Pokročilé metody analýzy dat v neurovědách

RNDr. Eva Koriťáková, Ph.D.

doc. RNDr. Ladislav Dušek, Dr.

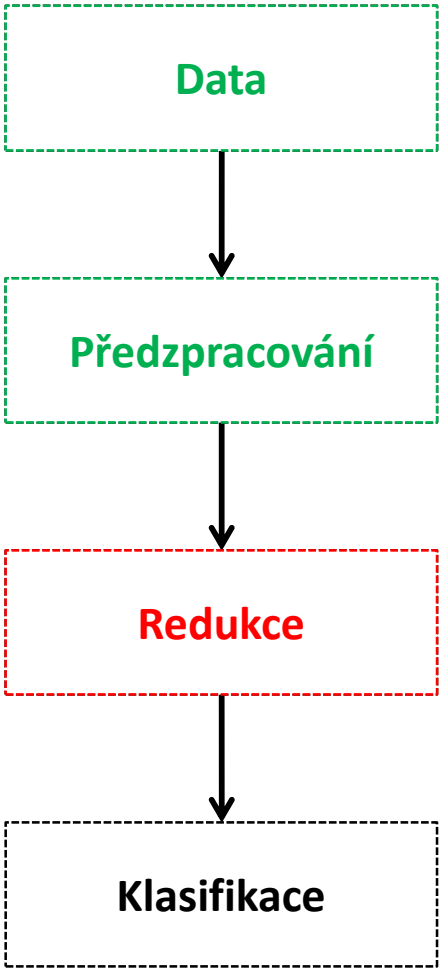
Blok 6

Ordinační analýzy II

Osnova

1. Analýza nezávislých komponent (ICA)
2. Vícerozměrné škálování (MDS)
3. Varietní učení (manifold learning)
4. Korespondenční analýza (CA)
5. Metoda parciálních nejmenších čtverců (PLS)
6. Redundanční analýza (RDA)
7. Kanonická korelační analýza (CCorA)

Schéma analýzy a klasifikace dat – opakování



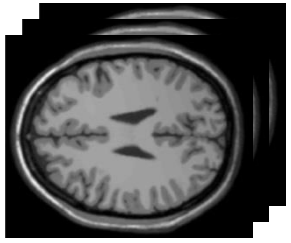
Ukázka - kognitivní data apod.

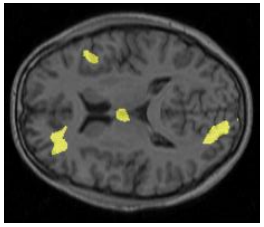
	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M		90
4	3	26	Z	178	70

	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

?

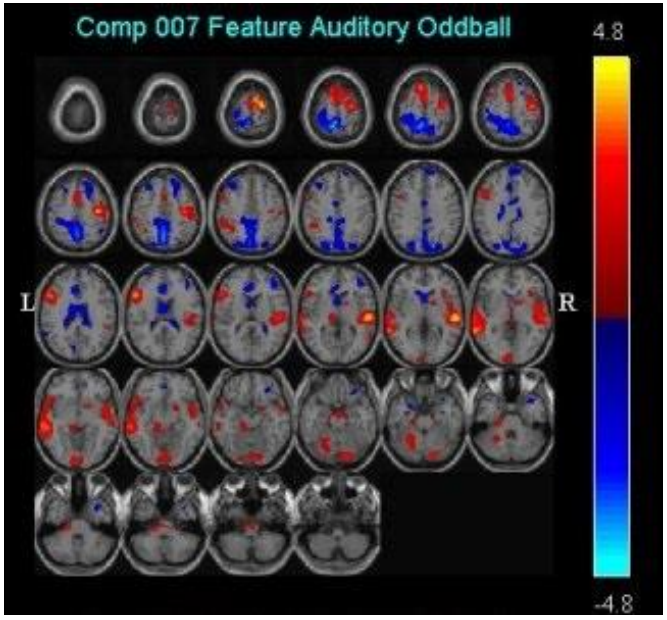
Extrakce proměnných – opakování

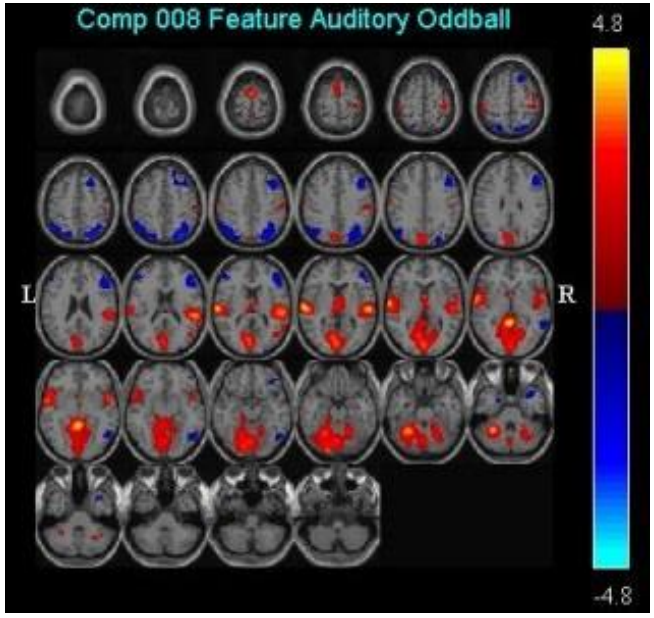
- jednou z možných přístupů redukce dat (vedle selekce)
- transformace původních proměnných na menší počet jiných proměnných $\Rightarrow$ tzn. hledání (optimálního) zobrazení Z, které transformuje původní *p*-rozměrný prostor (obraz) na prostor (obraz) *m*-rozměrný ($m \leq p$)
- pro snadnější řešitelnost hledáme zobrazení Z v oboru lineárních zobrazení
- 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

Analýza nezávislých komponent

Analýza nezávislých komponent (ICA)

Princip: Hledání statisticky nezávislých komponent v původních datech.





Výhody:

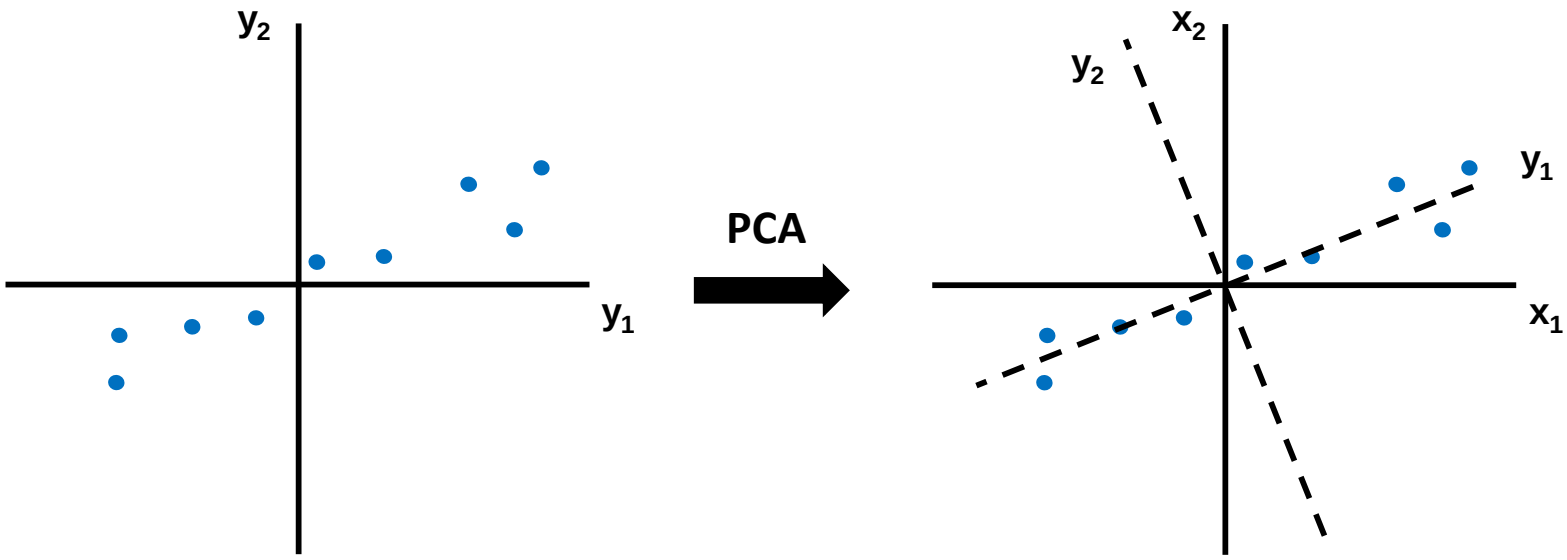
+ analýza na celém mozku, vícerozměrná metoda
+ dokáže vytvořit lépe interpretovatelné komponenty než PCA

Nevýhody:

- velmi časově náročná, předstupněm je redukce pomocí PCA
- je třeba expertní znalost pro výběr komponent

Srovnání s analýzou hlavních komponent (PCA)

Princip: Vytvoření nových proměnných (komponent) z původních proměnných tak, aby zůstalo zachováno co nejvíce variability.



Výhody:

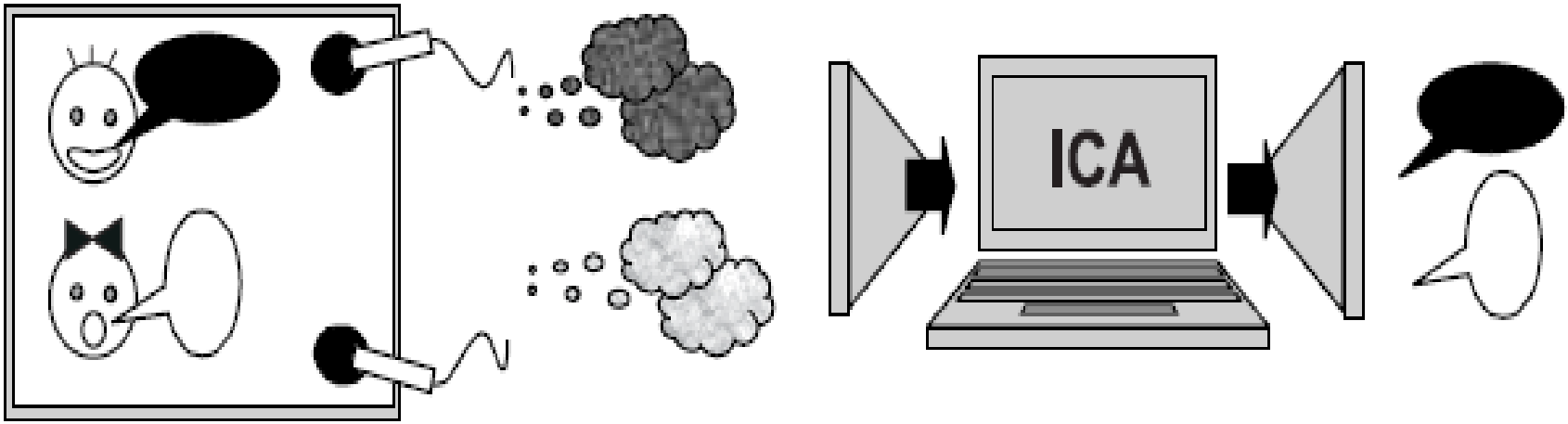
+ analýza na celém mozku
+ vícerozměrná metoda

Nevýhody:

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

Analýza nezávislých komponent

- anglicky *Independent Component Analysis* (ICA)



$$x_1(t) = a_{11}.s_1(t) + a_{12}.s_2(t)$$
$$x_2(t) = a_{21}.s_1(t) + a_{22}.s_2(t)$$

- úloha spočívá v nalezení originálních neznámých signálů z jednotlivých zdrojů $s_1(t)$ a $s_2(t)$ máme-li k dispozici pouze zaznamenané signály $x_1(t)$ a $x_2(t)$
- ICA umožňuje určit koeficienty a_{ij} za předpokladu, že známé signály jsou dány lineárních kombinací zdrojových, a za předpokladu statistické nezávislosti zdrojů v každém čase t

Analýza nezávislých komponent – model dat

- mějme $\mathbf{x} = T(x_1, x_2, \ldots, x_m)$, což je m-rozměrný náhodný vektor

$$x_i = a_{i1}^{orig}.s_1^{orig} + a_{i2}^{orig}.s_2^{orig} + \ldots + a_{im}^{orig}.s_m^{orig}, \; i = 1,2,\ldots,m$$

nebo maticově

$$\mathbf{x} = \mathbf{A}^{orig}.\mathbf{s}^{orig}$$

$\mathbf{s}^{orig}$ je vektor originálních skrytých nezávislých komponent a s_1^{orig} jsou nezávislé komponenty (předpoklad vzájemně statisticky nezávislosti)

$\mathbf{A}^{orig}$ je transformační matice

- skryté nezávislé komponenty je možno vyjádřit pomocí vztahu:

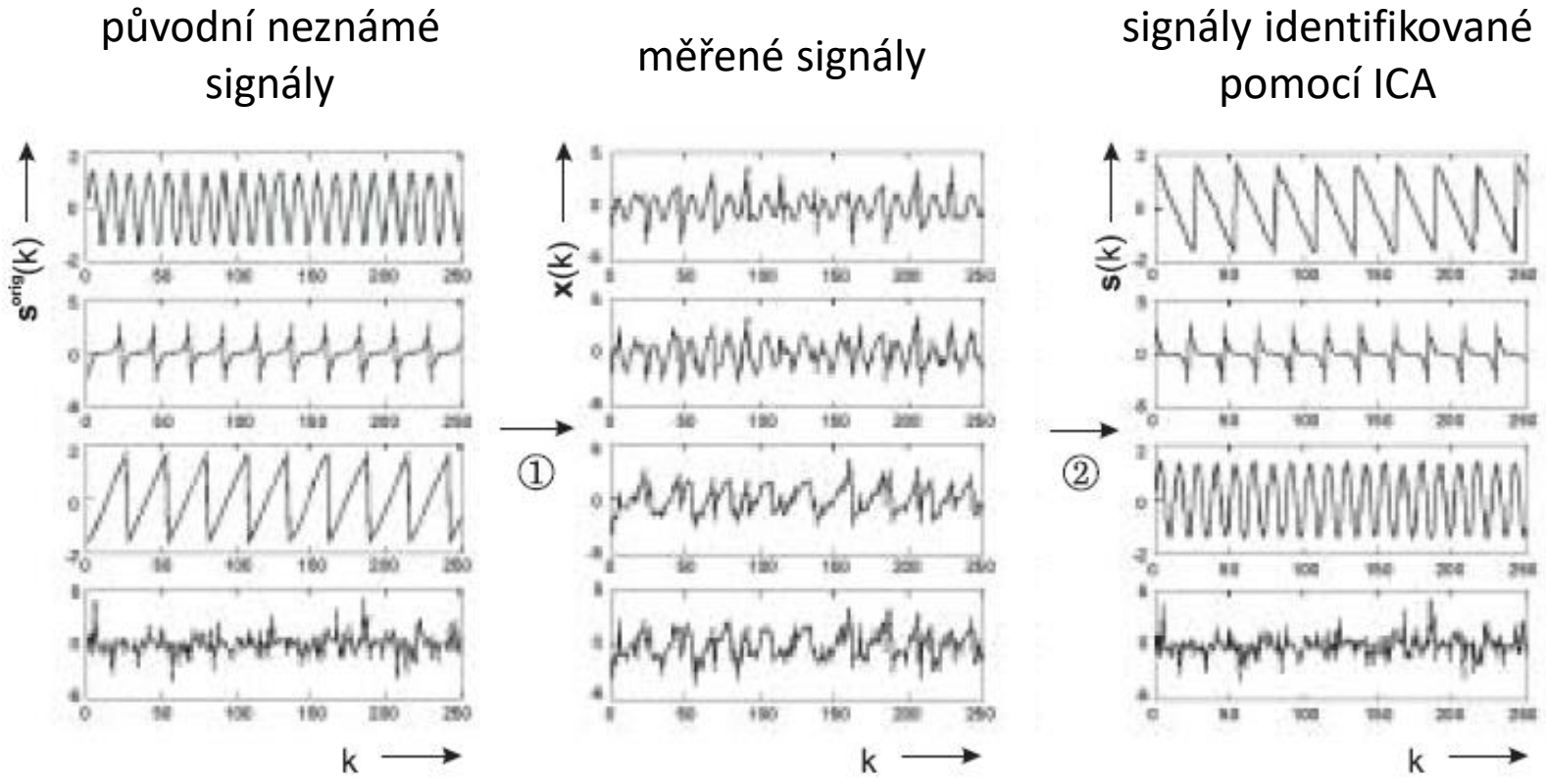
$$\mathbf{s} = \mathbf{W}.\mathbf{x}$$

- cíl: nalézt lineární transformaci (koeficienty transformační matice **W**) tak, aby vypočítané nezávislé komponenty s_i byly vzájemně statisticky nezávislé [$\mathbf{W} = \mathbf{A}^{-1}$]

Analýza nezávislých komponent - omezení

- pouze jedna originální nezávislá komponenta může mít normální rozložení pravděpodobnosti (pokud má více zdrojů normální rozložení, není ICA schopna tyto zdroje ze vstupních dat extrahovat)
- pro dané *m*-rozměrné obrazové vektory je ICA schopna najít pouze *m* nezávislých komponent
- nelze obecně určit polaritu nezávislých komponent
- nelze určit pořadí nezávislých komponent

Analýza nezávislých komponent - omezení



- jsou identifikovány správné původní signály, ale pořadí signálů a jejich polarita je jiná než v původních datech

Odhad nezávislých komponent

- optimalizace pomocí zvolené optimalizační (účelové, kriteriální, objektové) funkce

$$\Downarrow$$

a) nalézt kriteriální funkci

b) vybrat optimalizační algoritmus

ad a) možnost ovlivnit statistické vlastnosti metody

ad b) spojitá optimalizační úloha s „rozumnou“ kriteriální funkcí – gradientní metoda, Newtonova metoda – ovlivňujeme rychlost výpočtu (konvergenci), nároky na paměť,…

Odhad nezávislých komponent – základní úvaha

- nechť existuje *m* nezávislých náhodných veličin s určitými pravděpodobnostními rozděleními (jejich součet za obecných podmínek konverguje s rostoucím počtem sčítanců k normálnímu rozdělení – tzv. centrální limitní věta);
- o vektoru **x** (který máme k dispozici) předpokládáme, že vznikl součtem nezávislých komponent $\mathbf{s}^{orig}$

$$\Downarrow$$

jednotlivé náhodné veličiny x_i mají pravděpodobnostní rozdělení, které je „bližší“ normálnímu než rozdělení jednotlivých komponent s_i^{orig}

- používané míry „nenormality“:
  - koeficient špičatosti
  - negativní normalizovaná entropie
  - aproximace negativní normalizované entropie

Odhad nezávislých komponent – koeficient špičatosti

$$\text{kurt}(s) = \mathcal{E}\{s^4\} - 3(\mathcal{E}\{s^2\})^2$$

- **Gaussovo rozložení má koeficient špičatosti roven nule, zatímco pro jiná rozložení (ne pro všechna) je koeficient nenulový**
- při hledání nezávislých komponent hledáme extrém, resp. kvadrát koeficientu špičatosti veličiny $\mathbf{s} = \mathbf{w}_i.\mathbf{x}$
- **výhody:**
  - rychlost a relativně jednoduchá implementace
- **nevýhody:**
  - malá robustnost vůči odlehlým hodnotám (pokud v průběhu měření získáme několik hodnot, které se liší od skutečných, výrazně se změní KŠ a tím i nezávislé komponenty nebudou odhadnuty korektně)
  - existence náhodných veličin s nulovým KŠ, ale nenormálním rozdělením

Odhad nezávislých komponent – NNE

- Negativní normalizovaná entropie (NNE) = negentropy
- Informační entropie - množství informace náhodné veličiny
- pro diskrétní náhodnou veličinu s je: $H(s) = -\Sigma_i P(s=a_i).\log_2 P(s=a_i)$,
  kde $P(s=a_i)$ je pravděpodobnost, že náhodná veličina S je rovna hodnotě a_i
- pro spojitou proměnnou platí $$H(s) = -\int_{-\infty}^{\infty} p(s)\log_2 p(s)ds$$
- entropie je tím větší, čím jsou hodnoty náhodné veličiny méně predikovatelné
- **pro normální rozd. má entropie největší hodnotu ve srovnání v dalšími rozd.**
- NNE: $J(s) = H(s_{gauss}) - H(s)$, kde s_{gauss} je náhodná veličiny s normálním rozd.
- **výhody:**
  - přesné vyjádření nenormality
  - dobrá robustnost vůči odlehlým hodnotám
- **nevýhody:** časově náročný výpočet $\Rightarrow$ snaha o vhodnou aproximaci NNE, aby byly zachovány její výhody a současně byl výpočet méně náročný

Odhad nezávislých komponent – aproximace NNE

- použití momentů vyšších řádů $$J(s) \approx \frac{1}{12} E\{s^3\}^2 + \frac{1}{48} kurt(s)^2$$

  kde s je náhodná veličina s nulovou střední hodnotou a jednotkovým rozptylem

- **nevýhoda:**
  - opět menší robustnost vůči odlehlým hodnotám

- použití tzv. p-nekvadratických funkcí $$J(s) \approx \sum_{i=1}^{p} k_i . [\mathcal{E}\{G_i(s)\} - \mathcal{E}\{G_i(s_{gauss})\}]^2$$

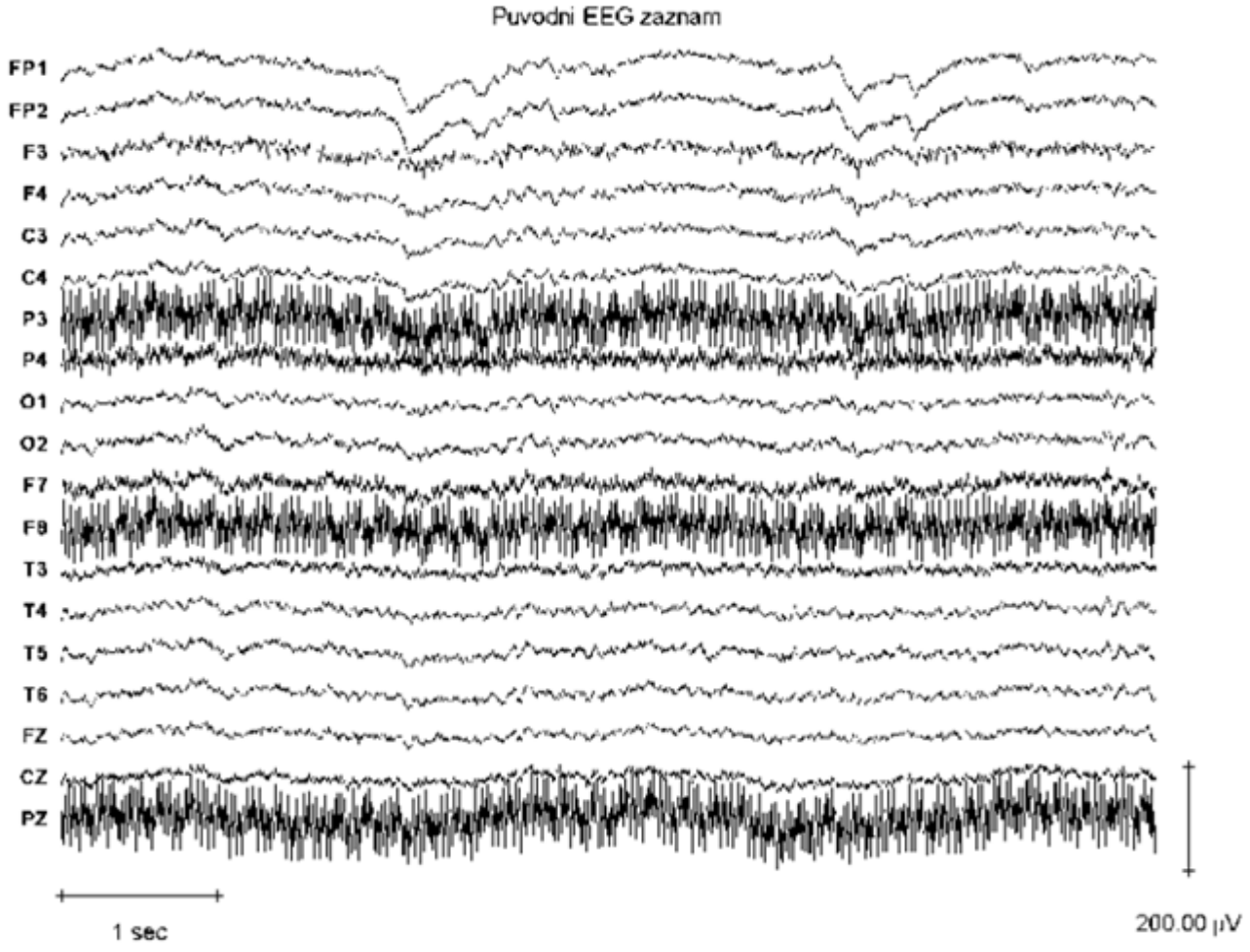
  kde $k_i>0$ je konstanta, G_i jsou šikovně navržené nelineární funkce a s_{gauss} je normální náhodná proměnná, která spolu s s má nulovou střední hodnotu a jednotkový rozptyl.

  Je-li použita pouze jedna funkce G, pak je

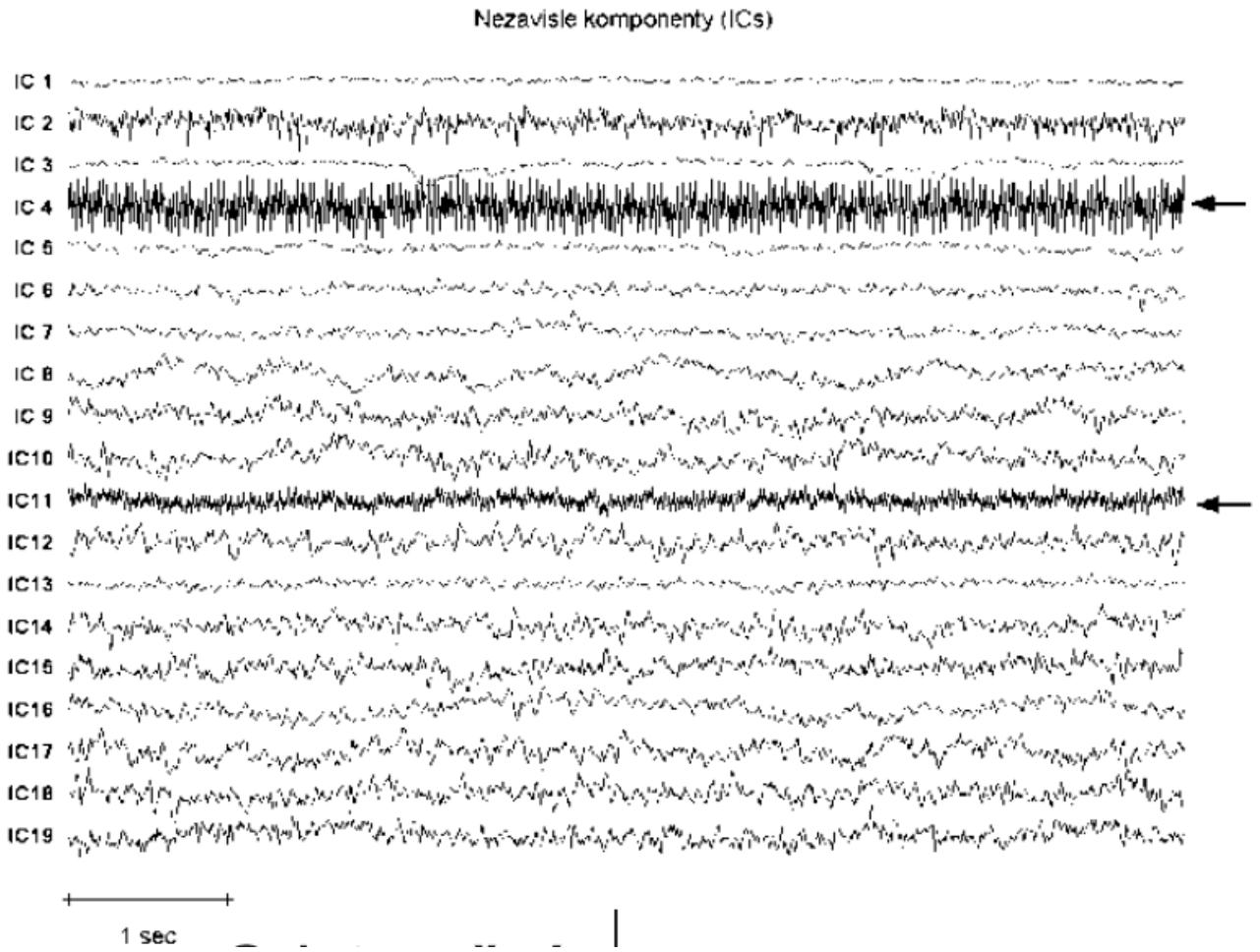
  $$J(s) \approx [\mathcal{E}\{G(s)\} - \mathcal{E}\{G(s_{gauss})\}]^2$$

- doporučuje se $G_1(s) \approx \frac{1}{a_1} \log(\cosh a_1 s)$ kde $a_1 \in \langle 1,2 \rangle$ nebo $G_2(s) \approx -\exp(-s^2/2)$

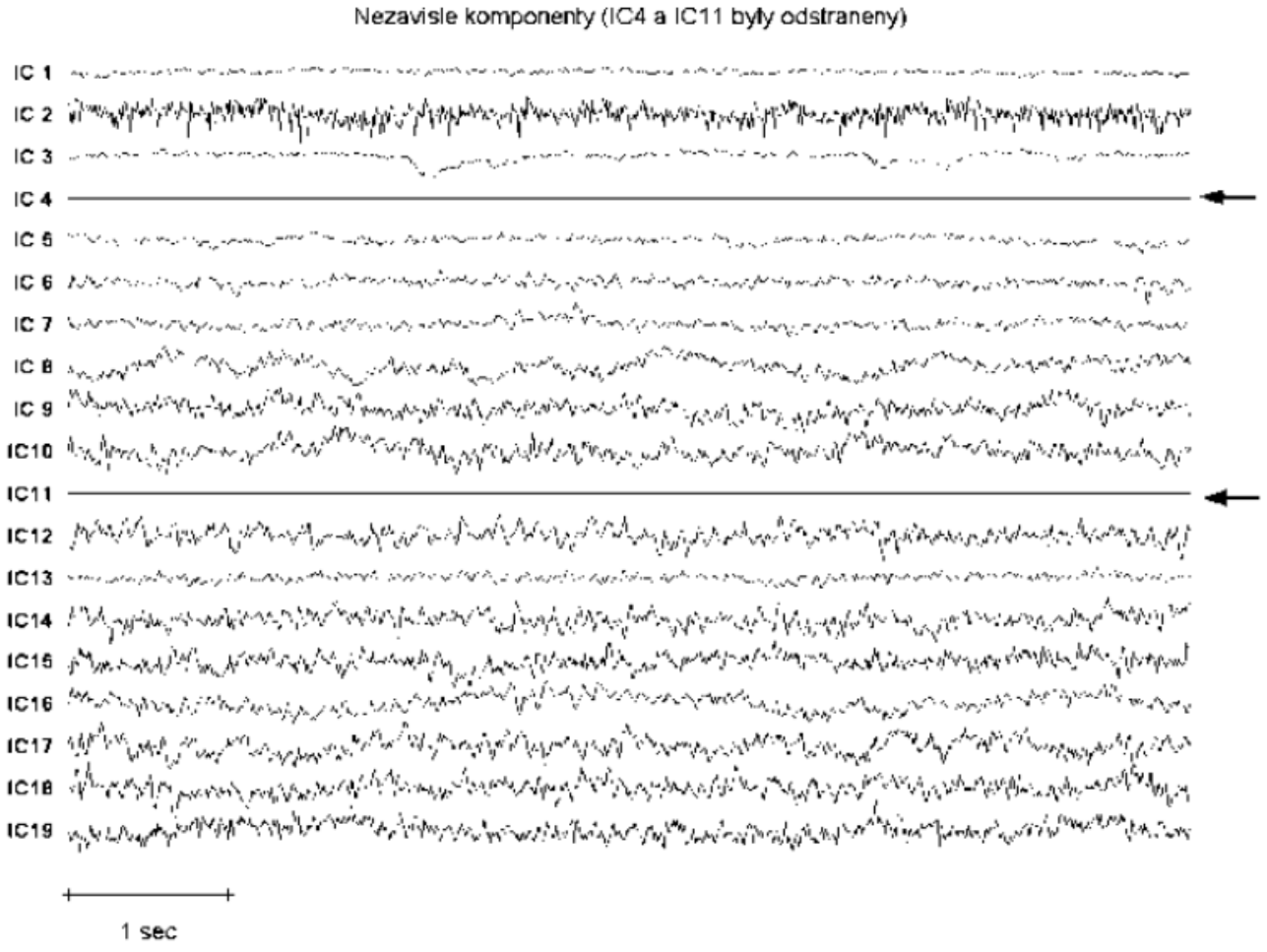
Analýza nezávislých komponent – příklad použití



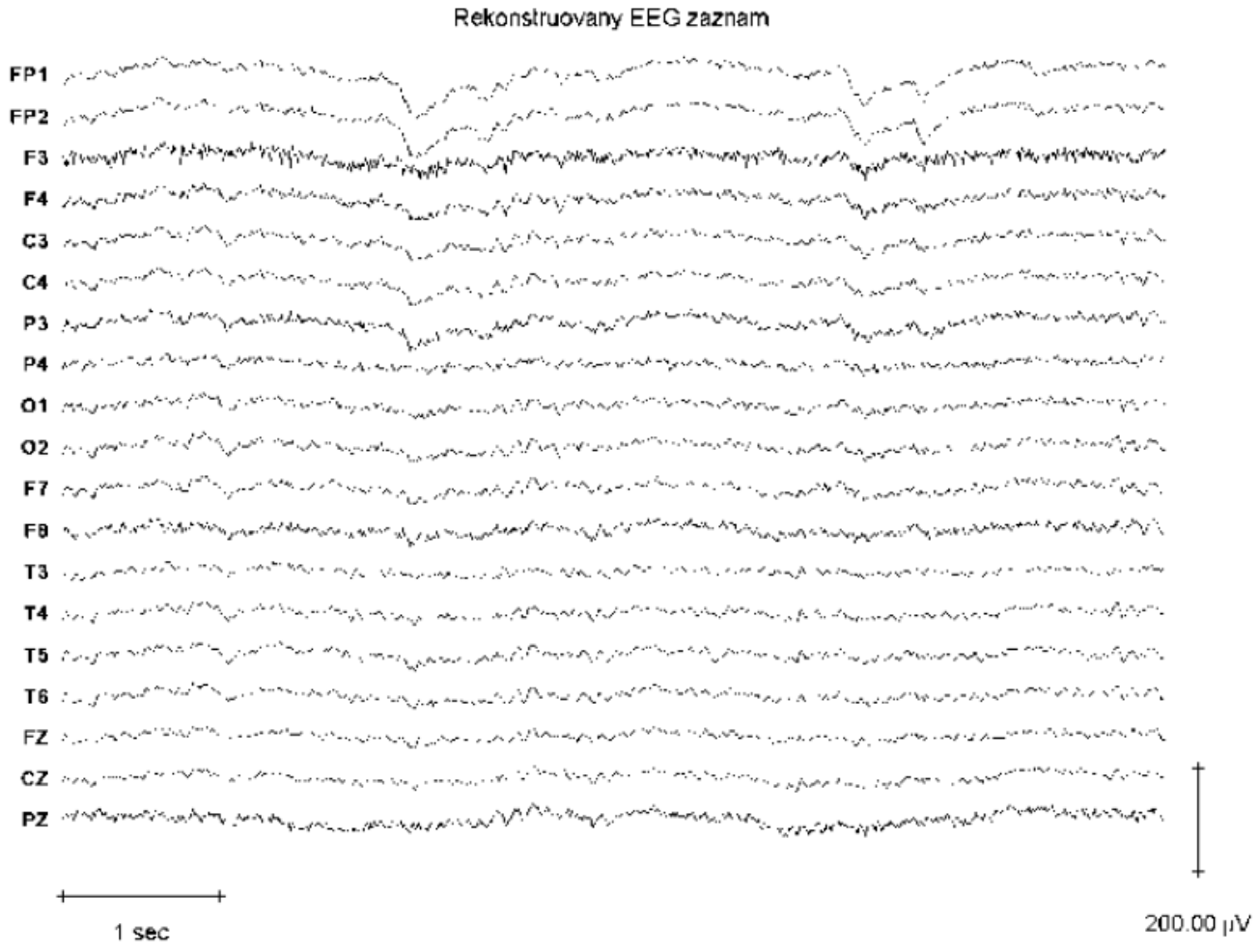
Analýza nezávislých komponent – příklad použití



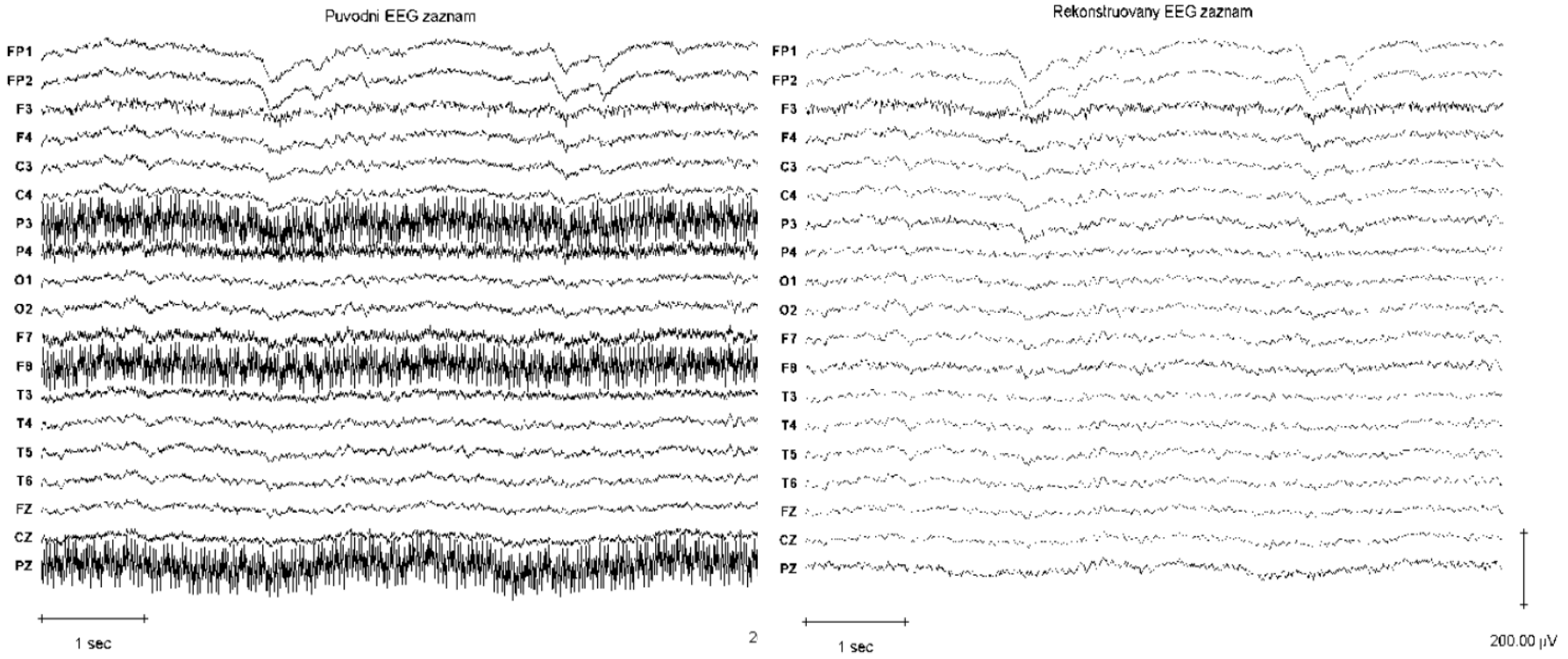
Analýza nezávislých komponent – příklad použití



Analýza nezávislých komponent – příklad použití

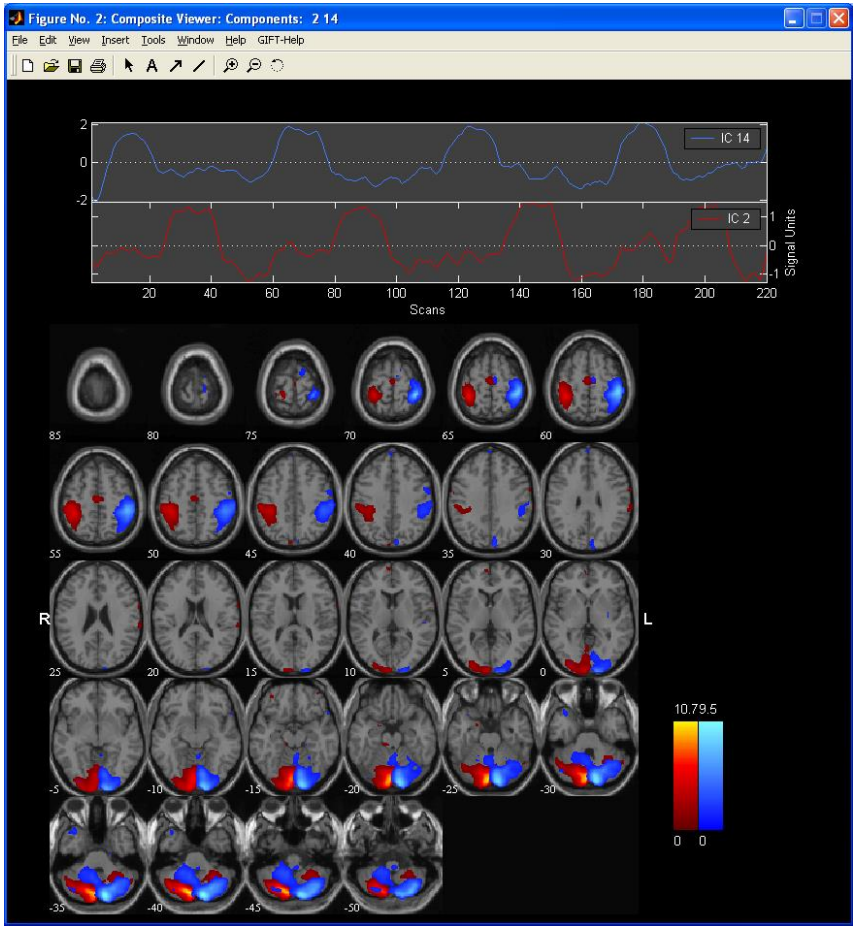


Analýza nezávislých komponent – příklad použití



Analýza nezávislých komponent – příklad 2

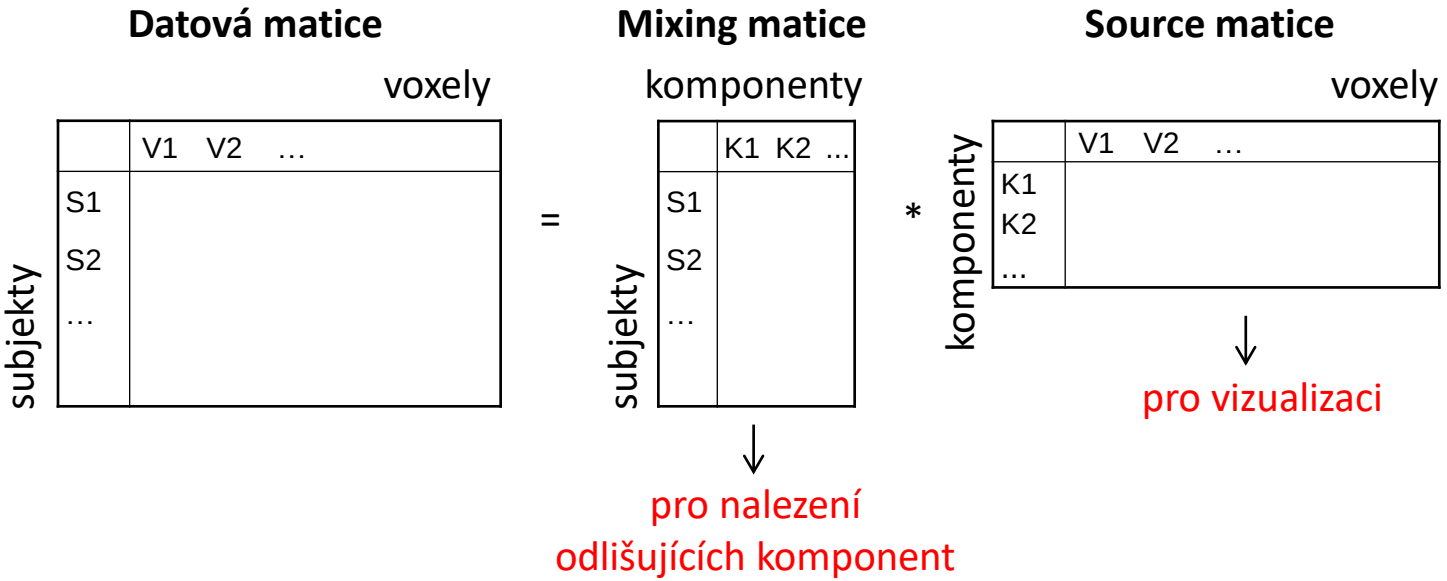
- Zadání: určete nezávislé komponenty ve fMRI datech zdravých subjektů, u nichž byl proveden vizuomotorický test.
- Řešení (s pomocí GIFT toolboxu v software MATLAB)
  http://mialab.mrn.org/software/gift/



Analýza nezávislých komponent – příklad 3

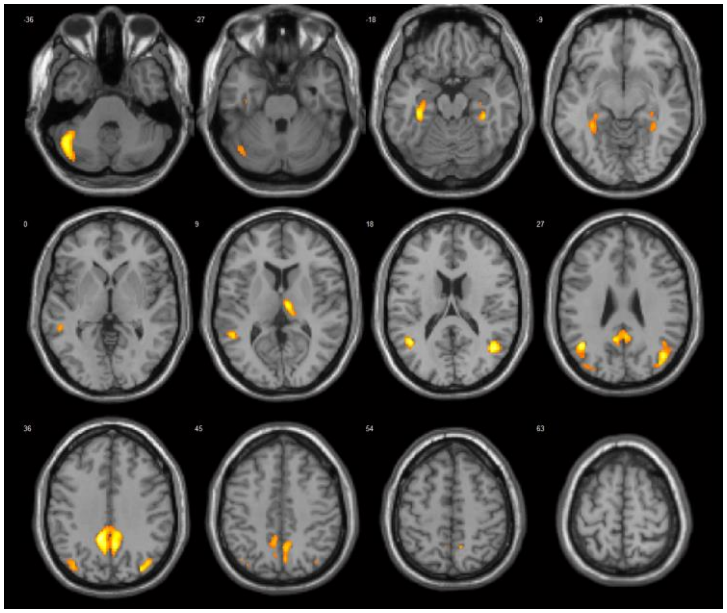
- Zadání: nalezněte nezávislé komponenty, které dokáží odlišit tři skupiny subjektů

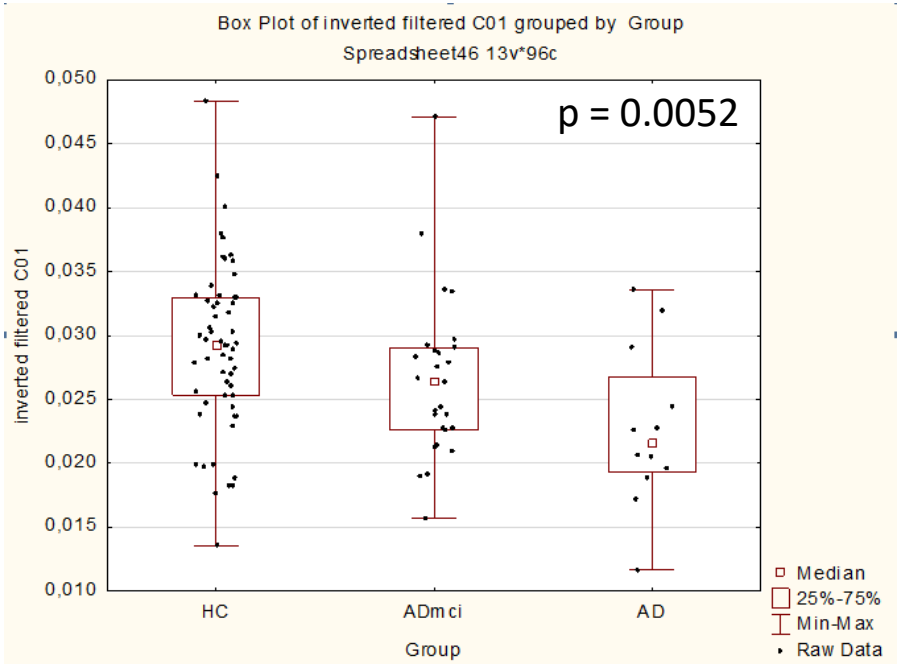
	#N	Age* [years]	Gender F / M	Education* [years]
HC	57	68 (47 – 81)	40 / 17	16 (12 – 21)
ADmci	27	69 (52 – 86)	17 / 10	13 (10 – 22)
AD	12	75 (55 – 88)	11 / 1	12 (8 – 25)



Analýza nezávislých komponent – příklad 3

- komponenta č. 1:

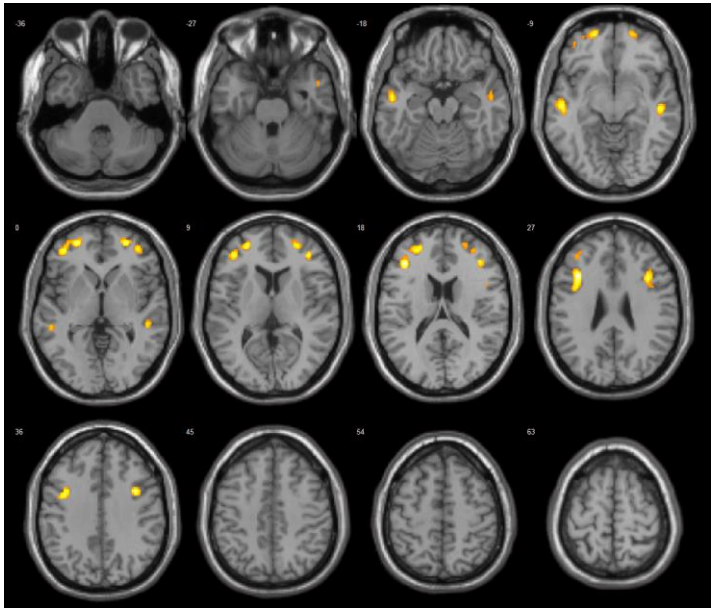


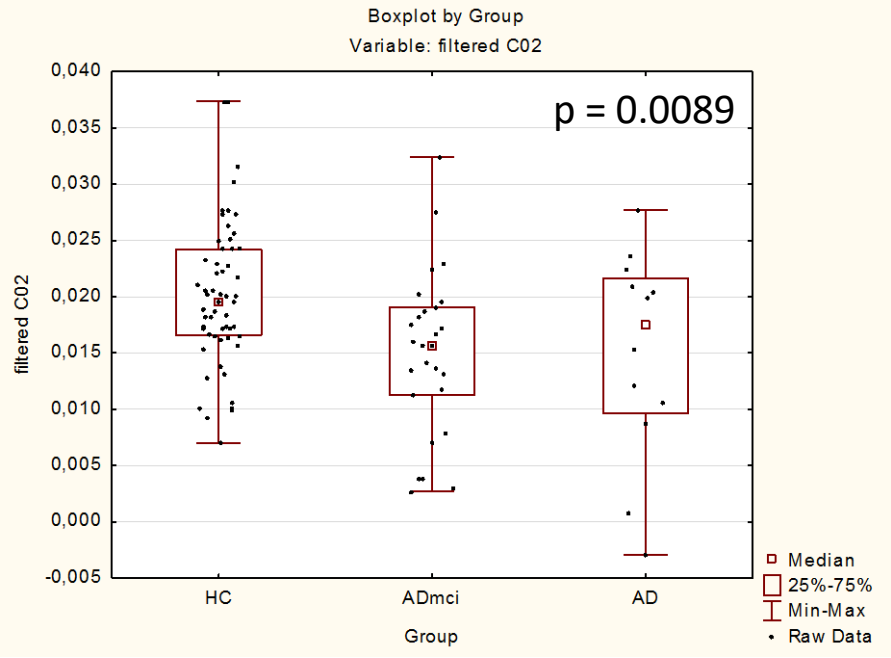


komponenta č.1 ukazuje místa, kde je úbytek šedé hmoty v ADmci a v AD, nicméně v AD větší

Analýza nezávislých komponent – příklad 3

- komponenta č. 2:

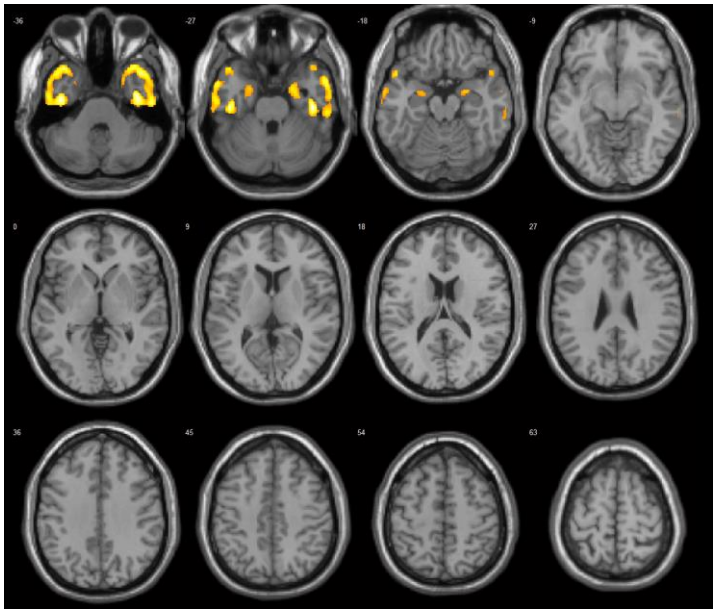


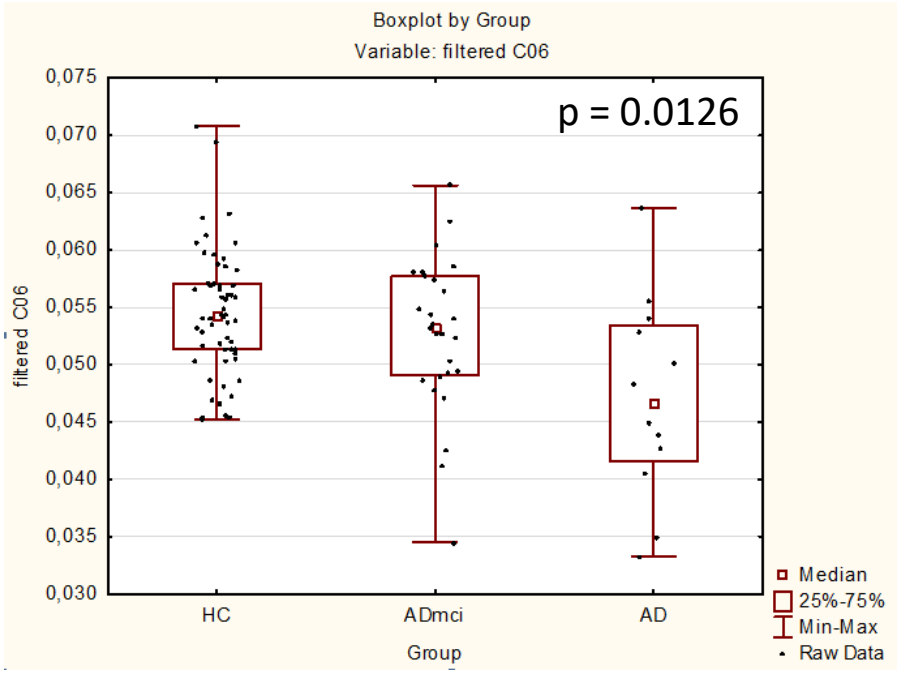


komponenta č.2 ukazuje místa, kde je úbytek šedé hmoty v ADmci a AD víceméně stejný

Analýza nezávislých komponent – příklad 3

- komponenta č. 6:





komponenta č.6 ukazuje místa, kde je úbytek šedé hmoty pouze u AD

Vícerozměrné škálování

Vícerozměrné škálování

- anglicky *Multidimensional Scaling* (MDS)
- přesnější název: nemetrické vícerozměrné škálování
- cíl: dosáhnout řešení, které při nejmenším počtu vytvořených os zachovává pořadí vzdáleností objektů v původní asociační matici
- jde o iterační algoritmus řešící převod libovolné asociační matice do Euklidovského prostoru (různé SW mohou dosahovat mírně odlišné výsledky)

- vstupem analýzy je libovolná asociační matice (včetně nemetrických koeficientů)
- výstupem je zadaný počet „faktorových os“

- pokud je vstupní asociační matice maticí Euklidovských vzdáleností, je MDS totožná s PCA

Vícerozměrné škálování – příklad

- Data vzdáleností evropských měst - > rekonstrukce mapy

STATISTICA - [Data: mesta_vzdalenosti (21v by 24c)]

File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help

Add to Workbook Add to Report Add to MS Word

Arial 10 Vars

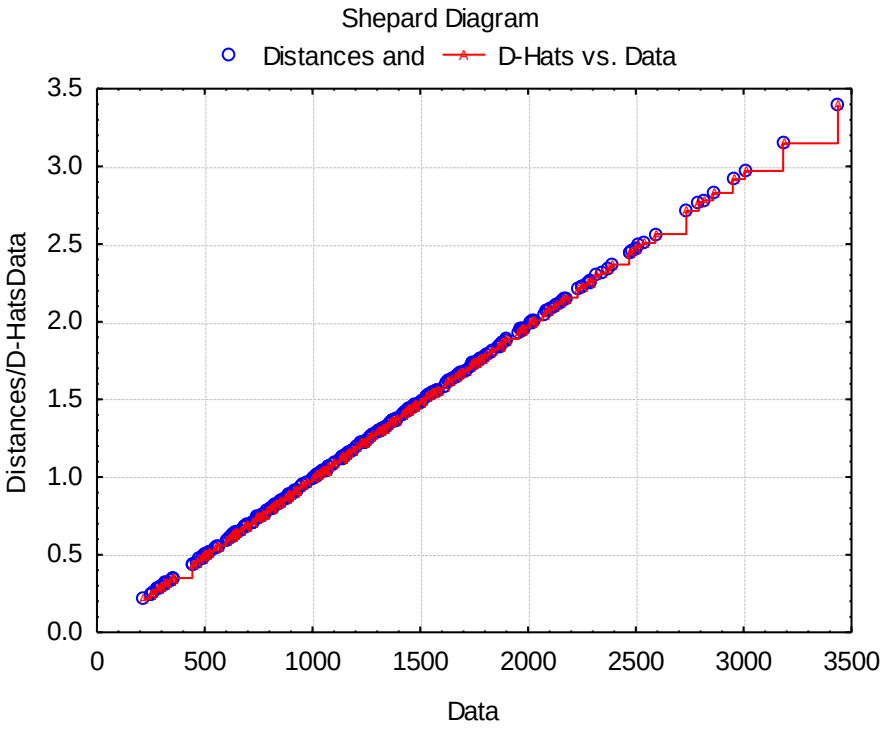
C:\Users\Jarkovsky\Desktop\FSTA\mesta_vzdalenosti.xlsx : Sheet1

	Barcelon	2 Bělehrad	3 Berlín	4 Brusel	5 Bukurešť	6 Budapešť	7 Kodaň	8 Dublin	9 Hamburg	10 Istanbul
Barcelona	0	1528	1497	1062	1968	1498	1757	1469	1471	2230
Bělehrad	1528	0	999	1372	447	316	1327	2145	1229	809
Berlín	1497	999	0	651	1293	689	354	1315	254	1735
Brusel	1062	1372	651	0	1769	1131	766	773	489	2178
Bukurešť	1968	447	1293	1769	0	639	1571	2534	1544	445
Budapešť	1498	316	689	1131	639	0	1011	1894	927	1064
Kodaň	1757	1327	354	766	1571	1011	0	1238	287	2017
Dublin	1469	2145	1315	773	2534	1894	1238	0	1073	2950
Hamburg	1471	1229	254	489	1544	927	287	1073	0	1983
Istanbul	2230	809	1735	2178	445	1064	2017	2950	1983	0
Kiev	2391	976	1204	1836	744	894	1326	2513	1440	1052
Londýn	1137	1688	929	318	2088	1450	955	462	720	2496
Madrid	504	2026	1867	1314	2469	1975	2071	1449	1785	2734
Miláno	725	885	840	696	1331	788	1157	1413	900	1669
Moskva	3006	1710	1607	2253	1497	1565	1558	2792	1779	1753
Mnichov	1054	773	501	601	1186	563	838	1374	610	1582
Paříž	831	1445	876	261	1869	1247	1025	776	744	2253
Praha	1353	738	280	721	1076	443	633	1465	492	1507
Řím	856	721	1181	1171	1137	811	1529	1882	1307	1373
Saint Petersburg	2813	1797	1319	1903	1740	1556	1143	2314	1414	2099
Sofia	1745	329	1318	1697	296	629	1635	2471	1554	502
Stockholm	2276	1620	810	1280	1742	1316	521	1626	809	2171
Vídeň	1347	489	523	914	855	216	868	1680	742	1273
Varšava	1862	826	516	1159	946	545	667	1823	750	1386

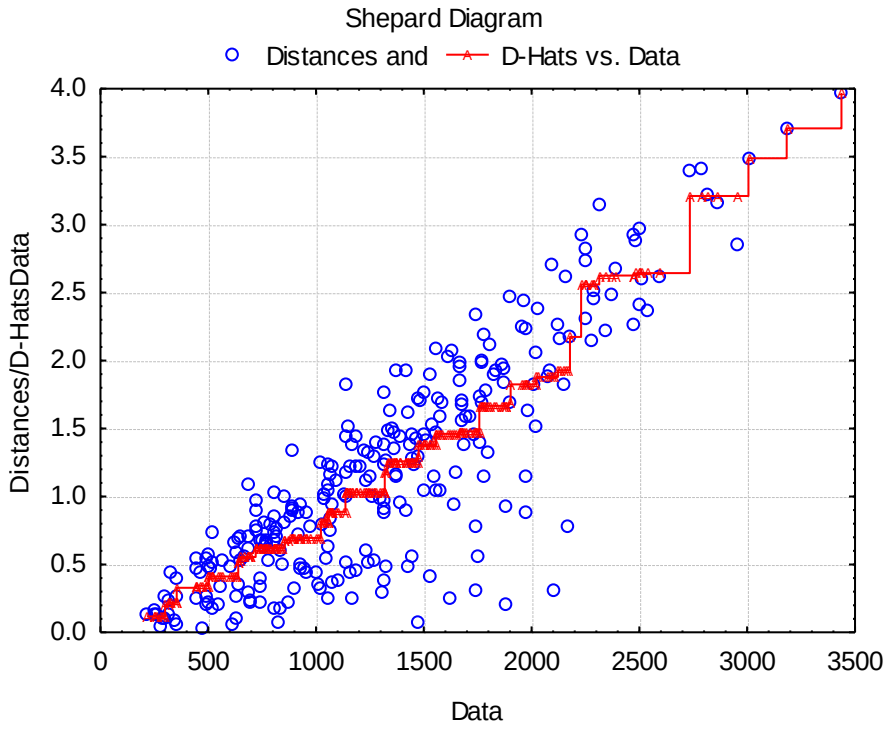
Vícerozměrné škálování – příklad

- kvalita dodržení pořadí vzdáleností v datech při daném počtu os je kontrolována Shepardovým diagramem

2 osy

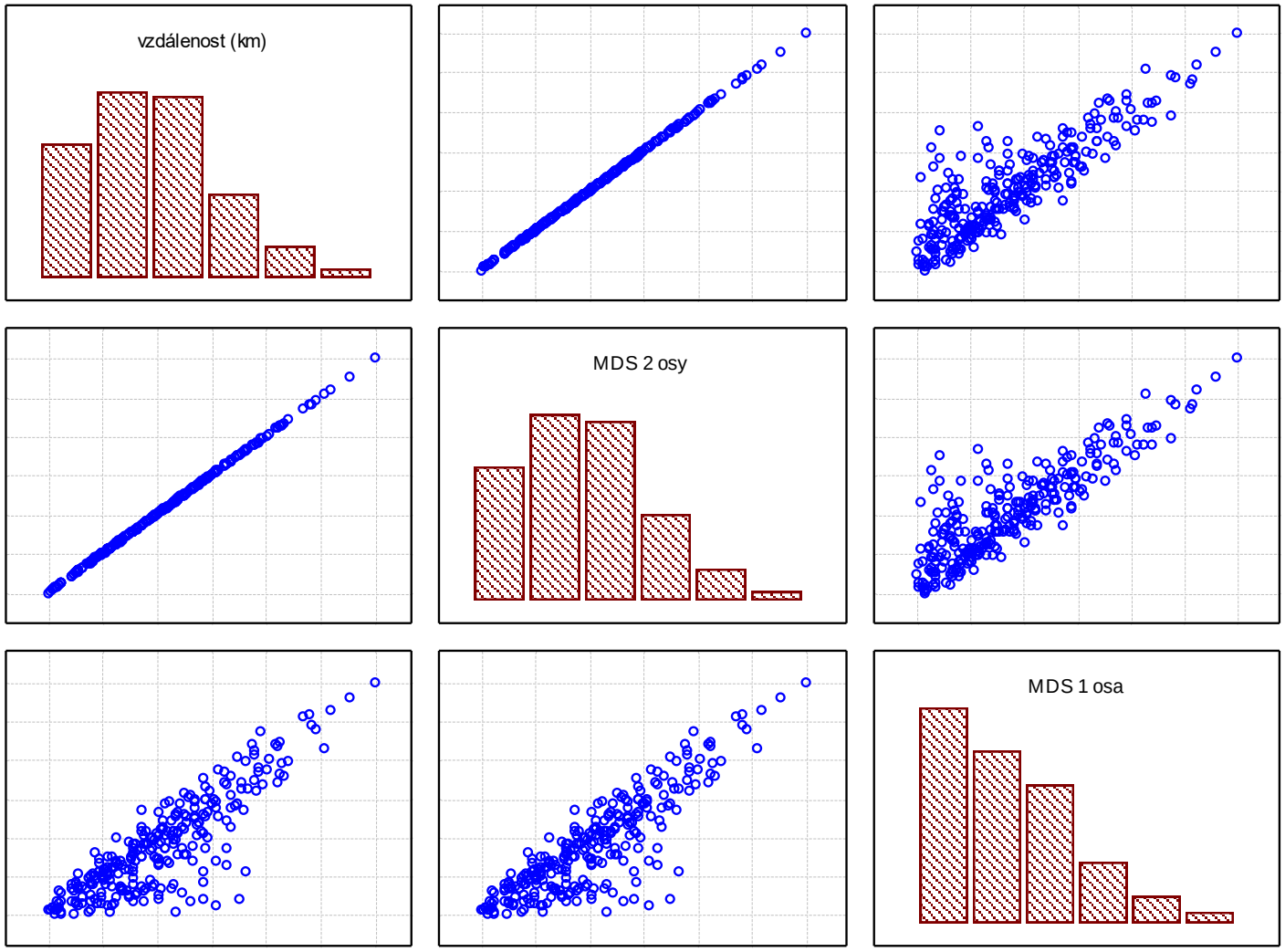


1 osa

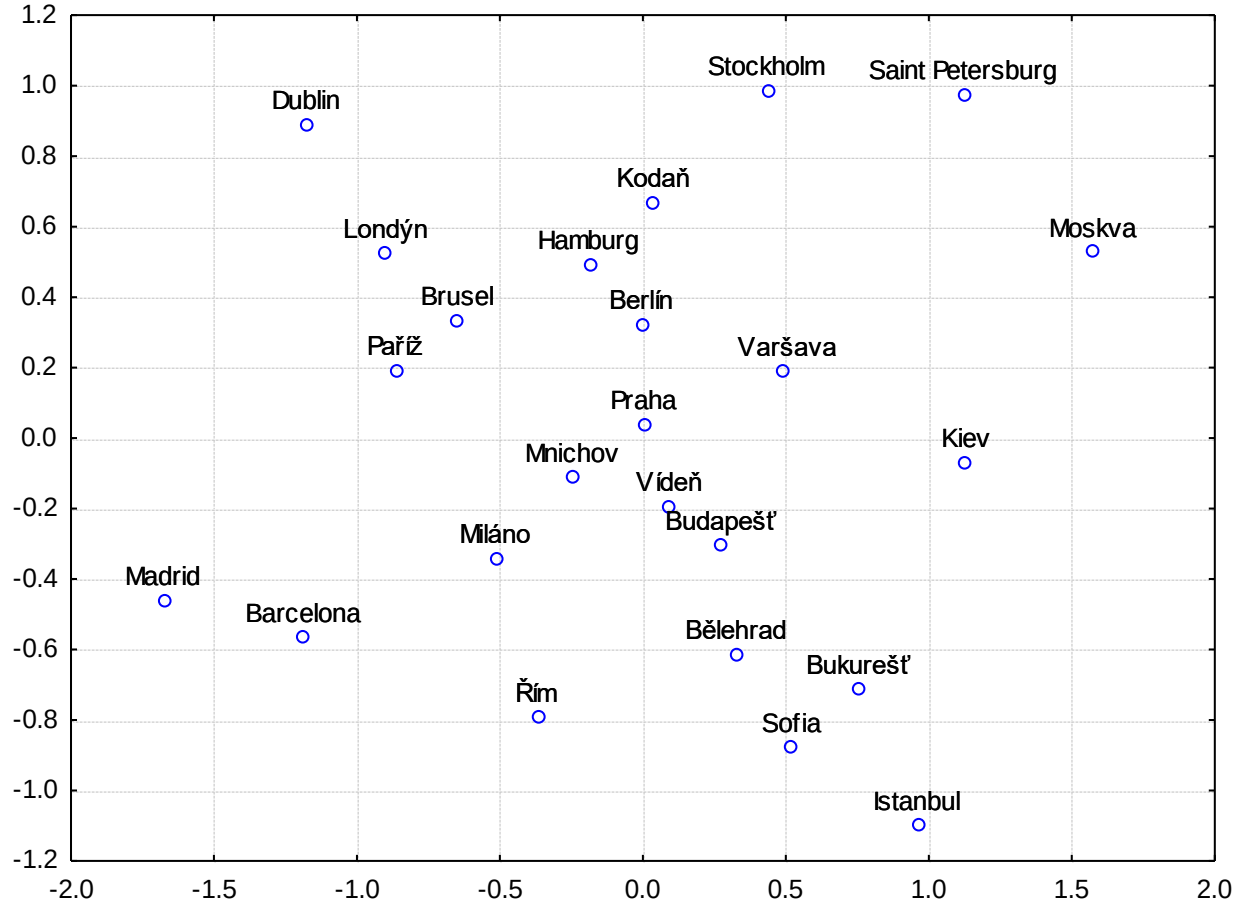


→ jedna osa není dostačující (data příliš daleko od diagonály), zatímco dvě osy jsou v tomto případě dostačující

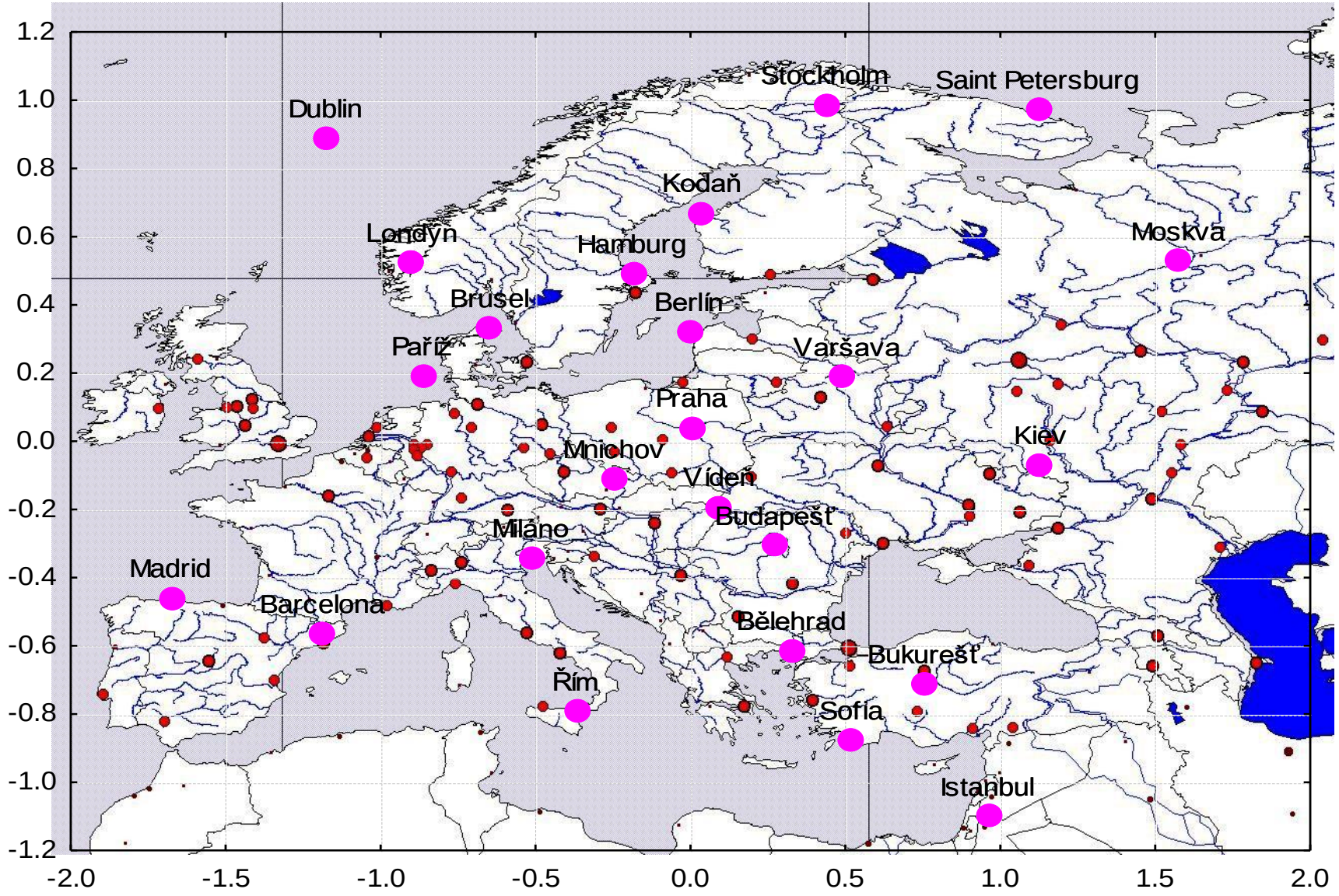
Vzdálenosti v původních datech a vytvořených faktorových osách



Reprezentace výstupu



Reprezentace výstupu



Varietní učení

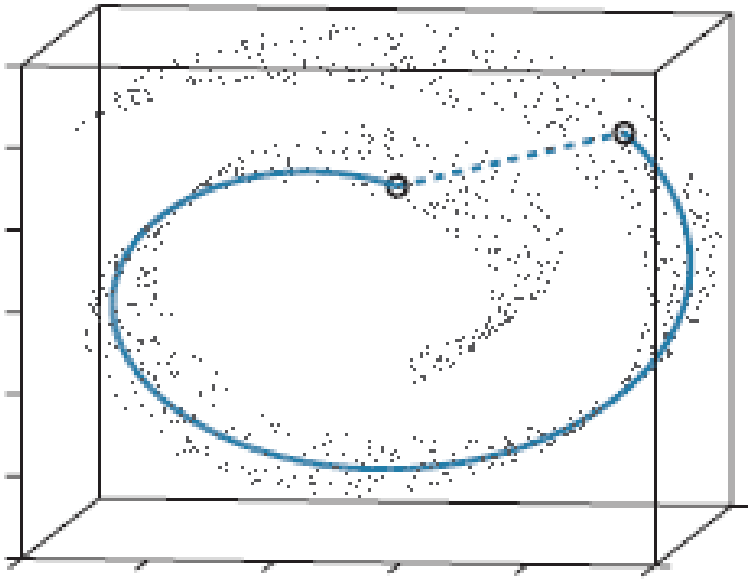
Úvod – redukce dimenzionality

- 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

- anglicky *Manifold Learning*
- 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)

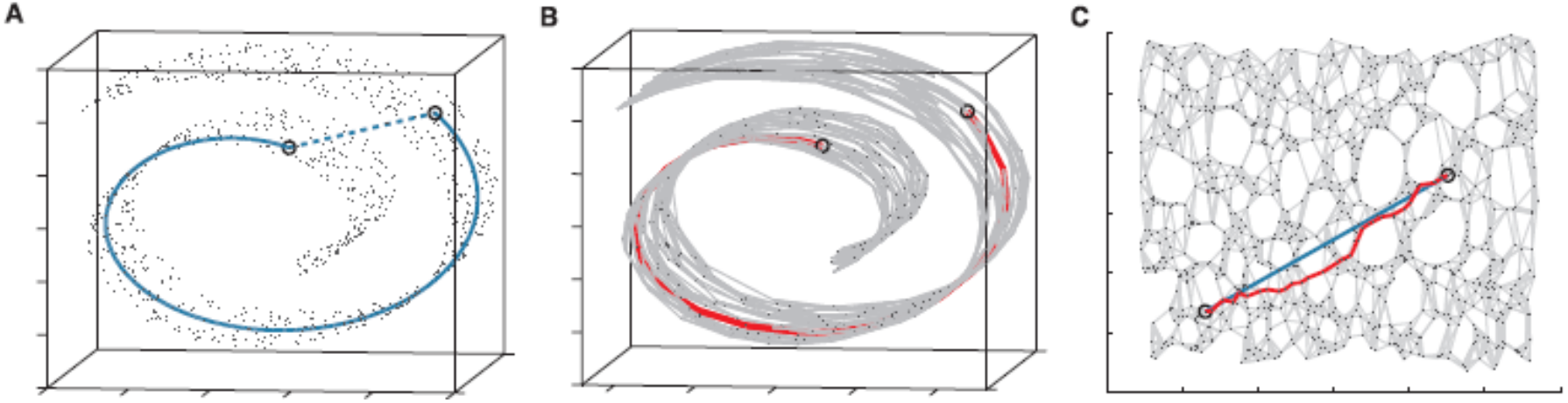
- další metody varietního učení:

  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: $D_G(\mathbf{x}_i, \mathbf{x}_j) = \infty$

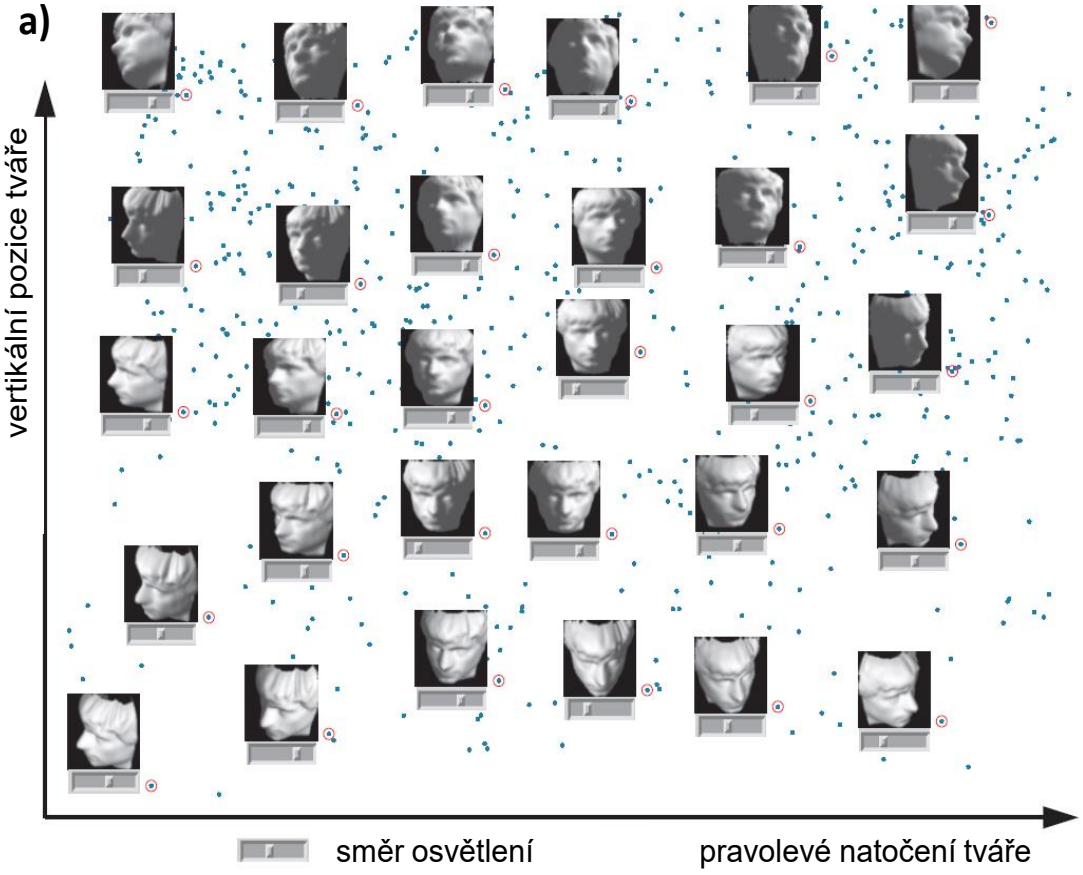
   poté je pro každé $k = 1,2, \ldots, N$ nahrazena vzdálenost $D_G(\mathbf{x}_i, \mathbf{x}_j)$ hodnotou $\min(D_G(\mathbf{x}_i, \mathbf{x}_j), D_G(\mathbf{x}_i, \mathbf{x}_k) + D_G(\mathbf{x}_k, \mathbf{x}_j))$.

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

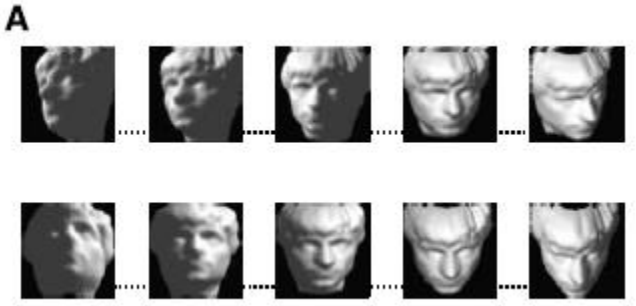
Tenenbaum et al. 2000 Science, A Global Geometric Framework for Nonlinear Dimensionality Reduction

ISOMAP metoda – ukázka 1

Výsledek *k*-ISOMAP algoritmu u 698 obrazů tváří



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

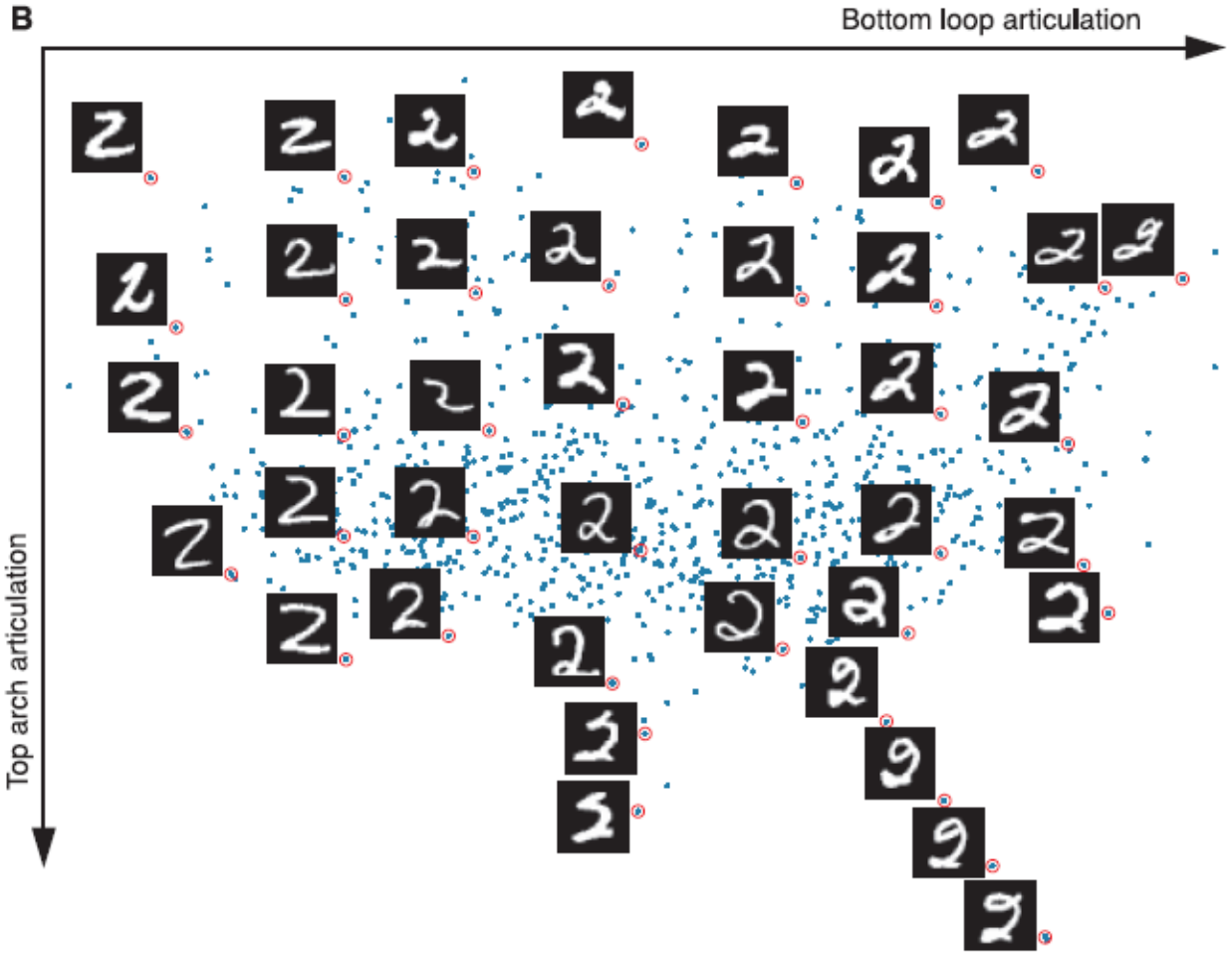


Výsledkem je redukce původních 4096 proměnných (obrazy měly rozměry 64 x 64 pixelů) na pouze tři komponenty

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

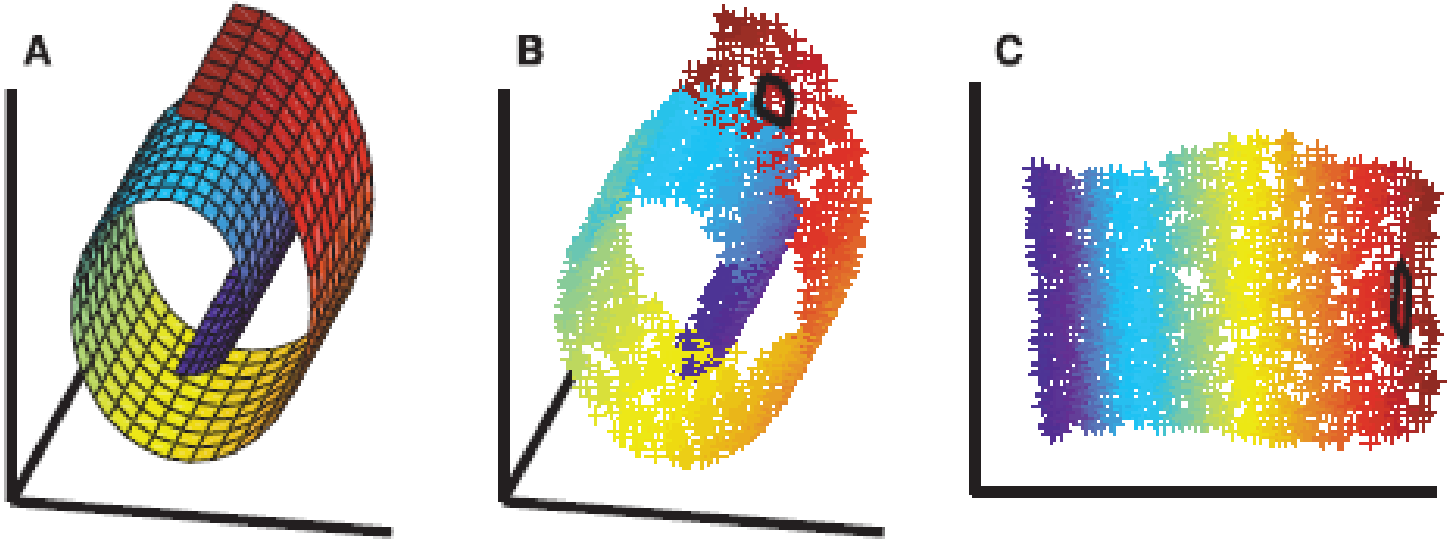


Interpolace podél os *x* a *y* v podprostoru obrazů číslic

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

Metoda lokálně lineárního vnoření (LLE)

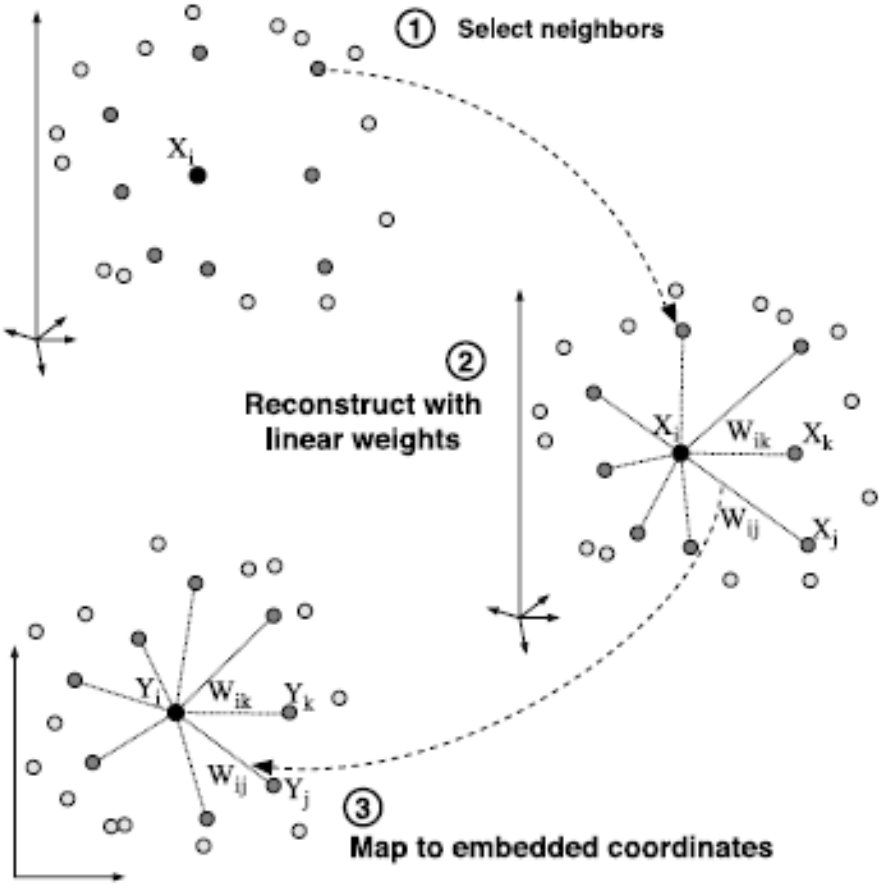
- Locally Linear Embedding (LLE)
- založena na zachování mapování sousedů (neighborhood-preserving mapping)
- LLE rekonstruuje globální nelineární struktury z lokálních lineárních fitů



Černě vyznačeno okolí (sousedi) jednoho bodu.

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

LLE - algoritmus



1. Výběr *k* nejbližších sousedů.

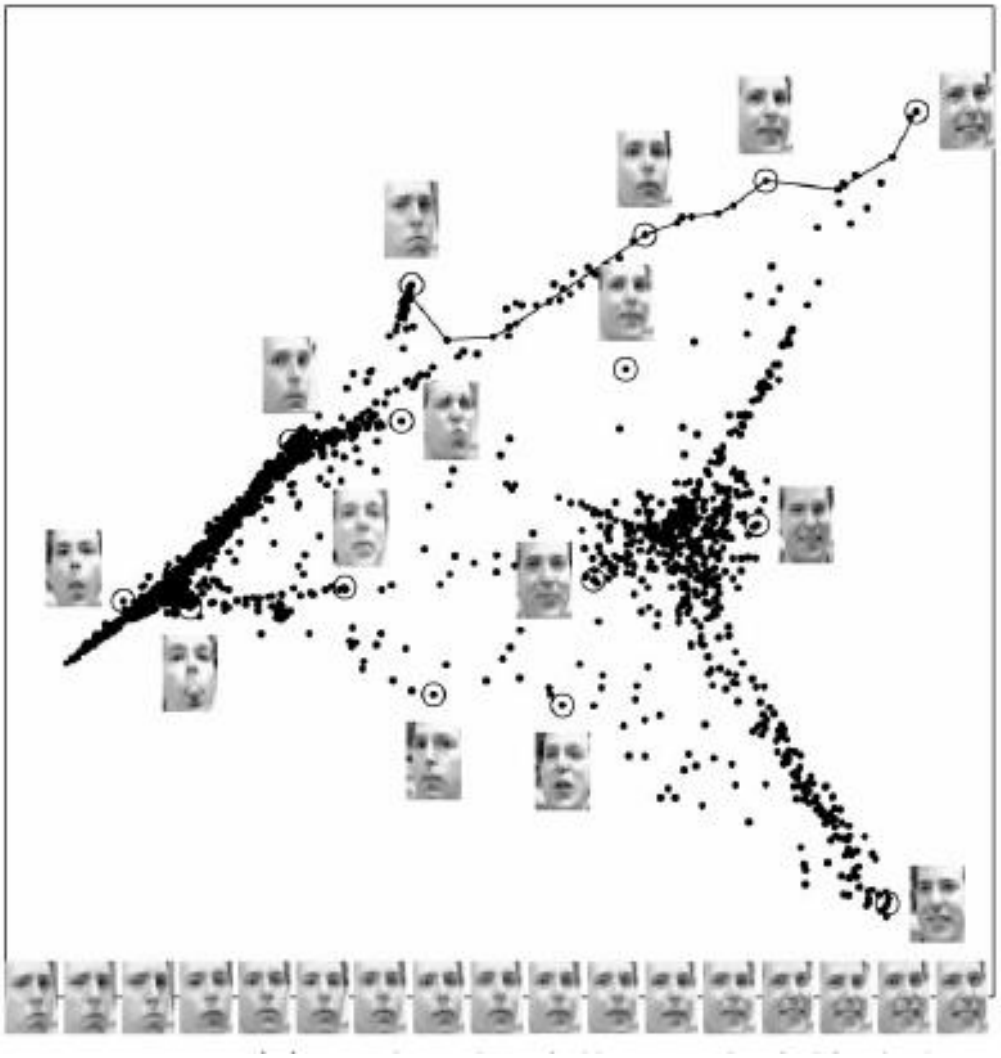
2. Rekonstrukce objektů z jejich sousedů – cílem je nalezení vah W_{ij} tak, aby rekonstrukční chyby byly co nejmenší, tzn. snažíme se minimalizovat výraz $\varepsilon(W) = \sum_i \left|\mathbf{x}_i - \sum_j W_{ij}\,\mathbf{x}_j\right|^2$, přičemž součet vah W_{ij} musí být roven 1; váhy jsou invariantní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

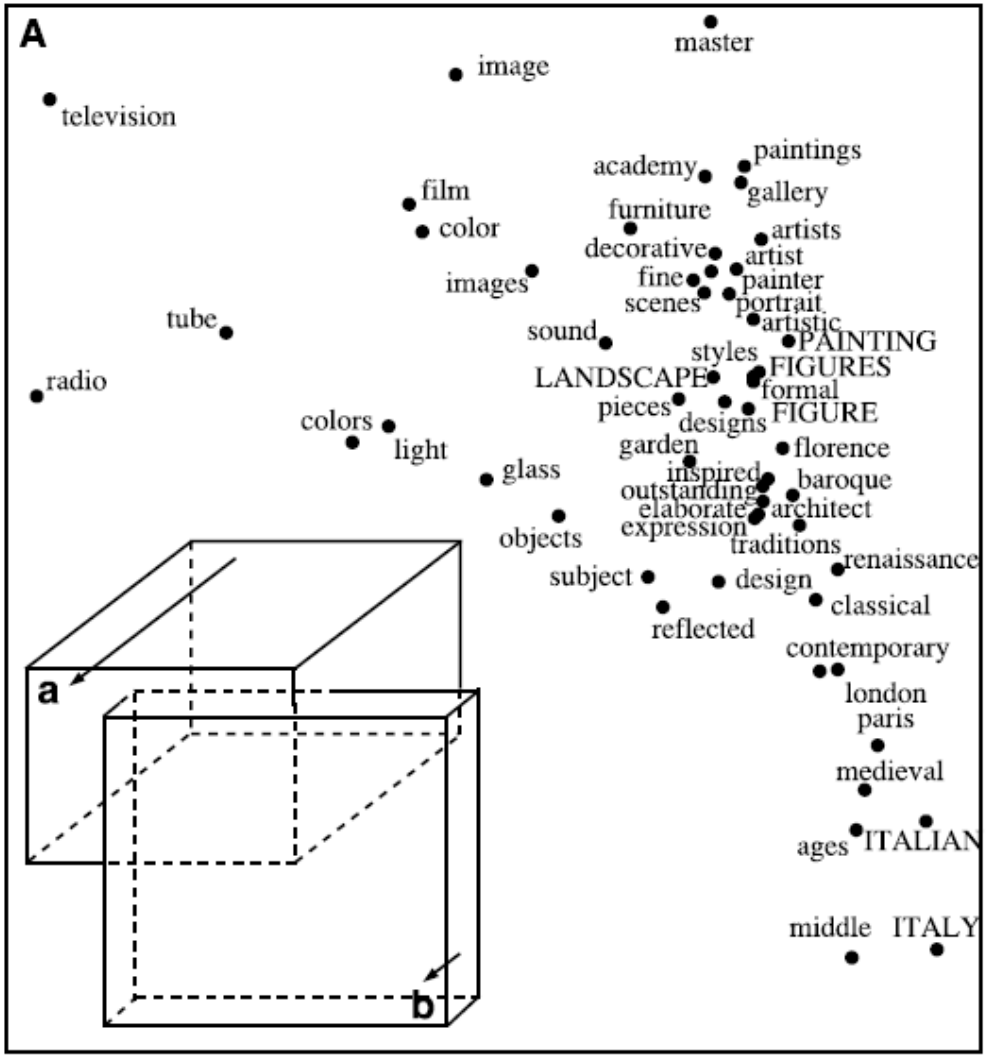
Výsledek LLE algoritmu u obrazů tváří



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)

Korespondenční analýza

Korespondenční analýza

- anglicky *Correspondence Analysis* (CA)
- cíl: nalezení vztahu mezi řádky a sloupci kontingenční tabulky
- vstupní data:
  - tabulka obsahující souhrny proměnných (počty, průměry) za skupiny subjektů/objektů
- výstupy analýzy:
  - vztahy všech původních faktorů a/nebo skupin subjektů v jednoduchém xy grafu
- kritické problémy analýzy
  - skupiny s malým počtem hodnot mohou být zatíženy značným šumem a náhodnou chybou
  - obtížná interpretace velkého množství malých skupin subjektů

- Výpočet probíhá prostřednictvím rozkladu na singulární hodnoty (*singular value decomposition*) na matici chí-kvadrát vzdáleností (tedy na matici příspěvků buněk tabulky k celkovému chí-kvadrátu obdobně jako v klasickém testu dobré shody na kontingenční tabulce)

Analýza kontingenčních tabulek jako princip výpočtu vícerozměrných analýz

- Počet pacientů s nežádoucími účinky na typu léčby lze brát jako kontingenční tabulku a mírou vztahu mezi řádky (typ léčby – lék A, lék B) a sloupci (nežádoucí účinky – ano, ne) je velikost chí-kvadrátu

$$\chi^2_{(1)} = \frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}$$

Počítáno pro každou buňku tabulky

	☠	☺
A	10	0
B	0	10

Pozorovaná tabulka

	☠	☺
A	5	5
B	5	5

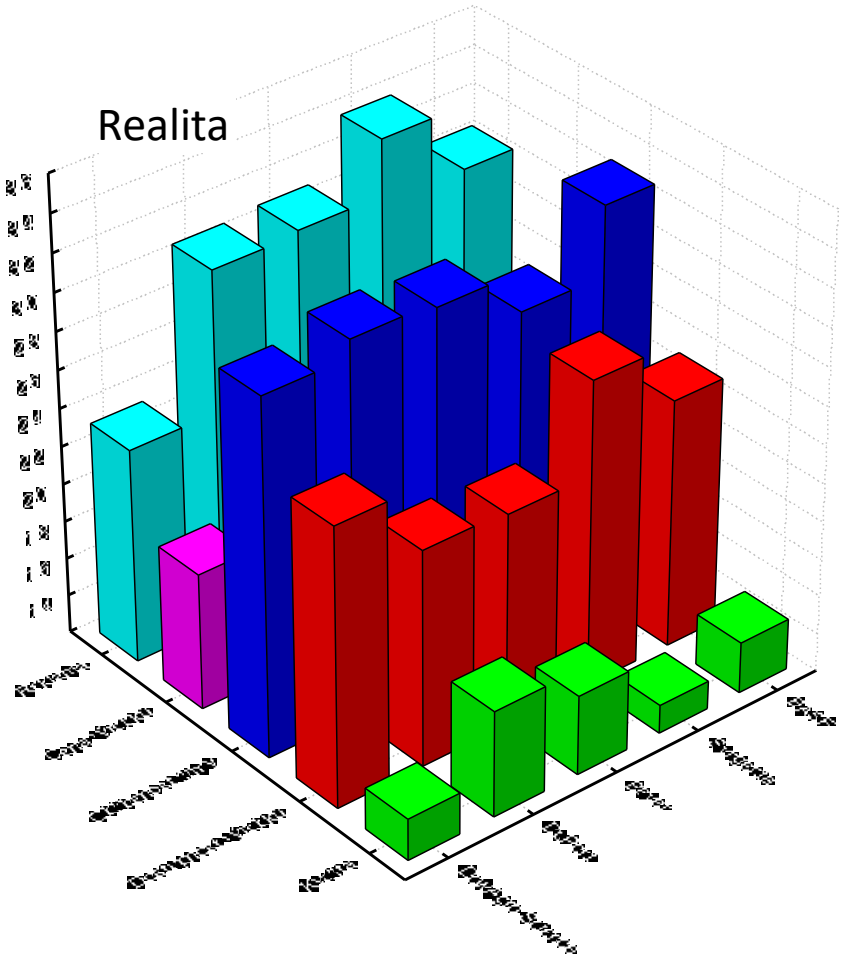
Očekávaná tabulka

Hodnota chí-kvadrátu definuje míru odchylky dané buňky (v našem kontextu vztahu nežádoucích účinků a typu léčby) od situace, kdy mezi řádky a sloupci (nežádoucími účinky a typem léčby) není žádný vztah

Princip korespondenční analýzy

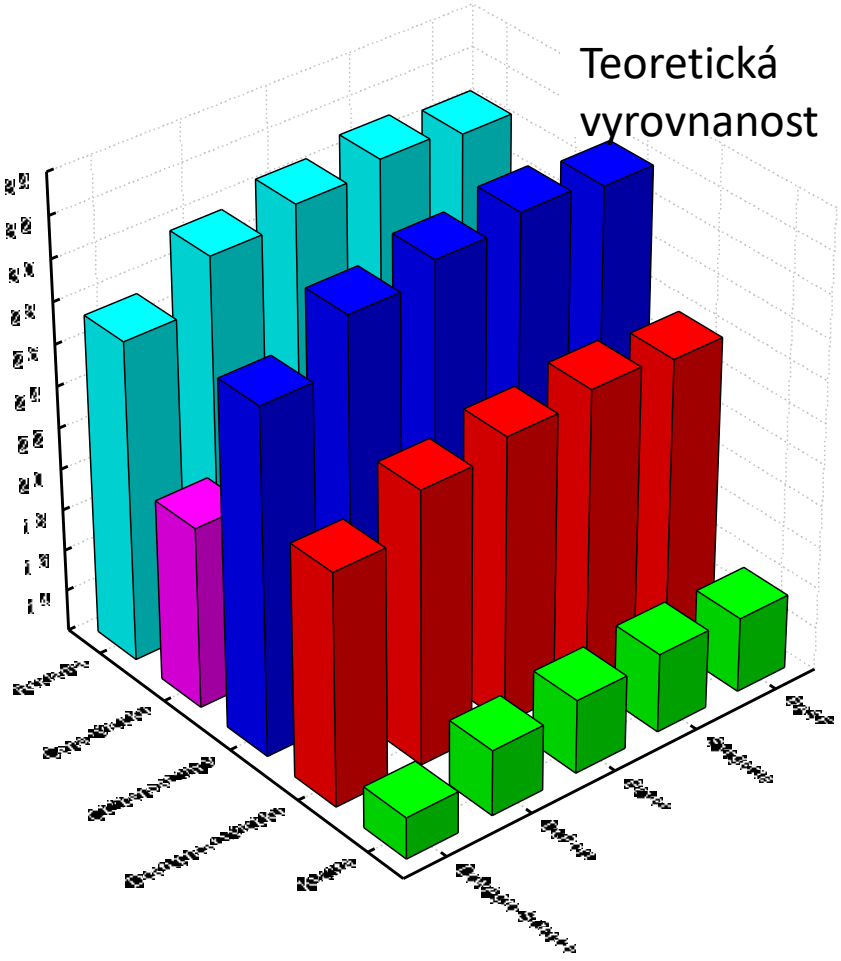
- Korespondenční analýza hledá, které kombinace řádků a sloupců hodnocené tabulky nejvíce přispívají k její variabilitě

Realita



Vs.

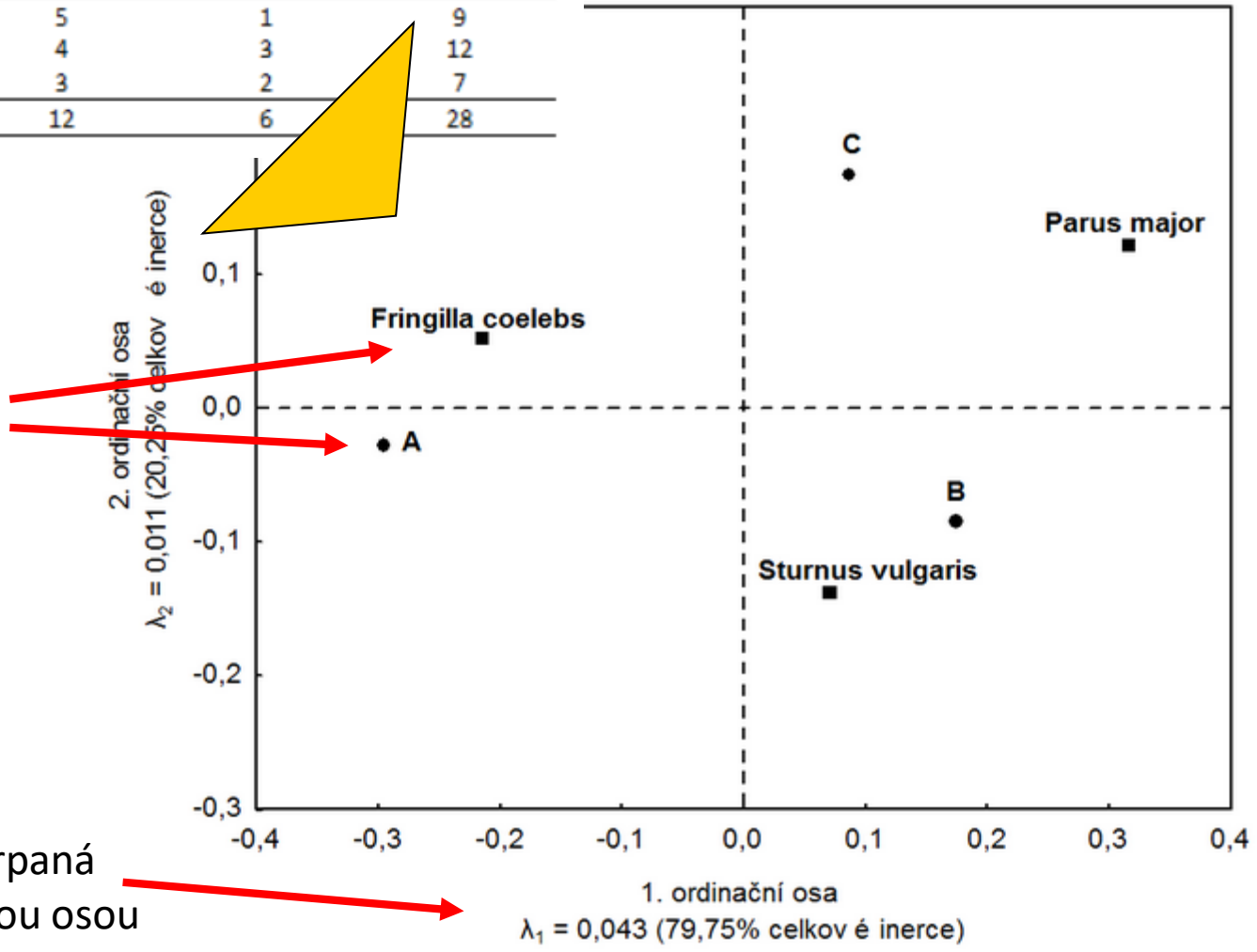
Teoretická vyrovnanost



Výstupy korespondenční analýzy

Tabulka 1: Zastoupení třech druhů ptáků na třech lokalitách.

	Druh 1 *Sturnus vulgaris*	Druh 2 *Fringilla coelebs*	Druh 3 *Parus major*	Celkem
Lokalita A	3	5	1	9
Lokalita B	5	4	3	12
Lokalita C	2	3	2	7
Celkem	10	12	6	28



Vzájemná pozice faktorů a skupin objektů/subjektů: vzájemnou pozici lze interpretovat

Variabilita vyčerpaná danou faktorovou osou

Kanonická korelační analýza

Kanonická korelační analýza

- anglicky *Canonical Correlation Analysis* (CCorA)
- cíl: nalezení maximální lineární korelace mezi dvěma sadami proměnných (tzn. zjištění, zda se jedna skupina proměnných chová stejně jako druhá skupina proměnných pro ty samé objekty, a pokud ano, co je podstatou této shody)
- vstupem do CCorA dvě matice:
  - se vzájemně závislými proměnnými
  - nebo jedna matice se závisle proměnnými a jedna s nezávisle proměnnými (v tom případě velmi podobné jako RDA)
- princip: CCorA hledá lineární kombinaci proměnných z první sady a lineární kombinaci proměnných z druhé sady, které mají maximální korelaci mezi sebou
- CCorA je zobecněním vícerozměrné lineární regrese, která hledá závislost pouze jedné závisle proměnné na sadě nezávislých proměnných
- příklad použití: hledání vztahu skupiny rizikových faktorů a skupiny symptomů nemoci

Kanonická korelační analýza – předpoklady

- data musí být kvantitativní
- data nesmí obsahovat odlehlé hodnoty (proměnné ale nemusí mít nutně normální rozdělení)
- počet proměnných první sady plus počet proměnných druhé sady musí být menší než počet objektů
- proměnné musí mít mezi sebou lineární vztah (ne nelineární)

Redundanční analýza

Redundanční analýza

- anglicky *Redundancy Analysis* (RDA)
- cíl: zjištění závislosti jedné skupiny proměnných na druhé skupině proměnných
- vhodná v případech, kdy mají dvě sady proměnných lineární vztah
- dává podobné výsledky jako kanonická korelační analýza
- princip: RDA je v podstatě vícerozměrnou regresní analýzou, která je následovaná analýzou hlavních komponent
- předpoklady: stejné jako u PCA

Redundanční analýza – označení a postup

Označení: **X** - matice nezávisle proměnných; **Y** - matice závisle proměnných

Postup:

1. regrese každé závisle proměnné Y_i na sadě nezávislých proměnných **X** pomocí vícerozměrné regrese a získání regresních koeficientů
2. PCA na sadě regresních koeficientů z vícerozměrné regrese a získání matice kanonických vlastních vektorů
3. použití kanonických vlastních vektorů k získání skóre objektů buď ve faktorovém prostoru **X** (skóre označovány jako lineární kombinace), nebo v prostoru závislých proměnných **Y** (skóre označovány jako vážené průměry)

Metoda parciálních nejmenších čtverců

Metoda parciálních nejmenších čtverců

- anglicky *Partial Least Squares* (PLS)
- cíl: zjištění vztahu (kovariance) mezi dvěma sadami proměnných (např. mezi funkčními obrazovými daty a behaviorálními daty)
- lze rovněž srovnávat skupiny mezi sebou – lze srovnat i více skupin (při porovnávání více skupin nedetekuje pouze rozdílné patterny mezi skupinami, ale i podobné či stejné)
- vhodné i pouze jako doplňková analýza, dokonce se doporučuje, aby byla v kombinaci s nějakým dalším typem analýzy
- reference pro využití PLS v neurozobrazování: McIntosh, A.R., Bookstein, F., Haxby, J., Grady, C., 1996. Spatial pattern analysis of functional brain images using partial least squares. Neuroimage 3, 143–157

PLS – metody

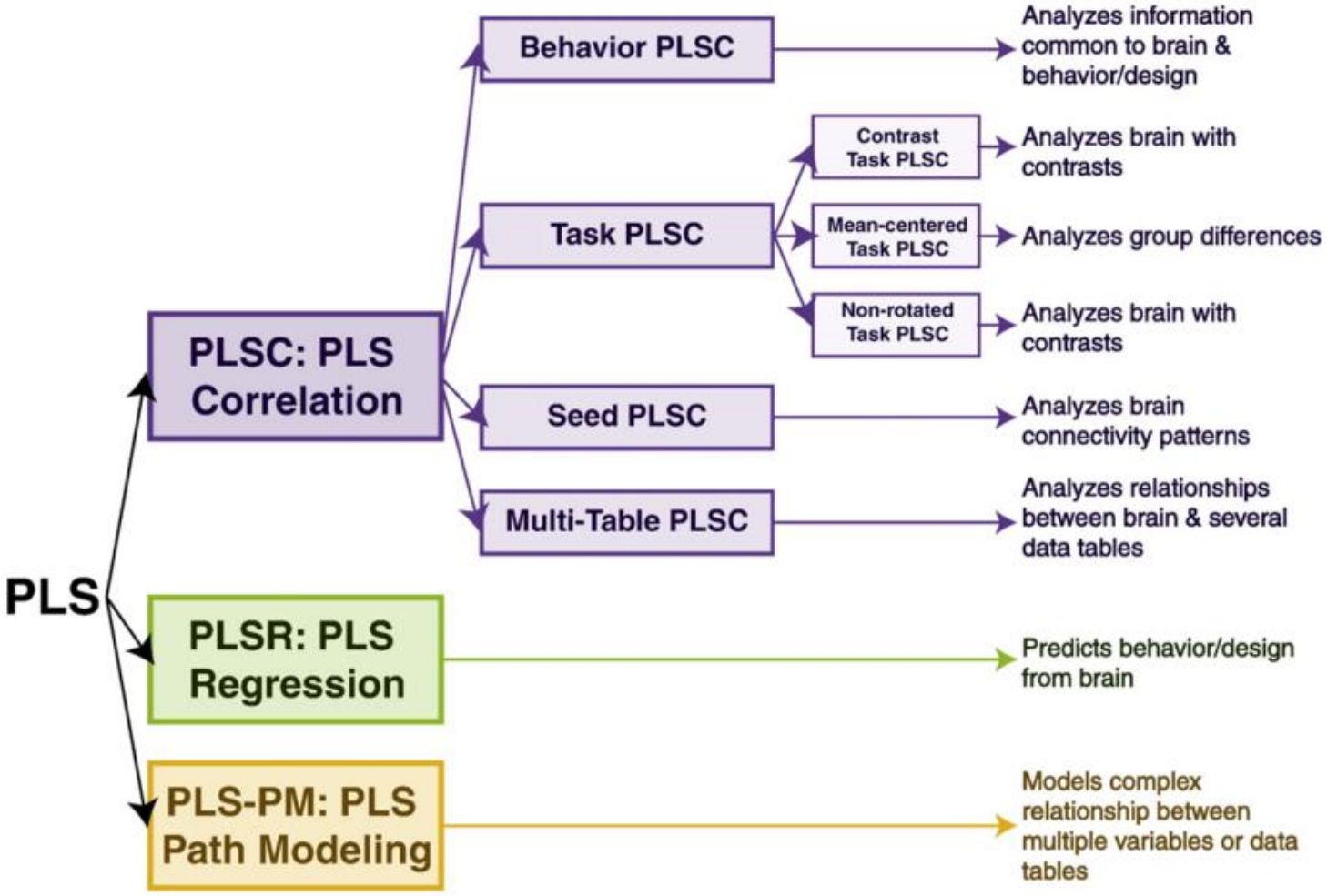
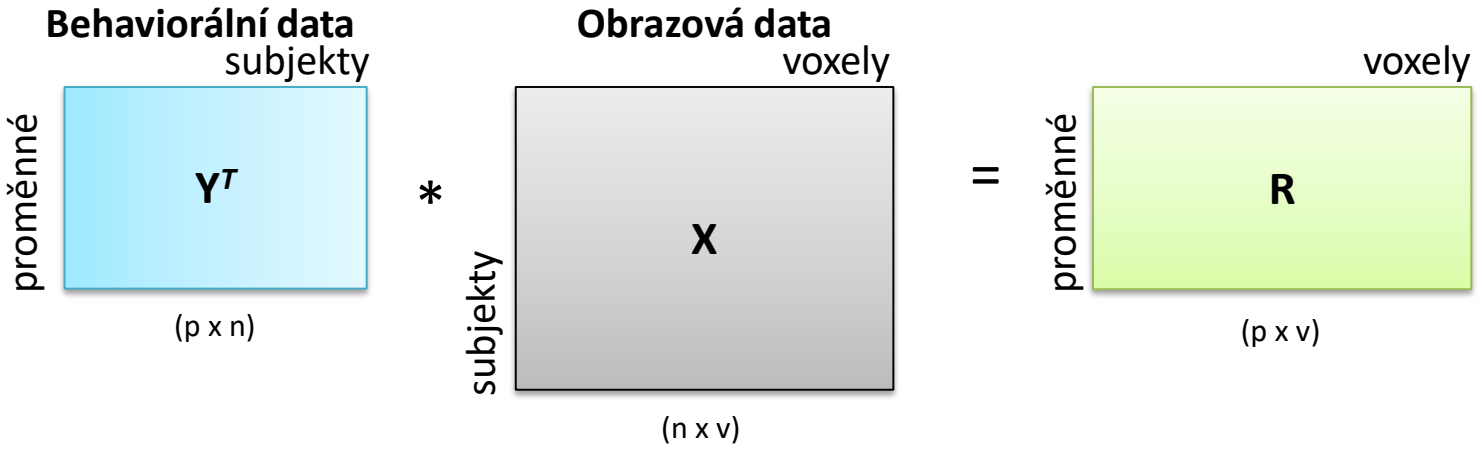


Fig. 1. The PLS family.

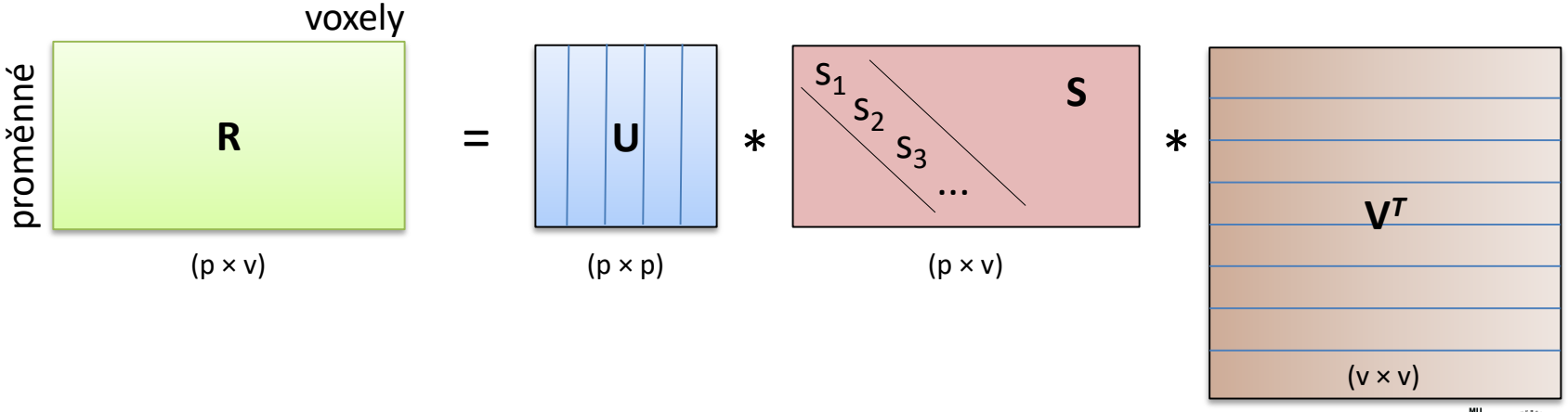
Krishnan, A., Williams, L.J., McIntosh, A.R., Abdi, H., 2011. Partial least squares (PLS) methods for neuroimaging: a tutorial and review. Neuroimage 56 (2), 455–475.

PLSC – princip

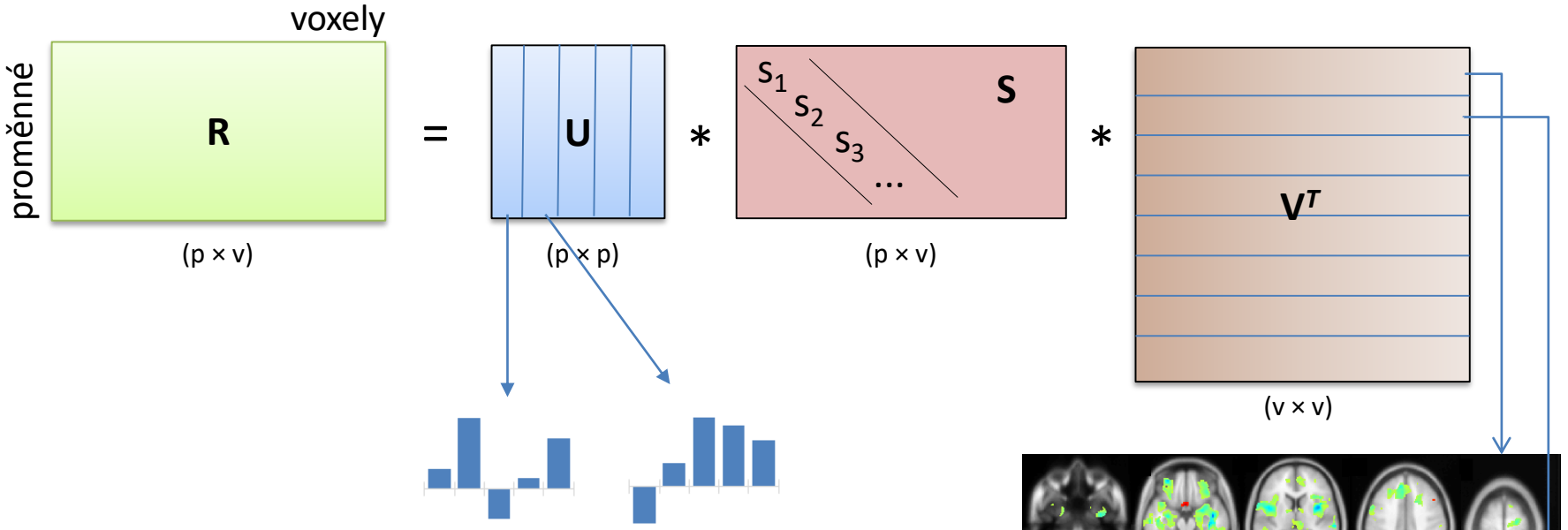
1. Výpočet matice **R** – ukazuje vztah mezi maticemi **X** a **Y** (korelace, pokud **X** a **Y** předem standardizovány; kovariance, pokud **X** a **Y** jen centrovány)



2. Rozklad matice **R** na $\mathbf{U \times S \times V^T}$ pomocí SVD (singular value decomposition)



PLSC – výstup



V každém sloupci váhy odpovídající nalezeným korelačním vzorům mezi obrazovými a behaviorálními daty

V každém řádku váhy pro korelaci s nalezeným behaviorálním vzorem v jednotlivých voxelech

S - diagonální matice singulárních hodnot ($s_1 \geq s_2 \geq ... \geq s_p$), odpovídajících kovarianci jednotlivých párů latentních proměnných. Z těchto hodnot lze odvozovat jakousi významnost dané latentní proměnné.

PLSC – optimalizace

Cílem PLSC je nalezení takových párů latentních proměnných, které:

1. Mají vzájemnou maximální **kovarianci**
2. Pro index l_1 a l_2 ,kdy $l_1 \neq l_2$, jsou latentní vektory **nekorelované**
3. Koeficienty u ***u*** a ***v*** jsou normalizovné

Latentní proměnné …. $\ell_{\mathbf{X},\ell} = \mathbf{X}\mathbf{v}_\ell \quad \text{and} \quad \ell_{\mathbf{Y},\ell} = \mathbf{Y}\mathbf{u}_\ell$

První podmínka …. $\operatorname{cov}\left(\ell_{\mathbf{X},\ell}, \ell_{\mathbf{Y},\ell}\right) \propto \ell_{\mathbf{X},\ell}^{\mathrm{T}} \ell_{\mathbf{Y},\ell} = \max$

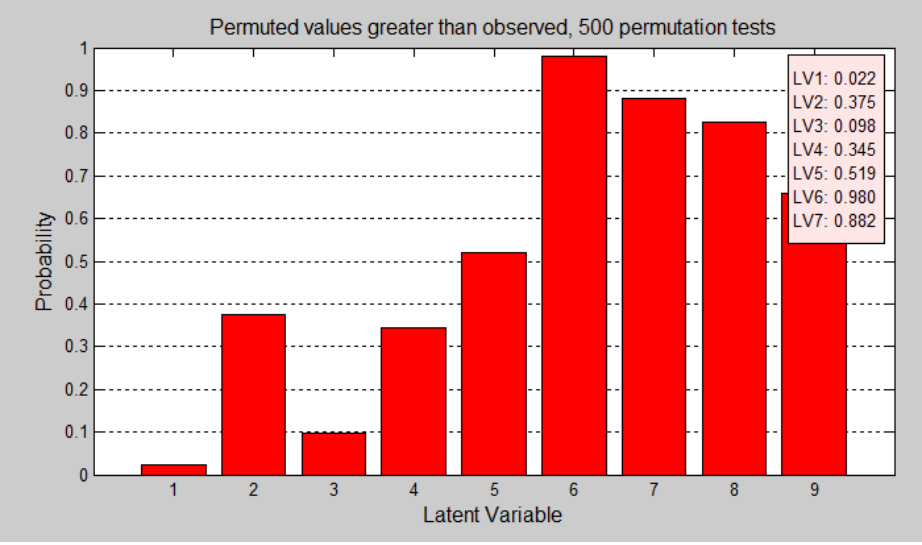
Druhá podmínka …. $\ell_{\mathbf{X},\ell}^{\mathrm{T}} \ell_{\mathbf{Y},\ell'} = 0 \text{ when } \ell \neq \ell'$

Třetí podmínka …. $\mathbf{u}_\ell^T \mathbf{u}_\ell = \mathbf{v}_\ell^T \mathbf{v}_\ell = 1$

Z SVD plyne, že kovariance mezi dvěma latentními proměnnými je rovna příslušné singulární hodnotě *s*. $\ell_{\mathbf{X},\ell}^{\mathrm{T}} \ell_{\mathbf{Y},\ell} = \delta_\ell$

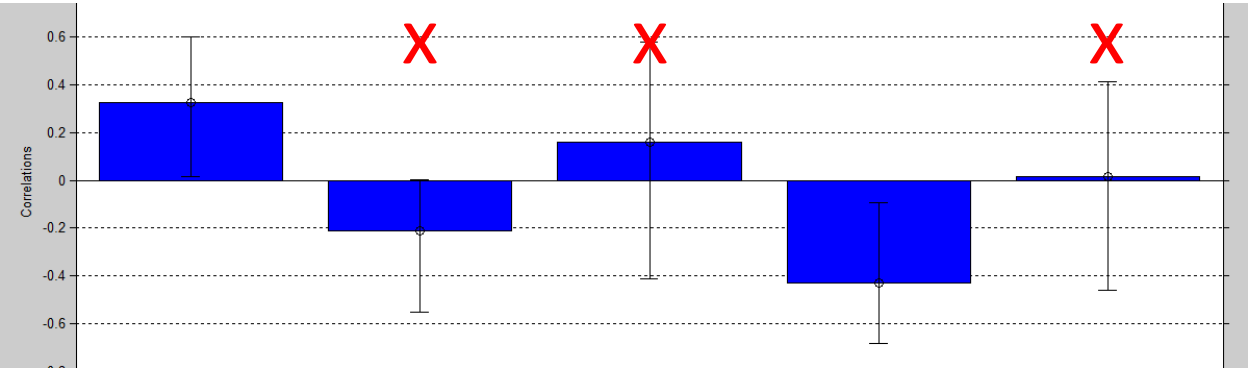
Významnost latentní proměnné

- umožní určit, jaké proměnné interpretovat
- významnost testována pomocí permutací – permutace v matici **Y** – matice s behaviorálními daty
- pro každou permutaci se opět vypočte PLSC a p-hodnota testu pak odpovídá pravděpodobnosti, že náhodně sestavená data měla vyšší singulární hodnotu u dané latentní proměnné než v originálním datovém souboru



Stabilita prostorového vzorce

- pro zjištění stability nalezených výsledků v závislosti na obrazech vstupujících do analýzy se dělají bootrapové výběry (řádově stovky až tisíce náhodných výběrů s vracením) – opět spočítáno PLSC
- stabilní latentní proměnné pak mají v daném voxelu přes všechny výběry menší směrodatnou odchylku
- poměr původní váhy z originálního PLSC k odhadnuté směrodatné odchylce se pak chová jako z-score → možnost vybrat pouze ty voxely, které jsou stabilní (např. s tímto poměrem ≥ 1.96)
- z boostrapových výběrů lze také odhadnout velikost intervalu spolehlivosti pro jednotlivé korelace behaviorálních dat s obrazovými → pokud obsahují 0, pak danou behaviorální proměnnou nemá moc cenu interpretovat



Ordinační analýzy – shrnutí

- analýza hlavních komponent, faktorová analýza, korespondenční analýza, multidimensional scaling a metody varietního učení 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, kanonická korelační analýza a metoda parciálních nejmenších čtverců se snaží nalézt vztah mezi dvěma sadami vícerozměrných dat

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“