Faculté des sciences **Année : 2020/2021**

1ère année Master Informatique (ASD)

Matière : Statistiques exploratoires

Mini-projet de TP :

Réalisation d'une méthode de réduction de dimension basée sur l'Analyse en Composantes Principales

Description du mini-projet :

Tout système complexe de prédiction ou de reconnaissance implique une phase de l'analyse de données manipulées. Ces données sont structurées dans des bases de données. En effet, les bases de données en général définies par des tableaux rectangulaires à deux entrées, ligne par ligne ou colonne par colonne, dont les lignes correspondent à des individus et les colonnes à des variables appelées caractères, caractérisant les données. Ces deux entrées (ou deux espaces) : individus et variables peuvent être deux dimensions importantes, ce qui peut poser un problème lors du stockage, de l'exploration et de l'analyse de données. Pour cela, il est important de mettre en place des outils automatiques de traitement permettant l'extraction des connaissances sous-jacentes.

L'extraction des connaissances s'effectue selon deux approches, la catégorisation des données (par regroupement en classes) et/ou la réduction de la dimension de l'espace de représentation de ces données.

La réduction de la dimension se pose comme une étape de prétraitement des données. En effet, pour des données appartenant à un espace de grande dimension, certaines variables n'apportent aucune information prédictive voire expriment du bruit, d'autres sont redondantes ou corrélées. Ceci rend les algorithmes de prédiction ou de reconnaissance complexes, inefficaces, moins généralisables et d'interprétation délicate. Les méthodes de réduction de la dimension servent principalement à :

- Réduire le volume de données à traiter, tout en conservant au mieux l'information utile.
- Supprimer les variables non pertinentes.
- Identifier les relations (les ressemblances/différences) entre les individus.
- Mettre en évidence et comprendre les relations (corrélations/anti-corrélation) entre les variables ou les groupes de variables.

Cependant, nous distinguons les méthodes de sélection de variables et celles de la transformation de variables. Ces dernières consistent à construire de nouvelles variables à partir des variables initiales.

Dans le cadre de ce travail, nous nous intéressons aux transformations de variables, particulièrement l'Analyse en Composantes Principales (ACP).

L'ACP est une méthode d'analyse exploratoire des données : à partir d'un ensemble d'observations caractérisées par un ensemble de variables quantitatives initiales, on cherche à condenser la représentation des données en conservant au mieux leur organisation globale. Donc, on obtient, comme résultat, de nouvelles variables (les composantes principales) où elles représentent des combinaisons linéaires des variables initiales, en conservant le plus de variance possible des projections.

Dans la figure 1, deux variantes de l'ACP sont présentées : l'ACP centrée et l'ACP normée.
Le processus de l'ACP est résumé comme suit :

1. Lire les données de départ (représentées par une matrice où chaque ligne correspond à un individu (observation) et chaque colonne à une variable initiale.

2.1 Pour l'ACP centrée, calculer la matrice centrée des données initiales. Cette variante de l'ACP est appliquée lorsque les variables initiales sont directement comparables (de même nature, intervalles de variation comparables).

2.2 Pour l'ACP normée, calculer la matrice centrée et réduite des données initiales. La deuxième variante de l'ACP est utilisée lorsque les variables sont de nature différente ou présentent des intervalles de variation très différente.

3.1 Calculer la matrice de variance/covariances des données centrées (ACP centrée).

3.2 Calculer la matrice de corrélation pour les données centrées et réduites (ACP normée).

4. Diagonaliser la matrice variance/covariante ou la matrice de corrélation i.e. chercher les valeurs propres.

5. Calculer les composantes (axes factoriels) principales.

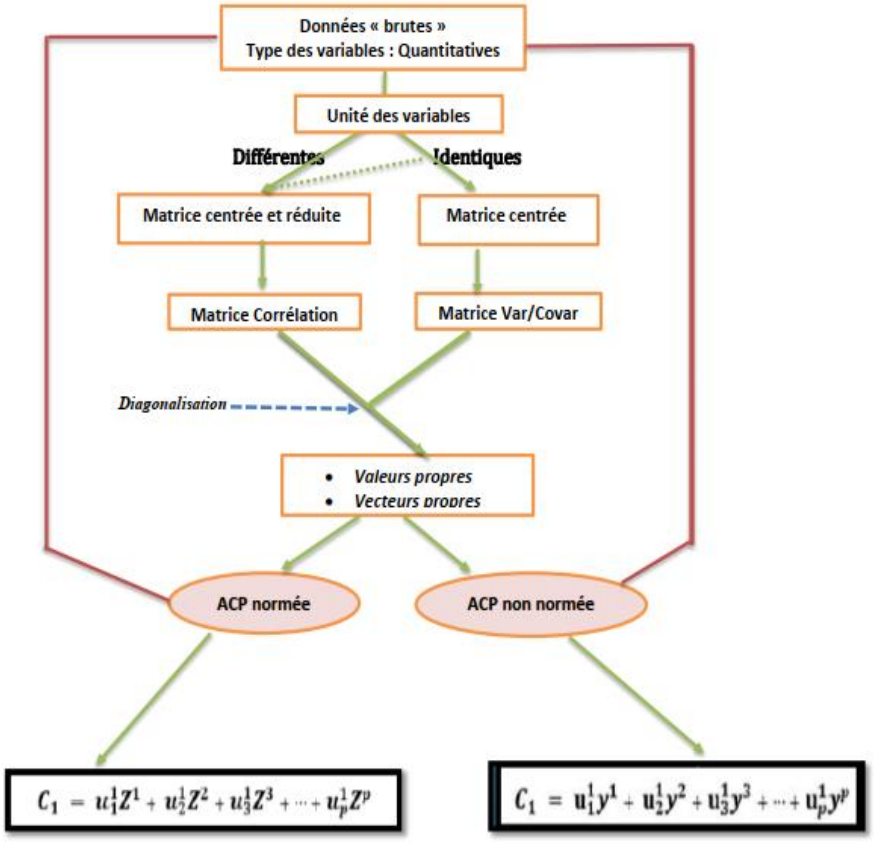

Figure 1. Analyse en Composantes Principales

Travail demandé :

Le but de ce mini-projet est de développer les deux variantes de l'ACP (centrée et normée) permettant, à travers une interface graphique, d'introduire le dataset et d'afficher, après les étapes de traitement, l'histogramme des valeurs propres et les deux graphiques des projections du nuage des individus et celui des variables sur les composantes principales. Les datasets utilisées sont publiés sur le lien suivant : (https://canvas.instructure.com/courses/2504960).

Le tableau 1 donne les détails des opérations à développer :

Phase	Méthode à utiliser	Implementation
Détection et prétraitement des données bruitées (valeurs null, valeurs invalides, ou liers, les données redondantes,…etc.)	Au choix (vue dans la partie 2 –support de TP)	Aucune bibliothèque autorisée
Analyse en composantes principales des données pré-traitées	✓ ACP centrée ✓ ACP normée	Aucune bibliothèque autorisée
Projection du nuage des individus et projection du nuage des variables sur les k composantes principales à choisir	/	Aucune bibliothèque autorisée
Interpréter les deux types de projection	/	/

Tableaux 1. Détails techniques des phases de la réduction de dimension

Important

- Chaque phase de la réduction de dimension basée ACP doit se faire indépendamment.

- L'interface graphique doit comporter un champ pour introduire le dataset, une zone pour afficher l'histogramme des valeurs propres et les deux graphiques des projections du nuage des individus et celui des variables sur les composantes principales à retenir.

- Chaque **trinôme** devra remettre :

  - Le code source Python commenté et un fichier exécutable JAR.
  - Un rapport d'environs 10 pages détaillants toutes les étapes de réalisation du mini-projet, et ce, **avant le 28 Mars 2021**.

Langage de développement : Python

Bon courage