SY09 Printemps 2019

TP 3
Représentation euclidienne des données

1 Représentation euclidienne

1.1 Représentation des données

Dans cette partie, on s'intéresse à un jeu de données contenant les notes de $n = 9$ individus pour $p = 5$ matières : mathématiques, sciences « naturelles » (physique-chimie), français, latin, « arts » (dessin-musique).

On souhaite étudier ces données et les représenter de manière à caractériser les individus en fonction de leur niveau scolaire. Plus particulièrement, on cherchera à déterminer une *base de représentation* permettant de visualiser « au mieux » les données.

On pourra charger les données au moyen des commandes suivantes :

```
notes <- read.table("donnees/notes.txt", header=T)
```

Le code suivant permet de représenter les individus par les valeurs des variables i et j, et de les identifier en fonction du nom renseigné dans le data.frame :

```
plot(varj~vari, data=notes, pch=20, asp=1)
text(notes[,c("vari","varj")], row.names(notes), pos=1)
```

Si les données sont simplement définies par une matrice et non un data.frame, on pourra utiliser les instructions suivantes :

```
plot(notes[,c(i,j)], pch=20, asp=1)
text(notes[,c(i,j)], row.names(notes), pos=1)
```

Analyse et représentation succinctes

1. Faire une brève analyse des données ; en particulier :
   (a) comment sont réparties les notes, dans chaque matière ?
   (b) Peut-on rapprocher certaines matières les unes des autres ?
2. On cherche à identifier des groupes d'élèves en fonction des résultats scolaires.
   (a) Représenter les élèves en fonction de leurs résultats dans les matières scientifiques ; quels sont les groupes d'individus qui semblent se détacher ?
   (b) Faire de même avec les matières littéraires, puis avec les arts.
3. Représenter les élèves par deux informations : leur moyenne dans les matières scientifiques, et leur moyenne dans les matières littéraires. Interpréter les résultats obtenus.

Projection et qualité de représentation

1. On considère la matrice suivante :

$$A_1 = \begin{pmatrix} 1/2 & 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & -1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & -1/2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

   (a) Définit-elle une base ?

   (b) Comment exprimer les notes du tableau X dans la nouvelle base A_1, et comment revenir dans la base canonique de $\mathbb{R}^5$?

   (c) Comment peut-on interpréter les coordonnées des individus exprimées dans cette nouvelle base ? Que permet de visualiser la représentation selon les composantes (X^1, X^2), ou (X^1, X^3), ou encore (X^2, X^4) ?

2. On considère à présent la matrice B_1 suivante :

$$B_1 = \begin{pmatrix} \sqrt{2}/2 & 0 & \sqrt{2}/2 & 0 & 0 \\ \sqrt{2}/2 & 0 & -\sqrt{2}/2 & 0 & 0 \\ 0 & \sqrt{2}/2 & 0 & \sqrt{2}/2 & 0 \\ 0 & \sqrt{2}/2 & 0 & -\sqrt{2}/2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

   (a) Définit-elle une base ?

   (b) Comment peut-on interpréter les coordonnées des individus exprimés dans cette nouvelle base ? Que permet de visualiser la représentation selon les composantes (X^1, X^2), ou (X^1, X^3), ou encore (X^2, X^4) ?

3. On considère à présent la matrice B_2 suivante :

$$B_2 = \begin{pmatrix} \sqrt{2}/2 & 0 & \sqrt{2}/2 & 0 & 0 \\ 0 & \sqrt{2}/2 & 0 & \sqrt{2}/2 & 0 \\ \sqrt{2}/2 & 0 & -\sqrt{2}/2 & 0 & 0 \\ 0 & \sqrt{2}/2 & 0 & -\sqrt{2}/2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

   (a) Définit-elle une base ?

   (b) Comment peut-on interpréter les coordonnées des individus exprimés dans cette nouvelle base ? Que permet de visualiser la représentation selon les composantes (X^1, X^2), ou (X^1, X^3), ou encore (X^2, X^4) ?

1.2 Choix d'une représentation

On définit la qualité de la représentation selon un axe comme étant la quantité d'inertie expliquée par cet axe.

1. Quelle est la quantité d'inertie du nuage de points expliquée par chacun des axes, si l'on considère la base canonique de $\mathbb{R}^5$? Quelle est la quantité d'inertie totale du nuage de points ?
2. Quelle est la quantité d'inertie expliquée par les axes si l'on considère la base A_1 ?
3. Qu'en est-il pour la base B_1, pour la base B_2 ?
4. On cherche à représenter le nuage de points dans un plan, au prix d'une perte d'information. Quels axes choisirait-on, parmi ceux définis par la base canonique, par B_1, ou par B_2 ? Pourquoi ? Interpréter.

2 Questions théoriques

On considère un nuage de points de coordonnées X exprimées dans un espace euclidien. On suppose que les individus ont tous le même poids.

1. Montrer que l'inertie I_X d'un nuage de points de coordonnées X est égale à la trace de sa matrice de covariance empirique (non corrigée) Σ_X.
2. Montrer l'invariance de la trace par permutation circulaire : $\mathrm{trace}(ABC) = \mathrm{trace}(CAB)$.
3. Montrer que le centrage est préservé par la projection, c'est-à-dire que la projection $X_c B$ d'un nuage de points X_c centré sur une base B est centrée.
4. Montrer que l'inertie totale des points représentés dans la base B est constante, quelle que soit la base orthonormée B considérée.