SROVNÁNÍ METOD PRO REDUKCI DIMENZIONALITY APLIKOVANÝCH NA ORDINÁLNÍ PROMĚNNÉ

Lukáš Sobíšek, Hana Řezanková*

Úvod

Důležitým nástrojem při různých výzkumech v ekonomické sféře, například při průzkumech trhu, jsou dotazníková šetření. Jejich pomocí jsou získávány názory a hodnocení respondentů. Analyzované datové soubory se vyznačují velkým zastoupením ordinálních proměnných, což je potřeba zohlednit při aplikaci statistických metod. Vzhledem k obvykle značnému počtu proměnných je při vícerozměrné analýze velmi významným faktorem jeho redukce. Kromě výběru nejvýznamnějších proměnných (podle určitého hlediska) hrají důležitou roli metody sloužící k redukci rozměru úlohy. Jejich výsledkem jsou odvozené proměnné, které zachycují vztahy mezi původními proměnnými. Při vícerozměrných analýzách pak lze respondenty charakterizovat pomocí menšího počtu těchto odvozených proměnných.

V článku se zaměříme na stanovování skupin podobných proměnných, které mohou být interpretovány jako faktory charakterizující sledované objekty. K tomuto účelu můžeme využít různé metody, jejichž výsledky nemusí být stejné. V dalším textu navrhneme jednak interpretovat získané výsledky pomocí fuzzy shlukové analýzy, jednak tyto výsledky kombinovat do konečného stanovení skupin proměnných. Součástí navrženého postupu je i zjištění vhodného počtu skupin. Pro redukci dimenzionality doporučujeme objekty charakterizovat pomocí proměnných, které reprezentují dané skupiny.

* Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky (xsobl06@vse.cz, hana.rezankova@vse.cz).

Článek vznikl v rámci grantů GAČR P202/10/0262 a IGA VŠE F4/3/2010.

Děkujeme PhDr. Marcelu Moulisové, Ph.D., a PaedDr. Zdeňku Valjentovi, Ph.D., za poskytnutá data a Ing. Michaele Ryšánkové a Ing. Vandě Vilhanové za pomoc při přípravě datových souborů.

Pro analýzu vztahů mezi proměnnými jsme využili klasickou a diskretní faktorovou analýzu, shlukové modely latentních tříd, kategoriální analýzu hlavních komponent, metrické i nemetrické vícerozměrné škálování. Pokud metody umožňují při analýze vycházet z matice podobností, byla využita matice vytvořená na základě Kendallova koeficientu pořadové korelace. Jednotlivé metody a jejich různé kombinace jsme porovnávali pomocí kvality shluků, vytvořených na základě interpretace získaných hodnot komponent (dimenzí) s využitím fuzzy shlukové analýzy. Shluky byly hodnoceny pomocí tří vybraných koeficientů.

Metody redukce dimenzionality jsme aplikovali na dva reálné datové soubory z dotazníkových šetření s proměnnými na pětibodové a sedmibodové ordinální škále. Využili jsme přitom programové systémy SPSS, STATISTICA a Latent GOLD. Fuzzy shluková analýza byla prováděna v systému S-PLUS. Výsledné přiřazení proměnných do skupin bylo znázorněno pomocí obrysového grafu.

Další text je uspořádán tak, že v sekci 1 jsou nejprve stručně charakterizovány vybrané statistické metody, které lze využít k redukci dimenzionality, a jiné použité postupy. V sekci 2 jsou popsány experimenty a jejich výsledky. Celkové zhodnocení je obsaženo v závěru článku.

1. Použité metody a jejich specifikace

Základními metodami, které umožňují redukovat dimenzi vektorů charakterizujících objekty zahrnuté do analýzy, jsou *analýza hlavních komponent* (PCA – *Principal Component Analysis*) a *faktorová analýza* (FA – *Factor Analysis*). Tyto metody umožňují nahradit větší počet původních proměnných menším počtem proměnných latentních, přičemž mezi původními proměnnými předpokládají vzájemné lineární vztahy. Tradiční model faktorové analýzy vychází z předpokladu, že latentní proměnné jsou spojité, normálně rozdělené náhodné veličiny. Výpočty jsou založeny na korelační matici pro původní proměnné. Pro účely tohoto článku jsme v systému SPSS využili matici vycházející z hodnot koeficientu pořadové korelace, viz níže.

Rozsáhlou skupinou metod pro nalezení skupin latentních proměnných jsou *modely latentních tříd* (též modely LC – *Latent Class*, či LCA – *Latent Class Analysis*). Z konkrétních metod lze uvést *shlukové modely latentních tříd* (modely LCC – *Latent Class Cluster*), *modely diskretní faktorové analýzy* (modely LC DFactor, resp. zkráceně DFactor), *analýzu latentních charakteristik* (LTA – *Latent Trait Analysis*), *analýzu latentních profilů* (LPA – *Latent Profile Analysis*), *regresní modely latentních tříd* (modely LCR – *Latent Class Regression*). Více o těchto metodách viz Vermunt (2005). Pro účely tohoto článku jsme v rámci systému Latent GOLD použili modely LCC a DFactor.

Protože metoda PCA je určena pro analýzu kvantitativních spojitých proměnných, pro kategoriální proměnné je vhodné využít jiných postupů. Jedním z nich je *kategoriální analýza hlavních komponent* (CATPCA – *CATegorical Principal Component Analysis*), která transformuje kategoriální proměnné na proměnné kvantitativní a lze uvažovat i nelineární vztahy mezi proměnnými. Analýzu pomocí stejnojmenné procedury jsme realizovali v systému SPSS.

Některé techniky redukce dimenze dat jsou založeny na projekci vícedimenziálního prostoru do prostoru s méně dimenzemi při zachování maximální variability dat. Každý z n objektů, charakterizovaný vektorem hodnot p proměnných, je zobrazen jako bod v k -rozměrném prostoru, tak aby platilo $k < p$. K těmto metodám patří již výše zmíněná metoda PCA a dále *vícerozměrné škálování* (MDS – *MultiDimensional Scaling*). Výběr k hlavních os (dimenzí) v klasické metodě MDS je ekvivalentní výběru p hlavních komponent v metodě PCA, viz např. Hebák (2007). Metoda MDS je zobecněním faktorové analýzy. Zatímco FA se zaměřuje na vztahy mezi proměnnými a vychází z korelační matice, MDS může vycházet z libovolné matice, která vyjadřuje vztahy buď mezi proměnnými, nebo mezi objekty. Pro účely tohoto článku jsme použili procedury ALSCAL a PROXSCAL v systému SPSS a nemetrické vícerozměrné škálování NMMDS v systému STATISTICA.

Metoda MDS (obdobně jako některé metody shlukové analýzy) vychází z matice vzdáleností (nepodobností, odlišností) pro všechny kombinace dvojic objektů či proměnných. Tato vzdálenost může být vypočtena na základě podobnosti. Pro dvě proměnné lze jejich podobnost ohodnotit pomocí některé z měr vzájemné (symetrické) závislosti. Základní mírou podobnosti dvou kvantitativních proměnných je výběrový Pearsonův korelační koeficient. Pro měření podobnosti ordinálních proměnných lze využít Spearmanův či Kendallův koeficient pořadové korelace, koeficient γ či τ_c , případně symetrické Sommersovo d . Podrobný výčet měr pro různé typy proměnných, společně s výpočetním postupem, je uveden např. v Řezanková (2009).

Pro aplikace popsané v tomto článku byl v rámci všech tří výše zmíněných metod vícerozměrného škálování a též ve faktorové analýze použit *Kendallův koeficient pořadové korelace* τ_b , založený na porovnávání dvojic objektů a počítaný podle vzorce

$$\tau_b = \frac{\Gamma - \Delta}{\sqrt{(\Gamma + \Delta + \Psi_k)(\Gamma + \Delta + \Psi_l)}}, \quad (1)$$

kde Γ je počet konkordantních párů, Δ je počet diskordantních párů, Ψ_k je počet párů, které obsahují stejnou hodnotu znaku X_k , ale různou hodnotu X_l , Ψ_l je počet párů, které obsahují stejnou hodnotu znaku X_l , ale různou hodnotu X_k .

Zařazení proměnných do skupin na základě získaných hodnot dimenzí (komponent) bylo provedeno pomocí *fuzzy shlukové analýzy*. Pro spojování výsledků získaných pomocí různých metod do jediného výsledku byl využit přístup *CSPA* (*Cluster-based Similarity Partitioning Algorithm*).

1.1 Faktorová analýza

Cílem faktorové analýzy je nalézt k latentních proměnných (*faktorů*) F_i , $i = 1, 2, \dots, k$, které co nejlépe vysvětlí závislosti mezi p pozorovanými proměnnými X_j , $j = 1, 2, \dots, p$. Na rozdíl od metody PCA, jež provádí ortogonální transformaci vzájemně lineárně závislých proměnných, FA identifikuje statistický model. Necht' $\mathbf{X}$ je p -rozměrný náhodný vektor uvažovaných veličin s vektorem středních hodnot μ , kovarianční maticí

Σ a korelační maticí $\mathbf{P}$. Model faktorové analýzy vychází z této matice $\mathbf{P}$ a můžeme ho zapsat ve tvaru

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{\Gamma}\mathbf{F} + \mathbf{E}, \quad (2)$$

kde $\mathbf{F}$ značí k -rozměrný náhodný vektor *společných faktorů*, $\mathbf{\Gamma}$ je matice *faktorových zátěží* typu $p \times k$ a $\mathbf{E}$ je p -rozměrný vektor *specifických* (chybových) *faktorů*. Předpokládá se, že společné faktory F_i a specifické faktory ε_i jsou navzájem nekorelované a mají nulovou střední hodnotu, tj. $E(F_i) = 0$, $E(\varepsilon_i) = 0$ a $\text{Cov}(\mathbf{F}, \mathbf{E}) = \mathbf{0}$, $\text{Cov}(\mathbf{F}) = \mathbf{I}$ a $\text{Cov}(\mathbf{E}) = \mathbf{\Psi}$ je diagonální matice. Za těchto předpokladů lze kovarianční matici vstupní matice proměnných zapsat jako

$$\mathbf{X} = \mathbf{\Gamma}\mathbf{\Gamma}^T + \mathbf{\Psi}. \quad (3)$$

Faktory F_i se interpretují pomocí korelace s původními proměnnými X_j . Korelační matici můžeme zapsat ve tvaru

$$\mathbf{P}_{XF} = \mathbf{D}^{-1/2} + \mathbf{\Gamma}, \quad (4)$$

kde $\mathbf{D}$ je diagonální matice, jejíž prvky jsou rozptyly původních pozorovaných proměnných, tj. $\mathbf{D} = \text{diag}(\sigma_{X_1 X_1}, \dots, \sigma_{X_p X_p})$.

Faktorová analýza může též vycházet z normovaných veličin Z_j s nulovými středními hodnotami a s jednotkovými rozptyly. Uvedený model pak zapíšeme ve tvaru

$$\mathbf{Z}_{XF} = \mathbf{\Gamma}^* \mathbf{F}^* + \mathbf{E}^*, \quad (5)$$

ve kterém je rozdíl především ve významu matice $\mathbf{\Gamma}^*$. Prvky této jsou matice korelačními koeficienty mezi proměnnými X_j a faktory F_i , tj. $\mathbf{P}_{XF} = \mathbf{\Gamma}^*$. Faktory interpretujeme přímo pomocí faktorových zátěží, více o metodách např. viz Hebák (2007).

Odhady $\hat{\mathbf{\Psi}}$, $\hat{\mathbf{\Gamma}}$ parametrů $\mathbf{\Psi}$, $\mathbf{\Gamma}$ faktorového modelu

$$\mathbf{S}_{XF} = \hat{\mathbf{\Gamma}} \hat{\mathbf{\Gamma}} + \hat{\mathbf{\Psi}} \quad (6)$$

vycházejí z odhadu kovarianční matice $\mathbf{S}$, resp. z odhadu korelační matice $\mathbf{R}$. Počet parametrů (stupňů volnosti) $d = 1/2(p - k)^2 - 1/2(p + k)$ zároveň definuje horní hranici počtu faktorů, které můžeme modelem identifikovat. Pokud $d = 0$, existuje jednoznačné řešení. V praxi však většinou platí, že $d > 0$, a řešení získáme aproximacemi. Pro odhad parametrů se zejména používají tři postupy: metoda maximální věrohodnosti, metoda hlavních faktorů a metoda hlavních komponent, jejichž výpočetní algoritmy jsou uvedeny např. v Mardia (1979). Při analýzách, jejichž výsledky jsou uvedené v tomto článku, byl použit třetí postup.

1.2 Modely latentních tříd

Modely latentních tříd (LC) předpokládají, že existuje relativně malý počet nepozorovaných (skrytých) proměnných Y_i , kde $i = 1, 2, \dots, k$. Tyto latentní proměnné stojí v pozadí vzájemných vztahů mezi sledovanými proměnnými X_j , $j = 1, 2, \dots, p$, kde X_j nabývá D_j odlišných hodnot. Modely LC se od sebe vzájemně liší v požadavcích na typ a pravděpodobnostní rozdělení pozorovaných a skrytých veličin, viz Vermunt (2005).

Pro účely tohoto článku jsou uvažovány modely, které předpokládají, že latentní proměnné jsou diskrétní náhodné veličiny. Uvažujme jednu nominální skrytou proměnnou Y obsahující k tříd. Jednotlivé třídy značíme symbolem y , kde $y \in \{1, 2, \dots, k\}$. Axiomem modelů LC je *lokální (podmíněná) nezávislost* v tom smyslu, že pokud je určena náležitost všech sledovaných objektů do jednotlivých tříd latentní proměnné, tyto latentní proměnné jsou navzájem nezávislé, viz Bartholomew (2002). Lokální nezávislost náhodných veličin X_i lze vyjádřit vztahem

$$P(\mathbf{X} = \mathbf{x} | Y = y) = \prod_{j=1}^p P(X_j = x_j | Y = y). \quad (7)$$

LC model je založen na předpokladu, že získaná data byla vygenerována „směsí“ pravděpodobnostních rozdělení. To lze vyjádřit pomocí Bayesova vzorce úplné pravděpodobnosti:

$$P(\mathbf{X} = \mathbf{x}) = \sum_{y=1}^k P(Y = y) \cdot P(\mathbf{X} = \mathbf{x} | Y = y), \quad (8)$$

jinými slovy, pravděpodobnost získání konkrétního vektoru $\mathbf{x}$ je váženým průměrem podmíněných pravděpodobností $P(\mathbf{X} = \mathbf{x} | Y = y)$, jejichž počet je k a kde $\sum_{y=1}^k P(Y = y) = 1$.

Kombinací vzorců úplné pravděpodobnosti a vzorce pro axiom lokální nezávislosti získáváme klasický LC model:

$$P(\mathbf{X} = \mathbf{x}) = \sum_{y=1}^k P(Y = y) \prod_{j=1}^p P(X_j = x_j | Y = y). \quad (9)$$

Odhad velikosti tříd $\pi = P(y)$ a odhady podmíněných pravděpodobností hodnot x_i v jednotlivých třídách, tj. $P(X_j = x_j | Y = y) = v_{y_j}$, získáváme metodou maximální věrohodnosti.

Použitý model LCC předpokládá takovýto případ s jednou nominální skrytou proměnnou Y obsahující k tříd. Podle Vermunt (2002) lze tento model vnímat jako alternativní modelový přístup ke klasickým nehierarchickým metodám shlukové analýzy (k -průměrů), kde jako míru podobnosti používáme odhad pravděpodobnosti.

Druhý již výše zmíněný model DFactor může obsahovat více latentních proměnných (binárních nebo ordinálních). V tomto článku jsou prezentovány výsledky získané na základě modelu se čtyřmi binárními faktory, a to z důvodu porovnatelnosti s ostatními metodami, pro které byly jako optimální vyhodnoceny právě čtyři skupiny proměnných.

1.3 Kategoriální analýza hlavních komponent

Standardní metoda PCA posuzuje vzájemné lineární vztahy mezi pozorovanými proměnnými. Algoritmus je navržen tak, aby první zjištěná latentní proměnná vysvětlovala co nejvíce původní variability. Tato latentní proměnná – *hlavní komponenta* – je vyjádřena vztahem

$$Y_1 = \omega_1^T(\mathbf{x} - \boldsymbol{\mu}), \quad (10)$$

kde ω_1^T je ortonormální charakteristický vektor charakteristického čísla λ_1 kovarianční matice Σ a $\boldsymbol{\mu}$ je vektor středních hodnot. Následně algoritmus hledá druhou lineární kombinaci (druhou hlavní komponentu), která vysvětlí co největší část zbývajících variability, atd. Komponenty jsou ortogonální (vzájemně nekorelované) a jsou vytvářeny postupně s klesajícím významem důležitosti. Jednotlivé prvky určitého charakteristického vektoru jsou vzájemně porovnatelné, nejsou však porovnatelné prvky z různých charakteristických vektorů. Pro lepší srovnání je vhodné použít vektory komponentních zátěží

$$\gamma_r = \omega_r \sqrt{\lambda_{(r)}} = C(X_j, Y_{r,N}). \quad (11)$$

Tyto vektory jsou zároveň kovariancí j -té proměnné X_j a r -té normované komponenty $Y_{r,N}$, definované jako

$$Y_{r,N} = \frac{\gamma_r^T}{\lambda_{(r)}} (\mathbf{x} - \boldsymbol{\mu}), \quad (12)$$

přičemž $Y_{r,N}$ mají jednotkové rozptyly.

Metoda CATPCA optimálním škálováním transformuje kategoriální proměnné na proměnné kvantitativní. Tato optimalizace (kvantifikace proměnných) vede k získání optimálních hlavních komponent. Škály jednotlivých proměnných se mohou lišit. Nejsou kladeny žádné požadavky na pravděpodobnostní rozdělení proměnných. Tato analýza, oproti standardní PCA, si navíc dokáže poradit s nelineárními vztahy mezi proměnnými. Jestliže se všechny proměnné řídí multinomickým pravděpodobnostním rozdělením, CATPCA je identická s vícerozměrnou korespondenční analýzou.

Iterativní postup v SPSS (De Leeuw, 1976, IBM SPSS Statistics) lze shrnout následovně. Proces je zahájen tak, že každému objektu je přiřazeno náhodné číslo (skóre objektu), není-li definováno jinak. Matici skóre objektů značíme $\mathbf{X}$. Poté je získána vážená ortonormální matice $\mathbf{X}_w^+ = \mathbf{W}\mathbf{X}$ tak, aby platilo

$$\boldsymbol{\mu}^T = \mathbf{M}_* \mathbf{W} \mathbf{X} = \mathbf{0} \quad \text{a} \quad \mathbf{X}^T \mathbf{M}_* \mathbf{W} \mathbf{X} = n_w m_w \mathbf{I}, \quad (13)$$

kde $\mathbf{M}_* = \Sigma_j \mathbf{M}_j$, přičemž $\mathbf{M}_j$ je diagonální matice s prvky $m_{(j)ii}$, vyjadřující váhy v_j jednotlivých proměnných pro každý objekt (implicitně $v_j = 1$, pokud váha není specifikována). Dále $\mathbf{W}$ značí diagonální matici, jejíž prvky w_i jsou váhy jednotlivých objektů (implicitně $w_i = 1$ pro nevážené objekty), n_w je součet vah objektů a m_w je součet vah proměnných. Kvantifikace proměnných začíná výpočtem prvotní vektorové zátěže

$$\mathbf{X}_w^+ \times (\mathbf{I} - \mathbf{M}_j \boldsymbol{\mu} \boldsymbol{\mu}^T \mathbf{W} / (\boldsymbol{\mu}^T \mathbf{M}_j \mathbf{W} \boldsymbol{\mu})) \mathbf{h}_j, \quad (14)$$

kde $\mathbf{h}_j$ je vektor indikátorů kategorií pro j -tou proměnnou.

Optimální vektor komponentních zátěží

$$\mathbf{a}_j^+ = n_w^{-1} \tilde{\mathbf{Y}}^T \mathbf{D}_j^+ \quad (15)$$

a kvantifikované kategorie (hlavní komponenty) původních proměnných¹

$$\mathbf{Y}_j^+ = \mathbf{y}_j^+ \mathbf{a}_j^{+T} \quad (16)$$

získáme metodou ALS (*Alternating Least Squares*). Původní odhad je vyjádřen jako $\mathbf{Y} = \mathbf{D}_j^{-1} \mathbf{G}_j^T \mathbf{X}_w^+$, kde $\mathbf{G}_j$ je indikátorová matice j -té proměnné s prvky $g_{(j)ir} = 1$, pokud i -tý objekt nabývá r -té kategorie pro j -tou proměnnou, a $\mathbf{D}_j$ je diagonální matice typu $k_j \times k_j$ (k_j je počet kategorií j -té proměnné) vážených sloupcových součtů $\mathbf{G}_j$, tj. $\mathbf{D}_j = \mathbf{G}_j^T \mathbf{W} \mathbf{G}_j$.

Dalším krokem je optimalizace skóre objektů, tj.

$$\mathbf{X}^* = (\mathbf{I} - \mathbf{M}_* \mathbf{u} \mathbf{u}^T \mathbf{W} / (\mathbf{u}^T \mathbf{M}_* \mathbf{W} \mathbf{u})) \mathbf{Z}, \quad (17)$$

kde $\mathbf{Z}$ je pomocná matice $\Sigma_j \mathbf{M}_j \mathbf{G}_j \mathbf{Y}_j^+$, ve které $\mathbf{Y}_j^+$ souhrně značí kvantifikace kategorií $\mathbf{Y}_j$ pro proměnné s vícenásobným nominálním škálováním a vektor hodnot $\mathbf{y}_j \mathbf{a}_j^{+T}$ pro jednoduché škálování. Ortonormální řešení pro optimalizované skóre objektů je nalezeno pomocí prokrustovské rotace, viz Cliff (1996).

1.4 Metody MDS a NMDS

Klasický Torgersonův euklidovský model (metrická metoda MDS), viz Torgerson (1958), předpokládá shodu euklidovských vzdáleností $d_{ii'}$ a měr nepodobnosti $\delta_{ii'}$ mezi i -tým a i' -tým objektem. Z matice $\mathbf{D}$ je možné odvodit souřadnice proměnných v redukovaném prostoru ve dvou krocích. V prvním kroku se matice $\mathbf{D}$ transformuje na matici skalárních součinů $\mathbf{B}$ (transformace vychází z kosinové věty), kterou lze maticově zapsat jako

$$\mathbf{B} = -\frac{1}{2} \left[\mathbf{I} - \frac{1}{n} \mathbf{i} \mathbf{i}^T \right] \mathbf{D}^2 \left[\mathbf{I} - \frac{1}{n} \mathbf{i} \mathbf{i}^T \right], \quad (18)$$

¹ Hodnoty hlavních komponent pro jednotlivé kategorie původních proměnných (*centroid coordinates*).

kde n označuje počet objektů, $\mathbf{I}$ je jednotková matice typu $n \times n$ a $\mathbf{i}$ je sloupcový vektor obsahující n jedniček. Ve druhém kroku je pak matice $\mathbf{B}$ rozložena pomocí spektrálního rozkladu na matici charakteristických čísel a charakteristických vektorů. Matice souřadnic $\mathbf{X}$ typu $n \times k$ může být vypočtena ze vztahu

$$\mathbf{X} = \mathbf{V}\mathbf{L}^{1/2}, \quad (19)$$

kde $\mathbf{L}$ označuje diagonální matici nenulových charakteristických čísel matice $\mathbf{B}$ a $\mathbf{V}$ matici odpovídajících charakteristických vektorů.

NMDS představuje metodu pro odhad souřadnic objektů za méně omezujících podmínek než Torgersonův metrický model. Rozdíl mezi metrickým a nemetrickým škálováním je ve způsobu zacházení s nepodobnostmi. V metrickém škálování se pracuje s jejich číselnými hodnotami, které vycházejí z euklidovského prostoru. V nemetrickém škálování se bere v úvahu pouze pořadí hodnot. Označme symbolem $\hat{d}_{ii'}$ odhad vzdálenosti mezi objekty o souřadnicích $\hat{x}_{ij}$ a $\hat{x}_{i'j}$ v obecném Minkowského prostoru, tj.

$$\hat{d}_{ii'} = \sqrt[k]{\sum_{j=1}^k |\hat{x}_{ij} - \hat{x}_{i'j}|^k}. \quad (20)$$

Základním předpokladem NMDS je zachování pořadí vzdáleností mezi body v konfiguraci $\delta_{ii'}$, tj.

$$\hat{d}_{ii'(1)} < \hat{d}_{ii'(2)} < \dots < \hat{d}_{ii'(n)}, \quad (21)$$

s pořadím nepodobností, tj.

$$\delta_{ii'(1)} < \delta_{ii'(2)} < \dots < \delta_{ii'(n)}. \quad (22)$$

Je zřejmé, že předpoklad zachování pořadí vzdáleností a nepodobností je mnohem mírnější než předpoklad shody jejich numerických hodnot. Nemetrické MDS tak oceníme hlavně v situacích, kdy přímé měření objektů není z jakéhokoli důvodu možné.

Nemetrický model hledá konfiguraci bodů v prostoru s cílem minimalizace tzv. *ztrátové funkce* pomocí iterativního algoritmu. Algoritmus začíná určením rozměrů matice souřadnic, pokračuje výpočtem tzv. disparit $\delta_{ii'}^2$, za kterými následuje výpočet odhadů souřadnic $\hat{x}_{ij}$ a vzdáleností mezi nimi $\hat{d}_{ii'}$. Na závěr se stanovují hodnoty ztrátové funkce. Celý proces končí, je-li dosaženo požadované hodnoty funkce stress, nebo nastaveného počtu iterací.

2 Disparita představuje funkční hodnotu vzdálenosti mezi objekty, tj. $\hat{\delta}_{ii'} = f(\hat{d}_{ii'})$. Jsou odhadovány tak, aby byly co nejbližší vzdálenostem $\hat{d}_{ii'}$ a přitom zachovávaly podmínku monotónnosti vzhledem k původním nepodobnostem.

V obecné rovině je definována ztrátová funkce podle J. Kruskala, který ji ve svém minimu nazval STRESS. Tuto funkci modifikoval a zavedl další funkce, které jsou minimalizovány, např. STRESS2:

$$STRESS2 = \sqrt{\frac{\sum_i \sum_{i'} (\hat{d}_{ii'} - \hat{\delta}_{ii'})^2}{\sum_i \sum_{i'} (\hat{d}_{ii'} - \hat{d}_{..})^2}}, \quad (23)$$

kde $\hat{d}_{..}$ odpovídá aritmetickému průměru všech odhadnutých vzdáleností mezi objekty v konfiguraci. Více k tomuto tématu viz Hebák (2007).

1.5 Fuzzy shluková analýza

Při shlukové analýze jsou objekty (v našem případě proměnné) přiřazovány do určitého počtu shluků. Obvykle se postupuje tak, že se objekty přiřadí do různých počtů shluků z určitého intervalu a následně se pomocí různých kritérií vyhodnocuje, který počet je výhodnější. Tato činnost většinou spočívá na uživateli, výjimečně je stanovení vhodného počtu shluků součástí metody shlukování.

Při fuzzy shlukové analýze se počítají stupně příslušnosti u_{ih} jednotlivých objektů k jednotlivým shlukům. Pro tyto stupně příslušnosti platí, že

$$\sum_{h=1}^k u_{ih} = 1 \text{ pro } i = 1, \dots, n \text{ a } \sum_{i=1}^n u_{ih} > 0 \text{ pro } h = 1, \dots, k,$$

kde u_{ih} hodnotí zařazení i -tého objektu k h -tému shluku a k je počet shluků. Jednou z výhod uvedeného přístupu je, že na základě stupňů příslušnosti lze pro každý shluk stanovit typický objekt (resp. objekty).

Existuje řada metod pro stanovení hodnot u_{ih} . Pro účely tohoto článku byla použita metoda FANNY, která je implementována v systému S-PLUS. Stupně příslušnosti se získají minimalizací účelové funkce

$$J_{FANNY} = \sum_{h=1}^k \frac{\sum_{i=1}^n \sum_{j=1}^n u_{ih}^2 u_{jh}^2 d_{ij}}{2 \sum_{j=1}^n u_{jh}^2}, \quad (24)$$

kde d_{ij} je známá (spočtená) vzdálenost (obvykle euklidovská) mezi i -tým a j -tým objektem.

Základním koeficientem pro stanovení vhodného počtu shluků v případě fuzzy shlukování, který je implementován v současném softwaru (systém S-PLUS), je *koeficient rozkladu*, vyjadřující míru překrytí mezi shluky. V systému S-PLUS je nazýván *Dunnův koeficient rozkladu*. Tento index se počítá podle vzorce

$$I_{DUNN} = \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^k u_{ih}^2 \quad (25)$$

a nabývá hodnot z intervalu $\langle 1/k; 1 \rangle$. Můžeme rozlišit dvě extrémní situace:

1. *úplné fuzzy shlukování*: všechna $u_{ih} = \frac{1}{k} \Rightarrow I_{DUNN} = nk \frac{1}{nk^2} = \frac{1}{k}$,
2. *pevné (disjunktní) shlukování*:
 pro jedno u_{ih} platí, že $u_{ih} = 1$,
 pro všechna ostatní: $u_{ih} = 0 \Rightarrow I_{DUNN} = \frac{n}{n} = 1$.

Ze zadaného intervalu počtu shluků je vybrán takový, pro který dosáhl Dunnův index nejvyšší hodnoty. Protože však se změnou počtu shluků se mění také interval, v němž se může hodnota koeficientu nacházet, je lépe porovnávat hodnoty *normalizovaného koeficientu rozkladu*, který je rovněž počítán v systému S-PLUS, a to podle vzorce

$$I'_{DUNN} = \frac{I_{DUNN} - \frac{1}{k}}{1 - \frac{1}{k}} = \frac{kI_{DUNN} - 1}{k - 1} \quad (26)$$

Uvedenou úpravou získáváme hodnoty z intervalu $\langle 0; 1 \rangle$. Pomocí Dunnova koeficientu lze také porovnávat různé metody, na jejichž základě byly shluky vytvořeny. Jako lepší je vyhodnocena ta metoda, pro kterou dosáhl Dunnův index vyšší hodnoty.

Kromě Dunnova koeficientu rozkladu jsme pro hodnocení získaných shluků použili obrysový koeficient, který souvisí s *obrysovým grafem*, rovněž využitým v tomto článku. Jak graf, tak koeficient lze získat v rámci systému S-PLUS, a to nejen pro fuzzy shlukovou analýzu. V obrysovém grafu jsou jednotlivé objekty znázorňovány pomocí obdélníků, jejichž šířka je dána hodnotou ψ_i z intervalu $\langle -1; 1 \rangle$, viz níže. V případě kladné hodnoty se obdélník zakresluje vpravo od osy Y , v případě záporné hodnoty pak vlevo. Zápornými hodnotami v případě fuzzy shlukové analýzy mohou být ohodnoceny objekty, pro které je míra příslušnosti přibližně stejná pro dva či více shluků. Je-li je ve shluku pouze jediný objekt, pak $\psi_i = 0$, a tudíž se žádný obdélník nezakresluje.

Výpočet hodnot ψ_i se provádí podle vzorce

$$\psi_i = \frac{\mu_i - \eta_i}{\max\{\eta_i, \mu_i\}} \quad (27)$$

kde η_i je průměrná vzdálenost i -tého objektu od ostatních objektů nacházejících se ve stejném shluku a μ_i je minimum z průměrných vzdáleností i -tého objektu od objektů

každého dalšího shluku, tj., $\eta_i = \sum_{j \in C_g} d_{ij} / (n_g - 1) \mu_i = \min_{h \neq g} \left(\sum_{j \in C_h} d_{ij} / n_h \right)$, kde C_g ,

resp. C_h , je označení g -tého, resp. h -tého shluku a n_g , resp. n_h , je počet objektů v g -tém, resp. h -tém shluku.

Obrysový koeficient je definován jako průměrná šířka obdélníků, tj. $\bar{\Psi}$. Vyšší hodnoty tohoto koeficientu vyjadřují lepší rozdělení objektů do shluků. Příklad viz např. Řezanková (2009).

1.6 Přístup CSPA

V praxi jsou řešeny případy, kdy jsou získána různá přiřazení objektů do skupin (například na základě různých skupin proměnných) a cílem je získat jeden výsledný způsob přiřazení. Za tímto účelem byly navrženy některé přístupy, viz Punera (2007). Pro kombinování přiřazení proměnných do shluků získaných na základě různých metod pro redukci dimenzionality jsme využili jeden z nich.

Předpokládejme, že bylo získáno r rozkladů (způsobů přiřazení proměnných do skupin). Označme symbolem $u_{ih}^{(q)}$ stupeň příslušnosti i -tého objektu k h -tému shluku v q -tém rozkladu. Pokud by tyto stupně příslušnosti byly pouze hodnoty 0 a 1, jednalo by se o pevné shlukování. Pro takové případy lze využít například *metodu CSPA* (*Cluster-based Similarity Partitioning Algorithm*), která je založena na matici podobnosti $\mathbf{A} = (1/r)\mathbf{U}\mathbf{U}^T$.

Pro kombinování výsledků fuzzy shlukové analýzy byla navržena *metoda sCSPA* (*soft version of CSPA*), viz Punera (2007). Při ní se jako matice podobnosti využívá buď matice $\mathbf{U}\mathbf{U}^T$, nebo matice vzdáleností vytvořená na základě euklidovských vzdáleností. Ve druhém zmíněném případě je vzdálenost mezi i -tým a j -tým objektem vypočtena podle vzorce

$$d_{ij} = \sqrt{\sum_{q=1}^r \sum_{h=1}^{k^{(q)}} \left(u_{ih}^{(q)} - u_{jh}^{(q)} \right)^2} . \quad (28)$$

Na základě vytvořené matice vzdáleností mohou být objekty shlukovány do k shluků vhodně zvolenou metodou. Při analýzách pro účely tohoto článku jsme aplikovali tento druhý uvedený způsob.

Pro kombinování výsledků v rámci našich analýz byla horizontálním spojením jednotlivých výsledných matic stupňů příslušnosti proměnných do shluků získaných pro jednotlivé metody vytvořena nová vstupní matice. Na jejím základě byla provedena fuzzy shluková analýza s využitím euklidovských vzdáleností, tj. vzorec (28). Pomocí ní byla získána jednak konečná matice stupňů příslušnosti, jednak jednoznačné přiřazené proměnných do skupin vždy na základě nejvyšší hodnoty příslušnosti.

2. Aplikace na reálná data a vyhodnocení

Metody popsané v sekci 1 jsme aplikovali na dva reálné datové soubory. Nejprve jsme na základě průzkumu nazvaného „*Výzkum percepce policisty*“ (výzkum z roku 2006, 356 respondentů, viz Moulisová, 2009) analyzovali soubor dat 24 proměnných charakterizujících typického policistu³. Tento soubor byl pracovně nazván „Policisté“. Respondenti vyjadřovali svá hodnocení na pětistupňové škále od 1 (nejpozitivnější hodnocení) do 5 (nejvíce negativní hodnocení).

Dále jsme na základě průzkumu nazvaného „*Aktivní životní styl vysokoškoláků (studentů FEL ČVUT v Praze)*“ (průzkum z roku 2008, 1453 respondentů, viz Valjent, 2010), dále jen „Studenti“, analyzovali soubor dat 15 proměnných vyjadřujících spokojenost s různými aspekty studentského života⁴. Respondenti vyjadřovali svou spokojenost na sedmistupňové škále od 1 (absolutně nespokojený) do 7 (absolutně spokojený).

V rámci metod FA a CATPCA byly pro oba datové soubory jako optimální vyhodnoceny čtyři komponenty (latentní proměnné, dimenze). Z důvodu porovnatelnosti jsme u všech ostatních metod zadávali analýzu rovněž pro čtyři komponenty. Abychom identifikovali skupiny podobných proměnných na základě výsledků (komponentních zátěží, faktorových zátěží), které jsme získali výše zmíněnými metodami, interpretovali jsme tyto výsledky pomocí fuzzy shlukové analýzy implementované v systému S-PLUS, viz Řezanková (2009), Řezanková (2010). Byly porovnávány výsledky shlukování do dvou, tří a čtyř shluků. Získané shluky byly hodnoceny pomocí obrysového koeficientu, Dunnova koeficientu rozkladu a jeho normovaného tvaru.

Při analýze proměnných ze souboru *Policisté* získala podle všech tří koeficientů nejlepší hodnocení metoda LCC při vytvoření dvou shluků (hodnota obrysového koeficientu 0,47, Dunnova koeficientu rozkladu 0,65 a jeho normovaného tvaru 0,3). Podle této metody by byly doporučeny dva shluky. Podle metod FA a CATPCA bychom mohli doporučit shluky čtyři. Při rozdělení proměnných do čtyř shluků se hodnocení metod lišilo v závislosti na použitém hodnotícím koeficientu, viz tabulka 1. Podle obrysového koeficientu byla nejlépe hodnocena metoda CATPCA (metoda normování *vprincipal*), podle Dunnova koeficientu rozkladu (jak základního, tak normovaného tvaru) metoda LCC. V pořadí třetí metodou byla faktorová analýza (metoda rotace *varimax*). Navržený postup analýzy se neosvědčil pro vícerozměrné škálování, kde pro čtyři shluky byly míry příslušnosti proměnných ke shlukům ve všech případech 0,25, takže hodnota Dunnova koeficientu rozkladu vyšla také 0,25, čemuž odpovídá nulová hodnota normovaného tvaru.

3 Dotazník byl zaměřen na zjišťování toho, jak je viděn typický policista a jaká je představa o jeho ideálních kvalitách. Respondenti netvořili homogenní a reprezentativní soubor. Šlo o 286 studentů Policejní akademie ČR, 48 studentů nepolicejních typů škol a 22 ostatních osob. V tomto výběrovém souboru bylo 177 mužů a 179 žen. Pokud jde o věk, pak medián byl 22 let, průměr 24,8 let, ale rozpětí bylo od 16 do 63 let.

4 Výběrový soubor obsahoval 1074 mužů a 312 žen, ve věku od 18 do 30 let s průměrem 22,1 let a mediánem 22 let. Šlo o studenty 1. až 6. ročníku magisterského studia (nejvíce z 2. ročníku – 650) a doktorandy.

Tabulka 1
Hodnocení přiřazení proměnných do 4 shluků získaných různými metodami pro soubor
Policisté

Metoda	Kritéria pro hodnocení výsledků shlukování		
	obrysový koeficient	Dunnův koeficient rozkladu	normovaný Dunnův koeficient
ALSCAL	0,21	0,25	0,00
DFactor	0,33	0,37	0,16
LCC	0,38	0,41	0,22
FA (metoda rotace: varimax)	0,38	0,33	0,11
CATPCA (metoda normování: vprincipal)	0,40	0,39	0,18
NMDS	0,19	0,25	0,00
PROXSCAL	0,02	0,25	0,00
Kombinace všech metod	0,43	0,46	0,28
Kombinace všech metod vyjma ALSCAL a PROXSCAL	0,43	0,46	0,28
Kombinace metod DFactor a LCC	0,53	0,56	0,41
Kombinace metod DFactor, CATPCA a NMDS	0,47	0,49	0,33

Pramen: Vlastní analýzy, data viz Moulisová (2009).

Tabulka 2
Hodnocení přiřazení proměnných do 4 shluků získaných různými metodami pro soubor
Studenti

Metoda	Kritéria pro hodnocení výsledků shlukování		
	obrysový koeficient	Dunnův koeficient rozkladu	normovaný Dunnův koeficient
ALSCAL	0,32	0,37	0,17
DFactor	0,48	0,52	0,36
LCC	0,45	0,54	0,39
FA (metoda rotace: varimax)	0,49	0,51	0,35
CATPCA (metoda normování: vprincipal)	0,46	0,50	0,33
NMDS	0,33	0,36	0,14
PROXSCAL	0,27	0,29	0,05
Kombinace všech metod	0,46	0,45	0,27
Kombinace všech metod vyjma ALSCAL a PROXSCAL	0,47	0,45	0,27
Kombinace metod DFactor a LCC	0,39	0,43	0,24
Kombinace metod CATPCA a NMDS	0,60	0,65	0,53
Kombinace všech metod vyjma DFactor a LCC	0,58	0,63	0,50
Kombinace metod DFactor, CATPCA a NMDS	0,59	0,66	0,54

Pramen: Vlastní analýzy, data viz Moulisová (2009).

Při analýze proměnných ze souboru *Studenti* byla na základě všech použitých metod nejlepší hodnocení dosažena při shlukování do čtyř shluků. Podle obrysového koeficientu byla dosažena nejvyšší hodnota u faktorové analýzy, za níž následují metody DFactor a CATPCA, podle Dunnova koeficientu byla nejlépe hodnocena metoda LCC, za níž následuje metoda DFactor, viz tabulka 2.

Dále jsme kombinovali výsledky různých skupin metod s cílem získat přiřazení proměnných do shluků na základě více metod současně. Za tímto účelem jsme aplikovali soft verzi algoritmu CSPA. Spojením matic stupňů příslušnosti získaných z fuzzy shlukové analýzy použité pro interpretaci výsledků jednotlivých metod jsme získali vstupní matici pro následnou analýzu opět pomocí fuzzy shlukování. Výsledné shluky jsme rovněž hodnotili podle již zmíněných koeficientů. Porovnávali jsme zejména výsledná shlukování do čtyř shluků. V tabulkách 1 a 2 jsou uvedeny nejzajímavější získané výsledky.

Na základě souboru *Policisté* byla vyhodnocena jako nejlepší kombinace metod DFactor a LCC (podle všech koeficientů), na základě souboru *Studenti* to byla podle obrysového koeficientu kombinace metod CATPCA a NMDS a podle Dunnova koeficientu ještě rozšíření této dvojice o DFactor (rozdíl v hodnotách je však pouhá jedna setina).

Přiřazení proměnných do shluků na základě zmíněných kombinací metod je znázorněno pomocí obrysového grafu na obrázcích 1 (soubor *Policisté*) a 2 (soubor *Studenti*). Skupinám proměnných lze přiřadit určité zobecnění. Z grafů lze také vyčíst, která proměnná nejvíce danou skupinu „reprezentuje“ (odpovídá jí nejvyšší hodnota koeficientu příslušnosti do daného shluku).

Obrázek 1

Zařazení proměnných ze souboru *Policisté* do čtyř shluků na základě metod DFactor a LCC

Pramen: Vlastní analýzy, data viz Moulisová (2009).

V souboru *Policisté* můžeme rozlišit skupinu týkající se především lidských vztahů (první shluk), v nichž je základní proměnná vyjadřuje názor na to, jak je typický policista přátelský. Druhý shluk zachycuje zejména sílu a aktivitu a základní proměnná vyjadřuje názor na to, jak je typický policista silný. Třetí skupina zachycuje vlastnosti související s tím, aby byly dodržovány zákony, a základní proměnná vyjadřuje názor na to, jak je typický policista odvážný. Čtvrtý shluk seskupuje morální vlastnosti a základní proměnná vyjadřuje názor na to, jak je typický policista férový.

Obrázek 2

Zařazení proměnných ze souboru *Studenti* do čtyř shluků na základě metod CATPCA a NMMDS

Pramen: Vlastní analýzy, data viz Valjent (2010).

V souboru *Studenti* můžeme v rámci čtyř shluků rozlišit skupinu proměnných, které se týkají fyzického vzhledu a zdraví. Tato skupina je reprezentována proměnnou týkající se spokojenosti se vzhledem. Druhá skupina se týká rodinného zázemí, kvality života, volného času a finanční situace. Reprezentuje ji proměnná týkající se kvality bydlení. Třetí skupina vyjadřuje spokojenost se sexem a s životním partnerem (spokojenosti jsou zřejmě velmi podobně hodnoceny) a čtvrtá se týká studia a studijních výsledků (i v tomto případě se odpovědi na tyto dvě otázky zřejmě dosti často shodují). Z obrázku 2 je dále zřejmé, že jedna proměnná je obtížně zařaditelná do některého ze čtyř shluků, i když je přiřazena k prvnímu z nich. Tato proměnná vyjadřuje spokojenost s úspěchem a uznáním od ostatních, což se může týkat i jiných oblastí, než je fyzický vzhled.

Závěr

Získané poznatky z analýzy pouze dvou souborů sice nelze zobecnit, mohou však být východiskem pro analýzy a směřování dalšího výzkumu. Shrňme-li zde uvedené poznatky, pak interpretace získaných hodnot komponentních (faktorových) zátěží, resp. souřadnic, pro jednotlivé metody pomocí fuzzy shlukové analýzy se neosvědčila u metod vícerozměrného škálování, pro které hodnotící koeficienty nabývaly nižších hodnot ve srovnání s ostatními metodami. Jako užitečné se však ukázaly kombinace výsledků metod vícerozměrného škálování a některých jiných metod.

Kombinací výsledků metod můžeme docílit vyšších hodnot sledovaných koeficientů, nemusí tomu však být vždy. Například u souboru Studenti hodnotící koeficienty pro kombinaci metod DFactor a LCC dosáhly nižších hodnot, než vyšly hodnoty koeficientů hodnotící výsledky získané pouze pomocí samostatných metod.

Z různorodosti výsledků je zřejmé, že při zpracování dat z průzkumů nelze spoléhat pouze na výsledky jedné metody. Analýza pomocí více metod by měla být aplikována zvláště v případě, kdy u některé z metod nejsou splněny předpoklady pro její použití, jak je tomu u faktorové analýzy, která je určena pro kvantitativní proměnné.

V dalším bádání bychom se chtěli zaměřit na vliv použitých měr závislosti na výsledné zařazení proměnných do skupin, na analýzu datových souborů s binárními a nominálními proměnnými a s proměnnými různých typů. Jako další zajímavý směr výzkumu se jeví zkoumání odlehklých proměnných, které v rámci fuzzy shlukové analýzy nejsou identifikovány, tj. v úvahu připadá využití novějších metod shlukování.

Literatura

- BARTHOLOMEW, D. J.; STEELE, F. et al. 2002. *The Analysis and Interpretation of Multivariate Data for Social Scientists*. Boca Raton : Chapman & Hall/CRC, 2002. ISBN 1-58488-295-6.
- CLIFF, N. 1996. Orthogonal rotation to congruence. *Psychometrika*. 1996, vol. 31, s. 33–42.
- DE LEEUW, J.; YOUNG, F. W.; TAKANE, Y. 1976. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*. 1976, vol. 41, s. 471–503.
- HÄRDLE, W.; SIMAR, L. 2007. *Applied Multivariate Statistical Analysis (Second edition)*. Berlin; Heidelberg; New York : Springer, 2007. ISBN 978-3-540-72243-4.
- HEBÁK, P.; HUSTOPECKÝ, J. et al. 2007. *Vícerozměrné statistické metody [3]*. 2. vyd. Praha : Informatorium, 2007. ISBN 978-80-7333-001-9.
- IBM SPSS Statistics, Help, Algorithms [online]. On-line manuál. [cit. 2011-01-16]. www.127.0.0.1:4004/help/index.jsp?topic=/com.ibm.spss.statistics.help/alg_introduction.htm.
- KAISER, H. F. 1985. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*. 1985, vol. 23, s. 187–200.
- MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. 1979. *Multivariate Analysis*. Duluth, London : Academic Press, 1979. ISBN 0-12-471252-5.
- MOULISOVÁ M. 2009. Výzkum percepce policisty. *Kriminalistika*. 2009, roč. 42, č. 1, s. 56–71.
- PUNERA, K.; GHOSH, J. 2007. Soft cluster ensembles. In OLIVEIRA, J. V.; PEDRYCZ, W. (eds.). *Advances in Fuzzy Clustering and Its Applications*. Chichester : John Wiley & Sons, 2007, s. 69–91. ISBN 978-0-470-02760-8.
- ŘEZANKOVÁ, H.; HÚSEK, D.; SNÁŠEL, V. 2009. *Shluková analýza dat*. 2. vyd. Praha : Professional Publishing, 2009. ISBN 978-80-86946-26-9.

- ŘEZANKOVÁ, H.; HÚSEK, D.; RYŠÁNKOVÁ, M. 2010. Grouping ordinal variables by using fuzzy cluster analysis. In *SMTDA 2010 – Stochastic Modeling Techniques and Data Analysis*. Chania : Swets, 2010, s. 83.
- TORGERSON, W. S. 1958. *Theory and Methods of Scaling*. New York : John Wiley & Sons, 1958.
- VALJENT, Z. 2010. *Aktivní životní styl vysokoškoláků (studentů FEL ČVUT v Praze)*. Praha : FEL ČVUT, 2010. ISBN 978-80-01-04669-2.
- VERMUNT, J. K.; MAGIDSON, J. 2002. Latent class cluster analysis. In HAGENAARS, J. A.; McCUTCHEON, A. L. (eds.). *Applied Latent Class Analysis*. Cambridge : Cambridge University Press, 2002, s. 89–106. ISBN 0-521-59451-0.
- VERMUNT, J. K.; MAGIDSON, J. 2005. *Technical Guide for Latent GOLD 4.0: Basic and Advanced* [online]. Statistical Innovations Inc., Belmont Massachusetts, 2005. [cit. 2011-01-16]. www.statisticalinnovations.com/products/LGtechnical.pdf.

COMPARISON OF DIMENSIONALITY REDUCTION METHODS APPLIED TO ORDINAL VARIABLES

Abstract: Questionnaire survey data are usually characterized by a great amount of ordinal variables. For multivariate analysis, it is suitable to reduce task dimensionality. The aim of this paper is a comparison of the results obtained by the analysis of data files with ordinal variables using selected methods for dimensionality reduction. The results are in the form of individual component values (e.g. factor loadings). For better interpretation and comparability, these component values were consequently analyzed by fuzzy clustering. On the basis of the obtained clusters of variables, we determined the optimal number of dimensions. We applied silhouette and Dunn's partition coefficients. Furthermore, we tried to merge the results received by individual methods on the basis of the sCSPA technique (soft version of cluster-based similarity partitioning algorithm). We considered groups of different methods and searched the best solution. The problems are illustrated by means of two real data files obtained from questionnaire surveys. We used SPSS, STATISTICA, Latent GOLD and S-PLUS systems.

Keywords: categorical principal component analysis, multidimensional scaling, latent class cluster models, discrete factor analysis, fuzzy cluster analysis

JEL Classification: C38