МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

Кафедра комп’ютерної математики і аналізу даних

ЗВІТ

про виконання Лабораторної роботи №3 за темою

«Компонентний аналіз даних»
з дисципліни «Аналіз даних і часових рядів»

Група КН-122а

Виконавець Жарський Н.Д.

Викладач Гардер С.Є.

Харків 2025

Лабораторна робота №3

Мета роботи

Вивчити метод головних компонент (Principal Component Analysis, PCA) для зменшення розмірності даних та побудови регресійної моделі на виділених компонентах.

Теоретичні відомості

- Стандартизація даних перетворення початкових ознак так, щоб кожна мала середнє 0 і дисперсію 1
- Кореляційна матриця R матриця попарних кореляцій між стандартизованими ознаками
- Власний розклад R=UΛU^T:
  - $\lambda^1 \geq \lambda^2 \geq \cdots \geq \lambda_k$ - власні значення (розподіл дисперсії)
  - стовпці U — власні вектори (напрямки головних компонент)
- Факторні навантаження $A = U\sqrt{\Lambda}$ характеризують, як початкові змінні «показані» в простір головних компонент
- **Вибір числа компонент**: за правилом «80 %» або «90 %» накопиченої дисперсії.
- **Регресія на ГК**: побудова моделі $Y = \beta 0 + \beta 1F1 + \cdots + \beta pFp + \varepsilon$

  де Fj — значення j-ої головної компоненти

Опис даних

- Кількість спостережень n=20
- Кількість початкових ознак k=5
- Вхідні змінні $X \in R^{20\times5}$
- Відгук $Y \in R^{20}$

Методика розв'язання

1. Стандартизація:

$Z = \frac{X-\overline{X}}{s_X}$, $\overline{X}$=середнє, s_X= стандартне відхилення

2. Кореляційна матриця:

$$R = \frac{1}{n-1} Z^T Z$$

3. Власний розклад:

$$R = U\Lambda U^T,\ \Lambda = diag(\lambda_1, \dots, \lambda_k)$$

4. Ранжування власних значень за спаданням
5. Факторні навантаження $A = U\sqrt{\Lambda}$
6. Частка поясненої дисперсії

ГК	Власне значення λ	% поясненої дисперсії	Накопичена %
1	2.0416	60.83 %	60.83 %
2	0.9454	28.16 %	88.99 %
3	0.2829	8.42 %	97.41 %
4	0.0749	2.23 %	99.64 %
5	0.0112	0.33 %	100.00 %

7. Побудова регресії на перших p ГК (обраних за накопиченою дисперсією)
8. Оцінка моделі R^2 , F-статистика, t-тести коефіцієнтів

Результати

1. Стандартизовані дані Z

Z =

[[-0.1383, -0.6338, -0.0511, -0.4523, -0.6710],

[-0.7264, -0.1967, -0.6383, 0.3870, 2.0027],

[0.3039, 0.4589, 0.1003, -0.5145, 0.2031],

[1.2918, 2.4259, 1.5700, -0.1104, 1.6428],

[0.1298, -0.6338, -0.0105, -0.2347, -0.5682],

[0.1298, 0.2404, 0.1852, -0.4523, -0.5167],

[-0.5664, 0.0219, -0.5386, -0.1414, -0.2082],

[-0.7217, -0.6338, -0.6531, -0.7942, -0.9795],

[-0.6417, -1.0709, -0.6272, -0.8253, -0.9795],

[0.7226, 0.4589, 0.3071, 2.8116, 2.3626],

[-0.0489, 0.2404, 0.3698, 0.8222, -0.5167],

[0.2427, 0.6775, 0.4105, -0.6699, -0.6710],

[3.5170, 2.4259, 3.4422, -0.2347, 0.5630],

[-0.0772, -0.1967, 0.0264, 1.1020, -0.5167],

[-0.6087, 0.0219, -0.5866, -1.0429, 0.4088],

[-0.7546, -0.6338, -0.6974, -1.1672, -0.6196],

[-0.1148, 0.0219, -0.2764, -0.8253, -0.9795],

[-0.8440, -1.5080, -0.7897, -0.1104, -0.3625],

[-0.8487, -0.8523, -0.7971, 0.8222, -0.3625],

[-0.2465, -0.6338, -0.7454, 1.6304, 0.7687]]

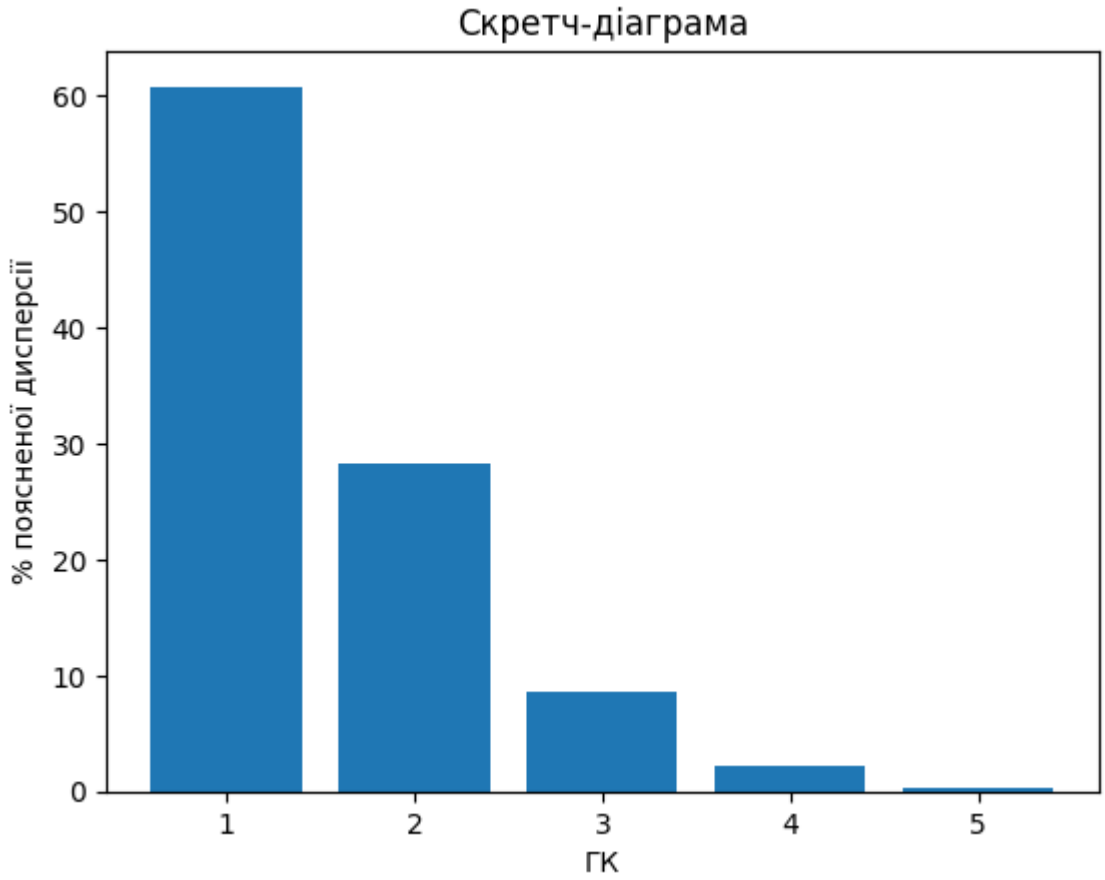
2. Кореляційна матриця R

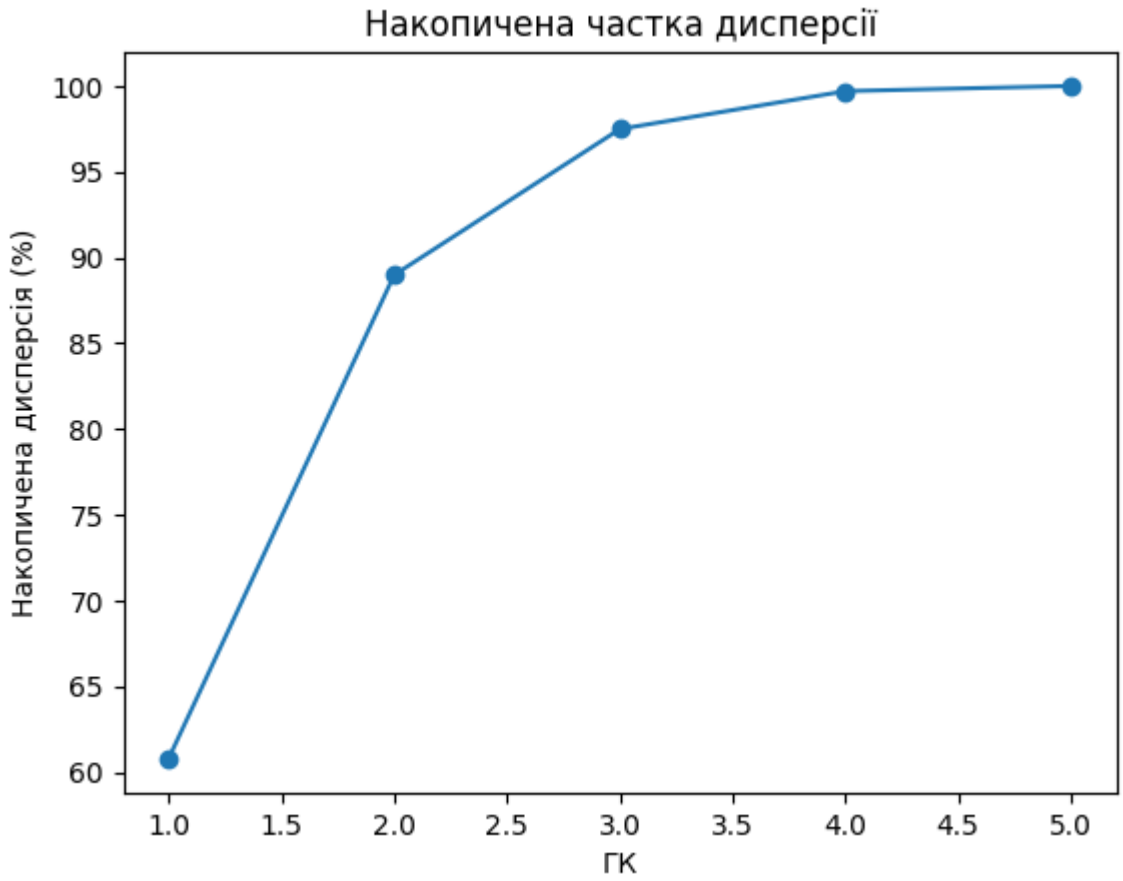
$$R = \begin{pmatrix} 1 & 0.8543 & 0.9779 & 0.1104 & 0.3410 \\ 0.8543 & 1 & 0.8819 & 0.0269 & 0.4596 \\ 0.9779 & 0.8819 & 1 & 0.1319 & 0.3187 \\ 0.1104 & 0.0269 & 0.1319 & 1 & 0.3278 \\ 0.3410 & 0.4596 & 0.3187 & 0.3278 & 1 \end{pmatrix}$$

3. Власні значення та частки дисперсії

ГК	**λ**	**% поясненої дисперсії**	**Накопичена %**
1	2.0416	60.83 %	60.83 %
2	0.9454	28.16 %	88.99 %
3	0.2829	8.42 %	97.41 %
4	0.0749	2.23 %	99.64 %
5	0.0112	0.33 %	100.00 %

Графіки:


Скретч-діаграма
% поясненої дисперсії
60
50
40
30
20
10
0
1
2
3
4
5
ГК



Обираємо перші дві компоненти (накоп. 88.99 % ≈ 90 %)

4. Факторні навантаження A

A =

[[-0.9506, -0.1931, 0.1776, 0.1459, -0.0787],

[-0.9357, -0.1651, -0.1753, -0.2571, -0.0186],

[-0.9411, -0.2794, 0.1543, 0.0626, 0.0921],

[-0.2355, 0.8850, 0.3929, -0.0837, 0.0012],

[-0.5642, 0.6960, -0.4299, 0.1110, 0.0092]]

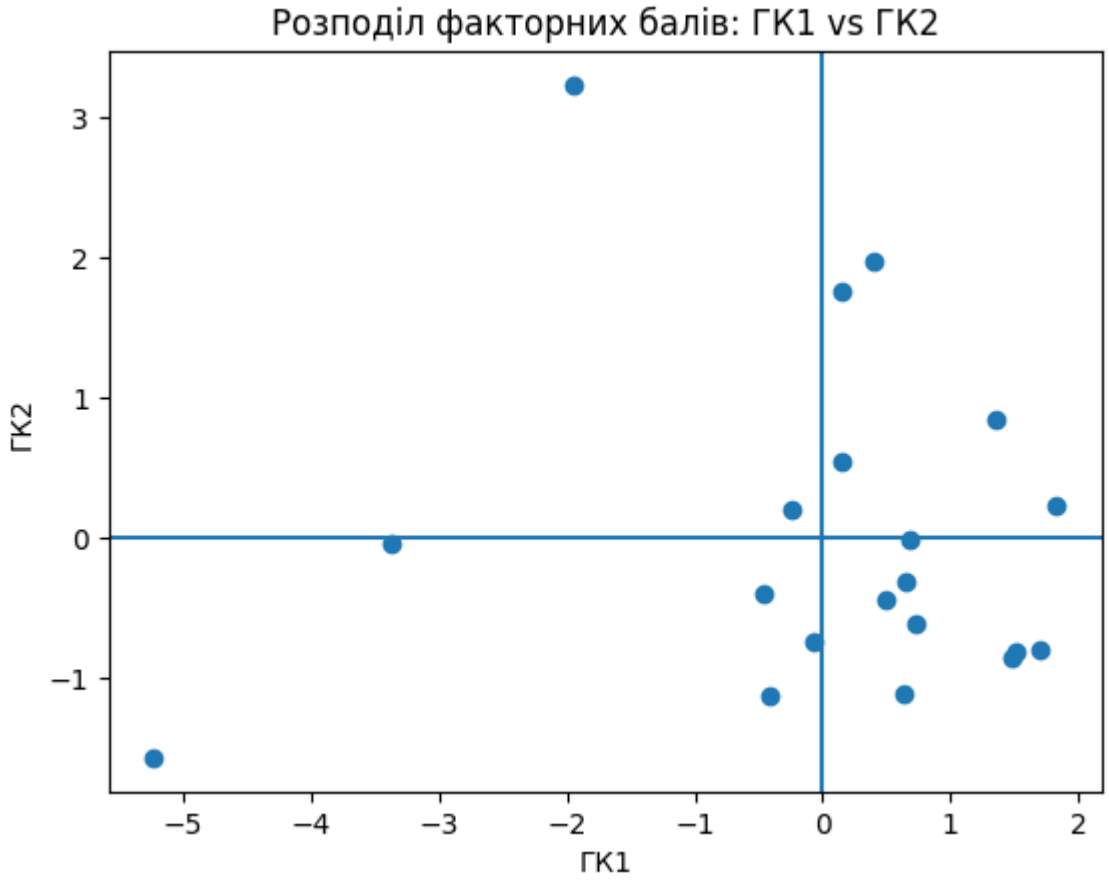
5. Регресія на ГК1 та ГК2

Оціночні коефіцієнти (інтерцепт + дві компоненти):

$$\hat{\beta} = [9.5250, -0.5452, 0.5716]^T$$

- $R^2 = 0.5173$

- F-статистика = 3.0008, Fcrit (1,17;α=0.05)=4.45 → модель помірно значуща
- Детальні t-тести:
    - Інтерцепт: t=26.64, p<0.001
    - ГК1: t=−2.59, p=0.021
    - ГК2: t=1.85, p=0.085



Опис результатів

- Перша головна компонента пояснює більшість дисперсії (> 60 %)
- Для моделі обрано дві компоненти: дають близько 90 % поясненої дисперсії
- У регресії ГК1 є статистично значущою при 5 % рівні, ГК2 — лише на 10 % рівні
- Отримане $R^2 \approx 0.52$ вказує на помірну якість апроксимації

Висновки

1. Метод головних компонент ефективно зменшує розмірність, зберігаючи основну інформацію

2. Обрані дві компоненти дозволяють пояснити близько 90 % дисперсії початкових ознак
3. Регресійна модель на ГК1–ГК2 показала статистичну значущість і може використовуватися для прогнозування Y
4. Для підвищення точності моделі слід оцінити можливість включення додаткових компонент або нелінійних перетворень