


Module : Analyse et fouille de données	
Responsables du Cours: Bouaziz Souhir, Abbès Amal Enseignants TP: Barhoumi Chawki, Rekik Amal, Njeh Maïssa	Auditoire: D-LSI-ADBD A-U: 2023-2024
TP2 : Analyse en Composantes Principales (ACP)	

Objectifs du TP

Ce TP vise à :

- Comprendre l'Analyse en Composantes Principales (ACP) et l'utiliser en se basant sur les fonctionnalités de la bibliothèque Scikit-learn
- Savoir fournir et interpréter la matrice de corrélations
- Savoir interpréter et calculer les valeurs et les vecteurs propres en utilisant les fonctions python
- Analyser et tracer l'épandage des valeurs propres et le cercle de corrélation
- Représenter et analyser les variables dans le plan factoriel

1. Description des données

Il s'agit d'analyser les données dans le fichier [seeds.csv](#). Celles-ci présentent des mesures des propriétés géométriques de grains appartenant à trois variétés différentes de blé : Kama, Rosa et Canadian. Chaque ligne du fichier décrit un grain. Sept variables caractéristiques sont utilisées à savoir :

1. Area A: surface du grain
2. Perimetre P: périmètre du grain
3. compactness : $C = 4 \cdot \pi \cdot A / P^2$
4. klength : longueur du noyau du grain (kernel)
5. kwidth : largeur du noyau du grain
6. AsymmetryCoef : coefficient d'asymétrie
7. kGrooveL : longueur du rainure du noyau

2. Travail à faire

(a) Préparation des données

1. Importer le jeu de données **seeds.csv** dans un DataFrame **X**
2. Sélectionner les colonnes nécessaires dans l'objectif d'analyser les dépendances entre les grains et la recherche des similarités. (Modification sur X)
3. Remplacer les valeurs manquantes dans chaque colonne par la moyenne de la variable
4. Standardiser les données en utilisant la classe **StandardScaler** de la bibliothèque **sklearn**.
5. Vérifiez les moyennes et les écarts types après la standardisation.

(b) Visualisation et analyse des données

6. Afficher et analyser la matrice de corrélation
7. Quels sont les couples de variables les plus corrélées.
8. Visualiser et analyser les dépendances des variables 2 à 2 en faisant le lien avec la matrice de corrélation.

(c) Application d'ACP

9. Réaliser sur Z une ACP normée en utilisant la méthode **pca** du module **sklearn.decomposition**
10. Quel est le nombre de composantes calculé par l'ACP construite
11. Calculer et afficher les coordonnées factorielles (matrice des scores des individus)
12. Afficher la matrice de corrélation des nouvelles composantes
13. Calculer de 2 façons les valeurs propres et les analyser. S'assurer que la somme des valeurs propres associées à toutes les composantes est égale au nombre de composantes ($k=7$)
14. Tracer l'éboulis des valeurs propres (scree plot)
15. Calculer la proportion des variances (parts d'inertie) expliquées par chaque Composante principale (CP)
16. Tracer le graphique des variances cumulées expliquées. Commenter.
17. Quelles sont les composantes à retenir afin de conserver 80% de l'information.
18. Projeter les individus dans le premier plan factoriel et proposer une typologie des individus
19. Afficher et analyser la matrice des vecteurs propres Q
20. Calculer la matrice de corrélation des anciennes variables (Z_j) et des nouvelles (Y_k) du plan factoriel.
21. Analyser la saturation des variables en projetant les variables (Z_j) sur le cercle de corrélation C

3. Exercice à rendre

Refaire le même travail sur les données fournies dans le fichier **image.csv**

Bon travail