PA081: Programování numerických výpočtů

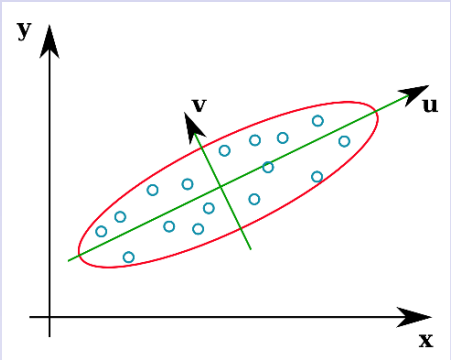
8. Analýza hlavních komponent

Aleš Křenek a Jan Fousek

jaro 2013

Úvodní příklad

- do zemské atmosféry proletěl velký meteorit
- ještě před dopadem explodoval a rozpadl se na mnoho úlomků
- na zemi najdeme jen část z nich



- je třeba určit
  - jak velký byl průměr shluku úlomků před dopadem
  - pod jakým úhlem meteorit dopadl
  - jakým směrem letěl

Obecná motivace

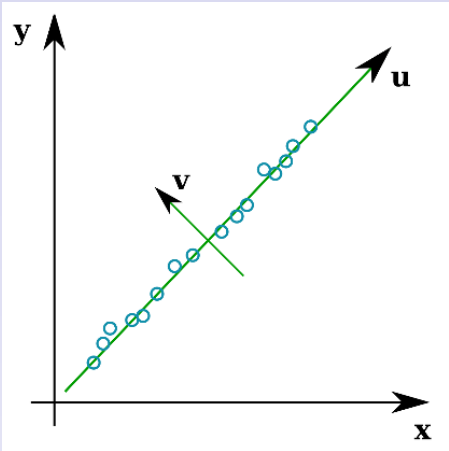
- soubor opakovaného měření nějakého jevu apod.
  - pro každý prvek souboru více měřených hodnot – n rozměrný prostor
  - hledáme nezávislé (kolmé) směry zodpovědné za proměnlivost hodnot

Obecná motivace

- soubor opakovaného měření nějakého jevu apod.
  - pro každý prvek souboru více měřených hodnot – n rozměrný prostor
  - hledáme nezávislé (kolmé) směry zodpovědné za proměnlivost hodnot
- v případě dopadu meteoritu
  - zaznamenáme souřadnice nálezu úlomků
  - směr největšího rozptylu určuje směr dopadu
  - rozptyl nejmenší ze všech směrů určuje poloměr shluku
  - poměr rozptylu ve směru dopadu a v kolmém směru odpovídá úhlu dopadu

Analýza hlavních komponent

- Principal Component Analysis (PCA)
- hledá nezávislé hlavní směry zodpovědné za proměnlivost dat a určuje jejich relativní význam
- některé směry nemají žádný nebo minimální význam
  - více veličin než prvků souboru (např. pixely obrázku)
  - významné závislosti v celém jevu
  - PCA je dokáže identifikovat – redukce počtu dimenzí



Trocha statistiky

- náhodná veličina, soubor vzorků, ...
  - viz základní kursy statistiky
- střední (očekávaná) hodnota náhodné veličiny X

$$EX = \bar{\mathbf{x}} = \mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i$$

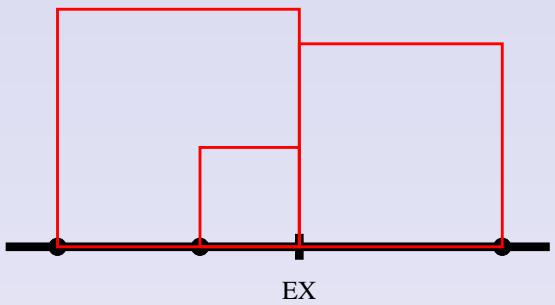
Trocha statistiky

- náhodná veličina, soubor vzorků, ...
  - viz základní kursy statistiky
- střední (očekávaná) hodnota náhodné veličiny X

$$EX = \bar{\mathbf{x}} = \mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- rozptyl (variance)

$$var(X) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - EX)^2$$



  - vystihuje, jak moc jsou prvky souboru vzdáleny od střední hodnoty

Trocha statistiky

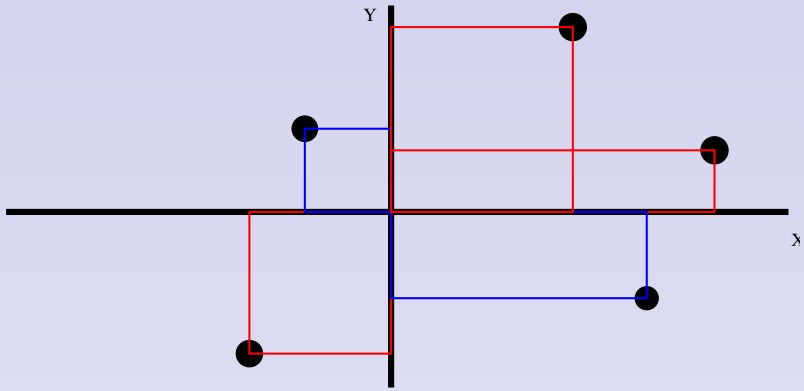
- kovariance
  - zobecnění rozptylu na vzájemný vztah dvou veličin

$$cov(X,Y) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - EX)(y_i - EY)$$

  - malý rozptyl v jedné veličině (rozložení vzorku podél druhé osy) kovarianci zmenšuje bez ohledu na chování druhé veličiny
  - současný výskyt (x, y) a $(x, -y)$ se ruší
  - nejvyšší absolutní hodnota při lineární závislosti $X = \pm Y$
  - triviálně $cov(X, X) = var(X)$ a $cov(X, Y) = cov(Y, X)$

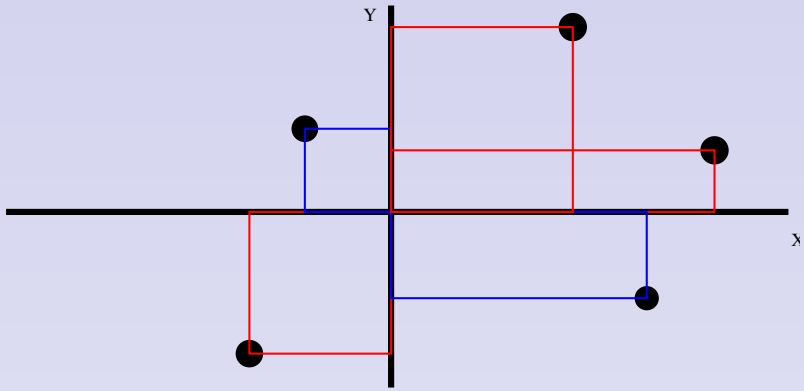
Trocha statistiky

- jednoduchý příklad graficky



Trocha statistiky

- jednoduchý příklad graficky



- matice kovariance veličin $X_1, \ldots, X_n$

$$\Sigma = \begin{pmatrix} var(X_1) & cov(X_1, X_2) & \ldots & cov(X_1, X_n) \\ cov(X_2, X_1) & var(X_2) & \ldots & cov(X_2, X_n) \\ & & \ddots & \\ cov(X_n, X_1) & cov(X_n, X_2) & \ldots & var(X_n) \end{pmatrix}$$

PCA formálněji

- prvky analyzovaného souboru chápeme jako vektory v n-rozměrném prostoru
- hledáme souřadný systém, jehož bázové vektory jsou hlavní směry daného souboru
- tedy kovariance mezi různými veličinami jsou nulové
- matice kovariance je diagonální

$$\Sigma = \mathbf{V}\Lambda\mathbf{V}^T$$

  - Σ je symetrická, proto jsou λ_i reálné a $\mathbf{V}$ ortogonální
  - $\mathbf{V}$ lze volit tak, že λ_i jsou uspořádány od největší

Aplikace – rozpoznávání tváří

- prvky souboru jsou normalizované fotografie
  - stejné rozlišení
  - obličej zabírá (přibližně) stejné pixely
  - pozadí je vždy stejné
  - černá a bílá odpovídá stejným hodnotám
- proměnné (složky vektoru) jsou jednotlivé pixely
  - tj. obrázek 256×256 znamená 65536 proměnných
- datová matice $\mathbf{D}_{N \times n}$ – N obrázků o n pixelech

Aplikace – rozpoznávání tváří

- vycentrování dat
  - spočteme „průměrný obrázek“ μ (průměr každého sloupce – pixelu)

$$\mu_j = \frac{1}{N} \sum_{k=0}^{N} d_{kj}$$

  - odečteme od datové matice $u_{ij} = d_{ij} - \mu_j$
- matice kovariance

$$\Sigma_{n \times n} = \frac{1}{N-1} \mathbf{U}^T \mathbf{U}$$

  - $\mathbf{U}$ je centrovaná datová matice
  - jen maticový zápis definice kovariance

Aplikace – rozpoznávání tváří

- výpočet vlastních hodnot Σ
- přímočarý postup
  - použijeme implementaci pro symetrické matice,
  - aplikujeme na $\frac{1}{N-1}\mathbf{U}^T\mathbf{U}$
  - zbytečně „nafouklá“ data ($n \gg N$)

Aplikace – rozpoznávání tváří

- výpočet vlastních hodnot Σ
- přímočarý postup
  - použijeme implementaci pro symetrické matice,
  - aplikujeme na $\frac{1}{N-1}\mathbf{U}^T\mathbf{U}$
  - zbytečně „nafouklá“ data ($n \gg N$)
- obrácený rozklad $\mathbf{U}\mathbf{U}^T = \mathbf{V}'\Lambda'\mathbf{V}'^T$
  - vynásobení $\mathbf{V}'$ zprava, $\mathbf{U}^T$ zleva a ozávorkování

$$\mathbf{U}^T\mathbf{U}(\mathbf{U}^T\mathbf{V}') = (\mathbf{U}^T\mathbf{V}')\Lambda'$$

  - tedy $\mathbf{U}^T\mathbf{V}'$ jsou vlastní vektory $\Sigma = \mathbf{U}^T\mathbf{U}$
  - Λ' jsou její vlastní hodnoty
  - ostatní vlastní hodnoty jsou nulové
- počítali jsme rozklad podstatně menší matice $N \times N$

Aplikace – rozpoznávání tváří

- vlastní vektory Σ jsou „typické tváře“
  - jsou vzájemně nezávislé
  - charakterizují nějaké rysy tváří
- ne všechny jsou stejně významné
  - přispívají k variabilitě vstupu úměrně velikosti vlastních hodnot
  - zvolíme (experimentálně) prahovou hodnotu $0 < \sigma < 1$ (např. 0.95)
  - uvažujeme pouze $\lambda_1, \ldots, \lambda_m$ tak, že

$$\sum_{i=1}^{m} \lambda_i = \sigma \sum_{i=1}^{N} \lambda_i$$

  - ostatní $\lambda_{m+1}, \ldots, \lambda_N$ zanedbáme
  - typicky pro rozumně velký počet N platí $m \ll N$

Aplikace – rozpoznávání tváří

- každý obrázek $\mathbf{p}$ lze reprezentovat souřadnicemi v prostoru PCA

$$\mathbf{p}' = \mathbf{V}^T(\mathbf{p} - \mu)$$

  - přitom posledních $n - m$ komponent $\mathbf{p}'$ lze zanedbat
  - lze použít k výrazné kompresi
- rozpoznání neznámé tváře
  - spočteme $\mathbf{p}'$ pro všechny vstupní obrázky
  - spočteme $\mathbf{p}'$ pro neznámý vstup
  - hledání nejbližšího prvku provedeme pouze v m dimenzích

Domácí úkol

- implementujte popsanou metodu rozpoznávání tváří
- použijte fotografie kamarádů/spolužáků z IS MU
  - jsou přijatelně uniformní
  - převeďte na černobílou, snažte se dosáhnout stejnoměrného jasu
- vyzkoušejte účinnost na jiných fotografiích stejných lidí
- zjistěte, jaké rysy tváře hlavní komponenty postihují
  - zvolíme libovolné hodnoty prvních m souřadnic, $\mathbf{p}'$ doplníme nulami
  - provedeme zpětnou transformaci na obrázek

$$\mathbf{p} = \mathbf{V}\mathbf{p}' + \mu$$

- pracujte i týmově, výsledky můžete předvést na některé z příštích přednášek

Ještě trocha statistiky

- praktické nevýhody kovariance
  - závislost na absolutní hodnotě veličin
  - neintuitivní jednotky (m^2 v úvodním příkladu, ale např. kgm, kdybychom uvažovali také hmotnost úlomku)

Ještě trocha statistiky

- praktické nevýhody kovariance
  - závislost na absolutní hodnotě veličin
  - neintuitivní jednotky (m^2 v úvodním příkladu, ale např. kgm, kdybychom uvažovali také hmotnost úlomku)

- zavádíme pojem korelace

$$corr(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}$$

  - existuje více podob, toto je nejběžnější tzv. „Pearsonova“
  - nadále vypovídá o vztahu veličin
  - je normovaná do intervalu $[-1,1]$ a bez jednotek
  - ± 1 – plná lineární závislost, 0 – zcela nezávislé

- PCA provádíme obdobně na korelační matici
  - numericky potenciálně lepší vlastnosti kvůli rovnoměrnému škálování
  - zpravidla stejně převáží nepoměr vlastních hodnot :-(

Zpracování dat fMRI

- funkční magenetická rezonance
- 3D diagnostická metoda funkce mozku
  - založena na NMR – rezonance atomových jader v magnetickém poli
  - v různém chemickém kontextu různé spektrum
  - specificky fMRI detekuje množství okysličené krve
- výstupní data
  - 3D struktura voxelů, typicky desítky až stovky v jednom rozměru
  - v každém voxelu jedna skalární hodnota
  - série desítek až stovek takových snímků
- použití PCA
  - detekce vzorů chování bez apriorního zavedení nějakého modelu
  - upozorní na artefakty v datech, které by jinak zkreslovaly další analýzu
  - může napovědět, jaký funkční model zvolit

Zpracování dat fMRI

- zpracovávaná data
  - série desítek až stovek 3D snímků
  - každý desítky až stovky voxelů ve všech dimenzích
  - zpravidla ruční výběr zpracovávané oblasti

Zpracování dat fMRI

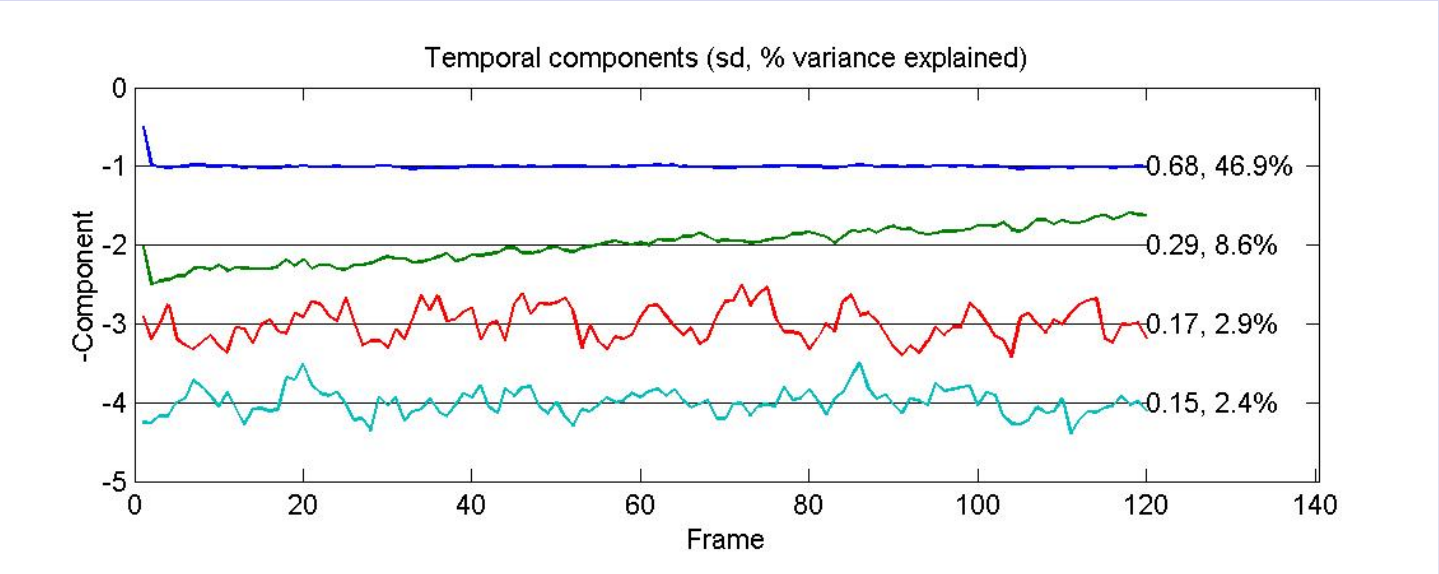
- zpracovávaná data
  - série desítek až stovek 3D snímků
  - každý desítky až stovky voxelů ve všech dimenzích
  - zpravidla ruční výběr zpracovávané oblasti
- prostorová korelace
  - matice korelace z $\mathbf{U}^T\mathbf{U}$ (odpovídá rozpoznávání tváří)
  - voxely jsou nezávislé náhodné veličiny
  - snímky chápeme jako jednotlivá pozorování
  - vysoká korelace – voxely se v čase chovají na sobě závisle
  - významnější vlastní vektory – reprezentativní snímky, z nichž dokážeme ostatní poskládat

Zpracování dat fMRI

- časová korelace
  - (podstatně méně intuitivní)
  - matice korelace z $\mathbf{U}\mathbf{U}^T$
  - body v čase jsou nezávislé náhodné veličiny
  - voxely jsou jednotlivá místa jejich pozorování
  - poloha bodu v prostoru vypovídá o jeho chování v čase
  - významné vlastní vektory – reprezentativní vzory chování voxelů

Zpracování dat fMRI

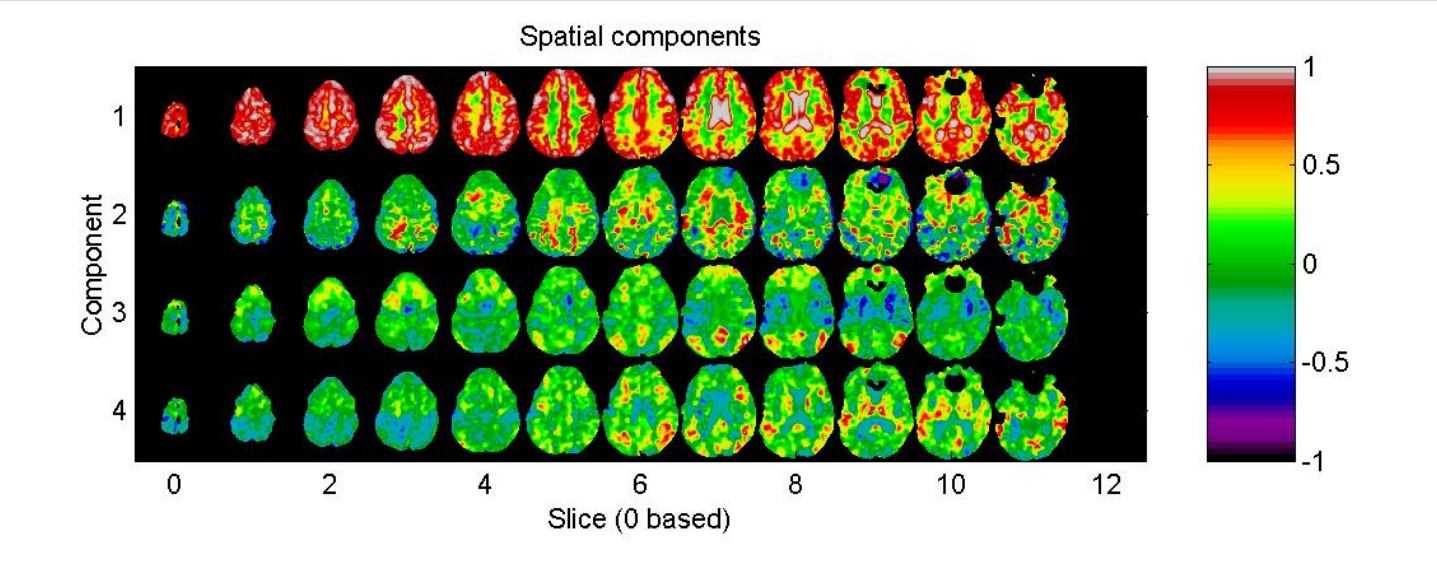
- série 120 snímků mozku
- nejvýznamnější komponenty časové korelace



- první zachycuje zjevně artefakt na začátku snímání

Zpracování dat fMRI

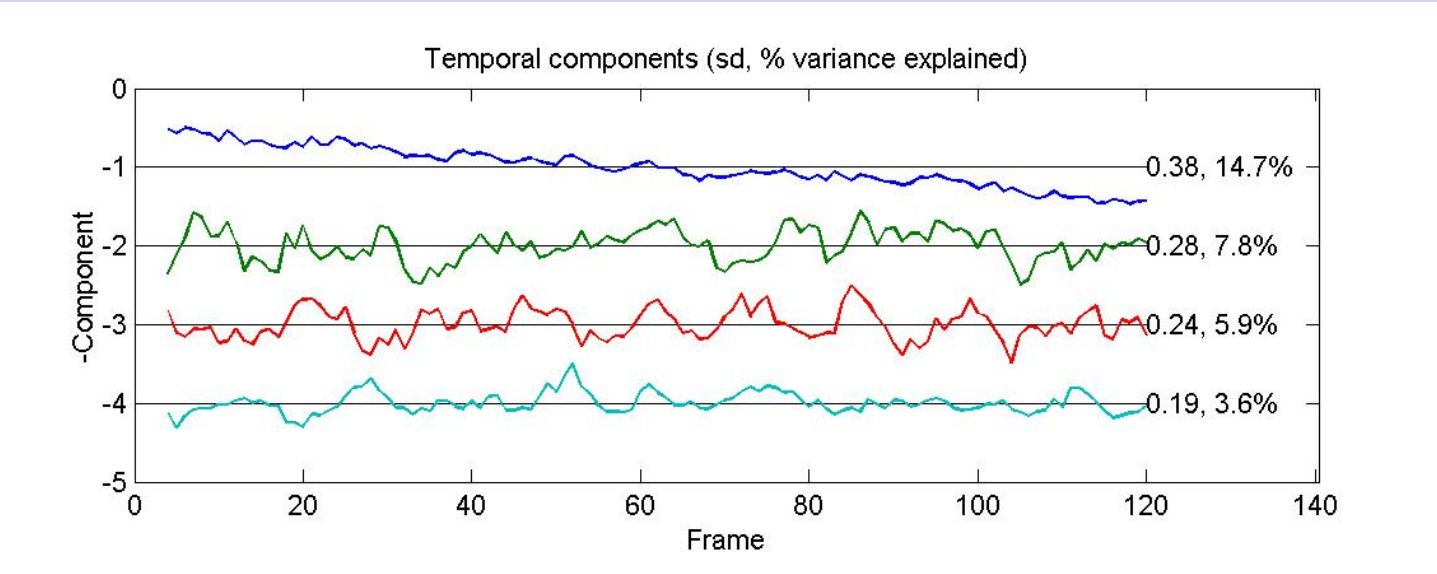
- stejný efekt je patrný na prostorových hlavních komponentách



- červené oblasti označují vzájemně závislé voxely
  - sklon chovat se v čase stejně
  - představte si na 3 voxelech, rozmístění 120 bodů odpovídajících snímkům v prosotoru, a vektorech hlavních komponent
- 1. komponenta: artefakt měření je v krajích a ve svislé ose

Zpracování dat fMRI

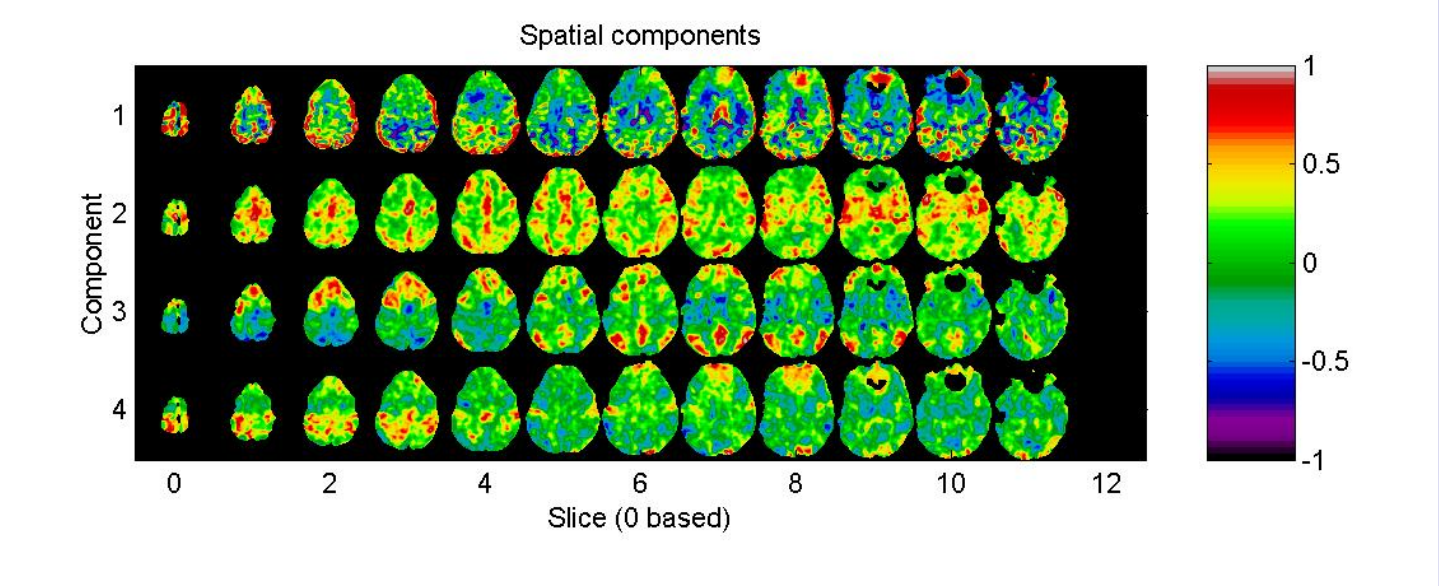
- ze souboru odstraníme prvních několik snímků



- první komponenta ukazuje už jasný lineární trend
- další jsou víceméně pravidelné oscilace

Zpracování dat fMRI

- hlavní komponenty prostorových korelací



- červená (resp. modrá) znamená voxely, které definují danou komponentu
- variabilita v čase v konkrétním hlavním směru je dána těmito voxely
- směry jsou vzájemně nezávislé
  - každý voxel je výrzně modrý nebo červený jen v jednom řádku

Shrnutí

- analýza hlavních komponent
- nalezení vzájemně nezávislých směrů variability vzorku
- identifikace jejich významu – snížení počtu dimenzí
- aplikace ve zpracování obrazu
- 2D – rozpoznávání tváří
- 3D – fMRI scany mozku