TP 2

Cordoval Chloe et Laabsi Zakaria

8 Novembre 2020

Énoncé

Charger dans R les données relatives aux vins de Loire (wine.csv).

Les données contiennent deux variables qualitatives (Appellation: "Label" = {Bourgueil, Chinon, Saumur } et Sol: "Soil" = { Env1, Env2, Env3=référence, Env4 }) et $p = 29$ variables quantitatives décrivant diverses intensités sensorielles (odeur, arôme, gout, couleur etc). Les vins seront traduits en nuage dans l'espace des 29 variables quantitatives, $\mathbb{R}^{29}$. On notera X la matrice dont les colonnes sont les 29 variables quantitatives centréesréduites, Y la matrice dont les colonnes sont les indicatrices d'appellations, et Z celle dont les colonnes sont les indicatrices de sols.

On notera $W = \frac{1}{n} I_n$ la matrice des poids des individus et $M = \frac{1}{p} I_p$ celle des poids des variables. Ici, tout est équipondéré.

Question 1

La matrice X dont les colonnes sont les 29 variables quantitatives centrées réduites :

Quatre fonctions à utiliser :

- écart - type : sd()

- racine carrée : sqrt()

- moyenne : means()

- effectif : longueur du vecteur : length()

Avec la fonction "apply", on évite l'utilisation explicite de construction de boucle avec MAGIN = 2 ce qui veut dire qu'elle fonctionne sur les colonnes et si on avait voulu sur les lignes, on aurait noté MAGIN = 1.

Ce qui donnera :

```
#Création de la matrice X
V.quantitative <- wine[4:32]

C_R <- function(x){
  w <- (x-mean(x))/(sd(x)*sqrt(1-1/length(x)))
return(w)
}
```

```r
X <- apply(as.matrix(V.quantitative ), 2, C_R)
```

Nous avons maintenant X , une matrice des variables quantitatives centrées réduites du tableau wine.

Nous allons maintenant construire la matrice Y dont les colonnes sont les indicatrices des appellations := "Label" = {Bourgueil, Chinon et Saumur}. Ce qui donnera :

```r
#Création de la matrice Y
Y <- matrix(c(wine$Label == "Saumur",wine$Label == "Chinon", wine$Label == "Bourgueuil"), nc
i <- which(Y==TRUE, arr.ind = FALSE)
colnames(Y) <- c("Saumur","Chinon", "Bourgueuil")
Y[i]<-1

print(Y)
```

```
   Saumur Chinon Bourgueuil
      1      0          0
      1      0          0
      0      0          1
      0      1          0
      1      0          0
      0      0          1
      0      0          1
      1      0          0
      0      1          0
      1      0          0
      1      0          0
      1      0          0
      1      0          0
      1      0          0
      0      0          1
      0      0          1
      0      1          0
      0      1          0
      0      0          1
      1      0          0
      1      0          0
```

Nous allons maintenant construire la matrice Z dont les colonnes sont les indicatrices de Sol := "Soils" = {Env1, Env2, Référence = Env3 et Env4}. Ce qui donnera :

```r
#Création de la matrice Z

Z <- matrix(c(wine$Soil=="Env1",wine$Soil=="Env2",wine$Soil=="Reference",wine$Soil=="Env4")
j <- which(Z==TRUE,arr.ind = FALSE)
colnames(Z) <- c("Env1","Env2","Env3","Env4")
Z[j] <- 1

print(Z)
```

```
   Env1 Env2 Env3 Env4
    1    0    0    0
```

```
1    0    0    0
1    0    0    0
0    1    0    0
0    0    1    0
0    0    1    0
0    0    1    0
1    0    0    0
1    0    0    0
0    1    0    0
0    1    0    0
0    1    0    0
0    0    1    0
0    0    1    0
1    0    0    0
0    0    1    0
0    0    1    0
0    1    0    0
1    0    0    0
0    0    0    1
0    0    0    1
```

Nous allons construire maintenant la matrice W qui est composé du poids des individus fois la matrice identité de taille du nombre des individus. Or on sait qu'il y a 21 variables quantitatives donc 21 individus. Ce qui donnera :

```r
#Création de la matrice W

W <- matrix(0, nrow = length(X[,1]), ncol = length(X[,1]))
diag(W) <- 1/length(X[,1])
```

Nous voulons calculer maintenant M qui vaut $\frac{1}{p}Id_p$ qui est la matrice des poids des variables = 29. Ce qui donnera :

```r
#Création de la matrice M

M <- matrix(0, nrow = length(X[1,]), ncol = length(X[1,]))
diag(M) <- 1/length(X[1,])
```

a) Rappeler pourquoi $\forall j,\ \Pi_Y x^j = \Pi_{Y^c} x^j$. Rappeler l'interprétation statistique de $\|\Pi_Y x^j\|_W^2$

On sait que

$$< Y >=< 1 > \oplus < Y^c > \Rightarrow\ \Pi_Y = \Pi_{\mathbf{1}} + \Pi_Y^c$$

et ainsi , on a

$$\Pi_Y \left(x^j\right) = \Pi_{\mathbf{1}} \left(x^j\right) + \Pi_{Y^c} \left(x^j\right)$$

Or on sait que $\Pi_1 \left(x^j\right) = 0$ car x^j est centrée, et donc $x^j \perp \mathbf{1}$

Donc on a bien $\boxed{\forall j, \Pi_Y x^j = \Pi_{Y^c} x^j}$.

Maintenant, lorsqu'on développe $\left\|\Pi_Y\left(x^j\right)\right\|_{\mathrm{W}}^2$, on trouve

$$\begin{aligned}\left\|\Pi_Y\left(x^j\right)\right\|_{\mathrm{W}}^2 &= \langle \Pi_Y\left(x^j\right) \mid \Pi_Y\left(x^j\right)\rangle_{\mathrm{W}} \\ &= \langle \sum_{j=1}^{J} \bar{x}^j y_j \mid \sum_{k=1}^{J} \bar{x}^k y_k\rangle_{\mathrm{W}} \\ &= \sum_{j=1}^{J}\sum_{k=1}^{J} \langle \bar{x}^j y_j \mid \bar{x}^k y_k\rangle_{\mathrm{W}} \\ &= \sum_{j=1}^{J}\sum_{k=1}^{J} \bar{x}^j \bar{x}^k \langle y_j \mid y_k\rangle_{\mathrm{W}}\end{aligned}$$

Nous avons deux cas:

- 1^{er} cas : $j \neq k$, alors $\langle y_j \mid y_k\rangle_w = 0$
- $2^{\text{ème}}$ cas : $j = k$, alors $\langle y_j \mid y_k\rangle_w = \|y_j\|_w^2$

On obtient alors:

$$\boxed{\left\|\Pi_Y\left(x^j\right)\right\|_{\mathrm{W}}^2 = \sum_{j=1}^{J} \bar{x}^j \bar{x}^j \|y_j\|_{\mathrm{W}}^2}$$

car X est centrée et donc sa moyenne est nulle.

On a réussi à exprimer $\left\|\Pi_Y\left(x^j\right)\right\|_{\mathrm{W}}^2$ comme la variance interclasse.

b) Programmer et calculer Π_Y, puis, pour chaque x^j : $\Pi_{x^j}, \mathrm{tr}\left(\Pi_{x^j}\Pi_Y\right)$.

On va introduire une fonction Projection qui nous permettra de calculer chaque projection au cours de ce TP 2.

```
Projection  <- function(B){
  Projection_B <- B %*% solve(aperm(B) %*% W %*% B)%*% aperm(B) %*% W
return(Projection_B)
}
```

Le projecteur d'une matrice Y est $\Pi_Y = Y(Y'WY)^{-1}Y'W$. Ce qui donnera :

```
Projection_Y <- Projection(Y)
```

Il s'agit içi de faire une boucle "for" qui calcule pour chaque colonne j la projection sur x^j en utilisant la fonction "Projection".

Pour concaténer des valeurs et variables dans R, il est possible d'utiliser la fonction paste() puis attribuer une valeur à un nom dans un environnement. Ce qui donnera :

```
for (j in 1:29){

  N <- paste("Le projecteur de X ", j ,collapse = NULL, recycle0 = FALSE )
```

```
  assign( N,value = Projection(matrix(X[,j])) , pos = -1, inherits = FALSE, immediate = TRUE
}
```

Maintenant nous allons créer une fonction "Trace" qui permet de calculer la trace d'un produit de deux matrices. Or on sait que la trace d'une matrice est la somme des valeurs sur sa diagonale. Ce qui donnera:

```
Trace <- function(A,B){

  C <- A %*% B
  trace_C <- sum(diag(C))

return(trace_C)
}
```

Nous allons initialiser un vecteur Trace_Projection_XY à 0; ce qui donnera:

```
Trace_projection_XY <- numeric(29)
```

Il faut maintenant introduire une boucle "for" . On utiliera la fonction "get0" qui recherche et appelle un objet de données avec l'option supplémentaire pour spécifier ce qui doit se passer au cas où l'objet de données que l'on recherche n'existe pas.

```
for (j in 1:29){

  Trace_projection_XY[j] <- Trace(get0(N ,ifnotfound = "pas disponible" ), Projection_Y)

}

Trace_projection_XY
```

```
 0.01119577 0.01119577 0.01119577 0.01119577 0.01119577 0.01119577
 0.01119577 0.01119577 0.01119577 0.01119577 0.01119577 0.01119577
 0.01119577 0.01119577 0.01119577 0.01119577 0.01119577 0.01119577
 0.01119577 0.01119577 0.01119577 0.01119577 0.01119577 0.01119577
 0.01119577 0.01119577 0.01119577 0.01119577 0.01119577
```

Rappelez l'interprétation statistique de cette dernière quantité :

Par définition d'un projecteur, on a :

$$\mathrm{tr}\left(\Pi_{x^j}\Pi_Y\right) = \mathrm{tr}\left(x^j\left(x^{j'}Wx^j\right)^{-1}x^{j'}W\Pi_Y\right)$$

Or on sait que

$$x^{j'}Wx^j = \langle x^j \mid x^j\rangle = \left\|x^j\right\|_{\mathrm{W}}^2$$

D'où l'expression

$$\begin{aligned}\operatorname{tr}\left(\Pi_{x^j}\Pi_Y\right) &= \operatorname{tr}\left(x^j \left\|x^j\right\|_{\mathrm{W}}^{-2} x^{j\prime} W \Pi_Y\right)\\ &= \operatorname{tr}\left(\left\|x^j\right\|_{\mathrm{W}}^{-2} x^{j\prime} W \Pi_Y x^j\right)\end{aligned}$$

Enfin, on a

$$\begin{aligned}x^{j\prime} W \Pi_Y x^j &= \langle x^j \mid \Pi_Y x^j\rangle_{\mathrm{W}}\\ &= \langle x^j \mid \Pi_Y^* \Pi_Y x^j\rangle_{\mathrm{W}}\\ &= \langle \Pi_Y x^j \mid \Pi_Y x^j\rangle_{\mathrm{W}}\\ &= \left\|\Pi_Y x^j\right\|_{\mathrm{W}}^2\end{aligned}$$

$$\text{car l'adjoint } \Pi_Y^* = \Pi_Y \text{ et } \Pi_Y^2 = \Pi_Y$$

En appliquant la trace d'un scalaire

$$\operatorname{tr}\left(\frac{\left\|\Pi_Y x^j\right\|_{\mathrm{W}}^2}{\left\|x^j\right\|_{\mathrm{W}}^2}\right) = \frac{\left\|\Pi_Y x^j\right\|_{\mathrm{W}}^2}{\left\|x^j\right\|_{\mathrm{W}}^2} = \left\|\Pi_Y x^j\right\|_{\mathrm{W}}^2 \quad \text{car } \left\|x^j\right\|_{\mathrm{W}}^2 = 1$$

On obtient ainsi

$$\boxed{\operatorname{tr}\left(\Pi_{x^j}\Pi_Y\right) = \left\|\Pi_Y x^j\right\|_{\mathrm{W}}^2 = R^2\left(x^j, Y\right)}$$

$\operatorname{tr}(\Pi_{x^j}\Pi_Y)$ est donc la liaison entre x^j et Y qui montre que chaque intensité sensorielle est plus ou moins liée aux appellations.

c) On note $R = XMX'W$. Programmer et calculer $\operatorname{tr}(R\Pi_Y)$.

Dans un premier temps, exprimons $R = XMX'W$; ce qui donnera :

```
R <-  X %*% M %*% aperm(X) %*% W
```

Comme nous avons définis une fonction Trace permettant de calculer la trace d'un produit matriciel ; On aura donc pour la $\operatorname{tr}(R\Pi_Y)$:

```
Trace_RY <- Trace(R,Projection_Y )

print(Trace_RY)
```

```
0.1092588
```

Interprétez statistiquement cette quantité :

On exprime R et on applique les propriétés de la trace :

$$\begin{aligned}\operatorname{tr}(R\Pi_Y) &= \operatorname{tr}(XMX'W\Pi_Y)\\ &= \operatorname{tr}(MX'W\Pi_Y X)\\ &= \operatorname{tr}\left(\frac{1}{p}I_p X'W\Pi_Y X\right)\\ &= \operatorname{tr}\left(\frac{1}{p}I_p X'W\Pi_Y X I_p^{-1}\right)\end{aligned}$$

On a donc

$$\operatorname{tr}\left(\frac{1}{p}I_p X'W\Pi_Y X I_p^{-1}\right) = \frac{1}{p}\operatorname{tr}(X'W\Pi_Y X)$$

Considérons la diagonale de la matrice $X'W\Pi_Y X$.

$$X'W\Pi_Y X = \begin{pmatrix} x^{1\prime}W\Pi_Y x^1 & \cdots & \cdots & \cdots & * \\ \vdots & \ddots & & & \vdots \\ \vdots & & \boxed{x^{j\prime}W\Pi_Y x^j} & & \vdots \\ \vdots & & & \ddots & \vdots \\ * & \cdots & \cdots & \cdots & x^{J\prime}W\Pi_Y x^J \end{pmatrix} \quad \forall j \in \{1, ..., J\}$$

On a ainsi :

$\forall j \in \{1, ..., J\}$

$$\begin{aligned}x^{j\prime}W\Pi_Y x^j &= \langle x^j \mid \Pi_Y x^j \rangle_{\mathrm{W}}\\ &= \langle x^j \mid \Pi_Y^* \Pi_Y x^j \rangle_{\mathrm{W}}\\ &= \langle \Pi_Y x^j \mid \Pi_Y x^j \rangle_{\mathrm{W}}\\ &= \left\| \Pi_Y x^j \right\|_{\mathrm{W}}^2\end{aligned}$$

$$\text{car l'adjoint } \Pi_Y^* = \Pi_Y \text{ et } \Pi_Y^2 = \Pi_Y$$

Ainsi

$$\operatorname{tr}(R\Pi_Y) = \frac{1}{p}\operatorname{tr}(X'W\Pi_Y X) = \frac{1}{p}\sum_{j=1}^{J}\left\| \Pi_Y x^j \right\|_{\mathrm{W}}^2$$

Or

$$\left\| x^j \right\|_{\mathrm{W}}^2 = 1$$

D'où

$$\boxed{\text{tr}(R\Pi_Y) = \frac{1}{p}\sum_{j=1}^{J} R^2(x^j, Y)}$$

Ainsi $\text{tr}\,(R\Pi_Y)$ correspond à **la moyenne arithmétique** de $R^2(x^j, Y)$. Ceci est la liaison entre le groupe des variables quantitatives X et Y, i.e les différentes intensités sensorielles qui sont plus ou moins liées à l'appellation.

d) Nous avons obtenu le résultat pour notre trace , or on sait que

$$R^2 = \frac{\text{Var}(inter - app)}{\text{Var}(total)}.$$

Ce qui a donné dans le TP1 , $R^2 = 0.109$ qui est l'indicateur de la qualité de la partition trouvée dans le TP 0, approximativement. Donc on a

$$\boxed{R^2 = \text{tr}\,(R\Pi_Y)}$$

Question 2

Programmez puis calculer chaque $\text{tr}\,(\Pi_{x^j}\Pi_Z)\,, \quad \text{tr}\,(R\Pi_Z)$

Dans un premier temps, nous allons calculer la projection de Z grâce à notre fonction Projection définit précédemment. On aura donc :

```
Projection_Z <- Projection(Z)
```

Nous allons initialiser un vecteur Trace_Projection_XZ à 0; ce qui donnera:

```
Trace_Projection_XZ <- numeric(29)
```

Calculons la trace de la projection sur x^j avec la projection sur Z en utilisant la même méthode pour calculer le Trace_projection_XY en utilisant la fonction "get0" et la fonction "Trace" définit précédemment;

Ce qui donnera;

```
for (j in 1:29){

  Trace_Projection_XZ[j] <- Trace(get0(N), Projection_Z)

}

Trace_Projection_XZ
```

```
0.351679 0.351679 0.351679 0.351679 0.351679 0.351679 0.351679 0.351679
0.351679 0.351679 0.351679 0.351679 0.351679 0.351679 0.351679 0.351679
0.351679 0.351679 0.351679 0.351679 0.351679 0.351679 0.351679 0.351679
0.351679 0.351679 0.351679 0.351679 0.351679
```

Il ne reste plus qu'à calculer $\text{tr}\,(R\Pi_Z)$ et pour cela, nous allons utiliser notre fonction Trace. Ce qui donnera :

```r
Trace_R_projection_Z <- Trace(R,Projection_Z)

print(Trace_R_projection_Z)
```

```
0.365302
```

Interprétez ces résultats statistiquement :

Via les résultats précédents, on obtient :

$$\boxed{\mathrm{tr}\left(\Pi_{x^j}\Pi_Z\right) = R^2\left(x^j, Z\right)}$$

$$\boxed{\mathrm{tr}(R\Pi_Z) = \frac{1}{p}\sum_{j=1}^{J} R^2(x^j, Z)}$$

Avec une analyse similaire à celle de la question précédente, on peut donc en déduire que $\mathrm{tr}\left(\Pi_{x^j}\Pi_Z\right)$ représente **la liaison entre** x^j **et** Z.

De même , on a $\mathrm{tr}\left(R\Pi_Z\right)$ qui correspond à **la moyenne arithmétique** de $R^2(x^j, Z)$. i.e la liaison entre le groupe des variables quantitatives X et Z

Enfin, via la valeur de $\mathrm{tr}\left(R\Pi_Z\right)$, on remarque que la liaison du groupe de variables avec la variable qualitatif Sol est meilleur que celle avec la variable Appellation.