Université de Bordeaux
M2 Master MAS-MSS - UE : Dataminig
Année 2016-2017

TP1 : Rappels sur l'ACP

Le but de ce TP est de proposer quelques rappels sur la notion d'Analyse en Composantes Principales (ACP) pour l'analyse de données quantitatives. Certaines analyses de données proposées dans ce TP sont inspirées des tutoriels du site WikiStat : `http://wikistat.fr`.

1 Quelques rappels sur l'ACP

L'analyse en composantes principales est une technique largement utilisée en statistique qui vise à représenter des individus observés dans un espace de grande dimension. Le but de l'ACP est de trouver un espace de plus petite dimension pour représenter ces individus tout en gardant au mieux l'information contenue dans les données initiales.

Observations : on dispose d'un ensemble de n individus qui sont décrits par p **variables quantitatives**. On peut modéliser (de façon équivalente) ces données sous la forme
- d'une suite de vecteurs $\mathbf{x}_i, i = 1, \dots, n$ de $\mathbb{R}^p$,
- d'une suite de vecteurs $\mathbf{x}^k, k = 1, \dots, p$ de $\mathbb{R}^n$.

Cet ensemble de données peut alors se représenter sous la forme d'un matrice $\mathbf{X}$ de taille $n \times p$ dont les éléments sont les $\mathbf{x}_i^k$ c'est à dire

$$\mathbf{X} = [\mathbf{x}^1 \cdots \mathbf{x}^p].$$

Espace des individus / variables : on utilisera les conventions suivantes
- à chaque individu i est associé le vecteur $\mathbf{x}_i$ de $\mathbb{R}^p$ qui est donc appelé l'espace des individus. L'espace $\mathbb{R}^p$ est muni d'une matrice diagonale M (à entrées positives), et de la métrique et produit scalaire associés

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle_M = \mathbf{x}_i^t M \mathbf{x}_j \quad \text{et} \quad \|\mathbf{x}_i\|_M^2 = \mathbf{x}_i^t M \mathbf{x}_i.$$

- à chaque variable k est associé le vecteur $\mathbf{x}^k$ de $\mathbb{R}^n$ qui est donc appelé l'espace des variables. L'espace $\mathbb{R}^n$ est muni d'une matrice diagonale $D = diag(\omega_1, \dots, \omega_n)$, à entrées positives, et de la métrique et produit scalaire associés

$$\langle \mathbf{x}^k, \mathbf{x}^\ell \rangle_D = \sum_{i=1}^n \omega_i \mathbf{x}_i^k \mathbf{x}_i^\ell \quad \text{et} \quad \|\mathbf{x}^k\|_D^2 = \sum_{i=1}^n \omega_i (\mathbf{x}_i^k)^2.$$

Barycentre des individus : le barycentre (ou moyenne des individus) et l'élément $\bar{\mathbf{x}} \in \mathbb{R}^p$ donné par

$$\bar{\mathbf{x}} = \mathbf{X}^t D \mathbb{1}_n,$$

où $\mathbb{1}_n$ est le vecteur de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1.

La **matrice de covariance** des variables / données est définie par

$$\mathbf{S} = \sum_{i=1}^{n} \omega_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^t = (\mathbf{X} - \mathbb{1}_n \bar{\mathbf{x}}^t)^t D (\mathbf{X} - \mathbb{1}_n \bar{\mathbf{x}}^t)$$

On notera également $\bar{X} = \mathbf{X} - \mathbb{1}_n \bar{\mathbf{x}}^t$ la matrice des données centrées.

Données centrées : afin de simplifier la présentation de l'ACP, il est parfois utile de supposer que les données ont été centrées de sorte que $\bar{\mathbf{x}} = 0$. On a alors l'expression simplifiée suivante pour la matrice de covariance

$$\mathbf{S} = \mathbf{X}^t D \mathbf{X}.$$

Quelques commentaires :

- afin de pouvoir définir une notion de distance entre deux individus, il est en général nécessaire d'introduire une métrique au travers de M matrice diagonale de taille $p \times p$. En effet, les variables observées peuvent correspondre à des unités de mesure complètement différentes (kilogrammes, mètres, temps), et la matrice M permet de quantifier la proximité entre deux individus en prenant en compte l'hétérogénéité des variables. Si les unités de mesure sont homogènes on peut choisir $M = \mathbf{I}$, sinon il est préférable de prendre $M = diag(\sigma_1^{-2}, \ldots, \sigma_p^{-2})$ où

  $$\sigma_k^2 = \|\bar{\mathbf{x}}^k\|_D^2 = \sum_{i=1}^{n} \omega_i (\mathbf{x}_i^k - \bar{\mathbf{x}}^k)^2$$

  est la variance empirique de la variable k. La distance (au carré) entre deux individus i et j s'exprime alors comme

  $$\|\mathbf{x}_i - \mathbf{x}_j\|_M^2 = (\mathbf{x}_i - \mathbf{x}_j)^t M (\mathbf{x}_i - \mathbf{x}_j).$$

- l'espace des variables est muni d'une métrique $\mathbf{D} = diag(\omega_1, \ldots, \omega_n)$ qui est une matrice diagonale dont les éléments caractérisent le poids de chaque individu et qui sont tels que $\sum_{i=1}^{n} \omega_i = 1$ et $w_i \geq 0$ pour tout $1 \leq i \leq n$. Dans le cas où tous les individus ont le même poids, on prend $\omega_i = 1/n$.

Composantes principales et représentation des individus/variables : L'inertie de la matrice des données $\mathbf{X}$ est définie comme :

$$\begin{aligned} I(\mathbf{X}) &= \sum_{i=1}^{n} \omega_i \|\mathbf{x}_i - \bar{\mathbf{x}}\|_M^2 = \sum_{i=1}^{n} \omega_i (\mathbf{x}_i - \bar{\mathbf{x}})^t M (\mathbf{x}_i - \bar{\mathbf{x}}) \\ &= trace(\bar{X}^t D \bar{X} M) = trace(\mathbf{S}M) \end{aligned}$$

La notion d'inertie généralise la notion de variance au cas multidimensionnel et permet de mesurer la dispersion du nuage des données. On peut également considérer l'inertie comme une mesure de l'information d'un nuage de points. L'ACP consiste à rechercher des nouveaux axes de représentation des individus dans $\mathbb{R}^p$ appelés *axes principaux* tel que l'inertie du nuage projeté sur ces nouveaux axes soit d'inertie maximale. Ces nouveaux axes sont engendrés par des *vecteurs principaux* qui engendrent donc chacun des sous-espace vectoriels de $\mathbb{R}^p$ de dimension 1 appelés *axes principaux*. On montre que la solution

obtenue correspond au calcul des vecteurs propres M-orthonormés, notés $\mathbf{v}_k, k = 1, \ldots, q$, de la matrice $\mathbf{S}M$ associés aux q plus grandes valeurs propres $\lambda_k, k = 1, \ldots, q$ de cette matrice, où q représente le nombre de vecteurs principaux choisis par le statisticien, c'est à dire

$$\mathbf{S}M\mathbf{v}_k = \mathbf{v}_k\lambda_k, \quad \langle \mathbf{v}_k, \mathbf{v}_\ell \rangle_M = \delta_{k,\ell}, \quad 1 \leq k, \ell \leq q.$$

Les coordonnées de l'individu i sur les q premiers vecteurs principaux sont données par :

$$c_i^k = (\mathbf{x}_i - \bar{\mathbf{x}})^t M \mathbf{v}_k, \quad k = 1, \ldots, q.$$

Quelques remarques et éléments de vocabulaire :

- les vecteurs $\mathbf{c}^k = \bar{X}M\mathbf{v}_k$ pour $1 \leq k \leq q$ sont appelés les *composantes principales*,
- chaque vecteur $\mathbf{c}^k$ est un élément de $\mathbb{R}^n$ qui constitue une nouvelle variable. Ces variables sont centrées, non-corrélées et de variance λ_k c'est à dire que

$$\mathbb{1}_n^t \mathbf{c}^k = 0, \quad \|\mathbf{c}^k\|_D^2 = \sum_{i=1}^n \omega_i (c_i^k)^2 = \lambda_k, \quad \langle \mathbf{c}^k, \mathbf{c}^\ell \rangle_D = \sum_{i=1}^n \omega_i c_i^k c_i^\ell = 0,$$

- la matrice $\mathbf{C} = [\mathbf{c}^1 \cdots \mathbf{c}^q] = \bar{X}\mathbf{V}$ où $\mathbf{V} = [\mathbf{V}^1 \cdots \mathbf{V}^q]$ est appelée *matrice des composantes principales*. Elle représentent les nouvelles coordonnées des individus sur les q premiers axes principaux,
- les axes définis par les vecteurs D-orthonormés $\mathbf{u}^k = \frac{1}{\sqrt{\lambda_k}}\mathbf{c}^k$ sont appelés *axes factoriels*.

Mesures de la qualité de représentation des individus/variables : afin de juger la qualité de la représentation des données sur ces q nouvelles variables, on peut utiliser les critères suivants :

— qualité globale (part de dispersion observée) : $\frac{\sum_{k=1}^q \lambda_k}{\sum_{k=1}^p \lambda_k}$,
— qualité de la représentation de l'individu i : $\frac{\sum_{k=1}^q (\mathbf{c}_i^k)^2}{\sum_{k=1}^p (\mathbf{c}_i^k)^2}$.

L'interprétation des nouvelles variables en fonction des variables initiales est donnée par

— projection de la variable initiale $\mathbf{x}^\ell$ sur l'axe factoriel $\mathbf{u}^k$: $\langle \bar{\mathbf{x}}^\ell, \mathbf{u}^k \rangle_D = \sqrt{\lambda_k}\mathbf{v}_k^\ell$,
— qualité de la représentation des variables : $\frac{\sum_{k=1}^q \lambda_k (\mathbf{v}_k^\ell)^2}{\sum_{k=1}^p \lambda_k (\mathbf{v}_k^\ell)^2}$,
— corrélation de la variable initiale $\mathbf{x}^\ell$ et la k-ème composante principale $\mathbf{c}^k$:

$$\frac{\langle \bar{\mathbf{x}}^\ell, \mathbf{c}^k \rangle_D}{\|\bar{\mathbf{x}}^\ell\|_D \|\mathbf{c}^k\|_D} = \frac{\sqrt{\lambda_k}}{\sigma_j}\mathbf{v}_k^\ell.$$

2 Analyse d'un jeu de données simulé

On propose revoir la notion d'ACP au travers de l'analyse d'un jeu de données simulées de petite dimension à l'aide du logiciel **R**. Ce jeu de données provient du site WikiStat : `http://wikistat.fr` et d'un cours sur l'ACP proposé par Philippe Besse (Professeur à l'INSA de Toulouse).

Q. 1 *Utiliser le code ci-dessous pour simuler un jeu de données qui contient les notes de $n = 9$ élèves dans $p = 4$ disciplines.*

```
# Donnnées : notes des élèves
p = 4
n = 9
notes = as.data.frame(matrix(0,n,p))

row.names(notes) = c('jean','alan', 'anni', 'moni',
                             'didi', 'andr', 'pier', 'brig', 'evel')
names(notes) = c('MATH','PHYS','FRAN','ANGL')

notes[,'MATH'] = c(6.00, 8.00, 6.00, 14.50, 14.00, 11.00, 5.50, 13.00, 9.00)
notes[,'PHYS'] = c(6.00, 8.00, 7.00, 14.50, 14.00, 10.00, 7.00, 12.50, 9.50)
notes[,'FRAN'] = c(5.00, 8.00, 11.00, 15.50, 12.00, 5.50, 14.00, 8.50, 12.50)
notes[,'ANGL'] = c(5.50, 8.00, 9.50, 15.00, 12.50, 7.00, 11.50, 9.50, 12.00)
```

Q. 2 *Utiliser les fonctions* `summary` *et* `boxplot` *pour effectuer une analyse statistique univariée de ces données. Que constatez-vous ?*

Q. 3 *Utiliser les fonctions* `cor` *et* `var` *pour calculer la matrice des corrélations et la matrice de covariance des variables de ce jeu de données. Que constatez-vous ?*

Q. 4 *Calculer les vecteurs propres et valeurs propres de la matrice de covariance à l'aide de la fonction* `eigen`*.*

Q. 5 *Utiliser le code ci-dessous pour effectuer l'ACP (non-normée i.e. avec $M = \mathbf{I}$) de ce jeu données "à la main" puis à l'aide de la fonction* `prcomp`*.*

```
# ACP "à la main"
# Calcul des Composantes principales
X = notes
Xbar = apply(X,2,'mean')
Y = X - matrix(1,n,1)%*%Xbar
CY = as.matrix(Y) %*% vecteurs_principaux
print(CY)

# ACP avec R
pca.notes = prcomp(notes, scale = FALSE)

# Composantes principales
C <- pca.notes$x
print(C)
```

Q. 6 *Vérifier que les composantes principales sont centrées, orthogonales et de variances égales aux valeurs propres de la matrice de covariance.*

Q. 7 *Utiliser le code ci-dessous pour représenter les individus sur les deux premiers axes principaux.*

```
dev.new()
plot(C[,1:2],type="n",xlab='PC1',ylab='PC2')
text(C[,1:2],labels=row.names(notes),cex=1.5)
title(main="Projection sur les 2 premiers axes principaux")
```

Q. 8 *Tracer le graphe de l'éboulis des valeurs propres. La qualité de représentation des données sur les deux premiers axes principaux est-elle satisfaisante ?*

Q. 9 *Utiliser le code ci-dessous pour tracer le graphe en biplot. Comment pouvez-vous interpréter les axes principaux en fonction des variables initiales ?*

```
# Graphe en biplot
dev.new()
biplot(pca.notes, xlabs = row.names(notes))
```

Q. 10 *Reprendre les questions précédentes pour faire une ACP normée des données, c'est à dire en prenant* $M = diag(\sigma_1^{-2}, \ldots, \sigma_p^{-2})$. *Que constatez-vous ?*

Q. 11 *Reprendre l'ensemble des questions précédentes en divisant au préalable par 5 les notes en Français et Anglais.*

3 Températures mensuelles en France

Le fichier `tempR.txt` contient les moyennes, entre 1931 et 1960, des températures mensuelles moyennes de 36 villes françaises. La première variable correspond au nom de la ville (4 caractères), les 12 suivantes représentent chacune un mois de l'année (source : Mémorial de la Météorologie nationale et `http://wikistat.fr/`).

Q. 12 *Utiliser le code suivant pour lire ces données et effectuer une analyse statistique univariée des villes.*

```
tempR = read.table('tempR.txt')

# Manipulation du dataframe tempR
tempR[1:3,]
rownames(tempR)
tempR["bres",]

# Représentation de chaque ville
dev.new()
i = 1
plot(1:12,tempR[i,],type="b",main=row.names(tempR)[i],
      xlab='Mois',ylab='Temperature',ylim=c(2,23))
```

```
dev.new()
i = 14
plot(1:12,tempR[i,],type="b",main=row.names(tempR)[i],
      xlab='Mois',ylab='Temperature',ylim=c(2,23))

# Boxplot
dev.new()
boxplot(tempR)

# Variance des variables
diag(var(tempR))
```

Q. 13 *Effectuer une ACP de ce jeu de données à l'aide de la fonction* `prcomp`*. Calculer les coordonnées des données centrées dans la nouvelle base obtenue par ACP, puis proposer une représentation graphique de ces données dans un plan. Que pouvez-vous dire de la qualité de cette représentation ? Comment interpréter les deux premiers axes de l'ACP ?*

Q. 14 *Reprendre la question précédente à l'aide d'une ACP normée. Ceci change-t-il l'interprétation de vos résultats ? Ceci vous semble-t-il cohérent avec la représentation en Boxplot des variables ?*

4 Arrestations aux Etats-Unis

Le fichier de données `USarrests`, accessible en R via la commande `data("USArrests")`, contient des statistiques collectées en 1973 sur les taux d'arrestation pour 100000 habitants pour agression, meurtre ou viol dans chacun des $n = 50$ états des USA. Une quatrième variable indique le pourcentage de résidents dans des zones urbaines pour chaque état.

Q. 15 *Effectuer une analyse descriptive de chaque variables, puis effectuer une ACP non-normée de ce jeu de données à l'aide de la fonction* `prcomp`*. Proposer une représentation graphique de ces données dans un plan. Que pouvez-vous dire de la qualité de cette représentation ? Comment interpréter les deux premiers axes de l'ACP ?*

Q. 16 *Effectuer une analyse de ce jeu de données à partir d'une ACP normée ? Quelles différences constatez-vous ?*

5 Analyse textuelle d'un corpus d'emails

Dans cette partie, il est proposé d'analyser un jeu de données textuelles dans le but de tenter de déterminer les caractéristiques de spams. Une telle analyse est classiquement basée sur la fréquence d'une sélection de mots dans un ensemble d'apprentissage constitué de courriels qui appartiennent à 2 catégories possibles : spam ou non-spam. Les données analysées dans ce TP sont publiques, et elles peuvent servir de "benchmark" pour la comparaison de méthodes d'apprentissage machine :

Frank, A. ; Asuncion, A. (2010). UCI Machine Learning Repository. Irvine, CA : University of California, School of Information and Computer Science.
`http://archive.ics.uci.edu/ml`

Il a été constitué un échantillon de messages électroniques dans chacun desquels a été évalué le nombre d'occurrences d'une sélection de mots et caractères. Les variables considérées sont des ratios qui correspondent au nombre d'occurrences d'un mot spécifique sur le nombre total de mots, ou nombre d'occurrences d'un caractère sur le nombre de caractères du messages. Il a également considéré trois variables prenant en compte la casse (majuscule / minuscule) des caractères et une dernière variable qualitative binaire indiquant le type de chaque message : `spam` ou `Nsp`.

Ces données peuvent être chargées en mémoire dans R à l'aide du code ci-dessous :

```
spam <- read.csv("data_spam.csv",header=FALSE,sep=";")
nom_spam <- read.csv("names_spam.csv",header=FALSE,sep=";")
names(spam) <- sapply((1:nrow(nom_spam)),function(i) toString(nom_spam[i,1]))
spam$y = as.factor(spam$y)
```

Q. 17 *Effectuer une analyse descriptive de chaque variable quantitative, puis effectuer une ACP de ce jeu de données à partir de ces variables. Proposer une représentation graphique de ces données dans un plan. Que pouvez-vous dire de la qualité de cette représentation ? Comment interpréter les deux premiers axes de l'ACP ? Etudier l'influence du choix d'une ACP normée ou non.*

Q. 18 *Pouvez-vous déterminer une représentation graphique et sélectionner des variables qui permettent d'expliquer la différence entre les deux catégories de messages à partir de leur contenu textuel ?*

Dans une deuxième approche, on décide de recoder les variables quantitatives sous la forme de facteurs avec peu de modalités du type "présence / absence de mots" ou "nombre de caractères dans un intervalle donné". On donne ci-dessous un exemple de recodage sous la forme de facteurs des variables liées au mot `make` et au comptage de majuscules.

```
# Variable word_freq_make
make=factor(spam[,"word_freq_make"] > 0, c(TRUE, FALSE),
                       labels=c("make", "Nmk"))
table(make)

# Variable du comptages de majuscules
CapLMq=cut(spam[,"capital_run_length_total"],
breaks=quantile(spam[,"capital_run_length_total"], probs = seq(0, 1, 1/3)),
labels = c("Mm1","Mm2","Mm3"),
      include.lowest = TRUE)
table(CapLMq)

# Table de contingence
table(make, CapLMq)
```

Q. 19 *Exécuter le code ci-dessus et expliquer le recodage effectué.*

Q. 20 *Effectuer une analyse descriptive des deux facteurs* `make` *et* `CapLMq`*, puis effectuer une ACP du jeu de données à partir de ces deux variables uniquement. Que constatez dans la représentation graphique obtenue ? L'ACP appliquée à des variables qualitatives vous semble-t-elle une approche pertinente ?*

6 Générateur aléatoire de visages

Utiliser le code ci-dessous pour récupérer et afficher en R un ensemble de $n = 10$ images du visage d'une même personne pris sous différentes poses.

```
n = 10;
img = list()
names = c('img1.dat','img2.dat','img3.dat','img4.dat','img5.dat',
'img6.dat','img7.dat','img8.dat','img9.dat','img10.dat')

for (i in 1:n) {
aux = t(as.matrix(read.table(names[i],sep=",")))
img[[i]] = aux[,112:1]
}

N1 = dim(img[[1]])[1]
N2 = dim(img[[1]])[2]

for (i in 1:n) {
dev.new()
image(img[[i]], col=gray(0:256/256) , xlab="", ylab="",axes=FALSE)
}
```

Le code ci-dessous permet de calculer une image moyenne et de construire une matrice de données centrées de taille $n \times p$ où p est le nombre de pixels dans les images.

```
# Image moyenne
moy = matrix(0,N1,N2)

for (i in 1:n) {
moy = moy + img[[i]]
}
moy = moy/n

dev.new()
image(moy, col=gray(0:256/256) , xlab="", ylab="",axes=FALSE)

# Matrice de données
X = matrix(0,n,N1*N2)
for (i in 1:n) {
```

```
X[i,] = as.vector(img[[i]]-moy)
}
```

Q. 21 *Utiliser les $q = 2$ premiers vecteurs principaux issues de l'ACP de* `X` *pour proposer un modèle de générateur aléatoire de la pose d'un visage. Visualiser les résultats obtenus sous la forme d'une série d'images aléatoires.*