**UT2J 2021-2022**

**L3 MIASHS**

**parcours**

**Mathématiques Appliquées - SHS**

**UE MI0A503T**

# Travaux Pratiques d'Analyses Factorielles (ACP, AFC et AFCM)

**Frédéric Ferraty**

# Table des matières

# Chapitre 1

# Analyse en Composantes Principales (ACP)

## 1.1 Réalisation d'une ACP avec le package R *FactoMineR* : étude des performances de décathloniens

L'objectif de cette section est de réaliser l'ACP du tableau de données "Decathlon" disponible à l'adresse

*http ://www.math.univ-toulouse.fr/∼ferraty/DATA/decathlon.csv*

Ces données fournissent les performances des participants à deux compétitions de décathlon (*decastars* et *jeux olympiques*) :

- 10 variables quantitatives correspondants aux résultats obtenus aux 10 épreuves sportives,
- 1 variable ordinale indiquant le rang final de l'athlète,
- 1 variable quantitative produisant le nombre de points obtenus par l'athlète,
- 1 variable qualitative donnant le nom de la compétition à laquelle les participants ont pris part (*decastars* ou *jeux olympiques*).

### 1.1.1 Ma première ACP avec *FactoMineR*

Pour faire l'ACP de ces données, il faut suivre méthodiquement les différentes étapes suivantes :

1. **Ouvrir une session R**,
2. **Télécharger le fichier des données**

   On peut télécharger le fichier contenant le tableau des données directement à partir de son adresse web :

   ```
   > DECATHLON = read.csv("http://www.math.univ-toulouse.fr/
   ~ferraty/DATA/decathlon.csv", header = TRUE, row.names =
   1, sep = ",", dec = ".", stringsAsFactors = T)
   ```

   *Remarque. La fonctionnalité "Import Dataset" de RStudio (sous menu de "Environnement" dans le panneau de droite en haut) permet de charger un tableau de données en pointant sur une adresse web (voir le sous-menu "From Text (readr)" de "Import Dataset"). Cependant cette fonctionnalité produit automatiquement un objet R de type* `tibble`*. Pour manipuler ce type d'objet, il est nécessaire d'installer au préalable deux packages (s'ils ne sont pas déjà présents dans votre installation de R) : "tidyverse" et "readr". Une fois ces deux packages installés, on peut utiliser la fonctionnalité "From Text (readr)" de "Import Dataset". Pour éviter ces complications (au moins dans un premier temps), on privilégiera la fonction* `read.csv` *comme indiqué précédemment.*
3. **Aperçu des données**

   Le tableau de données en bref avec la commande `summary`

   ```
   > summary(DECATHLON)
   ```

   On peut afficher le tableau des données à l'aide de la commande `View` :

   ```
   > View(DECATHLON)
   ```
4. **Réalisation de l'ACP.** Dans un 1er temps, il faut installer le package "factoMineR" grâce à la commande

   ```
   > install.packages("FactoMineR")
   ```

puis le charger afin que votre session R actuelle puisse y accéder

```
> library("FactoMineR")
```

Vous pouvez maintenant lancer l'ACP en utilisant la fonction "PCA" du package "FactoMineR" :

```
> res.pca = PCA(DECATHLON, quanti.sup = 11:12, quali.sup
= 13)
```

L'option `quanti.sup = 11:12` permet de déclarer comme supplémentaires les variables 11 et 12 (i.e. variables *Rank* et *Points*) ; l'option `quali.sup = 13` déclare la variable qualitative *compétition* comme supplémentaire. Ceci imlique que seules les 10 premières variables (c'est-à-dire les performances aux 10 épreuves) participent à la construction des axes de l'ACP. Grâce à cette commande, on obtient automatiquement le nuage des individus et le cercle des corrélations (i.e. nuage des variables) sur le plan principal (i.e. axes 1 et 2). Les deux variables quantitatives supplémentaires apparaissent en bleu sur le cercle des corrélations alors que les modalités de la variable qualitative supplémentaire sont positionnées au barycentre des individus qui la prennent (voir nuage des individus).

**Remarque 1** *De façon analogue, on peut déclarer des individus supplémentaires à l'aide de l'option* `ind.sup = ...`

On obtient un résumé des résultats de l'ACP à l'aide de la commande

```
> summary(res.pca)
```

Ce résumé fournit :

- les pourcentages d'inertie expliqués par chaque axe ainsi que les pourcentages d'inertie cumulés,
- les résultats pour les 10 premiers individus par défaut (distance de l'individu au centre de gravité du nuage, coordonnée sur l'axe 1, contribution à l'axe 1, qualité de représentation sur l'axe 1 évaluée par le cosinus carré, coordonnée sur l'axe 2, contribution à l'axe 2, qualité de représentation sur l'axe 2, idem pour l'axe 3),
- même chose pour les variables quantitatives actives (excepté que la distance à l'origine n'apparaît plus car cela n'a pas de sens pour les variables),

- même chose pour les variables quantitatives supplémentaires excepté les contributions qui n'ont pas de sens puisque ces variables illustratives ne participent pas à la construction des axes,
- les résultats pour les modalités de la variable qualitative supplémentaire. Pour chaque modalité on a : distance à l'origine, coordonnée sur l'axe 1, qualité de représentation sur l'axe 1 évaluée par le cosinus carré, coordonnée sur l'axe 2, qualité de représentation sur l'axe 2, idem pour l'axe 3. On a en plus une nouvelle colonne *v.test* qui correspond à une valeur test. Si cette valeur test est comprise en -1.96 et +1.96, alors on considère que sa coordonnée n'est pas significativement différente 0. Au contraire, si cette valeur test est en dehors de l'intervalle $[-1.96; +1.96]$, alors on considère que la coordonnée est significativement différente de 0. Ici ce n'est pas le cas pour la coordonnée sur la dimension 1 puisqu'on on a -1.43 pour la modalité *decastar* et 1.43 pour la modalité *jeux olympiques*.

**Remarque 2** *Si on veut obtenir les résultats pour tous les individus (au lieu des 10 premiers par défaut), il suffit de préciser l'option* `nbelements = Inf` *(Inf pour infinity) dans la commande summary :* `summary(res.pca, nbelements = Inf)`*. Cependant, cette commande doit être utilisée avec parcimonie, en particulier dans les situations où l'on dispose d'un très grand nombre d'individus. On peut également sauvegarder les résultats dans un fichier nommé* filename.txt *grâce à la commande* `summary(res.pca, nbelements = Inf, file = "filename.txt")`*. Ce fichier sera physiquement positionné dans le répertoire de travail de R par défaut (il suffit de taper dans R la commande* `getwd()` *pour connaître ce répertoire de travail par défaut).*

**Corrélations significatives entre variables et facteurs**

```
> dimdesc(res.pca)
```

**A propos des inerties.** On peut représenter la décroissance des inerties selon les axes à l'aide de la commande

```
> barplot(res.pca$eig[,2], main = "Décroissance des
inerties (%)", names.arg = 1:nrow(res.pca$eig))
```

Ce graphique est aussi appelé "éboulis des valeurs propres" (voir équivalence

entre inertie et valeurs propres de la matrice des corrélation vue lors de la détermination mathématique des axes factoriels)

### 1.1.2 Interprétation des sorties

- **Axe 1**. Au regard du cercle des corrélations (figure 1.1), l'axe 1 est relativement simple à interpréter. Le graphe des variables nous indique que les individus situés à droite de l'axe 1 ont globalement des scores élevés aux différentes épreuves de lancer et de saut (ces dernières étant corrélées positivment avec l'axe 1) et au contraire des temps faibles aux épreuves de courses (ces dernières étant corrélées négativement avec l'axe 1). Les individus situés à gauche de l'axe 1 ont des comportements opposés. En résumé, l'axe 1 oppose les athlètes bons dans toutes les épreuves (comme Karpov aux jeux Olympiques) aux individus qui sont "mauvais partout" (comme Bourguignon dans le "decastar"). Ces individus extrêmes sont visibles sur le graphe des individus (figure 1.2). On peut d'ailleurs extraire du tableau des données celles correspondant à ces deux individus particuliers et mesurer aisément les différences de scores :

| | 100m | Long jump | Shot put | High jump | 400m | 110m hurdle | Discus | Pole vault | Javeline | 1500m | Rank | Points |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BOURGUIGNON | 11.36 | 6.80 | 13.46 | 1.86 | 51.16 | 15.67 | 40.49 | 5.02 | 54.68 | 291.70 | 13 | 7313 |
| Karpov | 10.50 | 7.81 | 15.93 | 2.09 | 46.81 | 13.97 | 51.65 | 4.60 | 55.54 | 278.11 | 3 | 8725 |

Cette interprétation de l'axe 1 est aussi confirmé par la très forte corrélation positive entre la variable supplémentaire "Points" et l'axe 1.

- **Axe 2**. Examinons de plus près les sorties concernant les variables :

```
Variables
                Dim.1    ctr   cos2    Dim.2    ctr   cos2
X100m         | -0.775 18.344  0.600 |  0.187  2.016  0.035
Long.jump     |  0.742 16.822  0.550 | -0.345  6.869  0.119
Shot.put      |  0.623 11.844  0.388 |  0.598 20.607  0.358
High.jump     |  0.572  9.998  0.327 |  0.350  7.064  0.123
X400m         | -0.680 14.116  0.462 |  0.569 18.666  0.324
X110m.hurdle | -0.746 17.020  0.557 |  0.229  3.013  0.052
Discus        |  0.552  9.328  0.305 |  0.606 21.162  0.368
Pole.vault    |  0.050  0.077  0.003 | -0.180  1.873  0.033
Javeline      |  0.277  2.347  0.077 |  0.317  5.784  0.100
```

```
X1500m        | -0.058  0.103  0.003 |  0.474 12.946  0.225
```

Au regard des contributions et des cosinus carrés pour l'axe 2, on remarque que les variables "lancer du poids", "lancer du disque", "400m" et dans une moindre mesure "1500m" ont un impact important. Comme ces variables sont corrélées positivement avec l'axe 2, on peut dire que le second axe oppose en haut les athlètes puissants (fort au lancer du disque et du poids et plutôt faible sur les épreuves de longue distance comme le 400m et le 1500m) des autres situés en bas de l'axe 2.

- **Un peu plus loin dans l'interprétation.** Y-a-t-il une différence significative entre les résultats des décathloniens ayant participé aux deux compétitions "decastar" et "jeux olympiques" ? Pour cela, regardons les sorties fournies pour la variables qualitatives :

```
Supplementary categories
                 Dist    Dim.1   cos2 v.test    Dim.2   cos2 v.test
Decastar     |  0.946 | -0.600  0.403 -1.430 | -0.038  0.002 -0.123
OlympicG     |  0.439 |  0.279  0.403  1.430 |  0.017  0.002  0.123
```

Les valeurs correspondant à la colonne "v.test" sont toutes comprises entre -1.96 et +1.96. Cela signifie que les coordonnées des modalités "decastar" et "jeux olympiques" ne sont pas significativement différentes de 0.

Conclusion : bien que la représentation des modalités "decastar" et "jeux olympiques" fasse apparaître une différence donnant l'impression que les athlètes ont légèrement amélioré leur performance lors des jeux olympiques, les résultats obtenus sont néanmoins statistiquement les mêmes.

- **Comment atteindre directement les différents éléments contenus dans `res.pca` ?**

Pour comprendre comment atteindre les différents éléments contenus dans `res.pca` concernant les variables, exécuter les lignes de commandes suivantes :

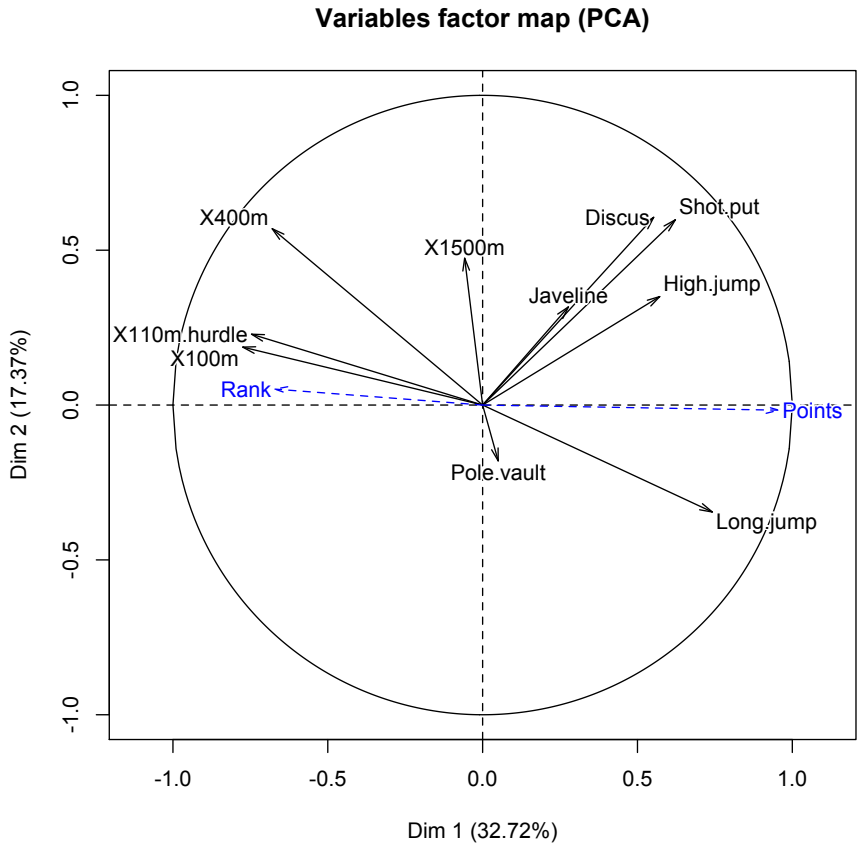


FIGURE 1.1 – Données sur les décathloniens : cercle des corrélations avec variables supplémentaires

```
> res.pca$var$coord[, "Dim.1"]
> res.pca$var$coord[, "Dim.2"]
> res.pca$var$cor[, "Dim.1"]
> res.pca$var$cor[, "Dim.2"]
> res.pca$var$cos2[, "Dim.1"]
> res.pca$var$cos2[, "Dim.2"]
> res.pca$var$cos2[,"Dim.1"] + res.pca$var$cos2[,"Dim.2"]
> res.pca$var$contrib[, "Dim.1"]
> res.pca$var$contrib[, "Dim.2"]
```

Pour comprendre comment atteindre les différents éléments contenus dans `res.pca` concernant les individus, exécuter les lignes de commandes suivantes :

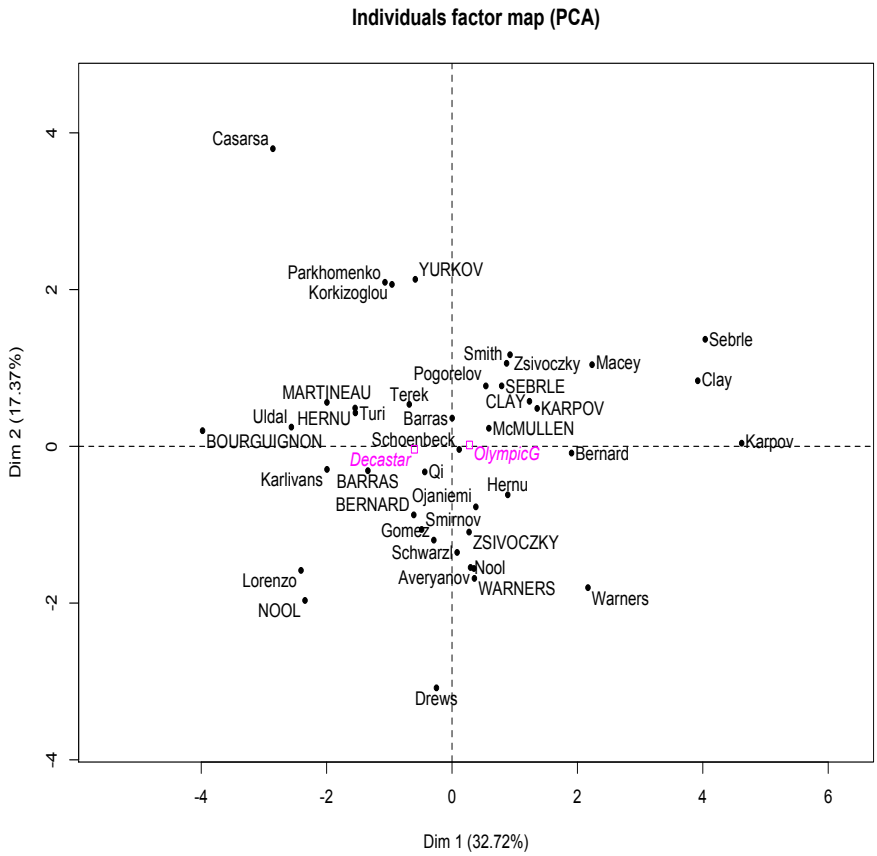


FIGURE 1.2 – Données sur les décathloniens : graphe des individus sur le plan principal accompagné de la représentation des modalités de la variable qualitative supplémentaire

```
> res.pca$ind$coord[, "Dim.1"]
> res.pca$ind$coord[, "Dim.2"]
> res.pca$ind$cos2[, "Dim.1"]
> res.pca$ind$cos2[, "Dim.2"]
> res.pca$ind$cos2[,"Dim.1"] + res.pca$ind$cos2[,"Dim.2"]
> res.pca$ind$contrib[, "Dim.1"]
> res.pca$ind$contrib[, "Dim.2"]
```

À propos des inerties et inerties cumulées :

```
> res.pca$eig[, "percentage of variance"]
> res.pca$eig[, "cumulative percentage of variance"]
```

### 1.1.3 Exercice

Reprenez les résultats de l'ACP précédente et répondez aux diverses questions suivantes (une fois l'exercice fait, vous pouvez consulter les réponses à la page suivante) :

1) Quel est le % d'inertie expliqué par l'axe 1 ? l'axe 2 ?
2) Quel est le % d'inertie expliqué par le plan principal (1,2) ?
3) Quel est l'individu le mieux représenté sur l'axe principal 1 ? le moins bien représenté sur l'axe principal 2 ?
4) Quelle est la corrélation entre la variable `X100m` et le facteur 1 ? Quelle est la variable la moins corrélée avec le facteur 1 ?
5) Quelle est la variable la mieux représentée sur l'axe factoriel 1 ? la moins bien représentée sur l'axe factotriel 2 ?
6) Quelles sont les 2 variables les mieux représentées sur le plan factoriel (1,2) ?
7) Quels sont les 3 individus qui contribuent le plus à l'axe principal 1 ?
8) Quels sont les 4 individus qui contribuent le plus à l'axe principal 2 ?
9) Quelle est la variable qui contribue le plus à l'axe factoriel 1 ? Le moins ?
10) Quelle est la variable qui contribue le plus à l'axe factoriel 2 ? Le moins ?

## 1.2 ACP des données de températures

Les données de températures ont déjà été abordées dans le cadre du cours. Votre travail est de mettre en oeuvre l'analyse en composantes principales sur ce jeu de données disponible à l'adresse

https ://www.math.univ-toulouse.fr/∼ferraty/DATA/acp-temperatures-villes-francaises.txt

### 1.2.1 Partie 1 : ACP des données de températures

Répondez aux questions suivantes :

1. Indiquer l'adresse de ce fichier dans votre navigateur pour voir le format des données,

2. Charger les données à l'aide de la fonctionnalité *Import Dataset* de RStudio en prenant soin de choisir les bonnes options de format,
3. Afficher un résumé des données de températures ; quelles sont les variables supplémentaires ?
4. Réaliser l'ACP à l'aide de la fonction `PCA` du package `FactoMineR`,
5. Afficher un résumé des sorties numériques de l'ACP en faisant apparaître tous les individus,
6. Représenter la décroissance des inerties associées à chaque axe factoriel (éboulis des valeur propres),
7. représenter les individus sur le plan principal,
8. Représenter le cercle des corrélations,
9. Quelles sont les variables qui différencient le plus les individus par rapport à l'axe 1 ?
10. Quelles sont les variables qui différencient le plus les individus par rapport à l'axe 2 ?
11. À partir des réponses précédentes, interpréter les résultats de l'ACP.

### 1.2.2 Partie 2 : réalisation d'un rapport `Rmarkdown` généré automatiquement avec la fonctionnalité `Knit` de RStudio

## 1.3 `FactoshinyACP` : un outil d'aide à l'interprétation des analyuses factorielles

`Factoshiny` est un package de R proposant des fonctionnalités complémentaires à celles de `FactoMineR`. `Factoshiny` propose une interface agréable pour mettre en oeuvre l'ensemble des analyses factorielles dont l'ACP. `Factoshiny` est aussi un outil d'aide à l'interprétation puisqu'il produit automatiquement des commentaires. Cet outil propose même d'aller plus loin en réalisant une classification des individus à partir des axes factoriels.

`FactoMineR` produit automatiquement des commentaires dont vous pourrez vous inspirer pour faire vos propres interprétations. Comme tout système expert, les commentaires fournis par `FactoMineR` sont imparfaits. Néanmoins il peuvent servir d'ancrage pour écrire des interprétations pertinentes.

Les 3 lignes de commandes suivantes suffisent pour lancer l'ACP avec `Factoshiny` sur le jeu de données concernant les températures mensuelles des villes françaises :

```
> TEMP = read.csv("http://www.math.univ-toulouse.fr/~ferraty/DATA
/acp-temperatures-villes-francaises.txt", header = TRUE, sep =
"", dec = ",", row.names = 1)
> library(Factoshiny)
> PCAshiny(TEMP)
```

Une fois la commande `PCAshiny(TEMP)` exécutée, l'interface graphique apparaît dans votre navigateur ; on peut alors réaliser l'ACP de façon interactive sous forme de menus permettant de choisir des options de représentation.

## 1.4 ACP : application au traitement d'images

On s'intéresse à une visualisation 3D de l'analyse en composantes principale d'un jeu de données appelé `MNIST` (*Modified ou Mixed National Institute of Standards and Technology*). Ce jeu de données contient 10000 images (= individus) représentant les chiffres (de 0 à 9) écrits à la main. Chaque chiffre écrit à la main est représenté par une image 28 x 28 pixels ou chaque pixel donne une intensité de gris (entre 0 et 255). En vectorisant les images (ici colonne par colonne), chaque chiffre (individu) est représenté par un vecteur de taille 784 = 28 x 28 (1 colonne supplémentaire fournit le chiffre qui est sensé être écrit). Au final, on dispose d'un tableau 10000 x 785 contenant les 10000 chiffres écrits à la main.

**Chargement des données**

```
mnist_test <- read.csv("http://www.math.univ-toulouse.fr/~ferraty/
DATA/mnist_test.csv")
nbpixels = ncol(mnist_test) - 1
# noms de variables attribués à chaque colonne de minst_test
dimnames(mnist_test)[[2]] = c("digit", paste0("pixel", 1:nbpixels))
# déclaration de "digit" comme variable catégorielle
mnist_test$digit = as.factor(mnist_test$digit)
```

### Représentation des 300 premiers chiffres écrits à la main (fig. 1.3)

```
DATA = as.matrix(mnist_test[, -1])
par(mar=c(0,0,0,0), mfrow = c(15,20))
for(number in 1:300){
  # chaque image est un vecteur de taille 784 = 28 * 28
  # "img" = version matricielle 28 lignes et 28 colonnes
  # (image 28 pixels par 28 pixels
  img = matrix(DATA[number, ], 28, 28, byrow=F)
  # la fonction "image" affiche l'image du chiffre écrit à la main
  image(255 - img[, nrow(img):1], axes=FALSE,
      col = gray.colors(255))
  }
```

### Mise en oeuvre de l'ACP

```
  library(FactoMineR)
res.PCA <- PCA(mnist_test,quali.sup=1,graph=FALSE)
```

### Représentation 3D "optimale" d'un échantillon de 1000 images (fig. 1.4)

```
# Sélection aléatoire d'un sous-échantillon "echan"
# La commande "set.seed(1)" permet de reproduire le même
# sous-échantillon chaque fois que l'on exécute le code
set.seed(1)
echan = sample(1:10000, 1000)
df = data.frame(pc1 = res.PCA$ind$coord[echan, 1],
```

```
                pc2 = res.PCA$ind$coord[echan, 2],
                pc3 = res.PCA$ind$coord[echan, 3],
                label = mnist_test$digit[echan],
                id = echan)
Colors <- c("black","blue","red","green","purple",
            "pink","orange","brown","cyan","gray50")
fig <- plot_ly()
fig <- fig %>%
  add_trace(
    type='scatter3d', mode = 'markers',
    x = df$pc1, y = df$pc2, z = df$pc3,
    text = df$id,
    color = df$label, colors = Colors,
    hoverinfo = 'text',
    showlegend = TRUE)
fig
```

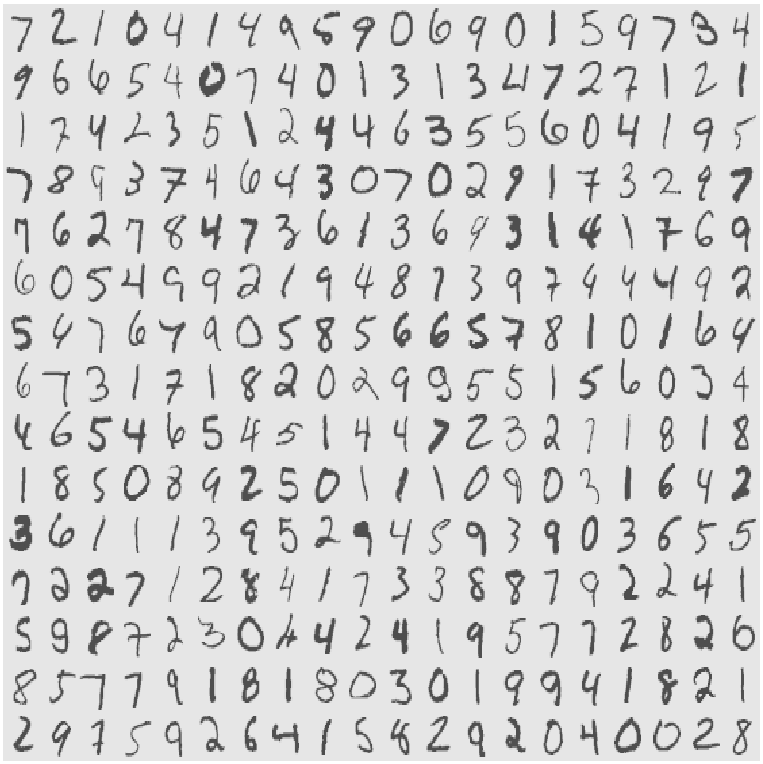

FIGURE 1.3 – 300 premiers chiffres écrits à la main

La figure 1.4 représente la meilleure représentation 3D issue de l'ACP d'un sous-

échantillon de 1000 images (rappel : chaque image est associée à un chiffre écrit à la main et vit dans un espace possédant 784 dimensions). Le code couleur correspond au chiffre représenté (10 chiffres = 10 couleurs). Lorsque ce graphique est réalisé dans RStudio, il devient interactif car on peut le faire pivoter avec la souris et on peut se promener dans le nuage de point. Il y a aussi une fonctionnalité "hover" (option hoverinfo de la fonction R *plot_ly* du package *plotly*) qui permet d'identifier chaque point (= image = individu) par du texte.

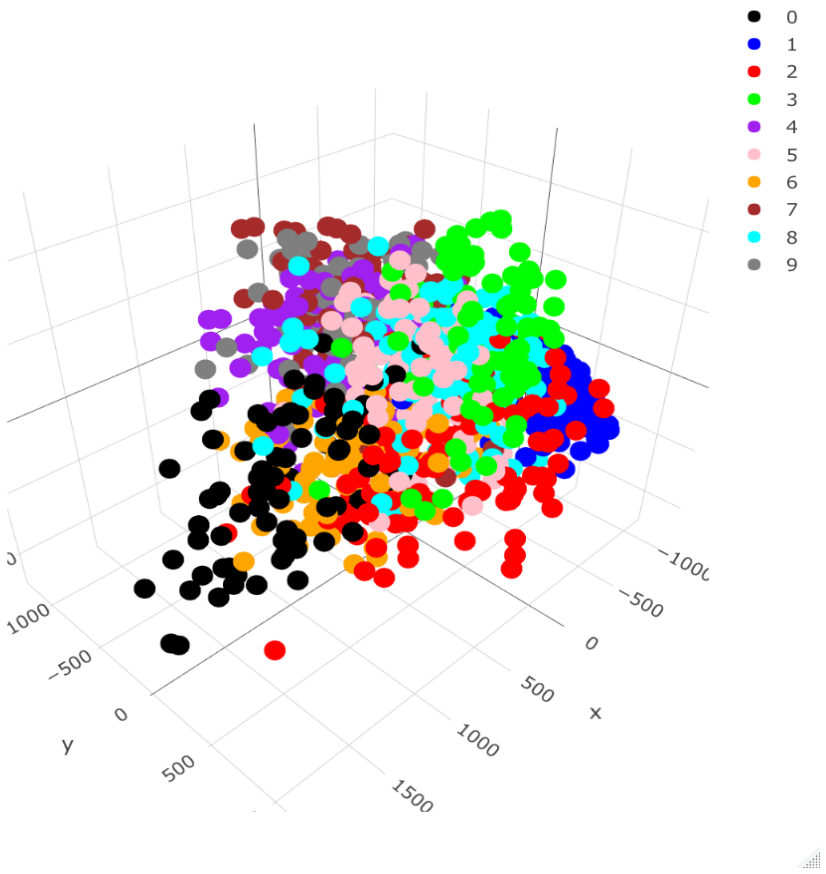


FIGURE 1.4 – Meilleure approximation 3D du nuage des individus (chaque point est une image)

### Reconstruction/compression des images à partir des premiers axes factoriels

Le paragraphe du polycopié de cours dédié à la reconstruction des données indique comment on peut obtenir une approximation des données en utilisant un procédé de reconstruction fondé sur les premiers axes factoriels et premières composantes principales. Cette technique s'applique à notre échantillon d'images représentant des

chiffres écrits à la main. Le code suivant produit la figure 1.5

```
nbcp = c(0, 3, 10, 50, 100, 200)
res.PCA = PCA(mnist_test, quali.sup = 1, graph = FALSE,
ncp = max(nbcp), scale.unit = FALSE)
MNIST =  as.matrix(mnist_test[, -1])
echan = c(1:5, 8, 9, 12, 19, 62)
nsample = length(echan)
par(mar=c(0,1,0,0), mfrow = c(nsample, length(nbcp)))
for(kk in nbcp){
  if(kk == 0){
    img = matrix(MNIST[echan[1], ], 28, 28, byrow=F)
    image(255 - img[, nrow(img):1], axes=FALSE,
        col = gray.colors(255), main = )
  }else{
    # "IMG_APPROX" contient les images (vectorisées) reconstruites
    # à partir des "kk" premiers axes factoriels
    IMG_APPROX = tcrossprod(res.PCA$ind$coord[, 1:kk],
         res.PCA$svd$V[, 1:kk])
    img = matrix(IMG_APPROX[echan[1], ], 28, 28, byrow=F)
    image(255 - img[, nrow(img):1], axes=FALSE,
        col = gray.colors(255))
  }
}
for(number in echan[-1]){
  for(kk in nbcp){
    if(kk == 0){
      img = matrix(MNIST[number, ], 28, 28, byrow=F)
      image(255 - img[, nrow(img):1], axes=FALSE,
          col = gray.colors(255))
    }else{
       IMG_APPROX = tcrossprod(res.PCA$ind$coord[, 1:kk],
           res.PCA$svd$V[, 1:kk])
```

```
        img = matrix(IMG_APPROX[number, ], 28, 28, byrow=F)
        image(255 - img[, nrow(img):1], axes=FALSE,
            col = gray.colors(255))
      }
    }
  }
```

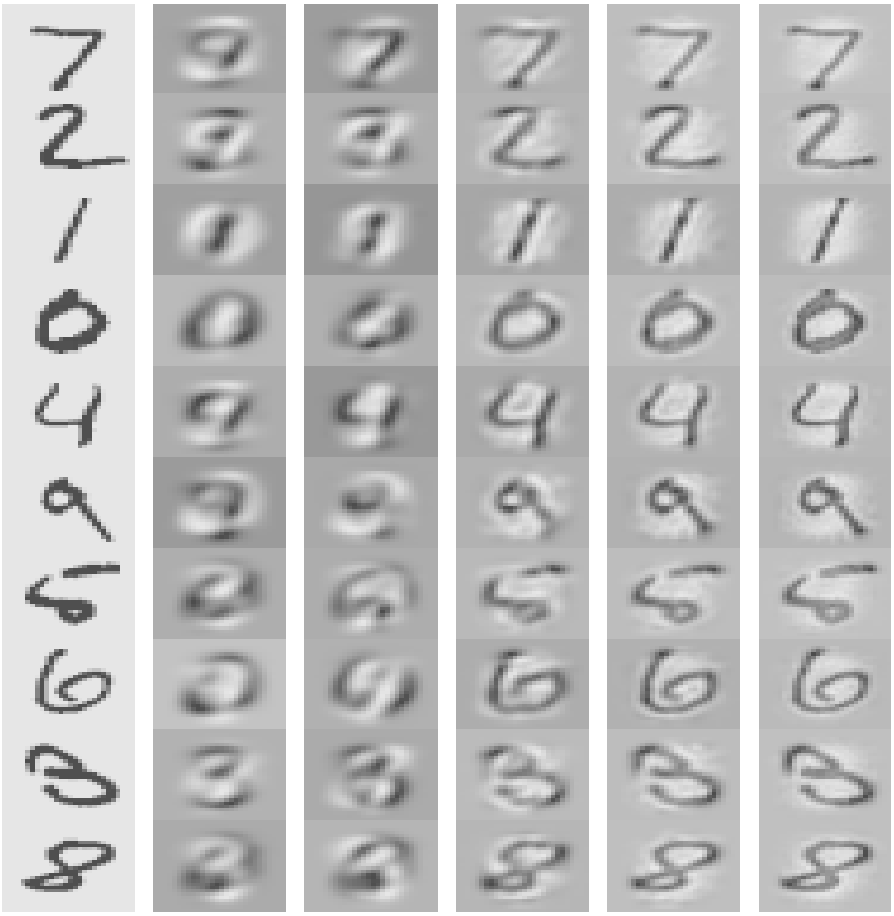

FIGURE 1.5 – Colonne 1 = images observées ; colonne 2 = approximation d'ordre 3 ; colonne 3 = approximation d'ordre 10 ; colonne 4 = approximation d'ordre 50 ; colonne 5 = approximation d'ordre 100 ; colonne 4 = approximation d'ordre 200

**Commentaires sur cette reconstruction d'images**. Il est intéressant de noter que l'information est quasi complète pour $q = 50$, c'est-à-dire qu'avec une approximation d'ordre 50, on est capable d'identifier les chiffres écrits. L'avantage d'une telle réduction de dimension est qu'elle permet de diminuer considérablement l'espace mémoire nécessaire pour sauvegarder les images ainsi approximées. En effet, pour ce jeu de données, on devra mémoriser seulement $539,200$ éléments au lieu des 7,840,000

(10,000 x 784). L'approximation des images ainsi obtenues occupe environ 15 fois moins d'espace mémoire que les images originales (pour plus de détails sur le calcul du gain en espace mémoire, voir paragraphe **??**).

## 1.5 ACP : un outil pour la reconnaissance de forme

L'objectif de cette section est de fournir, à travers un jeu de données simulées, la capacité de l'ACP à reconnaître des formes dans des nuages de points. Dans ce but, on considère les données contenues dans le fichier qui se trouve à l'adresse

*http ://www.math.univ-toulouse.fr/~ferraty/DATA/HiddenShape.txt*

Télécharger ce fichier qui est format `texte` (chaque élément d'une ligne étant séparé par un espace) dans un répertoire local `pathname` (qui représente ici le chemin du répertoire local dans lequel vous avez téléchargé le fichier `HiddenShape.txt`) puis chargez-le dans votre session R à l'aide de la commande

```
> HIDSHA = read.csv(paste(pathname, "HiddenShape.txt", sep =
"", header = FALSE, sep = " ")
```

Ce jeu de données correspond à un tableau comportant 375 lignes et 4 colonnes, lequel contient des valeurs quantitatives. Une forme se cache dans ce jeu de données. En réalisant une ACP non réduite et en représentant les individus sur le "bon" sous-espace principal, trouver cette forme cachée.

## 1.6 Enquêtes "Ces villes où il fait bon vivre"

### 1.6.1 Description des données

Les données concernent les villes de province (de plus de 100 000 habitants) et ont été publiées dans le "Nouvel Observateur" (1993) pour illustrer un article : "Ces villes où il fait bon vivre". Elles se composent des 12 variables quantitatives suivantes

observées sur 56 villes (i.e. individus) :

**m20ans** → Proportion de moins de 20 ans,

**mil** → Montant moyen des impôts locaux en francs,

**ncd** → Nombre de crimes et délits pour 1000 hab. par département,

**nce** → Nombre de créations d'entreprise pour 1000 hab.,

**nei** → Nombre d'entreprises industrielles,

**nes** → Nombre d'entreprises de service,

**nls** → Nombre de licenciés (sportifs) pour 100 hab.,

**qpo** → Quantité de polluants émis par les entreprises,

**rmc** → Recettes moyennes de cinéma par hab. en francs,

**sam** → Salaire annuel moyen par hab. en francs,

**tcd** → Taux de chomage du département,

**trb** → Taux de réussite au bac.

L'ensemble de ces données est contenu dans le fichier `province1.csv` disponible à l'adresse

*http ://www.math.univ-toulouse.fr/∼ferraty/DATA/province1.csv*

### 1.6.2 Importation des données

Exécuter les commandes suivantes :

```
>PROVINCE = read.csv("http://www.math.univ-toulouse.fr/
~ferraty/DATA/province1.csv", header = T, row.names = 1,
dec = ".", sep = ";")
```

### 1.6.3 Analyse en Composantes Principales

Réalisez l'ACP des données contenues dans `PROVINCE` puis répondez aux questions suivantes :

1. Combien de dimensions retiendriez-vous pour analyser ces données ? Quel pourcentage d'inertie est associé au plan factoriel (1, 2) ?

2. Quelle est la variable la moins bien représentée sur l'axe 1 ? Quelle est la variable la mieux représentée sur l'axe 2 ?
3. Quelle est la corrélation entre la variable **nce** et l'axe factoriel 1 ? Quelle est la corrélation entre la variable **tcd** et l'axe factoriel 2 ?
4. Un commentaire indique que les villes représentées à droite sur l'axe 1 possède un salaire moyen par hab. plus élevé que celles situées à gauche de l'axe 1. Êtes-vous d'accord avec cette affirmation ? (justifier brièvement votre réponse)
5. À votre avis : quelle est la ville potentiellement la plus polluée ? les 3 villes possédant potentiellement la plus forte proportion de jeunes de moins de 20 ans ?
6. Quelles sont les 4 variables qui contribuent le plus à la construction de l'axe 1 ? Quelles sont les 4 variables qui contribuent le plus à la construction de l'axe 2 ?
7. D'après ce qui précède, écrire un commentaire permettant de caractériser les villes Bethune, Calais et Lens.

# Chapitre 2

# Analyse Factorielle des Correspondances (AFC)

## 2.1 Reconnaissance de trois saveurs (sucré, acide, amer)

**cas 1 : très faible confusion des saveurs**

On considère la table de contingence suivante :

| | perçu sucré | perçu acide | perçu amer |
|---|---|---|---|
| sucré | 10 | 0 | 0 |
| acide | 0 | 10 | 1 |
| amer | 0 | 2 | 7 |

1. **Ouvrir une session R**.
2. **Chargement direct des données**

```
> SAVEURS1 = matrix(c(10, 0, 0, 0, 10, 2, 0, 1, 7), nrow
= 3)
```

Pour plus de lisibilité dans la lecture des résultats de l'AFC, il est important de nommer les modalités. Ceci se fait à l'aide la commande suivante :

```
dimnames(SAVEURS1) = list(c("sucré","acide", "amer"),
c("perçu_sucré", "perçu_acide", "perçu_amer"))
```

3. **Réalisation de l'AFC.** Charger le package "factoMineR" (normalement, vous avez déjà télégchargé le package "factoMineR" lors des séances concernant l'ACP) grâce à la commande

```
library("FactoMineR")
```

Lancer l'AFC en utilisant la fonction "CA" du package "FactoMineR" :

```
res.ca1 = CA(SAVEURS1)
```

La représentation simultanée des modalités apparaît automatiquement (en bleu les modalités lignes et en rouge les modalités colonnes). Pour obtenir un descriptif des axes (inerties, contributions et cosinus carrés), il suffit de taper la commande

```
summary(res.ca1)
```

Affichage des coordonnées des modalités qui participent le plus à la construction des axes :

```
dimdesc(res.ca1, axes = 1:2)
```

Remarque : dans cet exemple on a la totalité de l'inertie représentée sur les deux 1ers axes. C'est pour cela qu'on précise l'option "axes = 1 :2" de la function "dimdesc".

Représentation de la décroissance des inerties selon les axes :

```
barplot(res.ca1$eig[,2], main = "Décroissance des inerties
(%)", names.arg = 1:nrow(res.ca1$eig))
```

**cas 2 : confusion moyenne des saveurs**

Réaliser à nouveau l'AFC en considérant la table de contingence suivante :

| | perçu sucré | perçu acide | perçu amer |
|---|---|---|---|
| sucré | 10 | 0 | 0 |
| acide | 0 | 8 | 2 |
| amer | 0 | 4 | 6 |

Dans ce but, par analogie avec ce qui précède, vous créerez une nouvelle table de contingence "SAVEURS2" ainsi qu'un objet "res.ca2" résultant de l'AFC de "SAVEURS2".

**cas 3 : forte confusion des saveurs**

Réaliser à nouveau l'AFC en considérant la table de contingence suivante :

| | perçu sucré | perçu acide | perçu amer |
|---|---|---|---|
| sucré | 10 | 0 | 0 |
| acide | 0 | 6 | 4 |
| amer | 0 | 6 | 4 |

Dans ce but, par analogie avec ce qui précède, vous créerez une nouvelle table de contingence "SAVEURS3" ainsi qu'un objet "res.ca3" résultant de l'AFC de "SAVEURS3".

**Questions**

- Que remarquez-vous lorsque la confusion entre les saveurs acide/amer augmente du point de vue de la représentation simulatnée des modalités et du comportement de l'inertie ?
- Ce comportement était-il prévisible ?

### 2.1.1 Enquête sociologique sur le travail des femmes en 1974

L'objectif de cet exercice est de réaliser l'AFC de la table de contingence concernant la compatibilité du travail féminin avec la vie de famille à partir des données suivantes :

| | rester au foyer | trav. mi-temps | trav. plein-temps |
|---|---|---|---|
| 2 conj. trav. égal | 13 | 142 | 106 |
| trav. mari + absorbant | 30 | 408 | 117 |
| seul le mari trav. | 241 | 573 | 94 |

Pour cela, suivez méthodiquement les différentes étapes suivantes.

1. **Saisie de la table de contingence**

   Construisez "à la main" une matrice que vous nommerez `FEMMES` contenant ces données puis, à l'aide de la commande `dimnames`, affectez aux modalités les labels correspondant.

   Le tableau FEMMES contient maintenant la table de contingence. Pour plus de lisibilité dans la lecture des résultats, nommez les modalités :

2. **Réalisation de l'AFC.**

   Lancer l'AFC en utilisant la fonction "CA" du package "FactoMineR" :

   ```
   res.ca = CA(FEMMES)
   ```

   La représentation simultanée des modalités apparaît automatiquement (en bleu les modalités lignes et en rouge les modalités colonnes).
   Descriptif des axes (inerties, contributions et cosinus carrés) :

   ```
   summary(res.ca)
   ```

   Affichage des coordonnées des modalités qui participent le plus à la construction des axes :

   ```
   dimdesc(res.ca, axes = 1:2)
   ```

   La totalité de l'inertie est-elle représentée sur les deux 1ers axes ?

   Représentation de la décroissance des inerties selon les axes :

```
barplot(res.ca$eig[,2], main = "Décroissance des inerties
(%)", names.arg = 1:nrow(res.ca$eig))
```

3. **Questions**

   (a) Retrouver les interprétations données en cours,

   (b) Identifier la(les) modalité(s) de $X$ et $Y$ qui contribue(nt) le plus au 1er axe,

   (c) Identifier la(les) modalité(s) de $X$ et $Y$ qui contribue(nt) le plus au 2ème axe,

   (d) Identifier la(les) modalité(s) de $X$ et $Y$ la(les) mieux représentée(s) sur le 1er axe,

   (e) Évaluer l'intensité de la liaison entre $X$ et $Y$ pour chaque axe (est-elle faible ? moyenne ? forte ?)

   (f) À partir des relations de transition :
   - calculer les coordonnées des modalités de $X$ à partir de celles de $Y$,
   - calculer les coordonnées des modalités de $Y$ à partir de celles de $X$.

   (g) Déterminer la distance du khi-deux entre les 2 profils-colonnes "rester au foyer" et "travailler à plein-temps" puis entre "rester au foyer" et "travailler à mi-temps" ; les résultats obtenus vous paraissent-ils cohérents ?

## 2.2 "Ces villes où il fait bon vivre"

Les données concernent les villes de province (de plus de 100 000 habitants) et ont été publiées dans le "Nouvel Observateur" (1993) pour illustrer un article : “Ces villes où il fait bon vivre”. Elles se composent de 12 variables quantitatives observées sur 56 villes (i.e. individus). Un distributeur de films de cinéma s'intéresse particulièrement aux deux variables suivantes :

**m20ans** → Proportion de moins de 20 ans,

**rmc** → Recettes moyennes de cinéma par hab. en francs.

L'objectif de cet exercice est d'étudier la relation entre la variable **rmc** "recettes moyennes de cinéma par hab." et la variable **m20ans** "proportion de moins de 20 ans". Dans ce but, on découpe la variable **rmc** de sorte à obtenir la variable catégorielle **rmc_cat** possédant les 3 modalités **rmc_cat1** (faible recette), **rmc_cat2** (recette moyenne) et **rmc_cat3** (recette élevée). On réitère la même opération pour la variable **m20ans** ce qui nous permet d'obtenir la variable catégorielle **m20ans_cat** possédant les 3 modalités **m20ans_cat1** (faible proportion), **m20ans_cat2** (proportion moyenne) et **m20ans_cat3** (proportion élevée). On obtient ainsi la table de contingence suivante :

| | m20ans_cat1 | m20ans_cat2 | m20ans_cat3 |
|---|---|---|---|
| rmc_cat1 | 0 | 7 | 12 |
| rmc_cat2 | 8 | 6 | 5 |
| rmc_cat3 | 11 | 5 | 2 |

TABLE 2.1 – Table de contingence

1. Créer un objet R contenant cette table de contingence ; vous prendrez soin de nommer les modalités en utilisant les mêmes labels,
2. Réaliser l'AFC de cette table ; vous afficherez les différents éléments de l'analyse,
3. Déterminer le tableau des profils-lignes ; donner le profil-ligne moyen $\overline{PL}$,
4. Calculer $d_{\chi^2}(\text{"}rmc\_cat1\text{"}, \overline{PL})$ puis $d_{\chi^2}(\text{"}rmc\_cat2\text{"}, \overline{PL})$ ; les valeurs obtenues sont-elles en accord avec la représentation simultanée des modalités ?
5. À partir des résultats de l'AFC, déduisez la valeur de l'indicateur $\Phi^2$ ; l'axe 1 est-il le lieu d'une association exclusive ?
6. D'après vous, combien d'axes sont nécessaires pour représenter les modalités de ces 2 variables ? (justifier brièvement votre réponse)
7. À partir de la représentation simultanée des modalités, commentez brièvement la relation entre ces 2 variables catégorielles.

## 2.3 Enquête de satisfaction à l'hôpital

Une enquête de statisfaction a produit les données disponibles à l'adresse

*http ://www.math.univ-toulouse.fr/∼ferraty/DATA/satisfaction_hopital.csv*

Il s'agit d'une étude évaluant la qualité de relation et la quantité d'information reçue par le patient lors de son séjour à l'hôpital. 534 patients ont été recrutés sur plusieurs hôpitaux de la région parisienne. Ce fichier contient les variables suivantes :

- service : code (de 1 à 8) du service ayant accueilli le patient,
- sexe : sexe du patient (0 homme, 1 femme)
- age : âge en années
- profession :

  1 : agriculteur exploitant

  2 : artisan, commerçant, chef d'entreprise

  3 : cadre, profession intellectuelle ou artistique, profession libérale

  4 : profession intermédiaire de l'enseignement, de la santé, du travail social ou de la fonction publique, technicien, contremaître, agent de maîtrise, clergé

  5 : employé

  6 : ouvrier

  7 : étudiant, militaire, chômeur sans avoir jamais travaillé

  8 : autre
- amelioration.sante : impression d'amélioration de la santé du fait du séjour à l'hôpital (codé de 0 : aggravée, à 3 : nettement améliorée)
- amelioration.moral : impression d'amélioration du moral du fait du séjour à l'hôpital (codé de 0 : aggravé, à 3 : nettement amélioré)
- recommander : recommander le service à son entourage (codé 0 : non, 1 : oui, probablement, 2 : oui, sûrement)
- score.information : score relatif à la qualité de l'information reçue pendant le séjour (score variant de 10 à 40)
- score.relation : score relatif à la qualité des relations avec le personnel soignant pendant le séjour

**Questions**

1. Créer un objet `HOPITAL` contenant l'ensemble de ces données,
2. À l'aide de la fonction R "`as.factor`", transformer les variables qui le nécessitent,
3. Étudier la relation entre les variables `amelioration.moral` et `amelioration.sante`,
4. Recommencer avec un autre couple de variables catégorielles de votre choix.

# Chapitre 3

# Analyse Factorielle des Correspondances Multiples (AFCM)

## 3.1 Mise en oeuvre de l'AFCM sur l'étude concernant les activités de loisir des français

1. Ouvrir une session R.
2. Chargement des données dans R. Le chargement du fichier de données "loisir.csv" (ici au format ".csv") à partir d'une adresse web se fait comme suit :

```
> LOISIRS = read.table("http ://www.math.univ-toulouse.fr/
~ferraty/DATA/loisirs.csv", header = T, sep = ";")
```

L'objet `LOISIRS` que vous venez de créer dans R contient les données.

**Remarque 3** *On peut aussi télécharger le fichier à l'aide d'un moteur de recherche et le sauvegarder dans un répertoire de sa machine. Il suffit alors d'utiliser la ligne de commande R suivante*

```
> loisirs = read.table(paste(path, loisirs.csv, sep = ""), header = TRUE,
sep = ";")
```

*où* `path` *contient la chaîne de caractères donnant le chemin du fichier de données* `loisirs.csv`.

3. Descriptif des variables

```
> summary(LOISIRS)
```

Pb : la variable 18 "usage TV" n'est pas identifiée comme une variable catégorielle mais comme une variable quantitative. En effet, au lieu de fournir l'effectif pour chaque modalité, R affiche la moyenne, l'ecart-type, etc ce qui est le propre d'une variable quantitative. La commande qui suit permet de la déclarer comme variable catégorielle :

```
> LOISIRS[, 18] = as.factor(LOISIRS[, 18])
```

A l'aide de la commande

```
> summary(LOISIRS)
```

on vérifie bien que la variable 18 "usage TV" est bien identifiée comme une variable catégorielle.

4. Chargement de la librairie `FactorMineR` :

```
> library(FactoMineR)
```

5. Réalisation de l'AFCM : les 4 variables signalétiques 19, 20, 21 et 22 ainsi que la variable quantitative 23 sont déclarées comme variables supplémentaires :

```
> res.mca = MCA(LOISIRS, quali.sup = 19 :22, quanti.sup = 23)
```

L'objet qui vient d'être créé `res.mca` contient les résultats de l'AFCM. De plus, la fonction `MCA` ouvre autoamtiquement 4 fenêtres graphiques dans lesquelles vous pouvez observer :

- le graphe des modalités des variables actives (en rouge) et supplémentaires (en vert) sur le 1er plan factoriel,
- le graphe des individus sur le 1er plan factoriel,
- la représentation des variables actives (en rouge), catégorielles supplémentaires (en vert) et quatitative supplémentaire (en bleu) sur le 1er plan,
- la représentation de la variable quantitative supplémentaire.

La commande

```
> summary(res.mca)
```

fournit un résumé numérique de l'AFCM. On peut y lire les inerties exprimées par chaque axe (ligne `Variance`), contributions et qualités de représentation (cosinus carrés) pour chaque dimension, qu'il s'agissent des individus ou des modalités.

6. Graphe des pourcentages d'inertie associés à chaque axe.

```
> barplot(res.mca$eig[,2], main = "Pourcentage d'inertie", names.arg =
1 :nrow(res.mca$eig))
```

7. Graphe des individus sur les 2 premiers axes de l'analyse :

```
> plot(res.mca, choix = "ind", invisible = "var", label = "none", cex =
0.5)
```

L'option `label = "none"` efface le numéro identifiant chaque individu alors que l'option `cex = "0.5"` réduit de moitié la taille des points représentant les individus.

**Remarque 4** *Il est possible de colorer les individus selon les modalités d'une variable catégorielle. Par exemple, si on s'intéresse à l'activité* `Jardinage`*, observer ce que l'on obtient grâce à la commande :*

```
> plot(res.mca, choix = "ind", hab = "Jardinage", label = "none", cex =
0.5)
```

8. Graphe du nuage des modalités sur les 2 premiers axes de l'analyse :

```
> plot(res.mca, invisible = c("ind", "quali.sup"), hab = "quali")
```

L'option `hab = "quali"` permet d'affecter une couleur par variable (les modalités d'une variable possèdent la même couleur).

9. Repérage des individus extrêmes sur les axes de l'analyse.
   - **Axe 1**. La commande suivante permet d'identifier les 4 individus les plus à droite sur l'axe 1 puis d'afficher leurs réponses aux 18 questions concernant leurs activités de loisir :

```
> indexes = order(res.mca$ind$coord[, 1])[8400 :8403]
> LOISIRS[indexes, 1 :18]
```

La commande suivante permet de d'identifier les 4 individus les plus à gauche sur l'axe 1 puis d'afficher leurs réponses aux 18 questions concernant leurs activités de loisir :

```
> indexes = order(res.mca$ind$coord[, 1])[1 :4]
> LOISIRS[indexes, 1 :18]
```

- **Axe 2**. La commande suivante permet d'identifier les 2 individus les plus haut sur l'axe 2 puis d'afficher leurs réponses aux 18 questions concernant leurs activités de loisir :

```
> indexes = order(res.mca$ind$coord[, 2])[8402 :8403]
> LOISIRS[indexes, 1 :18]
```

La commande suivante permet d'identifier les 2 individus les plus bas sur l'axe 2 puis d'afficher leurs réponses aux 18 questions concernant leurs activités de loisir :

```
> indexes = order(res.mca$ind$coord[, 2])[1 :2]
> LOISIRS[indexes, 1 :18]
```

10. Représentation des variables sur le 1er plan.

```
> plot(res.mca, choix = "var")
```

Ce graphique permet de repérer les variables qui contribuent le plus à la construction des axes (en terme d'inertie).

11. Représentation des modalités des variables supplémentaires sur le 1er plan.

```
> plot(res.mca, invisible = c("ind", "var"), hab= "quali", palette = palette(c("blue", "orange", "darkgreen", "red")))
```

L'option `palette = palette(c("blue", "orange", "darkgreen", "red"))` permet de choisir la gamme de couleur pour chaque variable supplémentaire.

12. Représentation de la variable quantitative supplémentaire sur le 1er plan.

```
> plot(res.mca, choix = "quanti.sup")
```

On peut aussi raffiner les graphiques en utilisant des fonctionnalités supplémentaires de "FactoMineR". Pour vous familiariser avec ces nouvelles fonctionnalités, je vous propose d'exécuter les commandes suivantes :

- afficher les individus dont cosinus carré est plus grand que 0.5 :

  > plot(res.mca, invisible = c("var","quali.sup"), select = "cos2 0.5", unselect = "grey60", cex = 0.7, label = "none")

  L'option `unselect = "grey60"` permet de colorer en gris clair les individus dont le cosinus carré est plus petit que 0.5

- afficher les 1000 individus les mieux représentés sur le 1er plan factoriel :

  > plot(res.mca, invisible = c("var","quali.sup"), select = "cos2 1000", unselect = "green", cex = 0.5, label = "none")

- afficher les 1000 individus qui contribuent le plus :

  > plot(res.mca, invisible = c("var","quali.sup"), select = "contrib 1000", unselect = "green", cex = 0.5, label = "none")

- afficher les modalités dont le cosinus carré est plus grand que 0.4 :

  > plot(res.mca, invisible = c("ind","quali.sup"), selectMod = "cos2 0.4", unselect = "grey60")

- afficher les 10 modalités les mieux représentées :

  > plot(res.mca, invisible = c("ind","quali.sup"), selectMod = "cos2 10", unselect = "green")

- afficher les 10 modalités les plus contributives :

  > plot(res.mca, invisible = c("ind","quali.sup"), selectMod = "contrib 10", unselect = "green")

## 3.2 Réalisation d'une AFCM dans le domaine de l'économie

L'objectif de cette étude est de réaliser une AFCM sur un fichier de données portant sur les organismes de micro-financement, ceci à l'aide du package R *FactoMineR*. 492 organismes ont été observés suivant 18 variables ; 8 qualitatives et 7 quantitatives.

**Description des données**

Les organismes de micro-financement jouent un rôle économique important dans le développement de l'économie locale des pays émergeants. La caractérisation de ces organismes est un enjeu important. Ce tableau de données contient 492 lignes et 18 colonnes. Les 8 premières variables concernent différentes caractéristiques catégorielles ; les variables 9 à 14 portent sur des caractéristiques économiques quantitatives ; la variable quantitative 15 est un indicateur d'efficience (plus cet indicateur est faible, plus l'organisme est efficient). Au final, on observe pour chaque organisme de micro-financement 8 variables catégorielles (numérotées de 1 à 8) et 7 variables quantitatives (numérotées de 9 à 15) :

1. Country
2. Region (Africa / East Asia and the Pacific / Eastern Europe and Central Asia / Latin America and The Caribbean / Middle East and North Africa / South Asia)
3. Age (Mature / Young / New)
4. Current legal status (Bank / Credit-Union-Cooperative / NBFI / NGO / Rural Bank)
5. Financial Intermediation (High FI / Low FI)
6. Profit status (Non-profit / Profit)
7. Regulated (no / yes)
8. Scale (Large / Medium / Small)
9. Assets

10. Personnel expense / assets
11. Number of active borrowers
12. Number of depositors
13. Personnel expense
14. Personnel (number)
15. Efficience (indicateur économique évaluant la bonne santé d'une entreprise)

**Objectif**

L'objectif de cet exercice est de réaliser l'AFCM du tableau "micro_finance.csv" contenant les données portant sur les organismes de micro-financement disponible à l'adresse

***http ://www.math.univ-toulouse.fr/~ferraty/DATA/micro_finance.csv***

Pour cela, suivez méthodiquement les différentes étapes détaillées ci-après.

**Le fichier de données**

1. Ouvrir une session R,
2. Télécharger le fichier "micro_finance.csv" dans un répertoire local,
3. Charger dans R ce tableau de données que vous nommerez MICFIN,
4. Descriptif des variables

```
> summary(MICFIN)
```

5. Exploration des données quantitatives.
   (a) Histogrammes des 7 variables quantitatives initiales

```
> par(mfrow = c(2, 4))
> Varnames = names(MICFIN)
> for(j in 9:15) hist(MICFIN[, j], main =
paste(Varnames[j]), xlab = paste(Varnames[j]))
```

(b) Transformation logarithmique des variables quantitatives initiales

```
> MICFINLOG = cbind(MICFIN[, 1:8], log(MICFIN[, 9:15]))
```

(c) Histogrammes des nouvelles variables quantitatives ainsi transformées :

```
> par(mfrow = c(2, 4))
> for(j in 9:15) hist(MICFINLOG[, j], main =
paste(Varnames[j]), xlab = paste(Varnames[j]))
```

Que remarquez-vous lorsqu'on compare les histogrammes des variables quantitatives transformées avec ceux construits à partir des variables initiales ?

**Question 1**. Quelle est la taille (nombre de lignes et colonnes) du tableau disjonctif complet sur lequel sera fondé l'AFCM ?

### Mise en oeuvre de l'AFCM avec *FactoMineR*

1. Charger la librairie `FactoMineR`
2. Réalisation de l'AFCM. Les 7 variables quantitatives sont obligatoirement déclarées comme variables supplémentaires. On met aussi la variable qualitative 1 (i.e. "Country") qui comporte beaucoup trop de modalités comme variable supplémentaire. L'option "graph = F" produit aucun graphique (par défaut "graph = T").

```
> res.mca = MCA(MICFINLOG, quali.sup = 1, quanti.sup =
9:15, graph = F)
```

L'objet `res.mca` contient les résultats de l'AFCM qui produit aussi différents graphiques que vous pouvez observer. La commande

```
> summary(res.mca)
```

fournit un résumé numérique de l'AFCM.

3. Représentation du nuage des individus sur les 2 premiers axes de l'analyse :

```
> par(mfrow = c(1, 1))
> plot(res.mca, invisible = c("var", "quali.sup"), label =
"none", cex = 1)
```

**Question 2**. Observer ce nuage ; que peut-on dire ?

4. Représentation uniquement des variables actives.

```
> plot(res.mca, choix = "var", invisible = c("quanti.sup",
"quali.sup"))
```

Ce graphique nous permet de repérer les variables qui contribuent le plus à la construction des axes (en terme d'inertie).

**Questions 3**.

(a) Que représente l'abscisse de la variable "Current.legal.status" ?

(b) Que représente l'ordonnée de la variable "Region" ?

(c) Quelles sont les 2 variables qui contribuent le plus à l'axe 1 ?

(d) Quelles sont les 2 variables qui contribuent le plus à l'axe 2 ?

(e) Quelle est la contribution absolue de la variable "Regulated" à l'axe 1 ?

(f) À partir de ce graphique, comment peut-on déduire l'inertie associée à l'axe 1 ? au plan (1, 2) ?

5. Représentation du nuage des modalités sur les 2 premiers axes de l'analyse (les modalités d'une même variable possèdent la même couleur).

```
> plot(res.mca, invisible = c("ind", "quali.sup"), hab =
"quali")
```

**Questions 4**. Afin d'interpréter les axes de cette AFCM, que peut-on dire pour le 1er axe ? Même question pour le 2ème axe.

6. Repérage des individus extrêmes sur les axes de l'analyse.

   **Questions 5**. Afficher les réponses aux 8 premières variables catégorielles données par les 6 individus les plus extrêmes sur l'axe 1 (3 d'un côté et 3 de l'autre) puis comparer les tableaux de réponses obtenus ; que remarquez-vous ? Même question pour l'axe 2.

7. Représentation des modalités de la variable qualitative supplémentaire "Country"

```
> plot(res.mca, invisible = c("ind", "var"), hab= "quali",
cex = 0.75)
```

8. Représentation uniquement des 7 variables quantitatives supplémentaires sur le cercle des corrélations :

```
> plot(res.mca, choix = "quanti.sup")
```

9. Représentation simultanée de toutes les variables (actives et supplémentaires).

```
> plot(res.mca, choix = "var")
```

10. Pourcentages d'inertie.

```
> barplot(res.mca$eig[,2], main = "Pourcentage d'inertie",
names.arg = 1:nrow(res.mca$eig))
```

### Transformation de variables quantitatives en variables catégorielles

1. Transformation des deux variables quantitatives "Assets" et "Personnel.expense" en variables catégorielles afin d'enrichir l'AFCM précédente.

   (a) Rappel des quartiles de ces deux variables :

```
> summary(MICFINLOG$Assets)
> summary(MICFINLOG$Personnel.expense)
```

(b) Transformation de ces deux variables selon leurs quartiles :

```
> Ass_cat = cut(MICFINLOG$Assets, breaks = c(10, 15, 16.1,
17.7, 22), labels = paste("Ass_cat_", 1:4, sep = ""))
> Pers_exp_cat = cut(MICFINLOG$Personnel.expense, breaks =
c(8, 12.3, 13.6, 14.9, 19), labels = paste("Pers_exp_cat_",
1:4, sep = ""))
```

2. Création d'un nouveau tableau de données ; afin de rendre lisible la représentation des variables catégorielles supplémentaires on enlève "Country" :

```
> MICFINLOG_NEW = cbind(MICFIN[, c(-1, 2:8)], Ass_cat,
Pers_exp_cat)
```

3. Réalisation de l'AFCM avec cette nouvelle table :

```
> res.mca.new = MCA(MICFINLOG_NEW, quali.sup = c(8, 9),
graph = F)
```

4. Représentation des modalités des variables supplémentaires :

```
> plot(res.mca.new, invisible = c("ind", "var"), hab =
"quali", palette = palette(c("blue", "red")))
```

**Question 6**. Que pouvez-vous dire en terme d'interprétation ?

## 3.3 Pour conclure sur l'AFCM

- l'AFCM est une méthode factorielle adaptée aux tableaux de type individus $\times$ variables catégorielles,
- 1 individus est au pseudo-barycentre des modalités qu'il possède,
- 1 modalité est au pseudo-barycentre des individus qui la possèdent,
- l'AFCM est une méthode très générale avec des règles d'interprétation très simples,
- l'AFCM est particulièrement adaptée au traitement d'enquête,

- ne pas hésiter à revenir aux données en analysant des tableaux de contingence par AFC pour affiner/confirmer des liaisons suggérées par l'AFCM,
- l'AFCM permet de passer d'un tableau de variables catégorielles à un tableau de variables synthétiques quantitatives (facteurs). Ainsi, l'AFCM peut être vue comme un pré-traitement dans le but de réaliser une classification opérant sur les facteurs qu'elle fournit.