

# Pokročilé metody analýzy dat v neurovědách



RNDr. Eva Janoušová  
doc. RNDr. Ladislav Dušek, Dr.

Jaro 2016

# Blok 5

## Ordinační analýzy I

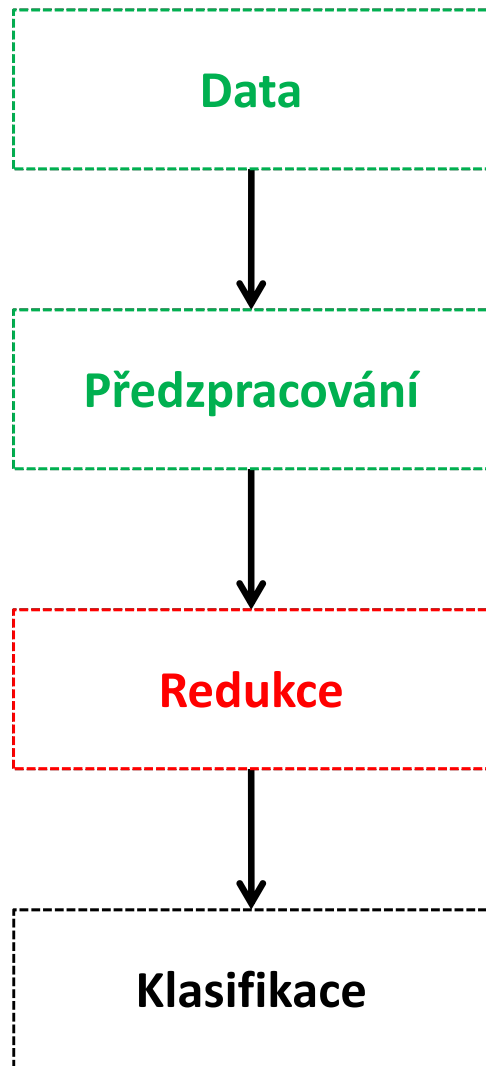
# Osnova

---

1. Principy redukce dimenzionality dat
2. Selekcce a extrakce proměnných
3. Analýza hlavních komponent (PCA)
4. Faktorová analýza (FA)

# Principy redukce dimenzionality dat

# Schéma analýzy a klasifikace dat



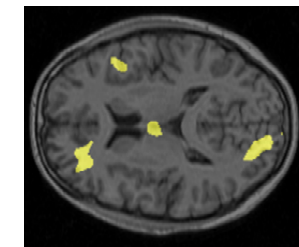
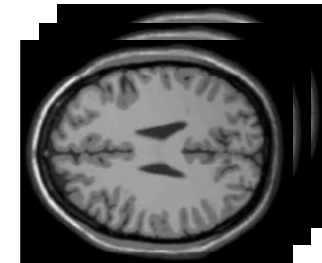
Ukázka - kognitivní data apod.

	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M		90
4	3	26	Z	178	70

	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M	167	90
4	3	26	Z	178	70

	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M	167	90
4	3	26	Z	178	70

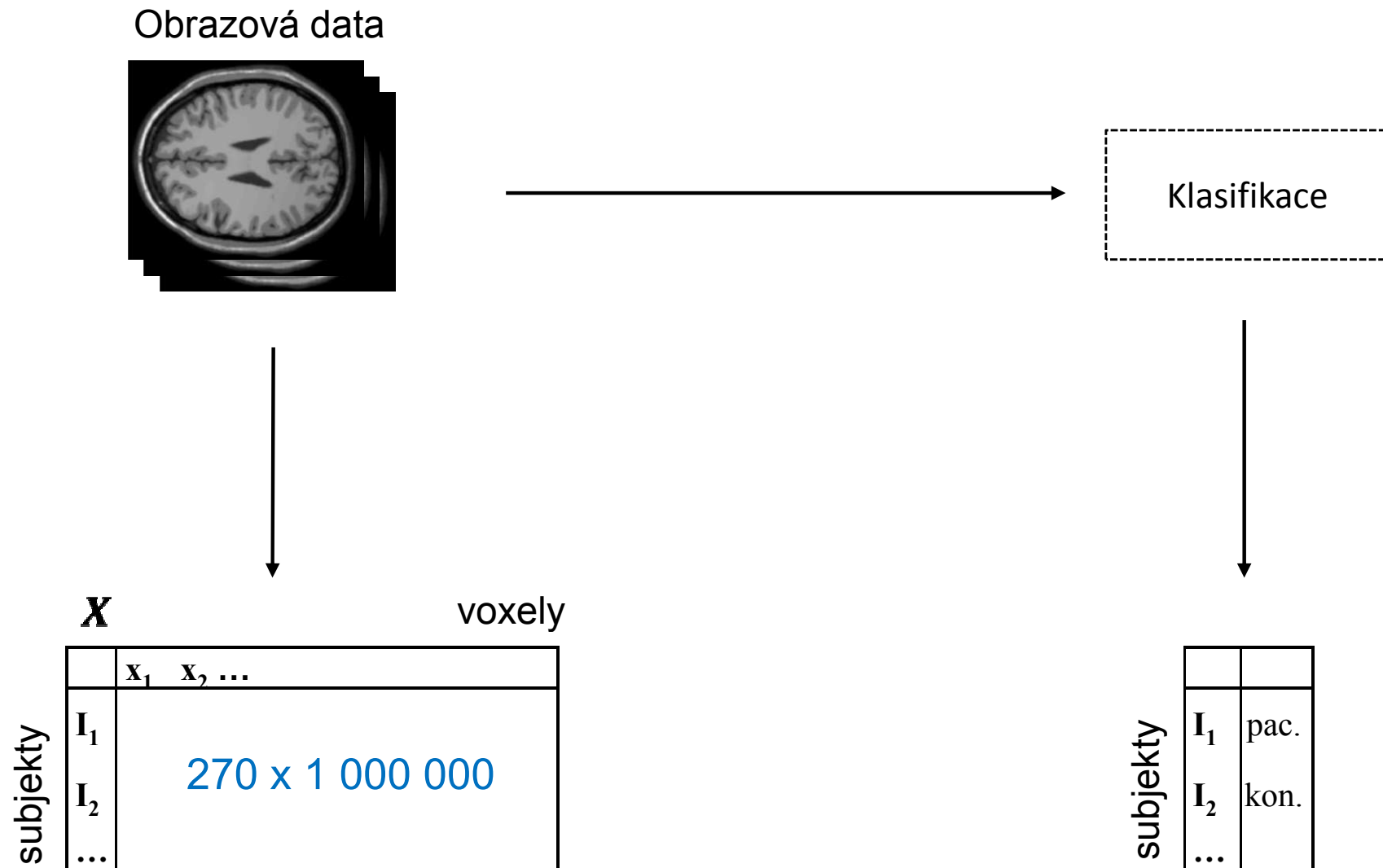
Ukázka - obrazová data



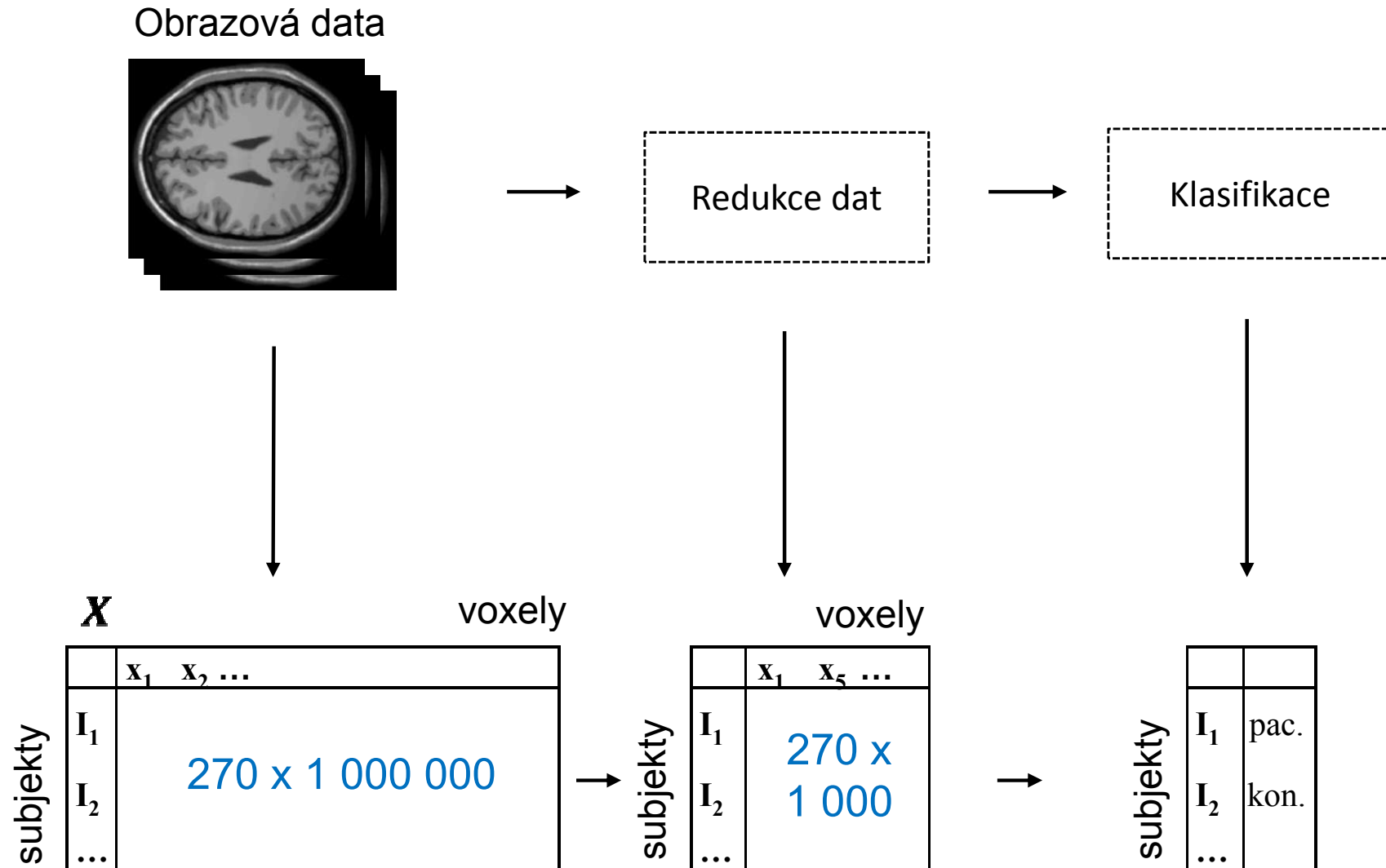
nebo



# Proč používat redukci dat?



# Proč používat redukci dat?



# Proč používat redukci dat?

---

- zjednodušení další práce s daty
- možnost použití metod analýzy dat, které by na původní data nebylo možno použít
- umožnění vizualizace vícerozměrných dat – může být nápomocné k nalezení vztahů v datech či k jejich interpretaci
- redukce dat může být i cílem analýzy (např. identifikace oblastí mozku, kde se nejvíce liší od sebe liší skupiny subjektů)



# Volba a výběr proměnných – úvod

- počáteční volba proměnných je z velké části empirická, vychází ze zkušeností získaných při empirické klasifikaci člověkem a závisí kromě rozboru podstaty problému i na technických (ekonomických) možnostech a schopnostech hodnoty proměnných určit
- kolik a jaké proměnné?
  - málo proměnných – možná nízká úspěšnost klasifikace či jiných následných analýz
  - moc proměnných – možná nepřiměřená pracnost, vysoké náklady

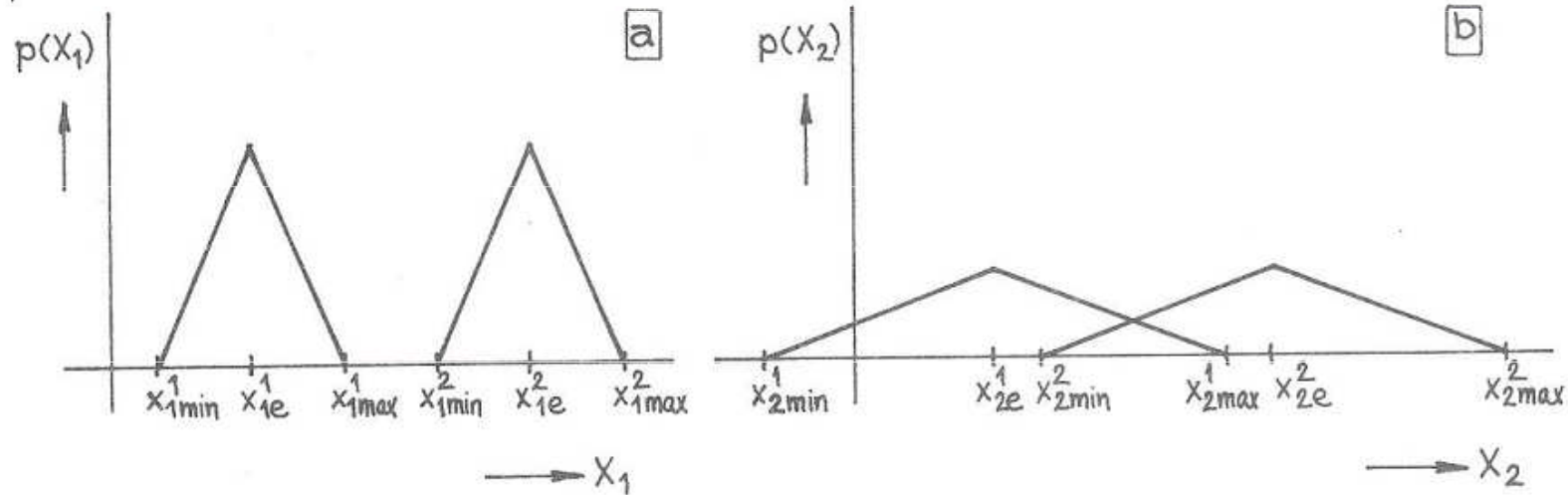


## **KOMPROMIS**

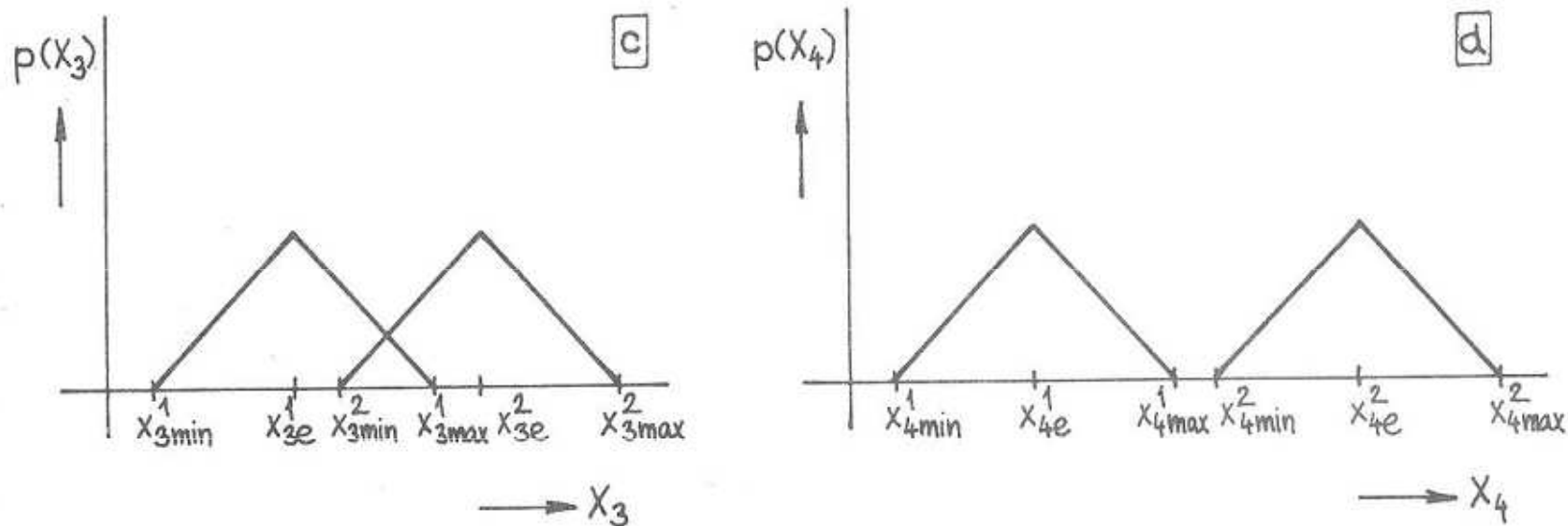
(určit ty proměnné, jejichž hodnoty nesou nejvíce informace z hlediska řešené úlohy, tj. např. ty proměnné, kterou jsou nejefektivnější pro vytvoření co nejoddělenějších klasifikačních tříd)

# Zásady pro volbu proměnných I

- výběr proměnných s minimálním rozptylem uvnitř tříd

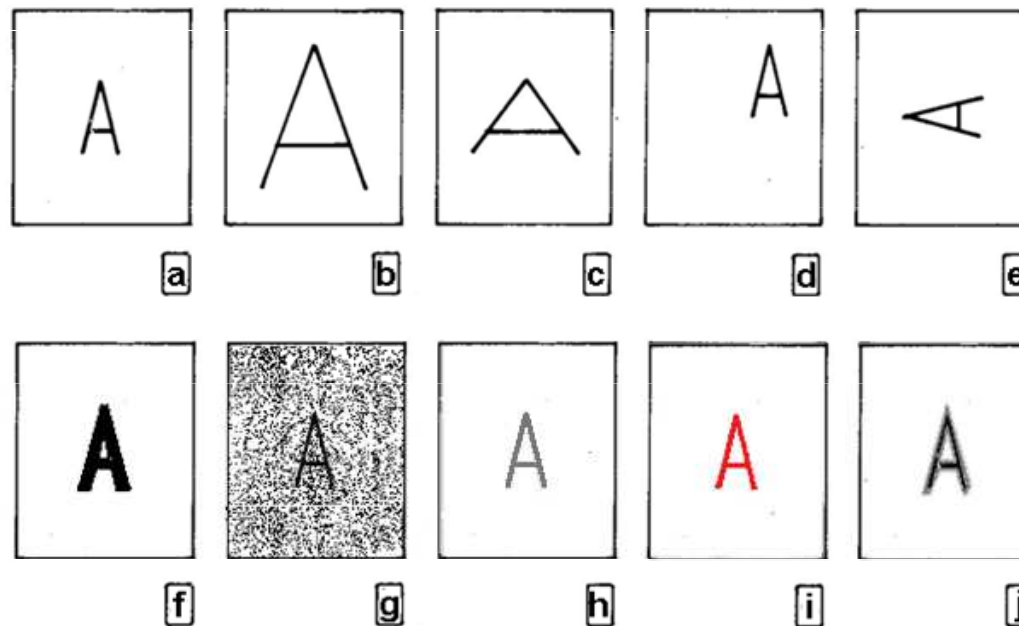


- výběr proměnných s maximální vzdáleností mezi třídami



# Zásady pro volbu proměnných II

- výběr vzájemně nekorelovaných proměnných
  - pokud jsou hodnoty jedné proměnné závislé na hodnotách druhé proměnné, pak použití obou těchto proměnných nepřináší žádnou další informaci – stačí jedna z nich, jedno která
- výběr proměnných invariantních vůči deformacím
  - volba elementů formálního popisu závisí na vlastnostech původních i předzpracovaných dat a může ovlivňovat způsob předzpracování



# Selekce a extrakce proměnných

# Selekce a extrakce proměnných

- popis objektu původně reprezentovaný  $p$  rozměrným vektorem se snažíme vyjádřit vektorem  $m$  rozměrným tak, aby množství diskriminační informace obsažené v původním vektoru bylo v co největší míře zachováno
- dva principiálně různé způsoby:
  1. **selekce** – výběr těch proměnných, které přispívají k separabilitě klasifikačních tříd nejvíce

proměnné

		$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	...
subjekty	$I_1$									
	$I_2$	pac.								
	$I_3$	pac.								
	...	kont.								

2. **extrakce** – transformace původních proměnných na menší počet jiných proměnných (které zpravidla nelze přímo měřit a často nemají zcela jasnou interpretaci)

proměnné

		$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	...
subjekty	$I_1$									
	$I_2$	pac.								
	$I_3$	pac.								
	...	kont.								

➔

		$y_1$	$y_2$	$y_3$	$y_4$
subjekty	$I_1$				
	$I_2$				
	$I_3$				
	...				

# Selekce proměnných

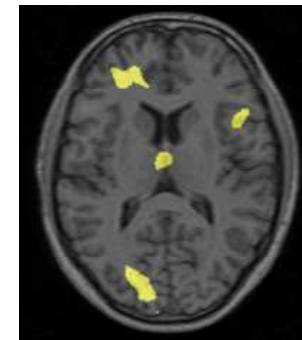
---

- cílem je výběr proměnných, které jsou nejužitečnější pro další analýzu (např. při klasifikaci výběr takových proměnných, které nejlépe od sebe dokáží oddělit skupiny subjektů/objektů)
- metod selekce je velké množství, nepoužívanější metody jsou:
  - výběr proměnných na základě statistických testů
  - výběr oblastí mozku (ROI) podle atlasu
  - algoritmy sekvenční selekce (dopředné či zpětné nebo algoritmus plus  $p$  minus  $q$ )

# Výběr proměnných na základě statistických testů

**Princip:** Výběr statisticky významných proměnných pomocí dvouvýběrového t-testu či Mannova-Whitneyova testu.

		proměnné								
		$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	...
subjekty	$I_1$									
	$I_2$	pac.								
	$I_3$	pac.								
	$I_4$	kont.								
	$I_5$	pac.								
	...	kont.								
p-hodnoty:		0,34	0,02	0,09	0,01	0,25	0,63	0,03	0,12	



## Výhody:

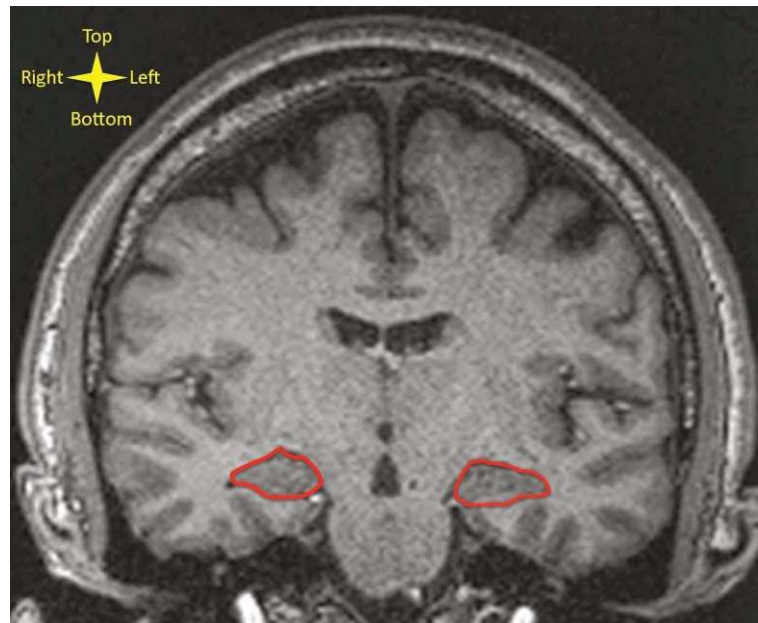
- + rychlé
- + u obrazů mozku výhodou, že je analýza provedena na celém mozku

## Nevýhody:

- jednorozměrná metoda (výběr proměnných bez ohledu na ostatní proměnné)
- potřeba použít metody korekce pro mnohonásobné testování (např. FDR)

# Výběr oblastí mozku (ROI) podle atlasu

**Princip:** Výběr oblastí mozku s využitím atlasu mozku podle expertní znalosti daného onemocnění (tzn. výběr oblasti postižené danou nemocí).



## Výhody:

- + anatomicky/funkčně relevantní – snadnější interpretace
- + zpravidla rychlé

## Nevýhody:

- ne vždy dopředu víme, která z oblastí je vhodná pro odlišení skupin osob
- některá onemocnění postihují celý mozek (např. schizofrenie)



# Algoritmy sekvenční selekce

- algoritmus sekvenční dopředné selekce:
  - algoritmus začíná s prázdnou množinou, do které se vloží proměnná s nejlepší hodnotou selekčního kritéria
  - v každém následujícím kroku se přidá ta proměnná, která s dříve vybranými veličinami dosáhla nejlepší hodnoty kritéria
- algoritmus sekvenční zpětné selekce:
  - algoritmus začíná s množinou všech proměnných
  - v každém následujícím kroku se eliminuje ta proměnná, která způsobuje nejmenší pokles kritériální funkce

**Výhody :** + dopředný algoritmus je výpočetně jednodušší, protože pracuje maximálně v  $n$ -rozměrném prostoru  
+ zpětný algoritmus umožňuje průběžně sledovat množství ztracené informace

**Nevýhody :** - dopředná selekce – nelze vyloučit ty veličiny, které se staly nadbytečné po přiřazení dalších veličin  
- zpětná selekce – neexistuje možnost opravy při neoptimálním vyloučení kterékoliv proměnné

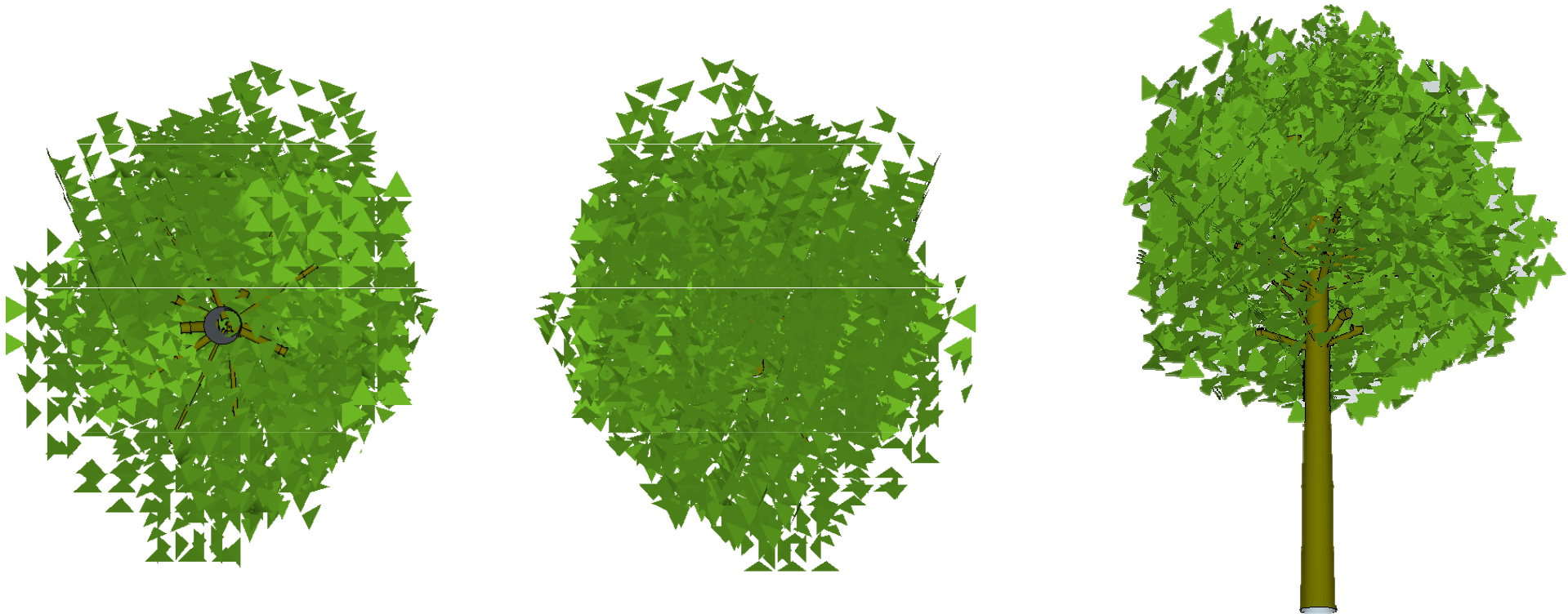
- algoritmus plus  $p$  mínus  $q$ :
  - po přidání  $p$  veličin se  $q$  veličin odstraní;
  - proces probíhá, dokud se nedosáhne požadovaného počtu příznaků

# Extrakce proměnných

- jednou z možných přístupů redukce dat
- transformace původních proměnných na menší počet jiných proměnných  
⇒ tzn. hledání (optimálního) zobrazení  $Z$ , které transformuje původní  $p$ -rozměrný prostor (obraz) na prostor (obraz)  $m$ -rozměrný ( $p \geq m$ )
- pro snadnější řešitelnost hledáme zobrazení  $Z$  v oboru lineárních zobrazení
- metody extrakce proměnných:
  - analýza hlavních komponent (PCA)
  - faktorová analýza (FA)
  - analýza nezávislých komponent (ICA)
  - korespondenční analýza (CA)
  - vícerozměrné škálování (MDS)
  - redundanční analýza (RDA)
  - kanonická korelační analýza (CCorA)
  - manifold learning metody (LLE, Isomap atd.)
  - metoda parciálních nejmenších čtverců (PLS)
- metody extrakce proměnných často nazývány jako metody ordinační analýzy

# Ordinační analýza dat = pohled ze správného úhlu

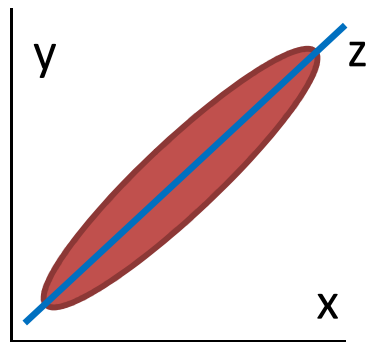
- Vícerozměrná analýza nám pomáhá nalézt v x-dimenzionálním prostoru nejvhodnější pohled na data poskytující maximum informací o analyzovaných objektech



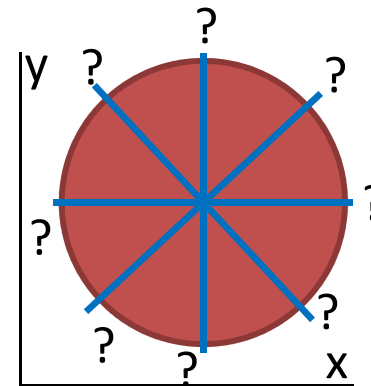
Všechny obrázky ukazují stejný objekt z různých úhlů v 3D prostoru.

# Obecný princip redukce dimenzionality dat pomocí extrakce

- V převážné většině případů existují mezi dimenzemi korelační vztahy, tedy dimenze se navzájem vysvětlují a pro popis kompletní informace v datech není třeba všech dimenzí vstupního souboru
- Všechny tzv. ordinační metody využívají principu identifikace korelovaných dimenzí a jejich sloučení do souhrnných nových dimenzí zastupujících několik dimenzí vstupního souboru
- Pokud mezi dimenzemi vstupního souboru neexistují korelace, nemá smysl hledat zjednodušení vícerozměrné struktury takového souboru !!!



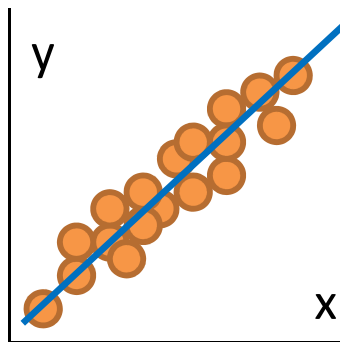
Jednoznačný vztah dimenzí x a y umožňuje jejich nahrazení jednou novou dimenzí z



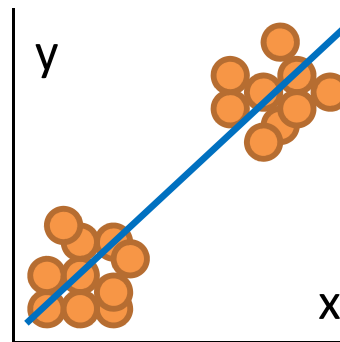
V případě neexistence vztahu mezi x a y nemá smysl definovat nové dimenze – nepřináší žádnou novou informaci oproti x a y

# Korelace jako princip výpočtu vícerozměrných analýz

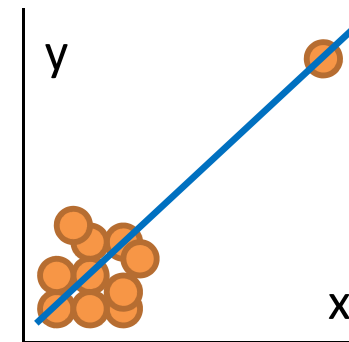
- Kovariance a Pearsonova korelace je základem analýzy hlavních komponent, faktorové analýzy jakož i dalších vícerozměrných analýz pracujících s lineární závislostí proměnných
- Předpokladem výpočtu kovariance a Pearsonovy korelace je:
  - Normalita dat v obou dimenzích
  - Linearita vztahu proměnných
- Pro vícerozměrné analýzy je nejzávažnějším problémem přítomnost odlehlých hodnot



Lineární vztah –  
bezproblémové použití  
Pearsonovy korelace



Korelace je dána 2 skupinami  
hodnot – vede k identifikaci  
skupin objektů v datech



Korelace je dána odlehlou  
hodnotu – analýza popisuje  
pouze vliv odlehlé hodnoty

# Typy ordinační analýzy

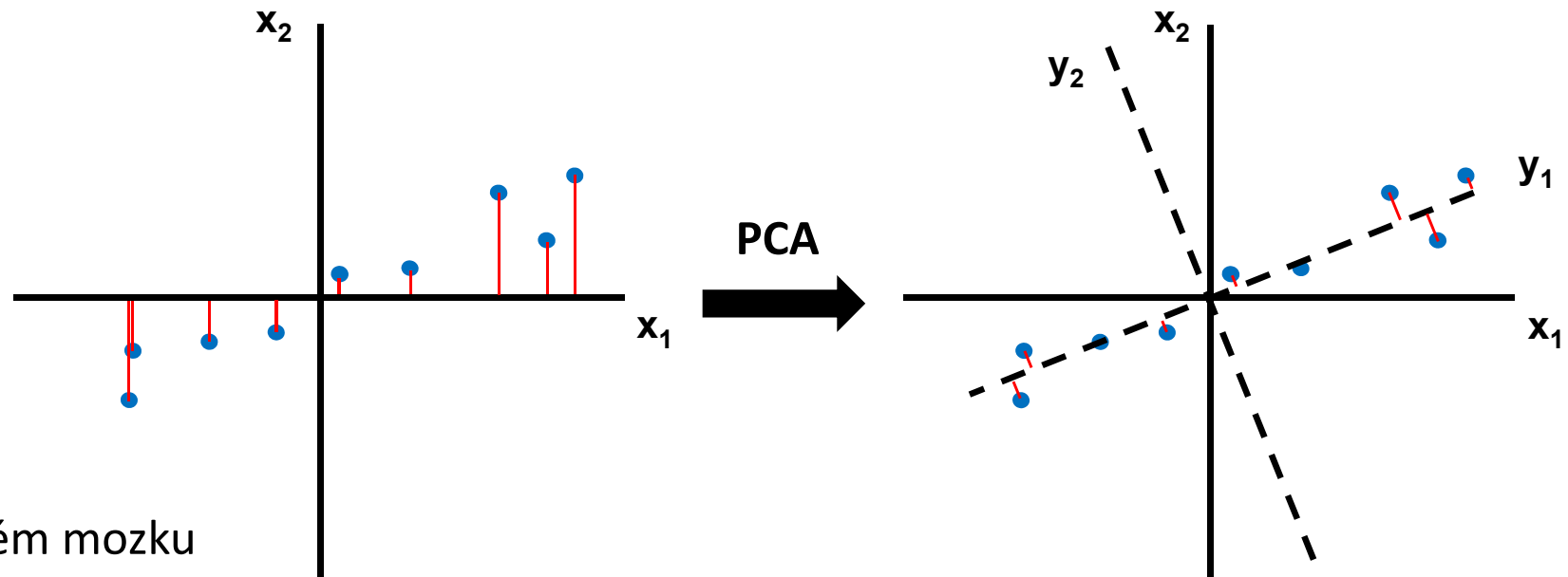
- Ordinačních analýz existuje celá řada, některé jsou spjaty s konkrétními metrikami vzdáleností/podobností
- V přehledu jsou uvedeny pouze základní typy analýz, nikoliv jejich různé kombinace hodnotící vztahy dvou a více sad proměnných (CCA, kanonická korelace, RDA, co-coordinate analysis, co-inertia analysis, diskriminační analýza apod.)

Typ analýzy	Vstupní data	Metrika
Analýza hlavních komponent (PCA)	NxP matice	Korelace, kovariance, Euklidovská
Faktorová analýza (FA)	NxP matice	Korelace, kovariance, Euklidovská
Analýza nezávislých komponent (ICA)	NxP matice	Korelace, kovariance, Euklidovská
Korespondenční analýza (CA)	NxP matice	Chi-square vzdálenost
Analýza hlavních koordinát (PCoA)	Asoc. matice	libovolná
Nemetrické mnohorozměrné škálování (MDS)	Asoc. matice	libovolná

# Analýza hlavních komponent (PCA)

# Analýza hlavních komponent

- anglicky Principal Component Analysis (PCA)
- snaha redukovat počet proměnných nalezením nových latentních proměnných (hlavních komponent) vysvětlujících co nejvíce variability původních proměnných
- nové proměnné ( $X_1, X_2$ ) lineární kombinací původních proměnných ( $Y_1, Y_2$ )



## Výhody:

- + analýza na celém mozku
- + vícerozměrná metoda

## Nevýhody:

- nevyužívá informaci o příslušnosti subjektů do skupin
- potřebné určit, kolik hlavních komponent se použije pro transformaci



# Analýza hlavních komponent – cíle

---

- Popis a vizualizace vztahů mezi proměnnými
- Výběr neredundantních proměnných pro další analýzy
- Vytvoření zástupných faktorových os pro použití v dalších analýzách
- Identifikace shluků v datech spjatých s variabilitou dat
- Identifikace vícerozměrně odlehlých objektů

# Analýza hlavních komponent – předpoklady

- vstupem do analýzy datová matice  $n \times p$  obsahující kvantitativní proměnné (s normálním rozdělením)
- předpoklady obdobné jako při výpočtu korelací a kovariancí:
  - nepřítomnost odlehlých hodnot (s výjimkou situace, kdy analýzu provádíme za účelem identifikace odlehlých hodnot)
  - nepřítomnost více skupin objektů (s výjimkou situace, kdy analýzu provádíme za účelem detekce přirozeně existujících shluků spjatých s největší variabilitou souboru)
- datový soubor by měl mít více objektů než proměnných, pro získání stabilních výsledků se doporučuje alespoň 10x tolik objektů než proměnných, ideální je 40-60x více objektů než proměnných

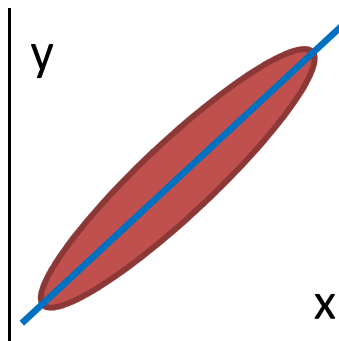
# Analýza hlavních komponent – volba asociační matice

- **autokorelační matice** – data nejsou nijak upravena (zohledňována průměrná hodnota i rozptyl původních dat)
- **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
- **každou úpravou původních dat ale přicházíme o určitou informaci !!!**

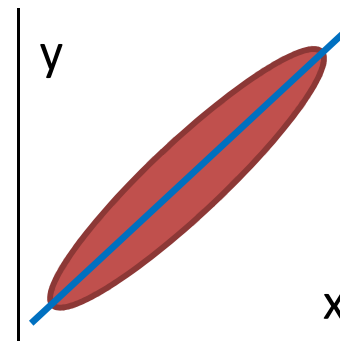
# Analýza hlavních komponent – volba asociační matice

- s jakými daty PCA pracuje v případě použití různých asociačních matic:

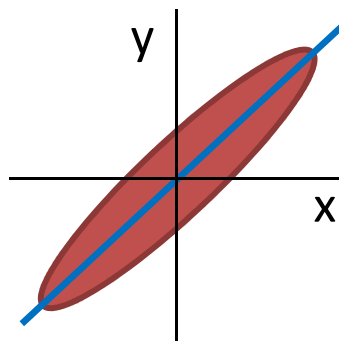
původní data



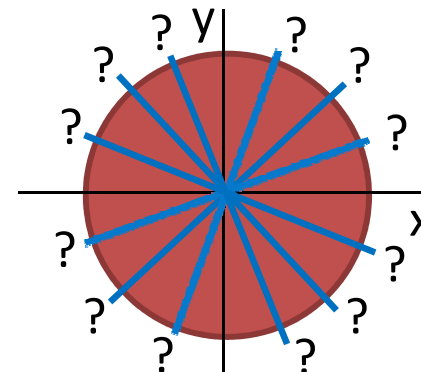
autokorelační matice  
(data nijak neupravována)



kovarianční matice  
(odečten průměr)



matice korelačních koeficientů  
(odečten průměr a podělení SD)



# Analýza hlavních komponent – postup

---

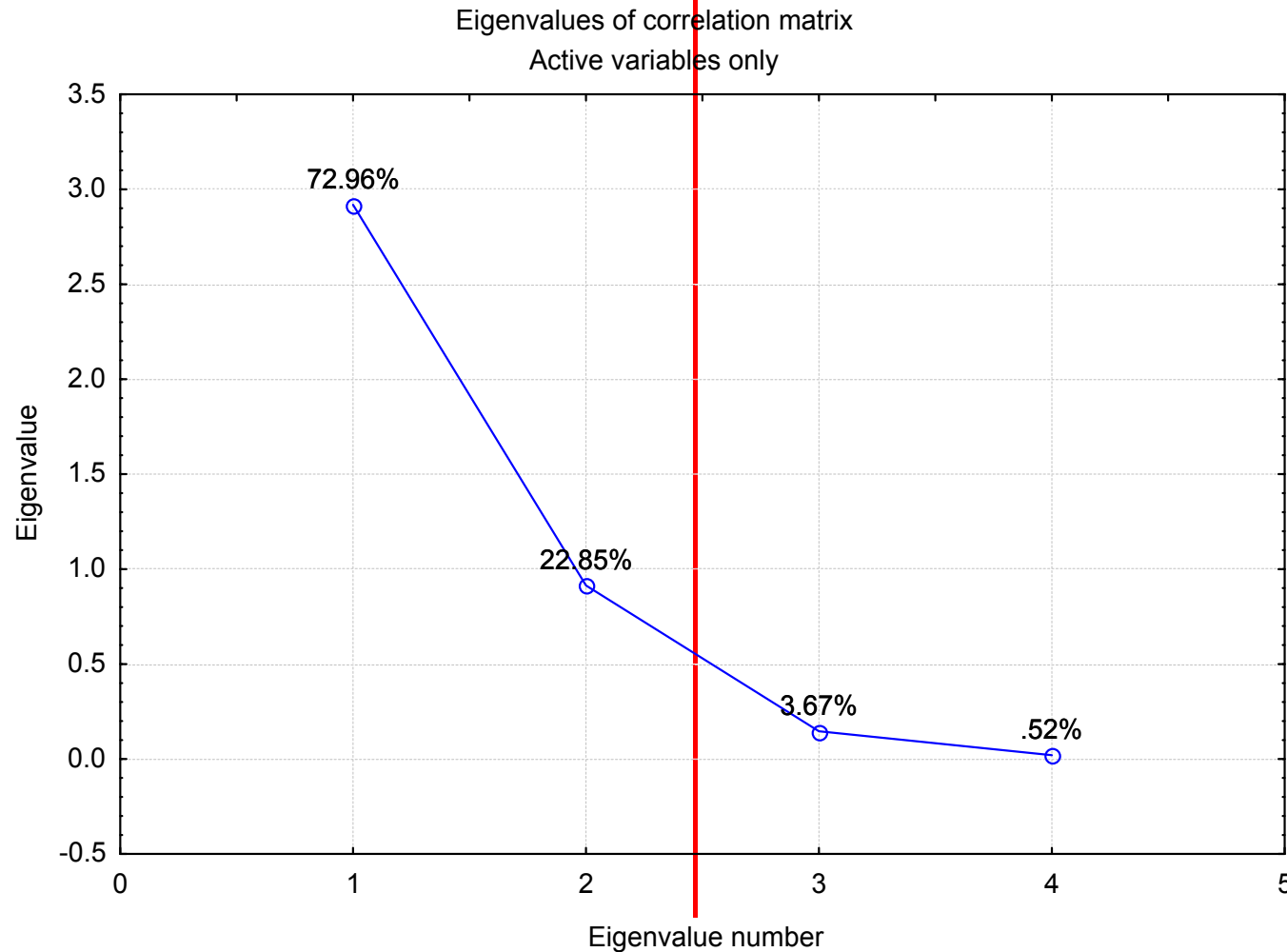
1. Volba asociační matice (autokorelační, kovarianční nebo kor. koeficientů)
2. Výpočet vlastních čísel a vlastních vektorů asociační matice:
  - vlastní vektory definují směr nových faktorových os (hlavních komponent) v prostoru
  - vlastní čísla odrážejí variabilitu vysvětlenou příslušnou komponentou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

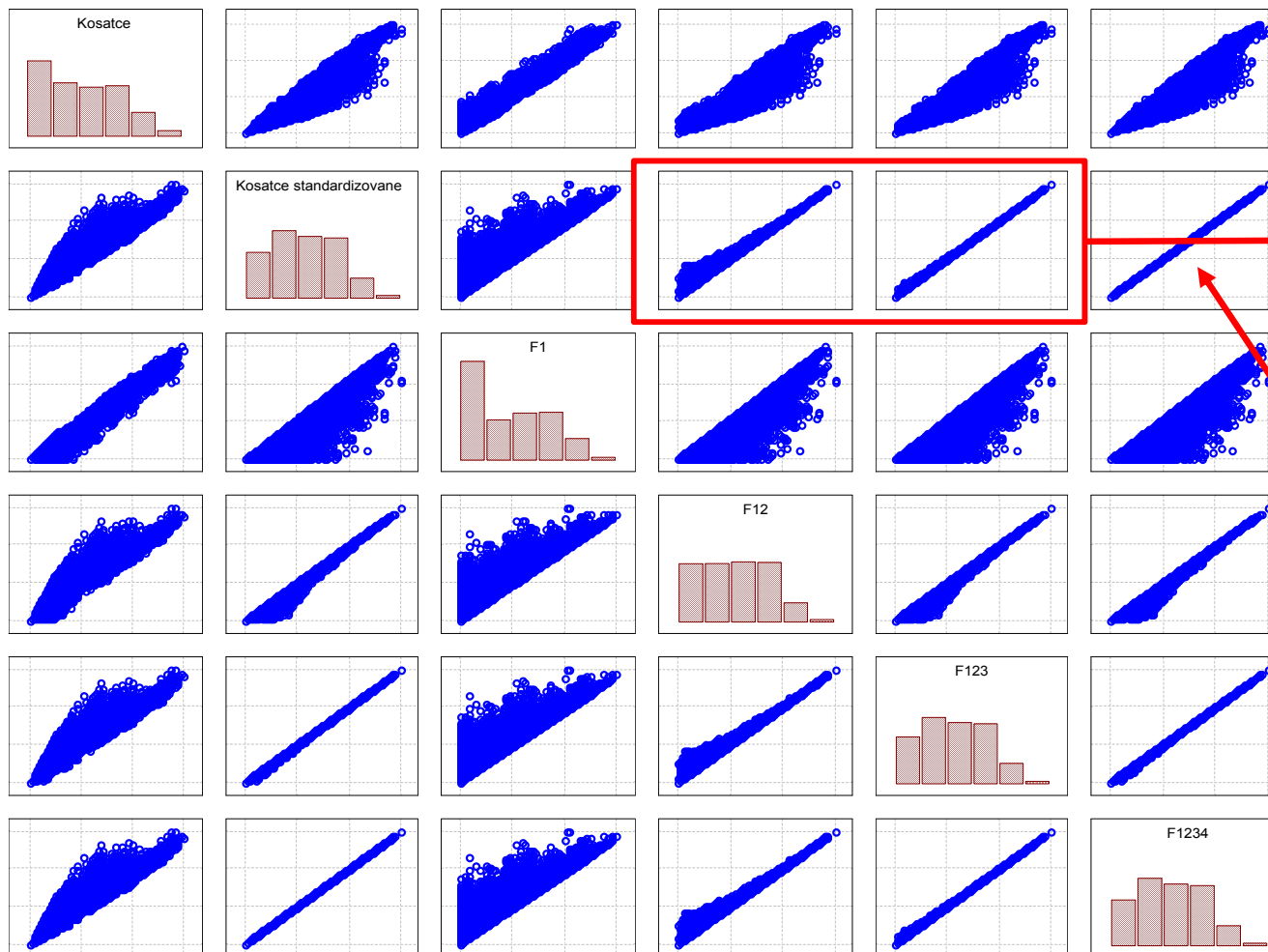
# Sutinový graf (scree plot)

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)

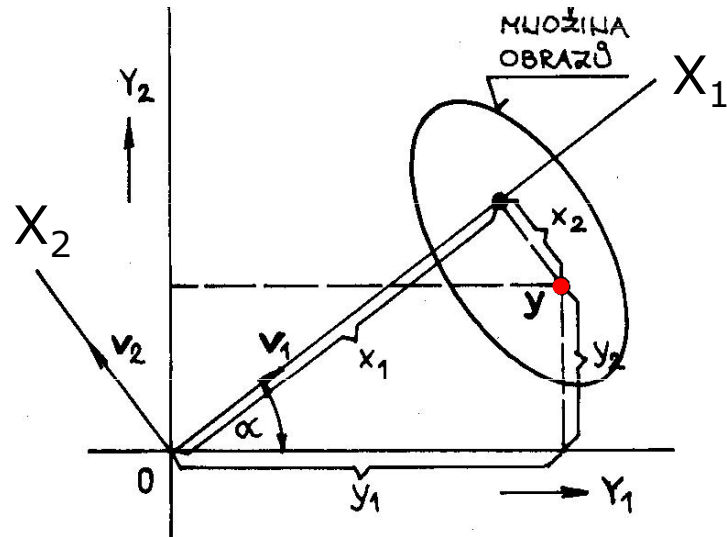


Za optimální z hlediska zachování vzdáleností objektů lze považovat dvě nebo tři dimenze

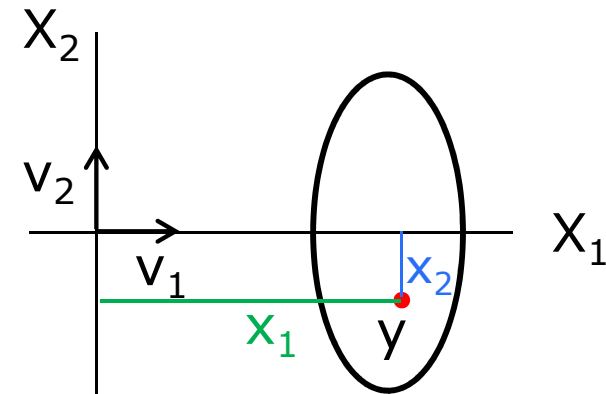
Při použití všech dimenzí jsou vzdálenosti perfektně zachovány



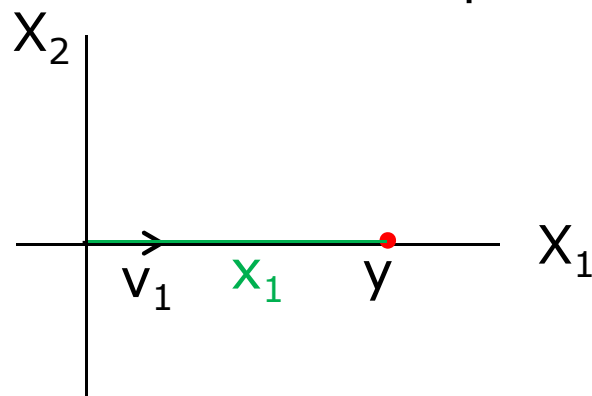
# PCA – geometrická interpretace



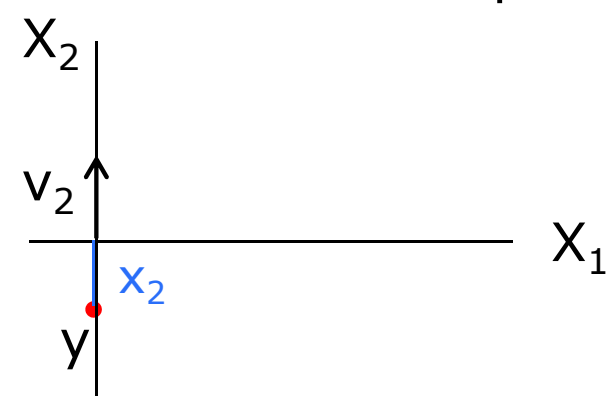
použití obou hlavních komponent



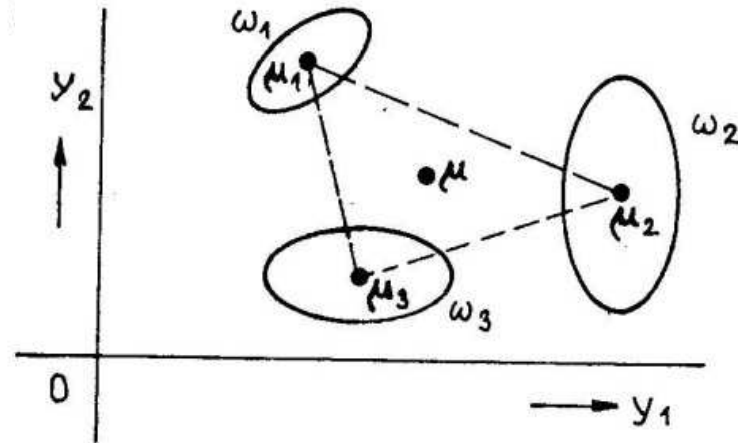
použití 1. hlavní komponenty



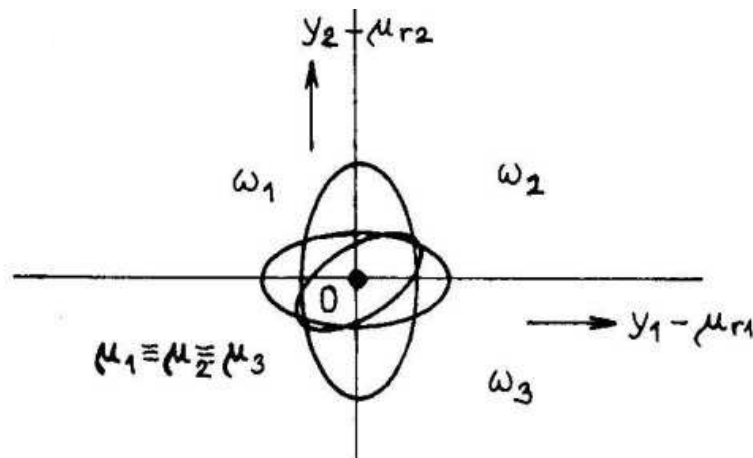
použití 2. hlavní komponenty



# PCA – rozdělení do tříd

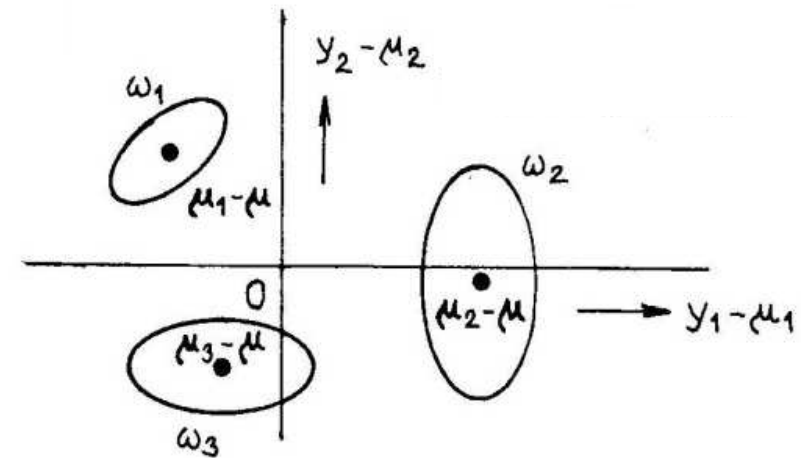


odečtení průměru každé skupiny zvlášť



→ není vhodné

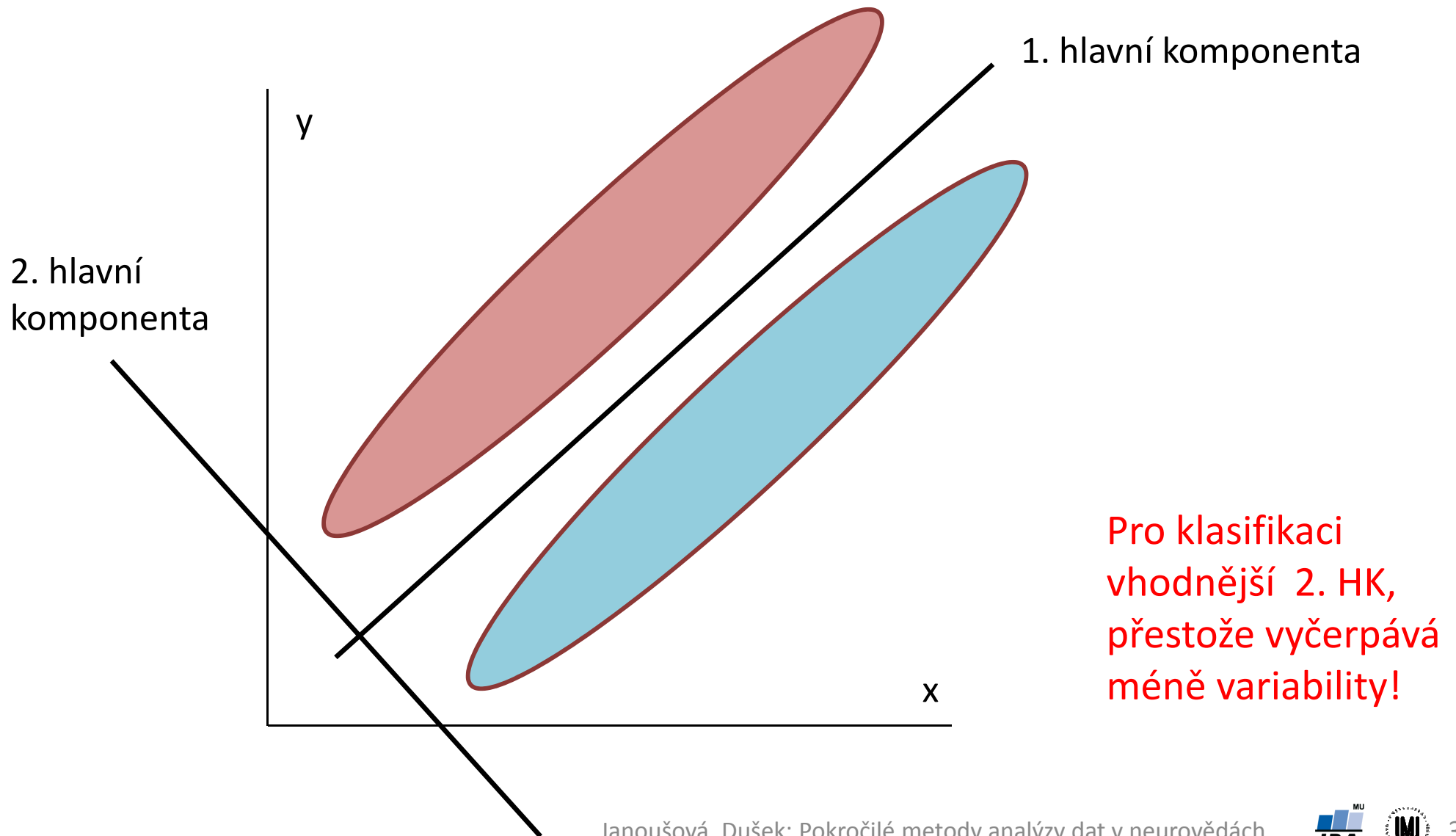
odečtení celkového průměru



→ je vhodné

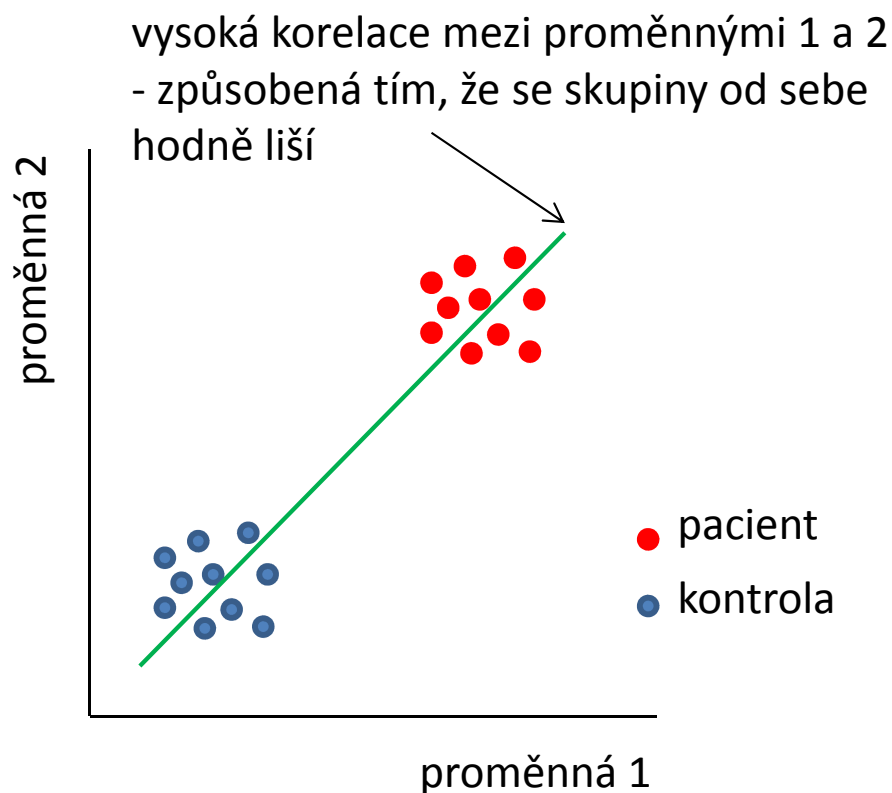
# PCA a klasifikace I

- PCA často nebývá vhodnou metodou redukce dat před klasifikací

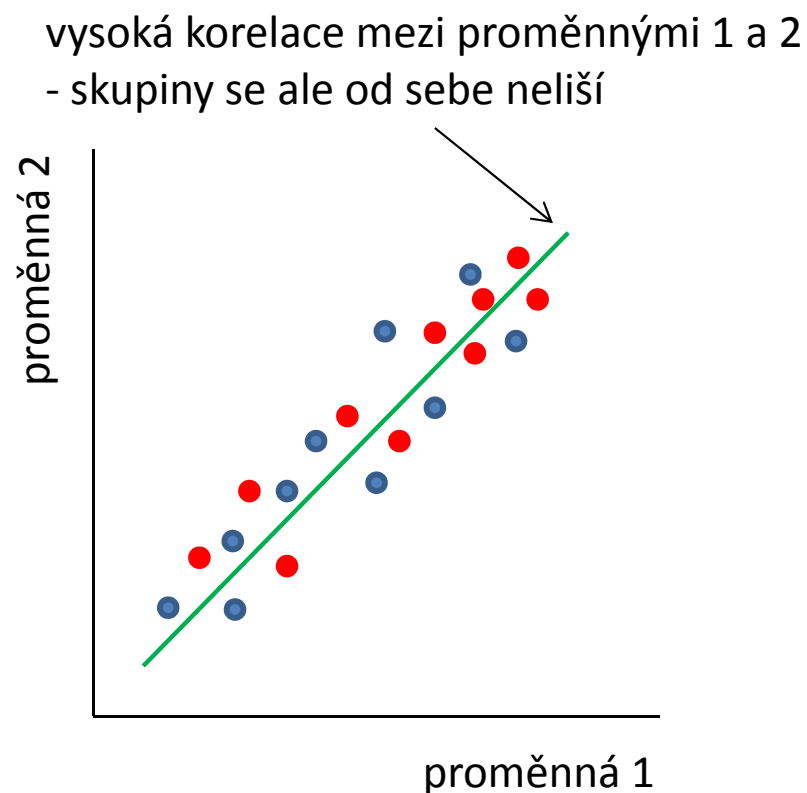


# PCA a klasifikace II

Když hlavní komponenta vyčerpává hodně variability, neznamená to, že musí rovněž dobře klasifikovat



→ v tomto případě obě proměnné budou korelovat s první hlavní komponentou a dokáží dobře diskriminovat pacienty a kontroly



→ v tomto případě obě proměnné budou také korelovat s první hlavní komponentou, ale nedokáží diskriminovat pacienty a kontroly

# PCA – rozšiřující poznatky I

**Výpočet PCA, když je počet proměnných mnohem větší než počet subjektů:**

- 1. způsob: iterativní postupný výpočet vlastních vektorů a vlastních čísel
- 2. způsob: pPCA - výpočet vlastních vektorů  $\mathbf{v}_i$  „velké“ kovarianční matice (proměnných)  $\mathbf{X}^T\mathbf{X}_{(p,p)}$  z vlastních vektorů  $\mathbf{w}_i$  „malé“ kovarianční matice (subjektů)

$\mathbf{X}\mathbf{X}^T_{(n,n)}$  pomocí: 
$$\mathbf{v}_i = \frac{\mathbf{X}^T \mathbf{w}_i}{\sqrt{\lambda_i(n-1)}}$$

**Datová matice:**

		proměnné		
		V1	V2	...
subjekty	S1	173 x 1 923 207		
	S2			
	...			



**Kovarianční matice subjektů:**

		subjekty		
		S1	S2	...
subjekty	S1	173 x 173		
	S2			
	...			



**Kovarianční matice proměnných:**

		proměnné		
		V1	V2	...
proměnné	V1	1 923 207 x 1 923 207		
	V2			
	...			

# PCA – rozšiřující poznatky II

Souvislost se singulárním rozkladem (SVD – Singular Value Decomposition):

$$\mathbf{X}_{(n,p)} = \mathbf{U}_{(n,k)} \mathbf{\Gamma}_{(k,k)} \mathbf{V}_{(k,p)}^T$$

- matice  $\mathbf{U}$  a  $\mathbf{V}$  jsou ortogonální a normované (ortonormální)
- matice  $\mathbf{U}$  složena z vlastních (charakteristických) vektorů matice  $\mathbf{X}\mathbf{X}^T_{(n,n)}$
- matice  $\mathbf{V}$  z vlastních vektorů matice  $\mathbf{X}^T\mathbf{X}_{(p,p)}$
- Matice  $\mathbf{\Gamma}$  je typu  $k \times k$  a její diagonála je tvořena singulárními hodnotami, které jsou na hlavní diagonále uspořádány podle klesající velikosti a které jsou rovny odmocninám vlastních čísel matice  $\mathbf{X}\mathbf{X}^T$  i  $\mathbf{X}^T\mathbf{X}$

# PCA – příklad – 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



	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

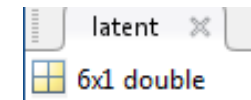
Matice vlastních vektorů



	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

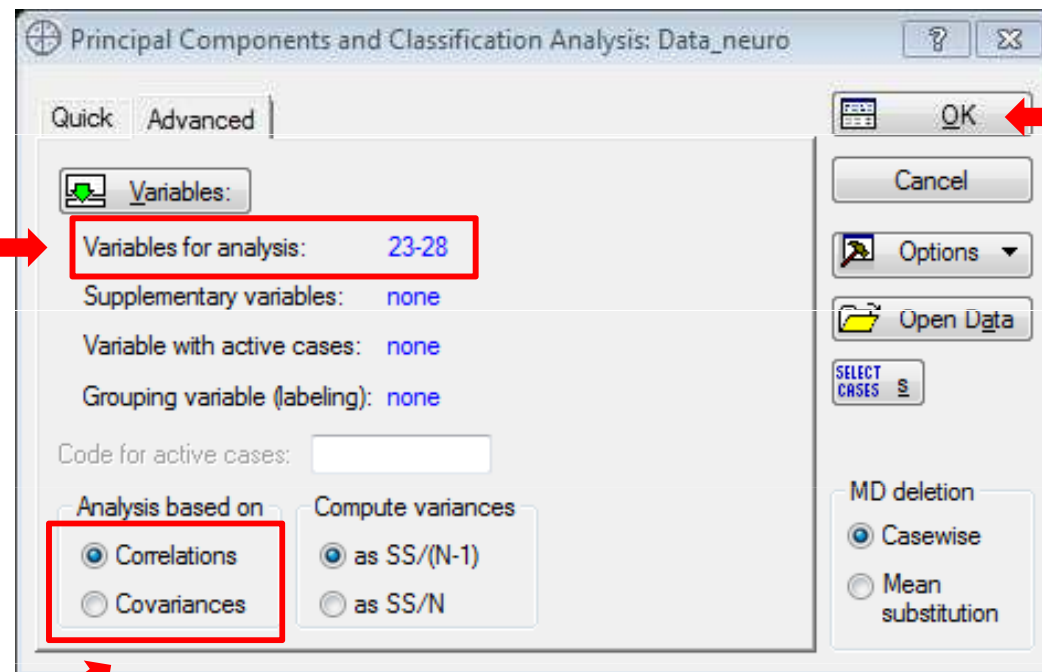


	1
1	4.0368e+05
2	1.3907e+05
3	7.0200e+04
4	4.1841e+04
5	4.0421e+04
6	3.2738e+04

# PCA – příklad – 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

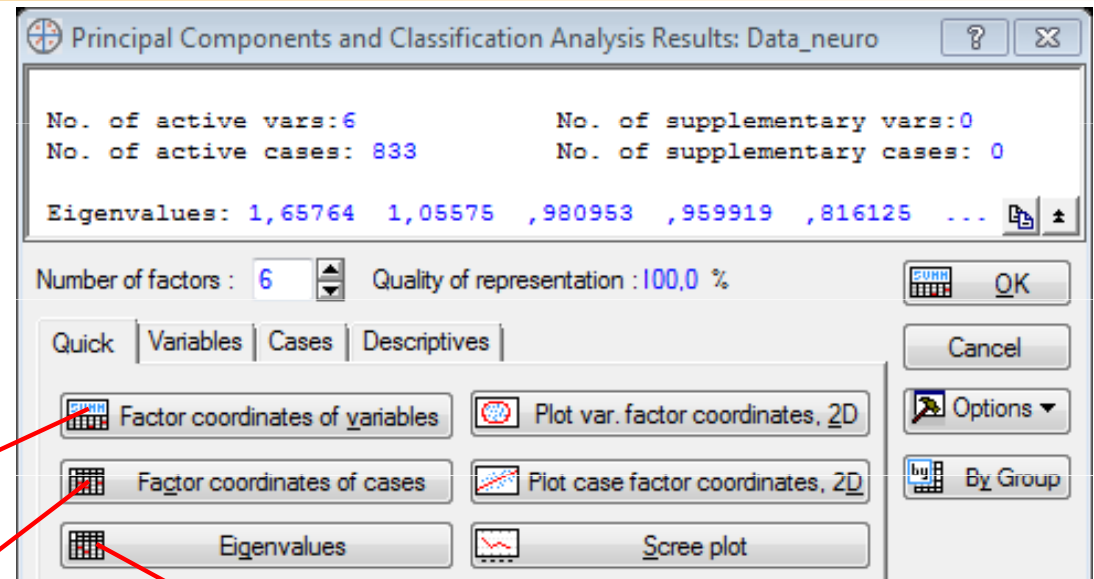
vybrat proměnné



zvolit, zda se má počítat  
kovarianční či korelační  
matice



# PCA – příklad – řešení v softwaru Statistica II



Matice vlastních vektorů

Variable	Factor coordinates of the variables, based on correlations (Data_neuro)					
	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	12,7090
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

Vlastní čísla

Eigenvalues of covariance matrix  
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

Souřadnice  
subjektů v  
novém  
prostoru

Case	Factor coordinates of cases, based on covariances (Data_neuro)					
	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

# PCA – příklad – 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

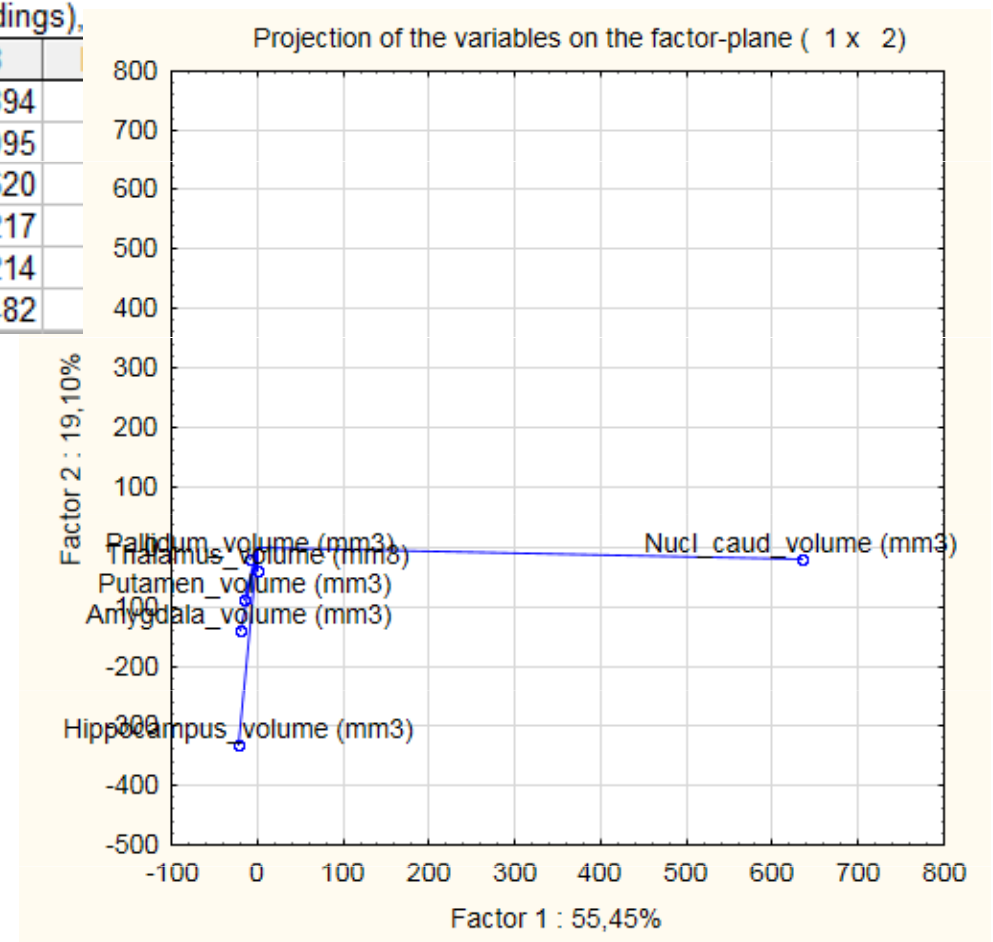
# PCA – příklad – řešení v softwaru Statistica IV

Záložka Variables:

Factor & variable correlations

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

Plot var. factor coordinates, 2D



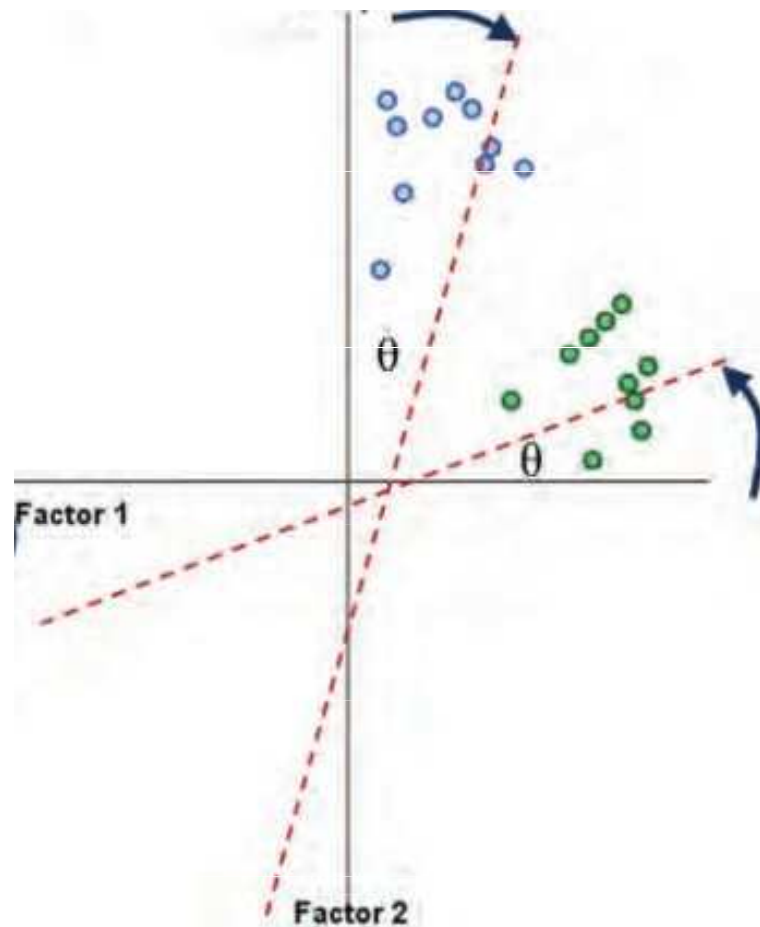
Z výsledků vyplývá, že:

- 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

# Faktorová analýza (FA)

# Faktorová analýza (FA)

- **Anglicky:** Factor Analysis
- **Princip:** Vytvoření nových proměnných (komponent, faktorů) z původních proměnných tak, aby zůstalo zachováno co nejvíce kovariance.



Stejný postup jako u PCA  
+ 1 krok navíc – rotace komponent

## Výhoda oproti PCA:

+ lepší interpretace nových proměnných

## Nevýhoda oproti PCA:

- prostor pro subjektivní názor analytika při výběru rotace

# Faktorová analýza

- faktorová analýza se snaží vysvětlit strukturu dat pomocí tzv. společných faktorů vysvětlujících sadu původních proměnných
- cíle, předpoklady, vstupní data a většina výpočtů obdobná jako u analýzy hlavních komponent
- čím se principiálně liší od analýzy hlavních komponent?
  - Analýza hlavních komponent – vysvětlení maxima variability v datech
  - Faktorová analýza – vysvětlení maxima kovariance mezi popisnými proměnnými
- čím se prakticky liší od analýzy hlavních komponent?
  - Hlavním praktickým rozdílem je rotace proměnných tak, aby se vytvořené faktorové osy daly dobře interpretovat
  - Výhodou je lepší interpretace vztahu původních proměnných
  - Nevýhodou je prostor pro subjektivní názor analytika daný výběrem rotace
- typy faktorové analýzy
  - Vysvětlující (Explanatory) – snaží se identifikovat minimální počet faktorů pro vysvětlení dat
  - Potvrzující (Confirmatory) – testuje hypotézy ohledně skryté struktury v datech

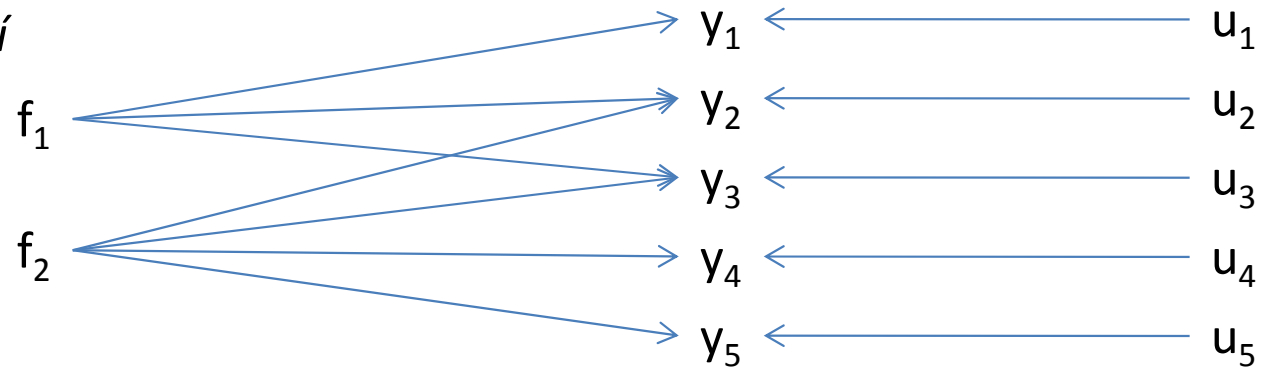
# Společné faktory a základní možné rotace

*Společný faktor*

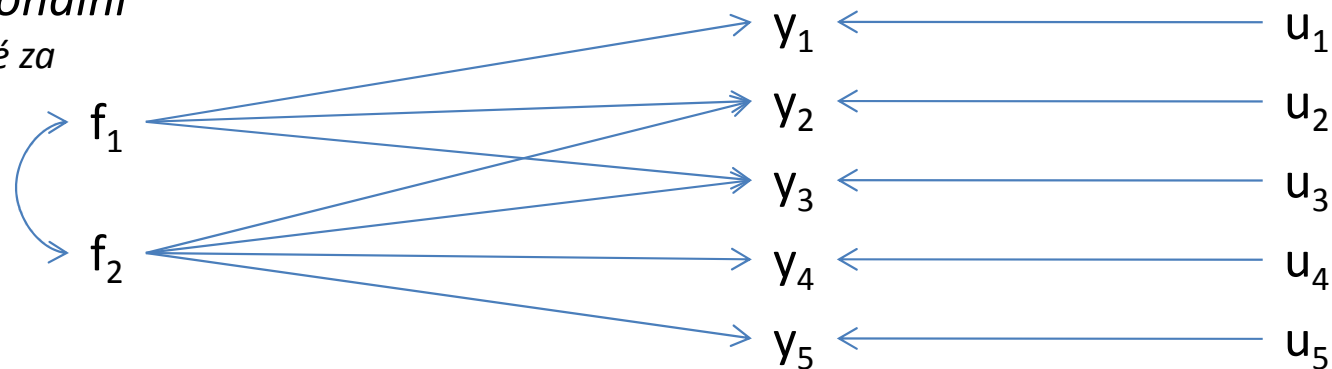
*Pozorovaná proměnná*

*Unikátní faktor*

*Rotace ortogonální*  
*- Nezávislé faktory*



*Rotace neortogonální*  
*- Faktory jsou závislé za účelem zvýšení interpretovatelnosti*



# Faktorová analýza – postup výpočtu

1. extrakce prvotních faktorů z kovarianční matice (analogie vlastních vektorů v PCA)
  - oproti PCA pracuje pouze s částí variability každé proměnné (tzv. communality), která je sdílena společnými faktory
  - několik možných algoritmů – principal factoring, metoda nejmenších čtverců, maximum likelihood apod.
  - výsledkem je komplexní struktura faktorů (obdobná PCA), kde řada faktorů má významné loadings (vztahy) k původním proměnným, počet takových faktorů je tzv. komplexita faktorů
2. v druhém kroku je rotací dosaženo zjednodušení struktury faktorů, tj. vztah mezi společnými faktory a původními proměnnými je zjednodušen (každá původní proměnná má hlavní vztah s jedním faktorem nebo malým počtem faktorů)
  - dva hlavní typy rotace:
    - ortogonální – faktory nemohou být korelovány, jsou tedy zcela nezávislé
    - neortogonální – faktory mohou být korelovány, nejsou tedy zcela nezávislé; vzhledem ke korelacím obtížnější interpretace



# Faktorová analýza - rotace

- Ortogonální rotace
  - Quartimax – minimalizuje sumu čtverců loadings původních proměnných na faktorových osách, tedy zjednodušuje řádky matice loadings (=každá původní proměnná má největší loadings na jedné faktorové ose)
  - Varimax – zjednodušuje sloupce matice loadings
  - Equimax – zjednodušuje řádky i sloupce matice loadings
  - Biquartimax – varianta equimax
- Neortogonální rotace
  - Oblimax
  - Quartimin
  - Oblimin
  - Covarimin
  - Biquartimin
  - Atd.

# Poděkování

Příprava výukových materiálů předmětu „DSAN02 Pokročilé metody analýzy dat v neurovědách“ byla finančně podporována prostředky projektu FRMU č. MUNI/FR/0260/2014 „Pokročilé metody analýzy dat v neurovědách jako nový předmět na LF MU“

