Unsupervised learning
"učenie bez učiteľa"
I. redukcia dimenzie
2. časť

Zuzana Rošťáková

14. október 2021

Seminár UM

Dáta

- vstupné údaje: na skupine **n** objektov pozorujeme **p** premenných

$$X = \begin{bmatrix} — & \mathbf{x_1^T} & — \\ — & \mathbf{x_2^T} & — \\ & \vdots & \\ — & \mathbf{x_n^T} & — \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \mathbf{x_1^\star} & \mathbf{x_2^\star} & \dots & \mathbf{x_p^\star} \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n \times p}$$

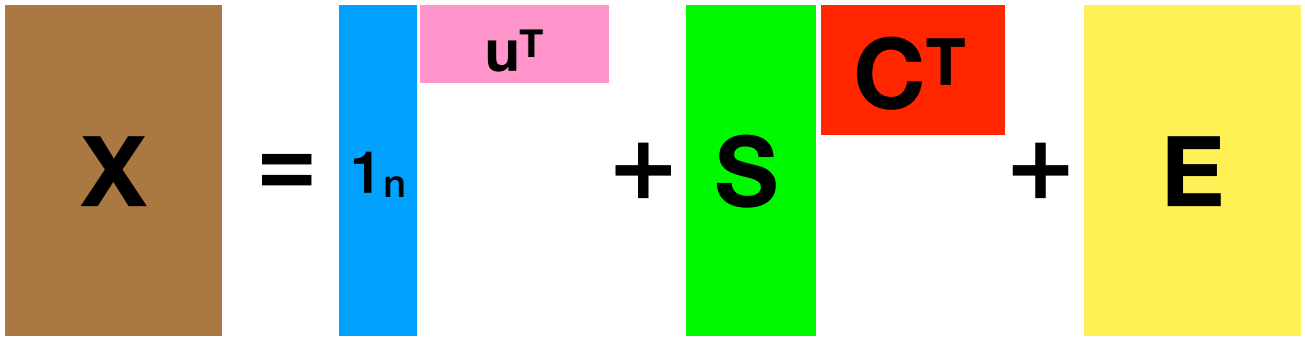
- $\mathbf{x_i} = (x_{i1}, x_{i2}, \dots, x_{ip})^T$

  $\rightarrow$ vektor hodnôt **p** premenných pre i-ty objekt, $i = 1, \dots, n$

- $\mathbf{x_l^\star} = (x_{1l}, x_{2l}, \dots, x_{nl})^T$

  $\rightarrow$ vektor hodnôt l-tej premennej, $l = 1, \dots, p$ pre **n** objektov

Redukcia dimenzie



$$x_{ij} = u_j + \sum_{k=1}^{K} s_{ik} c_{kj} + e_{ij}$$

- $S \in \mathbb{R}^{n \times K} \rightarrow$ hodnoty K "nových" premenných pre n pozorovaní
- $C \in \mathbb{R}^{p \times K} \rightarrow$ "vzťah" medzi K novými a p pôvodnými premennými
- $\mathbf{u} \in \mathbb{R}^p \rightarrow$ vektor priemerov jednotlivých premenných

$$u_j = \frac{1}{n} \sum_{l=1}^{n} x_{lj}, \qquad j = 1, \ldots, p$$

- $1_n = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$
- $E \in \mathbb{R}^{n \times p} \rightarrow$ matica chýb modelu

Predchádzajúci seminár

- **metóda hlavných komponentov**
  - nové premenné = lineárna kombinácia pôvodných premenných
  - len transformácia údajov
  - hlavné komponenty sú navzájom ortogonálne
  - nie je škálovo invariantná
- **faktorová analýza**
  - predpoklad existencie latentných (skrytých) premenných
  - nové premenné (faktory) sú nekorelované
  - rotácia faktorov - môže viesť k lepšej interpretácií

Metóda nezávislých komponentov

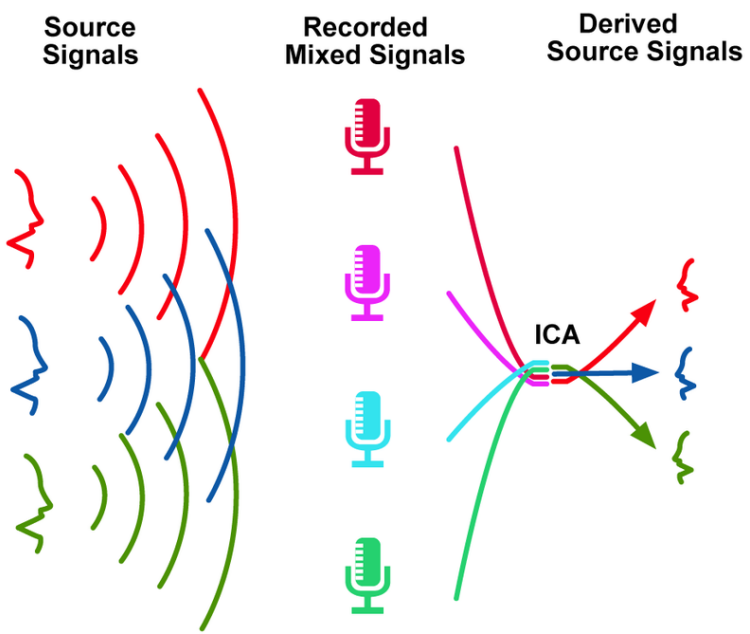
Independent component analysis (ICA)

ICA - Metóda nezávislých komponentov

- **cieľ:** faktorizácia dát na aditívne komponenty, ktoré sú:
  - negaussovské
  - nezávislé
- počet nových premenných K je vstupný parameter
- nové premenné môžu reprezentovať latentné komponenty rovnako ako aj šum
- nové premenné nie sú usporiadané, t.j. musíme prejsť všetky a nájsť tie, ktoré majú najlepšiu interpretáciu

ICA - Metóda nezávislých komponentov

$$X = \mathbf{1_n u}^T + SC^T + E \qquad x_{ij} = u_j + \sum_{k=1}^{K} s_{ik} c_{kj} + e_{ij}$$



[Karczewski et al., 2014]

- tzv. coctail party problem
- $S \in \mathbb{R}^{n \times K}$
  $\rightarrow$ K ľudí rozpráva v n čas. bodoch
  $\rightarrow$ naraz a nezávisle
- $X \in \mathbb{R}^{n \times p}$
  $\rightarrow$ zvuk z p mikrofónov v n čas. bodoch
- $C \in \mathbb{R}^{p \times K}$
  $\rightarrow$ matica parametrov závislých napr. na vzdialenosti medzi rečníkom a mikrofónom

ICA v MATLAB-e

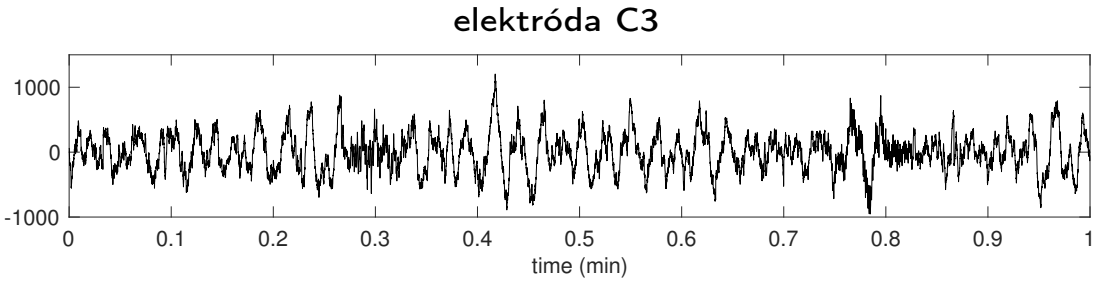
- nie je priamo v MATLAB-e $\Rightarrow$ balíky z MathWorks, GitHub, ...
- **joint approximation diagonalization of eigen-matrices (JADE)**
  - [Cardoso and Souloumiac, 1994]
  - https://github.com/ruohoruotsi/Riddim/blob/master/MATLAB/jade.m

  B = jader(Y,K)

  Y $= X^T$, matica *počet senzorov (p)* $\times$ *počet vzoriek (n)*
  K - počet komponentov; ak K nie je dané, tak $K = p$
  B $= C^{-1}$, tzv. $p \times p$ separačná matica, $S^T = BY$
- **FastICA**
  - [Hyvärinen and Oja, 2000]
  - napr. https://www.mathworks.com/matlabcentral/fileexchange/38300-pca-and-ica-package
- ...

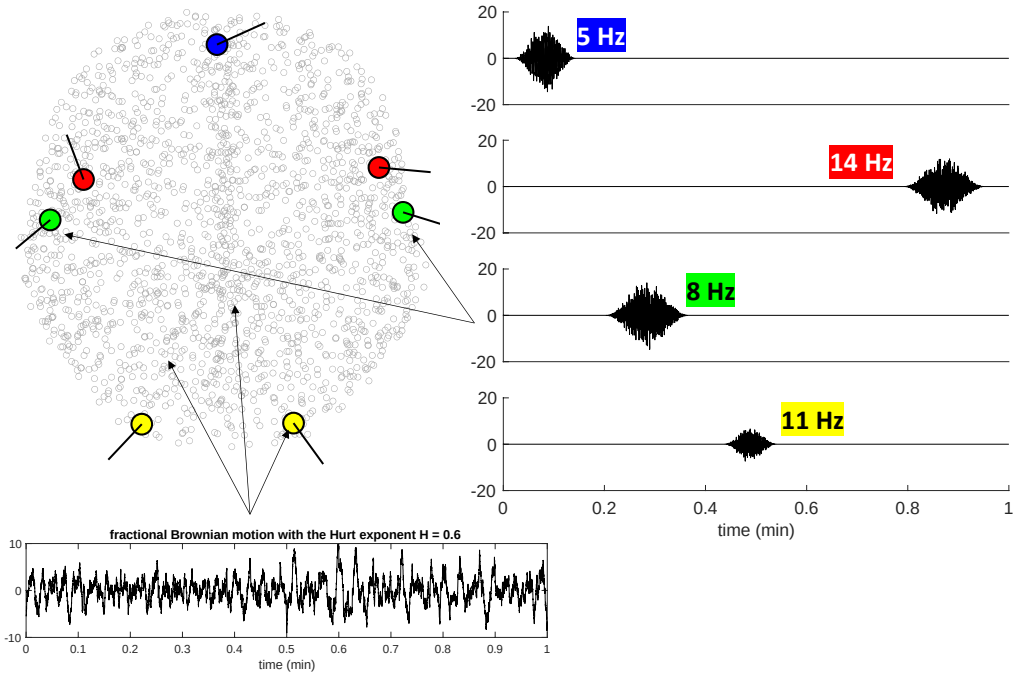
ICA - Príklad: Analýza EEG signálu

- simulovaný EEG signál (64 elektród, 1 minúta) $\rightarrow X \in \mathbb{R}^{n \times 64}$



$\rightarrow$ elektróda zachytáva signál z viacerých kortikálnych zdrojov

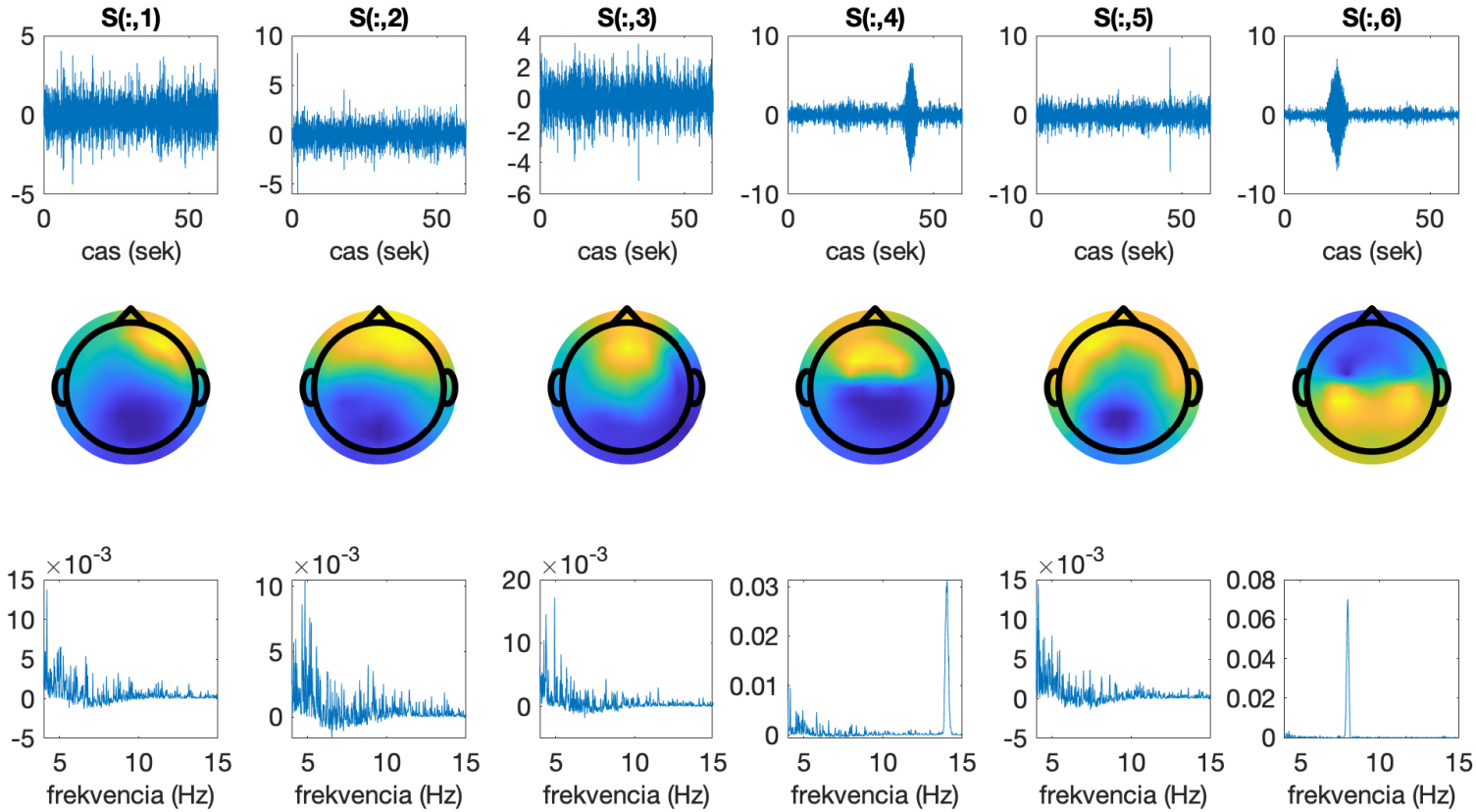
- **cieľ**: detekovať skryté zdroje oscilačnej aktivity



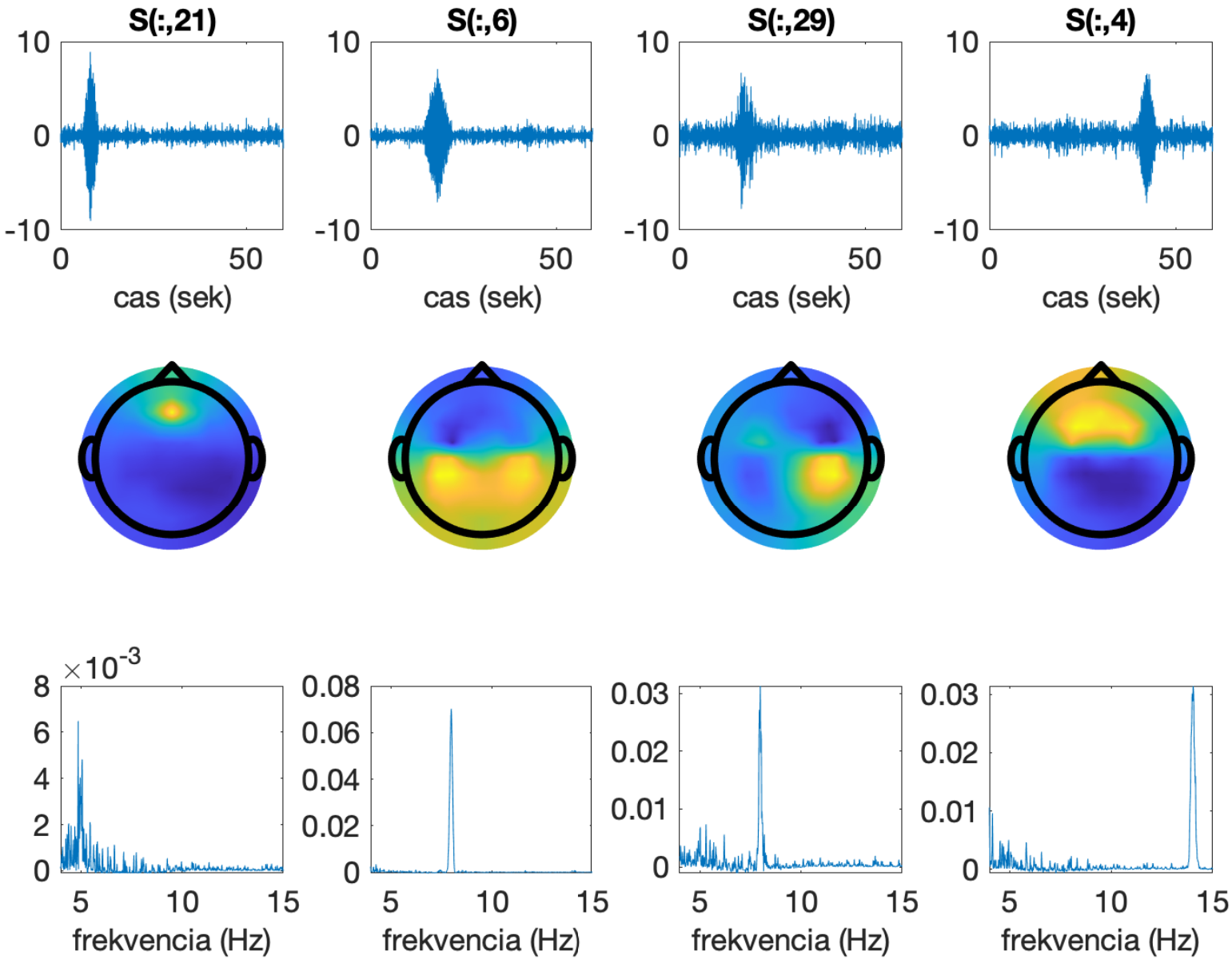
ICA - Príklad: Analýza EEG signálu

- $S \in \mathbb{R}^{n \times K}$ - časový priebeh K zdrojov oscilačnej aktivity
- $C \in \mathbb{R}^{64 \times K}$ - priestorové umiestnenie zdrojov oscilačnej aktivity

$\rightarrow$ interpretáciu majú komponenty 4 a 6 . . .



ICA - Príklad: Analýza EEG signálu



ICA - zhrnutie

- komponenty sú nezávislé
- nevieme odhadnúť varianciu nezávislých komponentov
  - $\rightarrow$ nevieme určiť poradie nezávislých komponentov
    - každý komponent treba "manuálne" prezrieť, akú má interpretáciu
- K je vstupný parameter
  - potrebné určiť vhodnú hodnotu

Nezáporná maticová faktorizácia

Nonnegative matrix factorisation (NNMF)

NNMF - Nezáporná maticová faktorizácia

- **cieľ:** nájsť faktorizáciu dát tak, aby matice S, C boli nezáporné
- **predpoklad:** nezápornosť skrytých faktorov
- **využitie:**
  - $\rightarrow$ dáta sú nezáporné
  - $\rightarrow$ ortogonalita / nekorelovanosť / nezávislosť faktorov nemá interpretáciu
- počet faktorov K je vstupný parameter

NNMF - Nezáporná maticová faktorizácia

$$X = \mathbf{1_n u}^T + SC^T + E$$

- X - $n \times p$ matica (nezáporných) dát
- S - nezáporná $n \times K$ matica hodnôt skrytých faktorov pre n pozorovaní
- C - nezáporná $p \times K$ matica faktorových nákladov
- K - počet skrytých faktorov

$\rightarrow$ nie je to presná faktorizácia X, len tzv. "low-rank" aproximácia

NNMF v MATLAB-e

[S, Ct, D] = *nnmf*(X,K,Name,Value)

X - dáta, matica $n \times p$

K - počet latentných faktorov

- voliteľné parametre
  - Algorithm - algoritmus na odhad S, C, napr. 'als' alebo 'mult'
  - w0, h0 - inicializačné odhady pre S (w0), C (h0)
  - Replicates - počet opakovaní algoritmu
    $\rightarrow$ algoritmus je iteračný, môže skončiť v lokálnom optime
  - ...

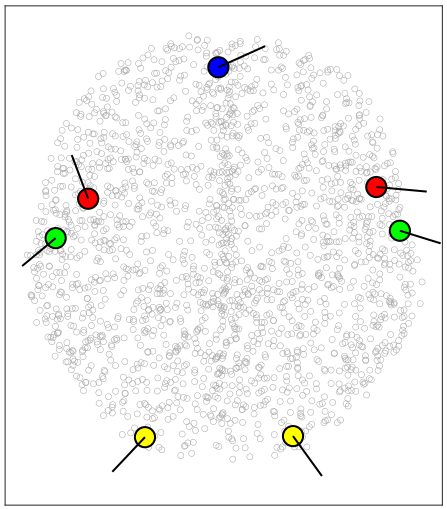
S - nezáporná $n \times K$ matica skrytých faktorov

Ct $= C^T$ - nezáporná $K \times p$ matica faktorových nákladov

D - druhá odmocnica súčtu štvorcov rezíduí

NNMF - Príklad: priemerné spektrum

- simulovaný EEG signál
  - obsahuje len 4 oscilácie
    $\rightarrow$ 5 Hz, 8 Hz, 11 Hz, 14 Hz
  - 64 elektród, 1 minúta



$\rightarrow$ pri ICA sme pracovali v časovo-priestorovej oblasti

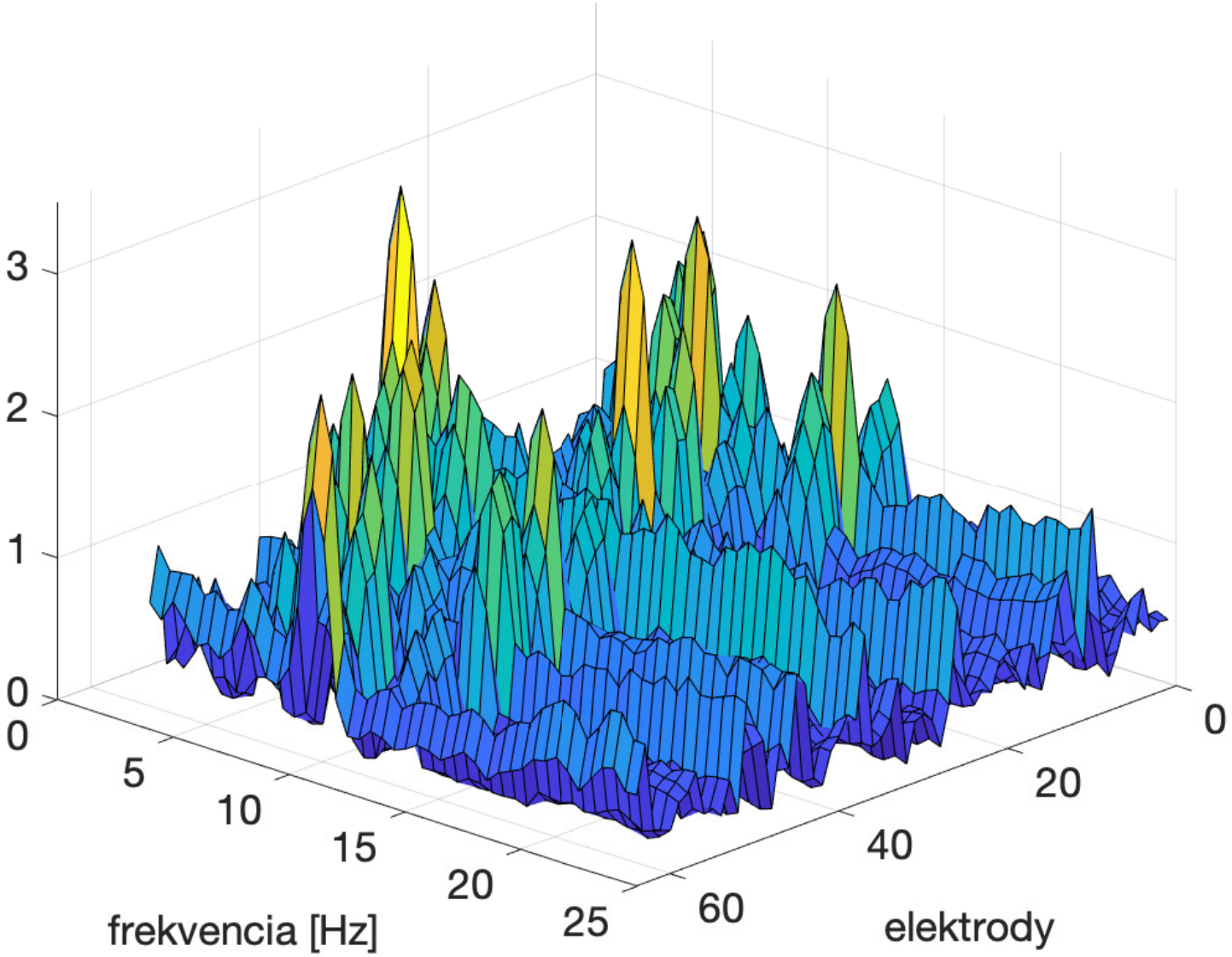
$\rightarrow$ teraz pracujeme vo frekvenčno-priestorovej oblasti

- pre každé 2-sekundové časové okno vypočítame spektrum v intervale 4 - 25 Hz s krokom 0.5 Hz (t.j. v 43 bodoch)

  $\rightarrow$ priemer spektier cez časové okná
- **dáta:** priemerné spektrum EEG signálu na 64 elektródach

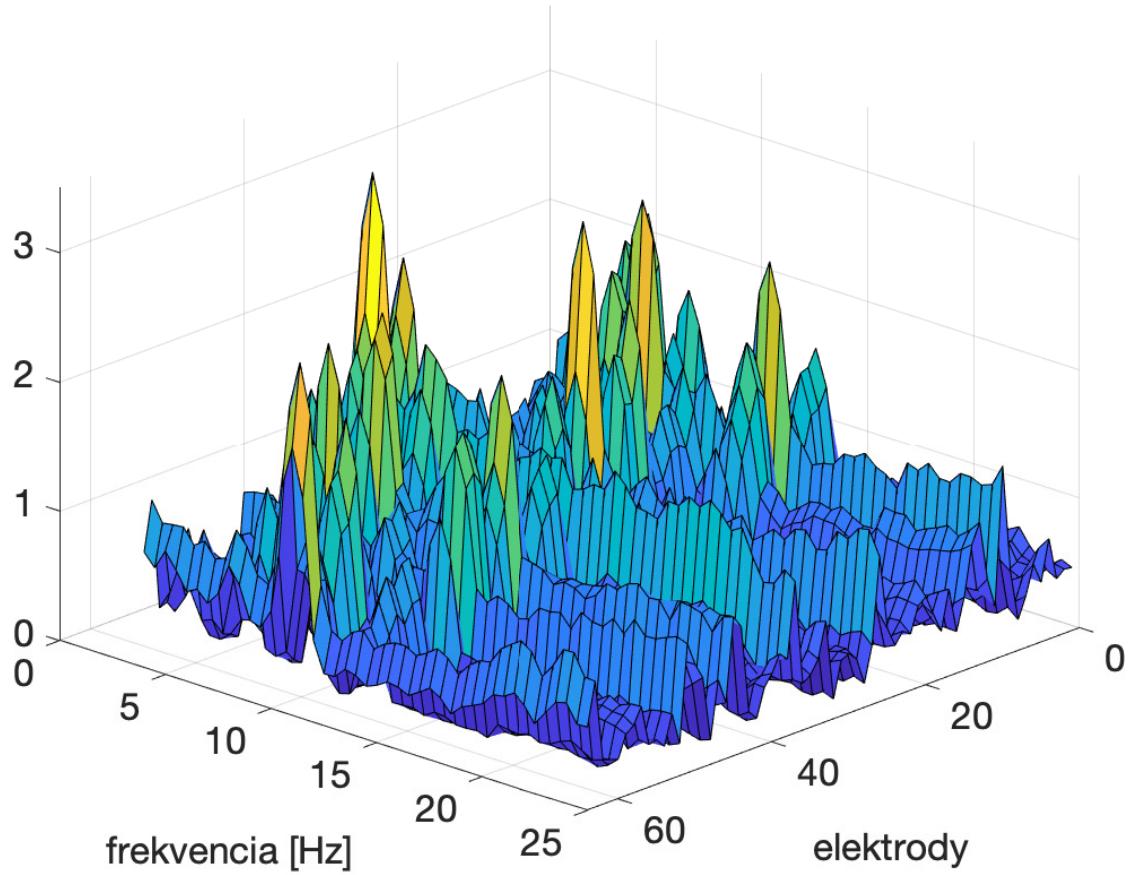
$$X \in \mathbb{R}_{+}^{43\times 64}$$

NNMF - Príklad: priemerné spektrum



NNMF - Príklad: priemerné spektrum

- **cieľ:** nájsť skryté zdroje oscilačnej aktivity
  $\rightarrow$ ich priestorové a frekvenčné charakteristiky



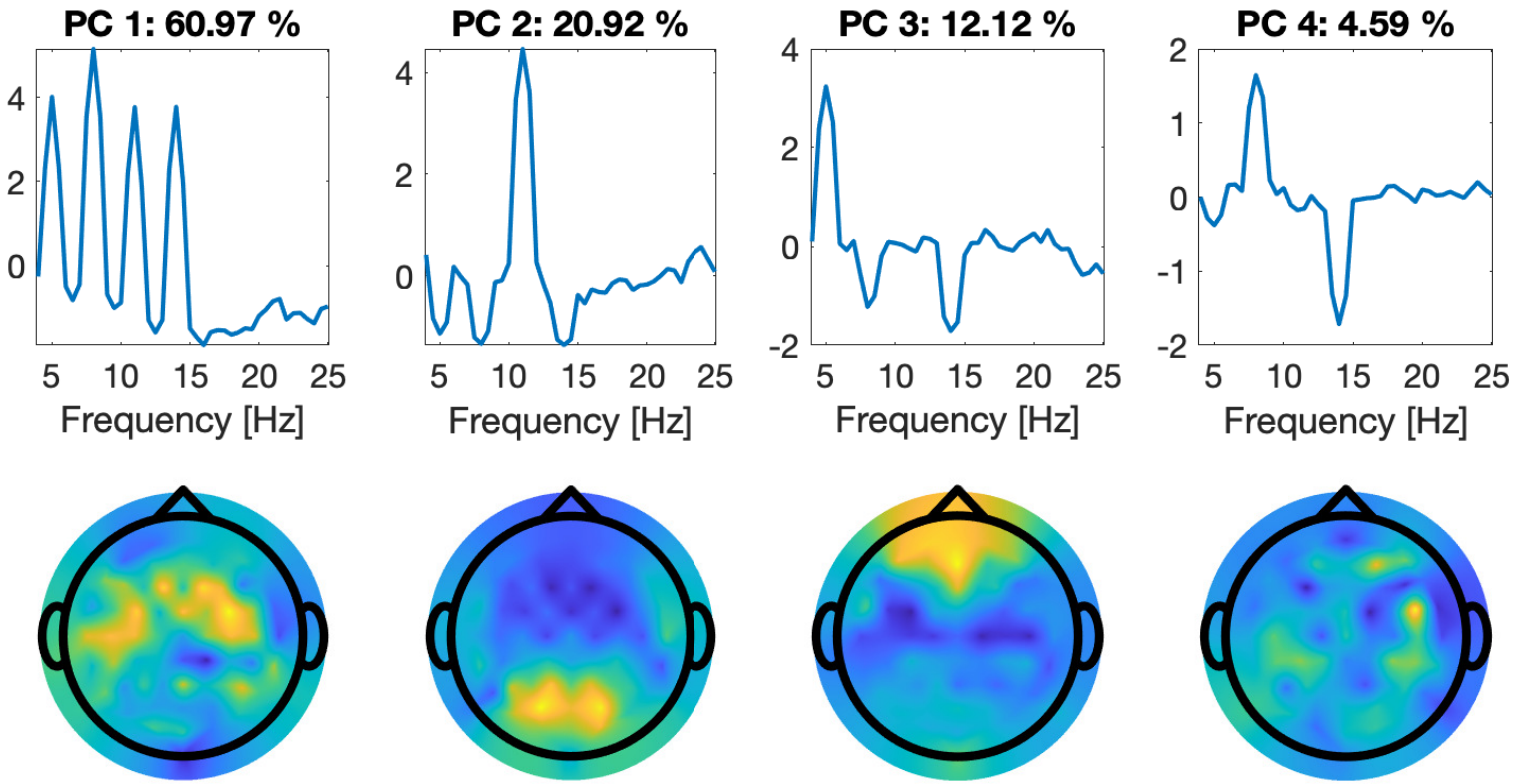
NNMF - Príklad: priemerné spektrum

1.) PCA

S = frekvenčné charakteristiky skrytých oscilačných rytmov

C = vzťah so 64 elektródami $\rightarrow$ priestorové charakteristiky

- prvé 4 komponenty vysvetľujú viac ako 98% variability v dátach



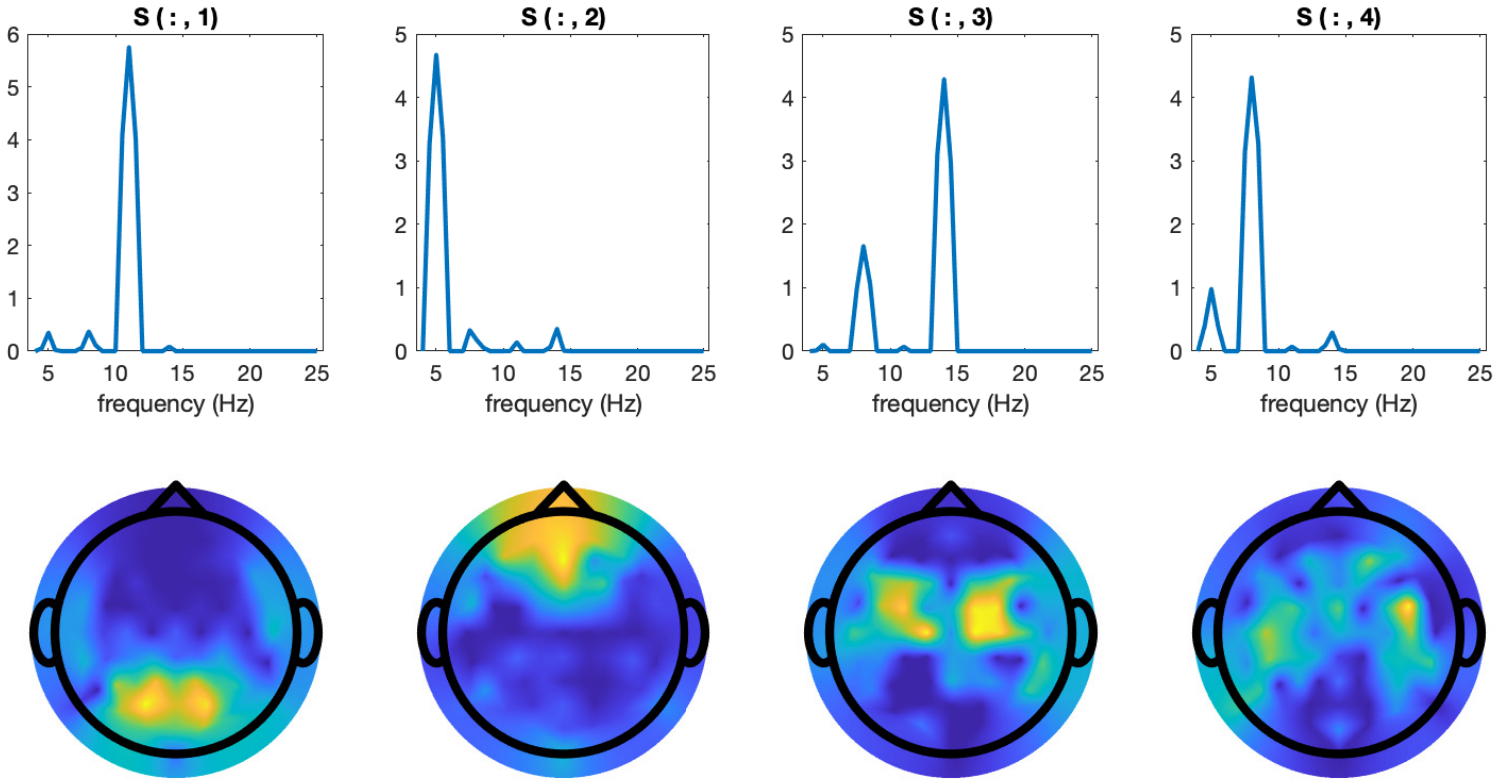
NNMF - Príklad: priemerné spektrum

2.) NNMF so 4 faktormi ($K = 4$)

S = frekvenčné charakteristiky skrytých oscilačných rytmov

C = vzťah so 64 elektródami $\rightarrow$ priestorové charakteristiky

- komponenty nie sú usporiadané



NNMF - zhrnutie

- **výhody:**
  - za predpokladu nezápornosti skrytých faktorov môže dať interpretačne lepšie riešenie ako PCA
- **"nevýhody":**
  - iteračný algoritmus na odhad $S, C \rightarrow$ môže skončiť len v lokálnom optime
    $\Rightarrow$ viacero spustení algoritmu
  - nie je to exaktná faktorizácia, ale len tzv. "low-rank" aproximácia
  - stanovenie počtu faktorov
    - obtiažnejšie než pri EFA, ICA
    - príliš veľké $K \rightarrow$ matice S, C môžu mať $rank << K$, t.j. obsahujú aj nulové stĺpce

Otázky ?

Literatúra I

Cardoso, J. and Souloumiac, A. (1994).

Blind beamforming for non gaussian signals.

Radar and Signal Processing, IEE Proceedings F, 140:362 – 370.

Carroll, J. D. and Chang, J.-J. (1970).

Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition.

Psychometrika, 35(3):283–319.

Harshman, R. A. (1970).

Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis.

UCLA Working Papers in Phonetics, 16:1–84.

Hyvärinen, A. and Oja, E. (2000).

Independent component analysis: algorithms and applications.

Neural Networks, 13:411–430.

Karczewski, K., Snyder, M., and Altman, R. (2014).

Coherent functional modules improve transcription factor target identification, cooperativity prediction, and disease association.

PLoS genetics, 10:e1004122.

Ramsay, J. O. and Silverman, B. W. (2005).

Functional data analysis.

Springer Series in Statistics, New York, second edition.

Literatúra II

Tucker, L. R. (1966).
Some mathematical notes on three-mode factor analysis.
Psychometrika, 31(3):279–311.