TP3 : Analyse en composantes principales

Durant cette séance de travaux pratiques, nous allons faire des analyses en composantes principales sur plusieurs jeux de données. Afin de faciliter la récupération de ces données et les représentations graphiques, un ensemble de fonctions a été défini et vous pouvez le charger grâce à la commande suivante,

```
source("http://www.math.univ-toulouse.fr/~xgendre/ens/l3sid/tpACP.R")
```

1 Notes

Pour introduire les outils de base, nous commençons par un jeu de données simple relatif aux notes obtenues dans $p = 4$ matières par $n = 9$ étudiants uniformément pondérés.

```
notes <- DataNotes()
```

Comme nous l'avons vu en cours, la première étape consiste à centrer les données pour construire la matrice X des données centrées.

- Que fait la fonction `colMeans` ? Utilisez-la pour obtenir le centre de gravité $g \in \mathbb{R}^4$.
- Calculez la matrice `X` des données centrées. Pour cela, vous pouvez écrire vos propres commandes ou vous renseigner sur la fonction `scale`.

Avant de passer à l'analyse en composantes principales, prenons le temps de comprendre un peu mieux ces données.

- Affichez la matrice `X`. Que vous donne la commande `t(X)` ?
- Utilisez `summary(X)` pour obtenir quelques informations sur le jeu de données.
- Affichez les boîtes à moustaches associées aux quatre variables avec la commande `boxplot`. Que pouvez-vous dire de la répartition des observations relatives à ces différentes variables ?

Nous rappelons que si W est la matrice $p \times p$ diagonale des poids alors la matrice de covariance vaut $\Sigma = {}^tXWX$. En langage R, la multiplication d'une matrice `A` par une matrice `B` s'obtient par `A %*% B`.

- Calculez la matrice de covariance de notre jeu de données.
- Que fait la fonction `eigen` ? Que contient la variable `V` après la commande suivante ?

```
V <- eigen(Sigma)
```

  Précisez en particulier ce que contiennent `V$values` et `V$vectors`.

- A partir des questions précédentes, déduisez la matrice `C` des composantes principales ainsi que la part d'inertie expliquée par le plan principal (commentez cette valeur).

Nous pouvons maintenant représenter les données dans le plan principal. Soit $i \in \{1, \dots, n\}$, pour rendre compte de la qualité de la représentation du $i^{\text{ème}}$ individu, nous utiliserons un symbole dont la taille sera proportionnelle au $\cos^2 \theta_i$ introduit en cours.

- Comment obtenir facilement une matrice de n lignes et 2 colonnes contenant les coordonnées des points dans le plan principal ? Affichez ces points avec la commande `plot`.
- Calculez un vecteur `cos2` de taille n contenant les mesures de qualité des représentations. Commentez les valeurs obtenues.
- Comment utiliser la paramètre `cex` de la commande `plot` pour afficher les points avec des tailles proportionnelles aux $\cos^2 \theta_i$? Vous pouvez utiliser le paramètre `pch` pour changer de symbole dans la commande `plot`.

Enfin, il nous faut interpréter les axes du plan principal. Pour cela, nous devons calculer les corrélations linéaires entre les variables initiales $x^1, \dots, x^p$ et les composantes principales $c^1, \dots, c^p$. Nous rappelons que, pour tout $j, j' \in \{1, \dots, p\}$, nous avons

$$\rho\left(x^j, c^{j'}\right) = \frac{\sqrt{\lambda_{j'}}}{\sqrt{\text{Var}(x^j)}} \times P_{jj'}$$

où les λ_j sont les valeurs propres de Σ et P est la matrice de changement de base.

- Construisez une matrice `rho` à p lignes et 2 colonnes telle que la $j^{\text{ème}}$ ligne contienne les coordonnées de la variable x^j dans le cercle des corrélations,

$$\big(\rho(x^j, c^1), \rho(x^j, c^2)\big) = \left(\frac{\sqrt{\lambda_1} P_{j1}}{\sqrt{\text{Var}(x^j)}}, \frac{\sqrt{\lambda_2} P_{j2}}{\sqrt{\text{Var}(x^j)}}\right) .$$

- La commande suivante vous permet d'afficher le cercle des corrélations,

```
rownames(rho) <- colnames(X) # Pour nommer correctement les variables
CercleCor(rho)
```

- Quelles sont les variables qui sont bien représentées ? Procédez à l'interprétation des axes par rapport aux variables initiales.
- Commentez la répartition des points dans le plan principal. Quels groupes d'étudiants pouvez-vous suggérer suite à la question précédente ?

2 Arcs et arbalètes

Le deuxième jeu de données que nous allons considérer est relatif aux arcs et aux arbalètes du jeu vidéo "The Elder Scrolls V : Skyrim". Il contient les statistiques de $n = 14$ armes relatives à $p = 4$ variables et peut être obtenu par la commande suivante,

```
arcs <-DataSkyrim()
```

Comme dans la section précédente, nous commençons par centrer les données et par visualiser leurs répartitions.

- Construisez la matrice `X` des données centrées à partir de `arcs`.

- Affichez les boîtes à moustaches des quatre variables. Commentez ce graphique. Quelle modification des données proposeriez-vous avant de commencer l'analyse en composantes principales ?
- Modifiez `X` pour qu'elle contienne dorénavant les données centrées et réduites. Pour cela, vous pouvez écrire vos propres commandes ou utiliser la fonction `scale`.
- Affichez les boîtes à moustaches des données centrées réduites et commentez ce nouveau graphique par rapport au précédent.

A partir d'ici, les étapes sont les mêmes que dans la section précédente. Construisez la matrice de covariance associée à `X`, représentez les points dans le plan principal et interprétez les axes grâce au cercle des corrélations. Quelles conclusions pouvez-vous tirer à propos des arcs et des arbalètes de ce jeu de données ?

3 Autres outils

Il existe des fonctions R permettant de réaliser directement l'analyse en composantes principales sans faire apparaître la matrice à diagonaliser. Il faut cependant mettre en garde l'utilisateur de ces commandes "boîtes noires" : la normalisation des axes n'est pas la même que celle choisie en cours.

Un exemple de telle fonction est `prcomp`. Lisez l'aide relative à cette fonction pour en comprendre le fonctionnement et les paramètres que vous pouvez utiliser. Voici un exemple sur les données de la première section avec la fonction d'affichage du diagramme biplot,

```
notes <- DataNotes()
ACP <- prcomp(notes)
biplot(ACP)
```

Comment utiliser cette fonction pour réaliser l'analyse en composantes principales de la deuxième section ? Affichez le diagramme biplot associé.