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

BAKALÁŘSKÁ PRÁCE

Korespondenční analýza

Vedoucí bakalářské práce:
doc. RNDr. Karel Hron, Ph.D.
Rok odevzdání: 2014

Vypracoval:
Denis Drzyzga
MATEKO, III. ročník

Prohlášení

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

V Olomouci dne 18. března 2014

Poděkování

Rád bych na tomto místě poděkoval vedoucímu bakalářské práce doc. RNDr. Karlu Hronovi, Ph.D. za trpělivost a čas, který mi věnoval při konzultacích. Dále bych rád poděkoval své rodině a přátelům, kteří mě po celou dobu mého studia podporovali.

Obsah

Úvod

Tato bakalářská práce se zabývá statistickou metodou zvanou korespondenční analýza. Je to mnohorozměrná statistická metoda, která je určena k popisu struktury závislostí dvou a více kategoriálních proměnných zpracovaných do kontingenční tabulky.

Existují dva typy korespondenční analýzy, a to jednoduchá a vícenásobná korespondenční analýza. Jednoduchá korespondenční analýza popisuje vztahy mezi dvěmi proměnnými v kontingenční tabulce. Naopak vícenásobná varianta této metody popisuje vztahy mezi třemi a více kategoriálními proměnnými. V této práci se zabývám popisem a využitím jednoduché korespondenční analýzy.

Práce je rozdělena do tří kapitol. První dvě kapitoly jsou spíše teoretické, zatímco třetí je čistě praktická.

První z nich popisuje, co to vlastně jsou kategoriální proměnné a kontingenční tabulky, do kterých se tyto proměnné zpracovávají. Dále pojednává o tom, jaké testy v kontingenčních tabulkách můžeme využít před použitím korespondenční analýzy.

Ve druhé kapitole je popsána samotná jednoduchá korespondenční analýza. Jsou zde vysvětleny základní pojmy spojené s touto metodou nutné pro její porozumění. Dále je zde popsán postup výpočtu a následné znázornění a zhodnocení vypočtených výsledků.

Konečně třetí kapitola je zaměřena již na samotnou aplikaci této metody. Je zde názorně na reálných datech ukázáno, k jakým výsledkům dojdeme při aplikování této metody na reálných datech a jak tyto výsledky máme interpretovat. Také zde najdeme zmínku o statistickém softwaru R a knihovně ca, které jsem často při vypracovávání této práce využíval.

1 Kategoriální proměnné a kontingenční tabulky

Jak již bylo řečeno v úvodu, korespondenční analýza je metoda určená k popisu vztahů mezi kategoriálními proměnnými, které jsou prezentovány ve formě kontingenční tabulky. Proto bych rád chtěl čtenáře seznámit s kategoriálními proměnnými a následně kontingenčními tabulkami ještě dříve, než začneme popisovat samotnou korespondenční analýzu. V této kapitole bylo využito zejména těchto zdrojů [1], [4], [5] a [7].

1.1 Kategoriální proměnné

Jako kategoriální jsou označovány zejména kvalitativní proměnné, avšak můžeme zde zahrnout také kvantitativní diskrétní proměnné. Hodnoty, kterých tyto proměnné nabývají, označujeme slovem kategorie, odtud pochází název kategoriální proměnná. Kvalitativní proměnné můžeme dále dělit na nominální a ordinální. Nominální proměnné nelze nijak uspořádat, pouze je možné o dvou hodnonách říci zda jsou stejné či nikoliv. Těmito proměnnými mohou být například jména a příjmení, barva vlasů, číselné kódy a jiné. Narozdíl od nominálních proměnných lze ordinální seřadit od nejmenší hodnoty po tu největší tzn. určit jejich pořadí. Příkladem těchto proměnných může být hodnocení výrobku nebo služby zákazníkem, datum a další. Kvantitativní diskrétní proměnné jsou takové proměnné, které nabývají pouze celočíselných hodnot, například počet válcu motoru, počet dětí žijících ve společné domácnosti a mnoho dalších.

1.2 Kontingenční tabulky

Pokud máme zadána kategoriální data, je pro názornost nejjednodušší je uspořádat do kontingenční tabulky. Vyjdeme ze situace, kdy máme dány dvě kategoriální proměnné. Provedeme n pozorování obou proměnných. Výsledky těchto pozorování můžeme rozdělit do tříd dle kombinací hodnot jednotlivých statistických znaků (proměnných). Přitom první proměnná, kterou můžeme označit jako A, bude nabývat hodnot $a_1, a_2, \dots\ a_r$ a druhá proměnná B bude nabývat

hodnot $b_1, b_2, ... b_s$. Takto dostaneme požadovanou kontingenční tabulku (viz tabulka 1).

Tabulka 1: Kontingenční tabulka.

	b_1	b_2	...	b_s	n_{i+}
a_1	n_{11}	n_{12}	...	n_{1s}	n_{1+}
a_1	n_{21}	n_{22}	...	n_{2s}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_r	n_{r1}	n_{r2}	...	n_{rs}	n_{r+}
n_{+j}	n_{+1}	n_{+2}	...	n_{+s}	n

V záhlaví této tabulky najdeme všechny hodnoty, neboli kategorie, našich proměnných. Uvnitř tabulky se nachází absolutní četnosti n_{ij}, kde $i = 1, 2, ..., r$ a $j = 1, 2, ..., s$, jednotlivých pozorování. Tyto absolutní četnosti odpovídají jednotlivým kategoriím proměnných A a B, proto je můžeme nazvat sdruženými četnostmi. Na okrajích této tabulky dále nalezneme řádkové absolutní četnosti n_{i+} a sloupcové absolutní četnosti n_{+j}, pro které platí

$$n_{i+} = \sum_{j=1}^{s} n_{ij} \tag{1.1}$$

a

$$n_{+j} = \sum_{i=1}^{r} n_{ij} \tag{1.2}$$

pro $i = 1, 2, ..., r$ a $j = 1, 2, ..., s$.

Často využíváme i tzv. relativní četnosti. Ty z tabulky vypočteme pomocí následujících vztahů,

$$p_{ij} = \frac{n_{ij}}{n} \tag{1.3}$$

pro relativní sdružené četnosti p_{ij},

$$p_{i+} = \frac{n_{i+}}{n} \tag{1.4}$$

pro jednotlivé relativní řádkové četnosti p_{i+} a

$$p_{+j} = \frac{n_{+j}}{n} \tag{1.5}$$

pro relativní sloupcové četnosti p_{+j}, kde $i = 1, 2, ..., r$ a $j = 1, 2, ..., s$.

Pro každou kontingenční tabulku navíc platí následující vztahy

$$\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = \sum_{i=1}^{r} n_{i+} = \sum_{j=1}^{s} n_{+j} = n \tag{1.6}$$

pro absolutní četnosti a

$$\sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij} = \sum_{i=1}^{r} p_{i+} = \sum_{j=1}^{s} p_{+j} = 1 \tag{1.7}$$

pro relativní četnosti.

Pro ilustraci zde uvedeme jednoduchý příklad, který nás bude provázet první a druhou kapitolou. Uvažujme, že máme dány dvě kategoriální proměnné, kde první z nich nabývá hodnot A, B a C. Druhá proměnná nabývá hodnot X, Y a Z. Řekněme, že jsme provedli průzkum a naměřili četnosti, které nyní uspořádáme do kontingenční tabulky (viz tabulka 2).

Tabulka 2: Kontingenční tabulka ilustračních dat.

	A	B	C	Celkem
X	22	10	25	57
Y	16	22	5	43
Z	5	16	4	25
Celkem	43	48	34	125

Vypočtěme nyní jednotlivé relativní četnosti, které jsme definovali v této podkapitole. Dostaneme tabulku 3.

Tabulka 3: Relativní četnosti ilustračních dat.

	A	B	C	Celkem
X	0.176	0.08	0.2	0.456
Y	0.128	0.176	0.04	0.344
Z	0.04	0.128	0.032	0.2
Celkem	0.344	0.384	0.272	1

Postupně se budeme k tomuto příkladu vracet a ukážeme si jednotlivé testy, které můžeme na tuto tabulku použít a jednotlivé pojmy, výpočty, znázornění výsledků korespondenční analýzy a jejich následné zhodnocení.

1.3 Analýza kontingenčních tabulek

Hlavní výhodou kontingenčních tabulek je možnost zjistit, jaké vztahy platí mezi jednotlivými proměnnými. V této podkapitole bych proto chtěl uvést některé testy, které dále k analýze tabulek využijeme, ještě předtím než se podíváme na samotnou korespondenční analýzu.

1.3.1 Test nezávislosti

Jako první uvedeme test nezávislosti. Ten vyjadřuje intenzitu závislosti kategoriálních proměnných v dané kontingenční tabulce. Využíváme zde Pearsonovy statistiky Z v tomto tvaru

$$Z = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - \frac{n_{i+}n_{+j}}{n})^2}{\frac{n_{i+}n_{+j}}{n}} = n\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}}, \tag{1.8}$$

která má za předpokladu platnosti nulové hypotézy H_0 o nezávislosti dvojice statistických znaků pro $n \to \infty$ χ^2 rozdělení. Stupeň volnosti tohoto rozdělení je roven

$$rs - 1 - [(r-1) + (s-1)] = (r-1)(s-1),$$

kde od celkového počtu tříd odečítáme počet neznámých parametrů za platnosti nulové hypotézy. Hypotéza bude zamítnuta, jestliže $z \geq \chi^2_{(r-1)(s-1),1-\alpha}$, tedy když se tato statistika bude realizovat v kritickém oboru $W = \langle \chi^2_{(r-1)(s-1),1-\alpha}, \infty)$. Dále musí platit podmínka

$$n\hat{p}_{ij} = \frac{n_{i+}n_{+j}}{n} \geq 5, \tag{1.9}$$

kde $\hat{p}_{ij}$ jsou očekávané relativní četnosti (za platnosti nulové hypotézy), pro $\forall i, j : i = 1, 2, ..., r;\ j = 1, 2, ..., s$.

Pokud bude tato hypotéza zamítnuta, můžeme prohlásit, že mezi proměnnými existuje nějaká závislost.

Nyní zkusme využít tento test pro náš ilustrační příklad. Nejdříve je nutné zkontrolovat, zda je splněna podmínka (1.9). V tomto případě je splněna, je tedy možné přistoupit k výpočtu statistiky Z. Pro naše data je $z = 25.2474$. Nyní

je třeba porovnat toto číslo s hodnotou $\chi^2_{4,0.95} = 9.488$. Naše statistika se tedy realizuje v kritickém oboru a proto můžeme zamítnout nulovou hypotézu. Z toho vyplývá, že naše proměnné nejsou nezávislé.

1.3.2 Znaménkové schéma

Pokud jsme zjistili, že naše proměnné nejsou nezávislé, bylo by užitečné vědět, které hodnoty z kontingenční tabulky nejvíce přispívají k dané závislosti. Právě k tomu nám slouží znaménkové schéma. Je to tabulka, kde nahrazujeme jednotlivé absolutní, či relativní četnosti jedním až třemi plusy, mínusy nebo nulou. Tato tabulka vypovídá o tom, jak se liší naměřené (empirické) četnosti od očekávaných. Tedy jak se liší četnosti našich závislých proměnných od četností, které bychom očekávali pokud by byly dané proměnné nezávislé.

Jak již bylo řečeno, každé pole tabulky obsahuje plusy, mínusy nebo nulu, kde znaménko plus značí, že naměřená četnost je vyšší než očekávaná četnost. Znaménko mínus nám říká, že naše naměřená četnost je nižší než ta očekávaná a nula vypovídá o tom, že obě četnosti jsou si přibližně rovny.

Abychom mohli toto schéma využít, je potřeba znát jednotlivé očekávané četnosti. Ty lze vypočítat jako

$$\hat{n}_{ij} = \frac{n_{i+}n_{+j}}{n} = np_{i+}p_{+j}. \tag{1.10}$$

Dále potřebujeme znát tzv. standardizovaná rezidua, která vypadají následovně,

$$u_{ij} = \frac{n_{ij} - \hat{n}_{ij}}{\sqrt{\hat{n}_{ij}(1 - \hat{p}_{ij})}} = \sqrt{n}\frac{p_{ij} - \hat{p}_{ij}}{\sqrt{\hat{p}_{ij}(1 - \hat{p}_{ij})}}, \tag{1.11}$$

pro $i = 1, 2, ..., r$ a $j = 1, 2, ..., s$.

Znaménkové schéma tedy vytváříme na základě hodnot těchto reziduí. Přiřadíme kladné znaménko, pokud naměřené hodnoty jsou vyšší než hodnoty očekávané a záporné znaménko v opačném případě. Počet znamének je dán kvantily normovaného normálního rozdělení podle tabulky 4.

Tabulka 4: Přiřazení znamének.

Znaménko	Intervaly				
$---$			u_{ij}	$\leq$	-3.29
$--$	-3.29	$<$	u_{ij}	$\leq$	-2.58
$-$	-2.58	$<$	u_{ij}	$\leq$	-1.96
0	-1.96	$<$	u_{ij}	$<$	1.96
$+$	1.96	$\leq$	u_{ij}	$<$	2.58
$++$	2.58	$\leq$	u_{ij}	$<$	3.29
$+++$	3.29	$\leq$	u_{ij}		

Nyní zkusme tento postup aplikovat na náš ilustrační příklad. Nejprve je třeba vypočítat očekávané hodnoty $\hat{n}_{ij}$ a následně jejich relativní četnosti $\hat{p}_{ij}$. Poté vypočteme standardizovaná rezidua a podle nich jednotlivým hodnotám tabulky přiřadíme znaménka. Výsledky můžeme shrnout do tabulky 5, kde vždy první řádek odpovídá očekávané hodnotě, druhý danému reziduu a třetí řádek odpovídá přiřazenému znaménku. V této chvíli tedy víme, které hodnoty tabulky nejvíce přispívají ke zjištěné závislosti.

Tabulka 5: Znaménkové schéma ilustračních dat.

	A	B	C
X	19.608	21.888	15.504
	0.588	-2.798	2.577
	0	$--$	$+$
Y	14.792	16.512	11.696
	0.335	1.450	-2.057
	0	0	$-$
Z	8.6	9.6	6.8
	-1.272	2.150	-1.104
	0	$+$	0

2 Jednoduchá korespondenční analýza

Korespondenční analýza vznikla v 60. až 70. letech minulého století zásluhou francouzského statistika Jeana-Paula Benzécriho. Z dalších osobností, které působí v této oblasti můžeme například zmínit Michaela Greenacra, z jehož prací bylo také při zpracovávání této práce čerpáno.

Z pohledu vícerozměrných statistických metod můžeme korespondenční analýzu zařadit mezi ordinační metody, tzn. metody, jejichž úkolem je zredukovat vícerozměrný charakter dat na menší počet rozměrů, kde budou data lépe čitelná. Jedná se o metodu popisnou a průzkumovou, kterou lze využít ke zjištění vlivů a podobností jednotlivých kategorií na jiné, avšak neobsahuje žádné nástroje pro zjištění statistické významnosti tohoto modelu.

Tato metoda je využívána v mnoha oblastech, zejména v marketingu, sociologii atd. Nyní již přistoupíme k samotnému popisu jednoduché varianty korespondenční analýzy. Tato kapitola vychází z literatury [2], [4] a [6].

2.1 Základní pojmy

Vycházíme z toho, že máme datovou matici $\mathbf{N}$ o r řádcích a s sloupcích, která odpovídá kontingenční tabulce. Celkově má tato matice n prvků, kde $n = rs$.

Nyní by bylo vhodné popsat základní pojmy využívané při výpočtech v korespondenční analýze.

Po vydělení jednotlivých prvků matice $\mathbf{N}$ číslem n dostaneme z původních absolutních četností v tabulce četnosti relativní. Tuto novou matici $\mathbf{P}$ relativních četností nazýváme korespondenční maticí

$$\mathbf{P} = \frac{\mathbf{N}}{n}, \tag{2.1}$$

kde jednotlivé prvky vypočteme takto

$$p_{ij} = \frac{n_{ij}}{n}. \tag{2.2}$$

Dalším pojmem potřebným při práci s touto metodou jsou řádkové a sloupcové zátěže. Řádkové zátěže, které označujeme r_i, vypočteme jako

$$r_i = \frac{n_{i+}}{n}; \tag{2.3}$$

odpovídají tedy dříve zavedeným relativním řádkovým četnostem p_{i+}. Jednotlivé vypočtené řádkové zátěže uspořádáme do vektoru $\mathbf{r}$ a dostaneme tzv. vektor řádkových zátěží

$$\mathbf{r} = \left(r_1, r_2, ..., r_r\right). \tag{2.4}$$

Sloupcové zátěže označujeme c_j a vypočteme je jako

$$c_j = \frac{n_{+j}}{n}; \tag{2.5}$$

odpovídají dříve zavedeným relativním sloupcovým četnostem p_{+j}. Opět můžeme jednotlivé zátěže uspořádat do vektoru

$$\mathbf{c} = \left(c_1, c_2, ..., c_s\right) \tag{2.6}$$

a dostaneme tzv. vektor sloupcových zátěží. Jak řádkové, tak sloupcové zátěže lze označit jako relativní okrajové četnosti.

Přichází na řadu další pojmy, a to řádkové a sloupcové profily. Řádkové profily vypočítáme takto

$$r_{j/i} = \frac{n_{ij}}{n_{i+}}. \tag{2.7}$$

Po vypočtení těchto jednotlivých profilů lze vytvořit novou matici, tzv. matici řádkových profilů $\mathbf{R}$, kterou je možné obecně vypočítat jako

$$\mathbf{R} = \mathbf{D}_r^{-1}\mathbf{P} = \begin{pmatrix} \mathbf{r}_1^T \\ \mathbf{r}_2^T \\ \vdots \\ \mathbf{r}_r^T \end{pmatrix}, \tag{2.8}$$

kde $\mathbf{D_r}$ je diagonální matice, která má na své diagonále prvky vektoru řádkových zátěží $\mathbf{r}$. Sloupcové profily vypočteme analogicky takto

$$c_{i/j} = \frac{n_{ij}}{n_{+j}}. \tag{2.9}$$

Opět lze jednotlivé profily uspořádat do matice sloupcových profilů $\mathbf{C}$, která je obecně dána jako

$$\mathbf{C} = \mathbf{D}_c^{-1}\mathbf{P}^T = \left(\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_s\right), \tag{2.10}$$

kde $\mathbf{D}_c$ je diagonální matice, která má na své diagonále prvky vektoru sloupcových zátěží $\mathbf{c}$.

Nyní lze na základě právě definovaných pojmů zapsat korespondenční matici jako

$$\begin{pmatrix} \mathbf{P} & \mathbf{r} \\ \mathbf{c}^T & 1 \end{pmatrix} = \begin{pmatrix} p_{11} & p_{12} & ... & p_{1s} & r_1 \\ p_{21} & p_{22} & ... & p_{2s} & r_2 \\ \vdots & \vdots & & \vdots & \vdots \\ p_{r1} & p_{r2} & ... & p_{rs} & r_r \\ c_1 & c_2 & ... & c_s & 1 \end{pmatrix}.$$

V tomto okamžiku se můžeme vrátit k našemu ilustračnímu příkladu a názorně na něm ukázat právě zavedené pojmy. Budeme vycházet z tabulky 2.

Matice $\mathbf{N}$ je pro tato data rovna

$$\mathbf{N} = \begin{pmatrix} 22 & 10 & 25 \\ 16 & 22 & 5 \\ 5 & 16 & 4 \end{pmatrix}.$$

Dále vypočteme korespondenční matici $\mathbf{P}$, která bude rovna

$$\mathbf{P} = \begin{pmatrix} 0.176 & 0.08 & 0.2 \\ 0.128 & 0.176 & 0.04 \\ 0.04 & 0.128 & 0.032 \end{pmatrix}.$$

Vektory řádkových a sloupcových zátěží budou vypadat takto

$$\mathbf{r} = \begin{pmatrix} 0.465 \\ 0.344 \\ 0.2 \end{pmatrix},$$

$$\mathbf{c}^T = \begin{pmatrix} 0.344 & 0.384 & 0.272 \end{pmatrix}.$$

Na základě těchto výpočtů lze korespondenční matici vyjádřit v podobě tabulky 3.

Zbývá výpočet matice řádkových profilů $\mathbf{R}$ a matice sloupcových profilů $\mathbf{C}$, které budou vypadat následovně

$$\mathbf{R} = (r_{ij}) = \begin{pmatrix} 0.386 & 0.175 & 0.439 \\ 0.372 & 0.512 & 0.116 \\ 0.2 & 0.64 & 0.16 \end{pmatrix},$$

$$\mathbf{C} = (c_{ij}) = \begin{pmatrix} 0.512 & 0.208 & 0.735 \\ 0.372 & 0.458 & 0.147 \\ 0.116 & 0.333 & 0.118 \end{pmatrix}.$$

2.2 Postup výpočtu

Ještě než uvedeme konkrétní postup korespondenční analýzy, je potřeba si říci, co je cílem těchto výpočtů. Řádky, resp. sloupce naší korespondenční tabulky si lze představit jako body v s-rozměrném, resp. r-rozměrném prostoru. Hovoříme zde o bodech, je tedy logické, že je možné mezi nimi počítat vzdálenosti. Ty odpovídají vzdálenostem řádkových a sloupcových profilů, tedy vzdálenostem mezi řádky a sloupci. Cílem korespondenční analýzy je převést tyto vzdálenosti do euklidovského prostoru. Zde tyto body nahrazují naše kategorie. Pro výpočet vzdálenosti mezi i-tým a i′-tým řádkem nejčastěji využíváme tzv. chí-kvadrát vzdálenost

$$V(i, i') = \sqrt{\sum_{j=1}^{s} \frac{(r_{ij} - r_{i'j})^2}{c_j}}, \tag{2.11}$$

analogicky též pro vzdálenosti mezi sloupci.

Přistupme nyní k samotným výpočtům korespondenční analýzy. Našim cílem je tedy redukovat mnohorozměrný charakter vektorů řádkových a sloupcových profilů, ale přitom zachovat co možná nejpřesněji informace v nich obsažené. Nejvyšší možný počet rozměrů, ve kterých můžeme zobrazit naše data je dán číslem

$$min\{r - 1, s - 1\}.$$

V této práci budeme pracovat v dvourozměrném prostoru, tedy v rovině. Body ležící v této rovině, které jsou nejblíže bodům ve vícerozměrném prostoru, nazývá-

me projekcemi. Hledáme tedy rovinu, resp. souřadnice bodů, které budou co možná nejblíže bodům, které zastupují.

Pro další postup je třeba aplikovat na matici standardizovaných reziduí $\mathbf{Z} = (u_{ij})$ singulární rozklad. Obecně je singulární rozklad součin

$$\mathbf{Z} = \mathbf{U}\boldsymbol{\Gamma}\mathbf{V}^T, \tag{2.12}$$

přičemž $\mathbf{U}$ je matice typu $r \times k$, kde sloupce této matice jsou levé zobecněné singulární vektory, $\boldsymbol{\Gamma}$ je diagonální matice typu $k \times k$, která má na diagonále zobecněné singulární hodnoty a $\mathbf{V}$ je matice typu $s \times k$, kde její sloupce jsou tvořeny pravými zobecněnými singulárními vektory. Také platí $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$.

Konkrétně pro matici standardizovaných reziduí $\mathbf{Z}$ typu $r \times s$ vypadá singulární rozklad takto

$$\mathbf{Z} = \mathbf{D}_r^{-\frac{1}{2}}(\mathbf{P} - \mathbf{r}\mathbf{c}^T)\mathbf{D}_c^{-\frac{1}{2}}, \tag{2.13}$$

kde jednotlivé prvky matice $\mathbf{Z}$ vypočteme jako

$$z_{ij} = \frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}}. \tag{2.14}$$

Od této chvíle bude celý výpočet vycházet z této matice.

Dříve než vypočítáme jednotlivé souřadnice je potřeba zvolit tzv. normalizační metodu. Znamená to, že se musíme rozhodnout, zda nás zajímají pouze řádkové kategorie, tehdy zvolíme analýzu řádkových profilů, nebo sloupcové kategorie, pak volíme analýzu sloupcových profilů. Avšak často požadujeme vzájemné srovnání řádkových a sloupcových kategorií. V tomto případě využijeme tzv. symetrickou normalizaci.

Pokud zvolíme analýzu řádkových profilů, je třeba vypočítat matici $\mathbf{F}$ danou tímto vztahem

$$\mathbf{F} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}\boldsymbol{\Gamma}, \tag{2.15}$$

která má ve svých sloupcích obsaženy jednotlivé souřadnice řádkových bodů, a dále též matici

$$\mathbf{Y} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}, \tag{2.16}$$

která obsahuje ve svých sloupcích souřadnice sloupcových kategorií.

V případě, že zvolíme analýzu sloupcových profilů, budeme počítat matici

$$\mathbf{G} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}\boldsymbol{\Gamma}, \tag{2.17}$$

která obsahuje souřadnice sloupcových kategorií a matici

$$\mathbf{X} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}, \tag{2.18}$$

která má ve svých sloupcích souřadnice řádkových kategorií.

A na závěr poslední varianta, kdy nás zajímají vzájemný vztah mezi řádkovými a sloupcovými kategoriemi. V tomto případě využijeme při znázornění výsledků matice $\mathbf{F}$ a $\mathbf{G}$.

Opět se vrátíme k ilustračnímu příkladu a vypočteme právě zavedené matice. Matice $\mathbf{Z}$ bude pro naše data rovna

$$\mathbf{Z} = \begin{pmatrix} 0.048 & -0.227 & 0.216 \\ 0.028 & 0.121 & -0.175 \\ -0.110 & 0.185 & -0.096 \end{pmatrix}$$

a její rozklad bude

$$\mathbf{U} = \begin{pmatrix} -0.732 & 0.090 & 0.675 \\ 0.459 & -0.667 & 0.587 \\ 0.503 & 0.740 & 0.447 \end{pmatrix},$$

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.433 & 0 & 0 \\ 0 & 0.121 & 0 \\ 0 & 0 & 2.113 * 10^{-17} \end{pmatrix}$$

a

$$\mathbf{V} = \begin{pmatrix} -0.180 & 0.727 & -0.662 \\ -0.790 & 0.295 & 0.538 \\ -0.587 & -0.620 & -0.522 \end{pmatrix}.$$

Dále vypočteme již konkrétní matice souřadnic $\mathbf{F}, \mathbf{Y}, \mathbf{G}$ a $\mathbf{X}$, které budou vypadat následovně

$$\mathbf{F} = \begin{pmatrix} -0.469 & 0.016 \\ 0.339 & -0.138 \\ 0.487 & 0.200 \end{pmatrix},$$

$$\mathbf{Y} = \begin{pmatrix} -0.306 & -1.347 \\ 1.174 & 0.476 \\ -1.270 & 1.031 \end{pmatrix},$$

$$\mathbf{G} = \begin{pmatrix} -0.132 & -0.163 \\ 0.508 & 0.058 \\ -0.550 & 0.125 \end{pmatrix},$$

a

$$\mathbf{X} = \begin{pmatrix} -1.084 & 0.133 \\ 0.783 & -1.137 \\ 1.125 & 1.654 \end{pmatrix}.$$

Naše kontingenční tabulka je rozměru 3×3, proto je maximální počet dimenzí, ve kterých můžeme tato data zobrazit, roven dvěma. To také odpovídá maticím $\mathbf{F}, \mathbf{Y}, \mathbf{G}$ a $\mathbf{X}$, které jsou rozměru 3×2. Obsahují tedy souřadnice pro nejvýše dvourozměrný prostor.

2.3 Znázornění výsledků

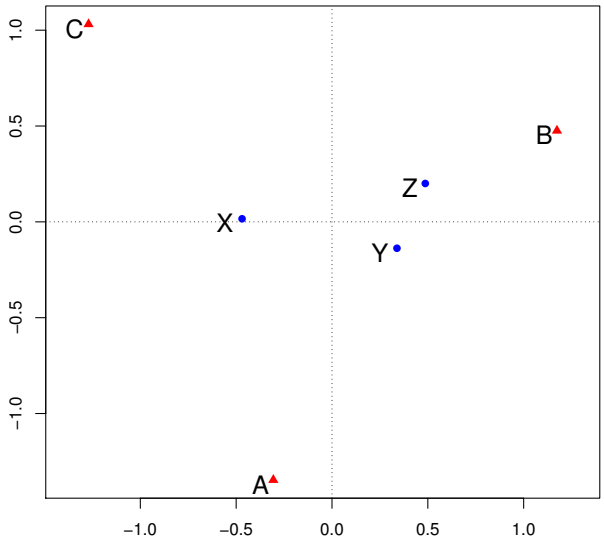
Dospěli jsme do situace, kdy máme souřadnice bodů a je třeba je graficky znázornit a vyvodit závěr. Tomuto grafickému znázornění říkáme korespondeční mapa. Obecně platí, že čím blíže k sobě jednotlivé body jsou, tím více spolu korespondují kategorie, které tyto body zastupují. Kromě korespondence vzájemných bodů lze sledovat například jejich polohu vůči hlavním osám.

Korespondenční mapy dělíme na dva typy. Prvním typem jsou asymetrické mapy, které využíváme v případě, kdy nás zajímají pouze řádkové nebo sloupcové kategorie. V asymetrických mapách jsou profily, které analyzujeme, koncentrovány kolem počátku. Výsledek je tedy špatně čitelný a proto častěji využíváme druhý typ korespondenčních map, a to mapy symetrické. V těch jsou všechny body rovnoměrně rozptýleny v prostoru. Proto je výsledek lépe interpretovatelný.

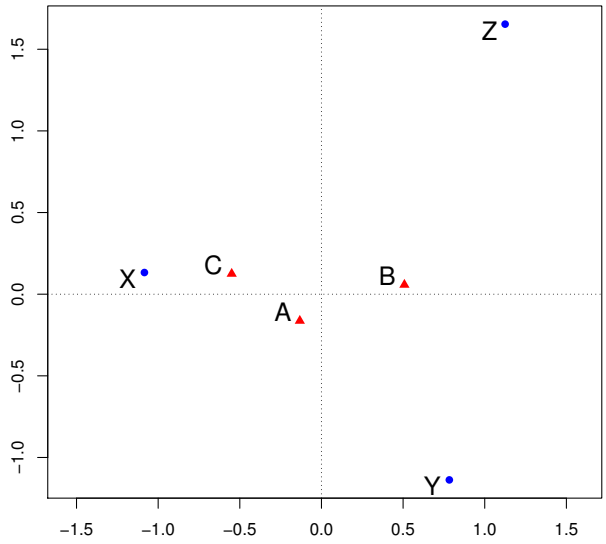
Nyní si toto grafické znázornění ukážeme na našich ilustračních datech. Pokud použijeme asymetrickou mapu řádkových profilů, výsledkem bude korespondenční mapa na obrázku 1. Pokud zvolíme asymetrickou mapu sloupcových profilů, bude

výsledkem mapa na obrázku 2. A na závěr, pokud zvolíme symetrickou mapu sloupcových profilů, dostaneme korespondenční mapu na obrázku 3. Na těchto mapách nyní přehledně vidíme, které kategorie spolu korespondují.

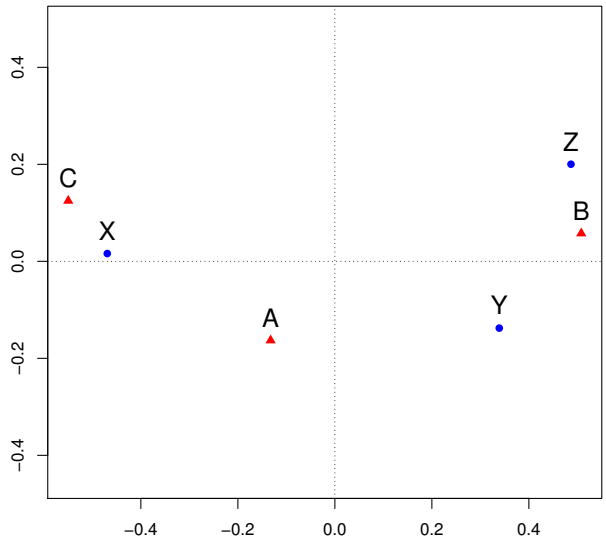
U map asymetrických jsou výsledky hůře čitelné, ale u mapy symetrické je velmi dobře vidět, že například sloupcová kategorie B koresponduje s řádkovými kategoriemi Y a Z. Stejnou situaci můžeme najít u kategorií C a X. Kategorie A je nejblíže kategoriím X a Y, i když je od zbytku kategorií poměrně vzdálená.



Obrázek 1: Korespondenční mapa řádkových profilů pro ilustrační data.



Obrázek 2: Korespondenční mapa sloupcových profilů pro ilustrační data.



Obrázek 3: Korespondenční mapa řádkových a sloupcových profilů pro ilustrační data.

2.4 Hodnocení výsledků

V tomto okamžiku již máme k dispozici výsledky, které jsme znázornili do korespondenční mapy a vyvodili určitý závěr. Ještě je však potřeba zhodnotit tyto výsledky a říci, jakou část informace obsažené v kontingenční tabulce zahrnují. Pokud bychom zjistili, že naše výsledky neobsahují dostatečnou část této informace, bylo by potřeba do korespondenční mapy přidat další rozměr.

První dvě veličiny se týkají celé kontingenční tabulky, zatímco ostatní jsou již spojeny s konkrétními řádkovými, resp. sloupcovými kategoriemi.

Základním ukazatelem, který popisuje rozptýlení bodů, které odpovídají našim kategoriím, je celková inerce. Čím vyšší hodnoty nabývá, tím je rozptýlení bodů větší. Vypočítáme ji podle vzorce

$$I = \sum_{i=1}^{r} p_{i+}(\mathbf{r}_i - \mathbf{c})^T \mathbf{D}_c^{-1}(\mathbf{r}_i - \mathbf{c}) = \sum_{j=1}^{s} p_{+j}(\mathbf{c}_j - \mathbf{r})^T \mathbf{D}_r^{-1}(\mathbf{c}_j - \mathbf{r}), \tag{2.19}$$

tedy jako vážený průměr chí-kvadrát vzdáleností řádkových, resp. sloupcových profilů od svého průměru, kterým je vektor $\mathbf{c}$, resp. vektor $\mathbf{r}$. Jako váhy používáme relativní řádkové četnosti p_{i+}, resp. relativní sloupcové četnosti p_{+j}.

Celková inerce se dá také vypočítat jako

$$I = \frac{Z}{n} = \sum_{i=1}^{k} l_i^2, \tag{2.20}$$

kde Z je Pearsonova statistika, l_i jsou zobecněné singulární hodnoty obsažené v matici $\mathbf{\Gamma}$ a k je počet rozměrů v našem řešení.

Příspěvek jednotlivých řádků a sloupců k celkové inerci pak spočítáme jako

$$\frac{\sum_{i=1}^{k} l_i^2}{\sum_{i=1}^{r} l_i^2}, \tag{2.21}$$

kde k opět odpovídá počtu rozměrů v řešení a r je počet singulárních hodnot. Tyto příspěvky nám dávají informaci o kvalitě řešení v závislosti na počtu rozměrů, který jsme zvolili. Pokud bude tento poměr kolem hodnoty jedna, je počet rozměrů dostačující a řešení považujeme za dobré.

Dále nás zajímají celkové řádkové, resp. sloupcové inerce jednotlivých kategorií, které vypočteme jako součet druhých mocnin prvků v daném řádku, resp. sloupci matice $\mathbf{Z}$. Tyto inerce lze taktéž vypočítat těmito vztahy

$$\sum_{l=1}^{k} r_i f_{il}^2, \tag{2.22}$$

pro řádkové kategorie a

$$\sum_{l=1}^{k} c_j g_{jl}^2, \tag{2.23}$$

pro sloupcové kategorie.

Příspěvky řádkových, resp. sloupcových bodů k inerci v daném rozměru vyjadřují relativní míru vlivu kategorií na orientaci os a vypočítáme je jako

$$\frac{r_i f_{ik}^2}{l_k^2}, \tag{2.24}$$

resp.

$$\frac{c_j g_{jk}^2}{l_k^2}, \tag{2.25}$$

kde r_i jsou prvky vektoru řádkových zátěží, c_j prvky vektoru sloupcových zátěží, f_{ik} je rovno prvkům matice $\mathbf{F}$, g_{jk} jsou prvky matice $\mathbf{G}$ a l_k jsou opět prvky matice $\mathbf{\Gamma}$.

Posledními ukazateli jsou příspěvky os k reprodukci řádkových, resp. sloupcových kategorií, které vyjadřují relativní podíl řádkové inerce na vysvětlení celkové řádkové inerce. Pokud sečteme prvních k příspěvků, dostaneme hodnotu kvality zobrazení dané řádkové, resp. sloupcové kategorie. Dané příspěvky vypočítáme jako

$$\frac{r_i f_{ik}^2}{\sum_{l=1}^{k} r_i f_{il}^2} \tag{2.26}$$

pro řádkové kategorie a

$$\frac{c_j g_{jk}^2}{\sum_{l=1}^{k} c_j g_{jl}^2} \tag{2.27}$$

pro sloupcové kategorie.

Nyní se již naposledy podíváme na náš ilustrační příklad a zhodnotíme získané výsledky. Celková inerce pro tato data je rovna 0,202. Příspěvky řádkových a sloupcových profilů k celkové inerci jsou pro první rozměr 0,93 a pro druhý je to 0,07, což dává dohromady součet 1. Vidíme, že dvourozměrný prostor je maximální možný rozměr, ve kterém lze tato data zobrazit, protože

$$min\{r-1, s-1\} = 2;$$

součet 1 tedy dává smysl. Ostatní ukazatele týkající se těchto dat jsou znázorněny v tabulkách 6 a 7.

Tabulka 6: Řádkové kategorie pro ilustrační data

Kategorie	X	Y	Z	Celkem
Celková řádková inerce	0.101	0.046	0.055	0.202
Příspěvky řádkových bodů	0.536	0.211	0.253	1
k inerci v odpovídajícím rozměru	0.008	0.445	0.547	1
Podíly hlavních os na	0.999	0.858	0.855	-
vysvětlení odpovídající inerce	0.001	0.142	0.145	-
Celkem	1	1	1	-

Tabulka 7: Sloupcové kategorie pro ilustrační data

Kategorie	A	B	C	Celkem
Celková sloupcové inerce	0.015	0.100	0.086	0.202
Příspěvky sloupcových bodů	0.032	0.529	0.439	1
k inerci v odpovídajícím rozměru	0.624	0.087	0.289	1
Podíly hlavních os na	0.398	0.987	0.951	-
vysvětlení odpovídající inerce	0.602	0.013	0.049	-
Celkem	1	1	1	-

Podle podílů hlavních os na vysvětlení odpovídající inerce vidíme, že jsme z našich dat získali maximum informací. Opět je to díky tomu, že tyto data lze zobrazit nejvýše v dvourozměrném prostoru, což jsme provedli. Více informací tedy získat nelze a proto můžeme naše řešení považovat za vyhovující.

3 Aplikace korespondenční analýzy

Tato kapitola je zaměřena na reálné aplikace korespondenční analýzy. Na úvod je zde menší podkapitola věnována softwaru R, který byl při zpracování této práce využíván. Dále zde najdeme praktické využití korespondenční analýzy na reálných datech, konkrétně na dvou datových souborech. V této kapitole bylo využito zdrojů [6], [8] a [10]. Data použitá v této kapitole pochází z [3] a [9].

3.1 Využití softwaru R v korespondenční analýze

R je programovací jazyk (resp. software) určený pro statistickou analýzu dat a grafické zobrazení. Původními autory tohoto prostředí jsou Ross Ihaka a Robert Gentleman. V současnosti je vyvíjen skupinou R Development Core Team. Vychází z programovacího jazyku S. Jeho největší výhodou je jeho bezplatnost a možnost dále rozšiřovat funkce pomocí tzv. knihoven, což jsou soubory dalších funkcí vytvářených komunitou, které lze v tomto prostředí využívat.

Při provádění výpočtů korespondenční analýzy lze docela efektivně tohoto softwaru využít. Pokud přitom využijeme pouze jeho základní funkce, výpočet je poměrně obsáhlý ve srovnání se situací při použití funkcí, které lze do R přidat v podobě knihoven. Existuje naštěstí knihovna, která se zaměřuje výhradně na funkce spojené s korespondenční analýzou. To nám značně ulehčí práci při zpracovávání kategoriálních dat touto metodou.

Knihovna `ca` je určena pro výpočet a zobrazení výsledků jak jednoduché korespondenční analýzy, tak její vícenásobné varianty. Autory jsou Oleg Nenadic a již dříve zmiňovaný Michael Greenacre. Ke své funkci vyžaduje instalaci knihovny `rgl`, která rozšiřuje prostředí R například o funkce související s 3D grafikou. Zájemce o tuto knihovnu bych odkázal na [10], kde jsou knihovny `ca` a `rgl` dostupné ke stažení včetně veškeré dokumentace potřebné k jejich pochopení. Při výpočtech v této práci bylo využito zejména funkcí `ca` a `plot.ca`.

3.2 Barva očí a barva vlasů

Byl proveden průzkum na souboru 592 lidí, kde byla zjišťována jejich barva očí a vlasů. Výsledkem je tabulka 8, kde vidíme četnosti jednotlivých kombinací barvy očí a vlasů. Použitá data pochází z [7].

Tabulka 8: Barva očí a barva vlasů

	Černá	Hnědá	Zrzavá	Blond
Hnědá	68	119	26	7
Modrá	20	84	17	94
Světle hnědá	15	54	14	10
Zelená	5	29	14	16

Ještě než přistoupíme k samotné korespondenční analýze, můžeme si připomenout nástroje, které jsme si uvedli v kapitole o kontingenčních tabulkách.

Prvním je test nezávislosti. Podmínka (1.9) pro použití tohoto testu je pro naše data splněna. Je tedy třeba vypočítat hodnotu statistiky Z. Ta je rovna podle vztahu (1.8) hodnotě 138.29. Hodnota kvantilu $\chi^2_{9,0.95}$ je rovna 16.919, naše statistika se tedy realizuje v kritickém oboru $W = \langle 16.919, \infty)$ a proto proměnné barva vlasů a barva očí nejsou nezávislé.

Dalším nástrojem je znaménkové schéma, díky kterému zjistíme, které kategorie způsobují zjištěnou závislost. Nejdříve je třeba vypočítat očekávané hodnoty těchto kategorií a příslušná rezidua. Následně jednotlivým hodnotám přiřadíme znaménka podle tabulky 4. Výsledky znázorníme do tabulky 9, kde vždy první řádek odpovídá očekávané hodnotě, druhý standardizovanému reziduu a třetí přiřazenému znaménku.

Tabulka 9: Barva očí a barva vlasů: Znaménkové schéma

	Černá	Hnědá	Zrzavá	Blond
Hnědá	40.135 4.556 +++	106.284 1.362 0	26.385 −0.077 0	47.196 −6.099 −−−
Modrá	39.223 −3.176 −−	103.868 −2.147 −	25.785 −1.769 0	46.123 7.341 +++
Světle hnědá	16.966 −0.484 0	44.929 1.408 0	11.154 0.860 0	19.951 −2.266 −
Zelená	11.676 −1.973 −	30.919 −0.354 0	7.676 2.298 +	13.730 0.620 0

Na místech, kde vidíme nejvíce znamének, jsou hodnoty, které přispívají k závislosti těchto proměnných nejvíce. Ukázali jsme tedy, že mezi proměnnými existuje závislost a zjistili jsme, které hodnoty k této závislosti nejvíce přispěly. Můžeme tedy přistoupit k samotné korespondenční analýze.

Matice $\mathbf{N}$ v tomto případě odpovídá tabulce 8. Korespondenční matice $\mathbf{P}$ bude podle vztahu (2.1) rovna

$$\mathbf{P} = \begin{pmatrix} 0.115 & 0.201 & 0.044 & 0.012 \\ 0.034 & 0.142 & 0.029 & 0.159 \\ 0.025 & 0.091 & 0.024 & 0.017 \\ 0.008 & 0.049 & 0.024 & 0.027 \end{pmatrix}.$$

Vektor řádkových zátěží bude podle vztahů (2.3) a (2.4) vypadat následovně

$$\mathbf{r} = \begin{pmatrix} 0.372 \\ 0.363 \\ 0.157 \\ 0.108 \end{pmatrix}.$$

Vektor sloupcových zátěží bude obdobně podle vztahů (2.5) a (2.6) ve tvaru

$$\mathbf{c}^T = \begin{pmatrix} 0.182 & 0.483 & 0.120 & 0.215 \end{pmatrix}.$$

Na základě těchto výpočtů můžeme korespondenční matici vyjádřit pomocí tabulky 10.

Tabulka 10: Barva očí a barva vlasů: Korespondenční matice.

	Černá	Hnědá	Zrzavá	Blond	Celkem
Hnědá	0.115	0.201	0.044	0.012	0.372
Modrá	0.034	0.142	0.029	0.159	0.363
Světle hnědá	0.025	0.091	0.024	0.017	0.157
Zelená	0.008	0.049	0.024	0.027	0.108
Celkem	0.182	0.483	0.120	0.215	1

Dále můžeme vypočítat matici řádkových profilů $\mathbf{R}$ podle vztahů (2.7), (2.8) a matici sloupcových profilů $\mathbf{C}$ podle vztahů (2.9), (2.10). Tyto matice budou vypadat následovně

$$\mathbf{R} = \begin{pmatrix} 0.309 & 0.541 & 0.118 & 0.320 \\ 0.093 & 0.391 & 0.079 & 0.437 \\ 0.161 & 0.581 & 0.151 & 0.108 \\ 0.078 & 0.453 & 0.219 & 0.250 \end{pmatrix},$$

$$\mathbf{C} = \begin{pmatrix} 0.630 & 0.416 & 0.366 & 0.055 \\ 0.185 & 0.294 & 0.239 & 0.740 \\ 0.139 & 0.189 & 0.197 & 0.079 \\ 0.046 & 0.101 & 0.197 & 0.126 \end{pmatrix}.$$

Nyní máme vypočteny základní matice a vektory a můžeme tedy určit jednotlivé souřadnice našich kategorií. K tomu ještě potřebujeme matici $\mathbf{Z}$, kterou vypočteme podle vztahu (2.13), kde jednotlivé prvky budou rovny (2.14). Matice $\mathbf{Z}$ bude vypadat takto

$$\mathbf{Z} = \begin{pmatrix} 0.181 & 0.051 & -0.003 & -0.240 \\ -0.126 & -0.080 & -0.071 & 0.290 \\ -0.020 & 0.056 & 0.035 & -0.092 \\ -0.080 & -0.014 & 0.094 & 0.025 \end{pmatrix}.$$

Dále je třeba provést singulární rozklad této matice podle (2.12). Ten bude roven

$$\mathbf{U} = \begin{pmatrix} -0.657 & -0.361 & 0.258 & 0.610 \\ 0.722 & -0.335 & -0.056 & 0.603 \\ -0.184 & 0.445 & -0.782 & 0.396 \\ 0.116 & 0.748 & 0.565 & 0.329 \end{pmatrix},$$

$$\mathbf{\Gamma} = \begin{pmatrix} 0.4569 & 0 & 0 & 0 \\ 0 & 0.1491 & 0 & 0 \\ 0 & 0 & 0.5097 & 0 \\ 0 & 0 & 0 & 9.533 * 10^{-18} \end{pmatrix}$$

a

$$\mathbf{V} = \begin{pmatrix} -0.472 & -0.615 & 0.465 & -0.427 \\ -0.226 & 0.152 & -0.665 & -0.695 \\ -0.098 & 0.742 & 0.565 & -0.346 \\ 0.847 & -0.216 & 0.147 & -0.463 \end{pmatrix}.$$

Za pomoci získaných údajů můžeme jednoduše dopočítat matice souřadnic $\mathbf{F}, \mathbf{Y}, \mathbf{G}$ a $\mathbf{X}$, podle vztahů (2.15), (2.16), (2.17) a (2.18). Matice souřadnic budou rovny

$$\mathbf{F} = \begin{pmatrix} -0.492 & -0.088 & 0.022 \\ 0.547 & -0.083 & -0.005 \\ -0.213 & 0.167 & -0.101 \\ 0.162 & 0.339 & 0.088 \end{pmatrix},$$

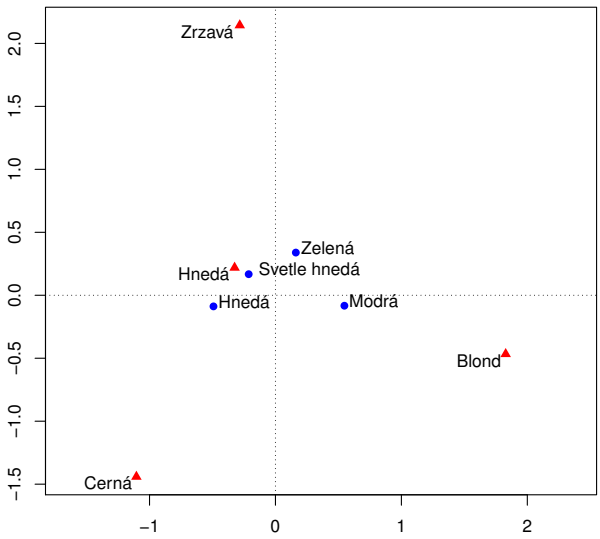
$$\mathbf{Y} = \begin{pmatrix} -1.104 & -1.441 & 1.089 \\ -0.324 & 0.219 & -0.957 \\ -0.283 & 2.144 & 1.631 \\ 1.828 & -0.467 & 0.318 \end{pmatrix},$$

$$\mathbf{G} = \begin{pmatrix} -0.505 & -0.215 & 0.056 \\ -0.148 & 0.033 & -0.049 \\ -0.130 & 0.320 & 0.083 \\ 0.835 & -0.070 & 0.016 \end{pmatrix}$$

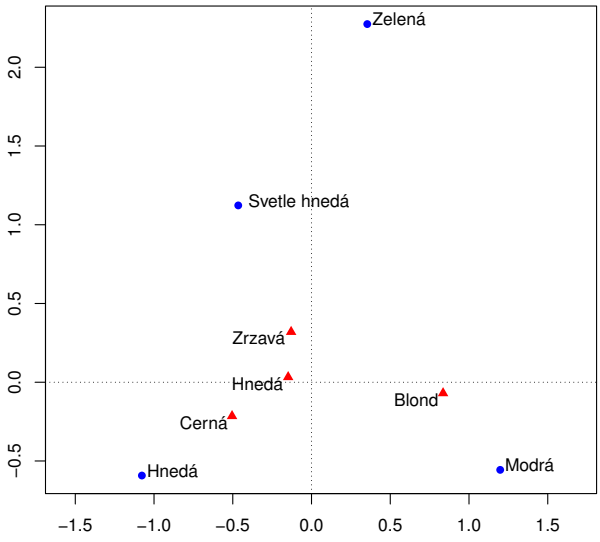
a

$$\mathbf{X} = \begin{pmatrix} -1.077 & -0.592 & 0.424 \\ 1.198 & -0.556 & -0.092 \\ -0.465 & 1.123 & -1.972 \\ 0.354 & 2.274 & 1.718 \end{pmatrix}.$$

Nyní můžeme přejít k samotnému znázornění výsledků do korespondenční mapy. Pokud zvolíme asymetrickou mapu řádkových profilů dostaneme výsledek na obrázku 4. Při volbě asymetrické mapy sloupcových profilů bude výsledkem obrázek 5.

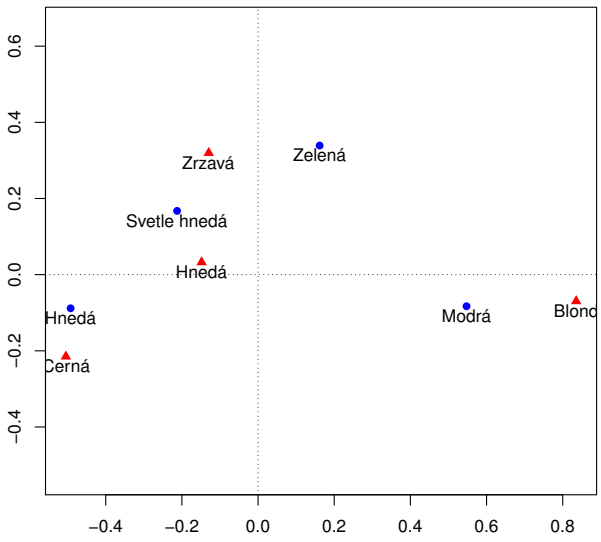


Obrázek 4: Barva očí a barva vlasů: Korespondenční mapa řádkových profilů.



Obrázek 5: Barva očí a barva vlasů: Korespondenční mapa sloupcových profilů.

A na závěr, pokud zvolíme symetrickou mapu sloupcových a řádkových profilů, dostaneme výsledek na obrázku 6.



Obrázek 6: Barva očí a barva vlasů: Korespondenční mapa řádkových a sloupcových profilů.

Nejlepší informaci o vztazích mezi hodnotami proměnných v naší tabulce nám dá symetrická mapa. Vidíme, že se z mapy dá celkem dobře vyčíst, které kategorie spolu korespondují. Například si můžeme povšimnout, že modré oči se nejčastěji vyskytují u lidí s blond vlasy. Dále vidíme, že hnědé oči najdeme nejčastěji u lidí s černými vlasy. Světle hnědé oči se nejčastěji vyskytly u lidí s hnědými a zrzavými vlasy. A nakonec, zelené oči byly nejvíce pozorovány u lidí se zrzavými vlasy. U asymetrických map můžeme pozorovat podobné výsledky. Nejsou však tak jasné jako u mapy symetrické.

Dále je potřeba zhodnotit výsledky, kterých jsme dosáhli. Celková inerce pro tato data je rovna 0,233. Příspěvky řádkových a sloupcových profilů k celkové inerci jsou 0,894 pro jednorozměrný prostor a 0,095 navíc pro ten dvourozměrný. Dohromady tedy dávají součet 0,989. Naše řešení tedy můžeme považovat za dobré. Další veličiny popisující kvalitu našeho řešení najdeme v tabulkách 11 a 12.

Tabulka 11: Barva očí a barva vlasů: Řádkové kategorie.

Kategorie		Hnědá	Modrá	Světle hnědá	Zelená	Celkem
Celková řádková inerce		0.093	0.111	0.013	0.016	0.233
Příspěvky řádkových bodů k inerci v odpovídajícím rozměru	1	0.431	0.521	0.034	0.014	1
	2	0.130	0.112	0.198	0.559	1
Podíly hlavních os na vysvětlení odpovídající inerce	1	0.967	0.977	0.542	0.176	-
	2	0.031	0.022	0.336	0.773	-
Celkem		0.998	0.999	0.879	0.948	-

Tabulka 12: Barva očí a barva vlasů: Sloupcové kategorie.

Kategorie		Černá	Hnědá	Zrzavá	Blond	Celkem
Celková sloupcová inerce		0.055	0.012	0.015	0.151	0.233
Příspěvky sloupcových bodů k inerci v odpovídajícím rozměru	1	0.222	0.051	0.010	0.717	1
	2	0.379	0.023	0.551	0.047	1
Podíly hlavních os na vysvětlení odpovídající inerce	1	0.838	0.864	0.133	0.993	-
	2	0.152	0.042	0.812	0.007	-
Celkem		0.990	0.906	0.945	0.999	-

Zde vidíme, že podíly hlavních os na vysvětlení odpovídající inerce jsou pro skoro všechny kategorie větší jak 0,9. Tedy více jak 90% informací o těchto kategoriích je zahrnuto v našem řešení. Nejmenší podíl najdeme u kategorie Světle hnědé oči, kde se nám podařilo získat přibližně 87,9% všech informací obsažených v těchto datech. Zbytek těchto informací bychom našli ve zbývajícím třetím rozměru. Řešení tedy můžeme i z tohoto hlediska považovat za vyhovující.

3.3 Zdraví člověka a jeho věk

V tomto případě byl proveden průzkum na skupině 6371 osob a zjišťovalo se, v jakém zdravotním stavu jsou tyto osoby a jejich věk. Výsledky jsou shrnuty v tabulce 13. Použitá data pochází z [8].

Tabulka 13: Zdraví člověka a jeho věk

	Velmi dobrý	Dobrý	Normální	Špatný	Velmi špatný
16-24	243	789	167	18	6
25-34	220	809	164	35	6
35-44	147	658	181	41	8
45-54	90	469	236	50	16
55-64	53	414	306	106	30
65-74	44	267	284	98	20
75+	20	136	157	66	17

Jako první budeme zkoumat vztah mezi proměnnými v této tabulce. Podmínka (1.9) je pro tato data splněna. Hodnota statistiky Z je podle vztahu (1.8) rovna 893.546 a $\chi^2_{24,0.95} = 36.415$. Statistika Z se tedy realizuje v kritickém oboru $W = \langle 36.415, \infty)$ a proto o kategoriích věk člověka a jeho zdravotní stav můžeme říci, že nejsou nezávislé.

Pomocí znaménkového schématu nyní zjistíme, které hodnoty přispívají k této závislosti. Znaménkové schéma je zpracováno v tabulce 14, kde opět první řádek odpovídá očekávané hodnotě, druhý standardizovanému reziduu a třetí již konkrétnímu přiřazenému znaménku.

Podle schématu vidíme, že tabulka je rozdělena na pět částí. Můžeme říci, že je v jistém smyslu symetrická. V první části jsou lidé ve věku 16-44 let s velmi dobrým až dobrým zdravím. Vidíme, že naměřené četnosti těchto kategorií jsou daleko vyšší než četnosti očekávané. U té stejné věkové kategorie můžeme vysledovat, že četnosti jedinců s normálním až velmi špatným zdravím byla naopak daleko menší než ta očekávaná. Třetí část tabulky odpovídá věkové skupině 45-54 let, kde se naměřené četnosti přibližně shodují s těmi očekávanými. Poslední dva řádky tabulky zahrnují věkovou kategorii 55 a více let, kde vidíme, že pozorované četnosti jedinců s vemi dobrým až dobrým zdravím jsou daleko menší jako ty

očekávané a naopak četnosti jedinců s normálním až velmi špatným zdravím jsou větší než očekávané. Tyto výsledky jsou celkem logické a odpovídají realitě.

Tabulka 14: Znaménkové schéma - Zdraví člověka a jeho věk

	Velmi dobrý	Dobrý	Normální	Špatný	Velmi špatný
16-24	158.117 6.836 +++	685.495 4.231 +++	289.332 −7.339 − − −	80.123 −6.976 − − −	19.934 −3.121 −−
25-34	158.245 4.995 +++	686.051 5.016 +++	289.567 −7.531 − − −	80.188 −5.068 − − −	19.950 −3.124 −−
35-44	132.726 1.271 0	575.415 3.652 +++	242.870 −4.026 − − −	67.256 −3.208 −−	16.733 −2.133 −
45-54	110.412 −1.945 0	478.679 −0.425 0	202.040 2.453 +	55.950 −0.788 0	13.920 0.564 0
55-64	116.568 −5.930 − − −	505.365 −4.202 − − −	213.303 6.484 +++	59.069 6.152 +++	14.696 4.006 +++
65-74	91.433 −4.985 − − −	396.397 −6.684 − − −	167.311 9.170 +++	46.332 7.635 +++	11.527 2.505 +
75+	50.782 −4.329 − − −	220.159 −6.684 − − −	92.924 9.170 +++	25.733 7.968 +++	6.402 4.198 +++

Nyní můžeme přejít na samotnou korespondenční analýzu. Matice $\mathbf{N}$ odpovídá původní tabulce 13. Po výpočtu relativních četností a vektorů řádkových a sloupcových profilů dostaneme korespondenční matici $\mathbf{P}$, která odpovídá tabulce 15.

Dále vypočítáme matici řádkových profilů $\mathbf{R}$ a matici sloupcových profilů $\mathbf{C}$. Poté můžeme jednoduše vypočítat matici $\mathbf{Z}$, která bude rovna

$$\mathbf{Z} = \begin{pmatrix} 0.0846 & 0.0495 & -0.0901 & -0.0869 & -0.0391 \\ 0.0615 & 0.0588 & -0.0924 & -0.0618 & -0.0391 \\ 0.0155 & 0.0431 & -0.0497 & -0.0401 & -0.0267 \\ -0.0243 & -0.0055 & 0.0299 & -0.0100 & 0.0070 \\ -0.0738 & -0.0509 & 0.0795 & 0.0765 & 0.0500 \\ -0.0621 & -0.0814 & 0.1130 & 0.0951 & 0.0313 \\ -0.0541 & -0.0711 & 0.0833 & 0.0994 & 0.0525 \end{pmatrix}.$$

Tabulka 15: Zdraví člověka a jeho věk: Korespondenční matice

	Velmi dobrý	Dobrý	Normální	Špatný	Velmi špatný	Celkem
16-24	0.0381	0.1238	0.0262	0.0028	0.0009	0.1935
25-34	0.0345	0.1270	0.0257	0.0055	0.0009	0.1937
35-44	0.0231	0.1033	0.0284	0.0064	0.0013	0.1625
45-54	0.0141	0.0736	0.0370	0.0078	0.0025	0.1351
55-64	0.0083	0.0650	0.0480	0.0166	0.0047	0.1427
65-74	0.0069	0.0419	0.0446	0.0154	0.0031	0.1119
75+	0.0031	0.0213	0.0246	0.0104	0.0027	0.0622
Celkem	0.1282	0.5560	0.2347	0.0650	0.0162	1

Za pomoci singulárního rozkladu matice $\mathbf{Z}$ dostaneme matice jednotlivých souřadnic, které budou vypadat následovně

$$\mathbf{F} = \begin{pmatrix} -0.3681 & 0.0422 & 0.0294 & 0.0202 \\ -0.3299 & 0.0195 & -0.0267 & -0.0117 \\ -0.1990 & -0.0407 & -0.0257 & -0.0236 \\ 0.0709 & -0.0709 & 0.0447 & 0.0170 \\ 0.3955 & -0.0332 & -0.0357 & 0.0192 \\ 0.5406 & 0.0343 & 0.0509 & -0.0372 \\ 0.6585 & 0.0835 & -0.0472 & 0.0229 \end{pmatrix},$$

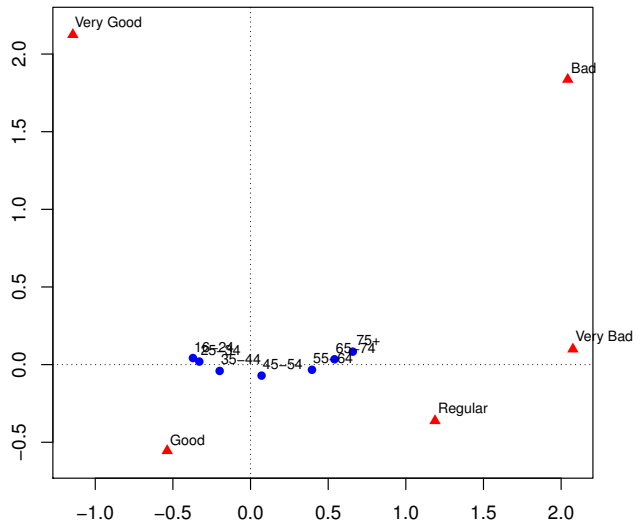
$$\mathbf{Y} = \begin{pmatrix} -1.1393 & 2.0913 & 0.8315 & 0.3948 \\ -0.5326 & -0.5859 & -0.4671 & -0.1923 \\ 1.1923 & -0.3943 & 1.2747 & -0.0922 \\ 2.0470 & 1.8051 & -2.1690 & -1.5750 \\ 2.0805 & 0.0678 & -2.5245 & 7.0180 \end{pmatrix},$$

$$\mathbf{G} = \begin{pmatrix} -0.4208 & 0.0956 & 0.0299 & 0.0086 \\ -0.1967 & -0.0268 & -0.0168 & -0.0042 \\ 0.4403 & -0.0180 & 0.0458 & -0.0020 \\ 0.7560 & 0.0825 & -0.0780 & -0.0343 \\ 0.7684 & 0.0031 & -0.0908 & 0.1530 \end{pmatrix}$$

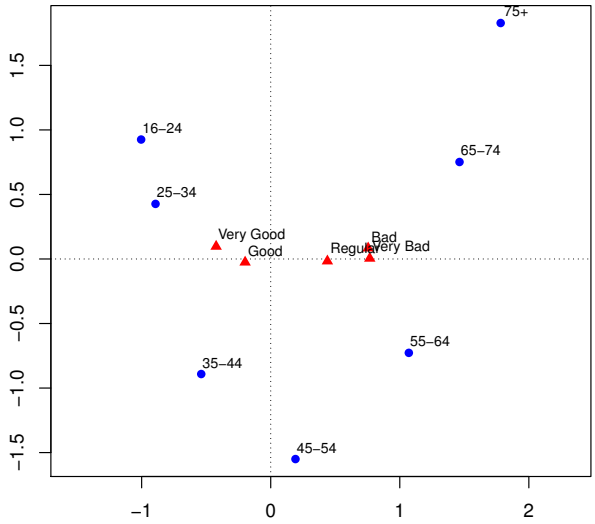
a

$$\mathbf{X} = \begin{pmatrix} -0.9967 & 0.9232 & 0.8172 & 0.9280 \\ -0.8932 & 0.4275 & -0.7418 & -0.5356 \\ -0.5387 & -0.8906 & -0.7142 & -1.0844 \\ 0.1920 & -1.5503 & 1.2436 & 0.7798 \\ 1.0709 & -0.7271 & -0.9932 & 0.8784 \\ 1.4639 & 0.7501 & 1.4155 & -1.7061 \\ 1.7830 & 1.8274 & -1.3140 & 1.0511 \end{pmatrix}.$$

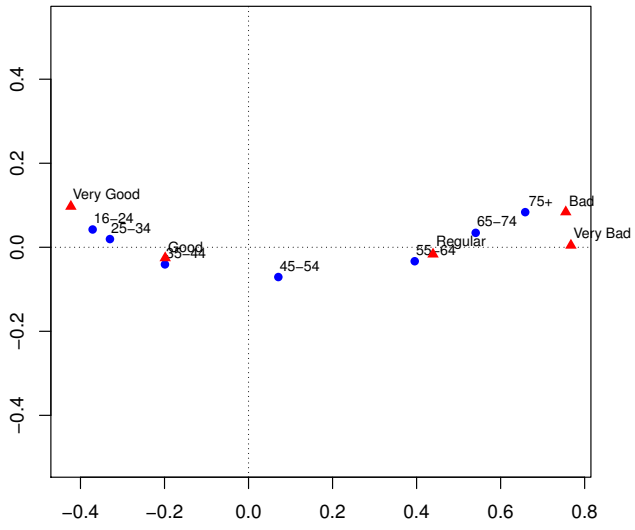
Nyní můžeme naše výsledky znázornit do korespondenční mapy. Jako první si zvolíme asymetrickou mapu řádkových profilů. Výsledek je na obrázku 7. Dále, pokud budeme uvažovat asymetrickou mapu sloupcových profilů, dostaneme obrázek 8. A na závěr vybereme symetrickou mapu (obrázek 9).



Obrázek 7: Zdraví člověka a jeho věk: Korespondenční mapa řádkových profilů.



Obrázek 8: Zdraví člověka a jeho věk: Korespondenční mapa sloupcových profilů.



Obrázek 9: Zdraví člověka a jeho věk: Korespondenční mapa řádkových a sloupcových profilů.

Ze symetrické mapy je patrné, že velmi dobrý a dobrý zdravotní stav měly osoby ve věku 16-34 let. Dobrý zdravotní stav mají také lidé ve věku 35-44 let. U věkové skupiny 45-54 let jsme nejčastěji pozorovali dobrý až normální zdravotní stav. Normální zdravotní stav se nejčastěji vyskytoval u věkové skupiny 55-64 let. U osob z věkové skupiny 65-74 let jsme si mohli nejčastěji všimnout normálního až špatného či dokonce velmi špatného zdravotního stavu. No a nakonec věková skupina 75+ korespondovala se špatným až velmi špatným zdravotním stavem.

Nyní opět přejdeme ke zhodnocení našich výsledků. Celková inerce je v tomto případě rovna 0,140. Příspěvky řádkových a sloupcových profilů k celkové inerci jsou pro jednorozměrný prostor 0,974 a pro prostor dvourozměrný navíc 0,015. Dohromady je to tedy 0,989. Další ukazatele kvality výsledků najdeme v tabulkách 16 a 17.

Tabulka 16: Zdraví člověka a jeho věk: Řádkové kategorie.

Kategorie		16-24	25-34	35-44	45-54	55-64	65-74	75+	Celkem
Celková řádková inerce		0.027	0.021	0.007	0.002	0.023	0.033	0.028	0.140
Příspěvky řádkových bodů k inerci v odpovídajícím rozměru	1	0.193	0.154	0.047	0.005	0.163	0.239	0.197	1
	2	0.164	0.035	0.129	0.325	0.075	0.063	0.208	1
Podíly hlavních os na vysvětlení odpovídající inerce	1	0.978	0.989	0.932	0.408	0.983	0.983	0.978	-
	2	0.013	0.003	0.039	0.407	0.007	0.004	0.016	-
Celkem		0.991	0.992	0.971	0.815	0.990	0.987	0.994	-

Tabulka 17: Zdraví člověka a jeho věk: Sloupcové kategorie.

Kategorie		VD	D	N	Š	VŠ	Celkem
Celková sloupcová inerce		0.024	0.022	0.046	0.038	0.010	0.140
Příspěvky sloupcových bodů k inerci v odpovídajícím rozměru	1	0.168	0.160	0.331	0.271	0.070	1
	2	0.579	0.171	0.031	0.219	$2 * 10^{-4}$	1
Podíly hlavních os na vysvětlení odpovídající inerce	1	0.944	0.978	0.987	0.976	0.949	-
	2	0.050	0.016	0.001	0.012	$3 * 10^{-5}$	-
Celkem		0.994	0.994	0.987	0.988	0.949	-

Můžeme si všimnout, že opět u většiny kategorií jsou podíly hlavních os na vysvětlení odpovídající inerce větší jak 0,9. Pouze u věkové skupiny 45 až 54 let je tento podíl roven 0,815. Tedy podařilo se nám získat pouze 81,5% informací o této kategorii z daných dat. Abychom zvýšili tento podíl, bylo by potřeba přejít do vyššího rozměru, v tomto případě lze až do čtvrtého, což by ovšem nebylo možné využít ke grafickému znázornění. Každopádně i v tomto případě lze tak dané výstupy považovat za vyhovující.

Závěr

Cílem této bakalářské práce bylo popsat postupy korespondenční analýzy a ukázat její využití na reálných datech. Před vypracováním této práce jsem se podrobněji s podobnou metodou nesetkal, proto pro mě bylo těžší do této problematiky proniknout. Ukázalo se ovšem, že po pochopení základních pojmů a postupů jsou výpočty spojené s touto metodou poměrně snadné.

Teoretická část této práce se podrobněji věnovala kategoriálním proměnným v kontingenčním tabulkách a tomu, jak popsat vztahy mezi proměnnými v dané tabulce. Dále byly podrobně popsány důležité pojmy a postupy korespondenční analýzy včetně grafického zobrazení výsledků a jejich následného zhodnocení. Praktická část byla zaměřena na demonstrování této metody na dvou reálných datových souborech. U obou se podařilo dojít k snadno interpretovatelným výsledkům a proto svou práci hodnotím jako úspěšnou.

Díky této práci jsem získal nové vědomosti z oblasti kategoriálních dat a vícerozměrných statistických metod. Také jsem se naučil pracovat se softwary R a TEX. Jako obohacující dále považuji práci s anglicky psanou literaturou, proto tuto práci celkově hodnotím jako přínosnou pro mé další studium.

Literatura

[1] Anděl, J., Statistické metody. Praha: Matfyzpress, 2007.

[2] Greenacre, M., Correspondence Analysis in Practice, Second Edition. London: Chapman & Hall/CRC, 2007.

[3] Greenacre, M., The Use of Correspondence Analysis in the Exploration of Health Survey Data. Bilbao: Fundación BBVA, 2002.

[4] Hebák, P. a kol., Vícerozměrné statistické metody (3), 2. vydání. Praha: Informatorium, 2007.

[5] Hron, K., Kunderová, P., Základy počtu pravděpodobnosti a metod matematické statistiky. Olomouc: Vydavatelství Univerzity Palackého, 2013.

[6] Nenadic, O., Greenacre, M., Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package, Journal of Statistical Software, 20, 3 (2007).

[7] Novák, T., Kopeček, M., Typy proměnných, Psychiatrie pro praxi, 11, 4, 176-177 (2010).

[8] CRAN - Package ca [online], dostupné z http://cran.r-project.org/web/packages/ca/index.html, [citováno 7.12.2013].

[9] Correspondence Analysis [online], dostupné z http://www.datavis.ca/courses/grcat/grc5.html, [citováno 7.12.2013].

[10] The R Project for Statistical Computing [online], dostupné z http://www.r-project.org/, [citováno 23.10.2013].