1.1 Метод головних компонент (Principal Component Analysis – PCA)

В аналізі даних, як і в будь-якому іншому аналізі, часом буває незайвим створити спрощену модель, що максимально точно описує реальний стан справ. Часто буває так, що ознаки досить сильно залежать одна від одної та їх одночасна наявність є надмірною. Якщо одна ознака строго залежить від іншої, то знаючи одну, ми завжди знаємо й іншу. Але набагато частіше буває так, що ознаки залежать одна від одної не так строго і (що важливо!) не так очевидно. Але як з'ясувати, який саме набір параметрів добре описує наш набір даних, але при цьому має невелику надмірність? Іншими словами, як зменшити розмірність простору, в якому "живуть" дані, втративши при цьому мінімум інформації?

Способи вирішення цього завдання називаються методами зменшення розмірності (dimensionality reduction). Метод головних компонент (principal components analysis, PCA) – один з них. Мета PCA – вилучення з цих даних потрібної інформації. Що є інформацією, залежить від суті завдання, що розв'язується. Дані можуть містити потрібну нам інформацію, вони можуть бути надлишковими. Однак, у деяких випадках, інформації в даних може не бути зовсім. Розмірність даних — кількість зразків та змінних -– має велике значення для успішного видобутку інформації. Зайвих даних немає — краще, коли їх багато, ніж мало.

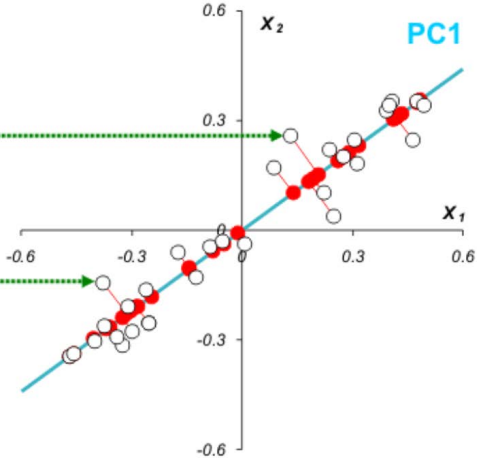
Дані завжди (або майже завжди) містять у собі небажану складову, яка називається шумом. Природа цього шуму може бути різною, але, у багатьох випадках, шум – це та частина даних, яка не містить інформації, що шукається. Що вважати шумом, а що інформацією, завжди вирішується з урахуванням поставлених цілей та методів, які використовуються для її досягнення.

Шум і надмірність у даних обов'язково виявляють себе через кореляційні зв'язки між змінними. Похибки даних можуть призвести до появи не систематичних, а випадкових зв'язків між змінними. Поняття ефективного рангу та прихованих, латентних змінних, кількість яких дорівнює цьому рангу, є найважливішим поняттям у PCA.

1.1.1 Інтуїтивний підхід

Розглянемо простіший випадок, коли об'єкти описуються лише двома ознаками: x_1 та x_2. Такі дані легко зобразити на площині (див. рисунок).

	X_1	X_2
1	0.407	0.353
2	0.475	0.355
3	-0.088	-0.045
4	0.394	0.325
5	0.274	0.202
6	0.131	0.258
7	-0.053	-0.031
8	-0.124	-0.128
9	-0.469	-0.344
10	0.088	0.171
11	-0.261	-0.162
12	0.401	0.341
13	-0.376	-0.143
14	-0.251	-0.255
15	-0.325	-0.316
16	0.464	0.248
17	-0.310	-0.207
18	0.307	0.247
19	-0.399	-0.303
20	-0.253	-0.253
21	-0.341	-0.291



Кожному рядку вихідної таблиці (тобто зразку) відповідає точка на площині з відповідними координатами. Вони позначені порожніми кружками. Проведемо через них пряму, так, щоб уздовж неї відбувалася максимальна зміна даних. На рисунку ця пряма виділена блакитним кольором; вона називається першою головною компонентою PC_1. Потім проектуємо всі вихідні точки на цю вісь (червоні кружки). Якщо дані описані не повністю (шум великий), то вибирається ще один напрямок (PC_2) – перпендикулярне до першого, так щоб описати зміну, що залишилася в даних і т.д. Таким чином, знаючи залежності та їх силу, ми можемо виразити кілька ознак через одну, злити докупи, так би мовити, і працювати вже з більш простою моделлю. Звичайно, уникнути втрат інформації, швидше за все, не вдасться, але мінімізувати її нам допоможе якраз метод PCA.

Отже, даний метод апроксимує n-розмірну хмару спостережень до еліпсоїда (теж n-вимірного), півосі якого і будуть майбутніми головними компонентами. І при проекції даних на такі осі (зниженні розмірності) зберігається найбільше інформації.

1.1.2 Зменшення розмірності даних

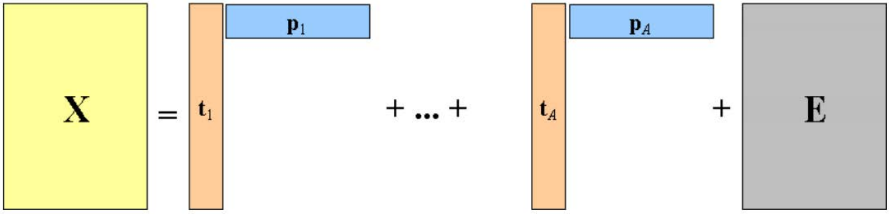
Метод основних компонентів застосовується до даних, записаним у вигляді матриці X -– прямокутної таблиці чисел розмірністю I рядків і J стовпців. Традиційно рядки цієї матриці називаються об'єктами. Вони нумеруються індексом $i = 1, \ldots, I$. Стовпці називаються ознаками і вони нумеруються індексом $j = 1, \ldots, J$. У методі головних компонент використовуються нові, формальні змінні t_a $(a = 1, \ldots, A)$, що є лінійною комбінацією вихідних змінних x_j $(j = 1, \ldots, J)$:

$$t_a = p_{a1}x_1 + p_{a2}x_2 + \cdots + p_{aJ}x_J$$

За допомогою цих нових змінних матриця X розкладається у добуток двох матриць T та P:

$$X = TP^T + E = \sum_{a=1}^{A} t_a p_a^T + E$$

Матриця T називається матрицею рахунків (scores). Її розмірність $(I \times A)$. Матриця P називається матрицею навантажень (loadings). Її розмірність $(J \times A)$. E – матриця залишків, розмірністю $(I \times J)$.



Нові змінні t_a називаються головними компонентами (Principal Components), тому сам метод називається методом головних компонент (PCA). Число стовпців t_a в матриці T, і p_a в матриці P дорівнює A, яке називається числом головних компонент (PC). Ця величина свідомо менша від числа змінних J і числа зразків I.

Важливою властивістю PCA є ортогональність (незалежність) основних компонент. Тому матриця рахунків T не перебудовується зі збільшенням числа компонент, а до неї просто додається ще один стовпець – відповідний новому напрямку. Теж відбувається і з матрицею навантажень P.

1.1.3 Статистичні основи

Для опису випадкової величини використовуються моменти. Потрібні нам – математичне очікування та дисперсія. Можна сміливо сказати, що математичне очікування це "центр тяжіння" величини, а дисперсія – її "розміри" (розкид). Сам процес проектування на вектор ніяк не впливає на значення середніх, тому що для мінімізації втрат інформації наш вектор має проходити через центр нашої вибірки. Тому зазвичай проводять центрування вибірки -– лінійно зміщуючи її так, щоб середні значення ознак дорівнювали 0. Оператор, зворотний зсуву дорівнюватиме вектору початкових середніх значень — він знадобиться відновлення вибірки у вихідній розмірності. При цьому, дисперсія залежить від порядків значень випадкової величини, тобто, є чутливою до масштабування. Тому якщо одиниці виміру ознак сильно відрізняються своїми порядками, рекомендується стандартизувати їх за формулою

$$x_i = \frac{x_i - \langle x_i \rangle}{\sigma_i},$$

де $\langle x_i \rangle$ середнє значення ознаки x_i, σ_i – її дисперсія.

Для опису форми випадкового вектора необхідна матриця коваріації. Це матриця, у якої елемент (i, j) є кореляцією ознак (x_i, x_j). Формула коваріації має вигляд:

$$\begin{aligned} Cov(x_i, x_j) =& E\left[(x_i - E(x_i)) \cdot (x_j - E(x_j))\right] \\ & E(x_i, x_j) - E(x_i) - E(x_j), \end{aligned} \tag{1.1}$$

де $E(\dots)$ – середнє від вектору. У нашому випадку вона спрощується, тому що $E(x_i) = E(x_j) = 0$ в силу проведеного нормування:

$$Cov(x_i, x_j) = E(x_i, x_j). \tag{1.2}$$

Зауважимо, що коли $x_i = x_j$:

$$Cov(x_i, x_i) = Var(x_i). \tag{1.3}$$

і це справедливо для будь-яких випадкових величин.

Таким чином, у коваріаційній матриці по діагоналі будуть дисперсії ознак (оскільки $i = j$), а в інших комірках — коваріації відповідних пар ознак. А в силу симетричності коваріації матриця теж буде симетрична. Коваріаційна матриця є узагальненням дисперсії на випадок багатовимірних випадкових величин вона так само описує форму (розкид) випадкової величини, як і дисперсія, яка для одновимірної випадкової величини має вигляд матриці розміру 1×1, в якій її єдиний член заданий формулою $Cov(x, x) = Var(x)$. Коваріаційна матриця описує форму нашої випадкової величини (ознаки). Узагальнення дисперсії на вищі розмірності – матриця коваріації, і ці два поняття еквівалентні. При проекції на вектор максимізується дисперсія проекції, при проекції на простір великих порядків – вся її коваріаційна матриця.

1.1.4 Власні значення та власні вектори

Тепер треба знайти такі вектори, при якому б максимізувався розкид (дисперсія) проекції вибірки на них. Слід зазначити, що узагальнення дисперсії на вищі розмірності – матриця коваріацій, і ці два поняття еквівалентні. При проекції на вектор максимізується дисперсія проекції, при проекції простору великих порядків – вся її коваріаційна матриця.

Отже, візьмемо одиничний вектор на який проектуватимемо наш випадковий вектор X. Тоді проекція на нього дорівнюватиме $v^T X$ (v^T – транспонування вектора v). Дисперсія проекції на вектор відповідно дорівнює $Var(v^T X)$. Загалом у векторній формі (для центрованих величин) дисперсія виражається так:

$$Var(X) = \Sigma = E(X \cdot X^T)$$

Відповідно, дисперсія проекції:

$$\begin{aligned} Var(X^*) = \Sigma^* = & E(X^* \cdot X^{*T}) = E\left((v^T X) \cdot (V^T X)^T\right) = \\ & E\left((v^T X \cdot X^T v\right) = v^T E\left(X \cdot X^T\right) v = v^T \Sigma v \end{aligned} \tag{1.4}$$

Таким чином, що дисперсія максимізується за умови максимального значення $v^T \Sigma v$. Далі зручно використати співвідношення Релея,

які для коваріаційних матриць M мають спеціальний випадок:

$$R(M, x) = \frac{X^T M X}{X^T X} = \lambda \frac{X^T X}{X^T X} \lambda \tag{1.5}$$

і відповідно маємо

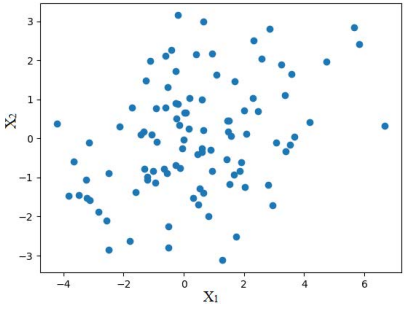
$$MX = \lambda X \tag{1.6}$$

Остання формула є розкладанням матриці на власні вектори та значення: X є власним вектором, а λ – власним значенням. Кількість власних векторів та значень дорівнюють розміру матриці (і значення можуть повторюватися).

Таким чином, приходимо до висновку, що напрямок максимальної дисперсії у проекції завжди збігається з власним вектором, що має максимальне власне значення, яке дорівнює величині цієї дисперсії. Це справедливо також для проекцій на більшу кількість вимірювань – дисперсія (коваріаційна матриця) проекції на m-вимірний простір буде максимальною у напрямку m власних векторів, що мають максимальні власні значення.

1.1.5 Зменшення розмірності дво-вимірної вибірки

Розглянемо простий приклад зменшення розмірності дво-вимірної вибірки, яку подано на рисунку.



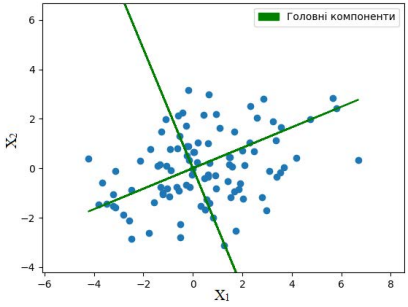
Видно, що дисперсія є максимальною у напрямку зростання (спадання) обох величин X_1 та X_2. Це і буде перша головна компонента PC_1, яка буде подаватися лінійною комбінацією X_1 та X_2. Відповідно друга головна компонента буде мати напрямок перпендикулярний по першої.

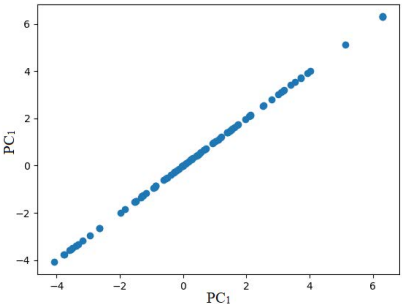
Для встановлення головних компонент розраховуємо матрицю коваріацій та знаходимо її власні значення та власні вектори. Власні вектори будуть вказувати напрямок зміни дисперсій нашої вибірки

– новий базис. Для нашої вибірки маємо:

$$\lambda = \begin{pmatrix} 4.41 \\ 1.58 \end{pmatrix} \qquad v = \begin{pmatrix} 0.92 & -0.38 \\ 0.38 & 0.92 \end{pmatrix} \tag{1.7}$$

Отже, перший вектор з максимальних власним значення визначає напрямок першої головної компоненти, другий – другої. Вибірку з власними векторами коваріаційної матриці наведено на рисунку зліва. Для проведення проекції, треба провести операцію $v^T X$ (вектор повинен бути довжини 1). Або якщо у нас не один вектор, а гіперплощина, то замість вектора v^T беремо матрицю базисних векторів V^T. Отриманий вектор (або матриця) буде масивом проекцій спостережень. Для нашої вибірки проекція на перший власний вектор з максимальним власним значенням наведено на рисунку справа.





Часто виникає потреба оцінити обсяг втраченої (і збереженої) інформації. Найзручніше це подавати у відсотках. Для цього використовуємо дисперсії по кожній осі і ділимо на загальну суму дисперсій по осях (тобто суму всіх власних чисел коваріаційної матриці). Таким чином, перший вектор вектор описує $(4.41/5.99) \times 100\% \simeq 74\%$, а менший, відповідно, приблизно 26%. Отже спроектувавши дані на перший вектор ми втратимо приблизно 26% інформації, проте зменшимо кількість даних для подальшого аналізу.

1.1.6 Постановка задачі

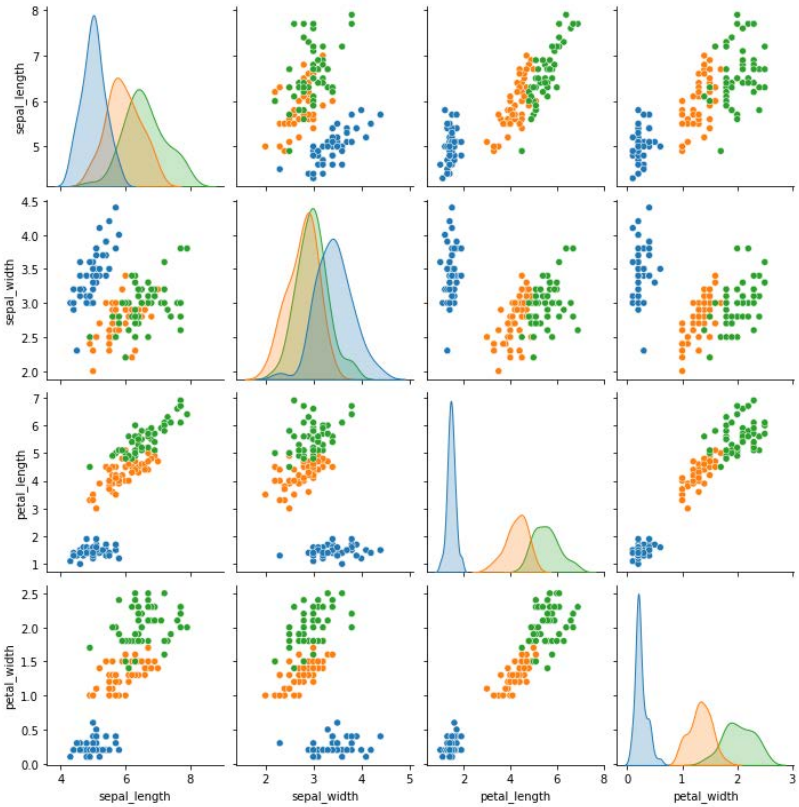
Провести реалізацію алгоритму PCA для зменшення розмірності вибірки до двох. Етапи розв'язання

1. Імпортувати дані з кількістю ознак більше трьох.
2. Побудувати матрицю коваріацій.
3. Розрахувати власні вектори та власні значення матриці
4. Відсортувати власні вектори в порядку спадання власних значень
5. Візуалізувати власні значення
6. Спроектувати дані на два перші власні вектори
7. Візуалізувати отриману дво-вимірну вибірку
8. Розрахувати принципові компоненти за допомогою вбудованої функції PCA з кількістю компонентів 2 з бібліотеки sklearn.
9. Порівняти результати роботи власного алгоритму з результатами функції PCA з бібліотеки sklearn
10. Візуалізувати різницю в отриманих вибірках
11. Оформити результати у вигляді звіту.

1.1.7 Приклад подання результатів

Розглянемо вибірку *iris* з бази даних sklearn.

1. Побудуємо парні залежності ознак, що представлено на рисунку.



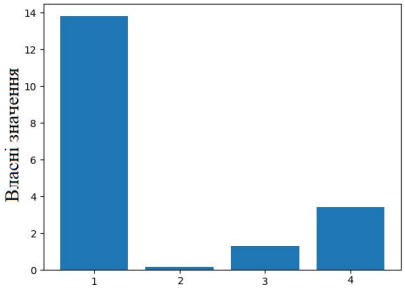
Оскільки перший клас добре виділяється від інших двох залишимо у вибірці лише дані для другого та третього класів. Отримуємо вибірку $X = [\ell \times n]$, де ℓ – кількість об'єктів, n – кількість ознак (для даного прикладу $n = 4$)

2. Проводимо нормалізацію даних: центрування та зведення до

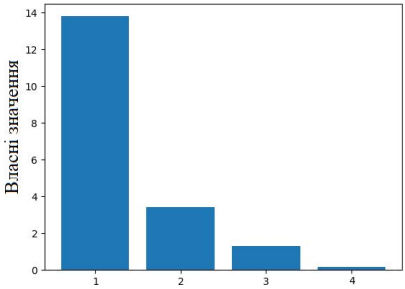
однакового розкиду (дисперсії):

$$X_i = (X_i - \langle X_i \rangle)/\sigma$$

3. Розрахунок матриці коваріацій та її власних векторів і власних значень

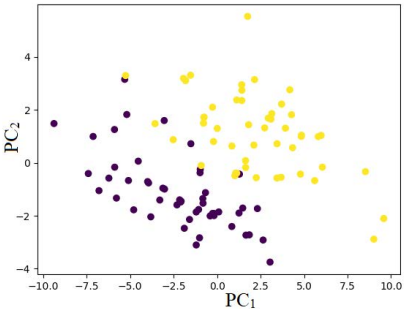


Розраховуємо матрицю коваріацій; розраховуємо власні вектори та власні значення. Візуалізуємо отримані власні значення.



Оскільки власні вектори не відсортовані за спаданням, проводимо їх сортування і відповідно сортуємо власні вектори. Візуалізуємо відсортовані власні значення.

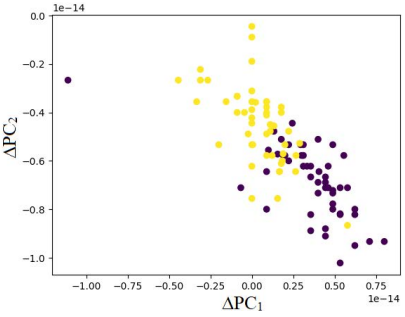
4. Встановимо кількість головних компонент $n_{PC} = 2$. Проектуємо дані на перші два власні вектори матриці коваріації, що характеризуються найбільшими власними значеннями



Множимо вихідну матрицю даних розміру $[\ell \times n]$ на транспоновану матрицю двох власних векторів розміру $[n \times n_{PC}]$ для отримання нової матриці даних розміром $[\ell \times n_{PC}]$. Візуалізуємо отримані дані.

5. Використовуючи вбудовану функцію PCA з пакету sclearn проводимо зменшення розмірності вибірки X до n_{PC} компонент.

6. Порівняння результатів



Розраховуємо для кожної з головних компонент PC_i, $i = 1, 2$ різницю між отриманими даними ΔPC_i за власним алгоритмом та вбудованою функцією PCA. Візуалізуємо отримані Результати.

7. Розраховуємо кількість втраченої інформації використовуючи власні значення матриці коваріацій при зменшенні розмірності з чотирьох ознак до двох:

$$\Delta I = 1 - \frac{\lambda_1 + \lambda_2}{\sum\limits_{i=1}^{n} \lambda_i} \times 100\% \simeq 7.9\%$$

Таким чином, проведена процедура зменшення розмірності дозволила зменшити кількість даних у два рази при цьому втрати інформації склали менше 8%.