INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 25, číslo 1, březen 2014

POROVNÁNÍ VYBRANÝCH ALGORITMŮ PRO OHODNOCENÍ ODLEHLOSTI VÍCEROZMĚRNÝCH POZOROVÁNÍ

COMPARISON OF SELECTED ALGORITHMS FOR COMPUTING THE LOCAL OUTLIER SCORE IN MULTIDIMENSIONAL DATA

Vanda Vintrová, Tomáš Vintr, Hana Řezanková, Vladimír Úradníček

Adresa: Vanda Vintrová, Tomáš Vintr, Hana Řezanková, Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky, Katedra statistiky a pravděpodobnosti, nám. W. Churchilla 4, CZ-130 67 Praha 3

E-mail: vanda.vintrova@vse.cz, tomas.vintr@vse.cz, hana.rezankova@vse.cz

Adresa: Vladimír Úradníček, Univerzita Mateja Bela, Ekonomická fakulta, Katedra kvantitatívnych metód a informačných systémov, Tajovského 10, SK-975 90 Banská Bystrica

E-mail: vladimir.uradnicek@umb.sk

Abstrakt: V tomto článku detailně ověřujeme a hodnotíme správnost algoritmů navržených k určení stupně lokální odlehlosti vícerozměrných pozorování. Úroveň odlehlosti je vyjadřována pomocí faktoru lokální odlehlosti, který je založen na tzv. hustotě okolí každého pozorování, jež může být vyjádřeno jako bod ve vícerozměrném prostoru. Hustota v tomto významu je definována jako relativní četnost výskytu pozorování v podmnožinách datového souboru. Na základě původního algoritmu zvaného LOF (local outlier factor) a algoritmů od něj odvozených jsme vytvořili 45 variant možných výpočtů zkoumaného faktoru. Všechny algoritmy jsme porovnali v experimentu, který odhalil základní rozdíl mezi nimi a dovolil nám vyjádřit se k základním vlastnostem této třídy algoritmů.

Klíčová slova: Faktor lokální odlehlosti, odlehlá pozorování založená na hustotě, algoritmus LOF, algoritmus LOF', algoritmus LOF", mnohorozměrná data.

Abstract: In this paper, we rate and verify the correctness of the algorithms for computing the local outlier score. These are based on the density of the neighborhood of every observation that can be depicted as a point in the multidimensional space. The density is defined as the relative frequency of

the observations in the subsets of the dataset. On the basis of the original algorithm for local outlier detection, the LOF (Local Outlier Factor) algorithm, and the algorithms derived from it we created 45 possible computations of the analyzed factor. We compared the algorithms in an experiment that revealed the basic differences between them and enabled us to discuss the basic characteristics of this class of algorithms.

Keywords: Local outlier factor, density-based outlier detection, LOF algorithm, LOF' algorithm, LOF" algorithm, multivariate data.

1. Úvod

V reálných souborech dat se mohou vyskytovat odlehlá pozorování, jakožto objekty výrazně nekonzistentní s ostatními objekty ze stejného souboru dat. Některé statistické metody se na takovéto odlehlé objekty dívají jako na šum, který by měl být identifikován a odstraněn, aby nezkresloval výsledky. Nicméně odlehlé objekty mohou být také nositeli užitečných informací, proto je vhodné je podrobně prozkoumat. Existují různé přístupy k detekci odlehlých pozorování. V článku se zaměřujeme na algoritmy z oblasti data miningu založené na tzv. hustotě, které mohou odhalit lokální odlehlá pozorování.

První, kdo přišel s konceptem algoritmů k vyhledávání lokálních odlehlých objektů založených na hustotě, byl Breunig et al. [1], který také zavedl pojem *local outlier factor* (LOF), což je skóre určující stupeň odlehlosti určitého objektu. Srovnává lokální hustotu objektu (četnost výskytu objektů ve sledovaném okolí) s lokálními hustotami jeho k nejbližších sousedů. Existuje několik modifikací tohoto algoritmu. Obecně lze říci, že algoritmy, které vyhledávají odlehlé objekty na základě hustoty, přiřazují každému objektu ze souboru objektů hodnotu, která kvantifikuje hustotu okolí tohoto objektu, a pak na základě srovnání této hodnoty s obdobně vypočtenými hodnotami pro objekty v okolí tohoto objektu přiřadí každému objektu skóre, které určuje stupeň jeho odlehlosti.

Skóre je obvykle vypočteno jako poměr hustoty okolí objektu a průměrné hustoty okolí objektů v okolí zkoumaného objektu. Důležitým tématem článků zabývajících se v poslední době problematikou algoritmů pro vyhledávání odlehlých pozorování na základě hustoty byl pokus o určení smysluplného okolí objektu, které by se mělo srovnávat. Je nutné zmínit, že přestože dochází k rozvoji těchto metod, základní úloha nebývá správně formalizována, čímž se objevuje mnoho složitých konstruktů, jejichž podstata bývá často těžko odhalitelná.

Původní LOF [1] algoritmus byl mnohokrát upravován, tento článek se zaměří zejména na srovnání s LOF' [3] algoritmem, který zjednodušuje vý-

počet LOF algoritmu tak, aby byl snadněji pochopitelný s tím, že by měl mít srovnatelné výsledky, a na srovnání s LOF" [3] algoritmem, který zavádí další parametr pro výpočet okolí objektu, a rozlišuje tak mezi okolím objektu pro výpočet hustoty a okolím objektu pro porovnání hustoty okolí jednotlivých bodů v okolí objektu.

Existují dva základní přístupy, jak určit hustotu okolí. První z nich určí počet objektů, které leží v ryzím okolí $\mathcal{O}(\mathbf{x}_p)$ sledovaného objektu $\mathbf{x}_p$, $p = 1, \ldots, n$, přičemž poloměr okolí R je parametr algoritmů, viz [6], [8], a analýza je obvykle prováděna na všech n objektech souboru. Druhý přístup nalezne prvních k objektů ležících v okolí objektů $\mathbf{x}_p$ a poloměr určí jako vzdálenost (nejčastěji euklidovskou) mezi objektem $\mathbf{x}_p$ a jeho k-tým nejbližším sousedem, viz [1], [2], [3], [5], [10]. Jednoduše řečeno, první způsob pracuje s okolím, které má pevně daný poloměr, a sleduje počet objektů v tomto okolí, zatímco druhý způsob pracuje s okolím, které má pevně daný počet objektů, a sleduje jeho poloměr. V obou případech potřebujeme apriorní znalost souboru. V prvním případě je to minimální variační rozpětí shluku v datovém souboru a v druhém je to minimální četnost shluku v datovém souboru. Určení minimálního variačního rozpětí shluku považujeme za mnohem složitější než určení minimální četnosti shluku, kdy pouze deklarujeme, kde je naše hranice rozlišitelnosti, tedy jak malé shluky již nebudeme brát v potaz během analýzy datového souboru.

2. Výpočet faktoru odlehlosti

Budeme se zabývat algoritmy, které velikost okolí $\mathcal{O}(\mathbf{x_p})$ kolem zkoumaného objektu $\mathbf{x}_p$ určují na základě vzdálenosti pevně zvoleného k nejbližších (ve smyslu zvolené metriky) objektů. Základním algoritmem této třídy je LOF, od kterého se odvíjejí další modifikace jako například LOF' a LOF". Významným neduhem této třídy algoritmů je složitý způsob, jakým jsou algoritmy definovány, což vychází právě z definice LOF [1]. Tento neduh bychom rádi napravili rozborem existujících definic a navržením zobecnění a zjednodušení, které, jak doufáme, přispějí k dalšímu rozvoji těchto metod.

Na hodnotu faktoru odlehlosti se nahlíží tak, že čím je vyšší hodnota faktoru, tím je objekt $\mathbf{x}_p$ podezřelejší z odlehlosti. Jde o bezrozměrnou veličinu, ze které lze vyčíst, zda je okolí zkoumaného objektu relativně velké nebo malé. Zjednodušeně řečeno, jedná se o poměr mezi velikostí okolí zkoumaného objektu a velikostí běžného okolí okolních objektů. Vyjádříme-li velikost okolí jako velikost poloměru R_p (hyper)kulového okolí zkoumaného objektu $\mathbf{x}_p$ a poloměr běžného okolí objektů v okolí zkoumaného objektu $\overline{R}_p$, lze výpočet zapsat následovně:

$$OF_p = \frac{R_p}{\overline{R}_p}. \tag{1}$$

Jakým způsobem spočítat veličiny R_p a $\overline{R}_p$ se zabývají následující části této kapitoly.

2.1. Rozprava o poloměru

Poloměrem okolí sledovaného objektu $\mathbf{x}_p$ je v algoritmu LOF' charakteristika polohy množiny vzdáleností $\mathcal{D}_p = \{d(\mathbf{x}_p, \mathbf{x}_{p,i})\}_{i=1}^{k}$, kde $d(\mathbf{a}, \mathbf{b})$ je euklidovská vzdálenost objektů $\mathbf{a}$ a $\mathbf{b}$ a množina objektů $\{\mathbf{x}_{p,i}\}_{i=1}^{k}$ je množina k nejbližších objektů k objektu $\mathbf{x}_p$. V případě algoritmu LOF' je zvolenou charakteristikou polohy maximum.

V algoritmu LOF je výpočet složitější, poloměr se vypočítává jako aritmetický průměr hodnot $q_{p,i}$ množiny $\mathcal{Q}_p = \{q_{p,i}\}_{i=1}^{k}$,, které jsou buď vzdálenosti $d(\mathbf{x}_p, \mathbf{x}_{p,i})$ nebo vzdálenosti $d(\mathbf{x}_{p,i}, \mathbf{x}_{p,i,k})$ mezi objektem $\mathbf{x}_{p,i}$ a jeho k-tým nejbližším sousedem $\mathbf{x}_{p,i,k}$. Výběr prvků množiny $\mathcal{Q}_p$ se řídí následujícím předpisem:

$$q_{p,i} = \max\big(\{d(\mathbf{x}_p, \mathbf{x}_{p,i}), d(\mathbf{x}_{p,i}, \mathbf{x}_{p,i,k})\}\big).$$

Poloměr okolí objektů uvnitř shluku je tímto způsobem navyšován s cílem, aby hodnota faktoru OF_p uvnitř shluku příliš neklesala pod hodnotu 1. Zde je na místě zdůraznit, že algoritmus LOF byl původně stavěn pro datové soubory tvořené shluky s vícerozměrným rovnoměrným rozdělením.

Lze se domnívat, že, v případě jednoho velmi početného shluku vygenerovaného rovnoměrným rozdělením (bez šumu) při použití kvantilu nebo jiné charakteristiky polohy množiny $\mathcal{D}_p$ k určení poloměru, poloměr okolí objektu na okraji shluku musí být alespoň dvakrát větší než poloměr jeho k nejvzdálenějšího souseda. Vzhledem k této úvaze doporučujeme určit poloměr nejen pomocí charakteristiky polohy vzdáleností, ale také použít některou míru variability. Inspirovat se lze například Boxovým M-testem a použít determinant kovarianční matice $\mathbf{C}_p$ množiny objektů $\mathcal{K}_p = \left\{\{\mathbf{x}_{p,i}\}_{i=1}^{k}, \mathbf{x}_p\right\}$,

$$R_p = \det \mathbf{C}_p. \tag{2}$$

Hledání odlehlých pozorování je i dlouhodobým cílem robustní statistiky, kde je také používán determinant kovarianční matice [4], [9].

2.2. Rozprava o průměrném poloměru

Původní algoritmus LOF při použití našeho značení vypočítává faktor odlehlosti následovně:

$$OF_p = \frac{R_p}{\left(\frac{\sum_{i=1}^{k} R_{p,i}^{-1}}{k}\right)^{-1}}. \tag{3}$$

Z tohoto vzorce vyplývá, že autoři zvolili jako charakteristiku polohy množiny poloměrů okolí objektů v okolí zkoumaného objektu harmonický průměr

$$\overline{R}_p = \left(\frac{\sum_{i=1}^{k} R_{p,i}^{-1}}{k}\right)^{-1}, \tag{4}$$

aniž je tento přístup řádně vysvětlen. My to nepovažujeme za geometricky šťastné řešení, protože je velmi obtížné interpretovat veličinu R_p.

Pokud bychom chápali veličinu R_p pouze jako vzdálenost ke k nejbližšímu objektu od zkoumaného objektu, pak by měl být použit aritmetický průměr. V případě, že se jedná o poloměr (hyper)kulového okolí, a veličinou, ze které vycházíme, je hustota tohoto okolí, byl by výpočet $\overline{R}_p$ složitější.

Hustotu d_p okolí $\mathcal{O}(\mathbf{x}_p)$ zkoumaného objektu $\mathbf{x}_p$ lze intuitivně definovat jako poměr počtu objektů v okolí zkoumaného objektu a objemu tohoto okolí,

$$d_p = \frac{k}{C_m R_p^m}, \tag{5}$$

kde k je počet objektů v okolí $\mathcal{O}(\mathbf{x}_p)$, m je počet proměnných a $C_m R_p^m$ je objem okolí (m-rozměrná hyperkoule), $C_m = \frac{\pi^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}+1\right)}$. Chápeme-li faktor odlehlosti jako poměr mezi poloměrem okolí o hustotě d_p a poloměrem průměrně hustého okolí $\overline{d}_{p,i}$ při konstantním k a C_m, platí

$$OF_p = \frac{R_p}{\left(\frac{\sum_{i=1}^{k} R_{p,i}^{-m}}{k}\right)^{-1/m}}, \tag{6}$$

a tedy

$$\overline{R}_p = \left(\frac{\sum_{i=1}^{k} R_{p,i}^{-m}}{k}\right)^{-1/m}. \tag{7}$$

2.3. Určení faktoru odlehlosti

Na základě výše uvedených úvah lze sestrojit poměrně velké množství výpočtů faktoru odlehlosti OF_p, z nichž mnohé jsme prezentovali již v článku [12].

Poloměr R_p okolí objektu $\mathbf{x}_p$ může být spočítán pomocí charakteristik polohy z množiny $\mathcal{D}_p$, nebo z množiny $\mathcal{Q}_p$. Vhodnými charakteristikami polohy pro obě množiny mohou být minimum, maximum, medián a aritmetický průměr. Devátou možností, jak spočítat poloměr R_p, je námi navrhovaný výpočet determinantu kovarianční matice $\mathbf{C}_p$ množiny objektů $\mathcal{K}_p$.

Výpočet R_p lze kombinovat s výpočtem $\overline{R}_p$, který lze vypočítat buď podle původních algoritmů (LOF, LOF',...) jako harmonický průměr (4) nebo jako průměr (7), případně jako minimum, maximum nebo medián množiny $\{R_{p,i}\}_{i=1}^{k}$. Podotkněme, že maximum, resp. minimum $\{R_{p,i}\}_{i=1}^{k}$ je ekvivalentní s minimem, resp. maximem hustot, bez ohledu na to, zda jsou hustoty počítány jako funkce R^{-1} nebo R^{-m}.

Algorithm 1 Algoritmus pro výpočet meanQhean (původní LOF).

1: **for** $p = 1$ **to** n **do**
2: vytvoř množinu $\mathcal{X}_p$ k nejbližších objektů k $\mathbf{x}_p$
3: **end for**
4: **for** $p = 1$ **to** n **do**
5: vytvoř množinu $\mathcal{Q}_p = \{q_{p,i}\}$, $q_{p,i} = \max\left(\{d(\mathbf{x}_p, \mathbf{x}_{p,i}), d(\mathbf{x}_{p,i}, \mathbf{x}_{p,i,k})\}\right)$
6: spočítej R_p, $R_p = \frac{\sum_{i=1}^{k} q_{p,i}}{k}$
7: **end for**
8: **for** $p = 1$ **to** n **do**
9: spočítej $\overline{R}_p$, $\overline{R}_p = \left(\frac{\sum_{i=1}^{k} R_{p,i}^{-1}}{k}\right)^{-1}$
10: urči faktor odlehlosti OF_p, $OF_p = \frac{R_p}{\overline{R}_p}$
11: **end for**

Vzhledem k množství kombinací ($9 \cdot 5 = 45$), je nutné jednotlivé výpočty od sebe nějak systematicky odlišit. Pro účely tohoto článku jsme se rozhodli složit názvy jednotlivých kombinací z anglických zkratek funkcí, které v nich vystupují. Na prvním místě bude vystupovat malými písmeny zkratka zvolené charakteristiky polohy při výpočtu poloměru R_p, na druhém velkým písmenem název množiny, ze které se tato charakteristika polohy počítá a na třetím místě je malými písmeny zkratka charakteristiky polohy, podle které počítáme $\overline{R}_p$. Minimum budeme značit jako *min*, maximum jako *max*, medián jako *med*, aritmetický průměr jako *mean*, harmonický průměr jako *hean*, a průměr definovaný v (7) jako *nean*. Výpočet poloměru R_p pomocí determinantu kovarianční matice bude značeno s předponou *det*. Například původní LOF algoritmus bude značen jako *meanQhean* (Alg. 1), protože poloměr R_p

je vypočten jako aritmetický průměr z množiny $\mathcal{Q}_p$ a $\overline{R}_p$ je vypočten jako harmonický průměr. LOF' algoritmus bude značen jako *maxDhean*, protože poloměr R_p je vypočten jako maximum z množiny $\mathcal{D}_p$ a $\overline{R}_p$ je vypočteno jako harmonický průměr. Algoritmy, které budou k výpočtu poloměru R_p používat determinant kovarianční matice množiny $\mathcal{K}_p$, budou značeny jako $detK \ldots$ apod.

3. Experimenty

Různé algoritmy pro výpočet faktoru odlehlosti jsme porovnali pomocí simulační studie v prostředí MATLAB. Pomocí vícerozměrného normálního rozdělení se střední hodnotou $\mathbf{0}$ a se směrodatnou odchylkou každé proměnné $\sigma = \frac{1}{3}$ jsme vygenerovali dva soubory o rozsahu 1000 vektorů. První soubor byl dvourozměrný a druhý byl desetirozměrný. Do každého souboru jsme přidali 7 vektorů $\mathbf{v}_j$. První vektor $\mathbf{v}_0$ byl nulový vektor $\mathbf{v}_0 = \mathbf{0}$, ostatní vektory $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_6$ obsahovaly postupně pro jednu proměnnou hodnotu o velikosti $1\sigma, 2\sigma, ..., 6\sigma$. Nastavili jsme $k = 20$. V tabulkách 1–3 jsou uvedeny hodnoty faktoru odlehlosti těchto přidaných vektorů pro jednotlivé testované algoritmy. Čára v tabulce představuje hypotetickou hranici shluku, kdy vektory nad touto čarou hypoteticky náležejí do shluku. To znamená, že vektory, které pro některou proměnnou nabývají hodnoty větší než 3σ, jsou s vysokou pravděpodobností vektory odlehlými, a tudíž by měl být vypočtený faktor odlehlosti pro tyto vektory výrazně vyšší než u ostatních vektorů.

Je zjevné, že není snadné najít rozdíl mezi normálním a odlehlým pozorováním. Mezi navrženými variantami se objevily takové, kde je hranice mezi normálním pozorováním a hypotetickým odlehlým pozorováním výraznější. Nicméně ve většině případů jsou rozdíly pro vyšší dimenze těžce rozeznatelné. Je těžké rozhodnout, zda by měla být výrazná hranice pro odlehlá pozorování mezi 2σ a 3σ nebo mezi 3σ a 4σ vzdálenými vektory od středu shluku vygenerovaného vícerozměrným normálním rozdělením, ale je jisté, že by tato hranice měla být někde mezi 2σ a 5σ. Na první pohled mají tuto vlastnost algoritmy, které použily k výpočtu poloměru determinant kovarianční matice, ale je důležité poznamenat, že tyto míry jsou výpočetně velmi nestabilní.

Tabulka 1: Faktory odlehlosti přidaných vektorů ve dvourozměrných (2D) a desetirozměrných (10D) souborech.

j	maxDmax		maxDmin		maxDmed		maxDhean		maxDnean	
	2D	10D	2D	10D	2D	10D	2D	10D	2D	10D
0	0,88	0,81	1,09	0,95	1,00	0,88	0,99	0,88	0,99	0,89
1	0,83	0,78	1,32	1,08	1,06	0,90	1,04	0,91	1,05	0,94
2	0,66	0,85	1,64	1,26	1,23	1,03	1,18	1,03	1,20	1,09
3	0,89	0,87	2,83	1,38	1,80	1,10	1,79	1,11	1,86	1,17
4	1,68	0,92	4,33	1,62	2,82	1,30	2,89	1,25	3,00	1,34
5	1,56	1,15	5,55	1,93	3,86	1,58	3,82	1,56	3,95	1,68
6	3,01	1,39	8,40	2,42	4,96	1,93	5,18	1,90	5,36	2,04

j	minDmax		minDmin		minDmed		minDhean		minDnean	
	2D	10D	2D	10D	2D	10D	2D	10D	2D	10D
0	0,13	0,66	1,00	1,26	0,19	0,81	0,27	0,86	0,33	1,03
1	0,87	0,55	6,40	1,15	2,07	0,67	2,36	0,74	2,82	0,93
2	0,41	0,77	2,11	1,75	1,33	0,99	1,27	1,07	1,36	1,40
3	0,32	0,89	3,47	1,80	1,70	1,26	1,79	1,30	2,01	1,48
4	2,56	1,00	19,98	2,12	5,99	1,55	6,89	1,59	7,79	1,79
5	2,78	0,90	29,70	1,56	15,91	1,29	16,70	1,28	18,71	1,36
6	6,26	1,50	46,80	2,98	21,84	2,07	23,53	2,23	26,89	2,53

j	medDmax		medDmin		medDmed		medDhean		medDnean	
	2D	10D	2D	10D	2D	10D	2D	10D	2D	10D
0	0,80	0,83	1,32	1,02	1,03	0,91	1,01	0,90	1,02	0,91
1	0,98	0,76	1,81	1,06	1,26	0,87	1,30	0,89	1,32	0,93
2	0,57	0,85	1,49	1,27	1,06	1,04	1,08	1,03	1,11	1,09
3	0,70	0,94	2,92	1,38	1,72	1,13	1,74	1,14	1,83	1,20
4	1,77	0,91	5,04	1,64	3,52	1,31	3,40	1,27	3,55	1,37
5	1,76	1,13	6,76	2,03	4,71	1,60	4,49	1,57	4,63	1,72
6	3,65	1,41	11,86	2,58	5,92	2,02	6,45	1,98	6,78	2,14

j	meanDmax		meanDmin		meanDmed		meanDhean		meanDnean	
	2D	10D	2D	10D	2D	10D	2D	10D	2D	10D
0	0,75	0,79	1,10	1,01	0,96	0,86	0,95	0,87	0,96	0,89
1	0,88	0,76	1,54	1,09	1,15	0,86	1,17	0,90	1,19	0,94
2	0,58	0,84	1,48	1,28	1,15	1,02	1,11	1,02	1,14	1,09
3	0,66	0,92	2,69	1,37	1,63	1,12	1,63	1,14	1,70	1,20
4	1,88	0,91	5,32	1,63	3,41	1,31	3,41	1,28	3,56	1,37
5	1,77	1,13	6,90	1,97	5,13	1,58	4,91	1,55	5,08	1,68
6	3,78	1,42	11,94	2,57	6,17	2,00	6,69	1,99	6,98	2,15

Tabulka 2: Faktory odlehlosti přidaných vektorů ve dvourozměrných (2D) a desetirozměrných (10D) souborech (pokračování).

j	maxQmax		maxQmin		maxQmed		maxQhean		maxQnean	
	2D	10D	2D	10D	2D	10D	2D	10D	2D	10D
0	0,81	0,79	1,02	0,97	0,99	0,85	0,97	0,86	0,97	0,87
1	0,91	0,85	1,27	1,13	1,00	0,95	1,03	0,95	1,03	0,98
2	0,76	0,76	1,78	1,21	1,30	0,96	1,29	0,97	1,33	1,03
3	1,00	0,92	2,48	1,20	1,13	1,03	1,44	1,04	1,51	1,07
4	1,00	1,00	3,24	1,45	1,89	1,16	2,08	1,18	2,16	1,23
5	0,97	0,94	3,65	1,62	3,16	1,37	2,81	1,35	2,89	1,44
6	3,01	1,35	5,53	1,96	4,04	1,62	3,84	1,64	3,91	1,70

j	minQmax		minQmin		minQmed		minQhean		minQnean	
	2D	10D	2D	10D	2D	10D	2D	10D	2D	10D
0	0,96	1,05	1,00	1,05	1,00	1,05	0,99	1,05	0,99	1,05
1	0,90	0,92	1,12	1,00	1,12	0,94	1,05	0,96	1,06	0,96
2	0,72	0,86	1,19	1,09	1,16	1,00	1,08	1,01	1,08	1,03
3	0,76	0,97	1,89	1,35	1,25	1,14	1,33	1,15	1,36	1,19
4	1,63	0,92	3,98	1,64	2,36	1,33	2,54	1,28	2,62	1,37
5	2,12	1,13	5,98	1,68	4,58	1,45	4,46	1,45	4,56	1,53
6	3,72	1,43	8,83	2,20	5,42	1,82	5,92	1,86	6,09	1,95

j	medQmax		medQmin		medQmed		medQhean		medQnean	
	2D	10D	2D	10D	2D	10D	2D	10D	2D	10D
0	0,89	0,88	1,01	0,99	0,99	0,94	0,98	0,94	0,98	0,94
1	0,89	0,89	1,15	1,07	1,01	1,00	1,04	0,98	1,04	0,99
2	0,65	0,86	1,30	1,11	1,07	0,99	1,04	0,99	1,05	1,01
3	0,79	0,93	2,18	1,21	1,42	1,10	1,48	1,08	1,53	1,11
4	1,68	0,93	4,41	1,45	2,82	1,28	2,90	1,23	3,00	1,29
5	1,76	1,10	5,12	1,74	3,84	1,49	3,86	1,47	3,96	1,54
6	3,33	1,40	9,35	2,31	5,10	1,90	5,52	1,87	5,73	1,97

j	meanQmax		meanQmin		meanQmed		meanQhean		meanQnean	
	2D	10D	2D	10D	2D	10D	2D	10D	2D	10D
0	0,91	0,88	1,02	0,98	0,99	0,93	0,98	0,94	0,98	0,94
1	0,90	0,89	1,17	1,05	1,03	0,96	1,04	0,96	1,04	0,97
2	0,68	0,85	1,34	1,12	1,11	1,00	1,09	0,99	1,10	1,02
3	0,83	0,93	2,18	1,22	1,30	1,08	1,46	1,07	1,52	1,10
4	1,59	0,92	4,13	1,42	2,57	1,25	2,66	1,21	2,75	1,26
5	1,69	1,11	4,93	1,68	3,80	1,46	3,73	1,45	3,81	1,52
6	3,41	1,39	8,44	2,21	4,89	1,84	5,27	1,84	5,43	1,92

Tabulka 3: Faktory odlehlosti přidaných vektorů ve dvourozměrných (2D) a desetirozměrných (10D) souborech (pokračování).

j	detKmax		detKmin		detKmed		detKhean		detKnean	
	2D	10D	2D	10D	2D	10D	2D	10D	2D	10D
0	0,44	0,01	1,41	0,57	1,01	0,09	1,01	0,14	1,05	0,44
1	0,52	0,04	4,03	5,97	1,48	0,46	1,79	1,02	2,10	4,43
2	0,16	0,02	4,22	17,39	1,67	0,50	1,84	2,28	2,18	12,90
3	0,48	0,07	12,65	15,47	2,21	1,02	3,53	2,63	4,77	11,52
4	1,72	0,04	54,57	34,77	6,55	7,69	11,34	13,67	17,19	28,03
5	1,39	0,36	47,42	645,02	23,83	68,16	23,44	117,10	26,55	478,17
6	10,87	7,91	245,56	8871,64	31,46	55,93	56,11	749,13	82,56	6575,11

4. Závěr

Provedli jsme experiment s různými variantami výpočtu faktoru lokální odlehlosti pro vícerozměrné pozorování z datového souboru. Geometricky složitější experiment pro množinu objektů se dvěma proměnnými jsme prezentovali a komentovali v [12].

Původní algoritmy *meanQhean* (LOF) a *maxDhean* (LOF') jsou srovnatelné, *maxDhean* je trochu rychlejší a *meanQhean* vykazuje lepší výsledky pro vektory na okraji shluku. Jsme přesvědčeni, že LOF by měl být nadefinován jako *neanQnean*, což je geometricky mnohem elegantnější řešení, navíc tento algoritmus zvyšuje hodnoty faktoru odlehlosti pro odlehlá pozorování, čímž je zvýrazní. Jak je zdůrazněno v [7], je tato vlastnost velmi žádoucí.

Lze říci, že užití množin $\mathcal{Q}_p$ ve srovnání s užitím množin $\mathcal{D}_p$ sníží rozptyl hodnot faktoru kolem hodnoty 1. Algoritmy, které pro výpočet používají determinanty kovariančních matic množin $\mathcal{K}_p$, vykazují velmi pěkné výsledky, přičemž nedochází k tak očividnému stírání rozdílu mezi vnitřním bodem a odlehlým bodem při vyšší dimenzi úlohy.

Vzhledem k tomu, že na základě výsledků experimentů je patrno, že hodnoty faktoru odlehlosti jsou jen nepatrně ovlivněny způsobem, jak je vypočten R_p, navrhujeme používat výpočetně méně nákladné kvantily množin $\mathcal{D}_p$. Vhodnými se jeví nízké kvantily při užití vysokého k. Zejména pokud předpokládáme, že soubor obsahuje hodně šumu a relativně řídké shluky, je důležité nastavit parametr k vysoký a použít nízké kvantily množin $\mathcal{D}_p$, což nelze v původním algoritmu LOF nastavit. Podobná myšlenka je součástí LOF" algoritmu.

Mnohem zásadnější je způsob výpočtu $\overline{R}_p$. Pokud chce výzkumník nalézt pouze výrazně odlehlá pozorování s nízkou pravděpodobností označení po-

zorování chybně jako odlehlé, potom by měl zvolit pro výpočet $\overline{R}_p$ vysoký kvantil množiny $\{R_{p,i}\}$ v okolí zkoumaného vektoru.

Pokud chceme označit za neodlehlá pozorování pouze taková, jejichž okolí je významně hustší, nebo pokud je snahou minimalizovat pravděpodobnost, že pozorování bude chybně označeno za neodlehlé, potom je vhodné počítat $\overline{R}_p$ jako nízký kvantil množiny $\{R_{p,i}\}$ v okolí zkoumaného vektoru.

V další práci se zaměříme na přípravu souborů dat pro shlukování, kde použijeme faktory odlehlosti jako váhy vstupující do algoritmů shlukové analýzy. Takovouto aplikaci by bylo lze označit jako robustní shlukovou analýzu. S tím souvisí nalezení způsobu, jak kvalitativně porovnat různé metody určování faktoru lokální odlehlosti [11].

Poděkování

Tento článek byl připraven za podpory projektu IGA F4/6/2012.

Literatura

[1] Breunig M. M., Kriegel H. P., Ng R. T., Sander J. et al.: LOF: identifying density-based local outliers. *Sigmod Record*, 29(2): 93–104, 2000. http://www.it.iitb.ac.in/~deepak/deepak/courses/mtp/papers/LOF-identifying%20density-based%20local%20outliers.pdf

[2] Cao H., Si G., Zhu W., Zhang Y.: Enhancing effectiveness of density-based outlier mining. In *International Symposiums on Information Processing (ISIP), 2008*, 149–154. IEEE, 2008. ISBN 978-0-7695-3151-9. doi: 10.1109/ISIP.2008.67

[3] Chiu A. L., Fu A. W.: Enhancements on local outlier detection. In *Seventh International Database Engineering and Applications Symposium, 2003. Proceedings*, 298–307. IEEE, 2003. ISBN 0-7695-1981-4. http://www.cse.cuhk.edu.hk/~adafu/Pub/ideas03-lm.ps

[4] Davies L., Gather U.: The identification of multiple outliers. *Journal of the American Statistical Association*, 88(423), 782–792, 1993. doi: 10.1080/01621459.1993.10476339

[5] Jin W., Tung A., Han J., Wang W.: Ranking outliers using symmetric neighborhood relationship. *Advances in Knowledge Discovery and Data Mining*, 577–593, 2006. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.66.9380&rep=rep1&type=pdf

[6] Knorr E., Ng R.: Finding intensional knowledge of distance-based outliers. In *Proceedings of the International Conference on Very Large Data Bases*, 211–222, 1999. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.45.9005&rep=rep1&type=pdf

[7] Kriegel H. P., Kröger P., Sander J., Zimek A.: Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3): 231–240, 2011. doi: 10.1002/widm.30

[8] Papadimitriou S., Kitagawa H., Gibbons P. B., Faloutsos C.: LOCI: Fast outlier detection using the local correlation integral. In *19th International Conference on Data Engineering, 2003. Proceedings*, 315–326. IEEE, 2003. http://www.bitquill.net/pdf/loci_icde03.pdf doi: 10.1109/ICDE.2003.1260802

[9] Rousseeuw, P. J., Driessen, K. V.: A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3), 212–223, 1999. doi: 10.1080/00401706.1999.10485670

[10] Tang J., Chen Z., Fu A., Cheung D.: Enhancing effectiveness of outlier detections for low density patterns. In *Advances in Knowledge Discovery and Data Mining*, 535–548, 2002. ISBN 3-540-43704-5. http://www.cse.cuhk.edu.hk/~adafu/Pub/pakdd02.pdf

[11] Vintr T., Vintrová V.: Use of local outlier factors as weights in fuzzy clustering. *Forum Statisticum Slovakum*, 54(6), 171–176, 2013. ISSN 1336-7420. http://www.ssds.sk/casopis/archiv/2013/fss0613.pdf#page=173

[12] Vintrová V., Vintr T., Řezanková H.: Comparison of different calculations of the density-based local outlier factor. In *IMMM 2012, The Second International Conference on Advances in Information Mining and Management*, 60–67. IARIA, 2012. http://www.thinkmind.org/index.php?view=article&articleid=immm_2012_3_30_20089

METODY PRO REDUKCI DIMENZE V MNOHOROZMĚRNÉ STATISTICE A JEJICH VÝPOČET

STANDARD MULTIVARIATE STATISTICAL METHODS FOR DIMENSION REDUCTION

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

E-mail: kalina@euromise.cz

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.

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.

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 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 supervidované a nesupervidované. Supervidovanými jsou takové, které jsou určeny pro

Tabulka 1: Přehled statistických metod pro redukci dimenze.

Metoda	Vlastnosti	
Analýza hlavních komponent	Lineární	Nesupervidovaná
Korespondenční analýza	Lineární	Nesupervidovaná
Mnohorozměrné škálování	Nelineární	Nesupervidovaná
Faktorová analýza	Lineární	Nesupervidovaná
Lineární diskriminační analýza	Lineární	Supervidovaná

Tabulka 2: Přehled metod shlukové analýzy.

Metoda	Vlastnost
Hierarchická shluková analýza	Nesupervidovaná
k průměrů (k-means)	Nesupervidovaná
k nejbližších sousedů (k-nearest neighbour)	Supervidovaná

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 nesupervidovaných metod není vhodná pro redukci dimenze 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 provádějí 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. To může napravit potíže se spolehlivostí závěrů následné statistické analýzy.

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, patří i kanonická korelační analýza, která zkoumá lineární vztah mezi dvěma skupinami proměnných, nebo shluková analýza, která provádí průzkum dat s cílem najít nějaké jejich shluky [16, 20, 30]. Na rozdíl od metod popsaných v tomto článku tedy shluková analýza 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 v analýze obrazu [19]. Z grafických metod pro exploratorní analýzu dat stojí za zmínku *biplot*, zobrazující současně data i proměnné a bývá použit např. pro interpretaci výsledků analýzy hlavních komponent [22].

V článku budeme používat následující značení. 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 označíme

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

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

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

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

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

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

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

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 $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times p}$ 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 $\mathbf{A}\mathbf{A}^T$ (i $\mathbf{A}^T\mathbf{A}$). Jsou tedy vždy reálná a kladná. Je konvencí uvádět singulární čísla matice $\mathbf{\Sigma}$ tak, aby byla sestupně uspořádaná. 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, kdežto 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 bázím. 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 matici $\mathbf{A}$. 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 úložného 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 [10, 16].

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

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

kde $\overline{\mathbf{X}}$ 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 jejích 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 obdobu 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

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

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.

Vektory marginálních pravděpodobností označíme jako

$$\mathbf{c} = (p_{\cdot 1}, \ldots, p_{\cdot J})^T \,, \quad \mathbf{r} = (p_{1\cdot}, \ldots, p_{I\cdot})^T \tag{20}$$

a definujeme řádkové a sloupcové profily jako

$$\mathbf{r}_i = \left(\frac{n_{i1}}{n_{i\cdot}}, \ldots, \frac{n_{iJ}}{n_{i\cdot}} \right)^T = \left(\frac{p_{i1}}{p_{i\cdot}}, \ldots, \frac{p_{iJ}}{p_{i\cdot}} \right)^T , \tag{21}$$

$$\mathbf{c}_j = \left(\frac{n_{1j}}{n_{\cdot j}}, \ldots, \frac{n_{Ij}}{n_{\cdot j}} \right)^T = \left(\frac{p_{1j}}{p_{\cdot j}}, \ldots, \frac{p_{Ij}}{p_{\cdot j}} \right)^T . \tag{22}$$

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 profily $\mathbf{r}_i$ a marginálními sloupcovými pravděpodobnostmi $\mathbf{c}$. Stejně tak lze na hodnotu χ^2 nahlížet jako na vážený průměr χ^2 vzdáleností mezi sloupcovými profily $\mathbf{c}_j$ a marginálními řádkovými pravděpodobnostmi $\mathbf{r}$.

V rámci redukce mnohorozměrného prostoru lze vzorec (18) s využitím toho, že čísla $p_{ij} - p_{i\cdot} p_{\cdot j}$ jsou prvky matice $\mathbf{P} - \mathbf{rc^T}$, 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 [22, 25].

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]. Jako výhodu metod mnohorozměrného škálování lze uvést i jejich schopnost odhalit shluky v datech.

6. Faktorová analýza

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

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

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 supervidovanou 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 $\overline{\mathbf{X}}_k$ a celkový průměr napříč skupinami jako $\overline{\mathbf{X}}$. 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 [8, 18]. Matice $\mathbf{B}$ o rozměrech $p \times p$ je definovaná jako

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

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

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

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 $\mathbf{v}^T\overline{\mathbf{X}}_1 > \mathbf{v}^T\overline{\mathbf{X}}_2$. Lineární diskriminační analýza klasifikuje pozorování $\mathbf{Z}$ do skupiny 1 právě tehdy, když

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

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

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

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 $\mathbf{S}^{-1}\mathbf{B}$ [10].

Poděkování

Práce vznikla za finanční podpory Nadačního fondu Neuron na podporu vědy. Druhý autor byl podpořen grantem GA13-06684S 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. In Pokorný J., Šaloun P., Paralič J., Horváth T. (Eds.): *Datakon a Znalosti 2013, Part II.* VŠB-Technická univerzita, Ostrava, 34 – 43.

[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 ze serveru `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.

OPTIMÁLNÍ ABSOLUTNÍ MOMENTY
OPTIMAL ABSOLUTE MOMENTS

Václav Čermák

Adresa: Severozapadní II/18, 141 00 Praha 4

Abstrakt: Uvažujme distribuční funkci F, pro kterou platí $\int |x|^r \,\mathrm{d}F(x) < 1$ pro dané $r > 0$. Optimální absolutní moment (r-tého řádu) je definován jako takové číslo A_r, pro které je výraz $\int |x - A_r|^r \,\mathrm{d}F(x)$ minimální. Jsou zkoumány vlastnosti optimálních absolutních momentů.

Klíčová slova: Optimální aboslutní moment, distribuční funkce.

Abstract: Let F be a distribution function for which $\int |x|^r \,\mathrm{d}F(x) < 1$ holds for a given $r > 0$. Optimal absolute moment (of the order r) is defined as the real number A_r which minimizes the integral $\int |x - A_r|^r \,\mathrm{d}F(x)$. The current paper deals with properties of optimal absolute moments.

Keywords: Optimal absolute moment, distribution function.

The absolute moments are usually defined – cf. *Kotz and Johnson (1982)* and/or some textbooks on the theory of statistics, e.g., *Stuart and Ord (1987)* – by means of the formula

$$\int_{-\infty}^{\infty} |x - \mu_1'|^r \,\mathrm{d}F(x), \quad r = 1, 2, \ldots, \tag{1}$$

i.e., as the moments about the expectation μ_1'. Only in the case $r = 1$, can another central value be used, namely the median $X_{0.5}$. Then, the mean absolute deviation

$$\int_{-\infty}^{\infty} |x - X_{0.5}|^r \,\mathrm{d}F(x) \tag{2}$$

has some advantages: its use in constructing the measures of relative variability and measures of skewness leads to standardised and well-behaved measures (functions of the shape parameters).

As well-established, the choice of the median $X_{0.5}$ instead of the expectation μ_1' minimalizes the mean absolute deviation. This idea can also be transposed to the absolute moments of higher orders, i.e. to the cases $r = 2, 3, \ldots$ Thus let us define the optimal absolute moments by means of the formula

$$\nu_r^* = \int_{-\infty}^{\infty} |x - A_r|^r \,\mathrm{d}F(x), \tag{3}$$

where A_r denotes the quantity minimizing the functional (3). Of course, in the case $r = 2$, we have $A_2 = \mu_1'$ and, therefore, ν_r^* turns into the variance μ_2.

Another interesting case is where $r = 3$. This is useful in constructing a moment-based measure of skewness, μ_3^*/ν_3^*, where $\mu_3^* = \int_{-\infty}^{\infty} (x - A_3)^3 \, \mathrm{d}F(x)$. It can be shown that the optimal absolute moment ν_3^* can be expressed in a simple form, despite the fact that the explicit expression of A_3 is somewhat complicated.

Example: let X be a continuous random variable with the power distribution, $F(x) = x^c, f(x) = cx^{c-1}, c > 0, 0 \leq x \leq 1$. Then $\nu_3^* = \frac{c}{c+3}(1 - A_3)^3$, where A_3 is the root of the equation

$$\frac{12}{(c+1)(c+2)} A_3^{c+2} - 3A_3^2 + \frac{6c}{c+1} A_3 - \frac{3c}{c+2} = 0.$$

As regards robust methods of statistics, a very interesting case is $r = 1.5$. This also leads to unambiguous and simple solutions. For example, in the case of power distribution with $c = 2$, i.e. $f(x) = 2x$, we get $\nu_{1.5}^* = \frac{4}{7}(1 - A_{1.5})^{1.5}$, where $A_{1.5} \doteq 0.68433$. Further, $\mu_{1.5}^*/\nu_{1.5}^* = 1 - 0.8A_{1.5}^2 - 1.2A_{1.5} \doteq -0.196$.

The optimal absolute moments have the folowing advantages:

(a) While the coefficient of variation, constructed as the ratio of the mean deviation about μ_1', and of the mean, take values within the interval $(0, 2)$, the analogous coefficient based on the mean deviation about the median $X_{0.5}$, takes values within the interval $(0, 1)$. This latter also has a smaller sampling variability.

(b) The family of skewness coefficients, the general formula of which is

$$Sk = \frac{\mu_r^*}{\nu_r^*} = \frac{-\int_{-\infty}^{A_r} (A_r - x)^r \mathrm{d}F + \int_{A_r}^{\infty} (x - A_r)^r \mathrm{d}F}{\int_{-\infty}^{A_r} (A_r - x)^r \mathrm{d}F + \int_{A_r}^{\infty} (x - A_r)^r \mathrm{d}F},$$

has very good properties both from the point of view of descriptive statistics and of inferential statistics.

Reference

[1] S. Kotz, N. L. Johnson – eds.: *Encyclopedia of Statistical Sciences*, Vol. 1. New York, Wiley, 1982

[2] A. Stuart, J. K. Ord: *Kendall's Advanced Theory of Statistics, Vol. 1: Distribution Theory*, 5th ed. London, Griffin, 1987

ZPRÁVA O ČINNOSTI ČESKÉ STATISTICKÉ SPOLEČNOSTI V ROCE 2013

1. Základní údaje o Společnosti

Dne 31. ledna 2013 se na Vysoké škole ekonomické v Praze konalo valné shromáždění České statistické společnosti, na němž byl zvolen nový výbor Společnosti na dvouleté funkční období. Předsedkyní Společnosti byla zvolena prof. Řezanková z katedry statistiky a pravděpodobnosti Vysoké školy ekonomické v Praze. Na první schůzi výboru ČStS, která se konala 27. února 2013, byli zvoleni další členové předsednictva, a to prof. Dohnal (FS ČVUT v Praze) – místopředseda, prof. Picek (TU Liberec) – vědecký tajemník, Ing. Löster (VŠE Praha) – hospodář. Dále výbor ustavil funkci odpovědného redaktora časopisu Informační Bulletin České statistické společnosti, do které byl zvolen Mgr. Vencálek (PřF UP Olomouc). Činnost technického redaktora tohoto časopisu stejně jako v předchozích letech vykonával Ing. Pavel Stříž, který v roce 2013 zabezpečoval též žádosti o registrace časopisu, viz níže. V roce 2013 dva členové ČStS požádali o ukončení členství a tři zájemci se přihlásili, takže ke konci roku měla Česká statistická společnost 233 členů. Ostatní údaje se vzhledem k předchozímu roku nezměnily, kromě průměrného věku, který se mírně zvýšil na 52 let.

2. Činnost výboru Společnosti

Celkem se v roce 2013 konala tři zasedání výboru ČStS (v únoru, červnu a listopadu). Kromě toho řada důležitých záležitostí byla diskutována prostřednictvím elektronické pošty. Jednou oblastí činnosti bylo „zviditelnění“ ČStS. Výbor realizoval vytvoření a spuštění nových webových stránek, a to jak v českém, tak anglickém jazyce. Jejich adresy jsou `www.statspol.cz/cs/` a `www.statspol.cz/en/`. Do české verze byl umístěn on-line formulář sloužící jako přihláška za člena Společnosti. Na Wikipedii byl založen článek o České statistické společnosti. Několik informací o aktivitách ČStS bylo publikováno v časopise *Statistika & My*. Členové výboru zastupovali ČStS na konferencích na Slovensku a na schůzce představitelů národních statistických společností V6 v Lublani. V červenci 2013 byla dojednána Dohoda o bezúplatném užívání nebytových prostor mezi Českým statistickým úřadem a Českou statistickou společností, kterou podepsala paní předsedkyně ČSÚ prof. Ritschelová a předsedkyně ČStS prof. Řezanková. Díky této dohodě bude mít ČStS po dobu 8 let svoji kancelář, a to místnost 126 v prostorách knihovny ČSÚ. Součásti dohody je spolupráce v oblasti statistiky těchto dvou institucí, zaměřené

zejména na spolupráci ve výzkumu a vývoji zaměřeném do oblasti statistiky, spolupráci při vydávání odborných periodik a společnou organizaci konferencí a odborných seminářů. ČStS se zavázala předat Českému statistickému úřadu svůj knihovní fond, který se stane součástí Ústřední statistické knihovny.

Byla vydána dvě čísla časopisu Informační Bulletin České statistické společnosti, v nichž byly zohledněny formální požadavky, aby se časopis mohl ucházet o zařazení do Seznamu recenzovaných neimpaktovaných periodik vydávaných v České republice (dále „Seznam"). IB ČStS byl dle zákona č. 46/2000 Sb. zaregistrován u Ministerstva kultury ČR (evidenční číslo registrace je E 21214) a byla podána žádost o zařazení časopisu do Seznamu. V časopise jsou nyní zřetelně odděleny vědecké a odborné statě od jiných článků, zpráv a informací, pozvánek na akce apod. Výbor ČStS doporučil redakční radě IB postup při sestavování jednotlivých čísel IB.

Na své první schůzi výbor ČStS schválil možnost přidělení tří finančních příspěvků na konference pro mladé vědecké pracovníky, a to v max. výši 5 tis. Kč na jednotlivce. Výbor schválil nové podmínky pro udělení jednorázového příspěvku. Předsedkyně obdržela pouze jednu žádost, a to od RNDr. Patrície Martínkové na konferenci IMPS 2013 (červenec, Holandsko). Členové výboru se vyjadřovali e-mailem; byl odsouhlasen žádaný příspěvek ve výši 5000,– Kč.

Byla aktualizována databáze členů ČStS, v plánu je její zpřístupnění členům výboru v rámci intranetu. Byla provedena aktualizaci údajů o ČStS v informačním systému Rady vědeckých společností ČR. Do tohoto systému byla též vložena výroční zpráva o činnosti ČStS.

3. Aktivity ČStS v oblasti konferencí a zastoupení ČStS na různých akcích

– Dne 31. ledna 2013 se na Vysoké škole ekonomické v Praze konalo valné shromáždění České statistické společnosti, na kterém byly předneseny zprávy o činnosti a o hospodaření ČStS a proběhla volba předsedy a výboru ČStS. Odbornou přednášku na téma *Generování náhody je příliš důležité, než abychom jej mohli přenechat náhodě* přednesl prof. Antoch z MFF UK Praha.
– Dne 20. března 2013 se ve Sládkovičovu předsedkyně ČStS zúčastnila slavnostní konference 45 rokov Slovenskej štatistickej a demografickej spoločnosti: minulost, prítomnosť, budúcnosť. Prof. Řezanková na ní vystoupila s příspěvkem informujícím o aktivitách České statistické společnosti na mezinárodní úrovni. Předsedkyně ČStS na této konferenci převzala pro Českou statistickou společnost pamětní list.
– Ve dnech 27. – 29. května 2013 se ve slovenských Kočovcích konala slovensko-

-česká konference PRASTAN 2013. Čtyři členové ČStS byli zastoupeni v programovém výboru této konference. Několik členů ČStS se této konference aktivně zúčastnilo, mezi nimi i předsedkyně a místopředseda ČStS. Hlavním tématem konference byly shlukovací metody a jiné vícerozměrné statistické techniky.
– Členové ČStS se zúčastnili také dalších statistických konferencí, z nichž nejvýznamnější byl 59. světový statistický kongres mezinárodního statistického institutu ISI, který se konal ve dnech 25. – 30. srpna 2013 v Hongkongu. Jednou z jeho mnoha akcí bylo mezinárodní kolo soutěže o nejlepší statistický plakát s tématem zemědělství, určené pro základní a střední školy, do které postoupily i plakáty z České republiky (organizátory národního kola byly ČSÚ a VŠE, prof. Řezanková byla předsedkyní hodnotící komise).
– Na dny 26. – 29. září 2013 byly naplánovány Statistické dny na Brejlově ve spojení s konferencí TEXperience 2013. K hezkému místu, příjemnému prostředí, dobrému počasí a skvělé kuchyni chyběla jen větší účast členů České statistické společnosti.
– Dne 23. října 2013 se místopředseda prof. Dohnal zúčastnil zasedání Rady vědeckých společností ČR a získal doplňující informace o možnosti získávání finančních příspěvků na činnost jednotlivých společností.
– Ve dnech 24. – 25. října 2013 se v Lublani konalo již deváté zasedání představitelů národních statistických společností ze zemí střední Evropy (V6). Českou statistickou společnost zastupovala na této akci předsedkyně prof. Řezanková. Výsledkem setkání je společné prohlášení, které mj. obsahuje příslib každoročního uspořádání společné virtuální konference společností V6. Dalším úkolem je potřeba získávání mladých statistiků pro aktivní členství v národních společnostech.
– Dne 5. prosince 2013 se v respiriu MFF UK uskutečnil již tradiční Mikuklášský den České statistické společnosti spočívající v celodenním odborném programu a závěrečném společenském posezení.
– Byla připravena konference (zimní škola) ROBUST 2014 na dny 19. – 24. ledna 2014 v Jetřichovicích.

4. Plán hlavních aktivit ČStS pro rok 2014

– První akcí bude již výše zmíněná zimní škola ROBUST 2014 ve dnech 19. – 24. ledna 2014 v Jetřichovicích.
– Druhou akcí bude valné shromáždění ČStS dne 12. února 2014 v budově ČSÚ v Praze s odbornou přednáškou Ing. Sixty na téma *Pokroky v měření ekonomiky – ESA 2010*.
– Na dny 9. – 10. října 2014 je naplánováno setkání zástupců statistických

společností V6 v Praze ve spolupráci s Českým statistickým úřadem.
– V prosinci 2014 bude uspořádán Mikuklášský den na MFF UK v Praze.
– Dále plánujeme zorganizování Statistických dnů, případně další odborná setkání.
– Budou vydána minimálně čtyři čísla Informačního Bulletinu České statistické společnosti, bude pokračovat údržba webových stránek ČStS a rozšířen článek o ČStS na Wikipedii.

V Praze dne 17. 1. 2014

prof. Ing. Hana Řezanková, CSc.
předsedkyně ČStS

ZASEDÁNÍ PŘEDSTAVITELŮ NÁRODNÍCH STATISTICKÝCH SPOLEČNOSTÍ V6 V LUBLANI

Hana Řezanková

Ve dnech 24. – 25. října 2013 se v Lublani konalo již deváté zasedání představitelů národních statistických společností ze zemí střední Evropy (V6). V této skupině jsou kromě České republiky zastoupeny Maďarsko, Rakousko, Rumunsko, Slovensko a Slovinsko. V letošním roce se bohužel nemohla zúčastnit delegace z Rakouska. Jednání se účastnili předsedové, příp. místopředsedové statistických společností, Českou statistickou společnost zastupovala její předsedkyně Hana Řezanková. Letošní setkání zorganizoval předseda Slovinské statistické společnosti (SSS) Andrej Blejec za podpory Statistického úřadu Slovinské republiky (SURS).

Oficiální část zasedání se konala v nové budově tohoto úřadu. Na úvod delegace statistických společností V6 přivítala náměstkyně generální ředitelky SURS Karmen Hren. O přestávce byla na programu prohlídka budovy SURS, například pracoviště pro telefonické dotazování respondentů a výpočetního střediska. Tuto část a technickou podporu zasedání zajišťovala místopředsedkyně SSS a zaměstnankyně SURS, paní Mojca Noč Razinger. V závěru jednání přítomní představitelé národních statistických společností V6 podepsali společné prohlášení, které je zaměřeno především na:
– potřebu intenzivnější výměny informací mezi statistickými společnostmi V6, zejména o konání národních statistických konferencí a dalších akcí,
– příslib zorganizovat alespoň jednu virtuální konferenci pro statistické společnosti V6,
– zvýšení úsilí při získání mladých statistiků pro aktivní členství. Příští jubilejní desáté zasedání představitelů statistických společností V6 se bude konat na podzim roku 2014 v Praze za podpory Českého statistického úřadu.

ZPRÁVA Z VÝROČNÍ KONFERENCE PSYCHOMETRICKÉ SPOLEČNOSTI

Patrícia Martinková

22. – 26. července 2013 Arnhem, Holandsko

Patrícia Martinková, Ústav informatiky AV ČR

V červenci tohoto roku se v holandském Arnhemu konalo již 78. výroční setkání Psychometrické společnosti, IMPS 2013. Zúčastnilo se ho na 330 zástupců z téměř 30 zemí. Byli zde zástupci univerzit – kateder psychometrie, psychologie, pedagogiky, sociologie, statistiky i matematiky. Zastoupeny byly také organizace zbývající se testováním znalostí jako např. Cambridge Assessment, Educational Testing Service (ETS), nebo konferenci pořádající CITO.

Přednášky probíhaly paralelně v šesti sálech místní opery (Musis Sacrum). Sekce se věnovaly různým tématům potřebným pro analýzu psychometrických dat: teorii odpovědi na položku (IRT), modelování času odpovědí, detekci odlišně fungujících položek (DIF), klastrové a korespondenční analýze, faktorové analýze, strukturálním modelům (SEM), bayesovským metodám, případu chybějících dat, odhadu reliability a validity, ale také praktickým otázkám testovaní jako je návrh testu nebo jeho praktické provedení, aj.

Obzvláště pro mladé vědce byla velmi přínosná úterní diskuse s editory 5 časopisů, plná doporučení a odpovědí na otázku „How to get published?“. Čtvrteční keynote lecture se věnovala historii sekvenční analýzy a připravila tak půdu pro odpolední proslov prezidenta společnosti. Ten mluvil o počítačovém adaptivním testování (computerized adaptive testing, CAT) umožňujícím individuální přístup k testovaným. V přednášce zazněly metody potřebné pro analýzu CAT, zdůrazněna byla aktuálnost tohoto tématu a potřeba většího počtu vědců, kteří by se tématu věnovali.

Pro studenty vypisuje psychometrická společnost každoročně tři ocenění (travel award) v hodnotě 500 $. Další podobné ocenění pro studenty poskytuje ETS, to navíc hradí také krátkou návštěvu ETS v Princetonu, letos ním byla oceněna doktorandka z katedry matematiky a statistiky z McGill University. Na místě pak byla oceněna nejlepší studentská prezentace a nejlepší poster. Poslední ze zmíněných ocenění putovalo do České republiky.

Autorka článku děkuje České statistické společnosti za poskytnutí finančního příspěvku na konferenci.

Obsah

Vědecké a odborné statě

Zprávy a informace

Informační bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo. Vydavatelem je Česká statistická společnost, IČ 00550795, adresa společnosti je Na padesátém 81, 100 82 Praha 10. Evidenční číslo registrace vedené Ministerstvem kultury ČR dle zákona č. 46/2000 Sb. je E 21214.

The Information Bulletin of the Czech Statistical Society is published quarterly.
The contributions in bulletin are published in English, Czech and Slovak languages.

DOI: 10.5300/IB, http://dx.doi.org/10.5300/IB
ISSN 1210–8022 (Print), ISSN 1804–8617 (Online)

Toto číslo bylo vytištěno s laskavou podporou Českého statistického úřadu.