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

Zuzana Rošťáková

23. september 2021

Seminár UM

Dáta

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

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

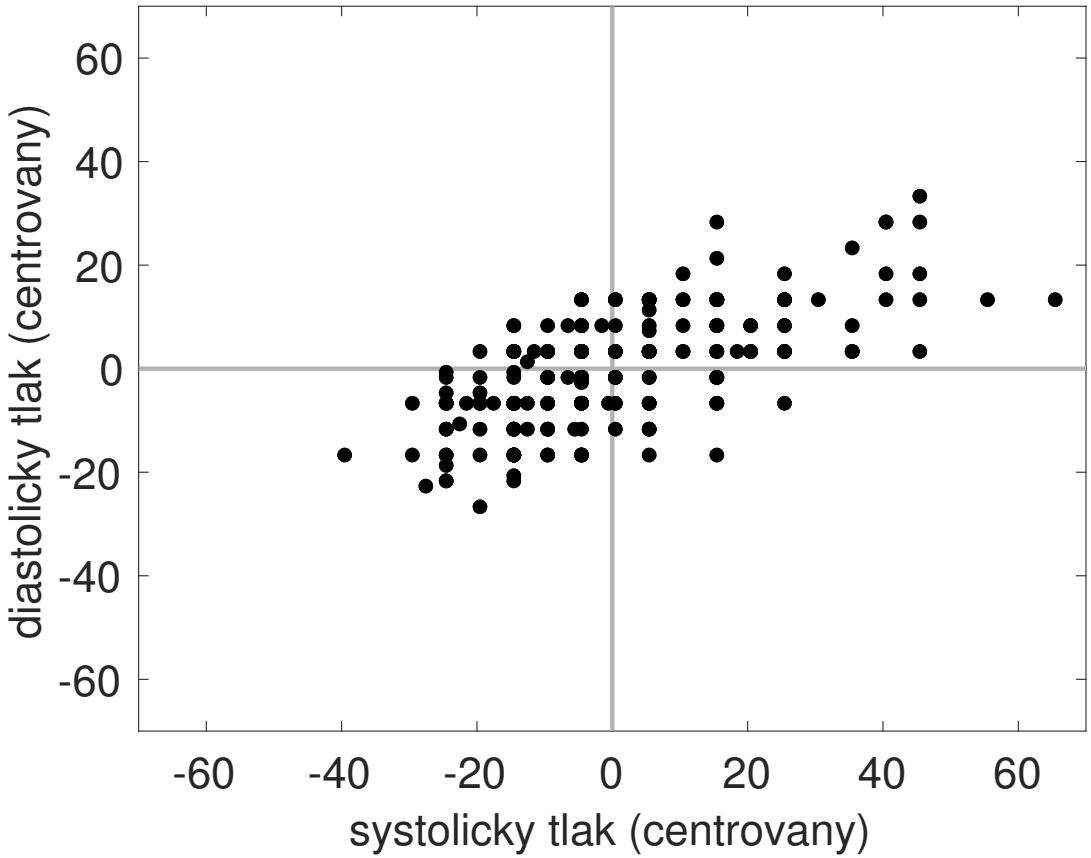
- $\mathbf{x_i} = (x_{i1}, x_{i2}, \dots, x_{ip})^T$

  $\rightarrow$ 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

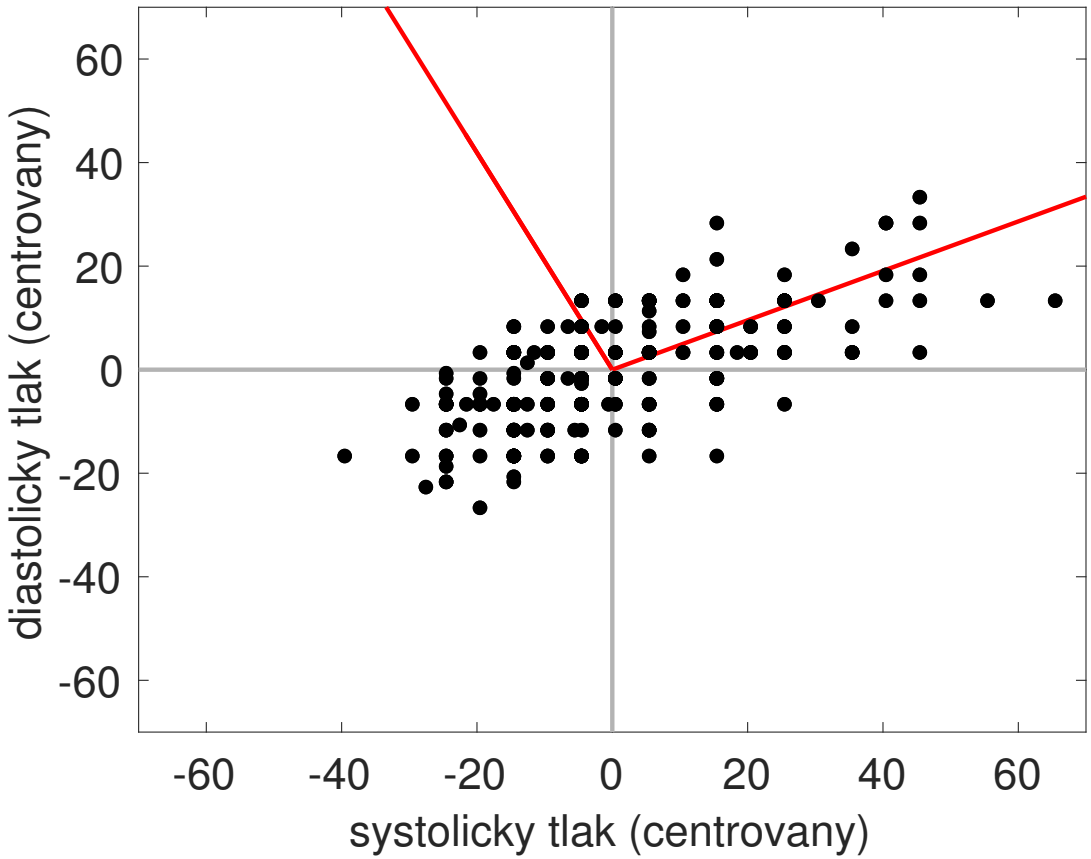
Redukcia dimenzie



→ premenné môžu byť nositeľmi "podobnej" informácie

⇒ stačí jedna (nová) premenná, napr. ich vhodná lineárna kombinácia

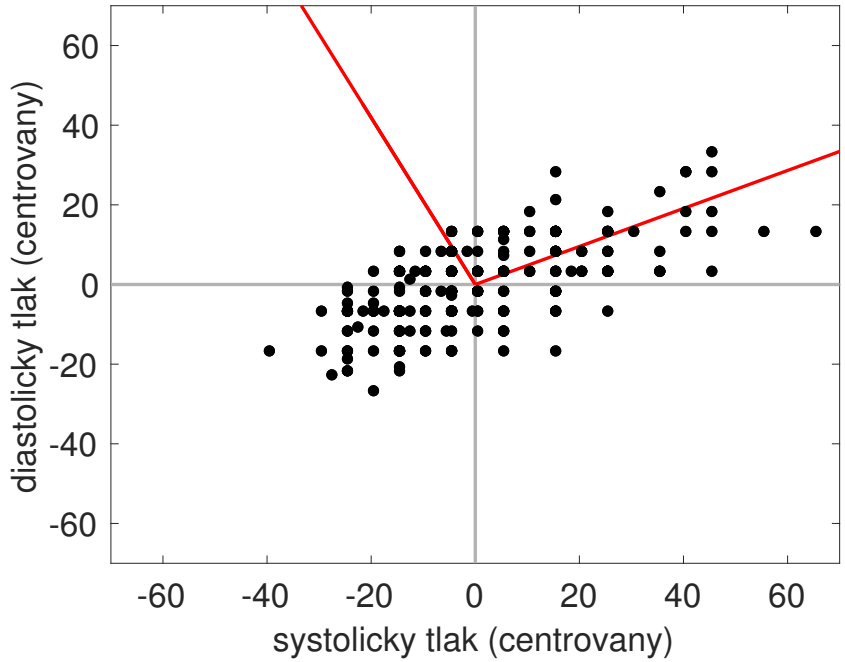
Redukcia dimenzie



→ premenné môžu byť nositeľmi "podobnej" informácie

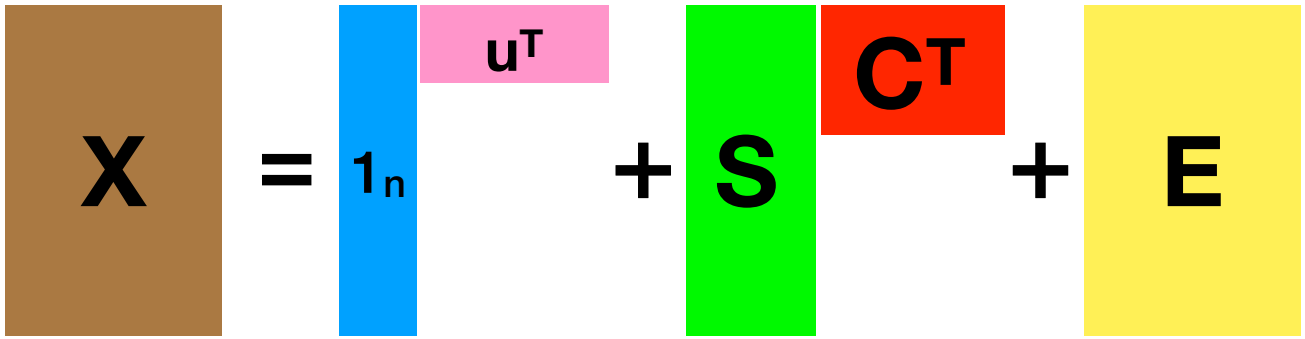
⇒ stačí jedna (nová) premenná, napr. ich vhodná lineárna kombinácia

Redukcia dimenzie



- **cieľ:** hľadanie **K** nových premenných
  - $K \leq p$
  - nové premenné dobre popisujú pôvodné dáta
  - nové premenné vieme interpretovať

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

Metódy redukcie dimenzie

$\rightarrow$ odlišnosti v predpokladoch vlastností a odhadoch matíc S, C

- **metóda hlavných komponentov** (principal component analysis)
  - klasická, jadrová (kernel), riedka (sparse), nelineárna, robustná ...
- **faktorová analýza** (exploratory factor analysis)
- **metóda nezávislých komponentov**
  (independent component analysis)
  - klasická, jadrová (kernel), ...
- **nezáporná maticová faktorizácia** (nonnegative matrix factorisation)
  - klasická, semi-ortogonálna, hladká, riedka, ...
- ...

Metóda hlavných komponentov

Principal component analysis (PCA)

PCA - Metóda hlavných komponentov

- **cieľ:** nájsť nové premenné, ktoré:
  - sú navzájom ortogonálne
  - sú lineárnou kombináciou pôvodných premenných
    $\Rightarrow$ dochádza len k transformácií údajov
  - "dobre" charakterizujú dáta
    $\Rightarrow$ vieme ich interpretovať

PCA - Metóda hlavných komponentov

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

$$x_{ij} = u_j + \sum_{l=1}^{p} s_{il} x_{lj} + e_{ij}$$

$\rightarrow$ $C = (\mathbf{c_1}, \ldots, \mathbf{c_p}) \in \mathbb{R}^{p \times p}$

- prepojenie medzi hl. komponentami (PC, stĺpce) a pôvodnými premennými (riadky)

$$C^T C = I_p$$

$\rightarrow$ $S = (\mathbf{s_1}, \ldots, \mathbf{s_p}) \in \mathbb{R}^{n \times p}$

- "nové" premenné = hlavné komponenty
- lin. kombinácia pôvodných (centrovaných) premenných

$$s_{il} = \sum_{j=1}^{p} c_{lj} (x_{ij} - u_j) \quad l = 1, \ldots, p$$

PCA - podiel vysvetlenej variancie

- $\mathbf{s_1}$ = **1. hlavný komponent (PC)**
  - vysvetľuje najväčší podiel variancie v dátach spomedzi všetkých normovaných lineárnych kombinácií pôvodných premenných

- $\mathbf{s_k}$ = **k–ty hlavný komponent (PC)**, k = 2,. . . ,p
  - vysvetľuje najväčšiu časť zostávajúcej variability v dátach, ktorá nebola vysvetlená pomocou $\mathbf{s_1}, \ldots, \mathbf{s_{k-1}}$, spomedzi všetkých normovaných lineárnych kombinácií pôvodných premenných, ktoré sú nekorelované s prvými $(k-1)$ hl. komponentami $\mathbf{s_1}, \ldots, \mathbf{s_{k-1}}$

$\rightarrow$ počet hlavných komponentov je p, ale posledné PC vysvetľujú len malý podiel variability v dátach

$\Rightarrow$ posledné PC môžeme zanedbať

$\Rightarrow$ pracujeme len s prvými $K << p$ PC

PCA - odhad hlavných komponentov I

1. odhad vektora $\mathbf{u}$
   - $\widehat{u}_j = \frac{1}{n}\sum_{l=1}^{n} x_{lj}$
   - $\widehat{\mathbf{u}} = mean(X)$
2. odhad kovariančnej matice
   - $\widehat{\Sigma}_{ij} = \frac{1}{n}\sum_{l=1}^{n}(x_{li} - \widehat{u}_i)(x_{lj} - \widehat{u}_j)$
   - $\widehat{\Sigma} = cov(X) \in \mathbb{R}^{p\times p}$

PCA - odhad hlavných komponentov II

3. stĺpce matice C = vlastné vektory matice $\widehat{\Sigma}$

$$\widehat{\Sigma}\widehat{C} = \widehat{C}\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_p\end{bmatrix}; \quad \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \lambda_{p-1} \geq \lambda_p$$

$\rightarrow$ vlastné hodnoty súvisia s podielom vysvetlenej variancie v dátach

$$\alpha_i = \frac{\lambda_i}{\sum_{k=1}^{p} \lambda_k} \text{ je podiel variance vysvetlenej i-tym PC}$$

$\rightarrow$ $[C, D] = eig(\widehat{\Sigma})$ $\qquad\qquad$ $\lambda = diag(D)$

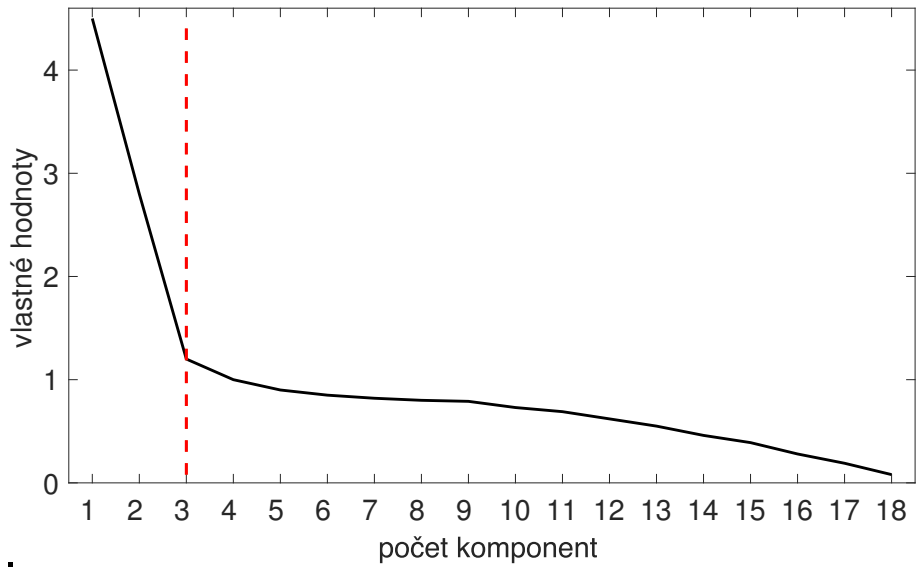
4. $\widehat{S} = (X - 1_n \widehat{\mathbf{u}}^T)\widehat{C}$

PCA - určenie počtu hlavných komponentov K

- podiel vysvetlenej variancie $\geq$ zvolený prah ρ, napr. $\rho = 0.95$

$$K \in \text{argmin}_{k \in \{1,\ldots,p\}} \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j} \geq \rho$$

- lakťový graf (elbow diagram)



- Kaiserovo pravidlo

$$K \in \text{argmax}_{k \in \{1,\ldots,p\}} \lambda_k > \frac{1}{p} \sum_{i=1}^{p} \lambda_i$$

PCA v MATLAB-e

$[\mathbf{C}, \mathbf{S}, \lambda, \sim, \mathbf{expl}, u] = \boldsymbol{pca}(\mathbf{X})$

X = matica pozorovaní $n \times p$, MATLAB ju automaticky centruje

C = matica $p \times p$

- i-ty stĺpec reprezentuje i-ty hlavný komponent
- j-ty riadok reprezentuje príspevok j-tej pôvodnej premennej k jednotlivým hlavným komponentom

S = $n \times p$ matica hodnôt hl. komponentov pre n pozorovaní

λ = vlastné hodnoty patriace k jednotlivým hl. komponentom

expl = $(\alpha_1, \ldots, \alpha_p)^T \times 100$, podiel vysvetlenej variancie jednotlivými hl. komponentami (v %)

$u \in \mathbb{R}^p$ - priemer po stĺpcoch matice X

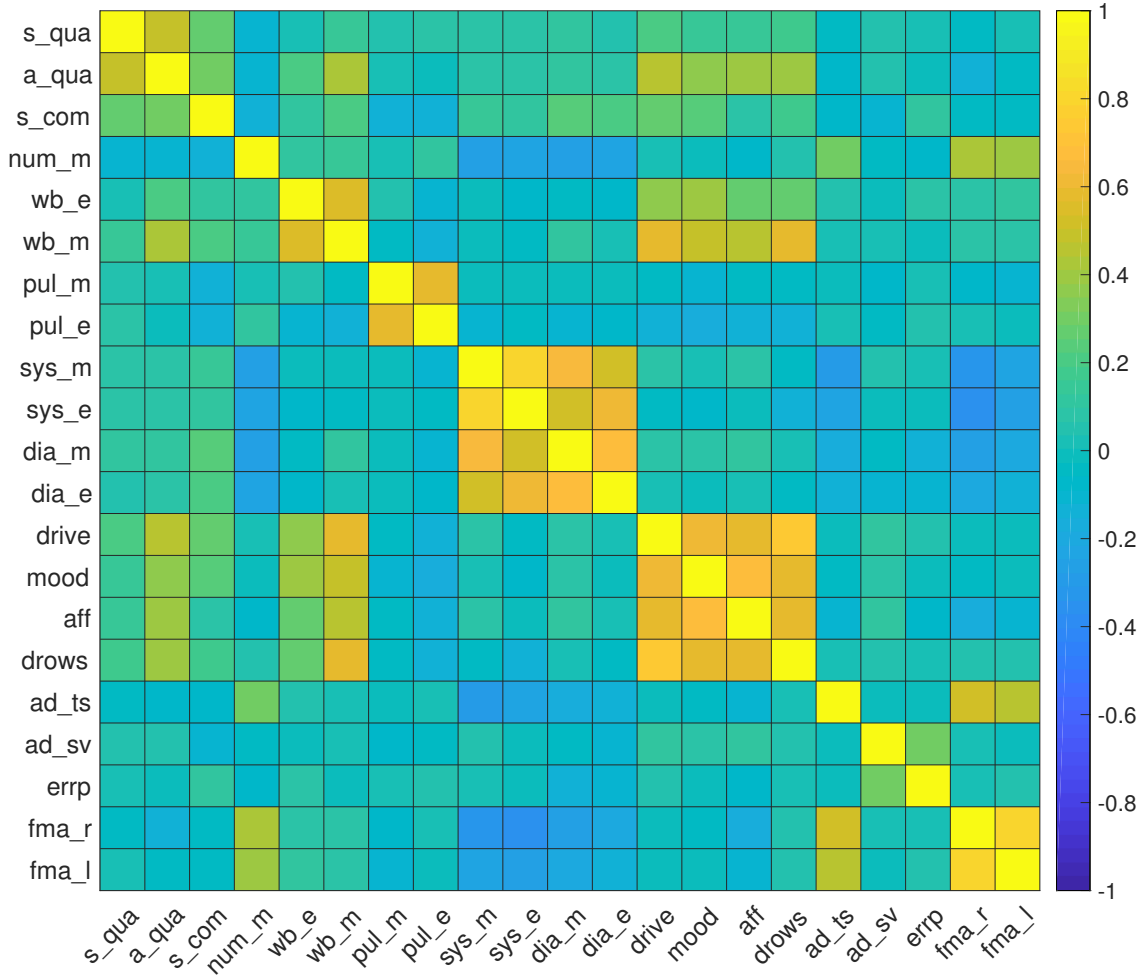
PCA - Príklad: Denné miery

- [Rosipal et al., 2013]
- 148 ľudí, 2 noci v spánkovom laboratóriu $\rightarrow$ 296 pozorovaní
- dotazníky a testy ohľadom ich subjektívneho a objektívneho stavu (ráno, večer)

$\rightarrow$ **21 premenných**

- Self-rating questionnaire for sleep quality, awakening quality and somatic complaints
- Numerical memory test
- Well-being self-assessment scale evening/morning
- Pulse rate evening/morning
- Systolic and diastolic blood pressure evening/morning
- Visual analog scale test for drive, mood, affectivity and drowsiness
- Alphabetical cross-out test - total score, variability, % or errors
- Fine-motor activity test for right/left hand

PCA - Príklad: Denné miery

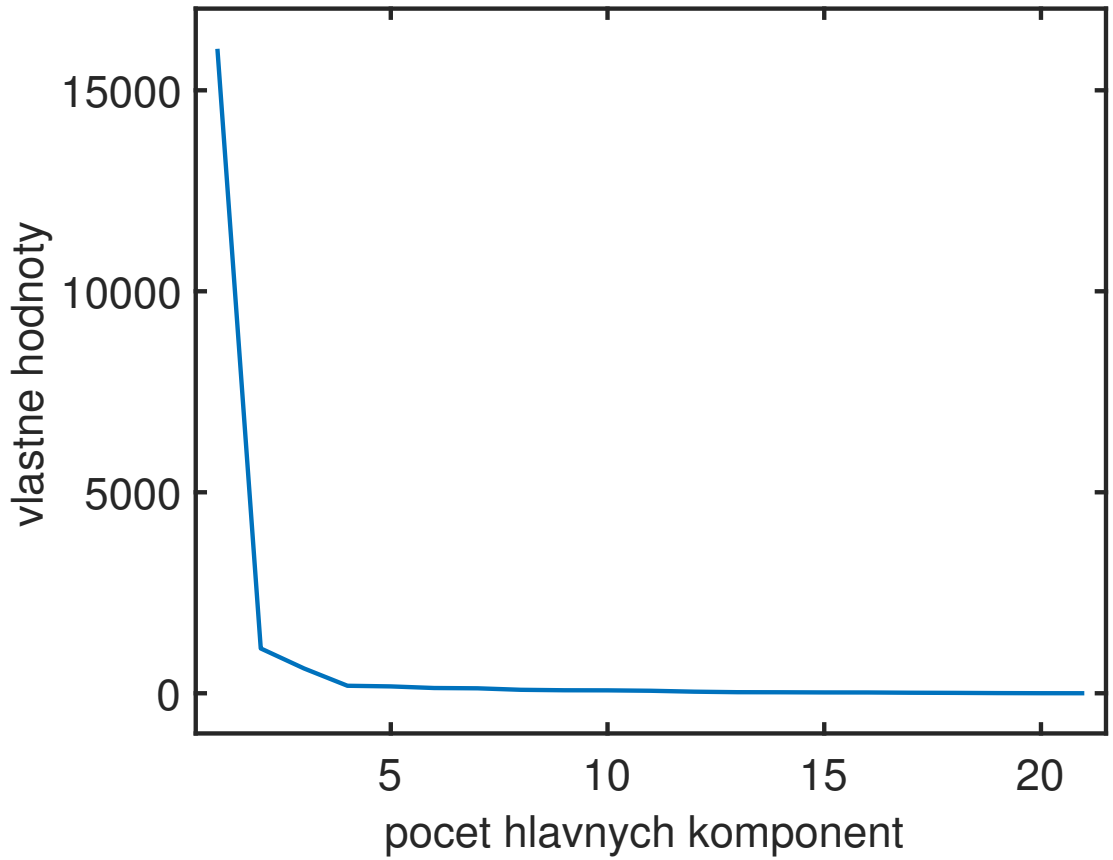


→ viaceré premenné sú korelované → nositelia podobnej informácie

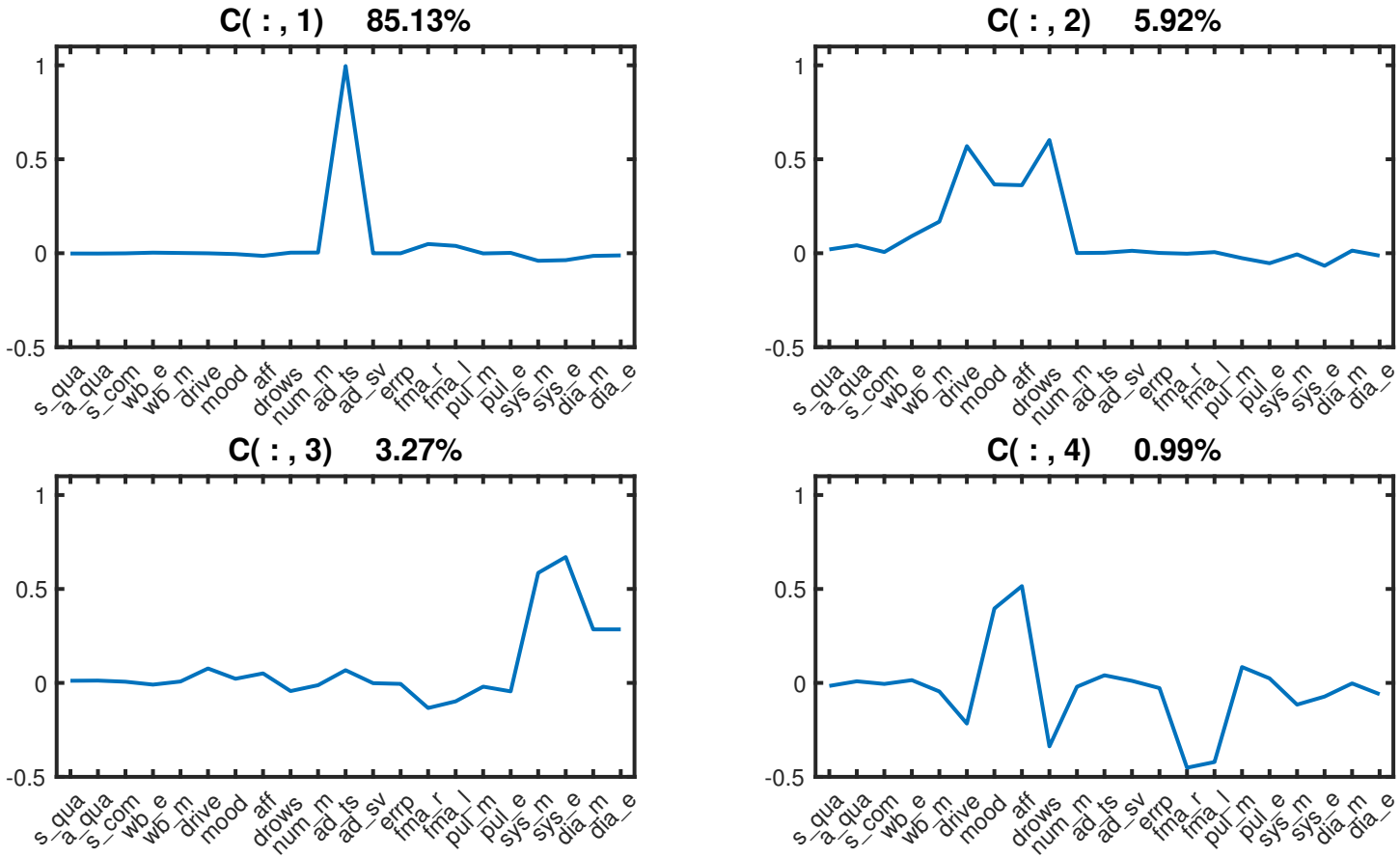
→ $R = corr(data)$
$heatmap(nazvy\_premennych, nazvy\_premennych, R)$

PCA - Príklad: Denné miery

- **pôvodné dáta**
- prvé 4 PC vysvetľujú viac ako 95% variability v dátach
- prvé 2 PC majú príslušné vlastné hodnoty vyššie ako priemer (Kaiserovo pravidlo)

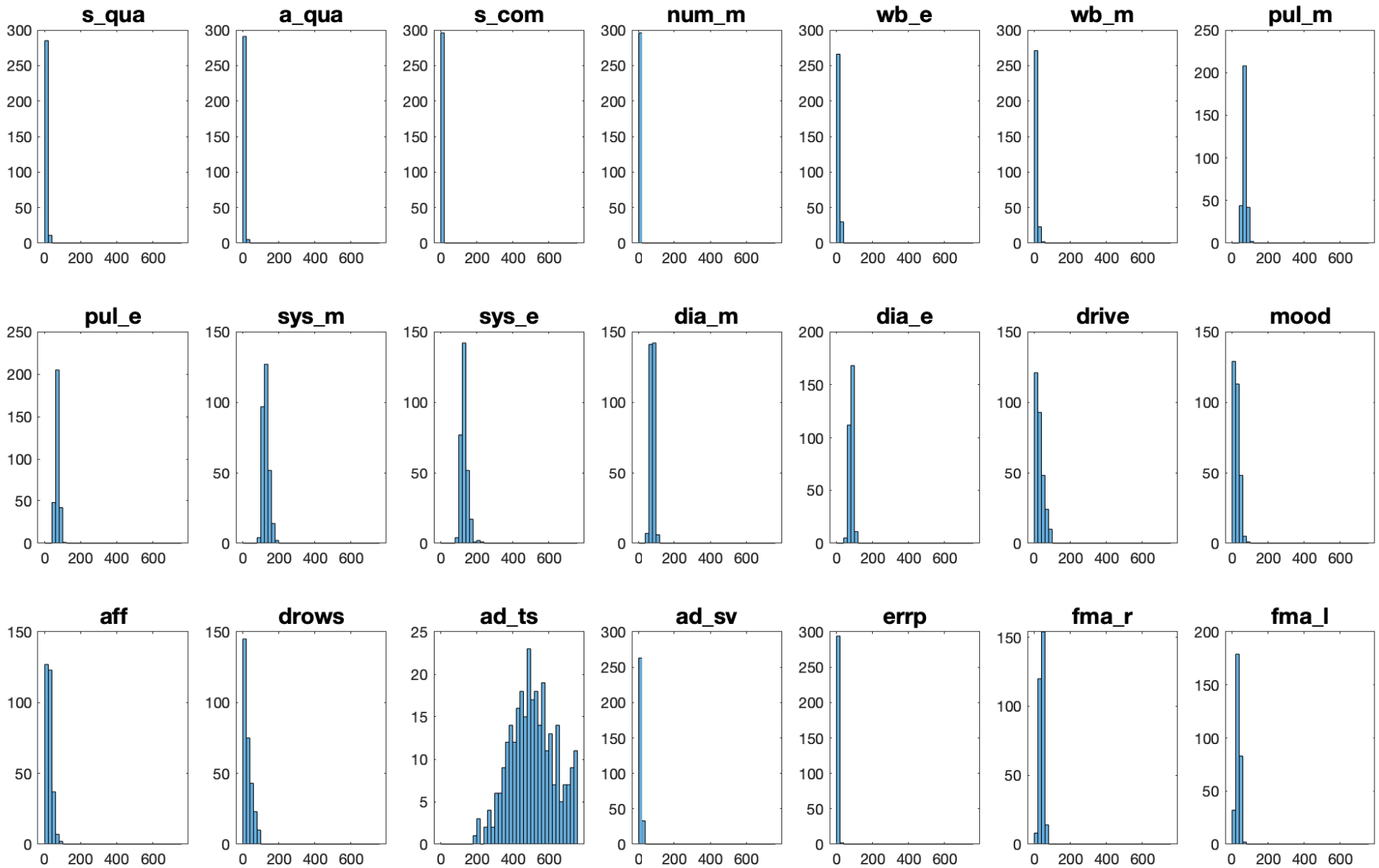


PCA - Príklad: Denné miery



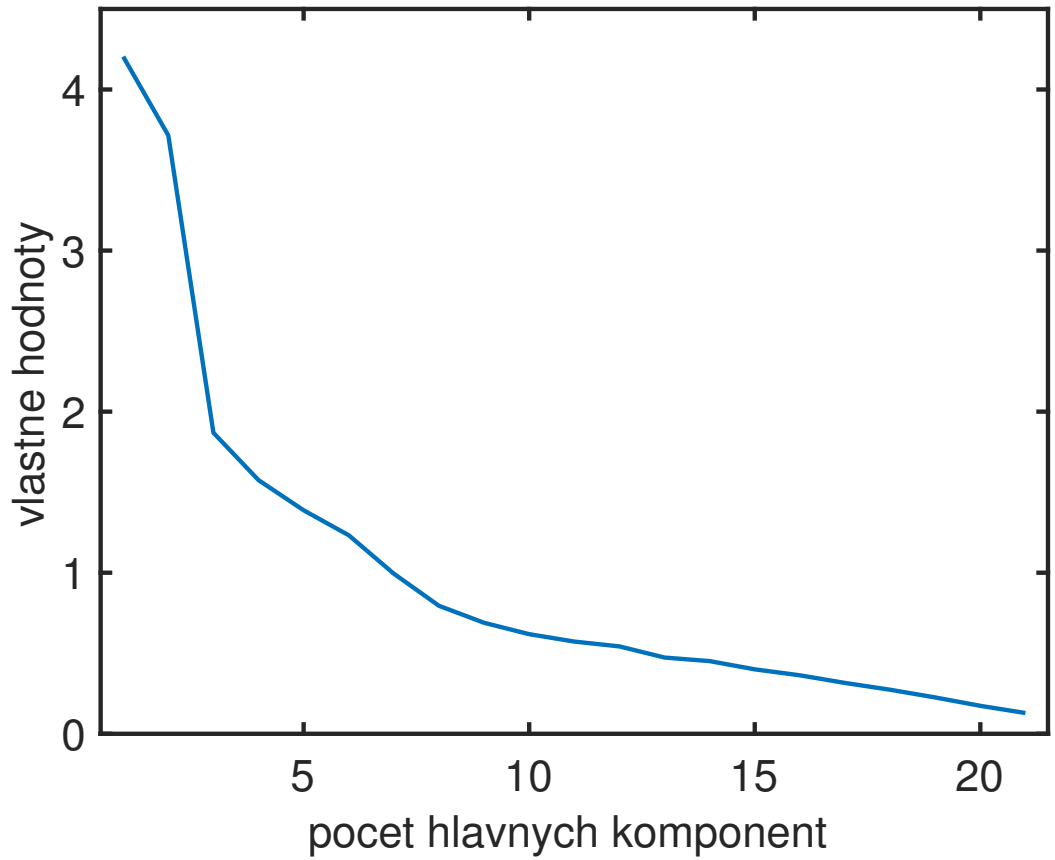
- C(:,1) - Abecedný pamäťový test $\rightarrow$ **výrazne vyššie hodnoty**
- C(:,2) - Visual analog scale for drive, mood, affectivity and drowsiness
- C(:,3) - krvný tlak

PCA - Príklad: Denné miery

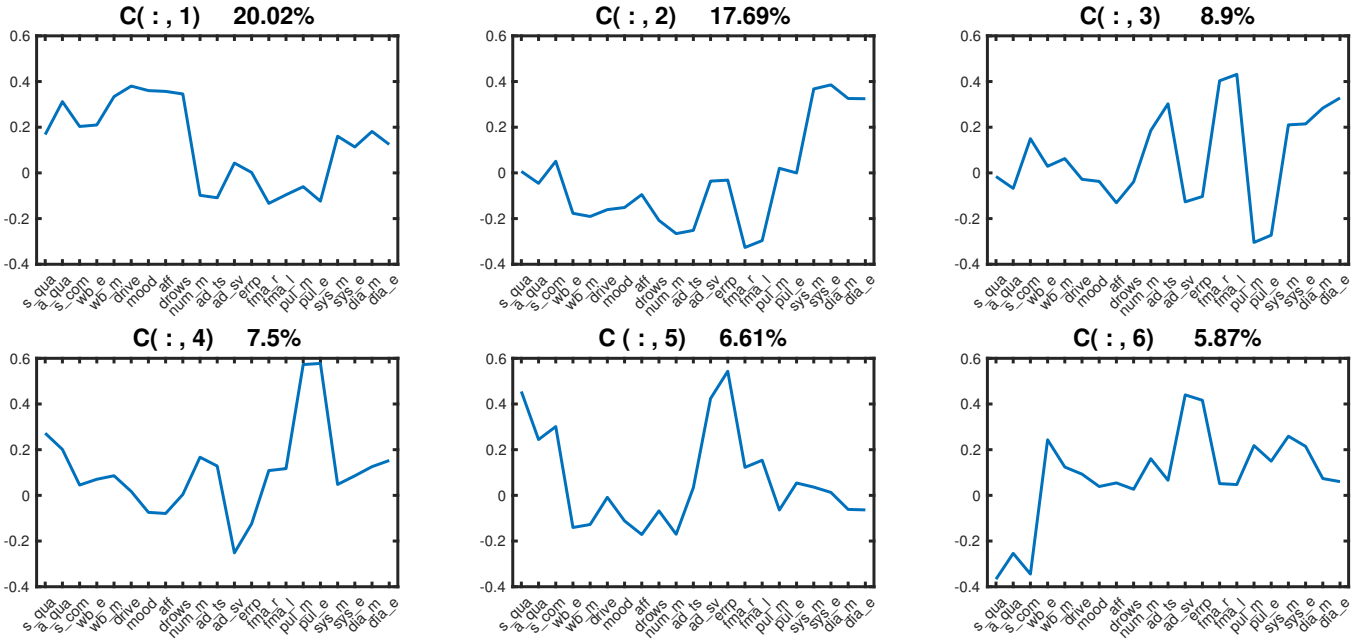


PCA - Príklad: Denné miery

- **štandardizované dáta** $\rightarrow$ $Xnew = zscore(X)$
- prvých 17 PC vysvetľuje viac ako 95% variability v dátach
- prvých 6 PC majú príslušné vlastné hodnoty vyššie ako priemer (Kaiserovo pravidlo)



PCA - Príklad: Denné miery



- $C(:, 1)$ - subjektívny stav organizmu po prebudení
- $C(:, 2)$ - fyziologický stav organizmu (krvný tlak)
- $C(:, 3)$ - kognitívno-fyziologický stav organizmu (num. test, test jemnej motoriky, krvný tlak)
- $C(:, 4)$ - pulz ráno a večer

PCA zhrnutie

- **výhody**
  - dochádza len k transformácií údajov, nie sú potrebné žiadne špeciálne predpoklady
  - vhodné na vizualizáciu dát, resp. lepšie pochopenie charakteru dát
  - odhadnuté PC možno použiť ako vstup v ďalších analýzach
- **nevýhody**
  - nie je škálovo invariantná $\rightarrow$ **preškálovanie vedie k inému riešeniu**
  - ak niektorá premenná vykazuje výrazne vyššie/nižšie hodnoty ako ostatné premenné
    - hl. komponenty budú v prvom rade zahŕňať informáciu z tejto premennej

    $\Rightarrow$ niekedy je lepšie pracovať so **štandardizovanými** premennými, t.j. zero-mean, unit-variance

Metóda hl. komponentov pre časové rady a funkcionálne dáta

Functional principal component analysis

PCA pre funkcionálne dáta a časové rady (FPCA)

- literatúra:
  - **Ramsay, J. O. and Silverman, B. W. (2005).** ***Functional data analysis.*** Springer Series in Statistics, New York, second edition.
  - **Yao, F. et al. (2003).** ***Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate.*** Biometrics, 59(3):676-685.
- n pozorovaní (časových radov) $x_1, \ldots, x_n$ na intervale T

$$x_i(t) = \mu(t) + \sum_{k=1}^{K} \alpha_{ik}\varphi_k(t) + \varepsilon_i(t), \quad t \in T,$$

$$\int_T \varphi_k^2(t) = 1, \quad k = 1, \ldots, K$$

$$\int_T \varphi_k(t)\varphi_l(t) = 0, \quad k \neq l$$

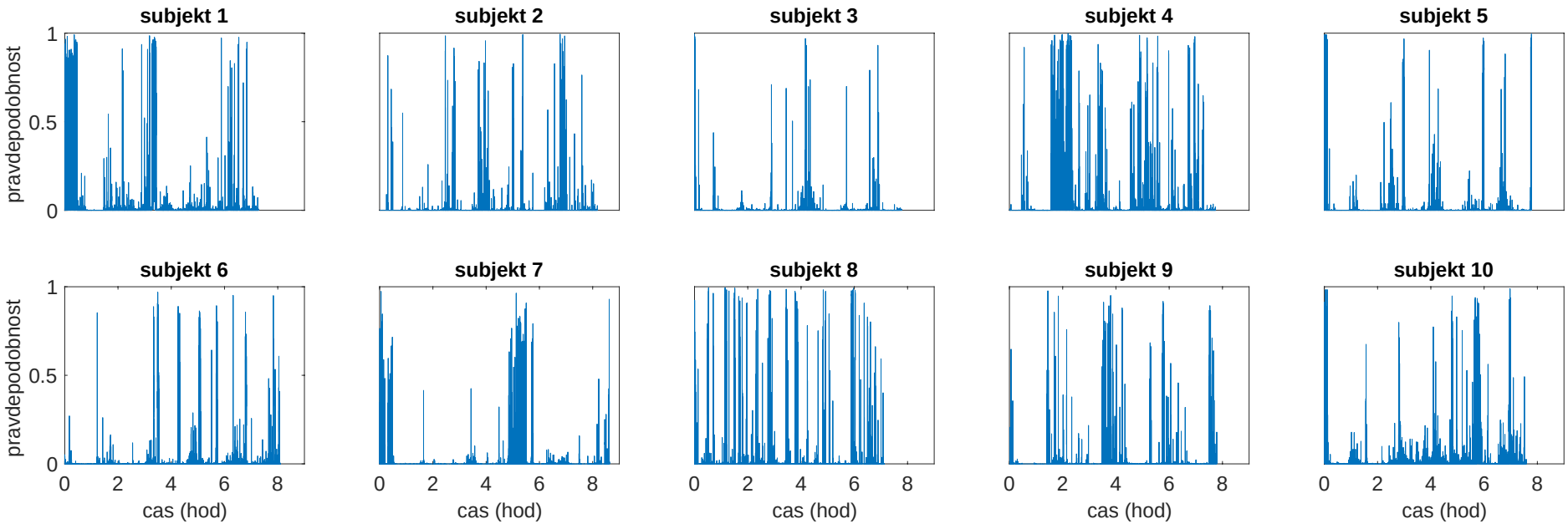
FPCA v MATLAB-e

- nie je priamo implementovaná v MATLAB-e
- balík PACE - Principal Analysis by Conditional Estimation
  - https://anson.ucdavis.edu/ mueller/data/pace.html
  - Yao, F., Müller, H.G., Clifford, A.J., Dueker, S.R., Follett, J., Lin, Y., Buchholz, B., Vogel, J.S. (2003). Shrinkage estimation for functional principal component scores, with application to the population kinetics of plasma folate. Biometrics 59, 676-685.
  - funkcia *FPCA*

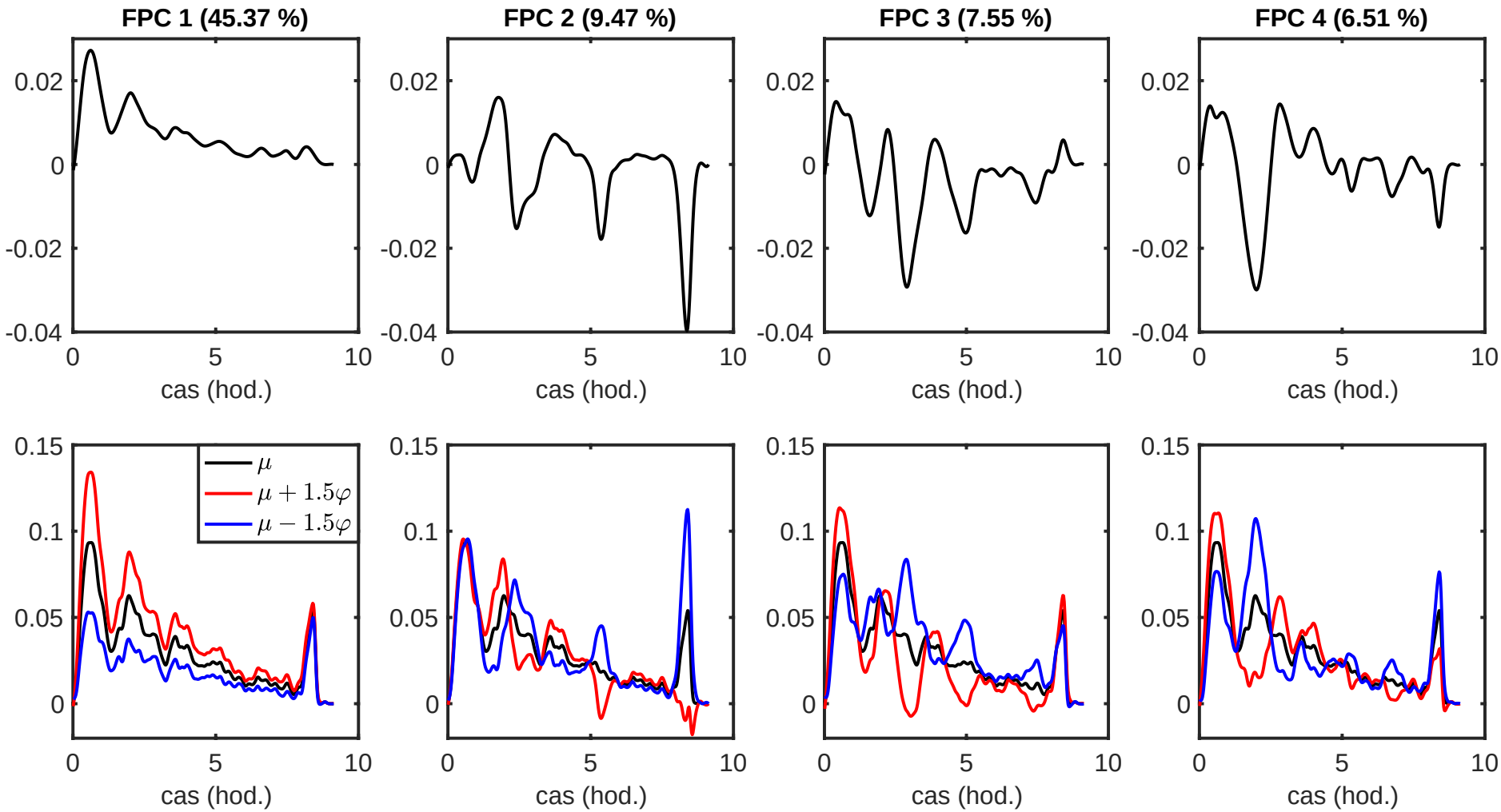
FPCA - Príklad: Spánkové krivky

- 146 ľudí
- 2 noci v spánkovom laboratóriu

$\rightarrow$ 292 kriviek reprezentujúcich stav bdelosti počas noci



FPCA - Príklad: Spánkové krivky



Faktorová analýza

Exploratory factor analysis (EFA)

EFA - Faktorová analýza

- **Spearman, C. (1904).** ***"General intelligence," objectively determined and measured.*** The American Journal of Psychology, 15(2):201-292.
- **cieľ:** nájsť **K** nových priamo nepozorovateľných (tzv. skrytých, latentných) premenných, ktoré
  - sú navzájom nekorelované
  - popisujú dáta rovnako "kvalitne" ako pôvodné premenné
  - vieme interpretovať
- počet nových premenných K je vstupný parameter
- skryté (latentné) premenné nazývame aj faktory

EFA - Faktorová analýza

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

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

→ $S \in \mathbb{R}^{n \times K}$

→ matica skrytých faktorov, ktoré sú navzájom nekorelované

→ $C \in \mathbb{R}^{p \times K}$

→ matica faktorových nákladov (váh)

→ prepojenie medzi pôvodnými premennými a skrytými faktormi

→ C, S sa odhadujú inak ako v PCA

- C - metóda hlavných faktorov; metóda maximálnej vierohodnosti; ...
- S - metóda regresnej analýzy, ...

EFA - odhad počtu faktorov

- apriórna vedomosť
- pokus-omyl s dôrazom na interpretáciu
- formálny postup
  1. korelačná matica dát $R = corr(X)$
  2. vypočítame tzv. redukovanú korelačnú maticu $R^\star$
     - r je diagonála inverznej korelačnej matice R^{-1}
     - hlavnú diagonálu v R nahradíme $1 - \frac{1}{r}$
     - v MATLAB-e:

$$R^\star = R - eye(p) + diag(1 - 1/diag(inv(R)))$$

  3. vypočítame vlastné hodnoty matice $R^\star$
     - v MATLAB-e: $e = eig(R^\star)$
     - elbow diagram - hľadáme výrazný zlom v grafe
     - Kaiserovo pravidlo - $K = \#\{e_i : e_i > 1\}$

$\rightarrow$ len ako odporúčané K, dôraz sa kladie na interpretáciu

EFA - rotácia faktorov

- **použitie:** pôvodné faktory sú ťažko interpretovateľné
- nech $\mathbf{A}$ je rotačná matica a $\mathbf{A}^{-1}$ reprezentuje inverznú rotáciu

$$X - 1_n \mathbf{u}^T = SC^T + E$$
$$X - 1_n \mathbf{u}^T = SA^{-1}AC^T + E$$
$$X - 1_n \mathbf{u}^T = (SA^{-1})(AC^T) + E$$
$$X - 1_n \mathbf{u}^T = S^\star C^{\star T} + E$$

$\rightarrow$ model s rovnakou chybou, ale **INOU** interpretáciou faktorov

$\rightarrow$ môže zlepšiť interpretáciu faktorov

- najčastejšie používané rotácie:
  - *varimax* - maximalizuje varianciu štvorcov faktorových nákladov
  - *orthomax, quartimax, equamax, parsimax*
  - "vlastná"
  - ...

EFA v MATLAB-e

[C, ∼, T, stats, S]=*factoran*(X,K,Name,Value)

X = $n \times p$ matica pozorovaní alebo $p \times p$ kovariančná matica pozorovaní

K = počet skrytých (latentných) faktorov

- voliteľné parametre
  - Xtype = 'data', 'covariance'
  - Rotate = bez rotácie ('none'); s rotáciou ('varimax','orthomax',...)
  - Scores = metóda na odhad S ('wls', 'Bartlett', 'regression',...)

C = $p \times K$ matica faktorových nákladov (váh)

T = rotačná matica

stats = test hypotézy, že v dátach máme K skrytých faktorov
- stats.p = p-hodnota (stats.p $< 0.05 \rightarrow$ hypotézu zamietame)

S = $n \times K$ matica skrytých faktorov

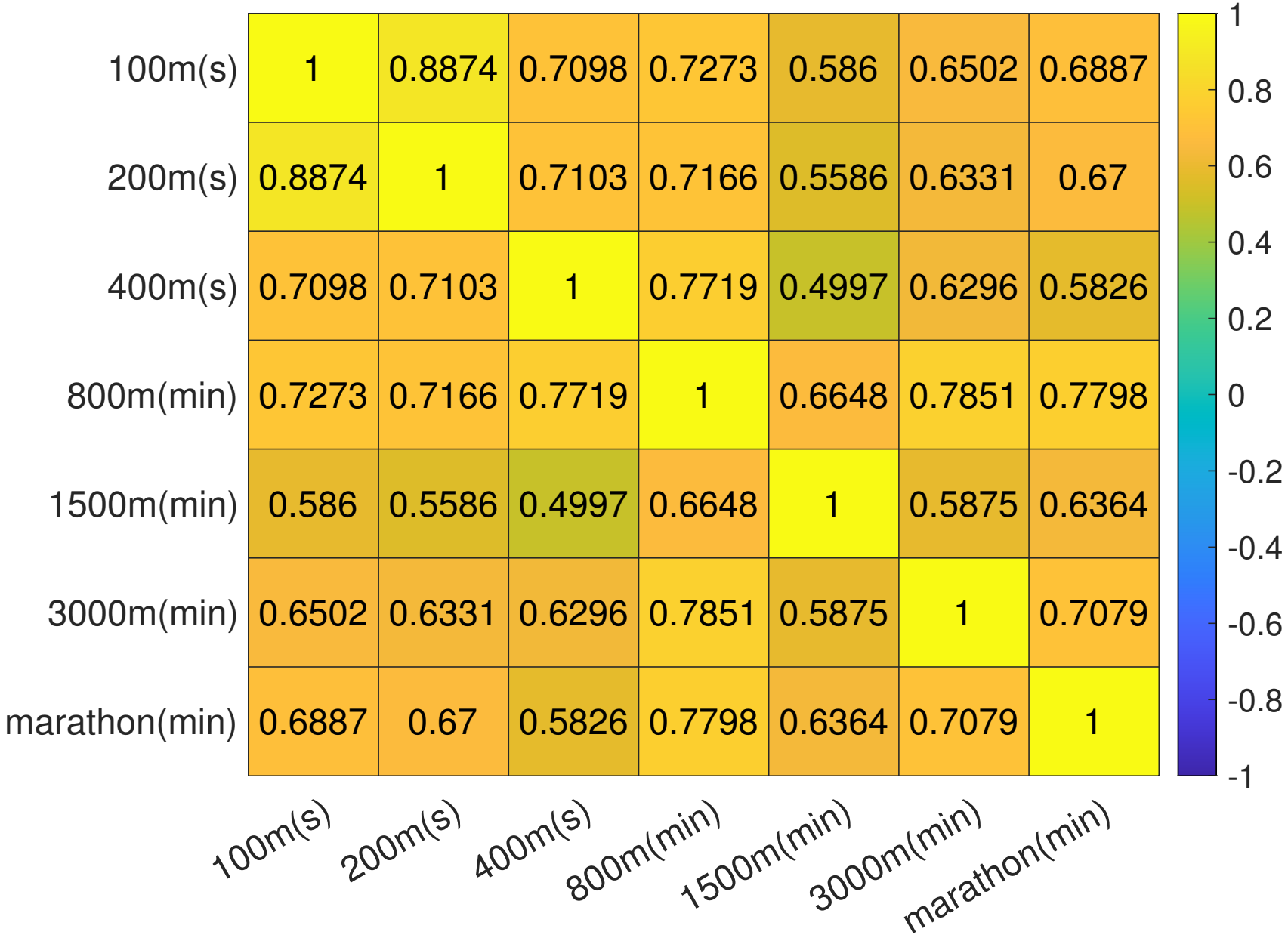
- podiel vysvetlenej variancie: $\alpha = diag(C^T C)/p$

EFA - Príklad: Výsledky bežeckých disciplín u žien

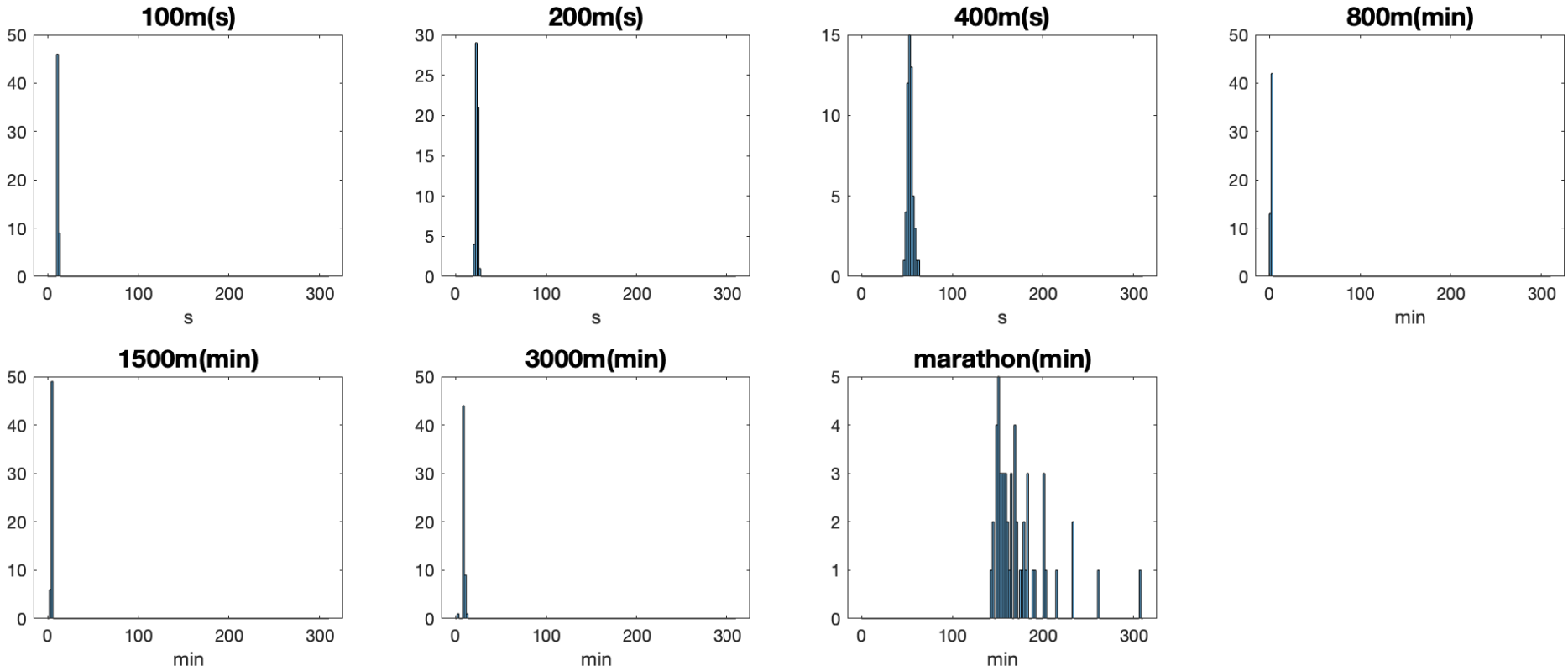
- najlepšie výsledky reprezentantiek 55 krajín v 7 disciplínach
  - 100 m (v s)
  - 200 m (v s)
  - 400 m (v s)
  - 800 m (v min)
  - 1500 m (v min)
  - 3000 m (v min)
  - maratón (v min)

$\rightarrow$ chceme nájsť skryté faktory súhrnne charakterizujúce tieto disciplíny

EFA - Príklad: Výsledky bežeckých disciplín u žien



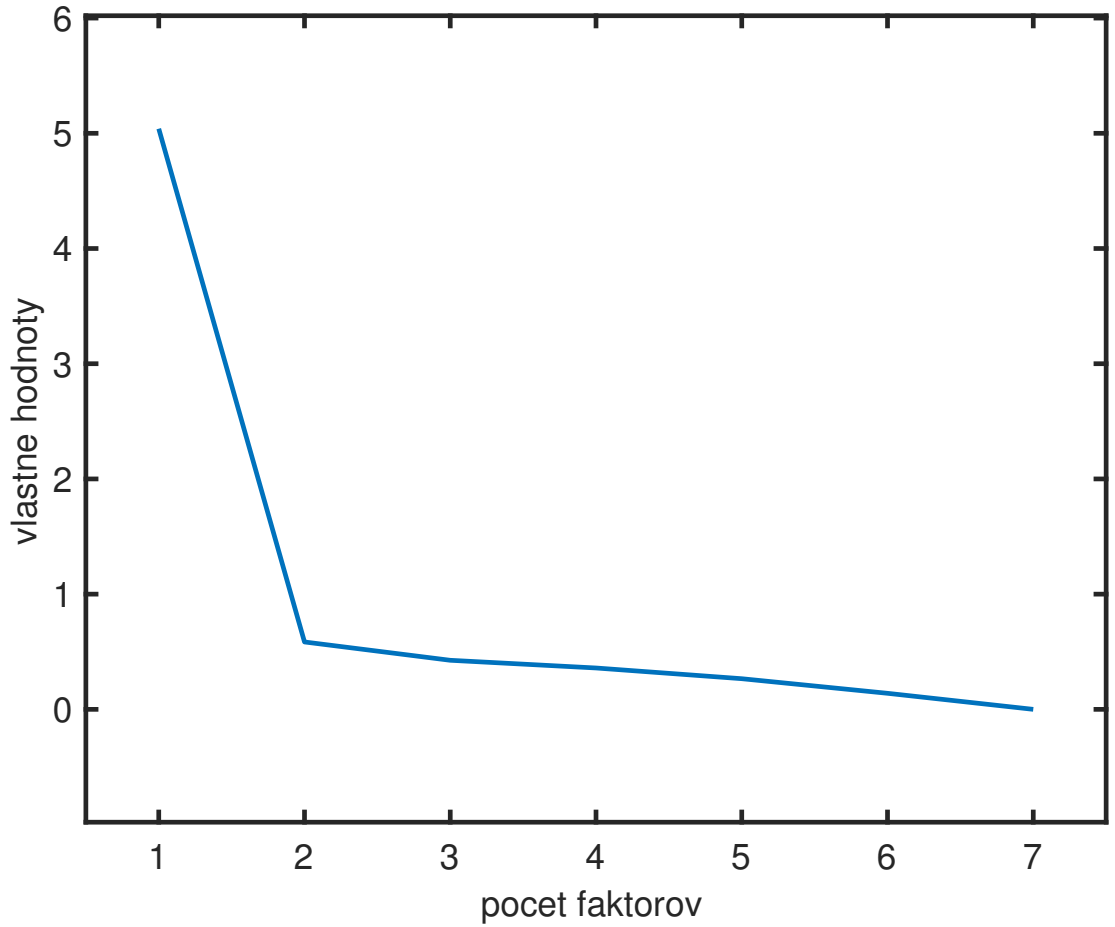
EFA - Príklad: Výsledky bežeckých disciplín u žien



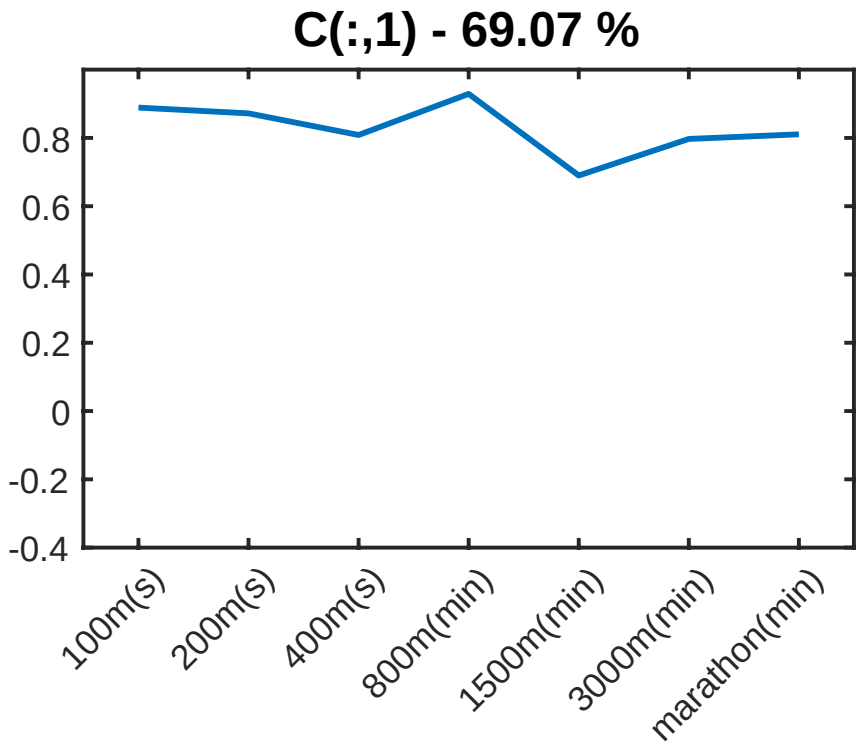
→ potrebná štandardizácia dát $X^{\star} = zscore(X)$

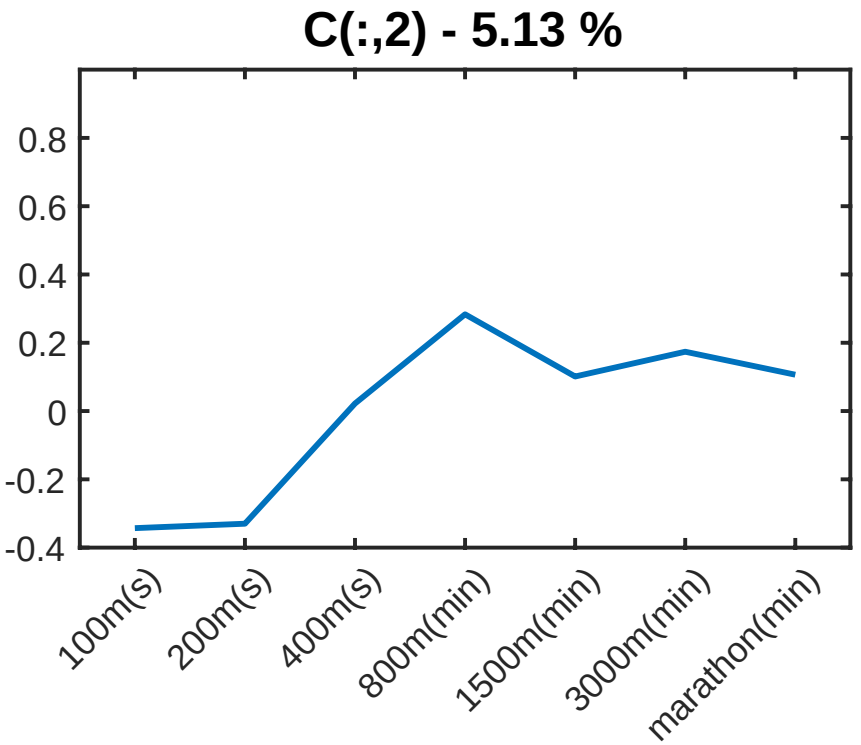
EFA - Príklad: Výsledky bežeckých disciplín u žien

- stanovenie počtu faktorov
  - Kaiserovo pravidlo: 1 faktor
  - elbow diagram: 1-2 faktory



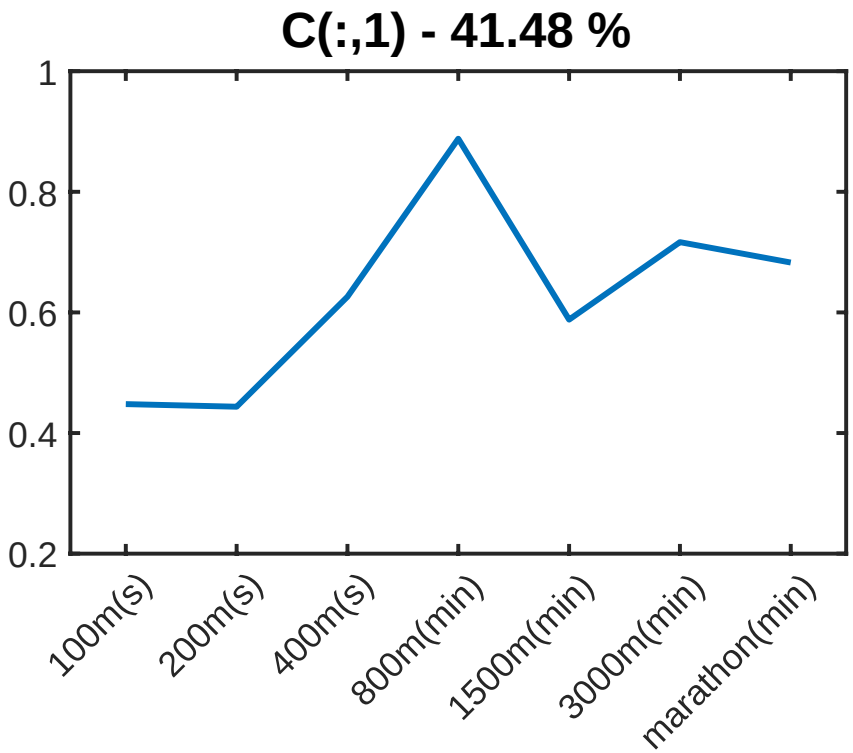
EFA - Príklad: Výsledky bežeckých disciplín u žien

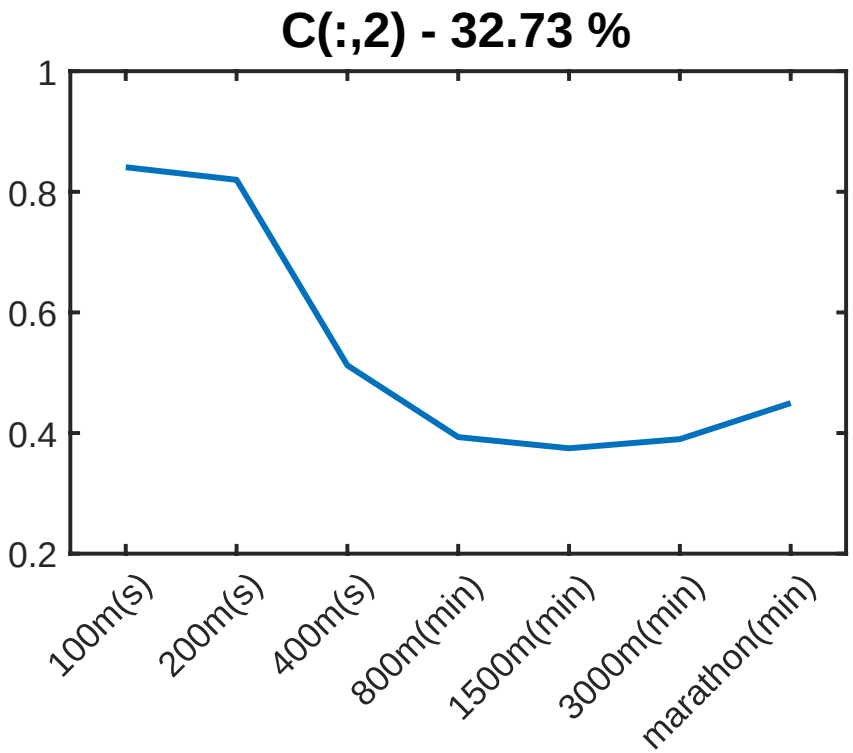




- model s 2 faktormi, bez rotácie faktorov
  - C(:,1) - (+) pre všetky disciplíny (rovnomerné váhy)
    $\rightarrow$ faktor by mal reprezentovať len malú skupinu premenných
  - C(:,2) - (+) pre vytrvalostné disciplíny, (-) pre rýchlostné disciplíny

EFA - Príklad: Výsledky bežeckých disciplín u žien





- model s 2 faktormi, *varimax rotácia faktorov*
  - C(:,1) - vyššie (+) hodnoty pre vytrvalostné disciplíny
  - C(:,2) - vyššie (+) hodnoty pre rýchlostné disciplíny

EFA - Príklad: Denné miery

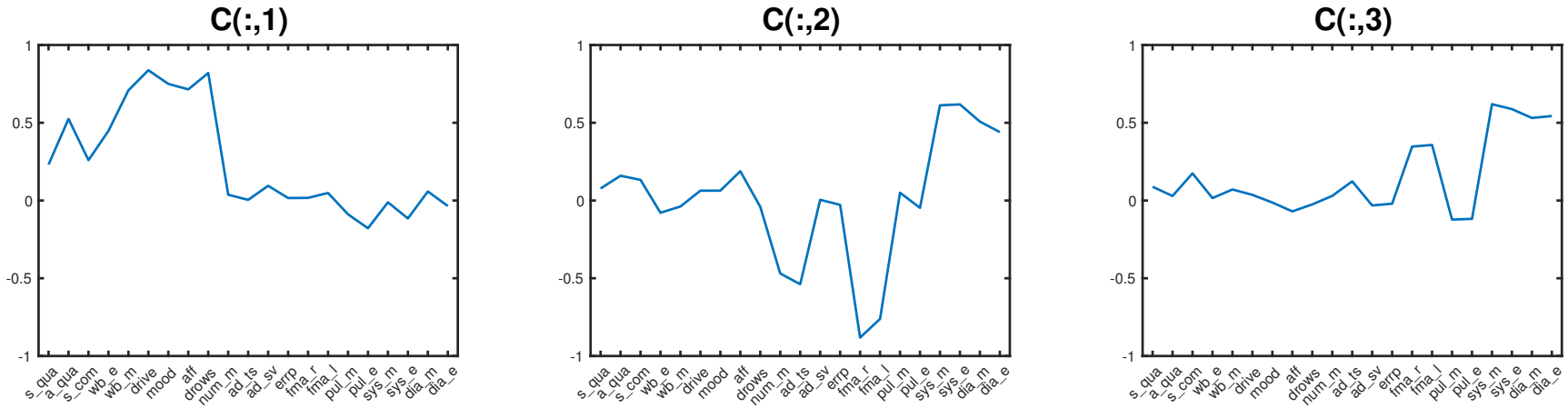
- [Rosipal et al., 2013]
- 148 ľudí, 2 noci v spánkovom laboratóriu $\rightarrow$ 296 pozorovaní
- dotazníky a testy ohľadom ich subjektívneho a objektívneho stavu (ráno, večer)

$\rightarrow$ **21 premenných**

- Self-rating questionnaire for sleep quality, awakening quality and somatic complaints
- Numerical memory test
- Well-being self-assessment scale evening/morning
- Pulse rate evening/morning
- Systolic and diastolic blood pressure evening/morning
- Visual analog scale test for drive, mood, affectivity and drowsiness
- Alphabetical cross-out test - total score, variability, % or errors
- Fine-motor activity test for right/left hand

EFA - Príklad: Denné miery

- štandardizácia dát $Xnew = zscore(X)$
- EFA s 3 faktormi, bez rotácie faktorov

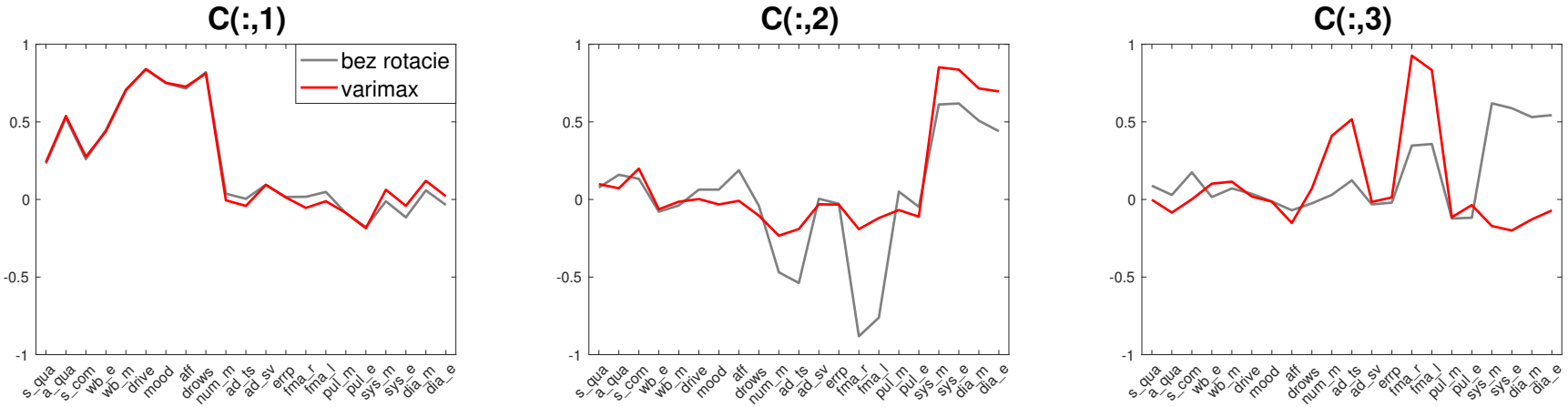


- $C(:, 1)$ (17,23%) - faktor subjektívneho hodnotenia kvality spánku
- $C(:, 2)$ (15,15%) - krvný tlak (+) + "niektoré" kognitívne testy (-)
- $C(:, 3)$ (7,87%) - krvný tlak (+) + "zvyšné" kognitívne testy (-)

→ **problém:** pôvodné premenné by mali mať "vysoké" (+/-) váhy len pre jeden z latentných faktorov

EFA - Príklad: Denné miery

- štandardizácia dát $Xnew = zscore(X)$
- výsledky v [Rosipal et al., 2013]
- EFA s 3 faktormi, rotácia faktorov pomocou metódy ***varimax***



- $C(:,1)$ (17,36%) - faktor subjektívneho hodnotenia kvality spánku
- $C(:,2)$ (12,63%) - faktor fyziologického stavu organizmu
- $C(:,3)$ (10,25%) - faktor kognitívneho stavu organizmu po prebudení

EFA a PCA

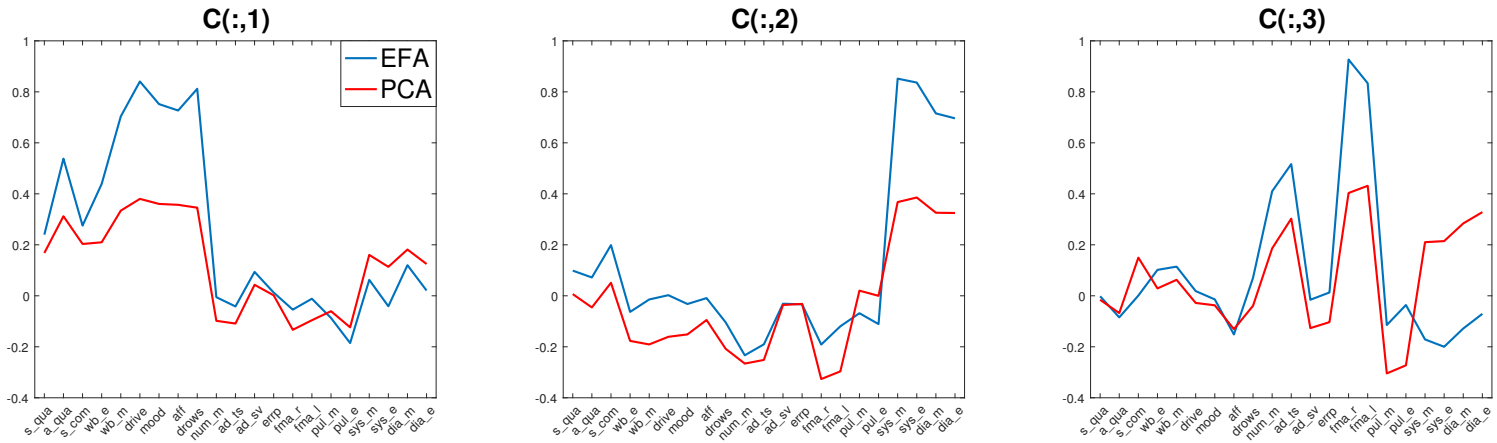
EFA

- predpoklad existencie latentných premenných
- K je vstupný parameter
- odhady - metóda hlavných faktorov, metóda maximálnej vierohodnosti

PCA

- nové premenné = lin. kombináciu pôvodných
- K sa určí až po analýze
- odhady - eigenanalysis, SVD,...

$\rightarrow$ výstup PCA = inicializačné odhady faktorov v EFA



Literatúra I

Rosipal, R., Lewandowski, A., and Dorffner, G. (2013).
In search of objective components for sleep quality indexing in normal sleep.
Biological Psychology, 94(1):210–220.