Nečas Center for Mathematical Modeling

Metody pro redukci dimenze v mnohorozměrné statistice a jejich výpočet

J. Kalina, and J. Duintjer Tebbens

Preprint no. 2013-027

http://ncmm.karlin.mff.cuni.cz/

METODY PRO REDUKCI DIMENZE V MNOHOROZMĚRNÉ STATISTICE A JEJICH VÝPOČET

Jan Kalina[a] , Jurjen Duintjer Tebbens[a,b]

Adresa: [a]Ústav informatiky AV ČR, v.v.i., Pod Vodárenskou věží 2, 182 07 Praha 8. [b]Univerzita Karlova v Praze, Farmaceutická fakulta v Hradci Králové, Heyrovského 1203, 500 05 Hradec Králové.

Abstract: The paper is devoted to standard multivariate statistical methods for dimension reduction. Their common basis is the eigendecomposition (i.e. computation of eigenvalues and eigenvectors) or, more generally, the singular value decomposition of specific matrices. These matrix decompositions possess excellent properties from the point of view of numerical mathematics. The paper overviews some results of numerical linear algebra on the numerical stability of methods for the computation of these decompositions. After that, it discusses various dimension reduction methods based on the decompositions, like principal component analysis, correspondence analysis, multidimensional scaling, factor analysis, and linear discriminant analysis for high-dimensional data.

Keywords: Dimension reduction, eigendecomposition, numerical stability.

Abstrakt: Článek je věnován standardním mnohorozměrným statistickým metodám pro redukci dimenze. Jejich společným rysem je spektrální rozklad určité matice (tj. výpočet vlastních čísel a vlastních vektorů) anebo obecněji singulární rozklad. Takové rozklady mají vynikající vlastnosti z hlediska numerické matematiky. Tento článek shrnuje některé výsledky numerické lineární algebry o numerické stabilitě metod pro výpočet popsaných rozkladů. Poté diskutuje možnosti použití metod pro redukci dimenze založených na těchto rozkladech jako analýzy hlavních komponent, korespondenční analýzy, mnohorozměrného škálování, faktorové analýzy a lineární diskriminační analýzy v kontextu vysoce dimenzionálních dat.

Klíčová slova: Redukce dimenze, spektrální rozklad, numerická stabilita.

1. Problematika redukce dimenzionality

Redukce dimenze představuje důležitý krok statistické analýzy či extrakce informace z mnohorozměrných dat, který může být v případě vysoce dimenzionálních dat zcela nezbytný. Jednotlivé metody slouží ke zjednodušení

Metoda	Vlastnosti	
Analýza hlavních komponent	Lineární	Nesupervizovaná
Korespondenční analýza	Lineární	Nesupervizovaná
Mnohorozměrné škálování	Nelineární	Nesupervizovaná
Faktorová analýza	Lineární	Nesupervizovaná
Lineární diskriminační analýza	Lineární	Supervizovaná

Tabulka 1: Přehled statistických metod pro redukci dimenze

dalších analýz (jako např. klasifikační nebo shlukové analýzy) a současně i umožňují přímo extrakci informace z dat, popisují rozdíly mezi skupinami, odhalují dimenzionalitu separace mezi skupinami i vyjadřují příspěvek jednotlivých proměnných k této separaci. Z anglicky psaných knih můžeme k danému tématu doporučit [15, 28] anebo [16, 22], kde jsou tytéž metody diskutovány spíše z hlediska dolování znalostí (*data mining*). Některé metody pro redukci dimenze jsou popsány i v českých učebnicích [2, 25, 33].

Metody pro redukci dimenze se někdy dělí do dvou rozsáhlých skupin, zejména v aplikacích dolování znalostí [22]:

1. Selekce proměnných (selekce příznaků, *variable selection*, *feature selection*, *variable subset selection*)
2. Extrakce příznaků (*feature extraction*)
   - Lineární
   - Nelineární

Selekcí proměnných se rozumí výběr jen těch proměnných, které jsou relevantní. Naproti tomu metody pro extrakci příznaků nahrazují pozorovaná data jejich kombinací. Jsou založeny na takovém zobrazení, které převádí pozorovaná data z vysoce rozměrného prostoru do prostoru menší dimenze. Výpočty tak sice proběhnou v prostoru menší dimenze, ale přesto je nutné napozorovat hodnoty všech proměnných. Podle charakteru tohoto zobrazení rozlišujeme metody lineární a nelineární [22].

Podle jiného kritéria dělíme metody pro redukci dimenze na supervizované a nesupervizované. Supervizovanými jsou takové, které jsou určeny pro data pocházející ze dvou nebo více skupin a současně využívají informaci o tom, které pozorování patří do které skupiny. To umožňuje zachovat oddělitelnost mezi skupinami. Někteří autoři varují, že kupříkladu analýza hlavních komponent jako příklad nesupervizovaných metod není vhodná pro redukci dimenze

Metoda	Vlastnost
Hierarchická shluková analýza	Nesupervizovaná
k průměrů (k-means)	Nesupervizovaná
k nejbližších sousedů (k-nearest neighbour)	Supervizovaná

Tabulka 2: Přehled metod shlukové analýzy

dat pocházejících ze dvou nebo více skupin v situaci, kdy cílem je klasifikační analýza [7].

Společným rysem lineárních metod pro redukci dimenze je skutečnost, že naleznou nejdůležitější transformace pozorovaných dat pomocí nástrojů lineární algebry a následně i nahradí původní data těmito novými transformovanými proměnnými. Všechny metody v tomto článku provádí transformaci dat pomocí výpočtu vlastních a/nebo singulárních vektorů. Přitom použití vlastních/singulárních vektorů způsobí i nekorelovanost transformovaných proměnných. Snížením počtu proměnných se sníží redundance v původních datech. Následná statistická analýza může profitovat ze snížení počtu stupňů volnosti [30] a obecně lze říci, že snížením počtu uvažovaných proměnných lze napravit potíže se spolehlivostí závěrů.

Tento článek stručně popisuje klasické mnohorozměrné statistické metody pro redukci dimenze uvedené v Tabulce 1 a věnuje se i jejich vhodnosti pro vysoce dimenzionální data. V kapitole 2 shrneme metody lineární algebry, které tvoří společný základ různých statistických metod pro redukci dimenze. Přitom pro vysoce rozměrná data je důležitý i pohled numerické lineární algebry, který se týká numerické stability výpočtu jednotlivých rozkladů matic. Jednotlivé metody pro redukci dimenze jsou pak popsány ve zbývajících kapitolách. Nejde přitom o vyčerpávající přehled poznatků o klasických metodách, ale spíše o zamyšlení nad jejich použitelností pro vysoce dimenzionální data, kdy počet proměnných p převyšuje počet pozorování n. Těmto aspektům se věnuje značná pozornost i v bioinformatice nebo analýze obrazu.

Mezi další metody, které by si zasloužily pozornost, stojí i shluková analýza, kterou lze také považovat za metodu pro redukci dimenze [16, 20]. Shluková analýza provádí průzkum dat s cílem najít nějaké jejich shluky. Na rozdíl od metod popsaných v tomto článku tedy není založena na převodu dat do prostoru menší dimenze za pomoci nějakého (lineárního či nelineárního) zobrazení [22]. Přehled různých metod shlukové analýzy uvádí Tabulka 2. Mezi další postupy pro redukci dimenze patří například metoda *locally linear embedding* anebo metoda *self-organizing maps* založená na neuronových sítích,

což jsou metody oblíbené například při analýze obrazové informace [19].

V článku používáme následující označení: Jednotkovou matici o velikosti p označíme $\boldsymbol{\mathcal{I}}_p$, případně $\boldsymbol{\mathcal{I}}$, pokud je dimenze jasná z kontextu. Prvek matice $\mathbf{M} \in \mathbb{R}^{n\times p}$ na i-tém řadku a v j-tém sloupci píšeme jako m_{ij}. Dále zavedeme značení

$$m_{i\cdot} = \sum_{j=1}^{n} m_{ij}, \quad m_{\cdot j} = \sum_{i=1}^{p} m_{ij}, \quad m_{\cdot\cdot} = \sum_{i=1}^{n}\sum_{j=1}^{p} m_{ij}. \tag{1}$$

Normou vektoru i matice vždy rozumíme euklidovskou normu.

2. Singulární a spektrální rozklad matice

Základem celé řady statistických metod pro redukci dimenze jsou singulární a spektrální rozklad matice. Oba pojmy se ve statistice často považují za synonyma, ale v lineární algebře představují odlišné matematické koncepty. V této sekci je popíšeme postupně a uvedeme vlastnosti (většinou bez důkazu), které jsou pro nás důležité. Pro odvození, důkazy a detailnější popis odkážeme na [32, kap. 4, 5 a 6] a [12, kap. 2.4, 7 a 8] anebo na české práce [11, kap. 2] a [9, kap. 2 a 5]. Různé algoritmy pro výpočet rozkladů matic byly popsány např. v přehledu výpočetní statistiky [6]. Naproti tomu v této kapitole klademe důraz na numerickou stabilitu algoritmů, což je klíčová vlastnost při aplikacích na vysoce rozměrná data.

2.1. Spektrální rozklad

Nejprve definujeme vlastní čísla a vlastní vektory. Pro reálnou čtvercovou matici $\mathbf{A} \in \mathbb{R}^{p\times p}$ nazveme číslo $\lambda \in \mathbb{C}$ vlastním číslem (charakteristickým číslem, *eigenvalue*) matice $\mathbf{A}$, pokud existuje nenulový vektor $\mathbf{q} \in \mathbb{C}^p$ takový, že $\mathbf{A}\mathbf{q} = \lambda\mathbf{q}$. Vektor $\mathbf{q}$ nazveme vlastním vektorem (*eigenvector*) matice $\mathbf{A}$. Každé vlastní číslo λ s příslušným vlastním vektorem $\mathbf{q}$ matice $\mathbf{A}$ zřejmě splňuje vztah $(\mathbf{A} - \lambda\boldsymbol{\mathcal{I}})\mathbf{q} = 0$. Matice $\mathbf{A} - \lambda\boldsymbol{\mathcal{I}}$ je proto singulární a její determinant je nulový. Rovnice $\det(\mathbf{A} - \lambda\boldsymbol{\mathcal{I}}) = 0$ představuje polynomiální rovnici s neznámou λ. Jelikož každý nekonstantní polynom má alespoň jeden komplexní kořen, víme, že existuje vždy alespoň jedno vlastní číslo. Vlastní čísla reálné matice mohou být komplexní, protože kořeny polynomu s reálnými koeficienty jsou obecně komplexní.

Z existence alespoň jednoho vlastního čísla vyplývá, že lze k danému lineárnímu zobrazení z lineárního prostoru do stejného prostoru vždy najít vektor, který při zobrazení nemění svůj směr. V případě, kdy existuje maximální počet p návzájem lineárně nezávislých vlastních vektorů, zřejmě existuje báze prostoru $\mathbb{R}^p$ složená z vlastních vektorů. Matici $\mathbf{A}$ můžeme potom

pomocí této báze transformovat na jakýsi kanonický tvar, tj. na diagonální matici, jejíž všechny vlastní vektory jsou jednotkové vektory (sloupce jednotkové matice). *Spektrální rozklad* čtvercové matice $\mathbf{A}$ popisuje právě tuto transformaci, a to ve tvaru

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}, \tag{2}$$

kde $\mathbf{\Lambda}$ je diagonální matice s vlastními čísly $\lambda_1, \ldots, \lambda_p$ na diagonále. Rozepíšeme-li ekvivalentní vztah $\mathbf{AQ} = \mathbf{Q\Lambda}$ po sloupcích, dostaneme

$$\mathbf{A}\mathbf{q}_i = \lambda_i \mathbf{q}_i, \quad i = 1, \ldots, p, \tag{3}$$

kde $\mathbf{q}_i$ značí i-tý sloupec matice $\mathbf{Q}$. Vidíme, že sloupce matice $\mathbf{Q}$ jsou tvořeny vlastními vektory $\mathbf{A}$.

Třídou matic, pro kterou spektrální rozklad vždy existuje, je třída symetrických pozitivně semidefinitních matic. Empirická varianční matice i korelační matice patří k této třídě. Pro symetrické pozitivně semidefinitní matice platí, že všechna vlastní čísla jsou reálná a nezáporná. Navíc existuje tím pádem vždy báze vlastních vektorů, které jsou navzájem ortogonální,

$$\mathbf{q}_i^T \mathbf{q}_j = 0, \quad i \neq j, \quad \mathbf{q}_i^T \mathbf{q}_i = 1, \quad i = 1, \ldots, p. \tag{4}$$

V tomto případě platí $\mathbf{Q}^T\mathbf{Q} = \boldsymbol{\mathcal{I}}$, tedy $\mathbf{Q}^{-1} = \mathbf{Q}^T$ a spektrální rozklad (2) lze psát jako

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T. \tag{5}$$

2.2. Singulární rozklad

Singulární rozklad (*singular value decomposition, SVD*) je narozdíl od spektrálního rozkladu definován pro libovolnou obdélníkovou matici $\mathbf{A} \in \mathbb{R}^{n\times p}$. Je-li $r \leq \min\{n, p\}$ hodnost matice $\mathbf{A}$, pak má tvar

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{6}$$

kde $\mathbf{U} \in \mathbb{R}^{n\times n}$ a $\mathbf{V} \in \mathbb{R}^{p\times p}$ jsou ortonormální matice, tj. $\mathbf{U}^T\mathbf{U} = \boldsymbol{\mathcal{I}}_n$ a $\mathbf{V}^T\mathbf{V} = \boldsymbol{\mathcal{I}}_p$ a matice $\mathbf{\Sigma} \in \mathbb{R}^{n\times p}$ má na prvních r diagonálních pozicích prvky

$$\sigma_{ii} > 0, \qquad i = 1, \ldots, r \tag{7}$$

a je nulová jinde. Čísla σ_{ii} se nazývají *singulární čísla* a platí, že se rovnají odmocnině nenulových vlastních čísel matice $\mathbf{AA}^T$ (i $\mathbf{A}^T\mathbf{A}$). Jsou tedy vždy reálná a kladná. Je konvencí, aby singulární čísla byla na diagonále

$\mathbf{\Sigma}$ uspořádana sestupně. Sloupce matice $\mathbf{U}$ obsahují takzvané levé singulární vektory a jsou zároveň vlastními vektory matice $\mathbf{A}\mathbf{A}^T$. Sloupce $\mathbf{V}$ se nazývají pravé singulární vektory a jsou i vlastními vektory $\mathbf{A}^T\mathbf{A}$. Není těžké vidět, že pro symetrické matice ($\mathbf{A} = \mathbf{A}^T$) představuje spektrální a singulární rozklad totéž. V lineární algebře se pro symetrické matice používá vždy pojem spektrální rozklad, kdyžto statistická literatura používá i pojem singulární rozklad.

Ze singulárního rozkladu (6) dostaneme po jednoduchém (ale dlouhém) rozepsání po prvcích ekvivalentní vyjádření

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_{ii}\mathbf{u}_i\mathbf{v}_i^T. \tag{8}$$

Pro jednotlivé členy přitom platí $\|\sigma_{ii}\mathbf{u}_i\mathbf{v}_i^T\| = \sigma_{ii}$ a tudíž

$$\|\sigma_{11}\mathbf{u}_1\mathbf{v}_1^T\| = \sigma_{11} \geq \|\sigma_{22}\mathbf{u}_2\mathbf{v}_2^T\| = \sigma_{22} \geq \cdots \geq \|\sigma_{rr}\mathbf{u}_r\mathbf{v}_r^T\| = \sigma_{rr}. \tag{9}$$

Ve vztahu (8) nahlížíme na matici $\mathbf{A}$ jako na součet celkového počtu r komponent. Jde vlastně o ekvivalentní vyjádření pozorovaných dat vůči jiným ortonormálním bázim. Lze říci, že komponenty příslušné největším singulárním číslům mají v pozorovaných datech největší váhu.

Statistické metody pro redukci dimenze založené na singulárním (popř. pro symetrické pozitivně semidefinitní matice spektrálním) rozkladu určité matice $\mathbf{A}$ lze interpretovat i tak, že samotnou matici $\mathbf{A}$ nahradí aproximací podle

$$\mathbf{A} \approx \sum_{i=1}^{s} \sigma_{ii}\mathbf{u}_i\mathbf{v}_i^T, \tag{10}$$

v němž $s < r$. To znamená, že se zcela ignoruje vliv takových komponent z (8), které přísluší nejmenším (a tedy v určitém smyslu nejméně důležitým) singulárním číslům.

2.3. Numerické vlastnosti

Singulární rozklad je velmi silný nástroj nejen z teoretického hlediska, ale i výpočetně, pakliže je správně implementován. Standardní metoda pro jeho výpočet je sice dražší (ale ne řádově) než metody pro jiné rozklady jako např. LU rozklad (Gaussova eliminace), ale ze všech rozkladů si nejlépe dokáže poradit v situacích, kdy daná matice je singulární nebo téměř singulární. Například určení hodnosti matice je v řadě případů možné jen pomocí SVD.

Správná implementace SVD je totiž schopná najít veškerá singulární čísla včetně těch nejmenších s přesností stejnou jako strojová přesnost [3].

Navíc singulární (v symetrickém pozitivně semidefinitním případě spektrální) rozklad lze spočíst tzv. *zpětně* stabilně. To znamená, že existují metody pro SVD dané matice $\mathbf{A}$, které spočtou v konečné aritmetice vždy rozklad, který je přesným rozkladem (tj. rozkladem v přesné aritmetice) pro matici velmi blízkou původní matice $\mathbf{A}$. Zdůrazníme, že tato vlastnost není vlastností singulárního rozkladu, ale vlastností metody jejího výpočtu.

Na výběru nejvhodnější metody maticového výpočtu velice záleží. Typickým příkladem je řešení problému nejmenších čtverců

$$\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|, \qquad \mathbf{A}\in\mathbb{R}^{n\times p}, \quad \mathbf{b}\in\mathbb{R}^n, \quad n>p. \tag{11}$$

Matematicky přesným řešením je $\mathbf{x}=\mathbf{A}^\dagger\mathbf{b}$, kde $\mathbf{A}^\dagger$ je Mooreova-Penroseova pseudoinverze k matici $\mathbf{A}$. Naivní přístup spočívá v tom, že se násobí inverze matice $\mathbf{A}^T\mathbf{A}$ s vektorem $\mathbf{A}^T\mathbf{b}$. Jsou-li sloupce matice $\mathbf{A}$ téměř lineárně závislé, pak může dojít k obrovským chybám při výpočtu $(\mathbf{A}^T\mathbf{A})^{-1}$. Vhodná implementace založená na SVD využívá vzorec

$$\mathbf{x}=\mathbf{A}^\dagger\mathbf{b}=\mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^T\mathbf{b}=\sum_{i=1}^{r}\frac{\mathbf{u}_i^T\mathbf{b}}{\sigma_{ii}}\mathbf{v}_i \tag{12}$$

obdobný vzorci (8), přičemž výpočet SVD je numericky stabilní.

Pro nejstabilnější a často zároveň nejrychlejší maticové metody doporučujeme sofware Matlab [23], který obsahuje nejmodernější implementace maticových metod přímo od vědecké komunity numerické lineární algebry, tedy komunity, která se specializuje právě na efektivní (*computationally efficient*) maticové výpočty. Alternativně lze použít populární statistický software R [26], který je narozdíl od Matlabu zdarma. I když samotné metody maticového počtu byly již dávno podrobně popsány v statistické literatuře [27], zatím se jim věnovalo málo pozornosti z numerického hlediska a není ani vždy zaručeno, že metody jsou efektivně implementovány v R. To platí zejména pro práci s vysoce dimenzionálními daty.

Výpočet singulárního rozkladu čtvercové matice velikosti p stojí přibližně $16/3\cdot p^3$ aritmetických operací. Pro vysoce dimenzionální data (např. $p\geq 10\,000$) obecně platí, že výpočet nelze provést v rozumném čase anebo dochází k vyčerpání uložného prostoru v paměti počítače. Často jsou data však *řídká*, tzn. mnoho prvků dané matice je nulových, což může výpočet usnadnit. Existují iterační metody, které vyžadují v každé iteraci jen jedno poměrně levné násobení vektoru s danou řídkou maticí. Výsledkem iteračního procesu jsou

aproximace singulárních čísel a vektorů s tím, že zpravidla nejrychleji konvergují největší singulární čísla. Takový postup je pro redukci dimenze vhodný vzhledem k (10) a často i umožňuje vyhnout se regularizaci [16, 10].

3. Analýza hlavních komponent

Analýza hlavních komponent (*PCA, principal component analysis*) představuje nejčastěji používanou metodu pro redukci dimenze. Předpokládáme, že máme k dispozici nezávislé stejně rozdělené p-rozměrné náhodné vektory $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p$.

Metoda je založena na spektrálním rozkladu empirické varianční matice

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T \in \mathbb{R}^{p \times p}, \tag{13}$$

kde $\bar{\mathbf{X}}$ označí vektor výběrového průměru. Tato matice je symetrická a pozitivně semidefinitní s nezápornými vlastními čísly. Hodnost $\mathbf{S}$ je nejvýše $\min\{n, p\}$. Protože součet vlastních čísel matice je roven součtu jejich diagonálních prvků (tj. její stopě), je v případě varianční matice roven součtu rozptylů jednotlivých proměnných.

Cílem analýzy hlavních komponent je nahradit p-rozměrná pozorování malým počtem s ($s < \min\{n, p\}$) hlavních komponent, které představují navzájem nekorelované lineární kombinace naměřených proměnných vysvětlující velkou (maximální možnou) část variability dat [2]. Hlavní komponenty lze interpretovat i tak, že jednotlivé naměřené pozorování se skládá z průměru spočítaného přes všechna pozorování plus nějaká lineární kombinace všech jednotlivých hlavních komponent. Přitom existují různá doporučení, jak volit vhodnou hodnotu s.

Analýza hlavních komponent promítá jednotlivá pozorování $\mathbf{X}_i$ na podprostor generovaný s vlastními vektory $\mathbf{q}_1, \ldots, \mathbf{q}_s$ matice $\mathbf{S}$, které přísluší největším vlastním číslům,

$$\mathbf{X}_i \quad \longrightarrow \quad [\mathbf{q}_1, \ldots, \mathbf{q}_s][\mathbf{q}_1, \ldots, \mathbf{q}_s]^T \mathbf{X}_i. \tag{14}$$

Následující výpočty pak probíhají v prostoru malé dimenze generovaném vektory $\mathbf{q}_1, \ldots, \mathbf{q}_s$ a místo $\mathbf{X}_i$ se pracuje s lineární kombinací

$$[\mathbf{q}_1, \ldots, \mathbf{q}_s]^T \mathbf{X}_i, \tag{15}$$

tedy se skalárními součiny s vlastními vektory $\mathbf{q}_1, \ldots, \mathbf{q}_s$. Příspěvek i-té hlavní komponenty ($i = 1, \ldots, p$), tj. komponenty příslušné i-tému největšímu

vlastnímu číslu, k vysvětlení celkové variability v datech přitom vyjádříme jako

$$\frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}, \tag{16}$$

kde $\lambda_1, \ldots, \lambda_p$ jsou vlastní čísla matice $\mathbf{S}$. Alternativně lze počítat hlavní komponenty z empirické korelační matice, což se doporučuje spíše jen v případě velkých odlišností ve variabilitě jednotlivých proměnných [28].

Výpočet hlavních komponent lze provést numericky stabilně i pro vysoce dimenzionální data ($n \ll p$). V softwaru je však možné narazit na takovou implementaci, která pro $n \ll p$ selhává. V softwaru R jsou k dispozici specializované knihovny HDMD či FactoMineR pro výpočet (nejen) hlavních komponent pro vysoce dimenzionální data, které lze doporučit před běžnými implementacemi [24].

4. Korespondenční analýza

Korespondenční analýza představuje obdobou analýzy hlavních komponent pro kategoriální data [14, 25, 28]. Někteří autoři ji poněkud překvapivě označují i jako korespondenční faktorovou analýzu [13].

Tzv. jednoduchá korespondenční analýza studuje vztah mezi dvěma kategoriálními proměnnými. Označme pomocí $\mathbf{N} \in \mathbb{R}^{I \times J}$ kontingenční tabulku pozorovaných četností n_{ij} s I řádky a J sloupci. Pomocí χ^2 budeme značit testovou statistiku Pearsonova χ^2 testu nezávislosti (nebo homogenity) pro stanovení vztahu mezi sloupci a řádky. Dejme tomu, že χ^2 test zamítá nulovou hypotézu nezávislosti mezi kategoriální proměnnou v řádcích tabulky a kategoriální proměnnou v jejích sloupcích. Cílem korespondenční analýzy je dále redukovat mnohorozměrný prostor řádkové a sloupcové proměnné a prostudovat interakci mezi oběma proměnnými.

Statistika χ^2 se obvykle počítá jako

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}} \right)^2 \Big/ \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}} \ . \tag{17}$$

To lze napsat pomocí relativních četností $p_{ij} = n_{ij}/n_{\cdot\cdot}$ jako

$$\chi^2 = n_{\cdot\cdot} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(p_{ij} - p_{i\cdot} p_{\cdot j} \right)^2 / p_{i\cdot} p_{\cdot j} \ . \tag{18}$$

Matice $\mathbf{P}$, jejíž prvky jsou relativní četnosti p_{ij}, se v tomto kontextu označuje

jako korespondenční matice. Další zápisy téhož jsou

$$\chi^2 = n_{..} \sum_{i=1}^{I} p_{i\cdot} \sum_{j=1}^{J} \left(\frac{p_{ij}}{p_{i\cdot}} - p_{\cdot j} \right)^2 / p_{\cdot j} = n_{..} \sum_{j=1}^{J} p_{\cdot j} \sum_{i=1}^{I} \left(\frac{p_{ij}}{p_{\cdot j}} - p_{i\cdot} \right)^2 / p_{i\cdot} \,, \tag{19}$$

kde čísla tvaru $p_{ij}/p_{i\cdot}$ a $p_{ij}/p_{\cdot j}$ jsou prvky tzv. profilů.

Řádkové a sloupcové profily jsou definovány jako

$$\mathbf{r}_i = \left(\frac{n_{i1}}{n_{i\cdot}}, \ldots, \frac{n_{iJ}}{n_{i\cdot}} \right)^T = \left(\frac{p_{i1}}{p_{i\cdot}}, \ldots, \frac{p_{iJ}}{p_{i\cdot}} \right)^T, \tag{20}$$

$$\mathbf{c}_j = \left(\frac{n_{1j}}{n_{\cdot j}}, \ldots, \frac{n_{Ij}}{n_{\cdot j}} \right)^T = \left(\frac{p_{1j}}{p_{\cdot j}}, \ldots, \frac{p_{Ij}}{p_{\cdot j}} \right)^T, \tag{21}$$

a jejich průměry

$$\mathbf{c} = (p_{\cdot 1}, \ldots, p_{\cdot J})^T \,, \quad \mathbf{r} = (p_{1\cdot}, \ldots, p_{I\cdot})^T \,. \tag{22}$$

Ze vztahu (19) dostaneme

$$\chi^2 = n_{..} \sum_{i=1}^{I} p_{i\cdot} (\mathbf{r}_i - \mathbf{c})^T \mathbf{D}_c^{-1} (\mathbf{r}_i - \mathbf{c}) = n_{..} \sum_{j=1}^{J} p_{\cdot j} \, (\mathbf{c}_j - \mathbf{r})^T \mathbf{D}_r^{-1} (\mathbf{c}_j - \mathbf{r}), \tag{23}$$

kde $\mathbf{D}_r = \mathsf{diag}(\mathbf{r})$, $\mathbf{D}_c = \mathsf{diag}(\mathbf{c})$ a diag značí diagonální matici. Zřejmě platí $\sum_{i=1}^{I} p_{i\cdot} = \sum_{j=1}^{J} p_{\cdot j} = 1$ a testovou statistiku χ^2 lze tedy interpretovat jako vážený průměr χ^2 vzdáleností mezi řádkovými průměry $\mathbf{r}_i$ a jejich průměrem $\mathbf{c}$ anebo jako vážený průměr χ^2 vzdáleností mezi sloupcovými průměry $\mathbf{c}_j$ a jejich průměrem $\mathbf{r}$. V rámci redukce mnohorozměrného prostoru lze vzorec (18) s využitím toho, že čísla $p_{ij} - p_{i\cdot} p_{\cdot j}$ jsou prvky matice $\mathbf{P} - \mathbf{r}\mathbf{c}'$, vyjádřit jako

$$\chi^2 = n_{..} \mathsf{tr} \left[\mathbf{D}_r^{-1} (\mathbf{P} - \mathbf{r}\mathbf{c}^T) \mathbf{D}_c^{-1} (\mathbf{P} - \mathbf{r}\mathbf{c}^T)^T \right], \tag{24}$$

kde tr značí stopu matice. Jsou-li $\lambda_1^2, \ldots, \lambda_r^2$ nenulová vlastní čísla matice

$$\mathbf{D}_r^{-1} (\mathbf{P} - \mathbf{r}\mathbf{c}^T) \mathbf{D}_c^{-1} (\mathbf{P} - \mathbf{r}\mathbf{c}^T)^T, \tag{25}$$

pak s využitím věty o stopě máme

$$\chi^2 = n_{..} \sum_{i=1}^{r} \lambda_i^2. \tag{26}$$

Další výpočty pak vedou ke grafickému znázornění vztahů mezi řádky a sloupci kontingenční tabulky za pomoci redukce dat do dvou dimenzí [25]. Potřebné dvě hlavní komponenty jsou přeškálovanými levými a pravými singulárními vektory v SVD rozkladu matice $\mathbf{D}_r^{-1/2}(\mathbf{P}-\mathbf{r}\mathbf{c}^T)\mathbf{D}_c^{-1/2}$, pro kterou lze dokázat, že má singulární čísla $\lambda_1, \ldots, \lambda_r$. Tyto dvě hlavní komponenty přísluší vlastním číslům λ_1^2 a λ_2^2 matice (25) a jejich příspěvek k vysvětlení všech dimenzí lze vyjádřit jako podíl

$$\frac{\lambda_1^2 + \lambda_2^2}{\sum_{j=1}^{r} \lambda_j^2}, \tag{27}$$

kde $\sum_{i=1}^{r} \lambda_i^2 = \chi^2/n_{..}$ se nazývá celková inercie. Zde je na místě upozornit na další nekonzistenci mezi terminologií ve statistice a lineární algebře. Inercie v lineární algebře je pojem spojený s *počty* (a ne součty) vlastních čísel, přesněji jde o počty záporných, nulových a kladných vlastních čísel. Celková inercie $\sum_{i=1}^{r} \lambda_i^2 = \chi^2/n_{..}$ je oblíbenou mírou asociace mezi dvěma kategoriálními proměnnými, která se běžně označuje jako ϕ (koeficient fí) [1]. Inercie odpovídá stupni rozptýlení bodů v mnohorozměrném prostoru a rovná se váženému průměru χ^2 vzdáleností řádkových profilů od svého průměru.

Jednotlivé úrovně řádkové i sloupcové proměnné se zobrazují do jediného společného grafu. Vodorovné ose v grafu odpovídá první hlavní komponenta a svislé ose druhá hlavní komponenta transformovaných dat. Řádky zobrazené blízko u sebe pak mají podobné profily a obdobně i sloupce zobrazené blízko sebe.

Současně platí, že bod odpovídající konkrétnímu řádku a bod odpovídající konkrétnímu sloupci jsou si blízko tehdy a jen tehdy, když se daná kombinace objevuje častěji, než by se očekávalo v modelu nezávislosti. Polohy bodů tak vyjadřují asociaci mezi konkrétními úrovněmi řádkového a sloupcového profilu a tato asociace se označuje jako stupeň korespondence. Grafický výstup tedy hodnotí rozdíl pozorované tabulky četností oproti situaci, kdyby platil model nezávislosti mezi oběma proměnnými. To následně umožňuje např. shlukování kategorií nominálních proměnných.

Zobecněním na více než dvě kategoriální proměnné je mnohorozměrná korespondenční analýza, kterou podrobně studuje kniha [29]. Korespondenční analýza je numericky stabilní i pro vysoce dimenzionální data [5].

5. Mnohorozměrné škálování

Mnohorozměrné škálování (vícerozměrné škálování) je metoda pro redukci dimenze mnohorozměrných dat a jejich grafické zobrazení, které co

možná nejpřesněji zachová vzdálenosti mezi pozorováními. Nejčastěji se jedná o dvourozměrnou vizualizaci [22], tedy o redukci dimenzionality dat do dvou dimenzí.

Předpokládejme, že jsou k dispozici mnohorozměrná spojitá data měřená na n objektech. Metoda pak pracuje s maticí euklidovských vzdáleností mezi objekty. Metodu však lze použít i v případě, že jsou k dispozici pouze vzdálenosti mezi objekty, zatímco původní naměřené hodnoty nejsou známy. Jinou možností je situace, kdy jsou naměřeny spíše podobnosti mezi objekty, pokud je lze snadno převést na nepodobnosti (vzdálenosti).

Nejjednodušším případem je tzv. klasické mnohorozměrné škálování, které se též označuje jako analýza hlavních souřadnic (též analýza hlavních koordinát, *principal coordinate analysis*). To představuje lineární metodu pro redukci dimenze [22]. Nechť δ_{ij} představuje vzdálenost mezi i-tým a j-tým pozorováním a nechť $\mathbf{D} \in \mathbb{R}^{n\times n}$ značí symetrickou čtvercovou matici s prvky

$$d_{ij} = -\frac{1}{2}\,\delta_{ij}^2, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n. \tag{28}$$

Pomocí $\mathbf{C} \in \mathbb{R}^{n\times n}$ označíme symetrickou čtvercovou matici s prvky c_{ij}, kde pro $i, j = 1, \ldots, n$,

$$c_{ij} = d_{ij} - (d_{i\cdot} - d_{\cdot\cdot}) - (d_{\cdot j} - d_{\cdot\cdot}) = d_{ij} - d_{i\cdot} - d_{\cdot j} + d_{\cdot\cdot}\,. \tag{29}$$

Klasické mnohorozměrné škálování je založeno na spektrálním rozkladu matice $\mathbf{C}$, která je pozitivně semidefinitní [15]. Vlastní vektory příslušné právě dvěma největším vlastním číslům poslouží k transformaci souřadnic dat a jejich následnému grafickému zobrazení [25].

Důležitou roli hraje i metrické mnohorozměrné škálování, které umožňuje transformovat vzdálenosti pomocí monotónní funkce, a je tedy nelineární metodou pro redukci dimenze. Výpočet je pak založen na iterativním řešení minimalizace ztrátové funkce, která vyjadřuje odlišnost mezi transformovanými vzdálenostmi (po redukci dimenze) a původními naměřenými vzdálenostmi. Kromě toho existuje nemetrické (ordinální) mnohorozměrné škálování, které bylo podrobněji popsáno např. v knihách [25, 22].

Obecně lze mnohorozměrné škálování popsat jako soubor mnoha různých metod a algoritmů. Přitom jsou některé z nich vhodné i pro analýzu vysoce dimenzionálních dat. V tomto kontextu se za nejvhodnější považují právě algoritmy založené na singulárním rozkladu [4].

6. Faktorová analýza

Faktorová analýza je založena na předpokladu, že lze pozorovaná data vysvětlit pomocí malého počtu latentních proměnných (faktorů) [2, 33]. Před-

stavuje častý nástroj při vyhodnocování psychologických testů či v ekonomii. Faktorová analýza vzbuzuje určité kontroverze i kvůli problematické interpretaci vzniklých faktorů.

Model faktorové analýzy pro i-té pozorování zapíšeme jako

$$\begin{aligned} X_{i1} - \mu_1 &= \gamma_{11} f_{i1} + \cdots + \gamma_{1t} f_{it} + e_{i1}, \\ &\vdots \qquad\quad \vdots \\ X_{ip} - \mu_p &= \gamma_{p1} f_{i1} + \cdots + \gamma_{pt} f_{it} + e_{ip}, \end{aligned} \tag{30}$$

kde $i = 1, \ldots, n$. Pozorování $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ zde vysvětlujeme pomocí latentních faktorů $f_{i1}, \ldots, f_{it}$, parametrů $\mu_i, \ldots, \mu_p$ a $\gamma_{11}, \ldots, \gamma_{pt}$ a šumu $e_{i1}, \ldots, e_{ip}$. Model můžeme vyjádřit maticově jako

$$\mathbf{X}_i - \boldsymbol{\mu} = \boldsymbol{\Gamma} \mathbf{f}_i + \mathbf{e}_i, \quad i = 1, \ldots, n. \tag{31}$$

Oproti analýze hlavních komponent se nepředpokládá, že by latentní proměnné vysvětlily veškerou variabilitu pozorovaných dat. Část variability jednotlivé proměnné, která je vysvětlena latentními proměnnými, se označuje jako komunalita.

Nyní stručně popíšeme možný způsob pro odhad parametrů v (30). Označme pomocí $\mathbf{S}$ empirickou varianční matici spočítanou ze všech pozorování $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Předpokládá se, že $\mathbf{S} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + \mathsf{var}\,\mathbf{e}$ a že matice $\mathsf{var}\,\mathbf{e}$ je diagonální. Díky tomu se odhadnou mimodiagonální prvky matice $\mathbf{T} = \mathbf{S} - \mathsf{var}\,\mathbf{e}$ přímo pomocí mimodiagonálních prvků $\mathbf{S}$. Pokud jde o diagonální prvky $\mathbf{T}$, lze je odhadnout iteračním postupem [2]. Tím se získá odhad celé matice $\mathbf{T}$ a dále se hledá taková matice $\boldsymbol{\Gamma}$, která splňuje vztah $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \mathbf{T}$. Komplikací je i to, že také pro matici $\boldsymbol{\Gamma}^* = \boldsymbol{\Gamma}\mathbf{U}$ platí $\boldsymbol{\Gamma}^*\boldsymbol{\Gamma}^{*T} = \mathbf{T}$, pokud $\mathbf{U}$ je (libovolná) ortonormální matice. To znamená, že latentní proměnné nejsou určeny jednoznačně. Byly navrženy různé metody pro odhad matice $\boldsymbol{\Gamma}$:

1. Metoda hlavních komponent
2. Metoda hlavních faktorů
3. Iterovaná metoda hlavních faktorů
4. Metoda maximální věrohodnosti
5. Metoda minimalizace reziduí

Metodu hlavních komponent pro odhad parametrů ve faktorové analýze lze nastínit pomocí spektrálního rozklad matice $\mathbf{T}$ ve tvaru

$$\mathbf{T} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T. \tag{32}$$

Označme napřed pomocí $\mathbf{Q}_t$ matici obsahující prvních t sloupců $\mathbf{Q}$ a pomocí $\mathbf{\Lambda}_t^{1/2}$ diagonální matici, jejíž diagonální prvky jsou rovny odmocninám t prvních diagonálních prvků matice $\mathbf{\Lambda}$. Matice $\mathbf{\Gamma}$ se pak určí jako

$$\mathbf{\Gamma} = \mathbf{Q}_t \mathbf{\Lambda}_t^{1/2}. \tag{33}$$

Alternativně lze matici $\mathbf{S}$ nahradit empirickou korelační maticí [28].

Pro vysoce dimenzionální data ($n \ll p$) lze faktorovou analýzu použít, pokud se zvolí vhodná metoda pro odhad matice $\mathbf{\Gamma}$. Často používaná metoda maximální věrohodnosti zde však selhává. Vhodně implementovanou metodu nabízí např. knihovna HDMD v softwaru R.

7. Lineární diskriminační analýza

Přestože lineární diskriminační analýza (LDA) představuje klasifikační metodu, lze ji interpretovat jako metodu, která již v sobě automaticky zahrnuje supervizovanou redukci dimenze [22].

Předpokládejme, že máme k dispozici celkový počet K různých skupin, v nichž jsou pozorovány nezávislé p-rozměrné náhodné veličiny $\mathbf{X}_1, \ldots, \mathbf{X}_n$ pocházející z p-rozměrného normálního rozdělení. Předpokládáme, že každé skupině přísluší odlišný vektor středních hodnot, ale varianční matice $\mathbf{\Sigma}$ je společná pro všechny skupiny. Její odhad označíme jako $\mathbf{S}$. V k-té skupině označíme výběrový průměr pozorovaných dat jako $\bar{\mathbf{X}}_k$ a celkový průměr napříč skupinami jako $\bar{\mathbf{X}}$. LDA zařadí nové pozorování do té skupiny, k jejímuž průměru má nejblíž ve smyslu Mahalanobisovy vzdálenosti.

Ekvivalentně lze výpočet LDA založit na tzv. diskriminačních skórech, kterých je právě $s = \min\{K-1, p\}$. Tento přístup se typicky využívá v softwarových implementacích [18, 8]. Matice $\mathbf{B}$ o rozměrech $p \times p$ je definovaná jako

$$\mathbf{B} = \sum_{k=1}^{K} (\bar{\mathbf{X}}_k - \bar{\mathbf{X}})(\bar{\mathbf{X}}_k - \bar{\mathbf{X}})^T. \tag{34}$$

Další výpočet je založen na spektrálním rozkladu matice $\mathbf{S}^{-1}\mathbf{B}$. Diskriminační skóry se rovnají těm hlavním vektorům $\mathbf{S}^{-1}\mathbf{B}$, které přísluší nenulovým vlastním číslům.

Věta: Uvažujme p-rozměrné pozorování $\mathbf{Z}$. Označme vlastní vektory matice $\mathbf{S}^{-1}\mathbf{B}$ příslušné nenulovým vlastním číslům jako $\mathbf{v}_1, \ldots, \mathbf{v}_s$. Pak lineární dis-

kriminační analýza klasifikuje pozorování $\mathbf{Z}$ do skupiny k právě tehdy, když

$$\sum_{j=1}^{s} \left[\mathbf{v}_j^T(\mathbf{Z} - \bar{\mathbf{X}}_k)\right]^2 \leq \sum_{j=1}^{s} \left[\mathbf{v}_j^T(\mathbf{Z} - \bar{\mathbf{X}}_i)\right]^2, \quad i = 1, \ldots, K. \tag{35}$$

Důkaz je uveden např. v knize [18]. Důsledkem věty je pak následující tvrzení, které ukazuje, jak LDA speciálně pro $K = 2$ redukuje dimenzi na hodnotu 1.

Věta: Uvažujme $K = 2$ a mějme k dispozici p-rozměrné pozorování $\mathbf{Z}$. Pak má matice $\mathbf{S^{-1}B}$ jediné nenulové vlastní číslo, jemuž přísluší vlastní vektor, který označíme $\mathbf{v}$. Předpokládejme dále $\mathbf{v}^T\bar{\mathbf{X}}_1 > \mathbf{v}^T\bar{\mathbf{X}}_2$. Lineární diskriminační analýza klasifikuje pozorování $\mathbf{Z}$ do skupiny 1 právě tehdy, když

$$\mathbf{v}^T\mathbf{Z} > \frac{\mathbf{v}^T\bar{\mathbf{X}}_1 + \mathbf{v}^T\bar{\mathbf{X}}_2}{2}\,. \tag{36}$$

Pokud však platí $\mathbf{v}^T\bar{\mathbf{X}}_1 < \mathbf{v}^T\bar{\mathbf{X}}_2$, lineární diskriminační analýza klasifikuje $\mathbf{Z}$ do skupiny 1 právě tehdy, když

$$\mathbf{v}^T\mathbf{Z} < \frac{\mathbf{v}^T\bar{\mathbf{X}}_1 + \mathbf{v}^T\bar{\mathbf{X}}_2}{2}\,. \tag{37}$$

Pro vysoce dimenzionální data za předpokladu $n \ll p$ trpí lineární diskriminační analýza tzv. prokletím dimenzionality. Při výpočtu diskriminačních skórů je velmi obtížné nejprve spočítat potřebná vlastní čísla, přičemž odpovídající vlastní vektory nemusí v tomto kontextu ani být definovány [10]. Možným řešením je použít regularizovaný odhad varianční matice [16] anebo vhodně modifikovat Fisherovo optimalizační kritérium, které stojí v pozadí metody LDA a vyžaduje výpočet vlastních čísel matice $\mathbf{S^{-1}B}$ [10].

Poděkování

Práce vznikla za finanční podpory Nadačního fondu Neuron na podporu vědy. Druhý autor byl též podporován grantem č. P201/ 13-06684S Grantové agentury České republiky.

Literatura

[1] Agresti A. (2002): *Categorical data analysis.* Second edition. Wiley, New York.

[2] Anděl J. (1978): *Matematická statistika.* SNTL, Praha.

[3] Barlow J.L., Bosner N., Drmač Z. (2005): A new stable bidiagonal reduction algorithm. *Linear Algebra and its Applications* **397**, 35 – 84.

[4] Bécavin C., Tchitchek N., Mintsa-Eya C., Lesne A., Benecke A. (2011): Improving the efficiency of multidimensional scaling in the analysis of high-dimensional data using singular value decomposition. *Bioinformatics* **27** (10), 1413 – 1421.

[5] Busygin S., Pardalos P.M. (2007): Exploring microarray data with correspondence analysis. In Pardalos P.M. et al. (Eds.): *Data Mining in Biomedicine.* Springer, New York, 2007, 25 – 38.

[6] Čížková L, Čížek P. (2012): Numerical linear algebra. In Gentle J.E., Härdle W.K., Mori Y. (Eds.): *Handbook of Computational Statistics.* Springer, Berlin, 105 – 137.

[7] Dai J.J., Lieu L., Rocke D. (2006): Dimension reduction for classification with gene expression microarray data. *Statistical Applications in Genetics and Molecular Biology* **5** (1), Article 6.

[8] Duda R.O., Hart P.E., Stork D.G. (2001): *Pattern Classification.* Second edition. Wiley, New York.

[9] Duintjer Tebbens J., Hnětynková I., Plešinger M., Strakoš Z., Tichý P. (2012): *Analýza metod pro maticové výpočty.* Matfyzpress, Praha.

[10] Duintjer Tebbens J., Schlesinger P. (2007): Improving implementation of linear discriminant analysis for the high dimension/small sample size problem. *Computational Statistics & Data Analysis* **52**, 423 – 437.

[11] Fiedler M. (1981): *Speciální matice a jejich použití v numerické matematice.* SNTL, Praha.

[12] Golub G., van Loan Ch. (1996): *Matrix computations.* Johns Hopkins University Press, Baltimore.

[13] Göhlmann H., Talloen W. (2009): *Gene expression studies using Affymetrix microarrays.* Chapman & Hall/CRC, Boca Raton.

[14] Greenacre M. (1984): *Theory and applications of correspondence analysis.* Academic Press, London.

[15] Härdle W.K., Simar L. (2007): *Applied multivariate statistical analysis.* Springer, Berlin.

[16] Hastie T., Tibshirani R., Friedman J. (2001): *The elements of statistical learning.* Springer, New York.

[17] Havel V., Holenda J. (1984): *Lineární algebra.* SNTL, Praha.

[18] Johnson R.A., Wichern D.W. (1982): *Applied multivariate statistical analysis.* Prentice-Hall, Englewood Cliffs.

[19] Kalina J. (2011): Facial image analysis in anthropology: A review. *Anthropologie* **49** (2), 141 – 153.

[20] Kalina J. (2013): Robustness aspects of knowledge discovery. Zasláno do sborníku *Znalosti 2013, 13.-15.10. 2013, Ostrava.*

[21] Ledoit O., Wolf M. (2004): A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88**, 365 – 411.

[22] Martinez W.L., Martinez A.R., Solka J.L. (2011): *Exploratory data analysis with MATLAB.* 2nd edn. Chapman & Hall/CRC, Boca Raton.

[23] MathWorks, Inc., 1984–2013. *MATLAB 8.1, http://www.mathworks.com/products/matlab.*

[24] McFerrin L. (2013): Package HDMD. Staženo z *http://cran.r-project.org/web/packages/HDMD/HDMD.pdf* (14.6.2013).

[25] Meloun M., Militký J. (2006): *Kompendium statistického zpracování dat. Metody a řešené úlohy.* Academia, Praha.

[26] R Development Core Team: *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, 2012, *http://www.R-project.org/.*

[27] Rao C.R. (1973): *Linear statistical inference and its applications.* Wiley, New York.

[28] Rencher A.C. (2002): *Methods of multivariate analysis.* Second edn. Wiley, New York.

[29] Řehák J., Řeháková B. (1986): *Analýza kategorizovaných dat v sociologii.* Academia, Praha.

[30] Řezanková H., Húsek D., Snášel V. (2007): *Shluková analýza dat.* Professional Publishing, Praha.

[31] Saad Y. (2011): *Numerical methods for large eigenvalue problems.* Revised edition. SIAM, Philadelphia.

[32] Watkins D.S. (2010): *Fundamentals of matrix computations.* Third edition. John Wiley & Sons, New York.

[33] Zvárová J., Svačina Š., Valenta Z., Berka P., Buchtela D., Jiroušek R., Malý M., Papíková V., Peleška J., Rauch J., Vajda I., Veselý A., Zvára K., Zvolský M. (2009): *Systémy pro podporu lékařského rozhodování.* Karolinum, Praha.