Analýza a klasifikace dat – přednáška 9

RNDr. Eva Janoušová

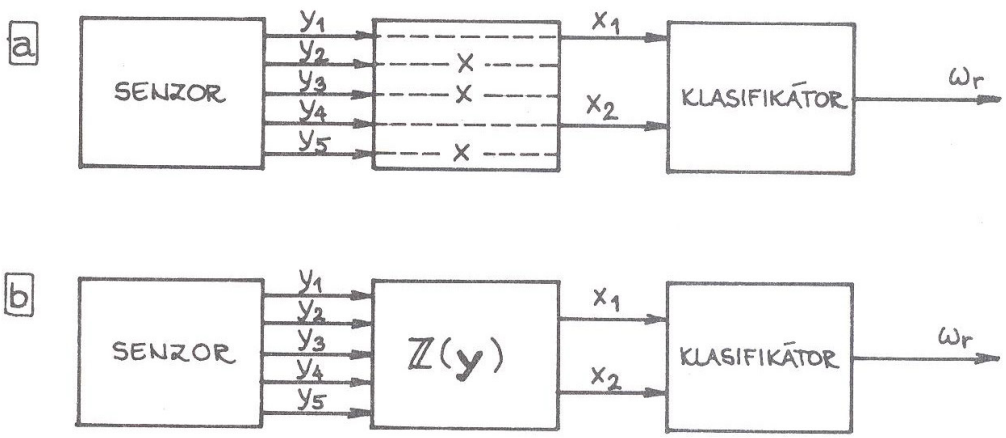
Podzim 2014

© Institut biostatistiky a analýz

Výběr příznaků

- formální popis objektu původně reprezentovaný *m* rozměrným vektorem se snažíme vyjádřit vektorem *n* rozměrným tak, aby množství diskriminační informace obsažené v původním vektoru bylo v co největší míře zachováno

- dva principiálně různé způsoby:
  1. **selekce** – nalezení a odstranění těch příznakových funkcí, které přispívají k separabilitě klasifikačních tříd nejméně
  2. **extrakce** – transformace původních příznakových proměnných na menší počet jiných příznakových proměnných

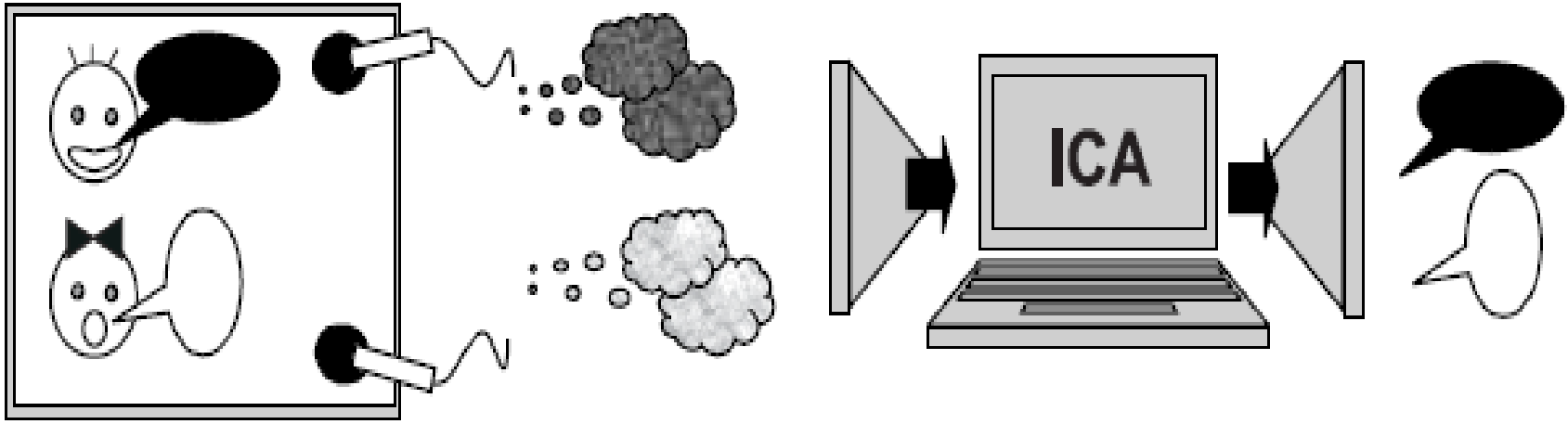


Extrakce příznaků

- jedním z principů výběru příznaků
- transformace původních příznakových proměnných na menší počet jiných příznakových proměnných $\Rightarrow$ tzn. hledání (optimálního) zobrazení Z, které transformuje původní *m*-rozměrný prostor (obraz) na prostor (obraz) *n*-rozměrný ($m \geq n$)
- 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
  - **rozložení pravděpodobnosti veličin v novém prostoru budou splňovat podmínky kladené na jejich pravděpodobnostní charakteristiky**
  - obrazy v novém prostoru budou minimalizovat odhad pravděpodobnosti chyby

Analýza nezávislých komponent

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

- nechť $\mathbf{x} = T(x_1, x_2, \ldots, x_m)$ je m-rozměrný náhodný vektor (s nulovou střední hodnotou $E(\mathbf{x})=0$).

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

nebo

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

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

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

- definice:

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

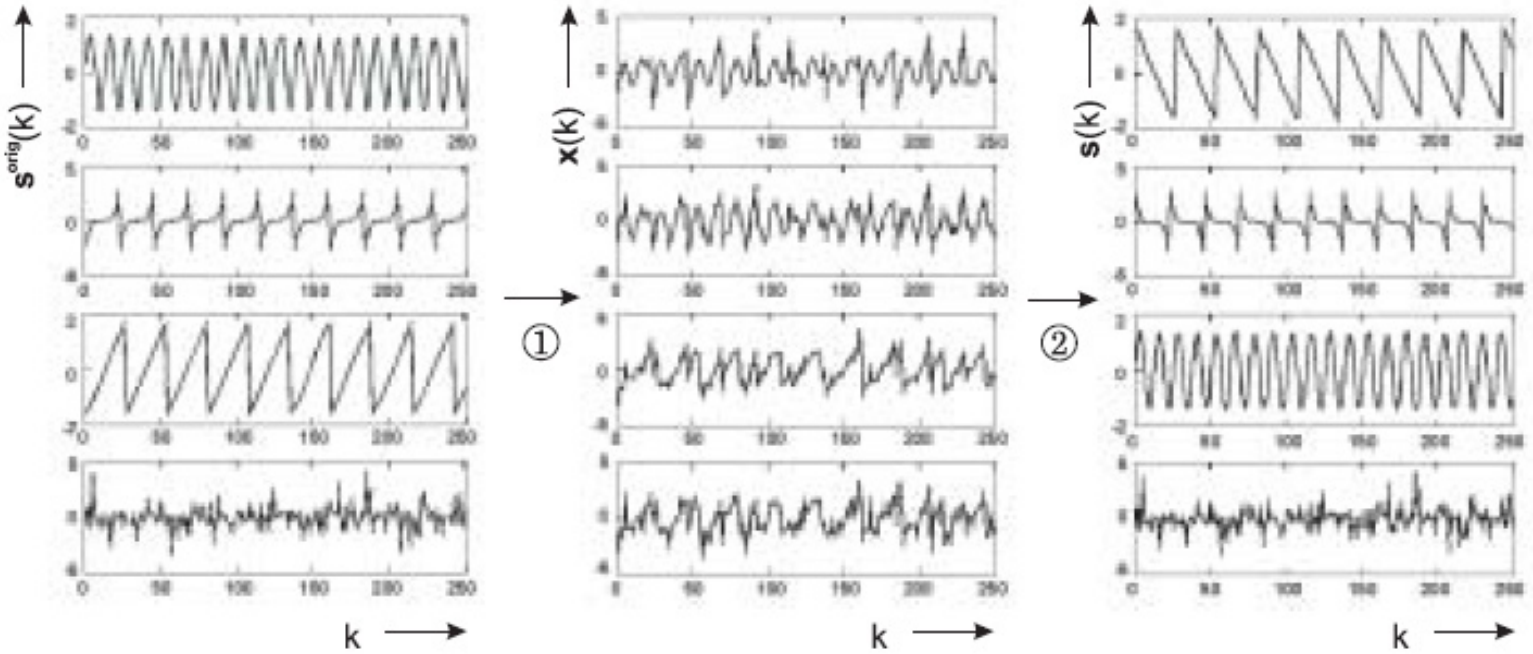
- cíl: nalézt lineární transformaci (koeficienty transformační matice **W** tak, aby vypočítané nezávislé komponenty s_i byly vzájemně statisticky nezávislé [$\mathbf{W} = \mathbf{A}^{-1}$]

$$[p(s_1,s_2,\ldots,s_m) = p_1(s_1).p_2(s_2)\ldots p_m(s_m)]$$

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



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 dosti obecných podmínek konverguje s rostoucím počtem sčítanců k normálnímu rozdělení – 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 odhadnut 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 nenáročný

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

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

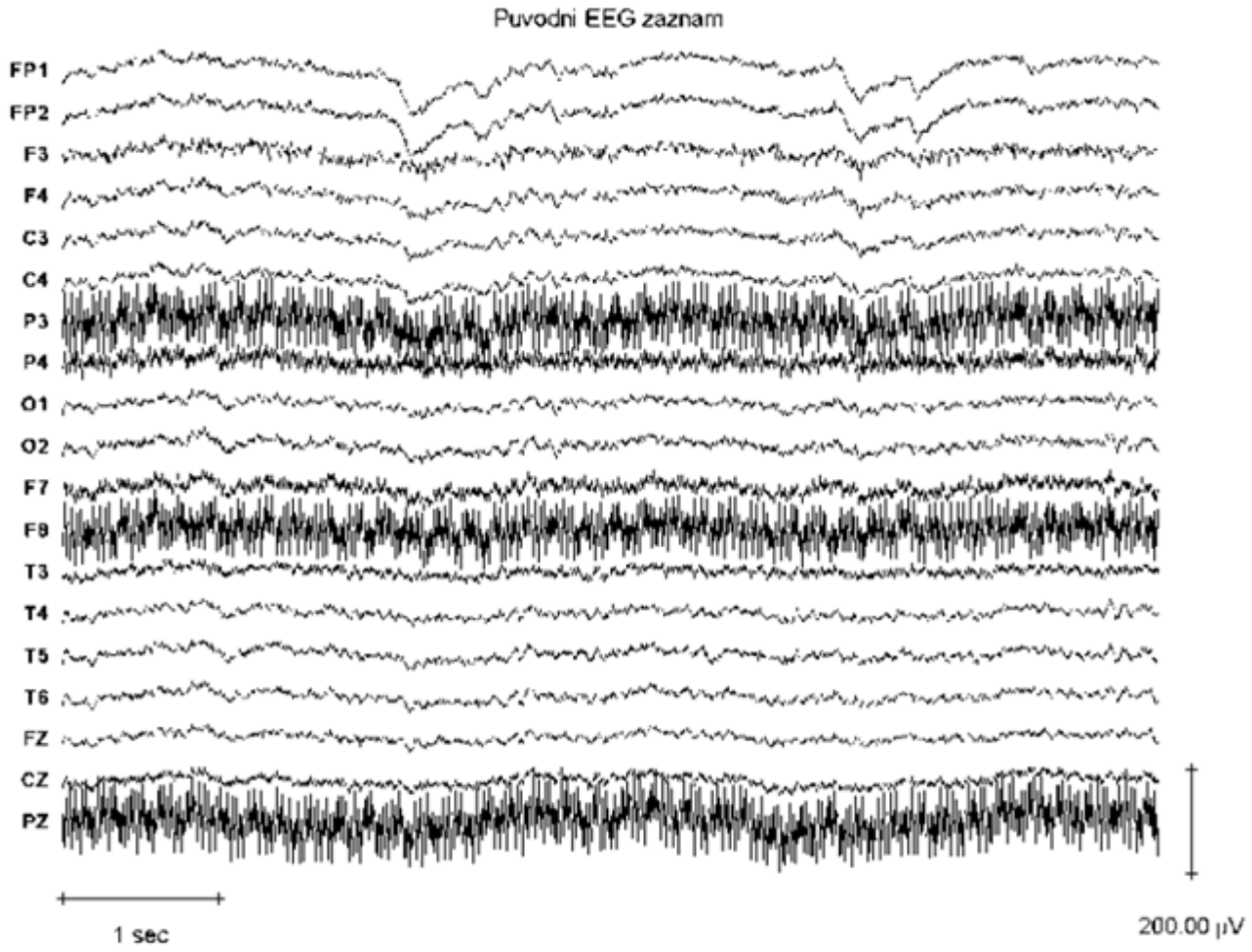
  kde $k_i>0$ je konstanta, G_i jsou šikovně navržené nelineární funkce a s_{gauss} je normální náhodná proměnná, která spolu s s má nulovou střední hodnotu a jednotkový rozptyl.

  Je-li použita pouze jedna funkce G, pak je

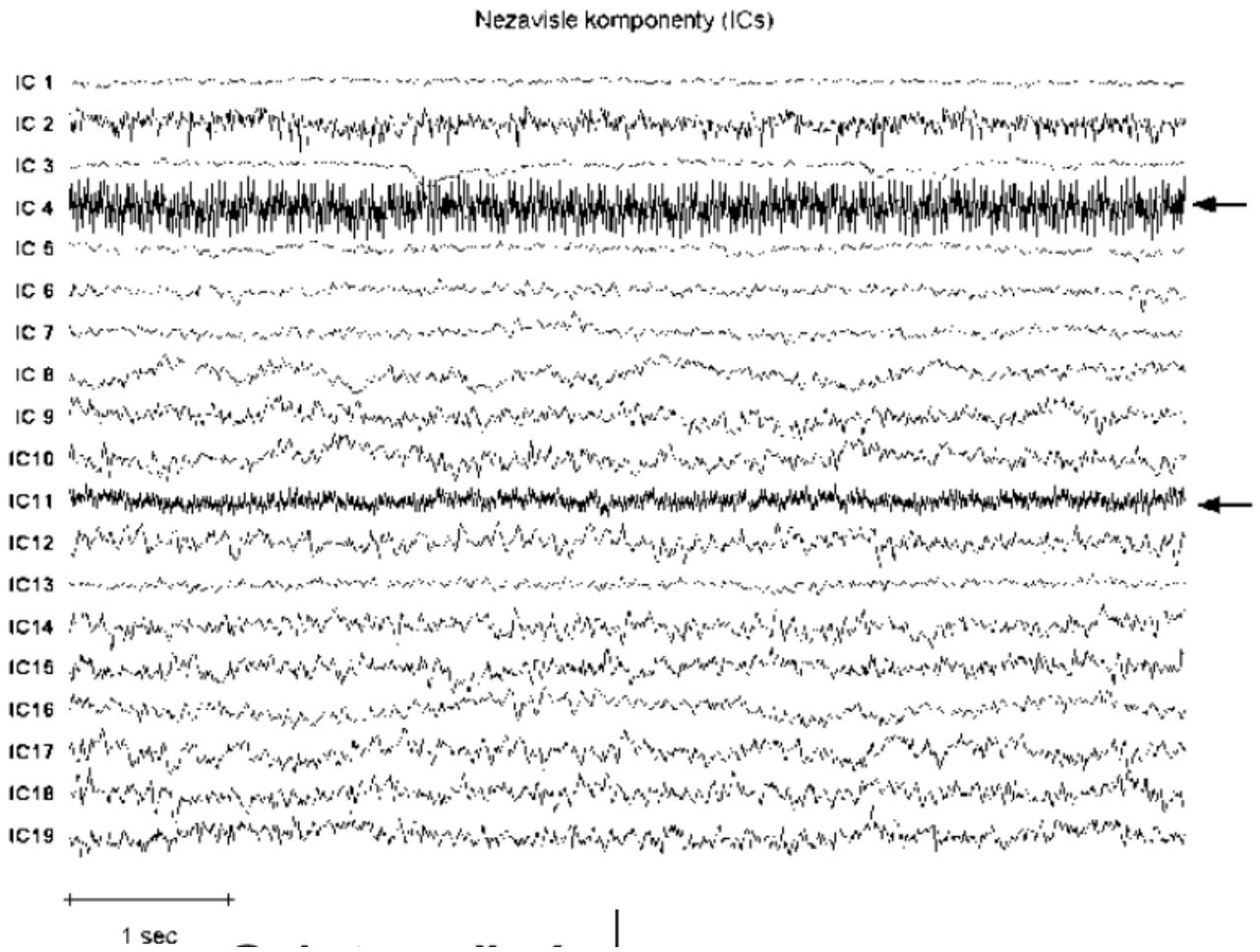
$$J(s) \approx [\mathcal{E}\{G(s)\} - \mathcal{E}\{G(s_{gauss})\}]^2$$

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

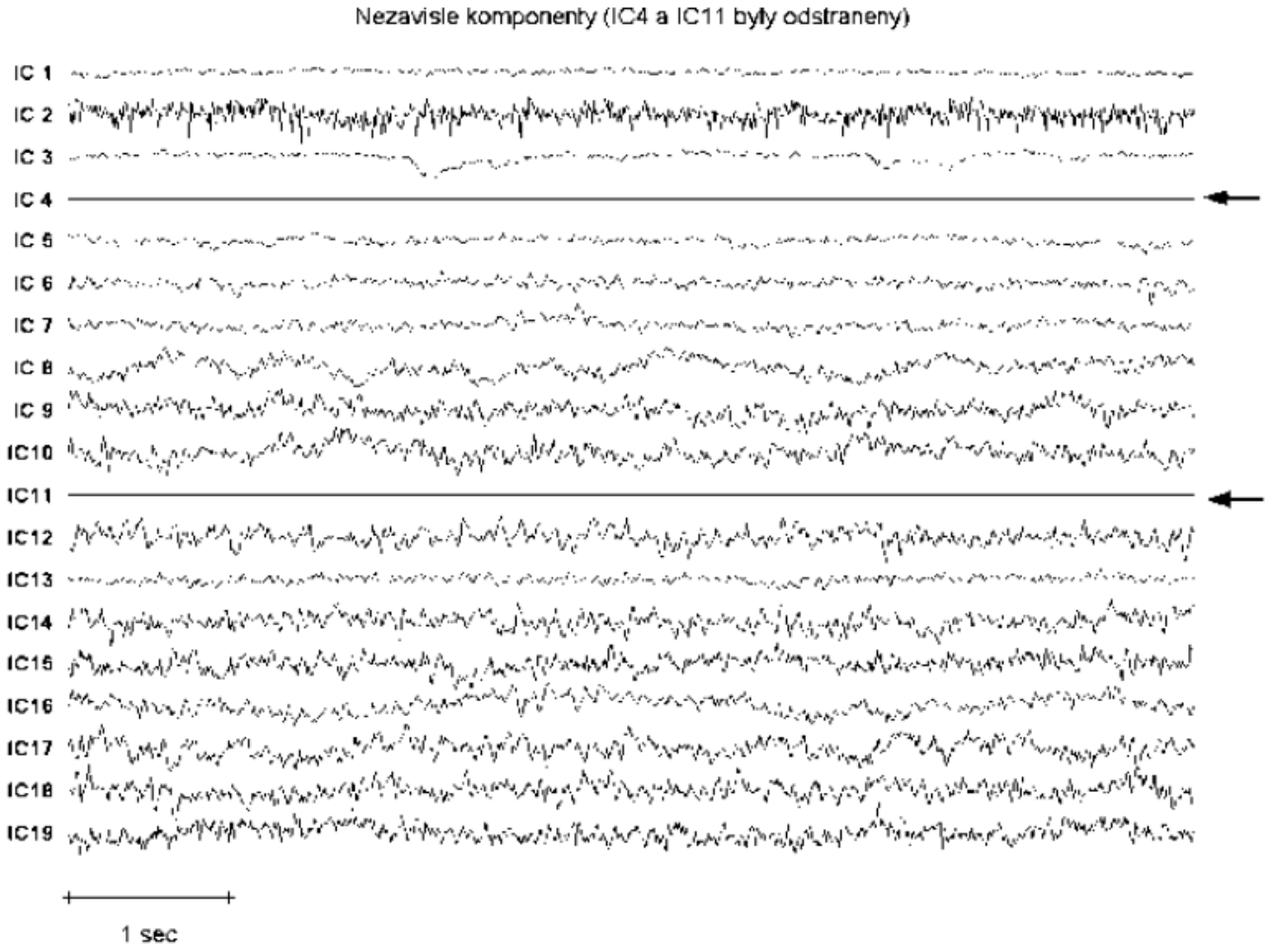
Analýza nezávislých komponent – příklad použití



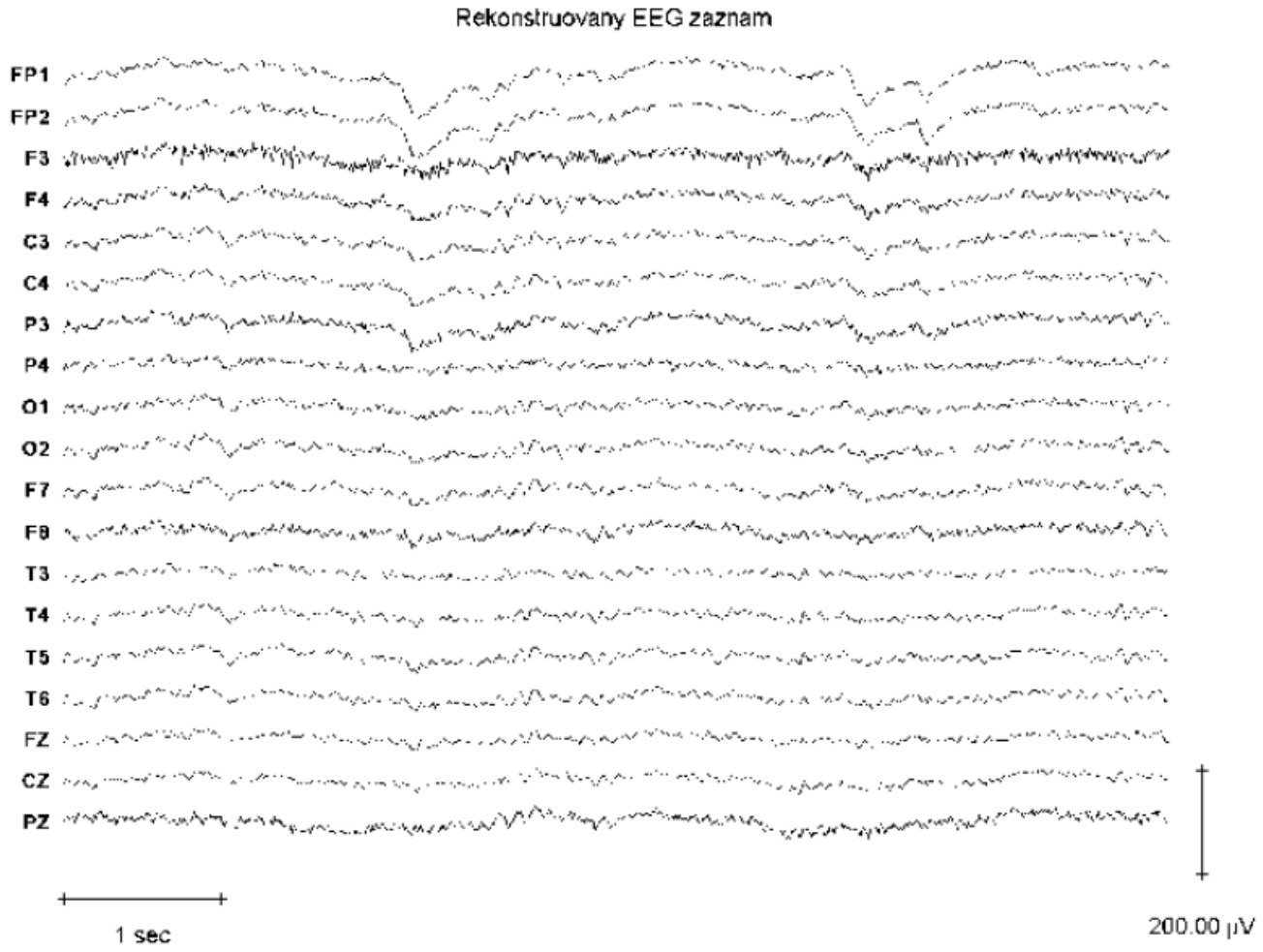
Analýza nezávislých komponent – příklad použití



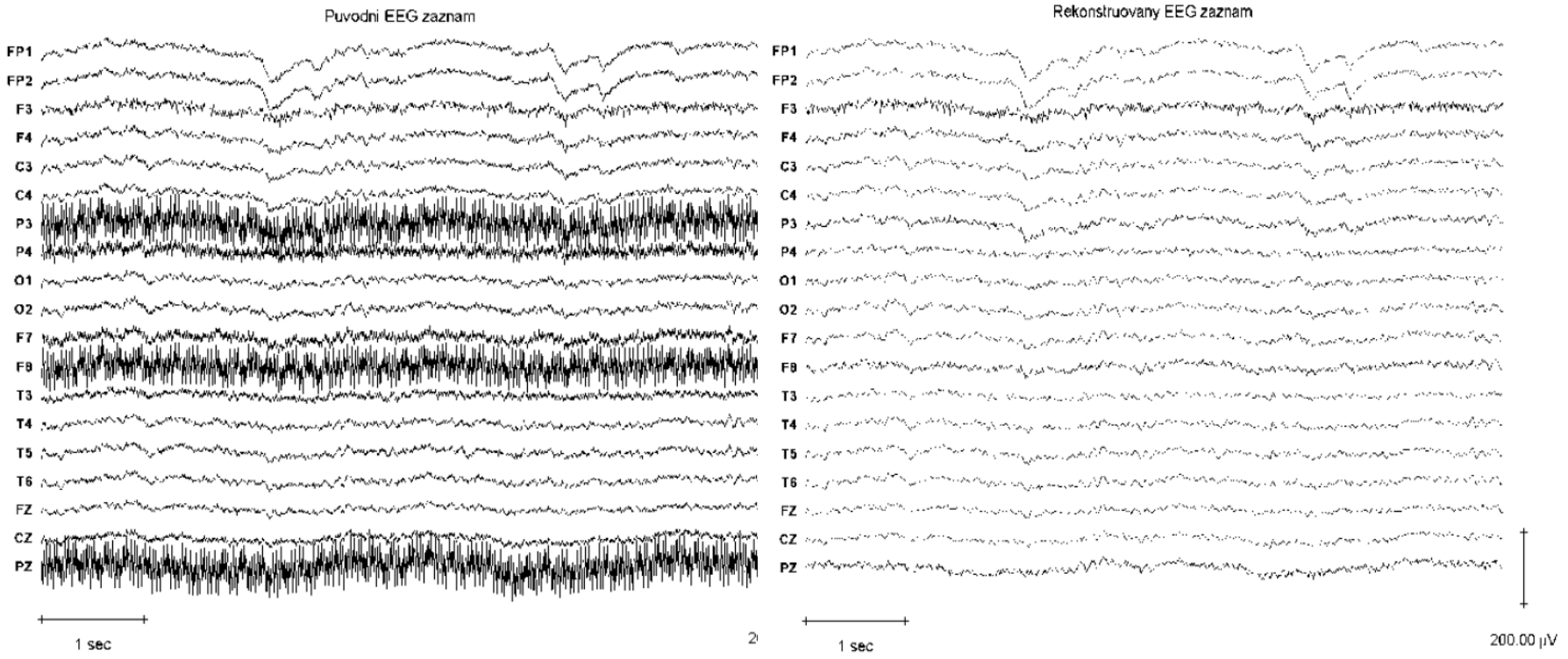
Analýza nezávislých komponent – příklad použití



Analýza nezávislých komponent – příklad použití



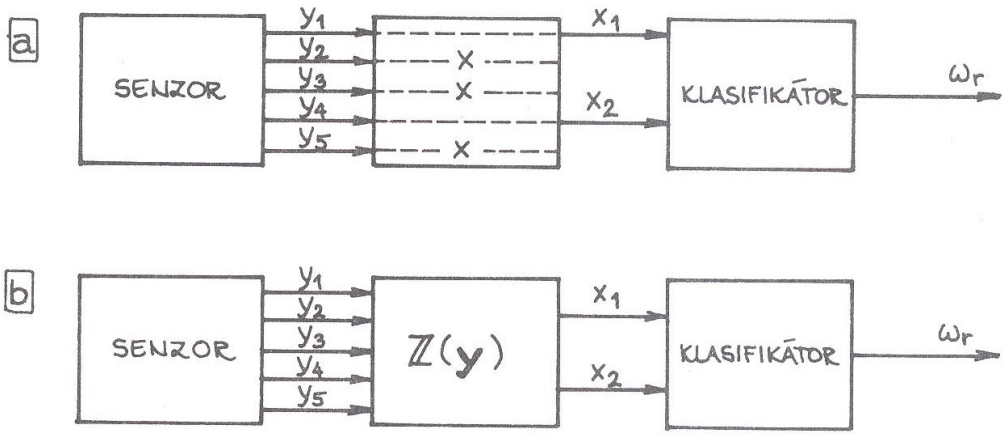
Analýza nezávislých komponent – příklad použití



Výběr příznaků

- formální popis objektu původně reprezentovaný *m* rozměrným vektorem se snažíme vyjádřit vektorem *n* rozměrným tak, aby množství diskriminační informace obsažené v původním vektoru bylo v co největší míře zachováno

- dva principiálně různé způsoby:
  1. **selekce** – nalezení a odstranění těch příznakových funkcí, které přispívají k separabilitě klasifikačních tříd nejméně
  2. **extrakce** – transformace původních příznakových proměnných na menší počet jiných příznakových proměnných



Algoritmy selekce příznaků

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

$$\Downarrow$$

hledáme jen kvazioptimální řešení

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

$$X_1 \subset X_2 \subset \ldots \subset X_j \subset \ldots \subset X_m$$

splňuje selekční kritérium vztah

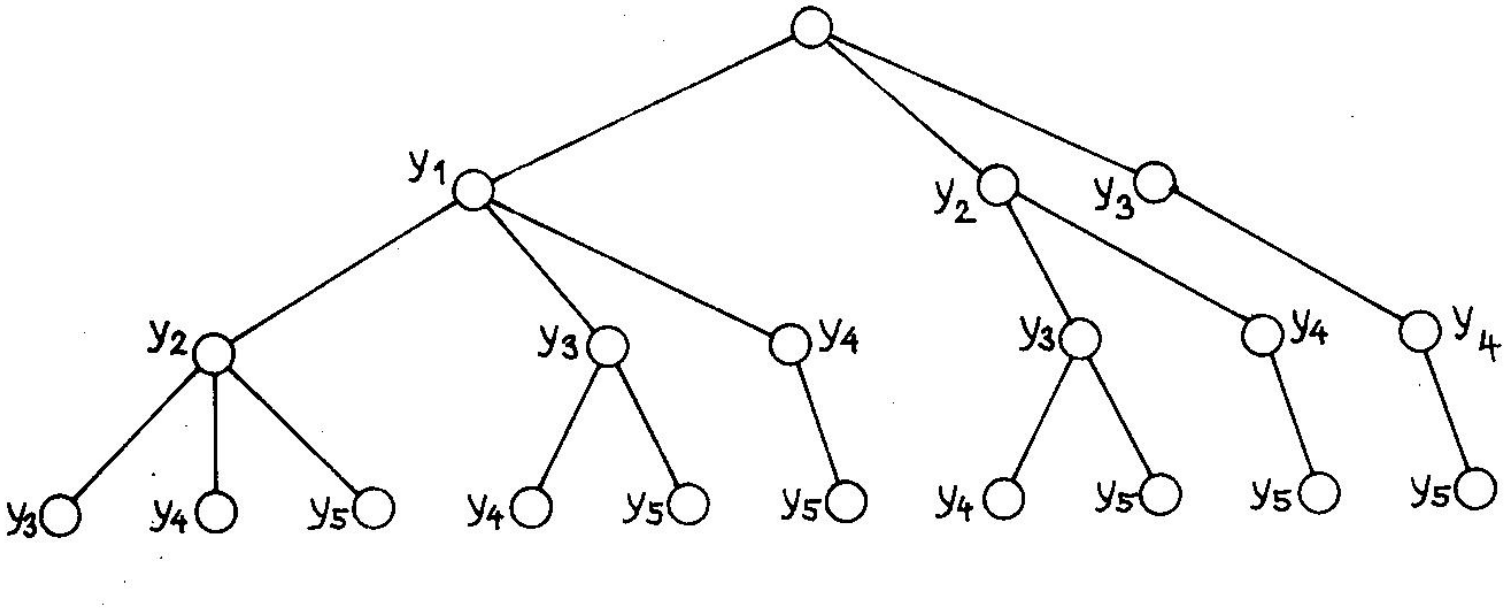
$$J(X_1) \leq J(X_1) \leq \ldots \leq J(X_m)$$

Algoritmy selekce příznaků

- Algoritmus ohraničeného větvení
- Algoritmus sekvenční dopředné selekce
- Algoritmus sekvenční zpětné selekce
- Algoritmus plus p mínus q
- Algoritmus min - max

Algoritmus ohraničeného větvení

- uvažme případ selekce dvou příznaků z pěti



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 příznakových veličin;
- v každém následujícím kroku se eliminuje ta proměnná, která způsobuje nejmenší pokles kriteriální funkce, tj. po (k+1). kroku platí

$$J(\{X_{m-k-1}\})=\max J(\{X_{m-k}-y_j\}),\ y_j \in \{X_{m-k}\}$$

- suboptimalita algoritmů sekvenční selekce:
  - dopředná selekce – suboptimalita způsobena tím, že nelze vyloučit ty veličiny, které se staly nadbytečné po přiřazení dalších veličin
  - zpětná selekce – neexistuje možnost opravy při neoptimálním vyloučení kterékoliv proměnné
- výhody sekvenčních algoritmů:
  - dopředný algoritmus je výpočetně jednodušší, protože pracuje maximálně v n-rozměrném prostoru
  - zpětný algoritmus umožňuje průběžně sledovat množství ztracené informace

Algoritmus plus p mínus q

- po přidání p veličin se q veličin odstraní;
- proces probíhá, dokud se nedosáhne požadovaného počtu příznaků;
- je-li p>q, pracuje algoritmus od prázdné množiny;
- je-li p<q, varianta zpětného algoritmu

Algoritmus min - max

- Heuristický algoritmus vybírající příznaky na základě výpočtu hodnot kriteriální funkce pouze v jedno- a dvourozměrném příznakovém prostoru.
- Předpokládejme, že bylo vybráno k příznakových veličin do množiny $\{X_k\}$ a zbývají veličiny z množiny $\{Y\text{-}X_k\}$. Výběr veličiny $y_j \in \{Y\text{-}X_k\}$ 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 veličiny již zahrnuté do množiny X_k. Vybíráme tedy veličinu y_{k+1}, pro kterou platí

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

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“

esf evropský sociální fond v ČR

EVROPSKÁ UNIE

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ