Fiche 9 - TIAD - M1 IM
Classification

Tran Viet Chi, `chi.tran@math.univ-lille1.fr`, bureau 316 (bâtiment M3).

1 Classification sous R

Exercice 1 (Simulations avec R)
1. Créer une fonction qui prend en arguments $n \in \mathbb{N}^*$ et $\delta > 0$, et qui simule 4 échantillons indépendants, chacun composé de n vecteurs aléatoires gaussiens indépendants de $\mathbb{R}^2$, de variance Id, et d'espérance respectivement $(-\delta, -\delta)$, $(-\delta, \delta)$, $(\delta, -\delta)$, (δ, δ). On prendra dans un premier temps $n = 5$ et $\delta = 3$.

2. En utilisant la fonction `agnes` du package `cluster`, réaliser une CAH avec la méthode de Ward. Visualiser le dendogramme associé ainsi que le graphique des partitions obtenues pour différents nombre de classes. Pour cela, on créera une fonction qui affiche le nuage de points et les centres de gravité en utilisant un symbole différent pour chaque classe.

3. Tracer la valeur du rapport variance inter sur variance intra en fonction du nombre de classes.

4. Puisque l'on connaît pour chaque observation la vraie classe à laquelle elle appartient, on peut calculer le pourcentage de concordance entre la prédiction et la vraie partition.

5. Pour $n = 100$ et $\delta = 0, 3$, relancer $N = 100$ fois les simulations (pour la distance de Ward) et donner un estimateur de la densité du pourcentage de concordance. Quel est le pourcentage de concordance moyen ?

6. Reproduire les résultats précédents avec d'autres méthodes que celle de Ward.

7. On choisit de nouveau $n = 5$ et $\delta = 3$. Avec la fonction `kmeans`, réaliser une classification par centre mobile. Reprendre les questions 3, 4, 6.

8. Les iris de Fisher sont un jeu de données classique pour illustrer la classification en statistique. Les données rassemblent 150 individus (les iris) décrits par quatre variables : largeur et longueur des pétales, largeur et longueur des sépales. Ces fleurs se regroupent en trois espèces différentes (Verginica, Setosa et Versicolor), chacune représentée par 50 individus. Sous R, ce jeu de données est disponible en faisant `data(iris)`. Réaliser une CAH et une classification par la méthode des K-moyennes.

9. Visualiser les données dans le premier plan factoriel de l'ACP, en utilisant un symbole différent pour chaque espèce.

10. Réaliser une CAH, en utilisant les projections sur le premier plan factoriel. Comparer.

2 Classification sous SAS

Exercice 2 (Mortalité infantile)
En 1994, les indicateurs démographiques donnaient les taux de mortalité infantile et de mortalité néonatale précoce (pour 1000 naissances vivantes) suivants :

Pays	Infantile	Néonatale
France	5.9	2.3
Allemagne	5.6	2.4
Royaume-Uni	5.9	3.4
Canada (hors Québec)	6.3	3.5
Québec	5.6	3.1
Etats-Unis	8.0	4.2

1. Créer une table SAS contenant ces données.

2. Effectuer une CAH, regarder le dendogramme et discuter le nombre de classes.

3. Visualiser graphiquement dans le plan la partition obtenue.