**Q1**: Perform a complete **Principal Components Analysis** for this data and interpret the result. Note: The number of PCs must be determined by a formal statistical hypothesis test, while the relationships among objects and variables can be interpreted by using a 2D plot.

先看看資料的樣式  
原始資料

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Country | Agr | Min | Man | PS | Con | SI | Fin | SPS | TC |
| Belgium | 3.3 | 0.9 | 27.6 | 0.9 | 8.2 | 19.1 | 6.2 | 26.6 | 7.2 |
| Denmark | 9.2 | 0.1 | 21.8 | 0.6 | 8.3 | 14.6 | 6.5 | 32.2 | 7.1 |
| France | 10.8 | 0.8 | 27.5 | 0.9 | 8.9 | 16.8 | 6.0 | 22.6 | 5.7 |
| W. Germany | 6.7 | 1.3 | 35.8 | 0.9 | 7.3 | 14.4 | 5.0 | 22.3 | .6.1 |

(僅顯示前四資料)

資料總共有10個變數，26筆樣本資料，其中Country為國家的名稱，其餘變數代表該國從事該職業的人口百份比(上表只顯示4筆國家與其產業人口百分比資料)。

變數介紹

|  |  |  |
| --- | --- | --- |
| 變數名稱 | 變數解釋 | 變數型態 |
| Country | 國家名稱 | 類別 |
| Agr | 該國從事農業的人口百分比 | 數值 |
| Min | 該國從事礦業的人口百分比 | 數值 |
| Man | 該國從事製造業的人口百分比 | 數值 |
| PS | 該國從事能源業的人口百分比 | 數值 |
| Con | 該國從事建築業的人口百分比 | 數值 |
| SI | 該國從事服務業的人口百分比 | 數值 |
| Fin | 該國從事金融業的人口百分比 | 數值 |
| SPS | 該國從事社會與個人服務的人口百分比 | 數值 |
| TC | 該國從事交通運輸業的人口百分比 | 數值 |

9個變數皆為數值型態，但每個變數的範圍都不大相等。因此，再進行分析前要先將資料標準化，藉此消除掉變異程度過大的變數，降低分析誤差。

資料標準化

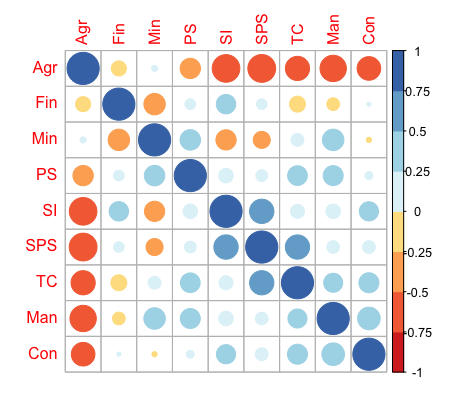
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Agr | Min | Man | PS | Con | SI | Fin | SPS | TC |
| Belgium | -1.0183 | -0.3648 | 0.0845 | -0.0204 | 0.0210 | 1.3425 | 0.7839 | 0.9630 | 0.4699 |
| Denmark | -0.6388 | -1.1895 | -0.7431 | -0.8179 | 0.0818 | 0.3590 | 0.8908 | 1.7830 | 0.3980 |
| France | -0.5359 | -0.4679 | 0.0703 | -0.0204 | 0.4464 | 0.8398 | 0.7126 | 0.3773 | -0.6081 |
| W. Germany | -0.7996 | 0.0476 | 1.2547 | -0.0204 | -0.5259 | 0.3152 | 0.3563 | 0.3334 | -0.3206 |

(僅顯示前面四筆)

相關係數矩陣

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Agr | Min | Man | PS | Con | SI | Fin | SPS | TC |
| Agr | 1.0000 | 0.0358 | -0.6711 | -0.4001 | -0.5383 | -0.7370 | -0.2198 | -0.7468 | -0.5649 |
| Min | 0.0358 | 1.0000 | 0.4452 | 0.4055 | -0.0256 | -0.3966 | -0.4427 | -0.2810 | 0.1566 |
| Man | -0.6711 | 0.4452 | 1.0000 | 0.3853 | 0.4945 | 0.2038 | -0.1558 | 0.1542 | 0.3507 |
| PS | -0.4001 | 0.4055 | 0.3853 | 1.0000 | 0.0599 | 0.2019 | 0.1099 | 0.1324 | 0.3752 |
| Con | -0.5383 | -0.0256 | 0.4945 | 0.0599 | 1.0000 | 0.3560 | 0.0163 | 0.1582 | 0.3877 |
| SI | -0.7370 | -0.3966 | 0.2038 | 0.2019 | 0.3560 | 1.0000 | 0.3656 | 0.5722 | 0.1876 |
| Fin | -0.2198 | -0.4427 | -0.1558 | 0.1099 | 0.0163 | 0.3656 | 1.0000 | 0.1076 | -0.2459 |
| SPS | -0.7468 | -0.2810 | 0.1542 | 0.1324 | 0.1582 | 0.5722 | 0.1076 | 1.0000 | 0.5679 |
| TC | -0.5649 | 0.1566 | 0.3507 | 0.3752 | 0.3877 | 0.1876 | -0.2459 | 0.5679 | 1.0000 |

以圖表表示相關係數矩陣



(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)

由相關係數表得知每兩個變數之間的相關係數都未呈現強力的正相關或負相關(相關係數 > 0.98或 <-0.98)，而從事農業人口百分比與服務業、社會與個人服務、交通運輸業、製造業與建築業呈現中度負相關(相關係數為-0.5 ~ -0.7)，推測可能與國家主要經濟來源以及開發程度有相關。

這筆資料用於分析的變數為9個，過多的變數要用於統整分析資料上會有困難，不易解釋變數之間相互的影響性，同時也不易用圖表呈現。因此，我們運用除了國家名之外的9個變數去做組成分分析(Principal Components Analysis, PCA)，將9個變數透過PCA的方式縮減維度，同時保持數據中的對變異數貢獻最大的特徵。

Principal Components Analysis 流程

Step1. 檢定資料是否為多維常態

Step2. 檢測是否有離群值離群值(outlier)存在

Step3. 將資料的相關係數矩陣進行特徵值分解，取得特徵值（）與特徵向量（）

Step4. 選擇合適的principal components組合

Step5. 計算principal components (由資料內積特徵向量)、將資料投影到新的components上，並解釋其意義。

(一) 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

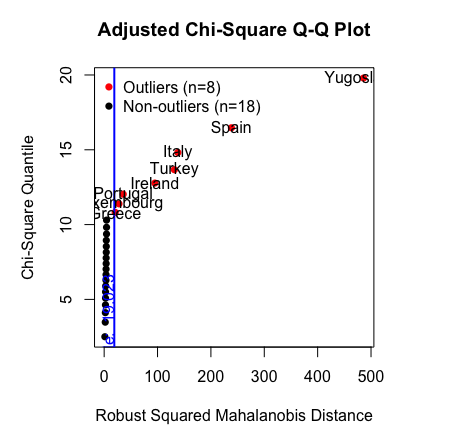
|  |  |  |
| --- | --- | --- |
| Test | p-value | Result |
| Mardia Skewness test | 0.021617 | Reject |
| Mardia Kurtosis test | 0.456601 | Do not reject |
| Henze-Zirkler multinormal test | 0.699563 | Do not reject |
| Royston multinormal test | 0.000084 | Reject |
| Dorni-Haansen's multinormal test | 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。因此我們假設資料為多元常態。

(二)檢測離群值

由於PCA是根據變異數去做縮間維度的方法，因此離群值對於PCA方法會有很大的影響。

接著使用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，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。

(三)計算特徵向量，也就是各個主成份，所對應的線性組合(linear combination)的係數、累積變異程度

接下來，我們透過Ｒ計算PCA，得到下列幾種組合以及他們解釋變異的程度

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 | Comp.7 | Comp.8 | Comp.9 |
| Agr | -0.5238 | -0.0536 | 0.0487 | -0.0288 | 0.2127 | -0.1533 | 0.0213 | 0.0079 | 0.8064 |
| Min | -0.0013 | -0.6178 | -0.2011 | -0.0641 | -0.1637 | 0.1006 | -0.7257 | 0.0884 | 0.0486 |
| Man | 0.3475 | -0.3551 | -0.1505 | 0.3461 | -0.385 | 0.2882 | 0.4794 | 0.1258 | 0.366 |
| PS | 0.2557 | -0.2611 | -0.5611 | -0.3933 | 0.2952 | -0.3573 | 0.2556 | -0.3412 | 0.0194 |
| Con | 0.3252 | -0.0513 | 0.1533 | 0.6683 | 0.4716 | -0.1304 | -0.2207 | -0.3557 | 0.0826 |
| SI | 0.3789 | 0.3502 | -0.1151 | 0.0502 | -0.2836 | -0.6148 | -0.2294 | 0.3875 | 0.2383 |
| Fin | 0.0744 | 0.4537 | -0.5874 | 0.0516 | 0.2796 | 0.5256 | -0.1875 | 0.1743 | 0.1452 |
| SPS | 0.3874 | 0.2215 | 0.3119 | -0.4122 | -0.2204 | 0.2629 | -0.1913 | -0.5062 | 0.3509 |
| TC | 0.3668 | -0.2026 | 0.3751 | -0.3144 | 0.5129 | 0.124 | 0.0682 | 0.5446 | 0.0721 |
| Standard deviation | 1.86739 | 1.45951 | 1.04831 | 0.99724 | 0.73703 | 0.61922 | 0.47514 | 0.36985 | 0.00675 |
| Eigenvalue（） | 3.48715 | 2.13017 | 1.09896 | 0.99448 | 0.54322 | 0.38343 | 0.22575 | 0.13679 | 0.00005 |
| Proportion of Variance | 0.38746 | 0.23669 | 0.12211 | 0.1105 | 0.06036 | 0.0426 | 0.02508 | 0.0152 | 5.1\* |
| Cumulative Proportion | 0.38746 | 0.62415 | 0.74625 | 0.85675 | 0.91711 | 0.95971 | 0.9848 | 0.99999 | 1 |

（四）選取合適components 係數

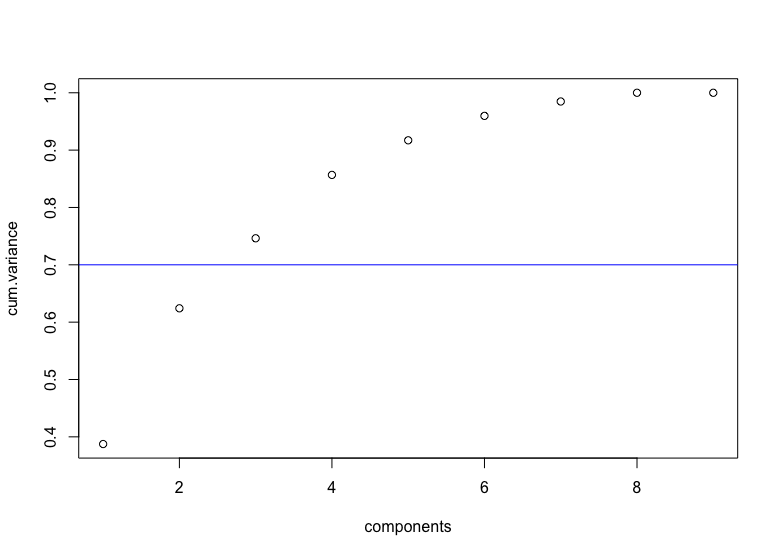
由上表去判斷要使用第幾個principal components:

1. 依據學者Kaiser提出的準則，取用 > 1的components組合，則會有components1, components2以及components3作為縮減後的新維度。



（藍色線為Eigenvalue為1）

2. 依據累積總變異數比例至少70%準則，去選取components係數組合，則會有components1, components2以及components3作為縮減後的新維度。



3. 使用Permutation test，將每一個變數的資料隨機排列（破壞correlation structure），重新計算每個特徵值（）並記錄起來，重複上述動作多次後，繪製長條圖去看原始資料的特徵值的位置。若是原始的特徵值越大，則所對應的p–value則會越小。

Permutation test的假設為

經過計算後的得到components1 ~ components9的p - value

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 | Comp.7 | Comp.8 | Comp.9 |
| P - value | 0.000 | 0.000 | 0.990 | 0.915 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

在給定顯著水準為0.05下，僅有component1以及component2有足夠的證據拒絕。因此在Permutation test下選擇component1與component2作為縮減後的二維度變數。

綜合這三中方法，前兩種方法在取捨component上較具有主觀意識，而Permutation test則是透過統計檢定方式去做principal components取捨。為了讓實驗更加客觀，我們選擇由Permutation test的取捨方式，選擇component1與component2為縮減後的二維度變數。

經過PCA縮減維度後，原始資料由9個變數構成(Agr, Min, … ,TC)降維度至2個變數（component1以及component2），而用這個兩個成分所能解釋的總變異數仍有62%，為可接受範圍。

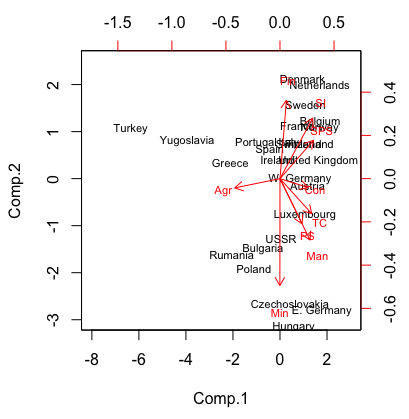
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Component | Agr | Min | Man | PS | Con | SI | Fin | SPS | TC |
| 1 | -0.5238 | -0.0013 | 0.3475 | 0.2557 | 0.3252 | 0.3789 | 0.0744 | 0.3874 | 0.3668 |
| 2 | -0.0536 | 0.6178 | 0.3551 | -0.2611 | -0.0513 | 0.3502 | 0.4537 | 0.2215 | -0.2026 |

（五）將資料投影至component1以及component2上

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Agr | Min | Man | PS | Con | SI | Fin | SPS | TC |
| Belgium | -1.01828 | -0.36477 | 0.08452 | -0.02045 | 0.02104 | 1.34251 | 0.78388 | 0.96301 | 0.46990 |
| Denmark | -0.63878 | -1.18948 | -0.74313 | -0.81786 | 0.08180 | 0.35895 | 0.89077 | 1.78298 | 0.39803 |
| France | -0.53586 | -0.46786 | 0.07025 | -0.02045 | 0.44642 | 0.83980 | 0.71262 | 0.37732 | -0.60810 |
| W. Germany | -0.79958 | 0.04758 | 1.25465 | -0.02045 | -0.52588 | 0.31524 | 0.35631 | 0.33339 | -0.32064 |
| Ireland | 0.26174 | -0.26169 | -0.90010 | 1.04277 | -0.40435 | 0.83980 | -0.42757 | 0.11376 | -0.32064 |
| Italy | -0.20781 | -0.67404 | 0.08452 | -1.08367 | 1.11487 | 1.12394 | -0.85514 | 0.01126 | -0.60810 |
| Luxembourg | -0.73526 | 1.90317 | 0.54116 | -0.28625 | 0.62872 | 1.21137 | 0.21378 | -0.12052 | -0.24877 |
| Netherlands | -0.82531 | -1.18948 | -0.64324 | 0.24536 | 1.05410 | 1.10208 | 0.99766 | 1.24121 | 0.18243 |
| United Kingdom | -1.05687 | 0.15067 | 0.45554 | 1.30858 | -0.76896 | 0.86166 | 0.60572 | 1.21193 | -0.10504 |
| Austria | -0.41365 | -0.15860 | 0.45554 | 1.30858 | 0.50718 | 0.83980 | 0.32068 | -0.47193 | 0.32616 |

|  |  |  |
| --- | --- | --- |
|  | Comp.1 | Comp.2 |
| Belgium | 1.71050 | 1.22179 |
| Denmark | 0.95290 | 2.12778 |
| France | 0.75463 | 1.12121 |
| W. Germany | 0.85255 | 0.01138 |
| Ireland | -0.10350 | 0.41399 |
| Italy | 0.37541 | 0.76955 |
| Luxembourg | 1.05944 | -0.75583 |
| Netherlands | 1.68822 | 2.00484 |
| United Kingdom | 1.63045 | 0.37313 |
| Austria | 1.17645 | -0.14310 |

繪製2維principal components圖並解釋



先看水平軸components1，數值越大表示從事農業人口百分比越低而從事建築業、服務業、製造業等的人口百分比越高，反之則相反，但從事金融業以及礦業則是不受影響。

Ex. Turkey中從事農業人口比例有66％其餘產業人口則不高，而United Kingdom從事農業人口僅有2.7%但從事製造業、服務業性質的人口比例就很高

接著看垂直軸components2，數值越大表示從事金融業、服務業、交通運輸的人口百分比越高而礦業、製造業、能源業等人口百分比越低，反之則相反，同樣的農業以及建築業則不受影響。

Ex. E. Germany是屬於製造業大國，約有41.2%的人從事該產業。

整個2維principal components圖來解釋，越靠近紅線表該產業人口所占比例越大，越右上角的國家，代表金融、服務業、交通運輸業較為發達，因此從事人口百分比較高，但礦業、製造業則越少，意味非勞力密集型的國家，可能為已開發國家居多。Ex. Sweden、France與United Kingdom。而右下角的國家則是顯示從事製造業、礦業以及能源業居多，可能開發中國家較多，勞力成本較低，大多公司的廠區多建在該國，但也有可能是該國本身的天然優勢使然。Ex. Rumania、Hungary與E. Germany。