ANALYSE DE DONNÉES
TRAVAUX DIRIGÉS ET PRATIQUES

Fiche n^{o}2 : Analyse Factorielle

PARTIE I : Analyse en Composantes Principales

Exercice 1

On considère dans $\mathbb{R}^2$ les points suivants :

$$A = \begin{pmatrix} 0 \\ 4 \end{pmatrix}, \ B = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \ C = \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \ D = \begin{pmatrix} 1 \\ 3 \end{pmatrix}.$$

1) Représenter ces quatre points dans un repère orthonormé.

2) Déterminer le barycentre g de ces quatre points.

3) On réalise un changement de repère qui consiste à prendre g comme nouveau centre.
- Exprimer les nouvelles coordonnées des points dans le nouveau repère.
- Quelle opération géométrique représente ce changement de repère ? Représenter sur le dessin le nouveau repère.

4) Considérons dans le nouveau repère le sous-espace vectoriel engendré par le vecteur $u = (1, 1)^t$. Que représente ce sous-espace vectoriel ? Représenter ce dernier sur le graphique.

5) Calculer pour chacun des points la coordonnée de son projeté orthogonal sur ce sous-espace et le représenter.

Exercice 2

On a fait subir à dix étudiants trois épreuves d'endurance physique ; les notes sur 10 obtenues sont données dans le tableau ci-dessous, sous les noms respectifs X, Y et Z.

Prénom	X	Y	Z
Abdel	2	4	8
Benoit	5	6	10
Charles	5	5	8
Driss	2	5	6
Eric	8	5	10
Florian	5	5	8
Gaëtan	8	6	8
Huang	8	6	10
Iwan	5	4	6
John	2	4	6

1) Représenter le nuage de points associé à ces données dans le plan (X, Y). Que remarquez-vous ?

2) Calculer les valeurs moyennes des variables X, Y et Z. En déduire la table des données centrées.

3) Calculer, sans calculatrice, les variances et les écart-type des variables X, Y et Z (exprimer les résultats en fonction de 0.6 et de $\sqrt{0.6}$). En déduire la table des données centrées et réduites, toujours exprimée en fonction de $\sqrt{0.6}$.

4) Montrer que la matrice des corrélations Σ des variables X, Y et Z est exactement égale à

$$\Sigma = \begin{pmatrix} 1 & 2/3 & 2/3 \\ 2/3 & 1 & 2/3 \\ 2/3 & 2/3 & 1 \end{pmatrix}$$

5) On admet que la plus grande valeur propre de Σ vaut $\lambda = 7/3$. Donner un vecteur propre associé. En déduire les coordonnées du premier vecteur principal.

6) Proposer une variable statistique F_1, combinaison linéaire des variables centrées et réduites, possédant une variance $> 3/2$. Exprimer cette nouvelle variable F_1 en fonction de X, Y, Z. Quelle serait la note F_1 d'un étudiant qui aurait obtenu 8 , 9 et 10 aux épreuves ?

Exercice 3

Une étude gastronomique donne les appréciations sur le service, la qualité et le prix de 4 restaurants. Ces appréciations sont notées de -3 (mauvais) à 3 (bon). Les résultats sont donnés dans le tableau suivant :

Restaurant	**Service**	**Qualité**	**Prix**
R_1	-2	3	-1
R_2	-1	1	0
R_3	2	-1	-1
R_4	1	-3	2

Les calculs des matrices des variances/covariances V, et des corrélations Σ, donnent les résultats suivants :

$$V = \begin{pmatrix} 5/2 & -3 & 1/2 \\ -3 & 5 & -2 \\ 1/2 & -2 & 3/2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & -0,85 & 0,26 \\ -0,85 & 1 & -0,73 \\ 0,26 & -0,73 & 1 \end{pmatrix}$$

1) Commenter le tableau et les matrices ci-dessus.

2) Calculer l'inertie totale du nuage de points.

3) On calcule les valeurs propres de la matrice V. On obtient $\lambda_1 = 7,625$ et $\lambda_3 = 0$. En déduire λ_2.

4) On calcule les vecteurs propres associés aux deux premières valeurs propres. On obtient

$$u_1 = \begin{pmatrix} 0,5 \\ -0,8 \\ 0,3 \end{pmatrix} \qquad u_2 = \begin{pmatrix} 0,65 \\ 0,11 \\ -0,75 \end{pmatrix}$$

Calculer les composantes principales.

5) Représenter les individus dans le premier plan factoriel.

6) Calculer les corrélations entre les variables et les facteurs.

7) Représenter les variables sur le cercle des corrélations projeté sur le premier plan factoriel.

8) Interpréter les résultats obtenus.

9) Calculer les pourcentages d'inertie de chaque facteur, ainsi que les pourcentages d'inertie cumulée.

10) Tracer l'ébouli des valeurs propres. Combien de facteurs est-il convenable de retenir ?

Exercice 4

Cet exercice sera effectué sous le logiciel R.

On cherche à décrire les cantons suisses aux alentours de l'année 1888 par : mesure de la fertilité (Fertility), pourcentage d'agriculteurs (Agriculture), pourcentage d'appelés ayant obtenu les meilleurs résultats lors de la visite médicale (Examination), pourcentage d'appelés ayant continué leur scolarité après l'école primaire (Education), pourcentage de catholiques (en opposition à protestants) (Catholic), et mortalité infantile (Infant.Mortality).
Charger les données dans `R` à l'aide de la commande :

```
> data(swiss)
```

1) Faire l'étude descriptive variable par variable de la table de données `swiss` (nom et type des variables, indicateurs centraux et de dispersion, boîtes à moustaches, scatterplot).

2) Effectuer une ACP sur la table `swiss`. Interpréter les résultats obtenus (interprétation des axes factoriels, description des individus, repérage des individus atypiques).

3) Tracer l'ébouli des valeurs propres.

4) Choisir le nombre de facteurs à retenir pour réduire la dimension des données en tenant compte de l'ébouli des valeurs propres, de la règle de Kaiser, ainsi que du pourcentage cumulé de variance expliquée.

Exercice 5

Cet exercice sera effectué sous le logiciel R.

On analyse les résultats du décathlon masculin des jeux olympiques et du Decastar de 2004. Les variables sont : 100 m, saut en longueur, poids, saut en hauteur, 400 m, 110 m haies, disque, perche, javelot, 1500 m, rang, nombre de points obtenus, évènement sportif (jeux olympiques ou Decastar). Charger les données dans `RStudio` à l'aide de la commande :

```
> data(decathlon)
```

1) Après les avoir nettoyées, effectuer une ACP sur les données de décathlon.
Indication : on peut traiter le rang, le nombre de points et l'évènement sportif comme des variables illustratives et les projeter sur le premier plan factoriel.

2) Choisir le nombre de facteurs à retenir afin de décrire ces données et interprèter les rèsultats obtenus.

PARTIE II : Analyse Factorielle des Correspondances

Exercice 6

On reprend les données de recensement aux Etats-Unis en 2012. Le croisement des variables *Statut marital du chef de foyer* et *Revenus du foyer* donne la répartition suivante :

Revenus / **Statut**	< 40	$\geqslant 100$	40 - 60	60 - 100	**Total**
Célibataire ou veuf	89	35	41	34	199
Divorcé ou séparé	35	10	14	16	75
Marié	59	80	76	110	325
Total	183	125	131	160	599

1) Donner les profils moyens (ligne et colonne) du tableau de contingence.

2) Construire les profils lignes et colonnes du tableau de contingence.

3) On souhaite faire une AFC sur le tableau de contingence. Quel est le nombre maximal de facteurs possibles dans cette analyse ? Sur quels profils l'ACP avec distance du χ^2 sera-t-elle appliquée ?

4) Les valeurs propres obtenues en diagonalisant la matrice des distances entre profils sont les suivantes : $\lambda_1 = 0,09$, $\lambda_2 = 0,002$. Calculer les pourcentages d'inertie expliquée par chaque facteur. Que peut-on en conclure ? Proposer un test afin d'appuyer cette conclusion.

5) Les contributions des modalités lignes et colonnes sont données dans le tableau suivant :

Axe / Modalité	C_1	C_2
Célibataire ou veuf	38,1	28,7
Divorcé ou séparé	16,2	71,3
Marié	45,7	0,01

Axe / Modalité	C_1	C_2
< 40	65,3	2
$\geqslant 100$	8,8	42,3
40 - 60	1,4	8,5
60 - 100	24,5	47,3

Commenter et interpréter.

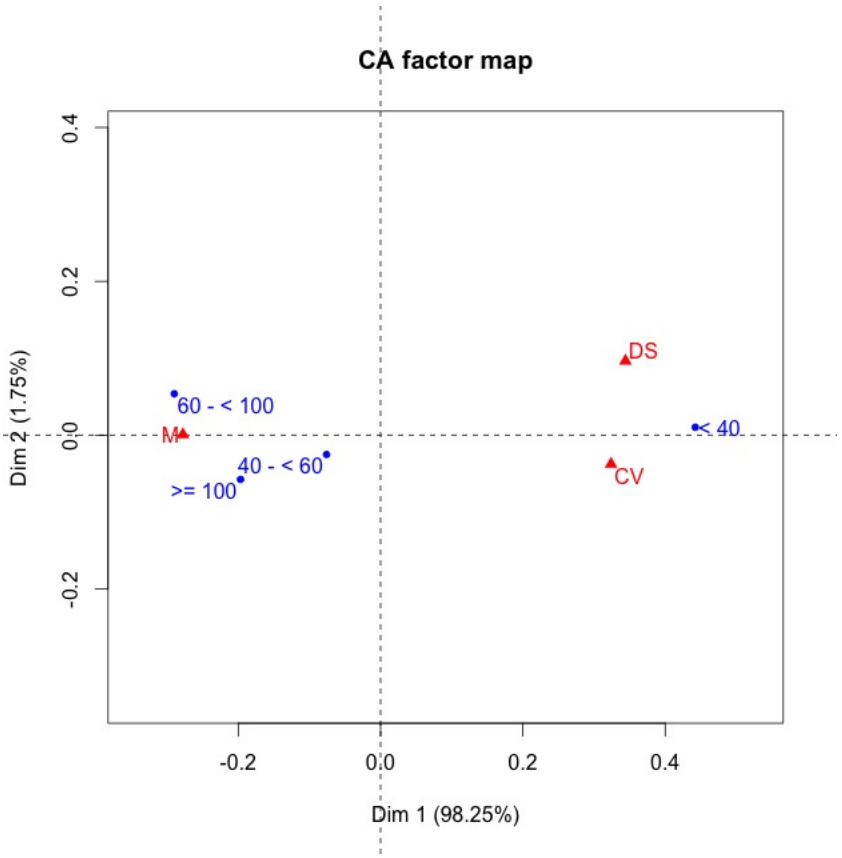
6) Les valeurs des cosinus carrés des modalités lignes et colonnes sont données dans le tableau suivant :

Axe / Modalité	C_1	C_2
Célibataire ou veuf	0,99	0,01
Divorcé ou séparé	0,93	0,07
Marié	1	0

Axe / Modalité	C_1	C_2
< 40	1	0
$\geqslant 100$	0,92	0,08
40 - 60	0,9	0,1
60 - 100	0,97	0,03

Commenter et interpréter.

7) La répartition croisée des modalités lignes et colonnes dans le premier plan factoriel est la suivante :



Commenter et interpréter. Quelles informations nous a apportées cette analyse ?

Exercice 7

Cet exercice sera effectué sous le logiciel R.

On étudie la relation potentielle entre la catégorie socio-professionnelle (CSP) et la principale source de renseignement sur les problèmes d'environnement. Pour cela, un échantillon de 1283 personnes a été interrogé.

Chaque individu interrogé devait renseigner sa CSP parmi 7 choix : agriculteur (AGRI), cadre supérieur (CSUP), cadre moyen (CMOY), employé (EMPL), ouvrier (OUVR), retraité (RETR) ou chômeur (CHOM).

Il devait également renseigner sa principale source de renseignement sur les problèmes d'environnement parmi 6 choix : télévision (TEL), journaux (JOU), radio (RAD), livres (LIV), associations (ASS) ou mairie (MAI).

Les données se trouvent dans le fichier `media.csv`.

1) Importer les données sous `RStudio`.

2) Effectuer l'AFC sur le tableau de contingence (on prendra bien soin de préciser le nombre maximal de facteurs possibles).

3) Afficher le tableau des valeurs propres et des pourcentages d'inertie expliquée.
Combien de facteurs proposez-vous de garder pour l'analyse ?

4) Calculer la moyenne des valeurs propres et tracer l'ébouli des valeurs propres. Confirmer ou infirmer le choix fait à la question précédente.

5) Afficher les coordonnées, les contributions et les cosinus carrés des profils lignes et colonnes.

6) Dans toute la suite, on ne considérera que les axes retenus à la question **4)**.

a- Quelles sont les modalités contribuant à la construction des axes retenus ? Préciser le signe de leurs coordonnées sur ces axes.

b- Quelles sont les modalités les mieux représentées sur les axes retenus ?

c- Déduire des questions précédentes s'il existe une ou plusieurs modalités atypiques dans ce jeu de données.

7) Faire la synthèse des analyses précédentes en donnant une interprétation des axes factoriels et une description des données à l'aide des plans factoriels.

Exercice 8 ACP vs AFC

Cet exercice sera effectué sous le logiciel R.

Des données sur 23 papillons ont été recueillies afin de les étudier suivant 4 caractéristiques notées Z_1, Z_2, Z_3, Z_4. Ces variables étant quantitatives discrètes, elles peuvent être traitées comme des variables numériques via une ACP, ou bien comme un tableau de contingence (avec une modalité par papillon et par caractéristique) via une AFC.

Le but de l'exercice est de comparer les deux analyses. Les données se trouvent dans le fichier `Papillon.csv`.

1) Effectuer une ACP sur le tableau de données et interpréter les résultats.

2) Effectuer une AFC sur le tableau de données et interpréter les résultats.

3) Comparer les résultats des deux études. Quel comportement illustre cet exemple ?

Rappel : aide à l'interprétation des résultats

— Déterminer le nombre d'axes à retenir afin de faciliter l'étude.
— Décrire les axes retenus à l'aide du cercle des corrélations (uniquement pour l'ACP) et/ou des contributions et cosinus carrés.
— Visualiser les données sur les 1 à 3 premiers plans factoriels suivant le nombre d'axes retenus.
— Détecter et traiter les individus, variables ou modalités atypiques.
— Décrire et synthétiser les analyses.