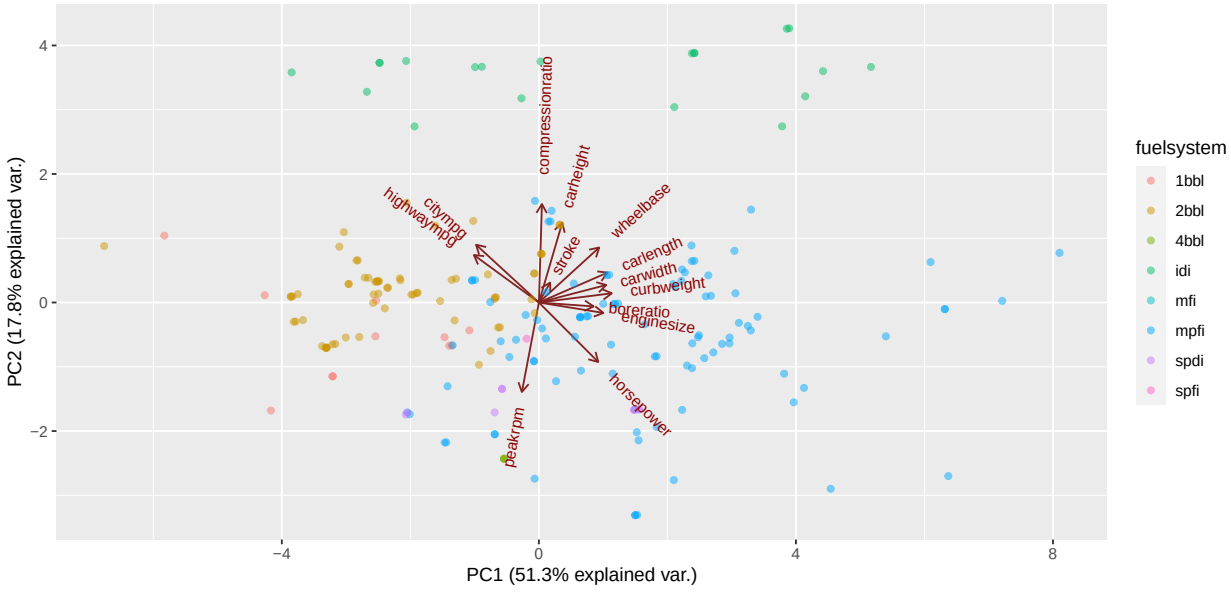
Applied Multivariate Analysis Homework 4

110024516 邱繼賢

(a)

將 13 個變數的 correlation matrix 做 eigen decomposition 獲得前兩個 principal components，並且繪製 biplot

	PC1	PC2
wheelbase	0.31	0.28
carlength	0.35	0.15
carwidth	0.35	
carheight	0.12	0.41
curbweight	0.37	
enginesize	0.33	
boreratio	0.28	
stroke		0.1
compressionratio		0.5
horsepower	0.3	-0.3
peakrpm		-0.46
citympg	-0.32	0.3
highwaympg	-0.33	0.24



- Biplot 是以前兩個 principal components 為軸，將資料點投影到兩軸所形成的二維平面，而 13 個變數也以各自所占前兩個 principal components 的比例形成在二維平面中的向量。
- 將資料點以類別型變數 *fuelsystem* 做分類，可以看到相同顏色的資料點在 biplot 明顯的聚集在一起。
- 變數 *fuelsystem* 中前三多的類別 *mpfi, 2bbl, idi*，分別聚集於 biplot 中的右方、左方以及上方。

由此可知，藉由 principal components analysis 將資料的維度降至二維，依舊可以保留資料依變數 *fuelsystem* 不同所形成的群聚特徵。

(b)

將 205 筆資料 (standardized) 的 correlation matrix $R_{205\times205}$ 和 1 的差距當作距離：$D_{205\times205} = 1_{205}1'_{205} - R_{205\times205}$，然後進行 multidimensional scaling：

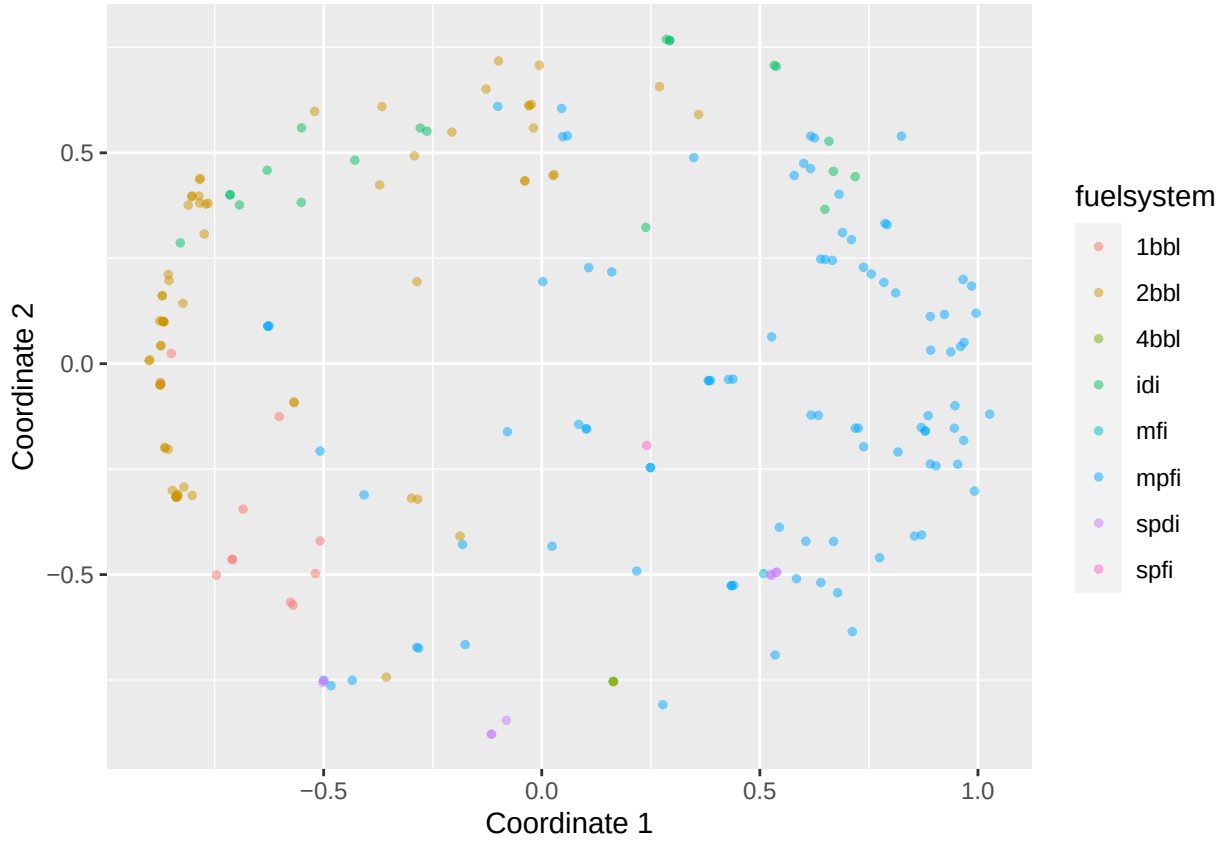
(1) 從距離矩陣 $D_{205\times205}$ 求得 data matrix 的 $n \times n$ inner products matrix $B_{205\times205} = XX^T$

(2) 然後對矩陣 B 做 eigen decomposition

$$B = U_1\Lambda_1U_1^T$$

where U_1 contains the first q eigenvectors and $\Lambda_1^{\frac{1}{2}} = diag(\lambda_1^{\frac{1}{2}}, ..., \lambda_q^{\frac{1}{2}})$ the q non-zero eigenvalues with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_q$.

(3) 只選取矩陣 $U_1\Lambda_1^{\frac{1}{2}}$ 的前兩個 columns，將資料降維到以前兩個 eigenvectors 為兩軸所形成的平面



- 兩軸的範圍都很小，這是因為我們計算距離是使用 $1 - r_{ij}$，此值最大值只為 2，並不是資料點之間真實的距離，故降維後的資料點就會十分集中在一個小範圍。

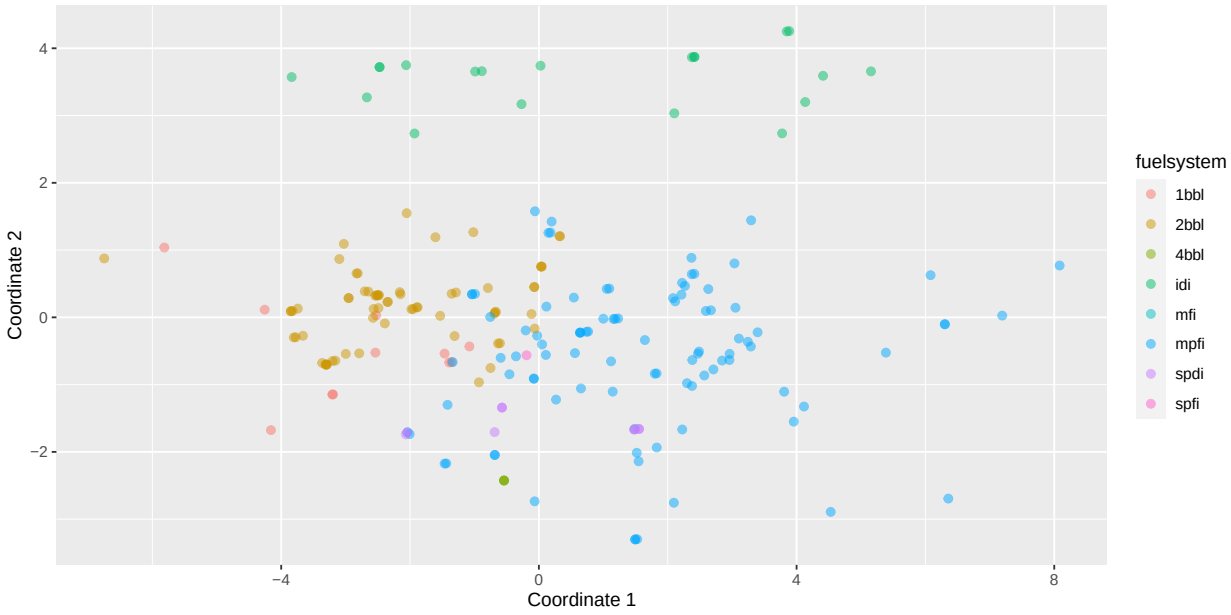
- 降維後顏色相同的資料點，依舊有聚集的現象。

- 變數 *fuelsystem* 中 *mpfi, 2bbl, idi* 三個類別在二維平面的分布位置一樣大致落在右方、左方、上方。

以 $1 - r_{ij}$ 為距離進行 multidimensional scaling analysis 的降維方式，雖然沒辦法呈現原始資料的真實距離，但還是可以呈現出資料依據變數 *fuelsystem* 不同所形成的聚集現象。

(c)

以 205 筆資料間的 Euclidean distance matrix D 進行和 **(b)** 一樣的 multidimensional scaling analysis 將資料降維至二維平面



- 資料分佈和 **(a)** 中的 biplot 完全一樣。

因為用 Euclidean distance 為距離所求出的 $B = XX^T = \left(U\Lambda^{\frac{1}{2}}\right)\left(U\Lambda^{\frac{1}{2}}\right)^T$，其中 X 就會是原始資料 standardized 後的 data matrix，然而 $X = U\Lambda^{\frac{1}{2}}V^T$，$U\Lambda^{\frac{1}{2}} = XV$ 就是 PCA scores，所以兩個圖形的資料點分佈才會完全一致。

```
library(vegan)
```

```
## 載入需要的套件：permute
```

```
##
## 載入套件：'permute'
```

```
## 下列物件被遮斷自 'package:devtools':
##
##     check
```

```
## 載入需要的套件：lattice
```

```
## This is vegan 2.5-7
```

```
iso = isomap(D_car, ndim = 2, k = 10)
graph_dis = as.matrix(isomapdist(D_car, k=10))
round(graph_dis[1:6,1:6],2)
```

```
##      1    2    3    4    5    6
## 1 0.00 0.00 7.79 6.51 7.47 6.54
## 2 0.00 0.00 7.79 6.51 7.47 6.54
## 3 7.79 7.79 0.00 4.22 2.81 2.78
## 4 6.51 6.51 4.22 0.00 1.96 1.44
## 5 7.47 7.47 2.81 1.96 0.00 0.93
## 6 6.54 6.54 2.78 1.44 0.93 0.00
```

```
ggplot(as.data.frame(iso$points)) +
    geom_point(aes(-Dim1, Dim2, color = car_data$fuelsystem), size = 2, alpha = 0.5) +
    labs(x = "Coordinate 1", y = "Coordinate 2", color = "fuelsystem")
```