Chapitre 1

TP ACP

1.1 Analyse en composantes principale

1.2 Package R

Plusieurs fonctions, de différents packages, sont disponibles dans le logiciel R pour le calcul de l'ACP

— prcomp() et princomp() [fonction de base, package stats],
— PCA() [package FactoMineR],
— dudi.pca() [package ade4],
— et epPCA() [package ExPosition]

Peu importe la fonction que vous décidez d'utiliser, vous pouvez facilement extraire et visualiser les résultats de l'ACP en utilisant les fonctions R fournies dans le package factoextra.

Ici, nous utiliserons les deux packages FactoMineR (pour l'analyse) et factoextra (pour la visualisation, des données, basée sur ggplot2). En mode console Installez les deux packages comme suit :
install.packages(c("FactoMineR", "factoextra")) puis chargez-les dans R, en tapant ceci : library("FactoMineR") library("factoextra")

Format des données

Nous utiliserons les jeux de données de démonstration decathlon2 du package factoextra :

data(decathlon2)

Ces données décrivent la performance des athlètes lors de deux évènements sportifs (Decastar et OlympicG). Elles contiennent 27 individus (athlètes) décrits par 13 variables.

Selon la terminologie ACP, nos données contiennent des :

1. Individus actifs (lignes 1 :23) : individus qui sont utilisés lors de l'analyse en composantes principales
2. Individus supplémentaires (lignes 24 :27) : les coordonnées de ces individus seront prédites en utilisant l'information et les paramètres de l'ACP

obtenue avec les individus/variables actifs.

3. Variables actives (colonnes 1 :10) : variables utilisées pour l'ACP.
4. Variables supplémentaires : comme les individus supplémentaires, les coordonnées de ces variables seront également prédites. On distingue des :
   (a) Variables quantitatives supplémentaires : les colonnes 11 et 12 correspondent respectivement au rang et aux points des athlètes.
   (b) Variables qualitatives supplémentaires : Colonne 13 correspondant aux deux rencontres sportives (Jeux olympiques de 2004 ou Décastar 2004). Il s'agit d'une variable catégorielle. Elle peut être utilisée pour colorer les individus par groupes.

Extraire les individus actifs et les variables actives pour l'ACP :

```
decathlon2.active <- decathlon2[1 :23, 1 :10]

head(decathlon2.active[, 1 :6], 4)
```

1.3 Standardisation des données

Dans l'analyse en composantes principales, les variables sont souvent normalisées. Ceci est particulièrement recommandé lorsque les variables sont mesurées dans différentes unités (par exemple : kilogrammes, kilomètres, centimètres, ...) ; sinon, le résultat de l'ACP obtenue sera fortement affecté.

L'objectif est de rendre les variables comparables. Généralement, les variables sont normalisées de manière à ce qu'elles aient au final
— un écart type égal à un
— une moyenne égale à zéro.
A l'issue de cette transformation les données obtenues sont dites données centrées-réduites. L'ACP appliquée à ces données transformées est appelée ACP normée. La standardisation des données est une approche beaucoup utilisée dans le contexte de l'analyse des données d'expression de gènes avant les analyses de type PCA et de clustering.

1.4 code R

Fonction R : PCA() [FactoMineR]. Format simplifié

```
PCA(X, scale.unit = TRUE, ncp = 5, graph = TRUE)
```

1. X : jeu de données de type data frame. Les lignes sont des individus et les colonnes sont des variables numériques
2. scale.unit : une valeur logique. Si TRUE, les données sont standardisées/normalisées avant l'analyse
3. ncp : nombre de dimensions conservées dans les résultats finaux.
4. graph : une valeur logique. Si TRUE un graphique est affiché. Calculer l'ACP sur les individus/variables actifs :

res.pca = PCA(decathlon2.active, graph = FALSE)

Le résultat de la fonction PCA() est une liste, contenant les éléments :

print(res.pca)
L'objet créé avec la fonction PCA() contient de nombreuses informations trouvées dans de nombreuses listes et matrices différentes. Ces valeurs sont décrites dans la section suivante

1.5 Visualisation et interprétation

Les fonctions suivantes, de factoextra, seront utilisées
- $get_eigenvalue(res.pca)$: Extraction des valeurs propres / variances des composantes principales
- $fviz_eig(res.pca)$: Visualisation des valeurs propres
- $get_pca_ind(res.pca)$, $get_pca_var(res.pca)$: Extraction des résultats pour les individus et les variables, respectivement.
- $fviz_pca_ind(res.pca)$, $fviz_pca_var(res.pca)$: visualisez les résultats des individus et des variables, respectivement.
- $fviz_pca_biplot(res.pca)$ Création d'un biplot des individus et des variables. Dans les sections suivantes, nous allons illustrer chacune de ces fonctions.

Valeurs propres / Variances

Que mesure les valeurs propres ?
Nous examinons les valeurs propres pour déterminer le nombre de composantes principales à prendre en considération. Les valeurs propres et la proportion de variances (i.e. information) retenues par les composantes principales peuvent être extraites à l'aide de la fonction $get_eigenvalue()$.
$library("factoextra")$

$eig.val = get_eigenvalue(res.pca)$
$eig.val$
Interpréter cette sortie R ?

Utiliser les valeurs propres pour déterminer le nombre d'axe ?
Une autre méthode pour déterminer le nombre de composantes principales est de regarder le graphique des valeurs propres (appelé scree plot). Le nombre d'axes est déterminé par le point, au-delà duquel les valeurs propres restantes sont toutes relativement petites et de tailles comparables (Jollife 2002, Peres-Neto, Jackson, and Somers (2005)). Le graphique des valeurs propres peut être généré à l'aide de la fonction $fviz_eig()ou fviz_screeplot()$.
$fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 50))$
Interpréter le scree graphe et déterminer le nombre d'axes.

1.6 Graphique des variables

Une méthode simple pour extraire les résultats, pour les variables, à partir de l'ACP est d'utiliser la fonction $get_pca_var()$. Cette fonction retourne une liste

d'éléments contenant tous les résultats pour les variables actives (coordonnées, corrélation entre variables et les axes, cosinus-carré et contributions)

$var = get_pca_var(res.pca)$

Que contient cette sortie var ?

$var\$coord$: coordonnées des variables pour créer un nuage de points.

Consulter les différents éléments.

1.6.1 Cercle de corrélation

La corrélation entre une variable et une composante principale (PC) est utilisée comme coordonnées de la variable sur la composante principale. La représentation des variables diffère de celle des observations : les observations sont représentées par leurs projections, mais les variables sont représentées par leurs corrélations (Abdi and Williams 2010).

$fviz_pca_var(res.pca, col.var = "black")$

Le graphique ci-dessus est également connu sous le nom de graphique de corrélation des variables. Il montre les relations entre toutes les variables.

Interpréter la sortie.

1.6.2 Qualité de représentation

La qualité de représentation des variables sur la carte de l'ACP s'appelle cos2 (cosinus carré) . Vous pouvez accéder au cos2 comme suit :

Vous pouvez visualiser le cos2 des variables sur toutes les dimensions en utilisant le package corrplot : library("corrplot") $corrplot(var\$cos2, is.corr = FALSE)$

Notez que,

- Un cos2 élevé indique une bonne représentation de la variable sur les axes principaux en considération. Dans ce cas, la variable est positionnée à proximité de la circonférence du cercle de corrélation.
- Un faible cos2 indique que la variable n'est pas parfaitement représentée par les axes principaux. Dans ce cas, la variable est proche du centre du cercle.

Pour une variable donnée, la somme des cos2 sur toutes les composantes principales est égale à 1.

Si une variable est parfaitement représentée par seulement deux composantes principales (Dim.1 & Dim.2), la somme des cos2 sur ces deux axes est égale à 1. Dans ce cas, les variables seront positionnées sur le cercle de corrélation.

Pour certaines des variables, plus de 2 axes peuvent être nécessaires pour représenter parfaitement les données. Dans ce cas, les variables sont positionnées à l'intérieur du cercle de corrélation.

Il est possible de colorer les variables en fonction de la valeur de leurs cos2 à l'aide de l'argument $col.var = "cos2"$. Cela produit un gradient de couleurs. Dans ce cas, l'argument gradient.cols peut être utilisé pour spécifier une palette de couleur personnalisée. Par exemple, gradient.cols = c("white", "blue", "red") signifie que :

• les variables à faible valeur de cos2 seront colorées en “white” (blanc) • les variables avec les valeurs moyennes de cos2 seront colorées en “blue” (bleu) • les variables avec des valeurs élevées de cos2 seront colorées en “red” (rouge)

$fviz_pca_var(res.pca, col.var = "cos2", gradient.cols = c("\#00AFBB", "\#E7B800","\#FC4E07"), repel = TRUE)$

Notez qu'il est également possible de modifier la transparence des variables en fonction de leurs valeurs de cos2 à l'aide de l'option $alpha.var = "cos2"$. Par exemple, tapez ceci :

$fviz_pca_var(res.pca, alpha.var = "cos2")$

1.6.3 Contributions des variables aux axes principaux

Les contributions des variables dans la définition d'un axe principal donné, sont exprimées en pourcentage

— Les variables corrélées avec PC1 (i.e., Dim.1) et PC2 (i.e., Dim.2) sont les plus importantes pour expliquer la variabilité dans le jeu de données.
— Les variables qui ne sont pas en corrélation avec un axe ou qui sont corrélées avec les derniers axes sont des variables à faible apport et peuvent être supprimées pour simplifier l'analyse globale.

Il est possible d'utiliser la fonction corrplot() [package corrplot] pour mettre en évidence les variables les plus contributives pour chaque dimension :

$corrplot(var\$contrib, is.corr = FALSE)$

La fonction$fviz_contrib()$ peut être utilisée pour créer un bar plot de la contribution des variables. Si vos données contiennent de nombreuses variables, vous pouvez décider de ne montrer que les principales variables contributives. Le code R ci-dessous montre le top 10 des variables contribuant le plus aux composantes principales :

\# Contributions des variables à PC1 $fviz_contrib(res.pca, choice = "var", axes = 1, top = 10)$ \# Contributions des variables à PC2 $fviz_contrib(res.pca, choice = "var", axes = 2, top = 10)$

La contribution totale à PC1 et PC2 est obtenue avec le code R suivant :

$fviz_contrib(res.pca, choice = "var", axes = 1:2, top = 10)$

La ligne en pointillé rouge, sur le graphique ci-dessus, indique la contribution moyenne attendue.

Comment peut-on définir cette ligne ?