主成分分析的原理與實驗

汪群超

January 31, 2023

觀察兩個變數間的相關性，可以畫散佈圖。觀察三個變數的相關性，也可以畫出 3-D 的立體圖來觀察。但是對於三個以上的變數，在視覺上便無從觀察起，即便是要計算變數間的相關係數，也顯得繁複許多。事實上，變數一多，就可能發生某些變數間其實存在著相依性，或是某些變數的影響程度非常微小，但在一般的應用上，往往因為人為的直覺判斷，造成挑選出過多的變數。多變量分析提供許多工具，試圖化繁為簡，降低變數的個數，並能抽離出真正的核心資訊，其中「主成分分析」極具代表性。在主成分分析的過程中，許多統計學與線性代數的基本觀念再度被應用到，這個單元要從這些基本觀念開始。

本章將學到關於程式設計：

⟨本章關於 Python 的指令與語法⟩

套件與指令：

scipy.linalg: `eig,pinv`

sklearn: `decompisition.PCA, preprocessing.StandardScaler`

1 背景介紹

1.1 從一個「評量表」說起

我們常常可以讀到有關城市評比的資料。譬如，舊金山是美國「生活品質」最好的城市、香港是亞洲生活費最高的城市。通常主辦單位會為評比的項目做一些定義，然後根據定義一一去評分。下列是一份針對全美三百多個城市做「生活品質」調查的 9 項評量項目（[1]），

> Climate / Housing Health / Crime / Transportation / education / Atrs / Research / Econmics

這些調查項目有些只要簡單的數據即可評分，有些或許需要經過比較嚴謹複雜的程序才能得到。可想而知這樣的調查工作所需的人力、物力及時間消耗甚鉅。但另一方面，從調查的項目來看，有些項目之間似乎存在『相關性』。這些相關性會讓所量測到的資料充斥著多餘的訊息 (Redundant information)。不過，調查項目在選擇之初通常只是表面上的認知，或不容易發現彼此間的關係，這些關係往往要透過相關性的分析才會出現。

「主成分分析」可以用來分析調查項目（或稱為變數）間的相關性。分析後的結果或許可以因為發現某些變數間的相關性，而縮減調查項目 (這當然進一步節省了調查資源的使用)，或是產生另一組數量較原變數少的新變數，這個過程即所謂的 Dimension-reduction。新變數常呈現出新的意義，是事先分析時不易或無法察覺的，主成分分析便是從原始變數的資料中，找到這層關係；不但保留大部分的「訊息」，也有效的降低變數的數量，對後續的統計分析，甚至圖表的表現都很大的助益。

從以上的案例，很清楚的可以知道，我們不能用任一變量來代表所有變量所呈現的資訊，這是常識。但是如果將所有變量以適度的比例組合，成為一新的變量，它能代表的資訊會比單一變量來得多。主成分分析便是在新變量的產生下功夫，試圖以最少的變數代表原始資料最大的「成分（變量）」，其原則如下：

- 新變數為原變數的線性組合。
- 保留原變數間的最大變異量 (variance)。

當一個新變數不足以代表於變數間的變異，主成分分析也會以相同的原則產生第二個、第三個... 新變數，直到新變數間的變異能涵蓋「大部分」原變數間的變異。這裡所謂的「大部分」無法定義的非常明確，需情況而定，通常在

70% ~ 90% 之間便能滿足需求。

1.2 理論基礎

假設將原始變數 $X_1, X_2, \cdots, X_p$ 做線性組合，轉換為一組新的變數 $Z_1, Z_2, \cdots, Z_p$，

$$
\begin{aligned}
Z_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Z_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
\vdots &= \vdots \\
Z_p &= a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p
\end{aligned}
$$

或表示為

$$
\mathbf{z} = A\mathbf{x} \tag{1}
$$

其中

$$
\mathbf{z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{bmatrix}, \mathbf{x} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}, A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix}
$$

從幾何的角度來看，矩陣 A 也稱為投射矩陣 (Projection matrix)，將資料向量 $\mathbf{x}$ 從原來的空間投射到另一個空間，投射的方式與投射到的空間大小決定了矩陣 A 的組成。資料經過投射或轉置之後，並不會損失或增加原有的「資訊」(線性的轉換不會使資料憑空增加或減少)，只是會改變資料在空間中的「長相」，藉此提供額外的資訊，供進一步資料處理的參考。

主成分分析的理論基礎可以從幾個面象來觀察；分別陳述如下：(為方便分析及符號的簡潔，原始變數均假設均數為零，即 $E(X_i) = 0, \forall i$ 。)

1.3 從 Uncorrelated Variables 的角度

假設新變數 $Z_1, Z_2, \cdots, Z_p$ 間彼此「不相關」(uncorrelated)，則其共變異矩陣為對角化矩陣，即

$$\Sigma_Z = E(\mathbf{z}\mathbf{z}^T) = AE(\mathbf{x}\mathbf{x}^T)A^T = A\Sigma_X A^T = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p^2 \end{bmatrix} \tag{2}$$

因已假設 $E(\mathbf{x}) = \mathbf{0}$ ，共變異矩陣與相關矩陣相同。下面這個定理讓上式得到一個幾何上的意義：

> 定理 1. A symmetric matrix Σ_X can be diagonalized by an orthogonal matrix containing normalized eigenvectors of Σ_X, and the resulting diagonal matrix contains eigenvalues of Σ_X.

假設對稱矩陣 Σ_X 的特徵值（eigenvalues）及特徵向量（eigenvectors）分別為 $\lambda_1 > \lambda_2 > \cdots > \lambda_p$（依大小），$\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p$，根據上述定理，新變數的共變異矩陣 (2) 可以改寫為

$$\Sigma_Z = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}, \text{並且} A^T = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_p \end{bmatrix}$$

從 $\mathbf{z} = A\mathbf{x}$，新變數可以寫成

$$\begin{aligned} Z_1 &= \mathbf{v}_1(1)X_1 + \mathbf{v}_1(2)X_2 + \cdots + \mathbf{v}_1(p)X_p = \mathbf{v}_1^T\mathbf{x} \\ Z_2 &= \mathbf{v}_2(1)X_1 + \mathbf{v}_2(2)X_2 + \cdots + \mathbf{v}_2(p)X_p = \mathbf{v}_2^T\mathbf{x} \\ \vdots &= \vdots \\ Z_p &= \mathbf{v}_p(1)X_1 + \mathbf{v}_p(2)X_2 + \cdots + \mathbf{v}_p(p)X_p = \mathbf{v}_p^T\mathbf{x} \end{aligned} \tag{3}$$

其變異數分別為 $\lambda_1, \lambda_2, \cdots, \lambda_p$。式 (2) 也可以改寫為

$$\Sigma_X = A^T\Sigma_Z A = \sum_{k=1}^{p} \lambda_k \mathbf{v}_k \mathbf{v}_k^T \tag{4}$$

又稱為原始變數共變異矩陣的頻普解構 (Spectral decomposition)。矩陣 $\mathbf{v}_k\mathbf{v}_k^T$（Rank=1）代表組成 Σ_X 的第 k 個「元素」，其相對的特徵值 (variance) λ_k 則表示

該「元素」所貢獻的比例。當 λ_k 相對太小時，甚至可以捨棄該「元素」，僅以「主要成分」（λ_k 相對大的）來近似原來的矩陣。譬如前面 $q(q < p)$ 個特徵值相對大於其餘的，可以下列矩陣近似 Σ_X

$$\Sigma_X \approx \sum_{k=1}^{q} \lambda_k \mathbf{v}_k \mathbf{v}_k^T \tag{5}$$

1.4 從最大變異量的角度

原變數的線性組合中，哪一種組合其變異數最大？假設新變數為

$$Z = u_1 X_1 + u_2 X_2 + \cdots + u_p X_p = \mathbf{u}^T \mathbf{x} \tag{6}$$

問題變為選擇一組組合係數，讓新變數 z 的變異數最大，即

$$\max_{\mathbf{u}} E(Z^2) \equiv \max_{\mathbf{u}} \mathbf{u}^T \Sigma_X \mathbf{u} \tag{7}$$

組合係數 **u** 必須有所限制，否則任意放大將使最大值趨近無限大而失去意義。一般假設 $\mathbf{u}^T\mathbf{u} = 1$ ，問題變為限制式最佳化問題

$$\max_{\mathbf{u}, \mathbf{u}^T\mathbf{u}=1} \mathbf{u}^T \Sigma_X \mathbf{u} \tag{8}$$

利用 Lagrangian multiplier 的方式去除限制式，上述問題進一步成為

$$\max_{\mathbf{u}} \mathbf{u}^T \Sigma_X \mathbf{u} - \lambda(\mathbf{u}^T\mathbf{u} - 1) \tag{9}$$

其最佳解如下：

$$\Sigma_X \mathbf{u}^o = \lambda \mathbf{u}^o$$

這恰是原始變數的共變異矩陣的特徵結構（eigen-structure）。此時，新變數的變異數為

$$var(Z) = E(Z^2) = \mathbf{u}^T \Sigma_X \mathbf{u} = \lambda \mathbf{u}^T \mathbf{u} = \lambda$$

換句話說，當 λ 等於 Σ_X 最大的特徵值時，其相對的特徵向量 $\mathbf{v}_1$ 便是最佳的組合係數。此時的新變數稱為第一個主成分，

$$Z_1 = \mathbf{v}_1(1)X_1 + \mathbf{v}_1(2)X_2 + \cdots + \mathbf{v}_1(p)X_p \tag{10}$$

第二個主成分 $Z_2 = \mathbf{v}^T\mathbf{x}$ 的推演類似上面的過程，但多一個條件：與第一個主成分不相關，即

$$E(Z_1Z_2) = E(Z_1)E(Z_2)$$

這個條件進一步為

$$\mathbf{v}^T\Sigma_X\mathbf{v}_1 = 0 \;\text{ 或是 }\; \mathbf{v}^T\mathbf{v}_1 = 0 \tag{11}$$

同樣利用 Lagrangian multiplier 的方式（此時有兩個限制條件），找到最佳的組合係數 $\mathbf{v}$，求最大的變異數 $var(Z_2)$ 。求解過程留待讀者親自演算，其解為：

$$Z_2 = \mathbf{v}_2^T\mathbf{x} \tag{12}$$

其中 $\mathbf{v}_2$ 為 Σ_X 第二大的特徵值相對的特徵向量。其餘的成分依此方式便可逐一呈現。以下的練習有助於瞭解主成分分析的原理及意義。

2 練習

範例 1. Rand McNally Places Rated Almanac [1] 提供了一組美國城市生活品質的調查資料，將針對美國 329 個城市的 9 項評比資料拿出來觀察，你可以從裡面看到什麼訊息？如何去觀察這麼多 (9×329) 的數字資料？要畫什麼樣的圖？計算哪些統計量呢？

藉著這個範例，不妨可以利用本單元說明的主成分分析原理，實際寫程式去計算主成分分析的所有結果，相信對主成分分析的原理與精神更能掌握。這比起直接執行 sklearn.decomposition 裡的 PCA 還要有感覺，瞭解更深刻。

資料中的 ratings 是 329 個城市的 9 項評比資料，針對大量資料的第一印象或是初期的瞭解可以畫盒鬚圖，程式片段如下，結果如圖 1 所示。

```
from scipy.io import loadmat
import matplotlib.pyplot as plt

D = loadmat('city_quality.mat')
ratings = D['ratings']
categories = D['categories']

fig, ax = plt.subplots()
boxprops = dict(linestyle = '--', linewidth = 1, \
        color = 'darkgoldenrod')
flierprops = dict(marker='o', markerfacecolor = 'green',
    markersize = 4, linestyle = 'none')

ax.boxplot(ratings, boxprops = boxprops, \
        flierprops = flierprops, \
    labels = categories, vert = False)
ax.set_xlabel('Values')
plt.show()
```

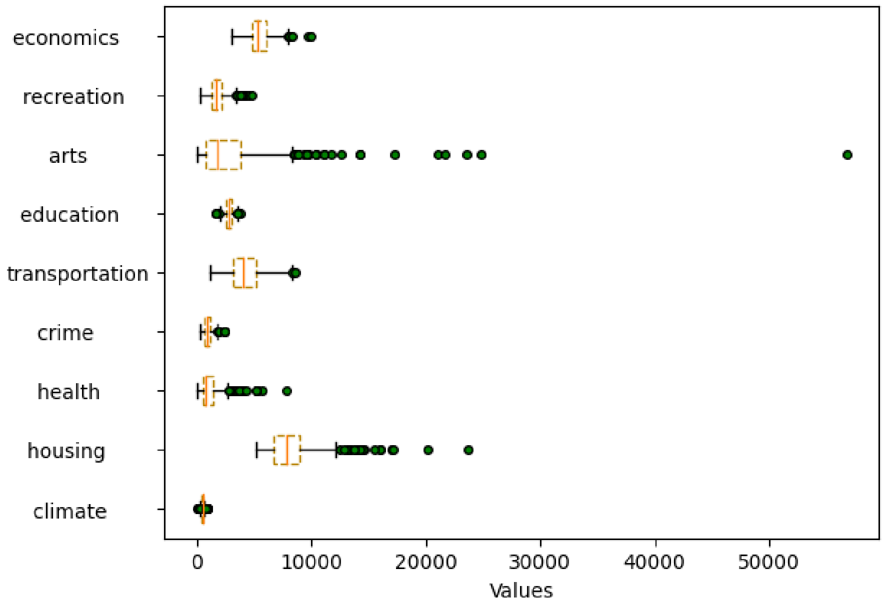


圖 1: 城市生活品質調查資料的盒鬚圖

這張圖對於資料分析很重要，可以看出資料間的差異性，譬如級距 (scale) 的差距及資料的散佈情況，這些都對於判斷資料是否需要做前置處理（pre-processing）很有幫助。從這組城市評比的資料來看，不同項目的數字大小與變異相差頗大，這對做主成分分析可能不利，因此有必要先將這些差距以標準化的方式拉近些。譬如 sklearn 對資料做標準化的指令如下，而標準化的結果如圖 2 所示。

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
# Compute the mean and std to be used for later scaling.
scaler.fit(ratings)
# Apply transform to dataset.
ratings_ = scaler.transform(ratings)
```

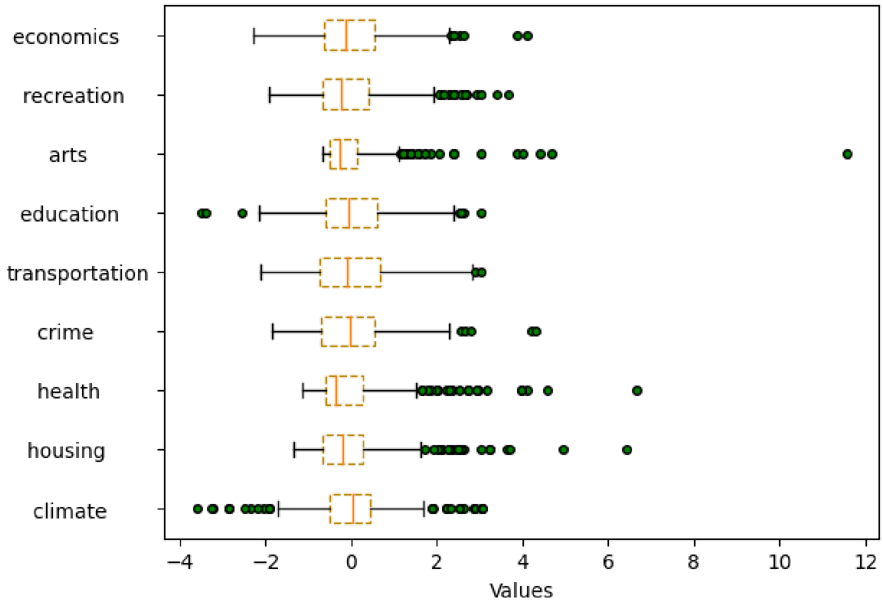


圖 2: 城市生活品質調查資料的盒鬚圖（標準化後）

範例 2. 9 項評比 (9 個變數) 資料是否彼此相關？彼此間的相關性有何差別？畫一張相關矩陣（correlation matrix）是多變量資料分析的基本動作。

比較簡單的做法是結合 pandas 與 seaborn 兩大套件的強大指令，如下。結果如圖 3 所示。當變數數量不多時，相關矩陣也可以同時呈現散佈圖，讀者不妨找找哪個套件的哪個指令可以做到，或是自己寫一個。

```
import seaborn
import pandas as pd

df = pd.DataFrame(ratings_, columns = categories)
R = df.corr()
mask = np.triu(np.ones_like(R, dtype=bool)) # diagonal mask
seaborn.heatmap(R, annot=True, mask = mask, cmap='vlag')
```

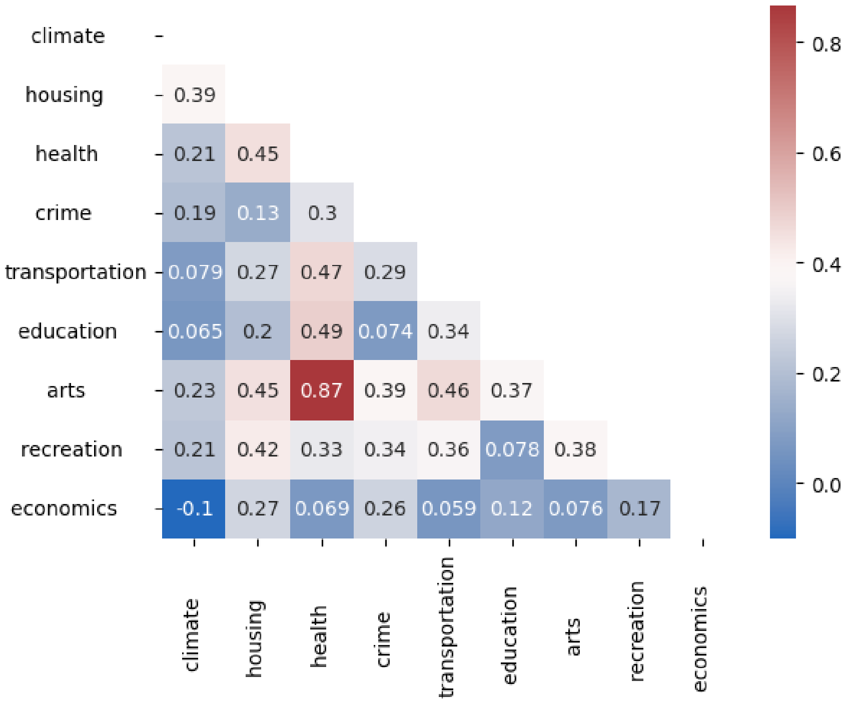


圖 3: 城市生活品質調查資料：不同項目資料間的相關係數

範例 3. 延續前範例的城市生活品質資料，利用 Numpy 指令 `cov` 計算樣本共變異矩陣 S_X，並使用樣本共變異矩陣的公式

$$S_X = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

驗證之。[a]其中 $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$。以城市生活品質資料而言，$N = 329$ 個城市，$\mathbf{x}_i$ 為 9×1 九項品質的向量。最後計算共變異矩陣 S_X 的特徵值（λ_i）及特徵向量（$\mathbf{v}_i$），並使用式 (4) 驗證之。

[a]作為統計專業人士，面對套件提供的統計相關指令都要先弄清楚到底計算了甚麼？用哪個公式？有些使用手冊會詳細載明，有些則沒有，此時最好親自利用公式計算並與使用指令的結果比對。確定後才能安心使用。

使用 `numpy.cov` 與上述公式的寫法如下。檢查變數 `Sx` 與 `Sx_formula` 的內容是否相同？

```
# ratings_ is a 329 by 9 data matrix

Sx = np.cov(ratings_.T, bias=False)

N = ratings_.shape[0]
mu_x = ratings_.mean(axis = 0)
```

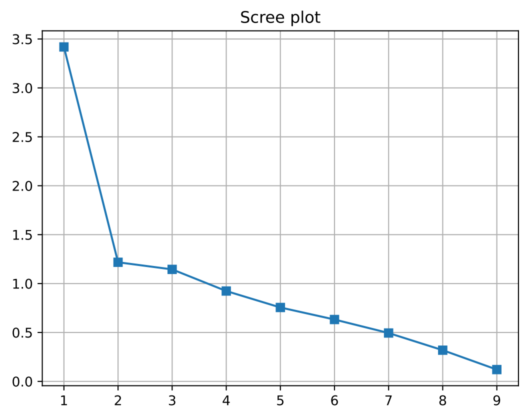
```
Tmp = ratings_ - np.tile(mu_x, (N, 1))
Sx_formula = Tmp.T @ Tmp / (N - 1)
```

接著進一步對樣本共變異矩陣 `Sx` 做特徵值與特徵向量分析（eigen analysis）取得由大而小排列的特徵值及相對應特徵向量，最後再將特徵值與特徵向量依式 (4) 合併回到原來的樣本共變異矩陣。

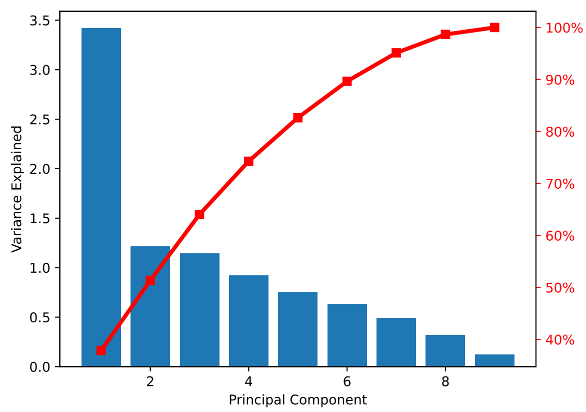
```
from numpy.linalg import eig

w, v = eig(Sx_numpy)
idx = np.argsort(w)[::-1]
eigvals = w[idx]
eigvecs = v[:, idx]
Sigma_x = eigvecs @ np.diag(eigvals) @ eigvecs.T
```

此外，特徵值的分布情況也值得印出來一看，如圖 4(a) 的 Scree plot 與 (b) 的 Pareto plot，參考程式如下：



(a) Scree Plot



(b) Pareto Plot

圖 4: 資料共變異矩陣的特徵值分布

```
from matplotlib.ticker import PercentFormatter

plt.figure()
x = np.arange(1, 1+len(eigvals))
plt.plot(x, eigvals, marker='s')
plt.title('Scree␣plot')
plt.grid(True)
plt.show()

fig, ax = plt.subplots()
```

```
x = np.arange(1, 1+len(eigvals))
ax.bar(x, eigvals)
ax2 = ax.twinx()
ax2.plot(x, eigvals.cumsum()/eigvals.sum()*100, \
    marker='s', color='red', lw=3)
ax2.tick_params(axis='y', colors='red')
ax2.yaxis.set_major_formatter(PercentFormatter())
ax.set_xlabel('Principal␣Component')
ax.set_ylabel('Variance␣Explained')
plt.show()
```

範例 4. 假設五個變數 X_1, X_2, X_3, X_4, X_5，其中 X_1, X_2, X_3 為線性獨立，$X_4 = X_1 + X_2$，$X_5 = X_2 + X_3$，由這 5 個變數構成的共變異矩陣有幾個值為 0 的特徵值呢？試著去模擬這個問題。從樣本共變異矩陣中看看 5 個變數的樣本變異數與特徵值的關係。

進行這個實驗的程序大約如下：

1. 從亂數產生器（譬如假設為標準常態）產生 X_1, X_2, X_3 的樣本值，樣本數 N 自訂。
2. $X_4 = X_1 + X_2, X_5 = X_2 + X_3$
3. 建立 $N \times 5$ 的資料矩陣 $X = [X_1\ X_2\ X_3\ X_4\ X_5]$
4. 計算共變異矩陣 S_x 及其特徵值與特徵向量。
5. 觀察特徵值。

非 0 的特徵值數量也可以從資料矩陣的 rank 得知。出現特徵值為 0，代表資料矩陣並非 full rank(rank < 5)，也就是資料矩陣的內容有部分相依，或說變數間有相依性。於是可以用比較少的變數或比較少的資料便能代表原來的資料矩陣。這便是主成分分析的能力。

範例 5. 主成分分析的在幾何上的概念是「Change of basis」，也就是座標軸的改變（旋轉與位移），如圖 5 之左圖所示，將座標軸從 X_1, X_2 轉換為資料共變異矩陣的特徵向量 Z_1, Z_2。這些特徵向量除了彼此正交之外，還依據資料矩陣的變異成分排列，譬如沿著 Z_1 軸的資料變異大於沿著 Z_2 軸。整個過程可以透過如圖 5 的二維座標轉換來展示。右圖則是將 Z_1, Z_2 軸轉正來看，其資料座標從 (X_1, X_2) 轉換為 (Z_1, Z_2) 。試著實作本範例想表達的意思並繪製圖 5 的左右兩張圖，藉以了解主成分分析的幾何意義。

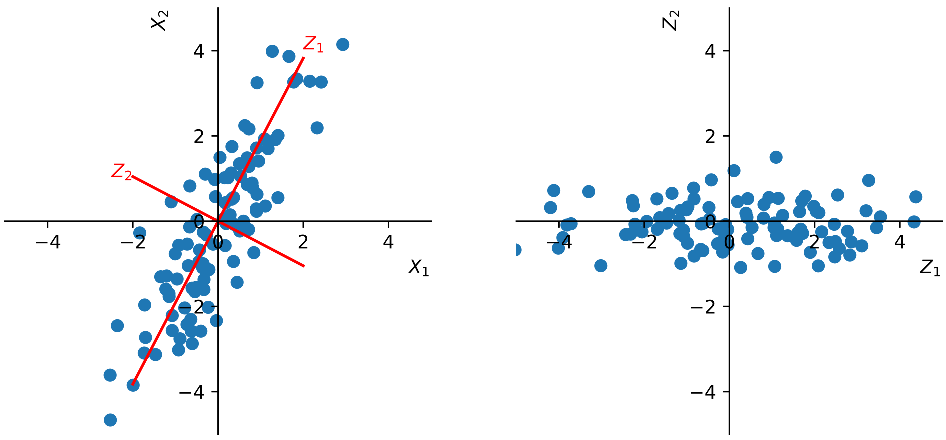


圖 5: 主成分分析的幾何意義：座標轉換

座標軸的改變即變數之變換，從 X_1, X_2 變換為 Z_1, Z_2，其中 Z_1 佔據了最大的資料變異成分，也可以說以單以變數所能涵蓋最多資料的「內涵」。圖 5 的實驗可以這樣做：

1. 產生兩組具相依性的模擬資料，畫出散佈圖。下列兩變數 X_1, X_2 的關係式是一個方式，其中 c 用來調節相關性。

$$X_2 = cX_1 + \epsilon, \quad c \in R, \quad X_1, \epsilon \in N(\mu, \sigma^2)$$

2. 建立兩變數的共變異矩陣 S_X，並計算其特徵值與特徵向量。

3. 第一個特徵向量（特徵值較大者）指向新的座標軸（以 Z_1 代表），第二個

特徵向量則指向與之垂直的另一個座標軸（以 Z_2 表示）。畫出這兩條軸線。

4. 建立矩陣 $A=[\mathbf{v}_1\ \mathbf{v}_2]^T$，其中 $\mathbf{v}_1,\mathbf{v}_2$ 為共變異矩陣 S_X 的兩個特徵向量。計算 $\mathbf{z}=A\mathbf{x}$，便是圖 5 右的座標值。

範例 6. 主成分分析是將原變數作線性組合，成為另一組變數，組合的原則是保留原變數間最大的變異，且新變數彼此不相關。這個練習想去瞭解不同的組合的變異量與幾何意義。

假定 X_1,X_2 兩個變數，樣本資料為 $x_1=[1\ 2\ 3\ 4\ 5]$，$x_2=[2\ 1\ 4\ 5\ 4]$，如果想要用一個新的變數 Z_1 來代表這兩個變數，在希望保留原變數最大變異（variance）的前提下，下列哪一個組合最理想：

a $Z_1=X_1$

b $Z_1=\frac{1}{\sqrt{5}}X_1+\frac{2}{\sqrt{5}}X_2$

c $Z_1=\frac{1}{\sqrt{2}}X_1+\frac{1}{\sqrt{2}}X_2$

d $Z_1=X_2$

問題：

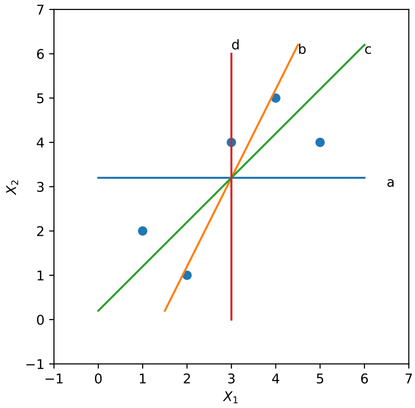
1. 變數 Z_1 的樣本值來自 x_1,x_2 兩變數資料的轉換，這相當前面練習所說的座標軸轉換，而且只代表轉換過後的一個座標軸。請根據上述的組合，分別畫出這個座標軸（含 X_1,X_2 的散佈圖）如圖 6 所示。
2. 分別計算新變數 Z_1 的變異數。哪一個最大？
3. 分別計算新的座標值與新座標軸 Z_1 垂直距離的平方和，哪一個最小？

主成分的來源是以保留原變數間最大的變異為原則，即式 (8) 所示。這個原則的另一面是

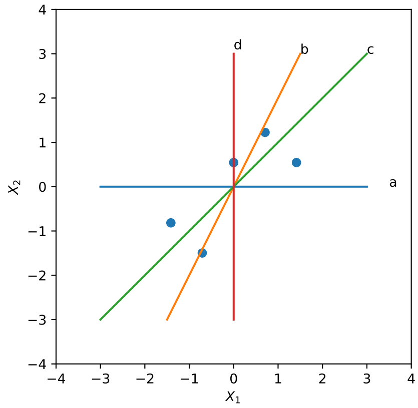
$$\min_P \sum_{k=1}^{n} ||(I-P)\mathbf{x}_k||^2 = \max_P \sum_{k=1}^{n} \mathbf{x}_k^T P \mathbf{x}_k \tag{13}$$

其中 P 即是所謂的 Orthogonal projection matrix。上式以樣本值為依據，若以變數型態則可寫成，

$$\min_P E(||(I-P)\mathbf{x}||^2) = \max_P E(\mathbf{x}^T P \mathbf{x}) \tag{14}$$



(a) 原始資料



(b) 標準化資料

圖 6: 不同線性組合在原始資料與標準化資料下的 Z_1 軸線。

式 (13)(14) 是一種觀念式的表示法，其中的 Orthogonal projection matrix P 便是由資料共變異矩陣的第一個特徵向量（最大特徵值的）$\mathbf{v}_1$ 組成，即 $P = \mathbf{v}_1\mathbf{v}_1^T$。此時，新的座標值與新座標軸 Z_1 垂直距離的平方和便是

$$\sum_{k=1}^{n} ||(I - \mathbf{v}_1\mathbf{v}_1^T)\mathbf{x}_k||^2$$

本範例的資料共變異矩陣的第一個特徵向量（最大特徵值的）也是變數 X_1, X_2 的線性組合，在此順便介紹 `sklearn.decomposition` 套件中對資料矩陣進行主成分分析的指令 `PCA`，`PCA` 的使用方式如下程式碼所示。讀者可以試著計算新的座標值與新座標軸 Z_1 垂直距離的平方和，是否比本範例的四個組合都小？

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA

x1 = np.array([1, 2, 3, 4, 5])
x2 = np.array([2, 1, 4, 5, 4])
X = np.c_[x1, x2] #資料矩陣

# pca = PCA(n_components=1).fit(X) #取第一個特徵向量
pca = PCA().fit(X) # 進行主成分分析
print(pca.explained_variance_ratio_) #共變異矩陣特徵值佔比
print(pca.explained_variance_) #共變異矩陣的特徵值
print(pca.components_)  #共變異矩陣的特徵向量

eigvals = pca.explained_variance_
eigvecs = pca.components_.T # by column [v1 v2]
```

3 觀察與延伸

1. 複迴歸分析所牽涉變數間的多重共線性，也可以運用主成分分析的方式來解決。
2. 轉換座標軸後的第一個軸（$\mathbf{v}_1$），像不像一條迴歸線？
3. 主成分分析通常做為其他資料處理方式的前置作業，能幫助去除多餘的資料、將變數量壓低。不過並非所有的應用都適合做這樣的處理，有些時候反而將有用的資料覆蓋或打亂（譬如，群組分析），未獲其利，先蒙其害。應用時機的選擇非常重要，需要經驗與審慎的態度。

4 習題

1. 利用 son.txt [2] 這組資料做主成分分析：

   a 共變異矩陣 (Covariance Matrix) 是觀察兩個變數之間關係較常用的統計量。取前兩欄資料，計算『頭部長度』與『頭部寬度』的樣本共變異矩陣 S_x (sample Covariance Matrix)。

   b 繪製兩者的散佈圖，圖形顯示的是否與共變異矩陣呼應？如何觀察？

   c 計算樣本共變異矩陣 S_x 的特徵值及相對的特徵向量。觀察特徵值的大小分佈，是否與兩變數間的相關程度有關？觀察特徵向量 $\mathbf{v}_1, \mathbf{v}_2$ 的關係，是否存在 orthogonal 的關係？即 $\mathbf{v}_1^T\mathbf{v}_2 = 0$ ？

   d 假設樣本共變異矩陣 S_x 的特徵值為 λ_1, λ_2，相對的特徵向量為 $\mathbf{v}_1, \mathbf{v}_2$。驗證 $S_x = \lambda_1\mathbf{v}_1\mathbf{v}_1^T + \lambda_1\mathbf{v}_2\mathbf{v}_2^T$

   e 當將資料的座標軸從 (X_1, X_2) 轉為 (Z_1, Z_2) 時，原資料的座標值將隨之改變。畫出如圖 7 的兩條垂直線。

   - 先計算中心點 (想想看這個中心點如何決定？)
   - Z_1, Z_2 軸就是 $\mathbf{v}_1, \mathbf{v}_2$ 的方向，透過向量與中心點便可以畫出如圖中的新座標軸 Z_1, Z_2。

   f 從座標軸 (Z_1, Z_2) 來看這些資料，似乎顯示出『比較散亂』的不相干關係。不過又扁向 Z_1 軸。這個『扁』的傾向或程度，可以畫一個橢圓來表示。

   g 以新的座標軸 (Z_1, Z_2) 來看這些資料，新的座標值如何計算？是不是可以找到一個轉換機制 (矩陣)？複習線性代數有關座標軸轉換的部分。

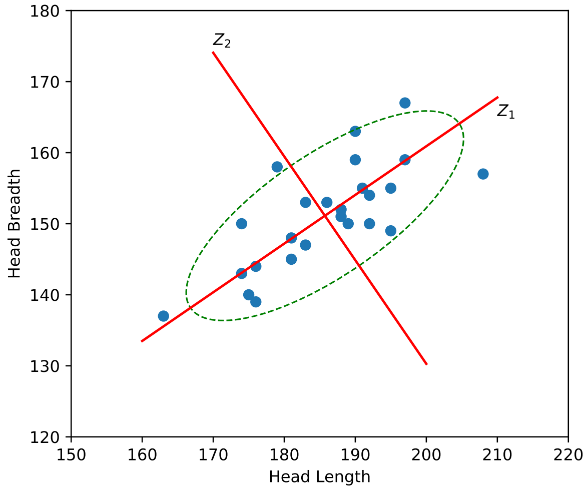


圖 7: 實際資料 son.txt 的主成分分析。

2. FOOTBALL.txt [4] 這組資料提供作為安全帽設計與頸部傷害的研究。研究的對象是美國大學 football 與非 football 球員共 60 名，並量測 6 種頭部相關的資料。選擇這 6 種頭部相關資料是否能反映出設計的關鍵，並不是本主題的興趣。本主題想探討這些資料彼此間是否有相關性？也就是說：或許更少的資料就能表達出這 6 種資料所能表達的意涵！如果是這樣，對於應用上的幫助不小，因為那代表需要花費的人力成本降低 (要量測的項目變少)，在分析上也比較容易 (變數少了)，結果也會比較『穩定』(獨立性強了)。這個練習要探討幾個理論與程式設計的技巧：

a 先簡單的觀察一下這 6 組資料的相關性，以得到一個初淺的變數間相關的程度。建議畫出每組資料的散佈圖。

b 計算並觀察原始資料的共變異矩陣 (Covariance Matrix)。從這個關聯性值的矩陣，能否看出初步的相依性，或變數個別的重要性？

c 執行主成分分析，觀察其特徵值的分布，並且求其比重的分布。可以畫所謂的 scree plot，即依特徵值大小做圖。或畫特徵值的 Pareto(柏拉圖) plot。

d 取前兩個主成分組成新的變數 Z_1, Z_2，即

$$\begin{aligned} Z_1 &= \mathbf{v}_1(1)X_1 + \mathbf{v}_1(2)X_2 + \cdots + \mathbf{v}_1(6)X_6 \\ Z_2 &= \mathbf{v}_2(1)X_1 + \mathbf{v}_2(2)X_2 + \cdots + \mathbf{v}_2(6)X_6 \end{aligned}$$

從由特徵向量組成的係數來看，觀察哪些變數 X_i 的重要性比較高？是不是可以據此說明只要這些變數即可表達所有的意義？

e 畫一張 Z_1, Z_2 的散佈圖，觀察他們的相關性及分佈的情況 (像常態嗎?)。另外值得觀察的是這些資料在 Z_1 軸及 Z_2 軸的變異性 (variance)，及是否存在群聚性 (grouping)。這個問題可以自行寫程式計算，也可以直接採用指令 princomp。

3. 證明式 (9) 與 (10) 是相同的問題。

4. 證明式 (12)。

References

[1] R. Boyer and D. Savageau, ”Rand McNally Places Rated Almanac”, 1985.

[2] G.P. Frets, ”Heredity of Head Form in Man”, Genetica, 3, 193–384, 1921.

[3] J. Latin, D. Carroll, P. E. Green, ”Analyzing Multivariate Data,”2003, Duxbruy.

[4] A. C. Rencher, ”Multivariate Statistical Inference and Applications,”1998, John Wily and Sons.