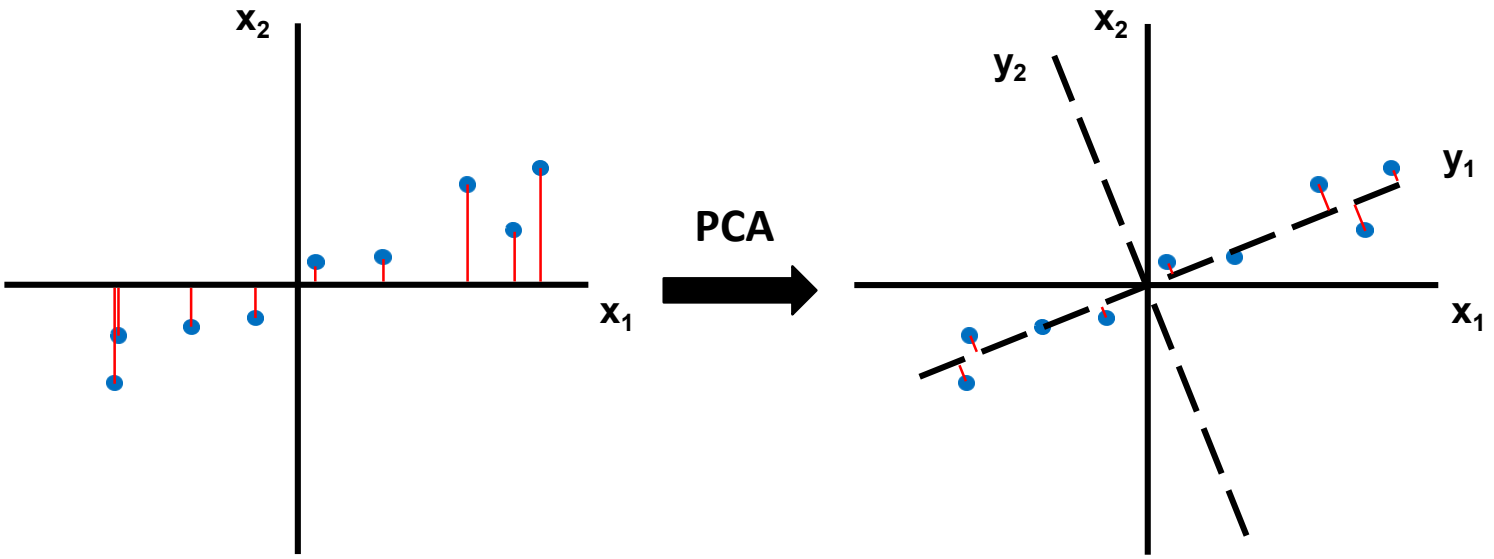
Vícerozměrné metody – cvičení

Mgr. Jan Zdražil

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 ($\mathbf{x}_1$, $\mathbf{x}_2$)



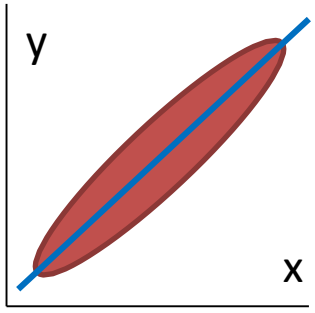
Analýza hlavních komponent – opakování II

- **vstup do PCA?**
  - kovarianční matice
  - matice korelačních koeficientů
- **hlavní komponenty odpovídají čemu?**
  - souřadnicím subjektů v novém prostoru s osami určenými vlastními vektory kovarianční matice (či matice korelačních koef.)
- **variabilita vysvětlená příslušnou komponentou odpovídá čemu?**
  - vlastním číslům
- **vlastní vektory seřazeny jak?**
  - podle vlastních hodnot (sestupně) $\Rightarrow$ vybráno prvních m komponent vyčerpávajících nejvíce variability původních dat
- **předpoklady?**
  - kvantitativní proměnné s normálním rozdělením

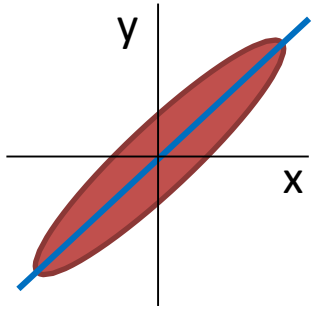
Analýza hlavních komponent – volba asociační matice

- **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)

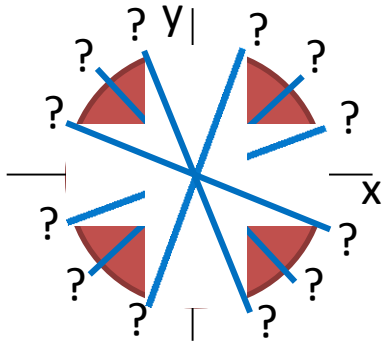
původní data



kovarianční matice (odečten průměr)



matice korelačních koeficientů (odečten průměr a podělení SD)



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

Analýza hlavních komponent – postup

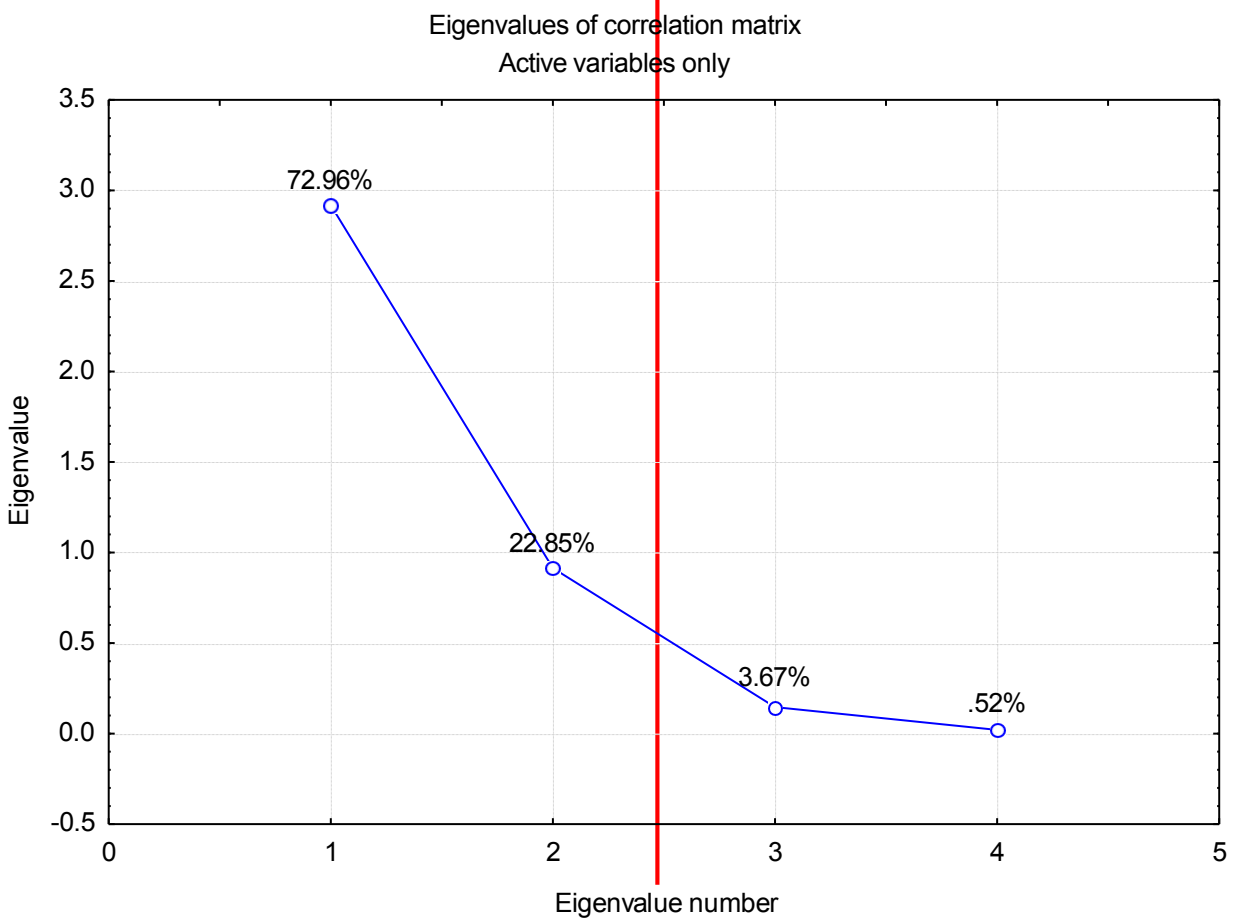
1. Volba asociační matice (kovarianční m. nebo m. korelačních koeficientů)
2. Výpočet vlastních čísel a vlastních vektorů asociační matice:
   - vlastní vektory definují směr nových faktorových os (hlavních komponent) v prostoru
   - vlastní čísla odrážejí variabilitu vysvětlenou příslušnou komponentou
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

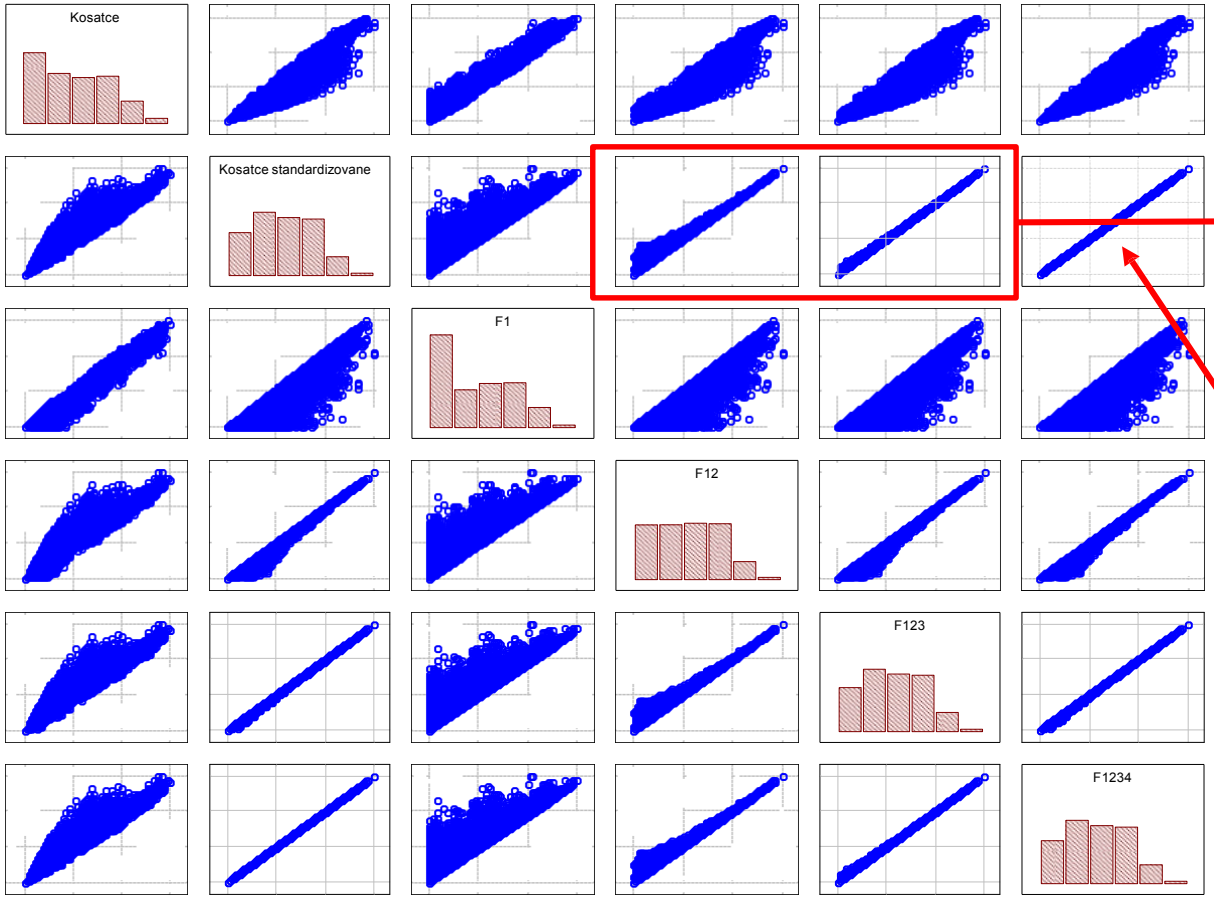
Sutinový graf (scree plot)

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 2 – řešení v softwaru SPSS

- **Zadání:** Proveďte PCA na objemech 6 mozkových struktur u 833 subjektů.
- **Řešení:** SPSS: Analyze – Dimension Reduction – Factor...
  - záložka Extraction:
    - volba metody (ponechat Principal components)
    - volba Correlation matrix či Covariance matrix (pozor, Correlation matrix je defaultní! tzn. přepnout na Covariance matrix)
    - možnost zatrhnout vykreslení Scree plotu
    - volba, kolik hlavních komponent se vytvoří (přepnout na Fixed number... a zvolit 6, když mám 6 vstupních proměnných)
  - záložka Rotation – ponechám zatržené „None“
  - záložka Scores... – zatrhnout „Save as variable“ a případně i zatrhnout „Display factor score coefficient matrix“

PCA – příklad 2 – řešení v softwaru R

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

```
library(readxl)
data <- read_excel('Data_neuro.xlsx',sheet="data")
data <- data[,24:29]   # vyber 6 promennych s objemy mozkovych struktur
pca <- prcomp(data)   # vypocet PCA s kovariancni matici; tzn. pouzito defaulni
center=TRUE a scale=FALSE; pro m. korel. koef. – prcomp(data,scale=TRUE)
pca$sdev^2   # vlastni cisla
```

```
> pca$sdev^2
[1] 403676.97 139067.09  70200.25  41840.70  40421.08  32737.94
```

```
pca$rotation   # vlastni vektory (ve sloupcich, serazene podle vlastnich cisel)
```

```
> pca$rotation
                              PC1         PC2         PC3          PC4          PC5         PC6
Hippocampus_volume (mm3) -0.035459125 0.88861834 -0.048506362  0.121740139  0.309258675 -0.31029927
Amygdala_volume (mm3)    -0.031283533 0.37476563 -0.095616471 -0.294217081 -0.866059128  0.11317002
Thalamus_volume (mm3)     0.001035499 0.10003061  0.986981343 -0.102255212 -0.021806247  0.07020677
Pallidum_volume (mm3)    -0.012014730 0.05596007 -0.104571564 -0.902442907  0.367642426  0.19032801
Putamen_volume (mm3)     -0.023074151 0.23311937 -0.058031628  0.271419287  0.136348899  0.92168098
Nucl_caud_volume (mm3)    0.998542011 0.04925323 -0.008340823 -0.009374972 -0.008553979  0.01604185
```

```
pca$x   # hlavni komponenty (tj. souradnice subjektu v novem prostoru)
```

```
> pca$x
              PC1          PC2          PC3          PC4          PC5          PC6
[1,] -5.416758e+02  322.0603857   90.54458062  -94.2298142  249.66114452  -27.3528609
[2,] -3.061072e+02  508.2458732 -423.53056436  204.0784644   40.59484197 -148.3389455
[3,]  2.180346e+02  473.6196500  192.81995921  163.2061839   82.36173899  128.0769292
[4,] -4.927048e+02  535.5032528 -267.88271465   74.2783108   56.03257012 -351.3861289
[5,] -3.463904e+02  240.7736931 -312.98274680  106.9214737    5.00591406   32.8322655
```

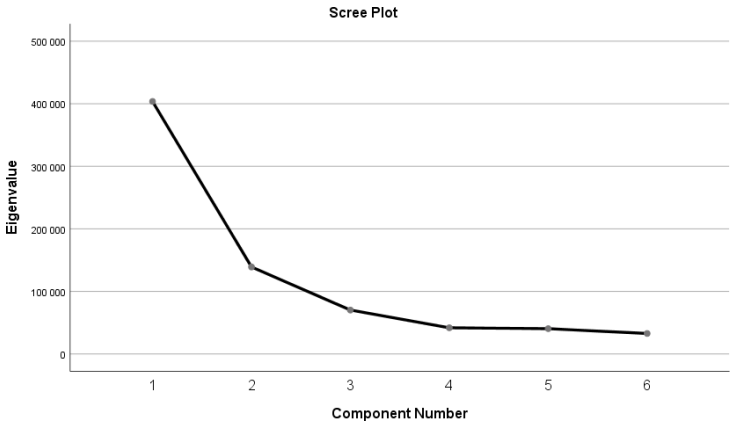
PCA – příklad 2 – řešení v softwaru SPSS

Vlastní čísla

Total Variance Explained

	Component	Initial Eigenvalues[a] Total	% of Variance	Cumulative %
Raw	1	403676,975	55,454	55,454
	2	139067,087	19,104	74,558
	3	70200,250	9,644	84,202
	4	41840,703	5,748	89,950
	5	40421,085	5,553	95,503
	6	32737,942	4,497	100,000

Sutinový graf



Matice vlastních vektorů *

Component Matrix[a]

	Raw Component 1	2	3	4	5	6
Hippocampus_volume (mm3)	-22,529	331,381	-12,852	-24,902	-62,176	-56,144
Amygdala_volume (mm3)	-19,876	139,756	-25,334	60,182	174,121	20,477
Thalamus_volume (mm3)	,658	37,303	261,504	20,916	4,384	12,703
Pallidum_volume (mm3)	-7,634	20,868	-27,707	184,595	-73,914	34,437
Putamen_volume (mm3)	-14,660	86,934	-15,376	-55,519	-27,413	166,766
Nucl_caud_volume (mm3)	634,429	18,367	-2,210	1,918	1,720	2,903

Extraction Method: Principal Component Analysis.

a. 6 components extracted.

Souřadnice subjektů v novém prostoru (jsou standardizované)

FAC1_1	FAC2_1	FAC3_1	FAC4_1	FAC5_1	FAC6_1
-,85256	,86362	,34174	,46067	-1,24179	-,15117
-,48179	1,36289	-1,59851	-,99769	-,20191	-,81984
,34317	1,27004	,72775	-,79788	-,40966	,70786
-,77548	1,43599	-1,01106	-,36313	-,27870	-1,94204
-,54519	,64565	-1,18128	-,52272	-,02490	,18146
-,19375	2,01086	-1,18890	-1,18152	,31479	-,25469

* normalizace vl. vektorů by se provedla v exelu (viz. slide 16)

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