

Analýza a klasifikace dat – přednáška 9



RNDr. Eva Koritáková, Ph.D.

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:
 - 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	pac.								
	I_2	pac.								
	I_3	kont.								
	...									

- 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	pac.								
	I_2	pac.								
	I_3	kont.								
	...									

➔

	y_1	y_2	y_3	y_4
subjekty	I_1			
	I_2			
	I_3			
	...			

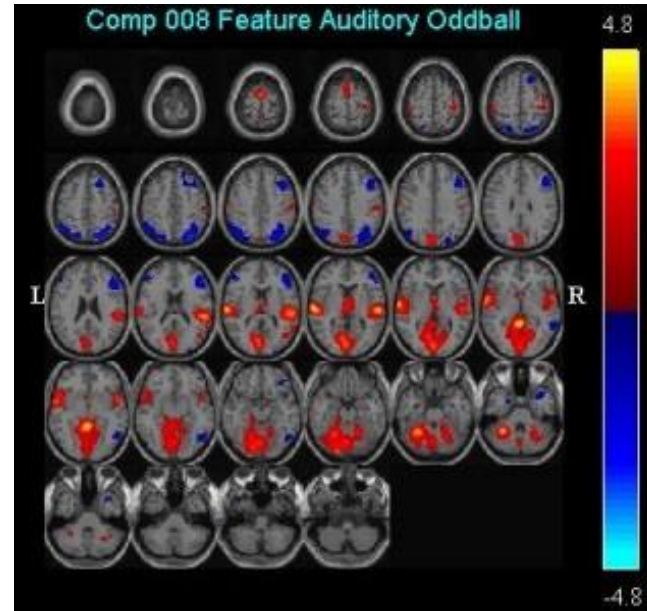
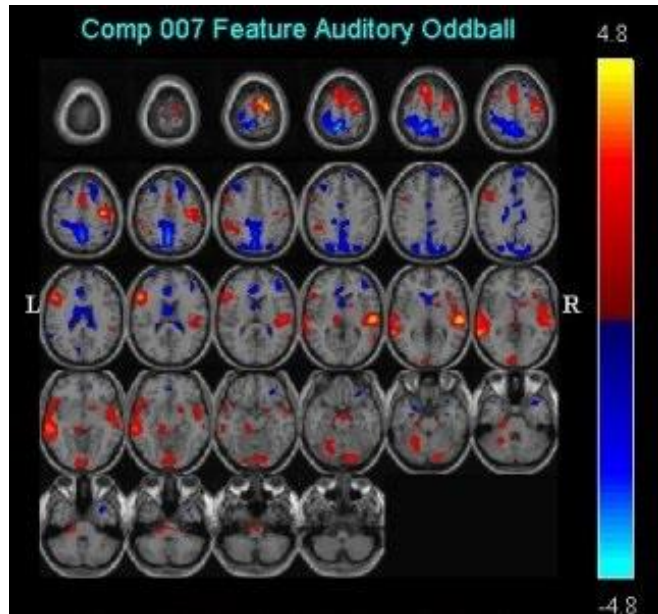
Extrakce proměnných

- 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 ($m \leq p$)
- 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 (\approx 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)

Analýza nezávislých komponent

Analýza nezávislých komponent (ICA)

Princip: Hledání statisticky nezávislých komponent v původních datech.



Výhody:

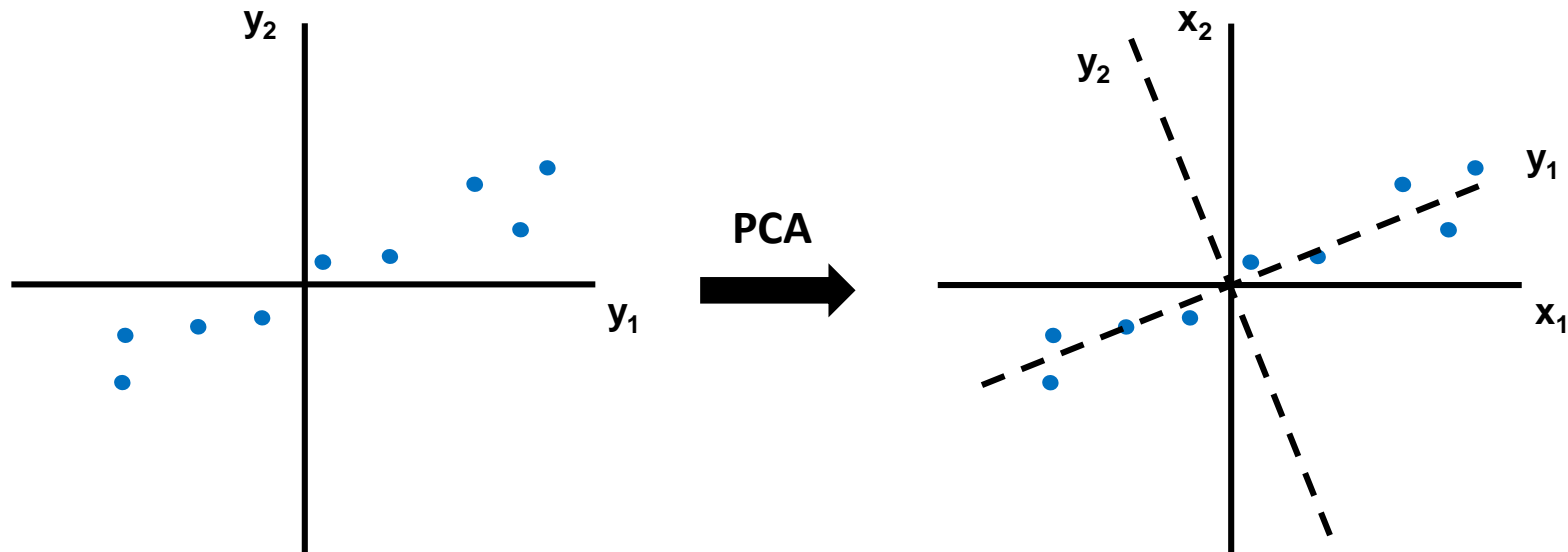
- + vícerozměrná metoda
- + dokáže vytvořit lépe interpretovatelné komponenty než PCA

Nevýhody:

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

Srovnání s analýzou hlavních komponent (PCA)

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



Výhody:

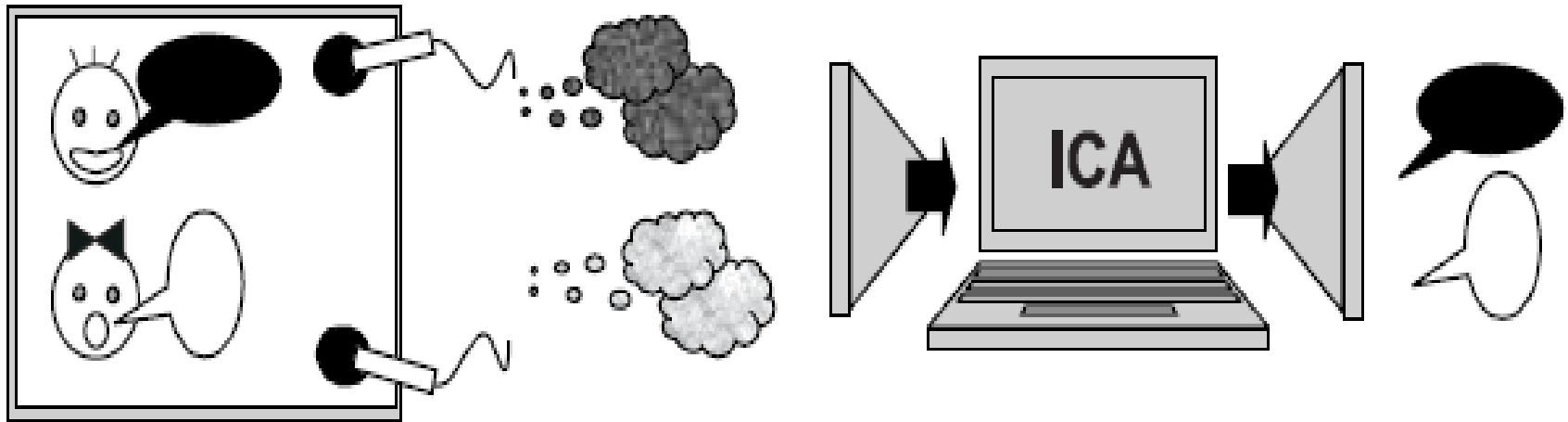
+ vícerozměrná metoda

Nevýhody:

- nevyužívá informaci o příslušnosti subjektů do skupin
- potřebné určit, kolik hlavních komponent se použije pro transformaci (avšak při změně počtu není nutné PCA přepočítat)

Analýza nezávislých komponent

- anglicky *Independent Component Analysis* (ICA)



$$x_1(t) = a_{11} \cdot s_1(t) + a_{12} \cdot s_2(t)$$

$$x_2(t) = a_{21} \cdot s_1(t) + a_{22} \cdot s_2(t)$$

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

$$x_i = a_{i1}^{\text{orig}} \cdot s_1^{\text{orig}} + a_{i2}^{\text{orig}} \cdot s_2^{\text{orig}} + \dots + a_{im}^{\text{orig}} \cdot s_m^{\text{orig}}, \quad i = 1, 2, \dots, m$$

nebo maticově

$$\mathbf{x} = \mathbf{A}^{\text{orig}} \cdot \mathbf{s}^{\text{orig}}$$

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

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

- skryté nezávislé komponenty je možno vyjádřit pomocí vztahu:

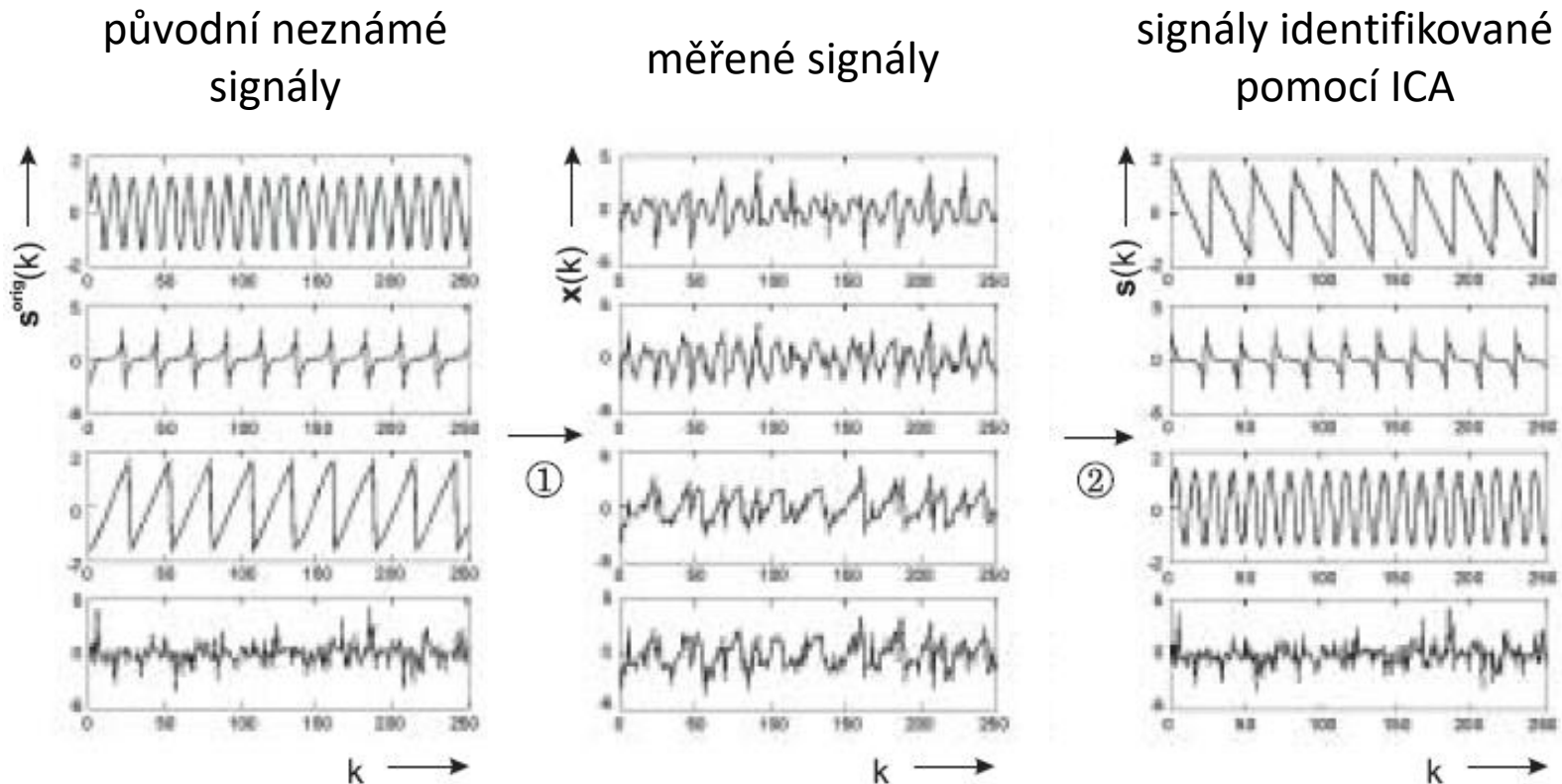
$$\mathbf{s} = \mathbf{W} \cdot \mathbf{x}$$

- cíl: nalézt lineární transformaci (koeficienty transformační matice \mathbf{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



- 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

- necht' 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 \mathbf{x} (který máme k dispozici) předpokládáme, že vznikl součtem nezávislých komponent \mathbf{s}^{orig}



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 $\mathbf{s} = \mathbf{w}_i \cdot \mathbf{x}$
- **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: $H(s) = -\sum_i P(s=a_i) \cdot \log_2 P(s=a_i)$,
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 s dalšími rozd.**
- NNE: $J(s) = H(s_{\text{gauss}}) - H(s)$, 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

$$J(s) \approx \frac{1}{12} E\{s^3\}^2 + \frac{1}{48} \text{kurt}(s)^2$$

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

- **nevýhoda:**

– opět menší robustnost vůči odlehlým hodnotám

- použití tzv. p-nekvadratických funkcí

$$J(s) \approx \sum_{i=1}^p k_i \cdot [E\{G_i(s)\} - E\{G_i(s_{\text{gauss}})\}]^2$$

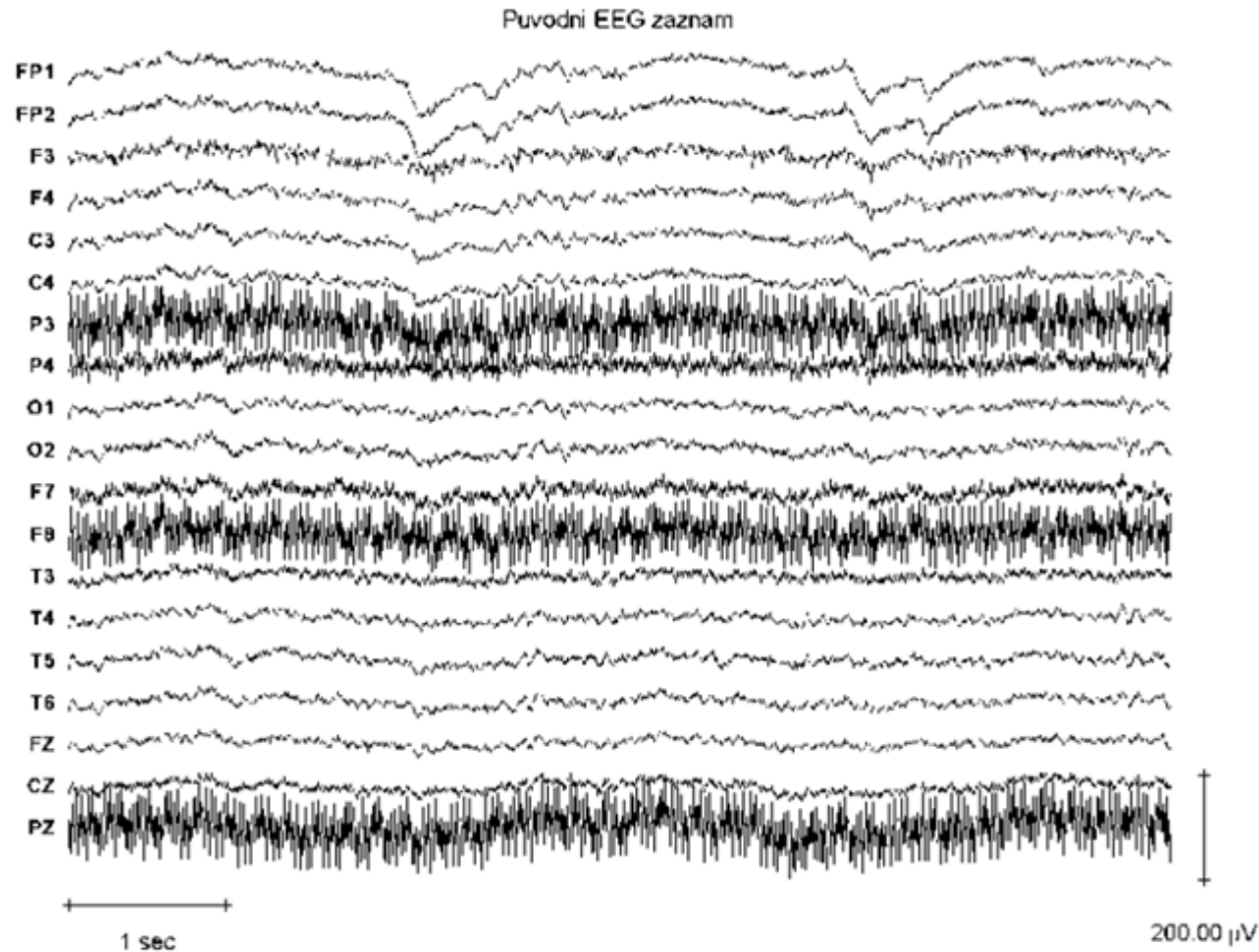
kde $k_i > 0$ je konstanta, G_i jsou šikovně navržené nelineární funkce a s_{gauss} je normální náhodná proměnná, která spolu s s má nulovou střední hodnotu a jednotkový rozptyl.

Je-li použita pouze jedna funkce G , pak je

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

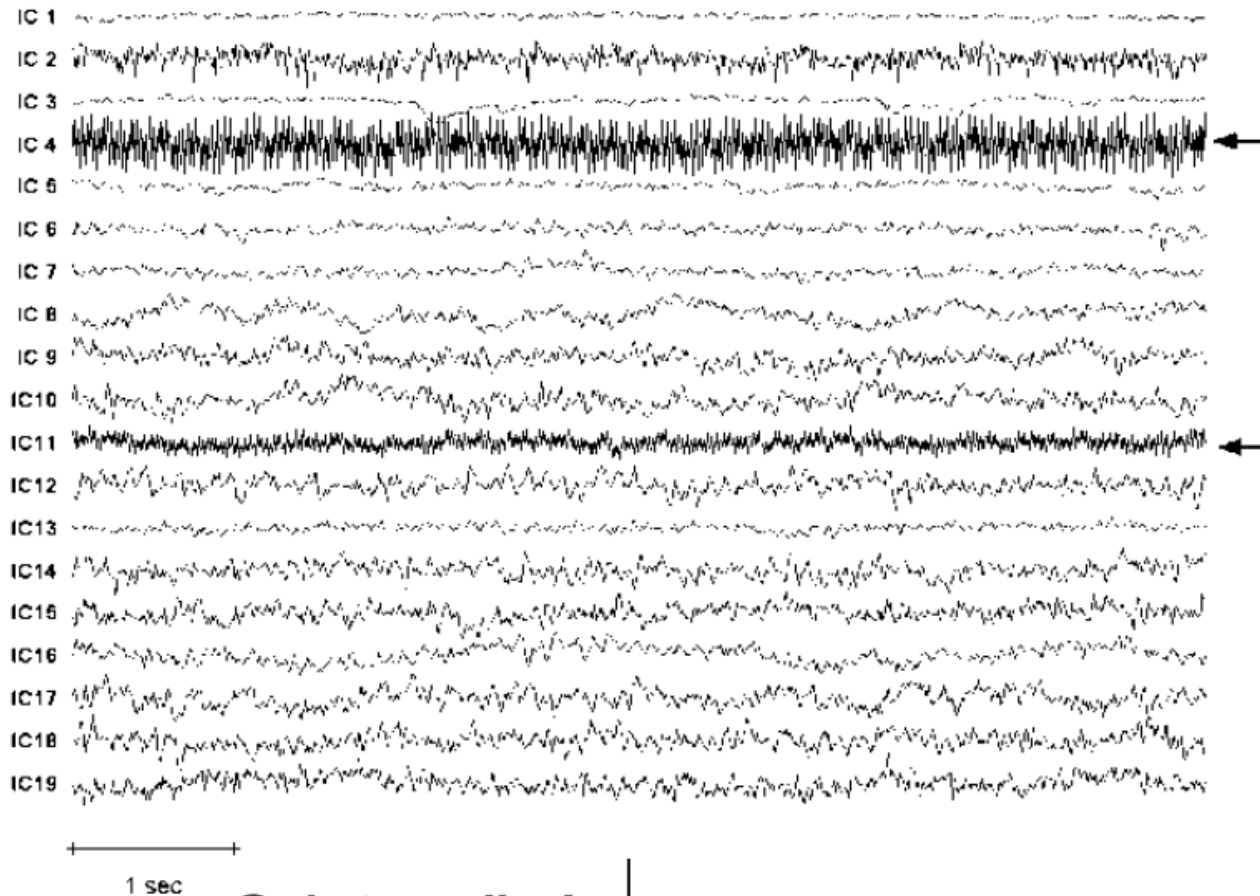
- doporučuje se $G_1(s) \approx \frac{1}{a_1} \log(\cosh a_1 s)$ kde $a_1 \in \langle 1, 2 \rangle$ nebo $G_2(s) \approx -\exp(-s^2/2)$

Analýza nezávislých komponent – příklad použití



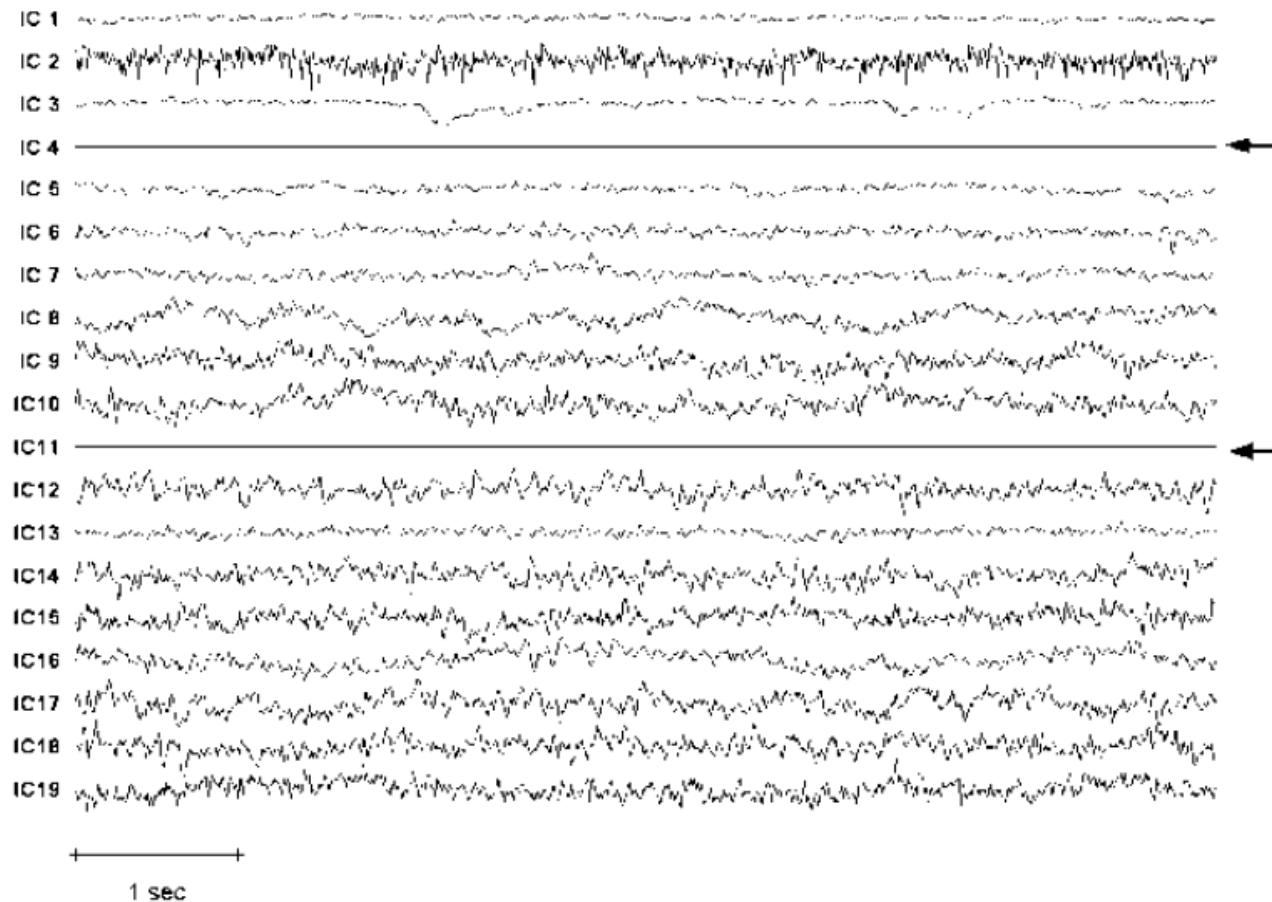
Analýza nezávislých komponent – příklad použití

Nezávislé komponenty (ICs)

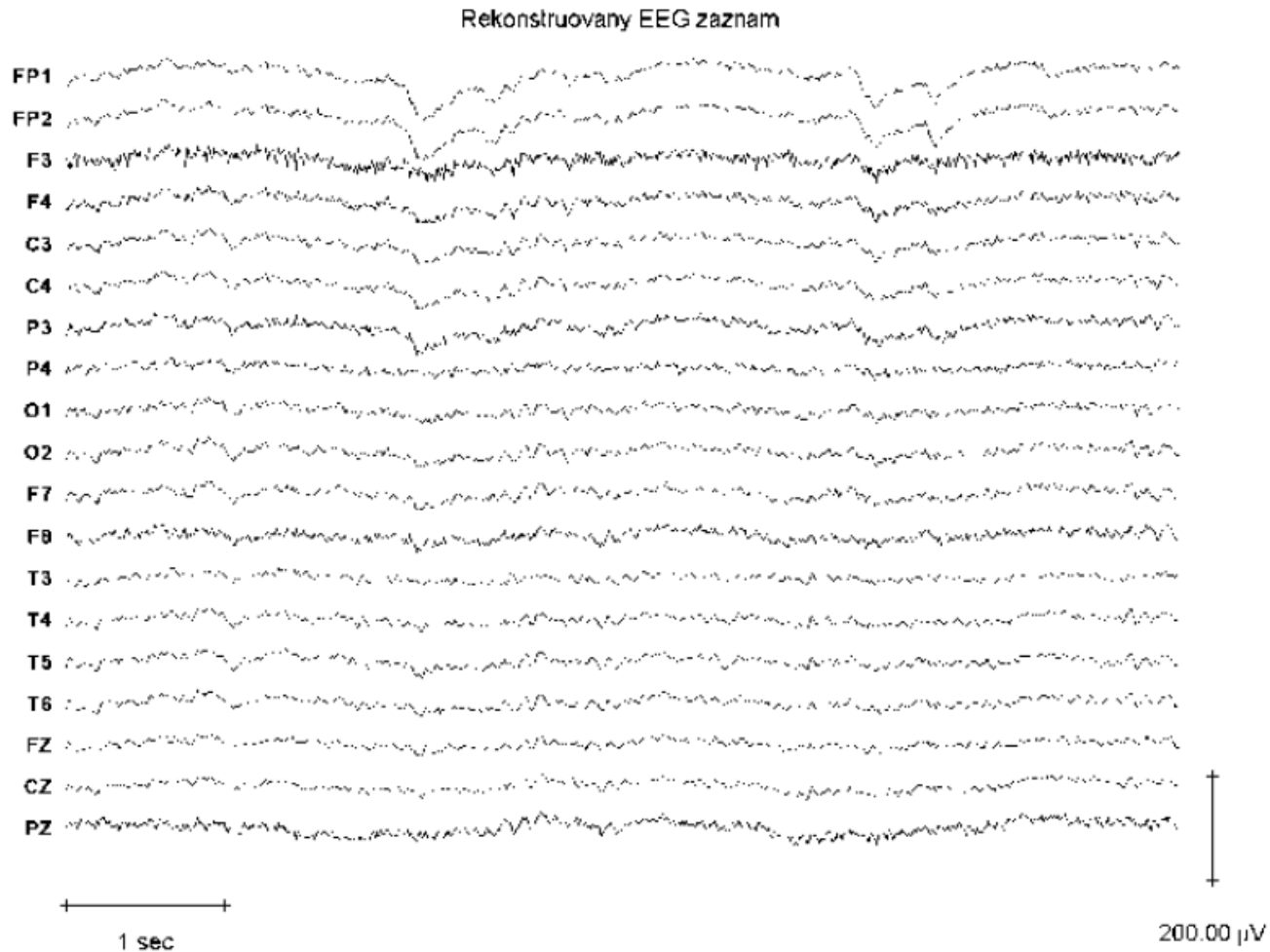


Analýza nezávislých komponent – příklad použití

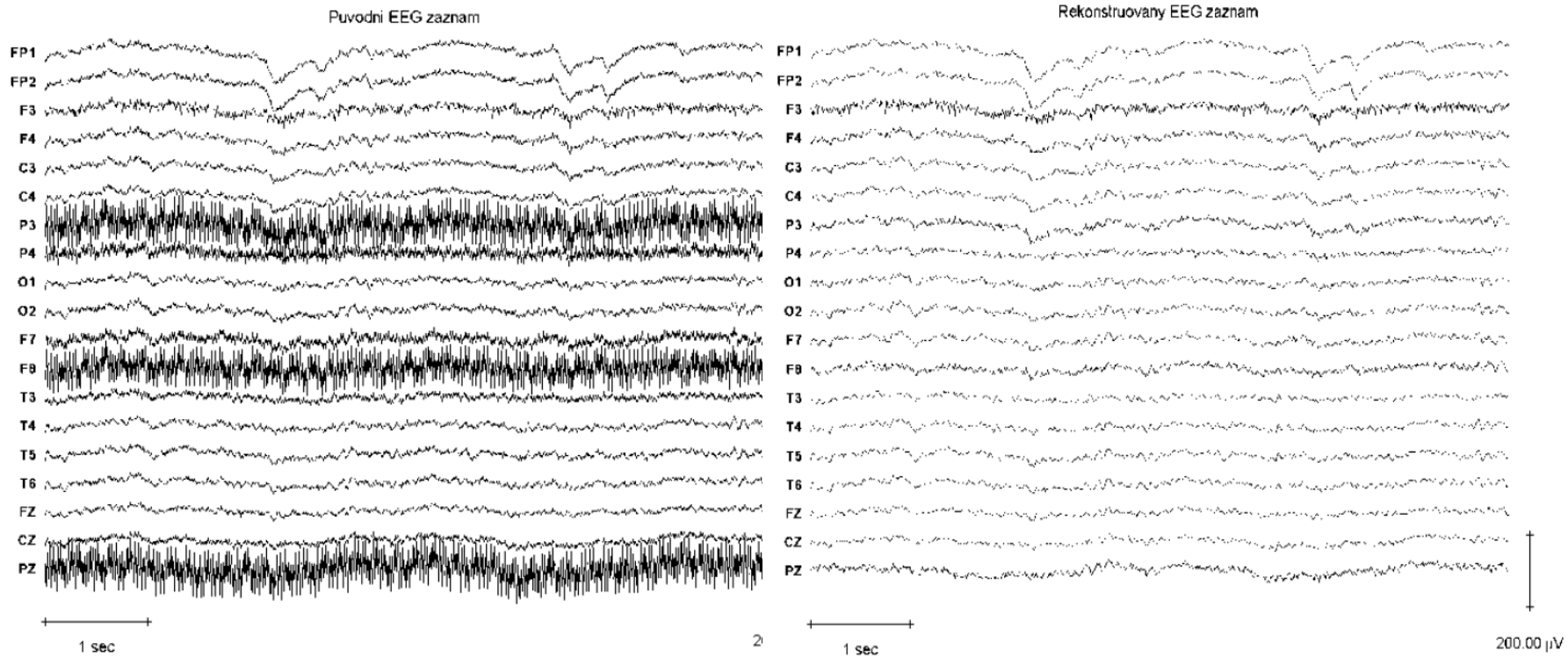
Nezávislé komponenty (IC4 a IC11 byly odstraněny)



Analýza nezávislých komponent – příklad použití



Analýza nezávislých komponent – příklad použití



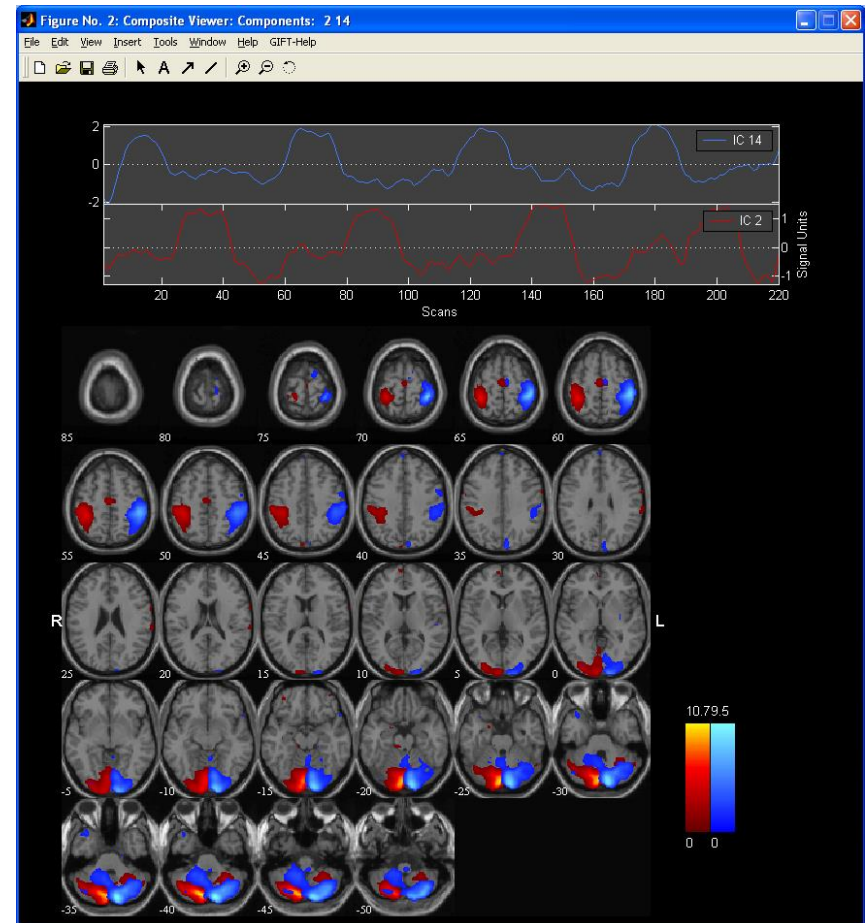
Analýza nezávislých komponent – příklad 2

- 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/>

Celkem 14 implementovaných ICA algoritmů v GIFT toolboxu:

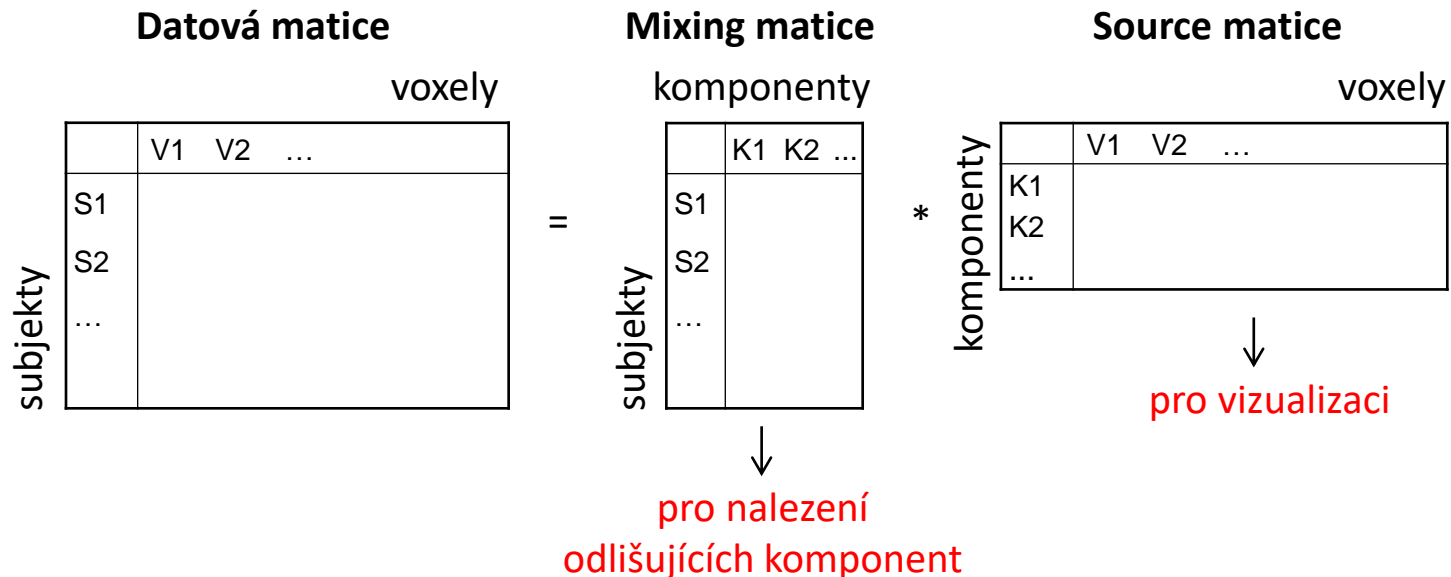
- | | |
|---------------------------------|---|
| <input type="radio"/> Infomax | <input type="radio"/> SDD ICA |
| <input type="radio"/> FastICA | <input type="radio"/> Semi-blind Infomax |
| <input type="radio"/> Erica | <input type="radio"/> Constrained ICA (Spatial) |
| <input type="radio"/> Simbec | <input type="radio"/> Radical ICA |
| <input type="radio"/> Evd | <input type="radio"/> Combi |
| <input type="radio"/> Jade Opac | <input type="radio"/> ICA-EBM |
| <input type="radio"/> Amuse | <input type="radio"/> FBSS |



Analýza nezávislých komponent – příklad 3

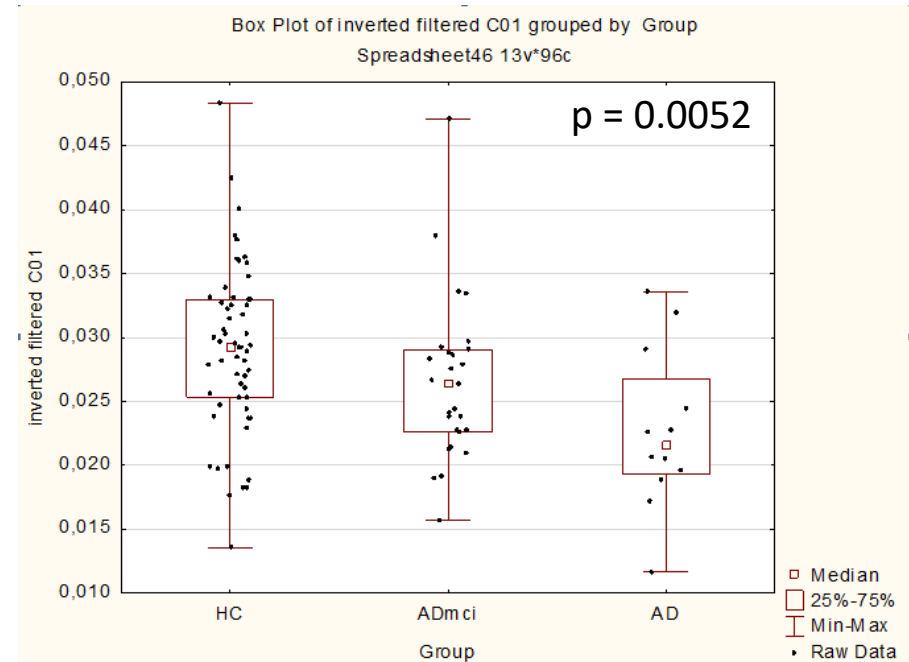
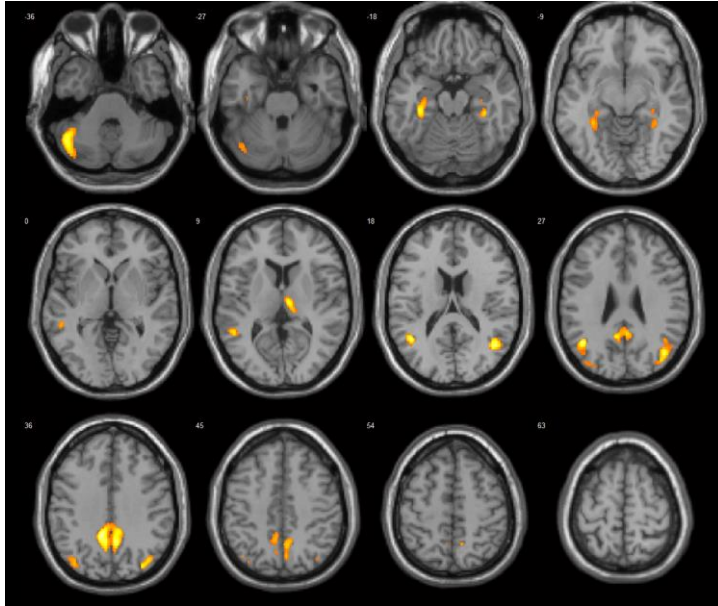
- Zadání: nalezněte nezávislé komponenty, které dokáží odlišit tři skupiny subjektů

	#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

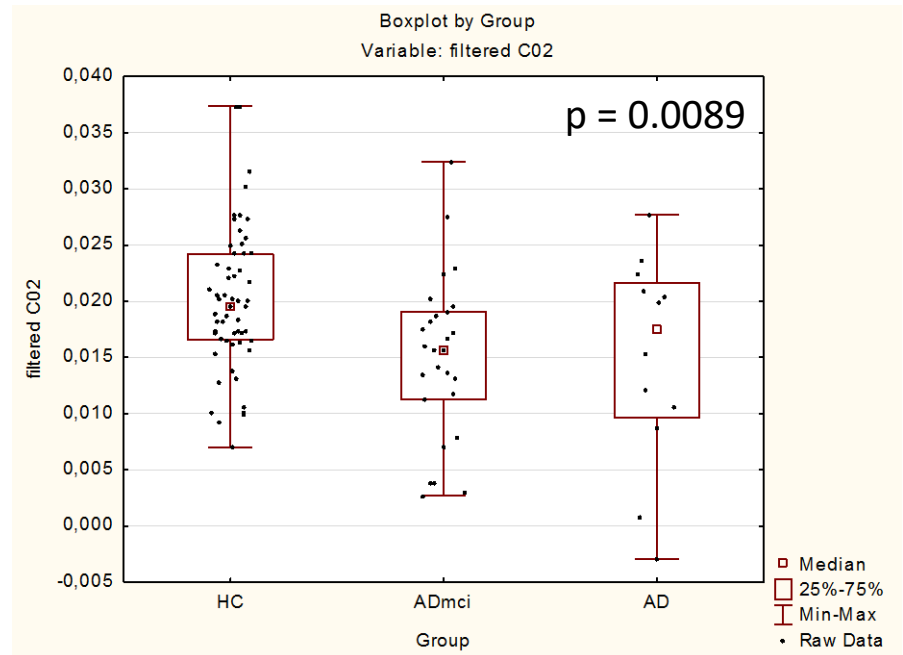
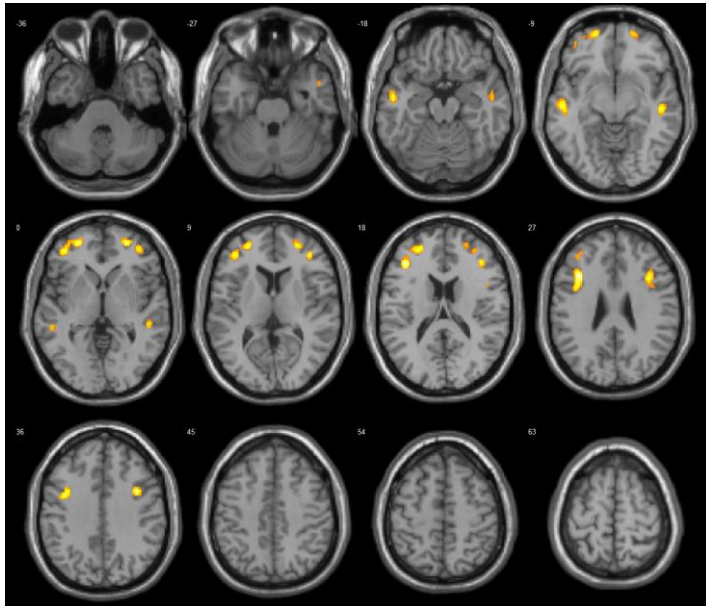
- komponenta č. 1:



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

Analýza nezávislých komponent – příklad 3

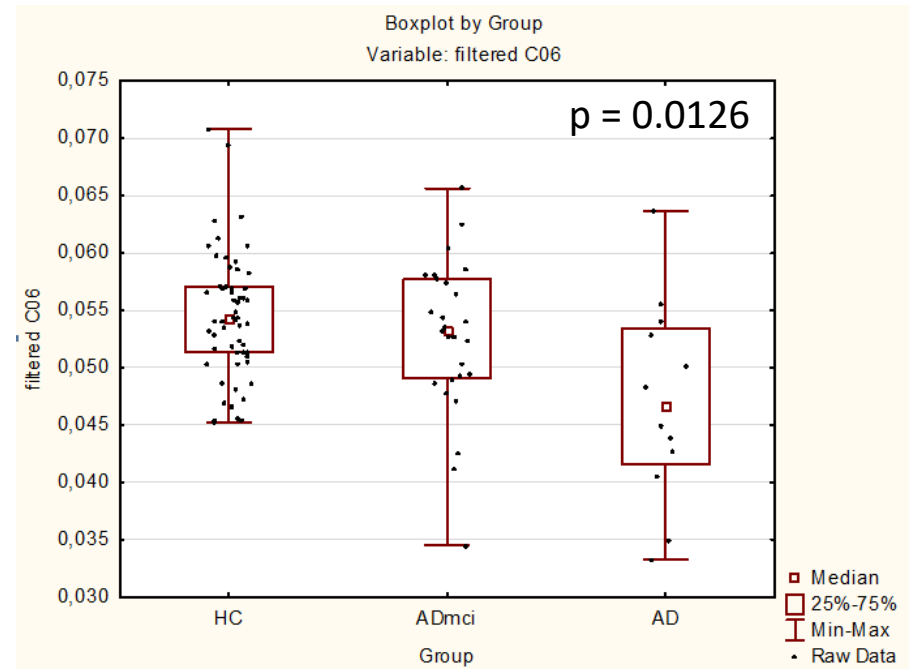
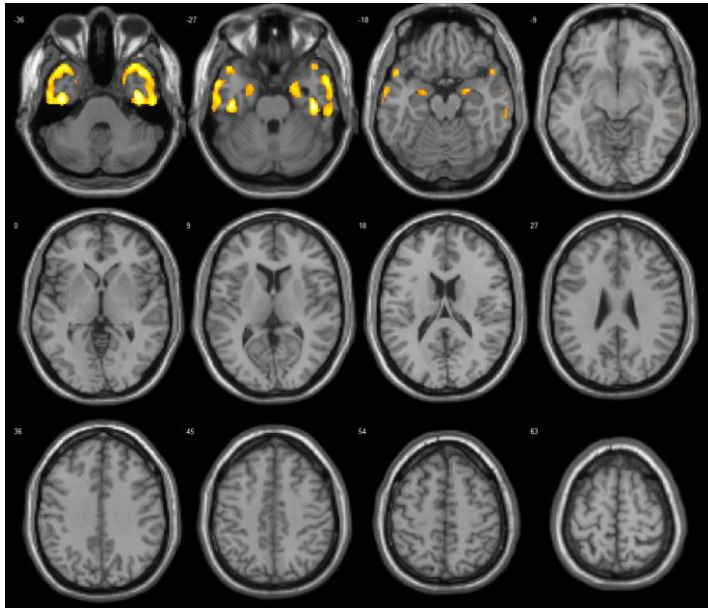
- komponenta č. 2:



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

Selekce proměnných

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:

- 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	pac.								
	I_2	pac.								
	I_3	kont.								
	...									

- 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	pac.								
	I_2	pac.								
	I_3	kont.								
	...									

→

		y_1	y_2	y_3	y_4
subjekty	I_1				
	I_2				
	I_3				
	...				

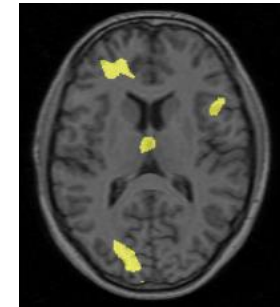
Selekce proměnných

- 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, nejpoužívanější metody jsou:
 - výběr proměnných na základě statistických testů
 - výběr oblastí mozku (ROI) podle atlasu
 - algoritmy sekvenční selekce (dopředné či zpětné; algoritmus plus p mínus q; algoritmus min-max)
 - algoritmus ohraničeného větvení

Výběr proměnných na základě statistických testů

Princip: Výběr statisticky významných proměnných pomocí dvouvýběrového t-testu či Mannova-Whitneyova testu (když máme 2 skupiny subjektů).

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



Výhody:

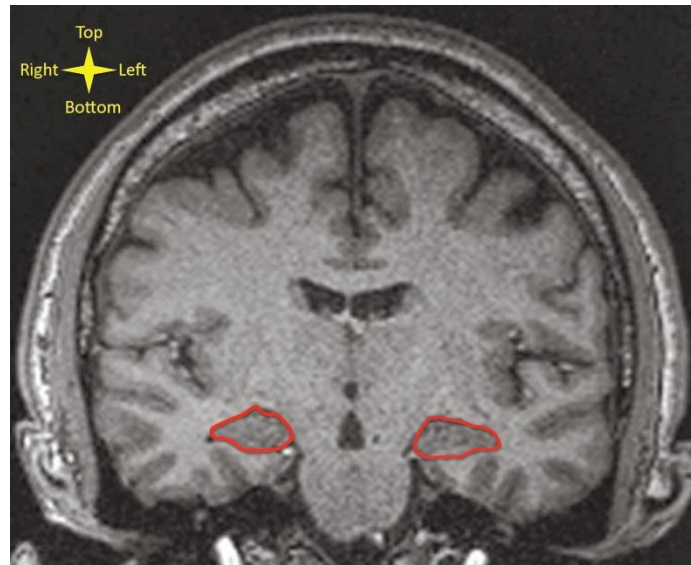
- + rychlé
- + u obrazů mozku výhodou, že je analýza provedena na celém mozku

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

- výběr optimální podmnožiny obsahující m ($m \leq p$) proměnných – kombinatorický problém – $p!/(p-m)!m!$ možných řešení!

- např. výběr 10 proměnných z 20: $\frac{p!}{(p-m)!m!} = \frac{20!}{10!10!} = 335\,221\,286\,400$



hledáme jen kvazioptimální řešení

- předpoklad: **monotónnost kritéria selekce** – označíme-li X_j množinu obsahující j proměnných, pak monotónnost kritéria znamená, že pro podmnožiny

$$X_1 \subset X_2 \subset \dots \subset X_j \subset \dots \subset X_m$$

splňuje selekční kritérium vztah

$$J(X_1) \leq J(X_2) \leq \dots \leq J(X_m)$$

- tedy: je nutno seřadit proměnné podle toho, jak dobře dokáží diskriminovat trénovací data

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, tj. $J(\{X_{k+1}\}) = \max J(\{X_k \cup y_j\})$, $y_j \in \{Y - X_k\}$
- **algoritmus sekvenční zpětné selekce:**
 - algoritmus začíná s množinou všech proměnných
 - v každém následujícím kroku se eliminuje ta proměnná, která způsobuje nejmenší pokles kriteriální funkce, tj. $J(\{X_{m-k-1}\}) = \max J(\{X_{m-k} - y_j\})$, $y_j \in \{X_{m-k}\}$

Výhody:

- + dopředný algoritmus je výpočetně jednodušší, protože pracuje maximálně v m -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 proměnných se q proměnných odstraní
- proces probíhá, dokud se nedosáhne požadovaného počtu proměnných
- je-li $p > q$, pracuje algoritmus od prázdné množiny
- je-li $p < q$, varianta zpětného algoritmu

Algoritmus min - max

- Heuristický algoritmus vybírající proměnné na základě výpočtu hodnot kriteriální funkce pouze v jedno- a dvourozměrném prostoru.
- Předpokládejme, že bylo vybráno k proměnných do množiny $\{X_k\}$ a zbývají veličiny z množiny $\{Y-X_k\}$. Výběr veličiny $y_j \in \{Y-X_k\}$ přináší novou informaci, kterou můžeme ocenit relativně k libovolné veličině $x_i \in X_k$ podle vztahu

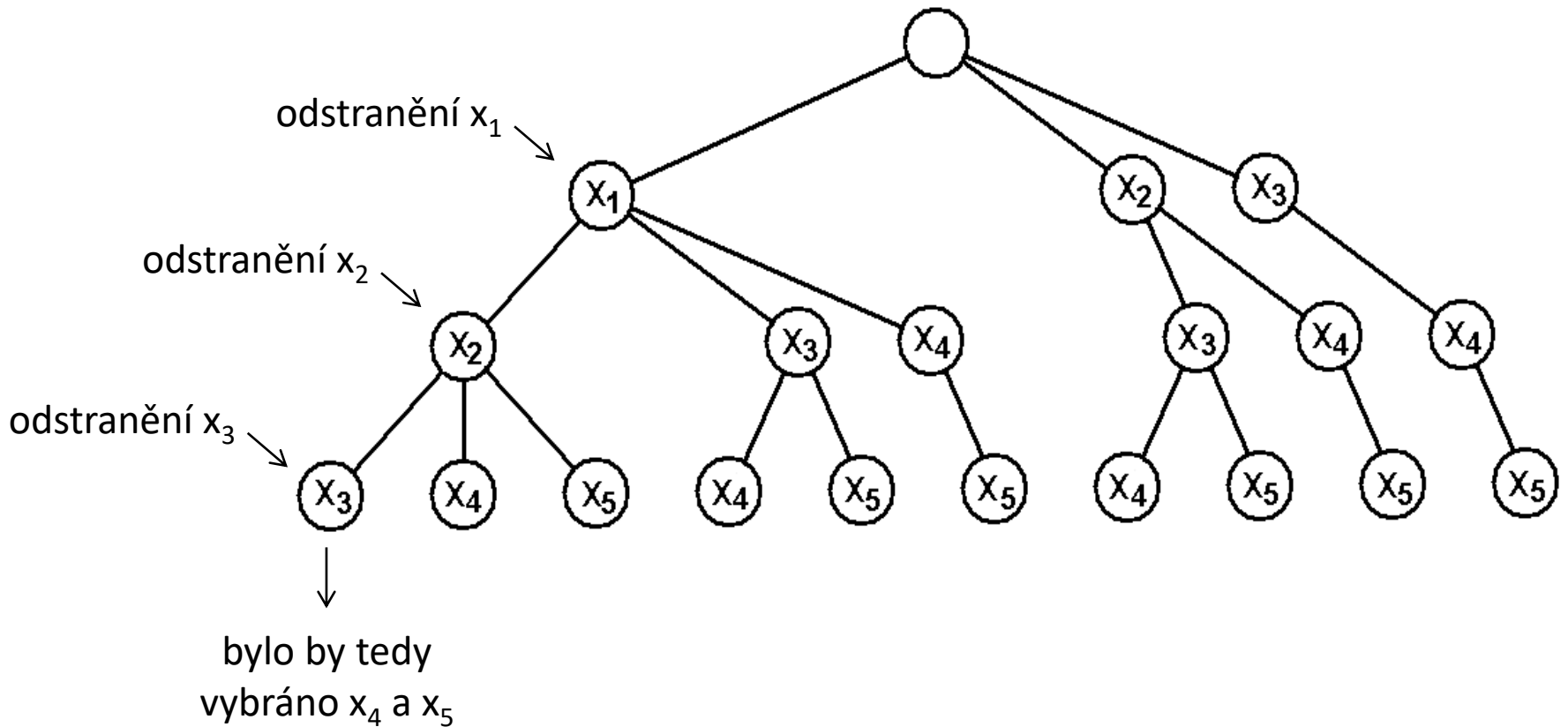
$$\Delta J(y_j, x_i) = J(y_j, x_i) - J(x_i)$$

- Informační přírůstek ΔJ musí být co největší, ale musí být dostatečný pro všechny proměnné již zahrnuté do množiny X_k . Vybíráme tedy veličinu y_{k+1} , pro kterou platí

$$\Delta J(y_{k+1}, x_k) = \max_j \min_i \Delta J(y_j, x_i), x_i \in X_k$$

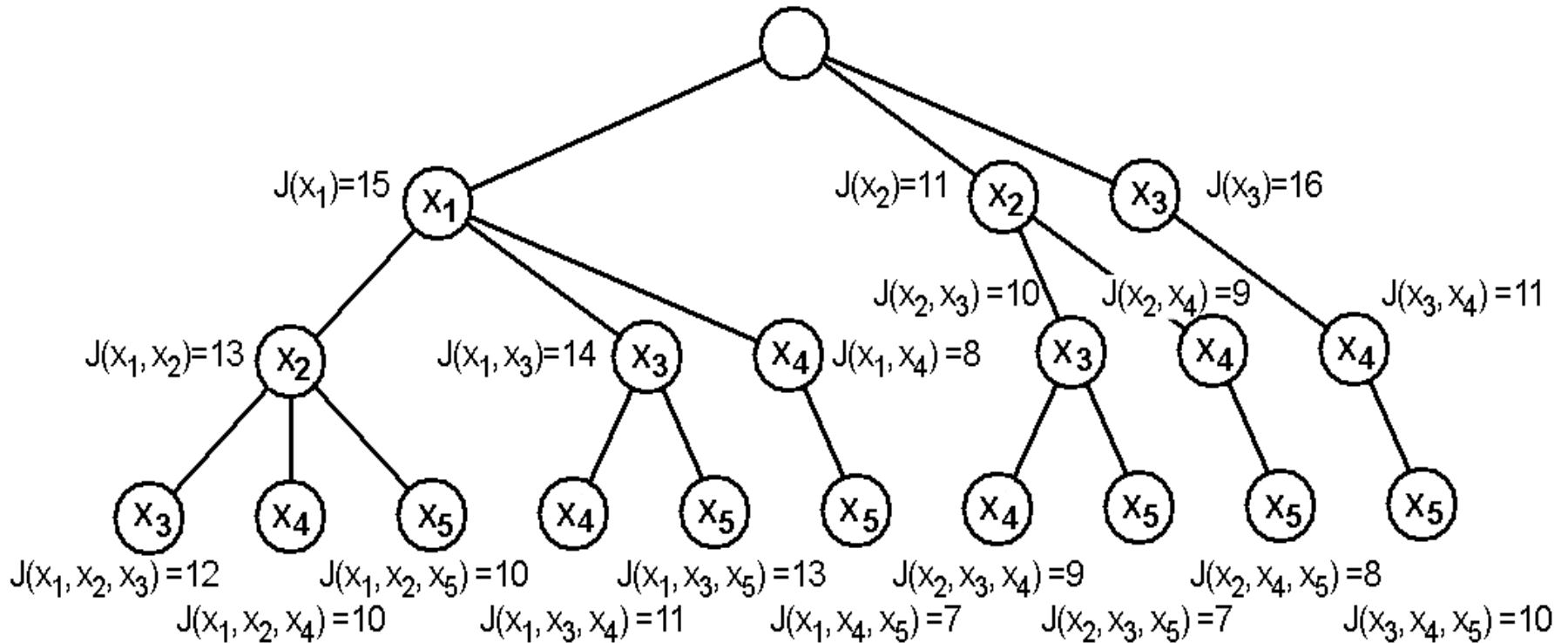
Algoritmus ohraničeného větvení

- uvažme případ selekce dvou proměnných z pěti:



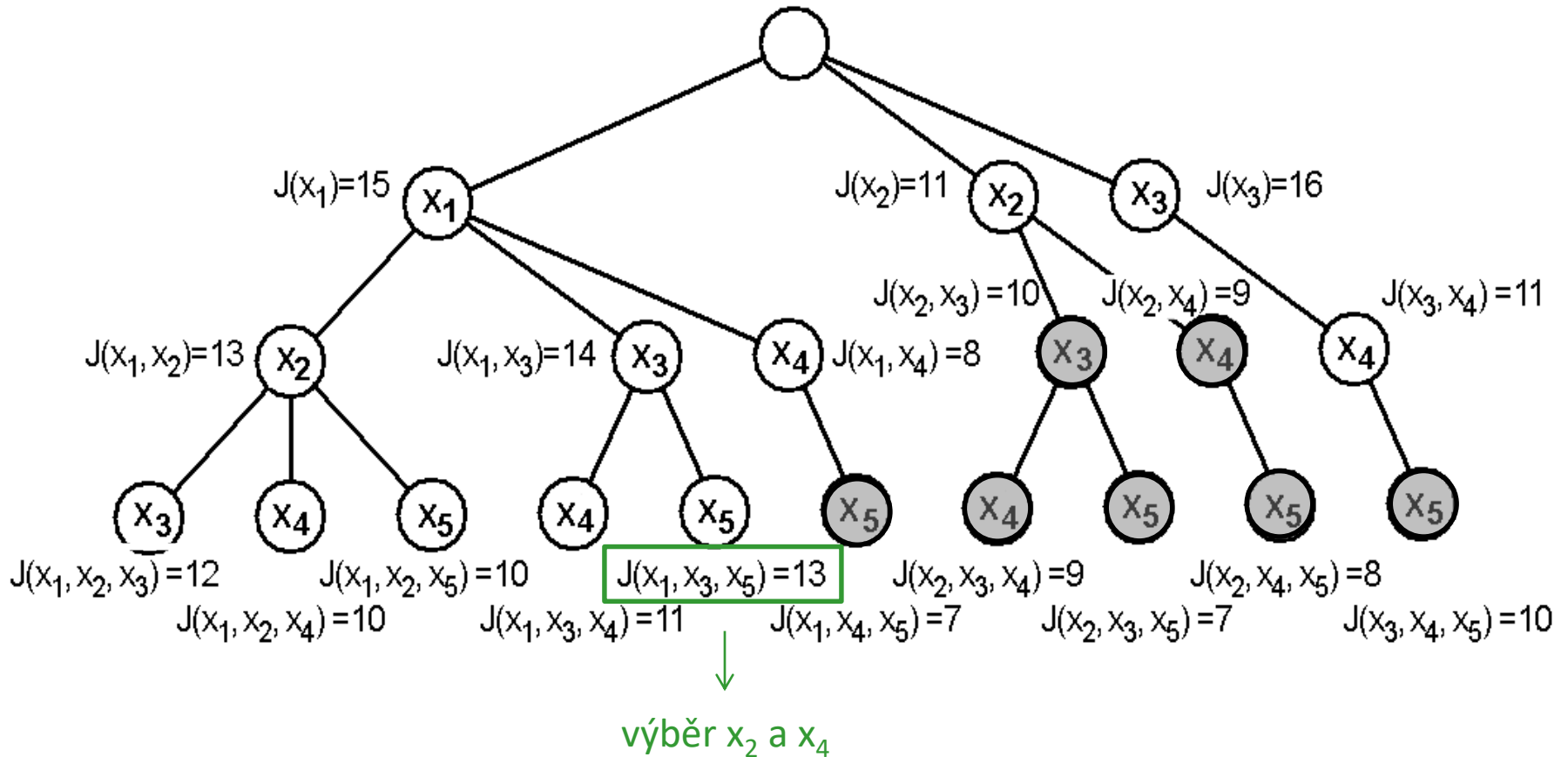
Algoritmus ohraničeného větvení

- příklad:



Algoritmus ohraničeného větvení

- příklad:



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“



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ