TP 1 - Analyse des données multi-dimensionnelle

Guillaume BOULAND
Camille MOTTIER

Toutes les valeurs présentées dans ce document seront données avec une précision de 10^{-3}.

1 Partie I

1.1 Données quantitatives : étude du nuage de points

Nous étudions ici des données relatives aux vins de Loire, contenues dans le fichier `wine.csv`. Dans ce fichier, sont présentés 21 vins selon le type d'appellations, de sol et de 29 paramètres quantitatifs d'intensités sensorielles.

Label	Soil	Odor.Int.bf.shake	Aroma.qual.bf.shake	Fruity.bf.shake	
2EL	Saumur	Env1	3.074	3.000	2.714
1CHA	Saumur	Env1	2.964	2.821	2.375
1FON	Bourgueil	Env1	2.857	2.929	2.560
1VAU	Chinon	Env2	2.808	2.593	2.417
1DAM	Saumur	Reference	3.607	3.429	3.154

Table 1: Extrait du tableau d'origine

Nous commençons par extraire du tableau les variables quantitatives, que nous centrons et réduisons afin de les ramener à une échelle comparable de valeurs. Nous considérons pour cela que les vins sont associés à un même poids : $\frac{1}{21}$. Nous appellons X ce nouveau tableau et (x_i^j) ses composantes.

Odor.Int.bf.shake	Aroma.qual.bf.shake	Fruity.bf.shake	Flower.bf.shake	Spice.bf.shake
-0.131	-0.228	-0.000	1.591	-0.135
-0.521	-1.120	-1.728	1.591	-1.332
-0.900	-0.582	-0.785	-0.692	0.366
-1.073	-2.256	-1.514	-1.028	0.721

Table 2: Extrait des valeurs centrées réduites

Barycentre du nuage de points :

Le barycentre $\overline{x}$ de ce nuage de points se trouve à l'origine. En effet, les variables étant centrées, nous avons :

$$\overline{x} = \left(\overline{x^j}\right)_{1\leqslant j\leqslant 29} = \left(\sum_{i=1}^{21} \frac{1}{21} x_i^j\right)_{1\leqslant j\leqslant 29} = 0.$$

En notant W la matrice diagonale de $\mathcal{M}_{21}(\mathbb{R})$ des poids de chaque vin et $\mathbb{1}$ la matrice colonne de $\mathcal{M}_{21,1}(\mathbb{R})$ dont toutes les composantes sont égales à 1, nous obtenons aussi $\overline{x}$ par le produit matriciel $\overline{x}' = X'W\mathbb{1}$, ce qui est exploité informatiquement pour vérifier le résultat précédent.

Odor.Int.bf.shake	Aroma.qual.bf.shake	Fruity.bf.shake	Flower.bf.shake
-0.000	0.000	-0.000	0.000

Table 3: Barycentre du nuage

Inertie du nuage de points :

Considérant la métrique canonique I sur $\mathbb{R}^{29}$, nous obtenons par le calcul suivant l'inertie du nuage de points :

$$\begin{aligned} In_0\{x_i, \tfrac{1}{21}\}_{1\leqslant i\leqslant 21} &= \sum_{i=1}^{21} \frac{1}{21}\|x_i\|_I^2 \\ &= \sum_{i=1}^{21}\frac{1}{21}\sum_{j=1}^{29}(x_i^j)^2 \\ &= \sum_{j=1}^{29}\sum_{i=1}^{21}\frac{1}{21}(x_i^j)^2 \\ &= \sum_{j=1}^{29} V(x^j) \\ &= \boxed{29} \qquad \text{car les } x^j \text{ sont centrées et réduites.} \end{aligned}$$

1.2 Partition par appellations

Les vins sont séparés en trois appellations : Bourgueil, Chinon et Saumur.
Nous noterons $Y = (y_i^k) \in \mathcal{M}_{21,3}(\mathbb{R})$ la matrice dont les colonnes sont les indicatrices d'appellations.
Les poids globaux des trois appellations sont obtenus par le produit matriciel $Y'W\mathbb{1}$:

	Poids
Bourgueil	0.286
Chinon	0.190
Saumur	0.524

Table 4: Poids par appellation

Les barycentres au sein des trois appellations sont obtenus par le calcul matriciel suivant :

$$\left(\overline{x}^k\right)_{1\leqslant k\leqslant 3} = (Y'WY)^{-1}(Y'WX) \in \mathcal{M}_{3,29}(\mathbb{R}).$$

En effet, $Y'WY$ est la matrice diagonale des poids des appellations et $Y'WX = \left(\sum_{i/y_i=k} w_i x_i^j\right)_{k,j}$ est la matrice des "pseudo-moyennes" par appellation.

	Odor.Int.bf.shake	Aroma.qual.bf.shake	Fruity.bf.shake	Flower.bf.shake	Spice.bf.shake
Bourgueil	-0.625	0.079	0.041	-0.182	0.037
Chinon	-0.595	-0.835	0.187	-0.287	-0.198
Saumur	0.558	0.260	-0.091	0.204	0.052

Table 5: Barycentres des variables sensorielles par appellation - Extrait

Ces trois barycentres ont pour norme :

	Normes
Bourgueil	3.604
Chinon	5.400
Saumur	2.120

Table 6: Normes des barycentres par appellation

On obtient alors l'inertie inter-appellations :

$$In_0\{\overline{x}^k, W^k\}_{1\leqslant k\leqslant 3} = \sum_{k=1}^{3} W^k\|\overline{x}^k\|_I^2 \simeq \boxed{3,169}$$

puis le coefficient R^2 de la partition des vins en appellations :

$$R^2 = \frac{In_0\{\overline{x}^k, W^k\}_{1\leqslant k\leqslant 3}}{In_0\{x_i, \frac{1}{21}\}_{1\leqslant i\leqslant 21}} \simeq \boxed{0,109}$$

Nous constatons ainsi que la part des appellations dans les disparités sensorielles des vins est d'environ 11%.

1.3 Influence de l'appellation sur les différentes variables sensorielles

Il s'agit ici d'étudier séparément l'influence des appellations sur chaque variable sensorielle. On calcule donc pour chacune d'entre elles le coefficient R^2 :

$$(R^2)^j = \frac{\sum_{k=1}^{3} W^k(\overline{x^j}^k)^2}{V(x^j)} = \sum_{k=1}^{3} W^k(\overline{x^j}^k)^2$$

Odor.Int.bf.shake	Aroma.qual.bf.shake	Fruity.bf.shake	Flower.bf.shake	Spice.bf.shake
0.342	0.170	0.011	0.047	0.009

Table 7: R^2 des variables sensorielles - Extrait

Nous constatons que l'appellation est la moins influente sur les variables intitulées "Spice.before.shaking", "Quality.of.odor", "Fruity" et "Flower", avec une influence de moins de 1%, tandis qu'elle est la plus influente sur les variables intitulées "Odor.intensity.before.shaking", "Odor.intensity" et "Phenolic", avec une influence de plus de 30%.

Spice.before.shaking	Quality.of.odour	Fruity	Flower
0.009	0.008	0.007	0.009

Table 8: Variables les moins liées à l'appellation

Odor.Int.before.shaking	Odor.Intensity	Phenolic
0.342	0.391	0.360

Table 9: Variables les plus liées à l'appellation

Remarquons que le R^2 de la partition est égal à la moyenne arithmétique des R^2 des variables sensorielles :

$$\begin{aligned} R^2 &= \frac{In_0\{\overline{x}^k, W^k\}_{1\leqslant k\leqslant 3}}{In_0\{x_i, \frac{1}{21}\}_{1\leqslant i\leqslant 21}} \\ &= \frac{1}{29}\sum_{k=1}^{3} W^k \sum_{j=1}^{29}(\overline{x^j}^k)^2 \\ &= \frac{1}{29}\sum_{j=1}^{29}\sum_{k=1}^{3} W^k(\overline{x^j}^k)^2 \\ &= \frac{1}{29}\sum_{j=1}^{29}(R^2)^j \end{aligned}$$

moy_R2var	R2
0.109	0.109

Table 10: Vérification de l'égalité entre le R^2 et la moyenne arithmétique du R^2 des variables

2 Partie II

Dans cette partie, nous étudions l'influence des appellations et de la nature du sol sur les différentes variables sensorielles. Pour ce faire, nous utilisons les projections orthogonales sur les espaces engendrés par les modalités. Nous notons ici Y (respectivement Z) la matrice dont les colonnes sont les indicatrices d'appellations (respectivement de sol).

2.1 Influence de l'appellation

1. $\sum_{k=1}^{3} y^k = \mathbb{1}$, donc $\mathbb{1} \in < Y >$. De plus, $< Y^c > \subset \mathbb{1}^\perp$. Donc $< Y > = < Y^c > \overset{\perp}{+} < \mathbb{1} >$.
   Or, pour tout $j \in \{1, \dots, 29\}$, x^j est centré donc est orthogonal à $\mathbb{1}$. On a ainsi :

$$\boxed{\Pi_Y x^j = \Pi_{Y^c} x^j + \Pi_{\mathbb{1}} x^j = \Pi_{Y^c} x^j}.$$

   De plus, $\Pi_Y x^j = \left(\overline{x^j}^{Y_i}\right)_{1 \leqslant i \leqslant 21}$ donc la norme de ce vecteur peut être interprêtée ainsi :

$$\|\Pi_Y x^j\|_W^2 = \sum_{k=1}^{3} W^k \left(\overline{x^j}^k\right)^2 = \boxed{(R^2)^j} \quad \text{(car } x^j \text{ est une variable réduite)}$$

2. On considère les matrices de $\mathcal{M}_{21}(\mathbb{R})$ des projections orthogonales sur Y et sur x^j :

$$\Pi_Y = Y(Y'WY)^{-1}Y'W \quad \text{et } \forall j \in \{1, 29\},\ \Pi_{x^j} = x^j \underbrace{(x^{j\prime} W x^j)}_{V(x^j)=1}{}^{-1} x^{j\prime} W = x^j x^{j\prime} W$$

   La trace du produit de ces deux matrices permet d'obtenir à nouveau les $(R^2)^j$ des variables sensorielles :

$$\mathrm{tr}(\Pi_{x^j}\Pi_Y) = \mathrm{tr}(\Pi_Y \Pi_{x^j}) = \mathrm{tr}(\Pi_{Y^c}\Pi_{x^j}) = [\Pi_{Y^c} | \Pi_{x^j}] = (R^2)^j$$

Odor.Int.bf.shake	Aroma.qual.bf.shake	Fruity.bf.shake	Flower.bf.shake	Spice.bf.shake
0.342	0.170	0.011	0.047	0.009

Table 11: R^2 de différentes variables sensorielles - Extrait

3. On introduit la matrice de $\mathcal{M}_{21}(\mathbb{R})$ définie par : $R = XMX'W$. On a alors :

$$\begin{aligned}
\mathrm{tr}(R\Pi_Y) &= \mathrm{tr}(XMX'W\Pi_Y) \\
&= \frac{1}{29}\mathrm{tr}(X'W\Pi_Y X) \\
&= \frac{1}{29}\mathrm{tr}\left(< x^j, \Pi_Y x^k >_W\right)_{j,k} \\
&= \frac{1}{29}\sum_{j=1}^{29} < x^j, \Pi_Y x^j >_W \\
&= \frac{1}{29}\sum_{j=1}^{29} \|\Pi_Y x^j\|_W^2 \qquad \text{(car } \Pi_Y \text{ est auto-adjoint et idempotent)} \\
&= \frac{1}{29}(R^2)^j \\
&= \boxed{R^2}
\end{aligned}$$

   La trace de ce produit matriciel permet de retrouver le coefficient R^2 de la partition par appellations.

traceRProjY	R2
0.109	0.109

Table 12: Vérification de l'égalité $\mathrm{tr}(R\Pi_Y) = R^2$

4. À l'aide des matrices de projections, et en exploitant les relations obtenues ci-dessus, nous retrouvons informatiquement les résultats de la partie I concernant l'influence des appellations sur les caractéristiques sensorielles des vins.

2.2 Influence du sol

En procédant de manière similaire, nous pouvons étudier l'influence du sol sur les caractéristiques sensorielles des vins. Pour ce faire, nous utilisons la projection sur l'espace engendré par les indicatrices de sols Π_Z. Nous obtenons alors les R^2 pour chaque variable sensorielle mais aussi le R^2 global pour les sols :

$$\forall j \in \{1, \dots, 29\},\ (R^2)^j = \mathrm{tr}(\Pi_{x^j} \Pi_Z) \quad \text{et} \quad R^2 = \mathrm{tr}(R\Pi_Z) \simeq \boxed{0,365}$$

Odor.Int.bf.shake	Aroma.qual.bf.shake	Fruity.bf.shake	Flower.bf.shake	Spice.bf.shake
0.534	0.415	0.323	0.347	0.558

Table 13: R^2 de différentes variables sensorielles pour les sols - Extrait

TraceRProjZ
0.365

Table 14: R^2 pour le sol

La nature du sol explique donc environ 37% des disparités sensorielles entre les vins de Loire donc est un élément fortement plus explicatif que l'appellation.

3 Annexe

```
wine <- read.csv("wine.csv")

X <- wine[4:32]
W <- diag(1/21 , nr=21)
M <- diag(1/29, 29)
Unit <- matrix(data=c(1), nrow=21, ncol=1)

Xc<- sqrt(21/20)*scale(X)
X <- Xc

#Matrice des indicatrices des appellations
Y <- matrix(0, nrow=21, ncol=3)
for (i in 1:21) {
  if (wine[i,2]=="Bourgueuil") {Y[i,1]<-1};
  if (wine[i, 2]=="Chinon") {Y[i,2]<-1};
  if (wine[i, 2]=="Saumur") {Y[i,3]<-1}
}

#Matrice des indicatrices des sols
Z <- matrix(0, nrow=21, ncol=4)
for (i in 1:21) {
  if (wine[i,3]=="Env1") {Z[i,1]<-1};
  if (wine[i, 3]=="Env2") {Z[i,2]<-1};
  if (wine[i, 3]=="Reference") {Z[i,3]<-1};
  if (wine[i, 3]=="Env4") {Z[i, 4]<-1}
}

#Barycentres des variables sensorielles
Bar <- t(X) %*% W %*% Unit

#Inertie totale du nuage
In = sum(apply(1/21*X^2 ,1, sum))
```

```r
#Poids et barycentre par appellation
Poids_app <- t(Y)%*%W%*%Unit

Moy_cat= solve(t(Y)%*%W%*%Y)%*%t(Y)%*%W%*%X

#Calcul normes euclidiennes
Normes<- apply(Moy_cat^2, 1, sum)

#Calcul de l'inertie inter-appellations
In_inter <- Normes%*%Poids_app

#Calcul du R^2
R2 <- In_inter/In

#Variance inter-appellation
var_inter <- t(Poids_app)%*%Moy_cat^2
R2var<- var_inter

#Variables les moins liées à l'appellation
subset(R2var, select = R2var<0.01)

#Variables les plus liées à l'appellation
subset(R2var, select = R2var>0.33)

#Calcul de la moy arithmétique du R2 des variables
moy_R2var<-mean(R2var)

#Vérification de l'égalité entre le R2 et la moy arithmétique du R2 des variables
delta<- moy_R2var - R2


#PARTIE 2

#Projection sur Y
projy <- Y %*% solve(t(Y) %*%W%*% Y) %*% t(Y)%*%W

#fonction de la projection sur les différrentes variables
projxj<-function(j) {
  proj <- X[,j] %*% solve(t(X[,j]) %*%W%*% X[,j]) %*% t(X[,j])%*%W
  return(proj)}

#Matrice de toutes les traces (pour tous les j) du produit Proj_xj*Proj_Y
V2 <- matrix( 0, nrow=1, ncol=29)
for (j in 1:29) {
  V2[,j]<- sum(diag(projxj(j)%*%projy))
}

#Trace du produit R*proj_Y
R = X%*%M%*%t(X)%*%W
R_Y = R%*%projy
traceRY <-sum(diag(R_Y))

#Trace de proj_xj*proj_Z
projz <- Z %*% solve(t(Z) %*%W%*% Z) %*% t(Z)%*%W
V3 <- matrix(0, nrow=1, ncol=29)
for (j in 1:29) {
  V3[,j]<- sum(diag(projxj(j)%*%projz))
}

#Trace de R*projz
traceRprojz <- sum(diag(R%*%projz))
```