TP5 classification: méthodologie

On utilise dans ce TP le fichier de données « fromages.txt » où 29 fromages sont décrits par 9 variables continues. Pour importer avec R ces données dans une matrice X :

```
> X <- read.table(file.choose(),sep="",header=TRUE,row.names=1)
```

Exercice 1 : Faut-t-il centrer-réduire les données avant de faire la classification ?

Dans cet exercice, il s'agira une classification hiérarchique de Ward avec la distance Euclidienne simple (directement sur X) et d'interpréter cette partition à l'aide d'une ACP non normée (analyse du nuage des individus centrés).

1- Avec la fonction var() de R, calculer les variances des 9 variables. Que pouvez-vous en conclure ?

2- Faire la classification hiérarchique de Ward avec la distance Euclidienne simple. Regarder le dendrogramme et la partition en 3 classes.

2- Avec la fonction PCA du package FactoMineR, faire l'ACP non normée (sur matrice des variances-covariances) de X.

3- Représentez le nuage des 29 fromages dans le premier plan factoriel des individus et le nuage des 9 variables dans le premier plan factoriel des variables. Commentez.

Indication : pour voir les deux graphiques en même temps: > par(mfrow=c(1,2))

4- Habillez les individus du graphique ci-dessus en fonction de la classe de la partition en trois classes à laquelle ils appartiennent.

5- Interprétez à l'aide de ces graphiques les classes de la partition en trois classes obtenue avec la méthode de Ward. En déduire pourquoi il peut-être utile ici de centrer-réduire les données avant de faire la classification.

Exercice 2 : Faut-t-il centrer-réduire les données avant de faire la classification ?

Dans cet exercice, il s'agira de faire une classification hiérarchique de Ward avec la distance Euclidienne normalisée par l'inverse de la variance (directement sur la matrice Z centrée-réduite) et d'interpréter les résultats à l'aide d'une ACP normée (analyse du nuage des individus centrés-réduits).

1- Créer la matrice Z des données centrées-réduites avec la fonction scale() (en utilisant la variance empirique « population » i.e. Non corrigée).

2- Faire la classification hiérarchique de Ward avec la distance Euclidienne normalisée par l'inverse de la variance et visualisez le dendrogramme.

3- Avec la fonction PCA du package FactoMineR, faire l'ACP normée (sur matrice des corrélations) de X.

4- Interprétez la partition en 4 classes issues de la classification de Ward.

Exercice 3 : Faire la classification sur les coordonnées factorielles des individus

On obtient les mêmes partitions si on applique une méthode de classification :

- Sur X ou sur toutes les composantes principales issues de l'ACP non normée,
- Sur Z ou sur toutes les composantes principales issues de l'ACP normée.

On va ici le vérifier dans le second cas avec comme méthode de classification une CAH avec la mesure d'agrégation de Ward.

1- Faire l'ACP normée de la matrice X.

2- Faire la CAH de Ward sur Z et sur la matrice de toutes les composantes principales. Vérifiez que les indices des deux hiérarchies sont identiques et assurez-vous visuellement que les dendrogrammes sont les mêmes.

3- Choisir le nombre q de composantes principales, qui résument « correctement » l'inertie initiale des données. Quel est le pourcentage d'inertie expliquée par ces q premières composantes ?

4- Refaire la CAH de Ward sur ces q premières composantes principales. Comparez aux résultats précédents. Les partitions en 4 classes sont-elles identiques ?