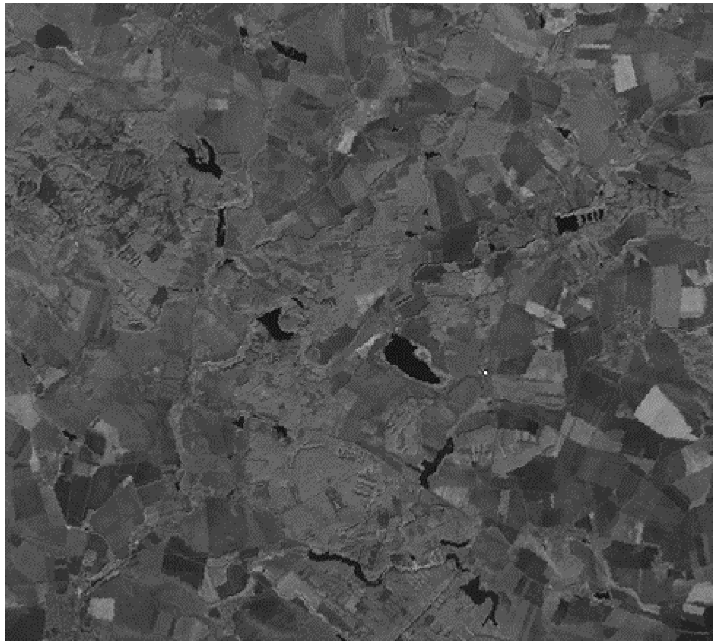
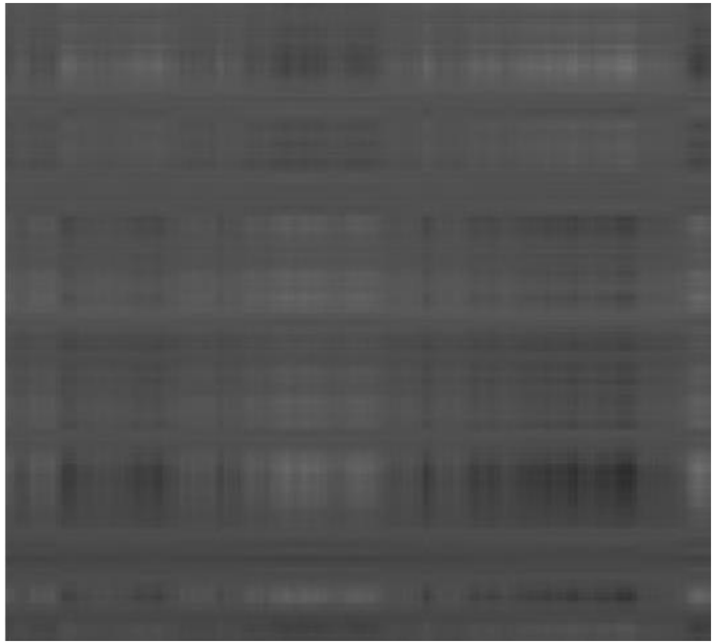


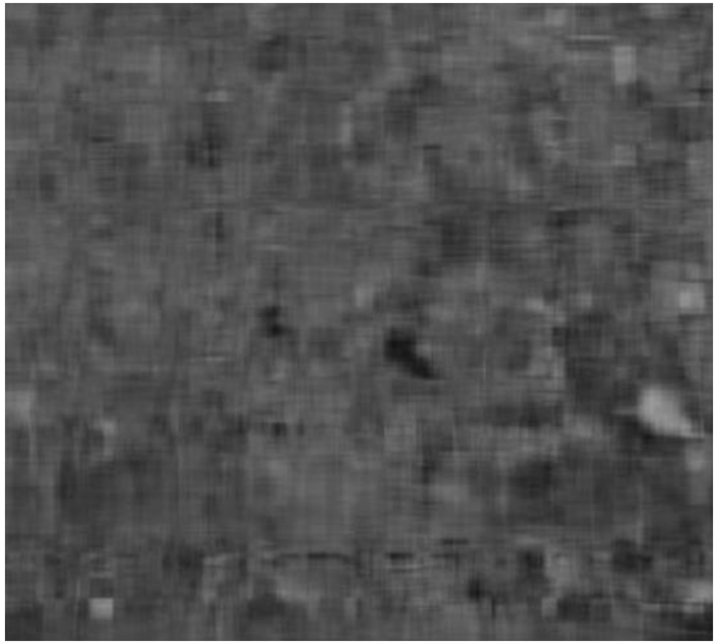
Geoinformatika

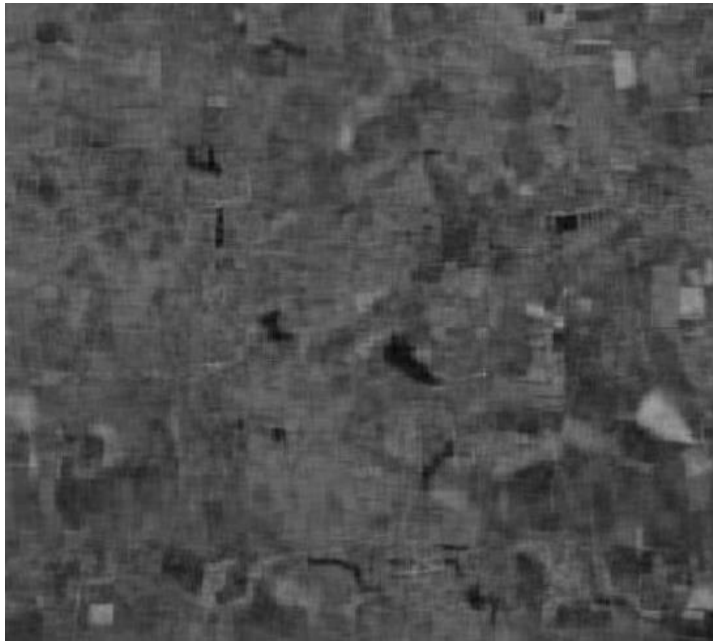
Analýza hlavních komponent

Markéta Potůčková

marketa.potuckova@natur.cuni.cz







Obsah přednášky

- Analýza dat a metoda hlavních komponent
- Hlavní komponenty – geometrický význam
- Metoda hlavních komponent – maticové vyjádření
    - Singulární rozklad
- Příklady

Pojmy

- Principal component analysis (PCA) – analýza hlavních komponent
- Eigen values/eigen vectors – vlastní čísla/vlastní vektory
- Single value decomposition – singulární rozklad

Analýza dat a metoda hlavních komponent

Na čem závisí kvalita výsledků použité metody zpracování dat?

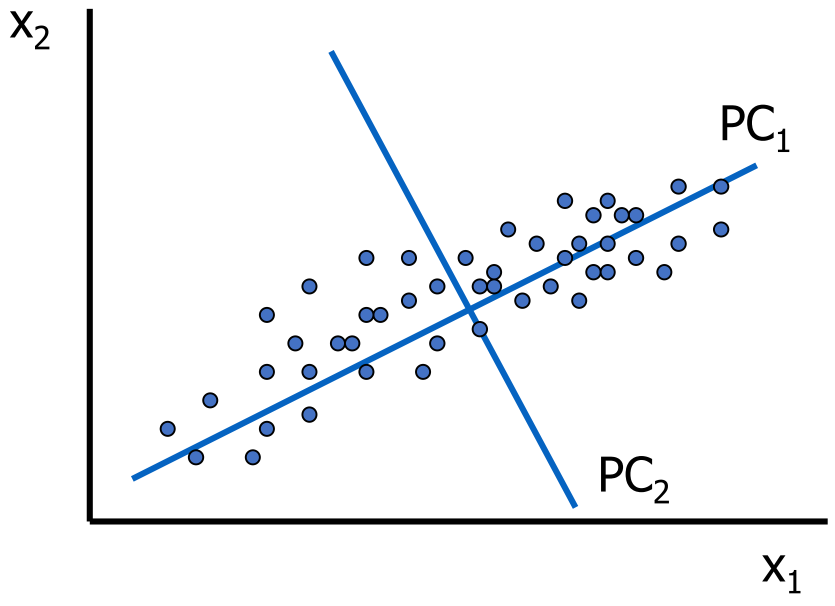
Oprávněnost a platnost předpokladů na

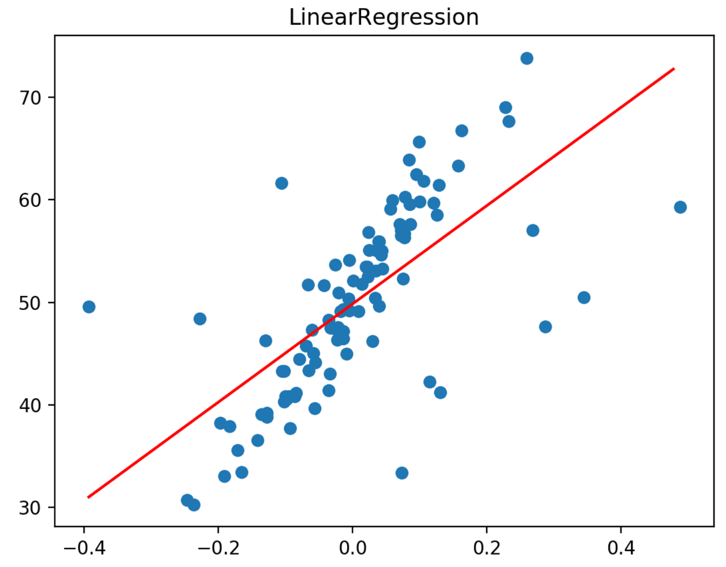
- Způsob pořízení dat
- Typ zvoleného statistického modelu/metody
- Použitý výpočetní postup

Co bychom měli o datech vědět?

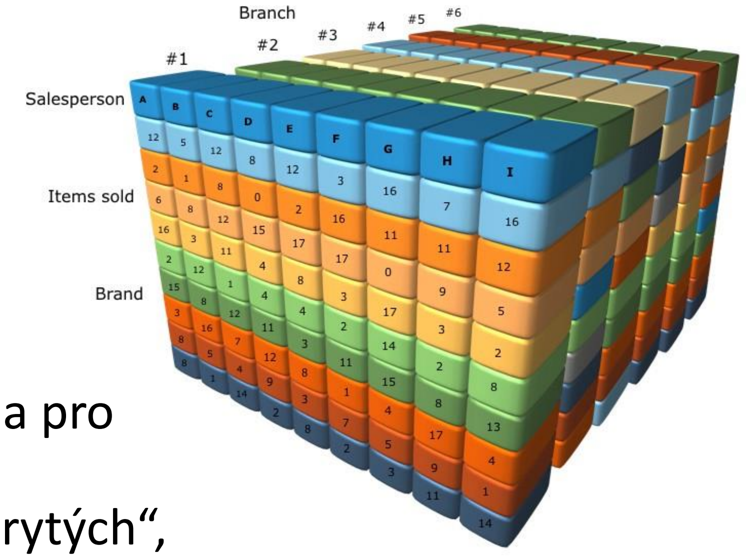
- Identifikace odlehlých/příliš vlivných pozorování (např. co způsobí při regresní analýze?, rozdíl PCA a regrese!)
- Statistické rozdělení dat (např. požadavek (vícerozměrného) normálního rozdělení)
- Lineární (obecně funkční) závislost vysvětlujících proměnných
    - Rozměr prostoru, v němž se data nacházejí, může být ve skutečnosti menší, než je počet sledovaných veličin

Řešení nabízí metoda hlavních komponent





Metoda hlavních komponent



- Nástroj pro posouzení a prověření kvality vícerozměrných dat

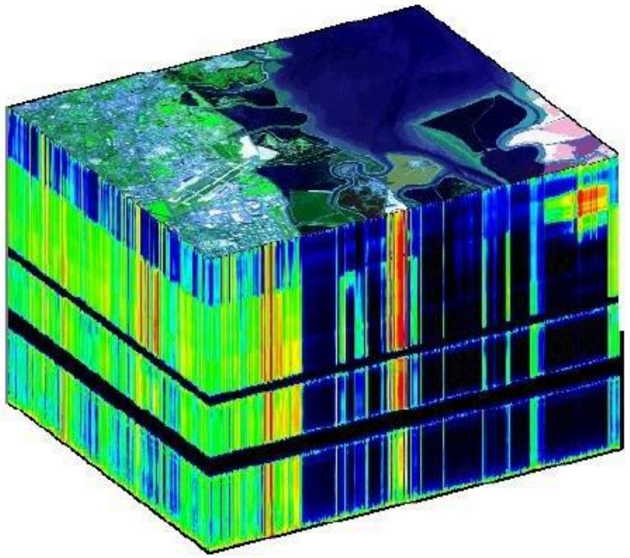
Východiska

- Výchozí počet proměnných u sledovaných jevů a procesů může být velký a pro interpretaci nepřehledný

Lze nahradit vlastnosti objektů menším počtem (třeba i „umělých“, „skrytých“, „neměřitelných“) proměnných shrnujících poznatky o výchozích proměnných, aniž by došlo k větší ztrátě informace?

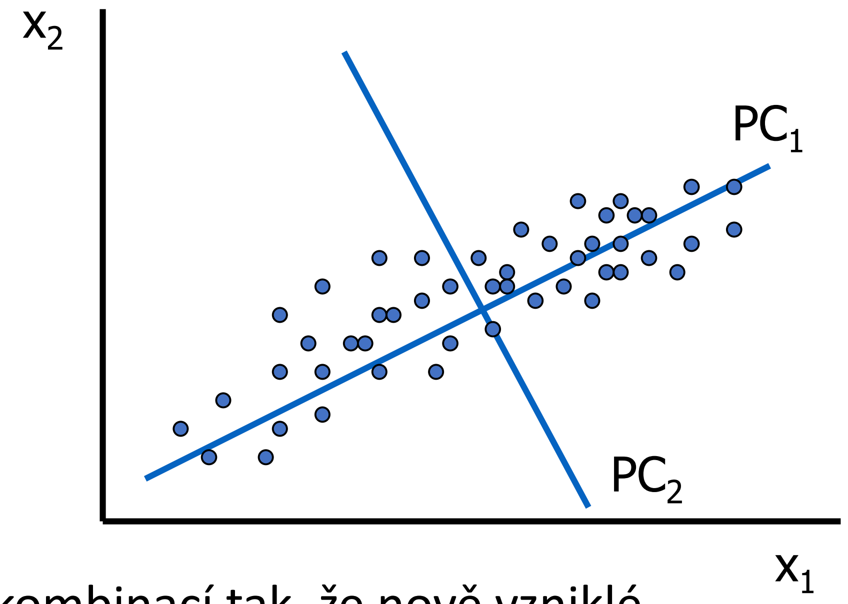
Možná řešení lineární kombinací původních/posuzovaných proměnných (ve šťastnějších případech mohou mít i určitou věcnou interpretaci ☺):

- Metoda hlavních komponent
- *Faktorová analýza* (oblíbená v sociálních vědách :)



Nalezené nové proměnné vysvětlující variabilitu, resp. *lineární závislost* původních proměnných označujeme jako komponenty, resp. nebo *faktory*

Metoda hlavních komponent



- K. Pearson 1901, H. Hotelling 1933
- Nazývaná také Hotelling nebo Karhunen–Loève transformace

Základní myšlenka

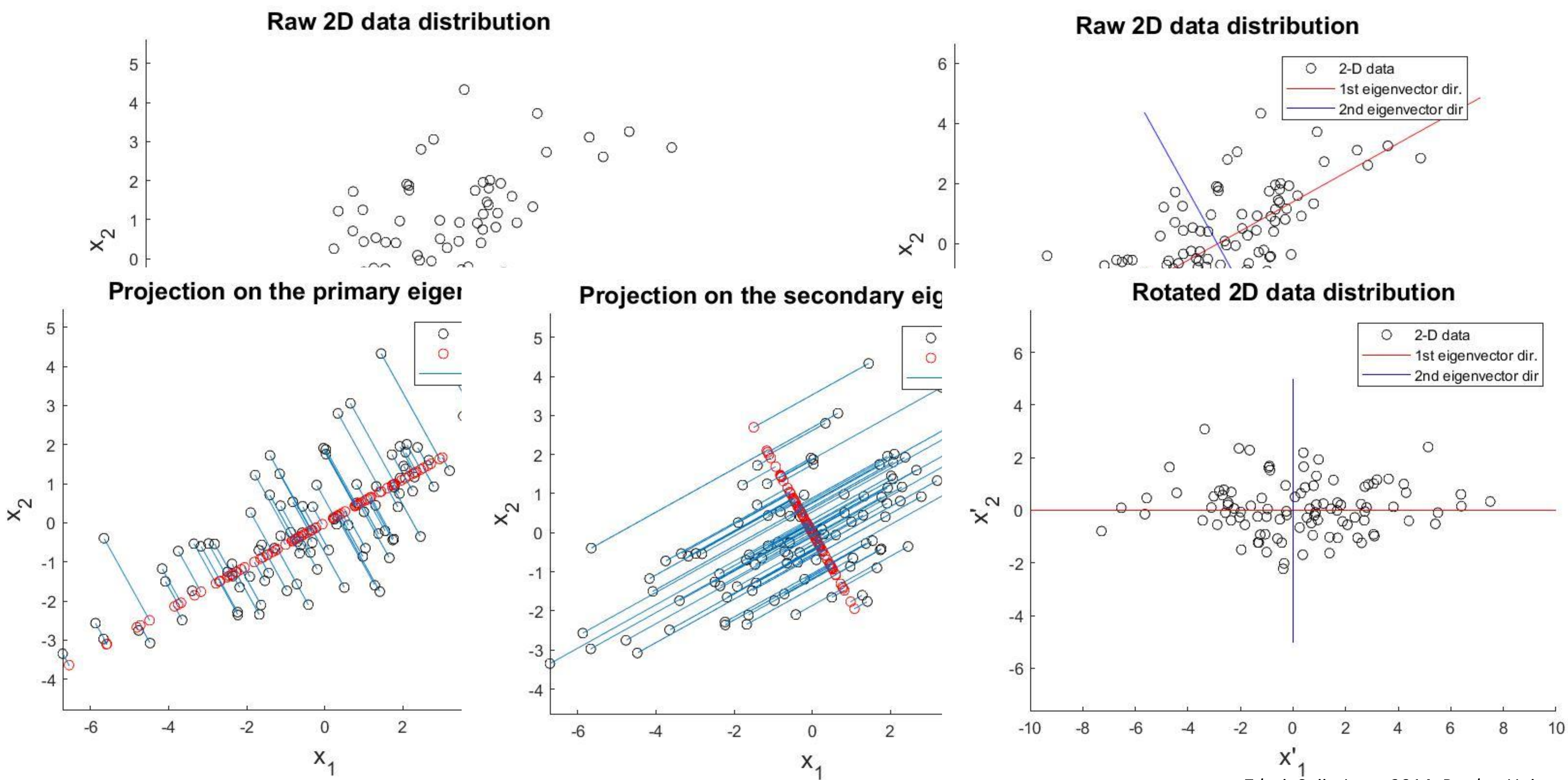
- Původní proměnné p-rozměrného prostoru nahrazuje jejich lineární kombinací tak, že nově vzniklé proměnné (komponenty) nejsou vzájemně korelované

Dále je výhodné, když

- Nové komponenty jsou uspořádány dle klesajícího významu své důležitosti; první komponenta vysvětluje co nejvíce z celkové variability (tj. součtu rozptylů zkoumaných proměnných)
- Pro p původních proměnných je $R \leq p$ „správný rozměr“ úlohy. Cílovým stavem je situace, v níž R (nejlépe výrazně menší než p) hlavních komponent dostatečně vysvětluje variabilitu původních proměnných.

Příklad korelační matice Landsat

Hlavní komponenty – geometrický význam



Zdroj: Sujin Jang, 2014, Purdue University

Metoda hlavních komponent – matematické vyjádření

Mějme náhodné veličiny $X_1, X_2, \ldots, X_p$ s vícerozměrným normálním rozdělením, p-členným vektorem středních hodnot $\boldsymbol{\mu}$ a kovarianční maticí $\boldsymbol{\Sigma}$

$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ jsou vlastní (charakteristická) čísla a jim odpovídající vlastní vektory $\boldsymbol{\omega_1, \omega_2, \ldots, \omega_p}$ kovarianční matice $\boldsymbol{\Sigma}$

První hlavní komponentou je veličina

$Y_1 = \boldsymbol{\omega}_1^T (\boldsymbol{x} - \boldsymbol{\mu})$

Vektor $\boldsymbol{\omega}_1$ je určen maximalizací rozptylu komponenty Y_1 přes všechny vektory $\boldsymbol{\omega}_i$ tak, aby byla splněna normalizační podmínka $\boldsymbol{\omega}_1^T \boldsymbol{\omega}_1 = 1.$

Maximální hodnota rozptylu Y_1 přes všechny vektory $\boldsymbol{\omega}_i$ při splnění normalizační podmínky je největší charakteristické číslo λ_1 kovarianční matice $\boldsymbol{\Sigma}$. Maxima je dosaženo jedině v případě, že $\boldsymbol{\omega}_1$ je charakteristický vektor odpovídající λ_1 a zároveň je splněna normalizační podmínka.

Stejným způsobem jsou definovány komponenty $Y_2, \ldots, Y_p$ a jim příslušející rozptyly odpovídající vlastním číslům $\lambda_2, \ldots, \lambda_p$ a vzájemně ortogonální charakteristické vektory $\boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_p$

R=p komponent vysvětluje celkový rozptyl původních proměnných: st($\boldsymbol{\Sigma}$) = $\sigma_1^2 + \sigma_2^2 + \ldots \sigma_p^2 = \lambda_1 + \lambda_2 + \ldots + \lambda_R$

Podíl λ_r / st($\boldsymbol{\Sigma}$) určuje význam r-té komponenty z hlediska celkového rozptylu proměnných X_j, j = 1, 2, …,p

Kovarianční nebo korelační matice?

- **Kovarianční matici** použijeme v případě, kdy sledované náhodné veličiny $X_1, X_2, \ldots, X_p$ jsou ve stejných nebo porovnatelných měřících jednotkách a rozptyly těchto veličin nejsou zásadně odlišné

- Při nesplnění **obou** uvedených podmínek se metoda hlavních komponent aplikuje s využitím **korelační matice**

Postup výpočtu je stejný, výsledky obou výpočtů se budou numericky lišit, jsou-li pro výpočet kovarianční matice použity původní či jen k průměru redukované vstupní veličiny. V případě normovaných veličin (redukované k průměru a normalizované směrodatnou odchylkou) je kovarianční a korelační matice totožná.

Metoda hlavních komponent – příklady výpočtu

Matlab – viz cvičení

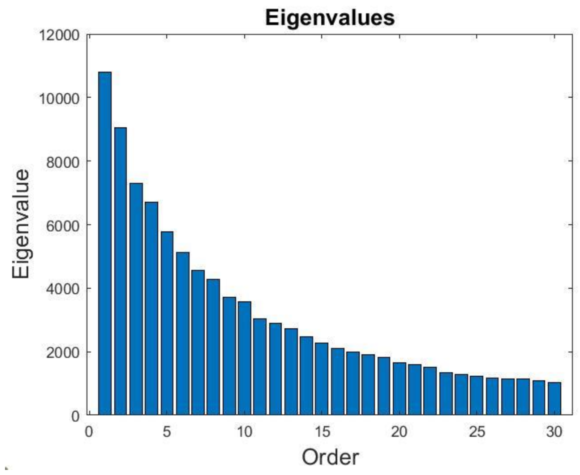
Hebák a kol. (2007): *Vícerozměrné statistické metody (3)*, Informatorium, Praha, str. 56 - 69

Kdy má smysl metodu hlavních komponent aplikovat?

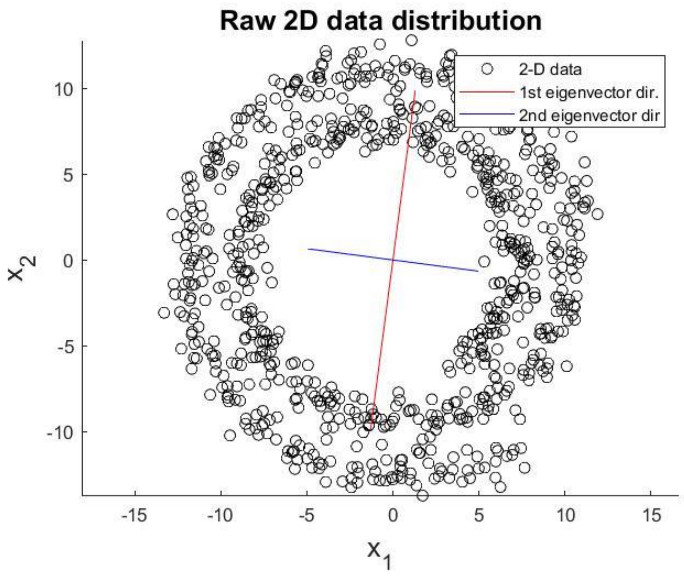
- Pro p původních proměnných je R ≤ p „správný rozměr“ úlohy. Cílovým stavem je situace, v níž R (nejlépe výrazně menší než p) hlavních komponent ***dostatečně vysvětluje*** variabilitu původních proměnných.

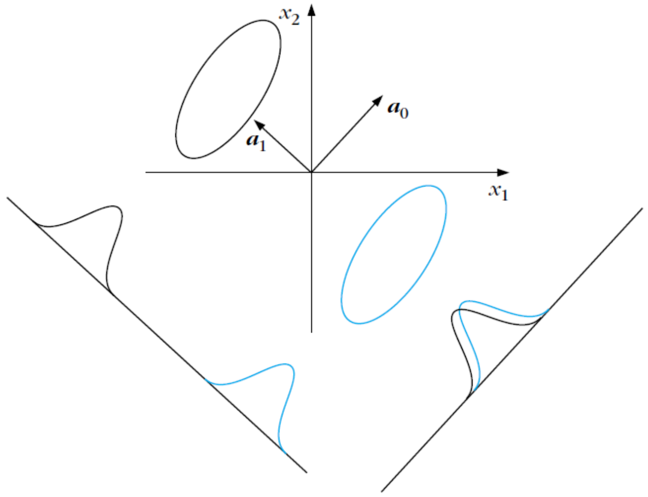
Dostatečně vysvětluje? 

- Testování hypotézy, že pro zbývajících p-R komponent platí rovnost vlastních čísel
- R komponent splňuje požadovaný podíl vysvětlené variability původních proměnných
- Scree-plot (subjektivní)



Kdy PCA nefunguje?





Příklady – viz cvičení a přednášky DPZ

- Komprese obrazu
- Omezení počtu příznaků pro klasifikaci (MS a HS data)
- Zvýraznění multispektrálního obrazu (decorrelation stretch)
- Fúze dat (pansharpening)

Odkazy a literatura

- Hebák a kol. (2007): *Vícerozměrné statistické metody (3)*, Informatorium, Praha, 271 s.
- Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics. New York: Springer-Verlag. doi:10.1007/b98835. ISBN 978-0-387-95442-4.
- Brunton, S. L., Kutz, J. N. (2022). *Data-driven science and engineering: Machine learning, dynamical systems, and control.* Cambridge University Press.
- Koutroumbas, K., Theodoridis, S. (2008): *Pattern Recognition*, 4th Edition, Academic Press, 984 p.
- Schowengerdt, R. A. (2007): *Remote Sensing Models and Methods for Image Processing,* Academic Press, 515 p.

- Videa:
- Statquest - PCA
- SVD

ÚLOHA

S využitím oblíbeného programovacího prostředí či výpočetního sw (Python, Matlab, Excel, …) vytvořte dva příklady dvourozměrných datových sad (např. po 20 pozorováních), na nichž ukážete význam transformace hlavních komponent. V prvním příkladě bude po transformaci první hlavní komponenta obsahovat alespoň 70% informace datového souboru. Ve druhém případě bude vliv transformace minimální (obsah informace v původních a transformovaných osách se nebude lišit více než o 10%). V obou případech spočítejte vlastní čísla a vlastní vektory kovarianční matice.