Suppose a human resource researcher would like to study the relationships between the employment proportions of various types of industry in Europe. He divides the variables in the data “European\_Jobs.txt” into two groups: (Agr, Min, Man, PS, Con) and (SI, Fin, SPS, TC). The first group represents the industries that are more , while the second group represents the industries that are less labor-intensive.

Q2: Perform a complete Canonical Correlation Analysis for these two groups of variables and interpret the result. Note: The number of canonical variates must be determined by a formal statistical hypothesis test, while the required model assumptions need to be validated.

本題使用的資料同第一題所使用，26筆樣本資料與10個變數，其中1個變數為國家名稱，其餘9個為該國從事該產業的人口百分比。本題將變數分為兩類：產業屬於勞力密集型的變數集合(Agr, Min, Man, PS, Con)與產業屬於勞動較少產業(SI, Fin, SPS, TC)的變數集合。

|  |  |
| --- | --- |
| 產業屬於勞力密集型的資料 | 產業屬於勞動較少行業的資料 |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Country | Agr | Min | Man | PS | Con | | Belgium | 3.3 | 0.9 | 27.6 | 0.9 | 8.2 | | Denmark | 9.2 | 0.1 | 21.8 | 0.6 | 8.3 | | France | 10.8 | 0.8 | 27.5 | 0.9 | 8.9 | | W. Germany | 6.7 | 1.3 | 35.8 | 0.9 | 7.3 | | Ireland | 23.2 | 1 | 20.7 | 1.3 | 7.5 | | |  |  |  |  |  | | --- | --- | --- | --- | --- | | Country | SI | Fin | SPS | TC | | Belgium | 19.1 | 6.2 | 26.6 | 7.2 | | Denmark | 14.6 | 6.5 | 32.2 | 7.1 | | France | 16.8 | 6 | 22.6 | 5.7 | | W. Germany | 14.4 | 5 | 22.3 | 6.1 | | Ireland | 16.8 | 2.8 | 20.8 | 6.1 | |

第一組資料包含6個變數，包含國家名稱，以及5個勞力密集型產業(Agr, Min, Man, PS, Con)

第二組資料包含5個變數，包含國家名稱，以及4個勞動較少產業(SI, Fin, SPS, TC)

在第一題之中，我們使用principle components方法縮減變異數，讓資料更為容易解釋、繪圖。而本題之中要使用的縮減變異數方法為Canonical Correlation Analysis(CCA)，透過事先將變數分為兩組並定義每一組變數集合，接著對這兩組變數組合進行CCA。CCA是研究兩組變數之間相關關係的一種多元統計方法，它能解釋兩組變數之間的內在關聯同時達到降低維度的目的。

Canonical Correlation Analysis(CCA)：基本原理

Canonical Correlation Analysis(CCA)是指利用變數組合之間的相關係數來反映兩組指標（變數線性組合）之間的整體相關性的多元[統計分析方法](http://wiki.mbalib.com/zh-tw/%E7%BB%9F%E8%AE%A1%E5%88%86%E6%9E%90%E6%96%B9%E6%B3%95" \o "统计分析方法)。它的基本原理是：為了從總體上分析兩組指標之間的相關關係，分別在兩組[變數](http://wiki.mbalib.com/zh-tw/%E5%8F%98%E9%87%8F" \o "变量)中提取有代表性的兩個變數組合U1和V1（分別為兩個變數組中各變數的線性組合），利用這兩個線性組合的相關關係來解釋兩組指標之間的整體相關性。

Canonical Correlation Analysis(CCA)流程

Step1. 資料標準化，並使用標準化的資料進行以下動作

Step2. 檢定資料有outlier以及是否為常態

Step3. 將資料分為兩組，並定義各組

Step4. 檢測各組變數之間是否存在高度相關係數

Step5. 進行Canonical Correlation Analysis(CCA)維度縮減

Step6. 選擇合適的成對線性組合並解釋其意義

從上表各組資料得知，每個變數的範圍皆不相同，為避免變數的變異數過大影響分析誤差，我們將資料標準化，再對各組做相關係數分析，若是任兩變數存在高度線性關係（相關係數 > 0.98 或

相關係數 < -0.98 )，則考慮是否要移除其中一變數。

資料標準化

第一組

|  |  |
| --- | --- |
| 原始資料 | 標準化後資料 |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Country | Agr | Min | Man | PS | Con | | Belgium | 3.3 | 0.9 | 27.6 | 0.9 | 8.2 | | Denmark | 9.2 | 0.1 | 21.8 | 0.6 | 8.3 | | France | 10.8 | 0.8 | 27.5 | 0.9 | 8.9 | | W. Germany | 6.7 | 1.3 | 35.8 | 0.9 | 7.3 | | Ireland | 23.2 | 1 | 20.7 | 1.3 | 7.5 | | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Country | Agr | Min | Man | PS | Con | | Belgium | -1.0183 | -0.3648 | 0.0845 | -0.0204 | 0.0210 | | Denmark | -0.6388 | -1.1895 | -0.7431 | -0.8179 | 0.0818 | | France | -0.5359 | -0.4679 | 0.0703 | -0.0204 | 0.4464 | | W. Germany | -0.7996 | 0.0476 | 1.2547 | -0.0204 | -0.5259 | | Ireland | 0.2617 | -0.2617 | -0.9001 | 1.0428 | -0.4043 | |

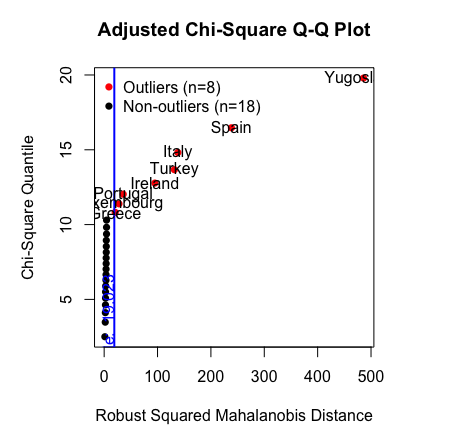
第二組

|  |  |
| --- | --- |
| 原始資料 | 標準化後資料 |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | Country | SI | Fin | SPS | TC | | Belgium | 19.1 | 6.2 | 26.6 | 7.2 | | Denmark | 14.6 | 6.5 | 32.2 | 7.1 | | France | 16.8 | 6 | 22.6 | 5.7 | | W. Germany | 14.4 | 5 | 22.3 | 6.1 | | Ireland | 16.8 | 2.8 | 20.8 | 6.1 | | |  |  |  |  |  | | --- | --- | --- | --- | --- | | Country | SI | Fin | SPS | TC | | Belgium | 1.3425 | 0.7839 | 0.9630 | 0.4699 | | Denmark | 0.3590 | 0.8908 | 1.7830 | 0.3980 | | France | 0.8398 | 0.7126 | 0.3773 | -0.6081 | | W. Germany | 0.3152 | 0.3563 | 0.3334 | -0.3206 | | Ireland | 0.8398 | -0.4276 | 0.1138 | -0.3206 | |

檢測資料是否為常態

(一)檢測離群值

使用Robust Squared Mahalanobis Distance方法檢測是否離群值存在，應盡量避免離群值存在，以避免資料解釋誤差產生。



圖上顯示有8筆資料為離群值，如Spain, Turkey, …等8筆資料(紅點)，但這筆資料的樣本數過少，若是任意將這8筆資料刪除，則資訊量會減少30％。由於刪除離群值會造成資訊量損失過多，因此我們考慮保留這8筆離群值。

(二) 檢定資料是否多維常態

在給定顯著水準為0.05下，分別以不同的檢測方式檢測

|  |  |  |
| --- | --- | --- |
| Test | p-value | Result |
| Mardia Skewness test | 0.021617 | Reject |
| Mardia Kurtosis test | 0.456601 | Do not reject |
| Henze-Zirkler multinormal test | 0.699563 | Do not reject |
| Royston multinormal test | 0.000084 | Reject |
| Dorni-Haansen's multinormal test | 0.331796 | Do not reject |
| E-statistic multinormal test | 0.028 | Reject |

上述方法有的拒絕的假設，有的則不拒絕的假設。為了接下來分析以及縮減為度方便，我們假設資料為多元常態。

看看各組之相關係數

第一組

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | Agr | Min | Man | PS | Con | | Agr | 1 | 0.0358 | -0.6711 | -0.4001 | -0.5383 | | Min | 0.0358 | 1 | 0.4452 | 0.4055 | -0.0256 | | Man | -0.6711 | 0.4452 | 1 | 0.3853 | 0.4945 | | PS | -0.4001 | 0.4055 | 0.3853 | 1 | 0.0599 | | Con | -0.5383 | -0.0256 | 0.4945 | 0.0599 | 1 | |  |

(圖形顏色越深越大代表相關係數越大，藍色為正相關，紅色為負相關)

第二組

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | SI | Fin | SPS | TC | | SI | 1 | 0.3656 | 0.5722 | 0.1876 | | Fin | 0.3656 | 1 | 0.1076 | -0.2459 | | SPS | 0.5722 | 0.1076 | 1 | 0.5679 | | TC | 0.1876 | -0.2459 | 0.5679 | 1 | |  |

(圖形顏色越深越大代表相關係數越大，藍色為正相關，紅色為負相關)

兩組之相關係數

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | SI | Fin | SPS | TC | | Agr | -0.7370 | -0.2198 | -0.7468 | -0.5649 | | Min | -0.3966 | -0.4427 | -0.2810 | 0.1566 | | Man | 0.2038 | -0.1558 | 0.1542 | 0.3507 | | PS | 0.2019 | 0.1099 | 0.1324 | 0.3752 | | Con | 0.3560 | 0.0163 | 0.1582 | 0.3877 | |  |

(圖形顏色越深越大代表相關係數越大，藍色為正相關，紅色為負相關)

從各組之相關係數觀察，並未出現任兩變數相關係大於0.98或是小於-0.98，僅有農業與服務業、社會與個人服務業呈現中度負相關。因此我們不考慮刪減變數。

進行Canonical Correlation Analysis(CCA)維度縮減分析

透過R運算CCA，結果如下

第一組變數組合之線性組合係數

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | [1] | [2] | [3] | [4] |
| Agr | -0.2658 | -0.1041 | 0.1771 | 0.0920 |
| Min | -0.0160 | -0.0016 | -0.2400 | -0.1238 |
| Man | -0.1206 | -0.0191 | 0.0564 | 0.2263 |
| PS | -0.0064 | -0.1490 | 0.1486 | -0.1078 |
| Con | -0.0272 | -0.1815 | 0.0289 | 0.0439 |

第二組變數組合之線性組合係數

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | [1] | [2] | [3] | [4] |
| SI | 0.0786 | -0.1165 | 0.0393 | 0.2195 |
| Fin | 0.0478 | -0.0313 | 0.1616 | -0.1527 |
| SPS | 0.1157 | 0.2453 | -0.0960 | -0.0786 |
| TC | 0.0238 | -0.2375 | -0.0301 | -0.1059 |

第一組與第二組對應線性組合之相關係數

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | [1] | [2] | [3] | [4] |
| Canonical Correlation() | 0.9999 | 0.6623 | 0.4352 | 0.2152 |

(六)

從Step5中得到4組成對線性組合以及其相關係數，但並非每一組成對線性組合都合適，要檢測每一組成對線性組合的Correlation是否顯著。

使用Wilk’s test進行檢定，而在使用Wilk’s test檢定之前要滿足幾個條件：(1)資料樣本夠多、(2)資料服從多維常態。由於這筆資料僅有26筆樣本，為了檢測方便，我們認為(1)有滿足，而資料服從多維常態則在之前檢測資料是否為多維常態時，假定資料為多維常態。

Wilk’s test: 檢定是否第i個Canonical Correlation為顯著

===0

==0

F =

Where

。

Reject when

檢測本題之Canonical Correlation是否顯著

===0

==0

Wilks' Lambda, using F-approximation (Rao's F):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | stat | approx | df1 | df2 | p.value |
| 1 to 4: | 5.3764E-05 | 52.6824124 | 20 | 57.3325 | 0 |
| 2 to 4: | 0.43395438 | 1.48132 | 12 | 47.91503 | 0.1644588 |
| 3 to 4: | 0.77307874 | 0.8697839 | 6 | 38 | 0.5259317 |
| 4 to 4: | 0.95367289 | 0.4857757 | 2 | 20 | 0.6222922 |

在給定顯著水準為0.05下，僅有1 to 4顯著，即只有為顯著，其餘都不顯著。

經過Wilk’s test我們選擇第一組成對線性組合係數

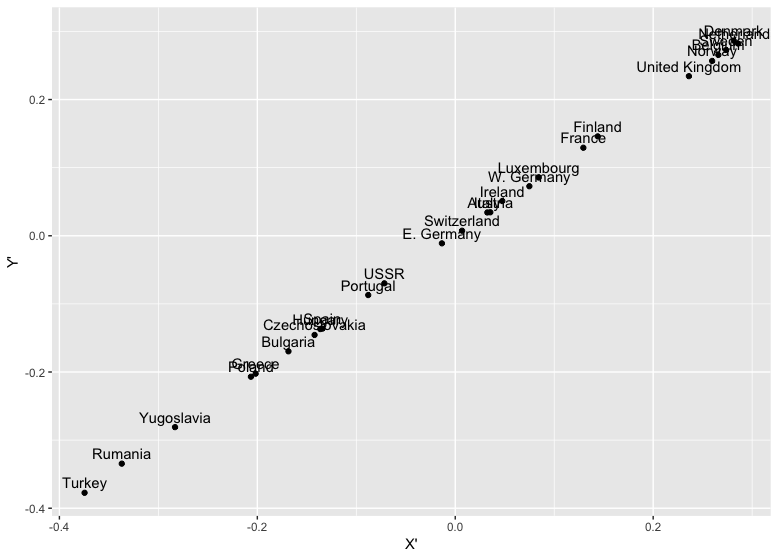
X’ = -0.2658Agr -0.016Min – 0.1206Man – 0.0064PS – 0.0272Con

Y’ = 0.0786SI + 0.0478Fin + 0.1157SPS + 0.0238TC

|  |  |
| --- | --- |
| X標準化後資料 | Y標準化後資料 |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Country | Agr | Min | Man | PS | Con | | Belgium | -1.0183 | -0.3648 | 0.0845 | -0.0204 | 0.0210 | | Denmark | -0.6388 | -1.1895 | -0.7431 | -0.8179 | 0.0818 | | France | -0.5359 | -0.4679 | 0.0703 | -0.0204 | 0.4464 | | W. Germany | -0.7996 | 0.0476 | 1.2547 | -0.0204 | -0.5259 | | Ireland | 0.2617 | -0.2617 | -0.9001 | 1.0428 | -0.4043 | | |  |  |  |  |  | | --- | --- | --- | --- | --- | | Country | SI | Fin | SPS | TC | | Belgium | 1.3425 | 0.7839 | 0.9630 | 0.4699 | | Denmark | 0.3590 | 0.8908 | 1.7830 | 0.3980 | | France | 0.8398 | 0.7126 | 0.3773 | -0.6081 | | W. Germany | 0.3152 | 0.3563 | 0.3334 | -0.3206 | | Ireland | 0.8398 | -0.4276 | 0.1138 | -0.3206 | |

|  |  |  |
| --- | --- | --- |
|  | X’ | Y’ |
| Belgium | 0.2658 | 0.2655 |
| Denmark | 0.2814 | 0.2865 |
| France | 0.1294 | 0.1293 |
| W. Germany | 0.0749 | 0.0728 |
| Ireland | 0.0475 | 0.0510 |

將資料投影到新的座標軸(X’, Y’)



解釋：

將X’以及Y’的係數取絕對值後，看大於0.1的部分，X’為Agr以及Man為最主要影響變數，而Y’則是SPS為最主要影響變數。在決定X’與Y’兩者線性組合係數時使得X’與Y’的相關係數最大（本題Corr(X’, Y’) = 0.9999），因此將資料投影到X’與Y’座標時，才會呈現斜直線的形式，才顯示這筆資料的X’與Y’呈現強力正相關。當X’越大時，Agr與Man越小，同時Y’與SPS也越大，即勞動力密集型的產業人口百分比越少時，從事勞動較少產業人口越多，反之則是。