ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE FAKULTA STAVEBNÍ KATEDRA GEOMATIKY					
Název předmětu: **Geoinformatika**					
Úloha: U5	**Název úlohy:** Metoda hlavních komponent				
Akademický rok: 2024/2025	**Semestr:** zimní	**Studijní skupina:** 102C	**Vypracoval:** Michal Kovář Filip Roučka Magdaléna Soukupová	**Datum:** 13. 12. 2024	**Klasifikace:**

1 Zadání

1. **Generování datových sad**: Vytvořit dva různé příklady dvourozměrných datových sad (například po 20 pozorováních), na nichž bude ukázán význam transformace hlavních komponent.
2. **Transformace hlavních komponent**: Na první datové sadě má první hlavní komponenta obsahovat alespoň 70% informace z původního datového souboru. Na druhé datové sadě má být vliv transformace minimální (obsah informace v původních a transformovaných osách se nebude lišit více než o 10%).
3. **Výpočet vlastních čísel a vektorů**: V obou případech mají být spočítána vlastní čísla a vlastní vektory kovarianční matice.

2 Bonus

V rámci této úlohy nebyla řešena žádná bonusová úloha.

3 Popis problému

Metoda hlavních komponent (PCA) je technika používaná pro dekorelaci a redukci rozměru dat. Cílem je najít nové osy (komponenty), které co nejvíce vysvětlují variabilitu původních dat. PCA je založena na transformaci původních proměnných do nových, vzájemně ortogonálních komponent. První komponenta je ta, která vysvětluje největší část variability v datech, druhá komponenta vysvětluje co nejvíce variability zůstávající po první a tak dále. Význam jednotlivých komponent je určen hodnotami vlastních čísel, která udávají, jaký podíl celkové variability daná komponenta vysvětluje.[1][2]

3.1 Dílčí kroky analýzy hlavních komponent

- **Příprava dat**: Ověření zda má PCA smysl a případné odstranění odlehlých hodnot.
- **Výpočet korelační/kovarianční matice**: Kovarianční matice se využívá v případě, kdy sledované náhodné veličiny 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.[1]
- **Vlastní čísla a vlastní vektory**: Výpočet vlastních čísel a vektorů kovarianční/korelační matice.
- **Transformace dat**: Původní data jsou transformována na nové osy, které odpovídají hlavním komponentám.

4 Postup

Úloha byla zpracována v softwaru MATLAB a programovacím jazyce Python.

4.1 Generování dat

Nejprve bylo vygenerováno 200 bodů s normálním rozdělením pro proměnné x a y.

Pro první datovou sadu byly souřadnice ve směru osy y vynásobeny dvěma, aby získaly protáhlý charakter. Poté byla data rotována maticí rotace o úhel $\pi/4$, aby se natočila vůči původním osám. Výsledkem je datová sada `points1`.

Pro druhou datovou sadu byl vytvořen náhodný úhel pro kruh pomocí $2\pi x$ a náhodný poloměr $r = y$. Uprostřed kruhově generovaných dat byla vytvořena „díra“ s poloměrem 0.8. Body byly kombinovány do matice `points2`.

4.2 Výpočet kovarianční a korelační matice

Byly vypočteny kovarianční a kovarianční matice pro obě datové sady `points1` a `points2`.

4.3 Výpočet vlastních čísel a vektorů

Pro obě datové sady byly vypočítány vlastní čísla a vlastní vektory korelačních matic. Vlastní čísla a vektory byly seřazeny sestupně podle velikosti vlastních čísel.

4.4 Analýza hlavních komponent

Bylo ověřeno, zda první hlavní komponenta obsahuje alespoň 70% informace pro první datovou sadu. Pro druhou datovou sadu bylo ověřeno, zda je vliv transformace minimální (rozdíl v informacích mezi hlavními komponentami není větší než 10%).

4.5 Vizualizace dat

Byly grafy pro obě datové sady. Hlavní komponenty byly vykresleny jako šipky, které ukazují směr a velikost jednotlivých komponent.

5 Popis metod

Postup analýzy hlavních komponent

$x \leftarrow \text{randn}(1, num\_points)$ // Generování dat pro x

$y \leftarrow \text{randn}(1, num\_points)$ // Generování dat pro y

$theta \leftarrow \pi/4$ // Úhel rotace

$rotation\_matrix \leftarrow \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$ // Definice rotační matice

$points1 \leftarrow rotation\_matrix \cdot \begin{bmatrix} x \\ y \cdot 2 \end{bmatrix}$ // Kombinace dat do points1

$circle\_angle \leftarrow 2 \cdot \pi \cdot x$ // Vytvoření náhodného úhlu pro kruh

$r \leftarrow y$ // Vytvoření náhodného poloměru pro kruh

while $\text{any}(r < 0.8)$ **do**

  $r(r < 0.8) \leftarrow \sqrt{\text{rand}(1, \text{sum}(r < 0.8))}$ // Vytvoření díry v datech

end while

$points2 \leftarrow \begin{bmatrix} r \cdot \cos(circle\_angle) \\ r \cdot \sin(circle\_angle) \end{bmatrix}$ // Kombinace dat do points2

$covMatrix1 \leftarrow \text{cov}(points1')$

$covMatrix2 \leftarrow \text{cov}(points2')$

$corrMatrix1 \leftarrow \text{corr}(points1')$

$corrMatrix2 \leftarrow \text{corr}(points2')$

$[eigenvectors1, eigenvalues1] \leftarrow \text{eig}(corrMatrix1)$

$[eigenvectors2, eigenvalues2] \leftarrow \text{eig}(corrMatrix2)$

$[eigenvalues\_sorted1, indices1] \leftarrow \text{sort}(\text{diag}(eigenvalues1), 'descend')$

$eigvecsort1 \leftarrow eigenvectors1(:, indices1)$

$[eigenvalues\_sorted2, indices2] \leftarrow \text{sort}(\text{diag}(eigenvalues2), 'descend')$

$eigvecsort2 \leftarrow eigenvectors2(:, indices2)$

$info\_pc1\_points1 \leftarrow eigenvalues\_sorted1(1)/\text{sum}(eigenvalues\_sorted1)$

$info\_pc1\_points2 \leftarrow eigenvalues\_sorted2(1)/\text{sum}(eigenvalues\_sorted2)$

$info\_pc2\_points2 \leftarrow eigenvalues\_sorted2(2)/\text{sum}(eigenvalues\_sorted2)$

$diff\_info\_points2 \leftarrow \text{abs}(info\_pc1\_points2 - info\_pc2\_points2)$

6 Vstupní data

Vstupní data se generují náhodně a jsou rozdělena do dvou datových sad s odlišnými charakteristikami.

6.1 První datová sada (Points1)

První datová sada má podlouhlý charakter a je natočená o úhel $\frac{\pi}{4}$. Data jsou generována z normálního rozdělení a následně transformována pomocí rotační matice a souřadnice y je násobena koeficientem 2.

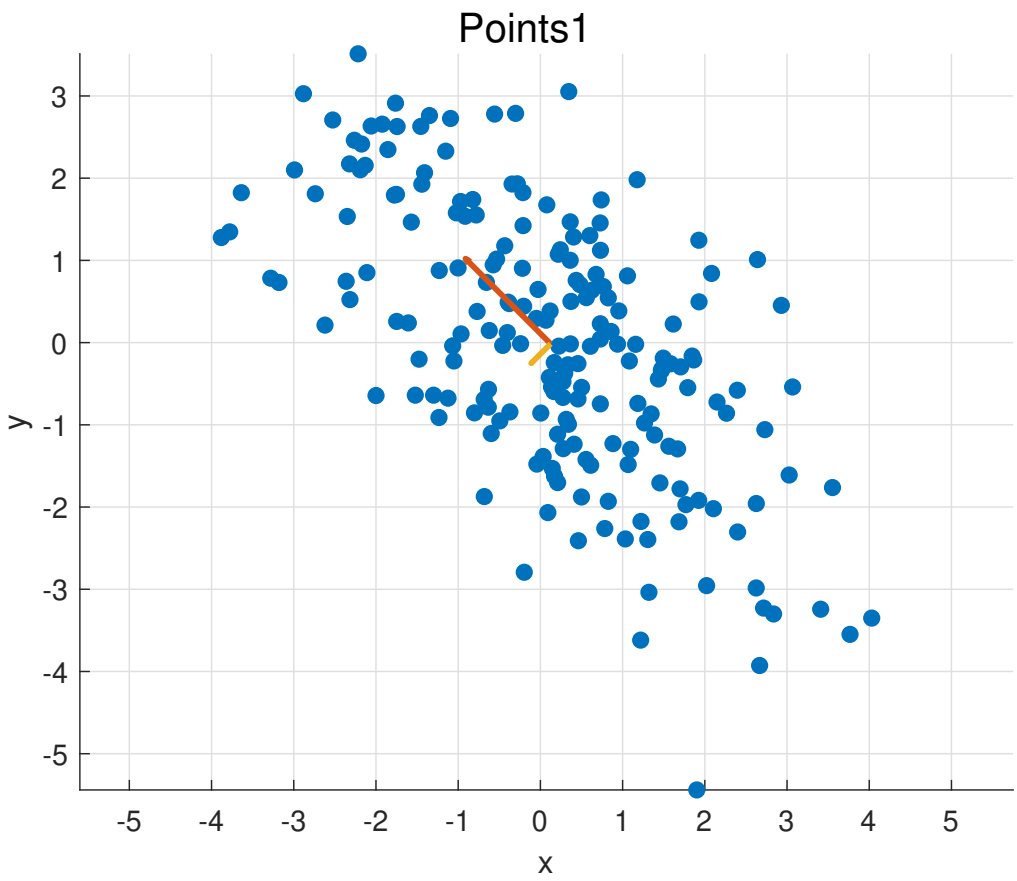
6.2 Druhá datová sada (Points2)

Druhá datová sada má charakter kružnice. Body jsou generovány tak, aby tvořily kruhový tvar s náhodnými úhly a poloměry. Navíc je v datech vytvořen otvor poloměrem 0.8.

7 Výstupní data

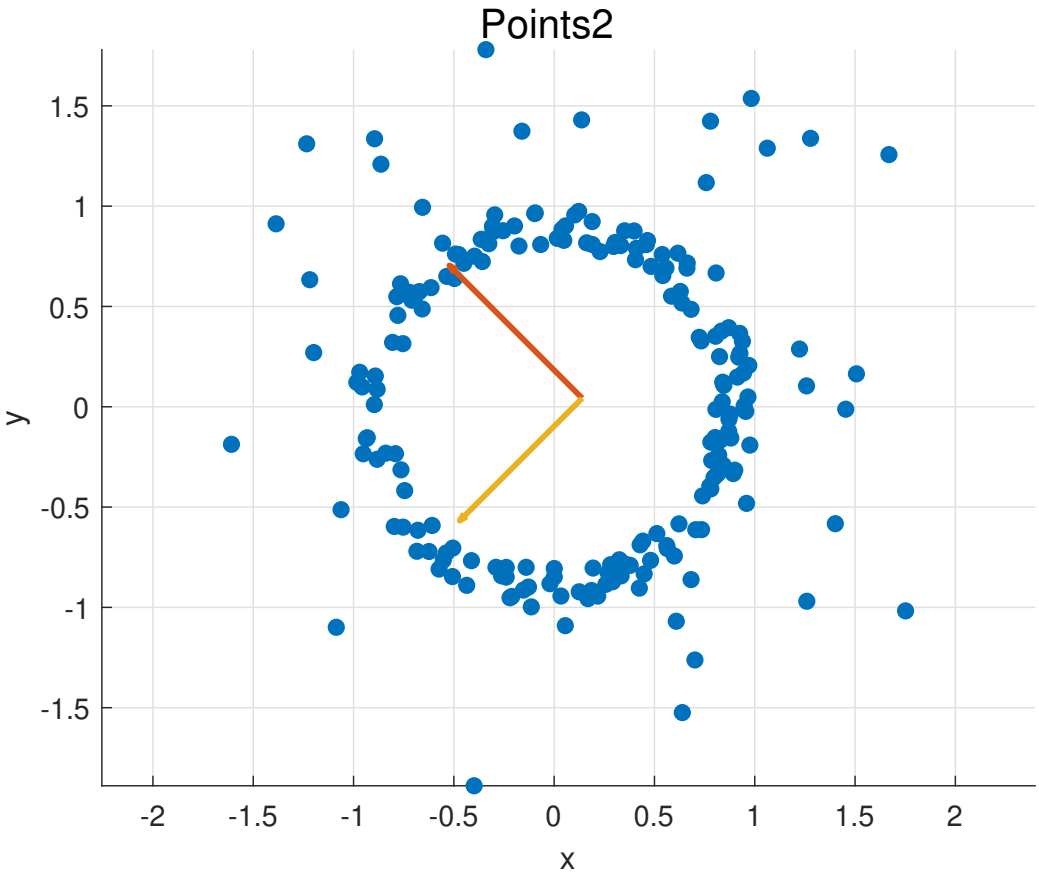
Výstupními daty jsou vizualizace vygenerovaných bodů a znázornění jejich hlavních komponent. Dále je výstupem kovarianční a korelační matice, vlastní čísla a vlastní vektory a informace o hlavních komponentách.

7.1 Body 1



Obrázek 1: Body 1 a jejich hlavní osy.

7.2 Body 2



Obrázek 2: Body 2 a jejich hlavní osy.

7.3 Matice kovariance a korelace

Kovarianční matice pro body 1:

$$\begin{bmatrix} 2.42 & -1.58 \\ -1.58 & 2.62 \end{bmatrix}$$

Kovarianční matice pro body 2:

$$\begin{bmatrix} 0.50 & -0.02 \\ -0.02 & 0.56 \end{bmatrix}$$

Korelační matice pro body 1:

$$\begin{bmatrix} 1 & -0.63 \\ -0.63 & 1 \end{bmatrix}$$

Korelační matice pro body 2:

$$\begin{bmatrix} 1 & -0.04 \\ -0.04 & 1 \end{bmatrix}$$

7.4 Vlastní čísla a vlastní vektory

Vlastní čísla pro body 1:

$$\begin{bmatrix} 1.63 \\ 0.37 \end{bmatrix}$$

Vlastní vektory pro body 1:

$$\begin{bmatrix} -0.71 & -0.71 \\ -0.71 & 0.71 \end{bmatrix}$$

Vlastní čísla pro body 2:

$$\begin{bmatrix} 1.04 \\ 0.96 \end{bmatrix}$$

Vlastní vektory pro body 2:

$$\begin{bmatrix} -0.71 & -0.71 \\ -0.71 & 0.71 \end{bmatrix}$$

7.5 Informace v hlavních komponentách

Informace v první hlavní komponentě pro body 1: 81.39%
Rozdíl v informaci mezi hlavními komponentami pro body 2: 4.01%

8 Závěr

Byl vyhotoven skript v MATLABu a pythonu, který generuje dvě dvourozměrné datové sady a provádí transformaci pomocí hlavních komponent (PCA). První datová sada má podlouhlý charakter a je natočená, zatímco druhá datová sada má charakter kružnice. Výsledky ukazují, že v první datové sadě obsahuje první hlavní komponenta alespoň 70% informace, zatímco ve druhé datové sadě je rozdíl v informaci mezi hlavními komponentami menší než 10%.

8.1 Možné oblasti pro vylepšení

- **Zadání uživatelských vstupů**: Uživatel by mohl zadávat parametry jako uživatelský vstup.
- **n-dimenzionální data**: Kód by mohl být rozšířen tak, aby fungoval s n-dimenzionálními daty.

Odkazy

[1] Markéta Potůčková. *GEOINFORMATIKA: Analýza hlavních komponent.* Prezentace k přednášce. Katedra aplikované geoinformatiky a kartografie, Přf UK, 2024. URL: `marketa.potuckova@natur.cuni.cz`.

[2] K. Koutroumbas a S. Theodoridis. *Pattern Recognition.* 2008. URL: `https://github.com/free-educa/books/issues/27` (cit. 05. 12. 2024).