Analýza a klasifikace dat – přednáška 8

RNDr. Eva Koriťáková, Ph.D.

Podzim 2018

Schéma analýzy a klasifikace dat

Data

↓

Předzpracování

↓

Redukce

↓

Klasifikace

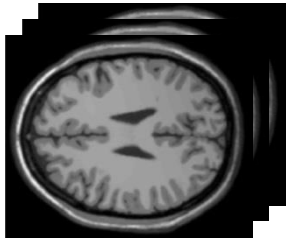
Ukázka - kognitivní data apod.

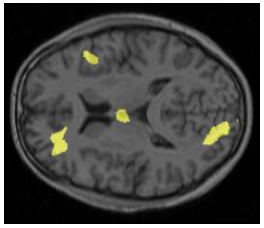
	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M		90
4	3	26	Z	178	70

	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





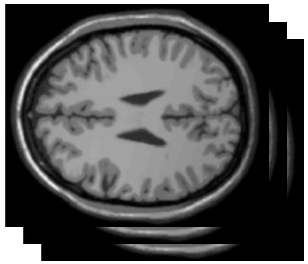
?

nebo

?

Proč používat redukci dat?

Obrazová data



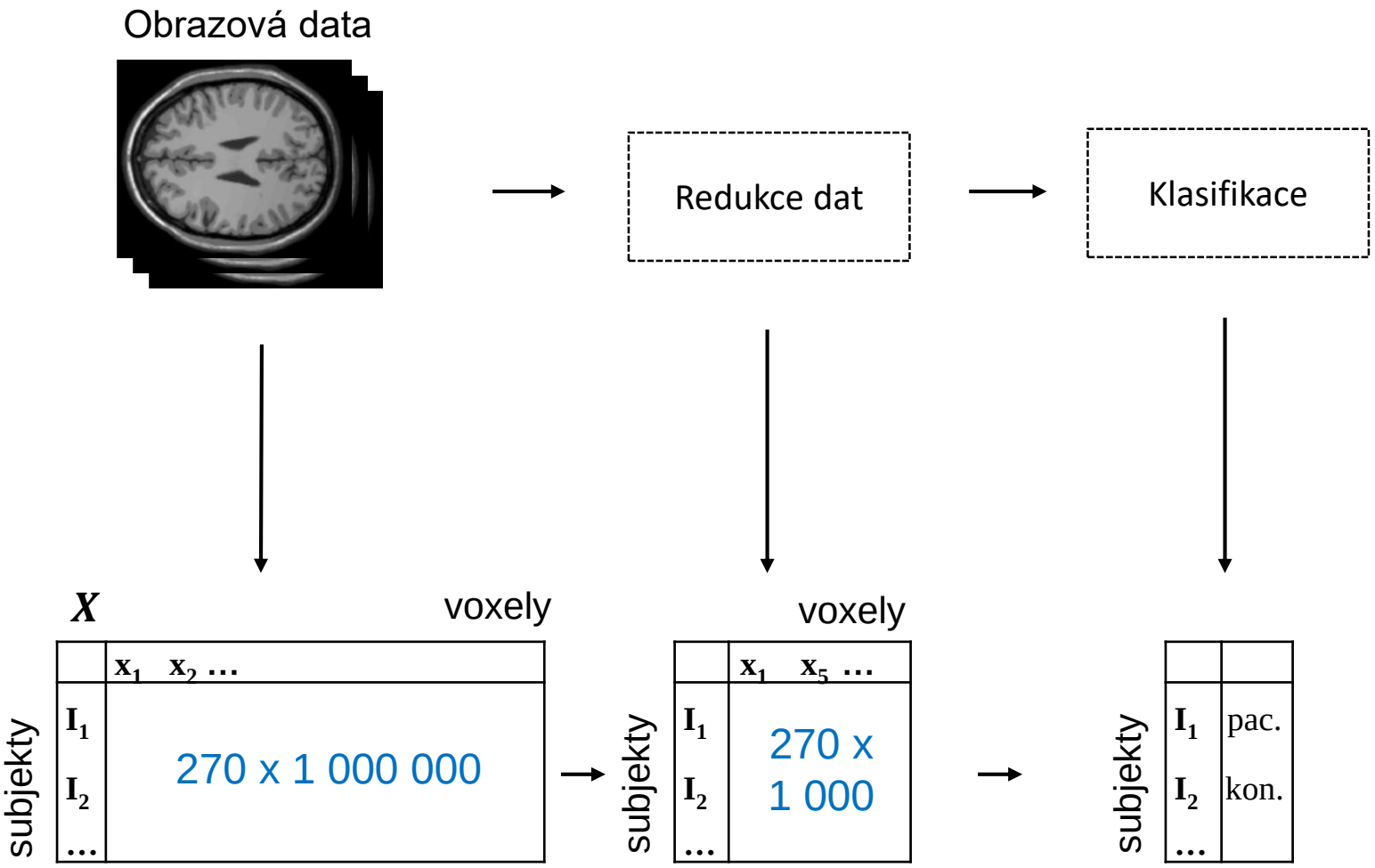
Klasifikace

$\boldsymbol{X}$ voxely

subjekty	x_1 x_2 …
I_1	270 x 1 000 000
I_2	
…	

subjekty	
I_1	pac.
I_2	kon.
…	

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 analýz
  - **moc proměnných** – možná nepřiměřená pracnost, vysoké náklady

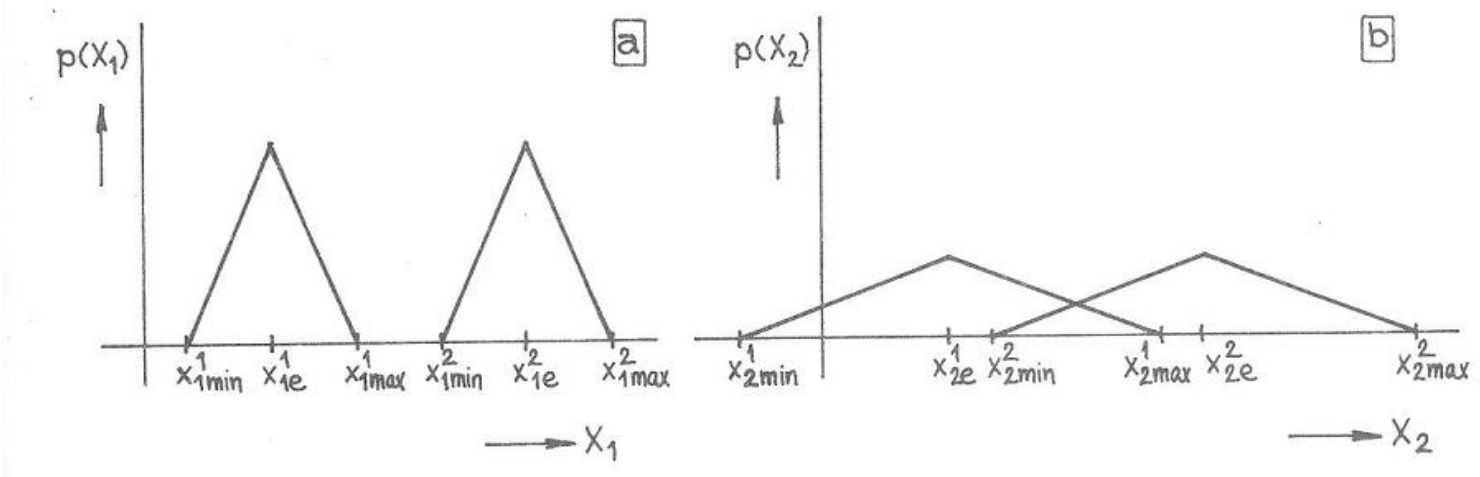
⇓

KOMPROMIS

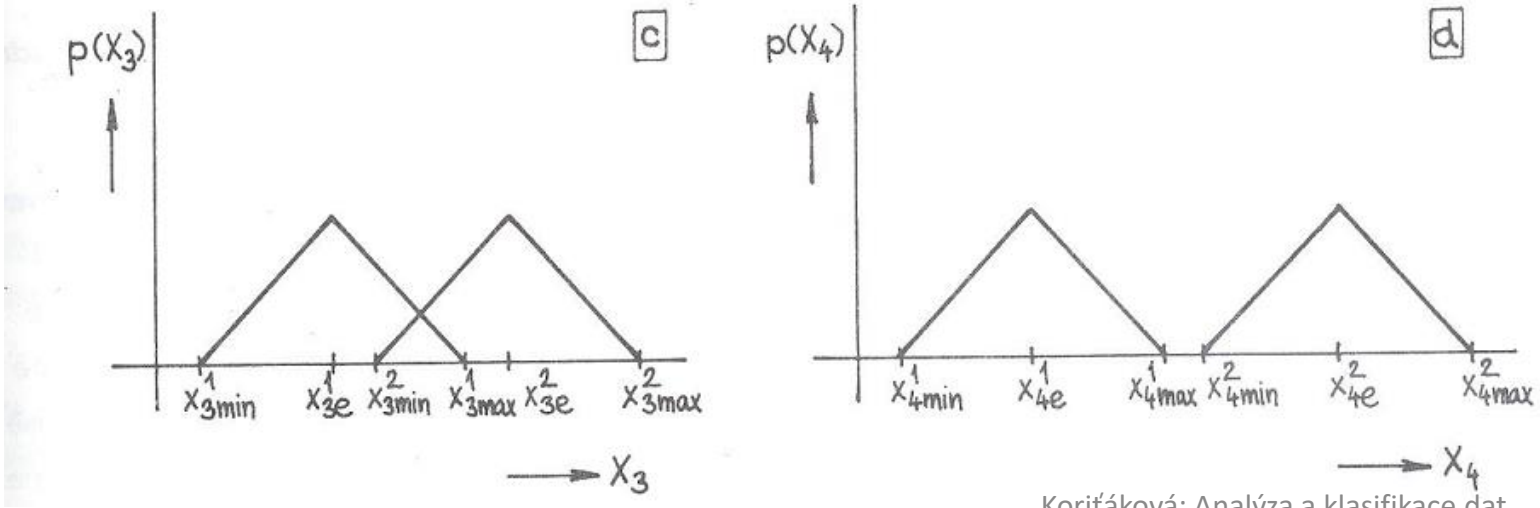
(určit ty proměnné, jejichž hodnoty nesou nejvíce informace z hlediska řešené úlohy, tj. např. ty proměnné, kterou jsou nejefektivnější pro vytvoření co nejodděleněji ších klasifikačních tříd)

Zásady pro volbu proměnných I

- výběr proměnných s minimálním rozptylem uvnitř tříd



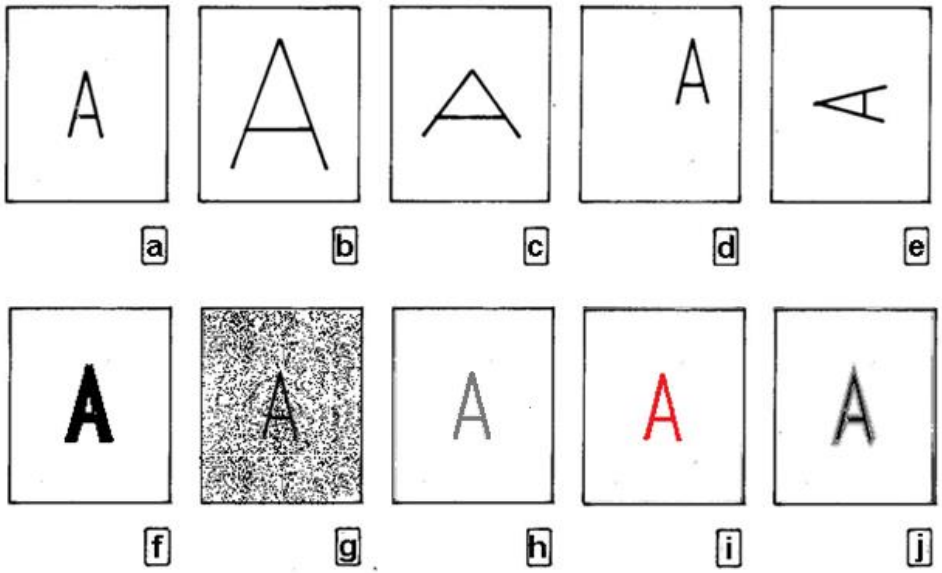
- výběr proměnných s maximální vzdáleností mezi třídami



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 pro správnou klasifikaci

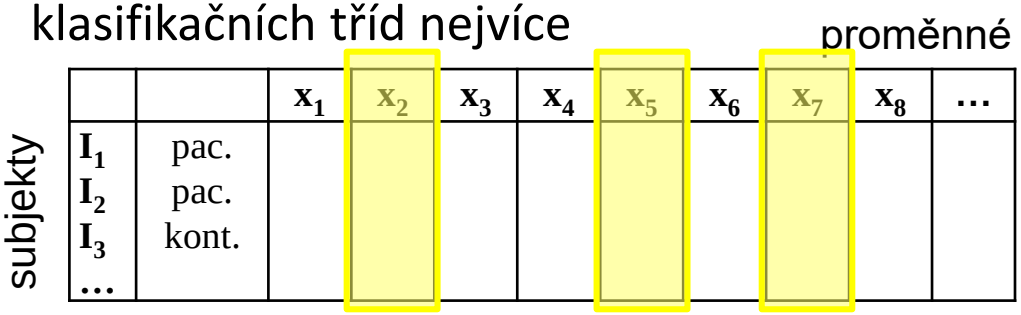
- výběr proměnných invariantních vůči deformacím
  - volba elementů formálního popisu závisí na vlastnostech původních i předzpracovaných dat a může ovlivňovat způsob předzpracování



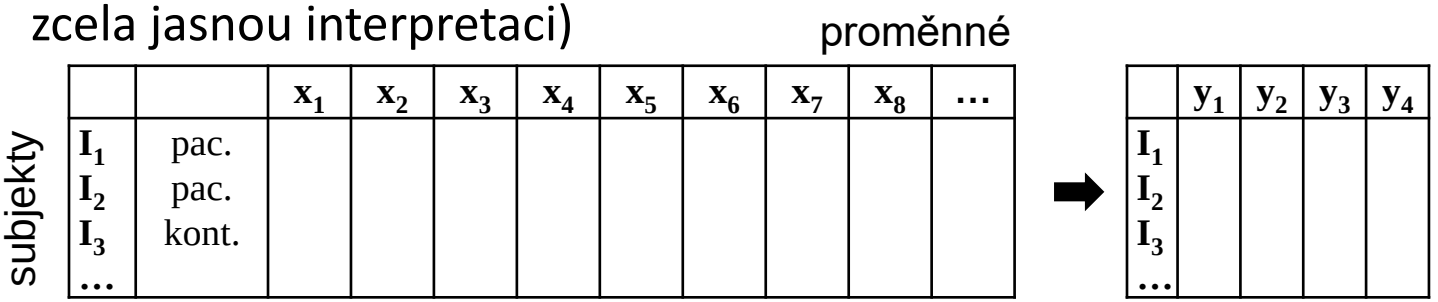
Selekce a extrakce proměnných

- formální popis objektu původně reprezentovaný *p*-rozměrným vektorem se snažíme vyjádřit vektorem *m*-rozměrným tak, aby množství diskriminační informace bylo co největší
- dva principiálně různé způsoby:

1. **selekce** – výběr těch proměnných, které přispívají k separabilitě klasifikačních tříd nejvíce



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



Extrakce proměnných

- transformace původních proměnných na menší počet jiných proměnných $\Rightarrow$ tzn. hledání (optimálního) zobrazení Z, které transformuje původní p-rozměrný prostor (obraz) na prostor (obraz) m-rozměrný ($m \leq p$)
- pro snadnější řešitelnost hledáme zobrazení Z v oboru lineárních zobrazení
- 3 kritéria pro nalezení optimálního zobrazení Z:
  - obrazy v novém prostoru budou aproximovat původní obrazy ve smyslu minimální střední kvadratické odchylky → **PCA**
  - rozložení pravděpodobnosti veličin v novém prostoru budou splňovat podmínky kladené na jejich pravděpodobnostní charakteristiky → **ICA**
  - obrazy v novém prostoru budou minimalizovat odhad pravděpodobnosti chyby
- metody extrakce proměnných (≈ metody ordinační analýzy):
  - **analýza hlavních komponent (PCA)**
  - faktorová analýza (FA)
  - **analýza nezávislých komponent (ICA)**
  - korespondenční analýza (CA)
  - vícerozměrné škálování (MDS)
  - **manifold learning metody (LLE, Isomap atd.)**
  - metoda parciálních nejmenších čtverců (PLS)

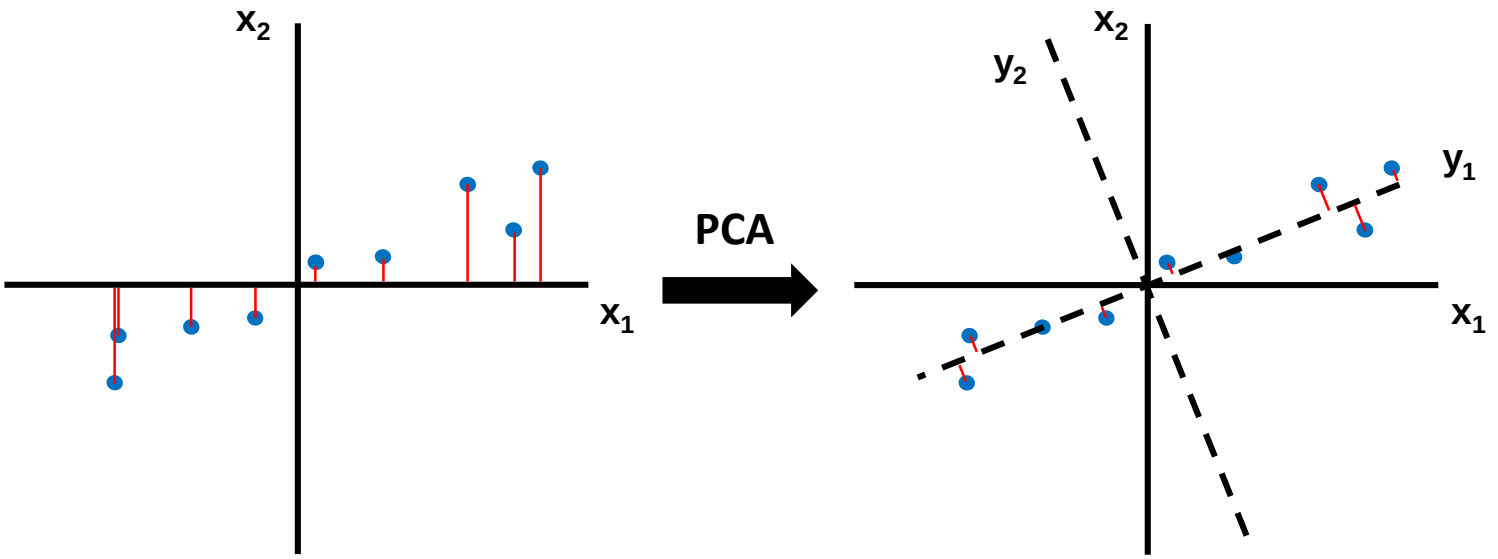
Metody ordinační analýzy – opakování

- analýza hlavních komponent, faktorová analýza, korespondenční analýza a vícerozměrné škálování se snaží zjednodušit vícerozměrnou strukturu dat výpočtem souhrnných os
- metody se liší v logice tvorby těchto os
  - maximální variabilita (analýza hlavních komponent, korespondenční analýza)
  - maximální interpretovatelnost os (faktorová analýza)
  - převod asociační matice do Euklidovského prostoru (vícerozměrné škálování)
- redundanční analýza a kanonická korelační analýza se snaží nalézt vztah mezi dvěma sadami vícerozměrných dat

Analýza hlavních komponent

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



- předpoklad: kvantitativní proměnné s normálním rozdělením

Postup PCA – opakování

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

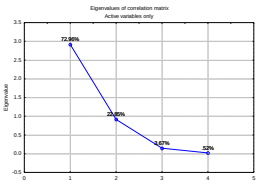
Identifikace optimálního počtu hlavních komponent pro další analýzu

- 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)

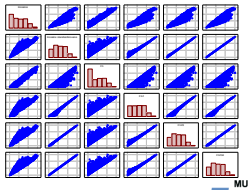
- kritéria pro výběr počtu komponent:

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



PCA – volba asociační matice

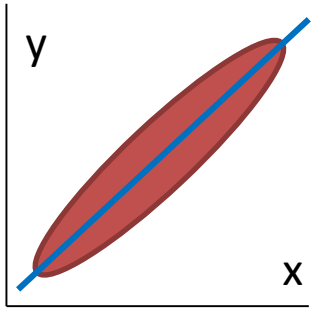
- **autokorelační matice** – data nejsou nijak upravena (zohledňována průměrná hodnota i rozptyl původních dat)
- **kovarianční (disperzní) matice** – data centrována (od každé proměnné odečtena její střední hodnota) – zohledňován rozptyl původních dat
- **matice korelačních koeficientů** – data standardizována (odečtení středních hodnot a podělení směrodatnými odchylkami) – použití, pokud mají proměnné různá měřítka a nám to v analýze vadí; avšak pozor na odlehlé hodnoty!

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

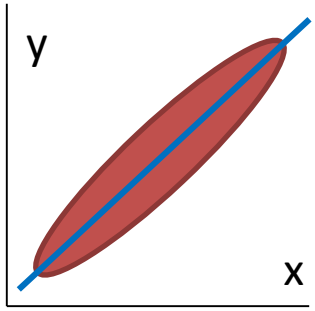
Analýza hlavních komponent – volba asociační matice

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

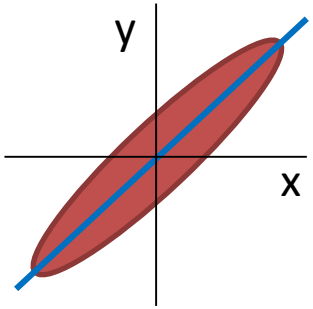
původní data



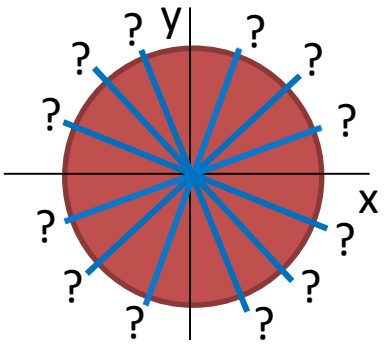
autokorelační matice
(data neupravována)



kovarianční matice
(odečten průměr)



matice korelačních koeficientů
(odečten průměr a podělení SD)



PCA – obecněji

- dáno *K* objektů (subjektů), *k*=1,…,*K,* charakterizovaných *p* proměnnými (objekty nejsou rozděleny do klasifikačních tříd)

proměnné

objekty	$\mathbf{v}_1$	$\mathbf{v}_2$	…	$\mathbf{v}_p$
$\mathbf{x}_1$				
$\mathbf{x}_2$				
…				
$\mathbf{x}_K$				

- aproximujme nyní kterýkoliv obraz $\mathbf{x}_k$ lineární kombinací *m* ortonormálních vektorů $\mathbf{e}_i$ ($m \leq p$)

$$\mathbf{y}_k = \sum_{i=1}^{m} c_{ki} e_i$$

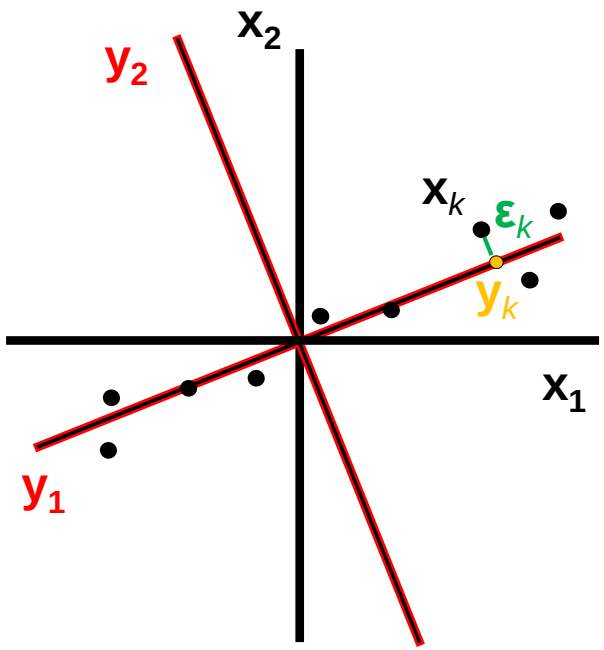
- koeficienty c_{ki} lze považovat za velikost *i*-té souřadnice vektoru $\mathbf{x}_k$ vyjádřeného v novém systému souřadnic s bází $\mathbf{e}_i$, *i*=1,2,…,*m*

$$c_{ki} = \mathbf{x}_k^T e_i$$

PCA – kritérium minimální střední kvadratické odchylky

- nalezení optimálního zobrazení pomocí **kritéria minimální střední kvadratické odchylky**:

$$\varepsilon_k^2 = \left\| \mathbf{x}_k - \mathbf{y}_k \right\|^2$$



- vztah lze pomocí dříve uvedených vztahů upravit na:

$$\varepsilon_k^2 = \left\| \mathbf{x}_k \right\|^2 - \sum_{i=1}^{m} c_{ki}^2$$

- střední kvadratická odchylka pro všechny objekty $\mathbf{x}_k$, k=1,…,K je:

$$\varepsilon^2 = \frac{1}{K} \sum_{k=1}^{K} \varepsilon_k^2 = \frac{1}{K} \sum_{k=1}^{K} \left\| \mathbf{x}_k \right\|^2 - \sum_{i=1}^{m} \mathbf{e}_i^T \left[\frac{1}{K} \sum_{k=1}^{K} \mathbf{x}_k \mathbf{x}_k^T \right] \mathbf{e}_i$$

PCA – kritérium minimální střední kvadratické odchylky

- musíme zvolit bázový systém $\mathbf{e}_i$ tak, aby střední kvadratická odchylka ε^2 byla minimální

- diskrétní konečný rozvoj podle vztahu $\mathbf{y}_k = \sum_{i=1}^{m} c_{ki} e_i$
  s bázovým systémem $\mathbf{e}_i$, optimálním podle kritéria minimální střední kvadratické chyby, nazýváme diskrétní Karhunenův – Loevův rozvoj

- střední kvadratická odchylka

$$\varepsilon^2 = \frac{1}{K}\sum_{k=1}^{K}\varepsilon_k^2 = \frac{1}{K}\sum_{k=1}^{K}\left\|\mathbf{x}_k\right\|^2 - \sum_{i=1}^{m}\mathbf{e}_i^T\left[\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_k\mathbf{x}_k^T\right]\mathbf{e}_i$$

je minimální, když je maximální výraz

$$\sum_{i=1}^{m}\mathbf{e}_i^T\boldsymbol{\kappa}(\mathbf{x})\mathbf{e}_i, \quad \text{kde} \quad \kappa(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_k.\mathbf{x}_k^T$$

je autokorelační matice řádu *m*. Protože je symetrická a semidefinitní, jsou její vlastní čísla λ_i, i=1,…,*m*, reálná a nezáporná a vlastní vektory $\mathbf{v}_i$, jsou buď ortonormální, nebo je můžeme ortonormalizovat (v případě násobných vlastních čísel).

PCA – kritérium minimální střední kvadratické odchylky

- uspořádáme-li vlastní čísla sestupně podle velikosti, tj.

  $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$

  a podle toho očíslujeme i odpovídající vlastní vektory, lze dokázat, že výše uvedený výraz dosahuje maxima, jestliže platí

  $$\mathbf{e}_i = \mathbf{v}_i,\ i=1,\ldots,m$$

  a pro velikost maxima je

  $$\max \sum_{i=1}^{m} \mathbf{e}_i^T . \kappa(\mathbf{x}) . \mathbf{e}_i = \sum_{i=1}^{m} \lambda_i$$

- pak pro minimální střední kvadratickou platí

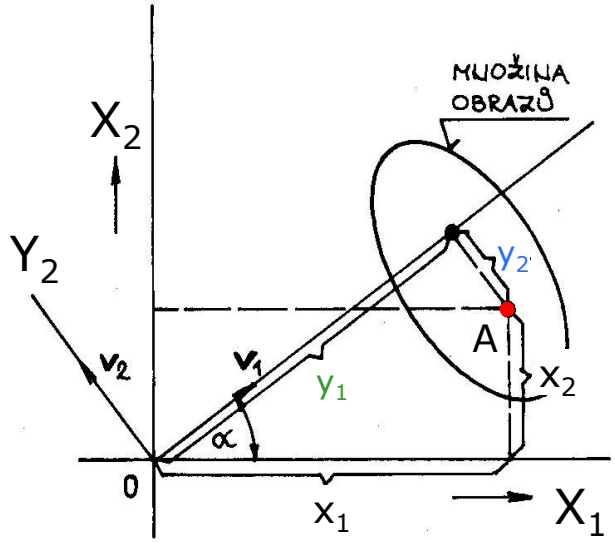
  $$\varepsilon_{\min}^2 = \frac{1}{K} \sum_{k=1}^{K} \left\| \mathbf{x}_k \right\|^2 - \sum_{i=1}^{m} \lambda_i = Tr(\kappa(\mathbf{x})) - \sum_{i=1}^{m} \lambda_i = \sum_{i=m+1}^{p} \lambda_i$$

- minimální střední kvadratickou je tedy rovna součtu těch vlastních čísel, jimž odpovídající vlastní vektory nebyly použity při aproximaci objektu

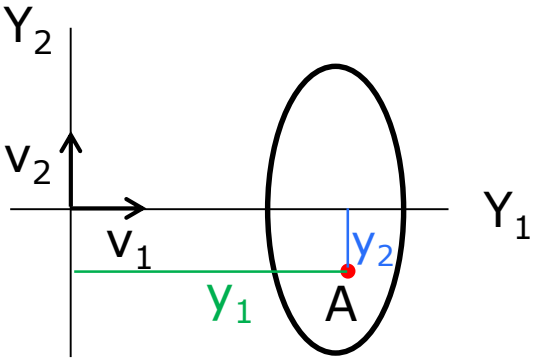
PCA – vlastnosti Karhunenova-Loevova rozvoje

- při daném počtu *m* členů rozvoje poskytuje ze všech možných aproximací nejmenší střední kvadratickou odchylku
- při použití kovarianční matice jsou transformované souřadnice nekorelované; pokud se výskyt obrazů řídí normálním rozložením zajišťuje nekorelovanost i jejich nezávislost
- vliv každého členu uspořádaného rozvoje se zmenšuje s jeho pořadím
- změna požadavků na velikost střední kvadratické odchylky nevyžaduje přepočítávat celý rozvoj, nýbrž jen změnit počet jeho členů

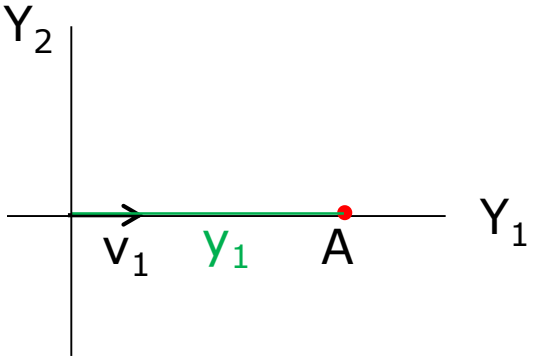
PCA – geometrická interpretace



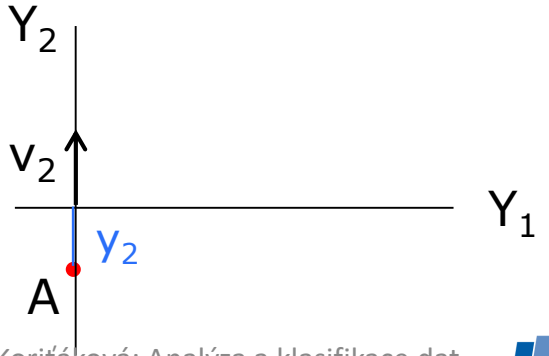
použití obou hlavních komponent



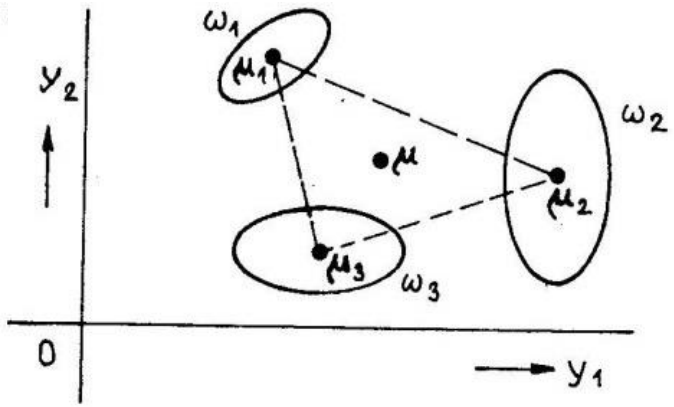
použití 1. hlavní komponenty



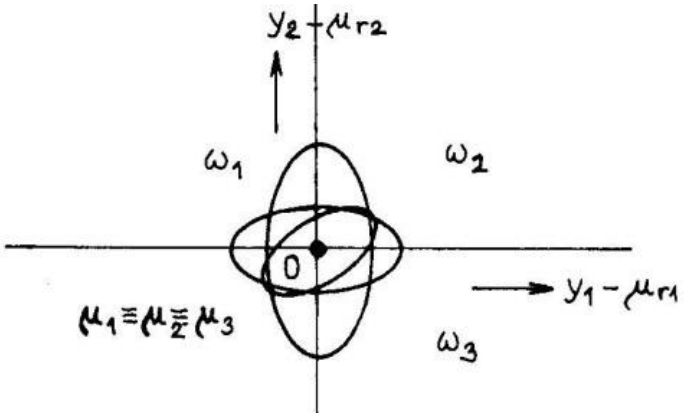
použití 2. hlavní komponenty



PCA – rozdělení do tříd

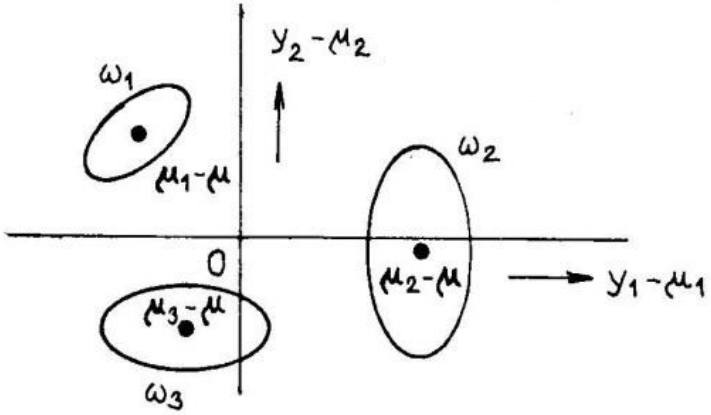


odečtení průměru každé skupiny zvlášť



→ není vhodné
(odlišení tříd jen podle rozptylu)

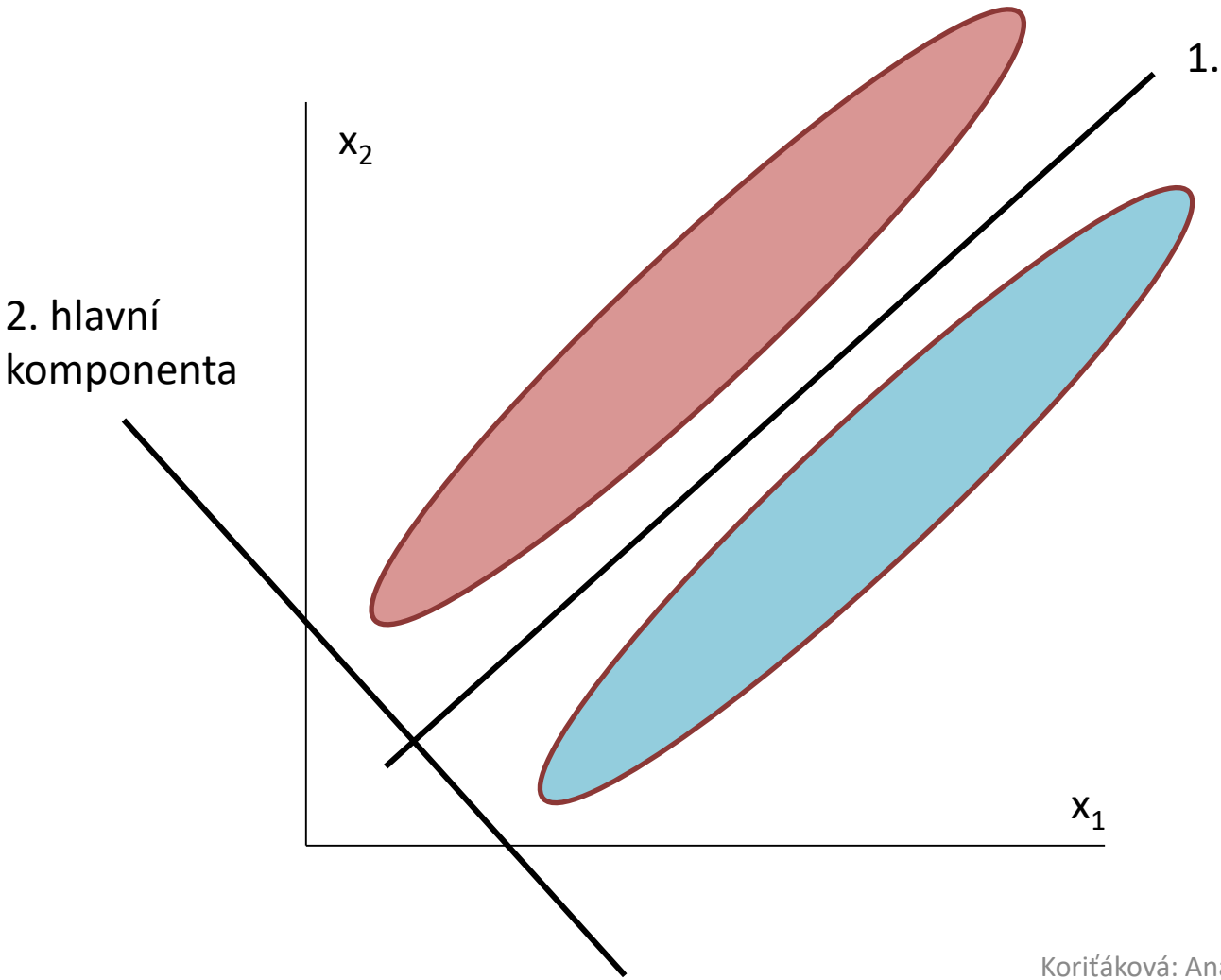
odečtení celkového průměru



→ je vhodné
(neodstraňuje vliv středních hodnot obrazů v jednotlivých třídách)

PCA a klasifikace

PCA často nebývá vhodnou metodou redukce dat před klasifikací

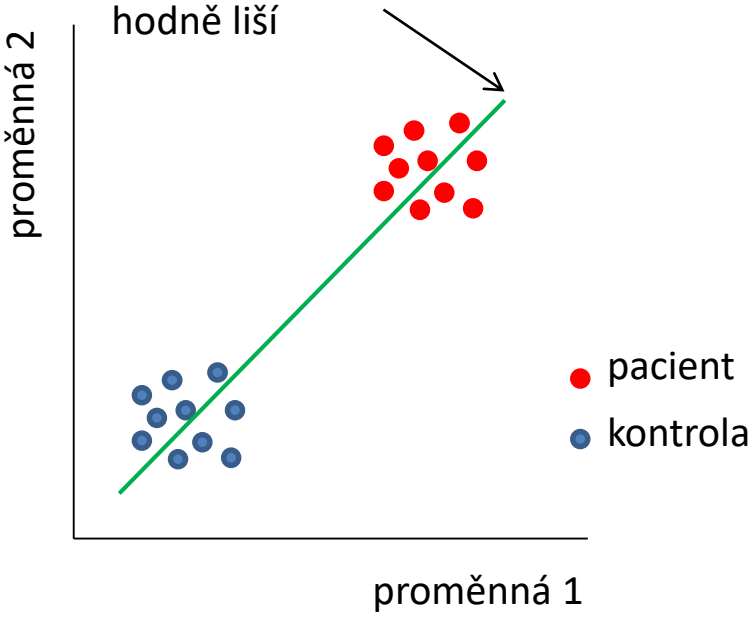


Pro klasifikaci vhodnější 2. HK, přestože vyčerpává méně variability!

PCA a klasifikace

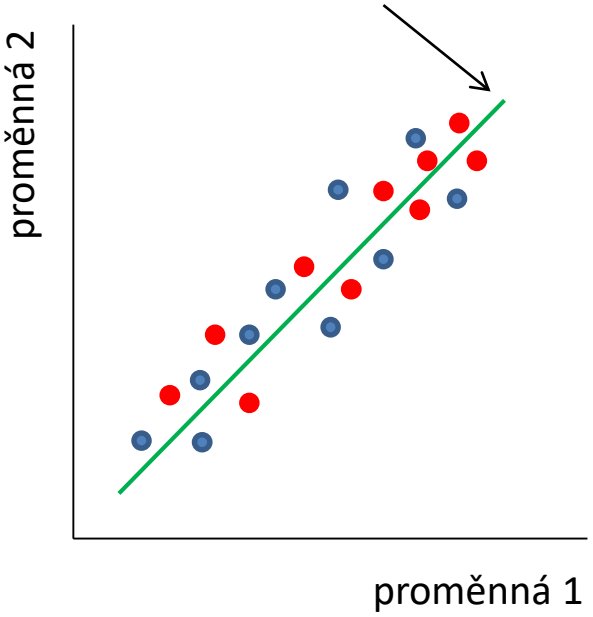
Když hlavní komponenta vyčerpává hodně variability, neznamená to, že musí rovněž dobře klasifikovat

vysoká korelace mezi proměnnými 1 a 2
- způsobená tím, že se skupiny od sebe hodně liší



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

vysoká korelace mezi proměnnými 1 a 2
- skupiny se ale od sebe neliší



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

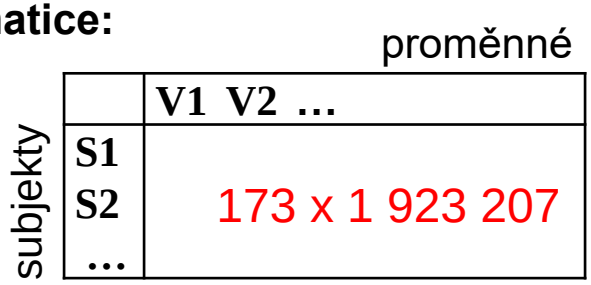
PCA – rozšiřující poznatky I

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

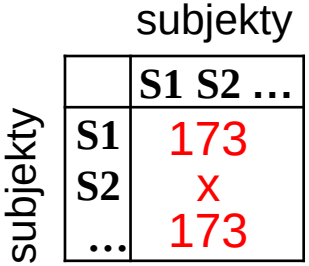
- 1. způsob: iterativní postupný výpočet vlastních vektorů a vlastních čísel
- 2. způsob: výpočet vlastních vektorů $\mathbf{v}_i$ „velké“ kovarianční matice (proměnných) $\mathbf{X}^T\mathbf{X}_{(p,p)}$ z vlastních vektorů $\mathbf{w}_i$ „malé“ kovarianční matice (subjektů) $\mathbf{X}\mathbf{X}^T_{(n,n)}$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

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

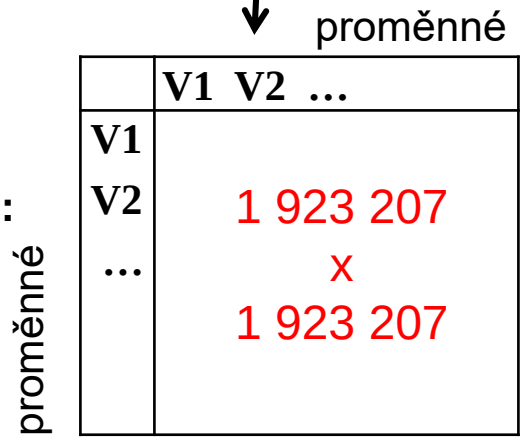
Datová matice:



Kovarianční matice subjektů:



Kovarianční matice proměnných:



PCA – rozšiřující poznatky II

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

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

- matice **U** a **V** jsou ortogonální a normované (ortonormální)
- matice **U** složena z vlastních (charakteristických) vektorů matice $\mathbf{X}\mathbf{X}^{\mathrm{T}}_{(n,n)}$
- matice **V** z vlastních vektorů matice $\mathbf{X}^{\mathrm{T}}\mathbf{X}_{(p,p)}$
- Matice **Γ** je typu *k x 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 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ i $\mathbf{X}^{\mathrm{T}}\mathbf{X}$

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(:,24:29); % 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

PCA – příklad – 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 – 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“

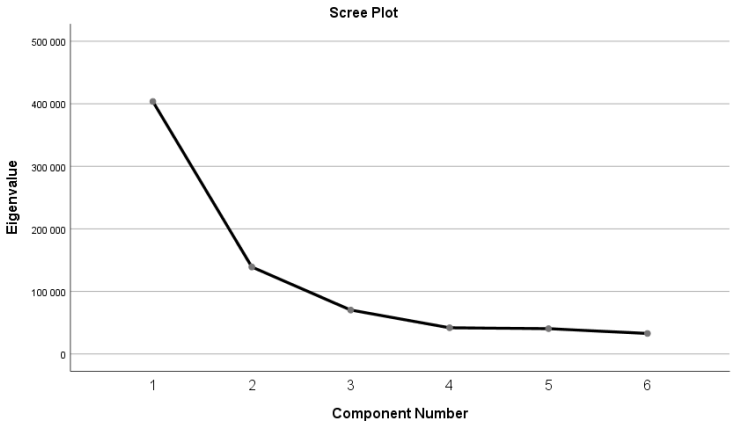
PCA – příklad – řešení v softwaru SPSS

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)

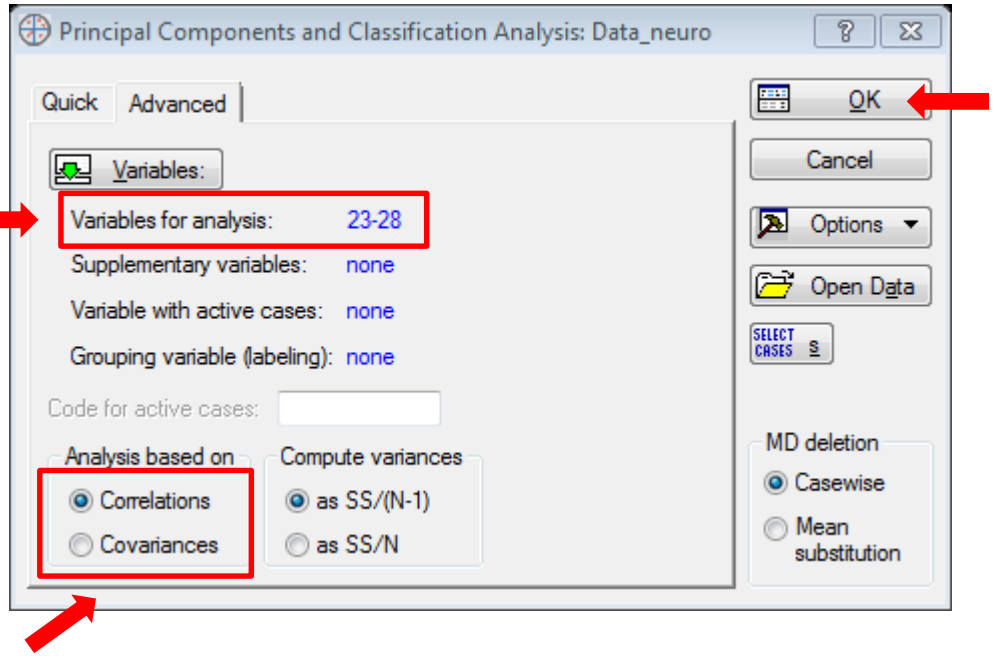
FAC1_1	FAC2_1	FAC3_1	FAC4_1	FAC5_1	FAC6_1
-,85256	,86362	,34174	,46067	-1,24179	-,15117
-,48179	1,36289	-1,59851	-,99769	-,20191	-,81984
,34317	1,27004	,72775	-,79788	-,40966	,70786
-,77548	1,43599	-1,01106	-,36313	-,27870	-1,94204
-,54519	,64565	-1,18128	-,52272	-,02490	,18146
-,19375	2,01086	-1,18890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

* normalizace vl. vektorů by se provedla v exelu (viz. slide 35)

PCA – příklad – řešení v softwaru Statistica

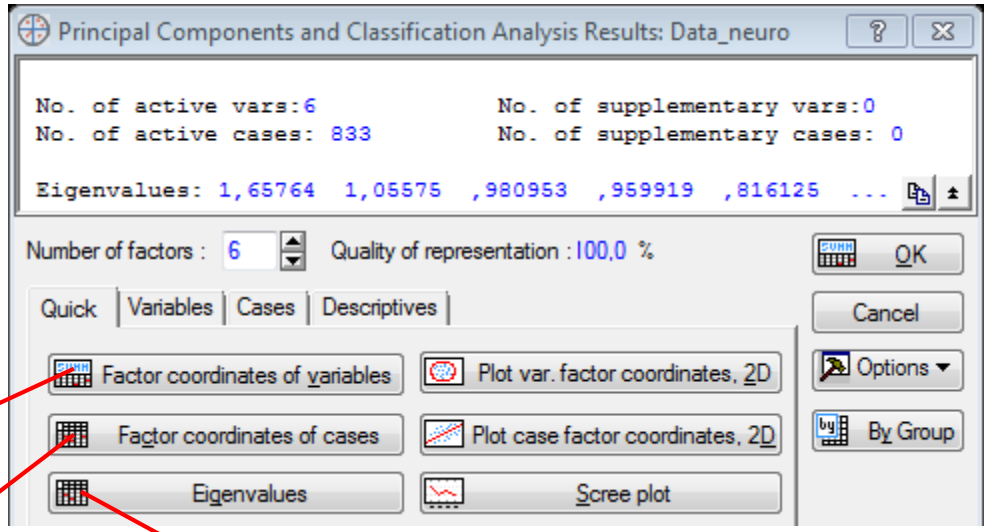
- Zadání: Proveďte PCA na objemech 6 mozkových struktur u 833 subjektů.
- Řešení: Statistics – Multivariate Exploratory Techniques – Principal Components & Classification Analysis

vybrat proměnné



zvolit, zda se má počítat kovarianční či korelační matice

PCA – příklad – řešení v softwaru Statistica



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 – řešení v softwaru Statistica

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

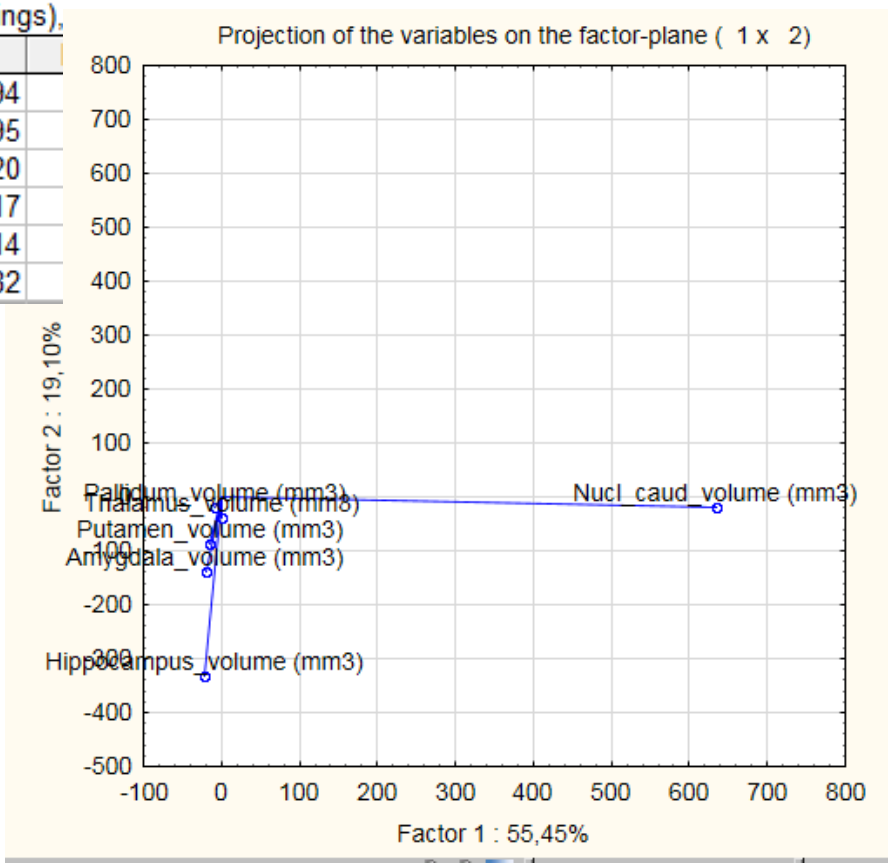
PCA – příklad – řešení v softwaru Statistica

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

Metody varietního učení (manifold learning)

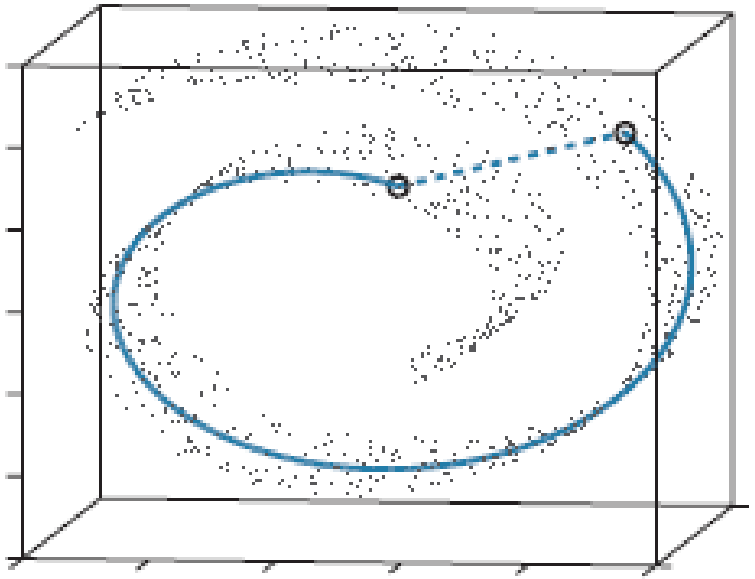
Úvod – redukce dimenzionality

- klasické metody redukce dimenzionality:
  - **PCA** (principal component analysis) – snaha o nalezení „podstruktury“ (embedding) v datech tak, aby byl zachován rozptyl
  - **MDS** (multidimensional scaling) – snaha o nalezení „podstruktury“ v datech tak, aby byly zachovány vzdálenosti mezi body; ekvivalentní s PCA při použití Euklidovské vzdálenosti

- tyto klasické metody redukce dimenzionality nedokáží zachytit složité nelineární struktury

  → **metody varietního učení**

Swiss roll



Tenenbaum et al. 2000, Science

Metody varietního učení

- metody pro nelineární redukci a reprezentaci dat
- manifold = „nadplocha“ – čáry a kruhy jsou 1D nadplochy, koule je příklad 2D nadplocha

- základní metody varietního učení:
  1. **ISOMAP** (Tenenbaum et al. 2000)
  2. **Metoda lokálně lineárního vnoření = LLE** (Roweis & Saul 2000)

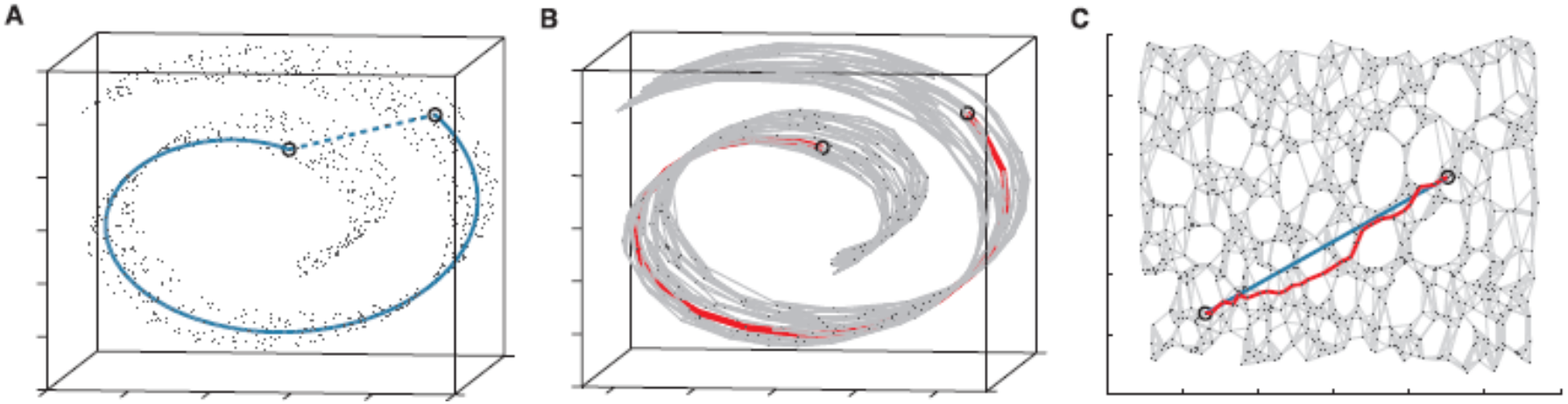
- další metody varietního učení:

  Laplacian Eigenmaps, Sammon's Mapping, Kohonen Maps, Autoencoders, Gaussian process latent variable models, Curvilinear component analysis, Curvilinear Distance Analysis, Kernel Principal Component Analysis, Diffusion Maps, Hessian LLE, Modified LLE, Local Tangent Space Alignment, Local Multidimensional Scaling, Maximum Variance Unfolding, Data-Driven High Dimensional Scaling, Manifold Sculpting, RankVisu

- některé z manifold learning metod implementovány v **mani.m** demu

ISOMAP metoda

- založena na MDS
- ISOMAP = isometric feature mapping
- snaha o zachování vnitřní geometrie dat, která je zachycena pomocí **geodézních vzdáleností** (geodesis distance) založených na hledání nejkratších cest v grafu s hranami spojujícími sousední datové body



Tenenbaum et al. 2000 Science, A Global Geometric Framework for Nonlinear Dimensionality Reduction

ISOMAP metoda – algoritmus se 3 kroky

1. **Vytvoření grafu spojujícího sousední objekty**:
   - nejprve nutno vypočítat vzdálenosti $D(\mathbf{x}_i, \mathbf{x}_j)$ mezi všemi objekty
   - poté dojde ke spojení objektů tak, že se *j*-tý objekt spojí s těmi objekty, jejichž vzdálenost je menší než ε (v případě ε-ISOMAP), nebo s jeho *k* nejbližšími sousedy (v případě *k*-ISOMAP)

2. **Výpočet geodézních vzdáleností** $D_G(\mathbf{x}_i, \mathbf{x}_j)$ mezi všemi objekty nalezením nejkratší cesty v grafu mezi danými objekty – iniciální nastavení $D_G(\mathbf{x}_i, \mathbf{x}_j)$ závisí na tom, jestli jsou objekty spojené hranou či nikoliv:
   - pokud objekty spojeny hranou: $D_G(\mathbf{x}_i, \mathbf{x}_j) = D(\mathbf{x}_i, \mathbf{x}_j)$
   - pokud ne: $D_G(\mathbf{x}_i, \mathbf{x}_j) = \infty$

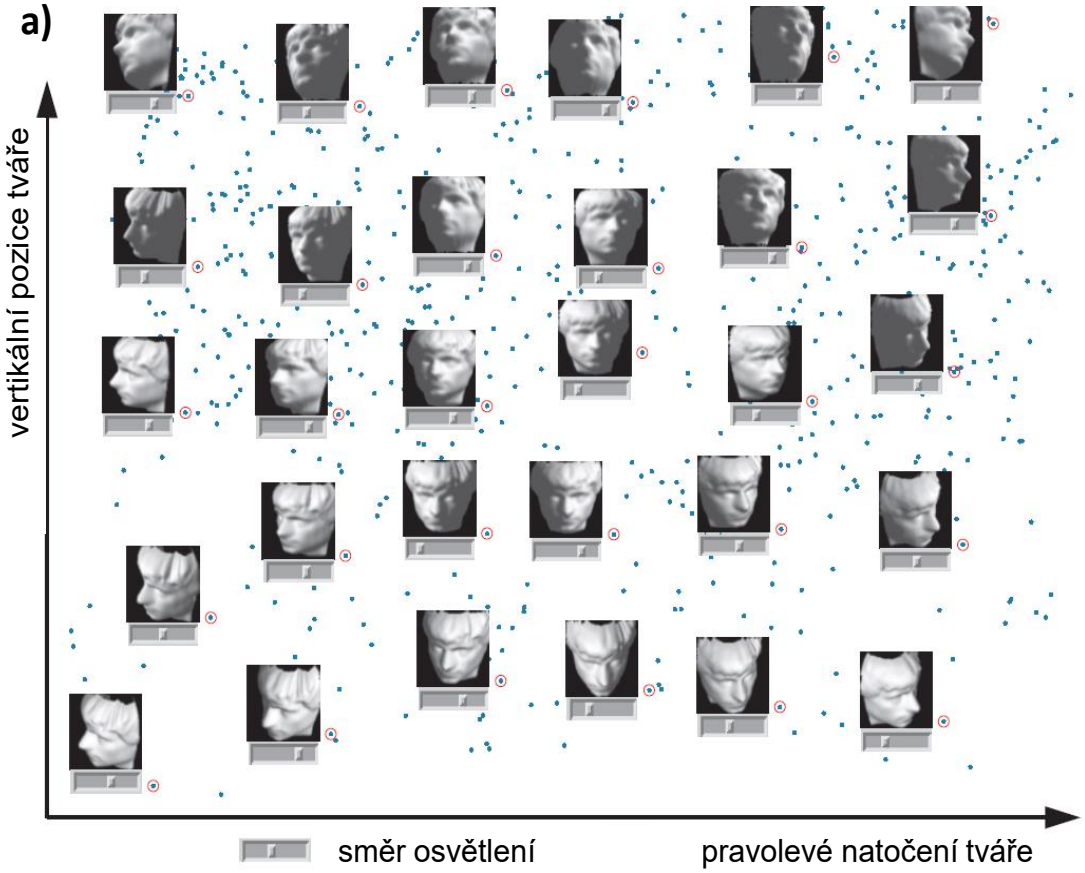
   poté je pro každé $k = 1,2, \dots, N$ nahrazena vzdálenost $D_G(\mathbf{x}_i, \mathbf{x}_j)$ hodnotou $\min(D_G(\mathbf{x}_i, \mathbf{x}_j), D_G(\mathbf{x}_i, \mathbf{x}_k) + D_G(\mathbf{x}_k, \mathbf{x}_j))$.

3. **Aplikace nemetrického vícerozměrného škálování (MDS) na matici geodézních vzdáleností** – tzn. transformace dat do Euklidovského prostoru tak, aby byly co nejlépe zachovány geodézní vzdálenosti.

Tenenbaum et al. 2000 Science, A Global Geometric Framework for Nonlinear Dimensionality Reduction

ISOMAP metoda – ukázka 1

Výsledek *k*-ISOMAP algoritmu u 698 obrazů tváří



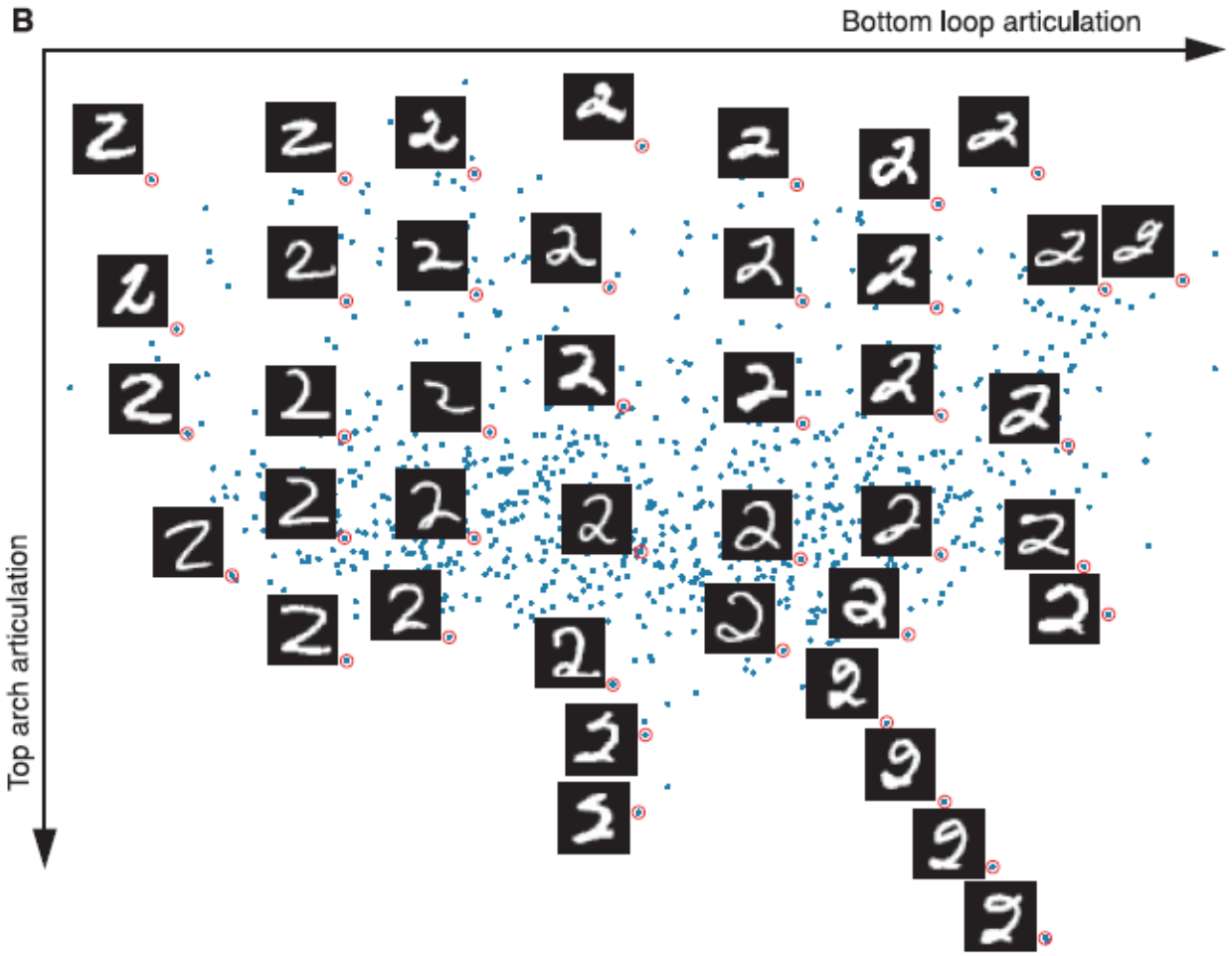
Interpolace podél os *x* a *y* v podprostoru obrazů tváří

Výsledkem je redukce původních 4096 proměnných (obrazy měly rozměry 64 x 64 pixelů) na pouze tři komponenty

Tenenbaum et al. 2000 Science, A Global Geometric Framework for Nonlinear Dimensionality Reduction

ISOMAP metoda – ukázka 2

Výsledek ISOMAP algoritmu u obrazů ručně psaných číslic

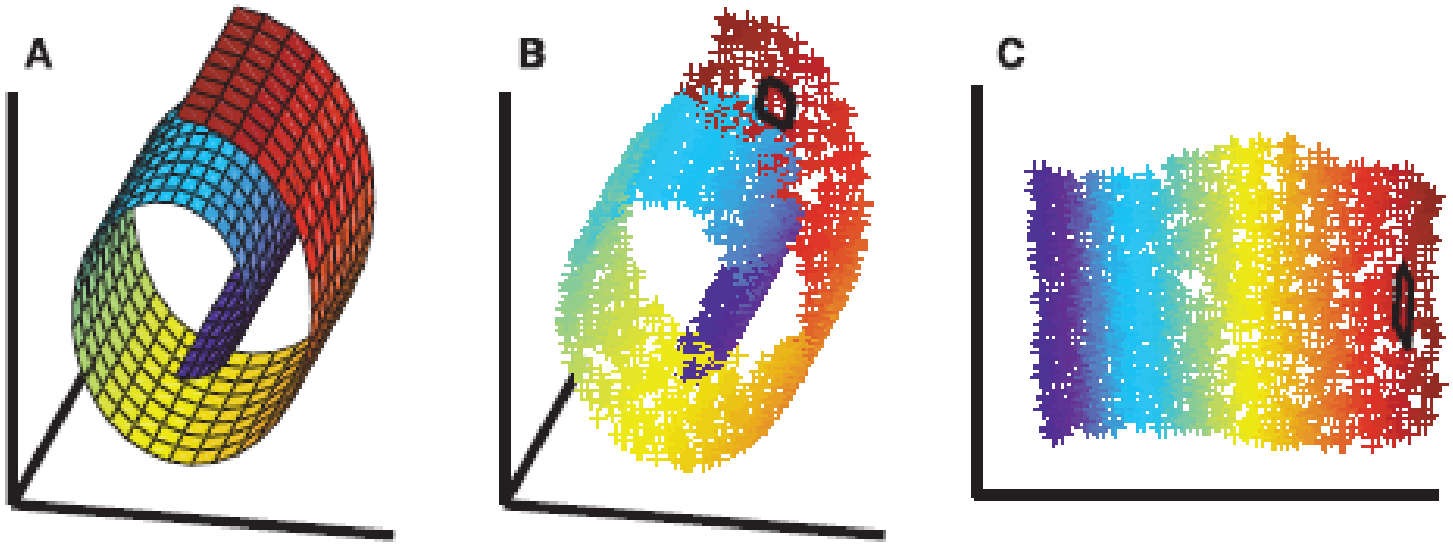


Interpolace podél os *x* a *y* v podprostoru obrazů číslic

Tenenbaum et al. 2000 Science, A Global Geometric Framework for Nonlinear Dimensionality Reduction

Metoda lokálně lineárního vnoření (LLE)

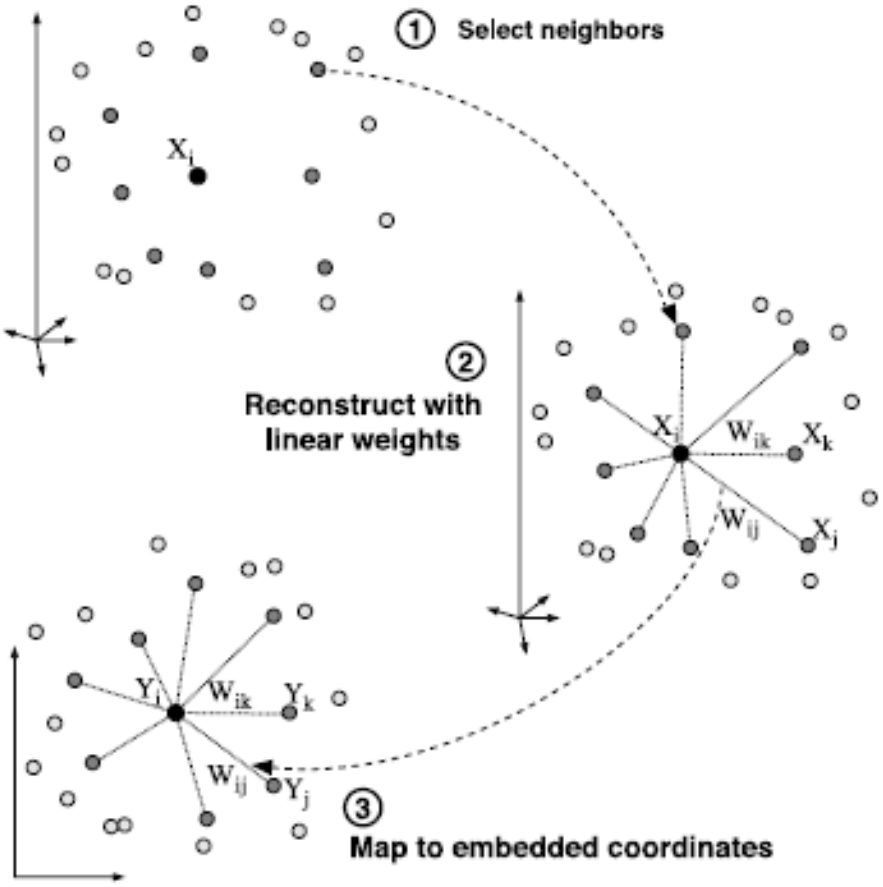
- Locally Linear Embedding (LLE)
- založena na zachování mapování sousedů (neighborhood-preserving mapping)
- LLE rekonstruuje globální nelineární struktury z lokálních lineárních fitů



Černě vyznačeno okolí (sousedi) jednoho bodu.

Roweis & Saul 2000 Science, Nonlinear Dimensionality Reduction by Locally Linear Embedding

LLE - algoritmus



1. Výběr *k* nejbližších sousedů.

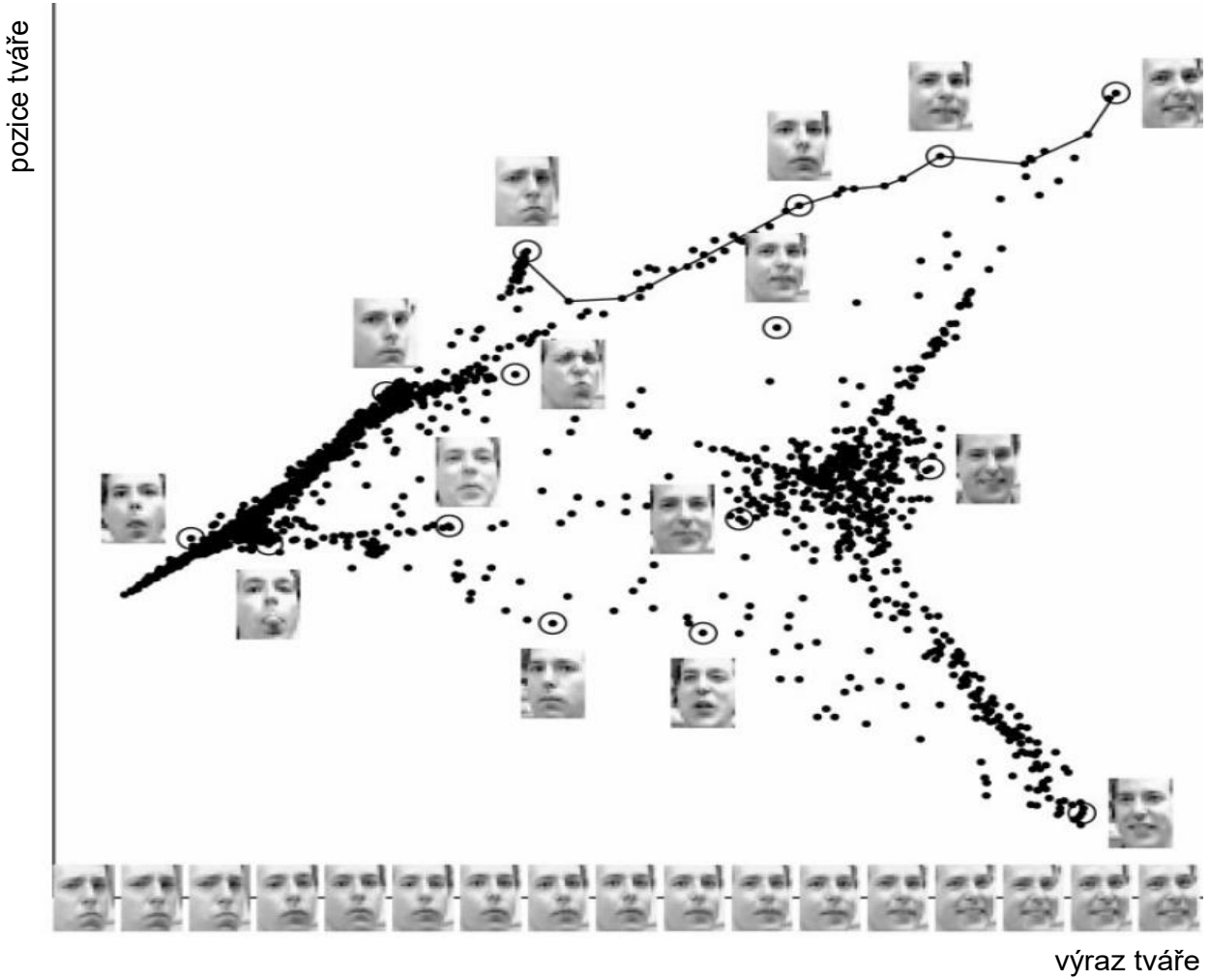
2. Rekonstrukce objektů z jejich sousedů – cílem je nalezení vah W_{ij} tak, aby rekonstrukční chyby byly co nejmenší, tzn. snažíme se minimalizovat výraz $\varepsilon(W) = \sum_i \left|\mathbf{x}_i - \sum_j W_{ij}\,\mathbf{x}_j\right|^2$, přičemž součet vah W_{ij} musí být roven 1; váhy jsou invariantní vůči rotaci, přeškálování a translaci objektů a jejich sousedů.

3. Mapování do „nadplochy“ s nižší dimenzionalitou (lineární mapování – skládající se z translací, rotací a přeškálování) pomocí výpočtu vlastních vektorů

Roweis & Saul 2000 Science, Nonlinear Dimensionality Reduction by Locally Linear Embedding

LLE – ukázka 1

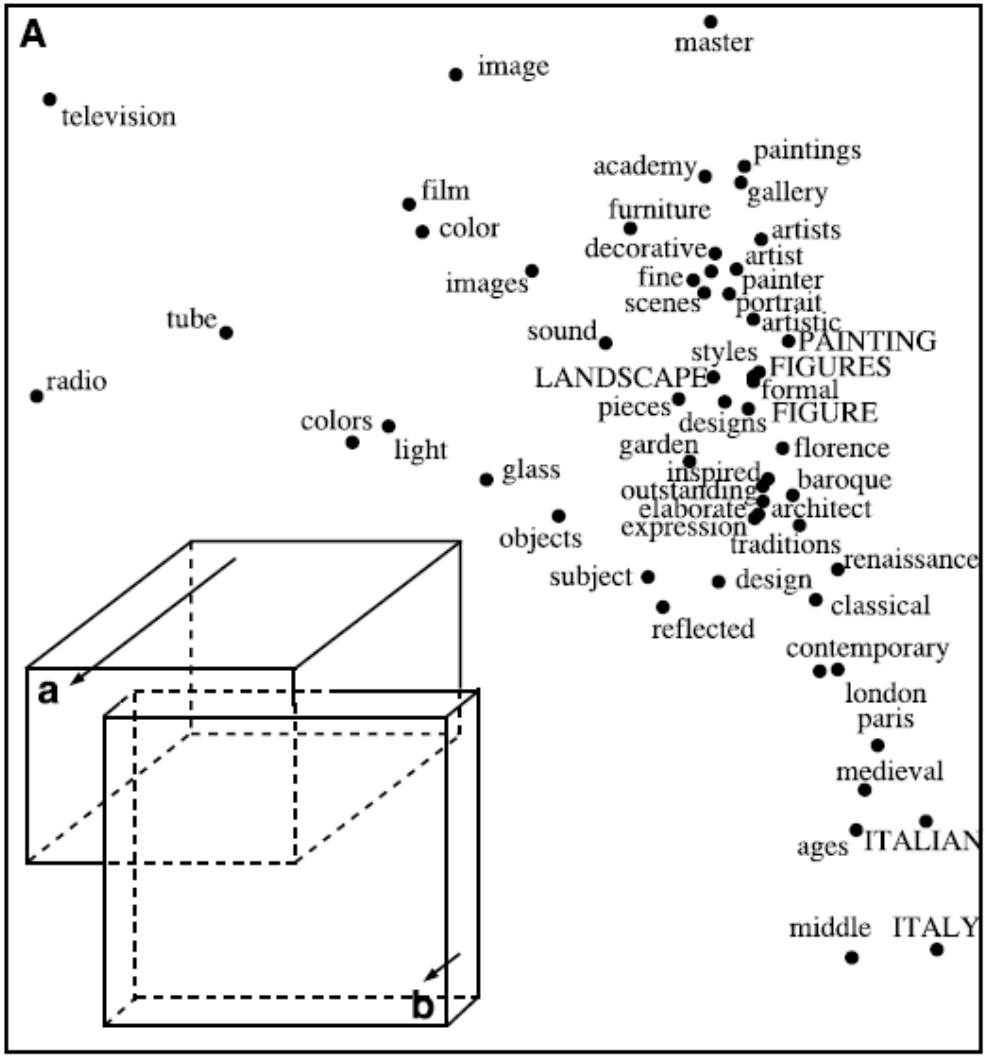
Výsledek LLE algoritmu u obrazů tváří



Roweis & Saul 2000 Science, Nonlinear Dimensionality Reduction by Locally Linear Embedding

LLE – ukázka 2

Výsledek LLE algoritmu u hodnocení počtu a výskytu slov v encyklopedii



Roweis & Saul 2000 Science, Nonlinear Dimensionality Reduction by Locally Linear Embedding

Výhody a nevýhody ISOMAP a LLE

- **výhody a nevýhody ISOMAP**:
  + zachovává globální strukturu dat
  + málo parametrů
  - citlivost k šumu
  - výpočetně náročné

- **výhody a nevýhody Locally Linear Embedding (LLE)**:
  + rychlý
  + jeden parametr
  + jednoduché operace lineární algebry
  - může zkreslit globální strukturu dat

Další práce

- **Laplacian Eigenmaps for Dimensionality Reduction and Data Representation** (Belkin & Niyogi 2003):
  - snaha o zachování mapování sousedů jako u Locally Linear Embedding
  - podobný algoritmus jako LLE, ale používá se zde výpočet vlastních vektorů a vlastních čísel s využitím Laplaciánu grafu
  - souvislost s klastrováním – lokální přístup k redukci dimenzionality způsobuje přirozené klastrování dat (klastrování tedy nastává u Laplacian Eigenmaps a LLE, nenastává u ISOMAP, protože to je globální metoda)

- **Manifold Learning for Biomarker Discovery in MR Imaging** (Wolz et al. 2010)
  - použití Laplacian eigenmaps u obrazů pacientů s Alzheimerovou chorobou (data ADNI)

Příprava nových učebních materiálů pro obor Matematická biologie

je podporována projektem OPVK

č. CZ.1.07/2.2.00/28.0043

„Interdisciplinární rozvoj studijního oboru Matematická biologie“

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ