MASARYKOVA UNIVERZITA

Z2069 Statistické metody a zpracování dat II

Vícerozměrné metody

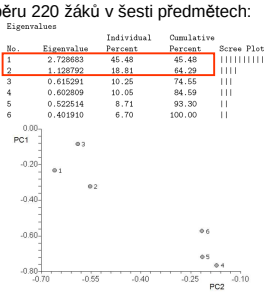
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

K čemu to slouží?

Vstupní data: výsledky dosažené ve výběru 220 žáků v šesti předmětech:

1. gaelština
2. angličtina
3. dějepis
4. aritmetika
5. algebra
6. geometrie

PCA



korelační matice vstupních dat

1.00					
0.44	1.00				
0.41	0.35	1.00			
0.29	0.35	0.16	1.00		
0.33	0.32	0.19	0.59	1.00	
0.25	0.33	0.18	0.47	0.46	1.00

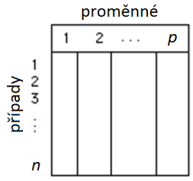
- Původních šest proměnných lze nahradit pouze dvěma novými proměnnými (faktory, hlavními komponentami) bez podstatné ztráty informace
- Dvě nové proměnné vysvětlují korelační strukturu pozorovaných dat; první faktor vyjadřuje matematickou dispozici žáka, druhý dispozici jazykově-humanitní

Úvod do vícerozměrných metod

O řadě jevů či procesů máme k dispozici ne jeden statistický znak, ale znaků několik.

Př. struktura obyvatelstva, vlastnosti povodí, klimatické poměry místa, prospěch v různých předmětech, …

Vstupní data: statistické jednotky např. městské obvody (případy – řádky) a k nim několik charakteristik např. demografická data (proměnné – sloupce).



Cíle prezentovaných metod:

1. **redukovat počet** proměnných
2. **detekovat strukturu** vztahů mezi proměnnými (klasifikovat, vytvořit typologii dat)

Analýza hlavních komponent (Principal Component Analysis – PCA)

Shluková analýza (Cluster Analysis)

Úvod do vícerozměrných metod

Literatura:

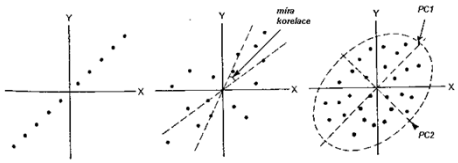
Heřmanová, E. (1991): Vybrané vícerozměrné statistické metody v geografii. SPN, Praha, 133 s.

Hendl, J. (2004): Přehled statistických metod zpracování dat. Portál, Praha, 583 s.

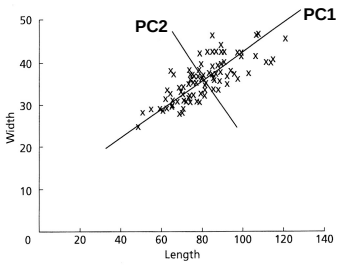
http://www.statsoft.cz/textbook/stathome.html

Princip analýzy hlavních komponent

- Máme-li pro soubor znaků dvě proměnné a ty spolu **vzájemně korelují** – potom vypovídají z velké části o tomtéž – jsou **redundantní**.
- Dvě (korelované) proměnné vyneseme do grafu a proložíme jím novou osu, která je orientována **ve směru maximálního rozptylu** původních dat.
- Druhá osa je na ni kolmá a je vedena ve směru největšího rozptylu nevysvětleného první osou.

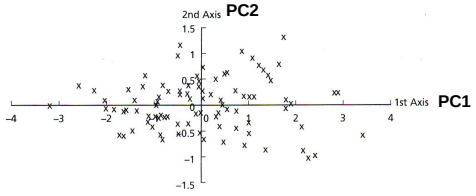


Princip analýzy hlavních komponent



- Nové osy představují nové proměnné (tzv. **hlavní komponenty** či **faktory**)
- Hlavní komponenta je **lineární kombinací původních** proměnných.
- Uvedený princip lze zobecnit na větší počet proměnných

Princip analýzy hlavních komponent



- Nové osy vytvářejí nový souřadný systém
- První hlavní komponenta (PC1) popisuje největší část proměnlivosti (rozptylu) původních dat
- Druhá hlavní komponenta (PC2) popisuj největší část proměnlivosti neobsažené v PC1 atd.
- Hlavní komponenty jsou nekorelované.

Princip analýzy hlavních komponent

Vstupní data představuje matice, která obsahuje *n* případů pro *p* proměnných. Běžně představují proměnné sloupce datové matice a případy její řádky.

První hlavní komponenta (*PC1*) je lineární kombinací proměnných X_1, X_2, …, X_p

$$PC_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$$

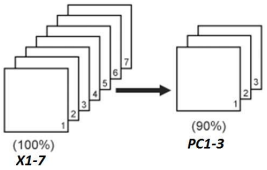
Pro koeficienty a_{ij} musí platit: $a_{11}^2 + a_{12}^2 + ... + a_{1p}^2 = 1$

Druhá hlavní komponenta (*PC2*) bude:

$$PC_2 = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p \qquad a_{21}^2 + a_{22}^2 + ... + a_{2p}^2 = 1$$

atd. Z původních *p* proměnných se vypočte *p* hlavních komponent

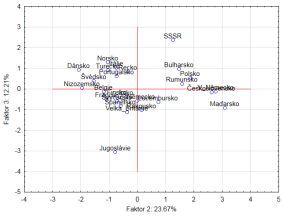
Princip analýzy hlavních komponent



- Cílem metody je zjednodušení popisu lineárně závislých tj. korelovaných znaků, tedy **redukce dat** bez podstatné ztráty
- Základní charakteristikou každé hlavní komponenty je její míra variability (**rozptyl**).
- Hlavní komponenty jsou **seřazeny dle důležitosti**, tj. klesajícího rozptylu, od největšího k nejmenšímu.
- **Většina informace** o variabilitě dat je přitom obvykle soustředěna do prvních dvou – tří komponent a ostatní obsahují epodstatné množství informace (představují „šum“).

Princip analýzy hlavních komponent

- Cílem metody je také **odhalení struktury v datech**
- Charakteristiky, které na jednotkách měříme, jsou jen určitou formou projevu tzv. **skrytých veličin**, které přímo měřit nemůžeme.



- Řada měřených charakteristik spolu do značné míry souvisí – vypovídá o stejné vlastnosti, koreluje spolu (mezi proměnnými existují „překryvy“).
- Tyto nové proměnné (hlavní komponenty) popisují soubor jednotek syntetičtěji a úsporněji.
- V některých případech však je problém tyto nové proměnné interpretovat – dát jim význam ve smyslu řešené úlohy a určit, co představují

Ilustrativní příklad – vstupní data

Podíl zaměstnaných v devíti odvětvích ve 26 evropských zemích (údaje z konce 70. let 20. století)

1. AGR = agriculture
2. MIN = mining
3. MAN = manufacturing
4. PS = power suplies
5. CON = construction
6. SER = service industries
7. FIN = finance,
8. SPS = social and personal services
9. TC = transport and communications

Stát	agro	doly	prumysl	energetik	stavebnic	mist_hosp	finance	sluzby	doprava
Belgie	3.3	0.9	27.6	0.9	8.2	19.1	6.2	26.6	7.2
Dánsko	9.2	0.1	21.8	0.6	8.3	14.6	6.5	32.2	7.1
Francie	10.8	0.8	27.5	0.9	8.9	16.8	6	22.6	5.7
Z_Německo	6.7	1.3	35.8	0.9	7.3	14.4	5	22.3	6.1
Irsko	23.2	1	20.7	1.3	7.5	16.8	2.8	20.8	6.1
Itálie	15.9	0.6	27.6	0.5	10	15.1	1.6	20.1	5.7
Lucembursko	7.7	3.1	30.8	0.8	9.2	18.5	4.6	19.2	6.2
Nizozemsko	6.3	0.1	22.5	1	9.9	18	6.8	28.5	6.8
Velká_Británie	2.7	1.4	30.2	1.4	6.9	16.9	5.7	28.3	6.4
Rakousko	12.7	1.1	30.2	1.4	9	16.8	4.9	16.8	7
Finsko	13	0.4	25.9	1.3	7.4	14.7	5.5	24.3	7.6
Řecko	41.4	0.6	17.6	0.6	8.1	11.5	2.4	11	6.7
Norsko	9	0.5	22.4	0.8	8.6	16.9	4.7	27.6	9.4
Portugalsko	27.8	0.3	24.5	0.6	8.4	13.3	2.7	16.7	5.7
Španělsko	22.9	0.8	28.5	0.7	11.5	9.7	0.5	11.8	5.5
Švédsko	6.1	0.4	25.9	0.8	7.2	14.4	6	32.4	6.8
Švýcarsko	7.7	0.2	37.8	0.8	9.5	17.5	5.3	15.4	5.7
Turecko	66.8	0.7	7.9	0.1	2.8	5.2	1.1	11.9	3.2
Bulharsko	23.6	1.9	32.3	0.6	7.9	8	0.7	18.2	6.7
Československo	16.5	2.9	35.5	1.2	8.7	9.2	0.9	17.9	7
V_Německo	4.2	2.9	41.2	1.3	7.6	11.2	1.2	22.1	8.4
Maďarsko	21.7	3.1	29.6	1.9	8.2	9.4	0.9	17.2	8
Polsko	31.1	2.5	25.7	0.9	8.4	7.5	0.9	16.1	6.9
Rumunsko	34.7	2.1	30.1	0.6	8.7	5.9	1.3	11.7	5
SSSR	23.7	1.4	25.8	0.6	9.2	6.1	0.5	23.6	9.3
Jugoslávie	48.7	1.5	16.8	1.1	4.9	6.4	11.3	5.3	4

Vstupní matice: 9 řádků (proměnných – odvětví) a 26 sloupců (případy – státy)

Cíl: Redukce počtu proměnných a odhalení typických znaků v zaměstnanosti jednotlivých států

Příklad – typický výstup PCA I.

No.	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	3.487151	38.75	38.75	||||||||
2	2.130173	23.67	62.41	|||||
3	1.098958	12.21	74.63	|||
4	0.994483	11.05	85.68	|||
5	0.543218	6.04	91.71	||
6	0.383428	4.26	95.97	|
7	0.225754	2.51	98.48	|
8	0.136790	1.52	100.00	|
9	0.000046	0.00	100.00	|

Význam jednotlivých sloupců:

- pořadové číslo nové proměnné (PC - hlavní komponenty)
- tzv. **vlastní hodnota** – část z celkového rozptylu původních dat vysvětlená každou z nových komponent
- procentuální vyjádření **množství rozptylu vysvětleného komponentou**
- **kumulativní** hodnota procentuálního podílu vysvětleného příslušnými komponentani (např. první 4 komponenty vysvětlují 85,68 % celkové variability původních dat)
- tzv. **sutinový graf** sloužící k určení počtu významných komponent

Příklad – typický výstup PCA II.

původní proměné

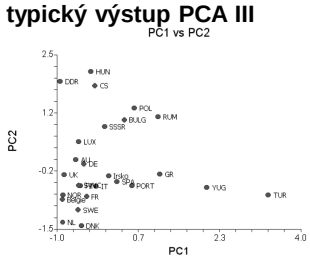
nové proměné (PC)

Variables	Factor1	Factor2	Factor3	Factor4
AGR	0.523791	0.053594	0.048674	0.028793
MIN	0.001323	0.617807	-0.201100	0.064085
MAN	-0.347495	0.355054	-0.150463	-0.346088
PS	-0.255716	0.261096	-0.561083	0.393309
CON	-0.325179	0.051288	0.153321	-0.668324
SER	-0.378920	-0.350172	-0.115096	-0.050157
FIN	-0.074374	-0.453698	-0.587361	-0.051567
SPS	-0.387409	-0.221521	0.311904	0.412230
TC	-0.366823	0.202592	0.375106	0.314372

Tzv. **zátěže** (loadings) - představují míru korelace mezi původními a novými proměnnými

Lze je využít k **interpretaci** nově vypočtených proměnných (faktorů, komponent)

Příklad – typický výstup PCA III



Struktura zaměstnanosti jednotlivých zemí vyjádřená polohou v grafu hodnot prvních dvou (nejvýznamnějších) hlavních komponent.

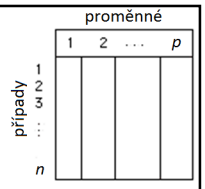
PC1 diferencuje země podle **rozsahu zemědělské výroby**, rozlišuje zemědělské a průmyslové země

PC2 rozlišuje země s velkým a malým sektorem služeb (Z předchozí tabulky je patrné, že PC2 má **záporné** zátěže u míst. hospodářství, služeb, financí – kapitalistické státy měly **rozvinutější** sektor služeb než socialistické

Obecný postup analýzy hlavních komponent

1. Sestavení matice o *p* proměnných (sloupcích) a *n* případech (řádcích) a její případná standardizace
2. Explorační analýza vstupních dat s cílem odhalení vztahů mezi proměnnými
3. Výpočet korelační matice typu *p,p,* identifikace silně korelovaných proměnných
4. Výpočet *p* nových ortogonálních proměnných (hlavních komponent)
5. Analýza tabulky s **vlastními čísly** a **sutinového grafu** za účelem rozhodnutí o počtu významných komponent
6. Interpretace významných komponent s využitím tabulky s tzv. **zátěžemi** a grafem komponentních **skóre**
7. (Případná **rotace** faktorů či komponent za účelem lepší interpretace)

Vstupní datová matice



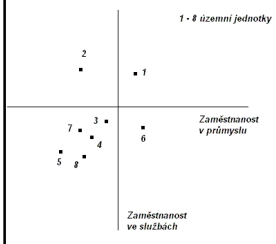
Vstupní data představuje matice, která obsahuje n případů pro m proměnných. V běžném případě představují proměnné sloupce datové matice a případy její řádky.

Charakteristiky vstupují do analýzy obvykle ve standardizovaném tvaru (ve formě normovaných normálních proměnných).

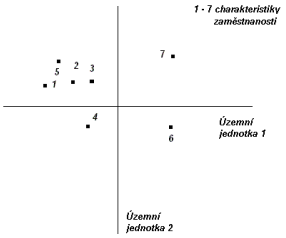
$$z_i = \frac{x_i - \mu}{\sigma}$$

Standardizaci provádíme proto, že různé proměnné ve vstupní datové matici mohou mít různý rozměr, různé jednotky.

Dva způsoby (módy) PCA



Analýza podobnosti jednotek (případů) – dimenze r-rozměrného prostoru jsou charakteristiky (proměnné). Cílem analýzy je redukovat sloupce datové matice



Analýza podobnosti proměnných - dimenze r-rozměrného prostoru jsou jednotky (případy). Cílem analýzy je redukovat dimensionalitu řádků.

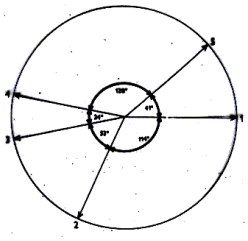
Grafické znázornění korelací mezi více proměnnými

Porovnáváme vzdálenosti mezi proměnnými, malá vzdálenost znamená silnou korelaci.

Hledáme shluk podobných proměnných, jež spolu korelují

Úhly mohou nabývat hodnot od 0 do 180 stupňů a *cos* úhlu odpovídá hodnotě korelačního koeficientu:

	V1	V2	V3	V4	V5
V1	1	-0,41	-0,97	-0,98	0,75
V2		1	0,60	0,22	-0,91
V3			1	0,91	-0,88
V4				1	-0,62
V5					1



cos 0 =1, $r_{xy} = 1$

cos 90 =0, $r_{xy} = 0$

cos 180 =-1, $r_{xy} = -1$

Komponentní váhy (zátěže, loadings)

Komponentní váhy informují o vztahu mezi původními *p* proměnnými a hlavními komponentami.

Factor Loadings

Variables	Factor1	Factor2
Gaelic	-0.660803	-0.444475
English	-0.688465	-0.289771
History	-0.516356	-0.639552
Arithmetic	-0.735620	0.417018
Algebra	-0.741868	0.372759
Geometry	-0.678168	0.354100

Zátěže ukazují, jakou měrou přispívají jednotlivé původní proměnné do hlavních komponent, představují míru korelace mezi původními proměnnými a novou komponentou.

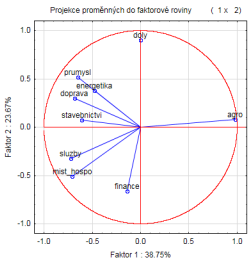
Zátěže informují o tom, které proměnné nejvíce „zatěžují“ jednotlivé nové komponenty (které v nich mají největší zastoupení).

Interpretace komponentních vah (zátěží)

- Pro vlastní interpretaci nových proměnných potřebujeme, aby původní proměnnou významně „zatěžoval“ pouze jeden faktor a aby u ostatních faktorů nabývaly zátěže malých hodnot.
- Pro identifikaci struktury v datech jsou důležité absolutní hodnoty zátěží.
- Cílem je dát vypočteným faktorům konkrétní význam, název, označení,…
- Strukturu lze odhalit i na základě zkušenosti.
- K lepší interpretaci výsledků PCA lze provést jejich rotaci

Graf komponentních vah

Na grafu komponentních vah pro dvě různé faktory či komponenty (např. PC1 a PC2) jsou na místo objektů jejich znaky a lze tak vyšetřovat závislosti a podobnosti mezi znaky.



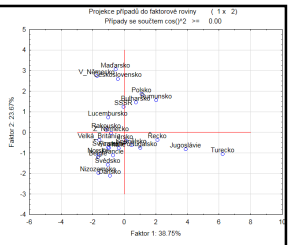
Porovnáváme vzdálenosti mezi proměnnými, malá vzdálenost znamená silnou korelaci.

Hledáme shluk podobných proměnných, jež spolu korelují

Z grafu je patrné, jakou měrou přispívají jednotlivé původní proměnné do hlavních komponent.

Komponentní skóre

Souřadnice každého objektu v prostoru hlavních komponent



Graf komponentních skóre

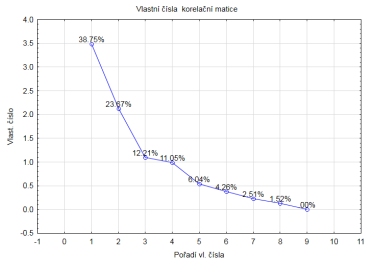
- Daleko od počátku jsou extrémy, objekty nejblíže počátku jsou nejtypičtější. Objekty blízko sebe si jsou podobné, daleko od sebe jsou si nepodobné.
- Objekty umístěné zřetelně v jednom shluku jsou si podobné a nepodobné objektům v ostatních shlucích.
- Umístění objektu na ploše v diagramu může být porovnáváno s komponentními vahami původních proměnných a slouží k interpretaci faktorů i shluků jednotlivých objektů.

Graf úpatí vlastních čísel (sutinový graf, scree plot)

Je to spojnicový graf vlastních čísel proti pořadovým číslům hlavních komponent

Vlastní číslo (eigenvalue) představuje hodnotu rozptylu vysvětleného komponentou

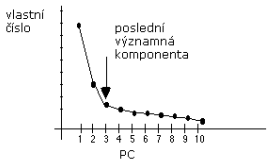
Graf slouží k určení počtu "užitečných" (významných) hlavních komponent.



Graf úpatí vlastních čísel (sutinový graf, scree plot)

K rozhodování o počtu významných a interpretovatelných nových komponent lze použít dvě základní kritéria:

1. Je-li hodnota vlastního čísla větší než 1, potom daná komponenta vysvětluje více celkového rozptylu než jedna původní proměnná.



2. Hledáme zřetelný zlom v průběhu křivky, která prezentuje spojnici hodnot celkového rozptylu vysvětleného jednotlivými komponentami.

Příklad

Vstupní data: výsledky dosažené ve výběru 220 žáků v šesti předmětech:

1. gaelština
2. angličtina
3. dějepis
4. aritmetika
5. algebra
6. geometrie

Korelační matice vstupních dat

1.00					
0.44	1.00				
0.41	0.35	1.00			
0.29	0.35	0.16	1.00		
0.33	0.32	0.19	0.59	1.00	
0.25	0.33	0.18	0.47	0.46	1.00

Příklad – výstup: vlastní čísla a zátěže

Eigenvalues

No.	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	2.728683	45.48	45.48	||||||||||
2	1.128792	18.81	64.29	||||
3	0.615291	10.25	74.55	|||
4	0.602809	10.05	84.59	|||
5	0.522514	8.71	93.30	||
6	0.401910	6.70	100.00	||

Factor Loadings

Variables	Factor1	Factor2
Gaelic	-0.660803	-0.444475
English	-0.688465	-0.289771
History	-0.516356	-0.639552
Arithmetic	-0.735620	0.417018
Algebra	-0.741868	0.372759
Geometry	-0.678168	0.354100

Příklad – výstup: vlastní čísla a zátěže (výsledek po provedení rotace)

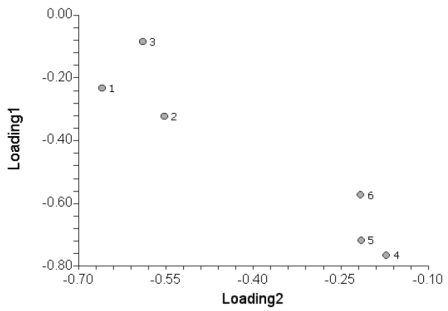
Eigenvalues after Varimax Rotation

No.	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	1.596863	56.94	56.94	|||||||||||
2	1.207981	43.08	100.02	||||||||
3	0.050820	1.81	101.83	|
4	0.011910	0.42	102.26	|
5	-0.008657	-0.31	101.95	|
6	-0.054642	-1.95	100.00	|

Factor Loadings after Varimax Rotation

Variables	Factor1	Factor2
Gaelic	-0.233132	-0.659253
English	-0.322810	-0.552071
History	-0.084713	-0.589192
Arithmetic	-0.765986	-0.170657
Algebra	-0.718105	-0.214689
Geometry	-0.573340	-0.214994

Příklad

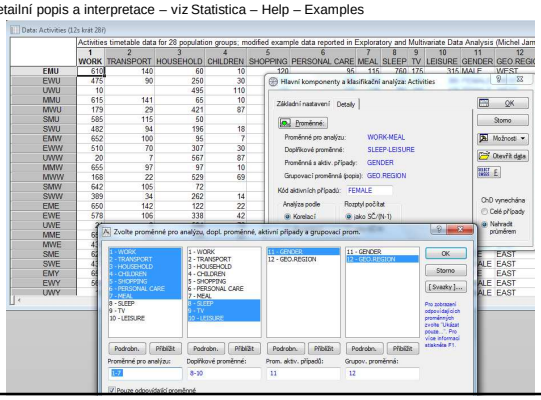


Korelační strukturu pozorovaných dat lze vysvětlit dvěma faktory. První faktor vyjadřuje matematickou dispozici žáka, druhý dispozici jazykově-humanitní.

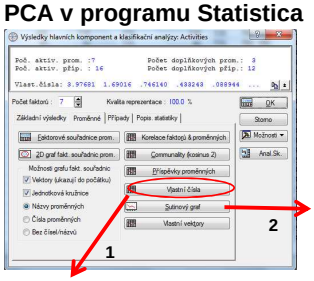
PCA v programu Statistica

Statistiky – Vícerozměrné průzkumné techniky – Hlavní komponenty & klasifikační analýza

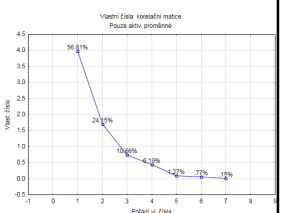
Detailní popis a interpretace – viz Statistica – Help – Examples

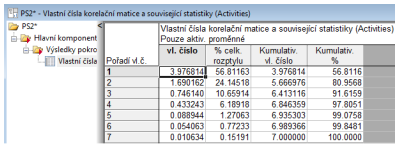


PCA v programu Statistica



Graf slouží k určení počtu významných komponent. V tomto příkladu mají první dvě komponenty vlastní číslo > 1

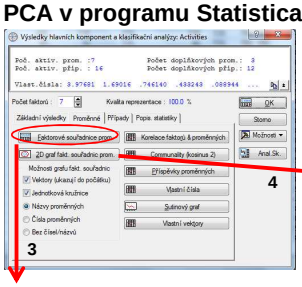


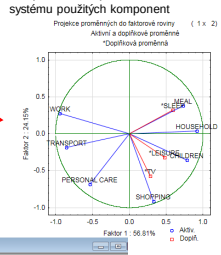


Pořadí vl. č.	vl. číslo	% celk. rozptylu	Kumulativ. vl. číslo	Kumulativ. %
1	3.976814	56.81163	3.976814	56.8116
2	1.690162	24.14518	5.666976	80.9568
3	0.746140	10.65914	6.413116	91.6159
4	0.433243	6.18918	6.846359	97.8051
5	0.088944	1.27063	6.935303	99.0758
6	0.054063	0.77233	6.989366	99.8481
7	0.010634	0.15191	7.000000	100.0000

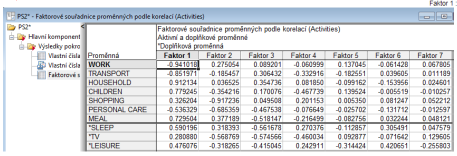
První dvě nové komponenty popisují téměř 81 % celkového rozptylu původních dat.

PCA v programu Statistica





Čím blíže je proměnná kružnici, tím lépe je tato proměnná reprezentována v souřadném systému použitých komponent



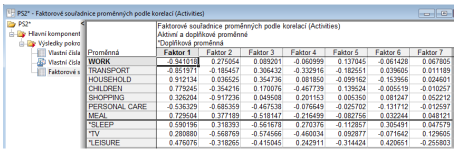
Proměnná	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5	Faktor 6	Faktor 7
WORK	-0.941018	0.275054	0.089201	-0.060999	0.137045	-0.061428	0.067805
TRANSPORT	-0.851971	-0.185457	0.306432	-0.332916	-0.182551	0.039605	0.011189
HOUSEHOLD	0.912134	0.036525	0.354736	0.081850	-0.099162	-0.153956	0.024601
CHILDREN	0.775245	-0.354216	0.170076	-0.467739	0.139524	-0.005519	-0.010257
SHOPPING	0.326204	-0.917236	0.045608	0.201153	0.005350	0.081247	0.052212
PERSONAL CARE	-0.536329	-0.685359	-0.467538	-0.076649	-0.025702	-0.131712	-0.012597
MEAL	0.729504	0.377189	-0.518147	-0.216499	-0.082756	0.032244	0.048121
*SLEEP	0.590196	0.318393	-0.561678	0.270376	-0.112857	0.305491	0.047579
*TV	0.280880	-0.568769	-0.574566	-0.460034	0.092877	-0.071642	0.129605
*LEISURE	0.476076	-0.318265	-0.415045	0.242911	-0.314424	0.420651	-0.255803

Zátěže - korelace mezi původními proměnnými a novými hlavními komponentami (faktory)

Vysoké či nízké hodnoty zátěží lze využít k „pojmenování" komponent

PCA v programu Statistica

3



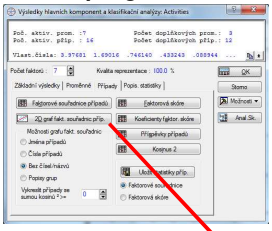
Proměnná	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5	Faktor 6	Faktor 7
WORK	-0.941018	0.275054	0.089201	-0.060999	0.137045	-0.061428	0.067805
TRANSPORT	-0.851971	-0.185457	0.306432	-0.332916	-0.182551	0.039605	0.011189
HOUSEHOLD	0.912134	0.036525	0.354736	0.081850	-0.099162	-0.153956	0.024601
CHILDREN	0.775245	-0.354216	0.170076	-0.467739	0.139524	-0.005519	-0.010257
SHOPPING	0.326204	-0.917236	0.045608	0.201153	0.005350	0.081247	0.052212
PERSONAL CARE	-0.536329	-0.685359	-0.467538	-0.076649	-0.025702	-0.131712	-0.012597
MEAL	0.729504	0.377189	-0.518147	-0.216499	-0.082756	0.032244	0.048121
*SLEEP	0.590196	0.318393	-0.561678	0.270376	-0.112857	0.305491	0.047579
*TV	0.280880	-0.568769	-0.574566	-0.460034	0.092877	-0.071642	0.129605
*LEISURE	0.476076	-0.318265	-0.415045	0.242911	-0.314424	0.420651	-0.255803

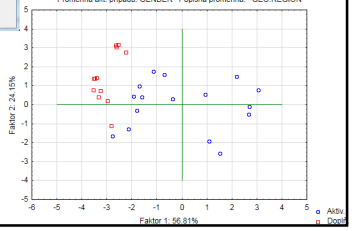
Interpretace hlavních komponent:

První komponenta nejvíce koreluje s proměnnými WORK a TRANSPORT (záporná korelace) a s proměnnými HOUSEHOLD a CHILDREN (positivní korelace) – nová osa diferencuje **pracovní vs. domácí aktivity**

Druhá komponenta vykazuje největší negativní korelaci s proměnnými SHOPPING a PERSONAL CARE - **activities required by modern organized life**

PCA v programu Statistica





První komponenta dobře diferencuje aktivity mužů a žen