TP2 : Analyse en Composantes Principales

A. L. N'Guessan, V. Roca & W. Heyse

Le 19 Septembre 2025

Contents

1 Rappels de cours **1**
1.1 Introduction à l'Analyse en Composantes Principales 1
1.2 Fonctionnement Mathématique de l'ACP 1
1.3 Individus 3
1.4 Variables 4

2 Exercice **6**

1 Rappels de cours

1.1 Introduction à l'Analyse en Composantes Principales

L'Analyse en Composantes Principales (ACP) est une méthode d'analyse multivariée qui permet de réduire la dimensionnalité d'un jeu de données quantitatives tout en conservant un maximum d'information. Elle est particulièrement utile lorsque les données comportent un grand nombre de variables, souvent corrélées entre elles, et que l'on souhaite obtenir une représentation graphique simplifiée des observations (individus) et des variables.

L'objectif principal de l'ACP est de transformer les variables initiales en un ensemble de nouvelles variables **non corrélées** appelées **composantes principales**. Ces composantes principales sont ordonnées de manière à capturer autant que possible de *l'information* (variance totale) des données.

L'idée mathématique va donc être de projeter le nuage de points sur des sous-espaces affines de dimension plus petite, tels que la projection selon chacun de ces sous-espaces soit la plus fidèle possible aux données initiales. Autrement dit, on souhaite minimiser la distance entre les points du nuage initial et leurs projections selon le sous-espace considéré.

1.2 Fonctionnement Mathématique de l'ACP

Pour comprendre comment fonctionne l'ACP, il est nécessaire d'examiner les étapes mathématiques impliquées, depuis la structuration de la matrice de données jusqu'à la détermination des axes principaux.

1.2.1 Matrice de Données

Soit une matrice de données $\mathbf{X}$ de taille $n \times p$ où n représente le nombre d'individus (ou observations) et p le nombre de variables.

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Avant de procéder à l'ACP, les données sont généralement **centrées** (soustraction de la moyenne) et **réduites** (division par l'écart-type) pour éviter que les variables ayant des unités différentes ou des ordres de grandeur différents n'influencent de manière disproportionnée les résultats. Pour la suite des formules, on considère que le nuage de données est centré-réduit, ainsi sont centre de gravité (centre du nuage) est l'origine du repère.

1.2.2 Distance Utilisée

L'ACP se base sur la **distance euclidienne** pour mesurer la similarité entre individus dans l'espace des variables :

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

1.2.3 Critère d'Optimisation

Le problème peut-être vu comme la recherche d'axes sur lesquels la projection orthogoinale des données maximise *l'information* projetée. Dans le cas de l'ACP, on considère que les axes qu'on recherche passent tous par le centre de gravité (i.e. l'origine du repère dans le cas d'une ACP centré-réduite) afin de faciliter les calculs (une simple translation des données suffit pour généraliser au cas non centré) l'ACP cherche à maximiser la **variance** des projections des individus sur les nouveaux axes (composantes principales). Pour cela, elle résout le problème d'optimisation suivant :

Soit $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_p$ les nouvelles variables (composantes principales). La première composante principale $\mathbf{Z}_1$ est définie par un vecteur directeur unitaire $\mathbf{u}_1$ de sorte à maximiser la variance des données projetées sur $\mathbf{Z}_1$, on a donc le problème d'optimisation suivant à résoudre :

$$\begin{cases} \max\limits_{\mathbf{u}_1} \quad \mathbf{u}_1^\top \mathbf{X}^\top \mathbf{X} \mathbf{u}_1 \\ \|\mathbf{u}_1\|^2 = \mathbf{u}_1^\top \mathbf{u}_1 = 1 \end{cases}$$

En effet, dans le cas centré-réduit, la matrice de variance-covariance des données est : $\frac{1}{n-1}\mathbf{X}^\top\mathbf{X}$. Ici le $\frac{1}{n-1}$ n'a pas d'importance dans l'optimisation.

Les composantes suivantes $\mathbf{Z}_2, \mathbf{Z}_3, \ldots$ sont déterminées de manière similaire, tout en étant **orthogonales aux précédentes**.

1.2.4 Détermination des Axes Principaux

La résolution du problème d'optimisation conduit à identifier les vecteurs propres $\mathbf{u}_k$ de la matrice de variance-covariance des données $\mathbf{X}^\top\mathbf{X}$:

$$\mathbf{X}^\top \mathbf{X} \mathbf{u}_k = \lambda_k \mathbf{u}_k$$

Chaque valeur propre λ_k correspond à la part de variance des données expliquée par la k-ième composante principale, et le vecteur propre associé $\mathbf{u}_k$ est un vecteur directeur de la k-ième composante.

Les composantes principales sont telles que $\mathbb{E}(Z_k) = 0$, $\mathrm{Var}(Z_k) = \lambda_k$ et $\mathrm{Cov}(Z_k, Z_j) = 0$

1.2.5 Contribution relative à l'Inertie

L'inertie totale est la somme des variances des variables initiales. Chaque composante principale contribue à une part de cette inertie. La contribution de la k-ième composante principale à l'inertie totale est donnée par :

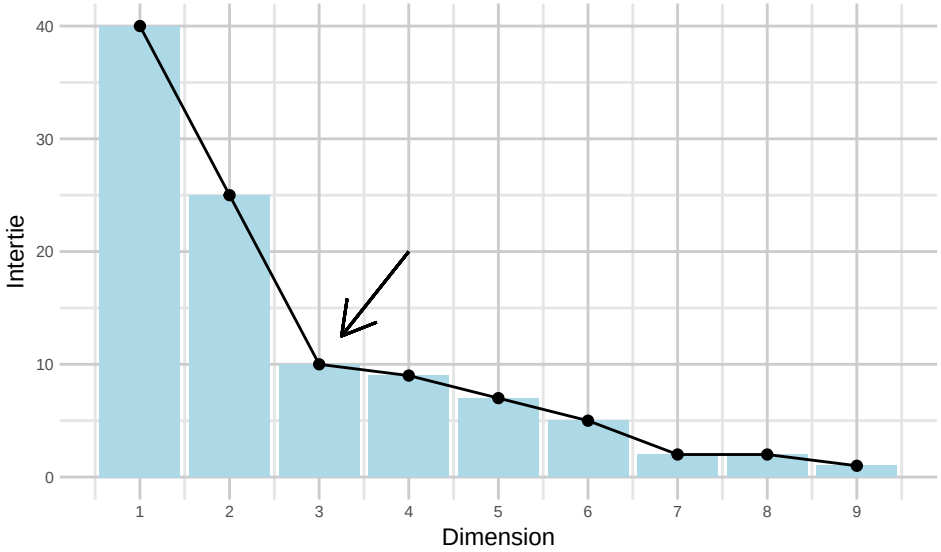
$$\text{Contribution relative de } \mathbf{Z}_k = \frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}$$

Plus la contribution relative est élevée, plus la composante principale est importante pour expliquer la variance des données.

En pratique :

Lors de la réalisation d'une ACP, le choix du nombre de composantes principales à conserver est crucial pour obtenir une représentation fidèle des données tout en réduisant la dimensionnalité. Plusieurs règles couramment utilisées aident à déterminer ce nombre :

1. **Critère de la Variance Cumulée** : On conserve les composantes principales qui expliquent un pourcentage cumulé significatif de la variance totale, souvent fixé entre 70 % et 90 %.
2. **Critère du Coude** : Le Scree Plot représente les valeurs propres en fonction des composantes. On choisit le nombre de composantes correspondant à la dernière forte diminution (coude) avant que les valeurs propres ne commencent à diminuer de manière régulière.



3. **Critère de Kaiser** : Ce critère recommande de conserver uniquement les composantes dont les valeurs propres sont supérieures à la moyenne des valeurs propres (supérieure à 1 dans le cadre d'une ACP centrée-réduite).

Ces règles ne sont pas exclusives et sont souvent utilisées conjointement pour prendre une décision éclairée sur le nombre de composantes à conserver.

1.3 Individus

1.3.1 Coordonnées des Individus dans les Composantes Principales

Lorsqu'on effectue une ACP, chaque individu du jeu de données est projeté dans le nouvel espace défini par les composantes principales. Les **coordonnées des individus** dans cet espace correspondent aux projections de ces individus sur les axes des composantes principales. Ces coordonnées sont essentielles pour interpréter la position des individus dans le nouvel espace réduit.

Pour un individu i et une composante principale k, la coordonnée z_{ik} est calculée comme suit :

$$z_{ik} = \mathbf{X}_{i\cdot} \cdot \mathbf{u}_k$$

où :

- $\mathbf{X}_{i\cdot}$ est le vecteur des p variables observées pour l'individu i (ligne i de la matrice de données centrée-réduite),
- $\mathbf{u}_k$ est le vecteur des coefficients de la k-ième composante principale (le k-ième vecteur propre de la matrice de covariance).

En pratique, ces coordonnées permettent de visualiser les individus dans un plan factoriel constitué par deux composantes principales, facilitant l'interprétation des relations entre individus et des composantes principales.

Pour évaluer la qualité de la représentation des individus dans l'espace des composantes principales, deux mesures clés sont utilisées : le **cosinus carré** ($\cos^2$) et la **contribution**.

1.3.2 Qualité de la représentation des individus : Cosinus Carré ($\cos^2$)

Le cosinus carré ($\cos^2$) mesure la qualité de la projection d'un individu sur un axe donné. Il est défini comme le carré du cosinus de l'angle entre le vecteur de l'individu et l'axe principal :

$$\cos^2(\text{Individu } i, \text{Axe } k) = \frac{z_{ik}^2}{\sum_{j=1}^{p} x_{ij}^2}$$

Où z_{ik} est la coordonnée de l'individu i sur l'axe k. Un $\cos^2$ **élevé** indique que l'angle entre l'individu et sa projection est faible donc que l'axe k **représente bien** l'individu i.

On note que le $\cos^2$ d'un individu sur toutes les composantes somme à 1 (i.e. l'individu est intégralement représenté dans le nouvel espace).

1.3.3 Contribution des individus

La contribution d'un individu à une composante principal indique dans quelle mesure cet individu influence la direction de l'axe. La contribution de l'individu i à l'axe k est donnée par :

$$\text{Contribution de l'Individu } i \text{ à l'Axe } k = \frac{z_{ik}^2}{(n-1)\lambda_k}$$

Plus la projection d'un individu sur un axe donné est grande, plus il contribuera à expliquer l'axe en question. Inversement, un individu dont la projection sur un axe est faible, et donc proche du centre de gravité, contribuera peu à l'inertie portée par cet axe. Une règle empirique pour l'analyse des contribution est de regarder les individus ayant un contribution plus grande que la contribution moyenne.

On note que la contribution de tous les individus à une composante somme à 1 (i.e. une composante est intégralement expliquée par les individus).

1.4 Variables

1.4.1 Coordonnées des variables / Cercle de Corrélation

Le **cercle de corrélation** est un graphique essentiel dans l'interprétation des composantes principales. Il permet de visualiser la relation entre les variables initiales et les nouvelles composantes principales, facilitant ainsi l'interprétation de ces dernières.

Dans un cercle de corrélation, les variables initiales sont représentées par des vecteurs. La position de chaque vecteur est déterminée par ses coordonnées sur les axes des composantes principales, qui sont les **coefficients de corrélation** entre la variable d'origine et les composantes principales.

Pour une variable j et une composante principale k, la corélation est donnée par :

$$\text{Coordonnée de la variable } j \text{ sur l'axe } k = a_{jk} = \text{corr}(X_{.j}, Z_k) = u_{jk}\sqrt{\lambda_k}$$

où :

- u_{jk} est le coefficient j du k-ième vecteur propre,
- λ_k est la valeur propre associée au k-ième vecteur propre.

On représente souvent ces points dans un plan factoriel muni d'un cercle de rayon 1. Les variables initiales sont alors représentées sous forme de vecteurs dont les coordonnées sont données par les corrélations.

Interprétation du Cercle de Corrélation :

- **Longueur du Vecteur** : La longueur du vecteur de chaque variable indique la qualité de sa représentation dans le plan factoriel formé par les composantes principales. Un vecteur proche du cercle unitaire (rayon = 1) signifie que la variable est bien représentée par les deux composantes principales choisies.
- **Angle entre les Vecteurs** : L'angle entre deux vecteurs représente la corrélation entre les variables correspondantes. Un angle proche de 0° ou 180° indique une forte corrélation (positive ou négative) entre les variables, tandis qu'un angle proche de 90° indique une faible ou nulle corrélation.
- **Position par Rapport aux Axes** : Les variables dont les vecteurs sont proches des axes des composantes principales sont celles qui contribuent le plus à ces axes et qui sont les mieux représentées par ces composantes.

Le cercle de corrélation est un outil visuel puissant pour comprendre comment les variables initiales se regroupent, s'opposent, ou se répartissent dans l'espace des composantes principales, offrant ainsi des insights précieux sur la structure des données.

1.4.2 Cosinus Carré ($\cos^2$) des Variables

Le $\cos^2$ d'une variable sur une composante principale mesure la part de la variance de cette variable expliquée par la composante. Il est donné par :

$$\cos^2(\text{Variable } j, \text{Axe } k) = \frac{a_{jk}^2}{\sum_{l=1}^{p} a_{jl}^2}$$

Où a_{jk} est la coordonnée de la variable j sur l'axe k.

1.4.3 Contribution des Variables

La contribution d'une variable à une composante principale représente l'importance de cette variable dans la définition de la composante. Elle est calculée par :

$$\text{Contribution de la Variable } j \text{ à l'Axe } k = \frac{a_{jk}^2}{\lambda_k \sum_{l=1}^{p} a_{jl}^2}$$

1.4.4 Interprétation des Axes

L'interprétation des axes en ACP est une étape cruciale pour tirer des conclusions significatives des résultats obtenus. Les axes représentent de nouvelles variables synthétiques qui sont des combinaisons linéaires des variables initiales. Comprendre ce que ces axes signifient permet de donner du sens aux données projetées dans cet espace réduit.

Méthode pour Interpréter les Axes

1. **Identifier les Variables Initiales Influentes** : Pour chaque axe, examinez les coordonnées, contributions et qualité de représentation des variables dans la composante principale. Ces coefficients indiquent combien chaque variable contribue à l'axe. Concentrez-vous sur les variables ayant des coefficients les plus grands en valeur absolue. Ce sont celles qui influencent le plus l'axe. Notez si elles sont positives ou négatives pour comprendre la direction de leur influence.
2. **Analyser les Indices de Corrélation (Cercle de Corrélation)** : Utilisez le cercle de corrélation pour visualiser comment les variables sont liées aux axes. Les variables proches d'un axe contribuent fortement à cet axe. Vérifiez également les angles entre les vecteurs des variables pour identifier les corrélations entre elles.
3. **Examiner les Individus Influents** : Étudiez les coordonnées, contributions et qualité de représentation des individus sur les axes pour identifier ceux qui se démarquent. Ces individus influants peuvent révéler des sous-groupes ou des tendances spécifiques dans les données et aider à donner du sens aux axes. Les individus situés aux extrémités d'un axe sont particulièrement importants pour comprendre l'interprétation de cet axe.
4. **Synthétiser les Informations** : Assemblez les informations obtenues en intégrant les contributions des variables et des individus pour chaque axe.
5. **Formuler des Interprétations Claires** : En utilisant les indices rassemblés, nommez ou décrivez chaque axe en fonction des variables qui le composent et des individus qu'il distingue. Cela permet de donner du sens aux données réduites tout en facilitant la communication des résultats.

2 Exercice

Dans cet exercice vous allez réaliser une Analyse en Composantes Principales "à la main" sur le jeu de données `Temperatures.csv`.

1. Chargez le jeu de données `Temperatures.csv`.
2. Réalisez une rapide analyse descriptive des données.
3. On s'intérèsse désormais uniquement aux températures mensuelles. Centrez et réduisez les données.
4. Calculez la matrice de variance-covariance des données sans utiliser la fonction `cor` (ou assimilié). Commentez cette matrice.
5. Déterminez les vecteurs propres et valeurs propres de la matrice de variance-covariance. (*Indication :* Utilisez la fonction `eigen`)
6. A l'aide des valeurs propres déterminez les contributions relatives de chaque axe et tracez un graphe de ces valeurs. Avec la méthode de votre choix, combien d'axes retiendriez-vous pour l'analyse ?
7. Déterminez les coordonées des individus dans les composantes principales. Représentez les individus dans les 3 premiers plans factoriels
8. Pour chaque individu, déterminez sa contribution et sa qualité de représentation. Quels individus sont les mieux représentés et les plus contributeurs sur les axes que vous avez sélectionnés.
9. Déterminez les coordonnées des variables. Représentez-les.
10. Pour chaque variable, déterminez sa contribution et sa qualité de représentation. Quelles variables sont les mieux représentées et les plus contributrices sur les axes que vous avez sélectionnés. Y a-t-il des similarités avec la question précédante ?
11. Interprétez les axes de votre ACP.