多變量期末報告

統研碩一
106354019 黃意珺
統研碩一
1063540 董承

目錄

(1)

The data set “Seeds.txt” describes seven measured geometric parameters of 205 wheat kernels which belong to three different types of wheat: Kama (Y=1), Rosa (Y=2) and Canadian (Y=3). Please download the file “Seeds_Variables.txt” for the detailed description.

Q1: Perform clustering analysis based on the attributes X1, …, X7 by using the following methods: (1) The hierarchical tree of Agglomerative Nesting by choosing the most adequate linkage; (2) The K-medoids; (3) The self-organizing maps (SOM).

Q2: Does the clustering result recover mostly the original wheat type for each of the above method?

題目描述

共有7個自變數及1個因變數，其中自變數為連續變數，因變數為類別變數。希望我們利用這些小麥仁的幾何變數將小麥的品種做分群。

資料描述

分析資料為205筆小麥仁的數據，小麥總共有三個品種Kama, Rosa, Canadian，以及七個小麥仁的幾何數據做為自變數，其中三個品種的各數分別為69,68及68個。

變數	變數解釋
X_1	面積(mm^2)
X_2	周長(mm)
X_3	緊密度$(4\pi$面積/周長$^2)$
X_4	仁的長度(mm)
X_5	仁的寬度(mm)
X_6	不對稱係數
X_7	仁的溝槽長度(mm)
Y	類別，Kama(Y=1)Rosa(Y=2)Canadian(Y=3)

表1. Seeds.txt 變數解釋.

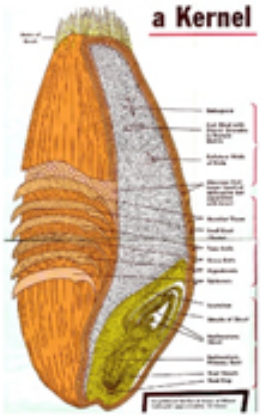


圖1. 小麥仁.

方法如何運作

Hierarchical clustering

階層式分群法（Hierarchical Clustering）透過一種階層架構的方式，將資料層層反覆地進行分裂（Divisive）或聚合（Agglomerative），以產生最後的樹狀結構。

分裂（Divisive）：由樹狀結構的頂端開始（Top Down），將群聚逐次分裂。

聚合（Agglomerative）：由樹狀結構的底部開始(Bottom Up)，將資料或群聚逐次合併。

本次題目使用聚合式階層分群法（Agglomerative Hierarchical Clustering）對資料做分群，也就是由樹狀結構的底部開始層層聚合。

操作步驟：

(1) 自己決定K個群數

(2) 將每筆資料視為一個群聚 $C_i, i = 1,2,...,n$

(3) *註找出所有群聚間，距離最接近的兩個群聚 C_i、C_j

(4) 合併 C_i、 C_j 成為一個新的群聚

(5) 重複(3)、(4)直到群數為K

*註距離(Metric)以及群間距離(Linkage Criteria)的定義會影響分群的結果，這裡使用歐式距離(Euclidean Distance)搭配四種群間距離做分群。

Linkage Criteria	群聚間的距離定義	計算式
Single-linkage agglomerative algorithm	不同群聚中最接近兩點間的距離	$\min_{a\in C_i, b\in C_j} d(a,b)$
Complete-linkage agglomerative algorithm	不同群聚中最遠兩點的距離	$\max_{a\in C_i, b\in C_j} d(a,b)$
Average-linkage agglomerative	不同群聚間各點與各點間距離總和的平均	$\sum_{a\in C_i, b\in C_j} \frac{d(a,b)}{\lvert C_i\rvert \lvert C_j\rvert}$
Ward's method	將兩個群聚合併後，個點到合併後的群中心的距離平方和	$\sum_{a\in C_i\cup C_j} \lvert\lvert a-\mu\rvert\rvert$

*μ是$C_i \cup C_j$的平均值

表2. 分群標準

K-medoids

不同於階層式分群法(Hierarchical Clustering)，分割式分群法 (Partitional Clustering)是事先指定群的數目後，再用一套疊代的數學運算法，找出最佳的分群方式以及相關的群中心。其中最有名的是K-means和K-medoids，而其差別在於計算到中心點的方式不同

(1) K-means 欲找到中心點μ_0，$\min \sum_{i=1}^{n} (x_i - \mu_0)^2$

(2) K-medoids 欲找到中心點μ_0，$\min \sum_{i=1}^{n} |x_i - \mu_0|$

假想我們現在給定其中一組數據$(x_i, \mu_0) = (0,2)$，使用K-means方法，則這項會使得全部增加4 ($\because (0-2)^2 = 4$)；同樣地，如果我們使用K-medoids方法，則這項會使得全部增加2 ($\because |0-2| = 2$)，從上述計算的方式可看出K-medoids相較於K-means而言，較不容易受noise及outlier影響，因此會更為穩健(Robust)，而K-medoid方法為：

操作步驟：

(1)隨機選擇K個點做為初始的medoid
(2)將每個object聚集到最近的medoid
(3)更新每個群的medoid，計算objective function
(4)選擇最佳參數
(5)重複(2)(3)(4)直到medoid不再變化

Kohonen's Self-Organizing Maps SOM

SOM的概念來源為人類大腦網路或生物網路在處理資訊時, 處理相同資訊的神經元會聚集在一起的特性, 會形成一個個區域的樣子。SOM有一個很重要的優點為, 將N維(N-dimension)的資料映射(mapping)到2維(2-dimension)的空間上並且維持資料中的拓撲(topology)特性，將資料映射到二維空間時則可以使用視覺化(visualization)的方式呈現, 以方便後續的觀察及分析。SOM實際操作如下：

操作步驟：
(1)初始化(initialization)每個節點(node)的權重
(2)從資料集挑選一個object並計算其到其他節點的距離
(3)尋找距離最近的節點Best Matching Unit(BMU)
(4)更新BMU和鄰近點的權重，越接近BMU的點權重更新越多
(5)重複步驟(2)~(4)N次

結果及意義

Hierarchical clustering

利用R中的指令rect.hclust在圖上直接分三群，Hierarchical Tree的Y軸代表距離，因此當群與群的聚合在Y軸越高的位置，代表兩群之間距離越遠，而不同的Linkage之下所測出來的距離尺度會有所差異，因此只看相對距離來做分群，再由另一種指標Agglomerative Coefficient來評估分群的好壞程度。其中Single-linkage Agglomerative與Average-linkage Agglomerative在圖上的分群結果看起來不如預期。

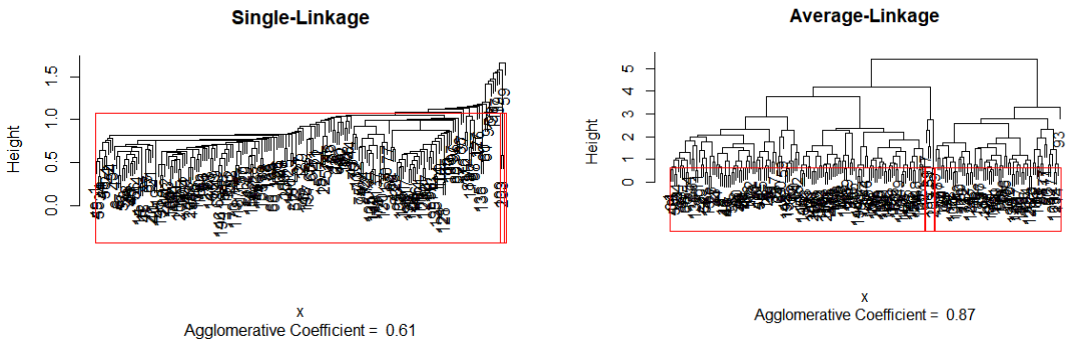


圖2. Single-linkage（左），Average-linkage（右）紅色的部分是我們對這些資料點所做的分群，使用R指令直接以套件指定分成三群。

因為我們這次使用R內部套件直接將結果以方形匡線的方式分群，我們可以看到紅框線以切出三群為目的標記。

可以清楚看到這兩中方法(如圖2.)相較於圖3. 其分群的資料個數比較不平均，Single-linkage左邊這群個數最多，右邊兩群個數都相當少，由於Single-Linkage的計算方式使他容易受到outlier影響。分群數的決定會對結果造成很大的影響，尤其是當離群值被獨立分為一群。圖2. 右邊即為所例，為了分成三群，其中一群的資料點離另外兩群較遠，導致被獨立出自己一群。

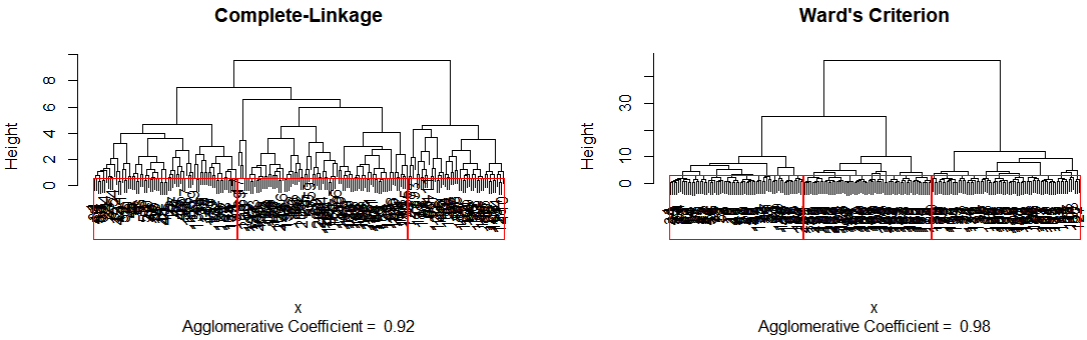


圖3. Complete-linkage（左），Ward’s method（右）紅色的部分是我們對這些資料所做的分群，使用R指令直接以套件分成三群。

相較於圖2.，可以清楚地看到在Complete-linkage 及Ward’s method的分群結果其資料個數較為平均，較接近我們理想中分群的方法。

Linkage Criterion	Agglomerative Coefficient
Single-linkage	0.6074219
Average-linkage	0.8668874
Complete-linkage	0.9218714
Ward’s method	0.9841322

表3. 四種Linkage Criterion的Agglomerative Coefficient

Agglomerative Coefficient(AC)的範圍介於0-1之間，係數越大代表模型越強(Strong Structure)，0.7以上為Strong Structure，0.51-0.7為Reasonable Structure，0.26-0.5為Weak Structure，0.25以下為No Substantial Structure。

因此從AC的指標來看，階層式分群採用Ward’s作為Linkage Criteria的分群能夠有最好的分群效果，而Single Linkage的分群效果最差。

K-medoids

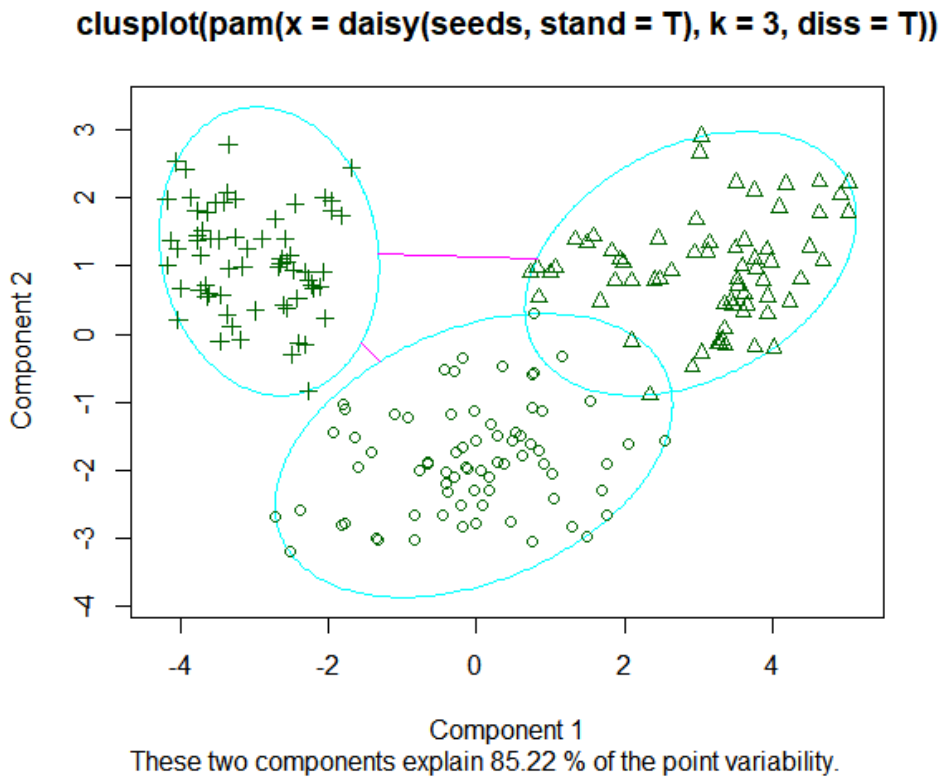


圖4. 資料縮減維度至PC1、PC2空間時的Clusplot

圖4. 分群的效果看起來還不錯，僅有群與群中間略有重疊，而前兩個Component的解釋能力可以達到88.21%。

Silhouette Coefficient

$$s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)}$$

b(i)代表(i)與最近的群的所有點的平均距離，而a(i)代表(i) 與群內所有其他點的平均距離，Silhouette Coefficient的值在-1~1之間，值越大分類效果越好。

從公式可知若s(i)大於0，代表(i)與其群內其他點的平均距離小於最近的其他群，表示分群效果較佳。從Silhouette Coefficient為0.4來看，分群效果似乎不太好，模型屬於Weak Structure。

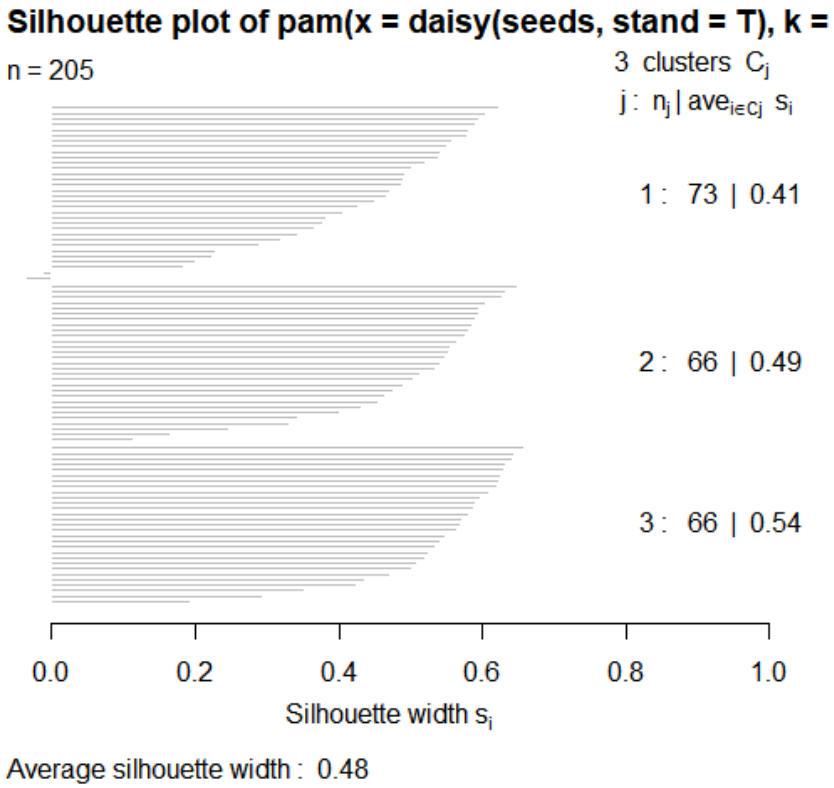


圖5. 資料的Silhouette Plot

群數K的決定可以透過選取使Silhouette Coefficient最大的值來決定，而經過幾次嘗試我們選擇K = 3能夠使Silhouette Coefficient最大。

Kohonen's Self-Organizing Maps SOM

由於資料一共有205筆，因此在neuron的選擇上選取10*10，並且設定重覆疊代100次。

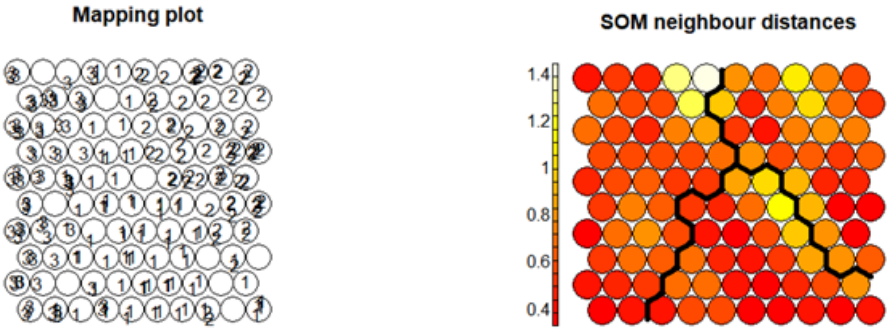


圖6. Mapping plot（左）SOM neighbor distance（右）

圖6. 可以看到三個品種的小麥分別位於圖的左方、右上方、中間及中下方，可以明顯看出三個群因此分的算是不錯。

接著再利用Agglomerative Hierarchical Clustering來協助在SOM的圖形上做分群，設定分為三群，其中顏色深淺的部份代表了該點與最近節點的距離，顏色越深(紅)距離越近，因此可以將淺黃色的部分當作分隔。

Accuracy Rate

利用R中cutree的指令，並令其分為三群，假設某一群中，小麥品種為Rosa的個數最多，則將該群當作Rosa，而其他品種被視為錯誤分群，以這樣的指標來比較各種分類方法。其中Hierarchical Clustering中Single-linkage的分群由於AC過小不在此做討論。

	Hierarchical Clustering			Others	
方法	Complete-linkage	Ward ’s method	Average-linkage	K-medoids	SOM
Accuracy	0.8195122	0.902439	*註	0.9707317	*註

表4. 各個方法的準確率

從表4. 可以得知，Complete Linkage, Ward ’s, K-medoids的分群都能大致將不同的品種做區分。由於SOM與Hierarchical clustering的Average Linkage無法由cutree的指令將其完好的分群，因此直接觀察所有樣本被分群的狀況，若可以判斷品種的出現次序有很明顯的分界，則可以說分群的效果是好的。

*註

		程式分群		
		1Kama	2Rosa	3Canadian
實際類別	1Kama	64	2	3
	2Rosa	4	64	0
	3Canadian	66	0	2
準確率		47.76%	94.12%	40.00%

表5. Average-linkage agglomerative 方法分群結果及其準確率

		程式分群		
		1Kama	2Rosa	3Canadian
實際類別	1Kama	40	22	7
	2Rosa	0	15	16
	3Canadian	0	0	0
準確率		100%	40.54%	0%

表6. SOM方法分群結果及其準確率

由表5.可以看到在這樣的分群方法底下，在第一群及第三群的表現正確率僅有約四成，相當不佳，導致整體準確過低。再來看到表6. SOM分群結果，在第二群及第三群的表現都相當不佳，第二

群準確率僅有四成，甚至到了第三群準確率到達0%，主要是因為SOM這個方法並不會對所有的資料點進行檢測，以我們這次實際操作來說，我們只選擇100(10*10)個資料點進行測試。由這兩個表可以觀察到：在程式分群的部分，第三群資料點皆相當的少。Average-linkage僅有5個資料點被分做這群，其中有3個資料點是被分錯的。SOM方法僅有23個資料點被分成第三群，雖說較Average-linkage多，但仍是三群當中最少的。

由於SOM與Hierarchical clustering的Average Linkage無法由cutree的指令將其完好的分群，其結果如下：

```
> seeds[,8][agn4$order]  # Average-Linkage
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1
 [21] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [41] 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [61] 1 3 3 1 1 1 1 3 1 3 3 1 3 3 3 3 3 3 3 3
 [81] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3
[101] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[121] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 3 3 1
[141] 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2
[161] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[181] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[201] 2 2 2 2 2
```

```
> som.hc2 %>% as.vector()
  [1] 1 1 1 1 1 1 1 1 1 2 2 2 2 2 1 1 1 1 1 1
 [22] 1 1 2 2 2 2 2 1 1 1 1 1 1 1 1 1 2 2 2 2 2
 [43] 1 1 1 1 1 1 1 1 1 2 2 2 2 2 1 1 1 1 1 3
 [64] 3 3 2 2 2 2 2 1 1 1 1 1 1 3 3 3 2 2 2 2 2
 [85] 3 3 3 1 1 3 3 3 3 2 2 2 2 2 3 3 3 3 3 3 3
[106] 3 3 3 2 2 2 2 3 3 3 3 3 3 3 3 3 3 2 2 2 2
[127] 3 3 3 3 3 3 3 3 3 3 2 2 2 2 3 3 3 3 3 3 3
[148] 3 3 3 2 2 2 2 3 3 3 3 3 3 3 3 3 3 2 2 2 2
[169] 3 3 3 3 3 3 3 3 3 2 2 2 2 2 3 3 3 3 3 3 3
[190] 3 3 3 2 2 2 2
```

左圖為Average Linkage的分群，可以看到品種大致是有順序的出現，代表分群做的不錯；而右圖為SOM的分群，其中品種2,3的出現順序不太能看出一個分界，因此SOM這個方法不太適合對小麥的資料做分群。

(2)

Perform the following multidimensional scaling (MDS) analysis for the data set “Seeds.txt” by using all variables X1, ..., X7 and Y:

(1) Classical Torgerson-Gower MDS;

(2) Kruskal and Shepard’s nonmetric MDS based on isotonic regression.

Comment on the quality of goodness of fit and compare the solutions of these two methods.

題目描述

題目欲利用MDS進行降維，使資料之間的相似性(或不相似性)得以在低維度的空間上表示。以Seeds這筆資料來說，即利用7個小麥的特徵（即其周長、面積、緊密度等等），在二維平面上能否表現出三個種類不同的小麥各成一群。其中小麥的特徵設為自變數$(X_i, i = 1 \sim 7)$為連續變數，而小麥的種類設為因變數$(y_j, j = 1,2,3)$且為類別變數

資料描述

如(1)

方法

MDS的目的是希望能夠維持原資料之間的關係(差異)之下，在低的維度上做呈現，因此這裡需要定義(不)相似性(dissimilarity)。

對於連續型資料、類別型資料、混和型資料的(不)相似性皆有其定義，而題目欲分析的資料Seeds之變數為小麥仁的幾何數據，因此這裡僅介紹連續型資料的不相似性。

Continuous Data: Euclidean distance

$$\Delta = [\delta_{ij}]_{NxN}$$

這裡的矩陣D代表Dissimilarity，是利用資料兩兩相減的Determine值做為指標，當我們以歐式距離(Euclidean Distance)計算這個值，我們則稱這類的MDS為classical(or Togerson-Gower)MDS。其中d_{ij}為第i筆資料與第j筆資料在p維空間中的距離，其中$p > 0$，在這個例子因為自變數一共有幾個，因此p=7。當不相似矩陣D裡面每個元素滿足對稱性：

$$\delta_{ij} \geq 0$$
$$\delta_{ii} = 0$$
$$\delta_{ij} = \delta_{ji}$$

及三角不等式

$$\delta_{ij} \leq \delta_{ik} + \delta_{jk}$$

當這個不相似矩陣D滿足對稱性，以及三角不等式時，則為metric；當有任一條件不符合，則為non-metric。降維後的矩陣如下：

$$D = [d_{ij}]_{NxN}$$

這裡引進一個loss function來評估由降維所引起的變異，它是由一個原資料的不相似矩陣(delta)與降低維度後的不相似矩陣(d)的方差，來估計這個變異量，透過最小化這個函數來找到最佳的近似結果，當這個函數越小代表這兩個距離矩陣越相近。Loss function有很多種，例如STRESS與normalized STRESS這個STRESS是會受資料尺度的影響，因此在做MDS之前需要先對資料做尺度標準化的修正。

$$STRESS(\tilde{X}) = \sum_{i=1}^{N} \sum_{j>i} w_{ij}(\delta_{ij} - d_{ij}(\tilde{X}))^2$$

這裡引進一個STRESS function來評估由降維所引起的變異，它是由一個原資料的不相似矩陣(delta)與降低維度後的不相似矩陣(d)的方差，來估計這個變異量，而目標是希望這個值越小越好。這個STRESS function是會受資料尺度的影響，因此在做MDS之前需要先對資料做尺度標準化的修正。

normalized(raw) STRESS:

$$\frac{STRESS(\tilde{X})}{\sum_{i,j} w_{ij}\delta_{ij}^2}$$

經過標準化後的STRESS function又稱為Kruskal’s Stress-1，可以用該值來評斷降維帶來的的變異是否足夠小，而Kruskal (1964)提出的判斷準則為下表：

表

Kruskal’s Stress-1	Goodness of Fit
20%	Poor
10%	Fair
5%	Good
2.5%	Excellent
0%	Perfect

而loss function另外的選擇為STRAIN，其利用資料間的內積來表達距離，則我們可以用類似PCA的概念，去解釋所能解釋變的變異比率，去衡量近似的距離矩陣所能解釋的變異百分比。

在metric之下，不相似性會滿足三角不等式，然而某些類型的資料的不相似性無法滿足三角不等式，例如順序尺度的資料，而這時的不相似性僅能以順序代表之，這種non-metric的不相似性則會以其他方式來取代原本的不相似性矩陣。

其中一種方式是利用Isotonic Regression，是一種嚴格遞增的函數可以保證順序尺度的順序關係，來取代原本的delta，並計做theta，因此新的STRESS function為

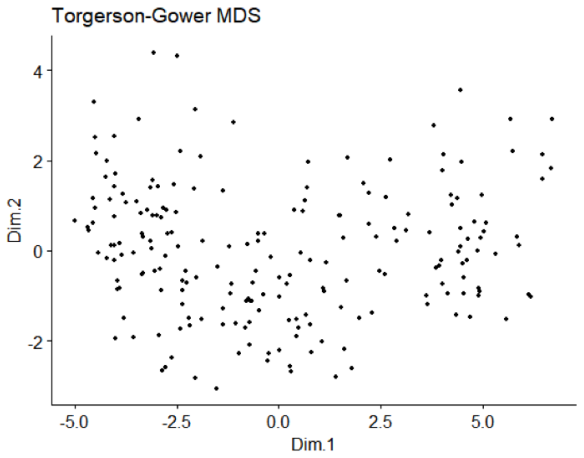
$$STRESS = \frac{\sum_{i,j}(\theta(\delta_{ij}) - d_{ij}(\tilde{X}))^2}{\sum_{i,j} d_{ij}^2(\tilde{X})}$$

而最後的目標一樣是希望可以讓STRESS function最小，透過給定降維後的不相似性矩陣，利用Isotonic Regression來估計原資料的不相似性矩陣，經過數次疊代後找到最佳答案。而除了用Isotonic Regression來尋找theta，也可以用嚴格遞增(以確保順序關係)迴歸、非線性等方式，可能會得到更小的STRESS值。

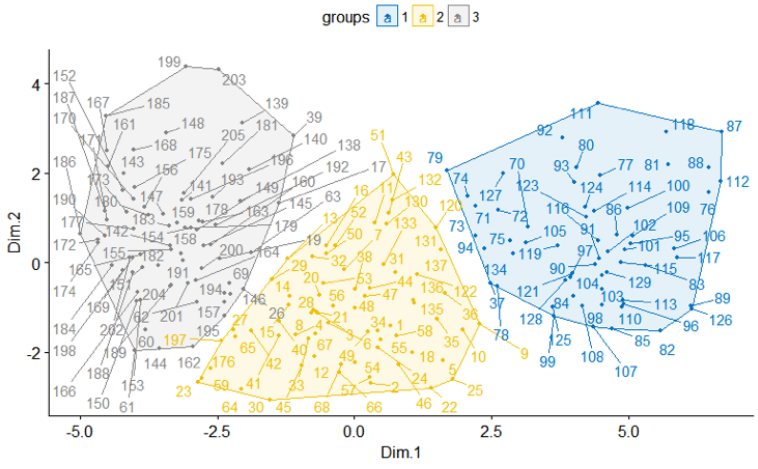
結果及意義

首先，我們對Seeds的資料做Classical (or Torgerson-Gower)MDS，以視覺化的方式嘗試能否由兩個dimension來呈現資料的不相似性，進而看出這樣的不相似性將三個小麥的品種做分群。

Classical (or Torgerson-Gower) MDS



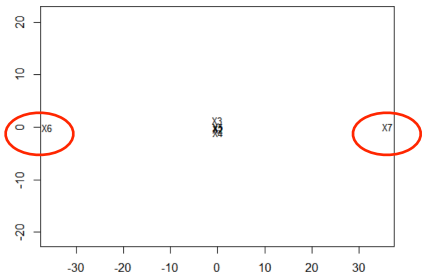
圖



圖

由前兩個Dimension所解釋的總變異量為88.85%，表示此模型配適的很好，因此我們可以透過兩個維度的圖來觀測資料的相似程度，左圖是將所有小麥資料呈現在這兩個Dimension上，右圖的部分是利用K-means將上圖的資料分為三群，可以大致看出左、中、右這樣三群的趨勢。

圖

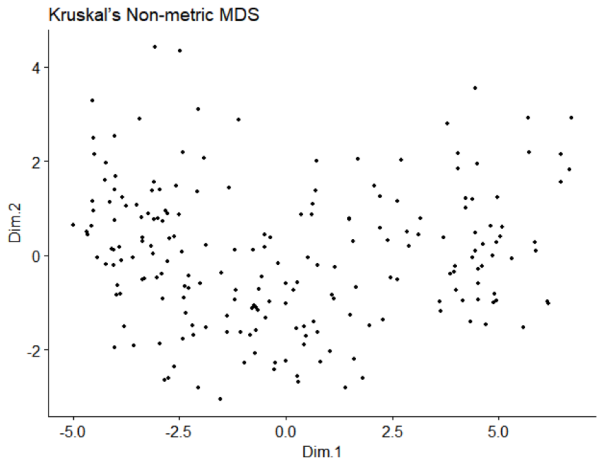


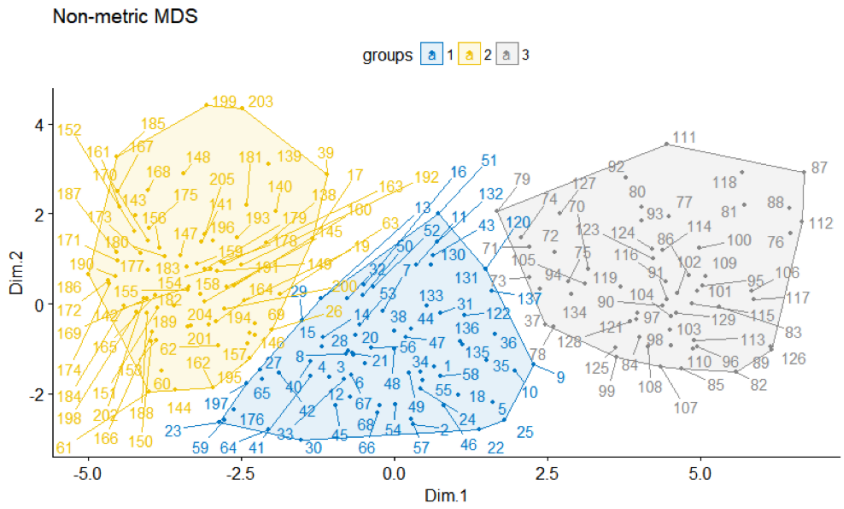
接著再以變數兩兩之間的相似度，其中以correlation coefficient來衡量彼此的關係，由MDS的特性，可以看到X1(面積),X2(周長),X3(緊密度),X4(仁的長度),X5(仁的寬度)是有高度相關的，而X6(不對稱係數)和X7(仁的溝槽長度)則各自與其它變數都不相關，如上圖紅色圈起來的部分。

Kruskal and Shepard's Non-metric MDS(Utilizing Isotonic Regression)

接著，我們嘗試以Non-metric MDS搭配Isotonic Regression在兩個維度的平面上觀察Seeds資料。

圖

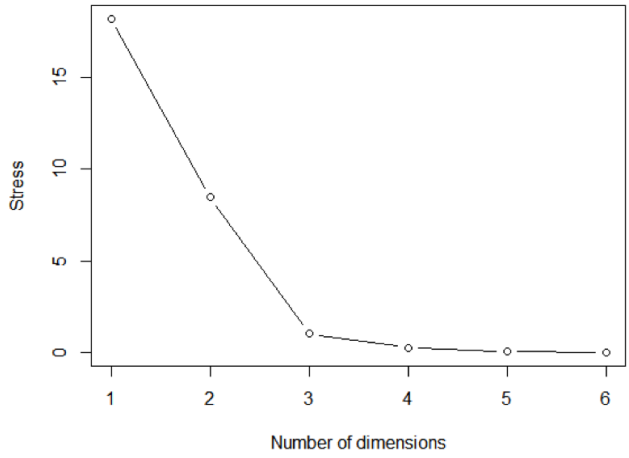




圖

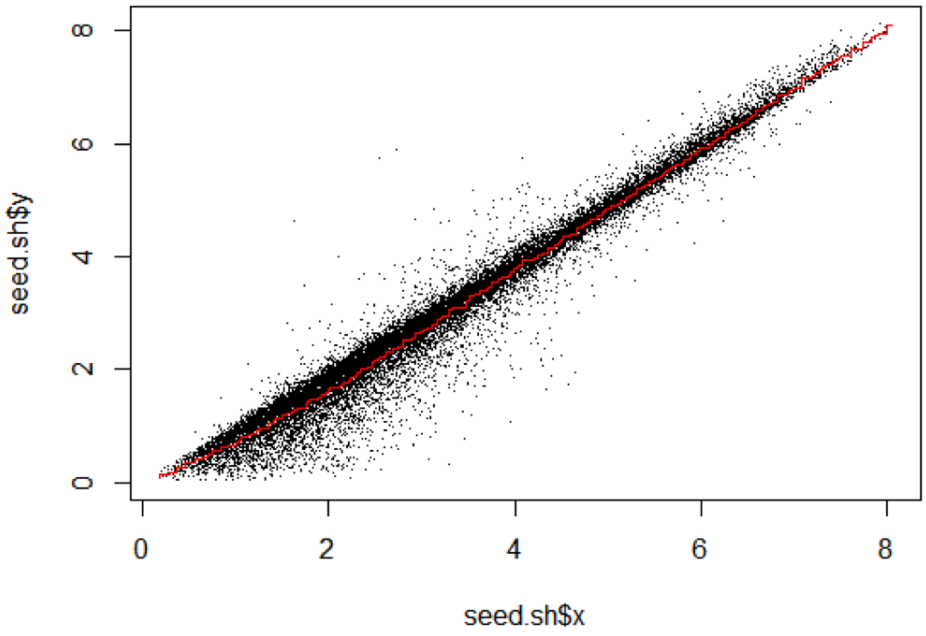
Kruskal and Shepard’s Non-metric MDS下的STREE值為8.46%，代表降低維度後引起的變異僅有8.46%，而對照Kruskal所提出的判斷準則，這是一個good的配適。

這裡一樣利用K-means來看Kruskal and Shepard’s Non-metric MDS的方法，在兩個維度的平面是否能夠明顯看出有三個群的趨勢，而從結果來看，似乎與Classical (or Torgerson-Gower) MDS差不多好。



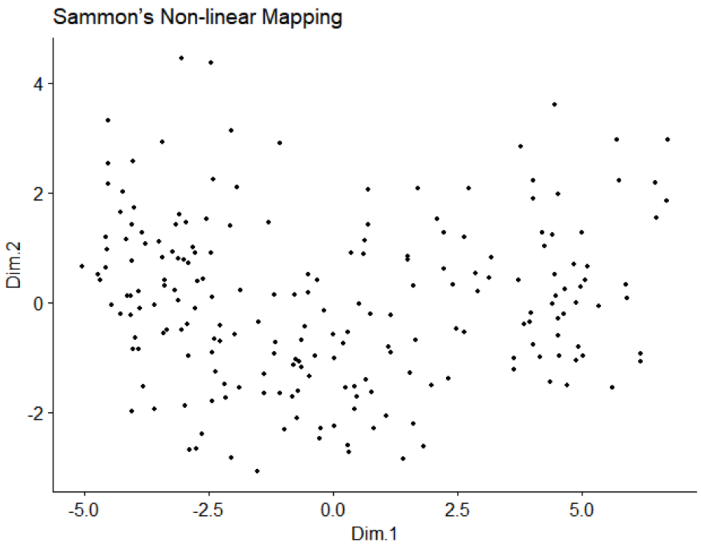
圖

接著，由Elbow Method來看我們需要多少的Dimension，由STRESS減少的幅度作為Elbow Method的選擇，而當Dimension為3的時候是最適合的，但是其實由2個Dimension已經有足夠小的變異量損失量。



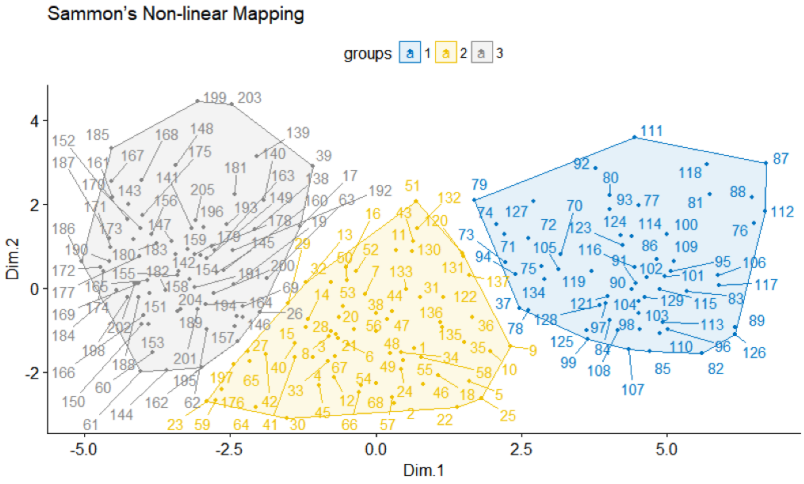
此圖可以看到僅有少數的點散佈在離配適的function較遠的地方，而此代表由兩個Dimention即可配適的不錯。

Sammon's Non-linear Mapping



圖

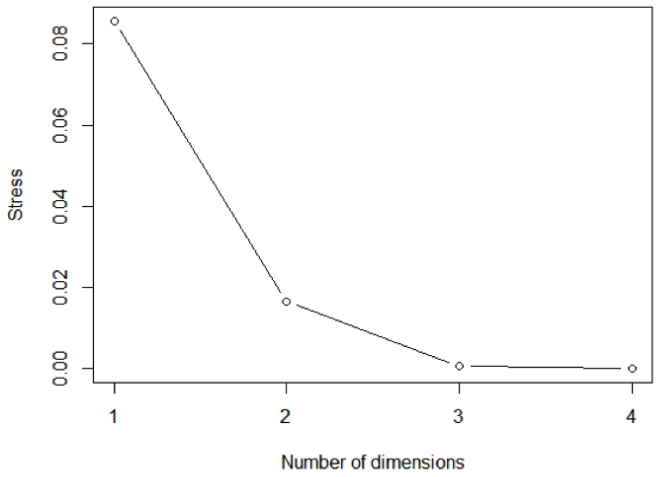
圖



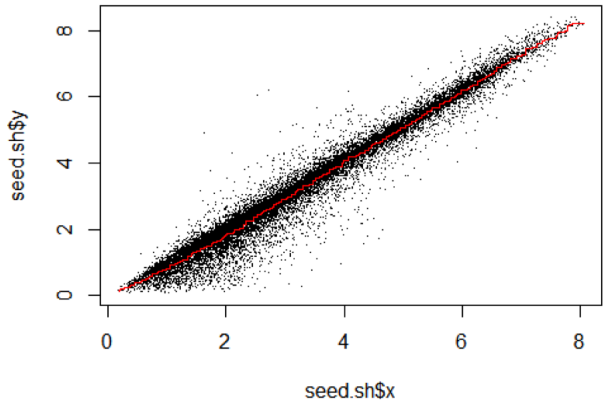
這裡以Non-linear Mapping的方式取代前一個Isotonic Regression的作法，來看是否能用兩個Dimension來呈現Seeds資料的三個群的趨勢。從圖中可以觀察到與前兩者的分群是差不多的。

而Non-linear Mapping的STRESS的值為1.6%，比在isotonic regression之下，因為降維而引起的變異量更少，因此Non-metric MDS用Non-linear Mapping取代Isotonic Regression的適合度更好。

圖



再次透過Elbow Method來判斷需要多少個Dimension，在Non-metric MDS with Non-linear Mapping之下，僅需要兩個維度即可。



圖

此圖可以看到僅有少數的點散佈在離配適的function較遠的地方，而此代表由兩個Dimention即可配適的不錯。

Summary

總結以上三種方法，使用兩個Dimension皆能良好的表現出Seeds三個品種的差異性。而Non-metric MDS的方法中，又以Non-linear Mapping的方式比Isotonic Regression之下，因為降維而導致的變異量更少，因此Non-linear Mapping是在Non-metric MDS是較好的方法。

(3)

Consider the ““new_wages.txt” data set in the midterm exam, which includes 11 categorical characteristics from a random sample of 534 persons.

(1) Perform a simple correspondence analysis (CA) for the first two variables “Wage” and “Education”.

(2) Perform a multiple correspondence analysis (MCA) for all the variables. By including all variables in this analysis, is the relationship between “Wage” and “Education” different from that observed from the simple correspondence analysis?Comment on the result.

資料描述

新薪資料包含534個人的薪水以及他們其他的類別變數，所有的自變數均為類別變數，一共有11個，詳細說明如下：

1)薪水種類：A=(時薪<=5美金),B=(5美金<時薪<10美金),C=(時薪>10美金)

2)教育程度：1=非常低、2=低、3=中等、4=非常高

3)是否居住南方：1=是、0=不是

4)生理性別：1=女生、0=男生

5)工作經驗：1=低,2=中, 3=高

6)是否為工會會員：1=是,0=不是

7)年紀：1=小於33歲, 2=33~48歲之間, 3=大於48歲

8)種族：1=其他, 2=拉丁裔,3=高加索人

9)職業性質：1=管理職務,2=業務,3=辦公室工作, 4=服務業,5=專業人員,6=其他

10)部分：0=其他,1=生產,2=製造

11)婚姻狀態：1=結婚,0=其他

資料前置處理

(1)將前兩欄(Wage, Education)統計其次數，並且繪製成列聯表。如下表：

		Wages		
		A	B	C
Education	ED1	3	3	1
	ED2	18	28	3
	ED3	93	160	86
	ED4	11	52	76

處理方法及為何選擇這個方法

透過列連表我們可以發現在原始資料的ED1(即教育程度較低)的觀測次數較少，為了做對應分析(Correspondence Analysis)，我們將ED1和ED2做合併。資料如下表：

		Wages		
		A	B	C
Education	ED1+ED2	21	31	4
	ED3	93	160	86
	ED4	11	52	76

方法

對應分析(Correspondence Analysis)是將一個列聯表的行與列之間的關係，並將列聯表的觀測次數的行、列轉成圖像呈現。數學上，對應分析分解χ^2，在低維度的空間中表現出來，因此當圖上的兩個點越接近，代表兩者關係相依程度(dependence)越大。對應分析透過各個類別在空間的分佈距離，探討類別彼此的關係，其中的分佈距離是χ^2距離：

$$\delta^2(i,i') = \sum_{j=1}^{J} \frac{[F(i,j)/r(i) - F(i',j)/r(i')]^2}{c(j)/N}$$

對應分析的特點在於沒有特殊模型導入、沒有任何假設，唯一的限制在於觀測的次數(Observations)不得過少（大部分以觀測次數≥5為標準，但本次實際操作僅以觀測次數≥4）。與PCA的概念相似，對應分析透過分解距離，即用以衡量關聯性的指標，作為維度並

且作圖觀察各個變數之間的關係。相較於PCA不同的是：PCA一般處理連續型資料，而CA可處理非連續型資料。

結果及意義

[合併前]

Principal Inertia (eigenvalue) :

	1	2
Value	0.118917	0.001013
%	99.16%	0.84%

Component 1 解釋了99.16%的total inertia

Rows:

	ED1	ED2	ED3	ED4
Mass	0.013109	0.091760	0.634831	0.260300
ChiDist	0.503916	0.553262	0.134413	0.544042
Inertia	0.003329	0.028088	0.011469	0.077044
Dim. 1	-1.331014	-1.593541	-0.388979	1.577443
Dim.2	6.534313	-2.017858	0.270767	-0.278094

Columns:

	A	B	C
Mass	0.234082	0.455056	0.310861
ChiDist	0.414262	0.123132	0.484126
Inertia	0.040171	0.006899	0.072859
Dim. 1	-1.194746	-0.343949	1.403150
Dim.2	1.358154	-1.038860	0.498035

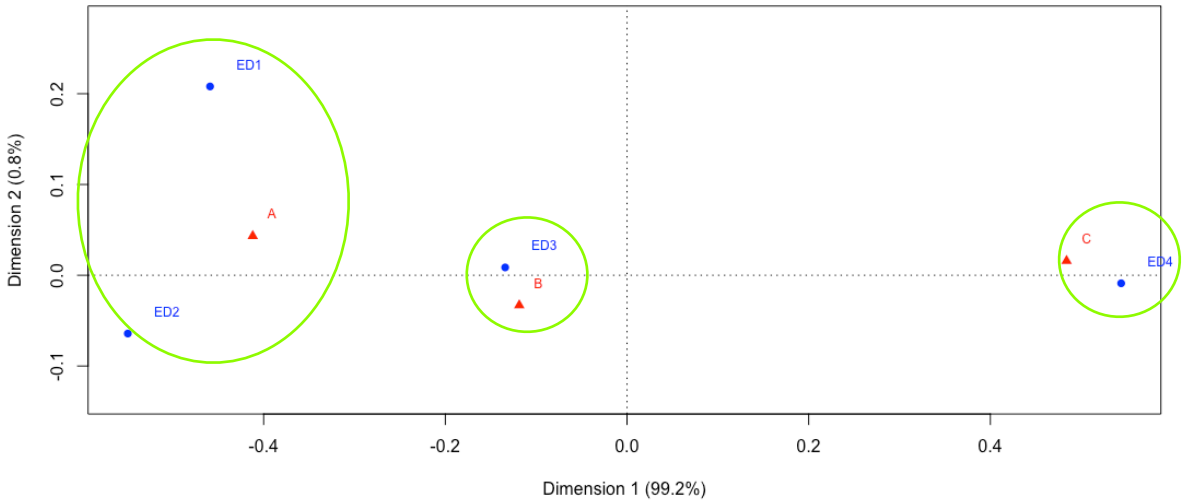


圖1. 二維關係圖

由圖1.可以看到ED3和B相當接近，其中ED3表示教育程度為中等者，標籤B代表時薪大於5美金，且小於等於10美金，也就是說，當教育程度為中等時，和所領的時薪有一定的關係。圖的最右邊可以發現ED4和C較為接近，也就是說當教育程度高和時薪大於10美金有關係。

[合併後]

Principal inertia(eigenvalues)：

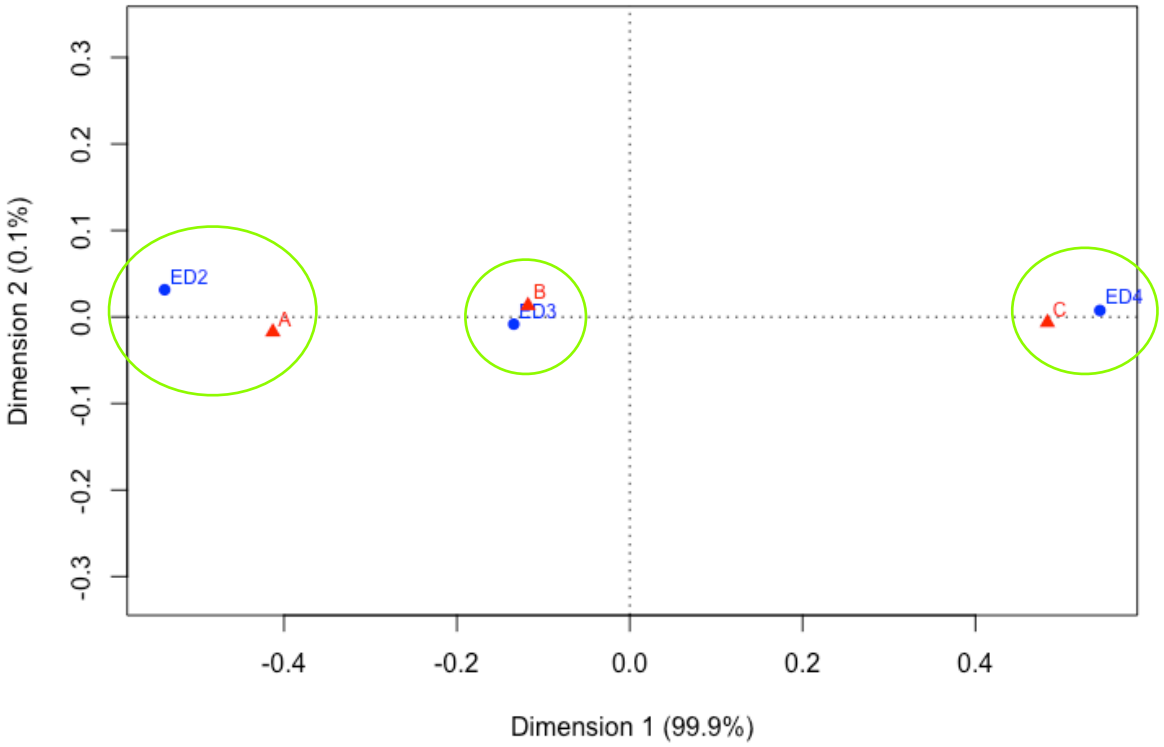
	1	2
Value	0.118823	0.000163
Percentage	99.86%	0.14%

Rows:

	ED1+ED2	ED3	ED4
Mass	0.104869	0.634831	0.2603
ChiDist	0.539052	0.134413	0.544042
Inertia	0.030472	0.011469	0.077044
Dim. 1	-1.561125	-0.38919	1.578119
Dim. 2	2.469535	-0.650962	0.592679

Columns :

	A	B	C
Mass	0.234082	0.455056	0.310861
ChiDist	0.413313	0.118452	0.483308
Inertia	0.039988	0.006385	0.072613
Dim. 1	-1.197978	-0.341473	1.401959
Dim.2	-1.355304	1.039676	-0.501376



↑圖. 在圖當中我們把ED1和ED2的觀測次數合併，統稱為ED2

可以看到能夠透過Dimension1解釋的資料量比合併前高出約0.7%，且ED3和B的距離較合併前相近許多，同理 ED4和C也是。換言之，可以清楚看到這些objectc和category之間的關係是比較相近的，也就是說兩者之間的關聯性較高。ED2和A類別之間的距離較遠，也就是說這個object和category的關係相較於另外兩者較沒有這麼高。

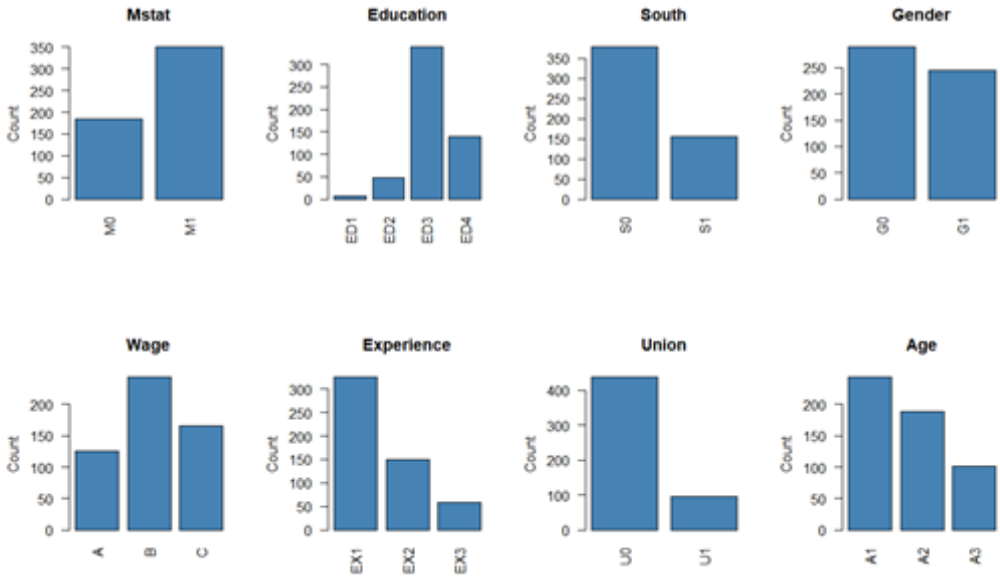
Multiple Correspondence Analysis (MCA)

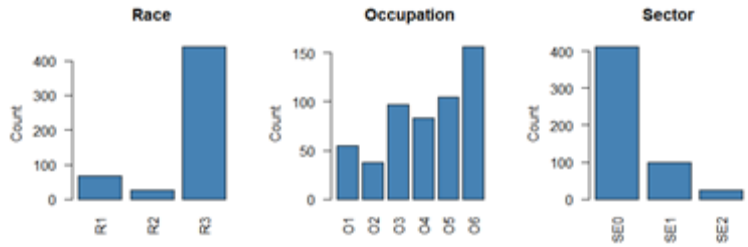
資料處理

變數類別	Wage	Education	South	Gender	Experience	Union
變數等級：次數	A:125	ED1:7	S0:378	G0:289	EX1:325	U0:438
	B:243	ED2:49	S1:156	G1:245	EX2:150	U1:96
	C:166	ED3:339			EX3:59	
		ED4:139				

變數類別	Age	Race	Occupation	Sector	Mstat
變數等級：次數	A1:243	R1:67	O1:55	SE0:411	M0:184
	A2:189	R2:27	O2:38	SE1:99	M1:350
	A3:102	R3:440	O3:97	SE2:24	
			O4:83		
			O5:105		
			O6:165		

首先，我們先對各個變數的類別做視覺化分析，觀察變數之下不同類別的個數關係，透過這個圖我們可以發現數量過少的類別，該類別可能會扭曲接下來的分析，因此可以考慮與其他類別合併。





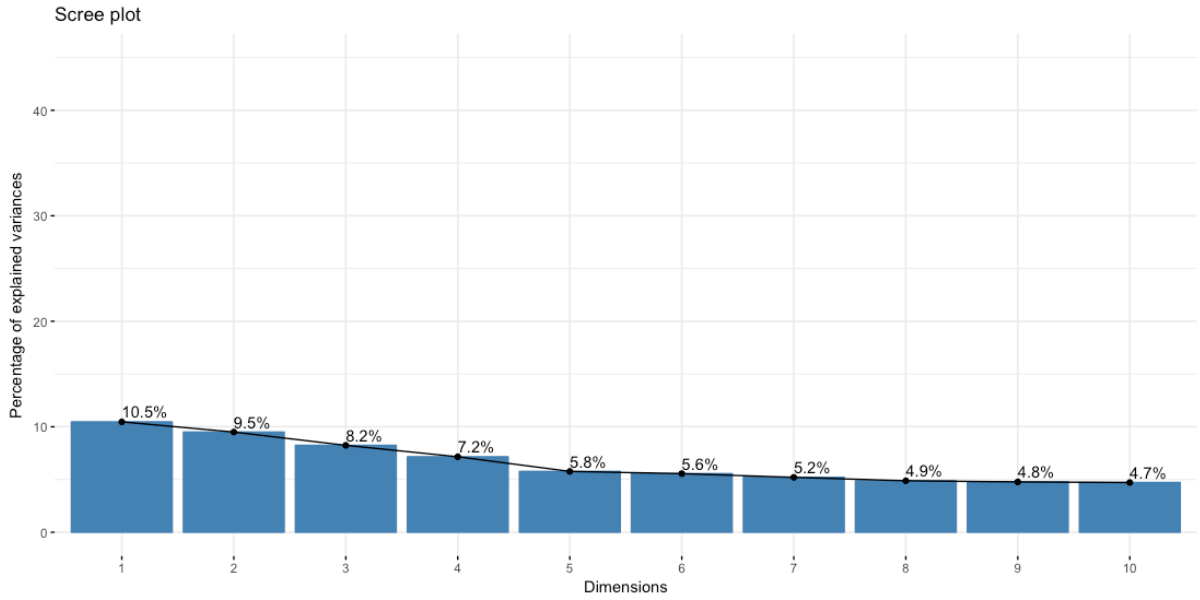
處理方法及為何選擇這個方法

選擇MCA(Multiple Correspondence Analysis)主要是因為此題類別變數總和大於2，因此我們不使用對應分析(Correspondence Analysis)。

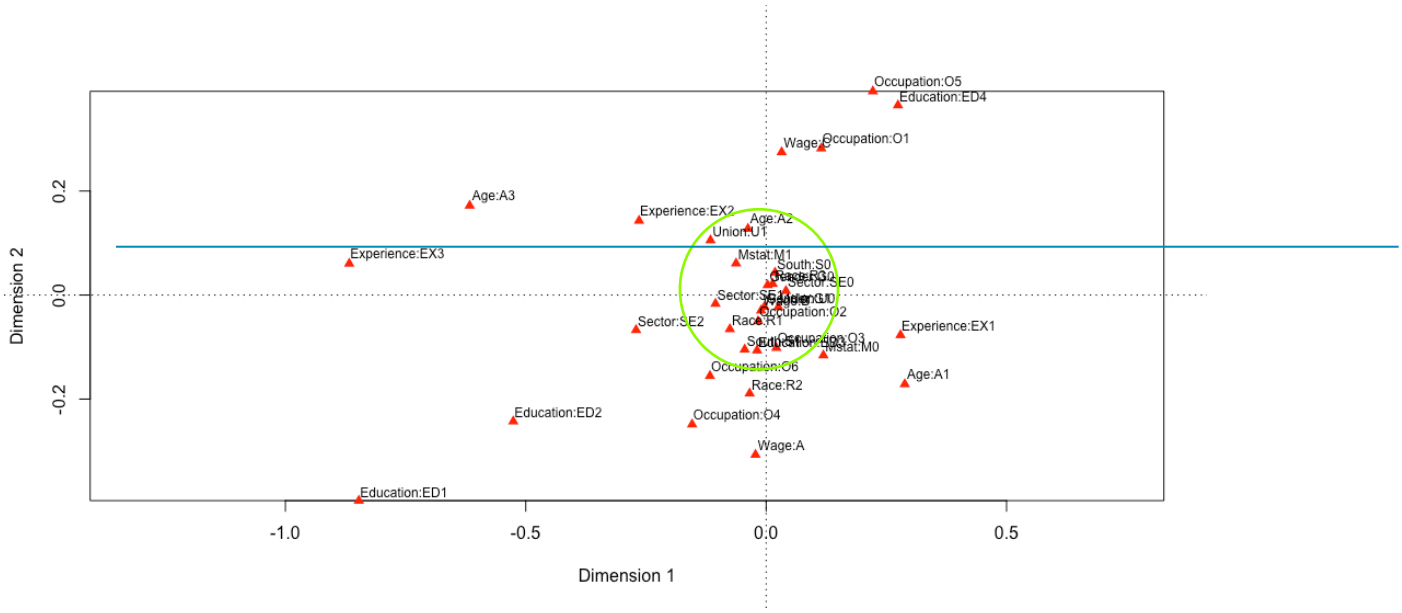
方法如何運作

可視為對應分析(Correspondence Analysis)的一種延伸，相較於主成份分析(Principal Component Analysis)，其中最大的不同在於主成份分析只能處理連續型變數，而MCA可以處理類別型變數。和對應分析相同(Correspondence Analysis)，有卡方距離、inertia、可以在二維空間圖上表示。不同的是，資料矩陣須先轉換成Burt table形式在生成資料矩陣。

結果及意義



將資料進行MCA可以由上圖看到需要十個維度才能解釋一定的資料比例，單看第一維度及第二維度，能被解釋的資料僅20%。MCA有個很大的缺點也就是總慣量(total inertia)容易被高估，尤其在我們使用Burt table的方法的時候，導致可解釋的資料比例傾向被低估。資料原本型態為11個變數，由上圖可以看到透過MCA仍需要10個維度才能解釋一定的資料量，也就是說這筆資料使用MCA並無法有效降低維度，因此我們採用另一個方法調整，JCA(Joint Correspondence Analysis)。



由上圖可以看到各個軸之間的相互關係，可以發現靠近座標軸圓點的地方有相當多紅點聚集，也就是說這幾個變數的類別相關性比較高。除了這張相關性的圖之外，我們仍然必須考量每一個點所得到的觀測值，也就是觀測次數必須夠多，在分析上才會比較有參考價值，資料點及其發生的次數如下：

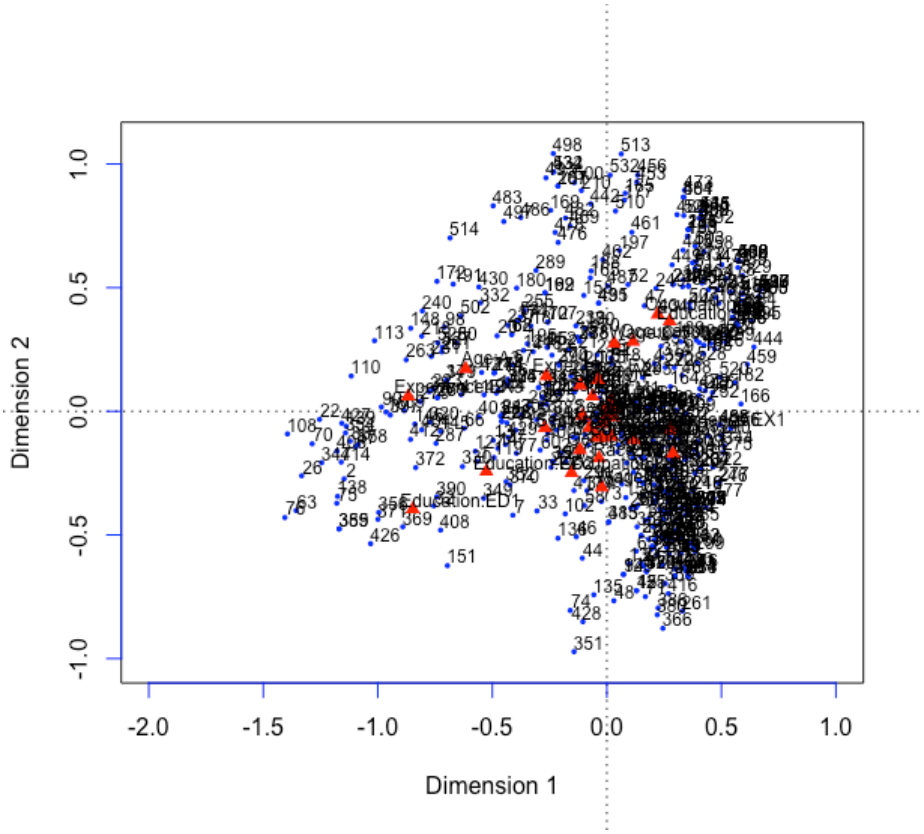
變數類別	Wage	Education	South	Gender	Experience	Union
變數等級：次數	A:125	ED1:7	S0:378	G0:289	EX1:325	U0:438
	B:243	ED2:49	S1:156	G1:245	EX2:150	U1:96
	C:166	ED3:339			EX3:59	
		ED4:139				

變數類別	Age	Race	Occupation	Sector	Mstat
變數等級：次數	A1:243	R1:67	O1:55	SE0:411	M0:184
	A2:189	R2:27	O2:38	SE1:99	M1:350
	A3:102	R3:440	O3:97	SE2:24	
			O4:83		
			O5:105		
			O6:165		

可以清楚看到中間的綠色部分可分為同一群，換句話說，這幾個資料點的類別比較相近，有較多的關係。值得注意的是，職業的第二種類別在這次觀測資料次數較少，再進行分類

的時候應該避免，主要是因為在分析這類資料時，我們會盡量避免使用次數較少的類別，以免在蒐集資料時造成的誤差（可能沒有隨機或者有其他誤差項）帶入資料分析中。

當我們放上資料點時，如下圖：



這個圖比較不好分析，主要是因為資料點過多，分析的時候會讓整體圖示不清楚，因此我們主要考慮各個變數不同類別的觀測次數，可用次數對照表相互對照判斷。

(4)

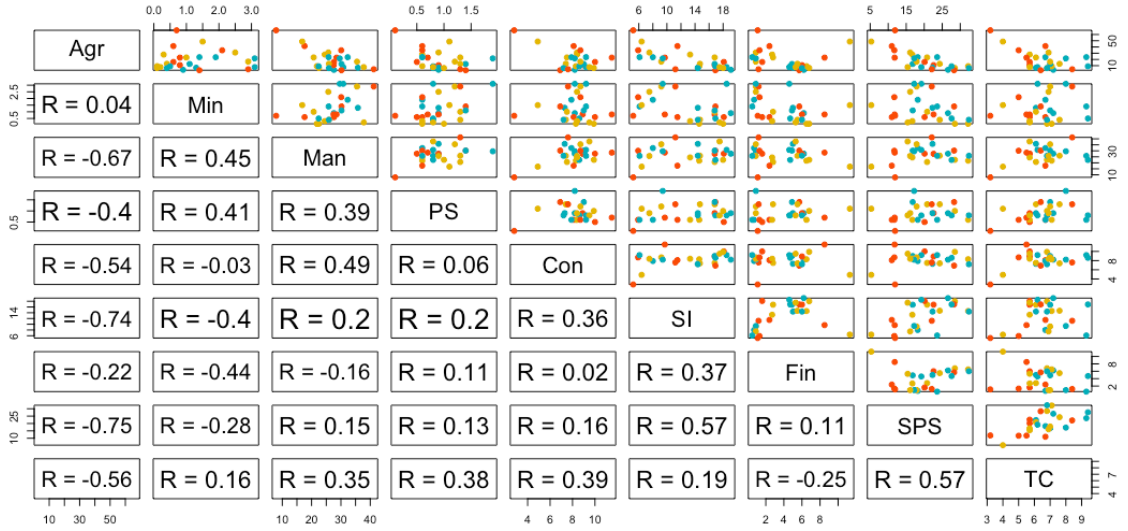
Consider the “European_Jobs.txt” data used in the midterm exam. Perform an Exploratory Factor Analysis for this data set (explain how many factors used, how to obtain the factors, and how to interpret the factors). Analogous to that done in PCA, make a “biplot” in the 2D factor-score space and interpret the result.

資料描述

原始資料為西元1979年(約略為第二次世界大戰三十年後)歐洲各國九種行業所佔之比率，總共26個歐洲國家，其中包含農業(Agr)、礦業(Min)、製造業(Man)、能源供給(PS)、營造業(Con)、服務業(SI)、金融業(Fin)、個人服務業(SPS)、交通運輸業(TC)。其中比較特殊的是服務業及個人服務業， 個人服務業是服務業的子集合，像是美容美髮業、乾洗店這類傾向客製化的服務業在這筆資料特別獨立出來討論。(Definition Source: Standard Industrial Classification)資料所提供的數值均為百分比。

資料前置處理

我們將所有資料進行相關性(Correlation)檢測，如下圖：



變數類別	Wage	Education	South	Gender	Experience	Union
變數等級：次數	A:125	ED1:7	S0:378	G0:289	EX1:325	U0:438
	B:243	ED2:49	S1:156	G1:245	EX2:150	U1:96
	C:166	ED3:339			EX3:59	
		ED4:139				

變數類別	Age	Race	Occupation	Sector	Mstat
變數等級：次數	A1:243	R1:67	O1:55	SE0:411	M0:184
	A2:189	R2:27	O2:38	SE1:99	M1:350
	A3:102	R3:440	O3:97	SE2:24	
			O4:83		
			O5:105		
			O6:165		

方法如何運作

因子分析(Factor Analysis)可以追朔至西元1905年。因子分析最主要的目的是透過少量無法觀測的(unobservable, latent)因子描述原始資料變數與變數之間的關係，並用這些潛在因子代替原本的資料，達到降維（意即變數縮減）的效果。

特別的是，因子分析剛好和主成份分析(Principal Component Analysis)相反。因子分析是指以潛在變數的線性組合表達原始變數；主成份分析是指以原始變數的線性組合來表達新的變數、主成份。兩者的共通點是：都可用來做多維資料降維。

基本假設：

(i) F and U are independent (Cov(F, U) = 0)

(ii) E(F) = 0 and $Cov(F) = I_k$(i.e.$Cov(F_i, F_j) = 0, i \neq j$)

(iii) E(U) = 0 and $Cov(U) = \Phi$, Φ是一個m*m的對角矩陣

(iv) X 為標準化之後的資料

其中X為

$$X = \Lambda F + U$$

而X的變異數$\Sigma = \Lambda\Lambda' + \Phi$

Λ：m*k 為Factor Loadings的矩陣，F：k*1 common factors的向量，U：m*1 誤差項的向量。目標是尋找合適的Λ, Φ, F去解釋原始的共變異數矩陣，但實際上這樣的矩陣並非唯一解，而是有相當多種組合。如何尋找最佳的組合：(1)principal factor analysis(2)maximum likelihood factor analysis。

最終k-factor model為

$$X_1 = \sum_{j=1}^{k} \lambda_{1j} F_j + U_1$$

$$X_2 = \sum_{j=1}^{k} \lambda_{2j} F_j + U_2$$

$$\cdots$$

(1)principal factor analysis

以樣本去估計原始資料。樣本的變異數矩陣S估計原始資料的變異數矩陣Σ，因此樣本Λ估計原始資料的$\hat{\Lambda}$，樣本Φ估計原始資料的$\hat{\Phi}$。使得$S = \hat{\Lambda}\hat{\Lambda}' + \hat{\Phi}$且$tr(S - \hat{S})'(S - \hat{S})$最小，即為所求。

(2)maximum likelihood factor analysis

假設F~N(0, I_k), U~N(0, Φ)，再透過最大概似估計法，求取適合的$\hat{\Lambda}, \hat{\Phi}$使得概似函數最大化。

結果及意義

首先我們進行一個factor的模型：

chi-square 檢定統計量	201.22
degree of freedom	27
p-value	1.45e-28

變數	Agr	Min	Man	PS	Con
可解釋資料比例	0.995000000	0.001381633	0.445902897	0.159673174	0.289381013

變數	SI	Fin	SPS	TC
可解釋資料比例	0.542011202	0.047503712	0.556508135	0.320064698

大約有37.3%的總變異能夠透過單一facctor模型解釋，解釋能力偏低。再來我們看到各的變數的解釋能力，農業(Agr)是裡面最高的(約99.5%)，服務業(SI)及個人服務業(SPS)可解釋變數大概佔五成(分別約是54.2%、55.6%)還算可以接受。製造業(Man)、交通運輸業(TC)、營造業(Con)、能源供給業(PS)可解釋變數的比例都偏低(分別約為44.6%、32.0%、28.9%、16.0%)，不甚理想。最後兩個產業的可解釋變數佔不到5%，分別是金融產業(Fin)約4.7%及礦業(Min)約0.14%。

再來我們進行兩個factor的模型：

chi-square 檢定統計量	158.11
degree of freedom	19
p-value	5.91e-24

變數	Agr	Min	Man	PS	Con
可解釋資料比例	0.995000000	0.4099547	0.9950000	0.1848297	0.3213772

變數	SI	Fin	SPS	TC
可解釋資料比例	0.6932345	0.2097172	0.7732976	0.3186261

約有54.5%的總變異數能夠透過2factor模型解釋，解釋能力尚可偏低，但相較於一個factor的模型，整體百分比提升許多。相較於1factor 模型，除了交通運輸業(TC)下降0.1%之外，各個變數可解釋變異比例都提升。交通運輸業(TC)的可解釋變數比例是31.9%，仍然有待加強。製造業(Man)更是從原本可解釋變異數由44.6%提升到99.5%，相當不錯。服務業(SI)及個人服務業(SPS)的部分，可解釋變異數都從大概五成分別提升到69.3%及77.3%，表現都算可以接受且偏好。礦業(Min)提升幅度不錯，但2factor模型可解釋變數比例才41.0%，還有進步的空間。金融業(Fin)及營造業(Con)在這個模型下也才分別是20.9%及32.1%，表現仍然不佳。

接著，我們進行三個factor的模型：

chi-square 檢定統計量	145.1
degree of freedom	12
p-value	5.57e-25

變數	Agr	Min	Man	PS	Con
可解釋資料比例	0.995000000	0.5342528	0.9950000	0.9950000	0.3629195

變數	SI	Fin	SPS	TC
可解釋資料比例	0.6942076	0.2201504	0.7824100	0.3489157

約有65.9%的總變異數能夠透過3factor模型解釋。相較於2factor模型，所有變數的可解釋變數比例均提升了。在農業(Agr)、製造業(Man)及能源供給業(PS)均達到99.5%，數據相當不錯。服務業(SI)及個人服務業(SPS)的可解釋變數比例均達到不錯的比例，分別是69.4%及78.2%。礦業(Min)因為我們使用3因子模型而達到53.4%，相較於雙因子模型算是表現尚可。營造業(Con)金融業(Fin)及交通運輸業(TC)可解釋變數比例皆不儘人意，分別是36.3%、22.0%及34.9%。

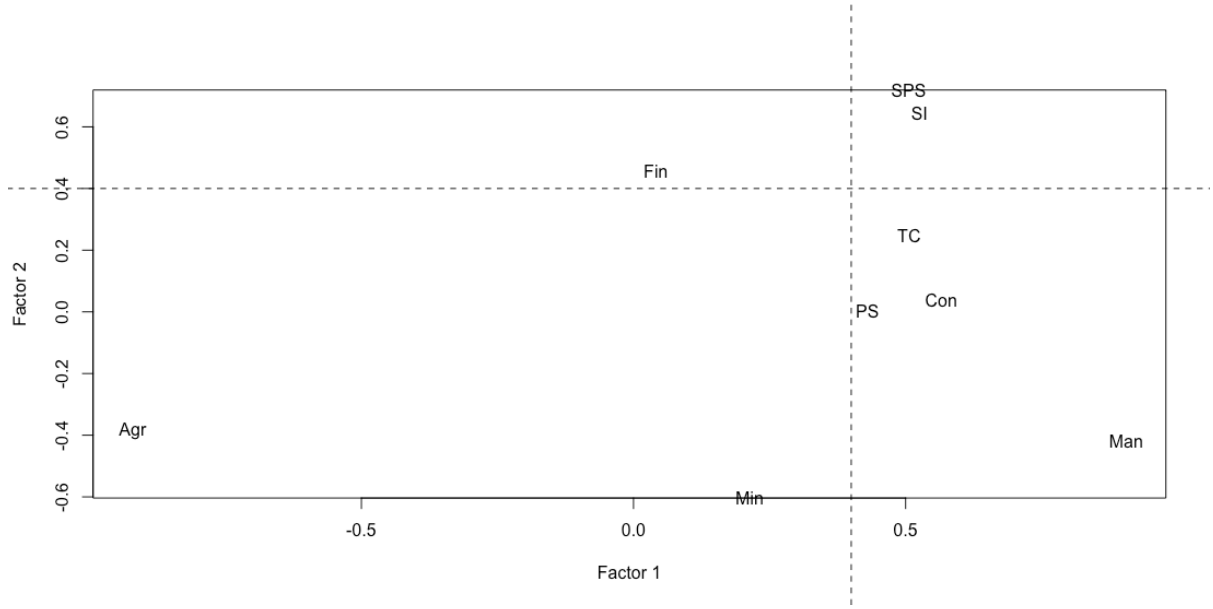
最終，我們進行第四個因子分析：

chi-square 檢定統計量	113.56
degree of freedom	6
p-value	3.66E-22

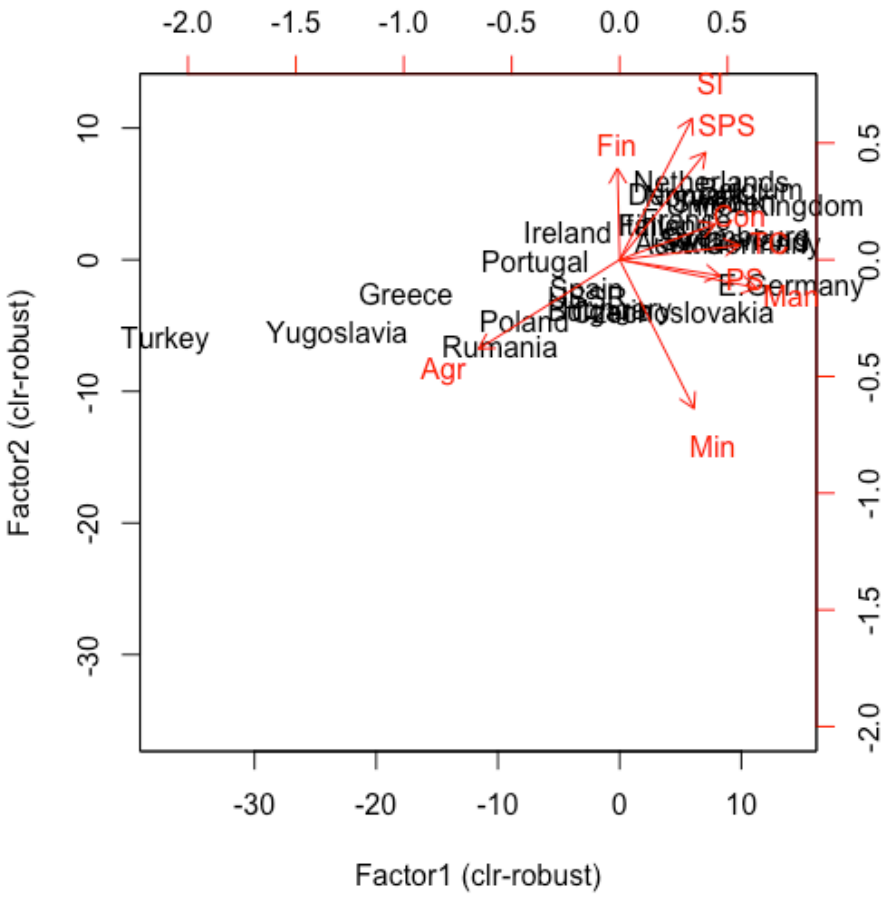
變數	Agr	Min	Man	PS	Con
可解釋資料比例	0.995000000	0.5852025	0.9950000	0.9950000	0.4853668

變數	SI	Fin	SPS	TC
可解釋資料比例	0.8366518	0.4758879	0.9950000	0.4682174

約有75.9%的總變異數能夠透過4factor模型解釋。其中比較特別的是p-value不論我們實驗幾個factor均無法達到顯著，但這次的p-value是四次裡面最大的，儘管如此，因為可解釋變異數比例不錯，因此我們認定此為我們使用的最終模型。相較於3factor模型，所有變數的可解釋變數比例均提升了。在農業(Agr)、製造業(Man)、個人服務業(SPS)及能源供給業(PS)均達到99.5%，數據相當不錯。服務業(SI)的可解釋變數比例均達到不錯的比例83.6%。礦業(Min)因為我們使用4因子模型而達到58.5%，相較於三因子模型算是表現尚可。營造業(Con)金融業(Fin)及交通運輸業(TC)可解釋變數比例皆不儘人意，分別是48.5%、47.6%及46.8%。



首先，以(0.4,0.4)作為中心，理想的狀況下應該是圖的左上及右下各一群。上圖我們可以看到右下一群所包含的變數包括：交通運輸(TC)、電力供給(PS)、營造業(Con)、製造業(Man)，而這些變數主要受到Factor 1所影響。而左上的變數僅有一個：金融業(Fin)，由圖上數值比例可以知道主要受到Factor 2所影響。右上跟左下這兩區的分群，代表這些區域的變數可能會有重疊因子(Overlapping)的情況出現。



水平方向主要受到Factor 1所影響，可以清楚看到金融業(Fin)幾乎垂直於x軸，表示這個變異數主要受到Factor 2 影響，跟我們前一個圖所分類出來的結果是可以相呼應的。我們在前一個圖所看到有四個產業交通運輸(TC)、營造業(Con)、能源供給(PS)、製造業(Man)受到Factor 1影響，因此他們在biplot的圖上，這四個產業就會比較接近水平。其他如礦業(Min)、服務業(SI)、個人服務業(SPS)同時受到Factor 1及Factor 2所影響，因此並無特別傾向哪一邊，其向量都有一定的角度。

關於國家所在的位置(即資料點)，可以發現在農業(Agr)方向包含土耳其、南斯拉夫、希臘、羅馬尼亞、波蘭，這些國家主要以農業為主，資料的時間點在1979年，也就是第二次世界大戰結束約三十年，這時候的希臘主要還是以第一級產業為主，其他南歐國家包括西班牙、葡萄牙也都傾向這樣的產業型態。捷克、保加利亞主要以礦業(Min)為主。東德(E. Germany)主要以製造業(Man)能源供給業(PS)為主。其他如丹麥、荷蘭、英國、法國、芬蘭等等都是屬於現在西歐或者北歐主要先進國家，他們所發展的產業主要也會以第二級產業及第三集產業為主。