako u segmentace popsané v předchozím odstavci. Další nejednoznačnosti v tomto případě nevznikají, což je oproti případu předešlému výhoda. Nevýhodou je jistě potřeba dalších výpočtů (pro každý vektor $\boldsymbol{x}_i$ a každý archetyp $\boldsymbol{z}_j$ je nutno nejprve dopočítat hodnoty $d(\boldsymbol{x}_i, \boldsymbol{z}_j)$ a ty následně porovnat jako v předešlém případě - to je jistě složitější než pouhé porovnání již vypočtených hodnot). Také je třeba poznamenat, že hodnota $d(\boldsymbol{x}, \boldsymbol{y})$ je významně závislá na volbě jednotek dat. Jelikož v definici 1 je užito eukleidovské normy, zaměříme se později ze zmíněných tří norem právě na eukleidovskou.

3.2 Segmentace reálných dat

V této kapitole se pokusíme na základě simulace na konkrétním datasetu demonstrovat některé vlastnosti metody archetypální analýzy, zmíněných segmentačních metod a metody (případně jejich modifikace) mezi sebou porovnat. Dataset byl představen v odstavci 2.2, kde bylo popsáno jeho předzpracování, nyní za pomoci programu R aplikujme metodu archetypální analýzy.

Nejprve je potřeba určit množství archetypů p, s nimiž budeme pracovat. Za tímto účelem opakujme algoritmus archetypů několikrát pro různá p. Obrázek 3.1 zachycuje výsledky výpočtu - graf závislosti nejnižší dosažené hodnoty $RSS(p)$ na počtu archetypů. Vyznačeny jsou body, které jsem na základě metody lokte vyhodnotila jako možné vhodné počty archetypů. K nejvýraznější změně sklonu vykreslené křivky dochází asi v bodě $p = 6$, přidáme-li však ještě další dva archetypy, hodnota $RSS(p)$ při přidání sedmého se sníží jen minimálně, při přidání osmého zase znatelně výrazněji. K první výraznější změně tvaru křivky však dochází pro $p = 3$.



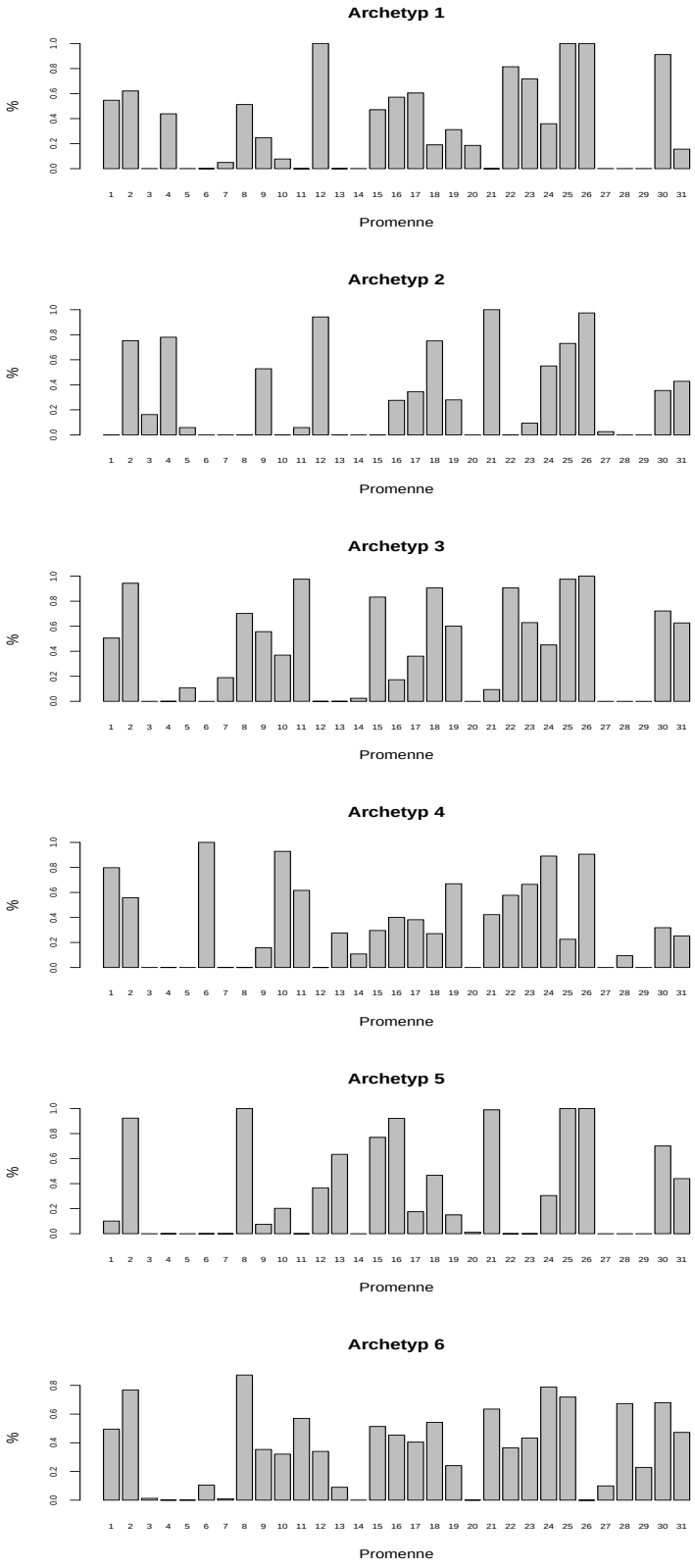
Obrázek 3.1: Závislost velikosti hodnoty $RSS(p)$ na počtu archetypů.

Obrázek 3.2 vykresluje archetypy, které byly vyhodnoceny jakožto nejlepší pro $p = 6$ (ze sta opakování algoritmu archetypů bylo při uvažování těchto archetypů dosaženo minimální hodnoty $RSS(p) \cong 0{,}08291$, a to po provedení deváté iterace tohoto algoritmu). Pro vykreslení archetypů jsem s ohledem na vysoký počet archetypů a vysokou dimenzi dat vybrala sloupcový graf, který zobrazuje hodnoty jednotlivých proměnných poměrně vzhledem z maximální hodnotě jednotlivých proměnných. Tuto skutečnost jsem se v popiskách os grafů rozhodla označit symbolem %. Proměnné jsou popsány z důvodu lepší čitelnosti pouze svým pořadím. Jejich význam je vysvětlen v tabulce 3.1 uvedené níže.

Z grafů v obrázku 3.2 lze vyčíst, že archetyp 4 je archetypem průměrně vzdělaného muže důchodce - jak napovídá proměnná *EA4*. Odpovídá tomu i výše proměnné *Age*. Jeho domácnost obývá nižší počet členů, přestože není svobodný. Jeho čistý měsíční příjem není příliš vysoký. Takový člověk preferuje zboží dle své vlastní osobní zkušenosti, nevěnuje pozornost reklamě ani veřejným zdrojům. V obchodě stráví méně času, nerad zkouší nové zboží, ale často srovnává ceny. Nepreferuje zahraniční značky zboží.

Oproti tomu archetyp 2 popisuje mladou svobodnou studentku, která žije v domácnosti s více členy (tedy buď s rodiči a sourozenci, nebo se spolubydlícími). Ráda nakupuje ve větších obchodech a zkouší novinky. České potraviny preferuje v případě nižší ceny a nevěnuje pozornost reklamě, veřejným zdrojům, ani doporučením známých.

Archetyp 5 vystihuje mladou zaměstnanou ženu s vysokým vzděláním a příjmem, která nejspíš žije sama - je buď rozvedená nebo svobodná. Žije ve větším městě. Nenakupuje často a v obchodě tráví průměrně dlouho. Zkouší nové věci a zboží si vybírá podle sebe, nekouká příliš na značku, ale nechá se ovlivnit veřejnými testy médií.



Obrázek 3.2: Míra zastoupení jednotlivých proměnných v konkrétních archetypech pro $p = 6$.

Pořadí	Zkratka	Význam proměnné
1	Sex	Pohlaví (0-žena, 1 - muž)
2	Edu	Vzdělání
3	EA1	Nezaměstnaný/á
4	EA2	Student/ka
5	EA3	Na mateřské dovolené
6	EA4	Důchodce
7	EA5	Podnikatel/ka
8	EA6	Zaměstnaný/á
9	HoM	Počet členů domácnosti
10	Age	Věk
11	MS1	Ženatý/Vdaná
12	MS2	Svobodný/á
13	MS3	Rozvedený/á
14	MS4	Vdovec/Vdova
15	Inc	Čistý příjem
16	ToS	Počet obyvatel obce
17	ShF	Frekvence nakupování
18	PSS	Preferovaná velikost obchodů
19	BrI	Důležitost značky zboží
20	CF1	Preference zahraničních potravin
21	CF2	Podmíněná preference českých potravin
22	CF3	Úplná preference českých potravin
23	Nqu	Počet upřednostňovaných značek kvality
24	PrC	Porovnávání cen podobného zboží
25	IoT	Zkoušení nového zboží
26	DPeE	Výběr zboží dle osobní zkušenosti
27	DCoR	Výběr zboží dle reklamy
28	DPeR	Výběr zboží dle doporučení známých
29	DPuR	Výběr zboží dle veřejných zdrojů
30	TiM	Vliv nezávislých testů médií
31	AST	Průměrná doba nákupu

Tabulka 3.1: Seznam označení a významu jednotlivých proměnných datasetu.

3.2.1 Segmentace datasetu podle koeficientů α

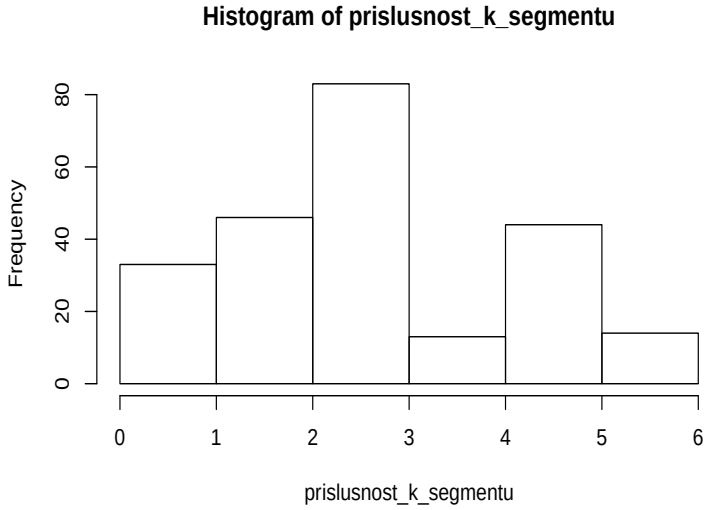
Aplikujme nyní na data segmentaci podle koeficientů α. Chceme-li pomocí programu R získat vektor, jehož každá složka ponese informaci o tom, kterému archetypu je přiřazen vektor $\boldsymbol{x}_i$ příslušného indexu, stačí zadat následující příkaz, kde proměnná *alphas* obsahuje koeficienty α (i-tý řádek obsahuje koeficienty pro vektor $\boldsymbol{x}_i$).

```
> apply(alphas, 1, which.max)
```

Jeho výstupem je řádkový vektor o n složkách, přičemž i-tá složka tohoto vektoru má hodnotu j právě tehdy, když vektor $\boldsymbol{x}_i$ přísluší do segmentu archetypu $\boldsymbol{z}_j$.

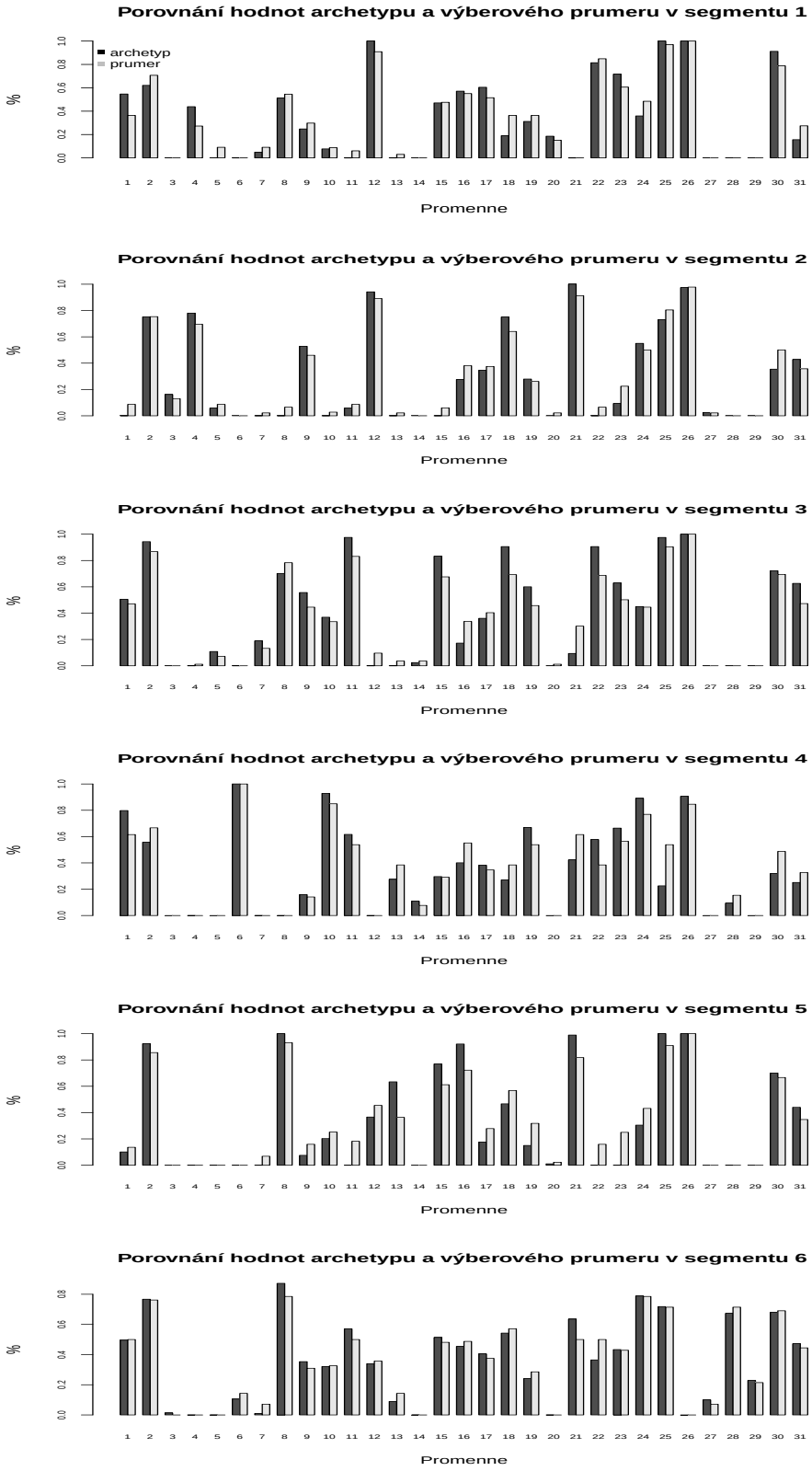
Poznámka. Funkce *which.max* je v programu R předdefinována. Vrací index prvního výskytu maximální hodnoty složky vektoru. To odpovídá deterministické volbě segmentu s nižším indexem při nejednoznačnosti plynoucí ze vzorce (3.2).

Graf na obrázku 3.3 zobrazuje počet vektorů přiřazených jednotlivým segmentům po aplikaci příslušné segmentace. Z grafu lze vyčíst, že asi třetina vektorů původních dat byla zařazena do segmentu 3. Naopak segmentům 4 a 6 bylo přiřazeno podstatně méně vektorů - konkrétně 13 a 14. Archetypy $\boldsymbol{z}_4$ a $\boldsymbol{z}_6$ budou tedy nejspíše ležet v blízkosti některých odlehlejších pozorování. Tento graf bude později porovnán s obdobnými grafy pro jiné metody segmentace.



Obrázek 3.3: Histogram počtu vektorů příslušných danému segmentu při segmentaci datasetu podle koeficientů α.

Srovnejme nyní ještě hodnoty archetypů a průměrné hodnoty příslušných segmentů zobrazené v grafech na obrázku 3.4. U proměnných, jejichž průměrná hodnota je u některého z archetypů výrazně vysoká či nízká, lze pozorovat, že hodnota téže proměnné příslušného archetypu je ještě výraznějším extrémem.



Obrázek 3.4: Porovnání archetypů s průměrnými hodnotami vektorů příslušných jednotlivým segmentům.

3.2.2 Segmentace datasetu pomocí eukleidovské metriky

Na shodná data se shodně zvolenými archetypy jako v předchozím případě nyní aplikujme segmentační metodu, která data přiřazuje do segmentů na základě vzorce (3.3). Reprezentace této metody v programu R je o něco složitější, výsledkem je však obdobný vektor jako v předešlém odstavci. Do výpočtu vstupuje matice, jejíž prvních p řádků tvoří zvolené archetypy a následujících n řádků vstupní data algoritmu. Tyto jsem do jedné matice spojila funkcí *rbind*. Program R dále obsahuje funkci *dist*, která počítá vzdálenost všech dvojic vektorů matice dané v jejím argumentu pomocí zvolené metriky (pro segmentaci na základě jiné metriky stačí argument funkce *dist "euclidean"* zaměnit za *"maximum"* pro metriku maximovou, či *"manhattan"* pro manhattanskou).Z matice je následně vybrána příslušná podmatice, na kterou stačí aplikovat obdobnou funkci jako v segmentaci na základě koeficientů α.

```
> K <- rbind(archetypes, data_in)
> M <- dist(K, method = "euclidean", diag=TRUE, upper=TRUE)
> M <- as.matrix(M)
> M <- head(M,6)
> M <- t(M[,c(7:239)])
> apply(M, 1, which.min)
```

Výstupem tohoto algoritmu je opět řádkový vektor o n složkách, které jsou definovány stejně jako pro předešlou segmentaci.

Poznámka. Číslo 6 v algoritmu je v roli počtu archetypů, čísla 7 a 239 vymezují část matice s původním vstupními daty. Obecně je můžeme po řadě zapsat jako $p, p+1$ a $n+p$.

Poznámka. Také v případě tohoto algoritmu pro segmentaci se při shodné eukleidovské vzdálenosti vektoru od dvou archetypů deterministicky volí archetyp s nižším indexem.

Na obrázku 3.5 je zachycen již známý graf zobrazující počet vektorů přiřazených jednotlivým segmentům po aplikaci segmentace eukleidovskou metrikou. Jeho tvar je téměř shodný s tvarem grafu z obrázku 3.3, který zobrazuje tutéž informaci pro segmentaci koeficienty α. To může svědčit o faktu, že tyto dvě metody segmentace vykazují jistou podobnost, nejedná se však o metody obecně ekvivalentní, což dokazuje následující zjednodušený příklad.

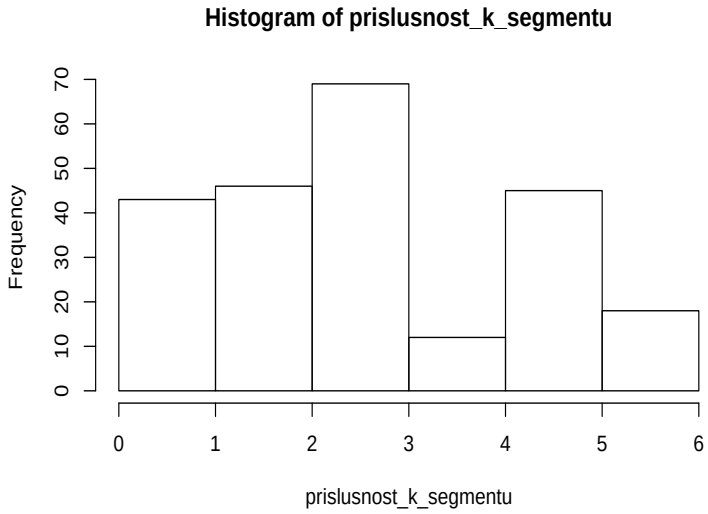
Příklad. Mějme dvojrozměrná data $\boldsymbol{x}_1 = (1,1), \boldsymbol{x}_2 = (6,1), \boldsymbol{x}_3 = (6,2)$ a $\boldsymbol{x}_4 = (8,1)$. Dataset může obsahovat i další vektory, pro jednoduchost však předpokládejme, že všechna další data jsou konvexními lineárními kombinacemi vektorů $\boldsymbol{x}_1, \boldsymbol{x}_3$ a $\boldsymbol{x}_4$. Zvolíme-li vhodný počet p archetypů pro tyto data roven 3 a navrhneme-li jako archetypy po řadě právě vektory $\boldsymbol{x}_1, \boldsymbol{x}_3$ a $\boldsymbol{x}_4$, potom všechna vstupní data lze vyjádřit jako konvexní lineární kombinaci těchto archetypů, čímž jsou splněny podmínky kladené na koeficienty α a zároveň můžeme tvrdit, že $RSS(3) = 0$, a tudíž byly archetypy zvoleny korektně. Ukažme nyní, že vektor $\boldsymbol{x}_2$ bude segmentací koeficienty α zařazen do jiného segmentu, než za užití eukleidovské segmentace.
Je zřejmé, že vektor $\boldsymbol{x}_2$ je konvexní lineární kombinací vektorů $\boldsymbol{x}_1$ a $\boldsymbol{x}_4$. Zvolíme-li

tedy koeficienty α_2 jako $(\frac{2}{7}, 0, \frac{5}{7})$, budou veškeré požadavky splněny a vektor $\boldsymbol{x}_2$ bude přiřazen do segmentu příslušnému vektoru (8,1).
Vypočteme-li však eukleidovskou vzdálenost vektoru $\boldsymbol{x}_2$ od archetypů $\boldsymbol{x}_1, \boldsymbol{x}_3$ a $\boldsymbol{x}_4$, dojdeme k jinému výsledku. Pro názornost proveďme výpočet:

$$d_e(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{(-5)^2 + (0)^2} = 5,$$
$$d_e(\boldsymbol{x}_3, \boldsymbol{x}_2) = \sqrt{(0)^2 + (1)^2} = 1,$$
$$d_e(\boldsymbol{x}_4, \boldsymbol{x}_2) = \sqrt{(2)^2 + (0)^2} = 2.$$

Eukleidovská vzdálenost je nejmenší pro vektor $\boldsymbol{x}_3$ a vektor $\boldsymbol{x}_2$ bude přiřazen do segmentu příslušnému vektoru (6,2).

Tím je prokázáno, že zmíněné dvě segmentace nejsou ekvivalentní.



Obrázek 3.5: Histogram počtu vektorů příslušných danému segmentu při segmentaci datasetu na základě eukleidovské metriky.

Závěr

V práci byla popsána metoda archetypální analýzy, která je sice známa již přes 20 let, ale jejímu rozmachu bránila nedostatečná výpočetní kapacita. V posledních několika letech se však výkonnost zvedla natolik, že se tento postup může začít prosazovat.

V první kapitole bylo podrobně dokázáno tvrzení, jehož úplný důkaz jsem v doposud publikovaných zdrojích nenašla. Ve druhé kapitole bylo názorně předvedeno zpracování reálných dat. Ve třetí kapitole byly představeny způsoby, jakými lze data roztřídit do jednotlivých segmentů, bylo ukázáno, že tyto způsoby nejsou ekvivalentní, ale na základě simulace se zdá, že k sobě mají velice blízko. Oba způsoby byly demonstrovány na reálných datech.

Sepsání práce vyžadovalo nastudování odborné literatury dostupné jen v angličtině, a proto jsem k některým termínům zavedla nové české označení. Veškeré obrázky a grafy uvedené v práci jsou přiloženy v plném rozlišení. Byly vytvořeny programem R na základě zpracovaného datasetu Analýza priorit spotřebitelů.

Seznam použité literatury

Anděl, J. (2011). *Základy matematické statistiky*. Třetí vydání. Matfyzpress, Praha. ISBN 978-80-7378-162-0.

Cutler, A. a Breiman, L. (1994). Archetypal Analysis. *Technometrics*, **36(4)**, 338–347.

Český Statistický Úřad (2014). Věková struktura k 31.12.2013. URL `https://www.czso.cz/staticke/animgraf/cz/`.

Eugster, M. J. A. a Leisch, F. (2009). From Spider-Man to Hero - Archetypal Analysis in R. *Journal of Statistical Software*, **30(8)**.

Jain, A., K. a Dubes, R., C. (1988). *Algorithms for Clustering Data*. Prentice Hall advanced refernce series. Prentice-Hall, Inc., New Jersey. ISBN 0-13-022278-X.

Kaufman, L. a Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction To Cluster Analysis*. John Wiley & Sons, New York. ISBN 0-471-73578-7.

Pražák, P. (2013). Analýza priorit spotřebitelů. URL `https://www.vyplnto.cz/realizovane-pruzkumy/analyza-priorit-spotrebitelu/`.

Seznam obrázků

Seznam tabulek