TP sur l'utilisation de méthodes d'analyse factorielle

Exercice 1 (Températures)

Les données. On s'intéresse au climat des différents pays d'Europe. Pour cela, on a recueilli les températures moyennes mensuelles (en degrés Celsius) pour les principales captiales européennes ainsi que pour certaines grandes villes (fichier temperatures.csv). En plus des températures mensuelles, on donne pour chaque ville, la température moyenne annuelle ainsi que l'amplitude thermique (différence entre la moyenne mensuelle maximum et la moyenne mensuelle minimum d'une ville). On donne également deux variables quantitatives de positionnement (la longitude et la latitude) ainsi qu'une variable qualitative (l'appartenance à une région de l'Europe, variable qualitative à quatre modalités: Europe du nord, du sud, de l'est et de l'ouest).

Problématique. On souhaite appréhender la variabilité des températures mensuelles d'un pays à l'autre de façon multidimensionnelle (*i.e.,* en prenant en compte simultanément les 12 mois de l'année).

Exercice 2 (Jeux Olympiques)

Les données. Le tableau de données croise en lignes les épreuves d'athlétisme et en colonnes les différents pays (données incluses dans le package R FactoMineR sous le nom JO). Chaque cellule contient le nombre total de médailles (or, argent et bronze) obtenues lors ques olympiades de 1992 à 2008 (Barcelone 1992, Atlanta 1996, Sydney 2000, Athènes 2004 et Pékin 2008).

Problématique. On cherche à savoir si il existe des associations "épreuves-pays" remarquables dans un sens (*i.e.,* tel pays n'obtient des médailles que dans telle épreuve) ou dans l'autre (*i.e.,* tel pays n'obtient pas de médailles dans telle épreuve alors qu'il en obtient beaucoup par ailleurs).

Exercice 3 (Crédits à consommation sous R)

Les données. Le jeu de données contient 66 clients ayant souscrit un crédit à la consommation dans un organisme de crédit (fichier `credit.csv`). Les 11 variables qualitatives et les modalités associées à cet exemple sont les suivantes:

- Marché: rénovation d'un bien, voiture, scooter, moto, mobilier-ameublement, side-car. Cette variable indique le bien pour lequel les clients ont réalisé un emprunt.
- Apport: oui,non. Cette variable indique si les clients possèdent un apport personnel avant de réaliser l'emprunt. Un apport personnel représente une garantie pour l'organisme de crédit.
- Impayé: 0, 1 ou 2, 3 et plus. Cette variable indique le nombre d'échéances impayées par le client.
- Taux d'endettement: 1 (faible), 2, 3, 4 (fort). Cette variable indique le niveau d'endettement du client. Le taux d'endettement est calculé comme le rapport entre les charges (ensemble des dépenses) et le revenu. Ce taux a été discrétisé en 4 classes.
- Assurance: sans assurance, AID (assurance invalidité et décès), AID + chômage, Senior (pour les plus de 60 ans). Cette variable indique le type d'assurance à laquelle le client a souscrit.
- Famille: union libre, marié, veuf, célibataire, divorcé.
- Enfants à charge: 0, 1, 2, 3, 4 et plus.
- Logement: propriétaire, accédant à la propriété (personne qui n'a pas encore fini de rembourser son emprunt immobilier), locataire, logé par la famille, logé par l'employeur.
- Profession: ouvrier non qualifié, ouvrier qualifié, retraité, cadre moyen, cadre supérieur.
- Intitulé: M, Mme, Melle.
- Age: 20 (18 à 29 ans), 30 (30 à 39), 40 (40 à 49), 50 (50 à 59), 60 et plus.

Problématique. Le but de cette étude est de caractériser la clientèle de l'organisme de crédit. Nous voulons dans un premier temps mettre en évidence différents profils de comportement bancaires, c'est-à-dire effectuer une typologie des individus. Nous voulons ensuite étudier la liaison entre la signalétique (CSP, âge, etc.) et les principaux facteurs de variabilité des profils de comportement bancaires (i.e. caractériser les clients aux comportements particuliers).

Exercice 4 (Températures (suite))

Les données. On considère à nouveau les données de l'exercice 1.

Problématique. L'objectif est de regrouper les capitales en classes homogènes de sorte que les capitales d'une même classe présentent des températures semblables tous les mois de l'année.

Exercice 5 (Races de chiens)

Les données. Les données présentés décrivent les caractéristiques de 27 races de chiens au moyen de variables qualitatives (fichier `chiens.txt`). Le codage des différentes modalités a la signification suivante:

	Modalité		
Variable	1	2	3
Taille	petite taille	taille moyenne	grande taille
Poids	petit poids	poids moyen	poids élevé
Vélocité	lent	assez rapide	très rapide
Intelligence	peu intelligent	intelligence moyenne	très intelligent
Affection	peu affectueux	affectueux	
Agressivité	peu agressif	agressif	
Fonction	chien de compagnie	chien de chasse	utilité

Table 1: Codage des différentes modalités pour les races de chiens.

Problématique. Effectuer puis visualiser une classification des données qualitatives `races de chiens`

Exercice 6 (Limites des K-means)

1. Télécharger le fichier `generedatasets.R` qui contient 4 fonctions permettant de générer différents jeux de données.
2. Générer les 4 jeux de données correspondants, en prenant n = 200. Tracer les nuages de points obtenus.
3. Sur chaque jeu de données, lancer l'algorithme des k-means à l'aide de la fonction kmeans en justifiant le choix du nombre de clusters. Dans un premier temps, laisser les valeurs par défaut pour les options de la fonction.
4. En comparant les résultats obtenus sur les différents jeux de données, pouvez-vous iden- tifier le type de distributions pour lesquelles l'algorithme des k-means n'est pas adapté ?
5. Comparer les résultats précédents avec ceux obtenus par classification ascendante hié- rarchique. Comparer les résultats obtenus avec différents critères d'aggrégation.
6. Les deux algorithmes (k-means et CAH) se comportent-ils de la même façon sur ces jeux de données ?

Exercice 7 (Segmentation d'image)

1. Télécharger le fichier `irm_small.jpeg` et l'importer sous R en utilisant la fonction `readJPEG` du package `jpeg`.
2. Afficher l'image à l'aide de la fonction `image`.

3. Appliquer l'algorithme de partitionnement spectral normalisé afin d'identifier différentes zones dans l'image. On utilisera la fonction de similarité Gaussienne et le graphe du ε-voisinage en choisissant un seuil ε égal au quantile d'ordre 75% des indices de similarité. On justifiera le nombre de classes sélectionnées.

4. Afficher les classes obtenues.

Exercice 8 (Modèles de mélange)
On considère que $X = [X_1, \ldots, X_d]$ est une variable aléatoire définie par la densité suivante

$$p(x;\theta) = \sum_{k=1}^{2} \pi_k \prod_{j=1}^{d} \phi(x_j; \mu_{kj}, \sigma_{kj}^2)$$

avec $\pi_1 = \pi_2 = 1/2$, $\sigma_{1j} = 1$, $\mu_{1j} = 0$,

$$\mu_{2j} = \begin{cases} 2.5 & \text{si } j < r \\ 0 & \text{sinon} \end{cases} \quad \text{et } \sigma_{2j} = \begin{cases} 2 & \text{si } j < r \\ 0 & \text{sinon} \end{cases}$$

1. Écrire la fonction `generdata` qui prend pour argument la taille de l'échantillon n, le nombre r et le nombre de variables d. Cette fonction retourne l'échantillon généré ainsi que la vraie partition.

2. On souhaite étudier l'importance de la sélection de variables en clustering. Pour cela, on calcule l'ARI entre la vraie partition et les partitions obtenues avec et sans sélection de variables par le package `VarSelLCM`, en considèrant 20 réplicats générés avec $n = 100$, $r = 3$ et $d \in \{3, 6, 20, 50\}$.

3. Modifier la fonction `generdata` pour que les donnée possèdent un taux τ% de valeurs manquantes (MCAR). La fonction retourne maintenant 3 éléments: l'échantillon sans valeurs manquantes, l'échantillon avec valeurs manquantes et la partition.

4. On souhaite étudier l'intérêt des mélanges lorsque les observations présentent des valeurs manquantes. Pour cela on compare les deux approches suivantes:

   - Imputation des valeurs manquantes par la fonction `imputePCA` puis K-means sur les données imputées.
   - Utilisation des mélanges pour l'imputation et le clustering.

   On considère le cas où $n = 100$ et $r = 3$. Montrez l'évolution de la qualité de la partition lorsque τ augmente pour $d \in \{3, 6, 20, 50\}$.