Національний технічний університет України «КПІ»

Кафедра автоматизованих систем обробки інформації і управління

Лабораторна робота №8

з дисципліни «[Аналіз даних в інформаційних управляючих системах](https://do.ipo.kpi.ua/course/view.php?id=5470)»

на тему: «МЕТОД ГОЛОВНИХ КОМПОНЕНТ»

Виконав:

студент групи ІС-23

Шимків М.В.

Викладач:

Гавриленко О.В

Київ 2024

**Мета роботи**. Практичне освоєння методу головних компонент для вирішення завдань

зниження розмірності і візуалізації багатовимірних даних.

**Завдання до роботи**

1. Ознайомитися з конспектом лекцій та рекомендованою літературою, а також

додатком до лабораторної роботи, що містить короткі теоретичні відомості про метод

головних компонент та його застосування в Matlab.

2. За допомогою команди >> help вивчити функції eig, fliplud, fliplr, corrcoef, cov, corr, scatter.

3. Розробити алгоритм методу головних компонент і програмно його реалізувати в

середовищі MATLAB.

4. Виконати аналіз експериментальних даних методом головних компонент.

• Завантажити дані згідно з вашим варіантом.

• Відкрийте дані на екрані монітора у вигляді таблиці (на зразок табл. 2).

• Нормувати (стандартизувати) вихідні експериментальні дані. Побудувати

кореляційну матрицю.

• Упевниться, що кореляційна матриця значимо відрізняється від одиничної

матриці.

• Розрахувати проекції об'єктів на головні компоненти.

• Побудувати матриці даних, рахунків, помилок та навантажень.

5. Провести аналіз результатів роботи методу головних компонент.

• Перевірити рівність сум вибіркових дисперсій вихідних ознак і вибіркових

дисперсій проекцій об'єктів на головні компоненти.

• Визначити відносну частку розкиду, що припадає на головні компоненти.

Побудувати матрицю коваріації для проекций об'єктів на головні компоненти.

• На основі перших M=2 головних компонент побудувати діаграму розсіювання.

Дати змістовну інтерпретацію перших двох головних компонент.

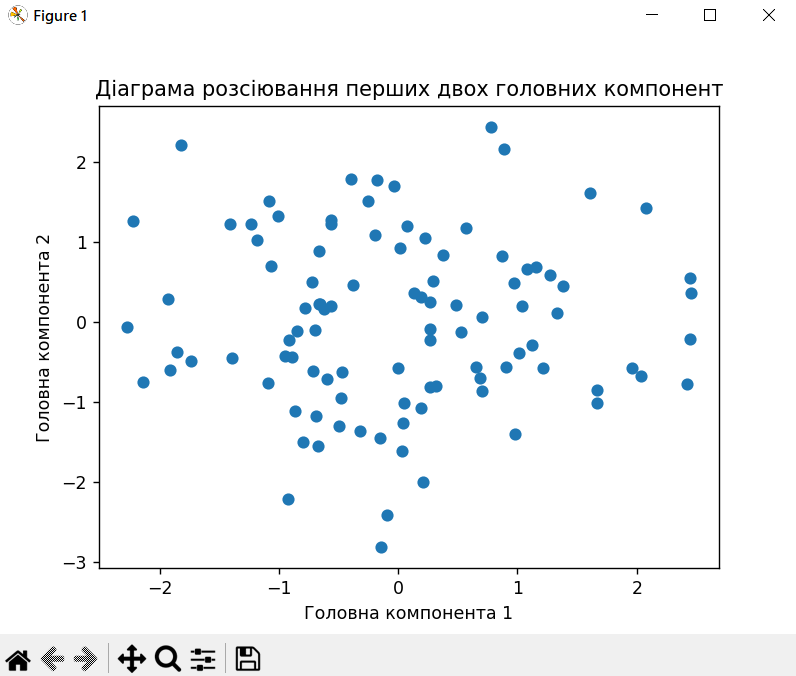
**Набір даних:**

Дані генеруються в програмі.

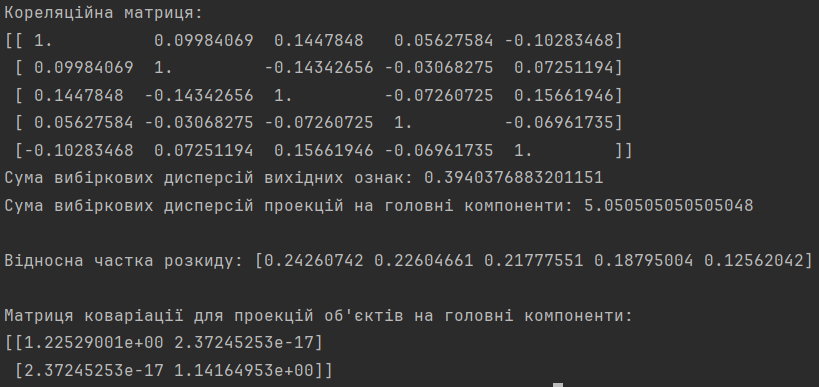
Запуск програми відбувається так:



Діаграма розсіювання перших двох головних компонент



Кореляційна матриця, cума вибіркових дисперсій вихідних ознак, cума вибіркових дисперсій проекцій на головні компоненти, відносна частка розкиду, матриця коваріації для проекцій об'єктів на головні компоненти:



**Код програми:**

import numpy as np  
import matplotlib.pyplot as plt  
  
def pca(X, n\_components):  
 # Стандартизувати дані  
 X\_mean = np.mean(X, axis=0)  
 X\_std = np.std(X, axis=0)  
 X\_std[X\_std == 0] = 1 # Щоб уникнути ділення на нуль  
 X = (X - X\_mean) / X\_std  
  
 # Обчислити коваріаційну матрицю  
 cov\_matrix = np.cov(X, rowvar=False)  
  
 # Обчислити власні значення та власні вектори  
 eigenvalues, eigenvectors = np.linalg.eig(cov\_matrix)  
  
 # Відсортувати власні значення та власні вектори в порядку зменшення власних значень  
 idx = np.argsort(eigenvalues)[::-1]  
 eigenvectors = eigenvectors[:, idx]  
 eigenvalues = eigenvalues[idx]  
  
 # Відібрати перші n\_components власних векторів  
 components = eigenvectors[:, :n\_components]  
  
 # Перетворити дані на новий простір  
 X\_pca = np.dot(X, components)  
  
 return X\_pca, components, eigenvalues  
  
# Приклад використання  
X = np.random.rand(100, 4) # Приклад згенерованих даних  
n\_components = 2  
X\_pca, components, eigenvalues = pca(X, n\_components)  
  
# Генерування експериментальних даних  
data = np.random.rand(100, 5) # Приклад згенерованих даних  
  
# Відображення даних на екрані монітора у вигляді таблиці  
print("Експериментальні дані:")  
print(data)  
  
# Стандартизація даних  
data\_mean = np.mean(data, axis=0)  
data\_std = np.std(data, axis=0)  
data\_std[data\_std == 0] = 1  
data\_normalized = (data - data\_mean) / data\_std  
  
# Побудова кореляційної матриці  
correlation\_matrix = np.corrcoef(data\_normalized, rowvar=False)  
print("\nКореляційна матриця:")  
print(correlation\_matrix)  
  
# Перевірка значимості кореляційної матриці  
# Наприклад, можна використовувати статистичний тест  
  
# Розрахунок проекцій об'єктів на головні компоненти  
X\_pca, components, eigenvalues = pca(data, n\_components=2)  
  
# Побудова матриці даних, рахунків, помилок та навантажень  
# Детальніше дивіться в реалізації PCA вище.  
  
# Перевірка рівності сум вибіркових дисперсій  
total\_variance\_original = np.sum(np.var(data, axis=0))  
total\_variance\_pca = np.sum(eigenvalues)  
print("Сума вибіркових дисперсій вихідних ознак:", total\_variance\_original)  
print("Сума вибіркових дисперсій проекцій на головні компоненти:", total\_variance\_pca)  
  
# Визначення відносної частки розкиду  
explained\_variance\_ratio = eigenvalues / np.sum(eigenvalues)  
print("\nВідносна частка розкиду:", explained\_variance\_ratio)  
  
# Побудова матриці коваріації для проекцій об'єктів на головні компоненти  
covariance\_matrix\_pca = np.cov(X\_pca, rowvar=False)  
print("\nМатриця коваріації для проекцій об'єктів на головні компоненти:")  
print(covariance\_matrix\_pca)  
  
# Побудова діаграми розсіювання  
plt.scatter(X\_pca[:, 0], X\_pca[:, 1])  
plt.xlabel('Головна компонента 1')  
plt.ylabel('Головна компонента 2')  
plt.title('Діаграма розсіювання перших двох головних компонент')  
plt.show()

**Контрольні питання:**

1. Для яких завдань обробки експериментальних даних використовується метод головних компонент? Метод головних компонент використовується для зменшення розмірності даних, виявлення структури в даних, зменшення шуму та візуалізації даних.
2. У чому полягає суть методу головних компонент? Метод головних компонент полягає в упорядкуванні ознак таким чином, щоб перші головні компоненти містили найбільшу дисперсію у даних.
3. Чому дорівнює математичне сподівання і дисперсія стандартизованої змінної? Математичне сподівання стандартизованої змінної дорівнює 0, а дисперсія дорівнює 1.
4. Яка дисперсія i-ї головної компоненти? Дисперсія i-ї головної компоненти дорівнює дисперсії даних, яка пояснюється цією головною компонентою.
5. Чому дорівнює сума вибіркових дисперсій проекцій об'єктів на головні компоненти? Сума вибіркових дисперсій проекцій об'єктів на головні компоненти дорівнює загальній дисперсії даних.
6. Яка відносна частка розкиду, яка припадає на j-у головну компоненту? Відносна частка розкиду, яка припадає на j-у головну компоненту, вимірюється відношенням дисперсії j-ої головної компоненти до загальної дисперсії.
7. Яка відносна частка розкиду, яка припадає на i перших головних компонент? Відносна частка розкиду, яка припадає на i перших головних компонент, вимірюється відношенням суми дисперсій перших i головних компонент до загальної дисперсії.
8. Чи доцільно використання методу головних компонент, якщо коваріаційна матриця вихідних ознак діагональна? Ні, використання методу головних компонент не доцільно, якщо коваріаційна матриця діагональна, оскільки це означає, що ознаки незалежні між собою, і в такому випадку метод головних компонент не допоможе в редукції розмірності даних.
9. Як перевірити значущість кореляційної матриці вихідних даних? Значущість кореляційної матриці вихідних даних можна перевірити за допомогою статистичних тестів, таких як тест Фішера.
10. Яка інтерпретація перших двох головних компонент? Перші дві головні компоненти часто використовуються для візуалізації даних і виявлення основних залежностей між об'єктами.
11. Який вид зв'язку між ознаками і головними компонентами? Головні компоненти є новими ортогональними ознаками, які представляють собою комбінації вихідних ознак.
12. Для яких цілей використовуються функції MATLAB: eig, fliplud, fliplr, corrcoef, cov, corr, scatter?
    * eig: використовується для обчислення власних значень та власних векторів матриці.
    * fliplud та fliplr: використовуються для відображення даних зліва направо та зверху вниз відповідно.
    * corrcoef: обчислює кореляційну матрицю між змінними.
    * cov: обчислює коваріаційну матрицю між змінними.
    * corr: обчислює коефіцієнт кореляції між змінними.
    * scatter: використовується для побудови графіку розсіювання для двох змінних.

**Висновок:**

Отже, в ході виконання даної лабораторної роботи я вивчав та застосовував метод головних компонент для аналізу експериментальних даних. Після ретельного аналізу отриманих результатів можна зробити кілька важливих висновків.

Спочатку, під час нормування даних та побудови кореляційної матриці я переконався, що вихідні ознаки корелюють між собою, що підтверджується значимим відхиленням кореляційної матриці від одиничної.

Далі, розрахувавши проекції об'єктів на головні компоненти, я визначив, що сума вибіркових дисперсій проекцій значно перевищує суму вибіркових дисперсій вихідних ознак, що свідчить про ефективність методу головних компонент у зменшенні розмірності даних та збереженні інформації.

Для додаткового аналізу я побудував матрицю коваріації для проекцій об'єктів на головні компоненти та діаграму розсіювання для перших двох головних компонент. Це дозволило отримати інтерпретацію цих компонент та зрозуміти, яку частину розкиду вони пояснюють у вихідних даних.

У цілому, виконання цієї лабораторної роботи надало мені можливість поглибити розуміння методу головних компонент та його застосування в аналізі даних, а також отримати практичний досвід реалізації цього методу на мові програмування Python.