Власні вектори та власні числа в лінійній алгебрі та статистиці за допомогою функцій MS Excel

О. В. Мулик[1]

[1]*КПІ ім. Ігоря Сікорського, Київ, Україна*

mulyk.olena@gmail.com

Анотація

Розглянуто зміст власних векторів і власних чисел для методу головних компонент через побудову канонічного рівняння еліпса методом повороту системи координат та розподілу вибіркових даних вздовж певного напрямку. Наведено приклад знаходження матриці Грама та можливість скорочення простору факторів відповідно до її власних значень.

Ключові слова: власні вектори, власні числа, матриця Грамма.

Прикладні статистичні методи, такі як метод головних компонент (PCA – Principal Component Analysis), методи факторизації матриць (NMF – Non-negative matrix factorization, ICA – Independent component analysis) базуються на різних концепціях лінійної алгебри таких, як власні числа, власні вектори, ортогональні матриці, що детально вивчаються в курсі лінійної алгебри, а також, спектральній теоремі (Spectral Decomposition Theorem), на яку не припадає великої уваги за браком часу. Більш загальний метод SVD (Singular Value Decomposition), матриця Грамма та інші методи факторизації матриць не розглядається в базовому курсі, але уявлення про можливості їх застосування є бажаною для розуміння популярних статистичних методів скорочення простору ознак без втрати інформативності. Процес розкладу матриці на добуток кількох простіших матриць, які містять корисну інформацію про вихідні дані використовується для спрощення обчислень, зменшення розмірності даних, знаходження прихованих закономірностей та виявлення найбільш вагомих факторів, що впливають на результат тощо. Аналіз головних компонент (PCA) є складовою частиною факторного аналізу і, загалом існує доволі простий спосіб показати студентам на прикладі застосування методу факторизації матриць (матриці Грамма), геометричний зміст власних векторів та власних чисел.

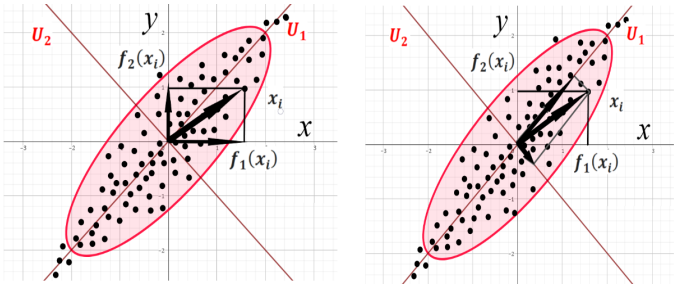

Рис. 1: Проєкція значення x_i вибіркових даних.

Розглянемо експериментальні вибіркові дані у вигляді матриці $F_{n\times 2}$, яка містить 2n елементів у двовимірному просторі ознак. Кожен стовпчик матриці $F_{n\times 2}$ описується функціями $f_1(x_i)$ та $f_2(x_i)$, $i = 1, ...n$ відповідно і які задають певні значення $f_1(x_i) = x_{i1}$ та $f_2(x_{i2})$. Хай вибіркові дані розподілені в напрямку осі Ox_1 під кутом 45^0 та нагадаємо знаходження канонічного рівняння еліпса:

$$F = \begin{pmatrix} f_1(x_1) & f_2(x_1) \\ f_1(x_2) & f_1(x_2) \\ \dots & \dots \\ f_1(x_n) & f_2(x_n) \end{pmatrix}, \quad 5x^2 + 5y^2 - 8xy - 8 = 0.$$

Знайдемо канонічне рівняння еліпса за відомим алгоритмом – повернемо систему координат на кут 45^0 та отримаємо канонічну систему координат. Власні числа матриці $\begin{pmatrix} 5 & -4 \\ -4 & 5 \end{pmatrix}$ є $L_1 = 1$, $L_2 = 9$ та відповідні орти власних векторів:

$V_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $V_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$, [Булдигін та ін., 20011].

Тоді шукане перетворення координат задає матриця H:

$H = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$, тоді $\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$ і $\begin{cases} x = \frac{1}{\sqrt{2}}x_1 + \frac{1}{\sqrt{2}}y_1 \\ y = \frac{1}{\sqrt{2}}x_1 - \frac{1}{\sqrt{2}}y_1 \end{cases}$

У нових координатах рівняння еліпса набуде вигляду $\frac{x^2}{8} + \frac{y^2}{0,(8)}$, де можна побачити, яка піввісь буде більшою за довжиною, в цьому випадку – це вісь OU_1. Тепер повернемось до вибіркових даних. Виберемо довільну точку x_i, як показано на малюнку та розглянемо вектор $\vec{V_i} = (f_1(x_i), f_2(x_i))$ з початком в точці O(0,0). Проєкції вектора $\vec{V_i}$ на осі Ox та Oy не однакові, але, в середньому, відрізняються не на значну величину і мають зміст відхилення (дисперсії) від осей Ox та Oy. Якщо розглянути проєкції цього ж вектора $\vec{V_i}$ на нові осі OU_1 та OU_2, то нерівність проєкцій стає значною і, зрозуміло, що, в середньому, проєкція на вісь OU_1 значно більша за проєкцію на OU_2. Тобто відхилення для експериментальних даних

(дисперсія) стає більш проявленим в певному напрямку. Позначимо нові координати для вектора експериментальних даних $\vec{V_i} = (g_1(x_i), g_2(x_i))$. Важливим фактом стає більша концентрація інформації на осі OU_1 з ознакою $g_1(x_i)$ і тепер кожен вектор $\vec{V_i} = (g_1(x_i), g_2(x_i))$. можна обчислювати за такими самими правилами, як і рівняння еліпса:

$$\begin{pmatrix} g_1(x_i) \\ g_2(x_i) \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} f_1(x_i) \\ f_2(x_i) \end{pmatrix},$$

Тобто маємо функціональні перетворення у вигляді скалярного добутку двох векторів які визначили проєкції вектора $\vec{V_i}$ на нові осі.

Якщо кількість ознак збільшується – простір ознак стає багатовимірним і матриця F буде розміру nxm, де рядки – це спостереження, а стовпці – ознаки. Тоді для знаходження власних чисел використовується матриця Грамма (Gram matrix), що визначається як:

$$G_n = (F_{nm})^T F_{mn} = \begin{pmatrix} f_1(x_1) & \dots & f_1(x_m) \\ \dots & \dots & \dots \\ f_n(x_1) & \dots & f_n(x_m) \end{pmatrix} \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_m) & \dots & f_n(x_m) \end{pmatrix} \text{[Zaiontz, 2023]}$$

G_n – це симетрична, додатно напіввизначена матриця (її власні значення невід'ємні), вона містить скалярні добутки всіх пар стовпців матриці F_{nm}. Фактично, це коваріаційна матриця (до нормалізації), яка показує, як змінні залежать одна від одної. Далі потрібно знаходити власні числа і власні вектори матриці G_n за відомим алгоритмом $det(G_n - LE_n) = 0$, тоді отримані значення власних чисел L_i будуть визначати відхилення (дисперсію), тобто варіацію проєкцій значень матриці F_{mn} на відповідну вісь OU_i ортонормованої системи координат. Схематично це можна зобразити так:

$$\begin{array}{cccc} L_1 \geqslant & L_2 \geqslant & \dots & \geqslant L_n \\ \downarrow & \downarrow & \dots & \downarrow \\ U_1 & U_2 & \dots & U_n \end{array}$$

Тобто, якщо розташувати власні значення L_i в порядку спадання, то їх значення по величині покажуть, наскільки важлива кожна компонента.

Звісно, обчислювати власні числа і власні вектори для великих масивів даних потрібно у певних середовищах. Реалізуємо практичну задачу функціями MS Excel. Сформуємо за допомогою Data Analysis (Random Number Generation) вибірку даних, що складається з двох незалежних ознак (факторів) $f_1(x_i)$ та $f_2(x_i)$, $i = 1, \dots, 10$, які є Гаусовими випадковими величинами з середнім 0 і дисперсією 1 та лінійною комбінацією $f_3(x_i) = \frac{f_1(x_i)+f_2(x_i)}{2}$. Для знаходження матриці Грамма використовуємо формулу `MMULT(TRANSPOSE(B3:D12);B3:D12)`, далі для знаходження власних чисел і власних векторів `eigVECTSym(F3:H5)` (Рис.2).

Запропонований простий приклад допоможе студентам першого курсу розширити погляд на застосування лінійної алгебри в їх майбутніх дослідженнях.

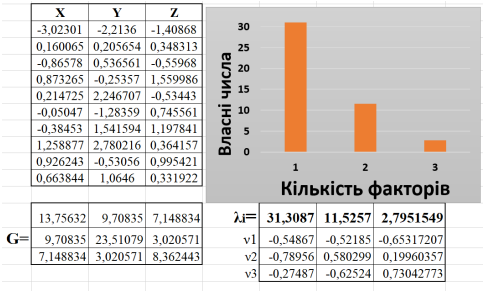

Рис. 2: Знаходження матриці Грамма та власних векторів та власних чисел функціями MS Excel.

Перелік посилань

Charles Zaiontz (2023). *The data analysis for this paper was generated using the Real Statistics Resource Pack software (Release 8.9.1). Copyright (2013–2023).*

Булдигін В. В., Алєксєєва І. В., Гайдей В. О., Диховичний О. О., Коновалова Н. Р., Федорова Л. Б. (2011). *Лінійна алгебра та аналітична геометрія. Київ:ТВіМС.*