ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE FAKULTA STAVEBNÍ KATEDRA GEOMATIKY					
Název předmětu Geoinformatika					
Úloha U5	Název úlohy: Analýza hlavních komponent				
akademický rok 2024/2025	semestr zimní	studijní skupina C102	vypracoval Matyáš Pokorný Tereza Černohousová	datum 13.12.2024	klasifikace

Technická zpráva

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.

2 Pracovní postup

Nejprve byla data načtena a centrálně posunuta odečtením jejich průměrů. Poté byly vypočteny kovarianční matice a z nich určena vlastní čísla a vlastní vektory, které reprezentují hlavní směry variability v datech. Tyto vlastní vektory byly seřazeny podle důležitosti (podle velikosti vlastních čísel) a využity pro projekci dat do nových souřadnicových os hlavních komponent. Byl také stanoven podíl vysvětlené variance pro každou hlavní komponentu, což ilustruje, jaká část variability dat je zachycena jednotlivými směry.

Následně byla provedena vizualizace výsledků, která zahrnovala zobrazení původních dat, směry hlavních komponent v původním prostoru a projekce dat do os hlavních komponent. Grafy ukázaly, jak data leží podél hlavních os, a umožnily posoudit, jak efektivně PCA redukuje rozměry a vysvětluje variabilitu dat. Podíl vysvětlené variance byl zobrazen pomocí sloupcových grafů, což poskytlo přehled o důležitosti jednotlivých hlavních komponent pro obě datové sady. Tento postup umožnil hlubší pochopení struktury a vztahů v analyzovaných datech.

3 Pseudokód

3.1 Generování dat

Matlab kód `GenPoints.m`

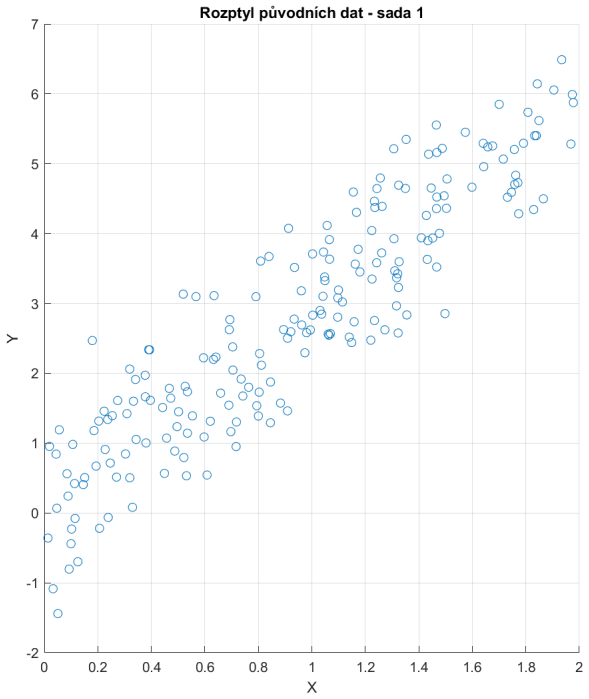
- Pro první datovou sadu ($sada1$):
  - Generuj náhodné hodnoty x s vysokou variabilitou.
  - Vypočítej y jako lineární kombinaci x s přidaným šumem.
- Pro druhou datovou sadu ($sada2$):
  - Generuj náhodné hodnoty q a w s malou variabilitou.
- Ulož obě datové sady do souborů (`sada1.mat`, `sada2.mat`).

3.2 Realizace analýzy hlavních komponent

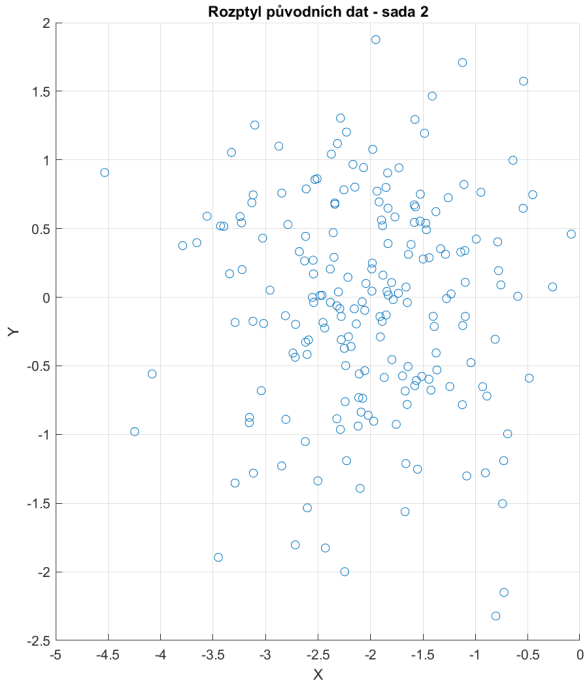
Matlab kód `MyPCA.m`

- **Vstup:** Dvě datové sady *sada*1 a *sada*2.
- **Výstup:**
  - Hlavní komponenty a vysvětlená variance.
  - Vizualizace dat a jejich projekcí na hlavní komponenty.
- **Načtení dat:**
  - Načti datové sady *sada*1 a *sada*2 ze souborů.
- **Předzpracování dat:**
  - Od každého sloupce dat odečti průměr (centrování dat).
- **Výpočet kovarianční matice:**
  - Pro každou datovou sadu spočítej kovarianční matici (`cov`).
- **Výpočet korelační matice:**
  - Pro každou datovou sadu spočítej korelační matici (`corrcoef`).
- **Spektrální analýza:**
  - Spočítej vlastní čísla a vlastní vektory kovarianční matice (`eig`).
  - Seřaď vlastní čísla sestupně a odpovídající vlastní vektory.
- **Projekce dat na hlavní komponenty:**
  - Vypočítej projekci dat na hlavní komponenty vynásobením datových bodů maticí vlastních vektorů.
- **Výpočet vysvětlené variance:**
  - Spočítej podíl variance vysvětlené jednotlivými hlavními komponentami.
- **Ověření podmínek zadání:**
  - Pro *sada*1: Zkontroluj, že první hlavní komponenta vysvětluje více než 70 % variance.
  - Pro *sada*2: Zkontroluj, že rozdíl mezi vysvětlenou variancí první a druhé komponenty je menší než 10 %.
- **Vizualizace:**
  - Vykresli původní data a směr hlavních komponent.
  - Vykresli projekci dat do os hlavních komponent.
  - Zobraz grafy podílu vysvětlené variance.

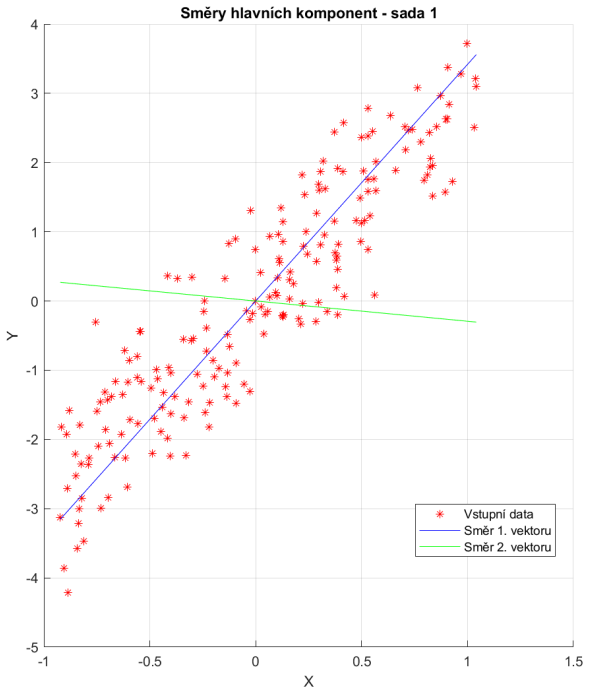
4 Výsledky



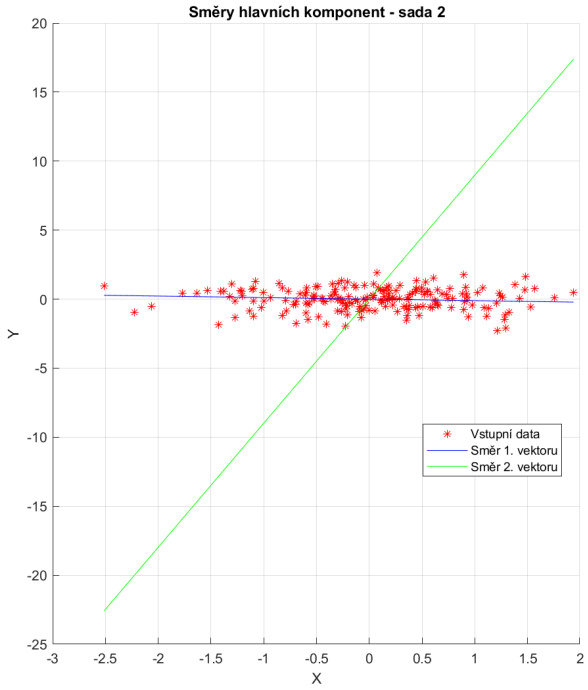
Obrázek 1: Vstupní data 1. sady



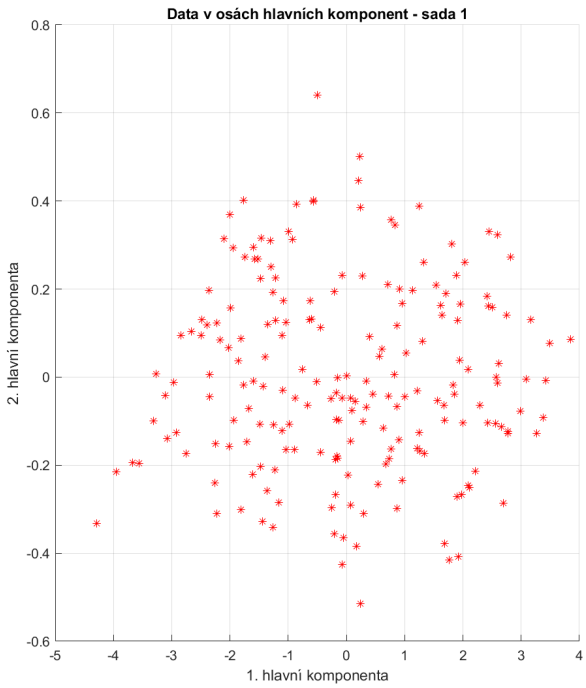
Obrázek 2: Vstupní data 2. sady



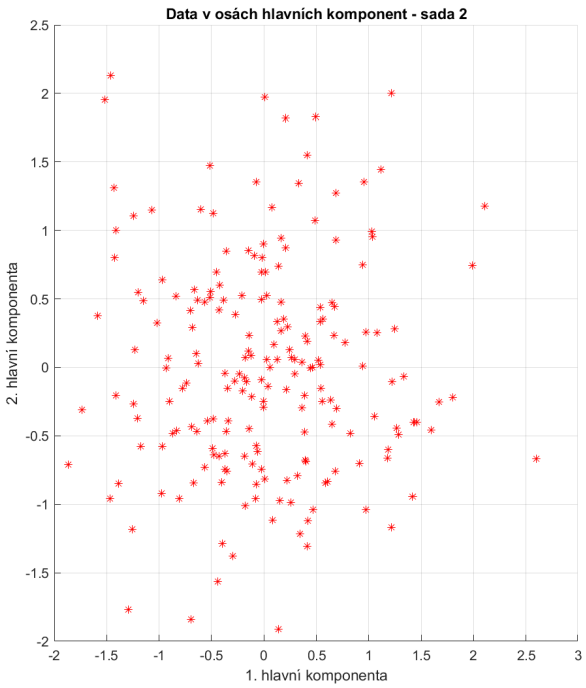
Obrázek 3: Směry hlavních komponent 1. sady



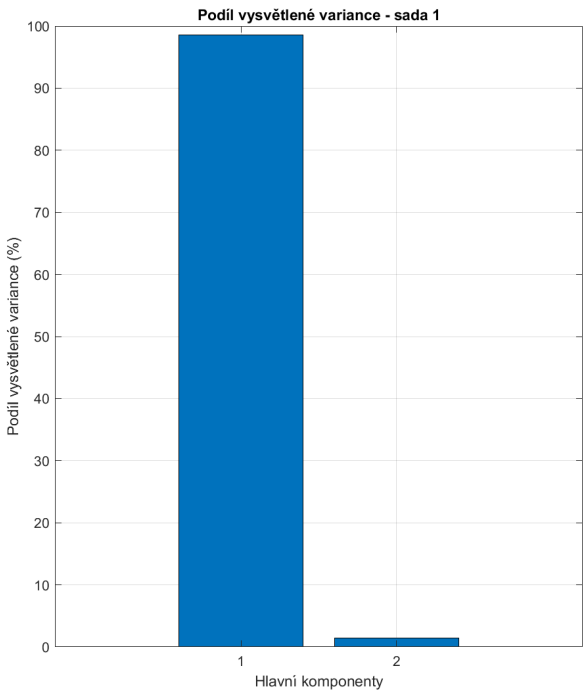
Obrázek 4: Směry hlavních komponent 2. sady



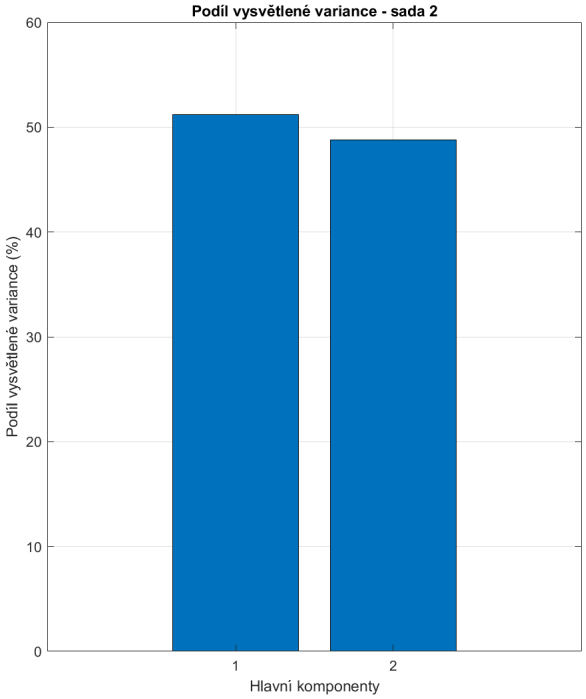
Obrázek 5: Data 1. sady v osách hlavních komponent



Obrázek 6: Data 2. sady v osách hlavních komponent



Obrázek 7: Kontrola splnění 1. sady



Obrázek 8: Kontrola splnění 2. sady

Hodnoty pro vlastní čísla a vlastní vektory kovarianční matice jsou součástí výpočetního skriptu v Maltabu.

5 Závěr

V této úloze jsme se zaměřili na analýzu hlavních komponent (PCA) dvou dvourozměrných datových sad. První datová sada byla navržena tak, aby první hlavní komponenta obsahovala alespoň 70 % informace datového souboru, zatímco ve druhé datové sadě byl vliv transformace minimální.

Postup zahrnoval načtení a centrální posunutí dat, výpočet kovariančních matic, určení vlastních čísel a vlastních vektorů, a následnou projekci dat do nových souřadnicových os hlavních komponent. Výsledky byly vizualizovány pomocí grafů, které ukázaly, jak data leží podél hlavních os a jak efektivně PCA redukuje rozměry a vysvětluje variabilitu dat.

Tato analýza poskytla hlubší pochopení struktury a vztahů v analyzovaných datech a ukázala, jak může být PCA efektivně použita pro redukci rozměrů a vysvětlení variability dat.

Hodnoty pro vlastní čísla a vlastní vektory kovarianční matice jsou součástí výpočetního skriptu v Maltabu.

V Praze dne: 13.12. 2024 **T. Černohousová**

M. Pokorný