TP1-ADM

Guillaume Bernard-Reymond et Lorenzo Gaggini

October 2023

1 Partie 1

Dans cette partie, nous désignerons par (x_i) où $i \in \{1;...;21\}$ les individus et (x^j) où $j \in \{1;...;32\}$ les variables, qu'elles soient quantitatives ou bien qualitatives. Enfin nous noterons (z^j) où $j \in \{4;...;32\}$ les variables quantitatives centrées-réduites.

1. Dans cette question nous placerons dans le cas général où $i \in \{1;...;n\}$ et $j \in \{1;...;p\}$

   - On considère $(w_i)_{i\in\{1...n\}}$ une suite de poids telle que $\sum_{i=1}^{n} = 1$ et soit $j \in \{1;...;p\}$. On peut alors écrire :

   $$\begin{aligned}\sum_{i=1}^{n} w_i z_i^j &= \sum_{i=1}^{n} w_i \left(\frac{x_i^j - \overline{x^j}}{\sigma_{x^j}}\right)\\ &= \frac{1}{\sigma_{x^j}} \sum_{i=1}^{n} \left(w_i x_i^j - w_i \overline{x^j}\right)\\ &= \frac{1}{\sigma_{x^j}} \left(\overline{x^j} - \overline{x^j} \sum_{i=1}^{n} w_i\right)\\ &= 0\end{aligned}$$

   Le barycentre du nuage est donc bien $0_{\mathbb{R}^p}$.
   Informatiquement, nous avons utiliser les commandes suivantes :

   ```
   Barycentre=(colMeans(CR[4:32]))
   print(Barycentre)
   ```

   Les résultats obtenus étant de l'ordre de 10^{-16}, on peut considérer qu'ils sont égaux à 0.

   - $$\begin{aligned}In_O\left(\{z_i; w_i\}_{i=1,...,n}\right) &= \sum_{i=1}^{n} w_i \left\|z_i\right\|^2\\ &= \sum_{i=1}^{n} w_i \left(\sum_{j=1}^{p} {z_i^j}^2\right)\\ &= \sum_{i=1}^{n} \left(\sum_{j=1}^{p} w_i {z_i^j}^2\right)\\ &= \sum_{j=1}^{p} \left(\sum_{i=1}^{n} w_i {z_i^j}^2\right)\end{aligned}$$

   Or l'expression $\sum_{i=1}^{n} w_i {z_i^j}^2$ n'est rien d'autre que l'expression de la variance de notre variable quantitative centrée réduite qui vaut donc 1.
   Ainsi : $In_O\left(\{z_i; w_i\}_{i=1,...,n}\right) = p$ c'est à dire le nombre de variables quantitatives.
   Informatiquement, nous avons utilisé les commandes :

```
#variance
Variance=diag(var(CR[4:32]))
print(Variance)

#inertia
Inertie=sum(Variance)
print(Inertie)
```

Le résultat obtenu est 29 comme attendu.

2. Pour chaque appellation, nous avons effectuer un tri afin d'obtenir pour chaque appellation un tableau des individus. A partir de là, nous avons calculé le poids, les barycentres et les normes euclidiennes carrées de ces trois barycentres. Voici le code utilisé où CR désigne le tableau dont les variables quantitatives sont centrées et réduites :

   ```
   "Q2"
   chinon=CR[CR$Label == 'Chinon',]
   saumur=CR[CR$Label == 'Saumur',]
   bourgueuil=CR[CR$Label == 'Bourgueuil',]

   #poids
   pchi=nrow(chinon)/nrow(CR)
   psau=nrow(saumur)/nrow(CR)
   pbou=nrow(bourgueuil)/nrow(CR)

   #barycentre
   mchi= unname(colMeans(chinon[4:32]))
   msau= unname(colMeans(saumur[4:32]))
   mbou= unname(colMeans(bourgueuil[4:32]))

   #carree des normes euclidiennes des barycentres
   nchi=sum(mchi^2)
   nsau=sum(msau^2)
   nbou=sum(mbou^2)
   ```

   Voici les résultats obtenus :

	Poids	Norme
Chinon	0,1905	5,1428
Saumur	0,5338	2,0188
Bourgueuil	0,2857	3,4319

   Pour chaque variable sensorielle, nous avons calculer le R^2. Nous avons donc obtenu 3 vecteurs lignes de 29 valeurs. Nous avons ensuite triés les valeurs et tracés les graphiques correspondant.

   Pour le Bourgueil :

   - la moins liée : Spice.before.shaking
   - la plus liée : Attack.intensity

   Pour le Chinon :

   - les moins liées : Spice.before.shaking et Astringency
   - les plus liées par ordre croissant : Fruity , Aroma.intensity , Acidity

   Pour le Saumur :

   - la moins liée : Surface.feeling
   - les plus liées par ordre croissant : Flower , Spice , Spice.before.shaking.

   A montrer : le R^2 de la partition est égal à la moyenne arithmétique des R^2 des variables.

2 Partie 2

1. (a) On considère le vecteur $\mathbb{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n$.

   On rappelle que $\mathbb{1}_n \in \text{Vect}\left(y^1; ...; y^p\right) = \langle Y \rangle$ car $\forall i \in \{1; ...; n\}$, $\sum_{j=1}^{p} \mathbb{1}_{Y_i=j} = 1$ donc $\mathbb{1}_n \in \langle Y \rangle$. Soit $j \in \{1; ...; p\}$:

$$\begin{aligned} \Pi_{\langle Y \rangle} x^j =& \Pi_{\langle \mathbb{1}_n \rangle} x^j + \Pi_{\langle \mathbb{1}_n^{\perp} \cap Y \rangle} x^j \text{ car } \langle Y \rangle = \langle \mathbb{1}_n \rangle \overset{\perp}{\oplus} \langle \mathbb{1}_n^{\perp} \cap Y \rangle \\ =& 0 + \Pi_{\langle Y^c \rangle} x^j \\ =& \Pi_{\langle Y^c \rangle} x^j \end{aligned}$$

   $\Pi_{\langle \mathbb{1}_n \rangle} x^j = 0$ car x^j étant centré réduit, il appartient à $\mathbb{1}_n^{\perp}$.

   $\left\| \Pi_Y x^j \right\|_W^2$ représente la variance de la variable x^j par rapport à l'appellation.

   (b) $\Pi_Y = Y\left(Y'WY\right)^{-1} Y'W$ est une matrice carrée de taille 21 et de même pour $\Pi_{x^j} = x^j \left(x^{j\prime} W x^j\right)^{-1} x^{j\prime} W$

   (c) $\text{tr}(R\Pi_Y)$ est le R^2 de la partition des vins en appellations.

   (d)

2. De la même façon, $\text{tr}(\Pi_{x^j}\Pi_Z)$ est ... et $\text{tr}(R\Pi_Z)$ est le R^2 de la partition des vins selon les types de sol.