Àlgebra Lineal (22.404), Pràctica 1, Curs 2020-21, Semestre 2

Mesures de qualitat de l'aire de Nova York, Estats Units d'Amèrica

Resolució

Part prèvia

Assignem el *data frame* `air quality` a la variable `nycair`:

```
nycair <- airquality
```

Pregunta 1

Generem el vector `label` i la matriu `W` en les següents línies de codi.

```
label <- nycair[,5]
W <- as.matrix(nycair[,1:4])
class(label)
```

```
## [1] "integer"
```

```
class(W)
```

```
## [1] "matrix" "array"
```

Pregunta 2

Comprovem que ens falten 44 **dades** amb la següent línia de codi.

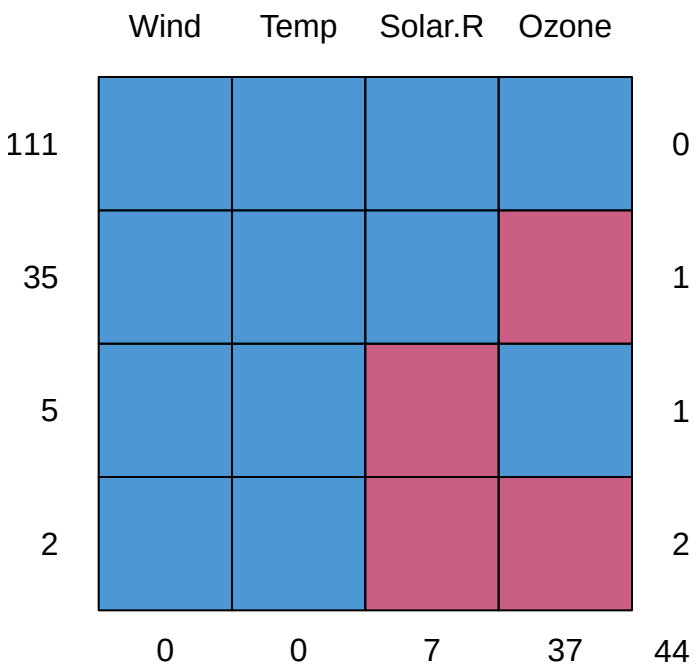
```
sum(is.na(W))
```

```
## [1] 44
```

Pregunta 3

Carreguem la llibreria `mice` i fem servir la instrucció `md.pattern(W)` com se'ns suggereix a l'enunciat de la pràctica.

```
library(mice)
md.pattern(W)
```




```
##     Wind Temp Solar.R Ozone
## 111    1    1       1     1  0
## 35     1    1       1     0  1
## 5      1    1       0     1  1
## 2      1    1       0     0  2
##        0    0       7    37 44
```

Podem veure que:

- hi ha 111 observacions on no hi falta cap dada;
- hi ha 35 observacions on només falta la dada de la variable `Ozone`;
- hi ha 5 observacions on només hi falta la dada de la variable `Solar.R`;
- hi ha 2 observacions on falten les dades de les variables `Solar.R` i `Ozone`.

Per tant, en total tenim les $35 + 5 + 2 \cdot 2 = 44$ dades que faltaven, de les que 37 corresponen a dades que falten de la variable `Ozone`.

Pregunta 4

Completem les dades (*data imputation*) amb el mètode anomenat *predictive mean matching* amb les següents línies de codi:

```
nycairi <- mice(data = W,m = 5, method= "pmm", maxit = 50, seed = 500)
W <- complete(nycairi,1)
```

Podem veure com ara ja no ens falta cap dada:

```
sum(is.na(W))
```

```
## [1] 0
```

Pregunta 5

Farem servir `R` per representar gràficament la primera i la segona variables en un diagrama de dispersió. Podeu veure el resultat a la Figura 1.

```
plot(W[,1],W[,2],type="p",col=label,xlab="Ozone",ylab="Solar.R")
```

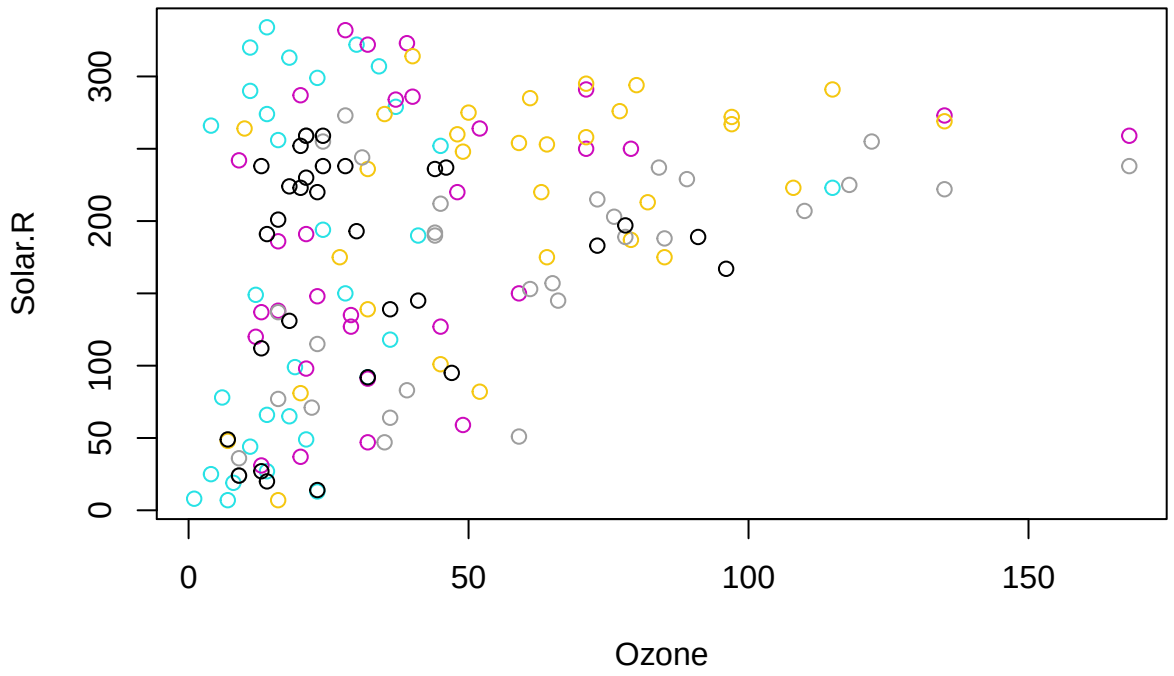


Figure 1: Diagrama de dispersió de la variable Ozone versus la variable Solar.R.

També farem servir `R` per representar gràficament la segona i la tercera variables en un diagrama de dispersió. Podeu veure el resultat a la Figura 2.

```
plot(W[,2],W[,3],type="p",col=label,xlab="Solar.R",ylab="Wind")
```

En cap dels dos casos podem veure una separació clara dels punts de colors, que representen cada un dels mesos de l'any entre maig i setembre.

Pregunta 6

Escalem les dades de la matriu `W` amb la següent línia de codi:

```
Ws<-scale(W,center = TRUE,scale = TRUE)
```

Per comprovar que cada variable té ara mitjana zero i desviació tipus 1, podem fer servir la instrucció `apply`. Si mirem l'ajuda que ens proporciona `R`:

```
help("apply")
```

podem veure que el segon argument ens permet escollir si apliquem la funció per files o per columnes:

> "... for a matrix 1 indicates rows, 2 indicates columns..."

Per tant, en el nostre cas, com que volem fer una comprovació per columnes, la instrucció que hem de fer servir és:

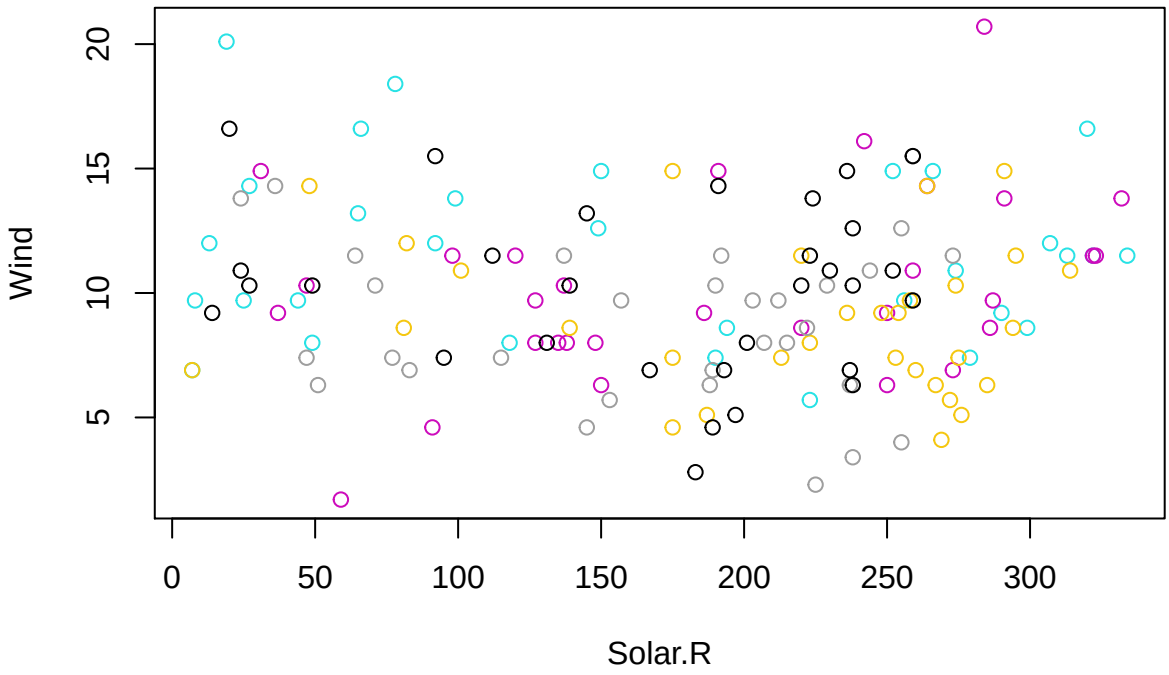


Figure 2: Diagrama de dispersió de la variable Solar.R versus la variable Wind.

```
apply(Ws,2,mean)
```

```
##         Ozone       Solar.R          Wind          Temp
##  4.535933e-18 -9.082957e-17 -3.732756e-17  6.973588e-16
```

```
apply(Ws,2,sd)
```

```
##   Ozone Solar.R    Wind    Temp
##       1       1       1       1
```

Noteu que per a les mitjanes aritmètiques no obtenim exactament zeros, però això és producte dels errors numèrics associats a les operacions (zero de màquina).

Pregunta 7

Calculem la matriu de covariàncies fent servir la instrucció `cov`.

```
CWs<-cov(Ws)
```

Com que volem saber quines dues variables (diferents) presenten la covariància més gran, per al càlcul del màxim, eliminem la diagonal de la matriu de covariàncies. Fem servir la instrucció `max` per calcular la màxima covariància i `which` per localitzar les dues variables. Això es pot veure a les següents línies de codi.

```
dim <- sqrt(length(CWs))
CWsABS<-abs(CWs-diag(dim))
which(CWsABS == max(CWsABS), arr.ind = TRUE)
```

```
##       row col
## Temp    4   1
## Ozone   1   4
```

També es pot veure a la Figura 3 el diagrama de dispersió d'aquestes dues variables, on es pot veure com el núvol de punts del diagrama de dispersió tendeix a distribuir-se de forma lineal.

```
plot(Ws[,1],Ws[,4],col=label,xlab="Ozone",ylab="Temp")
```

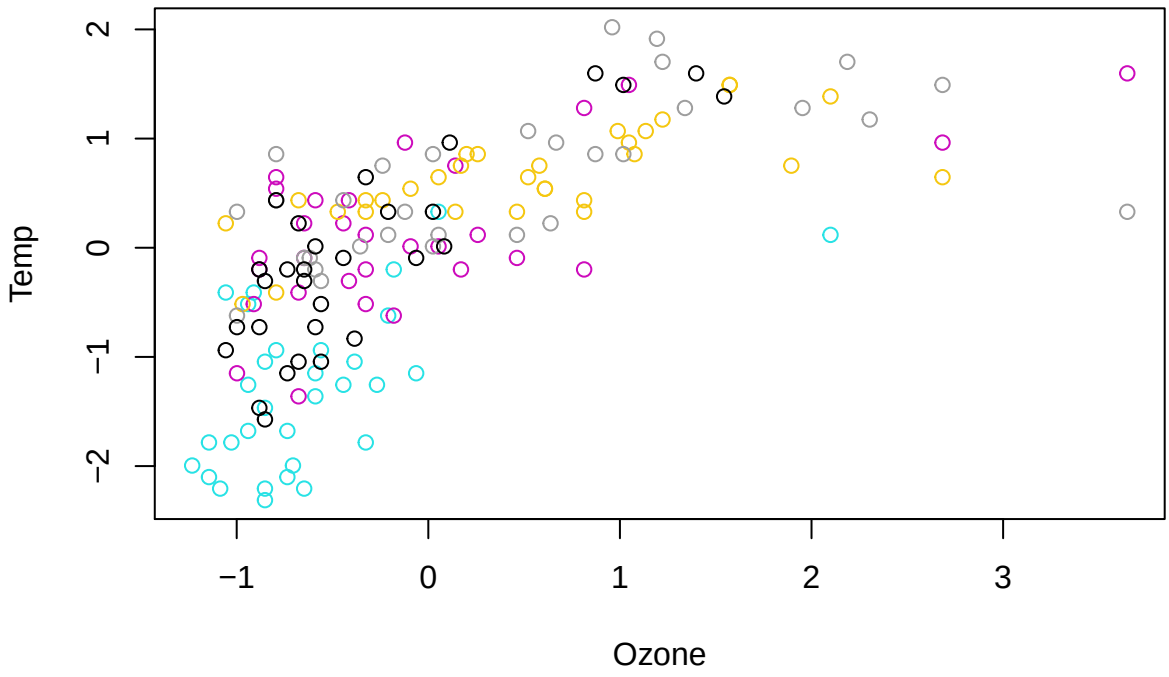


Figure 3: Diagrama de dispersió de la variable Ozone versus la variable Temp.

Pregunta 8

Com que volem saber quines dues variables (diferents) presenten la covariància més petita, per al càlcul del mínim, ho fem sobre tots els valors de la matriu de covariàncies `CWs`, en valor absolut. Fem servir la instrucció `min` per calcular la mínima covariància i `which` per localitzar les dues variables. Això es pot veure a les següents línies de codi.

```
CWsABS2<-abs(CWs)
which(CWsABS2 == min(CWsABS2), arr.ind = TRUE)
```

```
##         row col
## Wind      3   2
## Solar.R   2   3
```

També es pot veure a la Figura 4 el diagrama de dispersió d'aquestes dues variables, on es pot veure com el núvol de punts del diagrama de dispersió no segueix cap patró.

```
plot(Ws[,2],Ws[,3],col=label,xlab="Solar.R",ylab="Wind")
```

Pregunta 9

Calculem els valors i vectors propis de la matriu de covariàncies en les següents línies de codi.

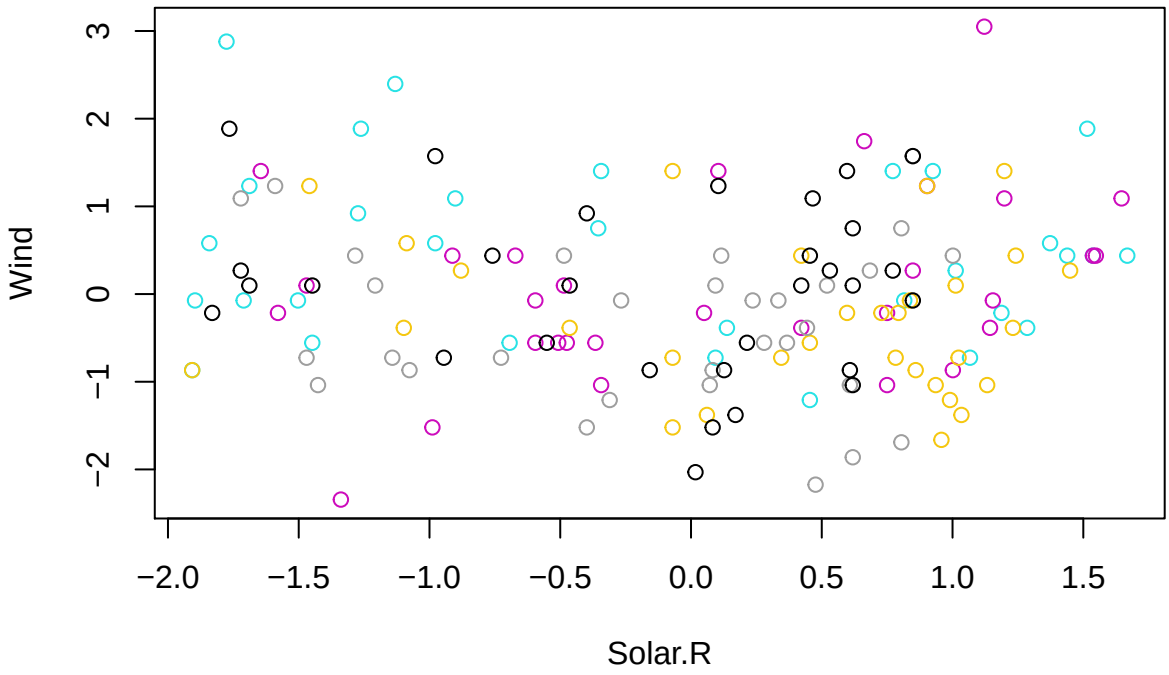


Figure 4: Diagrama de dispersió de la variable Solar.R versus la variable Wind.

```
eigCWs<-eigen(CWs)
P<-eigCWs$vectors
```

La matriu `P` conté, per columnes, les components principals. També calculem la variabilitat retinguda per les primeres `N` components principals (en percentatge), que mostrem a continuació. Com que hi ha quatre possibles valors de `N`, mostrem el resultats per als quatre valors:

```
N <- 1
sum(eigCWs$values[1:N])*100/4
```

```
## [1] 56.12794
```

```
N <- 2
sum(eigCWs$values[1:N])*100/4
```

```
## [1] 80.15364
```

```
N <- 3
sum(eigCWs$values[1:N])*100/4
```

```
## [1] 92.33167
```

```
N <- 4
sum(eigCWs$values[1:N])*100/4
```

```
## [1] 100
```

Evidentment, per a `N`= 4, obtenim una variabilitat acumulada del 100%.

Aquests quatre valors de la variabilitat retinguda es poden mostrar en una gràfica anomenada *scree plot* (Figura 5):

```
acc_var <- cumsum(eigCWs$values)*100/4
plot(1:4,acc_var,type="b",col="red",ylim=c(0,100),ann=FALSE)
```

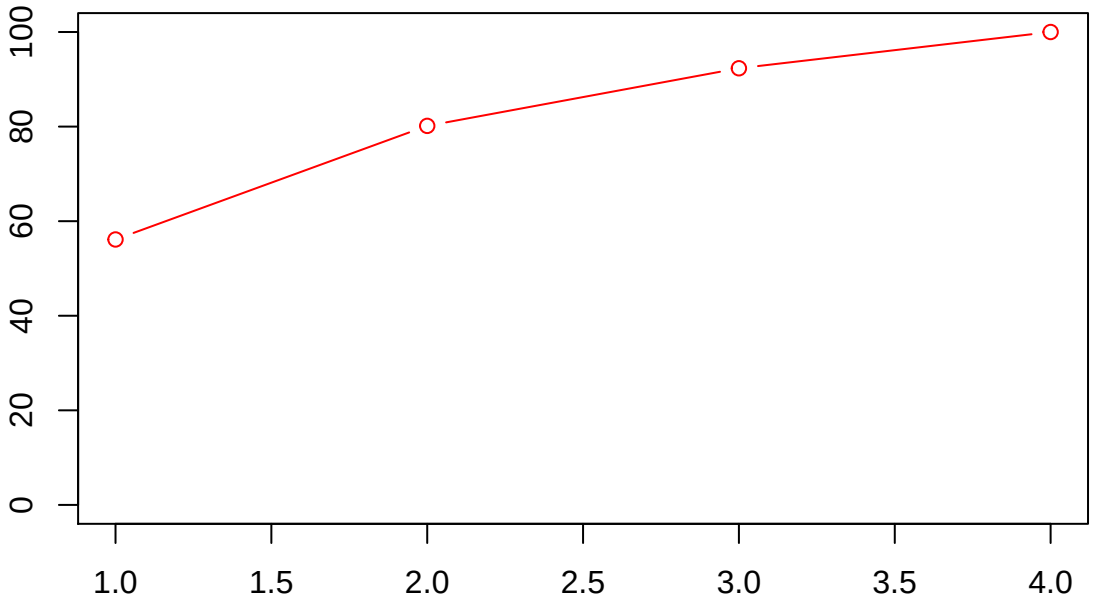


Figure 5: Variabilitat retinguda per les components principals.

Pregunta 10

Els coeficients (en valor absolut) de la primera component principal es poden calcular amb la següent línia de codi:

```
abs(P[,1])
```

```
## [1] 0.5939832 0.3258124 0.4606955 0.5734020
```

Es pot observar com la variable més important és la primera (`Ozone`, amb un valor de 0.5939832) seguida de la quarta (`Temp` amb un valor de 0.5734020). Noteu que, en aquest cas, aquestes dues variables coincideixen amb les variables amb la covariància màxima.

En el cas de la segona component principal, els coeficients (en valor absolut) són:

```
abs(P[,2])
```

```
## [1] 0.00645786 0.83372749 0.55083922 0.03785390
```

Es pot observar com la variable més important és la segona (`Solar.R`, amb un valor de 0.83372749) seguida de la tercera (`Wind` amb un valor de 0.55083922). En aquest cas, noteu dues coses:

- Aquestes dues variables coincideixen amb les variables amb la covariància mínima;
- En aquesta segona component principal, la contribució de la primera i de la quarta variables és pràcticament nul · la (0.00645786 i 0.03785390, respectivament).

Pregunta 11

La matriu `T` conté la projecció de les dades originals normalitzades (`Ws`) sobre les components principals. Amb la següent línia de codi, podem representar gràficament els 153 punts (Figura 6):

```
T <- Ws%*%P
plot(T[,1],T[,2],col=label,xlab="PC1",ylab="PC2")
```

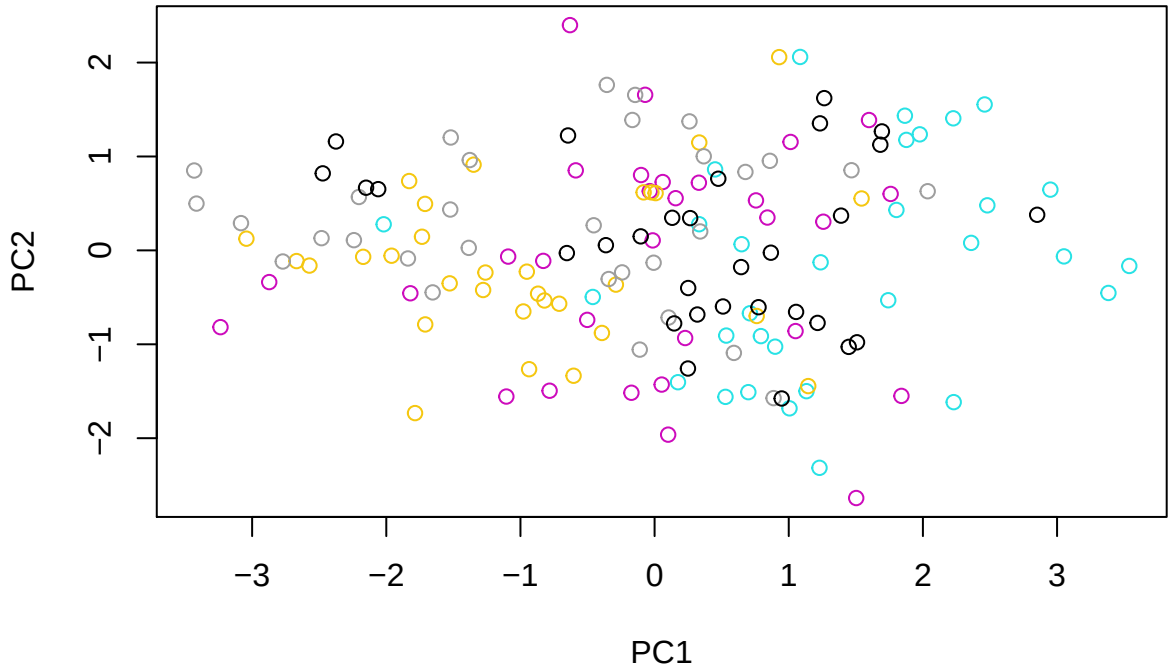


Figure 6: Projecció de les dades originals normalitzades sobre les dues primeres components principals.

Tot i que es pot intuir una certa separació entre els punts de color blau (més a la dreta) dels punts de color groc o gris (més a l'esquerra), la separació no és clara. Aquest és un exemple on PCA no ha estat capaç de separar clarament la nostra informació.

Pregunta 12

Fem servir la següent línia de codi per a representar gràficament **només** els 31 punts corresponents al mes de maig i els 31 punts corresponents al mes de juliol (Figura 7):

```
plot(T[c(1:31,62:92),1],T[c(1:31,62:92),2],col=label[c(1:31,62:92)],xlab="PC1",ylab="PC2")
```

En aquesta ocasió, sí és possible veure que els punts de color blau (corresponents al mes de maig) estan situats més a la dreta que els punts de color groc (corresponents al mes de juliol). Això és així perquè les característiques dels mesos de maig i juliol són significativament diferents.

Si representem gràficaments els mesos de maig i juliol (dades originals normalitzades) sobre les variables `Ozone` i `Solar.R` (vegeu la Figura 8), la separació també era possible tot i que una mica menys clara.

```
plot(Ws[c(1:31,62:92),1],Ws[c(1:31,62:92),2],col=label[c(1:31,62:92)],xlab="Ozone",ylab="Solar.R")
```

Pregunta 13

Els resultats demanats són:

- Falten 44 dades (pregunta 2).
- Falten 37 dades de la variable `Ozone`(pregunta 3).

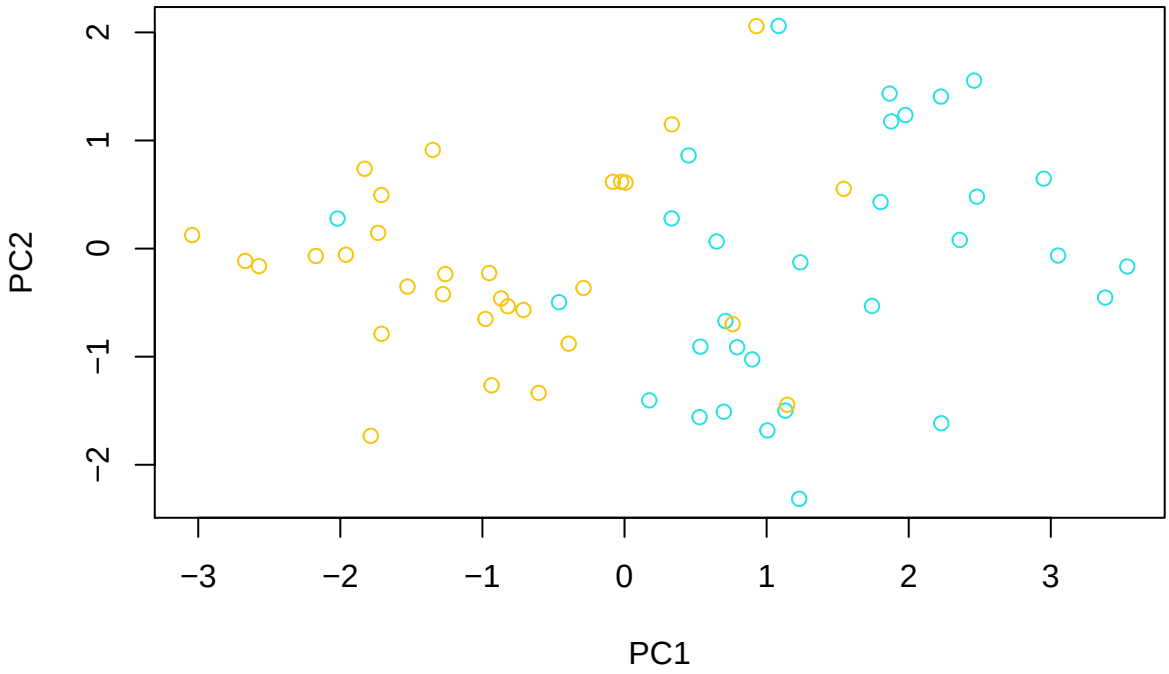


Figure 7: Projecció de les dades originals normalitzades corresponents als mesos de maig i juliol sobre les dues primeres components principals.

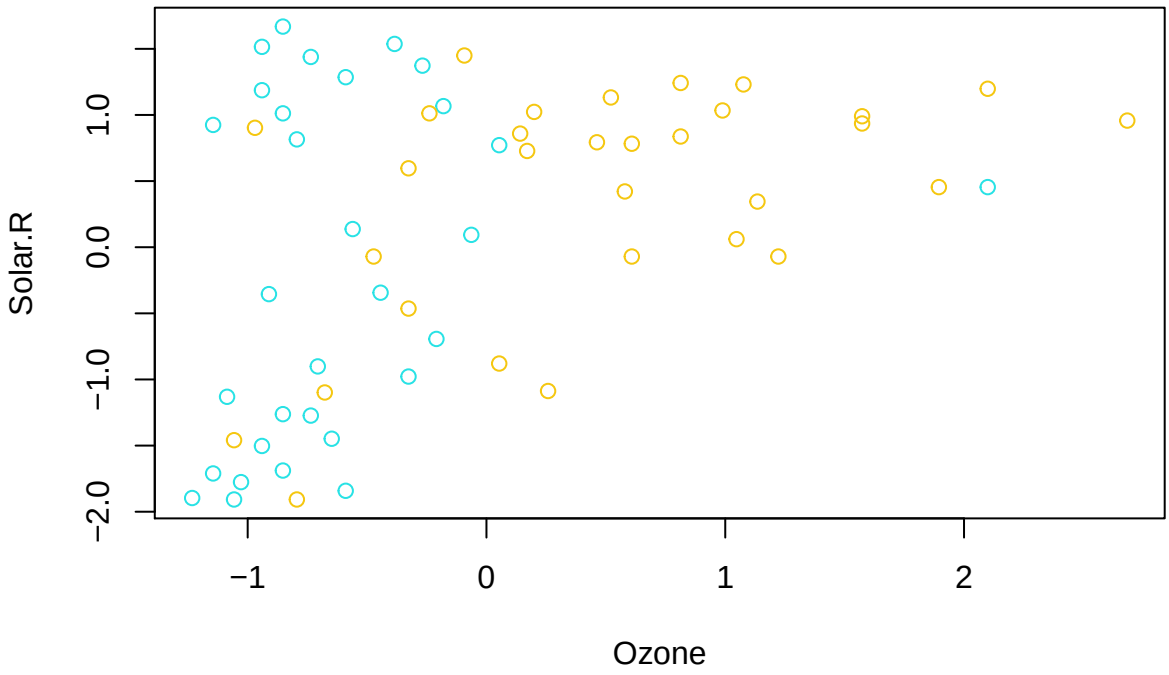


Figure 8: Diagrama de dispersió de la variable Ozone versus la variable Solar.R per a les dades originals normalitzades corresponents als mesos de maig i juliol.

```r
W[5,2] # pregunta 4
```

```
## [1] 27
```

- El segon argument de la instrucció `apply(Ws,?,mean)` és un `2`.

```r
Ws[3,2] # pregunta 6
```

```
##    Solar.R
## -0.3546998
```

```r
sum(CWs) # pregunta 7
```

```
## [1] 4.684205
```

- Les variables amb la màxima covariància són la primera (`Ozone`) i la quarta (`Temp`).
- Les variables amb la mínima covariància són la segona (`Solar.R`) i la tercera (`Wind`).
- La variabilitat retinguda per les primeres components principals és:
  - Si `N=1`, la variabilitat acumulada és 56.12794%.
  - Si `N=2`, la variabilitat acumulada és 80.15364%.
  - Si `N=3`, la variabilitat acumulada és 92.33167%.
  - Si `N=4`, la variabilitat acumulada és 100%.
- Les variables més importants en la primera component principal són, per aquest ordre, la *primera* (`Ozone`) i la *quarta* (`Temp`).
- Les variables més importants en la segona component principal són, per aquest ordre, la *segona* (`Solar.R`) i la *tercera* (`Wind`).