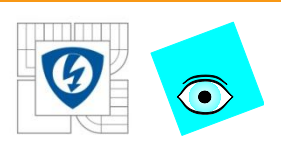


Počítačová analýza vícerozměrných dat v oborech přírodních, technických a společenských věd

Prof. RNDr. Milan Meloun, DrSc. (Univerzita Pardubice, Pardubice)

20.-24. června 2011

Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky.

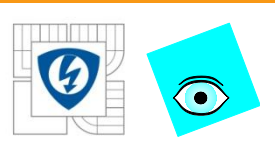


4.4 METODA HLAVNÍCH KOMPONENT (PCA)

24.2.2010

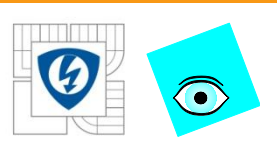
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





Určení vzájemných vazeb

- a) struktura a vazby v proměnných
 - b) struktura a vazby v objektech
- 1) Hledání struktury v proměnných (metrická škála): faktorová analýza FA, analýza hlavních komponent PCA a shluková analýza.
 - 2) Hledání struktury v objektech (metrická škála): shluková analýza
 - 3) Hledání struktury v objektech (nemetrická škála): korespondenční analýza
 - 4) Hledání lineárních vícerozměrných modelů (metrická i nemetrická škála): většina metod vícerozměrné statistické analýzy, kde závisle proměnné se uvažují jako lineární kombinace nezávisle proměnných.

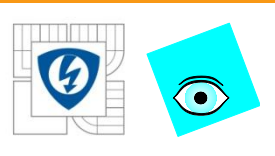


Analýza hlavních komponent (PCA)

Zaměření metody PCA: PCA jedna z nejstarších a nejvíce používaných metod vícerozměrné analýzy.

Zavedena: Pearsonem již v roce 1901 a nezávisle Hotellingem v roce 1933.

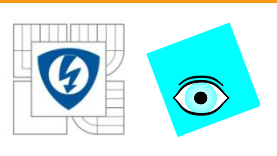
Cílem PCA: zjednodušení popisu lineárně závislých tj. korelovaných znaků, a to rozkladem matice dat do **matice strukturní** (=využité hlavní komponenty) a do **matice šumové** (=nevyužité hlavní komponenty).



Analýza hlavních komponent (PCA)

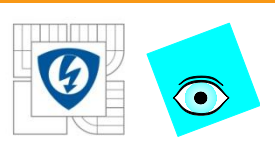
Popis metody PCA:

1. Lineární transformace původních znaků na nové, nekorelované proměnné, zvané *hlavní komponenty*.
2. Základní charakteristikou každé hlavní komponenty je její míra variability čili *rozptyl*.
3. Hlavní komponenty jsou *seřazeny dle důležitosti*, tj. klesajícího rozptylu, od největšího k nejmenšímu.
4. *Většina informace* o variabilitě dat je přitom soustředěna do první komponenty a nejméně informace je obsaženo v poslední komponentě.



Analýza hlavních komponent (PCA)

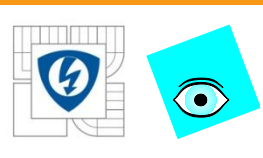
5. *Pravidlo*: má-li nějaký původní znak malý či dokonce žádný rozptyl, není schopen přispívat k rozlišení mezi objekty.
6. Využitím PCA je *snížení dimenze* úlohy čili redukce počtu znaků bez velké ztráty informace, užitím pouze prvních několika hlavních komponent.
7. *Nevyužité hlavní komponenty* obsahují malé množství informace, protože jejich rozptyl je příliš malý.
8. Hlavní komponenty jsou nekorelované.



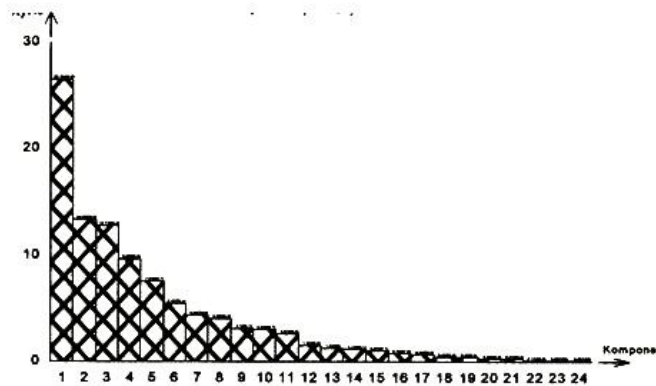
Analýza hlavních komponent (PCA)

9. První hlavní komponenta je například vhodným ukazatelem jakosti.

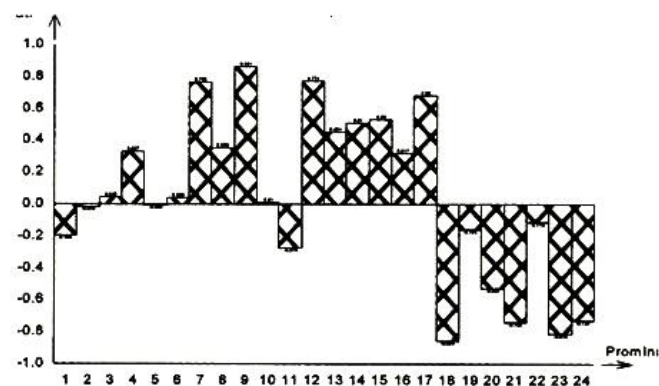
10. První dvě resp. tři hlavní komponenty se využívají především jako techniky zobrazení vícerozměrných dat v projekci do roviny (nebo do prostoru).



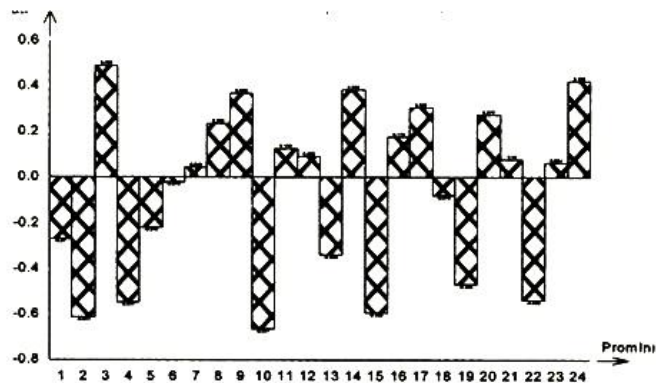
Analýza hlavních komponent (PCA)



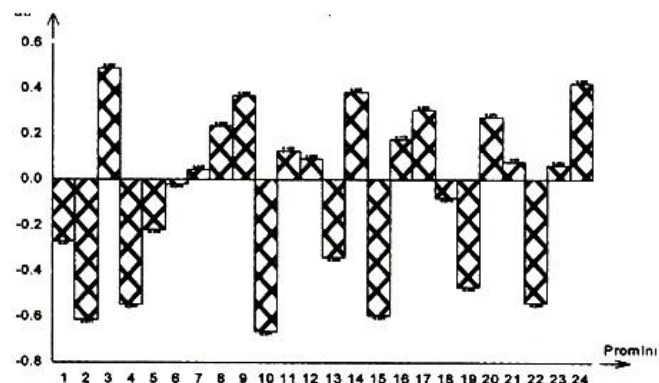
Obr. 1. Sloupový diagram indexového grafu úpatí pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



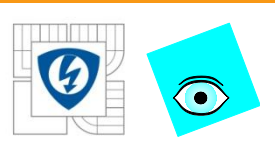
Obr. 2. Složení 1. hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



Obr. 3. Složení 2. hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



Obr. 4. Složení 3. hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



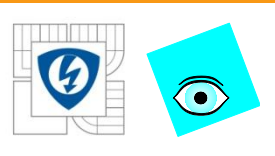
Podstata metody PCA

- **Zdrojová matice dat X** ($n \times m$): obsahuje n objektů a m znaků.

Objekty jsou pozorování, vzorky, experimenty, měření, pacienti, rostliny, atd.

Znaky či proměnné jsou druhy signálu měření, měřená veličina, vlastnosti (sladký, kyselý, hořký, slaný, cholerický, atd.), barva, apod.

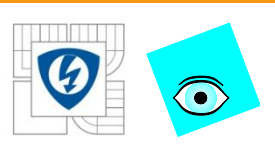
- Každý znak je znám pro všech n objektů.
- Cílem vícerozměrné analýzy je indikovat model a odkrýt skrytý jev.
- Myšlenka sledování rozptylu je velice důležitá, že „nalezené směry maximálního rozptylu“ jsou spjaté s



Podstata metody PCA

- **Zobrazení objektů v m rozměrech:** matice dat $X(n \times m)$ s m sloupci znaků a n řádky objektů může být zobrazena v m-rozměrném euklidovském prostoru, tj. ortogonálním systému souřadnic rozměru m.

Osy proměnných jsou ortogonální a mají společný počátek ale mají různé měrné jednotky.



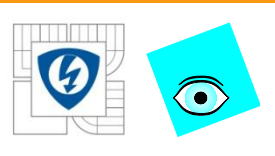
Analýza hlavních komponent (PCA)

Cílem PCA: transformace dat z původních proměnných $x_i, j = 1, \dots, m$, do menšího počtu latentních proměnných y_i .

PCA vystihuje téměř proměnlivost původních proměnných a jsou vzájemně nekorelované.

První hlavní komponenta y_1 popisuje největší část proměnlivosti čili rozptylu původních dat.

Druhá hlavní komponenta y_2 popisuje zase největší část rozptylu neobsaženého v y_1 atd.



Cíl metody hlavních komponent PCA

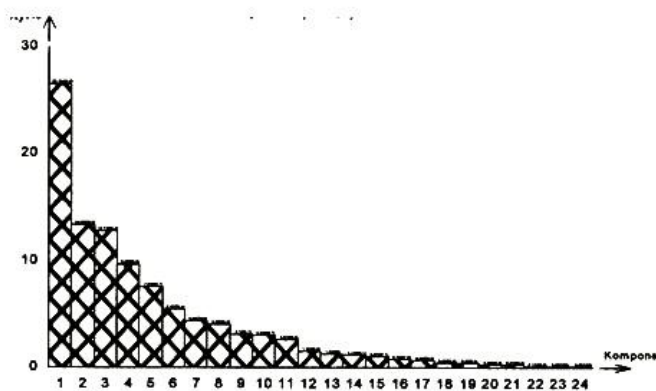
Cílem PCA je transformace původních znaků x_i , $j = 1, \dots, m$, do menšího počtu latentních proměnných y_j .

Latentní proměnné mají vhodnější vlastnosti:

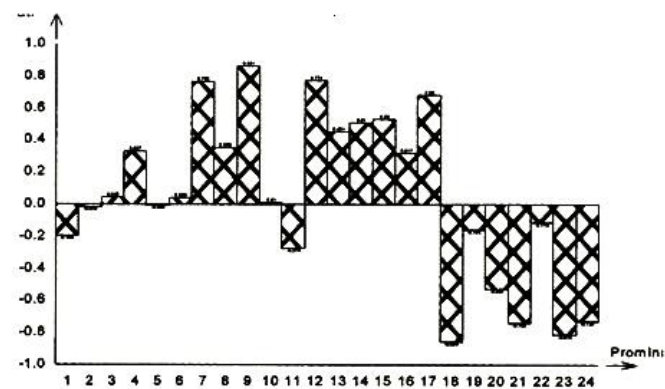
- je jich výrazně méně, vystihují téměř celou *proměnlivost* původních znaků,
- jsou vzájemně nekorelované,
- jsou nazvány *hlavními komponentami*, čili lineární kombinace původních proměnných.



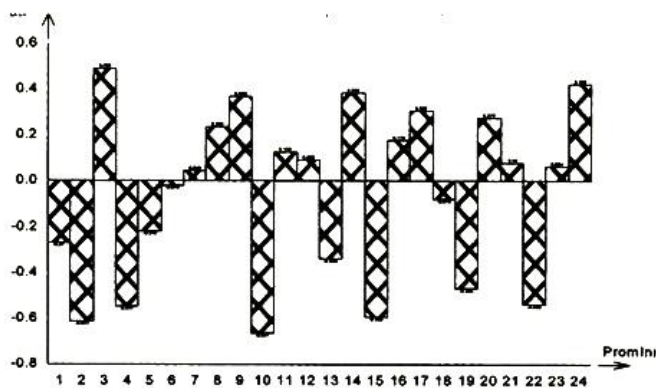
Cíl metody hlavních komponent PCA



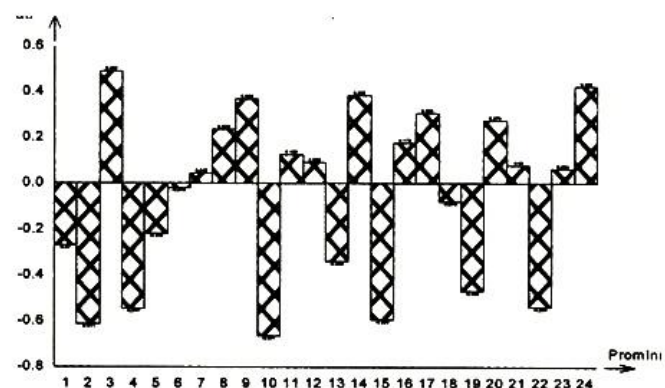
Obr. 1. Sloupcový diagram indexového grafu úpatí pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



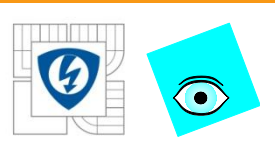
Obr. 2. Složení 1. hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



Obr. 3. Složení 2. hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



Obr. 4. Složení 3. hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



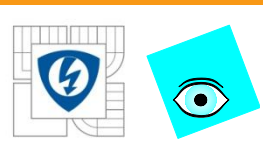
Hlavní komponenta

Matematicky: první hlavní komponenta je takovou lineární kombinací vstupních znaků, která pokrývá největší rozptyl mezi všemi ostatními lineárními kombinacemi.

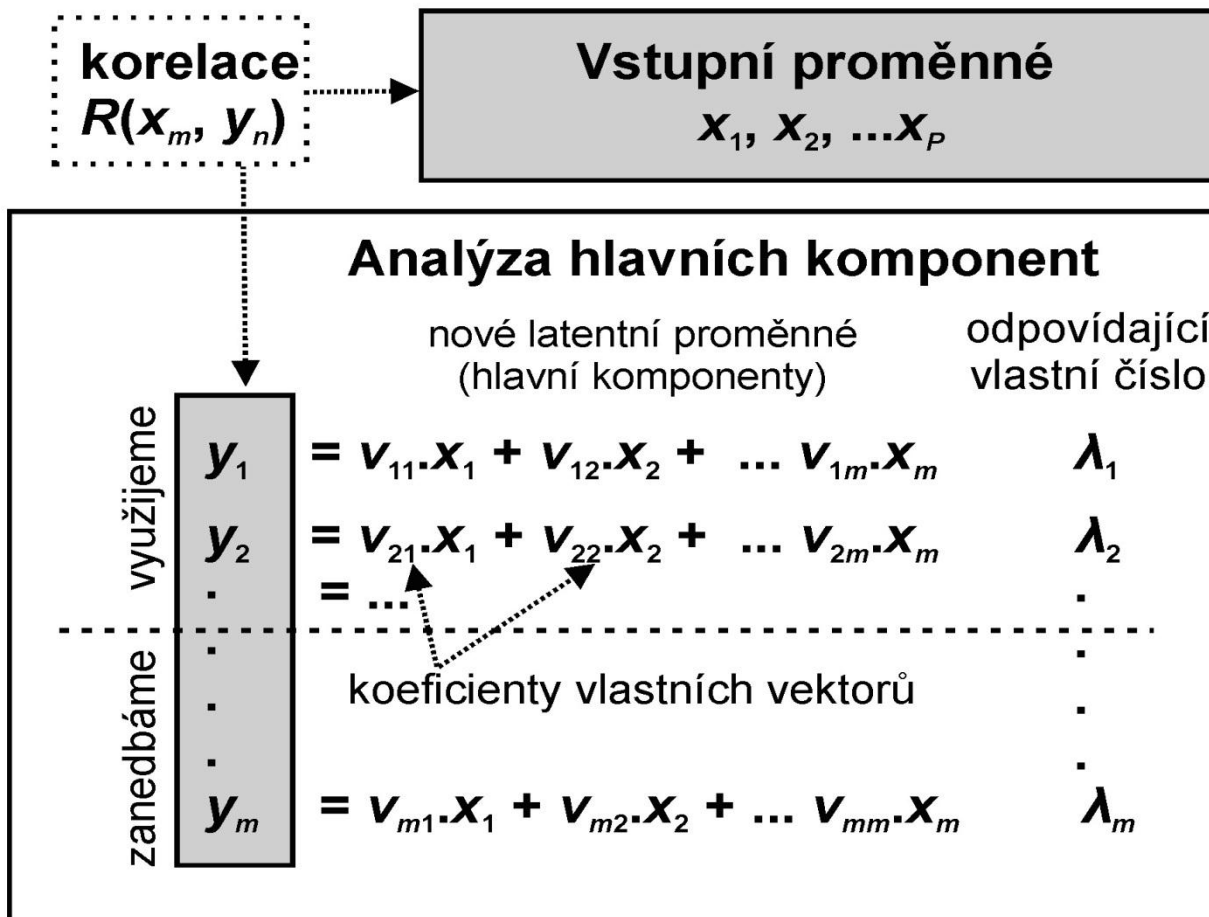
Ztráta informace projekcí do menšího počtu rozměrů se nazývá mírou těsností proložení modelu PCA nebo také chybou modelu PCA.

I při velkém počtu původních znaků m můžeme být k velmi malé, běžně 2 až 5.

Vysvětlení užitých hlavních komponent: jejich pojmenování a vysvětlení vztahu původních znaků $x_j, j = 1, \dots, m$, k hlavním komponentám $y_j, j = 1, \dots, k$, tvoří dominantní součásti analýzy modelu hlavních komponent PCA



Analýza hlavních komponent



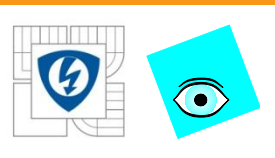
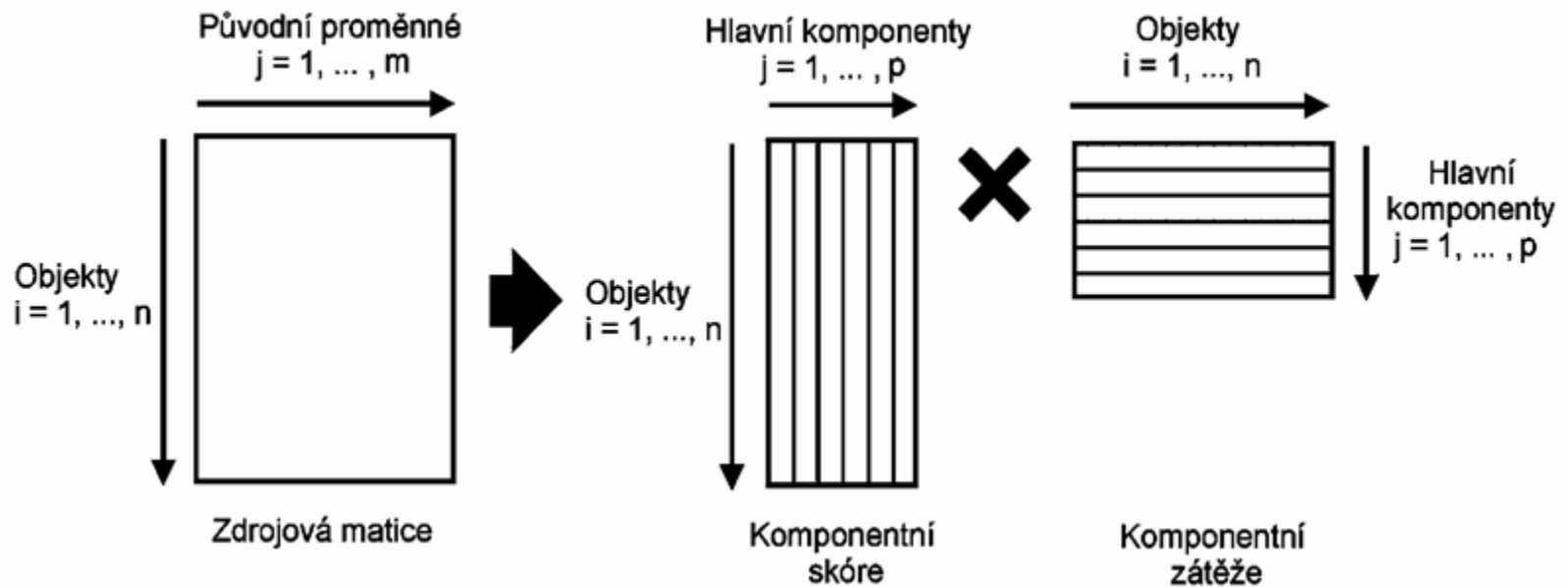
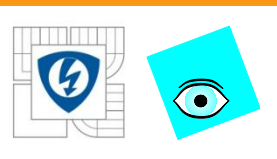


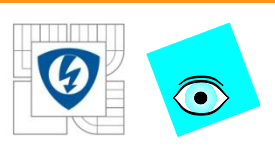
Schéma maticových výpočtů v PCA





Analýza hlavních komponent

- **Maximální počet hlavních komponent:** počet efektivních hlavních komponent se rovná hodnoti zdrojové matice X .
- **$X = \text{Struktura} + \text{šum}$:** všechny hlavní komponenty jsou vzájemně ortogonální a souvisí postupně se snižující hodnotou rozptylu objektů.
- Model hlavních komponent má tvar
$$X = TP^T + E = \text{struktura dat} + \text{šum}.$$
- **Střed modelu:** hlavní komponenty mají společný počátek, který odpovídá průměrnému objektu čili

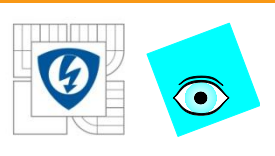


Analýza hlavních komponent

Po dosazení vyjde $D(y_1) = v_1^T S v_1 \lambda_1$ (to jsou koeficienty hlavních komponent $y_1, \dots, y_m$).

Pravidlo:

1. „Velké“ hlavní komponenty odpovídají nejdůležitější informaci.
2. „Malé“ hlavní komponenty odpovídají spíše šumu.



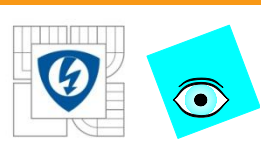
Komponentní váhy

zátěže – vztah mezi X a PC:

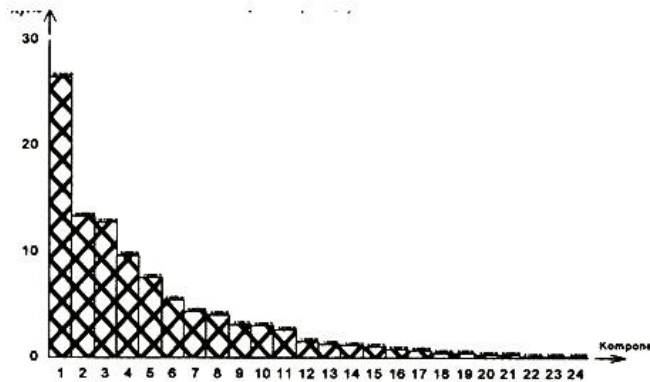
Hlavní komponenta představuje *lineární kombinaci* všech m vektorů v prostoru znaků v m rozměrném prostoru a jejich koeficienty se nazývají *komponentní váhy*.

Komponentní váhy informují o vztahu mezi původními m znaky a hlavními komponentami.

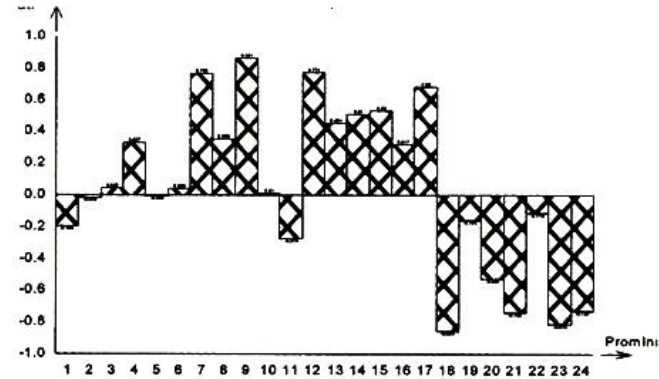
Na grafu komponentních vah p pro PC1 a PC2 jsou místo objektů jejich znaky a lze tak vyšetřovat závislosti a podobnosti mezi znaky.



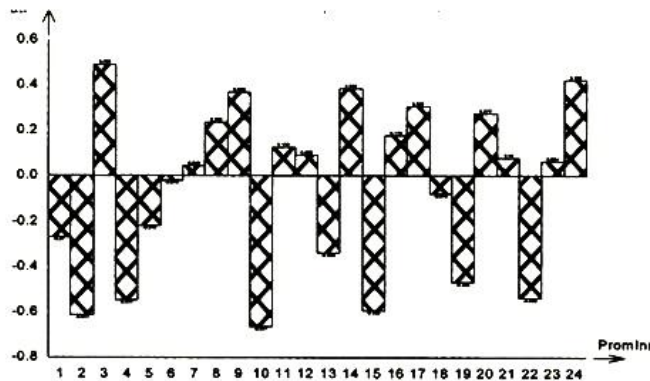
Komponentní váhy



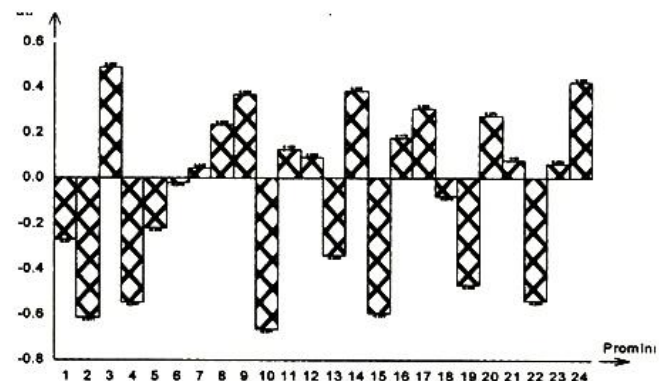
Obr. 1. Sloupceový diagram indexového grafu úpatí pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



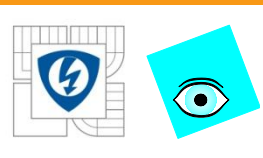
Obr. 2. Složení 1. Hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



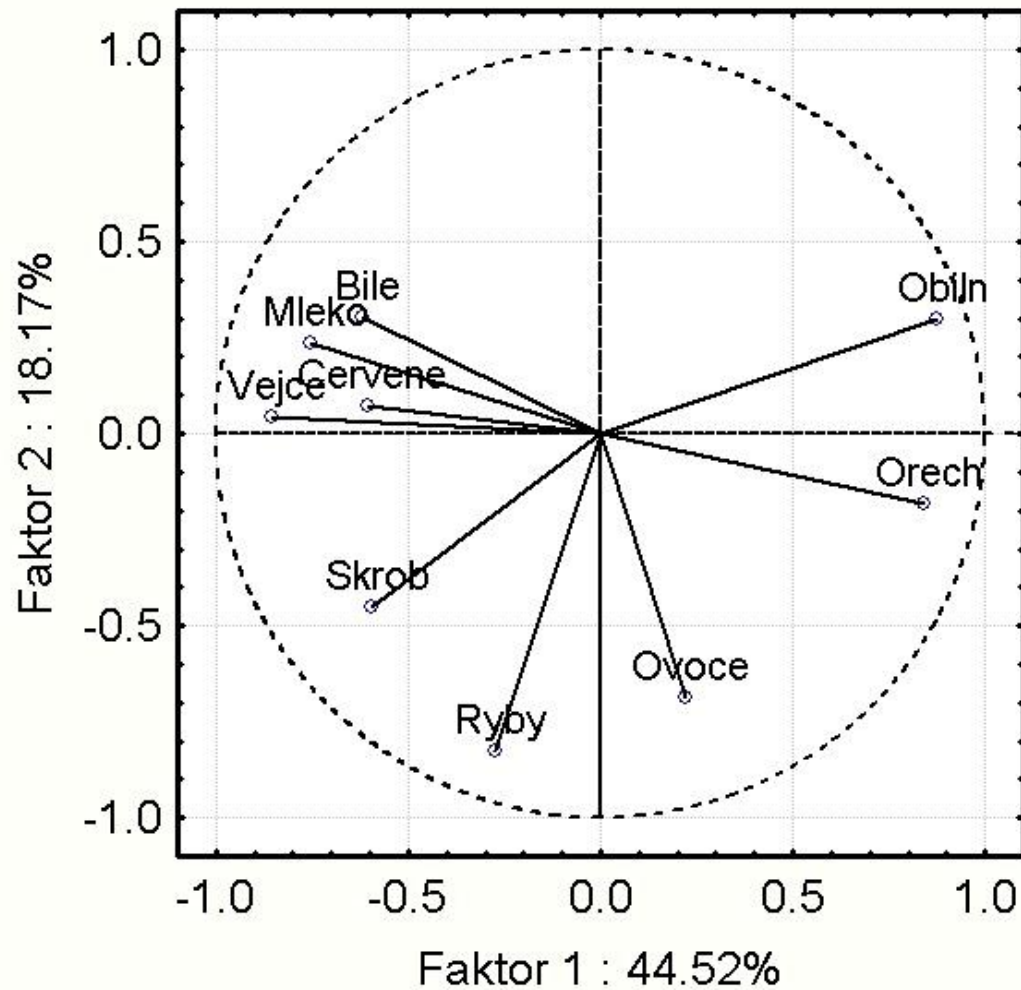
Obr. 3. Složení 2. Hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.

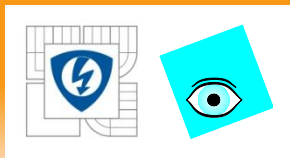


Obr. 4. Složení 3. Hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



Komponentní váhy





Graf komponentních vah, zátěží

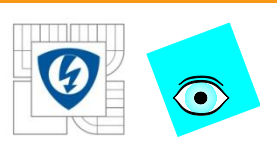
Zobrazí: komponentní váhy

Porovnávají se: vzdálenosti mezi proměnnými. Krátká znamená silnou korelaci.

Nalezneme: shluk podobných proměnných, jež spolu korelují.

Představuje: most mezi původními proměnnými a hlavními komponentami.

Ukazuje: jakou měrou přispívají jednotlivé původní proměnné do hlavních komponent.

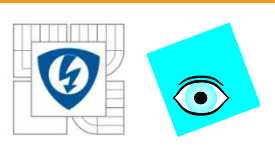


Graf komponentních vah, zátěží

Pojmenovat: podaří se hlavní komponenty y_1, y_2 a vysvětlit a přidělit jim fyzikální, chemický nebo biologický význam.

Původní proměnné x_j přispívají: kladnou vahou nebo zápornou.

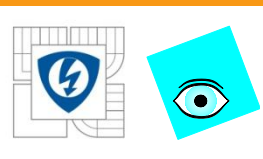
Sledujeme: kovarianci původních proměnných x_j v grafu komponentních vah y_1, y_2 a y_3 : jsou-li proměnné $x_j, j = 1, \dots, m$, blízko sebe v prostorovém zhluku, jde o silnou pozitivní kovarianci.



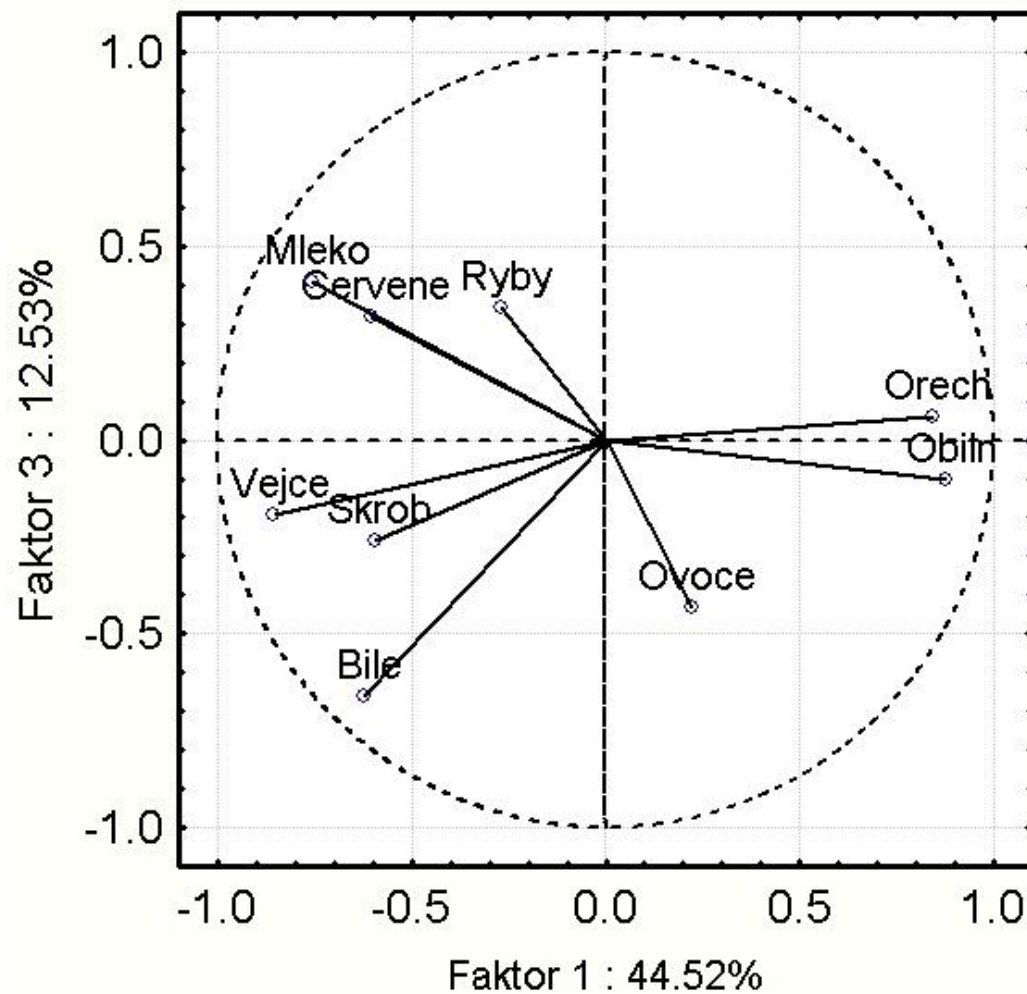
Korelace a kovariance

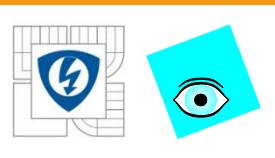
$x_j, j = 1, \dots, m$, blízko sebe anebo s malým úhlem mezi průvodiči mají vysokou kladnou kovarianci a korelaci.

$x_j, j = 1, \dots, m$, daleko od sebe anebo s velikým úhlem mezi průvodiči jsou negativně korelovány.

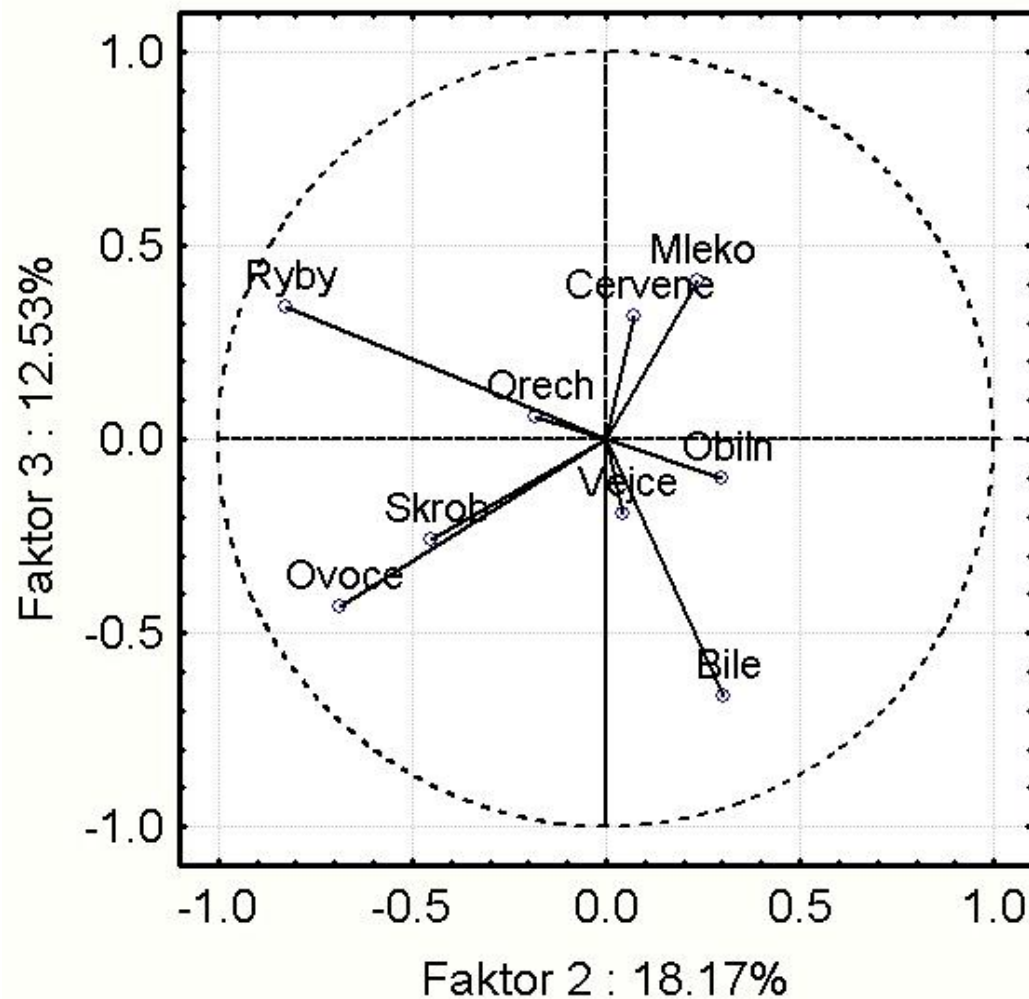


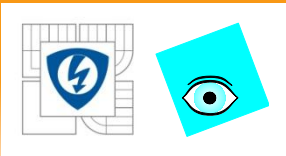
Korelace a kovariance





Korelace a kovariance





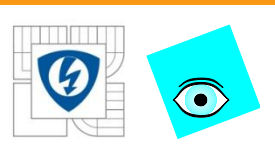
Komponentní skóre – souřadnice objektů v prostoru hlavních komponent

Souřadnice každého objektu na osách hlavních komponent nazýváme skóre.

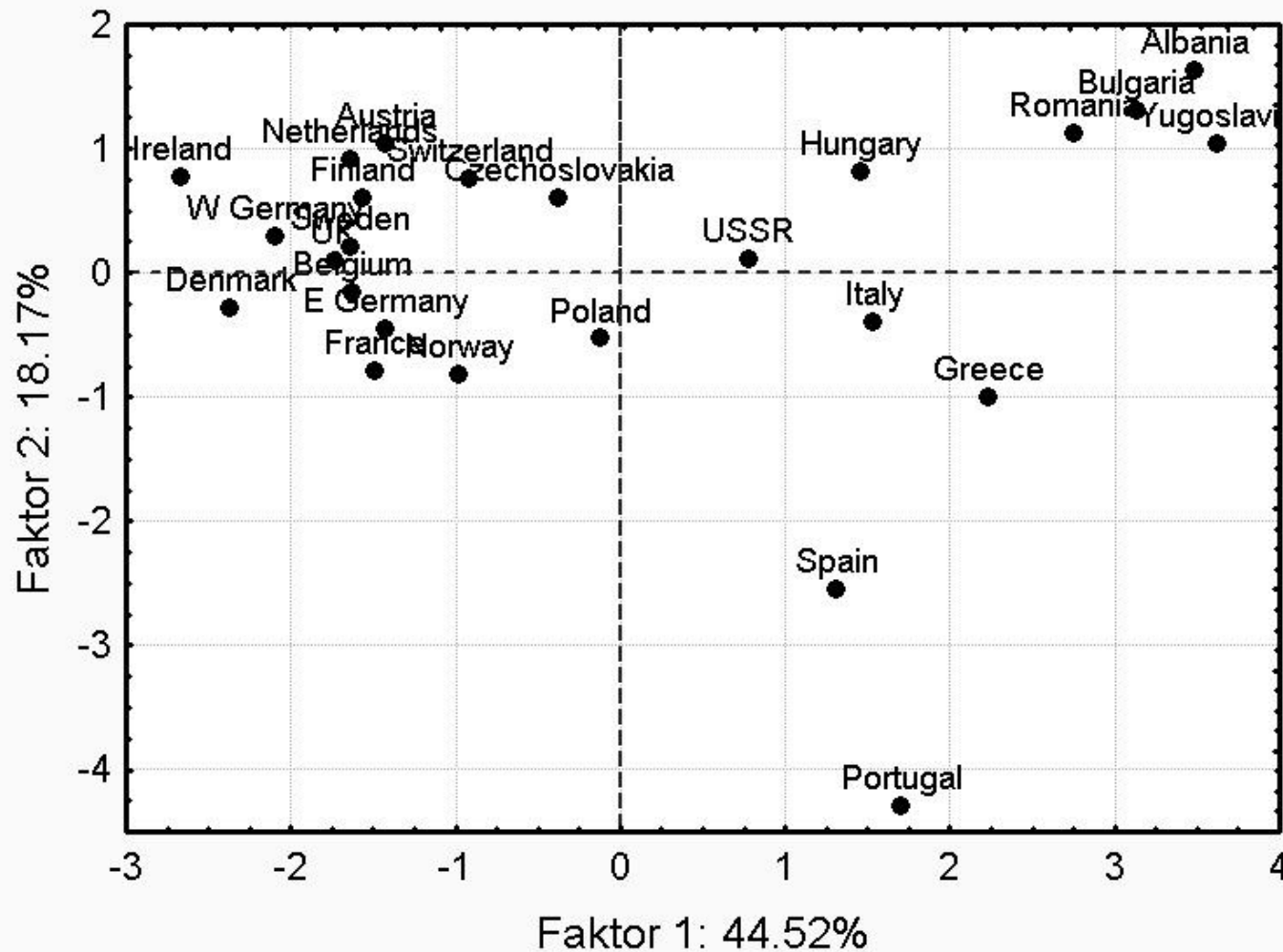
Graf komponentního skóre je zobrazení dvou skórových vektorů vynesných v systému kartézských os jeden proti druhému.

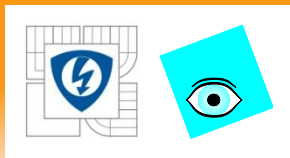
Pravidlo k volbě grafů:

1. Na x-ové ose užitíme vždy stejnou hlavní komponentu (obvykle první) u všech grafů komponentního skóre: t_1 proti t_2 , t_1 proti t_3 , t_1 proti t_4 , t_1 proti t_5 , atd., takže budeme vyšetřovat ostatní hlavní komponenty proti stále stejné, první.
2. Užitíme první hlavní komponentu, protože vykazuje největší hodnotu rozptýlení pro vyšetřovanou úlohu a vyneseme ji na



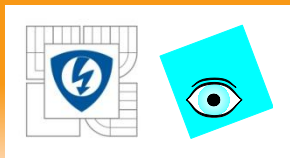
Komponentní skóre – souřadnice objektů v prostoru hlavních komponent





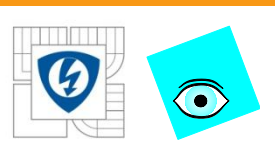
Rozptylový diagram komponentního skóre (Scatterplot)

1. **Umístění objektů:** daleko od počátku jsou extrémny. Objekty nejblíže počátku jsou nejtypičtější.
2. **Podobnost objektů:** objekty blízko sebe si jsou podobné daleko od sebe jsou si nepodobné.
3. **Objekty v shluku:** umístěné zřetelně v jednom shluku jsou si podobné a nepodobné objektům v ostatních shlucích. Jsou-li shluky blízko sebe, znamená to značnou podobnost objektů.

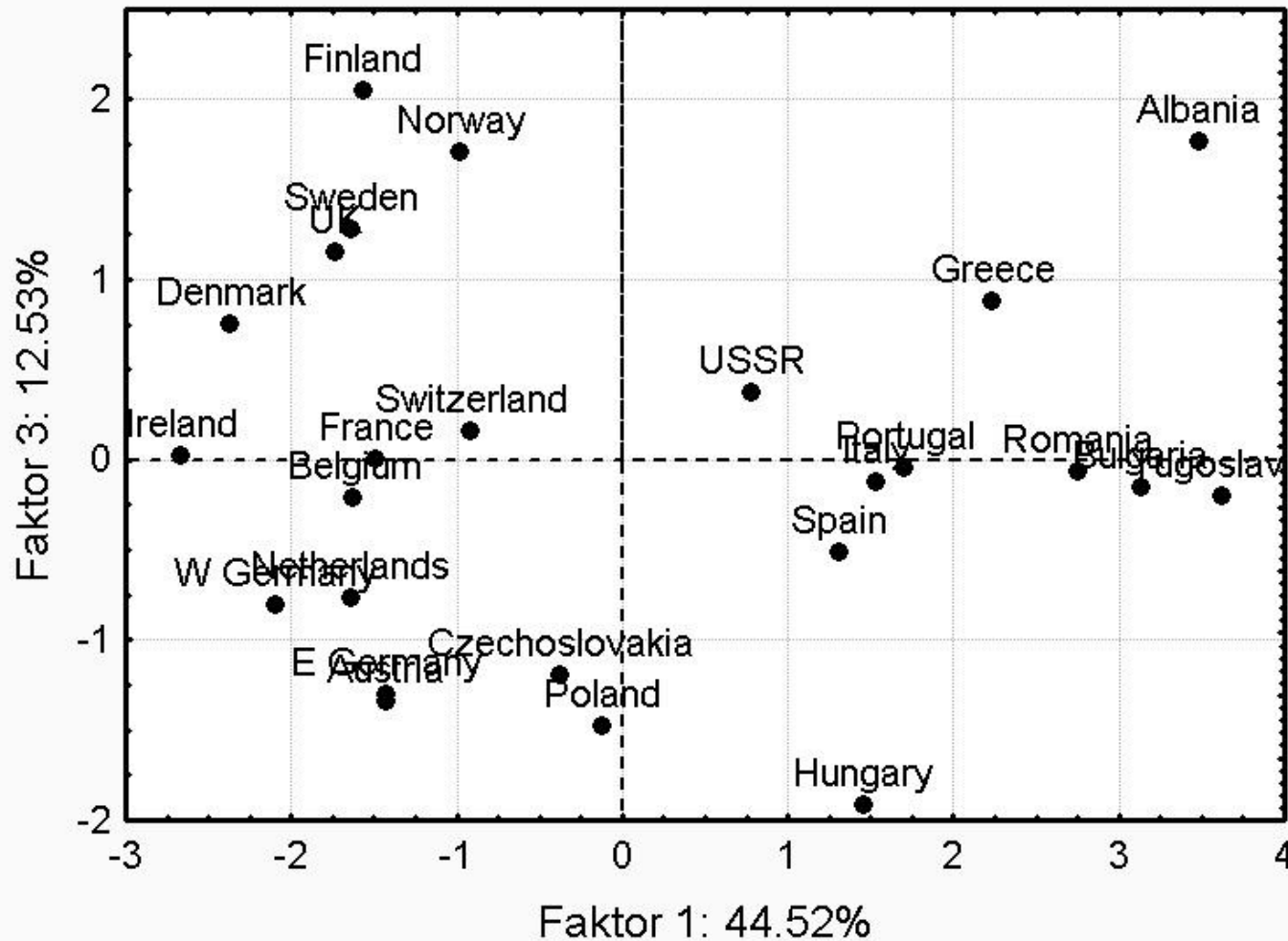


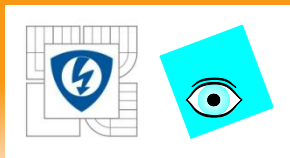
Rozptylový diagram komponentního skóre (Scatterplot)

4. Osamělé objekty: izolované objekty mohou být odlehlé.
5. Odlehlé objekty: ideálně bývají objekty rozptýlené po celé ploše diagramu. V opačném případě je špatný model.
6. Pojmenování objektů: výstižná jména objektů slouží k hledání hlubších souvislostí mezi objekty a vystihneme tak jejich fyzikální či biologický vztah.
7. Vysvětlení místa objektu: umístění objektu na ploše v diagramu může být porovnáváno s komponentními vahami původních proměnných ve dvojném grafu.



Rozptylový diagram komponentního skóre (Scatterplot)





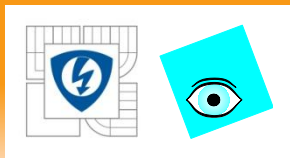
Rozptylový diagram komponentního skóre (Scatterplot)

Zobrazuje: komponentní skóre čili hodnoty obyčejně prvních dvou hlavních komponent u všech objektů.

Dokonalé rozptýlení objektů v rovině: vede k rozlišení objektů při jejich popisu pomocí y_1 a y_2 .

Lze nalézt: shluk vzájemně podobných objektů a dále objekty odlehlé a silně odlišné od ostatních.

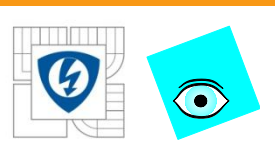
Využití: k identifikaci odlehlých objektů, identifikaci trendů, tříd, shluků objektů, k objasnění podobnosti objektů atd.



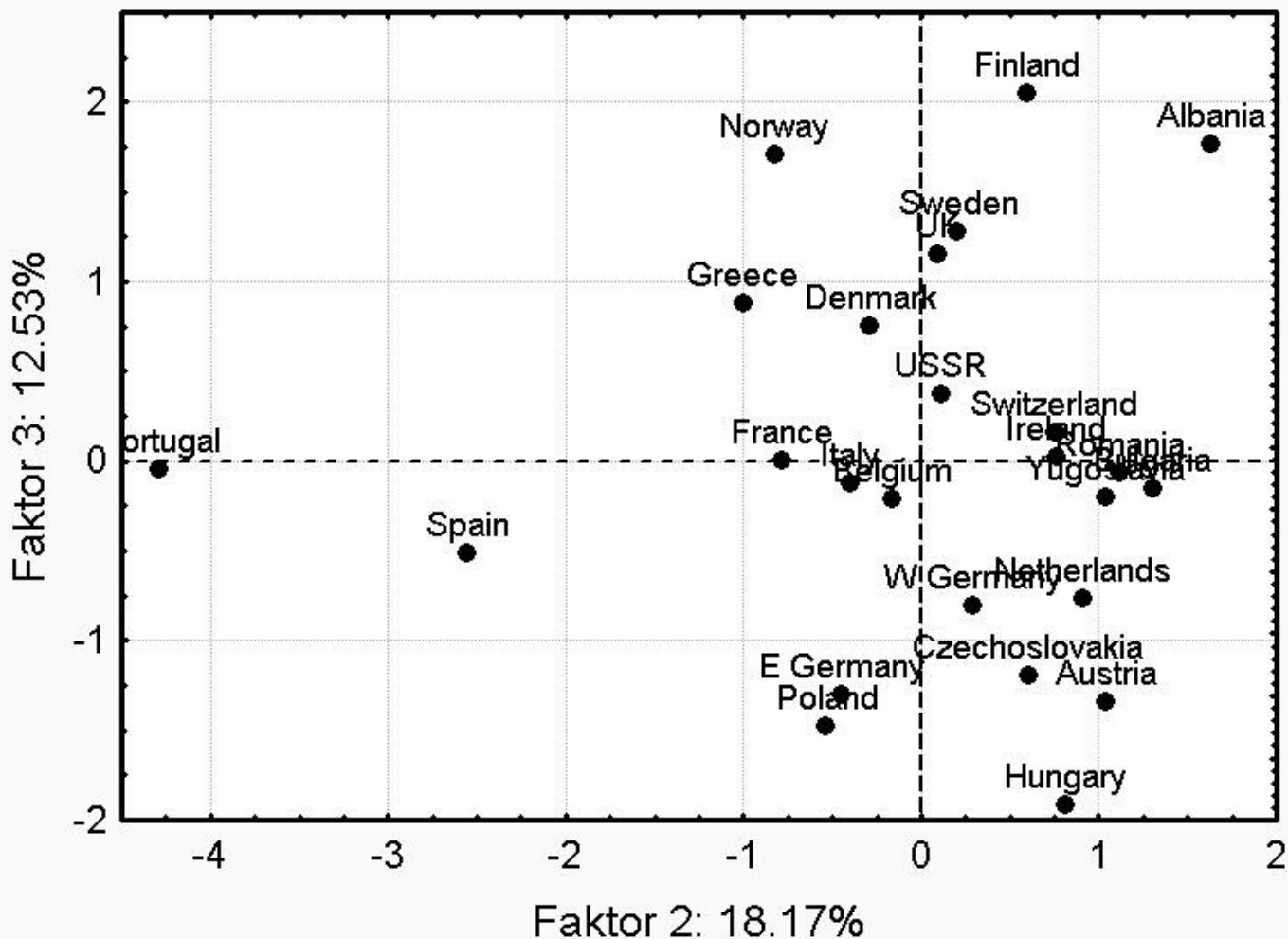
Rozptylový diagram komponentního skóre (Scatterplot)

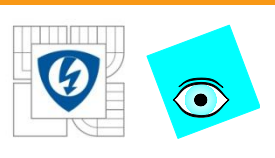
Nemožné analyzovat všechny diagramy: jich je mnoho:
např. $m < n$ pro $m=10$ existuje $m(m-1)/2=45$ diagramů,
pro $m=11$ pak 55 diagramů, pro $m=12$ pak 66 diagramů,
atd.

Vybíráme: diagramy y_1 vs. y_2 , y_1 vs. y_3 , y_1 vs. y_4 , atd. -
držíme se první hlavní komponenty y_1 protože v ní bývá
největší míra proměnlivosti v datech.

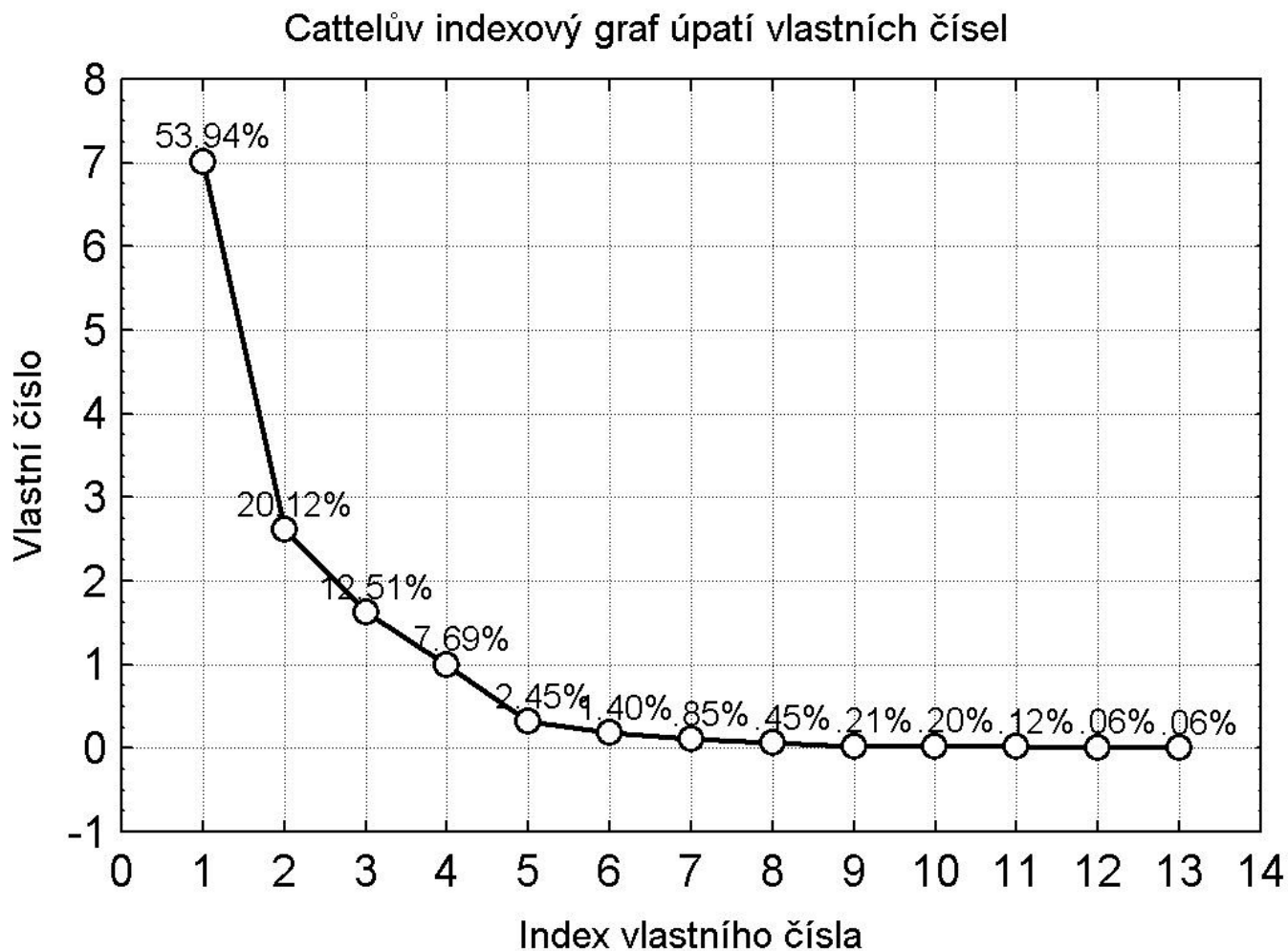


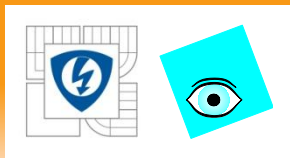
Rozptylový diagram komponentního skóre (Scatterplot)





Rozptylový diagram komponentního skóre (Scatterplot)





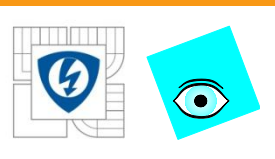
Indexový graf úpatí vlastních čísel (Scree Plot)

Je to sloupcový diagram vlastních čísel proti indexu A .

Zobrazuje: relativní velikost jednotlivých vlastních čísel.

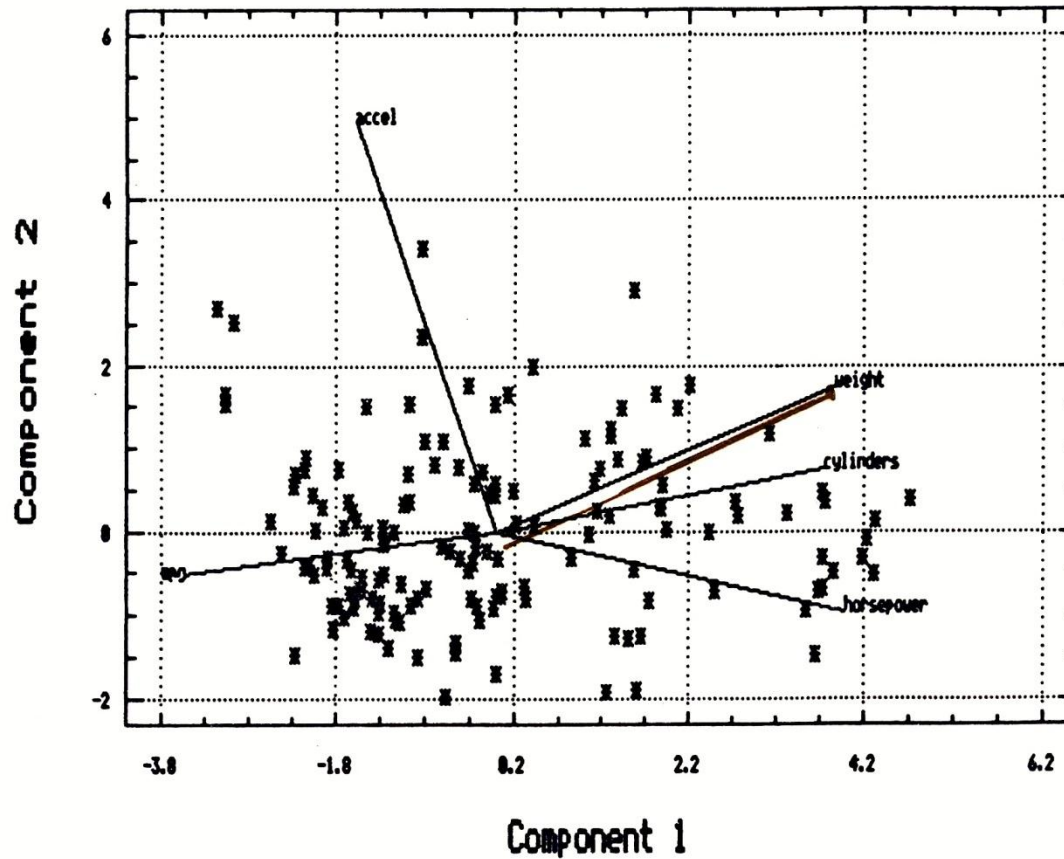
Využití: k určení počtu A "užitečných" hlavních komponent. Graf úpatí se jeví nejobektivnějším kritériem.

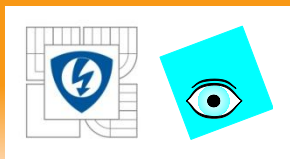
Kritérium "1": hrubším kritériem PC, jejichž vlastní číslo je větší než jedna. Graf úpatí se však jeví objektivnějším.



Biplot for First Two Principal Components

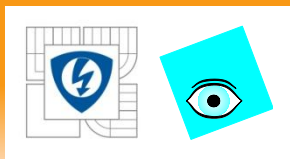
Biplot for First Two Principal Components





Dvojný graf (Biplot) kombinuje oba předchozí grafy

- **Úhel mezi průvodiči** x_j a x_k je nepřímo úměrný velikosti korelace mezi proměnnými x_j a x_k (čím je menší úhel, tím je větší korelace)
- **Délka souřadnice bodu:** je úměrná příspěvku x_j do hlavní komponenty, **čili** je úměrná komponentní váze.
- **Interakce:** objekt je v blízkosti proměnné x_j tzn. objekt "obsahuje" hodně této proměnné.
- **Umístění vpravo (nahore) od nuly:** Interakce proměnných a objektů vysvětluje umístění objektu vpravo od nuly na ose x_1 (či vlevo od nuly) resn



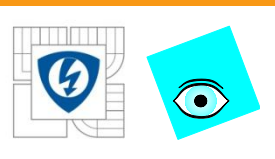
Vzorová úloha 4.2 *Postup metody hlavních komponent*

Na úloze B4.02 *Účinky neuroleptik při tlumení rozličných psychóz* si ukážeme pomůcky vícerozměrné analýzy dat. K analýze uijeme také škálovaná data.

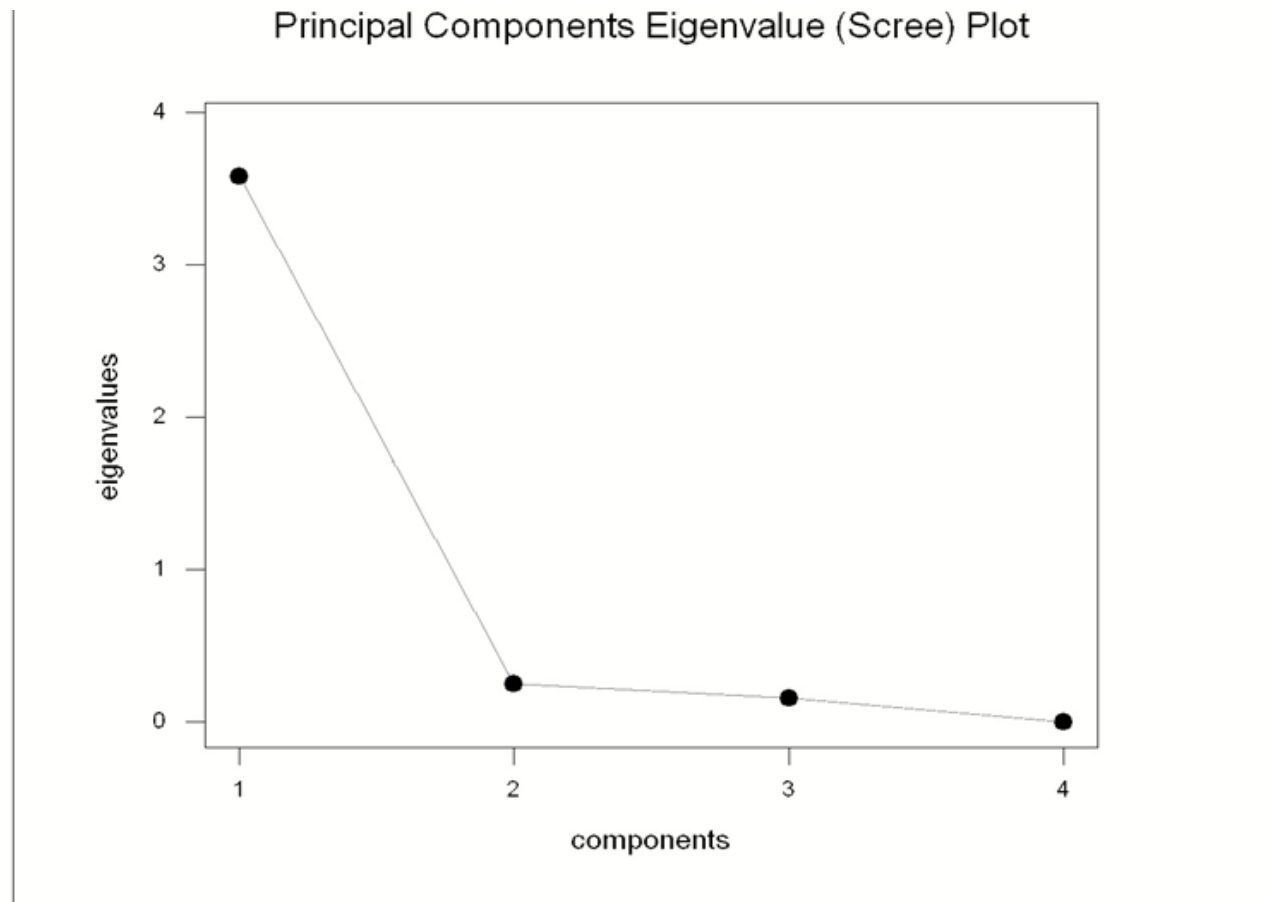
Řešení: k analýze byl použit program NCSS2000. Výstup metody PCA programu NCSS2000 pro nestandardizovaná data úlohy B402 obsahuje:

- 1. Vyšetření indexového grafu úpatí vlastních čísel:

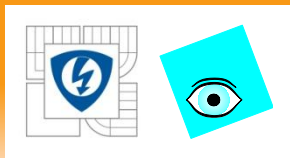
	Vlastní číslo	Individuální procento	Kumulativní procento	Kumulativní čárový graf úpatí
Index				
1	3394.339	92.62	92.62 mm	
2	252.286	6.88	99.50	
3	15.8825	0.43	99.94	
4	2.295	0.06	100.00	



Principle components eigenvalue (Scree) plot

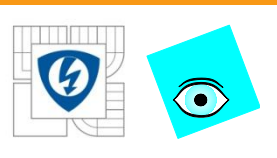


Obr. 4.7 Indexový graf úpatí vlastních čísel (Scree Plot) pro 20 objektů a 4 proměnné B402X1, B402X2, B402X3, B402X4 ze standardizovaných dat B402



Principle components eigenvalue (Scree) plot

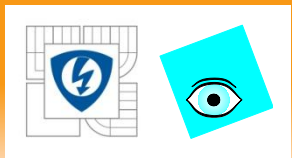
Indexový graf úpatí vlastních čísel je vlastně sloupcový diagram velikostí vlastních čísel proti stoupající hodnotě indexu, pořadového čísla. Určení počtu A „využitelných“ hlavních komponent.



2. (a) Vlastní vektory pro hlavní komponenty:

<i>Proměnná</i>	V_i	y_2	y_3	y_4
<i>B402X1</i>	-0.5684	0.4339	0.532	-0.4534
<i>B402X2</i>	-0.5688	0.4042	-0.584	0.4148
<i>B402X3</i>	-0.0758	-0.0286	0.6125	0.7863
<i>B402X4</i>	-0.5896	-0.8047	-0.0282	-0.0642

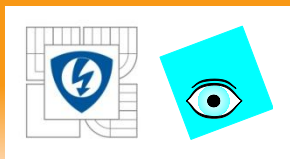
Vlastní vektory jsou váhy, jež umožňují kombinovat komponentní proměnné.



2. (a) Vlastní vektory pro hlavní komponenty:

Např. první hlavní komponenta y_1 je *vážený průměr normovaných proměnných*, kdy váha každé proměnné je dána odpovídajícím prvkem prvního vlastního vektoru.

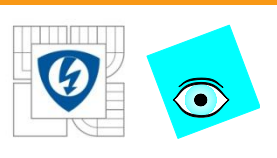
Koeficienty vystihují relativní důležitost každé proměnné při tvorbě hlavní komponenty.



(b) Čárový diagram absolutních hodnot vlastních vektorů:

Proměnná	y_1	y_2	y_3	y_4
<i>B402X1</i>				
<i>B402X2</i>				
<i>B402X3</i>				
<i>B402X4</i>				

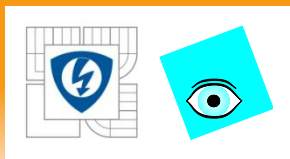
Diagram vysvětluje strukturu, která proměnná silně koreluje s hlavní komponentou.



3. (a) Komponentní váhy:

<i>Proměnná</i>	y_1	y_2	y_3	y_4
<i>B402X1</i>	-0.9769	0.2033	0.0625	-0.0203
<i>B402X2</i>	-0.9793	0.1897	-0.0688	0.01857
<i>B402X3</i>	-0.8486	-0.0874	0.4688	0.2288
<i>B402X4</i>	-0.9372	-0.3487	-0.0031	-0.0026

Ukazuje korelace mezi původními proměnnými a hlavními komponentami.

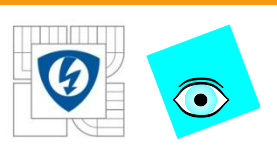


b) Čárový diagram absolutních hodnot komponentních vah:

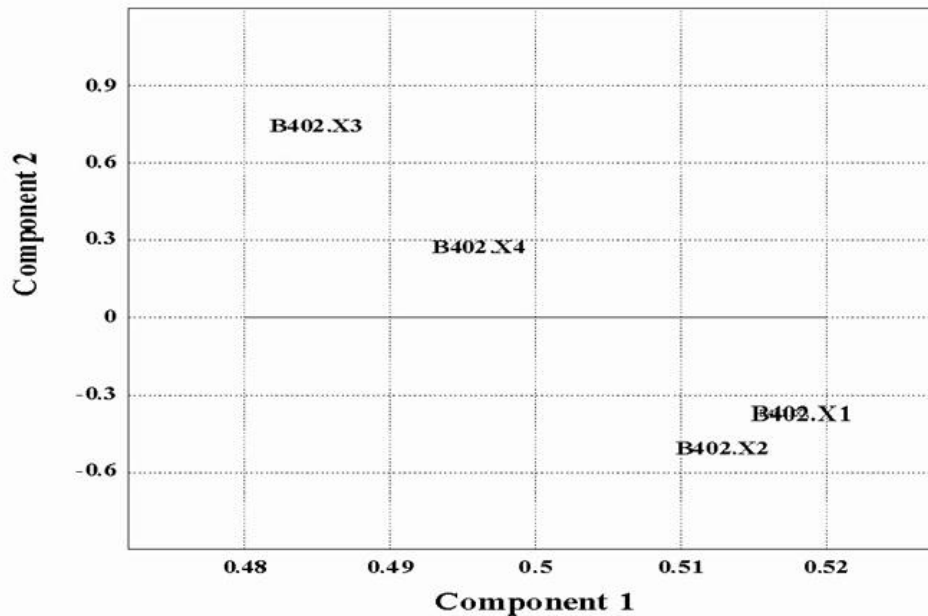
Proměnná	y_1	y_2	y_3	y_4
B402X1				
B402X2				
B402X3				
B402X4				

Diagram zobrazuje absolutní hodnoty komponentních vah a vysvětluje korelační strukturu, tj. která původní proměnná silně koreluje s hlavní komponentou.

1. hlavní komponenta silně koreluje se všemi proměnnými, zatímco 3. nebo 4. hlavní komponenta koreluje pouze s B402X3.

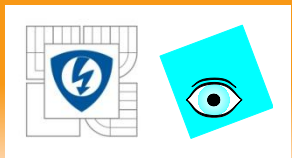


4. Vyšetření grafu komponentních vah:



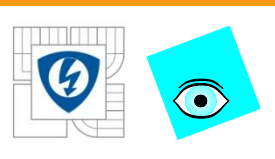
Obr. 4.8a Graf komponentních vah pro 20 objektů a 4 proměnné B402X1, B402X2, B402X3, B402X4 ze standard, dat

Proměnné B402X1 a B402X2 leží v diagramu blízko sebe a proto spolu korelují
Proměnné B402X3 a B402X4 jsou dál od sebe, proto méně korelují. Méně korelují rovněž se dvěma proměnnými B402X1 a B402X2, nacházejí se daleko od nich.

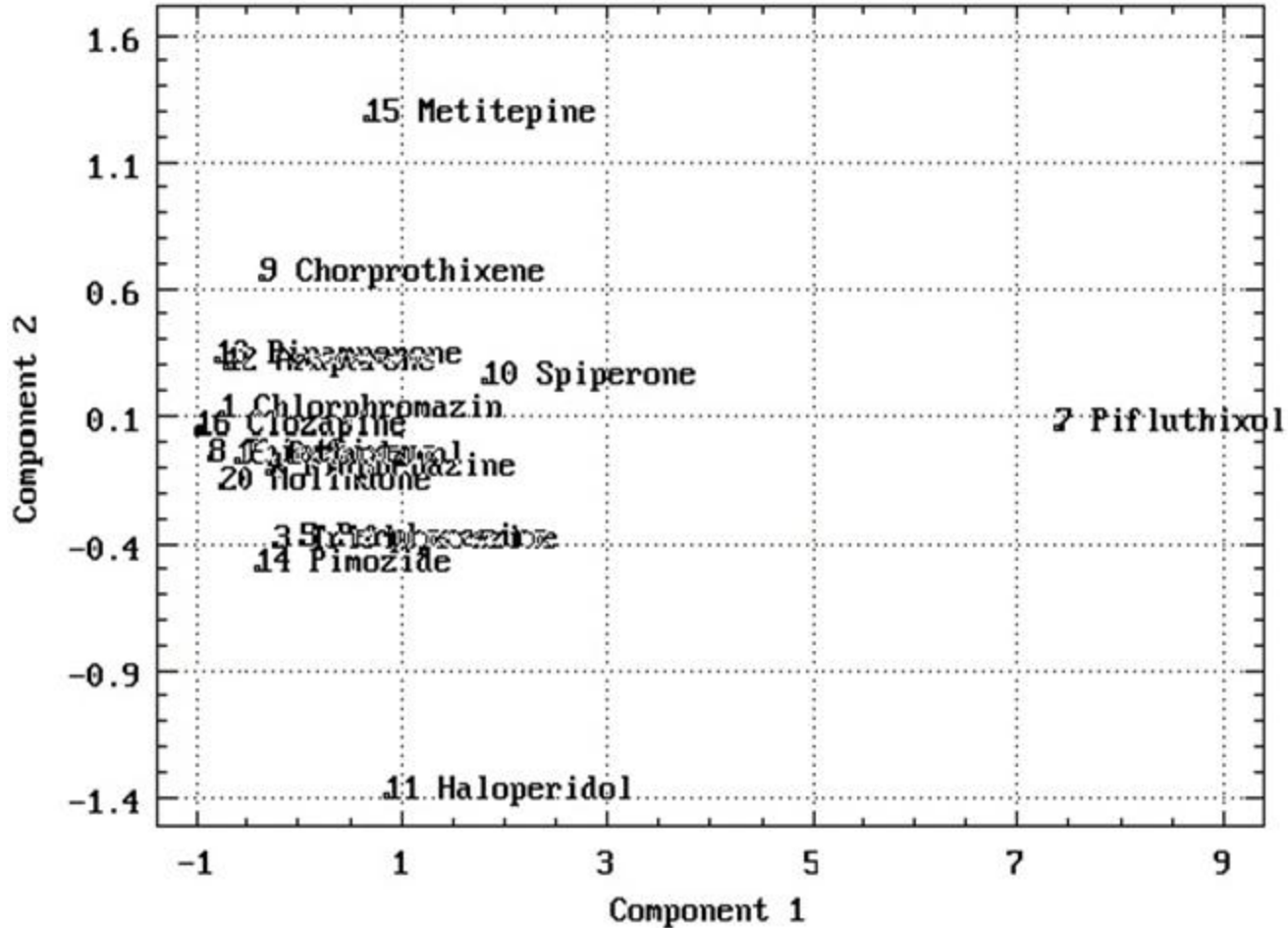


5. Vyšetření rozptylového diagramu komponentního skóre:

- Ukazuje vyšetřovanou strukturu objektů, tzn. shluky objektů, izolované objekty, odlehlé objekty, anomálie, atd.
- Objekty mohou být označeny textovým popisem nebo indexem.



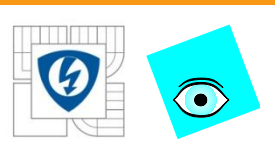
Plot of First Two Principal Components



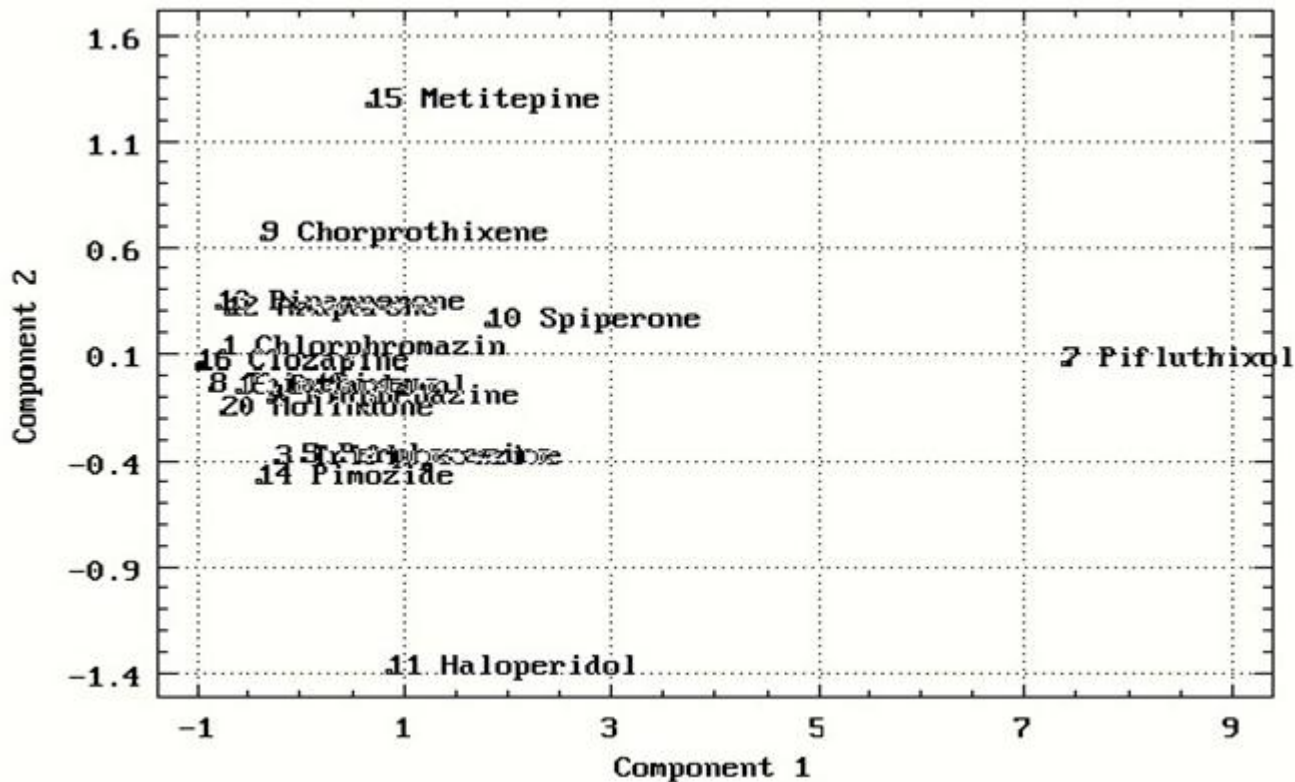
24.2.2010

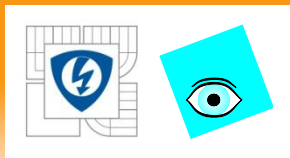
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





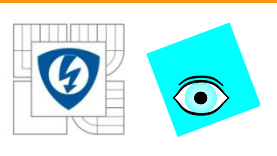
Rozptylový diagram komponentního skóre pro 20





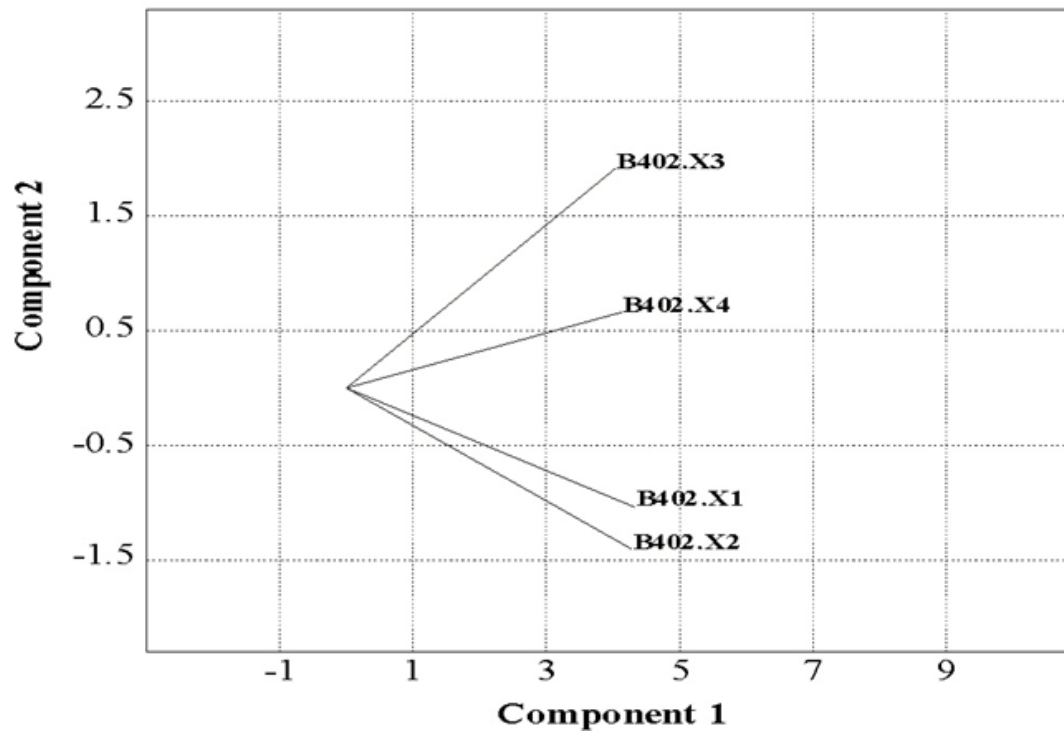
Rozptylový diagram komponentního skóre pro 20

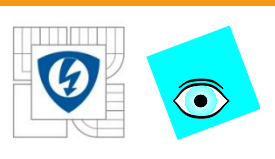
- Obr. 4.9a Rozptylový diagram komponentního skóre pro 20 objektů a 4 proměnné B4Ú2X1, B402X2, B402X3, B402X4 standardizovaných dat.
- ***Kromě tří objektů: 7, 11 a 15 leží zbývajících 17 objektů v jediném shluku.***
- ***Objekty 7, 11 a 15 jsou odlehlé body.***
- ***Nejvíce odlišný objekt od ostatních je 7, protože ten je odlehlý na hlavní komponentě 1 popisující většinu rozptylu.***
- ***První hlavní komponenta 1 popisuje rozdíl mezi Pifluthixolem a ostatními objekty.***
- ***Na druhé straně objekty 11a 15 jsou extrémy na druhé hlavní komponentě a udávají jej i směr.***
- ***Ostatní objekty tvoří v rovině prvních dvou hlavních komponent homogenní shluk***



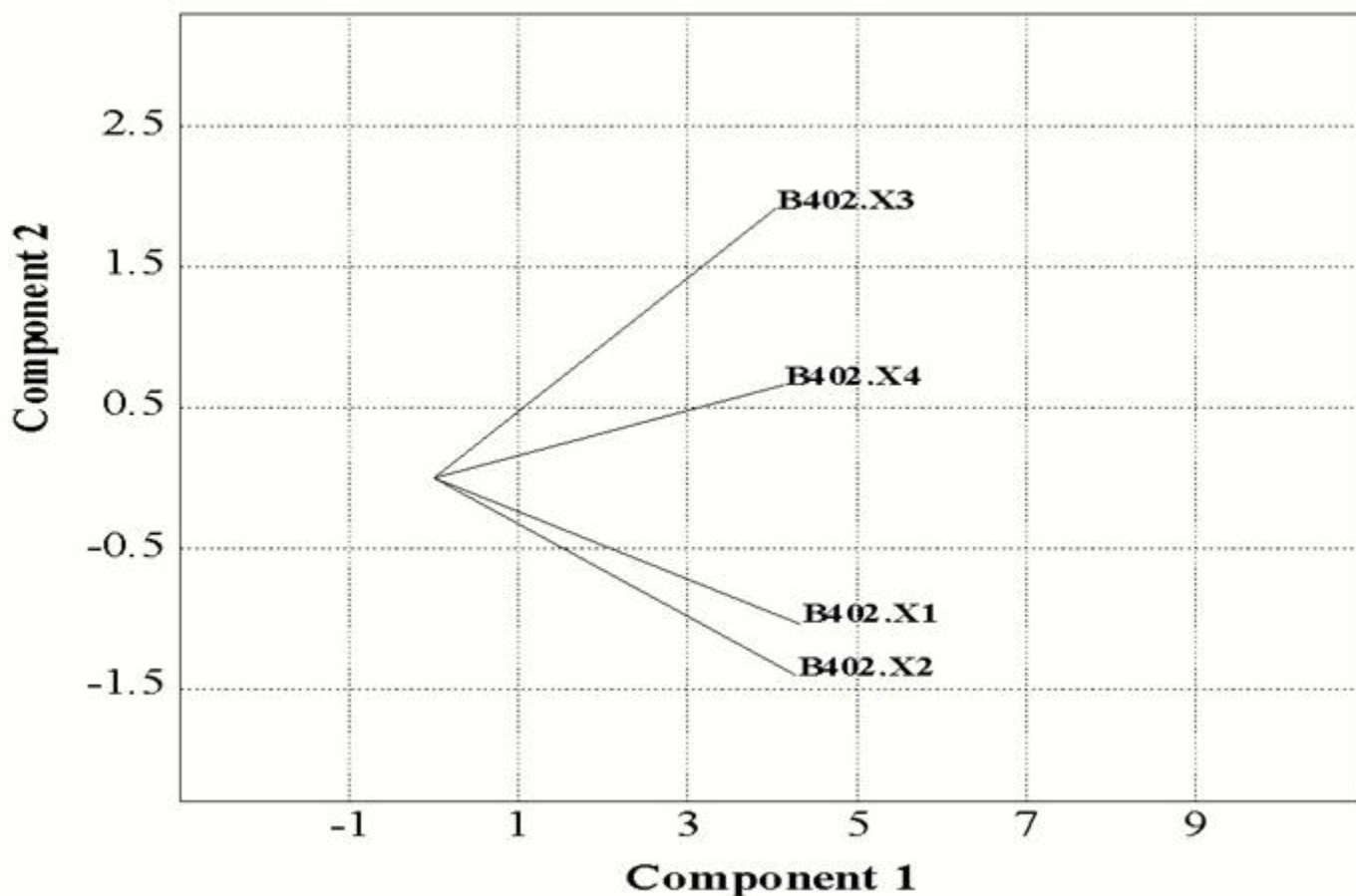
6. Vyšetření dvojného grafu: Biplot for First Two Principal Components

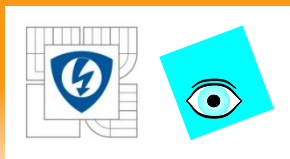
- Sledujeme interakci objektů a proměnných.
- Interakce poslouží interpretaci objektů.





Dvojný graf pro 20 objektů a 4 proměnné

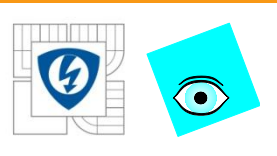




Dvojný graf pro 20 objektů a 4 proměnné

Obr. 4.10a Dvojný graf pro 20 objektů a 4 proměnné B402X1, B402X2, B402X3, B402X4 standardizovaných dat úlohy

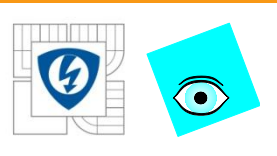
- *Uhel mezi dvěma průvodiči dvou proměnných je nepřímo úměrný velikosti korelace mezi těmito proměnnými.*
- *Mezi průvodiči B402X1 a B402X2 je malý úhel, což svědčí o silné korelaci.*
- *Uhly mezi těmito dvěma průvodiči a průvodiči B402X3 a B402X4 jsou pak větší, což ukazuje na jejich slabší korelaci.*



Diagnostika metody PCA

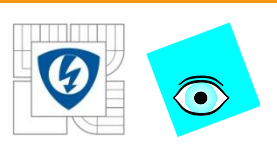
Maticový graf rozptylových diagramů znaků slouží k získání počáteční informace o datech, zda data potřebují škálování. V PCA postupně provádíme:

- 1. Vyšetření indexového grafu úpatí vlastních čísel- z hrany úpatí v tomto diagramu se určí vhodný počet hlavních komponent.*
- 2. Výpočet vlastních vektorů - vedle číselných hodnot se užívá i názorný čárový diagram hodnot vlastních vektorů, který přehledně informuje o relativním zastoupení původních znaků x_j , $j = 1, \dots, m$, v hlavních komponentách.*



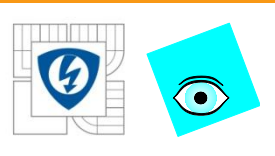
Diagnostika metody PCA

3. *Výpočet komponentních vah* - matice párových korelačních koeficientů obsahující korelace původních znaků s hlavními komponentami. Uživatel nyní vybere pouze prvních k hlavních komponent a vytvoří tak model PCA.
4. *Vyšetření grafu komponentních vah.*
5. *Vyšetření rozptylového diagramu komponentního skóre.*
6. *Vyšetření dvojného grafu.*



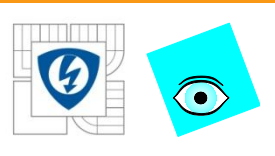
7. *Vyšetření reziduí* - rezidua objektů a rezidua proměnných by měla prokazovat dostatečnou těsnost proložení.

8. *Určení významných původních znaků* - je výhodné vyhledávat významné znaky, protože klasická metoda PCA umožňuje sice redukci počtu hlavních komponent, ale každá komponenta zůstává stále kombinací všech původních znaků.



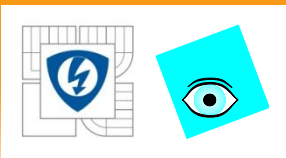
Řešení častých problémů v PCA

1. **Data neobsahují předpokládanou informaci.** Vysvětlení grafů a diagramů metody PCA nemá smysl, protože data neobsahují informaci popisující studovaný problém.
2. **Užito příliš málo hlavních komponent.** Nedostatečné vysvětlení dat vede ke ztrátě informace.
3. **Užito příliš mnoho hlavních komponent.** Příliš mnoho hlavních komponent může vyvolat vážnou chybu, protože šum je zahrnut do modelu.
4. **Neodstranění odlehlých objektů.** Odlehlé objekty mohou být důvodem hrubých chyb v datech a jsou vtahovány do dat.



Řešení častých problémů v PCA

5. **Odstraněné odlehlé objekty obsahovaly důležitou informaci.** Ztrátou určitých objektů se vytratila informace z dat a model je proto zkreslený.
6. **Komponentní skóre je nedostatečně analyzováno.** Nedostatečným rozbořem rozptylového diagramu byly zanedbány důležité rysy v datech.
7. **Vysvětlení komponentních vah se špatným počtem hlavních komponent.** Může vést k vážnému zkreslení interpretace.
8. **Přecenění standardních diagnostik.** Je třeba hodně rozvažovat a přemýšlet o úloze samé.
9. **Užití špatného předzpracování dat.** Chybná předúprava dat může vést ke zkresleným závěrům a nepochopení úloze.



CVIČENÍ V PROGRAMU STATISTICA

24.2.2010

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



3. PCA: V okně *Hlavní komponenty a klasifikační analýza* jsou podmínky výpočtu a dvě záložky: *Základní nastavení* a *Detaily*. Klikeneme na *Detaily* a naplníme *Proměnné*. Otevře se nové okno, ve kterém v záložce *Proměnné pro analýzu* zadáme prvních 7 znaků. V záložce *Doplňkové proměnné* zadáme další 3 doplňkové znaky. V záložce *Proměnná aktivních případů* zadáme rozlišovací znak GENDER o pohlaví respondenta. V poslední záložce *Grupovací proměnná* zadáme klasifikační znak GEO.REGION.

Hlavní komponenty a klasifikační analýza: Activities

Základní nastavení Detaily

Proměnné:

Proměnné pro analýzu:

Doplňkové proměnné:

Proměnná s aktiv. případy:

Grupovací proměnná (popis):

Kód aktivních případů:

Analýza podle

☒ Korelací

☐ Kovariancí

Rozptýl počítat

☒ jako SČ/(N-1)

☐ jako SČ/N

ChD vynechána

☒ Celé případy

☐ Nahradit průměrem

OK

Storno

Možnosti

Otevřít data

SELECT CASES E

Zvolte proměnné pro analýzu, dopl. proměnné, aktivní případy a grupovací prom.

1-WORK 2-TRANSPORT 3-HOUSEHOLD 4-CHILDREN 5-SHOPPING 6-PERSONAL CARE 7-MEAL 8-SLEEP 9-TV 10-LEISURE 11-GENDER 12-GEO.REGION	1-WORK 2-TRANSPORT 3-HOUSEHOLD 4-CHILDREN 5-SHOPPING 6-PERSONAL CARE 7-MEAL 8-SLEEP 9-TV 10-LEISURE 11-GENDER 12-GEO.REGION	1-WORK 2-TRANSPORT 3-HOUSEHOLD 4-CHILDREN 5-SHOPPING 6-PERSONAL CARE 7-MEAL 8-SLEEP 9-TV 10-LEISURE 11-GENDER 12-GEO.REGION	1-WORK 2-TRANSPORT 3-HOUSEHOLD 4-CHILDREN 5-SHOPPING 6-PERSONAL CARE 7-MEAL 8-SLEEP 9-TV 10-LEISURE 11-GENDER 12-GEO.REGION
--	--	--	--

DL. názvy

Detaily

Proměnné pro analýzu:

Doplňkové proměnné:

Prom. aktiv. případů:

Grupov. proměnná:

☐ Ukázat pouze odpovídající proměnné

OK

Storno

Volba "Ukázat pouze odpovídající proměnné" slouží k předzpracování seznamů proměnných a zobrazení kategorií a spojitéch proměnných. Pro více informací stiskněte F1.

100	95	7	57	85	150	807	115	330	MALE	WEST
70	307	30	88	85	118	845	87	383	FEMALE	WEST
7	567	87								
97	97	10								
22	529	69								
105	72									
34	262	14								
142	122	22								
106	338	42								
8	594	72								
133	134	22								
77	431	60								
148	68									
88	296	21								
140	120	15								
105	375	45								
10	710	55								
145	112	15								
52	576	59								
125	95									
89	318	23								

Výsledky hlavních komponent a klasifikační analýzy: Activities

Poč. aktiv. prom.: 7

Počet doplňkových prom.: 3

Poč. aktiv. příp.: 15

Počet doplňkových příp.: 8

Počet faktorů: 7

Kvalita reprezentace: 100,0 %

Základní výsledky

Proměnné

Případy

Popis. statistiky

Shrnutí popisných statistik

Korelační matice

Inverzní

Uložit korelační matici

Kovarianční matice

Inverzní

Uložit kovarianční matici

Do grafů zahrnout případy

☒ Všechny
 ☐ Pouze aktivní
 ☐ Pouze doplňkové

Krabicový graf

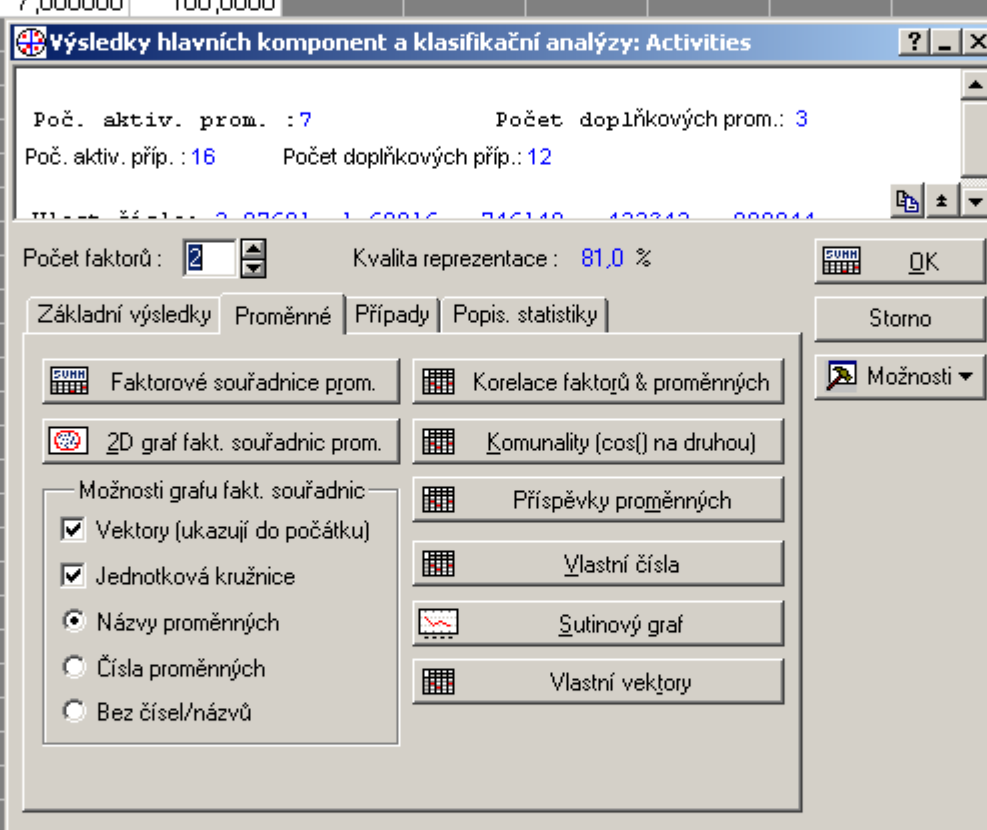
2D bodový graf

OK

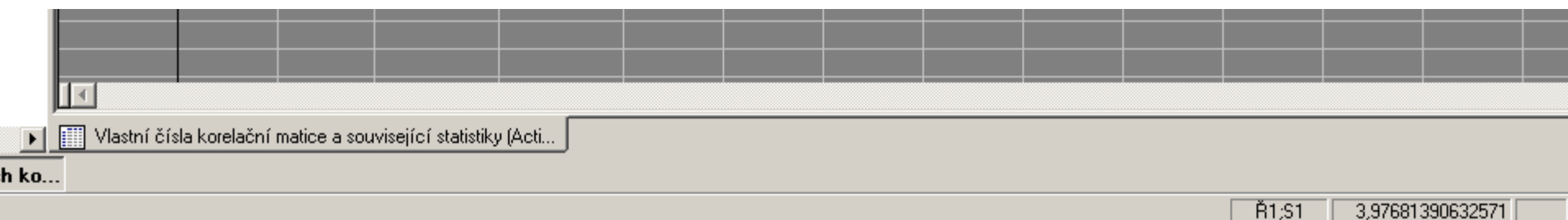
Storno

Možnosti

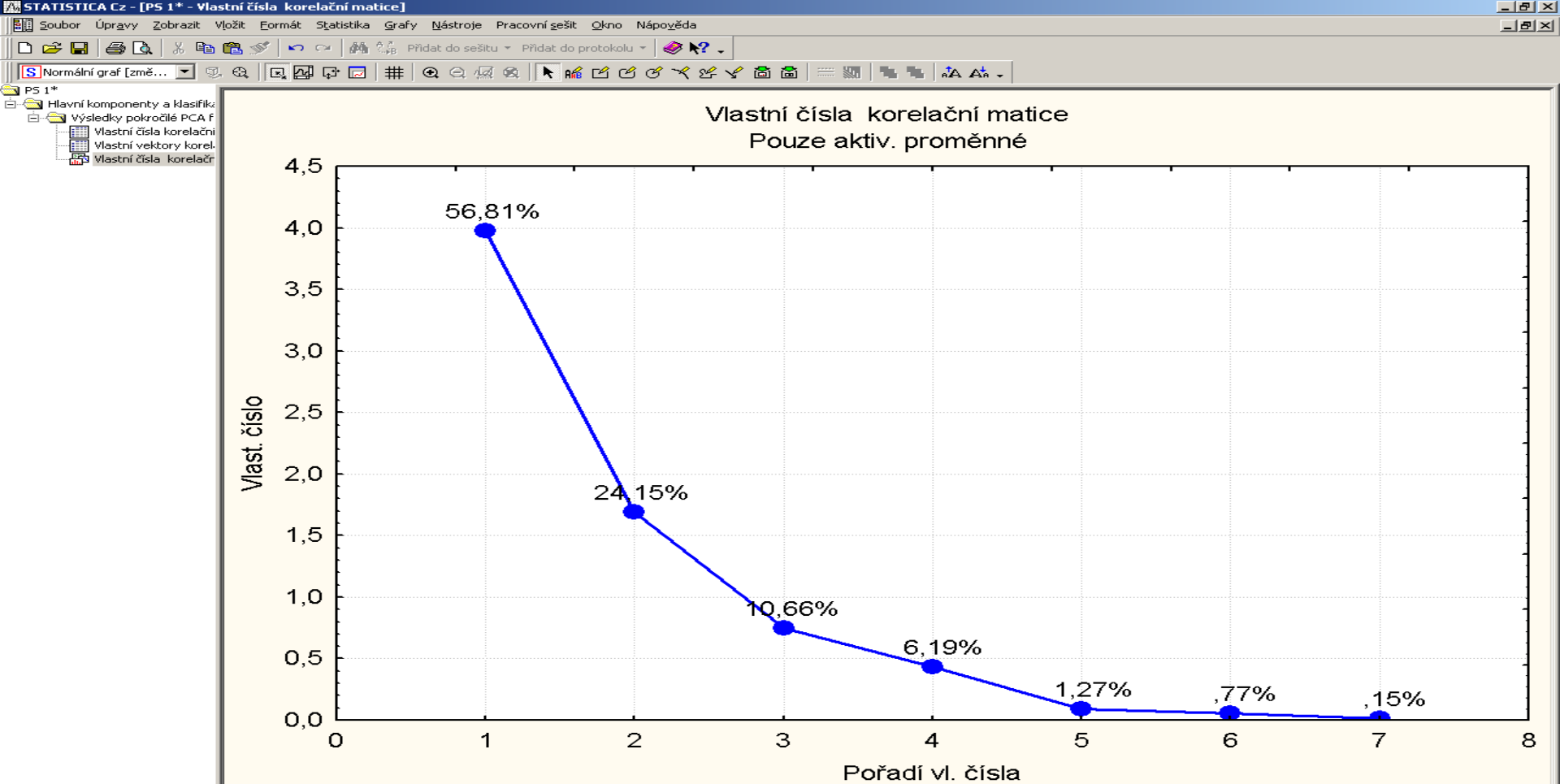
5. PCA: Otevře se okno *Výsledky hlavních komponent* se čtyřmi záložkami: *Základní výsledky*, *Proměnné*, *Případy* a *Popisné statistiky*. Zvolíme *Popisné statistiky*, kde bychom mohli provést obsáhlou exploratorní analýzu zdrojové matice formou korelační matice, kovarianční matice, krabicového grafu, histogramu, rankitového p -grafu, řady 2D-bodových rozptylových grafů, 3D-bodových rozptylových grafů a povrchových grafů. Lze také otevřít *Shrnutí popisných statistik*. Zde si vše nejprve vyzkoušíme.



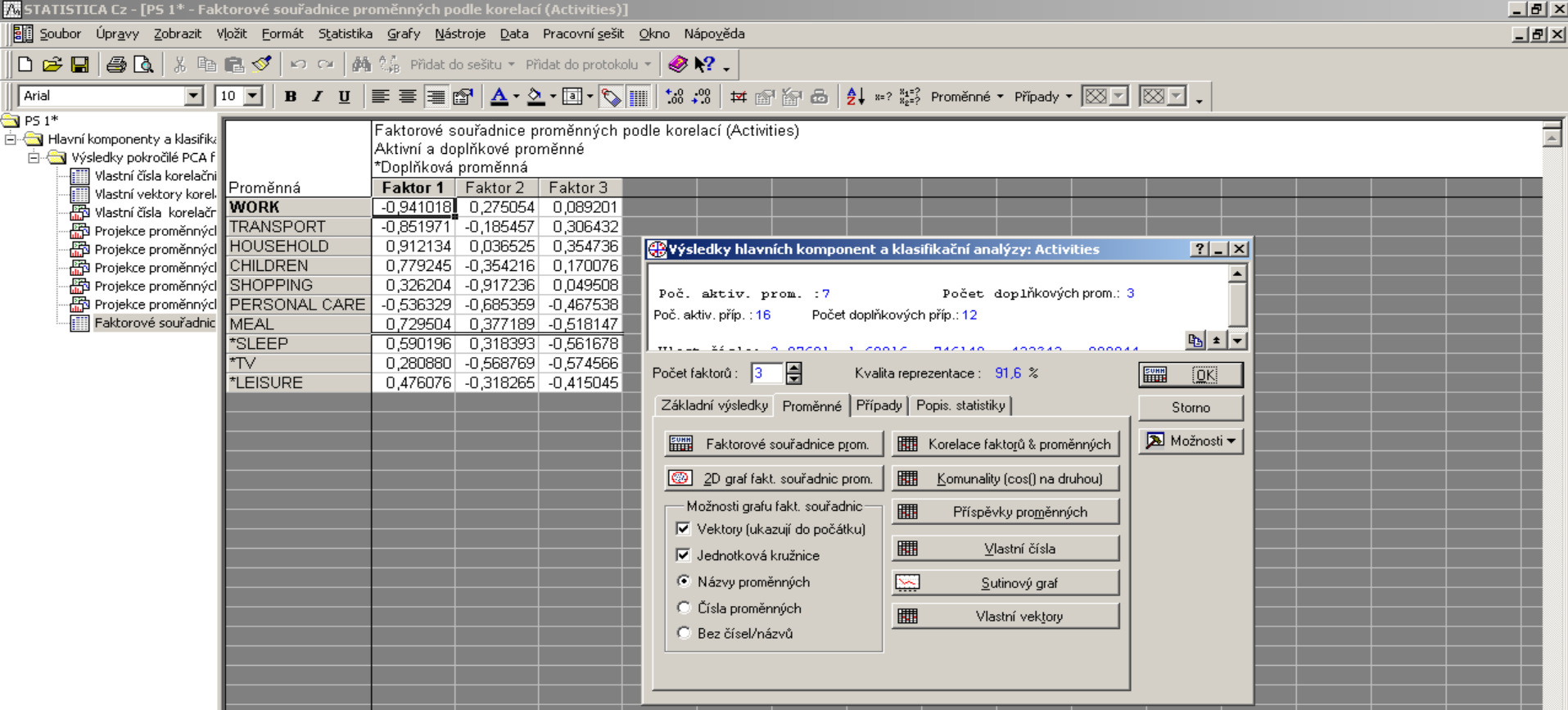
6. PCA: Dále otevřeme záložku *Proměnné*. Zadáme *Počet faktorů* rovný 7. V pravém sloupci klikneme na záložku *Vlastní čísla*.



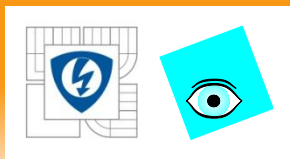
7. **PCA:** *Vlastní čísla:* tabelární podoba Cattelova indexového diagramu vlastních čísel ukazuje na zvolený počet hlavních komponent pomocí zlomu v diagramu. Největší vlastní číslo pokrývá 56.8% proměnlivosti v datech. Druhé největší vlastní číslo 24.15% proměnlivosti, což je dohromady s prvním 80.95% a značí, že první dvě hlavní komponenty dostatečně pokrývají proměnlivost v datech.



9. PCA: Velice důležitý je však *Cattelův indexový diagram vlastních čísel*: podle Kaiserova kritéria 1.0 pouze první dvě hlavní komponenty dosahují vlastních čísel větších než 1.0 čili hodnot zdrojové matice je rovna 2. Grafická analýza tohoto grafu však ukazuje na dva sloupce (=kolena) na křivce pro 2 a 3 hlavní komponenty.

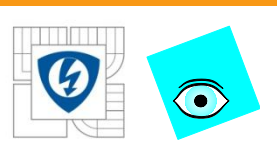


10. PCA: Volba tohoto grafu komponentních vah nabízí v levém sloupci v bloku *Možnosti grafu faktorových souřadnic* tři volby, kter nyní zaškrtneme: úsečky vektorů od počátku až ke znaku, vykreslení jednotkové kružnice korelace a zadání názvů znaků (proměnných).

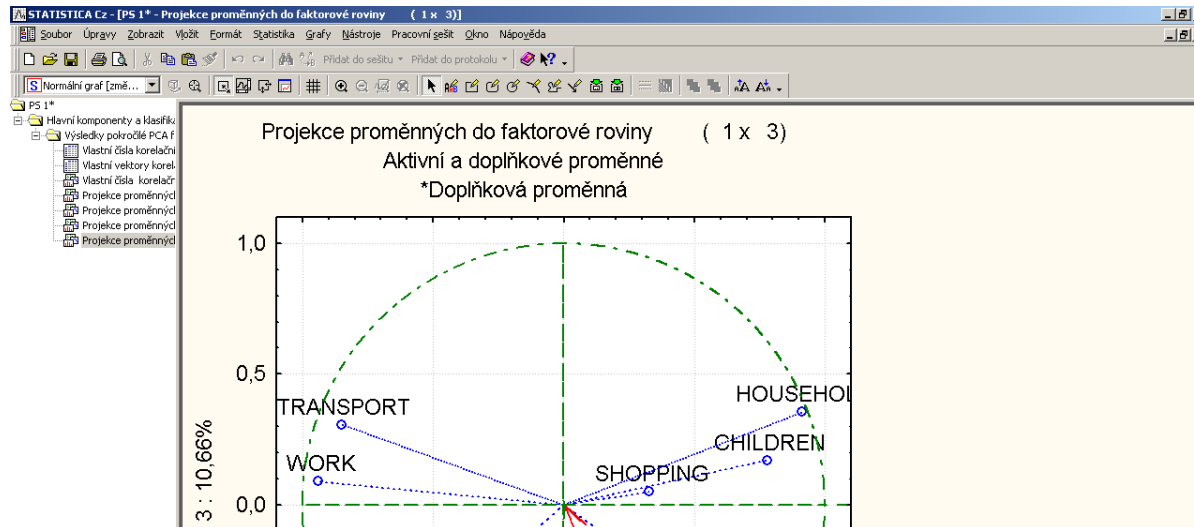


11. PCA: *Graf komponentních vah znaků 1 a 2:*

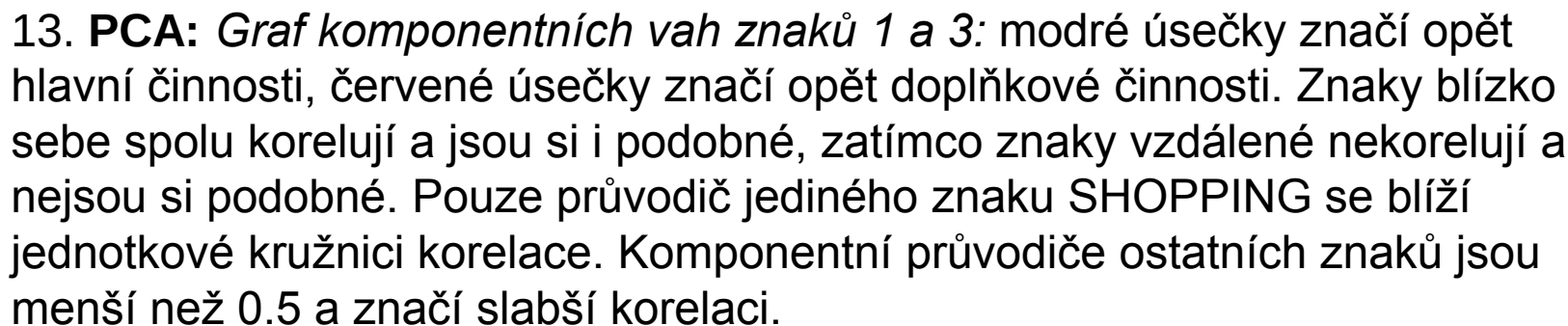
Souřadnice obou hlavních komponent jednotlivých znaků představují vlastně korelační koeficient mezi znakem a hlavní komponentou. Znaménko je velmi důležité: záporné je u negativně korelovaných znaků (čím více tím méně) s hlavní komponentou, a to jsou znaky WORK, TRANSPORT, PERSONAL CARE. Kladné je u i pozitivně korelovaných znaků (čím více, tím více) s 1. hlavní komponentou, a to jsou znaky HOUSEHOLD a CHILDREN. Druhá hlavní komponenta negativně koreluje se znaky každodenní činnosti SHOPPING a PERSONÁL CARE. Modré úsečky značí čas potřebný na hlavní činnosti, červené úsečky čas na doplňkové činnosti. Znaky blízko sebe značí, že znaky spolu korelují a znaky jsou si navíc i podobné. Znaky vzdálené nekorelují a nejsou si podobné. Protože tento graf je založen na korelacích, největší korelace mezi znakem a hlavní m komponentou může dosáhnout hodnoty 1.0. Proto také suma čtverců korelací mezi znakem a všemi faktory nepřekročí hodnotu 1.0. Čím více se blíží poloha znaku k jednotkové kružnici, tím lepší je jeho zobrazení v tomto souřadnicovém systému hlavních komponent. Komponentní průvodiče 2 znaků se dotýkají jednotkové korelační kružnice, a průvodiče 4 znaků dosahují téměř hodnoty korelace okolo 0.9.

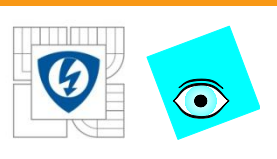


12. PCA: Graf komponentních vah znaků 1 a 3

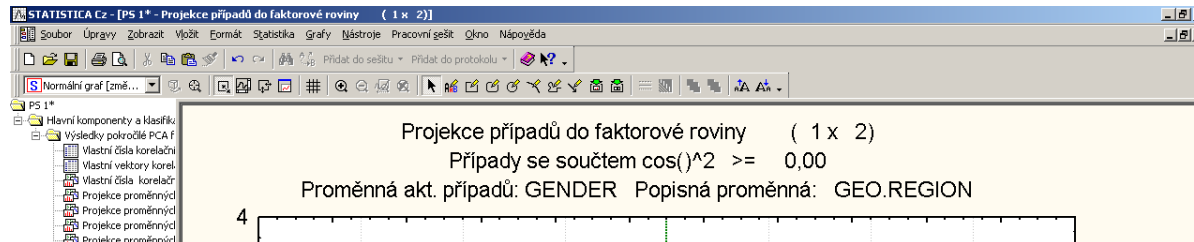


12. PCA: *Graf komponentních vah znaků 1 a 3*: modré úsečky značí opět čas potřebný na hlavní činnosti, červené úsečky opět značí čas na doplňkové činnosti, zatímco znaky blízko sebe spolu korelují a jsou si i podobné. Znaky vzdálené nekorelují a nejsou si podobné. I v tomto grafu jsou ještě komponentní průvodiče 3 znaků blízké jednotkové kružnici korelace a 7 menších s hodnotou korelace okolo 0.6.



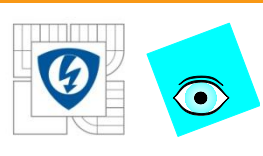


14. PCA: Rozptylový graf komponentního skóre komponenty 1 a 2:

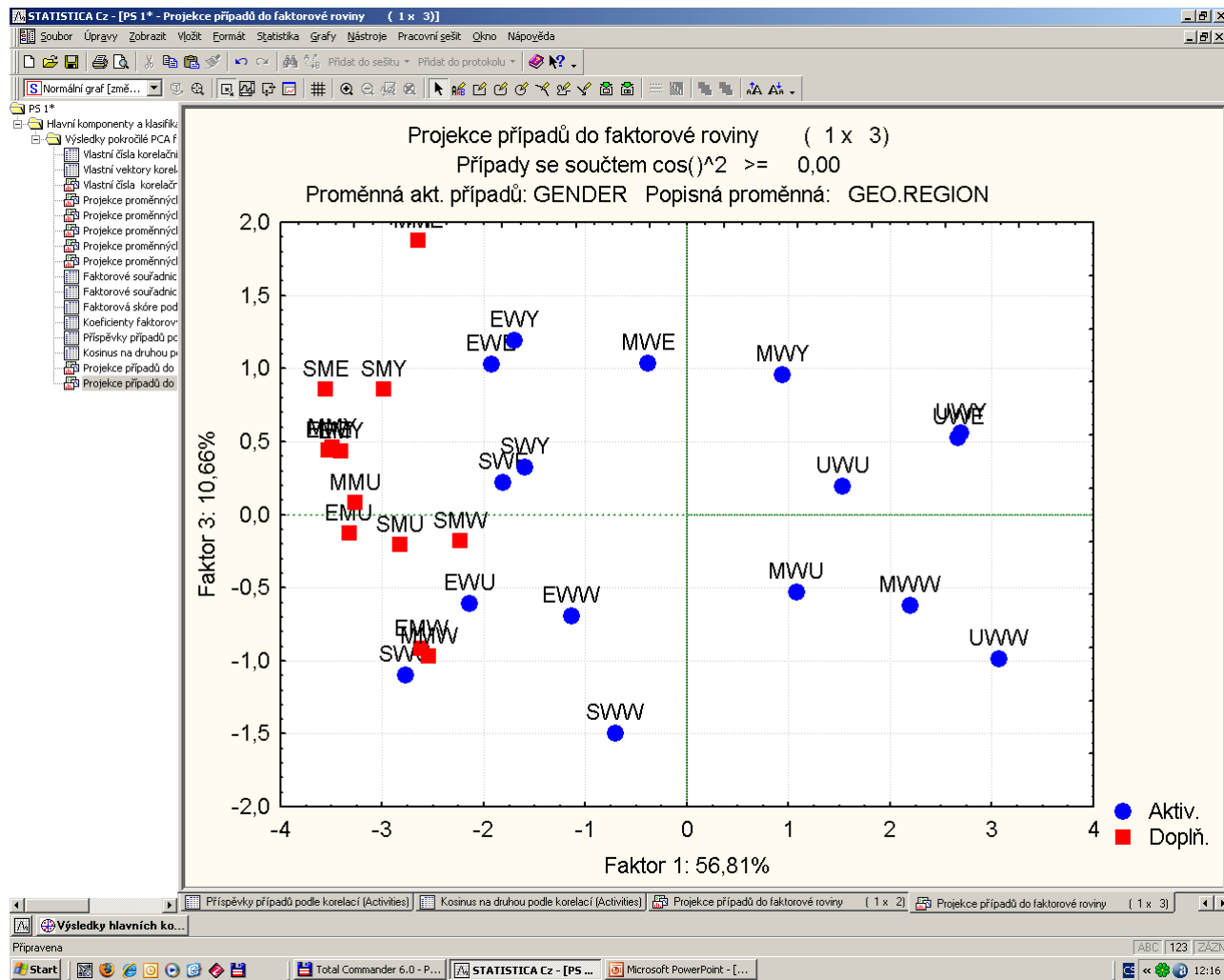


14. PCA: Rozptylový graf komponentního skóre komponenty 1 a 2: Graf ukazuje na komponentní skóre všech objektů (respondentů), a to aktivní činnosti (hlavně od žen, FEMALES) a pasivní činnosti (od mužů, MALES). Vznikají zde zajímavé shluky **aktivní** (modře) a **pasivní** (červeně) činnosti. Doplňková pasivní činnost, která je negativně korelována s 1. hlavní komponentou je umístěna hodně vlevo na x-ové ose 1. Hlavní komponenty. Červené čtverečky se týkají mužů (MALES) a jejich doplňkové činnosti s negativními hodnotami první hlavní komponenty. První hlavní komponenta je vysoce negativně korelována se znaky WORK a TRANSPORT na levém konci 1. komponenty a kladně korelována především s opakujícími se činnostmi moderního života jako jsou SHOPPING a PERSONÁL CARE ale také TV a LEISURE a kladně korelována se znaky MEAL a SLEEP..





15. PCA: Rozptylový graf komponentního skóre komponenty 1 a 3.



24.2.2010

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



STATISTICA Cz - [PS 1* - Faktorové souřadnice případů podle korelací (Activities)]

Soubor Úpravy Zobrazit Vložit Formát Statistika Grafy Nástroje Data Pracovní sešit Okno nápověda

Arial 10 B I U

PS 1*

Hlavní komponenty a klasifikace

Výsledky pokročilé PCA

Vlastní čísla korelační

Vlastní vektory korelací

Vlastní čísla korelační

Projekce proměnných

Projekce proměnných

Projekce proměnných

Projekce proměnných

Projekce proměnných

Faktorové souřadnice

Faktorové souřadnice

Faktorové souřadnice případů podle korelací (Activities)

Proměnná akt. případů: GENDER Popisná proměnná: GEO.REGION

Kód aktivních případů: FEMALE ; Hodn. dopl. příp. zvýrazněny

Případ	Faktor 1	Faktor 2	Faktor 3	GENDER	GEO.REGION
EMU	-3,32600	0,42156	-0,12408	MALE	WEST
EWU	-2,13833	-1,29595	-0,60855	FEMALE	WEST
UWU	1,53339	-2,57182	0,19267	FEMALE	WEST
MMU	-3,26751	0,74257	0,08472	MALE	WEST
MWU	1,07533	-1,91722	-0,53067	FEMALE	WEST
SMU	-2,82185	-1,10164	-0,20476	MALE	WEST
SWU	-2,77101	-1,64452	-1,09756	FEMALE	WEST
EMW	-2,61531	3,04182	-0,91696	MALE	WEST
EWV	-1,13678	1,77511	-0,69447	FEMALE	WEST
UWV	3,06790	0,78252	-0,98745	FEMALE	WEST
MMW	-2,54362	3,17956	-0,97200	MALE	WEST
MVW	2,19296	1,49985	-0,62079	FEMALE	WEST
SMW	-2,23206	2,76908	-0,18043	MALE	WEST
SVW	-0,70614	1,60208	-1,49698	FEMALE	WEST
EME	-3,52718	1,38023	0,43935	MALE	EAST
EWE	-1,92953	0,43054	1,03016	FEMALE	EAST
UWE	2,67143	-0,51161	0,52621	FEMALE	EAST
MME	-2,64299	3,15648	1,87398	MALE	EAST

17. PCA: Faktorové souřadnice objektů podle korelací obsahuje Proměnná aktivních případů, kam se dosadil znak GENDER, Popisná proměnná kam se dosadil znak GEO.REGION a Kód aktivních případů má dosazeno heslo FEMALE. Hodnoty doplňkových případů jsou zde červeně zvýrazněny.

STATISTICA Cz - [PS 1* - Faktorová skóre podle korelací (Activities)]

Soubor Úpravy Zobrazit Vložit Formát Statistika Grafy Nástroje Data Pracovní sešit Okno nápověda

Arial 10 B I U

PS 1*

- Hlavní komponenty a klasifikace
 - Výsledky pokročilé PCA
 - Vlastní čísla korelační
 - Vlastní vektory korelací
 - Vlastní čísla korelační
 - Projekce proměnných
 - Projekce proměnných
 - Projekce proměnných
 - Projekce proměnných
 - Faktorové souřadnice
 - Faktorové souřadnice
 - Faktorová skóre podle korelací

Faktorová skóre podle korelací (Activities)
 Proměnná akt. případů: GENDER Popisná proměnná: GEO.REGION
 Kód aktivních případů: FEMALE ; Hodn. dopl. příp. zvýrazněny

Případ	Faktor 1	Faktor 2	Faktor 3	GENDER	GEO.REGION
EMU	-1,66784	0,32426	-0,14365	MALE	WEST
EWU	-1,07228	-0,99684	-0,70451	FEMALE	WEST
UWU	0,76893	-1,97823	0,22305	FEMALE	WEST
MMU	-1,63851	0,57118	0,09808	MALE	WEST
MWU	0,53923	-1,47471	-0,61434	FEMALE	WEST
SMU	-1,41503	-0,84737	-0,23704	MALE	WEST
SWU	-1,38954	-1,26495	-1,27063	FEMALE	WEST
EMW	-1,31146	2,33975	-1,06155	MALE	WEST
EWV	-0,57004	1,36540	-0,80398	FEMALE	WEST
UWV	1,53842	0,60191	-1,14315	FEMALE	WEST
MMW	-1,27551	2,44570	-1,12527	MALE	WEST
MWV	1,09967	1,15367	-0,71868	FEMALE	WEST
SMW	-1,11928	2,12996	-0,20888	MALE	WEST
SWV	-0,35410	1,23231	-1,73303	FEMALE	WEST
EME	-1,76872	1,06166	0,50863	MALE	EAST
EWE	-0,96757	0,33117	1,19260	FEMALE	EAST
UWE	1,33960	-0,39353	0,60919	FEMALE	EAST
MME	-1,32534	2,42795	2,16948	MALE	EAST
MVE	-0,19184	0,24535	1,19988	FEMALE	EAST
SME	-1,78217	0,59822	0,99397	MALE	EAST
SWE	-0,90687	-0,23578	0,25174	FEMALE	EAST
EMY	-1,70945	1,09434	0,49989	MALE	EAST
EWY	-0,85378	0,76493	1,37883	FEMALE	EAST
UWY	1,35099	-0,07795	0,64341	FEMALE	EAST
MMY	-1,75101	1,07823	0,53013	MALE	EAST
MWY	0,47187	0,41801	1,11077	FEMALE	EAST
SMY	-1,49810	0,16940	0,99747	MALE	EAST
SWY	-0,80270	0,30924	0,37884	FEMALE	EAST

18. PCA: Faktorová skóre objektů seřazených podle korelací.

STATISTICA Cz - [PS 1* - Koeficienty faktorových skóre podle korelací (Activities)]

Soubor Úpravy Zobrazit Vložit Formát Statistika Grafy Nástroje Data Pracovní sešit Okno Nápověda

Arial 10 B I U

PS 1*

Hlavní komponenty a klasifikace

Výsledky pokročilé PCA

Vlastní čísla korelační

Vlastní vektory korelační

Vlastní čísla korelační

Projekce proměnných

Projekce proměnných

Projekce proměnných

Projekce proměnných

Projekce proměnných

Faktorové souřadnice

Faktorové souřadnice

Faktorové skóre podle

Koeficienty faktorových

Proměnná	Faktor 1	Faktor 2	Faktor 3
WORK	-0,236626	0,162738	0,119551
TRANSPORT	-0,214234	-0,109727	0,410689
HOUSEHOLD	0,229363	0,021610	0,475428
CHILDREN	0,195947	-0,209575	0,227942
SHOPPING	0,082026	-0,542691	0,066352
PERSONAL CARE	-0,134864	-0,405499	-0,626609
MEAL	0,183439	0,223167	-0,694437

19. PCA: *Koeficienty faktorových skóre znaků seřazených podle korelací.*

STATISTICA Cz - [PS 1* - Příspěvky případů podle korelací (Activities)]

Soubor Úpravy Zobrazit Vložit Formát Statistika Grafy Nástroje Data Pracovní sešit Okno Nápověda

Arial 10 B I U

PS 1*

Hlavní komponenty a klasifikace

Výsledky pokročilé PCA

Vlastní čísla korelační

Vlastní vektory korelační

Vlastní čísla korelační

Projekce proměnných

Projekce proměnných

Projekce proměnných

Projekce proměnných

Projekce proměnných

Faktorové souřadnice

Faktorové souřadnice

Faktorové skóre pod

Koeficienty faktorové

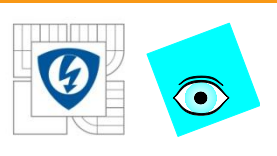
Příspěvky případů podle korelací (Activities)

Proměnná akt. případů: GENDER Popisná proměnná: GEO.REGION

Případ	Faktor 1	Faktor 2	Faktor 3	GEO.REGION
EWU	7,66517	6,62456	3,30887	WEST
UWU	3,94166	26,08921	0,33167	WEST
MWU	1,93847	14,49854	2,51612	WEST
SWU	12,87213	10,66738	10,76329	WEST
EWV	2,16632	12,42886	4,30919	WEST
UWV	15,77816	2,41531	8,71194	WEST
MWV	8,06183	8,87307	3,44330	WEST
SWV	0,83589	10,12394	20,02251	WEST
EWE	6,24131	0,73114	9,48194	EAST
UWE	11,96360	1,03244	2,47406	EAST
MWE	0,24534	0,40132	9,59814	EAST
SWE	5,48272	0,37061	0,42250	EAST
EWY	4,85960	3,90076	12,67440	EAST
UWY	12,16783	0,04051	2,75983	EAST
MWY	1,48443	1,16486	8,22547	EAST
SWY	4,29554	0,63751	0,95678	EAST

20. PCA: Příspěvky objektů seřazených podle korelací.

Připad	Faktor 1	Faktor 2	Faktor 3	GENDER	GEO.REGION
EMU	0,752436	0,012088	0,001047	MALE	WEST
EWU	0,683246	0,250961	0,055338	FEMALE	WEST
UWU	0,212589	0,598020	0,003356	FEMALE	WEST
MMU	0,709125	0,036624	0,000477	MALE	WEST
MWU	0,200835	0,638409	0,048910	FEMALE	WEST
SMU	0,657691	0,100238	0,003463	MALE	WEST
SWU	0,650255	0,229026	0,102015	FEMALE	WEST
EMW	0,365504	0,494439	0,044931	MALE	WEST
EWV	0,249937	0,609442	0,093280	FEMALE	WEST
UWV	0,821704	0,053460	0,085125	FEMALE	WEST
MMV	0,335141	0,523670	0,048939	MALE	WEST
MVV	0,623147	0,291490	0,049936	FEMALE	WEST
SMV	0,289582	0,445688	0,001892	MALE	WEST
SVV	0,088667	0,456411	0,398490	FEMALE	WEST
EME	0,765868	0,117273	0,011883	MALE	EAST
EWV	0,723131	0,036003	0,206121	FEMALE	EAST
UVE	0,812282	0,029792	0,031517	FEMALE	EAST
MME	0,273680	0,390355	0,137589	MALE	EAST
MVE	0,096166	0,066855	0,705867	FEMALE	EAST
SME	0,736173	0,035253	0,042965	MALE	EAST
SVE	0,850019	0,024420	0,012290	FEMALE	EAST
EMY	0,747432	0,130184	0,011992	MALE	EAST
EWY	0,507356	0,173083	0,248270	FEMALE	EAST
UWY	0,772474	0,001093	0,032873	FEMALE	EAST
MMY	0,744642	0,120001	0,012806	MALE	EAST
MVY	0,396336	0,132181	0,412050	FEMALE	EAST
SMY	0,731283	0,003974	0,060826	MALE	EAST
SVY	0,835235	0,052683	0,034905	FEMALE	EAST



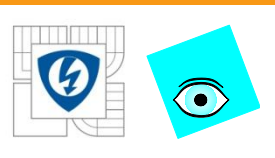
PŘÍKLAD 4.1 Interpretace grafů PCA demografického souboru Lidé

Cílem je interpretace grafu komponentního skóre a komponentních vah pro soubor Lidé

Data: použita data Lidé

Řešení: Na základě maticového grafu korelace byla získána počáteční informace o korelaci znaků. Grafy hlavních komponent zdrojové matice dat Lidé se týkají 32 osob (objektů) a 12 jejich vlastností (znaků).

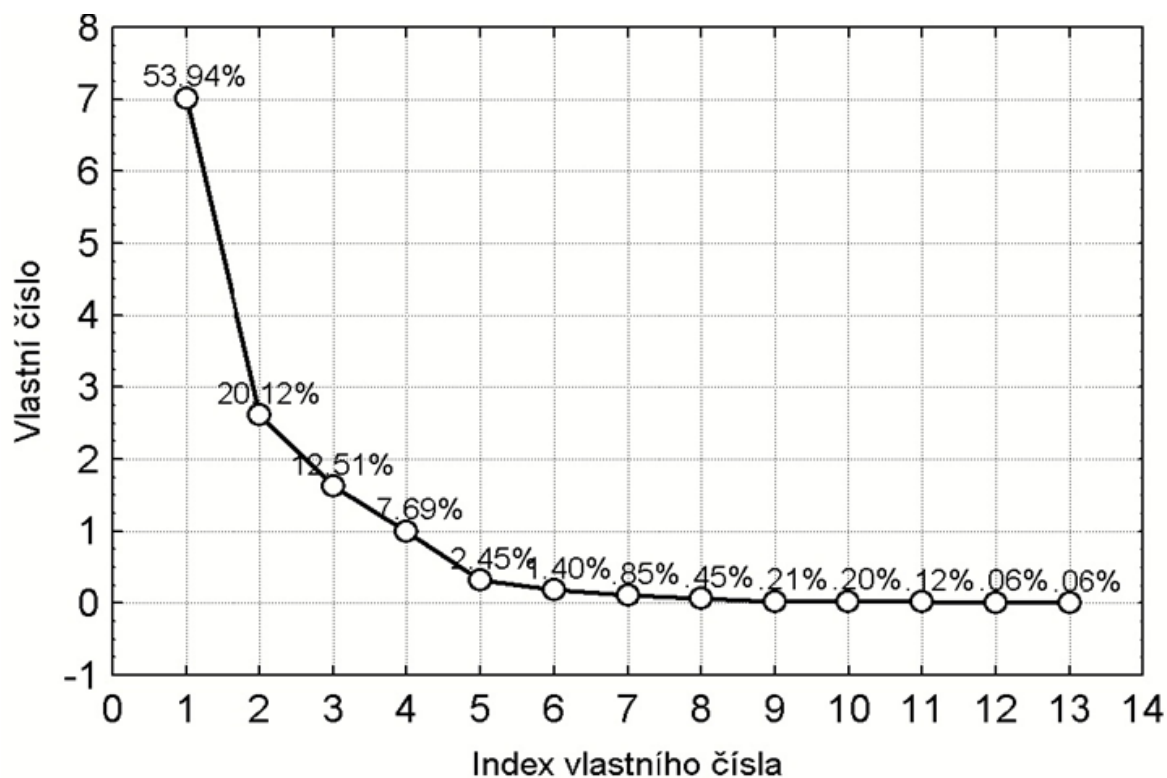
1. Počet potřebných hlavních komponent: indexový graf úpatí objasněného rozptylu ukazuje, že první dvě hlavní komponenty objasňují okolo 74% z celkového rozptylu.

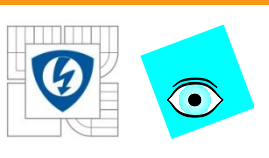


1. Počet potřebných hlavních komponent:

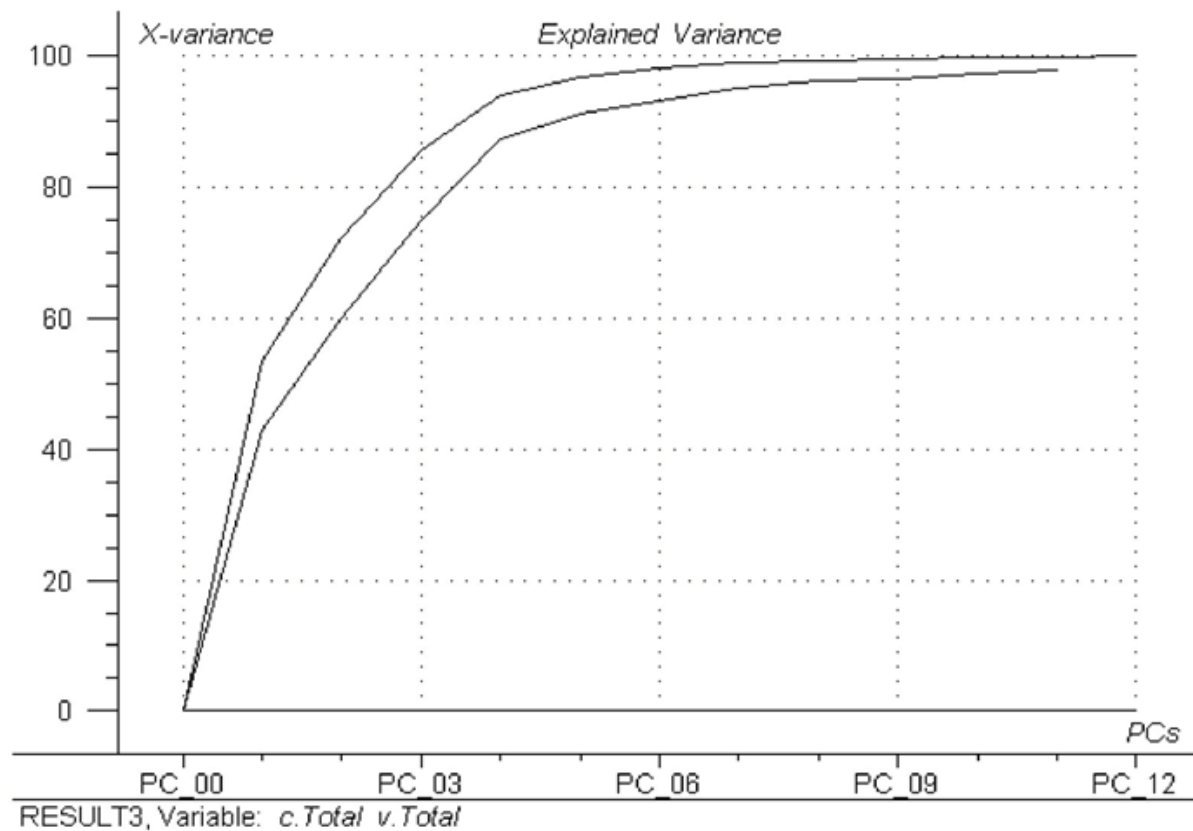
Indexový graf úpatí objasněného rozptylu ukazuje, že první dvě hlavní komponenty objasňují okolo 74% z celkového rozptylu.

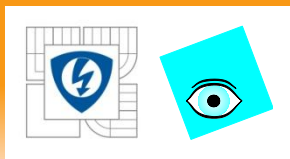
Obr. 4.2a Cattelův indexový graf úpatí vlastních čísel
Scree Plot zdrojové matice dat *Lidé*
(STATISTICA).



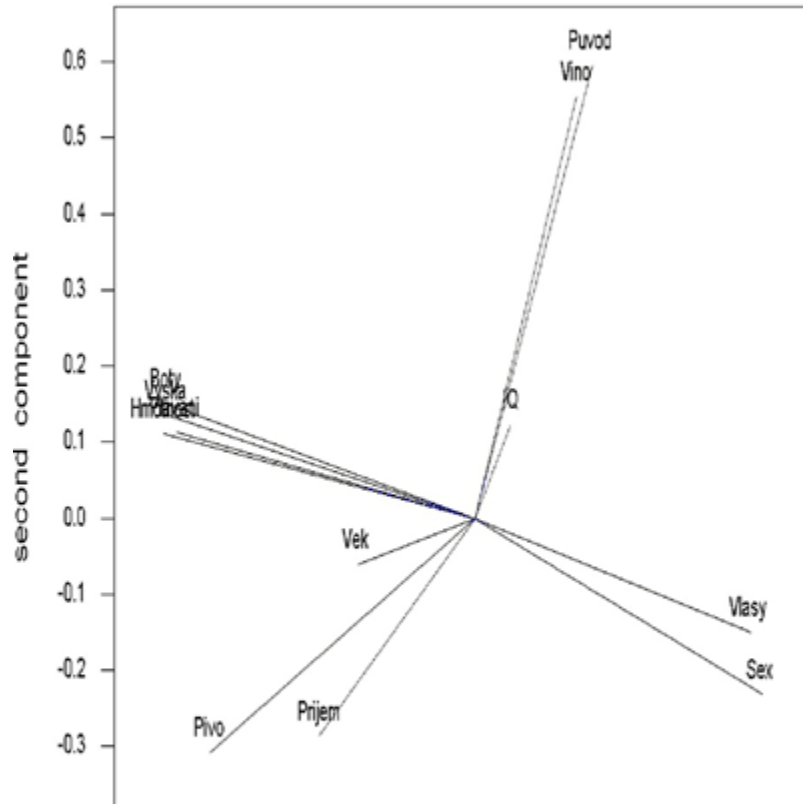


Obr. 4.2b Indexový graf úpatí
objasněného rozptylu
zdrojové matice dat *Lidé*.
Horní křivka značí značí
závislost *validačního rozptylu*
na počtu hlavních komponent
PC, zatímco dolní křivka
závislost *kalibračního rozptylu*
(UNSCRAMBLER).



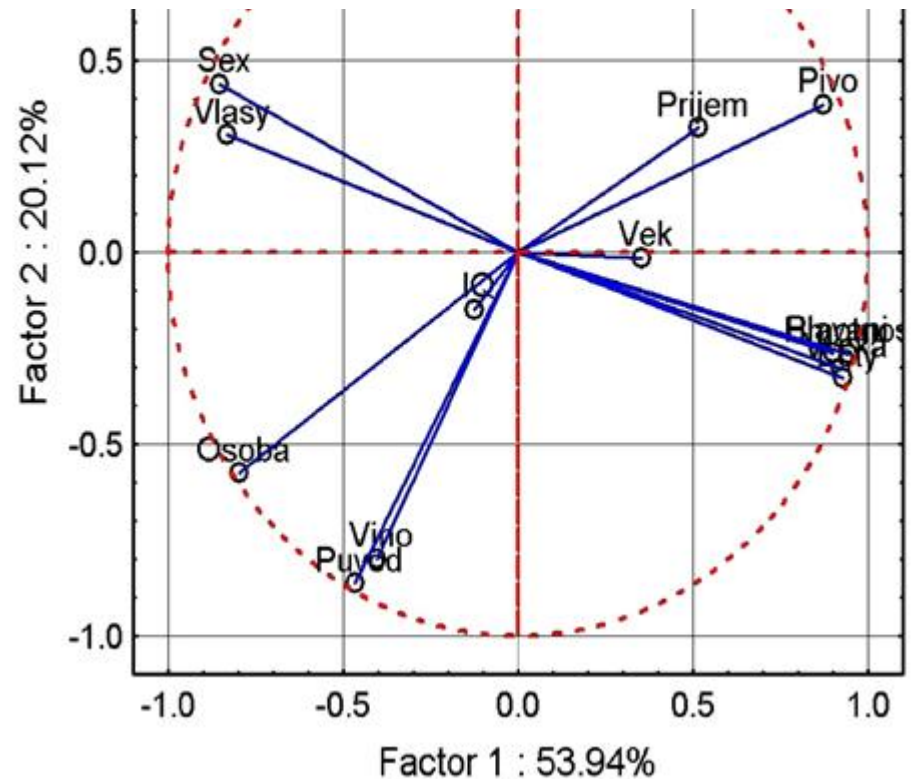


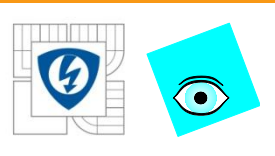
2. Graf komponentních vah: Graf komponentních vah vysvětlí, proč dochází v rozptylovém diagramu komponentního skóre ke vzniku čtyř shluků



Obr. 4.3b Graf komponentních vah 1 a 2 při popisu znaků jmény dat *Lidé* (STATISTICA).

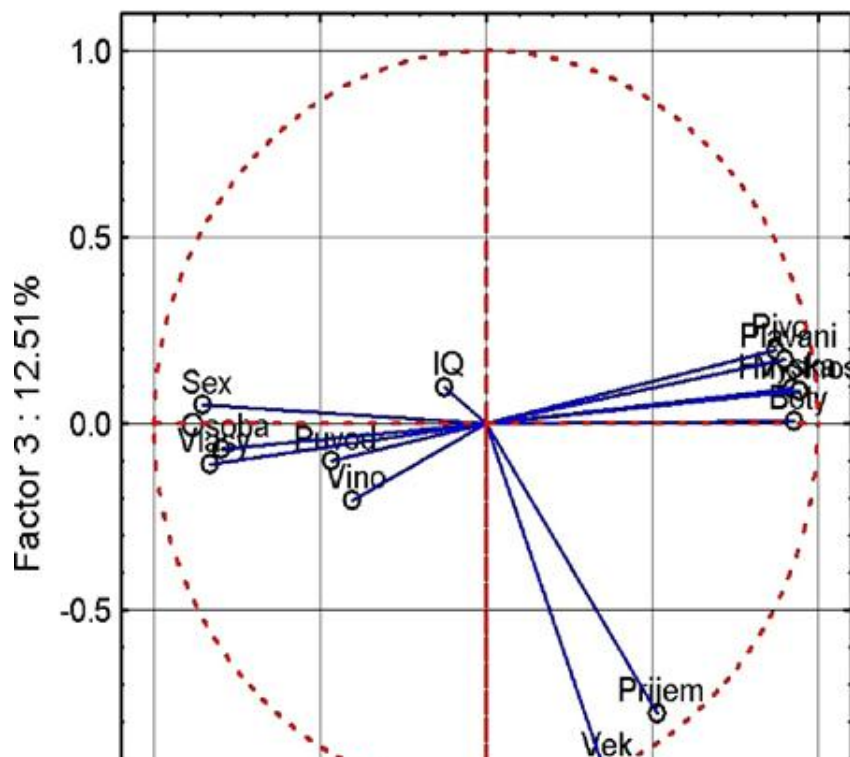
Obr. 4.3b Graf komponentních vah 1 a 2 při popisu znaků jmény dat *Lidé* (STATISTICA).



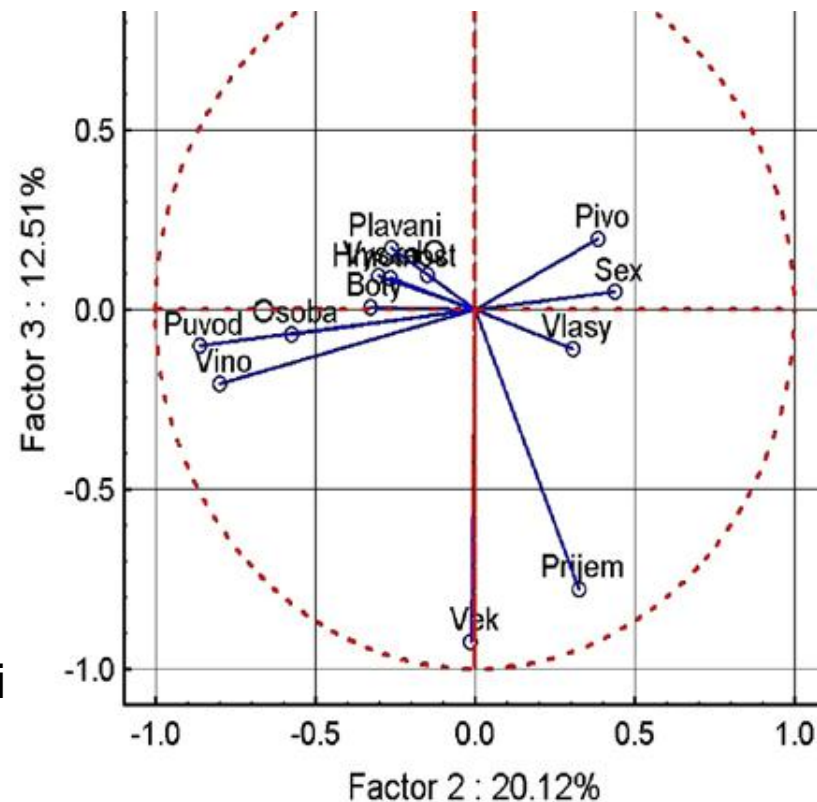


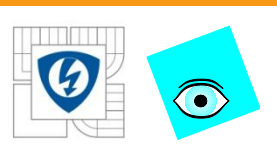
Obr. 4.3c a 4.3d

Obr. 4.3d Graf komponentních vah 2 a 3 při popisu znaků jmény dat *Lidé* (STATISTICA).

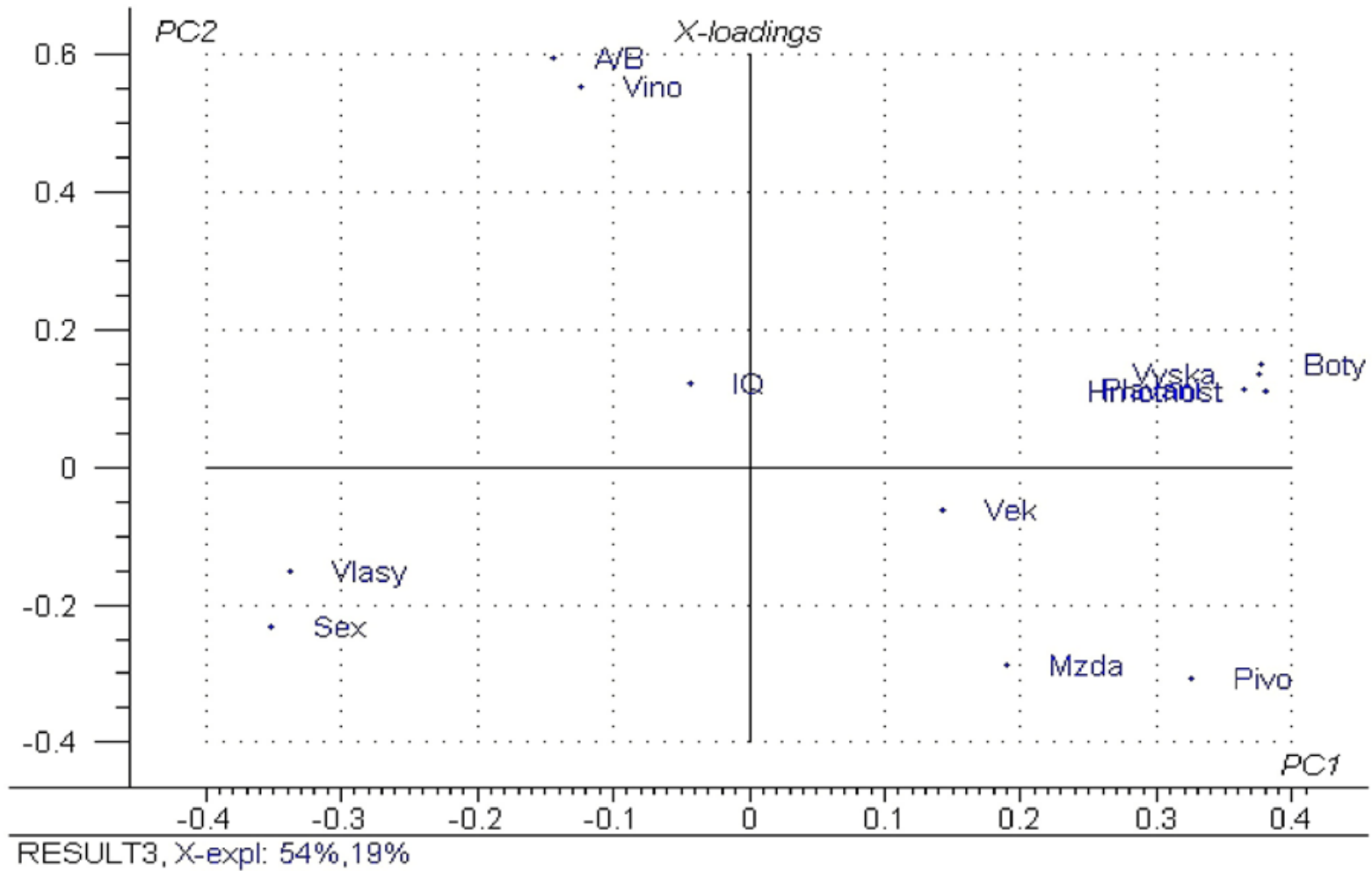


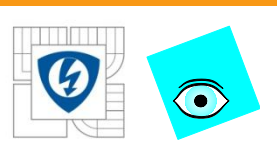
Obr. 4.3c Graf komponentních vah 1 a 3 při popisu znaků jmény dat *Lidé* (STATISTICA).





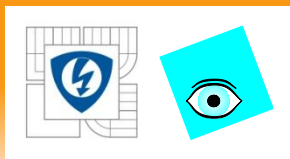
Obr. 4.3e Graf komponentních vah 1 a 2 při popisu znaků jmény dat *Lidé* (UNSCRAMBLER).



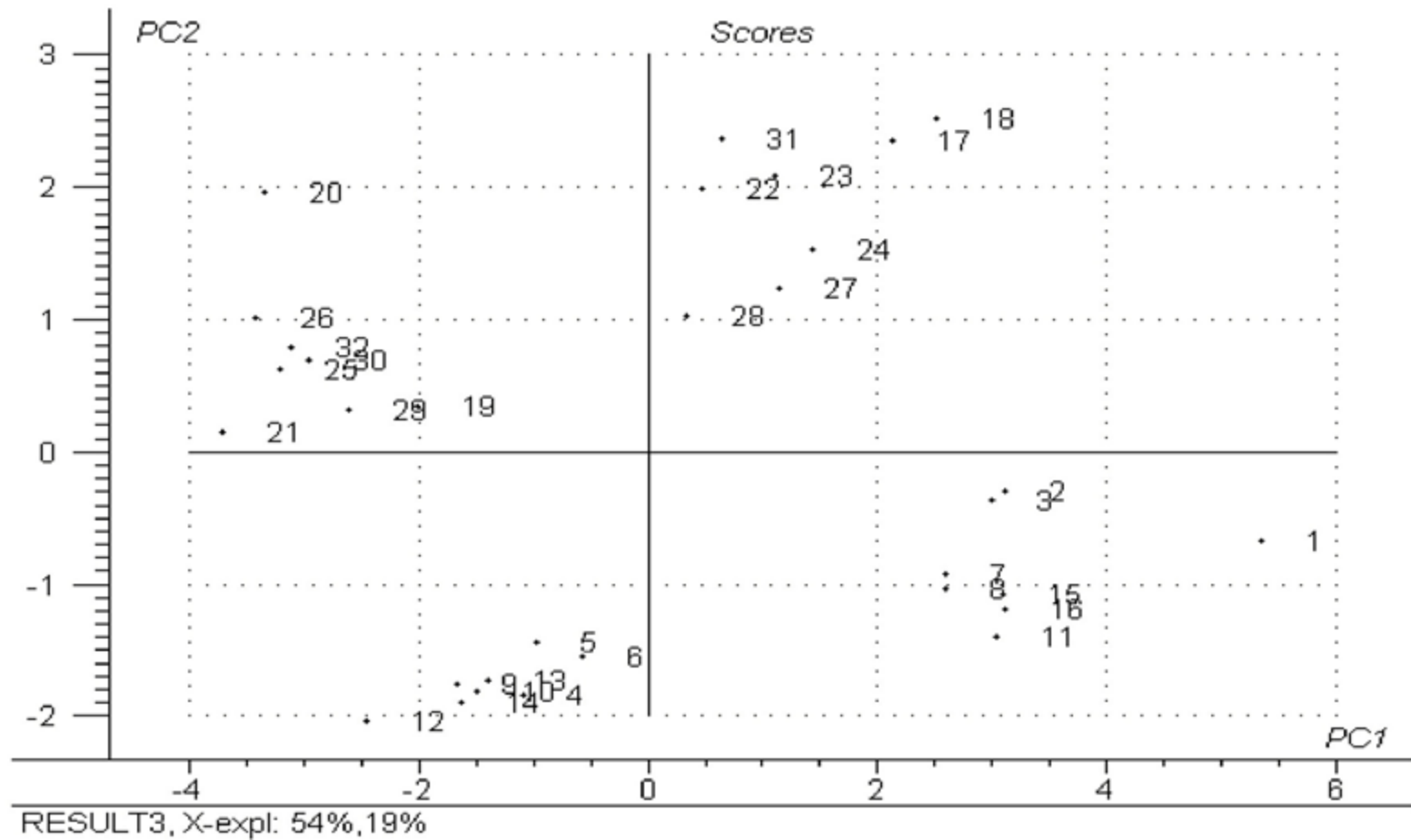


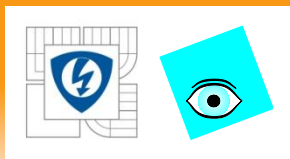
3. Rozptylový diagram komponentního skóre:

kóduje pohlaví (muž M a žena F) a zeměpisný původ (Skandinávie písmenem A a Středomoří písmenem B). Existují čtyři shluky, takže snadno odpovíme na otázky: Kdo jsou extrémní osoby? Které osoby jsou si navzájem podobné?

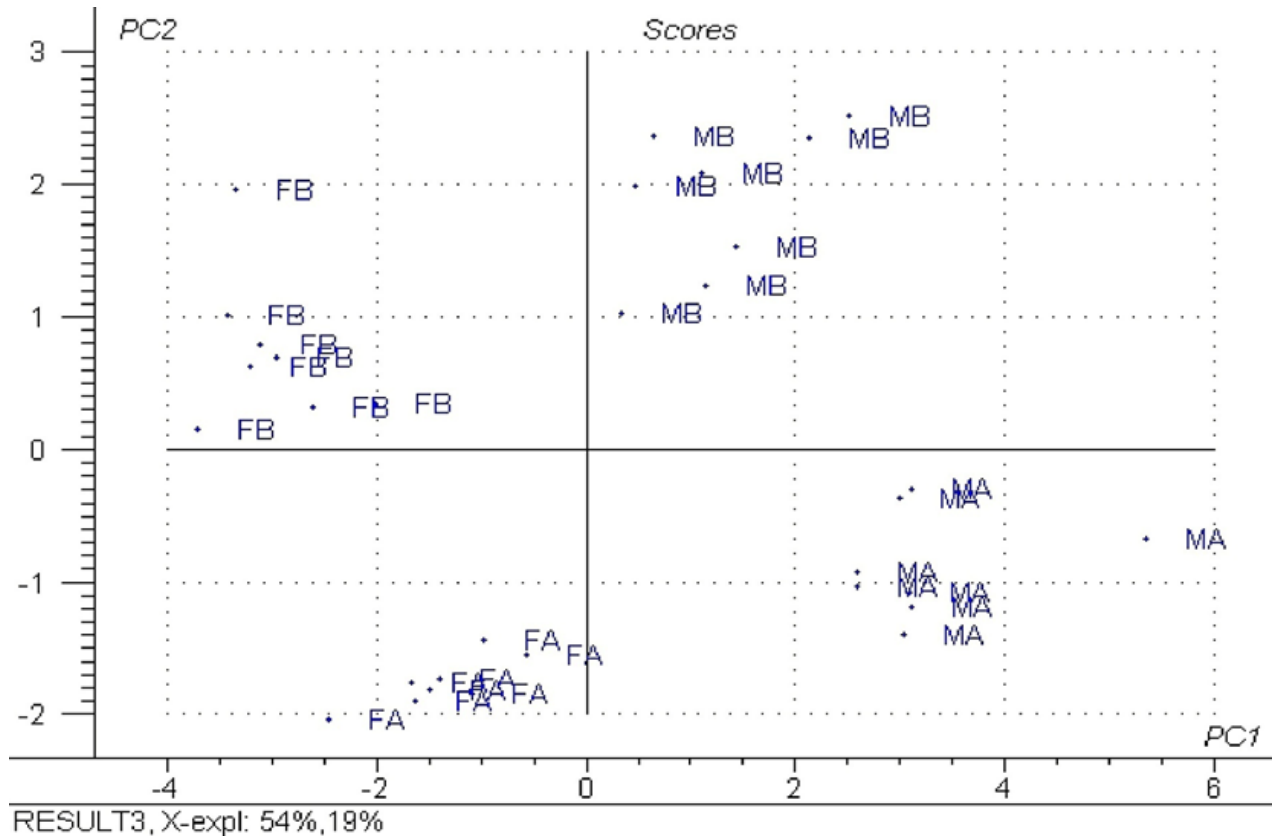


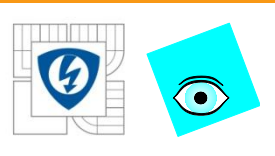
Obr. 4.4a Rozptylový diagram komponentního skóre při popisu objektů indexy dat Lidé (UNSCRAMBLER).





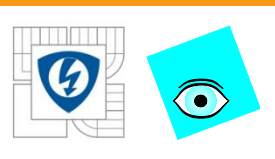
Obr. 4.4b Rozptylový diagram komponentního skóre při popisu objektů kódy dat Lidé: muži M, ženy F, A Skandinávie, B Středomoří (UNSCRAMBLER).



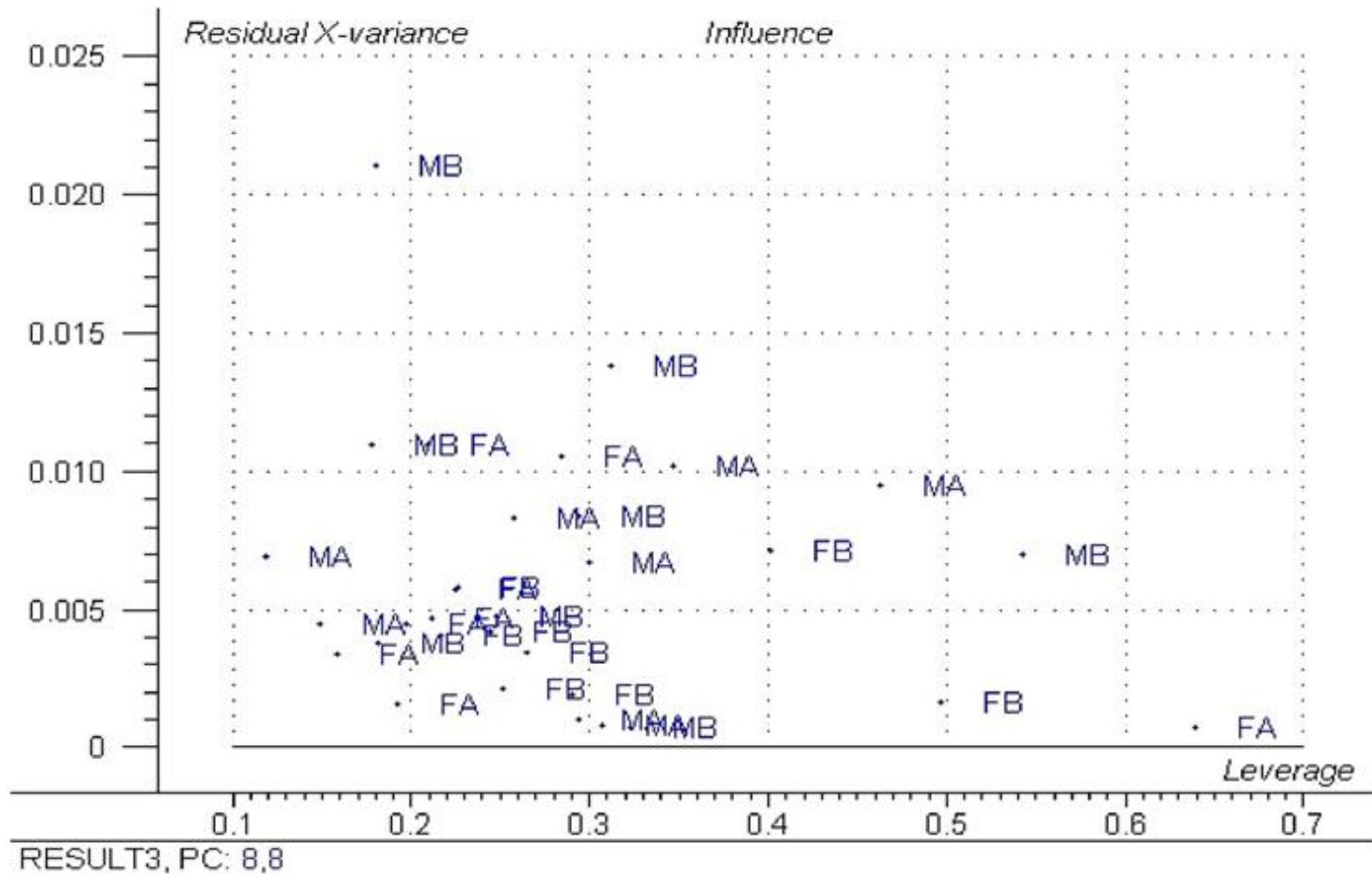


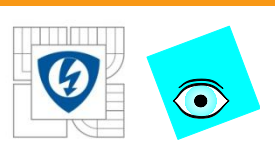
4. Vlivné body, objekty:

Z grafu vlivných bodů statistické analýzy reziduí objektů lze odhalit vlivné body tj. pořadová čísla vlivných objektů (obr. 4.5a). Tyto by měli být detailně zkoumány a určeny vybočující objekty (obr. 4.5b).

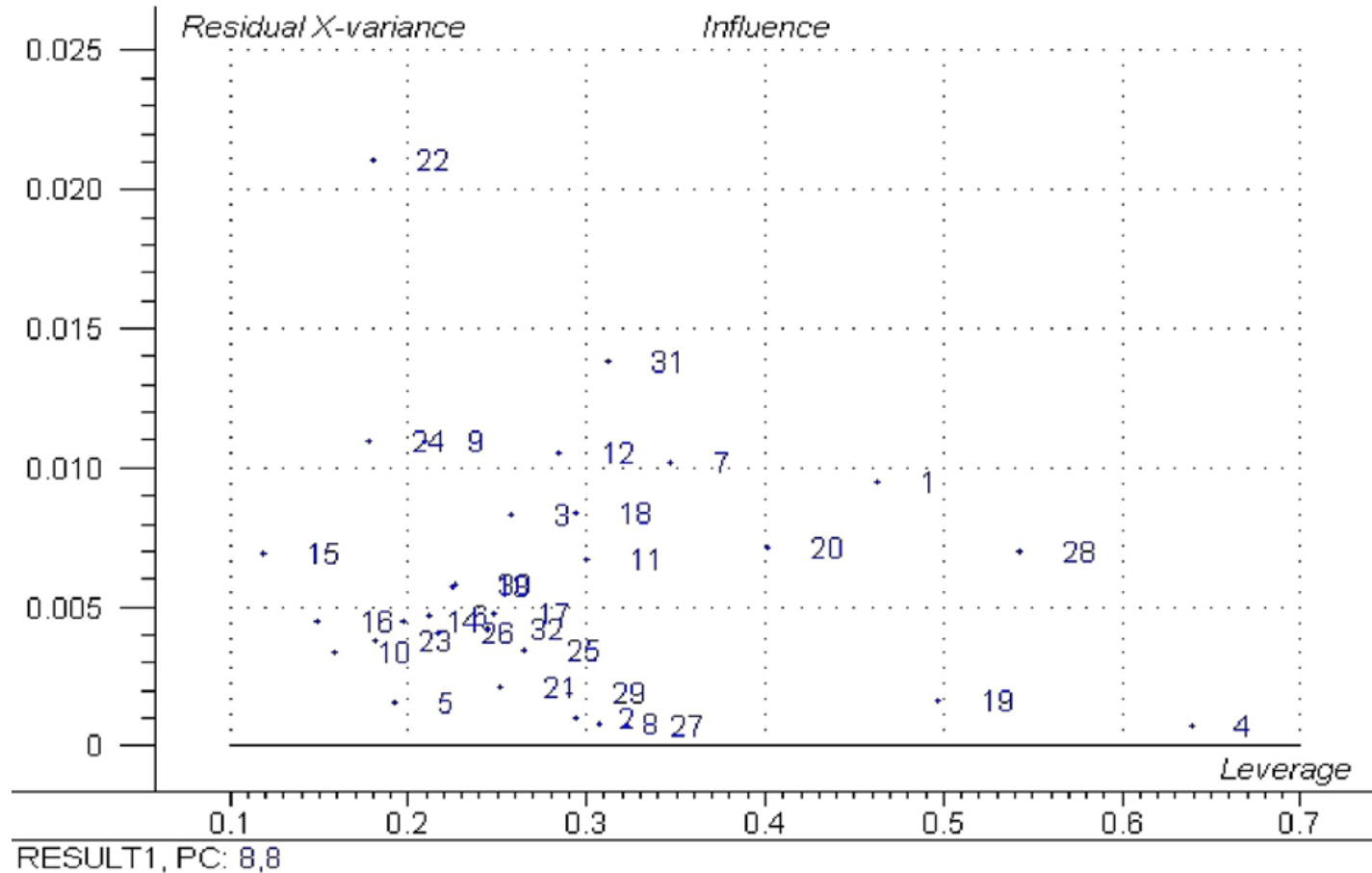


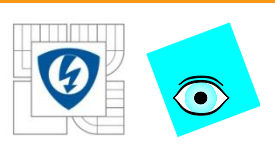
Obr. 4.5b Graf vlivných bodů statistické analýzy reziduí objektů při popisu objektů kódy dat Lidé (UNSCRAMBLER).





Obr. 4.5a Graf vlivných bodů statistické analýzy reziduí při indexovém značení objektů (UNSCRAMBLER).





5. Dvojný graf (Biplot):

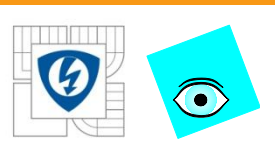
Vedle rozlišení osob do čtyř skupin se ukazuje, že osoby v levém dolním rohu pijí více víno a méně pivo než osoby v pravém horním rohu.

Tento trend na diagonálu je zřejmý přihlédnutím k hlavním komponentám $PC\ 1$ a $PC2$.

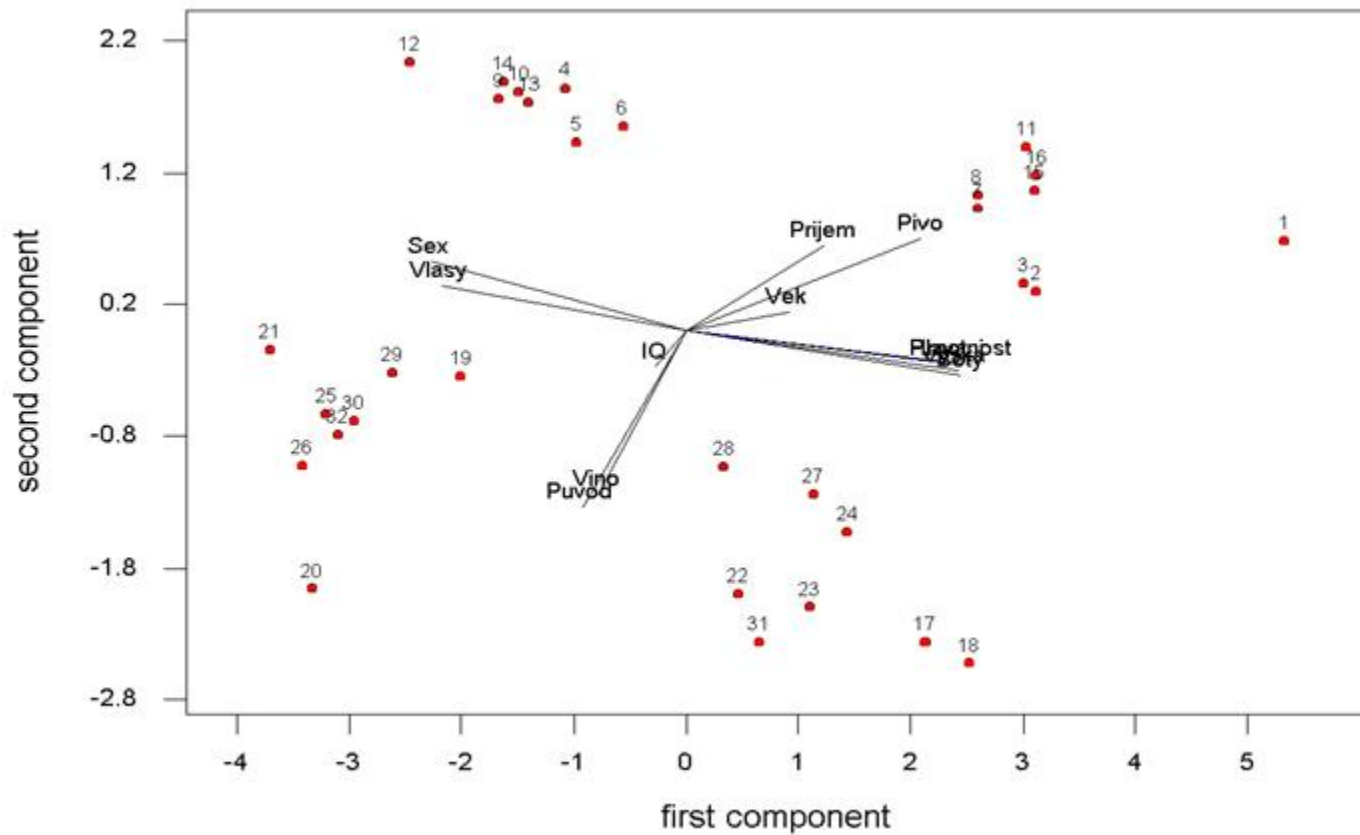
Znak víno má zřetelně vysoké záporné hodnoty $PC2$ a i $PC\ 1$, zatímco pivo má vysoké kladné hodnoty v komponentě $PC\ 1$ a dále kladné v komponentě $PC2$.

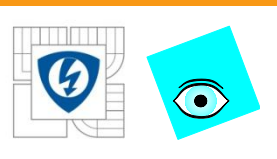
Dvojný graf také ukazuje na současnou tendenci lidí v regionu A pít více piva a méně vína a relativně vyšší mzdu než v regionu B.

Všichni muži vykazují relativně vysoké hodnoty znaků výška, hmotnost, velikost body, schopnost plavat a spotřeba piva.

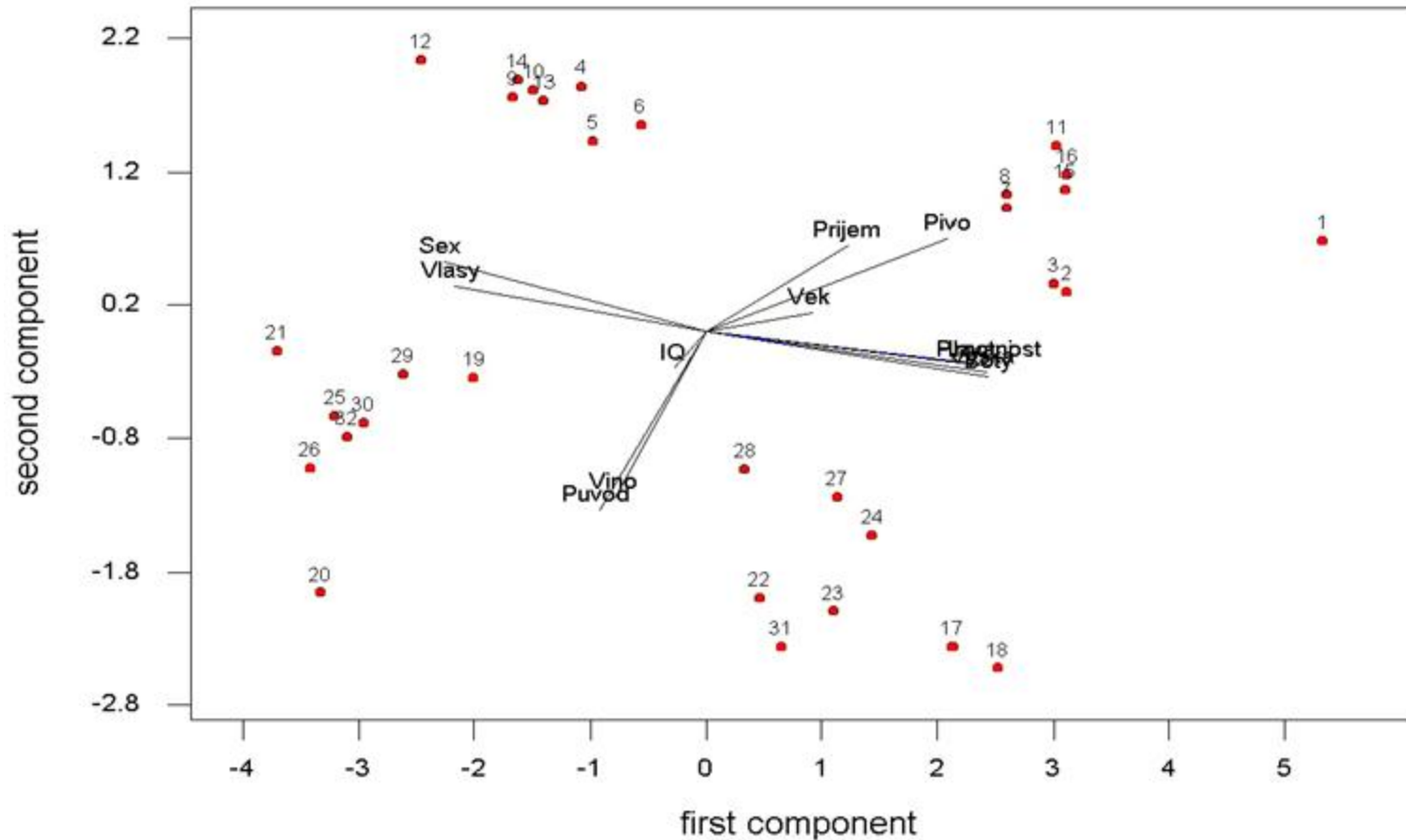


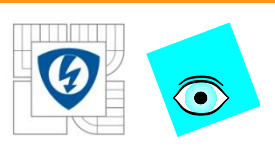
Biplot





Obr. 4.6 Dvojný graf Biplot zdrojové matice dat Lidé (SCAN).





Závěr

Analytická korelace neznamená nutně fyzikální korelaci v příčinném slova smyslu. Hmotnost, velikost bot a výška *musí* obecně korelovat u mladistvých ale věk nebude korelovat s výškou u osob zralého věku. IQ nezávisí na zeměpisném původu a fyzických charakteristikách. Je třeba zdůraznit, že metoda hlavních komponent poskytuje *úplný a nezaújatý pohled* na data, do kterého nevkládáme žádné předpojaté myšlenky.

STATISTICA Cz - [Data: 22Lide.sta* (13s krát 32ř)]

Soubor Úpravy Zobrazit Vložit Formát Statistika Grafy Nástroje Data Okno nápověda

Obnovit... Ctrl+R

Analýza skupin

- Základní statistiky/tabulky
- Vícerozměrná regrese
- ANOVA
- Neparametrická statistika
- Prokládání rozdělení
- Pokročilé lineární/nelineární modely
- Vícerozměrné průzkumné techniky
- Průmyslová statistika & Six Sigma
- Analýza gly testu
- Neuronové sítě
- Vytěžování dat
- QC Data mining & Analýza hlavních příčin
- Text & Document Mining, Web Crawling
- Statistiky bloku dat
- STATISTICA Visual Basic
- Pravděpodobnostní kalkulátor

Shluková analýza

Faktorová analýza

Hlavní komponenty & klasifikační analýza

Kanonická analýza

Analýza spolehlivosti/prvků

Klasifikační stromy

Korespondenční analýza

Vícerozměrné škálování

Diskriminační analýza

Modely obecné diskriminační analýzy

	1	2		7	8	9	10	11	12	13									
	Osoba	Vyska		Prijem	Pivo	Vino	Sex	Plavani	Puvod	IQ									
MA	MA	18		48	45000	420	115	-1	98	-1	100								
MA	MA	18		33	33000	350	102	-1	92	-1	130								
MA	MA	18		37	34000	320	98	-1	91	-1	127								
FA	FA	18		32	28000	270	78	1	75	-1	112								
FA	FA	18							81	-1	110								
FA	FA	18							82	-1	102								
MA	MA	18							85	-1	140								
MA	MA	18							84	-1	129								
FA	FA	18							78	-1	98								
FA	FA	18							78	-1	100								
MA	MA	18							90	-1	105								
FA	FA	18							70	-1	127								
FA	FA	18							76	-1	102								
FA	FA	18							75	-1	108								
MA	MA	18							88	-1	109								
MA	MA	18							86	-1	113								
MB	MB	18		26	18000	295	180	-1	92	1	109								
MB	MB	18		27	16500	299	178	-1	95	1	119								
FB	FB	188	50	1	37	49	34000	170	162	1	76	1	135						
FB	FB	186	49	1	36	21	14000	150	245	1	75	1	123						
FB	FB	158	46	1	34	30	18000	120	120	1	70	1	119						
MB	MB	177	65	-1	41	26	18000	209	160	-1	86	1	120						
MB	MB	180	72	-1	43	33	19000	236	175	-1	85	1	115						
MB	MB	181	75	-1	43	42	31000	198	161	-1	83	1	105						
FB	FB	163	50	1	36	18	11000	143	136	1	75	1	102						
FB	FB	162	50	1	36	20	11500	133	146	1	74	1	132						
MB	MB	176	68	-1	42	50	36000	195	177	-1	82	1	96						
MB	MB	175	67	1	42	55	38000	185	187	-1	80	1	105						
FB	FB	165	51	1	36	36	26000	121	129	1	76	1	126						
FB	FB	161	48	1	35	41	31500	116	196	1	75	1	120						
MB	MB	178	75	-1	42	30	24000	203	208	-1	81	1	118						
FB	FB	160	48	1	35	40	31000	118	198	1	74	1	129						

81. Načtení dat, překopírování 1. sloupce názvu objektů (respondentů, subjektů) do nultého sloupce, volba metody *Hlavní komponenty a klasifikační analýza* z menu *Statistika* a v okně *Vícerozměrné techniky*.

MA
MA
MA
FA
FA
FA
MA
MA
FA

FA	FA	168	52	1	37	27	23500	260	86	1	78	-1	100
MA	MA	183	81	4	42	27	25000	245	47	4	88	4	100
FA	FA	157	47	Zvolte proměnné pro analýzu, dopl. proměnné, aktivní případy a grupovací prom.									
FA	FA	164	50	1-Osoba	11-Plavani	1-Osoba	11-Plavani	1-Osoba	11-Plavani	1-Osoba	11-Plavani	1-Osoba	11-Plavani
FA	FA	162	49	2-Vyska	12-Puvod	2-Vyska	12-Puvod	2-Vyska	12-Puvod	2-Vyska	12-Puvod	2-Vyska	12-Puvod
MA	MA	180	82	3-Hmotnost	13-IQ	3-Hmotnost	13-IQ	3-Hmotnost	13-IQ	3-Hmotnost	13-IQ	3-Hmotnost	13-IQ
MA	MA	180	81	4-Vlasy		4-Vlasy		4-Vlasy		4-Vlasy		4-Vlasy	
MB	MB	185	82	5-Boty		5-Boty		5-Boty		5-Boty		5-Boty	
MB	MB	187	84	6-Vek		6-Vek		6-Vek		6-Vek		6-Vek	
FB	FB	168	50	7-Prijem		7-Prijem		7-Prijem		7-Prijem		7-Prijem	
FB	FB	166	49	8-Pivo		8-Pivo		8-Pivo		8-Pivo		8-Pivo	
FB	FB	158	46	9-Vino		9-Vino		9-Vino		9-Vino		9-Vino	
MB	MB	177	65	10-Sex		10-Sex		10-Sex		10-Sex		10-Sex	
MB	MB	180	72										
MB	MB	181	75										
FB	FB	163	50										
FB	FB	162	50										
MB	MB	176	68										
MB	MB	175	67										
FB	FB	165	51										
FB	FB	161	48										
MB	MB	178	75										
FB	FB	160	48										

Zvolte proměnné pro analýzu, dopl. proměnné, aktivní případy a grupovací proměnné.

1-Osoba	11-Plavari	1-Osoba	11-Plavari	1-Osoba	11-Plavari	1-Osoba	11-Plavari
2-Vyska	12-Puvod	2-Vyska	12-Puvod	2-Vyska	12-Puvod	2-Vyska	12-Puvod
3-Hmotnost	13-IQ	3-Hmotnost	13-IQ	3-Hmotnost	13-IQ	3-Hmotnost	13-IQ
4-Vlasy		4-Vlasy		4-Vlasy		4-Vlasy	
5-Boty		5-Boty		5-Boty		5-Boty	
6-Vek		6-Vek		6-Vek		6-Vek	
7-Prijem		7-Prijem		7-Prijem		7-Prijem	
8-Pivo		8-Pivo		8-Pivo		8-Pivo	
9-Vino		9-Vino		9-Vino		9-Vino	
10-Sex		10-Sex		10-Sex		10-Sex	

Vólba "Ukázat pouze odpovídající proměnné" slouží k předzpracování seznamu proměnných a zobrazení kategoričkových a spojitých proměnných. Pro více informací stiskněte F1.

Proměnné pro analýzu:
 Doplňkové proměnné:
 Prom. aktiv. případů:
 Grupov. proměnná:

☐ Ukázat pouze odpovídající proměnné

Hlavní komponenty a klasifikační analýza: 22Lide.sta

Základní nastavení Detaily

Proměnné:

Proměnné pro analýzu: **žádné**

Doplňkové proměnné: **žádné**

Proměnná s aktiv. případy: **žádné**

Grupovací proměnná (popis): **žádné**

Kód aktivních případů:

Analýza podle

☒ Korelaci

☐ Kovarianci

Rozptyl počítat

☒ jako SČ/(N-1)

☐ jako SČ/N

OK

Storno

Možnosti

Otevřít data

SELECT CASES

ChD vynechána

☒ Celé případy

☐ Nahradit průměrem

STATISTICA Cz - [Data: 22Lide.sta* (13s krát 32ř)]

Soubor Úpravy Zobrazit Vložit Formát Statistika Grafy Nástroje Data Okno Nápověda

Přidat do sešitu Přidat do protokolu

Arial 9 B I U

+.000 +.000

Proměnné Případy

	1	2	3	4	5	6	7	8	9	10	11	12	13
	Osoba	Vyska	Hmotnost	Vlasy	Boty	Vek	Prijem	Pivo	Vino	Sex	Plavani	Puvod	IQ
MA	MA	198	92	-1	48	48	45000	420	115	-1	98	-1	100
MA	MA	184										-1	130
MA	MA	183										-1	127
FA	FA	166										-1	112
FA	FA	170										-1	110
FA	FA	172										-1	102
MA	MA	182										-1	140
MA	MA	180										-1	129
FA	FA	169										-1	98
FA	FA	168										-1	100
MA	MA	183										-1	105
FA	FA	157										-1	127
FA	FA	164										-1	102
FA	FA	162										-1	108
MA	MA	180										-1	109
MA	MA	180										-1	113
MB	MB	185										1	109
MB	MB	187										1	119
FB	FB	168										1	135
FB	FB	166										1	123
FB	FB	158										1	119
MB	MB	177										1	120
MB	MB	180										1	115
MB	MB	181										1	105
FB	FB	163										1	102
FB	FB	162										1	132
MB	MB	176										1	96
MB	MB	175										1	105
			67	1	42	55	38000	185	187	-1	80		126
													120
													118
													129

Výsledky hlavních komponent a klasifikační analýzy: 22Lide.sta

Poč. aktiv. prom.: 12 Počet doplňkových prom.: 0
 Poč. aktiv. příp.: 32 Počet doplňkových příp.: 0

Počet faktorů: 12 Kvalita reprezentace: 100,0 %

Základní výsledky Proměnné Případy Popis. statistiky

☒ Faktorové souřadnice prom. ☐ Korelace faktorů & proměnných
☒ 2D graf fakt. souřadnic prom. ☐ Komunalita (cos²) na druhou
 Možnosti grafu fakt. souřadnic:
☒ Vektory (ukazují do počátku)
☒ Jednotková kružnice
☐ Názyv proměnných
☐ Čísla proměnných
☐ Bez čísel/názvů
☐ Příspěvky proměnných
☐ Vlastní čísla
☐ Sutinový graf
☐ Vlastní vektory

OK Storno Možnosti

83. Otevřeme *Výsledky hlavních komponent a klasifikační analýzy* a soustředíme se na podrobnější záložky. Nejprve proto otevřeme záložku *Proměnné* a v ní pak *2D-graf fakt. souřadnic prom.*

84. Otevřeme první graf komponentních vah 1. a 2. hlavní komponenty a kliknutím na kolečka znaků nastavíme jejich grafiku (velikost, tvar, barvu).