UNIVERZITA PALACKÉHO V OLOMOUCI
P Ř Í R O D O V Ě D E C K Á F A K U L T A
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

DIPLOMOVÁ PRÁCE

Metoda hlavních komponent a její aplikace

Vedoucí diplomové práce:
Mgr. Ondřej Vencálek, Ph.D.
Rok odevzdání: 2013

Vypracovala:
Bc. Zuzana Tonhauserová
AME, II. ročník

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením Mgr. Ondřeje Vencálka, Ph.D. s pomocí uvedené literatury a ostatních informačních zdrojů.

V Olomouci dne 3. prosince 2013

Poděkování

Ráda bych na tomto místě poděkovala především svému vedoucímu diplomové práce Mgr. Ondřeji Vencálkovi, Ph.D. za odbornou pomoc, cenné rady a čas, který mi věnoval při konzultacích. Dále bych chtěla poděkovat své rodině, která mě podporovala po celou dobu studia.

Obsah

Úvod

Metoda hlavních komponent (Principal Component Analysis, PCA) je důležitým nástrojem pro posouzení a prověření kvality mnohorozměrných dat. Řadí se mezi nejstarší a nejvíce používané metody vícerozměrné statistické analýzy a používá se jako pomocný nástroj v dalších metodách [3]. Téměř každou vícerozměrnou úlohu bychom měli začít výpočtem a zobrazením hlavních komponent, abychom poznali a pochopili vstupní data. Mnohorozměrná statistická analýza nebo také analýza mnohorozměrných dat se snaží řešit problém se zachycením více než třírozměrného prostoru při grafickém zobrazování dat. Pro znázornění mnohorozměrných dat je nutné hledat vhodné projekce, velikosti výběrů musí být značně větší než u jednorozměrných dat a mnoho statistických metod naráží na obtíže [4]. Tento problém bývá nejčastěji řešen redukcí dat.

Cílem této práce je seznámit čtenáře s metodou hlavních komponent a aplikovat tuto metodu na reálná data. Při tvorbě této práce mi jako základ posloužily knižní publikace „*Applied multivariate statistical analysis*“ [1] a „*Vícerozměrné statistické metody*“ [2].

Práce je tematicky rozdělena na tři kapitoly. V první kapitole seznámím čtenáře se základními pojmy důležitými pro pochopení problematiky metody hlavních komponent a také poskytnu stručný přehled nejpoužívanějších metod mnohorozměrné statistické analýzy. Druhá kapitola je zaměřena na všeobecný popis metody hlavních komponent. Závěr celé práce tvoří praktická část, ve které na reálných datech předvedu různé způsoby využití této metody. K výpočtům a grafickému vyjádření použiji statistický software R.

1 Základní pojmy

V této kapitole se pokusím čtenáři přiblížit nejdůležitější pojmy, které bychom měli znát pro pochopení metody hlavních komponent a pro orientaci v problematice statistické analýzy mnohorozměrných dat.

1.1 Mnohorozměrná data

Je třeba zmínit důležitý pojem, a tím je zkoumaná náhodná **veličina** (proměnná, statistický znak). Jedná se o vlastnost, kterou sledujeme na prvcích výběru či populace a nabývá různých hodnot. Je důležité uvědomit si, že veličiny, které zjišťujeme na sledovaných objektech, jsou různého typu. Informace k této kapitole jsou čerpány zejména z [5], [6].

Veličiny můžeme rozlišovat podle měřící stupnice, tzv. škály:

- **nominální** – veličiny měřené na nominální škále, tato škála je složena ze dvou či více vzájemně se vylučujících kategorií, které nemohou být seřazeny, tedy dále rozlišujeme nominální veličiny:
  a) binární (dvouhodnotové) – např. pohlaví,
  b) vícehodnotové – např. barva, rodinný stav,
- **ordinální** (pořadové) – stejně jako u nominálních veličin, jsou hodnoty ordinálních veličin seskupeny do neslučitelných kategorií, ale navíc lze tyto hodnoty vzájemně uspořádat, např. úroveň vzdělání, příjmová kategorie, společenská vrstva,
- **poměrové s absolutní nulou** – začíná skutečným nulovým bodem, což umožňuje vypočítat rozdíly (intervaly) mezi hodnotami a také podíly těchto hodnot. Zároveň ale absolutní nula vylučuje záporné hodnoty, např. teplota ve stupních Kelvina, hmotnost, objem,
- **intervalové s relativní nulou** – nemá jednoznačně stanovenou nulovou pozici, umožňuje stanovit vzdálenost (rozdíl) mezi hodnotami definovanou jednotkou měření, např. teplota ve stupních Celsia, stupnice výšky tónu.

Dále veličiny můžeme rozlišovat podle oboru jejich hodnot:

- **spojité** – mohou nabývat nespočetně mnoha hodnot (např. výška, váha, teplota),
- **diskrétní** – nabývají pouze spočetně mnoha hodnot, tyto hodnoty se nazývají varianty, kategorie nebo úrovně (např. počet dětí, počet zaměstnanců).

Také dělíme veličiny na kvantitativní a kvalitativní:

- **kvantitativní** (metrické) – označují veličiny diskrétní, spojité, poměrové a intervalové,
- **kvalitativní** (kategoriální) – označují pouze diskrétní veličiny, a to nominální a ordinální veličiny.

1.1.1 Typy statistických analýz

Rozlišujeme různé typy analýz mnohorozměrných dat, záleží na tom, zda sledované objekty můžeme považovat za nezávislé nebo zda data vznikla jako řada pozorování téhož objektu v různých obdobích. Informace jsou čerpány zejména z [5]. Nyní si zavedeme následující označení:

X_{ijt} → data,

$i = 1, \ldots, n$ → objekty (pozorování), kde n je počet pozorování,

$j = 1, \ldots, p$ → proměnné (vlastnosti), kde p je počet proměnných, tj. dimenze (rozměr) úlohy,

$t = 1, \ldots, T$ → čas, kde T je časový horizont.

Následující tabulka 1.1 znázorňuje různé typy úloh řešené analýzou dat.

Tabulka 1.1: Úlohy řešené analýzou dat

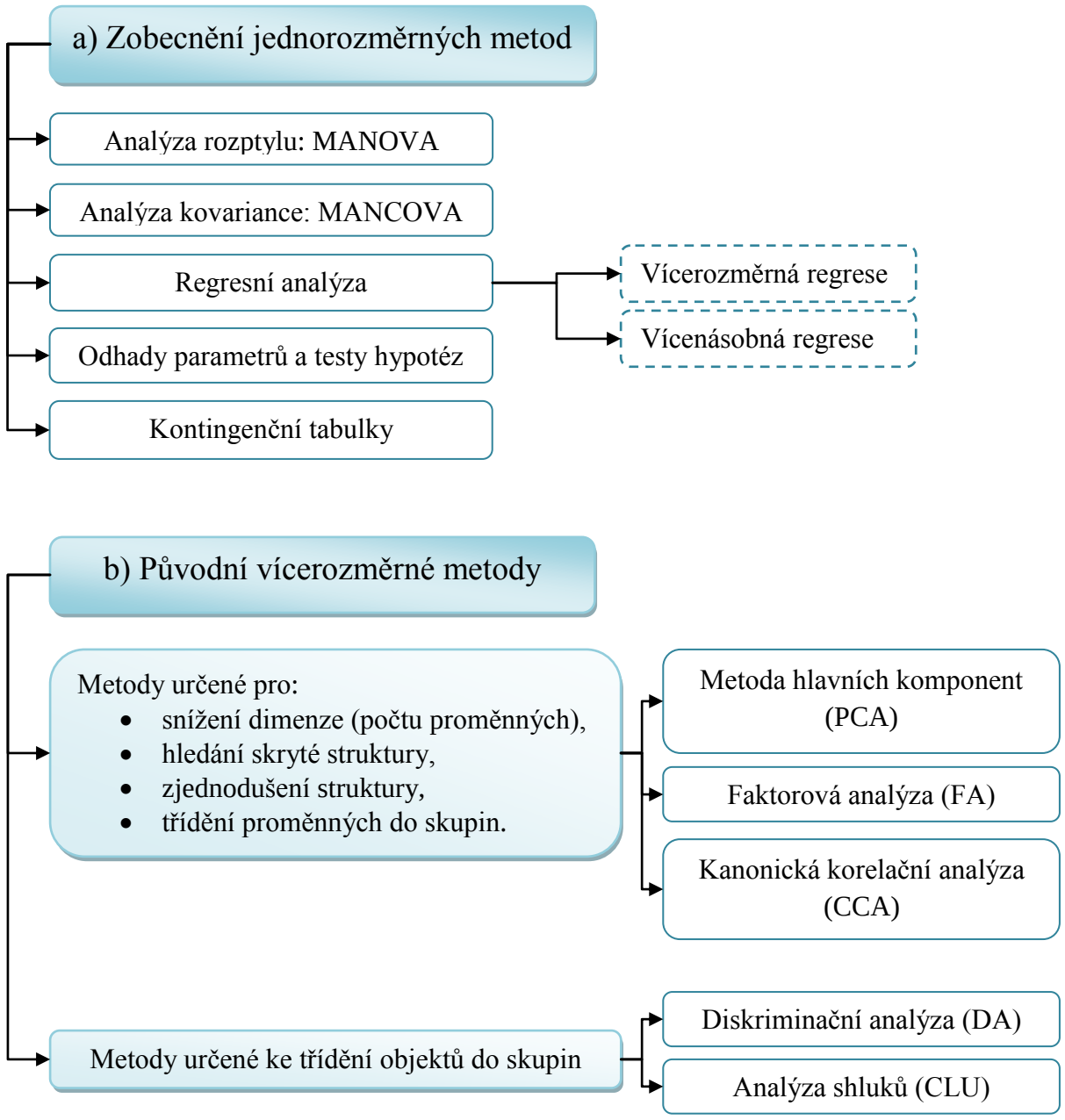
Počet pozorování	Počet proměnných	Časový horizont	Typ analýzy
n	1	1	jednorozměrná statistická analýza
1	p	1	kazuistika
1	1	T	jednorozměrná časová řada
n	1	T	jednorozměrné longitudinální sledování
n	p	1	mnohorozměrná statistická analýza
1	p	T	mnohorozměrná časová řada
n	p	T	mnohorozměrné longitudinální sledování

Povšimněme si, že mnohorozměrná statistická analýza, které se v této práci budeme věnovat, se zabývá situací, kdy na n objektech sledujeme p vlastností.

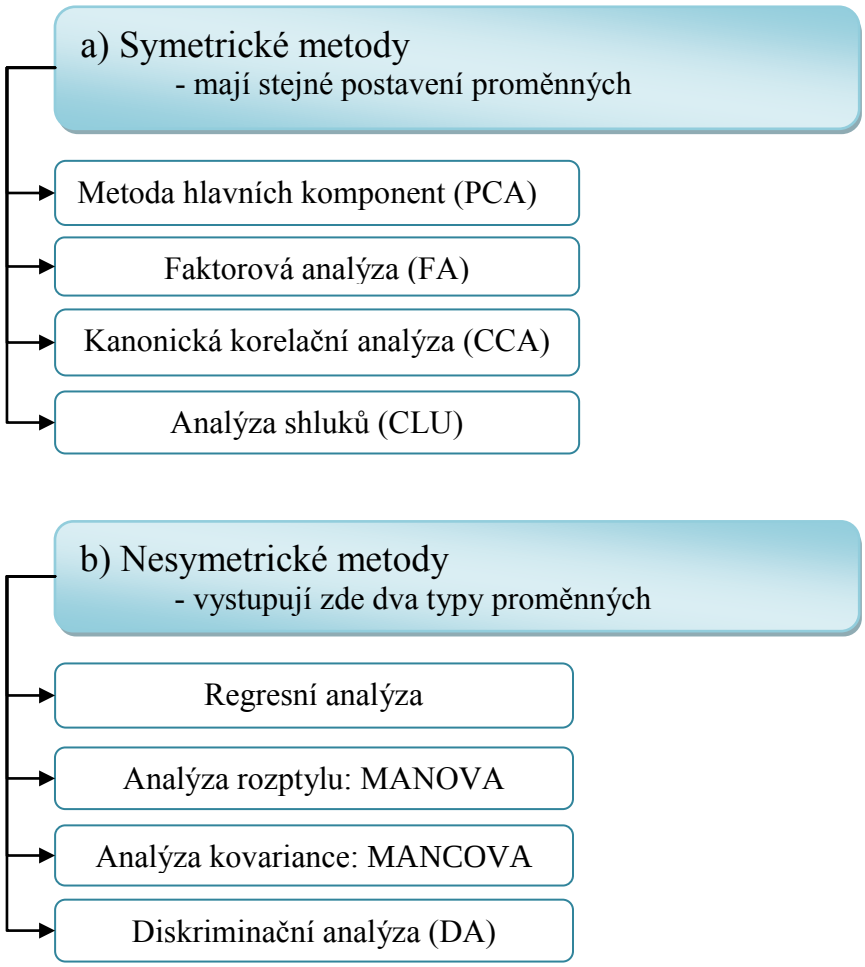
1.1.2 Přehled mnohorozměrných metod

Mnohorozměrná analýza zahrnuje velké množství metod. V této části poskytneme stručný přehled často používaných mnohorozměrných statistických metod. Existuje více možností jejich dělení. Na obrázcích 1.1 a 1.2 můžeme vidět dvojí způsob, jak metody rozlišit.

Obrázek 1.1: Schéma dělení mnohorozměrných metod podle jejich vzniku



Obrázek 1.2: Schéma dělení mnohorozměrných metod podle postavení proměnných



Dále se v práci budu zabývat pouze metodou hlavních komponent. Pro podrobný popis zmíněných metod mohu čtenáře alespoň nasměrovat na publikace, jakými jsou např. *Vícerozměrné statistické metody* P. Hebáka a kol. [2] a *Počítačová analýza vícerozměrných dat v příkladech* M. Melouna a kol. [3].

1.2 Vlastní čísla a vlastní vektory matice

Další pojem, se kterým se v textu setkáme, je tzv. vlastní číslo (charakteristické číslo, eigenvalue). Informace jsou čerpány z [7].

Nechť $\mathbf{A}= (a_{ij})$ je matice řádu n. Jestliže pro určitý parametr λ existuje vektor $\boldsymbol{\omega} \in \mathbf{R}^n$, $\boldsymbol{\omega} \neq 0$ takový, že

$$\mathbf{A}\,\boldsymbol{\omega} = \lambda\,\boldsymbol{\omega}, \tag{1.2.1}$$

potom λ se nazývá **vlastní (charakteristické) číslo** matice **A** a vektor $\boldsymbol{\omega}$ se nazývá **vlastní vektor** matice **A**. Říkáme pak, že vlastní vektor $\boldsymbol{\omega}$ přísluší k vlastnímu číslu λ.

Rovnici (1.2.1) můžeme zapsat v maticovém tvaru

$$(\mathbf{A} - \lambda\,\mathbf{I})\,\boldsymbol{\omega} = 0 \tag{1.2.2}$$

kde **I** je jednotková matice řádu n, která má prvky na hlavní diagonále rovny jedné a nediagonální prvky nulové, rovnici můžeme rovněž psát ve tvaru

$$\begin{pmatrix} a_{11} - \lambda & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} - \lambda \end{pmatrix} \cdot \begin{pmatrix} \omega_1 \\ \vdots \\ \omega_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

Rovnice (1.2.2) představuje soustavu homogenních lineárních rovnic o n neznámých, tj.

$$\begin{aligned} (a_{11} - \lambda)\,\omega_1 + a_{12}\,\omega_2 + \ldots + a_{1n}\,\omega_n &= 0 \\ a_{21}\,\omega_1 + (a_{22} - \lambda)\,\omega_2 + \ldots + a_{2n}\,\omega_n &= 0 \\ &\vdots \\ a_{n1}\,\omega_1 + a_{n2}\,\omega_2 + \ldots + (a_{nn} - \lambda)\,\omega_n &= 0 \end{aligned}$$

která má netriviální (nenulové) řešení právě tehdy, když je matice $(\mathbf{A} - \lambda\,\mathbf{I})$ singulární, tzn. $\det(\mathbf{A} - \lambda\,\mathbf{I}) = 0$. Polynom $\det(\mathbf{A} - \lambda\,\mathbf{I})$ se nazývá charakteristický polynom matice **A** a jeho kořeny jsou vlastní čísla matice **A**.

S vlastními čísly souvisí také pojem pozitivní definitnost. Čtvercová matice **A** se nazývá **pozitivně definitní**, jestliže pro každý vektor $\boldsymbol{x} \neq 0$ platí

$$\boldsymbol{x}^T\mathbf{A}\,\boldsymbol{x} > 0,$$

pokud platí neostrá nerovnost, tj. $\boldsymbol{x}^T\mathbf{A}\,\boldsymbol{x} \geq 0$, pak **A** je **pozitivně semidefinitní**. Pozitivně definitní matice má vždy všechna vlastní čísla kladná.

1.3 Základní číselné charakteristiky

V této části si uvedeme důležité populační a výběrové charakteristiky, které reprezentují mnohorozměrná data. Při zpracování této kapitoly bylo čerpáno zejména z [1], [5], [6], [8].

1.3.1 Číselné charakteristiky náhodného vektoru

Před uvedením základních populačních charakteristik neboli číselných charakteristik náhodného vektoru si přiblížíme pojem náhodný vektor, z něhož při výpočtu těchto charakteristik vycházíme.

- ***p* – rozměrný náhodný vektor**[1] – je uspořádaná *p*-tice náhodných veličin $X_1, \ldots, X_p$, odpovídající *p* vlastnostem a má následující tvar

$$\boldsymbol{x}_{(p \text{ x } 1)} = \left(X_1, \ldots, X_p\right)^{\mathrm{T}} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}. \tag{1.3.1}$$

- **Střední hodnota náhodného vektoru** – je vektor středních hodnot náhodných veličin X_j, $j = 1, \ldots, p$, tj. platí

$$\boldsymbol{\mu} = E\boldsymbol{x}_{(p \text{ x } 1)} = \left(EX_1, \ldots, EX_p\right)^{\mathrm{T}} = \begin{bmatrix} EX_1 \\ \vdots \\ EX_p \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}. \tag{1.3.2}$$

- **Kovarianční matice**[2] – jedná se o symetrickou, pozitivně semidefinitní matici typu *p* x *p*, jejíž prvky jsou příslušné kovariance dvou náhodných veličin $\mathrm{Cov}(X_i, X_j)$ a navíc platí

$$\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i), \quad i = 1, \ldots, p.$$

Kovarianční matice má tvar

$$\sum_{(p \text{ x } p)} = \mathrm{Var}\boldsymbol{x} = \begin{bmatrix} \mathrm{Var}X_1 & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}X_2 & \ldots & \mathrm{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p, X_2) & \ldots & \mathrm{Var}X_p \end{bmatrix},$$

[1] V některých publikacích bývá náhodný vektor označován velkým písmenem, např. [1]. Jelikož v této diplomové práci používáme velká písmena pro označení matice a náhodné veličiny, budeme náhodný vektor označovat malým písmenem.

[2] Některé publikace uvádí tuto matici pod názvem varianční matice.

Pro prvky kovarianční matice platí následující vztahy:

$$\mathrm{Cov}(X_i, X_k) = E(X_i - EX_i)(X_k - EX_k) = E(X_i - \mu_i)(X_k - \mu_k) = \sigma_{ik} \quad i,k = 1,\dots,p$$
$$\sigma_{ik} = \mathrm{Cov}(X_i, X_k) = \mathrm{Cov}(X_k, X_i) = \sigma_{ki} \quad i,k = 1,\dots,p$$
$$\mathrm{Var}X_i = \sigma_{ii} = \sigma_i^2 \quad i = 1,\dots,p$$

Kovarianční matici můžeme přepsat do tvaru

$$\sum_{(p\,\mathrm{x}\,p)} = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}. \qquad (1.3.3)$$

V souvislosti s výše uvedenou maticí se v textu setkáme s pojmy **hodnost matice** a **stopa matice**.

- Mějme matici **A** typu *n* x *p*. Hodností matice **A** (značíme rank(**A**)) rozumíme číslo, které udává maximální počet jejích lineárně nezávislých řádků, resp. sloupců, tj.

$$\mathrm{rank}(\mathbf{A}) = \min(n, p). \qquad (1.3.4)$$

- Nechť **A** je čtvercová matice typu *p* x *p*. Stopa matice **A** (značíme st(**A**)) vyjadřuje součet prvků na hlavní diagonále, tj.

$$\mathrm{st}(\mathbf{A}) = \sum_{i=1}^{p} a_{ii}. \qquad (1.3.5)$$

- **Populační korelační matice** – je symetrická matice, jejíž prvky odpovídají populačním korelačním koeficientům ρ_{ik}, které měří závislost náhodných veličin X_i, X_k v případě lineárního vztahu a mají následující tvar

$$\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}\sigma_{kk}}},$$

potom populační korelační matice má následující tvar:

$$\boldsymbol{\rho}_{(p\,\mathrm{x}\,p)} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}, \qquad (1.3.6)$$

kde

$$\rho_{ik} = \rho_{ki} \quad i,k = 1,\dots,p$$
$$\rho_{ii} = 1 \quad i = 1,\dots,p.$$

1.3.2 Číselné charakteristiky náhodného výběru

V této části si představíme základní výběrové charakteristiky. Nejdříve si však objasníme důležité pojmy, jakými jsou náhodný výběr a zdrojová matice dat, z nichž tyto charakteristiky počítáme.

- **Náhodný výběr** – jedná se o nezávislé, stejně rozdělené *p*-rozměrné náhodné vektory $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, které získáme nezávislým měřením *n* objektů.

- **Zdrojová matice dat** – je matice, ve které jsou zdrojová data uspořádána do rozměru *n* x *p*. Řádky tedy reprezentují *n* objektů a sloupce *p* znaků (proměnných), které se na objektech zkoumají. Zdrojová matice dat má tvar

$$\mathbf{X}_{(n\,\mathrm{x}\,p)} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{bmatrix}. \tag{1.3.7}$$

- **Výběrový průměr –** pokud máme náhodný výběr $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, potom výběrový průměr $\bar{\boldsymbol{x}}$ vypočítáme následujícím způsobem

$$\bar{\boldsymbol{x}}_{(p\,\mathrm{x}\,1)} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i, \tag{1.3.8}$$

tj. dostáváme náhodný vektor $\bar{\boldsymbol{x}} = \left(\bar{X}_1, \ldots, \bar{X}_p\right)^{\mathrm{T}}$.

- **Výběrová kovarianční matice** – matice S = (s_{jk}) typu *p* x *p*, jejíž prvky jsou výběrové kovariance mezi *j*-tou a *k*-tou proměnnou s_{jk}, pro které platí

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{ij} - \bar{X}_j\right)\left(X_{ik} - \bar{X}_k\right) \qquad j,k = 1, \ldots, p. \tag{1.3.9}$$

- **Výběrová korelační matice** – je symetrická matice, jejíž prvky odpovídají výběrovým korelačním koeficientům r_{jk}

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}} \qquad j,k = 1, \ldots, p. \tag{1.3.10}$$

Obdobně jako u populační korelační matice jsou diagonální prvky této matice vždy rovny jedné a pro nediagonální prvky platí $-1 \leq r_{jk} \leq 1$.

Výběrová korelační matice má tedy tvar

$$\mathbf{R}_{(p\,\mathrm{x}\,p)} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}. \tag{1.3.11}$$

2 Metoda hlavních komponent (PCA)

V následující kapitole se seznámíme s metodou hlavních komponent (Principal Component Analysis, PCA). Tuto metodu zavedl již v roce 1901 anglický matematik, filozof a zakladatel oboru matematické statistiky Karl Pearson jako popisnou statistickou metodu, která sloužila zejména k redukci mnohorozměrných dat. Na jeho práci navázal v roce 1933 statistik Harold Hotelling, a to tím způsobem, že zobecnil postup aplikací komponentní analýzy na náhodné vektory a navrhl použití metody hlavních komponent pro rozbor kovarianční struktury proměnných. Proto můžeme tuto metodu najít také pod názvem Hotellingova transformace (Hotelling Transform). V současné statistické literatuře je metoda hlavních komponent doporučována zejména jako význačný nástroj průzkumové analýzy dat pro ověřování předpokladů, dále jako samostatný nástroj rozboru struktury vztahů v množině vzájemně závislých pozorování a v neposlední řadě jako užitečný pomocník některých metod mnohorozměrné statistické analýzy. Metoda hlavních komponent pomáhá jako jedno z možných prvních řešení například diskriminační analýze v případě malého počtu pozorování a velkého počtu proměnných, dále umožňuje regresní analýze odstranit problémy s multikolinearitou a přebytečným počtem vzájemně závislých vysvětlujících proměnných, pomáhá také shlukové analýze při klasifikaci objektů do homogenních skupin na základě velkého počtu proměnných, ale i faktorové analýze a dalším mnohorozměrným metodám [2], [9]. Hlavním zdrojem pro tvorbu této kapitoly byly články a knihy [1], [2], [3], [5], [10], [11].

2.1 Cíle metody PCA

Základní cíl metody hlavních komponent spočívá zejména ve zjednodušení popisu skupiny vzájemně lineárně závislých, tedy korelovaných znaků. Metoda se snaží snížit dimenzi úlohy neboli redukovat počet znaků bez velké ztráty informace, což je výhodné především pro zobrazení mnohorozměrných dat. Jednotlivé měřené veličiny poměrně často vykazují silnou korelaci. Pro zjednodušení analýzy a snadnější hodnocení výsledků je vhodné zkoumat, zda je možné celou skupinu proměnných (tedy studované vlastnosti pozorovaných objektů) nahradit jedinou veličinou nebo menším počtem veličin, které budou nést o datech téměř stejnou informaci, jako nesly veličiny původní. Tento problém lze popsat jako metodu lineární transformace původních znaků na nové, vzájemně nekorelované proměnné, které mají vhodnější vlastnosti a je jich výrazně méně.

Metoda hlavních komponent se tedy snaží nalézt tyto skryté (umělé, latentní, neměřitelné) veličiny, nazvané jako hlavní komponenty. Nově vytvořené proměnné jsou lineární kombinací původních proměnných a požaduje se od nich, aby co nejlépe reprezentovaly původní proměnné, jinak řečeno, aby co nejlépe vysvětlovaly variabilitu původních proměnných.

Základní charakteristikou každé hlavní komponenty je její míra variability čili rozptyl. Hlavní komponenty jsou seřazeny podle důležitosti, tj. podle klesajícího rozptylu, od největšího k nejmenšímu. První hlavní komponenta obsahuje nejvíce informace o variabilitě původních dat, druhá hlavní komponenta zase největší část rozptylu původních dat neobsaženého v první komponentě. Nejméně informace je obsaženo v poslední komponentě. Pokud má nějaký původní znak malý či dokonce žádný rozptyl, potom není schopen přispívat k rozlišení mezi objekty. Metoda hlavních komponent umožňuje namísto vyšetřování velkého počtu původních znaků s komplexními vnitřními vazbami analyzovat pouze malý počet nekorelovaných hlavních komponent.

V praxi bývá metoda hlavních komponent využívána například pro efektivní rozpoznávání obrázků lidské tváře v tzv. facespace (prostoru tváří, kolekci obrázků tváří). V tomto případě metoda hlavních komponent redukuje původní prostor obrázků a poskytuje velmi rozumnou extrakci rysů. Praktickou úlohou je identifikace osob podle zvoleného biometrického rysu, jako je například oční duhovka či rysy tváře [13], [14]. Další praktické využití metody hlavních komponent si názorně ukážeme na konkrétních příkladech v kapitole 3.

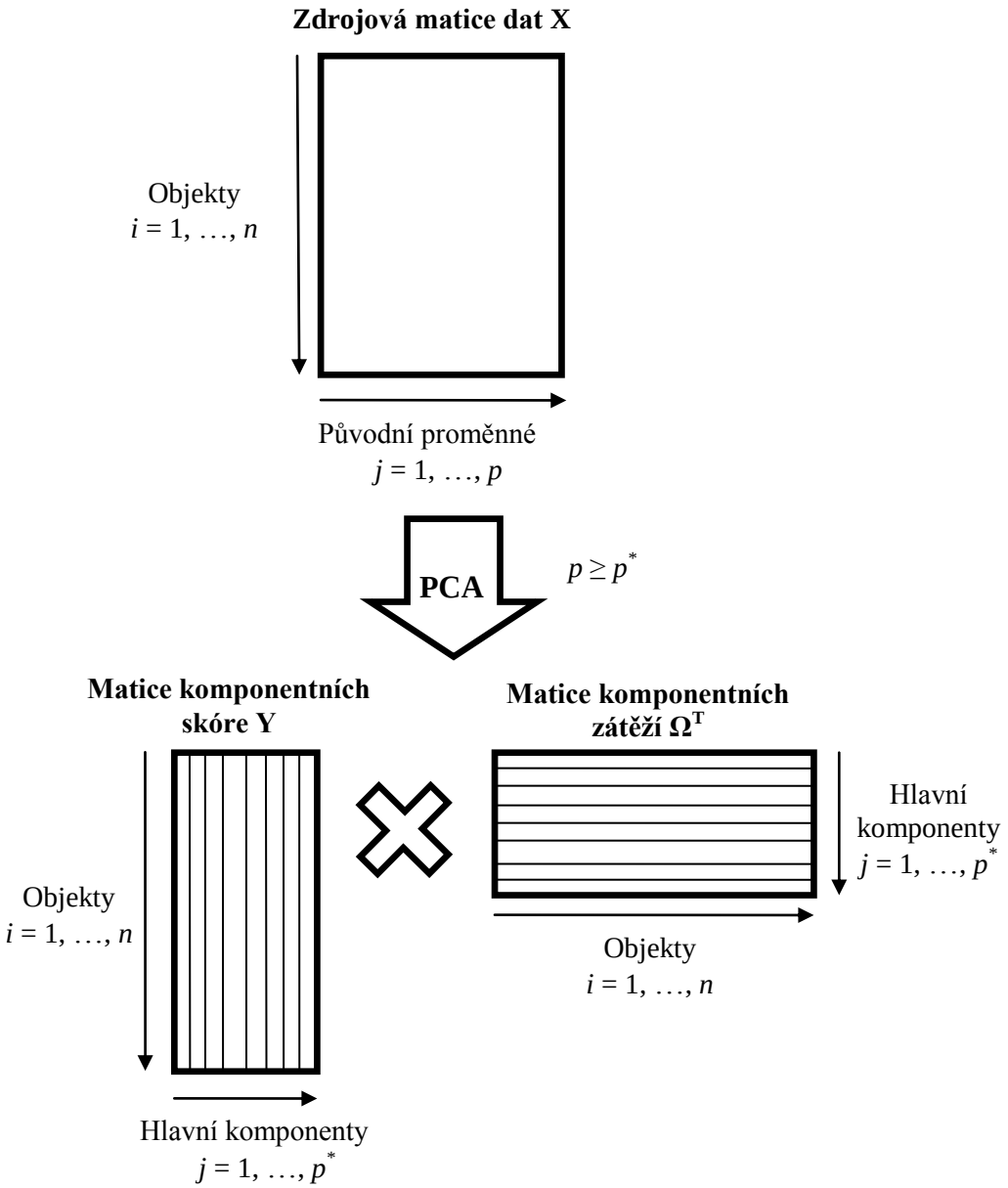
2.2 Podstata PCA

U metody hlavních komponent je vstupem zdrojová matice dat **X** typu n x p viz (1.3.7). Jak již bylo zmíněno v kapitole 1.3, tato matice obsahuje hodnoty n objektů (pozorování, měření) pro p původních proměnných (vlastností, zkoumaných znaků, měřených veličin). Důležitá je skutečnost, že pro každý objekt ze skupiny n objektů je známý zkoumaný znak. Výstupem je aproximace zdrojové matice dat **X**, obsahující hodnoty n měření pro p^* hlavních komponent. Informace k této kapitole jsou čerpány zejména z [2], [3], [12].

Na obrázku 2.1 je schematicky znázorněna situace, kde se zdrojová matice dat **X** rozkládá na tzv. matici komponentních skóre typu n x k a matici komponentních

zátěží typu k x p^*. Při změně souřadnic z objektů v původních znacích na objekty v hlavních komponentách dochází k rozdílu, tedy ke ztrátě informace projekcí do menšího počtu rozměrů. Tento rozdíl nazýváme mírou těsností proložení modelu PCA nebo také chybou modelu PCA. Jednou z nejdůležitějších součástí metody hlavních komponent je její interpretace, tj. vysvětlení užitých hlavních komponent, jejich pojmenování a vysvětlení vztahu původních veličin k hlavním komponentám. Aproximace zdrojové matice dat **X** má řadu výhod v interpretaci dat. Jedná se nejen o změnu souřadnicového systému, ale také především o nalezení tzv. šumu (ostatních zbývajících hlavních komponent) a jeho vypuštění.

Obrázek 2.1: Schéma maticových výpočtů v PCA (upraveno dle [3])



Prvky nově vzniklých matic nazýváme komponentní skóre a komponentní zátěže, jejichž význam si vysvětlíme v kapitole 2.3.1. Vzhledem k tomu, že se k rekonstrukci obecně používá k z p^* hlavních komponent, projeví se ztráta informace vznikem chybové matice neboli matice reziduí **E** rozměru n x p^*. Platí tedy vztah

$$\mathbf{X} = \mathbf{Y}\boldsymbol{\Omega}^{\mathrm{T}} + \mathbf{E} = \text{struktura dat} + \text{šum}$$

Matice reziduí **E** není objasněna modelem hlavních komponent. Souvisí s tzv. těsností proložení modelu a ukazuje, jak dobře jsou objekty proloženy modelem hlavních komponent.

Doposud jsme pro počet hlavních komponent používali označení $j = 1, \ldots, p^*$, dále budeme pro zjednodušení používat označení $j = 1, \ldots, p$, stejně jako u původních proměnných.

2.3 Hlavní komponenty v populaci

Informace k této kapitole jsou čerpány zejména z [1], [2], [5], [10].

Z algebraického hlediska jsou hlavní komponenty konkrétní lineární kombinace p náhodných veličin $X_1, X_2, \ldots, X_p$. Z geometrického hlediska tyto lineární kombinace jsou osy nového souřadnicového systému, jehož dimenze je „rozumně“ malá. Nové osy představují směrnice s maximální variabilitou a poskytují jednodušší a šetrný popis kovarianční struktury. Metoda hlavních komponent tedy vytvoří nový souřadnicový systém rotací původního systému a převede do něj původní proměnné $X_1, X_2, \ldots, X_p$. Rotaci provede tím způsobem, aby obrazy případů v nové souřadné soustavě vyhovovaly určitému kritériu.

Mějme náhodný vektor $\boldsymbol{x} = (X_1, X_2, \ldots, X_p)^{\mathrm{T}}$, jehož prvky jsou náhodné veličiny mající vícerozměrné normální rozdělení, s kovarianční maticí $\sum$ (viz 1.3.3) o hodnosti p, tj. $\mathrm{rank}(\sum) = p$. Označme postupně klesající charakteristická čísla matice $\sum$ jako $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ a jejich odpovídající charakteristické vektory $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_p$.

Uvažujme lineární kombinace:

$$\begin{aligned} Y_1 &= \boldsymbol{\omega}_1^{\mathrm{T}}\boldsymbol{x} = \omega_{11}X_1 + \omega_{12}X_2 + \ldots + \omega_{1p}X_p \\ Y_2 &= \boldsymbol{\omega}_2^{\mathrm{T}}\boldsymbol{x} = \omega_{21}X_1 + \omega_{22}X_2 + \ldots + \omega_{2p}X_p \\ &\vdots \\ Y_p &= \boldsymbol{\omega}_p^{\mathrm{T}}\boldsymbol{x} = \omega_{p1}X_1 + \omega_{p2}X_2 + \ldots + \omega_{pp}X_p \end{aligned} \qquad (2.3.1)$$

Platí (viz str. 19):

$$\mathrm{Var}(Y_j) = \boldsymbol{\omega}_j^{\mathrm{T}} \sum \boldsymbol{\omega}_j = \lambda_j \qquad j = 1, \ldots, p \qquad (2.3.2)$$

$$\mathrm{Cov}(Y_j, Y_k) = \boldsymbol{\omega}_j^{\mathrm{T}} \sum \boldsymbol{\omega}_k = 0 \qquad j, k = 1, \ldots, p \qquad (2.3.3)$$

Uvedené nekorelované lineární kombinace Y_1, Y_2, …, Y_p představují hlavní komponenty a jsou seřazeny podle důležitosti, tzn. podle klesajícího rozptylu od největšího k nejmenšímu.

První hlavní komponenta je lineární kombinace Y_1, kde vektor $\boldsymbol{\omega}_1$ je určen maximalizací rozptylu komponenty Y_1, tj. $\mathrm{Var}(\boldsymbol{\omega}_1^{\mathrm{T}}\boldsymbol{x})$, přes všechny vektory $\boldsymbol{\omega}_1$ tak, aby byla splněna normalizační podmínka $\boldsymbol{\omega}_1^{\mathrm{T}}\boldsymbol{\omega}_1 = 1$. V případě, že $\boldsymbol{\omega}_1$ je charakteristický vektor odpovídající λ_1 a při splnění normalizační podmínky je maximální hodnota rozptylu Y_1 rovna největšímu charakteristickému číslu λ_1 kovarianční matice $\sum$. První hlavní komponenta, s největším rozptylem, tedy obsahuje nejvíce informace o variabilitě původních proměnných.

Druhá hlavní komponenta je lineární kombinace Y_2, která maximalizuje rozptyl $\mathrm{Var}(\boldsymbol{\omega}_2^{\mathrm{T}}\boldsymbol{x})$ splňující podmínku $\boldsymbol{\omega}_2^{\mathrm{T}}\boldsymbol{\omega}_2 = 1$ a navíc musí být splněn požadavek nekorelovanosti veličin Y_1, Y_2, tj. $\mathrm{Cov}(Y_1, Y_2) = 0$, což zajišťuje kolmost obou hlavních komponent. Druhá hlavní komponenta popisuje největší část rozptylu neobsaženého v první komponentě.

j-tá hlavní komponenta je analogicky lineární kombinace Y_j, která maximalizuje rozptyl $\mathrm{Var}(\boldsymbol{\omega}_j^{\mathrm{T}}\boldsymbol{x})$ a splňuje omezující podmínky $\boldsymbol{\omega}_j^{\mathrm{T}}\boldsymbol{\omega}_j = 1$ a $\mathrm{Cov}(Y_j, Y_k) = 0$ pro $k < j$. Nejméně informace je tedy soustředěno v poslední komponentě.

Kovarianční matici $\sum$ můžeme napsat pomocí spektrálního rozkladu jako

$$\sum = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{\mathrm{T}} = \sum_{j=1}^{p} \lambda_j \boldsymbol{\omega}_j \boldsymbol{\omega}_j^{\mathrm{T}}, \qquad \mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{P}\mathbf{P}^{\mathrm{T}} = \mathbf{I}_{(p \times p)}, \qquad (2.3.4)$$

kde

$$\boldsymbol{\Lambda}_{(p \times p)} = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_p \end{bmatrix} \qquad (2.3.5)$$

je diagonální matice řádu p, která má na hlavní diagonále všechna sestupně uspořádaná charakteristická čísla matice $\sum$ a $\mathbf{P}_{(p \times p)} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_p)$ je ortogonální matice, jejíž prvky

jsou ortonormální vlastní vektory $\sum$, tj. k-tý sloupec matice $\mathbf{P}$ je vlastní vektor $\boldsymbol{\omega}_k$ matice $\sum$, který přísluší vlastnímu číslu λ_k a pro který platí normalizační podmínka $\boldsymbol{\omega}_k^{\mathrm{T}}\boldsymbol{\omega}_k = 1$.

Ze vztahu (1.3.5) plyne

$$\mathrm{st}(\sum) = \mathrm{st}(\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{\mathrm{T}}) = \mathrm{st}(\boldsymbol{\Lambda}\mathbf{P}^{\mathrm{T}}\mathbf{P}) = \mathrm{st}(\boldsymbol{\Lambda}) = \lambda_1 + \ldots + \lambda_p.$$

Pomocí následujících třech vztahů ([1], str. 432), které platí pro kovarianční matici $\sum$, si dokážeme platnost tvrzení (2.3.2) a (2.3.3):

1) Pro libovolný nenulový vektor $\boldsymbol{a}$ typu p x 1 platí

$$\max_{\boldsymbol{a}\neq 0} \frac{\boldsymbol{a}^{\mathrm{T}}\sum\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}} = \lambda_1. \qquad (2.3.6)$$

Důkaz:

Nechť $\mathbf{P}_{(p\ \mathrm{x}\ p)}$ je již zmíněná ortogonální matice a $\boldsymbol{\Lambda}_{(p\ \mathrm{x}\ p)}$ je diagonální matice s vlastními čísly $\lambda_1, \ldots, \lambda_p$ na hlavní diagonále. Dále nechť $\sum^{1/2} = \mathbf{P}\boldsymbol{\Lambda}^{1/2}\mathbf{P}^{\mathrm{T}}$ a $\boldsymbol{c}_{(p\ \mathrm{x}\ 1)} = \mathbf{P}^{\mathrm{T}}\boldsymbol{a}$. Tudíž je-li $\boldsymbol{a} \neq 0$, pak také $\boldsymbol{c} \neq 0$ a platí

$$\frac{\boldsymbol{a}^{\mathrm{T}}\sum\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}} = \frac{\boldsymbol{a}^{\mathrm{T}}\sum^{\frac{1}{2}}\sum^{\frac{1}{2}}\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\mathbf{P}\mathbf{P}^{\mathrm{T}}\boldsymbol{a}} = \frac{\boldsymbol{a}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{P}^{\mathrm{T}}\mathbf{P}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{P}^{\mathrm{T}}\boldsymbol{a}}{\boldsymbol{c}^{\mathrm{T}}\boldsymbol{c}} = \frac{\boldsymbol{c}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{c}}{\boldsymbol{c}^{\mathrm{T}}\boldsymbol{c}} = \frac{\sum_{j=1}^{p}\lambda_j\boldsymbol{c}_j\,\boldsymbol{c}_j^{\mathrm{T}}}{\sum_{j=1}^{p}\boldsymbol{c}_j\,\boldsymbol{c}_j^{\mathrm{T}}} =$$

$$= \frac{\sum_{j=1}^{p}\lambda_j\,\boldsymbol{c}_j^2}{\sum_{j=1}^{p}\boldsymbol{c}_j^2} \leq \lambda_1\frac{\sum_{j=1}^{p}\boldsymbol{c}_j^2}{\sum_{j=1}^{p}\boldsymbol{c}_j^2} = \lambda_1.$$

Potom pro volbu $\boldsymbol{a} = \boldsymbol{\omega}_1$ platí

$$\boldsymbol{c} = \mathbf{P}^{\mathrm{T}}\boldsymbol{a} = \mathbf{P}^{\mathrm{T}}\boldsymbol{\omega}_1 = \begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix},$$

odtud

$$\boldsymbol{\omega}_k^{\mathrm{T}}\boldsymbol{\omega}_1 = \begin{cases}1 & k = 1\\ 0 & k \neq 1\end{cases}$$

a také pro tuto volbu platí

$$\frac{\boldsymbol{c}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{c}}{\boldsymbol{c}^{\mathrm{T}}\boldsymbol{c}} = \frac{\lambda_1}{1} = \lambda_1 \quad \text{nebo} \quad \frac{\boldsymbol{a}^{\mathrm{T}}\sum\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}} = \frac{\boldsymbol{\omega}_1^{\mathrm{T}}\sum\boldsymbol{\omega}_1}{\boldsymbol{\omega}_1^{\mathrm{T}}\boldsymbol{\omega}_1} = \boldsymbol{\omega}_1^{\mathrm{T}}\sum\boldsymbol{\omega}_1 = \lambda_1$$

Rovnost (2.3.6) je splněna.

Nyní uvažujme, pokud je splněna normalizační podmínka $\boldsymbol{\omega}_1{}^{\mathrm{T}}\boldsymbol{\omega}_1 = 1$, pak pro volbu $\boldsymbol{a} = \boldsymbol{\omega}_1$ platí

$$\max_{\boldsymbol{a}\neq 0} \frac{\boldsymbol{a}^{\mathrm{T}}\sum\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}} = \lambda_1 = \frac{\boldsymbol{\omega}_1^{\mathrm{T}}\sum\boldsymbol{\omega}_1}{\boldsymbol{\omega}_1^{\mathrm{T}}\boldsymbol{\omega}_1} = \mathrm{Var}(Y_1)$$

protože

$$\mathrm{Var}(Y_1) = \mathrm{Var}(\boldsymbol{\omega}_1^{\mathrm{T}}\boldsymbol{x}) = \boldsymbol{\omega}_1^{\mathrm{T}}\,\mathrm{Var}\boldsymbol{x}\;\boldsymbol{\omega}_1 = \boldsymbol{\omega}_1^{\mathrm{T}}\sum\boldsymbol{\omega}_1 = \frac{\boldsymbol{\omega}_1^{\mathrm{T}}\sum\boldsymbol{\omega}_1}{\boldsymbol{\omega}_1^{\mathrm{T}}\boldsymbol{\omega}_1} = \frac{\boldsymbol{\omega}_1^{\mathrm{T}}\lambda_1\boldsymbol{\omega}_1}{\boldsymbol{\omega}_1^{\mathrm{T}}\boldsymbol{\omega}_1} = \frac{\lambda_1\boldsymbol{\omega}_1^{\mathrm{T}}\boldsymbol{\omega}_1}{\boldsymbol{\omega}_1^{\mathrm{T}}\boldsymbol{\omega}_1} = \lambda_1.$$

Je tedy zřejmé, že rozptyl první hlavní komponenty je roven vlastnímu číslu λ_1.

2) Pro libovolný nenulový vektor $\boldsymbol{a}$ platí

$$\min_{\boldsymbol{a}\neq 0} \frac{\boldsymbol{a}^{\mathrm{T}}\sum\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}} = \lambda_p.$$

Platnost této rovnosti, bychom dokázali analogicky jako v předchozím případě. Podobně, při splnění normalizační podmínky $\boldsymbol{\omega}_p{}^{\mathrm{T}}\boldsymbol{\omega}_p = 1$, je – li $a = \boldsymbol{\omega}_p$, platí

$$\min_{\boldsymbol{a}\neq 0} \frac{\boldsymbol{a}^{\mathrm{T}}\sum\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}} = \lambda_p = \frac{\boldsymbol{\omega}_p^{\mathrm{T}}\sum\boldsymbol{\omega}_p}{\boldsymbol{\omega}_p^{\mathrm{T}}\boldsymbol{\omega}_p} = \mathrm{Var}(Y_p).$$

3) Pro libovolný nenulový vektor $\boldsymbol{a}$ platí

$$\max_{\boldsymbol{a}\perp\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_k} \frac{\boldsymbol{a}^{\mathrm{T}}\sum\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}} = \lambda_{k+1} \qquad k = 1, 2, \ldots, p-1,$$

kde symbol $\perp$ označuje kolmost. Analogicky pro volbu $\boldsymbol{a} = \boldsymbol{\omega}_{k+1}$, je-li splněna podmínka $\boldsymbol{\omega}_{k+1}^{\mathrm{T}}\boldsymbol{\omega}_i = 0$, $i = 1, \ldots, k$ a $k = 1, \ldots, p-1$, potom

$$\frac{\boldsymbol{\omega}_{k+1}^{\mathrm{T}}\sum\boldsymbol{\omega}_{k+1}}{\boldsymbol{\omega}_{k+1}^{\mathrm{T}}\boldsymbol{\omega}_{k+1}} = \boldsymbol{\omega}_{k+1}^{\mathrm{T}}\sum\boldsymbol{\omega}_{k+1} = \mathrm{Var}(Y_{k+1}),$$

kde $\boldsymbol{\omega}_{k+1}^{\mathrm{T}}\sum\boldsymbol{\omega}_{k+1} = \lambda_{k+1}\boldsymbol{\omega}_{k+1}^{\mathrm{T}}\boldsymbol{\omega}_{k+1} = \lambda_{k+1}$, takže $\mathrm{Var}(Y_{k+1}) = \lambda_{k+1}$.

Zbývá dokázat, že vektory $\boldsymbol{\omega}_i$ a $\boldsymbol{\omega}_k$, které jsou na sebe kolmé (tzn. $\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{\omega}_k = 0, i \neq k$), splňují požadavek nekorelovanosti veličin, tj. $\mathrm{Cov}(Y_i, Y_k) = 0$. Jelikož vlastní čísla a vlastní vektory matice $\sum$ splňují vztah $\sum\boldsymbol{\omega}_k = \lambda_k\boldsymbol{\omega}_k$ podle (1.2.1), potom pro libovolné dva vektory $\boldsymbol{\omega}_i$ a $\boldsymbol{\omega}_k$ platí

$$\mathrm{Cov}(Y_i, Y_k) = \mathrm{Cov}\left(\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{x}, \boldsymbol{\omega}_k^{\mathrm{T}}\boldsymbol{x}\right) = \boldsymbol{\omega}_i^{\mathrm{T}}\,\mathrm{Var}\boldsymbol{x}\;\boldsymbol{\omega}_k = \boldsymbol{\omega}_i^{\mathrm{T}}\sum\boldsymbol{\omega}_k = \boldsymbol{\omega}_i^{\mathrm{T}}\lambda_k\boldsymbol{\omega}_k =$$
$$= \lambda_k\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{\omega}_k = 0 \qquad \text{pro } i \neq k.$$

Platnost tvrzení (2.3.2) a (2.3.3) vyjadřující, že hlavní komponenty jsou nekorelované veličiny s rozptyly rovnými vlastním číslům, je tímto ověřena.

Jak již bylo řečeno na začátku této kapitoly, z geometrického hlediska představují hlavní komponenty osy nového souřadnicového systému. Vycházíme z populace, ve které mají náhodné veličiny X_1, X_2, …, X_p vícerozměrné normální rozdělení, tj. předpokládáme, že náhodný vektor $\boldsymbol{x} = (X_1, X_2, \ldots, X_p)^{\mathrm{T}}$ má p-rozměrné normální rozdělení $N_p(\boldsymbol{\mu}, \sum)$. Oblast, ve které je hustota náhodného vektoru konstantní má tvar elipsoidu[3], tj. množiny všech vektorů $\boldsymbol{x}$ takových, pro které platí rovnost

$$(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \sum^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = c^2,$$

kde $\sum^{-1}$ je inverzní matice[4] ke kovarianční matici $\sum$. Tyto elipsoidy jsou centrované v $\boldsymbol{\mu}$ s osami $\pm c\sqrt{\lambda_i}\boldsymbol{\omega}_i$, kde λ_i jsou vlastní čísla matice $\sum$ a $\boldsymbol{\omega}_i$ jsou jim příslušející vlastní vektory. Bod ležící na i-té ose elipsoidu bude mít souřadnice úměrné vlastnímu vektoru $\boldsymbol{\omega}_i = (\omega_{i1}, \omega_{i2}, \ldots, \omega_{ip})^{\mathrm{T}}$ v souřadnicovém systému, který má počátek v $\boldsymbol{\mu}$ a osy rovnoběžné s původními osami X_1, X_2, …, X_p. Položíme-li $\boldsymbol{\mu} = 0$, posune se souřadnicový systém do počátku v nule, můžeme psát

$$c^2 = \boldsymbol{x}^{\mathrm{T}} \sum^{-1} \boldsymbol{x}.$$

Platí-li pro matici $\sum$ vztah (2.3.4), tak pro inverzní matici $\sum^{-1}$ platí

$$\sum^{-1} = \mathbf{P}\boldsymbol{\Lambda}^{-1}\mathbf{P}^{\mathrm{T}} = \sum_{j=1}^{p} \frac{1}{\lambda_j} \boldsymbol{\omega}_j \, \boldsymbol{\omega}_j^{\mathrm{T}}.$$

Můžeme tedy psát

$$c^2 = \boldsymbol{x}^{\mathrm{T}} \sum^{-1} \boldsymbol{x} = \frac{1}{\lambda_1} (\boldsymbol{\omega}_1^{\mathrm{T}} \boldsymbol{x})^2 + \frac{1}{\lambda_2} (\boldsymbol{\omega}_2^{\mathrm{T}} \boldsymbol{x})^2 + \cdots + \frac{1}{\lambda_p} \left(\boldsymbol{\omega}_p^{\mathrm{T}} \boldsymbol{x}\right)^2,$$

kde $\boldsymbol{\omega}_1{}^{\mathrm{T}}\boldsymbol{x}$, …, $\boldsymbol{\omega}_p{}^{\mathrm{T}}\boldsymbol{x}$ jsou hlavní komponenty. S využitím (2.3.1) platí

$$c^2 = \frac{1}{\lambda_1} Y_1{}^2 + \cdots + \frac{1}{\lambda_p} Y_p{}^2.$$

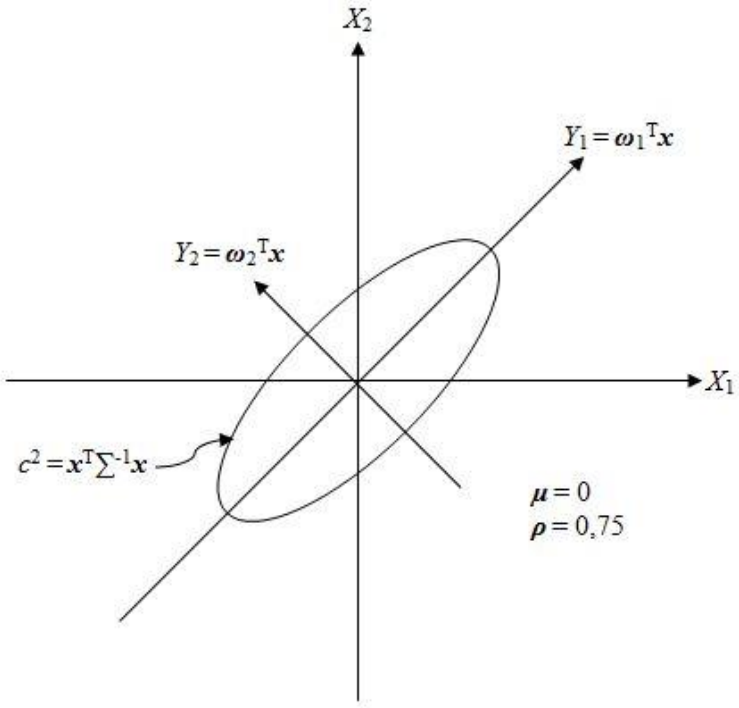
Tato rovnice definuje elipsoidy v souřadnicovém systému s osami Y_1, Y_2, …, Y_p ležícími ve směru příslušných vektorů $\boldsymbol{\omega}_1$, …, $\boldsymbol{\omega}_p$. Pokud je λ_1 největší vlastní číslo kovarianční matice $\sum$, hlavní osa leží ve směru vektoru $\boldsymbol{\omega}_1$. Zbývající osy menšího měřítka se nacházejí

[3] Ve dvourozměrném případě mluvíme o elipsách.
[4] Inverzní matice k dané matici je matice, která po vynásobení s maticí původní dá jednotkovou matici.

ve směru vlastních vektorů $\boldsymbol{\omega}_2$, …, $\boldsymbol{\omega}_p$. Hlavní komponenty Y_1, Y_2, …, Y_p tedy tvoří osy elipsoidu konstantní hustoty, tj. osy nového souřadnicového systému.

Elipsa konstantní hustoty pro dvourozměrný náhodný vektor $\boldsymbol{x} = (X_1, X_2)^T$, která má počátek v $\boldsymbol{\mu} = 0$ je znázorněna na obrázku 2.2. Můžeme vidět, že hlavní komponenty Y_1 a Y_2 jsme získali rotací původního souřadnicového systému os. Korelační koeficient $\boldsymbol{\rho}$ je zde roven 0,75 a udává míru zploštění elipsoidu (např. $\boldsymbol{\rho} = 0$ odpovídá kružnici, naopak $\boldsymbol{\rho} = 1$ přímce).

Obrázek 2.2: Elipsa konstantní hustoty (upraveno dle [1])



2.3.1 Počet hlavních komponent

Jak již bylo řečeno v kapitole 2.1, cílem metody hlavních komponent je „vměstnat“ co nejvíce informace do několika málo nových proměnných. Obecně lze dosáhnout stavu, ve kterém úplný systém všech p hlavních komponent Y_1, Y_2, …, Y_p dokonale (beze zbytku) vysvětlí celkový rozptyl původních proměnných. Jelikož hledáme nižší rozměr dat, uvítáme, když celkově bude hlavních komponent výrazně méně než p. První o co se tedy budeme ve výsledcích zajímat je, nakolik je taková „komprese“ možná a kolik budeme potřebovat výsledných komponent. Hlavním zdrojem informací o této problematice jsou publikace [1], [2], [10], [12].

Pro hodnocení a uspořádání objektů by bylo zřejmě nejvýhodnější mít pouze jednu hlavní komponentu, což lze velmi vzácně očekávat. Z hlediska grafického zobrazení dat by bylo vhodné mít nejvýše tři hlavní komponenty, ale samozřejmě záleží na počtu původních veličin. Jak je uvedeno v publikaci P. Hebáka a kol. [2], zkušenosti s používáním metody hlavních komponent ukazují, že se velmi často vyskytuje případ tří až čtyř hlavních komponent. Více než pět až šest hlavních komponent nelze považovat za příliš úspěšné řešení a nebývá ani potřebné.

Pro p komponent platí:

$$\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_p^2 = \sum_{i=1}^{p} \mathrm{Var}(X_i) = \lambda_1 + \ldots + \lambda_p = \sum_{i=1}^{p} \mathrm{Var}(Y_i), \qquad (2.3.7)$$

to znamená, že celkový rozptyl populace (tj. základního souboru či původních náhodných veličin $X_1, X_2, \ldots, X_p$) odpovídá celkovému rozptylu hlavních komponent.

Ze vztahu (1.3.5) plyne

$$\mathrm{st}(\sum) = \sum_{i=1}^{p} \mathrm{Var}(X_i) = \mathrm{st}(\boldsymbol{\Lambda}) = \sum_{i=1}^{p} \mathrm{Var}(Y_i).$$

Mírou významu (také označovanou jako podíl variability) k-té hlavní komponenty vzhledem k celkovému rozptylu vysvětlovaných veličin $X_1, X_2, \ldots, X_p$ je podíl

$$\frac{\lambda_k}{\lambda_1+\lambda_2+\ldots+\lambda_p} = \frac{\lambda_k}{\mathrm{st}(\sum)} = \frac{\lambda_k}{\mathrm{st}(\boldsymbol{\Lambda})} \quad k = 1, 2, \ldots, p\,. \qquad (2.3.8)$$

Jestliže součet prvních (nejvyšších) podílů variability je dostatečně blízký jedné, tj. vyjádřeno v procentech 100% (obvykle však stačí 80% až 90%), potom můžeme bez velké ztráty informace nahradit původních p proměnných těmito prvními k hlavními komponentami.

Vztahy mezi původními veličinami a hlavními komponentami vyjadřují následující korelační koeficienty [1]. Uvažujme hlavní komponenty

$$Y_1 = \boldsymbol{\omega}_1^{\mathrm{T}}\boldsymbol{x},\ Y_2 = \boldsymbol{\omega}_2^{\mathrm{T}}\boldsymbol{x},\ \ldots,\ Y_p = \boldsymbol{\omega}_p^{\mathrm{T}}\boldsymbol{x}$$

s kovarianční maticí $\sum$, potom

$$\rho_{Y_i,X_k} = \frac{\mathrm{Cov}(Y_i,X_k)}{\sqrt{\mathrm{Var}(Y_i)}\sqrt{\mathrm{Var}(X_k)}} = \frac{\lambda_i\boldsymbol{\omega}_{ik}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{\boldsymbol{\omega}_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad i, k = 1, 2, \ldots, p \qquad (2.3.9)$$

jsou korelační koeficienty mezi komponentami Y_i a veličinami X_k, kde $\mathrm{Cov}(Y_i, X_k)$ je kovariance mezi i-tou hlavní komponentou a k-tou původní veličinou.

V kapitole 2.2 jsme si ukázali rozklad zdrojové matice dat **X** na matici komponentních skóre **Y** a matici komponentních zátěží $\mathbf{\Omega}^{\mathrm{T}}$. Prvky těchto nově vzniklých matic souvisí s určením významnosti hlavních komponent. Jelikož některé publikace pojmy komponentní skóre a komponentní zátěže (váhy) neuvádějí, považuji za důležité tyto pojmy zmínit a vysvětlit jejich význam.

Komponentní skóre představuje souřadnice objektů v novém prostoru definovaném hlavními komponentami. Jedná se tedy o hodnoty hlavních komponent pro každý objekt, které se dále používají jako vstupní data i v dalších analýzách (například v analýze shlukové, diskriminační nebo mnohorozměrné regresi).

Označíme-li jako $\boldsymbol{x}_i$ vektor hodnot *i*-tého objektu $i = 1, \ldots, n$, potom projekce *i*-tého objektu na *j*-tou hlavní komponentu je komponentní skóre

$$Y_{ij} = \boldsymbol{\omega}_j^{\mathrm{T}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \qquad i = 1, \ldots, n,\ j = 1, \ldots, p. \tag{2.3.10}$$

Každý objekt má svůj soubor komponentních skóre, přičemž počet komponentních skóre v tomto souboru odpovídá počtu hlavních komponent. Řádky matice skóre tvoří skóre pro jeden objekt a sloupce jsou tvořeny vektory skóre, které jsou ortogonální a obsahují skóre všech objektů pro jednu hlavní komponentu. Skóre objektů na hlavních komponentách se používá k vykreslení dat do tzv. rozptylového diagramu komponentních skóre (Scatterplot), který si přiblížíme v kapitole 2.4. Jedná se o jeden z nejčastěji užívaných grafů metody hlavních komponent, pomocí kterého můžeme jednoduše určit počet významných komponent.

Matice komponentních vah $\mathbf{\Omega}^{\mathrm{T}}$ je transformační maticí, která převádí původní znaky zdrojové matice **X** na hlavní komponenty. Vztah mezi původními proměnnými a hlavními komponentami vyjadřují **komponentní váhy (zátěže).**

Komponentní váhy jsou prvky vektoru komponentních vah (sloupce matice komponentních vah), který slouží pro lepší srovnání vlastních vektorů kovarianční matice $\sum$. Zatímco prvky uvnitř každého vlastního vektoru jsou vzájemně porovnatelné, nejsou porovnatelné prvky z různých vlastních vektorů. Čím více prvků má vlastní vektor, tím jsou jeho jednotlivé prvky bližší nule. Z tohoto důvodu je pro lepší srovnání vhodné používat vektory komponentních vah, které získáme vynásobením prvků každého vlastního vektoru odmocninou příslušného vlastního čísla, tj.

$$\boldsymbol{\omega}_j\sqrt{\lambda_j} \qquad j = 1, \ldots, p.$$

Prvky vektoru komponentních vah jsou stále vlastní vektory odpovídající kovarianční matici, rozdíl je ale v tom, že jsou porovnatelné, protože čím větší jsou, tím větší význam má i odpovídající hlavní komponenta.

2.3.2 Ukázka redukce původních proměnných

Na následujícím jednoduchém příkladu si názorně ukážeme, jak postupovat při redukci souboru původních veličin, jinak řečeno, jak vytvořit hlavní komponenty bez velké ztráty informace.

Předpokládejme náhodné veličiny X_1, X_2, X_3 a jejich kovarianční maticí $\sum$, která je symetrická a pozitivně semidefinitní

$$\sum = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 1 & 5 \end{bmatrix}.$$

Hned na první pohled můžeme z kovarianční matice poznat, že náhodná veličina X_1 je jednou z hlavních komponent, protože je nekorelovaná s X_2, X_3, tzn. $\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_1, X_3) = 0$.

Nyní vypočítáme vlastní čísla a vlastní vektory matice $\sum$.

Charakteristická rovnice má tvar:

$$\begin{vmatrix} 4-\lambda & 0 & 0 \\ 0 & 3-\lambda & 1 \\ 0 & 1 & 5-\lambda \end{vmatrix} = 0$$

Dostaneme: $(4-\lambda)(3-\lambda)(5-\lambda) - (4-\lambda) = 0$

$(4-\lambda)(\lambda^2 - 8\lambda + 14) = 0$

Vlastní čísla matice $\sum$ jsou $\lambda_1 = 4 + \sqrt{2}$, $\lambda_2 = 4$, $\lambda_3 = 4 - \sqrt{2}$,

Nalezení vlastního vektoru příslušného k vlastnímu číslu $\lambda_1 = 4 + \sqrt{2}$ pak vede k řešení soustavy rovnic (1.2.2), tj.

$$\begin{pmatrix} -\sqrt{2} & 0 & 0 \\ 0 & -\sqrt{2}-1 & 1 \\ 0 & 1 & -\sqrt{2}+1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\omega}_1 \\ \boldsymbol{\omega}_2 \\ \boldsymbol{\omega}_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Provedením elementárních řádkových transformací dostaneme soustavu rovnic

$$\begin{pmatrix} \sqrt{2} & 0 & 0 \\ 0 & -\sqrt{2}-1 & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\omega}_1 \\ \boldsymbol{\omega}_2 \\ \boldsymbol{\omega}_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Pro $\lambda_1 = 4 + \sqrt{2}$ získáme vlastní vektor $\boldsymbol{\omega}_1 = (0, 1, \sqrt{2} + 1)^T$.

Analogicky postupujeme při nalezení vlastních vektorů příslušných k vlastním číslům λ_2 a λ_3. Pro $\lambda_2 = 4$ získáme vlastní vektor $\boldsymbol{\omega}_2 = (1, 0, 0)^T$ a pro $\lambda_3 = 4 - \sqrt{2}$ vektor $\boldsymbol{\omega}_3 = (0, -1, \sqrt{2} - 1)^T$.

Pro $\boldsymbol{\omega}_2 = (1, 0, 0)^T$ je splněna normalizační podmínka $\boldsymbol{\omega}_2^T \boldsymbol{\omega}_2 = 1$. Aby byla splněna normalizační podmínka u $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_3$, znormalizujeme tyto vlastní vektory a dostaneme:

$$\boldsymbol{\omega}_1 = (0, \frac{1}{\sqrt{4+2\sqrt{2}}}, \frac{\sqrt{2}+1}{\sqrt{4+2\sqrt{2}}})^T = (0; 0{,}383; 0{,}924)^T,$$

$$\boldsymbol{\omega}_3 = (0, \frac{-1}{\sqrt{4-2\sqrt{2}}}, \frac{\sqrt{2}-1}{\sqrt{4-2\sqrt{2}}})^T = (0; -0{,}924; 0{,}383)^T.$$

Shrnutí vlastních čísel a vlastních vektorů:

$\lambda_1 = 4 + \sqrt{2} = 5{,}414$, $\quad \boldsymbol{\omega}_1 = (0; 0{,}383; 0{,}924)^T$.

$\lambda_2 = 4$, $\quad \boldsymbol{\omega}_2 = (1, 0, 0)^T$.

$\lambda_3 = 4 - \sqrt{2} = 2{,}586$, $\quad \boldsymbol{\omega}_3 = (0; -0{,}924; 0{,}383)^T$.

Všechny 3 vektory splňují normalizační podmínku. Dle (2.3.1) dostáváme následující hlavní komponenty.

$Y_1 = \boldsymbol{\omega}_1^T \boldsymbol{x} = 0{,}383\, X_2 + 0{,}924\, X_3$

$Y_2 = \boldsymbol{\omega}_2^T \boldsymbol{x} = X_1$

$Y_3 = \boldsymbol{\omega}_3^T \boldsymbol{x} = -0{,}924\, X_2 + 0{,}383\, X_3$

Nyní si postupně ověříme, zda jsou splněny rovnice (2.3.2, 2.3.3).

Rozptyly komponent (2.3.2):

$\mathrm{Var}(Y_1) = \mathrm{Var}(0{,}383X_2 + 0{,}924X_3) = 5{,}414 = \lambda_1$,

$\mathrm{Var}(Y_2) = \mathrm{Var}(X_1) = 4 = \lambda_2$,

$\mathrm{Var}(Y_3) = \mathrm{Var}(-0{,}924\, X_2 + 0{,}383\, X_3) = 2{,}586 = \lambda_3$.

Kovariance komponent (2.3.3):

$\mathrm{Cov}(Y_1, Y_2) = \mathrm{Cov}(0{,}383\, X_2 + 0{,}924\, X_3, X_1) = 0$

$\mathrm{Cov}(Y_2, Y_3) = \mathrm{Cov}(X_1, -0{,}924\, X_2 + 0{,}383\, X_3) = 0$

$\mathrm{Cov}(Y_1, Y_3) = \mathrm{Cov}(0{,}383\, X_2 + 0{,}924\, X_3, -0{,}924\, X_2 + 0{,}383\, X_3) = 0$

Vidíme, že rozptyly komponent odpovídají vlastním číslům a je splněn požadavek nekorelovanosti veličin.

Také je zřejmé, že je splněn vztah (2.3.7):

$$\sigma_{11} + \sigma_{22} + \sigma_{33} = 4 + 3 + 5 = \lambda_1 + \lambda_2 + \lambda_3 = 5{,}414 + 4 + 2{,}586 = 12$$

Dle (2.3.8) zjistíme, že první hlavní komponenta Y_1 přibližně vysvětluje $\lambda_1/(\lambda_1 + \lambda_2 + \lambda_3) = 5{,}414 / 12 = 0{,}451 \Rightarrow 45{,}1$ % celkového rozptylu vysvětlovaných veličin X_1, X_2, X_3.

Dále uvažujme první dvě komponenty, které vysvětlují $(\lambda_1 + \lambda_2)/(\lambda_1 + \lambda_2 + \lambda_3) = 0{,}785 \Rightarrow 78{,}5$ % celkového rozptylu vysvětlovaných veličin. V tomto případě můžeme původní tři veličiny nahradit dvěma hlavními komponentami Y_1, Y_2, které bez velké ztráty informace z velké části vyčerpají celkovou variabilitu veličin X_1, X_2, X_3.

Vztahy mezi původními veličinami X_1, X_2, X_3 a hlavními komponentami Y_1, Y_2 vyjadřují příslušné kovariance a korelační koeficienty (2.3.9):

$$\rho_{Y_1,X_1} = \frac{\omega_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = 0$$

$$\rho_{Y_1,X_2} = \frac{\omega_{12}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}} = \frac{0{,}383\sqrt{5{,}414}}{\sqrt{3}} = 0{,}515$$

$$\rho_{Y_1,X_3} = \frac{\omega_{13}\sqrt{\lambda_1}}{\sqrt{\sigma_{33}}} = \frac{0{,}924\sqrt{5{,}414}}{\sqrt{5}} = 0{,}962$$

$$\rho_{Y_2,X_1} = \frac{\omega_{21}\sqrt{\lambda_2}}{\sqrt{\sigma_{11}}} = \frac{1\sqrt{4}}{\sqrt{4}} = 1$$

$$\rho_{Y_2,X_2} = \frac{\omega_{22}\sqrt{\lambda_2}}{\sqrt{\sigma_{22}}} = 0$$

$$\rho_{Y_2,X_3} = \frac{\omega_{23}\sqrt{\lambda_2}}{\sqrt{\sigma_{33}}} = 0$$

Je patrné, že původní veličiny X_2, X_3 jsou silně korelované s hlavní komponentou Y_1 a mezi veličinou X_1 a hlavní komponentou Y_2 je přímá lineární závislost. Zbylé korelace můžeme zanedbat, protože třetí komponenta Y_3 není důležitá.

2.3.3 Standardizace dat

Následující kapitola vychází z [1], [2].

Doposud popsaný postup určení hlavních komponent přímo z kovarianční matice je použitelný pouze v úlohách, ve kterých jsou sledované náhodné veličiny $X_1, \ldots, X_p$ měřeny ve stejných jednotkách, a navíc rozptyly těchto veličin nejsou zásadně odlišné. Pokud nejsou obě tyto podmínky splněny, můžeme použít standardizované veličiny, které získáme normalizací (lineární transformací) původních veličin, a to následovně

$$\begin{aligned} Z_1 &= \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}} \\ &\vdots \\ Z_p &= \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}} \end{aligned} \tag{2.3.11}$$

Z těchto normovaných (standardizovaných) veličin vypočítáme kovarianční matici $\sum$ dle (1.3.3). Jinak má smysl provádět metodu hlavních komponent na základě populační korelační matice $\boldsymbol{\rho}$ (1.3.6).

Standardizaci původních veličin můžeme zapsat maticově

$$\mathbf{Z} = \left(\mathbf{V}^{\frac{1}{2}}\right)^{-1} (\mathbf{X} - \boldsymbol{\mu}),$$

kde

$$\mathbf{V}^{\frac{1}{2}}{}_{(p \text{ x } p)} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix}$$

je diagonální matice řádu p, která má na hlavní diagonále směrodatné odchylky a $\boldsymbol{\mu}$ je vektor středních hodnot (1.3.2).

Dle [1, str. 437] je zřejmé, že

$$\mathrm{E}(\mathbf{Z}) = 0,$$

$$\mathrm{Cov}(\mathbf{Z}) = \left(\mathbf{V}^{\frac{1}{2}}\right)^{-1} \sum \left(\mathbf{V}^{\frac{1}{2}}\right)^{-1} = \boldsymbol{\rho}.$$

Numericky je tedy lhostejné zda hlavní komponenty získáme z kovarianční matice normovaných veličin či z korelační matice původních (netransformovaných) veličin.

Nicméně je nutné si uvědomit, že výsledky získané z kovarianční matice $\sum$ původních (netransformovaných) veličin se nedají převést na výsledky získané z korelační matice $\boldsymbol{\rho}$.

Pro i-tou hlavní komponentu získanou jako lineární kombinaci standardizovaných veličin $Z_1, \ldots, Z_p$ s kovarianční maticí $\mathrm{Cov}(\mathbf{Z}) = \boldsymbol{\rho}$ platí

$$Y_i = \boldsymbol{\omega}_i^{\mathrm{T}} \boldsymbol{z} = \boldsymbol{\omega}_i^{\mathrm{T}} \left(\mathbf{V}^{\frac{1}{2}}\right)^{-1} (\mathbf{X} - \boldsymbol{\mu}), \qquad i = 1, \ldots, p.$$

Kromě toho

$$\sum_{i=1}^{p} \mathrm{Var}(Y_i) = \sum_{i=1}^{p} \mathrm{Var}(Z_i) = p$$

a

$$\rho_{Y_i, Z_k} = \boldsymbol{\omega}_{ik} \sqrt{\lambda_i} \qquad i, k = 1, \ldots, p.$$

kde dvojice $(\lambda_1, \boldsymbol{\omega}_1), \ldots, (\lambda_p, \boldsymbol{\omega}_p)$ jsou sestupně uspořádaná vlastní čísla a jejich odpovídající vlastní vektory získané z korelační matice $\boldsymbol{\rho}$.

Mírou významu k-té hlavní komponenty vzhledem k celkovému rozptylu standardizovaných veličin $Z_1, Z_2, \ldots, Z_p$ je zde podíl $\frac{\lambda_k}{p}$, $k = 1, \ldots, p$.

2.4 Grafické zobrazení hlavních komponent

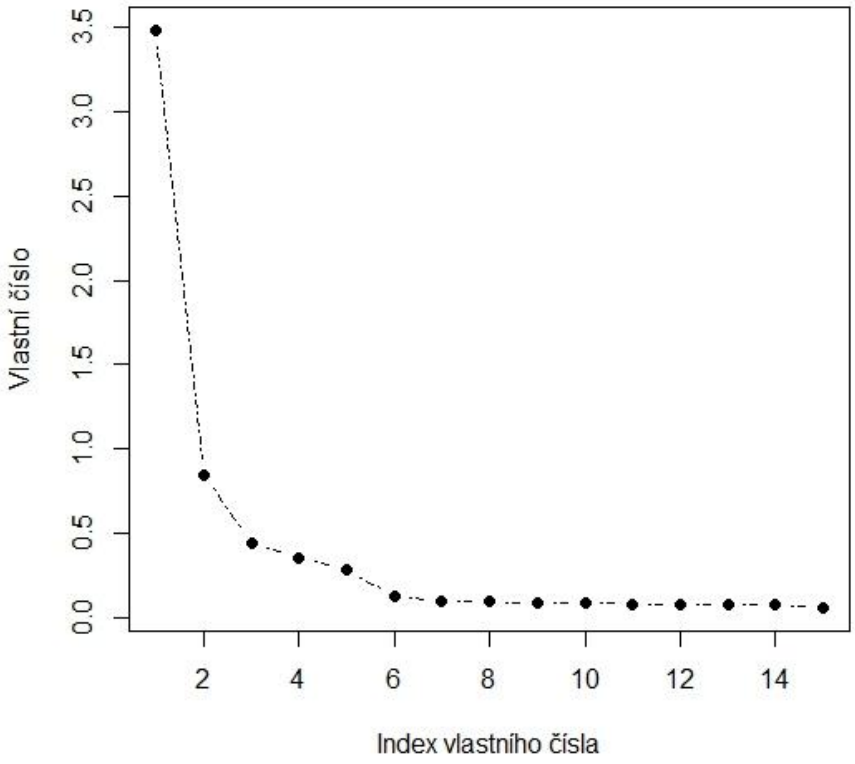
Výsledky metody hlavních komponent lze zobrazit pomocí různých grafů. V této části si ukážeme pouze některé typy grafů, které se pro interpretaci výsledků používají nejčastěji. Informace k této kapitole jsou čerpány zejména z [1], [3].

2.4.1 Cattelův indexový graf úpatí vlastních čísel (Scree plot)

Jedná se o graf, který zobrazuje relativní velikost jednotlivých vlastních čísel kovarianční matice uspořádaných sestupně, tj. $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Tento graf se často využívá pro určení počtu k významných komponent. Úpatí je zlomový bod v grafu, který odděluje významné komponenty od nevýznamných. Nachází se v místě, ve kterém se křivka dostává do rovnoběžné úrovně osy s indexy vlastních čísel. Hodnota indexu

tohoto zlomu udává počet významných komponent. Na obrázku 2.4 je uvedený příklad, ve kterém je z 15 hlavních komponent nejvýznamnější prvních 5 komponent.

Obrázek 2.4: Scree plot



2.4.2 Rozptylový diagram komponentních skóre (Scatterplot)

Používá se pro grafické znázornění komponentních skóre (2.3.10) čili hodnot obvykle pro první dvě hlavní komponenty u všech objektů (pozorování). Body v tomto grafu jsou Y_{i1}, Y_{i2}, $i = 1, \ldots, n$. Scatterplot ukazuje vztah mezi jednotlivými objekty. Také se užívá například k identifikaci odlehlých pozorování, trendů, shluků objektů a k vysvětlení podobnosti objektů. Pozorování nacházející se blízko sebe si jsou vzájemně velmi podobná, naopak odlehlá pozorování se silně odlišují. Pokud jsou pozorování umístěná zřetelně v jednom shluku, jsou si vzájemně podobná a přitom se liší od ostatních pozorování vyskytující se mimo shluk. Více shluků dobře oddělených od sebe nám dává najevo, že lze nalézt vlastní model pro samotné shluky. Tento diagram použijeme pro interpretaci hlavních komponent v praktickém příkladu v kapitole 3.

2.4.3 Dvojný graf (Biplot)

Jedná se o grafický nástroj mnohorozměrné statistické analýzy, který znázorňuje hodnoty prvních dvou hlavních komponent a zároveň jejich vztah k původním znakům. Biplot se dá také interpretovat jako graf znázorňující komponentní skóre a zátěže obou hlavních komponent. Jednotlivá pozorování jsou zobrazena pomocí bodů, šipky (vektory) vycházející z počátku reprezentují jednotlivé původní statistické znaky. Úhel mezi šipkami je nepřímo úměrný velikosti korelace mezi původními veličinami. Kosinus tohoto úhlu aproximuje jejich korelační koeficient. Svírají-li původní znaky úhel 90°, potom kosinus tohoto úhlu odpovídá korelačnímu koeficientu o velikosti 0, to znamená, že tyto znaky jsou nekorelované. Platí, že čím menší je úhel mezi nimi, hodnota korelačního koeficientu se blíží k 1, tj. tím silnější je korelace. Naopak blíží-li se úhel ke 180°, hodnota korelačního koeficientu klesá k -1. V tom případě se jedná o zápornou korelaci, tedy s rostoucí hodnotou jedné proměnné, klesá hodnota druhé proměnné. Biplot také znázorňuje, jakou měrou přispívají jednotlivé původní znaky do hlavních komponent. Každý vektor má své souřadnice na první a druhé hlavní komponentě. Délka této souřadnice odpovídá příspěvku původního znaku do hlavní komponenty (tj. je úměrná komponentní váze). Pokud se v biplotu nachází pozorování v blízkosti určitého statistického znaku, pak obsahuje velký podíl tohoto znaku. Stejně jako u diagramu komponentních skóre platí, že nacházejí-li se jednotlivá pozorování blízko u sebe, jsou si podobná. Praktické využití dvojného grafu si ukážeme na konkrétním příkladu v následující kapitole.

3 Příklady

V této kapitole si na konkrétních příkladech ukážeme využití metody hlavních komponent v praxi. K výpočtům a grafickému vyjádření nám poslouží především statistický software R (www.r-project.org) [15]. Pouze pro realizaci druhého příkladu použijeme navíc software Mathematica. Dalšími zdroji pro tvorbu této kapitoly, zejména z oblasti softwaru, byly [16], [17], [20].

3.1 Složení píce

První datový soubor použitý k ukázce práce s metodou hlavních komponent obsahuje 48 pozorování chemického složení píce od roku 2003 do roku 2010. Data jsou převzata od společnosti Agrovýzkum Rapotín s.r.o., která byla naměřena na území Rapotín při sledování vlivu různých úrovní pastevního využívání a hnojení trvalých travních porostů na botanické složení z hlediska nutričních hodnot. Datový soubor tvoří 6 pozorování složení píce v každém roce, přičemž se od sebe tato pozorování v rámci jednoho roku vzájemně liší počtem pastevních cyklů za rok a hnojením. Počet pastevních cyklů je rozdělen do 3 úrovní (2, 3 a 4 pastevní cykly za rok), přičemž pro jednotlivé úrovně je naměřeno chemické složení píce ve dvou variantách, a to u píce hnojené a nehnojené.

Úkolem je zjistit, jaký vliv má hnojení, rok, ve kterém bylo provedeno měření, a počet pastevních cyklů na složení píce, které je definováno pomocí pěti složek v různých množstvích. Abychom mohli vliv jednotlivých faktorů podrobněji zkoumat, chceme složení píce zjednodušit tím, že pětici složek charakterizujeme jednou či dvěma veličinami, tj. hlavními komponentami.

Na původní datový soubor tvořený složkami píce aplikujeme metodu hlavních komponent. Proměnnými jsou zde množství složek uvedených v příslušných jednotkách, tedy vláknina [$g.kg^{-1}$ sušiny], NDF[5] [$g.kg^{-1}$ DM], ADF[6] [$g.kg^{-1}$ DM], tuk [$g.kg^{-1}$ sušiny] a popel [$g.kg^{-1}$ sušiny]. Tento datový soubor, pomocí výše uvedených postupů, nejprve zredukujeme do prvních dvou hlavních komponent s největší variabilitou. Jak již bylo řečeno v kapitole 2.3.3, hlavní komponenty můžeme získat z kovarianční matice normovaných veličin či z korelační matice původních (netransformovaných) veličin. V tomto případě budeme vycházet z kovarianční matice normovaných veličin, protože

[5] Neutral Detergent Fibre = neutrálně detergentní vláknina.
[6] Acid Detergent Fibre = acidodetergentní vláknina.

budeme tyto veličiny později potřebovat i pro grafické znázornění hlavních komponent. Normalizaci provedeme dle vztahu 2.3.11 a získáme normované složky píce (Příloha A), které budeme považovat za zdrojová data.

K níže uvedeným výpočtům a grafickému znázornění použijeme statistický software R. V prvním kroku si vždy nastavíme pracovní adresář v daném počítači. K tomu použijeme příkaz:

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

Zdrojová data jsou uložena v tabulce v souboru MS Excel `normov.veliciny.csv`. Tato data načteme pomocí příkazu:

```
> H=read.csv("normov.veliciny.csv", header = FALSE, sep = ";", dec=",",
fill = TRUE).
```

Následujícím příkazem vytvoříme kovarianční matici normovaných veličin:

```
> R=var(H)
> R
           V1         V2         V3         V4         V5
V1  1.0000000  0.8820202  0.8536411 -0.5879256 -0.4144194
V2  0.8820202  1.0000000  0.8442962 -0.6734658 -0.4333213
V3  0.8536411  0.8442962  1.0000000 -0.6594064 -0.2196677
V4 -0.5879256 -0.6734658 -0.6594064  1.0000000  0.4376037
V5 -0.4144194 -0.4333213 -0.2196677  0.4376037  1.0000000
```

Je patrné, že korelace mezi prvními třemi veličinami, tj. množstvím vlákniny, NDF a ADF, je velmi silná. Množství těchto složek bude na sobě silně záviset. Z korelační matice vyjádříme vlastní čísla a jim odpovídající vlastní vektory:

```
> eigen(R)$values

[1] 3.48848011 0.84503651 0.44370420 0.12385974 0.09891944

> eigen(R)$vectors

            [,1]       [,2]       [,3]       [,4]       [,5]
[1,] -0.4940208 0.1495464  0.4055558 -0.1329793  0.7425767
[2,] -0.5052598 0.1070360  0.1899859  0.7696960 -0.3236190
[3,] -0.4807321 0.3815879  0.0293280 -0.5941234 -0.5190804
[4,]  0.4315914 0.1212693  0.8662008 -0.0387203 -0.2173003
[5,]  0.2885854 0.8976991 -0.2196945  0.1881558  0.1648836
```

Dle vztahu 2.3.1 vytvoříme první dvě hlavní komponenty:

$$Y_1 = -0{,}4940208z_1 - 0{,}5052598z_2 - 0{,}4807321z_3 + 0{,}4315914z_4 + 0{,}2885854z_5$$

$$Y_2 = 0{,}1495464z_1 + 0{,}1070360z_2 + 0{,}3815879z_3 + 0{,}1212693z_4 + 0{,}8976991z_5$$

Po dosazení příslušných hodnot normovaných veličin (Příloha A) do vytvořených hlavních komponent získáme hodnoty těchto komponent pro jednotlivá pozorování, které použijeme pro další analýzu. Obě hlavní komponenty lze interpretovat jako lineární kombinaci složek píce. Přičemž platí, že pokud klesá množství vlákniny, NDF a ADF nebo roste množství tuku a popela obsaženého v píci, roste hodnota první hlavní komponenty. Je zde viditelná závislost prvních třech složek, jak již vyplynulo z hodnot korelačních koeficientů mezi těmito proměnnými. Menší informaci o variabilitě složek v píci nese druhá hlavní komponenta. Zvyšuje-li se množství všech složek v píci, nejvíce však množství popela a ADF, roste hodnota druhé hlavní komponenty. Ze vztahu (2.3.8) vyplývá, že první hlavní komponenta Y_1 vysvětluje přibližně 69,8% a druhá hlavní komponenta Y_2 dalších 16,9% celkové variability původních veličin, tj. vlákniny, NDF, ADF, tuku a popela. V tomto případě můžeme bez velké ztráty informace nahradit původních pět veličin dvěma hlavními komponentami Y_1 a Y_2.

3.1.1 Analýza rozptylových diagramů komponentních skóre

V této kapitole se pokusíme analyzovat vliv hnojení, roku, ve kterém bylo provedeno měření, a počtu pastevních cyklů na složení píce pomocí grafického znázornění hodnot první a druhé hlavní komponenty u všech pozorování. Použijeme rozptylový diagram komponentních skóre (Scatterplot) uvedený v kapitole 2.4. Nejdříve načteme zdrojová data, která jsou tvořena první a druhou hlavní komponentou pomocí příkazu `read.csv` a označíme hlavní komponenty následovně:

```
> x=Data[[1]]         ## 1.hlavní komponenta
> y=Data[[2]]         ## 2.hlavní komponenta
```

Pro vykreslení bodového grafu použijeme příkaz:

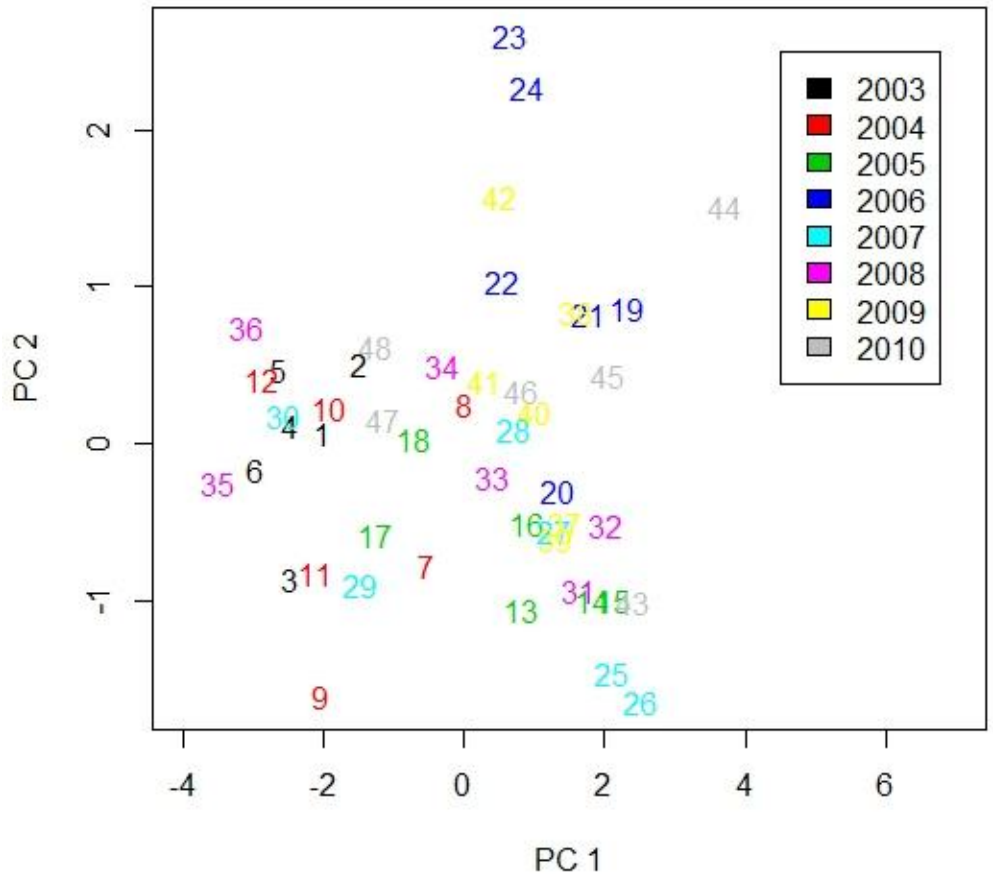
```
> plot(x,y,xlab="PC 1",ylab="PC 2",type="n")
```

Díky zvolenému typu (`"n"`) výstupního grafu nám software R nevykreslí žádná data, ale pouze osy *x* a *y*, které představují první a druhou hlavní komponentu. Pro zjištění vlivu jednotlivých proměnných na první hlavní komponentu zkusíme jednotlivá pozorování v grafu barevně odlišit podle zvolených proměnných. Postupujeme tak, že zadáme příkaz, pomocí kterého software R vypíše všechna pozorování do grafu na pozice [*x*, *y*] jako textový řetězec obsahující číslo pozorování a následně je odliší podle zvolené proměnné.

Jako první barevně odlišíme jednotlivá pozorování podle roku, ve kterém byla provedena analýza složení píce. Připomeňme, že v každém roce bylo provedeno 6 měření chemického složení píce, a to v průběhu osmi let (2003 - 2010). Zadáme tedy příkaz, pomocí kterého barevně odlišíme jednotlivé roky a vykreslíme pozorování do grafu jako textový řetězec obsahující číslo pozorování.

```
> r=rep(1:8,each=6)
> text(x,y,1:48,col=r)
```

Obrázek 3.1: Scatterplot – pozorování rozlišená podle roku analýzy



Výsledné grafické znázornění uvedené na obrázku 3.1 se nám může na první pohled zdát chaotické. Ale při podrobnějším zkoumání, si můžeme všimnout, že pozorování v roce 2003, která jsou vykreslená černou barvou, se nachází převážně v levé části grafu a červená pozorování pro rok 2004 jsou lehce posunutá doprava. Tato tendence přibližně pokračuje až do posledního pozorování v roce 2010. Z obrázku

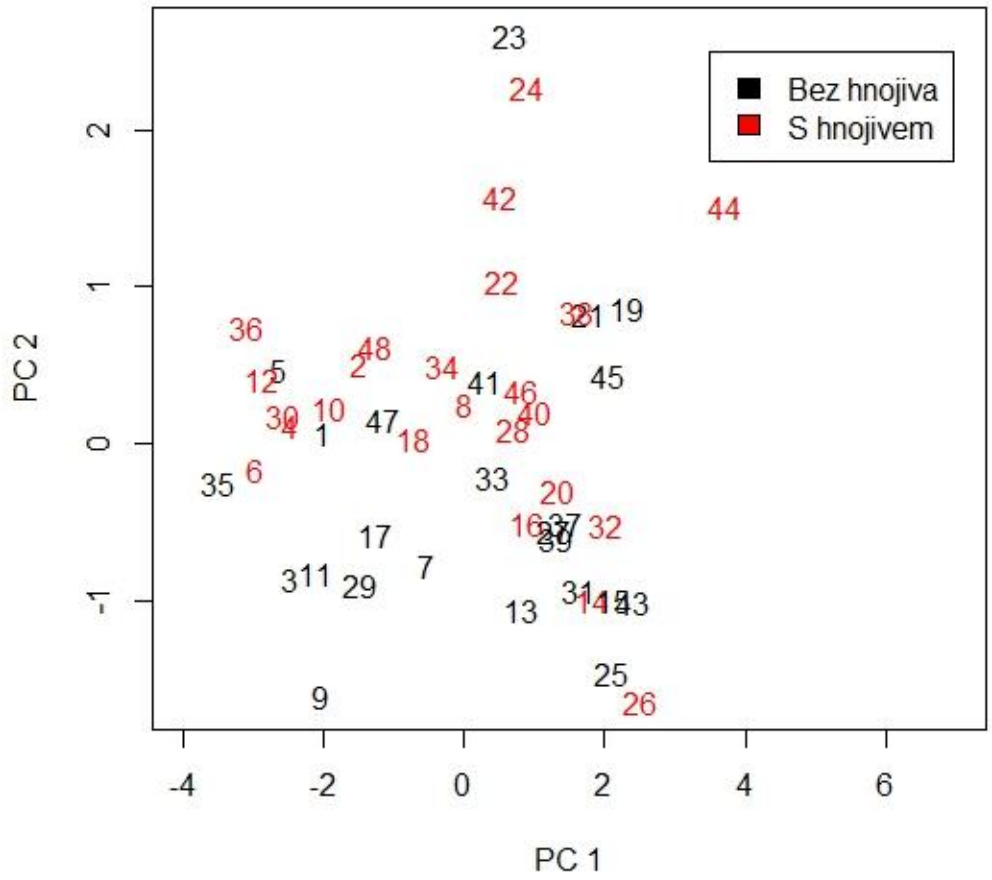
lze, sice s větší obtížností, vyčíst rostoucí tendenci první hlavní komponenty (PC 1). Platí tedy, že s rostoucím rokem roste i hodnota první komponenty, tj. klesá množství vlákniny, NDF a ADF nebo roste množství tuku a popela obsaženého v píci. Tuto tendenci nám ale narušuje pozorování v roce 2008, značené růžovou barvou. Je patrné, že tato pozorování jsou rozprostřená po celém grafu, to znamená, že se od sebe výrazně liší. Naopak zelená pozorování v roce 2005, konkrétně 14 a 15, jsou si z hlediska první komponenty velice podobná. Přičemž pozorování č. 14 bylo provedeno u pastevního porostu, který byl hnojený a počet pastevních cyklů odpovídal čtyřem, zatímco u pozorování č. 15 hnojivo aplikováno nebylo a má pouze 3 pastevní cykly. Můžeme tedy konstatovat, že v roce 2005 nemá na složení píce vliv hnojení ani počet pastevních cyklů. Při orientaci v grafu zdola nahoru si můžeme všimnout, že například pozorování v roce 2005 mají nižší hodnoty druhé hlavní komponenty oproti pozorováním v roce 2006. To znamená, že v období 2005 - 2006 roste hodnota druhé hlavní komponenty (PC 2), tj. zvyšuje se množství všech složek v píci, nejvíce však množství popela a ADF. Opačnou tendenci mají pozorování v období 2006 - 2007, ve kterém hodnota druhé hlavní komponenty klesá. Můžeme konstatovat, že rok, ve kterém bylo provedeno měření, do jisté míry ovlivňuje druhou hlavní komponentu.

Nyní barevně odlišíme jednotlivá pozorování podle hnojení. Postupujeme obdobně jako v předchozím případě pomocí příkazu:

```
> r2 = rep(c(1,2),24)
> text(x,y,1:48,col=r2)
```

Grafický výstup ukazuje obrázek 3.2, ve kterém odpovídají černým bodům pozorování bez aplikace hnojiva, naopak pozorování, v nichž bylo hnojení aplikováno, odpovídá červeným bodům. Při pohledu na graf si můžeme všimnout, že jsou černá i červená pozorování vzhledem k první komponentě rozmístěná po celém jejím intervalu. To svědčí o tom, že hnojení nemá vliv na první komponentu, což je překvapivý výsledek, jelikož by se dalo očekávat, že právě hnojení bude nějakým způsobem ovlivňovat kvalitu i složení píce. Podíváme-li se na druhou komponentu, můžeme zpozorovat její rostoucí tendenci. Hnojení tedy nemá vliv na první komponentu, ale může mít vliv na druhou hlavní komponentu, a to takový, že při aplikaci hnojiva hodnota druhé hlavní komponenty roste. To ale neplatí vždy, tento vliv nám vyvrací například pozorování č. 23, které i přes to, že u něj nebylo hnojení aplikováno, má největší hodnotu druhé komponenty. Může to být způsobeno větším vlivem ostatních proměnných (rok, počet pastevních cyklů za rok).

Obrázek 3.2: Scatterplot – pozorování rozlišená podle hnojení

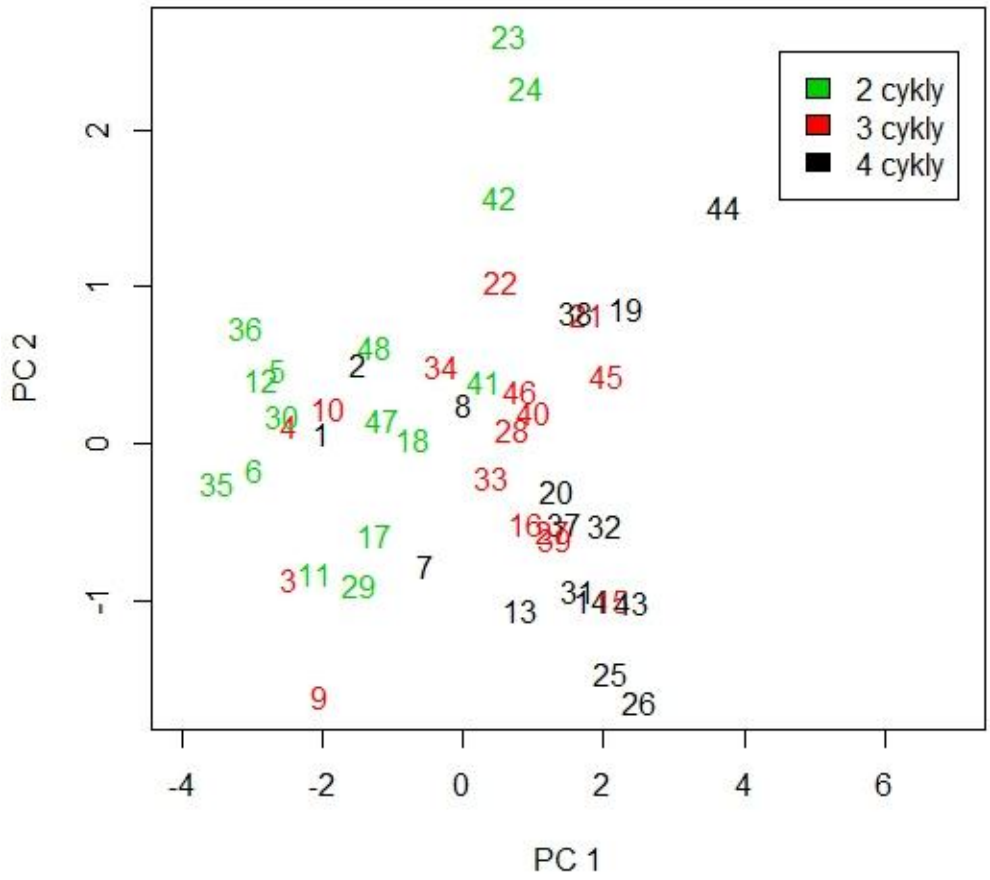


Nakonec odlišíme pozorování podle počtu pastevních cyklů příkazem:

```
> r3 = rep(c(1,1,2,2,3,3),8)
> text(x,y,1:48,col=r3)
```

Ve výsledném grafu na obrázku 3.3 jsou znázorněna pozorování s dvěma cykly (zelená), třemi cykly (červená) a čtyřmi cykly za rok (černá). Již na první pohled je v tomto grafu viditelná rostoucí tendence první komponenty, tj. s rostoucím počtem pastevních cyklů roste i hodnota první komponenty. Grafické znázornění není tak chaotické jako u prvního obrázku 3.1, ve kterém jsme odlišili pozorování podle roku analýzy, což může být způsobeno tím, že proměnná počet pastevních cyklů má větší vliv na první komponentu než proměnná rok. Z toho důvodu může být barevné odlišení v grafu podle počtu cyklů přehlednější než podle roku analýzy. Při pohledu na druhou hlavní komponentu můžeme sledovat její opačnou tendenci, kde s rostoucím počtem pastevních cyklů klesá hodnota druhé hlavní komponenty.

Obrázek 3.3: Scatterplot – Pozorování rozlišená podle počtu pastevních cyklů



3.1.2 Biplot

V této části graficky znázorníme vztah mezi jednotlivými složkami píce pomocí biplotu, který jsme si uvedli v kapitole 2.4. K vykreslení biplotu v softwaru R musíme jako zdrojová data použít normované veličiny (Příloha A), protože příkaz, který použijeme pro vykreslení biplotu, neumí původní data automaticky znormovat. Nejprve načteme zdrojová data, která jsou uložena v tabulce v souboru MS Excel `normov.veliciny.csv`, stejným způsobem, jako tomu bylo u výpočtu hlavních komponent na začátku příkladu.

```
> H=read.csv("normov.veliciny.csv", header = FALSE, sep = ";", dec=",",
fill = TRUE)
```

Dále označíme jednotlivé sloupce názvy složek píce:

```
> colnames(H)=c("Vlaknina","NDF","ADF","Tuk","Popel")
```

Následuje příkaz, který shrnuje důležité informace o hlavních komponentách

```
> summary(princomp(H))

Importance of components:
                          Comp.1   Comp.2   Comp.3   Comp.4   Comp.5
Standard deviation     1.848189 0.909632 0.659136 0.348251 0.311221
Proportion of Variance 0.697696 0.169007 0.088740 0.024771 0.019783
Cumulative Proportion  0.697696 0.866703 0.955444 0.980216 1.000000
```

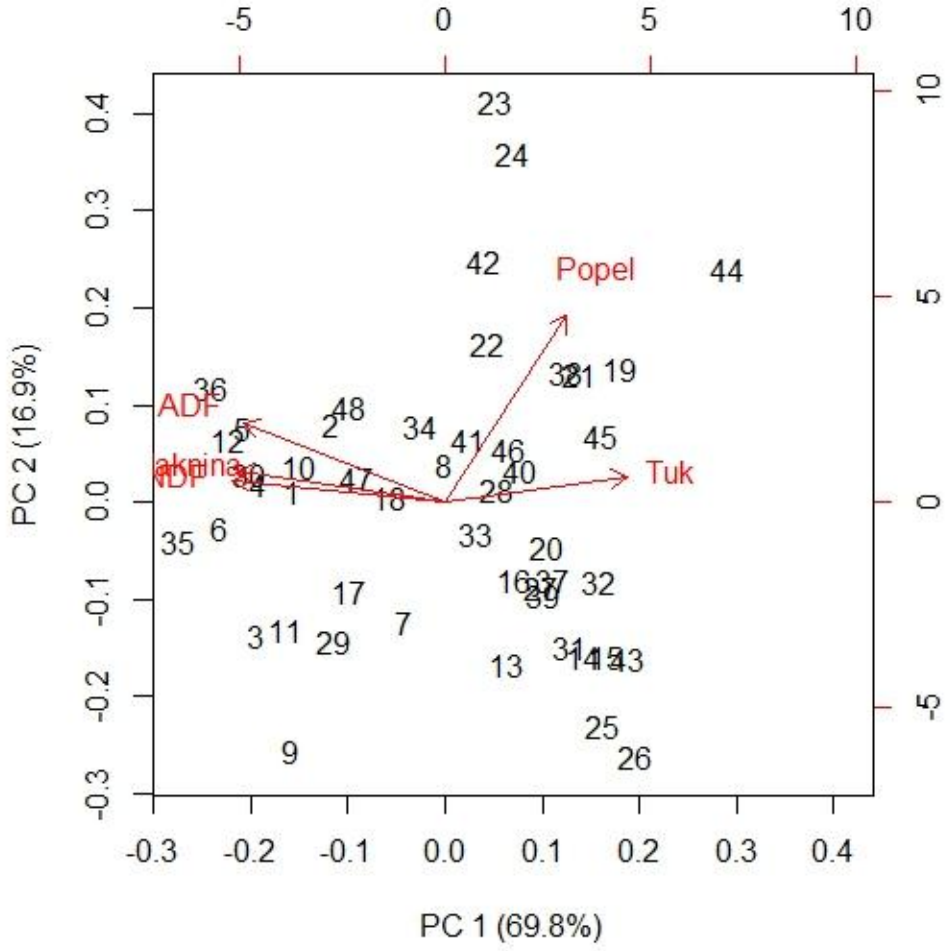
První řádek tabulky s názvem `Standard deviation` nám ukazuje směrodatné odchylky nově vzniklých veličin, tj. hlavních komponent. Řádek s názvem `Proportion of Variance` vyjadřuje (po vynásobení 100), kolik procent celkového rozptylu je vyčerpáno pomocí hlavních komponent. V našem případě potvrzujeme, že první hlavní komponenta vysvětluje 69,8% a druhá hlavní komponenta dalších 16,9% celkové variability, což dá v součtu 86,7%, jak je uvedeno v posledním řádku tabulky, ve sloupci `Comp.2`. Jak již bylo řečeno na začátku tohoto příkladu, původní datový soubor tvořený pěti složkami pastevního porostu můžeme bez velké ztráty informace, nahradit pouze prvními dvěma hlavními komponentami. Biplot tedy bude velmi dobře odrážet skutečnou strukturu vícerozměrného datového souboru.

Pro vykreslení biplotu použijeme příkaz:

```
> biplot(princomp(H),xlab="PC 1 (69.8%)",ylab="PC 2 (16.9%)")
```

Grafický výstup je znázorněn na obrázku 3.4. Připomeňme, že první i druhou hlavní komponentu jsme interpretovali jako danou kombinaci složek píce, přičemž pro první hlavní komponentu platí, že pokud klesá množství vlákniny, NDF a ADF nebo roste množství tuku a popela obsaženého v píci, její hodnota roste. Pozorování umístěná v levé části grafu na obrázku 3.4 odpovídají nízkým hodnotám první hlavní komponenty, tudíž mají velké zastoupení vlákniny, NDF a ADF a malé zastoupení tuku a popela v píci, naopak vysoké hodnoty hlavní komponenty nesou pozorování napravo. Uspořádání zdola nahoru ukazuje rostoucí hodnoty druhé komponenty, při kterých roste množství všech složek v píci, nejvíce však množství popela a také množství ADF.

Obrázek 3.4: Biplot – složení píce



Jak již bylo řečeno v kapitole 2.4, vztahy mezi jednotlivými proměnnými můžeme interpretovat pomocí šipek. Kosinus úhlů, které šipky svírají, vyjadřuje korelaci mezi proměnnými. Při pohledu na směry šipek vidíme, že nejmenší úhly mezi sebou mají složky NDF, vláknina a ADF, tudíž tyto složky jsou vzájemně silně korelované. Naopak úhel mezi složkou popel a složkami ADF, vláknina, NDF se pohybuje kolem 90 stupňů, to znamená, že jsou slabě korelované až nekorelované. Množství popela v píci tedy nezávisí na množství ostatních složek. Dále úhel mezi složkou tuk a složkami NDF, vláknina a ADF je přibližně roven 180 stupňům. Kosinus tohoto úhlu odpovídá korelačnímu koeficientu s hodnotou -1, tuk je tedy záporně korelovaný s uvedenými třemi složkami píce. Pozorování, u nichž se v píci vyskytuje vysoký podíl tuku, obsahují malé množství NDF, vlákniny a ADF či naopak. Díky biplotu můžeme hned na první pohled vidět, která pozorování jsou si podobná (čím menší vzdálenost je mezi nimi, tím jsou si podobnější) a dále jaká složka píce u jednotlivých pozorování převažuje.

Je zřejmé, že shluk hodnot, vytvořený okolo šipek NDF, vláknina, ADF, vykresluje pozorování s nejvyšším obsahem těchto složek v píci. Zajímavé je sledovat, že například pozorování 23 a 24, v horní části grafu, se svými vlastnostmi výrazněji liší od ostatních, ale vzájemně jsou si velmi podobná. Stejně tak je tomu i u pozorování 25 a 26 ve spodní části grafu. Rozdílná pozice těchto dvojic je způsobena počtem pastevních cyklů a rokem analýzy (viz Příloha B). Pozorování 23 a 24 odpovídá dvěma pastevním cyklům za rok 2006, zatímco pozorování 25 a 26 má čtyři pastevní cykly za rok 2007. Tato skutečnost potvrzuje předchozí analýzu provedenou pomocí rozptylových diagramů komponentních skóre, tedy, že s rostoucím počtem pastevních cyklů a rokem měření roste hodnota první, v tomto případě i hodnota druhé hlavní komponenty. V rámci těchto dvojic se od sebe pozorování vzájemně liší pouze použitím hnojiva. U pozorování 23 a 25 nebylo hnojivo aplikováno, zatímco u pozorování 24 a 26 aplikováno bylo, což se ale nijak výrazně neprojeví na poměru chemických složek v píci, který je reprezentovaný první hlavní komponentou. To svědčí o tom, že hnojení nezávisí na první komponentě. Tento vliv jsme zjistili již v předchozí kapitole při rozboru obrázku 3.2. Dále jsme zjistili, že hnojení může mít vliv na druhou hlavní komponentu, a to takový, že při aplikaci hnojiva hodnota druhé hlavní komponenty roste. Z obrázku 3.4 můžeme vidět, že se hnojení odrazí v poměru složek druhé hlavní komponenty, konkrétně v množství popela. Při aplikaci hnojiva tedy z velké části roste množství popela obsaženého v píci.

3.1.3 Mnohonásobná lineární regrese

Připomeňme, že cílem je zjistit, jaký má vliv hnojení, rok, ve kterém bylo provedeno měření, a počet pastevních cyklů na složení píce. Přičemž proměnná hnojení má 2 úrovně (s hnojivem/ bez hnojiva), délka sledovaného období odpovídá 8 rokům (2003 - 2010) a proměnná počet pastevních cyklů za rok má 3 úrovně (2/ 3/ 4). Datový soubor obsahuje tedy celkem 48 možností. Abychom mohli vliv jednotlivých faktorů podrobněji zkoumat, zjednodušili jsme složení píce tím, že jsme původní pětici složek charakterizovali níže uvedenými dvěma proměnnými, tj. hlavními komponentami.

$$Y_1 = -0{,}4940208z_1 - 0{,}5052598z_2 - 0{,}4807321z_3 + 0{,}4315914z_4 + 0{,}2885854z_5$$

$$Y_2 = 0{,}1495464z_1 + 0{,}1070360z_2 + 0{,}3815879z_3 + 0{,}1212693z_4 + 0{,}8976991z_5$$

Pro zjištění vlivu uvedených faktorů použijeme v této kapitole model mnohonásobné lineární regrese, který je podrobněji popsán v publikacích [1], [8], [18].

Tento model je velmi účinný pro analýzu vztahů mezi skupinou nezávisle proměnných a jednou závisle proměnnou.

Předpokládejme, že nekorelované náhodné veličiny $Y_{11}, \ldots, Y_{1i}$, $i = 1, \ldots, 48$, tj. hodnoty první hlavní komponenty, lze popsat lineární funkcí 4 neznámých parametrů (regresních koeficientů). Lineární regresní model vypadá následovně:

$$Y_{1i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + E_i \qquad i = 1, \ldots, 48 \qquad (3.1.1)$$

Zde první hlavní komponenta Y_{1i} vystupuje jako závislá (vysvětlovaná) proměnná a x_{ij}, $j = 1, 2, 3$ jsou nezávislé proměnné, v našem případě rok (x_{i1}), počet pastevních cyklů za rok (x_{i2}) a hnojení (x_{i3}) a E_i jsou nezávislé náhodné veličiny.

Pro snadnější výpočet regrese v softwaru R hodnoty nezávislých proměnných upravíme. Každému roku přiřadíme hodnotu na stupnici od 0 do 7 (tj. hodnota 0 odpovídá pozorování v roce 2003, hodnota 7 je přiřazena pozorováním v roce 2010). U proměnné hnojení přiřadíme pozorováním, ve kterých bylo hnojivo aplikováno hodnotu 1, u ostatních pozorování použijeme hodnotu 0. Upravená tabulka pro výpočet regrese je uvedena v Příloze B. Takto upravená zdrojová data, která jsou uložena v tabulce `regrese_hl.komp.csv`, načteme v softwaru R příkazem:

```
>  Data=read.csv("regrese_hl.komp.csv",  header  =  FALSE,  sep  =  ";",
dec=",",fill = TRUE)
```

Jednotlivé proměnné označíme pomocí příkazu:

```
> y=Data[[4]]         ##první hlavní komponenta
> x1=Data[[1]]        ##rok
> x2=Data[[2]]        ##počet pastevních cyklů
> x3=Data[[3]]        ##hnojení
```

Na regresním modelu nás budou nejprve zajímat jeho parametry. Regresní model aplikujeme na naše data takto:

```
> regrese1=lm(y~x1+x2+x3)
> regrese1

Call:
lm(formula = y ~ x1 + x2 + x3)

Coefficients:
(Intercept)           x1           x2           x3
    -5.3517       0.3914       1.3458      -0.1113
```

Ve výstupu `regrese1` jsme získali odhady regresních parametrů. Další základní informace o modelu získáme použitím funkce `summary` na objekt představující regresní model:

```
> summary(lm(y~x1+x2+x3))

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-2.80759 -0.90043  0.00668  0.83255  2.65679

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.35170    0.75352  -7.102 8.10e-09 ***
x1           0.39142    0.07788   5.026 8.84e-06 ***
x2           1.34580    0.21854   6.158 1.97e-07 ***
x3          -0.11134    0.35687  -0.312    0.757
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.236 on 44 degrees of freedom
Multiple R-squared: 0.5899,     Adjusted R-squared: 0.5619
F-statistic: 21.09 on 3 and 44 DF,  p-value: 1.279e-08
```

Sekce `Call` uvádí způsob, jakým byl model vytvořen. Sekce `Residuals` uvádí základní distribuční vlastnosti reziduí. Ty představují rozdíl mezi hodnotou, která byla pro dané pozorování vysvětlované proměnné (tj. první hlavní komponenty) změřena a hodnotou, kterou odhaduje regresní model. V sekci `Coefficients` ve sloupci `Estimate` můžeme opět vidět odhady regresních parametrů. Po dosazení odhadnutých parametrů do regresního modelu (3.1.1) dostáváme

$$Y_{1i} = -5{,}3517 + 0{,}3914x_{i1} + 1{,}3458x_{i2} - 0{,}1113x_{i3} + E_i \qquad i = 1, \ldots, 48.$$

Hodnoty ve výstupu `summary` označené jako `Signif. codes` vyjadřují statistickou signifikanci jednotlivých proměnných. Je-li tato hodnota u dané proměnné nižší než 0,05, znamená to, že je proměnná statisticky signifikantní, tj. má vliv na vysvětlovanou proměnnou. V sekci `Coefficients` ve sloupci `Pr(>|t|)` jsou uvedeny p-hodnoty jednotlivých regresních parametrů. Testujeme nulovou hypotézu H_0, která říká, že regresní parametr β_j, $j = 1, 2, 3$ je nulový, tzn. j-tá proměnná nemá vliv na vysvětlovanou proměnnou. Je-li příslušná p-hodnota menší než 0,05, tj. zamítáme-li nulovou hypotézu na hladině významnosti 0,05, říkáme, že j-tá proměnná statisticky významně souvisí s vysvětlovanou proměnnou. Z výše uvedeného výstupu příkazu `summary` vyplývá, že na hlavní komponentu nejsilněji působí počet pastevních cyklů, jelikož má největší odhad regresního parametru ($\beta_2 = 1{,}3458$), dále na hladině významnosti 0,05 zamítáme nulovou

hypotézu H_0: $\beta_2 = 0$ a signifikance této proměnné je nulová. Platí tedy, že s rostoucím počtem pastevních cyklů roste i hodnota hlavní komponenty, tj. zvyšuje se obsah tuku a popela a snižuje se množství vlákniny, NDF a ADF v píci. Stejným způsobem ovlivňuje hlavní komponentu rok, ve kterém byla analýza složení píce provedena, tento vliv ale není tak silný. Tím jsme potvrdili naši domněnku, z kapitoly o analýze rozptylových diagramů komponentních skóre, že větší vliv, než rok analýzy, má na první hlavní komponentu proměnná počet pastevních cyklů. Jelikož u proměnné hnojení nulovou hypotézu H_0: $\beta_3 = 0$ nezamítáme a také platí, že je signifikance vyšší než 0,05, nemá hnojení žádný vliv na první hlavní komponentu, což potvrzuje výsledky získané v předchozích kapitolách. Další důležitou částí výstupu `summary` je hodnota označená jako `Multiple R-squared`, která představuje takzvaný koeficient determinace, který může nabývat hodnot od nuly do jedničky. V našem případě nabývá hodnoty 0,5899, to znamená, že proměnné rok, počet cyklů a hnojení vysvětlují 58,99% rozptylu první hlavní komponenty, což je docela hodně.

Jelikož se neprokázala závislost hnojení na první hlavní komponentě, zkusíme zjednodušit model (3.1.1) tím způsobem, že vypustíme z regrese proměnnou hnojení:

```
> regrese2=lm(y~x1+x2)
```

Následně porovnáme tyto dva modely pomocí příkazu:

```
> anova(regrese1,regrese2)

Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3
Model 2: y ~ x1 + x2
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     44 67.243
2     45 67.392 -1  -0.14877 0.0973 0.7565
```

Ve výsledné tabulce jsou řádky s uvedenými čísly 1 a 2, které odkazují na dva srovnávané modely. Je zřejmé, že odebráním proměnné hnojení se zvýšila hodnota RSS (reziduální součet čtverců) z 67,243 na 67,392 (tj. o 0,14877, jak tabulka také ukazuje). Nárůst RSS je malý, druhý model je však jednodušší. Největší díl variability první hlavní komponenty lze tedy vysvětlit jen pomocí roku měření (x_{i1}) a počtu pastevních cyklů (x_{i2}).

Doposud jsme aplikovali model mnohonásobné lineární regrese na první hlavní komponentu. Dále se pokusíme studovat vztah mezi závisle proměnnou, kterou nyní bude druhá hlavní komponenta a stejnou skupinou nezávisle proměnných, kterými tedy jsou rok (x_{i1}), počet pastevních cyklů za rok (x_{i2}) a hnojení (x_{i3}). Lineární regresní model vypadá následovně:

$$Y_{2i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + E_i \qquad i = 1, \ldots, 48, \qquad (3.1.2)$$

kde nekorelované náhodné veličiny $Y_{21}, \ldots, Y_{2i}$, $i = 1, \ldots, 48$ jsou hodnoty druhé hlavní komponenty, kterou jsme si vyjádřili na začátku příkladu ve tvaru

$$Y_2 = 0{,}1495464z_1 + 0{,}1070360z_2 + 0{,}3815879z_3 + 0{,}1212693z_4 + 0{,}8976991z_5.$$

V softwaru R postupujeme analogicky, pouze místo hodnot první hlavní komponenty načteme hodnoty druhé hlavní komponenty. Nejdříve odhadneme regresní parametry:

```
> regrese1=lm(y~x1+x2+x3)
> regrese1

Call:
lm(formula = y ~ x1 + x2 + x3)

Coefficients:
(Intercept)           x1           x2           x3
    0.57978      0.06618     -0.37196      0.60894
```

Po dosazení odhadnutých parametrů do regresního modelu (3.1.2) dostáváme

$$Y_{2i} = 0{,}57978 + 0{,}06618x_{i1} - 0{,}37196x_{i2} + 0{,}60894x_{i3} + E_i \qquad i = 1, \ldots, 48.$$

Další základní informace o modelu získáme použitím funkce `summary` na objekt představující regresní model:

```
> summary(regrese1)

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6122 -0.5451 -0.1943  0.4880  2.5713

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.57978    0.50108   1.157   0.2535
x1           0.06618    0.05179   1.278   0.2080
x2          -0.37196    0.14532  -2.559   0.0140 *
x3           0.60894    0.23731   2.566   0.0138 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.8221 on 44 degrees of freedom
Multiple R-squared: 0.2513,     Adjusted R-squared: 0.2002
F-statistic: 4.923 on 3 and 44 DF,  p-value: 0.004917
```

Z výše uvedeného výstupu příkazu `summary` vyplývá, že na druhou hlavní komponentu nejsilněji působí proměnná hnojení, protože má největší odhad regresního parametru (β_3 = 0,60894), dále nulovou hypotézu H_0: $\beta_3 = 0$ na hladině významnosti 0,05 zamítáme a signifikance je menší než 0,05. Platí tedy, že hnojený travní porost způsobuje nárůst hodnot druhé hlavní komponenty, při kterém roste množství všech složek obsažených v píci, nejvíce však množství popela a také množství ADF. Druhou hlavní komponentu také ovlivňuje počet pastevních cyklů, tento vliv je ale opačný. Jelikož je odhad regresního parametru β_2 roven -0,37196, platí, že s rostoucím počtem pastevních cyklů klesá hodnota druhé hlavní komponenty. Dále si můžeme všimnout, že u proměnné rok nulovou hypotézu nezamítáme a signifikance je vyšší než 0,05. Rok, ve kterém bylo provedeno měření, tedy celkově nemá žádný vliv na složení píce reprezentované druhou hlavní komponentou, což vyvrací naši domněnku u analýzy grafu Scatterplot na obrázku 3.1, kde jsme konstatovali, že rok do jisté míry druhou komponentu ovlivňuje. Koeficient determinace, ve výstupu funkce `summary`, nabývá hodnoty 0,2513, to znamená, že proměnné rok, počet cyklů a hnojení vysvětlují 25,13% rozptylu druhé hlavní komponenty.

Jelikož se neprokázala závislost roku měření na druhé hlavní komponentě, zkusíme zjednodušit model (3.1.2), tím způsobem, že vypustíme z regrese proměnnou rok a tento nově vytvořený model porovnáme s původním:

```
> regrese2=lm(y~x2+x3)
> anova(regrese1,regrese2)

Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3
Model 2: y ~ x2 + x3
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     44 29.736
2     45 30.840 -1   -1.1037 1.6331  0.208
```

Je zřejmé, že odebráním proměnné rok se zvýšila hodnota RSS z 29,736 na 30,840. Tento nárůst je malý, můžeme tedy nahradit původní model modelem jednodušším. Tím vysvětlíme největší díl variability druhé hlavní komponenty jen pomocí počtu pastevních cyklů (x_{i2}) a hnojení (x_{i3}).

3.2 Užití PCA v rozpoznání obrazu

V následujícím příkladu máme k dispozici obrázek 3.5. Jedná se o barevný RGB obrázek, u něhož je barva každého obrazového bodu (pixelu) určena kombinací hodnot třech základních barev, tedy červené (Red), zelené (Green) a modré (Blue). Mícháním těchto základních barev získáme spektrum všech barev. Intenzita každé barvy se nejčastěji udává binárně určitým počtem bitů podle použité barevné hloubky. Nejčastěji se používá 8 bitů na každou základní barvu, přičemž pro 8 bitů je rozsah hodnot 0 – 255 (čím vyšší hodnota, tím vyšší intenzita zobrazované barvy). Jeden pixel tedy potřebuje pro zobrazení 8 x 3 = 24 bitů. Čím větší je výsledný součet hodnot (0 – 255) pro 24 bitů, tím světlejší je výsledná barva. Pro ukázku jsou v tabulce 3.1 uvedené základní směsi barev.

Obrázek 3.5: Vstupní obrázek č. 1

Tabulka 3.1: Základní kombinace barev (upraveno dle [19])

R	G	B	výsledná barva
0	0	0	černá
255	0	0	červená
0	255	0	zelená
0	0	255	modrá
255	255	0	žlutá
255	0	255	purpurová
0	255	255	azurová
255	255	255	bílá
100	100	100	tmavě šedá

Úkolem je zjistit, jak se při převodu do obrazové podoby projeví redukce datového souboru tvořeného třemi náhodnými veličinami (Red, Green, Blue) na jedinou proměnnou, tj. první hlavní komponentu.

Datový soubor získáme vygenerováním obrazových bodů pomocí softwaru Mathematica. Pro názornou ukázku, uvedu použitý postup i v tomto softwaru.

Nejprve načteme obrázek 3.5, se kterým budeme pracovat, pomocí příkazu:

```
obr=ImageData[ ]
```


Software nám následně vypíše obrazové body znázorněné v tabulce 3.2. Jedná se o datový soubor, který obsahuje 28 203 pixelů, v němž je hodnota barev RGB rozmístěna v rozsahu 0 – 1. Rozměr obrázku 3.5 zjistíme příkazem:

```
Dimensions[obr]
{119, 237, 3}
```

Výstup v závorce {} udává rozměr 119 x 237 pixelů.

Tabulka 3.2: Vstupní obrázek č. 1 vyjádřený pomocí pixelů

	1	2	…	237
1	(0.529412, 0.862745, 0.843137)	(0.509804, 0.843137, 0.823529)	…	(0.513725, 0.847059, 0.827451)
2	(0.505882, 0.839216, 0.819608)	(0.498039, 0.831373, 0.811765)	…	(0.517647, 0.85098, 0.831373)
⋮	⋮	⋮	⋱	⋮
119	(0.545098, 0.858824, 0.847059)	(0.54902, 0.862745, 0.85098)	…	(0.541176, 0.854902, 0.843137)

Jednotlivé obrazové body této tabulky pro nás představují pozorování. Proměnnými jsou 3 základní barvy RGB. Abychom mohli z datového souboru v tabulce 3.2 vypočítat hodnoty první hlavní komponenty, musíme jej převést pomocí softwaru Mathematica do požadovaného formátu 3 x 28 203 a následně ho exportovat do souboru MS Excel. Pro převedení do požadovaného formátu použijeme příkaz:

```
data = Flatten[obr, 1];
```

Následně ověříme správný rozměr dat

```
Dimensions[data]
{28203, 3}
```

a data převedeme do souboru s názvem `obrazekRGB.xlsx` příkazem:

```
Export["obrazekRGB.xlsx ", data]
```

Tímto získáme zdrojový soubor dat, ve kterém jsou sloupce tvořeny proměnnými RGB a řádky jednotlivými pozorováními, tj. pixely.

V softwaru R vypočítáme první hlavní komponentu. Postup je analogický jako v předchozím příkladu. Při výpočtu nesmíme zapomenout na normalizaci veličin RGB v souboru `obrazekRGB.xlsx`. Výsledná hlavní komponenta má následující tvar:

$$Y_1 = -0{,}5399864z_1 - 0{,}6087001z_2 - 0{,}5812906z_3$$

Po dosazení normovaných veličin (z_1, z_2, z_3) získáme hodnoty první hlavní komponenty, které uložíme do souboru MS Excel pod názvem `PC1.xlsx`.

Nyní opět použijeme software Mathematica, do kterého importujeme výsledné hodnoty první hlavní komponenty. Pro import dat zadáme příkaz:

```
x = Import["Cesta k datovému souboru/PC1.xlsx", "Data"];
```

Zdrojová data musíme pro zpětný převod do RGB obrázku standardizovat tak, abychom získali rozsah hodnot pro jednotlivé barvy RGB na intervalu 0 - 255. Pro standardizaci jednotlivých hodnot první hlavní komponenty aplikujeme následující postup:

```
z = x/(Max[x] - Min[x]);
z = z - Min [z];
z2 = Floor[255*z];
```

Pro ověření rozměru zdrojových dat opět použijeme příkaz:

```
Dimensions[x]
```

```
{1, 28203, 1}
```

Výstup udává, že nový soubor dat, tvořený hodnotami první hlavní komponenty, má rozměr 1 x 28 203. Pro vykreslení do obrazové podoby musíme data naformátovat do původního rozměru obrázku (119 x 237) následovně:

```
z2 = z2[[1]];
z2 = Partition[z2, 237];
Dimensions[z2]
{119, 237, 1}
```

Na závěr takto převedená data vykreslíme do grafické podoby pomocí příkazu:

```
Image[z2, "Byte"]
```

Tímto získáme výstup obsahující informaci o první hlavní komponentě, který je uvedený na obrázku 3.6.

Obrázek 3.6: Grafický výstup první hlavní komponenty

Z výsledného výstupu můžeme vidět, že se původní obrázek 3.5 převedl do černobílé podoby. Zjistili jsme tedy, že se dá první hlavní komponenta použít ke kompresi dat z podoby barevného obrázku na černobílý.

Nyní převedeme obrázek 3.5 do standardní podoby černobílého obrázku. Ke zjištění zastoupení jednotlivých barev RGB pro standardní přepis do stupňů šedi můžeme použít například programové prostředí Matlab, ve kterém pomocí příkazu `RGB2gray` získáme následující výstup v podobě lineární kombinace barev RGB:

$$0{,}2989\ R + 0{,}587\ G + 0{,}114\ B \qquad (3.2.1)$$

Po dosazení původních hodnot RGB uvedených v tabulce 3.2 do lineární kombinace (3.2.1) získáme datový soubor, který importujeme do softwaru Mathematica. Postupujeme stejným způsobem jako při importu první hlavní komponenty. Výsledkem bude grafický výstup hodnot RGB, vypočítaných ze vztahu (3.2.1), který je uvedený na obrázku 3.7.

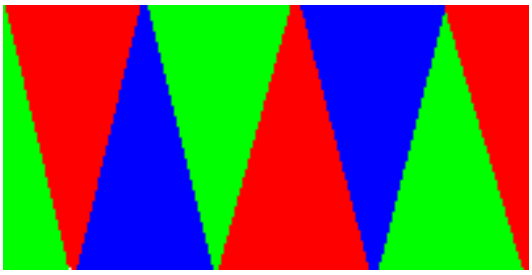
Obrázek 3.7: Standardní převod do černobílé podoby

Při porovnání obrázků získaných z první hlavní komponenty (3.6) a z lineární kombinace pro standardní převod do stupňů šedi (3.7) můžeme konstatovat, že výsledné

výstupy jsou si velice podobné. V tomto případě tedy první hlavní komponenta pracuje obdobně jako algoritmus (3.2.1).

Pro další analýzu použití první hlavní komponenty ke kompresi dat z podoby barevného obrázku na černobílý jsme vytvořili obrázek 3.8 obsahující pouze základní barvy, tj. červenou, zelenou a modrou.

Obrázek 3.8: Vstupní obrázek č. 2



Na barevný obrázek 3.8 aplikujeme celý výše uvedený postup, zkráceně tedy vygenerujeme hodnoty RGB a získáme datový soubor se třemi proměnnými, ze kterých vypočítáme první hlavní komponentu. Nově vzniklá proměnná má tvar:

$$Y_1 = 0{,}7601257z_1 - 0{,}6315112z_2 - 0{,}1529788z_3$$

Hodnoty první hlavní komponenty importujeme do softwaru Mathematica, pomocí kterého následně získáme grafický výstup znázorněný na obrázku 3.9.

Obrázek 3.9: Grafický výstup první hlavní komponenty

Výsledek potvrzuje, že metodu hlavních komponent můžeme použít pro převedení dat z barevného obrázku na jeho černobílou podobu. Je zajímavé, že hlavní komponenta aplikovaná na obrázek 3.8 převedla zelenou a červenou barvu do barvy bílé a černé, přičemž redukce hodnot RGB definujících modrou barvu způsobila převod této barvy na středně šedou.

Nyní převedeme obrázek 3.8 do černobílé podoby standardně, tj. pomocí algoritmu (3.2.1). Grafický výstup je znázorněn na obrázku 3.10.

Obrázek 3.10: Standardní převod vstupního obrázku č. 2 do černobílé podoby

Při porovnání obrázků 3.9 a 3.10 můžeme vidět, že zredukované proměnné RGB do první hlavní komponenty v tomto případě sice přemění barevný obraz na obraz v různých stupních šedi, převod se ale velmi liší od standardního převodu do černobílé podoby. První hlavní komponentu tedy můžeme použít k přepisu barevného obrazu do černobílého, převod bude ale u jednotlivých případů proveden jiným způsobem z důvodu rozdílných výsledků první hlavní komponenty. Pokud budeme mít k dispozici dva odlišné obrázky, u nichž se vyskytuje stejná barva, bude tato barva pomocí první hlavní komponenty převedena u každého obrázku do jiného stupně šedi.

Dále uvažujme, co se stane, nahradíme-li původní proměnné RGB tvořící barevný obrázek 3.8 pouze jednou z těchto proměnných, například Red, a ostatní proměnné vynecháme. Datový soubor bude tedy tvořen pouze hodnotami proměnné Red. Po importu tohoto souboru a následném převodu do grafické podoby získáme obrázek 3.11.

Obrázek 3.11: Grafický výstup proměnné Red

Na obrázku si můžeme všimnout, že software Mathematica převedl útvary v červené barvě do barvy černé. Proměnné Green a Blue, které byly vynechány, se v obrázku nijak neprojevily, tudíž oblasti vynechaných proměnných se zobrazily bíle. Stejným způsobem program postupuje, nahradíme-li původní soubor RGB pouze proměnnou Green (resp. Blue), kdy útvary v obrázku 3.8 v zelené (resp. modré) barvě převede do černé barvy a zbylé dvě barvy vynechá.

3.3 Religiozita

K dalšímu šetření nás inspirovala bakalářská práce Jany Dvořákové s názvem „*Víc než mrtví: Analýza postoje veřejnosti k pacientům v permanentním vegetativním stavu“* [21]. Autorka se ve své práci, mimo jiné, zabývá otázkou, jak jsou lidé v ČR nábožensky či duchovně založení. Pro zjištění tohoto faktu vytvořila dotazník, obsahující čtyři otázky:

č. 1: Jsem duchovně založený člověk?

č. 2: Jsem věřící?

č. 3: Věřím na posmrtný život?

č. 4: Duše žije i poté, co člověk zemře?

Na tyto otázky reagovalo 206 respondentů, kteří byli náhodně vybráni na ulici, v knihovnách či na fakultách v Olomouci. Jednotliví účastníci ankety odpovídali dle svého názoru na škále od -3 (zcela nesouhlasím) do 3 (zcela souhlasím). Pro vyjádření náboženského založení vybraného vzorku populace použila autorka práce tzv. index religiozity, který u každého respondenta vytvořila jako aritmetický průměr odpovědí na čtyři otázky.

V našem případě se nebudeme zabývat otázkou náboženského založení, ale pokusíme se na soubor dat, tvořený odpověďmi na výše uvedené otázky, aplikovat metodu hlavních komponent. Cílem bude zjistit, jak by mohla první hlavní komponenta reprezentovat danou kombinaci odpovědí každého respondenta a označíme ji jako nový index religiozity.

Nejdříve vypočítáme první hlavní komponentu a pomocí grafu ji porovnáme s původním indexem religiozity, tj. průměrem. Při výpočtu hlavní komponenty postupujeme stejně jako v předchozích příkladech. Nejprve datový soubor znormujeme a takto upravená data načteme v softwaru R, poté zadáme příkaz pro výpočet kovarianční matice normovaných veličin:

```
> H=read.csv("normovane_veliciny.csv", header = FALSE, sep = ";", dec=",", fill = TRUE)


> var(H)
          V1        V2        V3        V4
V1 1.0000000 0.7209401 0.5991575 0.6153421
V2 0.7209401 1.0000000 0.5827607 0.5910287
V3 0.5991575 0.5827607 1.0000000 0.8434365
V4 0.6153421 0.5910287 0.8434365 1.0000000
```

Z výsledné matice můžeme vidět, že korelace mezi veličinami jsou poměrně velké, což se projevuje na způsobu, jakým respondenti odpovídali na jednotlivé otázky. Největší korelace jsou mezi první a druhou otázkou (0,72) a dále mezi třetí a čtvrtou otázkou (0,84). Respondenti mají tendenci odpovídat na tyto dvojice otázek obdobně. Což může znamenat, že jsou si otázky v rámci těchto dvojic významově podobné. Zbylé korelace mají také poměrně vysokou hodnotu. U většiny respondentů tedy platí, že pokud se rozhodnou odpovídat kladně či záporně na jednu otázku, odpoví tak i na ostatní otázky.

Dále vyjádříme vlastní čísla a jim odpovídající vlastní vektory pomocí příkazů:

```
> eigen(var(H))$values
[1] 2.9780084 0.5872868 0.2784935 0.1562113

> eigen(var(H))$vectors
           [,1]       [,2]       [,3]       [,4]
[1,] -0.4908751  0.4801863 -0.7260239 -0.0367688
[2,] -0.4831362  0.5440535  0.6858908  0.0117875
[3,] -0.5105303 -0.5005668  0.0494245 -0.6973871
[4,] -0.5147619 -0.4720803 -0.0004358  0.7156537
```

Výsledná první hlavní komponenta (tj. nový index religiozity) má tvar:

$$Y_1 = -0{,}4908751z_1 - 0{,}4831362z_2 - 0{,}5105303z_3 - 0{,}5147619z_4 \qquad (3.3.1)$$

Je patrné, že všechny otázky mají přibližně stejnou váhu, tudíž aritmetický průměr, v práci Jany Dvořákové, je dobře zvoleným indexem religiozity.

Pomocí vlastního vektoru, uvedeného ve sloupci `[,2]` ve výstupu funkce `eigen(var(H))$vectors`, vyjádříme druhou hlavní komponentu:

$$Y_2 = 0{,}4801863z_1 + 0{,}5440535z_2 - 0{,}5005668z_3 - 0{,}4720803z_4.$$

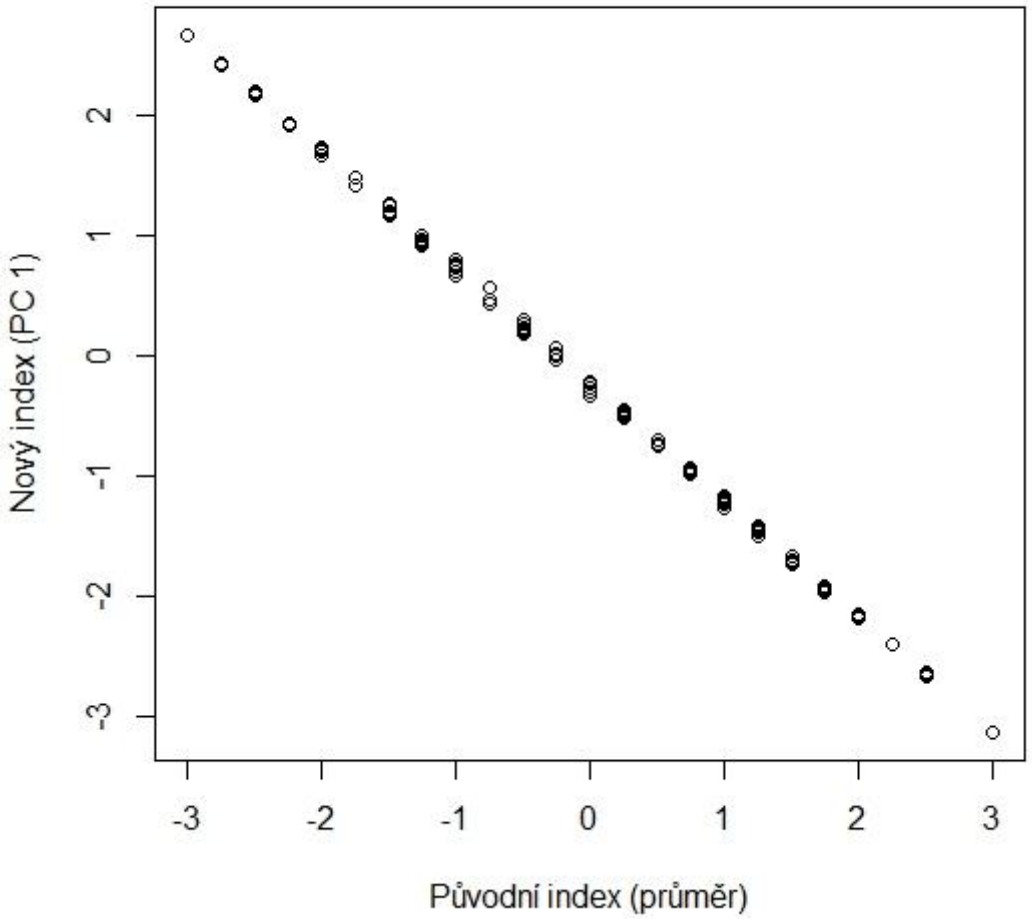
Tato komponenta nese jemnější informaci o variabilitě původních proměnných, kterými jsou uvedené otázky. Je zřejmé, že první dvě složky vlastního vektoru jsou kladné a druhé dvě složky záporné. To znamená, že druhá hlavní komponenta rozděluje otázky na dvě dvojice, přičemž provázanost otázek v těchto dvojicích je těsná, což dokazují i zmíněné hodnoty korelačních koeficientů. Respondenti mají tendenci odpovídat podobně v rámci těchto dvojic otázek a přitom odlišně mezi dvojicemi otázek. Jinak řečeno, respondenti přiřazující k prvním dvěma otázkám kladnou odpověď (např. „souhlasím“) odpovídají na další dvě otázky záporně („nesouhlasím“) a naopak.

Nyní porovnáme průměr odpovědí na dané otázky s hodnotami první hlavní komponenty. Vytvoříme graf, ve kterém budou na ose *x* hodnoty původního a na ose *y* hodnoty nového indexu religiozity, a to pomocí příkazu uvedeného na následující straně.

```
> plot(x,y,type="p",xlab="Původní index(průměr)",ylab="Nový index(PC 1)")
```

V grafickém výstupu tohoto příkazu na obrázku 3.12 jsou pomocí bodů znázorněny oba indexy religiozity jednotlivých respondentů. Můžeme si všimnout, že i přes to, že někteří respondenti mají stejný průměr odpovědí, hodnota hlavní komponenty je u nich různá, to dokazuje i tabulka 3.3 uvedená na str. 59. Z toho vyplývá, že aritmetický průměr nedává přesnou informaci o tom, jak daní respondenti odpovídali na jednotlivé otázky.

Obrázek 3.12: Srovnání indexů religiozity



Dále se pokusíme zjistit, zda lze pomocí nového indexu religiozity jednoznačně určit, jak respondenti odpovídali na uvedené otázky, tj. zda existuje jednoznačný vztah mezi první hlavní komponentou a odpověďmi jednotlivých respondentů. Pro zjištění tohoto jednoznačného vztahu vytvoříme matici M, která bude obsahovat všechny možné

kombinace odpovědí na uvedené otázky, přičemž každá kombinace se bude v matici vyskytovat právě jednou. Jelikož máme k dispozici 4 otázky s odpověďmi na škále -3 až 3, existuje celkem 7^4 = 2 401 různých kombinací odpovědí. Matice bude obsahovat 4 sloupce a 2 401 řádků a získáme ji v softwaru R postupným definováním jednotlivých sloupců následovně:

```
> M=matrix(nrow=2401,ncol=4)
> w=(-3):3
> M[,1]=rep(w,each=2401/7)
> M[,2]=rep(rep(w,each=49),7)
> M[,3]=rep(rep(w,each=7),49)
> M[,4]=rep(w,2401/7)
```

Výslednou matici M si můžeme představit jako skupinu respondentů, v níž každý odpověděl na uvedené 4 otázky jinak. Tuto matici budeme nyní považovat za nová zdrojová data. V dalším kroku do lineární kombinace (3.3.1) dosadíme hodnoty vygenerované matice M. Můžeme postupovat tak, že nejprve označíme vlastní vektor odpovídající největšímu vlastnímu číslu

```
> A=eigen(R)$vectors
> a=A[,1]
> a
[1] -0.4908751 -0.4831362 -0.5105303 -0.5147619
```

Následně dle (3.3.1) vypočítáme hodnoty první hlavní komponenty pro nová data:

```
> PC1=a[1]*M[,1]+a[2]*M[,2]+a[3]*M[,3]+a[4]*M[,4]
```

Pro zjištění jednoznačného vztahu mezi hlavní komponentou a danou kombinací odpovědí na 4 otázky použijeme příkaz:

```
> length(levels(as.factor(PC1)))
[1] 2401
```

Výsledek říká, že počet odlišných hodnot první hlavní komponenty (získaných postupným dosazením jednotlivých kombinací odpovědí) odpovídá počtu řádků matice M. To znamená, že každé kombinaci odpovědí přísluší jiná hodnota hlavní komponenty. Tímto jsme dokázali, že mezi hodnotou hlavní komponenty a danou kombinací odpovědí respondenta existuje jednoznačný vztah. První hlavní komponentu, tj. nový index religiozity, můžeme považovat za číselnou charakteristiku, jejíž hodnota přesně reprezentuje kombinaci odpovědí každého respondenta. V případě, že se v původním

souboru dat vyskytují respondenti se stejnou hodnotou první hlavní komponenty, znamená to, že odpověděli na výše uvedené otázky shodně. Tabulka 3.3 uvádí část respondentů z původního datového souboru, na níž je názorně vidět vztah, který jsme dokázali. Můžeme si všimnout, že u respondentů 16 a 17 je hodnota první hlavní komponenty stejná, tudíž odpověděli na čtveřici otázek shodně, tím pádem je i hodnota průměru z těchto odpovědí stejná. Podíváme-li se na sloupec s průměry odpovědí, přestože mají například respondenti 17, 26 a 28 stejnou hodnotu průměru, mají jinou hodnotu hlavní komponenty, tudíž odpověděli na dané otázky jinak. Nový index religiozity, na rozdíl od aritmetického průměru, tedy podává přesnou informaci o tom, jak účastníci ankety reagovali na jednotlivé otázky.

Tabulka 3.3: Ukázka původního datového souboru

Respondenti	Otázka č. 1	Otázka č. 2	Otázka č. 3	Otázka č. 4	Průměr	PC1
16	1	-2	-2	-2	-1,25	0,991756165
17	1	-2	-2	-2	-1,25	0,991756165
26	-3	-2	0	0	-1,25	0,937353981
27	-3	-3	-2	-2	-2,5	2,166778873
28	-2	1	-2	-2	-1,25	1,005557262
47	1	-2	0	2	0,25	-0,506417885
49	1	0	0	0	0,25	-0,468992976
63	3	3	3	3	3	-3,131965917
142	-1	2	0	0	0,25	-0,459792245
187	-3	-3	-3	-3	-3	2,665829195

Při podrobnější analýze získaných hodnot první hlavní komponenty v souboru všech 206 respondentů jsme zjistili, že k největší shodě došlo u 17 respondentů, u nichž je hodnota hlavní komponenty 2,665829165. Tato hodnota říká, že účastníci ankety přiřadili ke všem otázkám odpovědi s číslem -3, tj. zcela nesouhlasím. Naopak 13 respondentů, z celého souboru, reagovalo na čtveřici otázek kladně číslem 3, tj. zcela souhlasím. Shoda se projevila i u 6 respondentů odpovídající na všechny otázky číslem 2 (souhlasím) a dále u 5 respondentů, kteří odpovídali pouze číslem -2 (nesouhlasím). Dalších 7 účastníků ankety odpovědělo jednotně na všechny otázky, a to buď číslem -1 (mírně nesouhlasím), 0 (nemám názor) nebo 1 (mírně souhlasím). Z výše uvedeného vyplývá, že celkem 48 respondentů ke každé otázce ze čtveřice vždy přiřadily stejnou hodnotu ze škály -3 až 3. Potvrzuje se tedy, že jsou si uvedené otázky významově podobné a u většiny respondentů platí, že pokud se rozhodnou odpovídat kladně či záporně na jednu otázku, odpoví stejným

způsobem i na ostatní otázky. Další shody se objevují nejčastěji u dvou až čtyř respondentů a někteří se svými názory od ostatních zcela liší.

Závěr

Cílem této práce bylo seznámit se s metodou hlavních komponent a aplikovat ji na reálná data. Při psaní této práce jsem se nejdříve věnovala vysvětlení důležitých pojmů souvisejících s touto metodou. Poté jsem se zaměřila na teoretickou stránku metody hlavních komponent a nakonec jsem získané poznatky využila pro aplikaci metody na reálné příklady. Pro použití metody hlavních komponent v praktické části práce jsme hledali vhodné příklady, které budou odlišné a zajímavé nejen z matematického, ale také interpretačního hlediska.

V prvním příkladu jsme zkoumali vliv tří faktorů na složení píce, konkrétně vliv roku, ve kterém byla provedena analýza složení píce, vliv hnojení a počtu pastevních cyklů za rok. Složení píce bylo charakterizováno pomocí pěti složek v různých množstvích, tj. pomocí vlákniny [$g.kg^{-1}$ sušiny], NDF [$g.kg^{-1}$ DM], ADF [$g.kg^{-1}$ DM], tuku [$g.kg^{-1}$ sušiny] a popela [$g.kg^{-1}$ sušiny]. Abychom mohli vliv jednotlivých faktorů podrobněji zkoumat, zjednodušili jsme složení píce tak, že jsme pětici složek nahradili dvěma veličinami, tj. hlavními komponentami. Obě hlavní komponenty jsme interpretovali jako danou lineární kombinaci složek píce. Přičemž pro první hlavní komponentu platí, že pokud klesá množství vlákniny, NDF a ADF nebo roste množství tuku a popela obsaženého v píci, roste hodnota této komponenty. Menší informaci o variabilitě složek v píci nese druhá hlavní komponenta, pro kterou platí, že pokud se zvyšuje množství všech složek v píci, nejvíce však množství popela a ADF, roste hodnota druhé hlavní komponenty. Dále jsme příklad rozdělili na tři části, ve kterých jsme podrobně zkoumali vliv jednotlivých faktorů různými způsoby. V první části jsme všechna pozorování vykreslili pomocí softwaru R do rozptylových diagramů komponentních skóre (Scatterplot) a postupně odlišili tato pozorování podle jednotlivých faktorů. Z pozorování rozlišených podle roku analýzy jsme zjistili rostoucí tendenci první hlavní komponenty. Při rozlišení pozorování podle hnojení jsme došli k závěru, že tento faktor nemá vliv na první komponentu, ale může mít vliv na druhou hlavní komponentu, a to takový, že při aplikaci hnojiva hodnota druhé hlavní komponenty roste. U posledního faktoru, tj. počtu pastevních cyklů, bylo z grafu zřejmé, že zvyšuje-li se počet pastevních cyklů, roste také hodnota první komponenty. Při pohledu na druhou hlavní komponentu byla patrná její opačná tendence.

V další části příkladu jsme analyzovali vzájemné vztahy mezi původními pěti složkami píce pomocí dvojného grafu (Biplot). Zjistili jsme, že složky NDF, vláknina

a ADF jsou vzájemně silně korelované a také, že množství popela v píci nezávisí na množství ostatních složek. Pozorování, u nichž se v píci vyskytuje vysoký podíl tuku, obsahují malé množství NDF, vlákniny a ADF či naopak. Pomocí biplotu jsme potvrdili zjištěný vliv tří faktorů na složení píce z první části příkladu, platí tedy, že s rostoucím počtem pastevních cyklů a rokem měření roste hodnota první hlavní komponenty, tj. zvyšuje se obsah tuku a popela a snižuje se množství vlákniny, NDF a ADF v píci. Dále platí, že hnojení nezávisí na první hlavní komponentě, ale odrazí se v poměru složek druhé hlavní komponenty, konkrétně v množství popela. Při aplikaci hnojiva tedy z velké části roste množství popela obsaženého v píci.

V poslední části příkladu jsme pro zjištění vlivu uvedených faktorů použili model mnohonásobné lineární regrese. Model jsme aplikovali nejdříve na první hlavní komponentu a výsledky nám opět potvrdili, že roste-li počet pastevních cyklů, roste i hodnota hlavní komponenty. Stejným způsobem ovlivňuje první hlavní komponentu rok, ve kterém byla analýza složení píce provedena, tento vliv ale není tak silný. Hnojení na první hlavní komponentu nemá žádný vliv. Při aplikaci modelu mnohonásobné lineární regrese na druhou hlavní komponentu jsme zjistili, že píce, u které bylo aplikováno hnojivo, způsobuje nárůst hodnot této komponenty, při kterém roste množství všech složek obsažených v píci, nejvíce však množství popela a také množství ADF. Druhou hlavní komponentu také ovlivňuje počet pastevních cyklů, tento vliv je ale opačný. S rostoucím počtem pastevních cyklů tedy klesá hodnota druhé komponenty. Rok, ve kterém byla provedena analýza složení píce, nemá vliv na druhou hlavní komponentou.

Na druhém příkladu jsme si ukázali další využití metody hlavních komponent. Pokusili jsme se aplikovat tuto metodu na barevný RGB obrázek, u něhož je barva každého obrazového bodu určena kombinací hodnot třech základních barev, tedy červené, zelené a modré. Úkolem bylo zjistit, jak se při převodu do obrazové podoby projeví redukce datového souboru tvořeného třemi náhodnými veličinami (Red, Green, Blue) na jedinou proměnnou, tj. první hlavní komponentu. Pro vygenerování obrazových bodů barevného obrázku a zpětný převod vypočítaných hodnot první komponenty do grafické podoby jsme použili software Mathematica. Výsledný grafický výstup nám ukázal, že se první hlavní komponenta dá použít ke kompresi dat z podoby barevného obrázku na černobílý. Při porovnání tohoto výstupu se standardním převodem do černobílé podoby jsme došli k závěru, že první hlavní komponenta pracuje oproti standardnímu převodu odlišně, přičemž každý převod pomocí hlavní komponenty je u jednotlivých případů proveden do jiných stupňů šedi, z důvodu rozdílných výsledků redukce dat.

V posledním příkladu jsme se věnovali výsledkům dotazníkového šetření náboženského či duchovního založení obyvatel ČR, které provedla ve své bakalářské práci [21] Jana Dvořáková. Výsledky tohoto šetření byly tvořeny odpověďmi na čtyři otázky celkem 206 respondentů. V našem případě jsme se nezabývali otázkou náboženského založení účastníků ankety, cílem bylo zjistit, jak by mohla první hlavní komponenta reprezentovat danou kombinaci odpovědí každého respondenta. Zkoumali jsme, zda lze pomocí první hlavní komponenty jednoznačně určit, jak respondenti odpovídali na uvedené otázky. Došli jsme k závěru, že mezi hodnotou první hlavní komponenty a danou kombinací odpovědí respondenta existuje jednoznačný vztah. První hlavní komponentu tedy můžeme považovat za číselnou charakteristiku, jejíž hodnota přesně reprezentuje kombinaci odpovědí každého respondenta. Tento poznatek se dá obecně využít u jakéhokoliv dotazníkového šetření. V případě, že se ve výsledcích šetření vyskytují respondenti se stejnou hodnotou první hlavní komponenty, znamená to, že odpověděli na otázky uvedené v dotazníku shodně.

Těší mě, že jsem zvolila právě toto téma diplomové práce a prohloubila své vědomosti v oblasti mnohorozměrné statistické analýzy. Také jsem se přesvědčila, že využití metody hlavních komponent v praxi je velmi rozmanité. Doufám, že se mi podařilo vytvořit ucelený pohled na tuto metodu, který bude přínosný i pro další zájemce o danou problematiku.

Příloha A

Normované složky píce: vláknina (z_1), NDF (z_2), ADF (z_3), tuk (z_4), popel (z_5).

	z_1	z_2	z_3	z_4	z_5
1	0,7211849	0,3909822	1,5246919	-1,2781912	-0,5707546
2	0,3626524	0,3470353	1,1813909	-1,4013656	0,1463860
3	0,8631040	1,2006804	0,7470752	-1,6107621	-1,3530898
4	0,9751454	1,2477173	1,2488092	-1,3890481	-0,5272915
5	1,1600137	1,1137022	1,5912232	-1,5614923	-0,2592693
6	1,1656158	1,3860813	1,5042890	-1,5614923	-0,9764099
7	0,1348348	0,1134531	-0,0764924	-0,4898749	-0,8025576
8	-0,0350946	0,1889868	-0,0782666	-0,0833993	0,3057506
9	0,5157756	0,7028447	0,4788214	-1,4383179	-1,9760604
10	1,1562790	0,5655108	0,6438188	-1,7093016	-0,0491978
11	0,6801030	0,9889572	0,5978679	-1,4875877	-1,2009690
12	1,6529960	1,4433037	1,6142873	-0,9086679	-0,5562669
13	-1,4598879	-0,0982701	-0,8039003	-0,4406051	-0,5200477
14	-1,1685802	-1,1717640	-1,4975990	-0,0341296	-0,1361239
15	-0,9220891	-1,5528657	-1,2101842	0,9635832	-0,3896585
16	-0,9202218	-0,2504819	-0,8447061	-0,2065738	0,0087530
17	0,5643269	0,2965651	0,6846246	-0,4898749	-1,0053852
18	0,4410814	-0,1154368	0,5089822	-0,6623191	-0,1506116
19	-1,4113366	-1,4727542	-0,1279432	1,0128529	1,2909134
20	-1,1704476	-0,4404606	-0,6495479	0,4339332	0,1391422
21	-0,8361907	-0,5972502	-0,4011647	1,2592018	1,1315489
22	-0,7484250	-0,2642153	0,8762345	0,4955204	0,8707705
23	-0,0967174	-0,3580602	0,4717247	-0,3790179	2,8121208
24	-0,1079216	0,0093082	0,3439848	0,8034564	2,2905640
25	-0,6683153	-1,5047988	-1,5295339	1,1606622	-0,8387768
26	-1,2864105	-1,4578764	-1,7530788	1,2345669	-0,8677522
27	-0,2750500	-0,7116952	-0,8713185	0,9882181	-0,2447816
28	-0,0561958	-0,1566370	-0,4490672	0,7172343	0,2188245
29	1,0464784	0,9466125	0,4078548	0,3353937	-1,5052105
30	1,7263831	1,6939382	1,3676783	0,0644100	-0,8749960

	z_1	z_2	z_3	z_4	z_5
31	-1,1555087	-0,8902293	-1,3130857	0,1013623	-0,2013185
32	-1,0621409	-0,7849399	-1,3538916	0,9389483	0,1463860
33	-0,2087588	0,1123087	-0,7045470	0,0028227	0,0884353
34	-0,2909225	1,2510362	0,2517282	0,4708855	0,2840191
35	2,0115285	2,0052286	1,5646107	-0,7978109	-1,4110405
36	2,0992942	1,8060943	1,4368708	-0,9209854	-0,2230500
37	-1,2486899	-1,0773469	-0,7132404	-0,2435261	0,1101668
38	-1,0423469	-0,8414758	-0,7098695	-0,2804784	1,5516918
39	-0,8012712	-0,3652702	-1,1773621	0,3846634	-0,0419539
40	-0,4307876	-0,1249358	-0,7785297	0,3969809	0,5882605
41	0,3352022	-0,4260405	-0,2510703	-0,0833993	0,5592852
42	0,2913193	0,0812941	-0,1919906	0,2368541	1,7400318
43	-1,1069575	-1,5566424	-1,4363903	1,0867576	-0,2954885
44	-1,4094693	-1,8046447	-1,0015423	2,2569145	2,2543448
45	-0,4851276	-1,3678082	-0,4565187	1,5917727	0,7041620
46	0,4205404	0,0604651	-0,4123419	1,8504390	0,2260683
47	1,1432075	0,6339489	0,6180934	0,2861239	-0,3896585
48	0,9377983	0,8058452	1,1285208	0,3846634	-0,0999047

Příloha B

Vstupní data pro výpočet regrese.

i	Rok x_{i1}	Počet past. cyklů/rok x_{i2}	Hnojení x_{i3}	Y_{1i}	i	Rok x_{i1}	Počet past. cyklů/rok x_{i2}	Hnojení x_{i3}	Y_{1i}
1	0	4	0	-2,003164	25	4	4	0	2,084645
2	0	4	1	-1,485005	26	4	4	1	2,497288
3	0	3	0	-2,477862	27	4	3	0	1,270208
4	0	3	1	-2,464176	28	4	3	1	0,695486
5	0	2	0	-2,649479	29	4	2	0	-1,480965
6	0	2	1	-2,955033	30	4	2	1	-2,590947
7	1	4	0	-0,530194	31	5	4	0	1,637534
8	1	4	1	0,011716	32	5	4	1	2,019664
9	1	3	0	-2,031135	33	5	3	0	0,411823
10	1	3	1	-1,918377	34	5	3	1	-0,324196
11	1	2	0	-2,111691	35	5	2	0	-3,510590
12	1	2	1	-2,874601	36	5	2	1	-3,102250
13	2	4	0	0,817088	37	6	4	0	1,430785
14	2	4	1	1,835278	38	6	4	1	1,608105
15	2	3	0	2,125330	39	6	3	0	1,300306
16	2	3	1	0,900614	40	6	3	1	0,991304
17	2	2	0	-1,259317	41	6	2	0	0,295768
18	2	2	1	-0,733577	42	6	2	1	0,511675
19	3	4	0	2,312536	43	7	4	0	2,407649
20	3	4	1	1,340467	44	7	4	1	3,714230
21	3	3	0	1,777723	45	7	3	0	2,040431
22	3	3	1	0,547154	46	7	3	1	0,823793
23	3	2	0	0,649876	47	7	2	0	-1,171175
24	3	2	1	0,891036	48	7	2	1	-1,275782

Literatura

[1] Johnson, R. A., Wichern, D. W.: *Applied multivariate statistical analysis*. Sixth edition, Pearson education, New Jersey, 2007.

[2] Hebák, P. a kol.: *Vícerozměrné statistické metody [3]*. 2. doplň. vyd. Praha: Nakladatelství Informatorium, 2007.

[3] Meloun, M., Militký, J., Hill, M.: *Počítačová analýza vícerozměrných dat v příkladech*. Academia, Praha, 2005.

[4] Meloun, M., Militký, J.: *Analýza vícerozměrných dat*. [online], dostupné z: http://centrum.tul.cz/centrum/centrum/1Projektovani/1.2_publikace/[1.2.37].pdf, [cit. 2012-09-07].

[5] Tvrdík, J.: *Analýza vícerozměrných dat*. Ostravská univerzita, Přírodovědecká fakulta, 2003. [online], dostupné z: http://prf.osu.cz/doktorske_studium/dokumenty/Multivariable_Data_Analysis.pdf, [cit. 2012-09-07].

[6] Tonhauserová, Z., *Bakalářská práce: Statistické chyby v medicínském výzkumu*, Olomouc: UPOL, 2010.

[7] Matematika I, [online], dostupné z: http://www.studopory.vsb.cz/studijni-materialy/MatematikaI/13_MI_KAP%202_6.pdf, [cit. 2012-09-14].

[8] Anděl, J.: *Matematická statistika*. 1. vyd. Praha: SNTL/ALFA, 1978.

[9] Umetrics.com: *History*. [online], dostupné z: http://www.umetrics.com/default.aspx?id=7943, [cit. 2012-09-14].

[10] Sebera, M.: *Vícerozměrné statistické metody*. Masarykova univerzita v Brně, Fakulta sportovních studií, 2006. [online], dostupné z: http://is.muni.cz/el/1451/jaro2010/bk948/sebera-vicerozmerna.pdf, [cit. 2012-09-15].

[11] Meloun, M., Militký, J.: *Metoda hlavních komponent a exploratorní analýza vícerozměrných dat.* [online], dostupné z: http://meloun.upce.cz/docs/publication/127a.pdf, [cit. 2012-09-20].

[12] Meloun, M., Militký, J.: *Metoda hlavních komponent v laboratorní praxi.* [online], dostupné z: http://meloun.upce.cz/docs/publication/127b.pdf, [cit. 2012-10-03].

[13] Fritsch, L.: *Metoda PCA a její implementace v jazyce C++*. ČVUT v Praze, Fakulta elektrotechnická. [online], dostupné z: http://dsp.vscht.cz/konference_matlab/MATLAB07/prispevky/fritsch_l/fritsch_l.pdf, [cit. 2012-11-07].

[14] Krátký, M., Skopal, T., Snášel, V.: *Efektivní vyhledávání v kolekcích obrázků tváří*. VŠB - Technická univerzita Ostrava, Katedra informatiky. [online], dostupné z: http://www.cs.vsb.cz/kratky/courses/2003-04/dis/reference/effface.pdf, [cit. 2012-11-07].

[15] *The R Project for Statistical Computing* [online], dostupné z: http://www.r-project.org/ [cit. 2013-10-11].

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

[17] Šmilauer, P., *Moderní regresní metody*. České Budějovice: Biologická fakulta JU, 2007, [online], dostupné z: http://regent.jcu.cz/MRM.pdf, [cit. 2013-10-12].

[18] Víšek, J. Á., *Ekonometrie I*, Karolinum, Praha 1997.

[19] RGB [online], dostupné z: http://cs.wikipedia.org/wiki/RGB, [cit. 2013-16-10].

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

[21] Dvořáková, J., *Bakalářská práce: Víc než mrtví: Analýza postoje veřejnosti k pacientům v permanentním vegetativním stavu*, Olomouc: UPOL, 2012.