

## 5. Metoda hlavních komponent

Matěj Klimeš, Tomáš Zbíral

ZS 2024/25, číslo skupiny: 2, datum zpracování: 13.12.2024

---

### 1 Zadání

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.

#### 1.1 Řešené bonusové úlohy

V rámci této úlohy nebyly zadány žádné bonusové úlohy.

### 2 Teoretický úvod

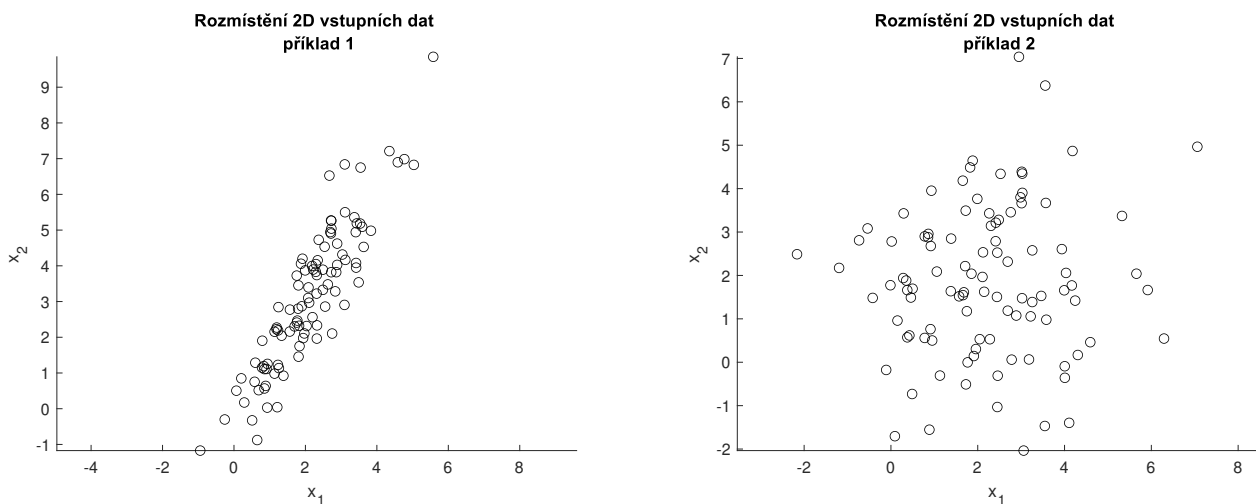
Metoda (analýza) hlavních komponent (PCA - Principal Component Analysis) je statistická technika používaná pro analýzu a prověření kvality vícerozměrných dat. Cílem PCA je transformovat data do nových os, které jsou lineárními kombinacemi původních proměnných, a tak zachovat co nejvíce informací v co nejmenším počtu nových proměnných.

### 3 Pracovní postup

- Byl vytvořen skript, do kterého uživatel zadá volbu, zda chce aplikovat příklad, kde má transformace hlavních komponent význam, nebo příklad, kde transformace nemá význam. Tuto volbu uživatel provádí prostřednictvím vstupního dialogu.
- Data byla generována na základě výběru uživatele pomocí funkce mvnrnd - Multivariate normal random numbers
  - Pokud je vybrán první příklad (příklad 1), byla generována data se středem v bodě  $\mu = [2, 3]$  a kovarianční maticí  $\Sigma = \begin{bmatrix} 1 & 1.5 \\ 1.5 & 3 \end{bmatrix}$ . Tento příklad je navržen tak, že první hlavní komponenta bude obsahovat podstatnou část variability dat.
  - Pokud je vybrán druhý příklad (příklad 2), byla generována data se středem v bodě  $\mu = [2, 2]$  a kovarianční maticí  $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$ . Tento příklad je navržen tak, že transformace hlavních komponent nebude mít velký vliv, protože data jsou téměř symetrická a nezávislá.

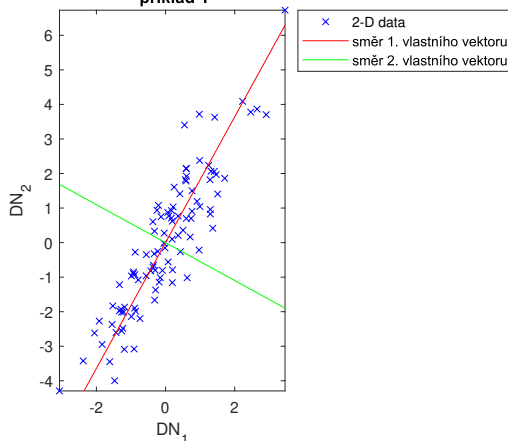
- Vstupní data byla zobrazena pomocí scatter plotu.
- Dále byla data redukována k těžišti a normalizována tak, aby variance normalizovaných dat byly rovny 1.
- Výpočet kovarianční a korelační matice
- Dále proběhl výpočet vlastních čísel a vlastních vektorů kovarianční matice
- Byla zobrazena původní data spolu se směry hlavních komponent, které ukazují, jak jsou data orientována v prostoru.
- Byla provedena transformace dat do nových os, které odpovídají hlavním komponentám.
- Následně byla data zobrazena v nových osách hlavních komponent, což ukazuje, jak jsou data rozložena po první a druhé hlavní komponentě.
- Byl spočítán informační obsah první a druhé hlavní komponenty, který ukazuje, jak velkou část celkové variability dat je zachycena jednotlivými komponentami.

## 4 Výsledky

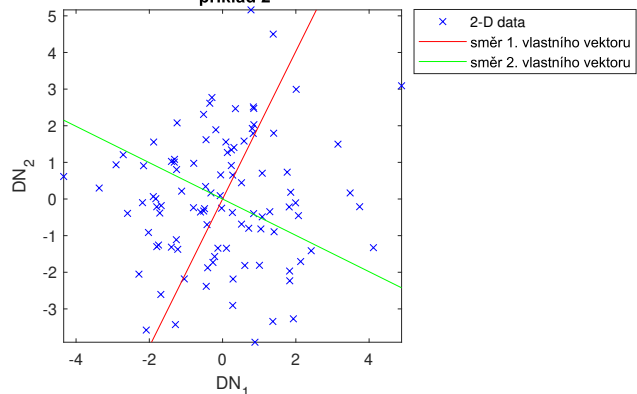


Obrázek 1: Rozmístění vstupních (generovaných) dat

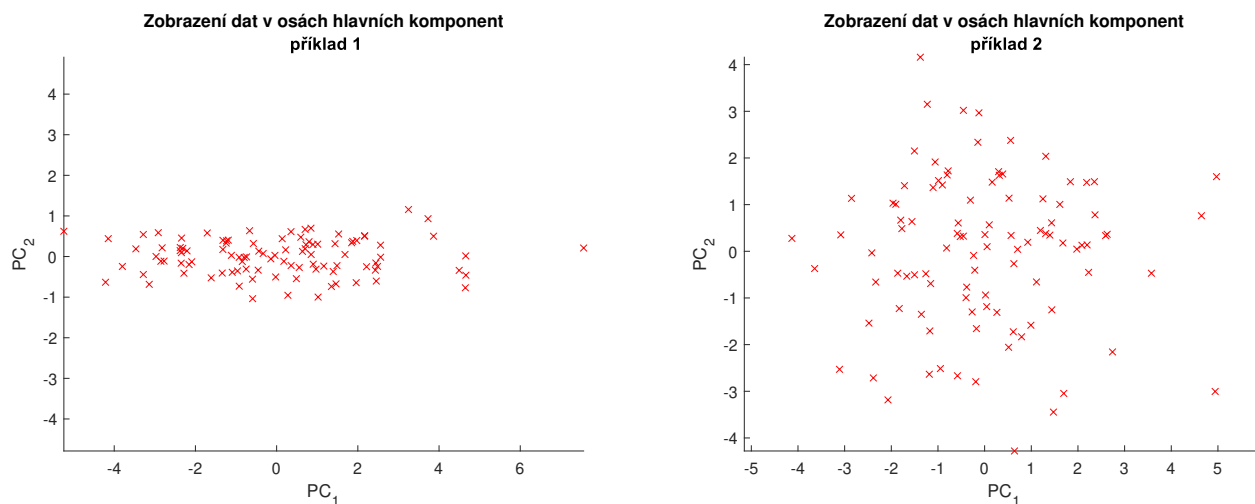
Zobrazení původních dat a směrů hlavních komponent  
příklad 1



Zobrazení původních dat a směrů hlavních komponent  
příklad 2



Obrázek 2: Zobrazení vstupních dat a směrů hlavních komponent



Obrázek 3: Zobrazení dat v osách hlavních komponent

Příklad	Komponenta	Vlastní číslo	Vlastní vektor	Informační podíl (%)
1	1. komponenta	5.1814	$\begin{bmatrix} 0.4813 \\ 0.8766 \end{bmatrix}$	96.35
	2. komponenta	0.1964	$\begin{bmatrix} -0.8766 \\ 0.4813 \end{bmatrix}$	3.65
2	1. komponenta	3.1375	$\begin{bmatrix} 0.4444 \\ 0.8958 \end{bmatrix}$	54.73
	2. komponenta	2.5953	$\begin{bmatrix} -0.8958 \\ 0.4444 \end{bmatrix}$	45.27

Tabulka 1: Vlastní čísla, vlastní vektory a informační podíl hlavních komponent

## 5 Závěr

V rámci této úlohy byly vytvořeny dvě datové sady, na které byla aplikována metoda hlavních komponent (PCA). Práce probíhala v sw MATLAB.

- V prvním příkladě obsahovala první hlavní komponenta přibližně 96 % informačního podílu.
- V druhém příkladě byl rozdíl mezi informačními podíly první a druhé komponenty přibližně 9 %.
- Pro oba příklady byly spočítány vlastní čísla a vlastní vektory kovarianční matice.

## 6 Přílohy

Příloha 1 - MATLAB Live script

V Praze dne 13.12.2024

Matěj Klimeš, Tomáš Zbíral