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 ?

   Il est possible de faire des histogrammes ou des diagrammes en bâtons à partir des tableaux de contingence. Pour faire court, pas de notes très basses (rien sous 5) ni très hautes (à l'exception d'un 18 en dessin-musique), avec des notes bonnes, moyennes et mauvaises dans chacune des matières.

   (b) Peut-on rapprocher certaines matières les unes des autres ?

   Un bref coup d'œil sur la matrice de corrélations indique que les matières scientifiques (maths-sciences) sont très corrélées, de même que les matières littéraires (français-latin) ; maths et sciences semblent assez corrélées (bien que moins) au latin ; les « arts » sont quant à eux décorrélés des autres matières.

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 ?

   Il suffit de sélectionner les deux variables correspondant aux matières scientifiques et de représenter les individus correspondants.

   (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.

```
plot((notes$math+notes$scie)/2,(notes$fran+notes$lati)/2,
    pch=20, xlim=c(5,15), ylim=c(5,15), asp=1)
text((notes$math+notes$scie)/2,(notes$fran+notes$lati)/2,
    row.names(notes), pos=1)
```

Grosso modo, 1er quadrant : bons en sciences et en lettres, 2e quadrant : bons en lettres mais pas en sciences, 3e quadrant : mauvais en sciences et en lettres, 4e quadrant : bons en sciences mais pas en lettres.

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 ?

   Elle définit une base orthogonale (mais les vecteurs la composant ne sont pas normés).

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

   En notant X_{A_1} les coordonnées dans la base A_1 :

$$X_{A_1} = X \left(A_1^{-1}\right)^T, \text{ et } X = X_{A_1} A_1^T.$$

   (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) ?

   La 1re composante représente la moyenne dans les matières scientifiques, la 2e la moyenne dans les matières littéraires. L'affichage dans le plan (X^1, X^2) permettra donc de distinguer les élèves selon leur niveau dans les matières scientifiques ou littéraires.

   La 3e représente la différence entre le niveau de maths et le niveau de sciences naturelles — couplée avec la 1re, elle permettra de distinguer, parmi les élèves moyens en sciences, ceux moyens partout de ceux bons dans une matière et mauvais dans l'autre. La 4e composante peut être interprétée comme la 3e (mais pour les matières littéraires : il conviendra donc de l'utiliser conjointement avec la seconde).

   La 5e composante, enfin, représente les élèves bons en arts, comme dans la base initiale (et ne communique aucune information concernant le niveau dans d'autres matières).

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 ?

La matrice B_1 correspond évidemment à la matrice A_1 après normalisation des vecteurs la composant. Elle définit donc une base orthonormée.

(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) ?

On peut interpréter les projections des individus de la même manière que dans la base A_1, si ce n'est que les vecteurs définissant B_1 étant normés, l'éloignement des points sera différent. Par exemple, dans le plan (X^1, X^5), le même élève sera désormais plus éloigné de l'origine qu'auparavant.

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 ?

La matrice B_2 définit une base orthonormée (on vérifie facilement que $u_i^T u_j = 0$ pour tout $i \neq j$ et que $u_i^T u_i = 1$ pour tout i).

(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) ?

1re composante : moyenne de maths et de français, 2e : moyenne de sciences naturelles et latin, 3e : différence entre maths et français, 4e : différence entre sciences naturelles et latin, 5e : arts.

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 ?

```
diag(cov.wt(Xc, method='ML')$cov)
```

La quantité totale d'inertie expliquée par les axes est égale à 48.97531.

2. Quelle est la quantité d'inertie expliquée par les axes si l'on considère la base A_1 ?

```
diag(cov.wt(Xc%*%A1, method='ML')$cov)
```

3. Qu'en est-il pour la base B_1, pour la base B_2 ?

```
diag(cov.wt(Xc%*%B1, method='ML')$cov)
diag(cov.wt(Xc%*%B2, method='ML')$cov)
```

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.

On choisirait les axes définis par les deux premiers vecteurs de B1 : ce sont ceux qui expliquent la plus grande partie de l'inertie du jeu de données (ils expliquent respectivement 41.01% et 39.37% de l'inertie totale du nuage de points ; en comparaison, avec deux axes de B2, on ne pourrait expliquer au maximum que 29.37% et 28.40% d'inertie, ou 24.63% et

23.25% avec deux axes du repère initial).

On notera qu'avec cette stratégie (représentation dans un plan choisi en fonction de l'inertie), on perd complètement l'information portée par le 5$^{\text{e}}$ axe (c'est-à-dire le résultat en arts). Il convient donc de garder à l'esprit que le critère de « qualité » utilisé ici est quantitatif. Pour donner davantage de poids à cette matière, il aurait été possible d'utiliser une métrique donnant une pondération plus importante à la 5$^{\text{e}}$ variable.

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.

   L'inertie est la moyenne des distances pondérées des points à leur centre de gravité μ :

   $$I_X = \frac{1}{n}\sum_{i=1}^{n} d^2(\boldsymbol{x}_i, \mu) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij}-\mu_j)^2 = \sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\mu_j)^2 = \sum_{j=1}^{p} s_j^2 = \text{trace}(\Sigma_X).$$

2. Montrer l'invariance de la trace par permutation circulaire : $\text{trace}(ABC) = \text{trace}(CAB)$.

   Soit $m_{uv} = [M]_{uv}$ le terme de la matrice M situé sur la u^{e} ligne et la v^{e} colonne ; on a :

   $$[ABC]_{i\ell} = \sum_{j,k} a_{ij}b_{jk}c_{k\ell}, \text{ d'où } \text{trace}(ABC) = \sum_{i,j,k} a_{ij}b_{jk}c_{ki} \text{ et } \text{trace}(CAB) = \sum_{i,j,k} c_{ij}a_{jk}b_{ki}.$$

   Un changement d'indices permet de montrer que $\text{trace}(ABC) = \text{trace}(CAB)$.

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.

   Le centrage de X s'écrit matriciellement :

   $$X_c = Q_n\,X,$$

   où $Q_n = I_n - \frac{1}{n}U_n$, U_n étant la matrice $n \times n$ remplie de 1. On a donc

   $$X_c\,B = (Q_n\,X)\,B = Q_n\,(X\,B).$$

4. Montrer que l'inertie totale des points représentés dans la base B est constante, quelle que soit la base orthonormée B considérée.

   Soit Σ_{XB} la matrice de covariance empirique des points représentés dans la base B ; l'inertie du nuage X représenté dans la base B s'écrit :

   $$\begin{aligned} I_{XB} &= \text{trace}(\Sigma_{XB}) = \text{trace}\left(\frac{1}{n}\,B^T\,(Q_n\,X)^T\,(Q_n\,X)\,B\right), \\ &= \text{trace}(B^T\,\Sigma_X\,B) = \text{trace}(B\,B^T\,\Sigma_X) = \text{trace}(\Sigma_X) = I_X, \end{aligned}$$

   par invariance de la trace par permutation circulaire puis orthogonalité de B.