Midterm exam. PCA, FA, CCA, CA and MCA

109354003 統研所一 吳書恆

April 27, 2021

Q1. The "Sales.txt" data set contains the following 7 variables:

表 1: Sales Variables

Variable	Description
V1	Sales growth
V2	Sales profitability
V3	New account sales
V4	Creativity test
V5	Mechanical reasoning test
V6	Abstract reasoning test
V7	Mathematics test

A firm is attempting to evaluate the quality of its sales staff and is trying to find some tests that may reveal the potential for good performance in sales. The firm has selected 50 of its employees at random and has evaluated each employee on 3 measures of performance (the first 3 variables that have been converted to a scale, on which 100 indicates average performance) and on 4 tests (the remaining 4 variables, measured on a scale of 0-50).

1. Perform a complete Principal Components Analysis for this data and interpret the result.
2. Perform a complete exploratory Factor Analysis based on MLE and interpret the result.
3. How do the factors obtained in Q2 compare to the principal components obtained in Q1?

Ans.

1. 首先，我們先看一下圖 1 的相關係數矩陣，發先這七個變數彼此間相關性都蠻高的，最低有 0.15，最高有 0.94，在尚未進行維度縮減前可以推測降爲的效果會不錯，下面是 R 的 PCA 執行的結果。

```
Importance of components:
                        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
Eigen values            5.0345 0.9335 0.4979 0.4212 0.0810 0.0203 0.0113
Standard deviation      2.2437 0.9661 0.7056 0.6490 0.2846 0.1426 0.1064
Proportion of Variance 0.7192 0.1333 0.0711 0.0601 0.0115 0.0029 0.0016
Cumulative Proportion  0.7192 0.8525 0.9237 0.9838 0.9954 0.9983 1.0000

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
V1  0.434  0.112                0.632  0.337  0.528
V2  0.420         0.442               -0.785
V3  0.421        -0.204  0.325 -0.701  0.157  0.399
V4  0.294 -0.668 -0.451  0.303  0.261 -0.114 -0.300
V5  0.349 -0.295        -0.847 -0.174  0.197
V6  0.289  0.642 -0.604 -0.154        -0.236 -0.228
V7  0.407  0.200  0.434  0.246         0.371 -0.636
```

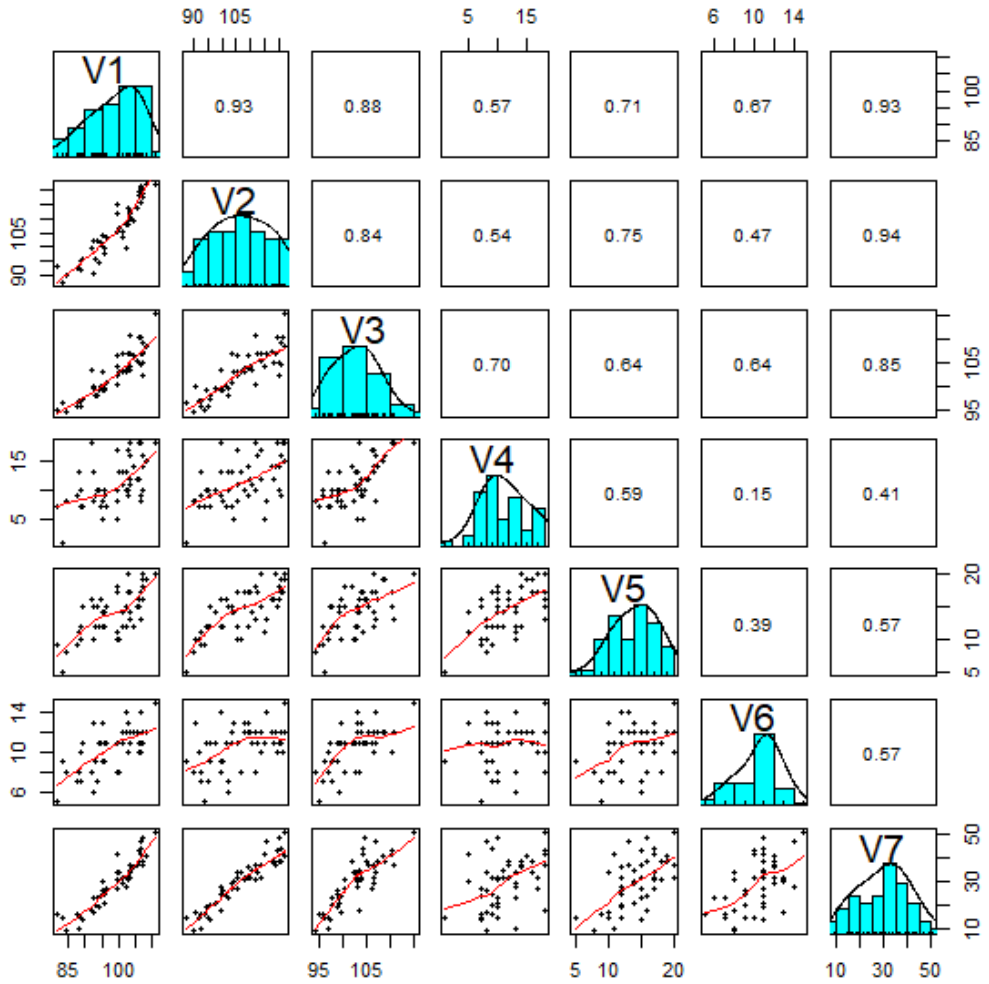

圖 1: Scatter Plot Matrices of Sales

第一主成分已經解釋總變異的 75%，到第二主成分已經了 85% 且特徵值已小於 1，透過 Permutation Test 也發現一個主成分已足夠（p 值在第二個主成分納入時大於 0.05），但爲了在二維上觀看與比較，這邊取首兩個主成分。根據 loading 這兩個主成分分別代表是總表現、邏輯思維能力。

圖 2 的左邊是 PCA 在兩個主成分的 PCscore 散布圖，可以看到 8 號在各種表現上都是較優秀的，44 號、48 號和 16 號則是較差，但是 44 號的邏輯思維能力較強。

2. 爲了比較 PCA 和 FA 分析結果，在因素分析中也取兩個首要因素，由於使用 MLE 的方式，得以檢定出選兩個因素是充分的（p 值 < 0.001），占總變異的 80%，剩餘 20% 推測是 V5 和 V6 的資訊。根據 loading 這兩個因素分別代表是除了創造力之外的總表現、創造力和創造新客戶能力。

```
Uniquenesses:
   V1    V2    V3    V4    V5    V6    V7
0.069 0.070 0.123 0.005 0.474 0.614 0.029

Loadings:
   Factor1 Factor2
V1 0.852   0.452
V2 0.868   0.419
V3 0.717   0.602
```

```
V4 0.148   0.987
V5 0.501   0.525
V6 0.619
V7 0.946   0.277

                Factor1 Factor2
SS loadings       3.545   2.071
Proportion Var    0.506   0.296
Cumulative Var    0.506   0.802

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 117.2 on 8 degrees of freedom.
The p-value is 1.25e-21
```

3. 圖 2 的右邊是 FA 在兩個因素的散布圖，可以看到 8 號在整體表現上仍是較優秀的，28 號、30 號和 36 號在因素一的尺度上比起 PCA 更凸顯了出來；32 號、48 號的的能力較差，44 號隨然在整體上已不再那麼低分，但其邏輯能力就在 FA 中消失而無法看出。

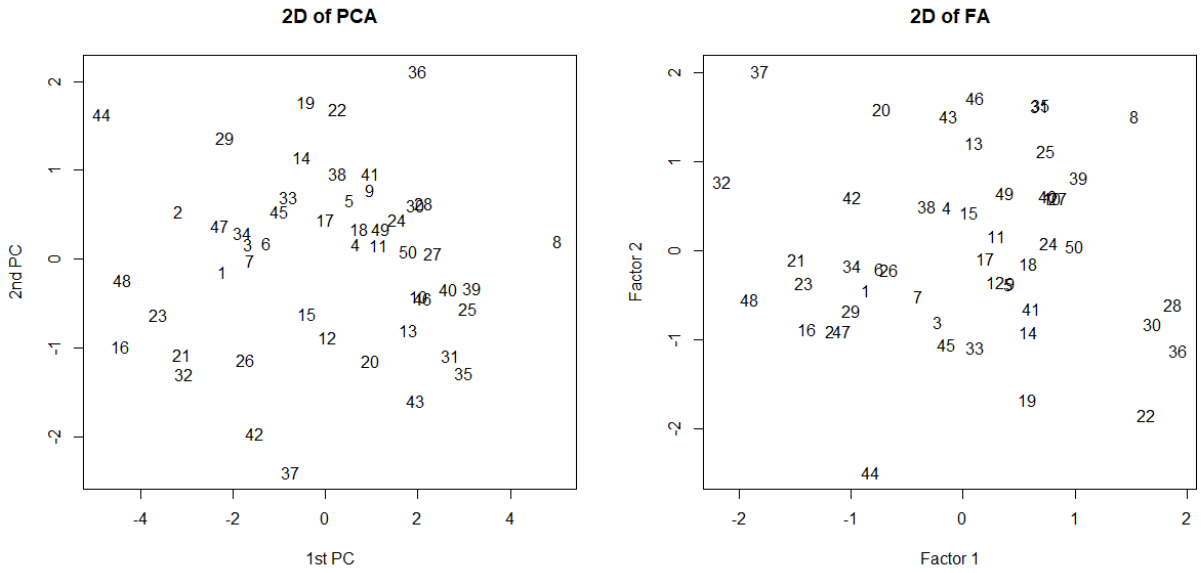

圖 2: 2D plot of PCA and FA in Sales

Q2. The ”Air Pollution.txt” data set contains the following 8 variables and 41 US cities:

表 2: Air Pollution Variables

Variable	Description
PT	Particulates content of air (in 10^{-6} grams per cubic meter)
CO	Carbon monoxide content of air (in ppm)
SO2	Sulfur dioxide content of air (in ppb)
PSI	Air pollution index, PSI=0 means ”good”, PSI=1 means ”moderate”, PSI=2 means ”Unhealthy”.
Temp	Average annual temperature in degrees Fahrenheit
Man	Number of manufacturing enterprises employing 20 or more workers
Pop	Population size in thousands from the 1970 census
Rain	Average annual precipitation in inches

Suppose a researcher is interested in finding how the density of air pollution contents (PT, CO, SO2) is related

to the set of variables (Temp, Man, Pop, Rain). Perform a complete Canonical Correlation Analysis for these two groups of variables and interpret the result.

Ans.

1. 一樣先對資料做簡單的探索性分析，如圖 3，除了發現變數與變數之間的相關，如 CO 與 SO2 有高度相關，Man 和 Pop 有高度相關，更發現所有變數似乎都非常態。

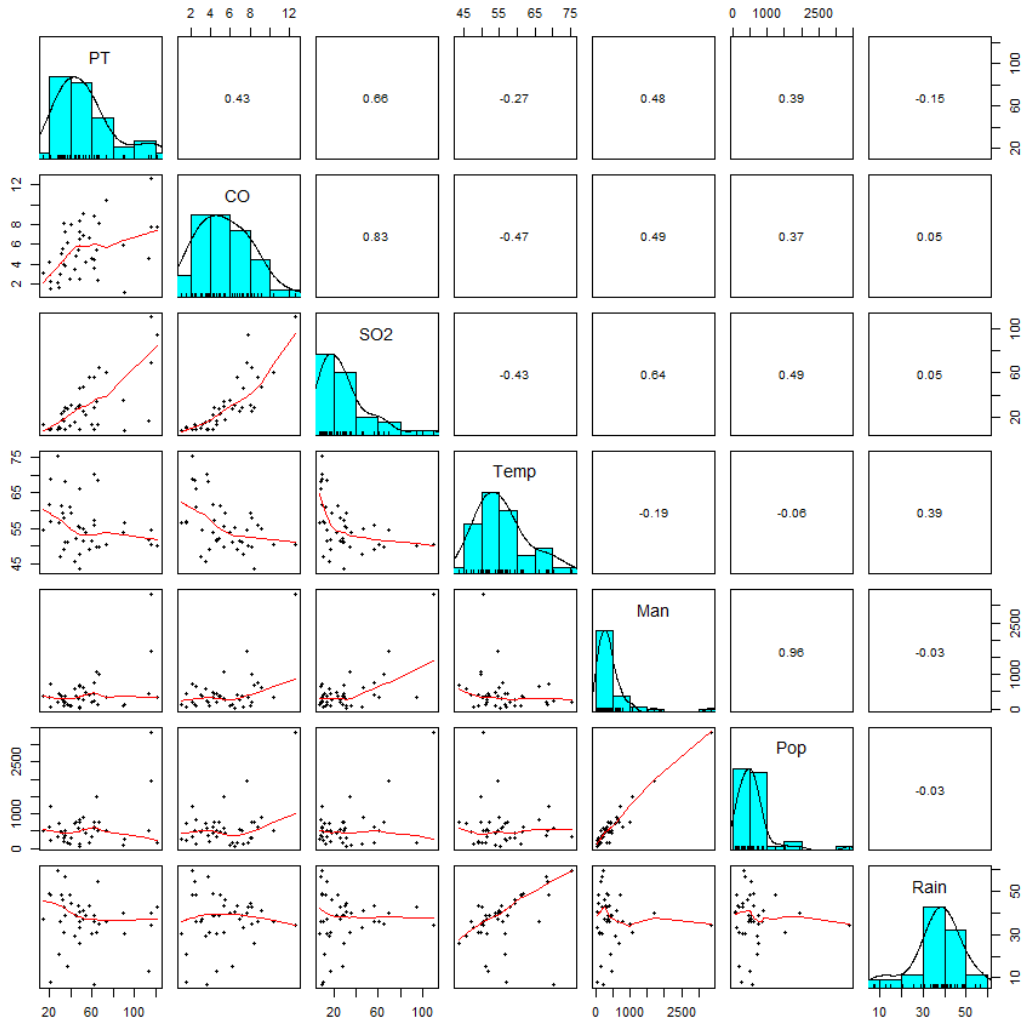

圖 3: Scatter Plot Matrices of Sales

針對右偏的變數做對數轉換後有好轉跡象，至於左偏的 Rain 則是勉強將降雨量極低的地區退出分析，這邊刪去 Phoenix 樣本，如此在常態的檢定才會通過，但四個檢定多維常態方式只有 royston 通過。

```
> mvn(newair, mvnTest = c('royston'), desc = FALSE)
$multivariateNormality
     Test       H   p value MVN
1 Royston 8.54569 0.2059667 YES

$univariateNormality
          Test  Variable Statistic   p value Normality
1 Shapiro-Wilk    PT        0.9801    0.6931    YES
2 Shapiro-Wilk    CO        0.9560    0.1221    YES
3 Shapiro-Wilk    SO2       0.9593    0.1581    YES
```

```
4 Shapiro-Wilk   Temp       0.9655    0.2565    YES
5 Shapiro-Wilk    Man       0.9779    0.6116    YES
6 Shapiro-Wilk    Pop       0.9740    0.4778    YES
7 Shapiro-Wilk   Rain       0.9569    0.1313    YES
```

接著進行 CCA 分析，以下是分析結果，可以發現第一組的相關性高達 0.68，第二組則降低至 0.31，透過 Wilk's test 檢定出只要選第一組即可（p 值 = 0.014）。解釋的部分可從 `$xcoef` 和 `$ycoef` 去看，在 *X* 的線性組合中，CO 和 SO2 相對較低，表示 *X* 代表的是空氣酸性不高的程度；在 *Y* 的線性組合中則是 Temp 較高，因此 *Y* 表示溫度高低。由圖 4 中可以看到一些城市落在右上的區域空氣汙染不高。

```
$cor
[1] 0.67968431 0.30599675 0.03676353

$xcoef
            [,1]        [,2]       [,3]
PT  -0.01056525  0.09084529  0.1862861
CO  -0.08665714 -0.23455362  0.1951337
SO2 -0.07361103  0.18053550 -0.3055074

$ycoef
             [,1]       [,2]        [,3]        [,4]
Temp  0.15983148  0.1167473 0.035648655 -0.10102992
Man  -0.09648190  0.2661367 0.001992716 -0.24160617
Pop   0.05109307 -0.2195795 0.150927407  0.24473497
Rain -0.09465305 -0.1803454 0.015444997 -0.04977727

Wilks' Lambda, using F-approximation (Rao's F):
              stat     approx df1     df2     p.value
1 to 3:  0.4869923 2.28142122  12 87.6013 0.01419681
2 to 3:  0.9051410 0.57907434   6 68.0000 0.74571224
3 to 3:  0.9986484 0.02368426   2 35.0000 0.97660965
```

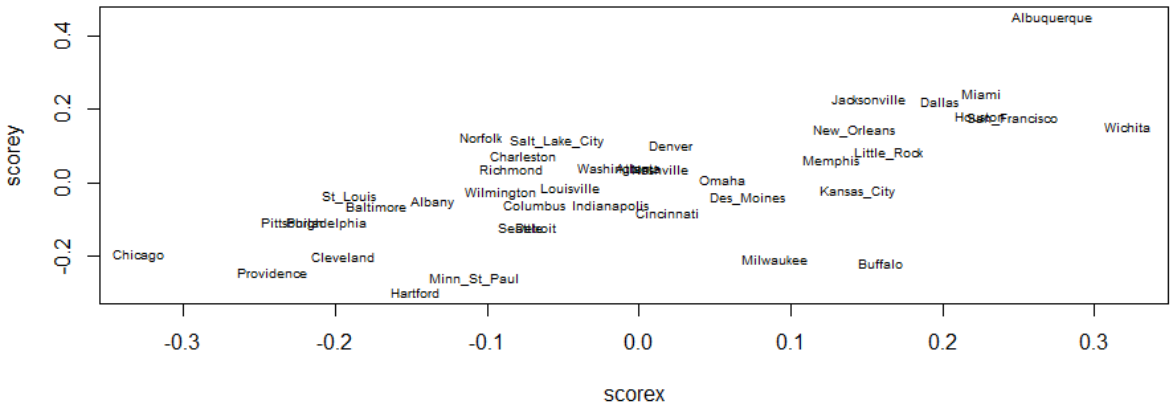

圖 4: 2D plot of CCA