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

109354003 統研所一 吳書恆

April 28, 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
```

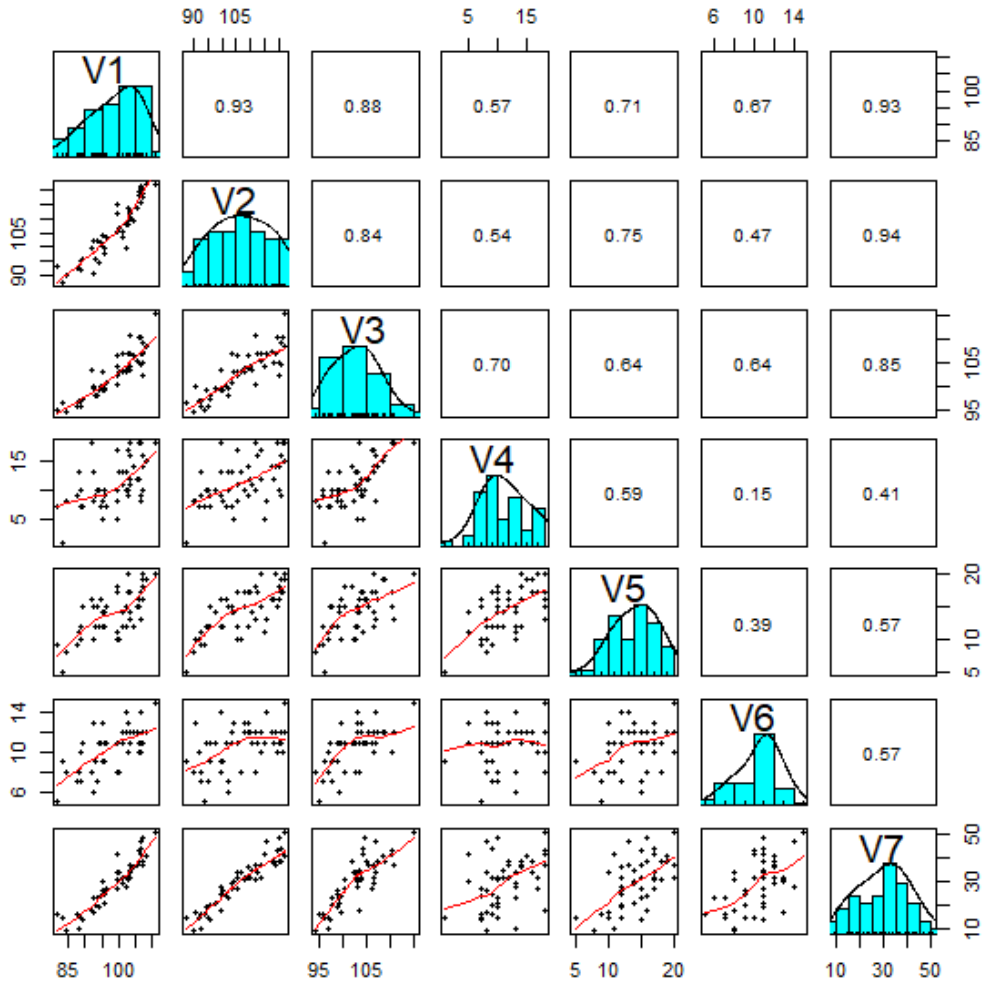


圖 1: Scatter Plot Matrices of Sales

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。

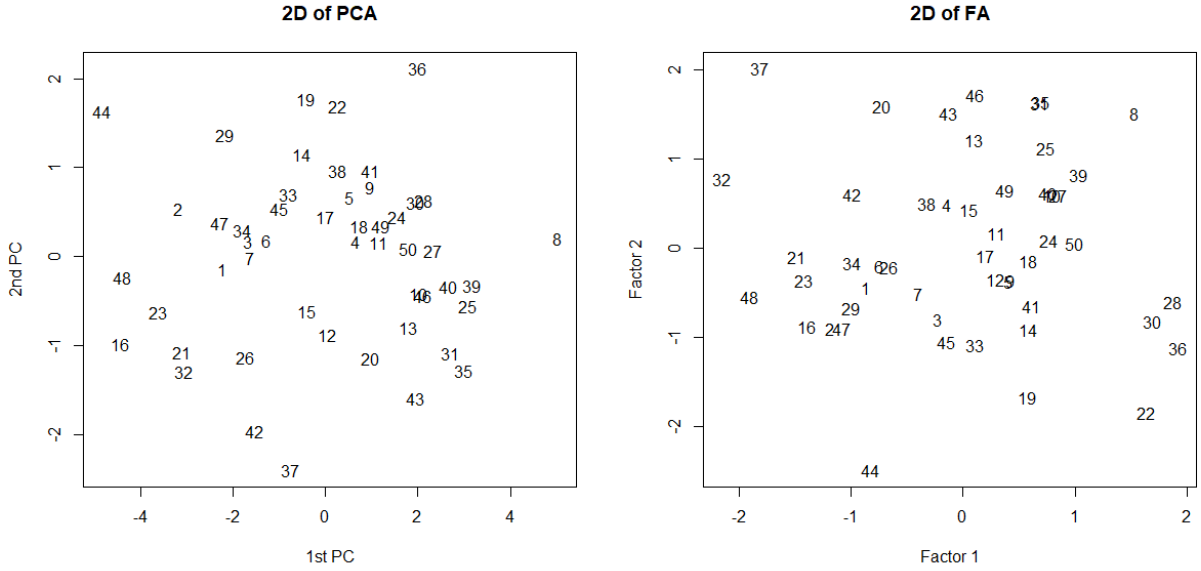


圖 2: 2D plot of PCA and FA in Sales

**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。

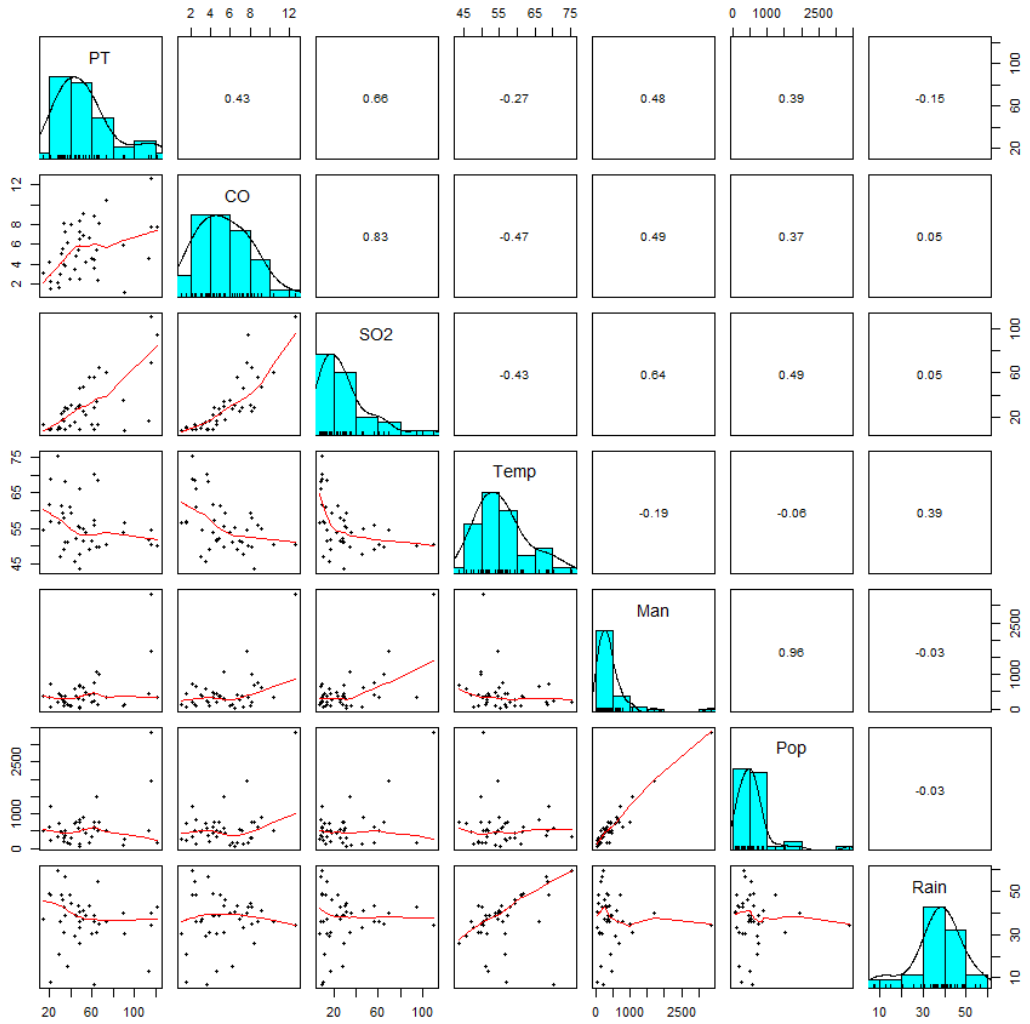


圖 3: Scatter Plot Matrices of Sales

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
```

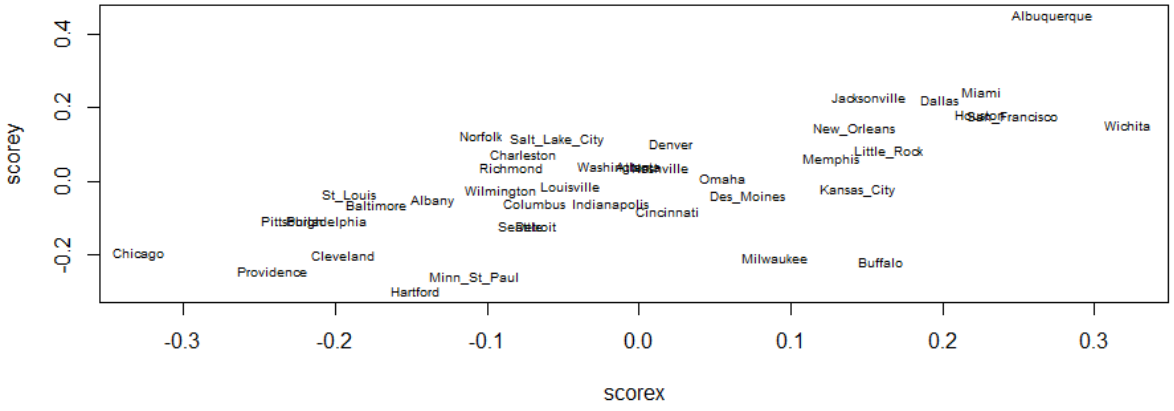


圖 4: 2D plot of CCA

**Q3.** The "Car Family.txt" data in the $4 \times 3$ contingency table describes the frequencies regarding "origin of cars" and "the owners' family status". The detailed description of the variables is given below.

The following three variables indicate the origin of the car:

- American
- Japanese

- European

The following four variables indicate the family status:

- Married
- Married living with kids
- Single
- Single living with kids

Perform a Simple Correspondence Analysis on this data set and interpret the result.

**Ans.** 資料呈現如下，由於此資料在某個細格的數量極低，進行卡方檢定可能不是好的選擇，爲了近一步了解兩個變數的關係，我們進行 CA 分析。

```
-------------------------------------------------------
                          |        Origin of Cars       |
Family                    |=============================|
Status                    | American  European Japanese |
-------------------------------------------------------|
Married                   |    37        14       51     |
Married living with kids  |    52        15       44     |
Single                    |    33        15       63     |
Single living with kids   |     6         1        8     |
-------------------------------------------------------

 Principal inertias (eigenvalues):
            1        2
Value       0.022866 0.001764
Percentage  92.84%   7.16%
```

從上方列連表中可以看出一些端倪，像是美國家庭型態大多是有家庭且有小孩，相較之下日本是單身居多，可以預計此資料降維的結果還不錯。從第一個 inertias 所解釋的比例高達 93% 可以應證。爲了將資料呈現在 2D 上，這邊取兩個 inertias，其與各個水準的相關距離如下。

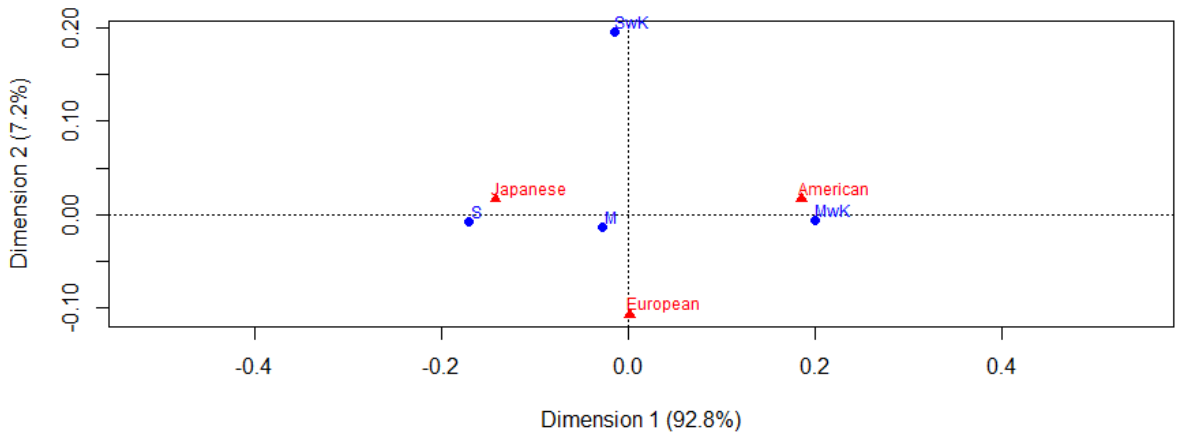


圖 5: 2D plot of CA

先看 Family 的水準之間關係，可以發現單親家庭 (SwK) 比起其他家庭來說是相對較不一樣的群體，而 Origin 的水準之間關係，發現美國與日本差距最大，歐洲則是介於兩者之間。若是同時考慮 Family 和 Origin 的關係，可以發現單親家庭幾乎都不會有車子，而美國通常有車子的家庭都是有小孩的家庭，日本則是單身。

**Q4.** The file ”Internet Shopping.txt” contains 1127 internet consumers’ responses to a survey with 19 questions. Each response in Questions 1-17 has 5 levels, while each response in Questions 18-19 has 6 levels. Perform a trustful Multiple Correspondence Analysis on this data set and interpret the result.

The data come from answering the following 19 survey questions.

A. Possible answers for Q1): 1 means "<$50", 2 means "$50-$100", 3 means "$100-$500", 4 means ">$500", 5 means "don't know".
   q1. What is the total amount you spent on purchases through internet during the past 6 months?

B. Possible answers for Q2)-Q6): 1 means "very unlikely", 2 means "somewhat unlikely", 3 means "neither unlikely nor likely", 4 means "somewhat likely", 5 means "very likely".
   q2. You provide credit card and purchase information through a toll call/fax.
   q3. You provide credit card and purchase information through a toll-free call/fax.
   q4. You set up an account with the vendor once, then provide an account number and purchase information each time you make a purchase.
   q5. You provide credit card and purchase information through e-mail.
   q6. Any subsequent transmissions of your credit card and purchase information would be secure.

C. Possible answers for Q7)-Q17): 1 means "strongly disagree", 2 means "somewhat disagree", 3 means "neither disagree nor agree", 4 means "somewhat agree", 5 means "strongly agree".
   q7. Providing credit card information through the Web is the single most important reason I don't buy through the Web often.
   q8. Providing credit card information through the Web is not riskier than providing it over the phone.
   q9. I would be more willing to provide my credit card information through the Web if the prices are much lower.
   q10. I would be more willing to provide my credit card information through the Web if the products/services were of a higher quality.
   q11. I would be more willing to provide my credit card information through the Web if the Web vendor was well known and reliable.
   q12. WWW vendors offer more useful information about the choices available.
   q13. It is easier to place orders with WWW vendors.
   q14. It is easier to cancel orders placed with WWW vendors.
   q15. It is just as safe to use credit cards when making purchases from WWW vendors.
   q16. WWW vendors offer better prices.
   q17. It is easier to contact WWW vendors.

D. Possible answers for Q18)-Q19): 1 means "never", 2 means "<1/month", 3 means "1-2/month", 4 means "3-5/month", 5 means "6-9/month", 6 means ">10/month".
   q18. How often do you use the Web for shopping for personal reasons?
   q19. How often do you use the Web for shopping for professional reasons?

圖 6: Internet Shopping Variables

**Ans.** 此題含有 19 個變數，每個變數的水準至少有 5 個，想要直接探求這 19 個變數之間的關係可能有點難，因此我們進行 MCA 去降維，用兩個維度去看這 19 個變數之間的關係，以下是使用「JCA」方法運行出來的結果，在四個 MCA 方法中，這個方法在兩個維度解釋的總變異最高，爲 64%。

```
Principal inertias (eigenvalues):

 dim     value
 1       0.056894
 2       0.022822
 .         ...
 28      1e-05000
```

```
        --------
Total: 0.121052

Diagonal inertia discounted from eigenvalues: 0.0061507
Percentage explained by JCA in 2 dimensions: 64%
(Eigenvalues are not nested)
[Iterations in JCA: 11 , epsilon = 3.64e-05]
```

接著進一步分析題項本身的差異，圖 7 中大多藍色度部分都是選擇 1；綠色是選擇 5 左右，說明 X 軸應表示每題符合不符合、或是同意不同意的差距，其中又以認同 Q15 與反對 Q11、Q13 的差距極大，剛好這幾題皆是是否信任網站購物的正反敍述，因此 X 軸也反映了網路信任之差異性，而 Y 軸則稍加紊亂些，但仍可以看出反對 Q11 與符合 Q6 差異很大，但大多數人都支持 Q6，少數人反對 Q11，因此 Y 或許反映的是人數之間的差異。

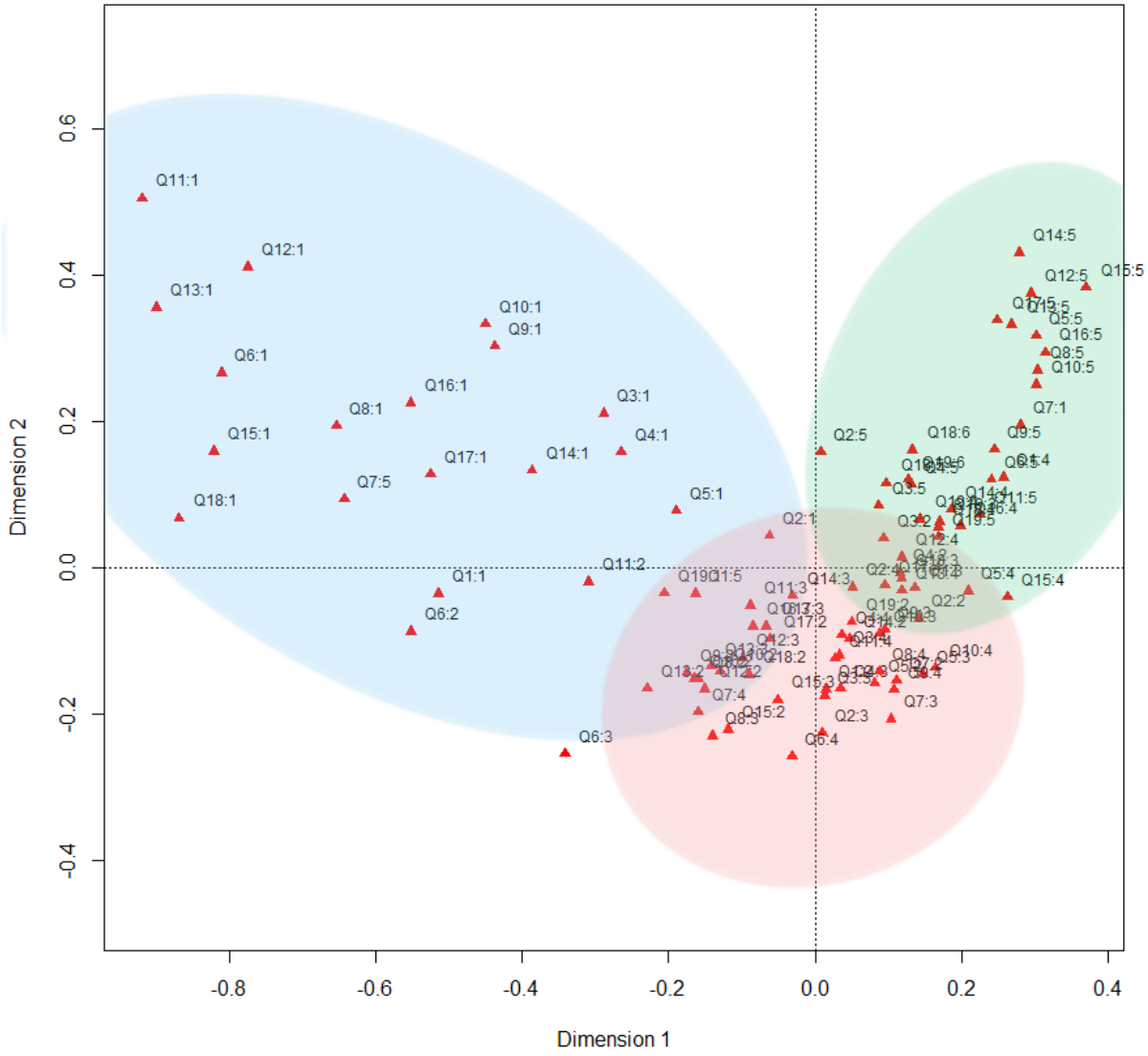


圖 7: 2D plot of MCA