SEMESTRÁLNÍ PRÁCE

Licenční studium

STATISTICKÉ ZPRACOVÁNÍ DAT

11. kurs licenčního studia

Metody s latentními proměnnými a klasifikační metody

Univerzita Pardubice
Fakulta chemicko – technologická
Katedra analytické chemie

Policie České republiky
Správa Severomoravského kraje
Odbor kriminalistické techniky a expertiz - chemie
pracoviště Frýdek - Místek
O S T R A V A

Ing. Jan Aufart

str.

Obsah …… 1

3.2.1. Výpočet 1. latentní proměnné …… 2
3.2.2. Určení nezbytného počtu latentních proměnných …… 5
3.2.3. Odhad chybějícího prvku matice …….. 6
3.2.4. Výpočet komunality ……. 7
3.2.5. Rozdíl mezi PCA a FA …… 8
3.2.6 Korelační koeficient (kanonické korelace) …….. 8
3.2.7. Příklad možného použití metody PLS …… 9
3.2.8 Euklidovská metrika ……. 10
3.2.9 Příklad aplikace klasifikačních metod …….. 11

Předmět: **Metody s latentními proměnnými a klasifikační metody**
Přednášející: **Prof. Ing. Oldřich Pytela, DrSc.**
Úloha 3.2.1.: **Vypočtěte algoritmem NIPALS 1.latentní proměnnou z matice A [řádek,sloupec]: A[1,1]=1, A[2,1]=2, A[3,1]=3, A[1,2]=1, A[2,2]=2, A[3,2]=0, A[1,3]=6, A[2,3]=4, A[3,3]=2. Matici před zpracováním standardizujte.**

Program: MS Excel, OPstat

Řešení:

Zadaná matice A

1 1 6
2 2 4
3 0 2

Standardizace

matice se provede odečtením sloupcového průměru od každého prvku sloupce a získanou hodnotu podělíme směrodatné odchylky sloupce

$y_{ij} = (x_{ij} - x_j)/s_j$

A	1	1	6
	2	2	4
	3	0	2
sloupcový průměr	2	1	4
směrodatná odchylka	1	1	2
A po standardizaci	-1	0	1
	0	1	0
	1	-1	-1

Metoda hlavních komponent rozkládá původní matici A na součin matice latentních proměnných T a matice zátěží P^T a matici nevysvětlené variability E

$$\mathbf{A = T * P^T + E}$$

Algoritmus NIPALS je jedním ze způsobů rozkladu matice A vycházející ze vztahu

$$\mathbf{P^T = (T^T * T)^{-1} * T^T * A^T}$$

při postupném hledání vektorů matic T a P^T (t_1;$p_1{}^T$…..), tedy

$$\mathbf{p_1{}^T = (t_1{}^T * t_1)^{-1} * t_1{}^T * A} \quad ; \quad \mathbf{t_1{}^T = (p_1{}^T * p_1)^{-1} * p_1{}^T * A^T}$$

t_1 - první sloupec standardizované matice A, podle vztahu $p_1^T = (t_1^T * t_1)^{-1} * t_1^T * A$ získáme vektor p_1^T

$t_1^T =$ | -1 0 1 |

$t^T t$	-1	0	1	x	-1 0 1	=	2			
$(t_1^T * t_1)^{-1} * t_1^T * A$	0,5	x	-1	0	1	x	-1 0 1 0 1 0 1 -1 -1	=	1	-0,5 -1

$p_1^T =$ | 1 -0,5 -1 |

a normováním tohoto vektoru $\mathbf{p_1^N = (p_1^T * p_1)^{-1/2} * p_1}$ dostaneme první odhad vektoru p_1,

(p1T* p1)-1/2	1,0000	-0,5000	-1,0000	x	1,0000 -0,5000 -1,0000	=	0,6667		
p_1^N	0,6667	x	1,0000	-0,5000	-1,0000	=	0,6667	-0,3333	-0,6667

který dosadíme do vztahu $t_1^T = (p_1^T * p_1)^{-1} * p_1^T * A^T$

$(p_1^T * p_1)^{-1}$	0,6667	-0,3333	-0,6667	x	0,6667 -0,3333 -0,6667	=	1,0000						
$(p_1^T * p_1)^{-1} * A^T$	1,0000	x	0,6667	-0,3333	-0,6667	x	-1,0000	0,0000	1,0000 -1,0000 -1,0000	=	-1,3333	-0,3333	1,6667
							0,0000	1,0000					
							1,0000	0,0000					

$t_1 =$ |-1,3333 -0,3333 1,6667|

Výše uvedený postup opakujeme pokud platí, že $d/N > 10^{-10}$, kde d je konvergenční kriterium dané vztahem

$$\mathbf{d = (t_{nový} - t_{starý})^T * (t_{nový} - t_{starý}) * (t_{nový} * t_{straý})}$$

Opakovaným postupem získáme tyto hodnoty:

N	vektor latentních proměnných t_1			normovaný vektor zátěží p_1^N			d	d/N
1	-1,33333	-0,33333	1,66667	0,66667	-0,33333	-0,66667	1,43E-01	1,43E-01
2	-1,27920	-0,42640	1,70561	0,63960	-0,42640	-0,63960	2,77E-03	1,39E-03
3	-1,26234	-0,45083	1,71317	0,63117	-0,45083	-0,63117	1,98E-04	6,61E-05
4	-1,25766	-0,45733	1,71499	0,62883	-0,45733	-0,62883	1,42E-05	3,56E-06
5	-1,25639	-0,45907	1,71546	0,62820	-0,45907	-0,62820	1,02E-06	2,04E-07
6	-1,25605	-0,45953	1,71558	0,62803	-0,45953	-0,62803	7,34E-08	1,22E-08
7	-1,25596	-0,45966	1,71561	0,62798	-0,45966	-0,62798	5,27E-09	7,53E-10
8	-1,25594	-0,45969	1,71562	0,62797	-0,45969	-0,62797	3,78E-10	4,73E-11

Konečným výsledkem je stabilní rozklad

p_1^T = | 0,62797 -0,45969 -0,62797 | t_1^T = | -1,25594 -0,45969 1,71562 |

Ze vztahu $\mathbf{E = A - t_1 * p_1^T}$ dostaneme matici variability, která není vysvětlena pomocí vektorů t_1 a p_1^T a která je výchozí pro výpočet vektorů t_2 a p_2^T.

t1 * p1T	-1,25594 -0,45969 1,71562	x | 0,62797 -0,45969 -0,62797 | =	-0,78869 0,57734 0,78869 -0,28867 0,21131 0,28867 1,07736 -0,78865 -1,07736

t_1 **p_1^T** **$t_1 * p_1^T$**

A - t1 * p1T	-1,00000 0,00000 1,00000 0,00000 1,00000 0,00000 1,00000 -1,00000 -1,00000	-	-0,78869 0,57734 0,78869 -0,28867 0,21131 0,28867 1,07736 -0,78865 -1,07736	=	-0,21131 -0,57734 0,21131 0,28867 0,78869 -0,28867 -0,07736 -0,21135 0,07736

A **$t_1 * p_1^T$** **E**

Výpočet je obsažen v souboru „U1_NIPALS.xls“ na přiloženém CD

Předmět: **Metody s latentními proměnnými a klasifikační metody**
Přednášející: **Prof. Ing. Oldřich Pytela, DrSc.**
Úloha 3.2.2.: **S použitím vhodných kritérií určete nezbytný počet latentních proměnných, když bylo z dat určeno: PRESS(0) = S(0) = 100; PRESS(1) = 20; S(1) = 10; PRESS(2) = 3,5; S(2) = 3,4; PRESS(3) = 3,45; S(3) = 3,39.**

Program: MS Excel

Řešení:

Woldovo kritérium $PRESS_{(P)} / S_{(P-1)}$ se vzhledem k zadání jeví jako nejvhodnější pro určení nezbytného počtu latentních proměnných. Pokud je $PRESS_{(P)} / S_{(P-1)} > 0,95$ není vhodné zavádět další (P+1)ní proměnnou.

P	$PRESS_{(P)}$	$S_{(P-1)}$	Woldovo kritérium $PRESS_{(P)} / S_{(P-1)}$
1	20,00	100,00	0,20
2	3,50	10,00	0,35
3	3,45	3,40	1,01

Hodnota Woldova kritéria pro P = 3 (1,01) je větší než 0,95 a zavedení čtvrté proměnné není tedy vhodné.

Nezbytný počet latentních proměnných podle Woldova kritéria je tři.

Výpočet je obsažen v souboru „U2_WOLD.xls“ na přiloženém CD

Předmět: **Metody s latentními proměnnými a klasifikační metody**
Přednášející: **Prof. Ing. Oldřich Pytela, DrSc.**
Úloha 3.2.3.: **Odhadněte hodnotu chybějícího prvku matice A [2,2] , jesliže výpočtem z nekompletní matice byly určeny vektory p: 0,541; 0,423; 0,514; 0,514 a t: -1,340; -0,735; 2,076**

Program: MS Excel

Řešení:

Rekonstrukce zdrojové matice A s použitím latentních proměnných popisujících podstatnou část variability matice A bez zahrnutí experimentálních chyb se označuje pojmem krátký cyklus. Rekonstrukce je popsána vztahem $\mathbf{A^{pred} = TP^T}$ a vychází z rozkladu matice (viz. vztah $\mathbf{A = T * P^T + E}$ v úloze č. 1 na str. 2).

t =	-1,340	-0,735	2,076

p =	0,541	0,432	0,514	0,514

$t*p^T$	-1,34 -0,735 2,076	x	0,541	0,423	0,514	0,514	=	-0,72494 -0,39764 1,12312	-0,56682 -0,31091 0,87815	-0,68876 -0,37779 1,06706	-0,68876 -0,37779 1,06706

Odhadnutá hodnota chybějícího prvku matice A [2,2] je -0,31091.

Výpočet je obsažen v souboru „U3_MMAT.xls“ na přiloženém CD.

Předmět: **Metody s latentními proměnnými a klasifikační metody**
Přednášející: **Prof. Ing. Oldřich Pytela, DrSc.**
Úloha 3.2.4.: **Výpočtem metodou PCA byly určeny vektory**
p1 : 0,012 0,458 -0,352 0,987 p2: 0,926 -0,238 0,872 -0,115
Vypočtěte komunality a vyberte sloupec, který nejlépe charakterizuje celou matici.

Program: MS Excel

Řešení:

Komunalita h_i^2 je část rozptylu příslušející použitým faktorům (objasněná část rozptylu, variabilitu společná všem manifestním proměnným). Prvek matice zátěží p_{ip} přísluší i-tému sloupci zdrojové matice a je mírou variability tohoto sloupce popsaného p-tou latentní proměnnou. Podíl variability daného sloupce popsaný společnými latentními proměnnými lze pak vyjádřit jako součet příspěvků jednotlivých latentních proměnných vztahem

$$\mathbf{h_i^2 = p^N * (p^N)^T = \sum_{p=1}^{P} (p_{ip}^N)^2}$$

kde p_{ip}^N jsou normované zátěže podle vztahu

$$\mathbf{p_{ip}^N = p_{ip} / \sum_{i=1}^{M} p_{ip}^2}$$

p_1	**i**	1	2	3	4	**Σ**
	p_{i1}	0,012	0,458	-0,352	0,987	
	p_{i1}^2	0,000144	0,209764	0,123904	0,974169	1,307981
	p_{i1}^N	0,00917445	0,350158	-0,26912	0,754598	
	$(p_{i1}^N)^2$	8,417E-05	0,122611	0,072424	0,569418	
p_2	**i**	1	2	3	4	**Σ**
	p_{i2}	0,926	-0,238	0,872	-0,115	
	p_{i2}^2	0,857476	0,056644	0,760384	0,013225	1,687729
	p_{i2}^N	0,54866628	-0,14102	0,516671	-0,06814	
	$(p_{i2}^N)^2$	0,30103469	0,019886	0,266949	0,004643	

i	p_1	p_1^N	$(p_1^N)^2$	p_2	p_2^N	$(p_1^N)^2$	h^2
1	0,012	0,009174	8,41704E-05	0,926	0,548666	0,30103469	0,301119
2	0,458	0,350158	0,122610619	-0,238	-0,14102	0,01988605	0,142497
3	-0,352	-0,26912	0,072423991	0,872	0,516671	0,26694854	0,339373
4	-0,352	0,754598	0,569418319	-0,115	-0,06814	-0,0681389	0,574061

Nejvyšší komunalitu vykazuje čtvrtý sloupec (h^2 = 0,574061), který nejlépe charakterizuje zdrojovou matici.

Výpočet je obsažen v souboru „U4_KOM.xls“ na přiloženém CD.

Předmět: **Metody s latentními proměnnými a klasifikační metody**
Přednášející: **Prof. Ing. Oldřich Pytela, DrSc.**
Úloha 3.2.5.: **Vysvětlete proč vysvětlená variabilita je při výpočtu metodou FA vždy nižší než při výpočtu metodou PCA.**

Vysvětlení:

FA a PCA se liší rozkladem variability zdrojové matice. FA uvažuje specifitu tedy část variability, která není vysvětlena společnými faktory (část variability „zbývá“), zatímco PCA předpokládá úplné vysvětlení variability latentními proměnnými (vysvětluje veškerou variabilitu). FA je neúplnou komponentní analýzou, PCA je úplnou komponentní analýzou.

Předmět: **Metody s latentními proměnnými a klasifikační metody**
Přednášející: **Prof. Ing. Oldřich Pytela, DrSc.**
Úloha 3.2.6.: **Výpočtem metodou kanonických korelací bylo zjištěno:**
0.297 X1 + 0.298 X2 + 0.050 X3 + 0.256 X4 = 0.493 Y1 - 0.213 Y2
r1 = 0.830
0.006 X1 - 0.115 X2 + 0.950 X3 + 0.056 X4 = 0.493 Y1 + 0.213 Y2
r1 = 0.512
Vypočtěte skupinový korelační koeficient a interpretujte výsledky.

Program: MS Excel

Řešení:

Skupinový korelační koeficient: $R^2 = 1-(1-r_1^2) * (1-r_2^2)$

0,830	0,512
=	
77,05%	

Model vysvětluje 77,05 % variability. Y1 není ovlivněna proměnnými X, Y2 závisí na proměnných X. Snížení X1, X2 a X4 zvyšuje Y2 při zvýšení X3.

Výpočet je obsažen v souboru „U6_KOR.xls“ na přiloženém CD.

Předmět: **Metody s latentními proměnnými a klasifikační metody**
Přednášející: **Prof. Ing. Oldřich Pytela, DrSc.**
Úloha 3.2.7.: **Uveďte nějaký konkrétní příklad vhodný pro zpracování metodou metodou PLS.**

PLS (projekce latentních struktur) je metoda k popisu vztahu mezi náhodnými vektory - závislým vektorem Y a vysvětlujícím vektorem X - cestou regresního vztahu mezi latentními proměnnými příslušejícími těmto vektorům. (obdoba NIPALS)

Příklad:

a) Určení vztahu např.mezi chemickým složením produktu a jeho vlastnostmi. (Predikce vlastností na základě složení pro vlastnosti jejichž stanovení je komplikované.)

b) Problematika efektivní separace složek analyzované směsi metodou GC, např. vztah mezi nastavitelnými parametry a výslednými kritérii.

tlak, průtok, náplň kolony, průměr kolony teplota => rozlišení píků, tvar píků , retenční čas

Předmět: **Metody s latentními proměnnými a klasifikační metody**
Přednášející: **Prof. Ing. Oldřich Pytela, DrSc.**
Úloha 3.2.8.: **Jeden objekt je charakterizován metrickými znaky (2,10), druhý (3,8), třetí (4,9), čtvrtý (10,4) a pátý (11,5). Vypočtěte matici vzdáleností v Euklidově metrice a dokumentujte výpočet shlukování některou z používaných metod. Výsledky interpretujte graficky.**

Program: MS Excel, OPstat

Řešení:

	ZNAK 1	ZNAK 2
PRVNÍ	2	10
DRUHÝ	3	8
TŘETÍ	4	9
ČTVRTÝ	10	4
PÁTÝ	11	5

Eukleidova metrika je definována vztahem:

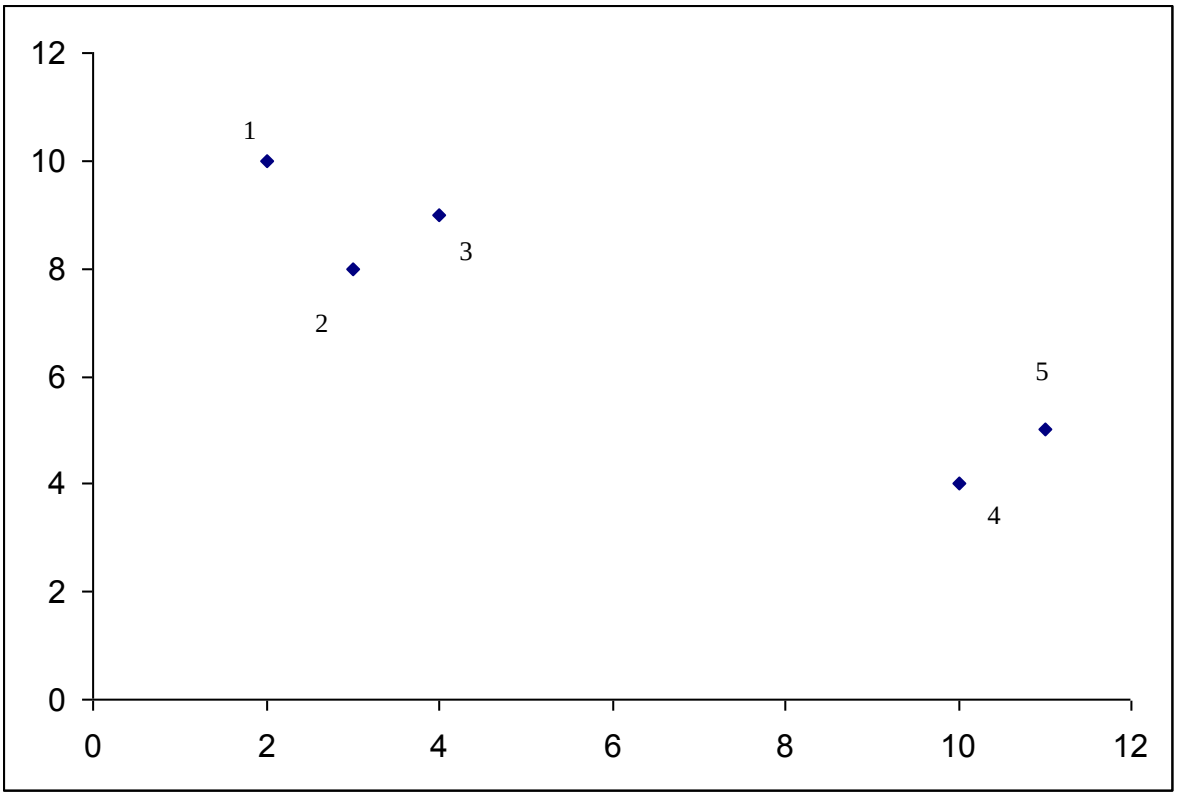
$$d_E = (X_k, X_l) = \left[\sum_{p=1}^{p} (x_{kp} - x_{lp})^2 \right]^{1/2}$$

výpočtem dostaneme

OBJEKT	PRVNÍ	DRUHÝ	TŘETÍ	ČTVRTÝ	PÁTÝ
PRVNÍ	0				
DRUHÝ	2,236068	0			
TŘETÍ	2,236068	1,4142	0		
ČTVRTÝ	10	8,0623	7,81025	0	
PÁTÝ	10,29563	8,544	8,062258	1,414214	0

Euklidovská vzdálenost	OBJEKT				
1,41421356	DRUHÝ	TŘETÍ			
1,41421356	ČTVRTÝ	PÁTÝ			
2,23606798	PRVNÍ	DRUHÝ	TŘETÍ		
7,81024968	PRVNÍ	DRUHÝ	TŘETÍ	ČTVRTÝ	PÁTÝ

Matice vzdáleností ukazuje stejné vzdálenosti u dvojic objektů
DRUHÝ, TŘETÍ a ČTVRTÝ, PÁTÝ
Shlukováním je k dvojici DRUHÝ, TŘETÍ přiřazen i PRVNÍ objekt. V daném případě jde tedy o dva shluky PRVNÍ, DRUHÝ, TŘETÍ a ČTVRTÝ, PÁTÝ (viz. následující graf).



Předmět: **Metody s latentními proměnnými a klasifikační metody**
Přednášející: **Prof. Ing. Oldřich Pytela, DrSc.**
Úloha 3.2.9.: **Popište slovně postup aplikace metod s latentními proměnnými nebo klasifikačních metod na nějakém konkrétním příkladu ze své praxe.**

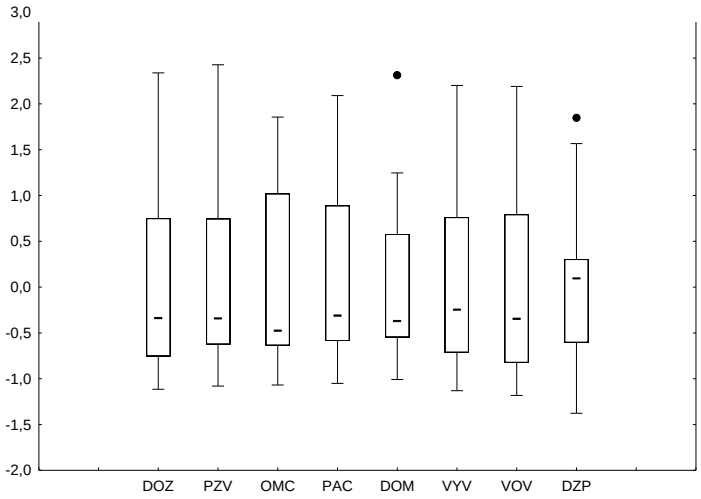
Příklad: Možnost utřídění pracovních skupin podle vybraných kriterií.

		1	2	3	4	5	7	8	9	10
	SKU	DOZ	PZV	OMC	PAC	DOM	VYV	VOV	DZP	HUB
1	BR	1532	966	731	12	12	645	1451	51	64
2	FM	5971	3801	1439	111	85	2926	5494	70	180
3	KA	6450	3952	3176	94	83	3496	6132	39	824
4	NJ	3214	1832	772	49	100	1558	3105	49	176
5	OL	4301	2545	2775	32	470	1913	4129	74	156
6	OP	2927	1610	412	30	76	1551	2794	40	159
7	OV	10298	6846	2355	146	123	4882	8805	28	1541
8	PR	2998	2054	887	31	129	1332	2776	33	157
9	PV	2699	1525	763	42	231	1179	2578	39	144
10	SU	1870	1322	1279	53	323	917	1660	52	97
11	JE	885	496	306	10	30	348	816	49	60

Popis:

1. Vyhodnocení dostupných dat – variabilita – výběr proměnných (kritérií)

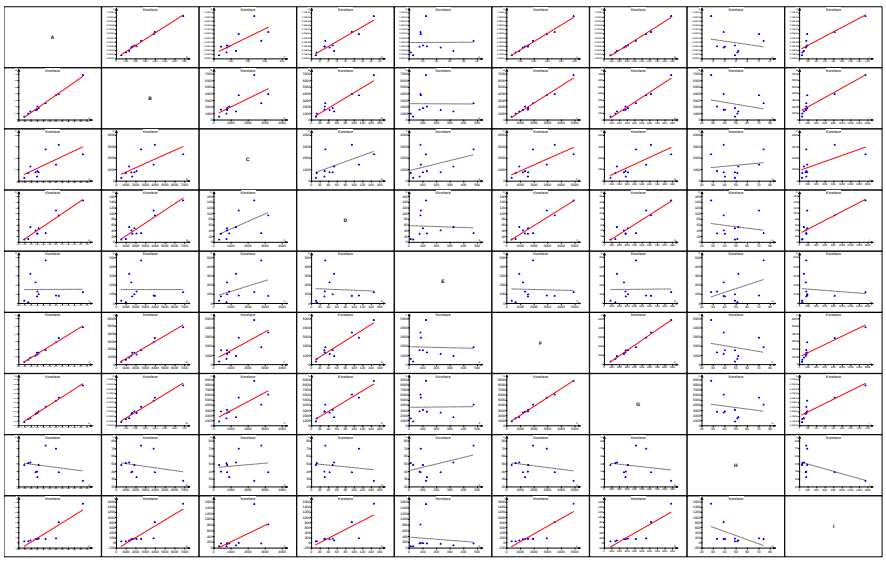
Krabicový graf ukáže variabilitu proměnných (STATISTICA)



2. Případná standardizace dat pokud jsou proměnné různého druhu v různých škálách např. $y_{ij} = (x_{ij} - x_j)/s_j$

3. Zjištění vztahu mezi proměnnými (korelace)

   Korelace (QC Expert)



4. Určení počtu latentních proměnných

(Cattelův indexový graf úpatí vlastních čísel, Woldovo kriterium, Kaiserovo pravidlo)

5. Graf komponentních vah a graf komponentního skóre.

6. Závěry a interpretace výsledků.

Data je možno podrobit také shlukové analýze (případně i diskriminační analýze).