1

Determine the population principal component Y_1 and Y_2 for the covariance matrix

$$\Sigma = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$$

Also, calculate the proportion of total population variance explained by the first principal component.

透過求解

$$det(\Sigma - \lambda I) = 0$$

可得其 eigenvalues 分別爲 $\lambda_1 = 6, \lambda_2 = 1$，並個別從其 eigenspace 中取對應的 unit eigenvectors

$$e_1 = \begin{bmatrix} 0.894 \\ 0.447 \end{bmatrix} \quad e_2 = \begin{bmatrix} -0.447 \\ 0.894 \end{bmatrix}$$

可得到兩個 PC 分別爲

$$Y_1 = 0.894X_1 + 0.447X_2$$
$$Y_2 = -0.447X_1 + 0.894X_2$$

而第一個 PC 解釋的部分佔總母體變異的比例爲

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{6}{6+1} \approx 0.857$$

2

Convert the covariance matrix in Exercise 8.1 to a corrlation matrix ρ

(a) Determine the principal component Y_1 and Y_2 from ρ and compute the proportion of total population variance explained by Y_1

(b) Compare the components calculated in Part a with those obtained in Exercise 8.1. Are they the same? Should they be?

(c) Compute the correlations ρ_{Y_1,Z_1}, ρ_{Y_1,Z_2} and ρ_{Y_2,Z_1}.

先將 covariance matrix 轉成 correlation matrix

$$\rho = \begin{bmatrix} 1 & 0.632 \\ 0.632 & 1 \end{bmatrix}$$

(a) 透過

$$det(\rho - \lambda I) = 0$$

求得 $\lambda_1 = 1.632, \lambda_2 = 0.368$，並個別從其 eigenspace 中取對應的 unit eigenvectors

$$e_1 = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix} \qquad e_2 = \begin{bmatrix} -0.707 \\ 0.707 \end{bmatrix}$$

可得到兩個 PC 分別爲

$$Y_1 = 0.707Z_1 + 0.707Z_2$$
$$Y_2 = -0.707Z_1 + 0.707Z_2$$

而第一個 PC 解釋的部分佔總母體變異的比例爲

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1.632}{1.632 + 0.368} = 0.816$$

(b) 徂結果可發現兩組 PC 並不一樣，基本上使用 Σ 和 ρ 計算的結果不會一樣，這是由於 Σ 本身包含數據的尺度，而 ρ 則是經過標準化後的結果。

(c) 根據定義

$$\rho_{Y_i,X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sigma_k}$$

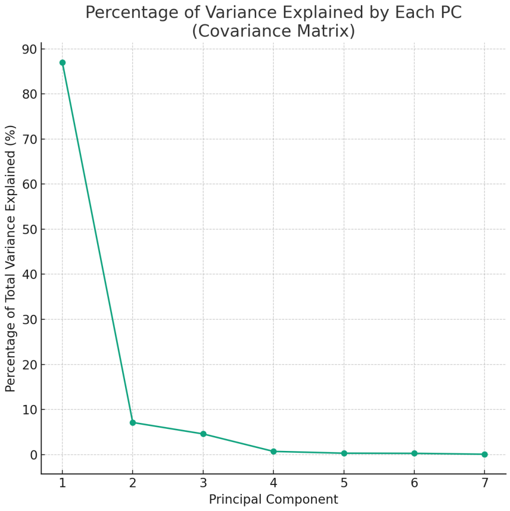
可得

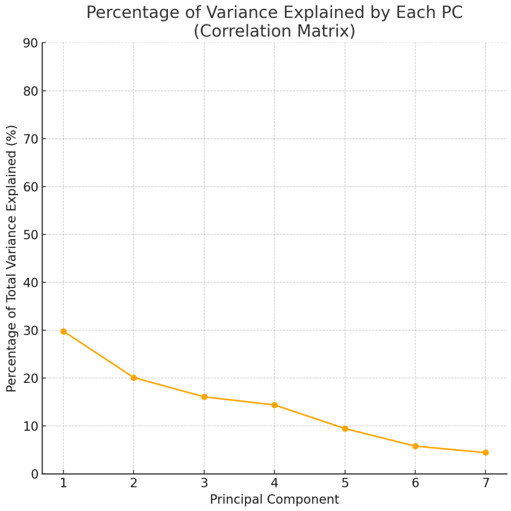
$$\rho_{Y_1,Z_1} = \frac{e_{11}\sqrt{\lambda_1}}{\sigma_1} = \frac{0.707 \times \sqrt{1.632}}{1} \approx 0.903$$
$$\rho_{Y_1,Z_2} = \frac{e_{12}\sqrt{\lambda_1}}{\sigma_2} = \frac{0.707 \times \sqrt{1.632}}{1} \approx 0.903$$
$$\rho_{Y_2,Z_1} = \frac{e_{21}\sqrt{\lambda_2}}{\sigma_1} = \frac{-0.707 \times \sqrt{0.368}}{1} \approx -0.429$$

3

Consider the air-pollution data listed in Table 1.5. Your job is to summarize these data in fewer than $p = 7$ dimensions if possible. Conduct a principal component analysis of the data using both the covariance matrix S and the correlation matrix R. What have you learned? Does it make any difference which matrix is chosen for analysis? Can the data be summarized in three or fewer dimensions? Can you interpret the principal components?

透過將 python 執行結果視覺化可得





透過觀察可發現用 covariace matrix 所產生的第一個 PC 就足以解釋大部分的變異(約佔 88% 的總變異)，而大致上取到第四個 PC 就足以解釋接近 100% 的變異；反觀用 correlation matrix 做出來的 PC 每一個佔的解釋量相對來說就較為均匀，從圖中情況來看大致上要取到第六才足以解釋接近全部的變異。

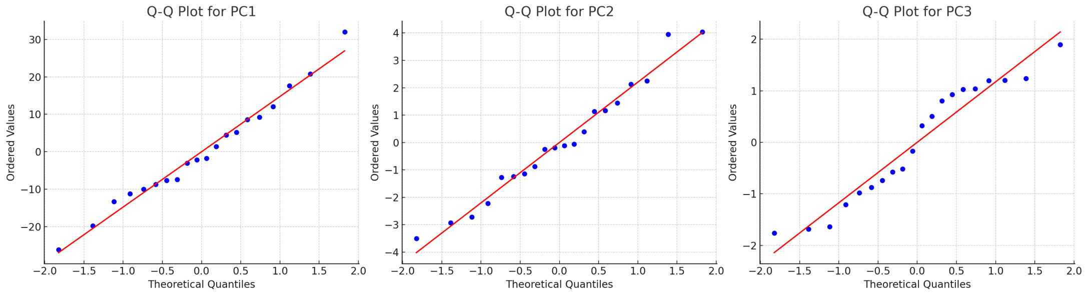
因此，對於使用 covariace matrix 所產生的 PC 用三個就可以解釋 (甚至兩個)，但對於用 correlation matrix 做出來的 PC 則無法只透過三個或更少的 PC 就解釋大部分變異。

If we have variables with widely varying scales for raw, using correlation-based PCA would be a better choice, since it is equivalent to standardizing the varuables before doing PCA. Otherwise, covariance-based PCA is fine. We will reconmand using covariance one in this case since the range and the scale of these variables are relatively similar.

4

Perform a principal component analysis using the sample covariance matrix of the sweat data given in Example 5.2. Construct a Q-Q plot for each of the important principal components. Are there any suspect observations? Explain.

透過 python 將結果視覺化



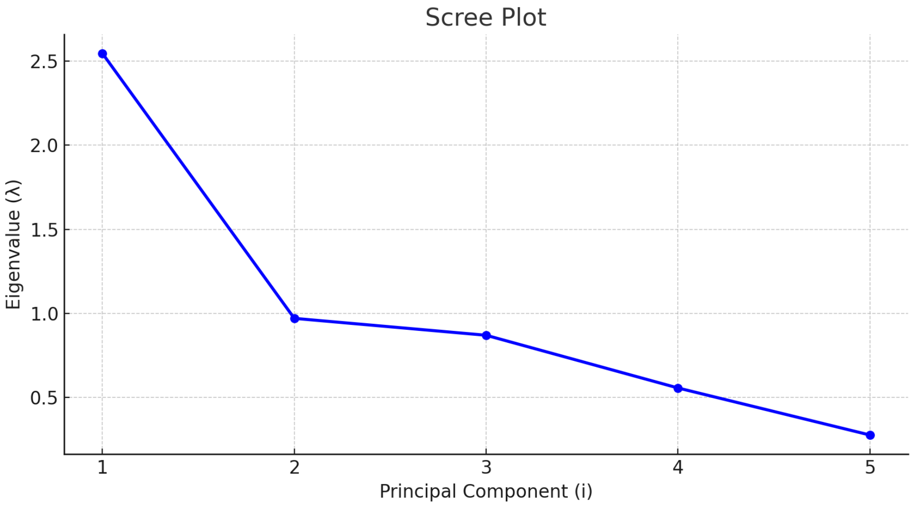
可發現並沒有過於特殊的 outlier，而第一和第二個 PC 大致上符合常態分配，第三個則稍微偏離常態分配 1[1]

5

File FOODP.DAT gives the average price in cents per pound of five food items in 24 U.S. cities. Use principal components analysis to analyze the data (correlation matrix) and determine the appropriate number of principal components.

(a) Show Scree plot, principal components and interpret their meanings
(b) Show the first two principal components score
(c) Plot the first two components scores in a scatter plot. How many “clusters” can be identified visually?
(d) Which city is the least expensive? Which city is the most expensive?

(a) 根據題意，透過 python 繪製 Scree Plot



[1] Choose eigenvalues that are larger than 1 to be important principal components

從這張圖可以得知，當 PCA 數量爲 2 的時候已解釋大部分的變異，

PC	Bread	Hamburger	Butter	Apples	Tomatoes
PC1	0.509927	0.520072	0.397311	0.290942	0.476442
PC2	-0.056496	0.277616	-0.099401	-0.876753	0.375714
PC3	-0.401716	-0.407437	0.768477	-0.071363	0.277433
PC4	-0.531979	0.071481	-0.430414	0.372578	0.622750
PC5	-0.540745	0.693787	0.237592	0.052439	-0.408723

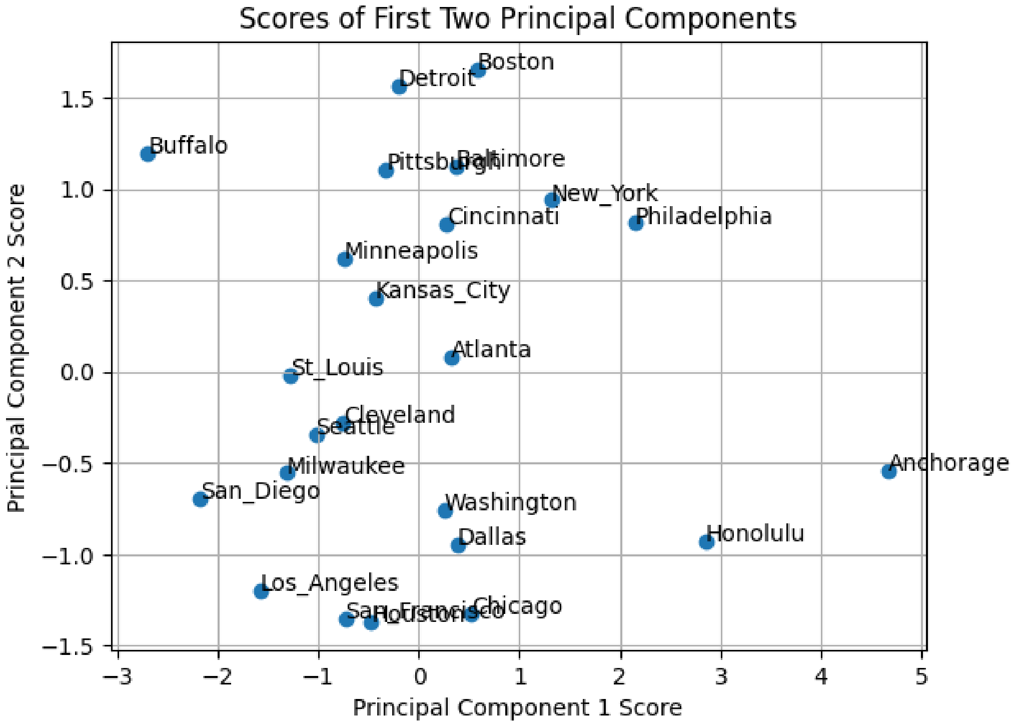
- PC1: 總體平均物價
- PC2: 速食價格
- PC3: 食品佐料、配菜價格
- PC4: 蔬果類價格
- PC5: 高油脂類食品價格

(b) 透過計算可得

City	PC1	PC2
Anchorage	4.674532	-0.539971
Atlanta	0.323536	0.081827
Baltimore	0.366146	1.125827
Boston	0.586307	1.652933
Buffalo	-2.691778	1.193800
Chicago	0.519218	-1.320803
Cincinnati	0.271076	0.804306
Cleveland	-0.753021	-0.280745
Dallas	0.390528	-0.943408
Detroit	-0.202286	1.563685
Honolulu	2.856135	-0.929948

City	PC1	PC2
Houston	-0.471509	-1.369421
Kansas City	-0.427955	0.403235
Los Angeles	-1.579078	-1.198470
Milwaukee	-1.311458	-0.555591
Minneapolis	-0.743544	0.616150
New York	1.316738	0.938323
Philadelphia	2.146621	0.812629
Pittsburgh	-0.326972	1.107534
St Louis	-1.279764	-0.020950
San Diego	-2.168506	-0.693957
San Francisco	-0.728630	-1.348053
Seattle	-1.023051	-0.344750
Washington	0.256718	-0.754180

(c) 根據題意，透過 Python 繪製 Scatter Plot：



(d) The city with the smallest PC1 score (which is the cheapest) is: Buffalo, with a PC1 score of -2.691778

The city with the largest PC1 score (which is the most expensive) is: Anchorage, with a PC1 score of 4.674532

6

Following the 1973-74 Arab oil embargo and the subsequent dramatic increase in oil prices, a study was conducted in three cities to estimate the potential demand for mass transportation. Five hundred and ninety-seven (597) responded to the twenty statements on a five-point scale (1=disagree strongly to 5=agree strongly). Use principal components analysis to analyze the data (correlation matrix) and identify the key perceptions about the energy crisis

(a) Show the first five sample principal components loadings. What is the percentage of total variance they account for?

(b) Interpret the first five sample principal components.

(a) 透過 R 計算可得各個 PCA 對應的 loadings

	PC1	PC2	PC3	PC4	PC5
X_1	0.14657024	0.09227945	0.024085542	0.429222864	0.28002768
X_2	-0.25970643	-0.12633077	-0.083626424	-0.213743149	-0.31082614
X_3	-0.31506778	-0.03036303	-0.059268131	-0.119943568	-0.12269589
X_4	0.05532526	0.11627308	-0.170475070	0.375466097	0.14991207
X_5	0.38024152	0.07125801	-0.132902774	0.183775941	0.08544616
X_6	-0.12187538	0.51858030	-0.234598908	-0.002330035	-0.00519485
X_7	0.01349917	0.21789159	0.299532194	0.216424722	-0.44725056
X_8	-0.07242297	-0.23380266	-0.392537092	0.026630266	-0.19549725
X_9	0.24026346	-0.05516351	-0.221235125	0.324417466	-0.27141076
X_{10}	0.28265396	0.12960761	-0.106081068	-0.185042017	-0.12814937
X_{11}	-0.04514394	-0.12295423	-0.498252387	-0.011924778	0.07322244
X_{12}	-0.34122018	-0.14554310	0.063172407	0.345433483	-0.10148906
X_{13}	-0.36314371	-0.13416633	-0.044714076	0.298074461	0.07889284
X_{14}	0.02676672	-0.22546943	-0.357762816	0.117148136	-0.33485677
X_{15}	0.27086172	-0.01914114	-0.007048647	0.205795185	-0.23135948
X_{16}	0.17688533	0.04534911	-0.121896746	-0.127136048	-0.18521808
X_{17}	-0.20866004	0.46343210	-0.187702592	-0.023519765	0.01631630
X_{18}	-0.25611231	-0.08802342	-0.016629035	0.259018169	0.22645267
X_{19}	-0.18108088	0.47042091	-0.221611135	0.018747984	-0.08985568
X_{20}	-0.09284673	0.11715413	0.318403228	0.212970372	-0.41268469

前五因素的總變異數占比爲

	PC1	PC2	PC3	PC4	PC5
Proportion of Variance	0.1468115	0.1049417	0.1018391	0.0830500	0.05971987
Cumulative Proportion	0.1468115	0.2517532	0.3535923	0.4366423	0.49636217

(b) 前五個主成份代表意義如下

(a) PC1 能源問題應被重視

(b) PC2 污染問題應被重視

(c) PC3 能源企業與政府單位應該爲此負責

(d) PC4 節省能源爲當務之急

(e) PC5 政府比起能源廠商更應該負責