TP2 : Analyse des Correspondences Multiples

1 ACM sur données fictives

Récupérer les jeux de données `chiens.rda`. Il s'agit de données fictives où 27 races de chiens sont décrites avec 7 variables qualitatives.

1. Charger le jeu de données `chiens` dans R avec la commande `load`. Afficher les données. Quelle est la classe de cet objet ?
2. Créez une matrice H contenant la description des $n = 27$ races canines sur uniquement les $p = 6$ premières variables.
3. On veut effectuer l'ACM de cette matrice H.
   (a) Quelle décomposition en valeurs singulières généralisée (GSVD) faut-il faire ? Réaliser cette DSVG avec R.
   (b) Montrer qu'en ACM, l'inertie totale des données vaut toujours $\frac{m}{p} - 1$ où m est le nombre total de modalités et p le nombre de variables qualitatives. Vérifiez ensuite avec R que la somme des valeurs singulières trouvées à la question précédente vaut bien $\frac{m}{p} - 1$.
   (c) Vérifiez également que le nombre maximum de dimension de cette ACM vaut bien $\min(n - 1, m - p)$.
   (d) Représenter dans un diagramme en barre les pourcentages d'inertie expliquée par les dimensions de l'ACM.
   (e) Determiner les matrices X et Y des coordonnées factorielles des races de chiens et des modalités des variables qualitatives sur les $k = 3$ premières dimensions. Modifier les noms des lignes et des colonnes dans X et Y afin qu'ils soient parlants.
   (f) Faire un plot des individus et des modalités dans le premier plan factoriel.
   (g) Utiliser la relation quasi-barycentrique pour retouver les coordonnées factorielles de la modalité T++ à partir des coordonnées factorielles des races de chiens.
   (h) Quels est le rapport de corrélation entre la variable `taille` avec la première composante principale ? Entre la variable `taille` et la seconde composante principale ?

```r
eta2 <- function(x, gpe) {
  moyennes <- tapply(x, gpe, mean)
  effectifs <- tapply(x, gpe, length)
  varinter <- (sum(effectifs * (moyennes - mean(x)) ^ 2))
  vartot <- (var(x) * (length(x) - 1))
  res <- varinter / vartot
  return(res)
}
```

4. On veut maintenant utiliser la fonction `MCA` du package `FactoMineR`.

(a) Faire l'ACM des données sur les races canines en mettant la variable `fonction` en illustratif.

(b) Retrouvez les résulats numériques et les graphiques de la question 2.

(c) Retrouver les rapports de corrélations entre les variables qualitatives et les deux premières composantes principales. Faire le plot des variables en fonction de ces rapports de corrélation en utilisant la fonction `plot.MCA`.

(d) Mettre des données manquantes dans les données avec le code suivant :

(e) Faire l'ACM de `chiensNA`. Comment les données manquantes sont-elles prises en compte dans la fonction `MCA` du package `FactoMineR` ?

5. On veut maintenant comparer l'ACM et l'AFC dans le cas particulier de deux variables qualitatives.

   (a) Avec la fonction `CA` de `FactoMineR`, effectuer l'AFC du tableau de contingence croisant les variables `taille` et `poids`.

   (b) Avec la fonction `MCA`, effectuer l'ACM des deux premières colonnes des données `chiens`.

   (c) Comparez les valeurs propres des deux analyses et vérifiez que vous retrouvez les relations du cours.

2 ACM avec données manquantes et choix du nombre de composantes

Le package R `missMDA` permet de gérer les données manquantes en ACP et en ACM, et de choisir le nombre de composantes par validation croisée. Ce travail est à réaliser en binôme et à me rendre.

1. Regarder les vidéos concernant ce package : `https://www.youtube.com/user/HussonFrancois`
2. Préparer un document avec `Rmardown` qui décrit les principales fonctionnalités de ce package, avec à chaque fois une explication de la méthode, des exemples et du code.

3 ACM et clustering

1. On reprend ici l'exercice 3 du devoir surveillé de décembre 2013.

   (a) Répondre aux questions.

   (b) Retrouver le code R de cet exercice.

2. On veut ici vérfier qu'il est équivalent de réaliser une classification ascendante hiérarchique sur la matrice des distances du χ^2 entre les données brutes recodées dans un tableau disjonctif complet, ou sur la matrice des distances Euclidiennes entre les données décrites avec toutes les composantates de l'ACM. Ce travail est à réaliser en binôme et à me rendre.

   (a) Faire une fonction pour calculer la matrice des distances du χ^2 entre les n lignes d'une matrice de données qualitatives. Appliquez cette fonction à un jeu de données de votre choix.

   (b) Faire l'ACM de ce jeu de données et conserver toutes les composantes principales. Calculer la matrice des distances Euclidiennes entre les n observations décrites par toutes les composantes principales.

   (c) Comparez la hiérarchie de Ward obtenue avec ces deux matrices de distances.