Introduction à L'analyse en composantes principales

L'analyse en composantes principales (ACP), est la procédure statistique qui permet de résumer le contenu de l'information dans de grands tableaux de données pour être plus facilement visualisés et analysés. Les données sous-jacentes peuvent être des mesures décrivant les propriétés d'échantillons, des moments d'un processus continu ou des individus.

L'analyse en composantes principales, est une méthode de réduction de dimension de grands ensembles de données, en transformant un grand ensemble de variables en un plus petit qui contient toujours la plupart des informations du grand ensemble. Parce que les ensembles de données réduites sont plus faciles à explorer, à visualiser et plus facile et rapide pour les algorithmes de traitement. Donc, pour résumer, l'idée de l'ACP est simple : réduire le nombre de variables d'un ensemble de données, tout en préservant autant d'informations que possible.

L'analyse en composantes principales est l'une des méthodes statistique de l'analyse factorielle. Cette méthode est aujourd'hui l'une des techniques statistiques multivariées les plus populaires. Elle a été largement utilisée dans les domaines de la reconnaissance de formes et du traitement du signal.

L'utilisation de l'ACP peut aider à identifier les corrélations entre les points de données, par exemple s'il existe une corrélation entre la consommation d'aliments comme le poisson congelé et le pain croustillant.

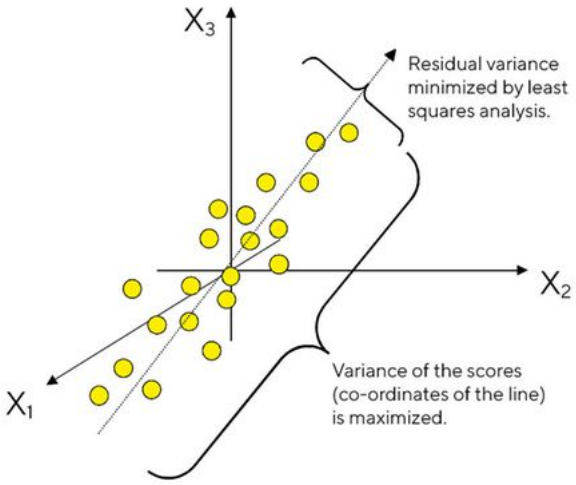
Historique de l'analyse en composantes principales

L'ACP constitue la base de l'analyse de données multivariée basée sur des méthodes de projection. Cette méthode statistique remonte à Cauchy mais a été formulée pour la première fois dans les statistiques par Pearson, qui a décrit l'analyse comme la recherche de « lignes et plans d'ajustement le plus proche aux systèmes de points dans l'espace» [Jackson, 1991].

Le but principale de l'ACP est d'extraire les informations les plus importantes des données et d'exprimer ces informations sous la forme d'un ensemble d'indices sommaires appelés **composantes principales** .

Statistiquement, l'ACP trouve des lignes, des plans et des hyper-plans dans l'espace K-dimensionnel qui se rapprochent aussi bien que possible des données dans le sens des

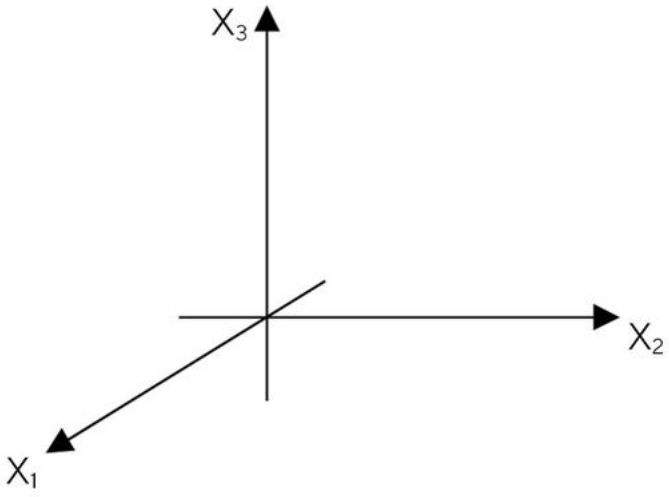
moindres carrés. Une ligne ou un plan qui est l'approximation par les moindres carrés d'un ensemble de points de données rend la variance des coordonnées sur la ligne ou le plan aussi grande que possible.



L'ACP crée une visualisation des données qui minimise la variance résiduelle dans le sens des moindres carrés et maximise la variance des coordonnées de projection.

Comment fonctionne ACP

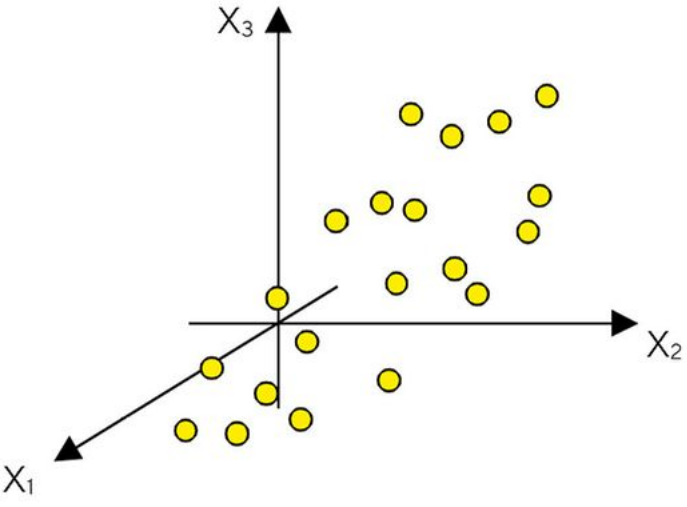
Considérons une matrice X avec N lignes (observations) et K colonnes (variables). Pour cette matrice, nous construisons un espace variable avec autant de dimensions qu'il y a de variables (voir figure ci-dessous). Chaque variable représente un axe de coordonnées. Pour chaque variable, la longueur a été normalisée selon un critère de mise à l'échelle, normalement par mise à l'échelle à la variance unitaire.



2.jpeg

Un espace de variable en K dimensions. Pour simplifier, seuls trois axes de variables sont affichés. La longueur de chaque axe de coordonnées a été normalisée selon un critère spécifique, généralement l'échelle de variance unitaire.

Dans l'étape suivante, chaque observation (ligne) de la matrice X est placée dans l'espace de variable à K dimensions. Par conséquent, les lignes du tableau de données forment un essaim de points dans cet espace.



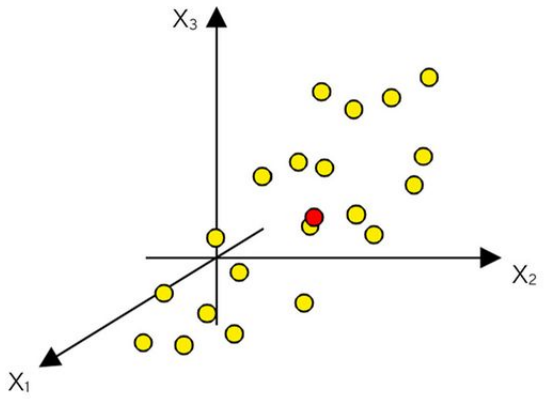
3.jpeg

Les observations (lignes) dans la matrice de données X peuvent être comprises comme

un essaim de points dans l'espace variable (espace K).

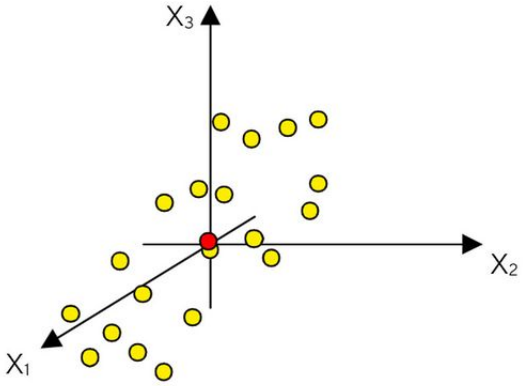
Centrage de la moyenne

Dans la procédure de centrage de la moyenne, vous calculez d'abord les moyennes des variables. Ce vecteur de moyennes est interprétable comme un point (voir figure il est en rouge) dans l'espace. Le point est situé au milieu de l'essaim de points (au centre de gravité).



4.jpeg

La soustraction des moyennes des données correspond à un repositionnement du système de coordonnées, de sorte que le point moyen est maintenant l'origine.

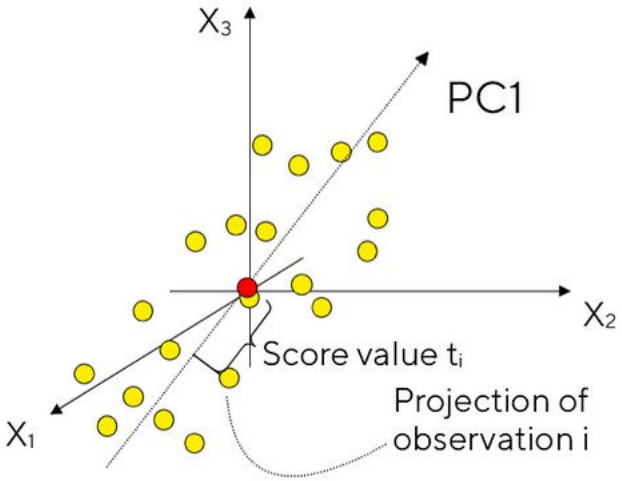


5.jpeg

La procédure de centrage moyen correspond au déplacement de l'origine du système de coordonnées pour qu'elle coïncide avec le point moyen (ici en rouge).

La première composante principale

Après centrage de la moyenne et mise à l'échelle de la variance unitaire, l'ensemble de données est prêt pour le calcul du premier indice récapitulatif, la première composante principale (PC1 en figure 6). Cette composante est la ligne dans l'espace de variable à K dimensions qui se rapproche le mieux des données au sens des moindres carrés. Cette ligne passe par le point moyen. Chaque observation (point jaune au figure 6) peut maintenant être projetée sur cette ligne afin d'obtenir une valeur de coordonnées le long de la ligne PC. Cette nouvelle valeur de coordonnée est également connue sous le nom de score .

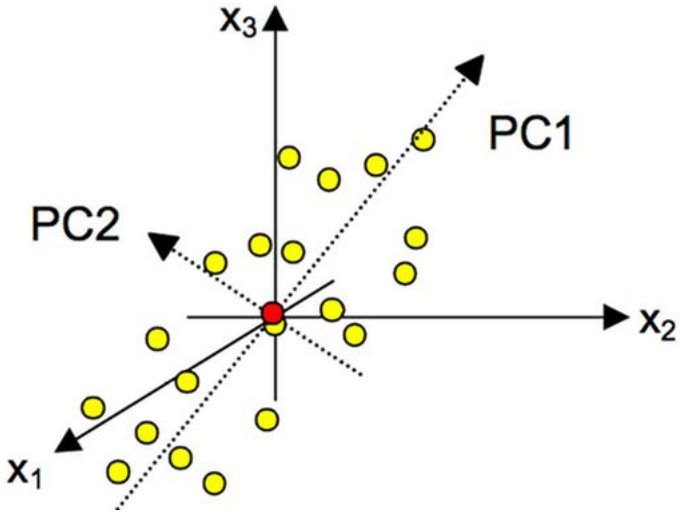


6.jpeg

La première composante principale (PC1) est la ligne qui représente la direction de variance maximale dans les données.

La deuxième composante principale

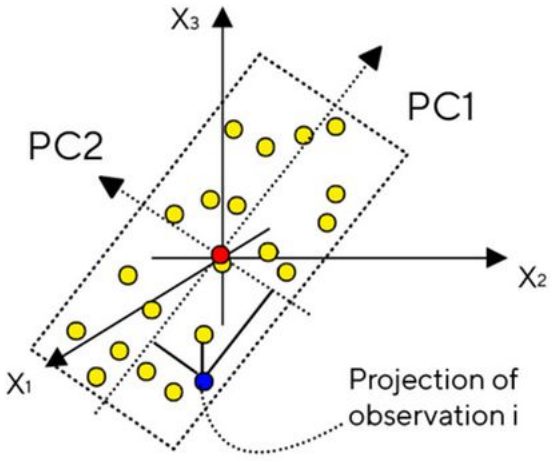
Habituellement, un indice de synthèse ou une composante principale est insuffisante pour modéliser la variation systématique d'un ensemble de données. Ainsi, un deuxième indice de synthèse - une deuxième composante principale (PC2) - est calculé. La deuxième CP est également représenté par une ligne dans l'espace variable K-dimensionnel, qui est orthogonal au premier CP. Cette ligne passe également par le point moyen, et améliore autant que possible l'approximation des données X. (Voire figure 7)



7.jpeg

Deux composantes principales définissent un plan et un modèle

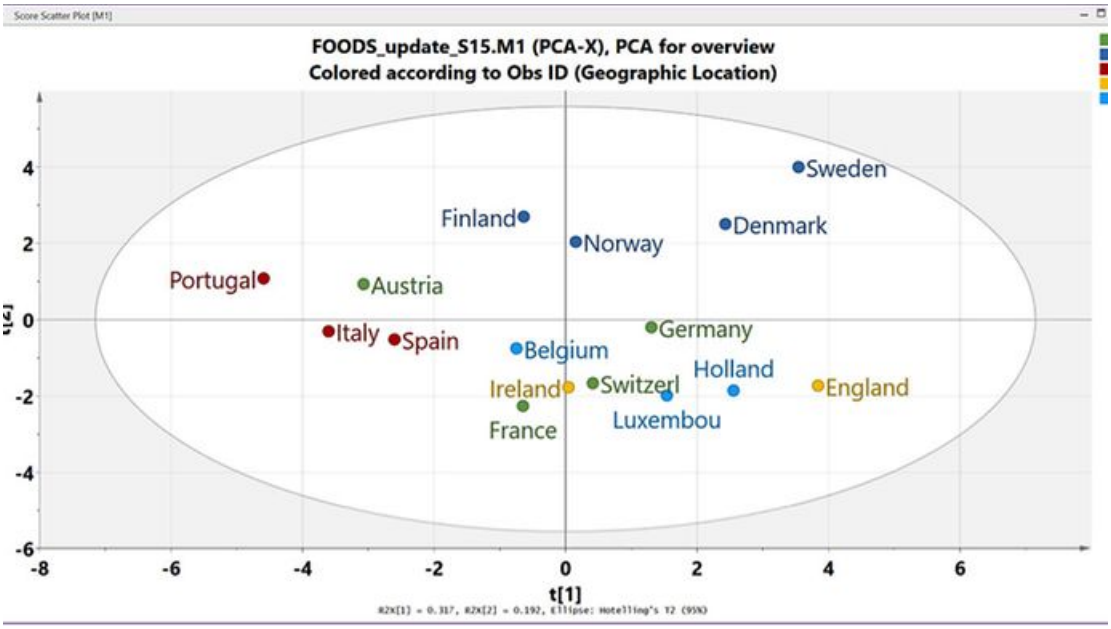
Lorsque deux composantes principales ont été dérivés, ils définissent ensemble un lieu, une fenêtre dans l'espace multidimensionnel de variable K-dimensionnel. En projetant toutes les observations sur le sous-espace de faible dimension et en traçant les résultats, il est possible de visualiser la structure de l'ensemble de données étudié. Les valeurs de coordonnées des observations sur ce plan sont appelées **scores**, et par conséquent, le tracé d'une telle configuration projetée est connu sous le nom de **graphique de score** (voire figure 8).



8.jpeg

Modélisation d'un ensemble de données

Voyons maintenant à quoi cela ressemble en utilisant un ensemble de données d'aliments couramment consommés dans différents pays européens. La figure ci-dessous affiche le **graphique des scores** des deux premières composantes principales. Ces scores sont appelés **t1 et t2**. Le graphique des scores est une carte de 16 pays. Les pays proches les uns des autres ont des profils de consommation alimentaire similaires, tandis que ceux qui sont éloignés les uns des autres sont différents. Les pays nordiques (Finlande, Norvège, Danemark et Suède) sont situés ensemble dans le coin supérieur droit, représentant ainsi un groupe de nations présentant une certaine similitude dans la consommation alimentaire. La Belgique et l'Allemagne sont proches du centre (origine) de la parcelle, ce qui indique qu'elles ont des propriétés moyennes.



9.jpeg

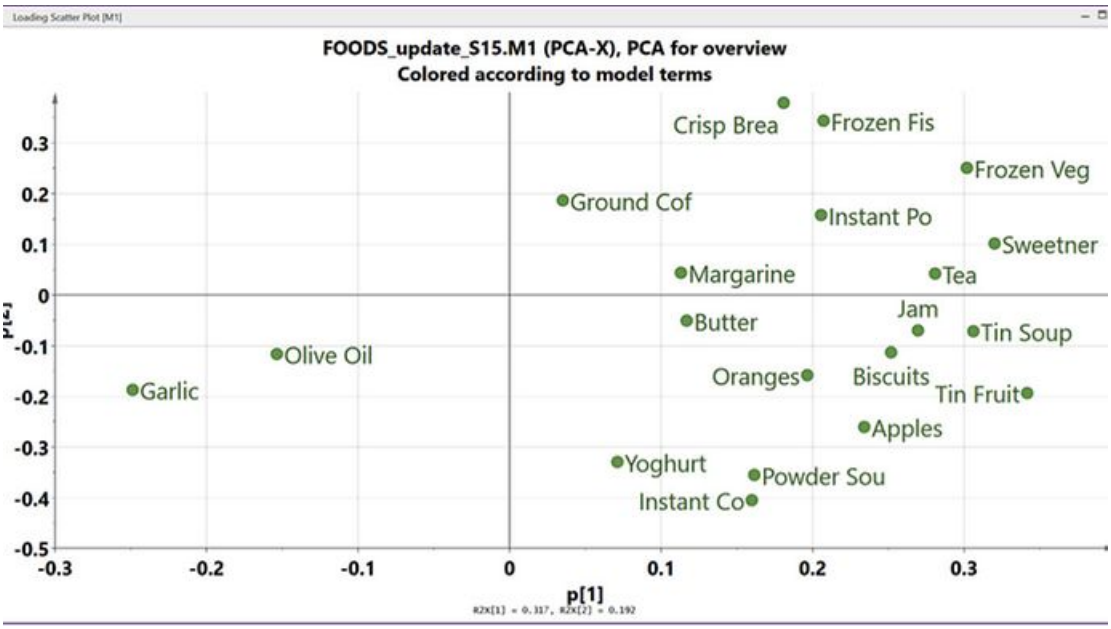
Le graphique des scores ACP des deux premiers CP d'un ensemble de données sur les profils de consommation alimentaire fournit une carte de la façon dont les pays sont liés les uns aux autres. La première composante explique 32% de la variation et la deuxième composante 19%. Couleur par emplacement géographique (latitude) de la capitale respective.

Comment interpréter le graphique des scores

Dans un modèle ACP à deux composantes, on pose la question suivante : quelles variables (provisions alimentaires) sont responsables des tendances observées parmi les observations (pays) ? Nous aimerions savoir quelles variables sont influentes, et aussi comment les variables sont corrélées. Cette connaissance est donnée par les chargements des composantes principales (graphique ci-dessous). Ces vecteurs de chargement sont appelés **p1 et p2**.

La figure ci-dessous affiche les relations entre les 20 variables en même temps. Les variables fournissant des informations similaires sont regroupées, c'est-à-dire qu'elles sont corrélées. Le pain croustillant (crips-br) et le poisson congelé (Fro-Fish) sont des exemples de deux variables qui sont positivement corrélées. Lorsque la valeur numérique d'une variable augmente ou diminue, la valeur numérique de l'autre variable a tendance à changer de la même manière.

Lorsque les variables sont corrélées négativement (inversement), elles sont positionnées sur les côtés opposés de l'origine de la parcelle, dans des quadrants opposés en diagonale. Par exemple, les variables ail et édulcorant sont inversement corrélées, ce qui signifie que lorsque l'ail augmente, l'édulcorant diminue, et vice versa.

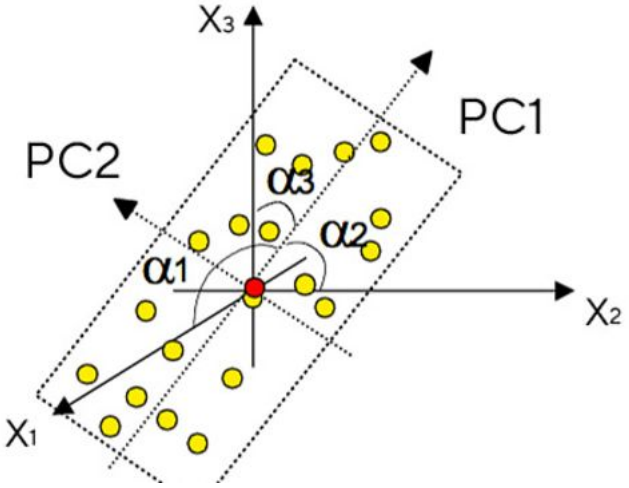


10.jpeg

Plus une variable est éloignée de l'origine du tracé, plus son impact sur le modèle est important. Cela signifie, par exemple, que les variables pain croustillant (Crisp-br), poisson congelé (Fro-Fish), légumes surgelés (Fro-Veg) et ail (ail) séparent les quatre pays nordiques des autres. Les quatre pays nordiques se caractérisent par des valeurs élevées (consommation élevée) des trois premières dispositions et une faible consommation d'ail. De plus, l'interprétation du modèle suggère que des pays comme l'Italie, le Portugal, l'Espagne et, dans une certaine mesure, l'Autriche ont une consommation élevée d'ail et une faible consommation d'édulcorant, de soupe en conserve (Ti-soup) et de fruits en conserve (Ti-Fruit).

Géométriquement, les chargements des composantes principales expriment l'orientation du plan du modèle dans l'espace de variable K-dimensionnel. La direction de PC1 par rapport aux variables d'origine est donnée par le cosinus des angles a1, a2 et a3. Ces valeurs indiquent comment les variables d'origine x1, x2 et x3 «se chargent» dans (ce qui signifie contribuent à) PC1. Par conséquent, ils sont appelés **chargements** .

Le deuxième ensemble de coefficients de chargement exprime la direction de PC2 par rapport aux variables d'origine. Par conséquent, étant donné les deux PC et les trois variables d'origine, six valeurs de chargement (cosinus des angles) sont nécessaires pour spécifier comment le plan du modèle est positionné dans l'espace K.



11.jpeg

Les chargements des composantes principales révèlent comment le plan du modèle ACP est inséré dans l'espace de variable. Les chargements sont utilisés pour interpréter la

signification des scores.

Pour plus d'information consulter la page suivante :

https ://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186

Explication étape par étape de l'ACP

ÉTAPE 1 : Normalisation

Le but de cette étape est de standardiser les variables initiales continues afin que chacune d'elles contribue de manière égale à l'analyse. Plus précisément, la raison pour laquelle il est essentiel d'effectuer une standardisation avant l'ACP, est que cette dernière est assez sensible aux variances des variables initiales. Autrement dit, s'il y a de grandes différences entre les distance de variables initiales, les variables avec des distance plus grandes domineront sur celles avec de petites distance (par exemple, une variable comprise entre 0 et 100 dominera sur une variable comprise entre 0 et 1), ce qui entraînera des résultats biaisés. Ainsi, la transformation des données à des échelles comparables peut éviter ce problème.

Mathématiquement, cela peut être fait en soustrayant la moyenne et en divisant par l'écart type pour chaque valeur de chaque variable.

$$z = \frac{valeur - moyenne}{ecart - type}$$

Une fois la standardisation terminée, toutes les variables seront transformées à la même échelle.

ÉTAPE 2 : Calcul de la matrice de covariance

Le but de cette étape est de comprendre comment les variables de l'ensemble de données d'entrée diffèrent de la moyenne les unes par rapport aux autres, ou en d'autres termes, de voir s'il existe une relation entre elles. Parce que parfois, les variables sont fortement corrélées de telle sorte qu'elles contiennent des informations redondantes. Ainsi, afin d'identifier ces corrélations, nous calculons la matrice de covariance.

La matrice de covariance est une matrice symétrique $p \times p$ (où p est le nombre de dimensions) qui a comme entrées les covariances associées à toutes les paires possibles des variables initiales. Par exemple, pour un ensemble de données à trois dimensions avec 3 variables x , y et z , la matrice de covariance est une matrice 3×3 de ceci à partir de :

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

Puisque la covariance d'une variable avec elle-même est sa variance ($Cov(a,a) = Var(a)$), dans le diagonale (en haut à gauche en bas à droite), nous avons en fait les variances de chaque variable initiale. Et comme la covariance est commutative ($Cov(a,b) = Cov(b,a)$), les entrées de la matrice de covariance sont symétriques par rapport à la diagonale principale, ce qui signifie que les parties triangulaires supérieure et inférieure sont égales.

Si le signe de la covariance positif alors : les deux variables augmentent ou diminuent ensemble (corrélées).

Si le signe de la covariance négatif alors : l'un augmente lorsque l'autre diminue (corrélation inverse).

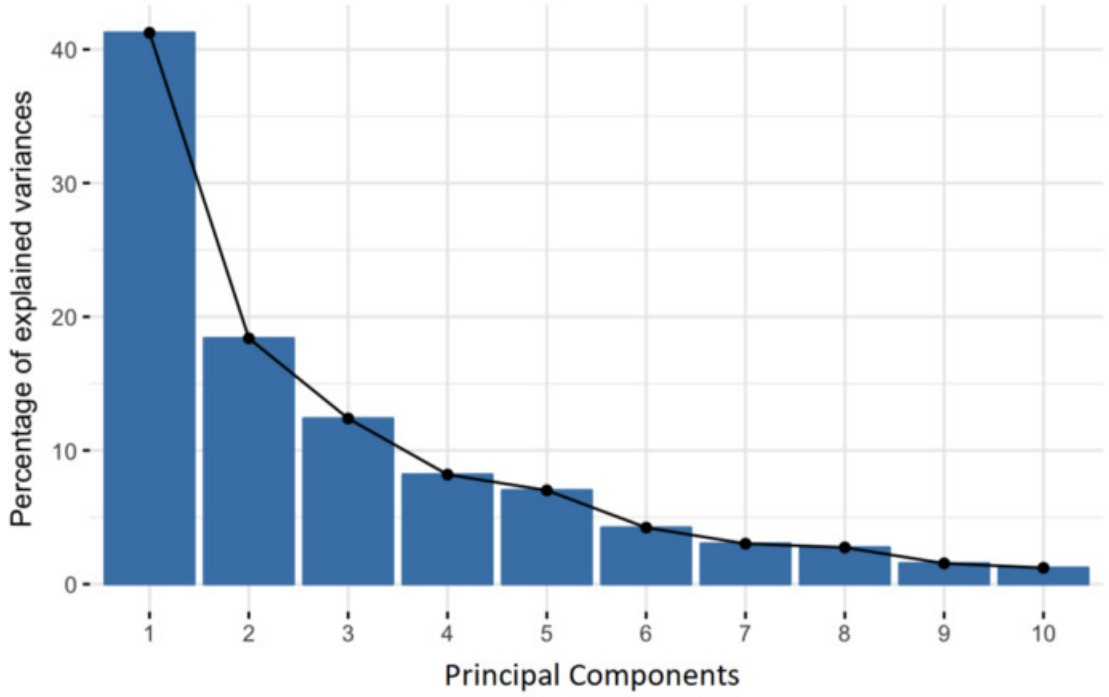
Donc, la matrice de covariance n'est rien de plus qu'un tableau qui résume les corrélations entre toutes les paires de variables possibles.

ÉTAPE 3 : Calcule des vecteurs propres et valeurs propres de la matrice de covariance pour identifier les composantes principales

Les vecteurs propres et les valeurs propres sont les concepts d'algèbre linéaire que nous devons calculer à partir de la matrice de covariance afin de déterminer les composantes principales des données.

Les composantes principales sont de nouvelles variables construites sous forme de combinaisons linéaires ou de mélanges des variables initiales. Ces combinaisons sont effectuées de telle sorte que les nouvelles variables (les composantes principales) ne sont pas corrélées et que la plupart des informations contenues dans les variables initiales sont

pressées ou compressées dans les premières composantes. Donc, l'idée est que les données à 10 dimensions vous donnent 10 composantes principaux, l'ACP essaie de mettre le maximum d'informations possibles dans la première composante, puis le maximum d'informations restantes dans le second et ainsi de suite. L'objectif de l'ACP est de réduire la dimensionnalité sans perdre beaucoup d'informations. Voir la figure ci-dessous.



Une chose importante à réaliser ici est que les composantes principales sont moins interprétantes et n'ont aucune signification réelle puisqu'elles sont construites comme des combinaisons linéaires des variables initiales.

Géométriquement, les composantes principales représentent les directions des données qui expliquent une quantité maximale de variance, ce sont les lignes qui capturent la plupart des informations des données. On peux dire que les composantes principales sont de nouveaux axes qui fournissent le meilleur angle pour voir et évaluer les données, afin que les différences entre les observations soient mieux visibles.