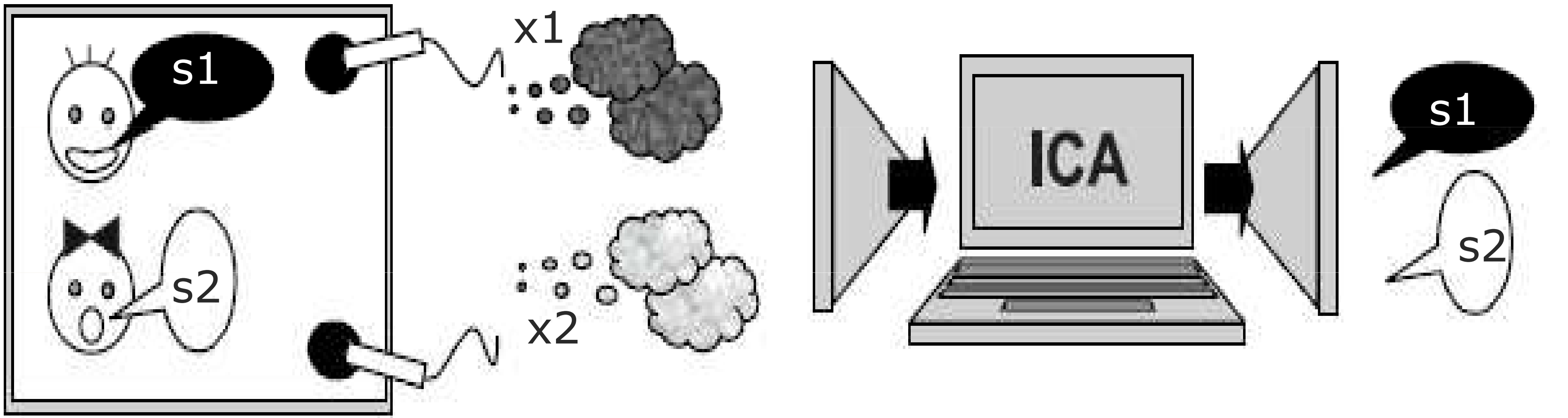
ANALÝZA A KLASIFIKACE DAT

prof. Ing. Jiří Holčík, CSc.

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

ANALÝZA NEZÁVISLÝCH KOMPONENT

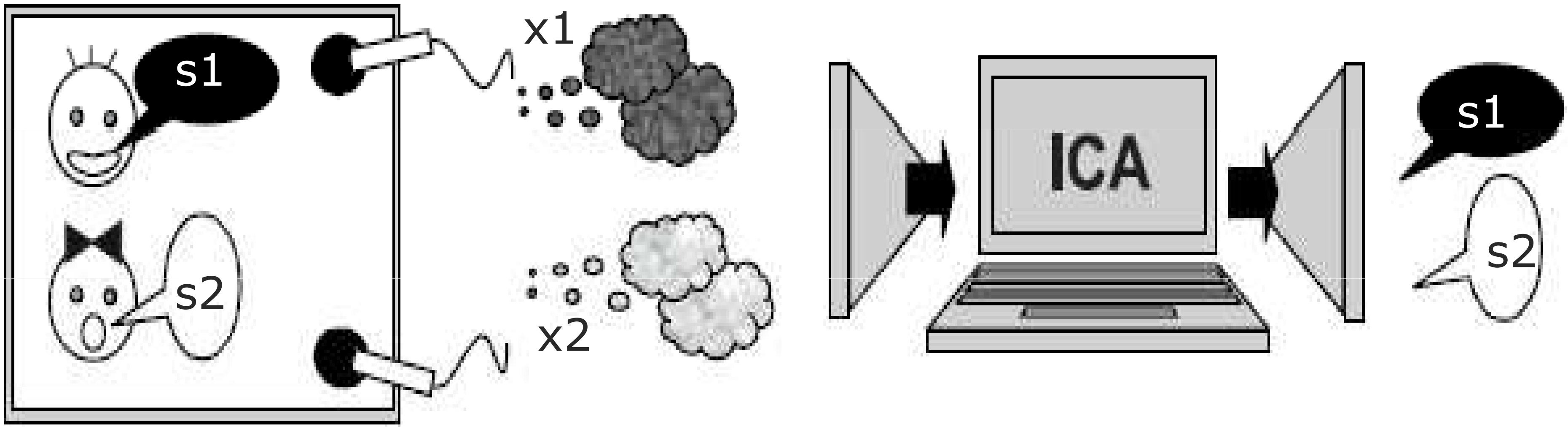
ANALÝZA NEZÁVISLÝCH KOMPONENT
PRINCIP METODY



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

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 $x_2(t)$.

ANALÝZA NEZÁVISLÝCH KOMPONENT
PRINCIP METODY



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 $\mathcal{E}(\mathbf{x})=0$).

$$x_i = a_{i1}^{orig}.s_1^{orig} + a_{i2}^{orig}.s_2^{orig} + \ldots + a_{im}^{orig}.s_m^{orig}$$

$$i = 1,2,\ldots,m$$

nebo

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

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

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

ANALÝZA NEZÁVISLÝCH KOMPONENT
MODEL DAT

- ☑ definice

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

- ☑ cíl:

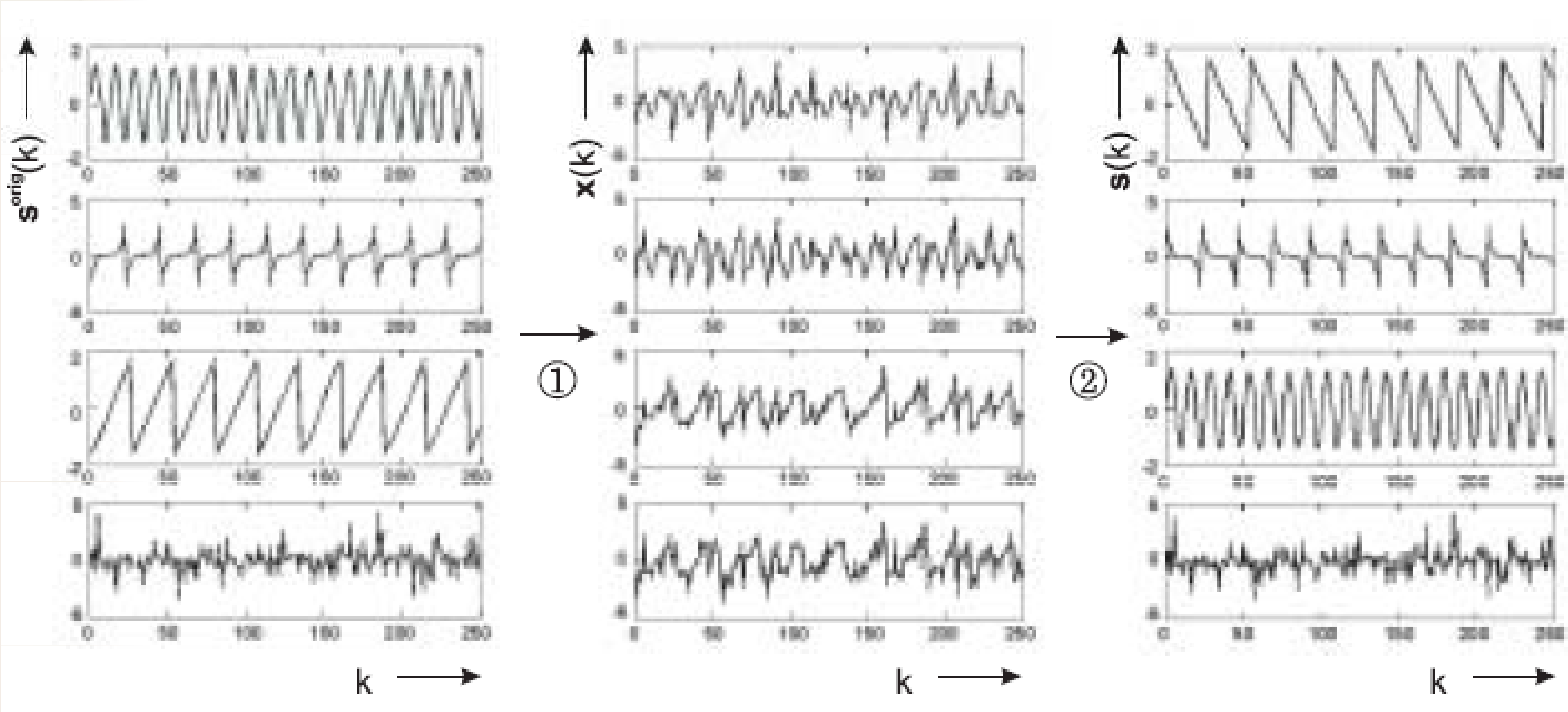
nalézt lineární transformaci (koeficienty transformační matice **W** tak, aby vypočítané nezávislé komponenty s_i byly vzájemně statisticky nezávislé [$\mathbf{W} = \mathbf{A}^{-1}$]

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

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}

ODHAD NEZÁVISLÝCH KOMPONENT
ZÁKLADNÍ ÚVAHA

- ☑ odhad nezávislých komponent si probíhá tak, že hledáme takové řádkové vektory $\mathbf{w}_i$ transformační matice **W**, aby pravděpodobnostní rozdělení součinu $\mathbf{w}_i.\mathbf{x}$ bylo „co nejvíce **nenormální**"

$$\Downarrow$$

tj. nalézt takovou transformační matici **W**, aby proměnné $\mathbf{w}_i.\mathbf{x}$ měly pravděpodobnostní rozdělení, které se co nejvíce liší od normálního

$$\Downarrow$$

potřeba nalézt míru náhodné veličiny, která by mohla být použita pro kvantifikaci míry (podobnost, vzdálenost) nenormality

ODHAD NEZÁVISLÝCH KOMPONENT
POUŽÍVANÉ MÍRY NENORMALITY

- ☑ koeficient špičatosti
- ☑ negativní normalizovaná entropie;
- ☑ aproximace negativní normalizované entropie;

ODHAD NEZÁVISLÝCH KOMPONENT
KOEFICIENT ŠPIČATOSTI

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

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 $\mathbf{s} = \mathbf{w}_i.\mathbf{x}$

ODHAD NEZÁVISLÝCH KOMPONENT
KOEFICIENT ŠPIČATOSTI

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
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

$$H(s) = -\int_{-\infty}^{\infty} p(s).\log_2 p(s)\,ds$$

ODHAD NEZÁVISLÝCH KOMPONENT
NEGATIVNÍ NORMALIZOVANÁ ENTROPIE

- ☑ entropie je tím větší, čím jsou hodnoty náhodné veličiny méně predikovatelné;
- ☑ pro normální rozdělení má entropie největší hodnotu ve srovnání v dalšími rozděleními

NNE

$$J(s) = H(s_{gauss}) - H(s),$$

kde s_{gauss} je náhodná veličiny s normálním rozdělením

ODHAD NEZÁVISLÝCH KOMPONENT
NEGATIVNÍ NORMALIZOVANÁ ENTROPIE

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 NEGATIVNÍ NORMALIZOVANÉ ENTROPIE

- ☑ použití momentů vyšších řádů

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

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

nevýhoda:

- ☑ opět menší robustnost vůči odlehlým hodnotám

ODHAD NEZÁVISLÝCH KOMPONENT
APROXIMACE NEGATIVNÍ NORMALIZOVANÉ ENTROPIE

- ☑ 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

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

ODHAD NEZÁVISLÝCH KOMPONENT
APROXIMACE NEGATIVNÍ NORMALIZOVANÉ ENTROPIE

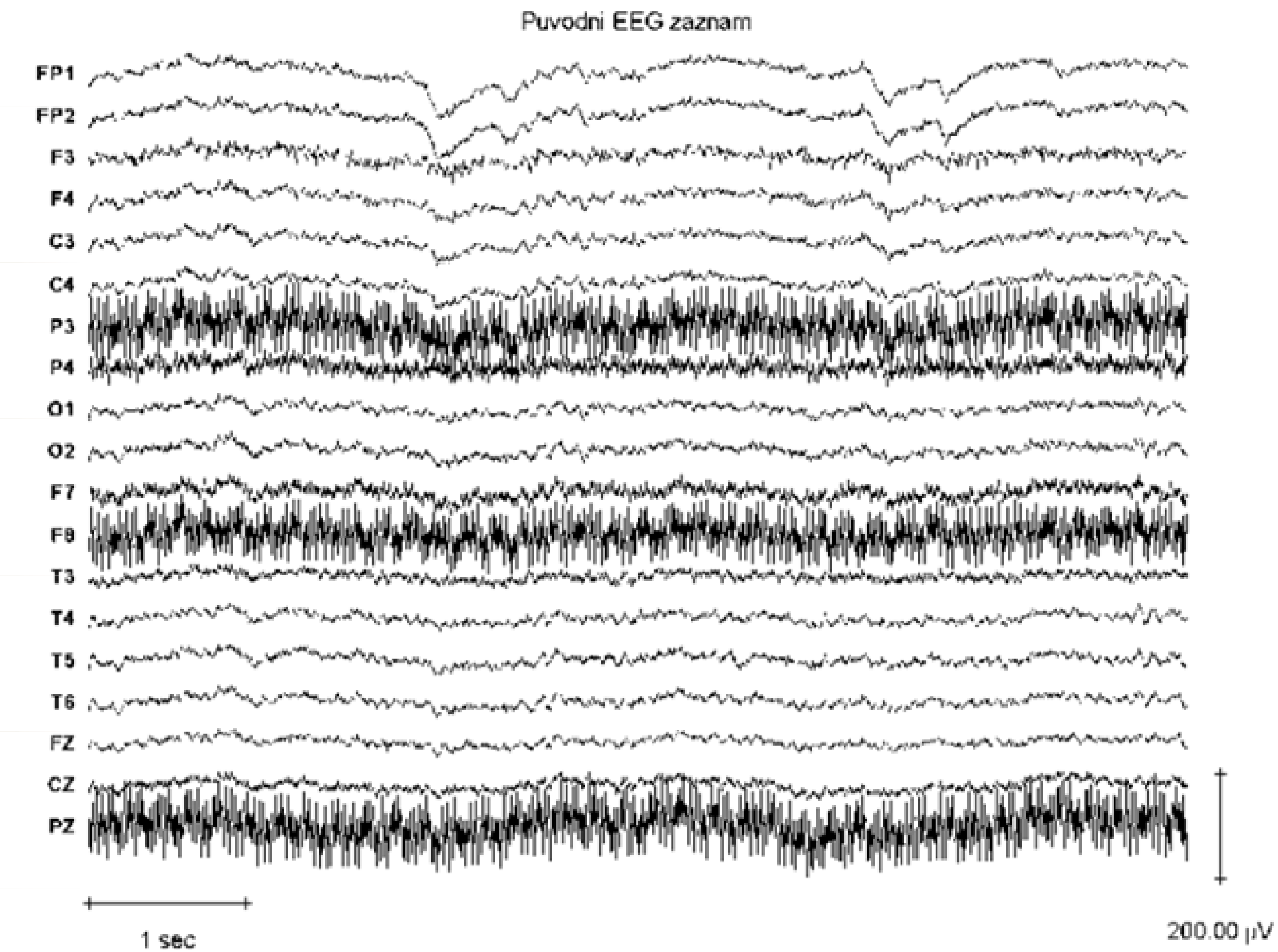
- ☑ doporučujeme:

$$G_1(s) \approx \frac{1}{a_1} \log(\cosh a_1 s)$$

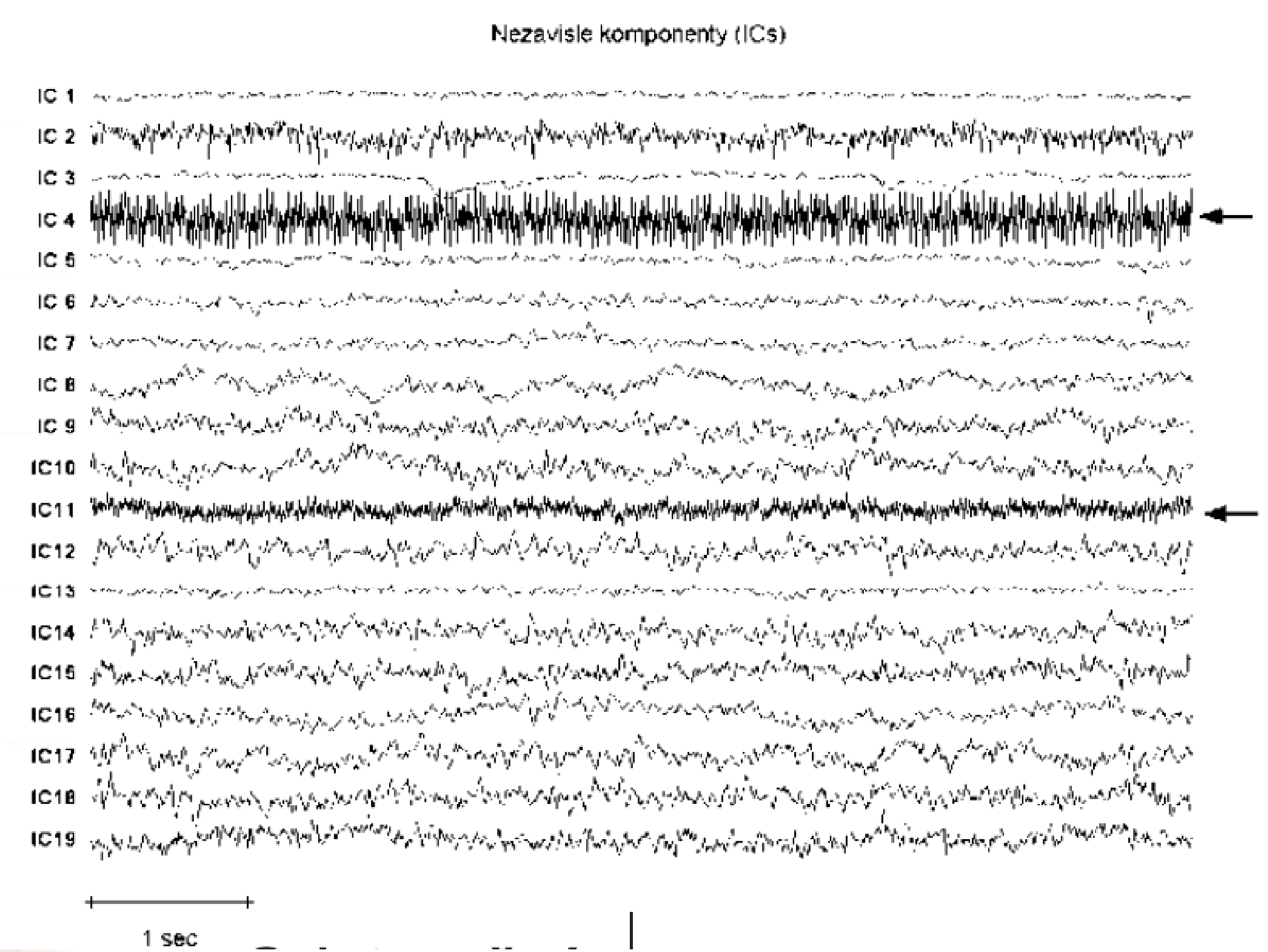
kde $a_1 \in \langle 1,2 \rangle$ nebo

$$G_2(s) \approx -\exp(-s^2/2)$$

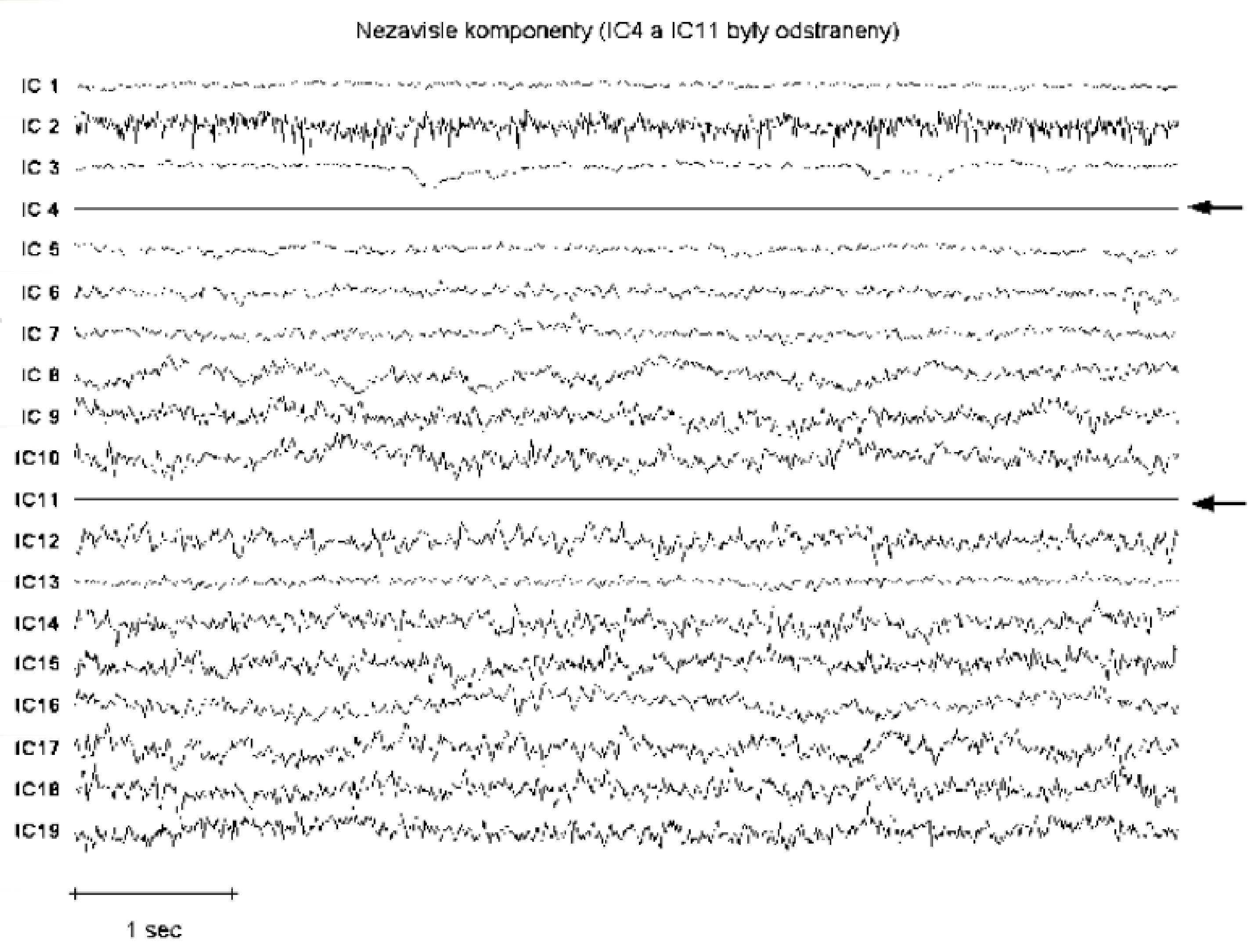
ODHAD NEZÁVISLÝCH KOMPONENT
PŘÍKLAD POUŽITÍ



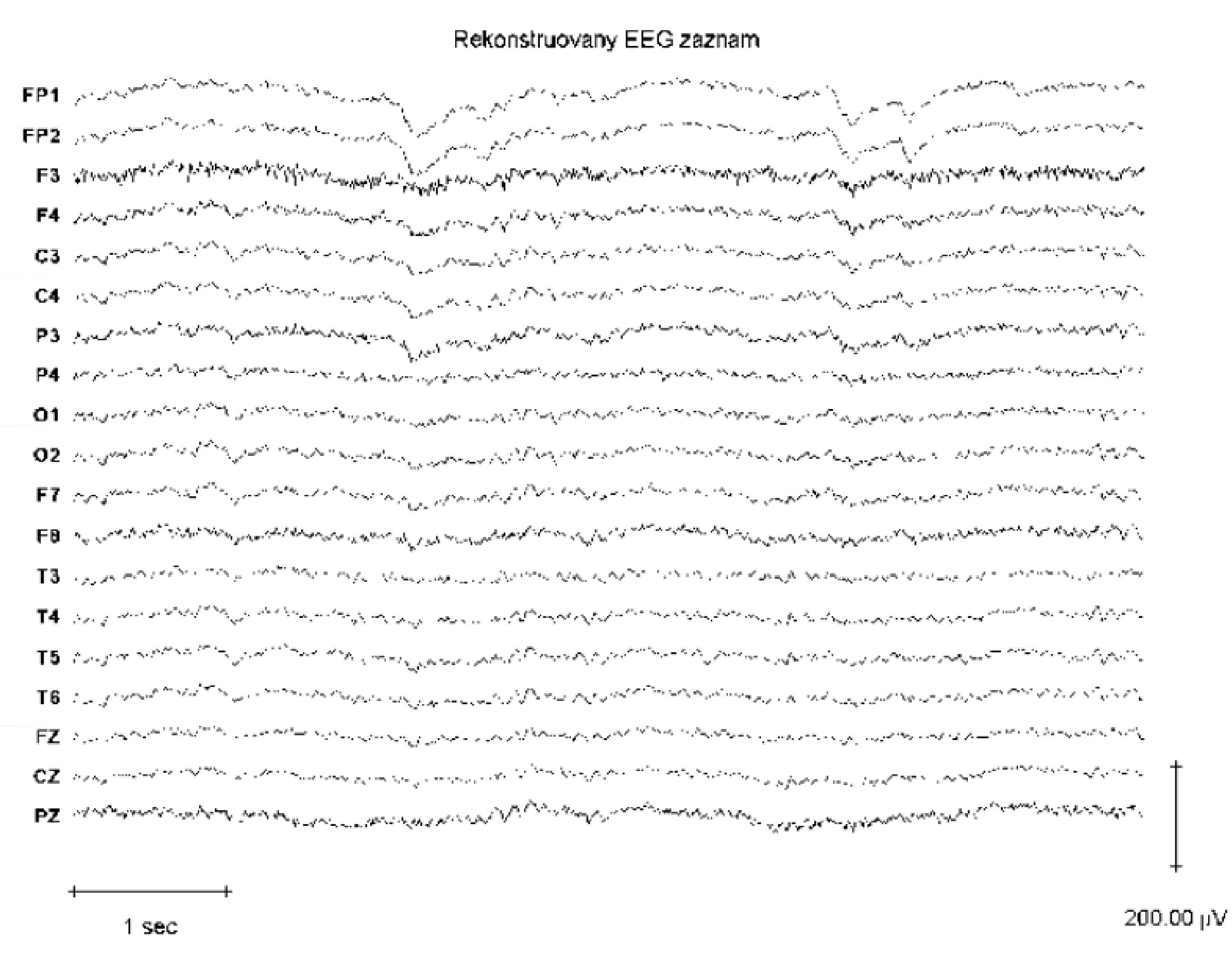
ODHAD NEZÁVISLÝCH KOMPONENT
PŘÍKLAD POUŽITÍ



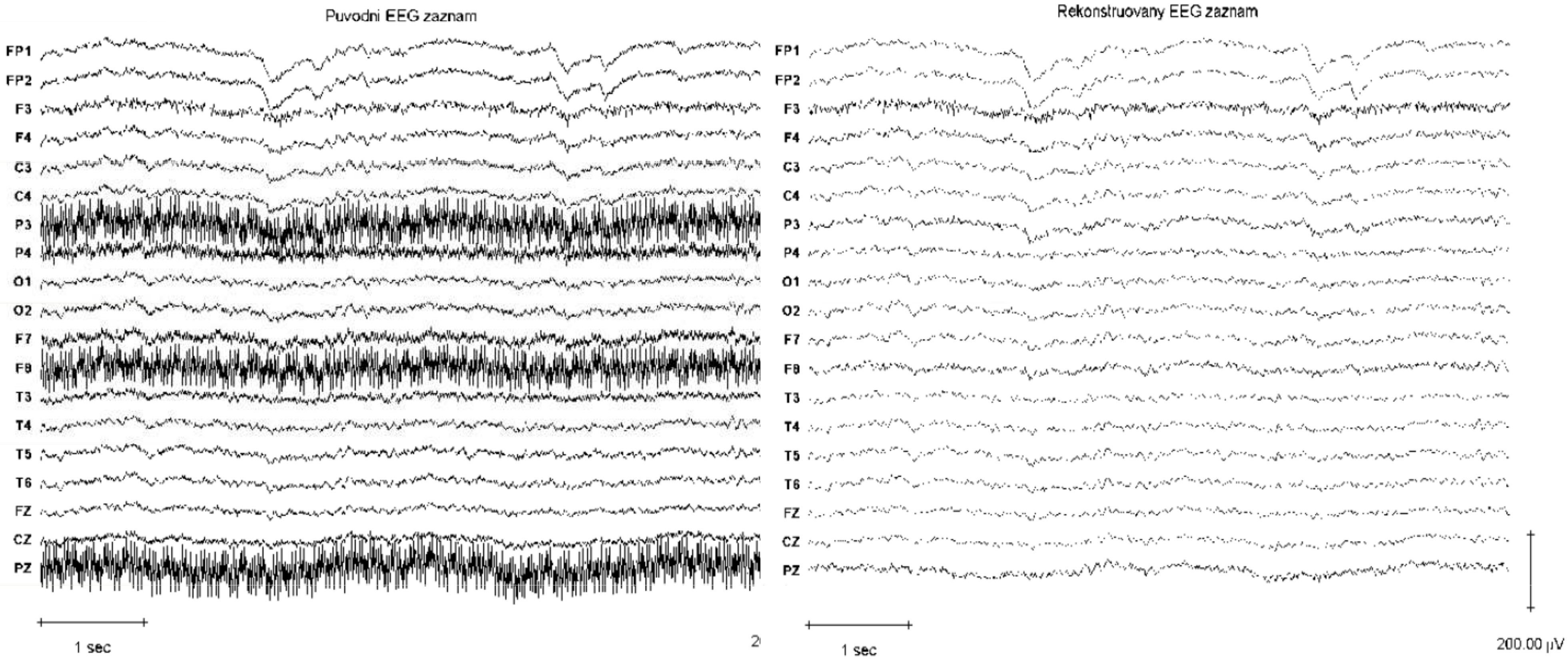
ODHAD NEZÁVISLÝCH KOMPONENT
PŘÍKLAD POUŽITÍ



ODHAD NEZÁVISLÝCH KOMPONENT
PŘÍKLAD POUŽITÍ



ODHAD NEZÁVISLÝCH KOMPONENT
PŘÍKLAD POUŽITÍ



Příprava nových učebních materiálů oboru Matematická biologie

je podporována projektem ESF

č. CZ.1.07/2.2.00/07.0318

„VÍCEOBOROVÁ INOVACE STUDIA MATEMATICKÉ BIOLOGIE“

EVROPSKÁ UNIE

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ