Thématique 5 - Analyse multivariée et analyse en composantes principales

Data Sciences Revision

Mention Mathématiques et Informatique

Parcours HPDA

02/10/2024

Rédigé par :
PAULY Alexandre
alexandre.pauly@cy-tech.fr

L'objectif d'une analyse en composantes principales est de projeter des individus d'un espace d'une certaine dimension dans une dimension moindre tout en présenvant la corrélation des données afin de réduire le nombre de variables à analyser. L'ACP est une méthode linéaire.

Remarque : L'ACP est une des nombreuses méthodes de Machine Learning pour réduire les dimensions d'un ensemble de données, mais il en existe d'autres comme LDA et t-SNE (méthode non linéaire).

1 Pre-processing

Lors de la phase préparatoire à l'analyse des données, l'une des premières étapes, et probablement la plus importante avant de commencer, consiste à calculer certains indicateurs permettant de mettre en lumière la répartition du jeu de données : effectif, moyenne, variance, médiane, Q1 et Q3.

Ces indicateurs peuvent être restitués sur une boîte à moustache pour offrir une meilleure représentation d'une variable. Il est plus intéressant de représenter sur un même graphique les différentes modalités d'une variable qualitative quand cela est possible.

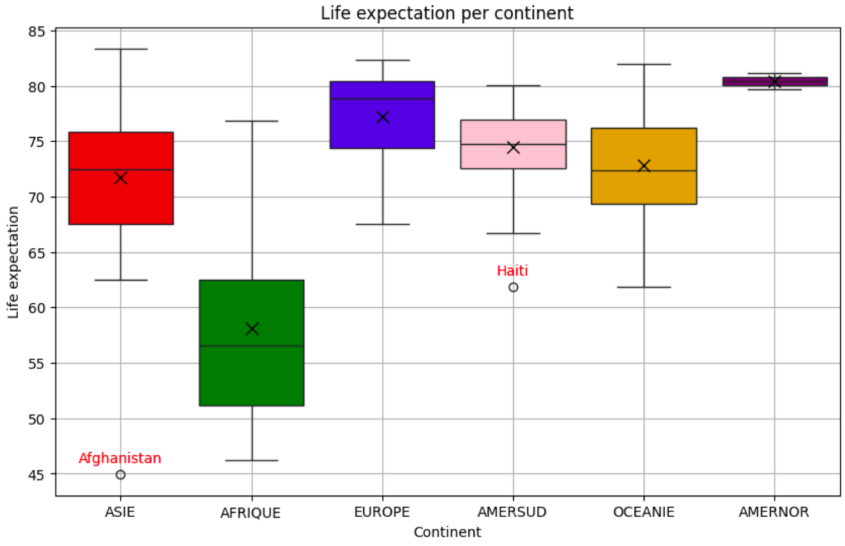

Figure 1: Boîtes à moustaches pour chaque modalité d'une variable

Mais dans un tel contexte où l'on souhaite comparer chaque variable pour connaître l'impact de chacune, il est intéressant d'afficher les nuages de points 2 à 2 des variables.

Remarque : Centrer et réduire les données n'impactera pas la visualisation du plot, seulement les échelles de grandeur.

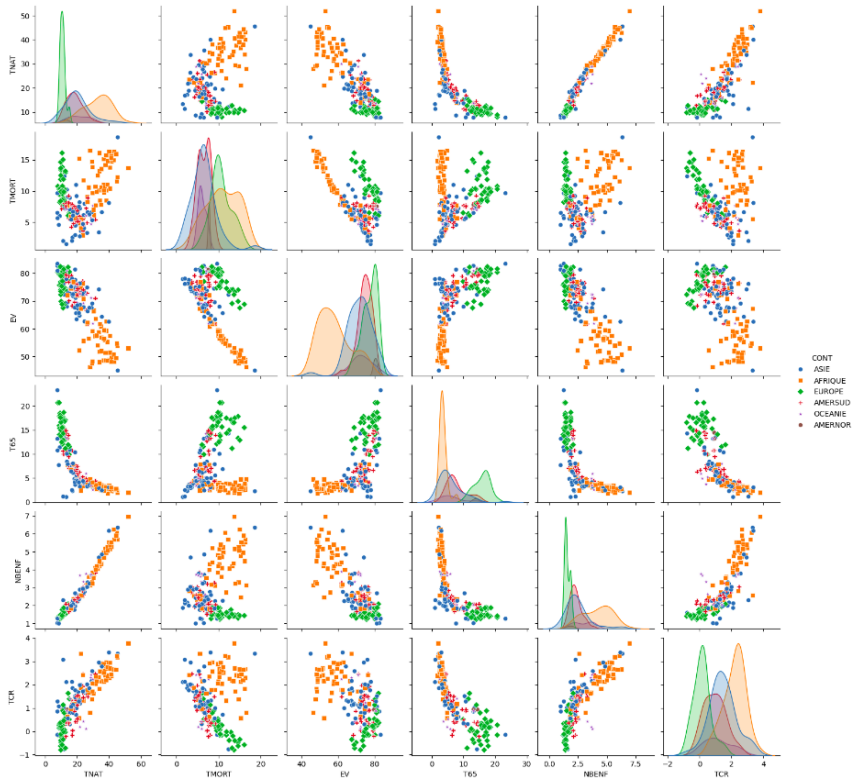

Figure 2: Nuage de points 2 à 2

2 Analyse en composantes principales

L'une des premières étapes lorsque l'on applique une ACP est de centrer et réduire les variables car l'ACP nécessite de calculer les distances entre observations. Sauf que si les ariables n'ont pas le même ordre de grandeur, certaines variables à valeurs faibles « disparaitrons » de l'information au profit de celles ayant de fortes valeurs.

De la même façon la quantification de l'information au travers de l'inertie, privilégie les variables fortement disperséses. D'où l'importance de centrer et réduire les variables.

2.1 Réduction de dimension

Le principe de l'ACP est de trouver des espaces de petites dimensions sur lesquels les projections des observations minimisent la déformation de la réalité. On cherche donc un sous-espace F_q de R^p de dimension q (q=2,3,..) sur lequel projeter le nuage de points. Les axes de ce sous-espace sont des combinaisons linéaires des axes d'origine (c.-à-d. les variables). Les nouveaux axes s'appellent les composantes principales.

Une fois les composantes définies, nous allons nous interroger sur le nombre de composantes à retenir.

2.2 Vecteurs propres et contribution

Pour déterminer le nombre d'axes à garder, nous allons comparer leur contribution. Pour cela, certains graphiques mettent facilement les résultats en valeur tel que l'histogramme suivant. Il permet de visualiser la part de chaque composante une à une ou alors des variances cumulées.

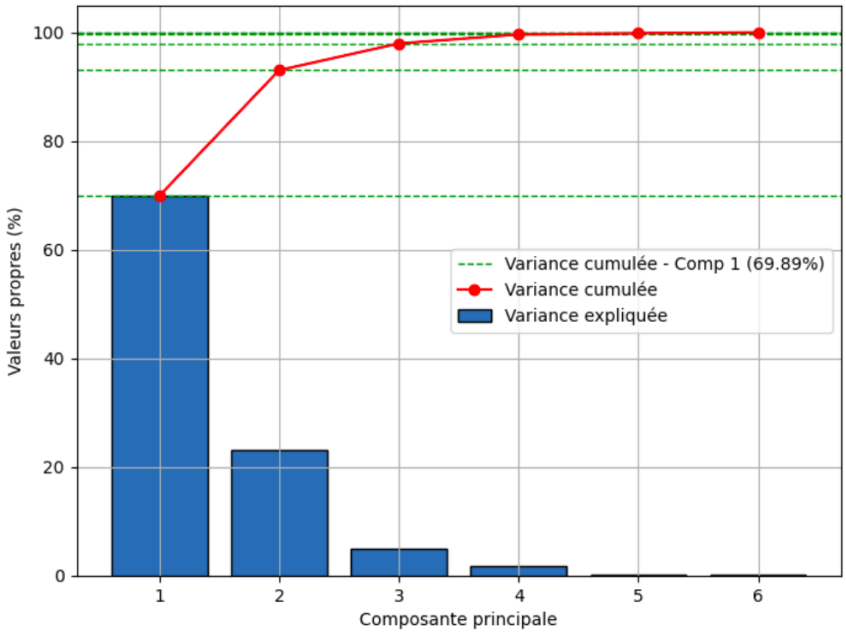

Figure 3: Diagramme des valeurs propres avec variance cumulée

A savoir qu'il y a deux règles pour le choix du nombre d'axes :

- Garder un maximum d'information contenu dans ces axes (pourcentage cumulé d'inertie)
- Couper sur le dernier grand saut d'information entre les axes (elbow rule)

Remarque : On adapte en fonction du nombre d'axes que l'on souhaite, mais en général on essaie de se ramener à 2. Surtout dans des cas comme l'exemple ci-dessus ou 2 composantes expliquent plus de 70% de la variance.

Afin de représenter les observations dans les axes gardées, il est commun d'afficher un nuage de points comme nous le faisons habituellement.

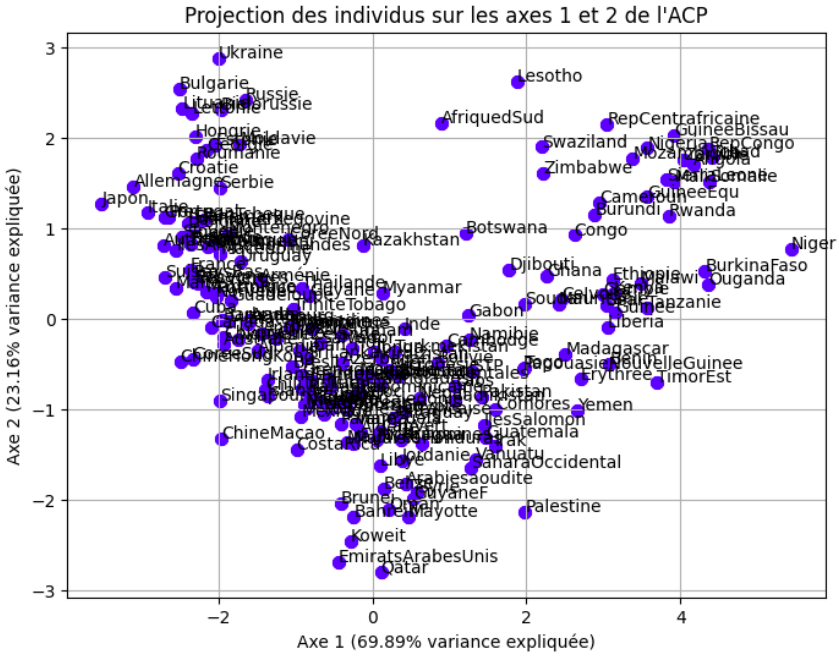

Figure 4: Projection des individus sur les axes 1 et 2 de l'ACP

Mais pour analyser plus précisément la corrélation des variables sur les axes en question, le cercle de corrélation est bien plus pratique. Il s'analyse de la façon suivante :

- Les variables sont bien représentées si elles sont proches du cercle. A contrario celles qui sont proches de l'origine sont peu corrélées avec les axes (pas d'interprétation possible pour ces variables).
- Les individus sont bien représentés s'ils ne sont pas trop éloignés de l'axe sur lequel on les projette (vérifier le cosinus entre l'individu et l'axe : proche de 1).
- Lorsque 2 axes sont perpendiculaires, ils ne sont pas corrélés (équivalent à un coeff de corrélation de 0).
- Lorsque 2 axes sont parallalèles (ou quasi), ils sont fortement corrélés positivement ensemble (coeff de corrélation de 1).
- Lorsque 2 axes sont aux opposés, ils sont fortement corrélés négativement ensemble (coeff de corrélation de -1).

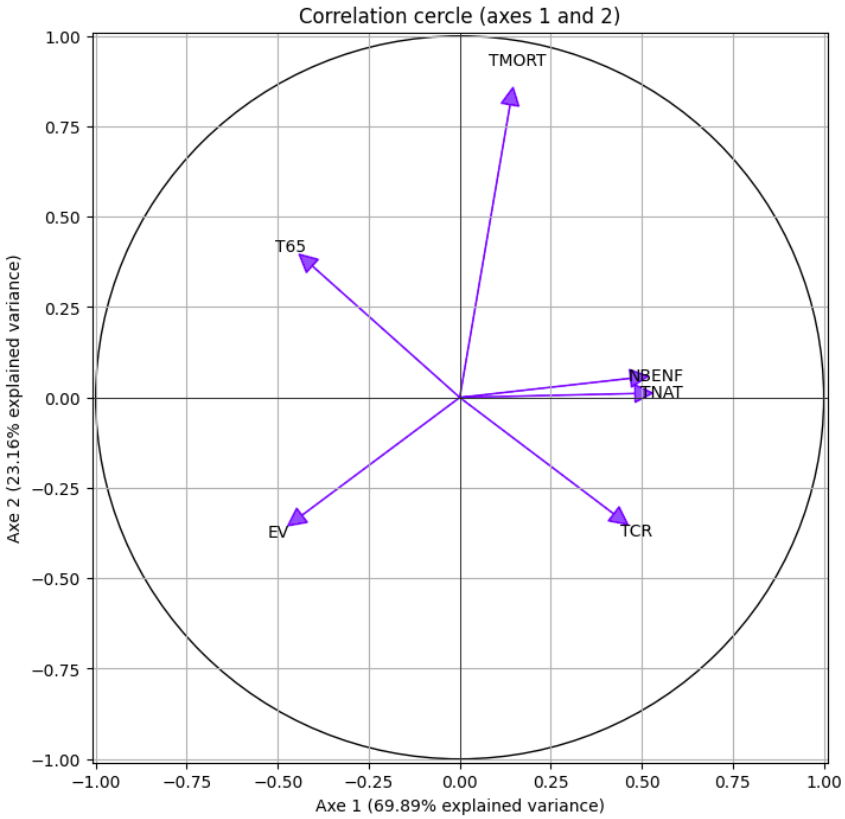

Figure 5: Cercle de corrélation des axes 1 et 2 de l'ACP

Exemple d'analyse : NBENF (Nombre d'enfants par famille) et TNAT (Taux de natalité) sont fortement corrélés positivement. Ce qui veut dire que plus il y a d'enfants par famille, plus le taux de natalité est élevé. A l'inverse, le taux de natalité et le taux de mort sont perpendiculaires, c'est-à-dire qu'il n'ont aucune corrélation entre eux.