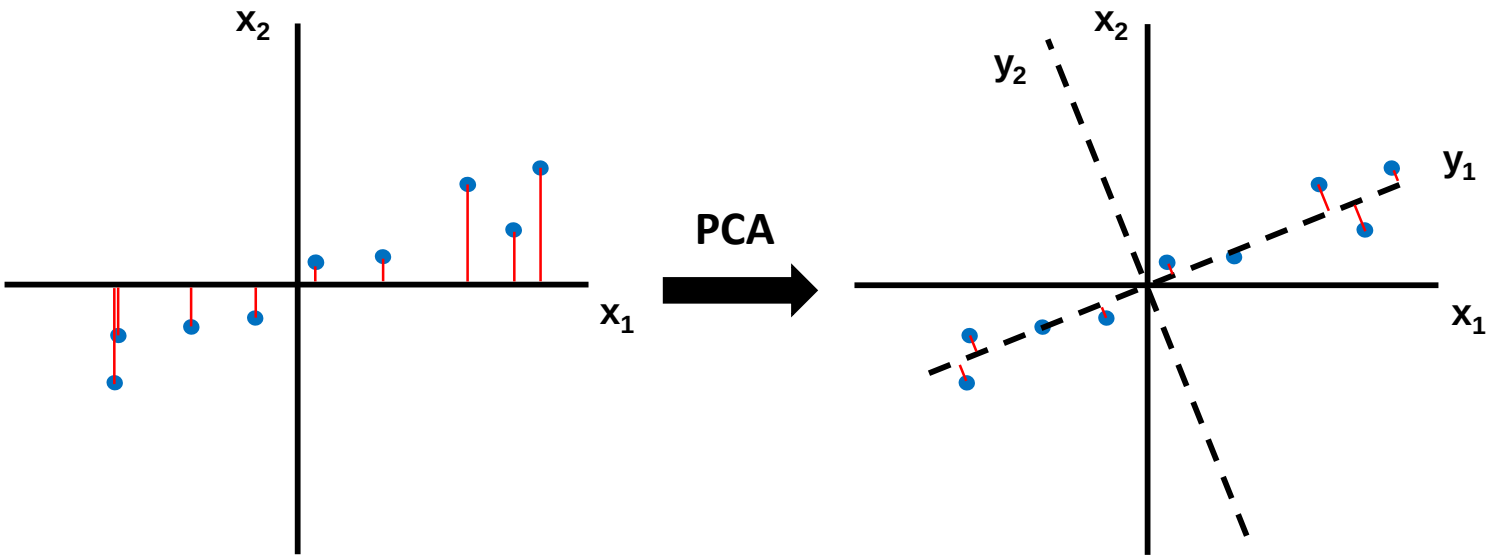
Vícerozměrné metody – cvičení

RNDr. Eva Koriťáková

Podzim 2016

Analýza hlavních komponent – opakování

- anglicky Principal component analysis (PCA)
- snaha redukovat počet proměnných nalezením nových latentních proměnných (hlavních komponent) vysvětlujících co nejvíce variability původních proměnných
- nové proměnné ($\mathbf{y}_1$, $\mathbf{y}_2$) lineární kombinací původních proměnných ($\mathbf{x}_1$, $\mathbf{x}_2$)



Analýza hlavních komponent – opakování II

- vstup do PCA?
- hlavní komponenty odpovídají čemu?
- variabilita vysvětlená příslušnou komponentou odpovídá čemu?
- vlastní vektory seřazeny jak?
- předpoklady?

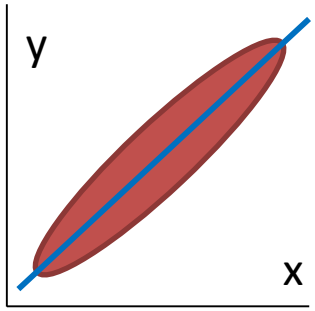
Analýza hlavních komponent – opakování II

- vstup do PCA:
  - kovarianční matice
  - matice korelačních koeficientů
- hlavní komponenty odpovídají vlastním vektorům kovarianční matice (či matice korelačních koef.)
- variabilita vysvětlená příslušnou komponentou odpovídá vlastním číslům
- vlastní vektory seřazeny podle vlastních hodnot (sestupně) $\Rightarrow$ vybráno prvních *m* komponent vyčerpávajících nejvíce variability původních dat
- předpoklady: kvantitativní proměnné s normálním rozdělením

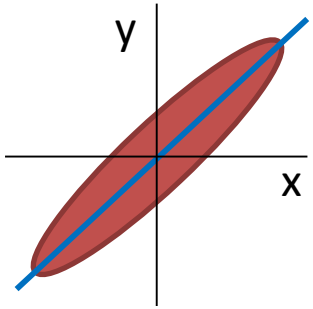
Analýza hlavních komponent – volba asociační matice

- **kovarianční (disperzní) matice** – data centrována (od každé příznakové proměnné odečtena její střední hodnota) – zohledňován rozptyl původních dat
- **matice korelačních koeficientů** – data standardizována (odečtení středních hodnot a podělení směrodatnými odchylkami) – použití pokud mají proměnné různá měřítka

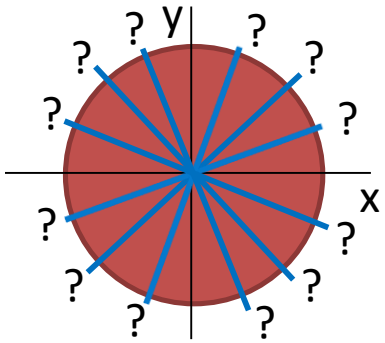
původní data



kovarianční matice (odečten průměr)



matice korelačních koeficientů (odečten průměr a podělení SD)



- **každou úpravou původních dat přicházíme o určitou informaci !!!**

Analýza hlavních komponent – postup

1. Volba asociační matice (autokorelační, kovarianční nebo kor. koeficientů)
2. Výpočet vlastních čísel a vlastních vektorů asociační matice:
   - vlastní vektory definují směr nových faktorových os (hlavních komponent) v prostoru
   - vlastní čísla odrážejí variabilitu vysvětlenou příslušnou komponentou
3. Seřazení vlastních vektorů podle hodnot jim odpovídajících vlastních čísel (sestupně)
4. Výběr prvních *m* komponent vyčerpávajících nejvíce variability původních dat

Identifikace optimálního počtu hlavních komponent pro další analýzu

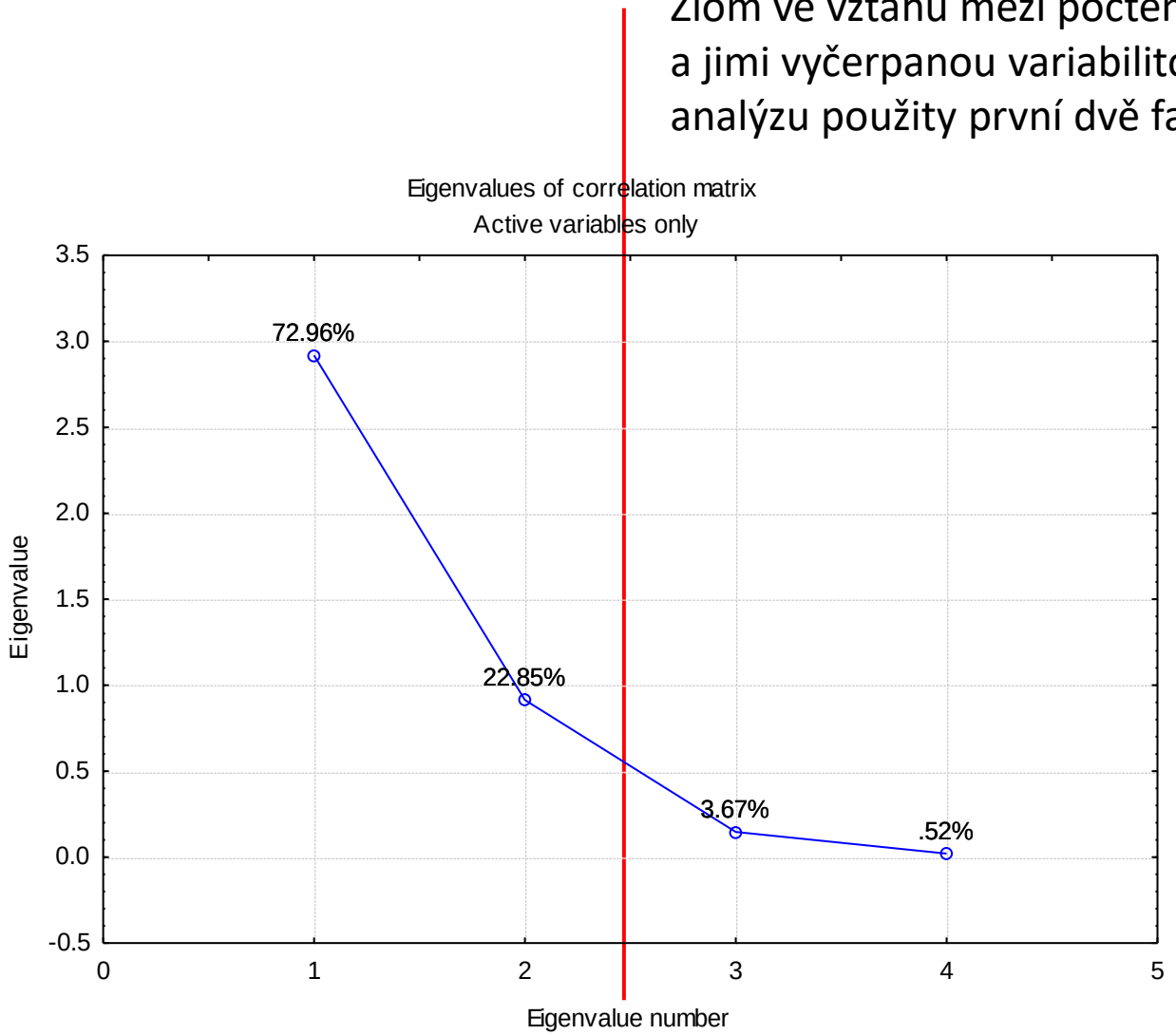
- pokud je cílem ordinační analýzy vizualizace dat, snažíme se vybrat 2-3 komponenty
- pokud je cílem ordinační analýzy výběr menšího počtu dimenzí pro další analýzu, můžeme ponechat více komponent (např. u analýzy obrazů MRI je úspěchem redukce z milionu voxelů na desítky)

- kritéria pro výběr počtu komponent:

1. Kaiser Guttmanovo kritérium:
   - pro další analýzu jsou vybrány osy s vlastním číslem >1 (při analýze matice korelačních koeficientů) nebo větším než průměrná hodnota vlastních čísel (při analýze kovarianční matice)
   - logika je vybírat osy, které přispívají k vysvětlení variability dat více, než připadá rovnoměrným rozdělením variability
2. Sutinový graf (scree plot)
   - grafický nástroj hledající zlom ve vztahu počtu os a vyčerpané variability
3. Sheppardův diagram
   - grafická analýza vztahu mezi vzdálenostmi objektů v původním prostoru a redukovaném prostoru o daném počtu dimenzí

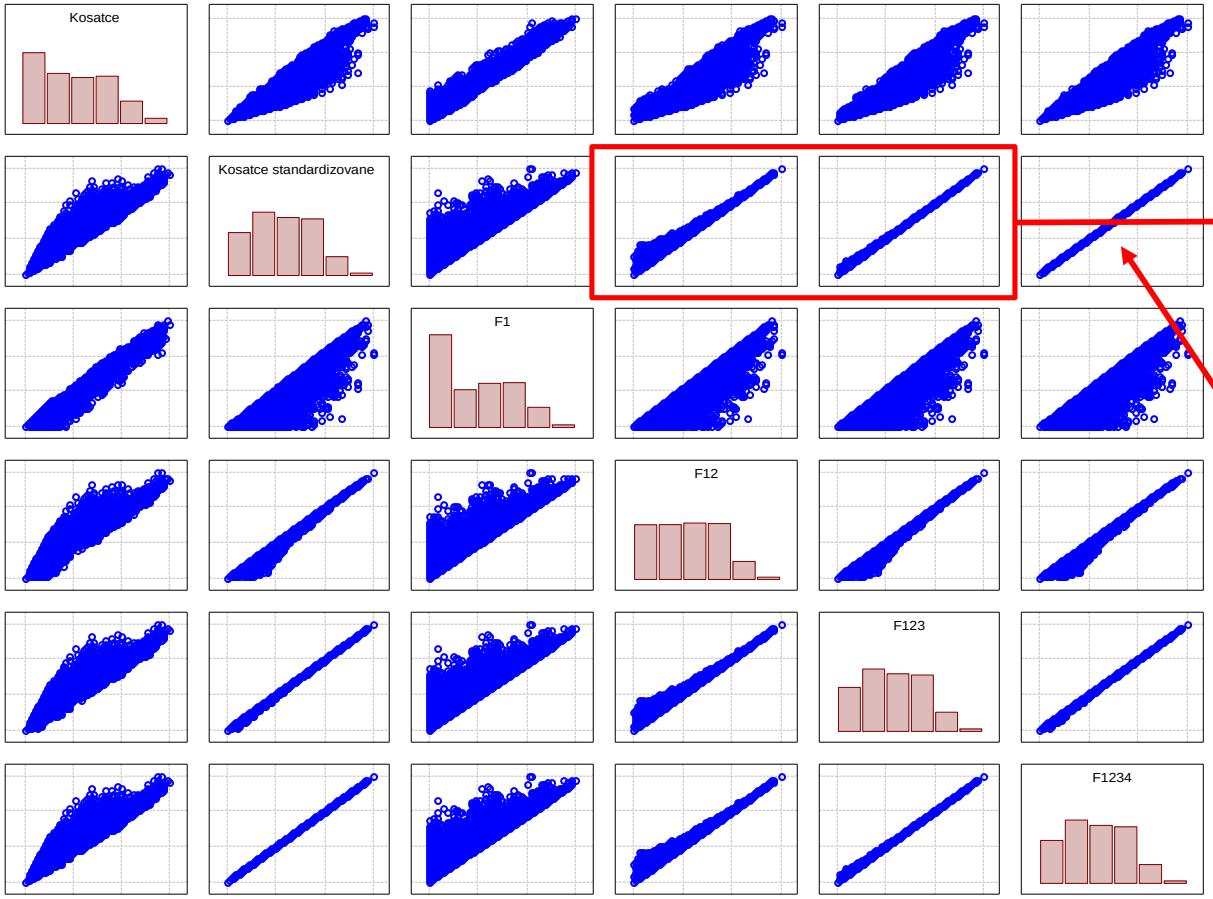
Sutinový graf (scree plot)

Zlom ve vztahu mezi počtem vlastních čísel a jimi vyčerpanou variabilitou – pro další analýzu použity první dvě faktorové osy



Sheppardův diagram

- Vztahuje vzdálenosti v prostoru původních proměnných ke vzdálenostem v prostoru vytvořeném PCA
- Je třeba brát ohled na typ PCA (korelace vs. kovariance)
- Obecná metoda určení optimálního počtu dimenzí v ordinační analýze (třeba respektovat použitou asociační metriku)



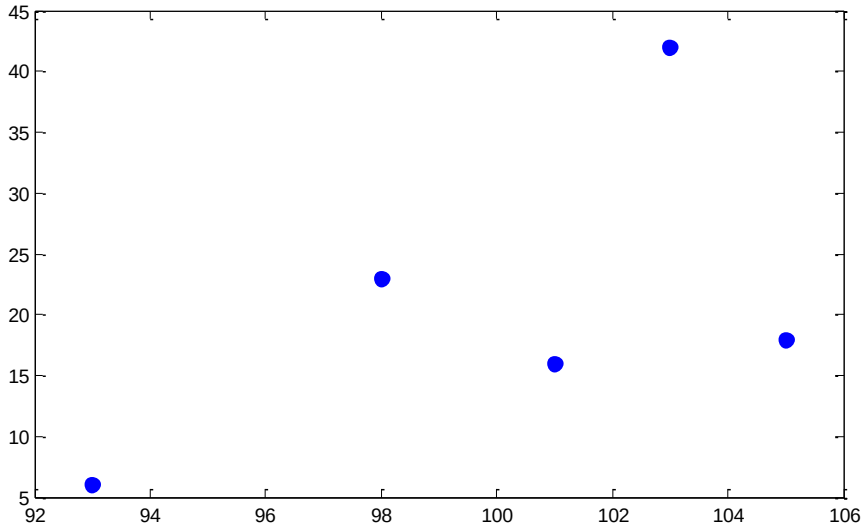
Za optimální z hlediska zachování vzdáleností objektů lze považovat dvě nebo tři dimenze

Při použití všech dimenzí jsou vzdálenosti perfektně zachovány

PCA – příklad

- data:

A	101	16
B	105	18
C	103	42
D	98	23
E	93	6



PCA – příklad – ruční počítání I

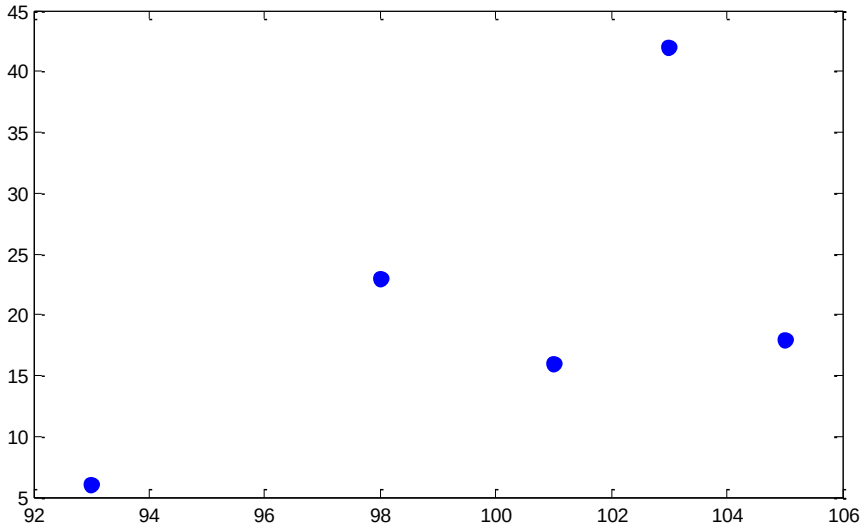
- data:

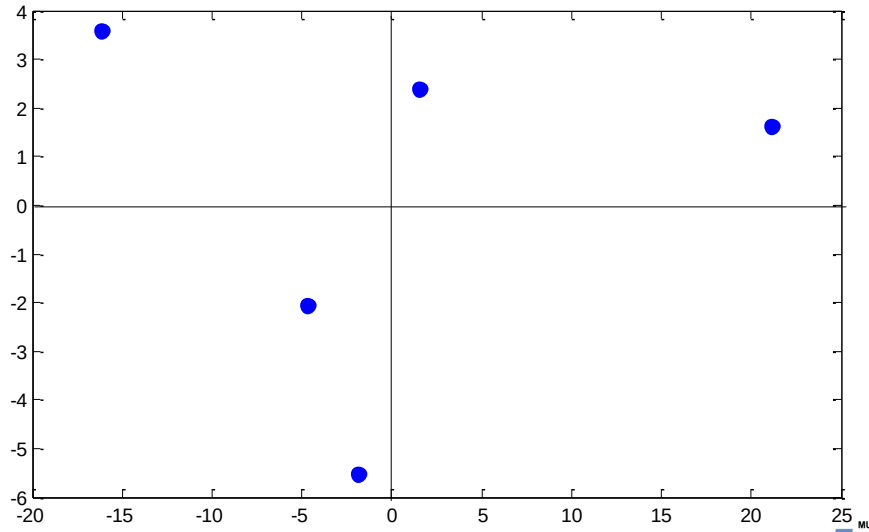
A	101	16
B	105	18
C	103	42
D	98	23
E	93	6

PCA – příklad – ruční počítání II

- data:

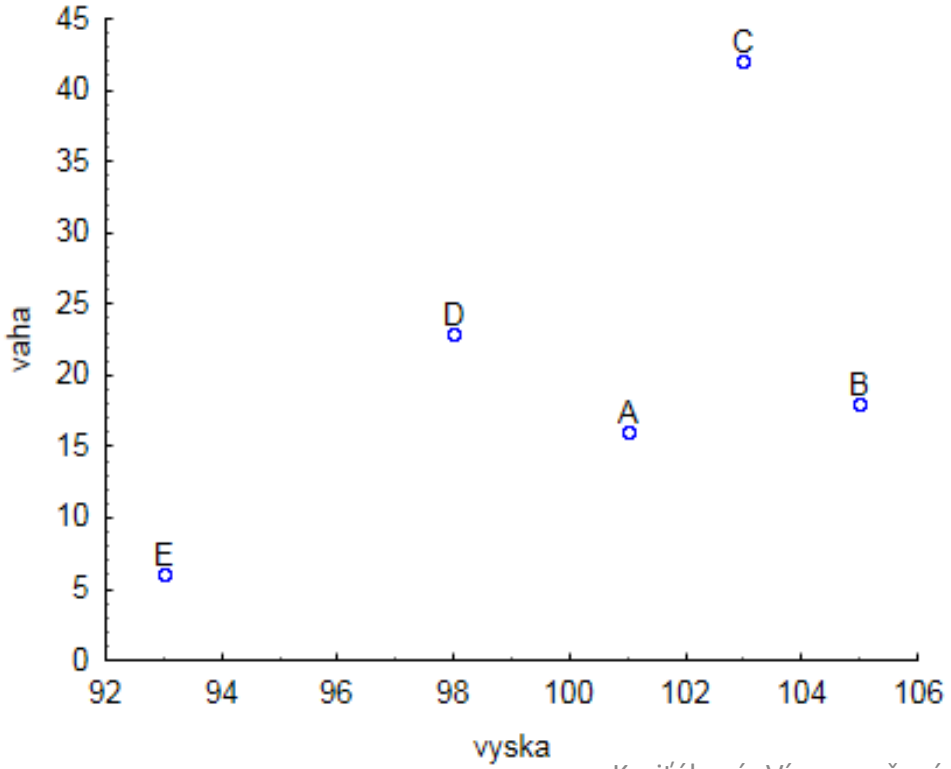
A	101	16
B	105	18
C	103	42
D	98	23
E	93	6





PCA – příklad – řešení v softwaru Statistica I

- vykreslení datového souboru včetně textových popisků:
  Graphs – Scatterplots... – zvolit proměnné (výška jako X, váha jako Y) – OK – vypnout zatržení „Fit type“ Linear – na záložce Options 1 zatrhnout „Display case labels“ – Case labels ze Spreadsheet změnit na Variable a vybrat proměnnou id – OK



PCA – příklad – řešení v softwaru Statistica II

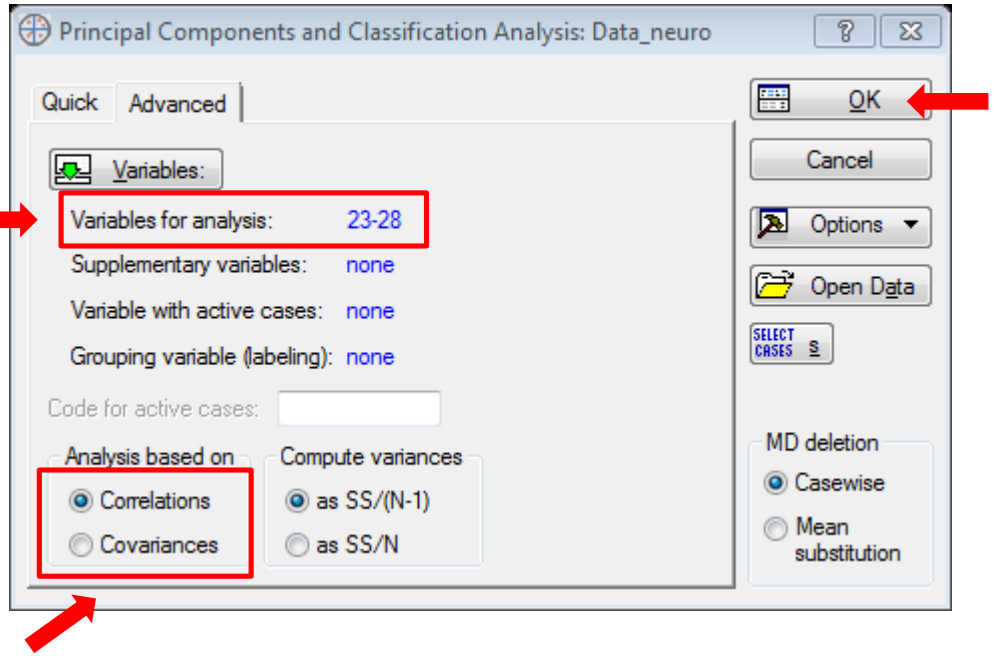
Výpočet PCA – postup:

- Statistics – Multivariate Exploratory Techniques – Principal Components & Classification Analysis – zvolit proměnné jako “Variables for analysis” , na záložce Advanced zvolit „Correlations“ nebo “Covariances” – OK
- záložka Descriptives – „Covariance matrix“
- záložka Variables:
  - “Eigenvectors” (vlastní vektory), “Eigenvalues” (vlastní čísla včetně procenta vyčerpané variability), “Scree plot” (sutinový graf)
  - “Factor & variable correlations” (korelace nových faktorových os s původními proměnnými)
  - “Factor coordinates of variables” jsou souřadnice vlastních vektorů, kterým odpovídá graf “Plot var. factor coordinates, 2D” (čím menší úhel svírá původní proměnná s novou faktorovou osou, tím větší vztah má nová faktorová osa s původní proměnnou)
- záložka Cases:
  - „Factor coordinates of cases“ (souřadnice dat v novém prostoru – data jsou centrována), „Plot case factor coordinates, 2D“ (vykreslení dat)

PCA – příklad 2 – řešení v softwaru Statistica I

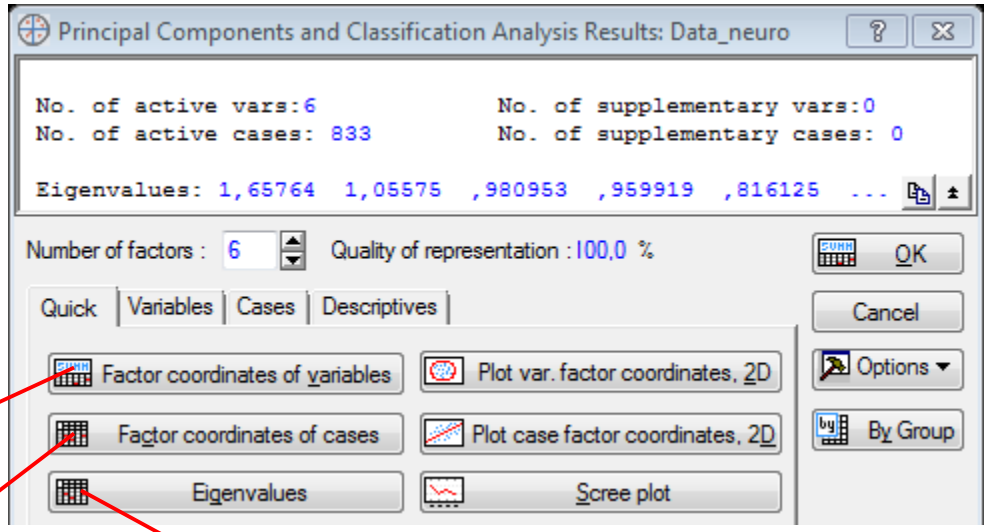
- Zadání: Proveďte PCA na objemech 6 mozkových struktur u 833 subjektů.
- Řešení: Statistics – Multivariate Exploratory Techniques – Principal Components & Classification Analysis

vybrat proměnné



zvolit, zda se má počítat kovarianční či korelační matice

PCA – příklad 2 – řešení v softwaru Statistica II



Matice vlastních vektorů

Factor coordinates of the variables, based on correlations (Data_neuro)

Variable	**Factor 1**	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Hippocampus_volume (mm3)	-22,5292	-331,381	12,852	-24,9019	-62,1764	-56,1444
Amygdala_volume (mm3)	-19,8762	-139,756	25,334	60,1821	174,1211	20,4766
Thalamus_volume (mm3)	0,6579	-37,303	-261,504	20,9163	4,3841	12,7030
Pallidum_volume (mm3)	-7,6336	-20,868	27,707	184,5947	-73,9145	34,4372
Putamen_volume (mm3)	-14,6603	-86,934	15,376	-55,5188	-27,4129	166,7655
Nucl_caud_volume (mm3)	634,4294	-18,367	2,210	1,9177	1,7198	2,9026

Souřadnice subjektů v novém prostoru

Factor coordinates of cases, based on covariances (Data_neuro)

Case	**Factor 1**	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
1	-541,68	-322,060	-90,545	94,230	-249,661	-27,353
2	-306,11	-508,246	423,531	-204,078	-40,595	-148,339
3	218,03	-473,620	-192,820	-163,206	-82,362	128,077
4	-492,70	-535,503	267,883	-74,278	-56,033	-351,386
5	-346,39	-240,774	312,983	-106,921	-5,006	32,832
6	-123,10	-749,883	315,002	-241,681	63,288	-46,083
7	-1179,78	-76,816	150,773	321,967	-182,452	162,240
8	-321,21	-8,941	255,254	151,791	-36,504	192,658

Vlastní čísla

Eigenvalues of covariance ma
Active variables only

Value number	**Eigenvalue**	% Total variance
1	403677,0	55,45440
2	139067,1	19,10409
3	70200,2	9,64363
4	41840,7	5,74779
5	40421,1	5,55277
6	32737,9	4,49732

PCA – příklad 2 – řešení v softwaru Statistica III

Normalizace vlastních vektorů:

- zkopírovat do Excelu („Copy with headers“)
- použití vzorce: =B3/ODMOCNINA(SUMA.ČTVERCŮ(B$3:B$8))

	A	B	C	D	E	F	G
1		Factor coordinates of the variables, based on correlations (Data_neuro)					
2	Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
3	Hippocamp	-22,5292	-331,381	12,852	-24,9019	-62,1764	-56,1444
4	Amygdala_	-19,8762	-139,756	25,334	60,1821	174,1211	20,4766
5	Thalamus_	0,6579	-37,303	-261,504	20,9163	4,3841	12,7030
6	Pallidum_v	-7,6336	-20,868	27,707	184,5947	-73,9145	34,4372
7	Putamen_v	-14,6603	-86,934	15,376	-55,5188	-27,4129	166,7655
8	Nucl_caud	634,4294	-18,367	2,210	1,9177	1,7198	2,9026
9							
10		-0,035459125	-0,88862	0,048506	-0,12174	-0,30926	-0,3103
11		-0,031283533	-0,37477	0,095616	0,294217	0,866059	0,11317
12		0,001035499	-0,10003	-0,98698	0,102255	0,021806	0,070207
13		-0,01201473	-0,05596	0,104572	0,902443	-0,36764	0,190328
14		-0,023074151	-0,23312	0,058032	-0,27142	-0,13635	0,921681
15		0,998542011	-0,04925	0,008341	0,009375	0,008554	0,016042

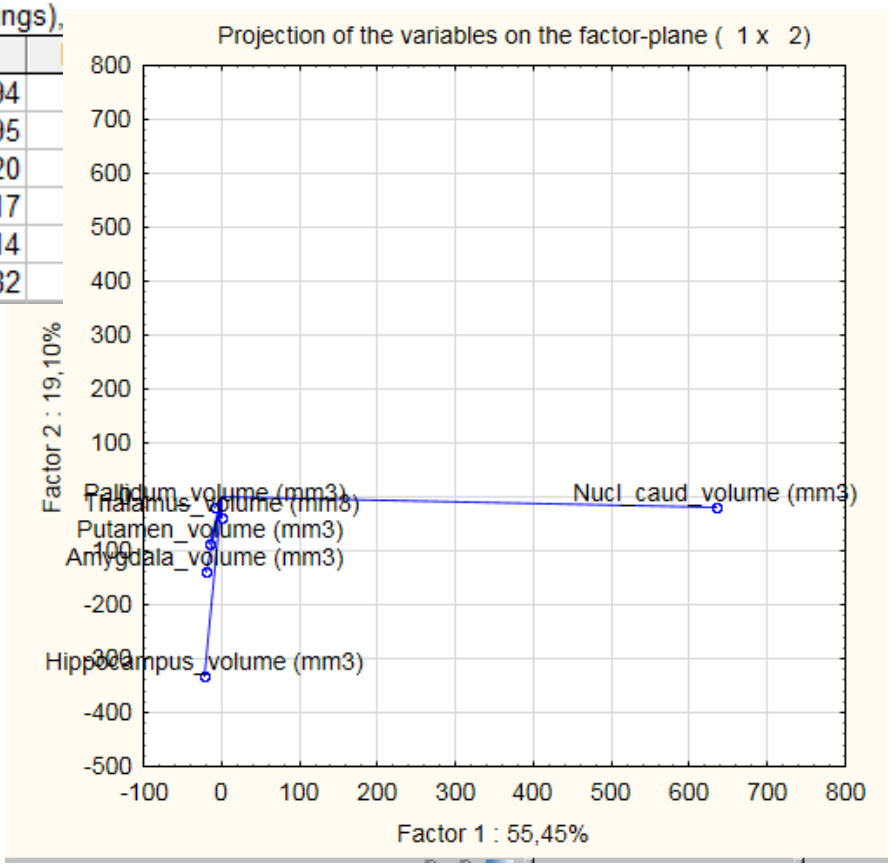
PCA – příklad 2 – řešení v softwaru Statistica IV

Záložka Variables:

Factor & variable correlations

Plot var. factor coordinates, 2D

Variable	Factor-variable correlations (factor loadings). Factor 1	Factor 2	Factor 3
Hippocampus_volume (mm3)	-0,065550	-0,964180	0,037394
Amygdala_volume (mm3)	-0,084808	-0,596314	0,108095
Thalamus_volume (mm3)	0,002480	-0,140597	-0,985620
Pallidum_volume (mm3)	-0,037255	-0,101845	0,135217
Putamen_volume (mm3)	-0,073621	-0,436566	0,077214
Nucl_caud_volume (mm3)	0,999556	-0,028938	0,003482



Z výsledků vyplývá, že:

- 1. hlavní komponenta je nevíce korelovaná s objemem Nucleus caudatus
- 2. hlavní komponenta je korelovaná s objemem hipokampu a také s objemem amygdaly a putamenu

PCA – příklad 2 – řešení v Matlabu

- Zadání: Proveďte PCA na objemech 6 mozkových struktur u 833 subjektů.
- Řešení:

```
[num, txt, raw] = xlsread('Data_neuro.xlsx',1);
data = num(:,23:28); % vyber 6 promennych s objemy mozkovych struktur
[coeff,score,latent] = pca(data);
```

Souřadnice subjektů v novém prostoru

score

833x6 double

	1	2	3	4	5	6
1	-541.6758	322.0604	90.5446	94.2298	-249.6611	-27.3529
2	-306.1072	508.2459	-423.5306	-204.0785	-40.5948	-148.3389
3	218.0346	473.6196	192.8200	-163.2062	-82.3617	128.0769
4	-492.7048	535.5033	-267.8827	-74.2783	-56.0326	-351.3861
5	-346.3904	240.7737	-312.9827	-106.9215	-5.0059	32.8323
6	-123.1009	749.8831	-315.0017	-241.6806	63.2878	-46.0834
7	-1.1798e+03	76.8159	-150.7726	321.9671	-182.4523	162.2400
8	-321.2074	8.9410	-255.2537	151.7913	-36.5035	192.6580
9	-345.8090	464.1571	-374.4555	11.8603	-5.8649	91.6828
10	-1.4653e+03	697.7425	-380.2903	267.2337	-19.2383	-81.4055

hlavní komponenty jsou ve sloupcích (jsou seřazené podle vlastních čísel);
v řádcích jsou subjekty

Matice vlastních vektorů

coeff

6x6 double

	1	2	3	4	5	6
1	-0.0355	0.8886	-0.0485	-0.1217	-0.3093	-0.3103
2	-0.0313	0.3748	-0.0956	0.2942	0.8661	0.1132
3	0.0010	0.1000	0.9870	0.1023	0.0218	0.0702
4	-0.0120	0.0560	-0.1046	0.9024	-0.3676	0.1903
5	-0.0231	0.2331	-0.0580	-0.2714	-0.1363	0.9217
6	0.9985	0.0493	-0.0083	0.0094	0.0086	0.0160

vlastní vektory jsou ve sloupcích (jsou seřazené podle vlastních čísel)

Vlastní čísla

latent

6x1 double

	1
1	4.0368e+05
2	1.3907e+05
3	7.0200e+04
4	4.1841e+04
5	4.0421e+04
6	3.2738e+04