UNIVERSITE DE PARAKOU

ECOLE NATIONALE DE STATISTIQUE, DE PLANIFICATION ET DE DEMOGRAPHIE

(ENSPD/UP)

Département de Statistique Appliquée

ANALYSE DESCRIPTIVE MULTIDIMENSIONNELLE (ACP, AFC, AFCM, ANALYSE DISCRIMINANTE, CLASSIFICATION)

MASTER 1

28 03 2025

Dr. DICKO Aliou

Objectif général de l’ECU

L'objectif de ce cours est de fournir aux apprenants les concepts fondamentaux et les outils méthodologiques permettant d'explorer et d'interpréter des données complexes à travers des techniques d’analyse descriptive multidimensionnelle. Il vise à développer la capacité à identifier les structures sous-jacentes des données, réduire leur dimensionnalité et faciliter leur interprétation visuelle et statistique.

Objectifs spécifiques de l’ECU

À la fin de ce cours, les étudiants seront capables de :

- Comprendre les principes de l'analyse descriptive multidimensionnelle, notamment les différences avec les autres méthodes statistiques.
- Appliquer les principales méthodes d’analyse factorielle (ACP, AFC, AFCM, ANALYSE DISCRIMINANTE, CLASSIFICATION) pour explorer et résumer des données multidimensionnelles.
- Interpréter les résultats des analyses en utilisant des indicateurs tels que les inerties, les contributions et les qualités de représentation.
- Utiliser des outils logiciels comme R pour réaliser des analyses et visualiser les résultats sous forme de graphiques factoriels.Identifier les situations où ces méthodes sont pertinentes et savoir choisir la technique appropriée en fonction des données.

Prérequis

- Bonne connaissance en statistiques descriptive et inférentielle
- Bonne connaissance en informatique (notamment le logiciel Excel).

Contenu de la formation

Le cours est structuré autour des grandes thématiques suivantes :

- Introduction à l'analyse descriptive multidimensionnelle
- Définition et cadre d'application
- Différences avec les méthodes univariées et bivariées
- L'Analyse en Composantes Principales (ACP)
  - Objectifs et principes
  - Interprétation des axes factoriels
  - Visualisation des individus et des variables
- L'Analyse des Correspondances (AFC et AFCM)
  - Tableau de contingence et distances du khi-deux
  - Interprétation des projections des individus et des modalités
  - Application aux données qualitatives

Contenu de la formation

- Les critères d'interprétation des projections
- Contributions, qualités de représentation ($\cos^2$)
- Indicateurs d'inertie et choix du nombre d'axes
- Application pratique avec le logiciel R
- Importation et manipulation des données
- Réalisation d'analyses factorielle
- Interprétation et mise en forme des résultats
- Comparaison des différentes méthodes d'analyse multidimensionnelle
- ACP vs AFC vs AFCM
- Cas pratiques et études de données réelles

Méthodes d’enseignement/apprentissage

- Cours théorique ;
- Travaux pratiques ;
- Travaux dirigés ;
- TPE (Etude de cas, exposé).

Lieu d’apprentissage

- Ecole/salle de cours ;
- Salle informatique.

Matériel pédagogique

- Projecteur ;
- Ordinateur ;
- Note de cours ;
- Tableau et craie.

Compétences générales

- Prendre des initiatives et décisions ;
- Croire que l'on peut surmonter tous les obstacles ;
- Mettre en œuvre des ressources organisationnelles pour produire des résultats ;
- Analyser les problèmes pour y trouver des solutions ;
- Apprendre.
- Ne pas remettre à plus tard ;
- Être une personne fiable et responsable ;
- Rédiger des rapports.

Mode d’évaluation

- Evaluation formative ;
- Devoir de table et oral.

Références bibliographiques

- Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer.
- Abdi, H., & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics, 2*(4), 433-459.
- Jackson, J. E. (2005). *A User's Guide to Principal Components*. John Wiley & Sons.
- Dunteman, G. H. (1989). *Principal Components Analysis*. Sage Publications.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2*(11), 559-572.
- Saporta, G. (2011). *Probabilités, analyse des données et statistique*. Technip.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis* (8th ed.). Cengage Learning.
- Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*(1-3), 37-52.
- Lever, J., Krzywinski, M., & Altman, N. (2017). Principal component analysis. *Nature Methods, 14*(7), 641-642.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.

Thème d’exposés :

1. Introduction à l’Analyse Factorielle des Correspondances (AFC) et ses applications
2. Analyse Factorielle des Correspondances Multiples (AFCM) : Étude des variables qualitatives
3. Comparaison entre ACP et AFC : Quand utiliser l’une ou l’autre ?
4. Analyse Discriminante : Classification supervisée et interprétation des résultats
5. Méthodes de Classification : Classification Ascendante Hiérarchique (CAH) vs K-Means
6. ACP et régression : Utilisation des Composantes Principales dans la prédiction
7. Critères d’évaluation des classifications en Analyse Multidimensionnelle
8. Applications de l'Analyse Multidimensionnelle dans la recherche en biologie et en agronomie
9. L’Analyse en Composantes Indépendantes (ICA) : Différence avec l’ACP et applications

Thème d’exposés :

10. Big Data et Analyse Multidimensionnelle : Défis et solutions
11. Implémentation en R et Python des méthodes d’Analyse Descriptive Multidimensionnelle
12. Sélection des variables en Analyse en Composantes Principales (ACP) : Méthodes et impacts
13. Interprétation des axes factoriels en ACP, AFC et AFCM : Bonnes pratiques et erreurs à éviter
14. Rôle et impact du choix des poids dans l’Analyse Factorielle des Correspondances (AFC et AFCM)
15. Analyse Discriminante Linéaire (LDA) vs Analyse Discriminante Quadratique (QDA) : Comparaison et applications
16. Intégration de l’Analyse Multidimensionnelle en Intelligence Artificielle et Machine Learning

1. Description

- **1. Description**
- Les analyses descriptives multidimensionnelles regroupent un ensemble de méthodes statistiques permettant d'explorer et de visualiser des ensembles de données complexes en plusieurs dimensions. Ces méthodes visent à réduire la dimensionnalité des données tout en préservant l'information essentielle pour faciliter l'interprétation.
- Parmi les principales techniques utilisées, on trouve :
- **L'Analyse en Composantes Principales (ACP)** : utilisée pour résumer l'information contenue dans un grand nombre de variables quantitatives en un nombre réduit de composantes principales.
- **L'Analyse Factorielle des Correspondances (AFC)** : adaptée aux tableaux de contingence et permettant d'étudier les relations entre catégories de variables qualitatives.
- **L'Analyse Factorielle des Correspondances Multiples (AFCM)** : extension de l'AFC pour analyser plusieurs variables qualitatives simultanément.
- **L'Analyse Discriminante** : utilisée pour classifier des observations en groupes prédéfinis en maximisant la séparation entre ces groupes.
- Ces méthodes sont largement utilisées en économie, en biologie, en marketing et dans bien d'autres domaines pour explorer des relations cachées dans les données et simplifier leur visualisation.

Chapitre 1 : Analyse en Composantes Principales (ACP)

1. Définition

L'**Analyse en Composantes Principales (ACP)** est une méthode statistique utilisée pour explorer et simplifier des ensembles de données complexes et multidimensionnels. Elle permet de réduire la dimensionnalité d'un jeu de données tout en conservant le maximum d'information possible.

- Une bonne présentation des données est cruciale pour une ACP réussie. Une fois les données bien préparées (choix des variables, standardisation, analyse de corrélation), on peut passer à l'application de l'ACP pour extraire les composantes principales et les interpréter.

2. Pourquoi utiliser l'ACP ?

- Dans de nombreux domaines (agronomie, économie, biologie, etc.), les données collectées comportent plusieurs variables qui peuvent être corrélées entre elles. L'ACP aide à :
- Identifier des structures sous-jacentes dans les données.
- Réduire le nombre de variables tout en minimisant la perte d'information.
- Visualiser des tendances et des regroupements dans un espace de plus faible dimension.

3. Clarification des Concepts Clés de l'Analyse en Composantes Principales (ACP)

- L'ACP repose sur plusieurs concepts fondamentaux qui permettent de réduire la dimensionnalité d'un jeu de données tout en conservant l'essentiel de l'information. Voici une clarification des termes et notions clés.

3.1. Composantes Principales

- Les composantes principales sont de nouvelles variables obtenues comme des combinaisons linéaires des variables initiales. Elles sont choisies de manière à maximiser la variance des données.
- La première composante principale (CP1) est celle qui capture la plus grande part de la variance totale des données.
- La deuxième composante principale (CP2) est orthogonale à la première et capture la plus grande variance restante, et ainsi de suite.
- Chaque composante principale est une combinaison linéaire des variables initiales :

$$CP_k = a_{k1}X_1 + a_{k2}X_2 + \dots + a_{kp}X_p$$

où a_{kj} sont les **coefficients des vecteurs propres** (poids de chaque variable dans la composante).

3.2. Valeurs Propres et Vecteurs Propres

- L'ACP repose sur la **décomposition en valeurs propres** de la matrice de covariance ou de corrélation.
- Les **valeurs propres** indiquent la part de variance expliquée par chaque composante principale.
- Les **vecteurs propres** donnent les coefficients permettant de calculer les nouvelles variables (composantes principales).
- Une valeur propre élevée signifie que la composante associée contient beaucoup d'information.

3.3. Variance Expliquée et Choix du Nombre de Composantes

- L'objectif de l'ACP est de représenter les données avec un plus petit nombre de variables tout en gardant un maximum d'information.
- La **variance expliquée** par une composante principale est donnée par :

$$\frac{\lambda_k}{\sum \lambda_j}$$

où λ_k est la valeur propre associée à la k-ième composante.

- Un **scree plot** (ou diagramme de l'éboulis) permet de visualiser la décroissance de la variance expliquée et d'identifier le "point de coude" où ajouter plus de composantes n'apporte que peu d'information supplémentaire.
- Une règle courante est de conserver les composantes qui expliquent au moins **70-80%** de la variance totale.

3.4. Cercle des Corrélations et Interprétation des Variables

Le **cercle des corrélations** est une représentation graphique qui permet d’analyser comment les variables initiales sont liées aux composantes principales.

Si une variable est proche du bord du cercle, elle est bien représentée par les composantes principales.

Si deux variables sont proches et alignées, elles sont **corrélées positivement**.

Si elles sont opposées, elles sont **corrélées négativement**.

Cela permet d'identifier des groupes de variables ayant des comportements similaires.

3.5. Représentation des Individus et Groupements

- L'ACP permet aussi de visualiser **les individus** dans un nouvel espace réduit (ex. en 2D ou 3D).
- Les **individus proches** sur le graphique ont des profils similaires.
- Les **groupements d'individus** peuvent révéler des structures cachées dans les données (ex. segmentation de clients, groupes de variétés végétales, etc.).

3.6. ACP Normale vs ACP sur Corrélation

L'ACP peut être réalisée sur :

- La **matrice de covariance** si les variables sont sur des échelles comparables.
- La **matrice de corrélation** si les variables sont sur des échelles différentes (cas courant car elle standardise les variables).

Chapitre 1 : Analyse en Composantes Principales (ACP)

- **Résumé**

Concept	Explication
Composantes principales	Nouvelles variables qui maximisent la variance des données
Valeurs propres	Indiquent la part de variance expliquée par chaque composante
Vecteurs propres	Déterminent le poids de chaque variable dans les composantes
Variance expliquée	Pourcentage d'information conservée par un ensemble de composantes
Scree plot	Aide à choisir le nombre de composantes pertinentes
Cercle des corrélations	Visualise les relations entre variables et composantes
Représentation des individus	Permet d'analyser les proximités et regroupements d'individus
ACP sur covariance vs corrélation	Utilisation de la covariance pour variables comparables, corrélation pour standardisation

4. Principe de l'ACP

- L’ACP repose sur le calcul des **composantes principales**, qui sont des combinaisons linéaires des variables d’origine. Ces composantes sont obtenues en maximisant la variance expliquée et sont orthogonales entre elles.
- Les étapes principales de l’ACP sont :
- **Standardisation des données** (si nécessaire) pour éviter qu’une variable à grande échelle domine l’analyse.
- **Calcul de la matrice de covariance ou de corrélation** entre les variables.
- **Détermination des valeurs propres et des vecteurs propres**, qui permettent d’obtenir les axes principaux.
- **Interprétation des composantes principales**, en analysant les contributions des variables initiales à chaque axe.
- **Visualisation des résultats** à l’aide de graphiques comme le cercle des corrélations et le plan factoriel.

5. Applications de l'ACP

- L'ACP est utilisée dans divers domaines, notamment :
- **Agriculture et biologie** : étude des caractéristiques des variétés végétales.
- **Économie et finance** : analyse des performances des entreprises ou des marchés.
- **Marketing** : segmentation des clients selon leurs comportements d'achat.
- **Médecine** : identification de facteurs influençant certaines maladies.

6. Présentation des données en Analyse en Composantes Principales (ACP)

- Avant d'appliquer l'ACP, il est essentiel de bien préparer et structurer les données. Voici comment les données sont généralement présentées et préparées pour une ACP.

6.1. Structure des Données

- L'ACP est appliquée sur un **tableau de données** où :
- **Les individus** (observations) sont placés en lignes.
- **Les variables** (caractéristiques mesurées) sont placées en colonnes.
- Exemple de tableau de données :

Individus	Variable 1	Variable 2	Variable 3	...	Variable p
Obs 1	5.2	3.1	7.8	...	2.5
Obs 2	6.1	2.9	8.0	...	3.1
Obs 3	4.9	3.5	7.2	...	2.8
...	...	...	...	...	...

- Chaque ligne correspond à un individu, et chaque colonne correspond à une variable quantitative.

6.2. Prétraitement des Données

Avant d'appliquer l'ACP, plusieurs étapes de traitement sont nécessaires :

a) Vérification du type de variables

L'ACP ne s'applique qu'aux variables quantitatives continues. Les variables qualitatives doivent être transformées en indicatrices (ex. ACP sur les correspondances multiples).

b) Standardisation des données (si nécessaire)

Si les variables sont mesurées sur des échelles différentes (ex. une variable en mètres, une autre en kilogrammes), il est important de centrer et réduire les données pour éviter qu'une variable domine l'analyse.

- Centrage : Soustraire la moyenne de chaque variable.
- Réduction : Diviser par l'écart-type de chaque variable.

- **Formule de normalisation d'une variable X_j :**

$$Z_j = \frac{X_j - \bar{X}_j}{\sigma_j}$$

où $\bar{X}_j$ est la moyenne et σ_j l'écart-type de la variable X_j.

6.3. Matrice des Données et Matrice de Corrélation/Covariance

- Après standardisation (si nécessaire), l'ACP utilise :
- **La matrice de covariance** (si les variables sont sur des échelles comparables).
- **La matrice de corrélation** (si les variables ont des échelles différentes).
- Exemple de matrice de corrélation (où chaque valeur représente la corrélation entre deux variables) :

Variables	Var1	Var2	Var3	...
Var1	1.00	0.85	0.63	...
Var2	0.85	1.00	0.72	...
Var3	0.63	0.72	1.00	...
...	...	...	...	...

- Une forte corrélation entre deux variables indique qu'elles apportent une information redondante, ce qui justifie l'ACP pour réduire la dimensionnalité.

6.4. Visualisation des Données Initiales

Avant d’appliquer l’ACP, il est utile de visualiser les données avec :

- **Des histogrammes ou boxplots** pour analyser la distribution des variables.
- **Des matrices de dispersion (scatter plots)** pour voir les corrélations entre les variables.
- **Un heatmap de corrélation** pour détecter les variables fortement liées.

7. Présentation de la Matrice X dans une Analyse en Composantes Principales (ACP)

Dans une ACP, la matrice des données X joue un rôle fondamental. C'est à partir de cette matrice que l'on calcule les composantes principales.

7.1. Définition de la Matrice X

La matrice X représente le **tableau des données initiales**, contenant n individus (observations) et p variables quantitatives. Elle est généralement de taille **n×p** :

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

où :

- xij représente la valeur de la variable j pour l'individu i.
- n est le nombre d'individus.
- p est le nombre de variables quantitatives.

Exemple pour n=4 individus et p=3 variables :

$$X = \begin{bmatrix} 5.2 & 3.1 & 7.8 \\ 6.1 & 2.9 & 8.0 \\ 4.9 & 3.5 & 7.2 \\ 5.5 & 3.0 & 7.5 \end{bmatrix}$$

7.2. Centrage et Réduction de la Matrice X

- Dans l'ACP, il est courant de **centrer et réduire** la matrice X pour éviter que des variables ayant des unités différentes n'influencent trop l'analyse.
- **Centrage** : On soustrait la moyenne de chaque variable pour obtenir une moyenne nulle.
- **Réduction** : On divise par l'écart-type pour que toutes les variables aient une variance égale à 1.
- La matrice **centrée et réduite** Z est donnée par :

$$Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix}$$

Avec $$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$ où :

- $\bar{x}_j$ est la moyenne de la variable j.
- σ_j est son écart-type.

Exemple : Si une variable a une moyenne de **5** et un écart-type de **2**, alors une valeur de **7** devient :

$$z = \frac{7-5}{2} = 1.0$$

7.3. Matrice de Covariance ou de Corrélation

- Une fois les données transformées en Z, l'ACP utilise :
- **La matrice de covariance** S, si les variables sont comparables en termes d'unités.
- **La matrice de corrélation** R, si les variables sont sur des échelles différentes.
- Ces matrices sont calculées comme suit :
- **Matrice de covariance** S : $$S = \frac{1}{n} Z^T Z$$
- **Matrice de corrélation** R : $$R_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\sigma_i \sigma_j}$$
- Exemple de **matrice de corrélation** pour 3 variables : $$R = \begin{bmatrix} 1.00 & 0.85 & 0.63 \\ 0.85 & 1.00 & 0.72 \\ 0.63 & 0.72 & 1.00 \end{bmatrix}$$

La **matrice X** est le point de départ de l'ACP, contenant les valeurs des variables pour chaque individu.

- On la transforme en **matrice centrée-réduite Z** pour éviter les biais dus aux différences d'échelle.
- L'ACP utilise ensuite la **matrice de covariance ou de corrélation** pour extraire les composantes principales.

8. Propriétés des Composantes Principales lors d'une Analyse en Composantes Principales (ACP)

- Les **composantes principales (CPs)** sont des variables synthétiques obtenues par l'ACP à partir des données initiales. Elles possèdent plusieurs propriétés fondamentales qui garantissent leur utilité et leur interprétabilité.

8.1. Orthogonalité et Indépendance

Les composantes principales sont **orthogonales** entre elles :

- Cela signifie que chaque composante principale est indépendante des autres et ne contient pas d'information redondante.

$$CP_i \perp CP_j \quad \text{pour } i \neq j$$

- Mathématiquement, les CPs sont des **combinaisons linéaires orthogonales** des variables initiales.

8.2. Maximisation de la Variance

- La **première composante principale (CP1)** est choisie de manière à capturer le maximum de variance des données.
- La **deuxième composante principale (CP2)** capture la variance restante tout en étant orthogonale à CP1.
- Chaque composante supplémentaire suit cette logique, ce qui garantit que les premières CPs contiennent l'essentiel de l'information.

$$\mathrm{Var}(CP_1) \geq \mathrm{Var}(CP_2) \geq \cdots \geq \mathrm{Var}(CP_p)$$

8.3. Somme de la Variance Conservée

- La somme des variances des CPs est égale à la somme des variances des variables initiales.
- Cela garantit que l'ACP ne perd pas d'information, mais la redistribue différemment.

$$\sum_{k=1}^{p} \lambda_k = \sum_{j=1}^{p} \sigma_j^2$$

où λ_k sont les **valeurs propres** associées aux CPs et σ^2_j sont les variances des variables initiales.

8.4. Combinaisons Linéaires des Variables Initiales

- Chaque CP est une combinaison linéaire des variables d'origine :

$$CP_k = a_{k1}X_1 + a_{k2}X_2 + \ldots + a_{kp}X_p$$

- Les **coefficients a_{kj}** sont les **poids des vecteurs propres** et indiquent l'importance de chaque variable dans la composante principale.
- Plus un coefficient est grand en valeur absolue, plus la variable contribue à cette composante.

8.5. Somme des Carrés des Coefficients Égale à 1

- Chaque vecteur propre est **normalisé**, ce qui signifie que la somme des carrés de ses coefficients est égale à 1 :

$$\sum_{j=1}^{p} a_{kj}^2 = 1$$

- Cela garantit que les CPs sont des transformations linéaires bien définies et comparables.

8.6. Interprétation Géométrique : Changement de Base

- L'ACP effectue une **rotation du système de coordonnées** vers une nouvelle base où les axes sont alignés avec les directions de plus grande variance.
- Les **axes principaux** sont donnés par les vecteurs propres de la matrice de covariance/corrélation.

8.7. Invariance par Transformation Linéaire

- Si les données sont transformées par une translation ou une mise à l'échelle uniforme, les résultats de l'ACP restent valables.
- En revanche, si les variables sont exprimées dans des unités très différentes, il est préférable d'utiliser l'**ACP sur la matrice de corrélation** plutôt que sur la covariance.

Résumé des propriétés

Propriété	Explication
Orthogonalité	Les CPs sont perpendiculaires et indépendantes.
Maximisation de la variance	Chaque CP capte le maximum de variance restante.
Somme de la variance conservée	Pas de perte d'information, seule la distribution de variance change.
Combinaisons linéaires des variables initiales	Chaque CP est définie par des poids appliqués aux variables.
Somme des carrés des coefficients = 1	Normalisation des vecteurs propres pour une interprétation standardisée.
Interprétation géométrique	Changement de base optimisé pour maximiser l'information.

9. Calcul des Valeurs Propres et des Vecteurs Propres en ACP

- L'**Analyse en Composantes Principales (ACP)** repose sur la **décomposition spectrale** de la matrice de covariance ou de corrélation. Cela signifie que nous devons **calculer les valeurs propres et les vecteurs propres**, qui permettent d'obtenir les composantes principales.

9.1. Étapes du Calcul

Étape 1 : Construction de la Matrice des Données X

- On commence avec une matrice X de taille $n \times p$, où :
- n est le nombre d'individus (observations).
- p est le nombre de variables quantitatives.
- Si les variables sont sur des échelles différentes, on standardise X pour obtenir la **matrice centrée-réduite Z**.

- **Étape 2 : Calcul de la Matrice de Covariance ou de Corrélation**
- **Si les variables ont la même unité**, on utilise la **matrice de covariance** S :

$$S = \frac{1}{n} Z^T Z$$

- **Si les variables sont sur des échelles différentes**, on préfère la **matrice de corrélation** R :

$$R_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\sigma_i \sigma_j}$$

Étape 3 : Détermination des Valeurs Propres et Vecteurs Propres

On résout l'équation caractéristique : $$\det(A - \lambda I) = 0$$

où :

- A est la matrice de covariance ou de corrélation (S ou R).
- λ est une **valeur propre**.
- I est la matrice identité.
- Les solutions λ_k obtenues sont les **valeurs propres**, qui indiquent la variance expliquée par chaque composante principale.
- Une fois les valeurs propres déterminées, on calcule les **vecteurs propres** V_k associés :

$$(A - \lambda_k I)V_k = 0$$

- Chaque vecteur propre correspond aux **coefficients des variables** pour une composante principale.

2. Exemple de Calcul à la Main

- Soit la matrice de covariance suivante :

$$S = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

- On résout l'équation caractéristique :

$$\det \begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix} = 0$$

- Développement du déterminant :

$$(2-\lambda)(2-\lambda) - 1 \times 1 = 0$$

$$(2-\lambda)^2 - 1 = 0$$

$$4 - 4\lambda + \lambda^2 - 1 = 0$$

$$\lambda^2 - 4\lambda + 3 = 0$$

$$(\lambda - 3)(\lambda - 1) = 0$$

Les valeurs propres sont **λ1=3 et λ2=1**

- Pour λ1=3 :

$$\begin{bmatrix} 2-3 & 1 \\ 1 & 2-3 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = 0$$

$$\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = 0$$

On trouve **v1=(1,1)**.

- Pour λ2=1 :

$$\begin{bmatrix} 2-1 & 1 \\ 1 & 2-1 \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = 0$$

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = 0$$

On trouve **v2=(−1,1)**.

9.2. Implémentation dans R

1. Définition de la Matrice de Covariance

- On prend la même **matrice de covariance** que dans l'exemple précédent : $S = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

2. Implémentation en R

On utilise la fonction eigen() pour calculer les valeurs propres et les vecteurs propres.

```
# Définition de la matrice de covariance
S <- matrix(c(2, 1,
              1, 2),
            nrow = 2, byrow = TRUE)
# Calcul des valeurs propres et des vecteurs propres
resultat <- eigen(S)
# Affichage des valeurs propres
print("Valeurs propres :")
print(resultat$values)
 # Affichage des vecteurs propres
print("Vecteurs propres :")
print(resultat$vectors)
```

Chapitre 1 : Analyse en Composantes Principales (ACP)

3. Explication du Code

- matrix(c(2, 1, 1, 2), nrow = 2, byrow = TRUE): Crée la matrice de covariance S.
- eigen(S): Calcule les valeurs propres et les vecteurs propres.
- resultat$values: Affiche les **valeurs propres**.
- resultat$vectors: Affiche les **vecteurs propres**.

4. Résultats Attendus

L'exécution du code affiche :

Valeurs propres :

```
[1] 3 1
```

Vecteurs propres :

```
          [,1]      [,2]
[1,]  0.707107 -0.707107
[2,]  0.707107  0.707107
```

Ce qui signifie :

- **Valeurs propres** : $\lambda 1=3$ et $\lambda 2=1$, donc la première composante principale capture plus de variance.
- **Vecteurs propres** :
    - CP1 = **(0.7, 0.7)** ⟶ Elle donne le même poids aux deux variables.
    - CP2 = **(-0.7, 0.7)** ⟶ Elle les oppose.

4. Interprétation des Résultats

- Les **valeurs propres** indiquent la quantité de variance capturée par chaque composante principale.
- Les **vecteurs propres** donnent les coefficients des variables initiales dans les nouvelles composantes.
- Dans notre exemple :
- **La première composante principale (CP1)** est une combinaison **égale** des deux variables initiales.
- **La deuxième composante principale (CP2)** oppose les deux variables, ce qui signifie qu’elle distingue les individus selon leur écart entre les deux variables.

5. Application sur des Données Réelles

```
# Charger les données iris
data(iris)
# Supprimer la colonne catégorielle (espèce)
iris_num <- iris[, 1:4]
# Calcul de la matrice de covariance
S_iris <- cov(iris_num)
# Calcul des valeurs propres et vecteurs propres
resultat_iris <- eigen(S_iris)
# Affichage des résultats
print("Valeurs propres de iris :")
print(resultat_iris$values)
print("Vecteurs propres de iris :")
print(resultat_iris$vectors)
```

EXERCICE

$$X = \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix}$$

- Calculer X'X correctement et vérifier qu'il est bien carré et symétrique.
- Déterminer ses valeurs propres λ_i.
- Trouver les vecteurs propres associés.
- Construire la matrice de passage A composée des vecteurs propres normalisés.

RESULTAT :

1. Définition de la Matrice X

- La matrice donnée est : $$X = \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix}$$

- Sa **transposée** est : $$X' = \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 1 \end{bmatrix}$$

Le **produit matriciel** X'X est donc une **matrice carrée 3×3**.

2. Implémentation en R

```
# Définition de la matrice X
X <- matrix(c(-1, 0, 1,
              0, -1, 1),
            nrow = 2, byrow = TRUE)
# Calcul de X'X
XtX <- t(X) %*% X
# Affichage du produit matriciel
print("Produit matriciel X'X :")
print(XtX)
# Vérification de la symétrie
print("Vérification de la symétrie de X'X :")
print(all(XtX == t(XtX)))  # Doit retourner TRUE
# Calcul des valeurs propres et vecteurs propres
resultat <- eigen(XtX)
# Affichage des valeurs propres
print("Valeurs propres de X'X :")
print(resultat$values)
```

```
# Affichage des vecteurs propres
print("Vecteurs propres de X'X :")
print(resultat$vectors)
# Matrice de passage A (vecteurs propres normalisés)
A <- resultat$vectors
print("Matrice de passage A :")
print(A)
# Vérification de la diagonalisation (XtX doit être diagonal dans la base A)
D <- diag(resultat$values)
XtX_reconstruit <- A %*% D %*% solve(A)
print("Vérification de la diagonalisation (A D A^-1) :")
print(XtX_reconstruit)  # Doit être égal à XtX
```

2. Calcul du produit matriciel X'X

- Le produit X'X est donné par :

$$X'X = \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix}$$

- Effectuons les produits élément par élément :

$$X'X = \begin{bmatrix} (-1)(-1)+(0)(0) & (-1)(0)+(0)(-1) & (-1)(1)+(0)(1) \\ (0)(-1)+(-1)(0) & (0)(0)+(-1)(-1) & (0)(1)+(-1)(1) \\ (1)(-1)+(1)(0) & (1)(0)+(1)(-1) & (1)(1)+(1)(1) \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1+0 & 0+0 & -1+0 \\ 0+0 & 0+1 & 0-1 \\ -1+0 & 0-1 & 1+1 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 2 \end{bmatrix}$$

3. Vérification des Propriétés de X'X

- **Matrice carrée** (3×3).
- **Matrice symétrique** car X'X = (X'X)'.

4. Calcul des Valeurs Propres et Vecteurs Propres

Nous cherchons les **valeurs propres** λ en résolvant : $\det(X'X - \lambda I) = 0$

- où I est la matrice identité 3×3.

$$\begin{vmatrix} 1-\lambda & 0 & -1 \\ 0 & 1-\lambda & -1 \\ -1 & -1 & 2-\lambda \end{vmatrix} = 0$$

- Développement du déterminant :

$$(1-\lambda)\begin{vmatrix} 1-\lambda & -1 \\ -1 & 2-\lambda \end{vmatrix} - 0 + 1\begin{vmatrix} 0 & 1-\lambda \\ -1 & -1 \end{vmatrix}$$

- Calcul des déterminants mineurs :

$$\begin{vmatrix} 1-\lambda & -1 \\ -1 & 2-\lambda \end{vmatrix} = (1-\lambda)(2-\lambda) - (-1)(-1) = 2 - 3\lambda + \lambda^2 - 1 = \lambda^2 - 3\lambda + 1$$

$$\begin{vmatrix} 0 & 1-\lambda \\ -1 & -1 \end{vmatrix} = 0(-1) - (1-\lambda)(-1) = \lambda - 1$$

- Donc l'équation caractéristique devient :

$$(1-\lambda)(\lambda^2 - 3\lambda + 1) + (\lambda - 1) = 0$$

$$\lambda^3 - 3\lambda^2 + \lambda - \lambda + 1 = 0$$

$$\lambda^3 - 3\lambda^2 + 2\lambda = 0$$

- Factorisation :

$$\lambda(\lambda^2 - 3\lambda + 2) = 0$$

$$\lambda(\lambda - 1)(\lambda - 2) = 0$$

- Les **valeurs propres** sont donc :

$$\lambda_1 = 2, \quad \lambda_2 = 1, \quad \lambda_3 = 0$$

5. Calcul des Vecteurs Propres

On résout (X'X−λI)v=0 pour chaque λ.

- **Pour λ1=2**:

$$\begin{bmatrix} -1 & 0 & -1 \\ 0 & -1 & -1 \\ -1 & -1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Solution normalisée : $v_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$

- **Pour λ2=1** :

$$\begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & -1 \\ -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Chapitre 1 : Analyse en Composantes Principales (ACP)

Solution normalisée : $v_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$

Pour $\lambda_3 = 0$ **:**

$$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Solution normalisée : $v_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

6. Matrice de Passage A

$$A = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

10. Choix de la distance entre deux individus dans une ACP

Dans une **Analyse en Composantes Principales (ACP)**, la notion de **distance entre individus** est essentielle pour analyser les similarités et structurer les données. Le choix de la distance dépend du type d'ACP réalisée et des caractéristiques des variables.

10.1. Distance Euclidienne (Classique)

La **distance Euclidienne** est la plus couramment utilisée lorsque les variables sont **quantitatives** et **centrées-réduites** (ce qui est souvent le cas en ACP) :

$$d(i,j) = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2}$$

- **Avantages :**

Simple et intuitive.

Utilisée par défaut en ACP.

- **Inconvénients :**

Sensible aux différences d'échelles (c'est pourquoi on standardise souvent les données avant ACP).

10.2. Distance de Mahalanobis

- La **distance de Mahalanobis** prend en compte la structure de corrélation des données :

$$d(i,j) = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)}$$

où S est la matrice de covariance des variables.

- **Avantages :**

Corrige l'effet des corrélations entre variables.

Mieux adaptée lorsque certaines dimensions sont plus importantes que d'autres.

- **Inconvénients :**

Nécessite l'inversion de la matrice de covariance, ce qui peut être problématique si elle est mal conditionnée (si certaines variables sont très corrélées).

10.3. Distance du Khi-Deux (Pour les Données Catégoriques)

- Utilisée en **Analyse des Correspondances (AC)** (analogue de l'ACP pour des variables qualitatives). Elle compare les profils de fréquences attendues et observées :

$$d(i,j) = \sum_k \frac{(X_{ik} - X_{jk})^2}{X_{ik} + X_{jk}}$$

- **Avantages :**

Adaptée aux **tableaux de contingence** et données qualitatives.

- **Inconvénients :**

Non applicable aux données quantitatives classiques.

10.4. Distance Cosinus (ou Similarité Cosinus)

- Mesure l'**angle** entre deux vecteurs plutôt que la distance en ligne droite :

$$d(i,j) = 1 - \frac{X_i \cdot X_j}{\|X_i\| \times \|X_j\|}$$

- **Avantages :**

Utile lorsque l'**amplitude** des valeurs importe moins que leur **direction** (par ex. en **traitement du langage** ou sur des données haute dimension).

- **Inconvénients :**

Moins intuitive qu'une distance Euclidienne.

10.5. Quel Choix pour une ACP ?

Type de Données	Distance Recommandée
Quantitatives, normalisées	Euclidienne
Corrélées, avec variances très différentes	Mahalanobis
Qualitatives (catégoriques)	Khi-Deux
Données en grande dimension (ex. texte)	Cosinus

11. Représentation des Variables dans une Analyse en Composantes Principales (ACP)

Dans une ACP, les **variables** sont projetées dans un nouvel espace défini par les **composantes principales**. Cette représentation est essentielle pour interpréter les relations entre variables et comprendre leur contribution aux axes factoriels.

11.1. Cercle des Corrélations

La représentation graphique des variables dans l'ACP se fait généralement à l'aide du **cercle des corrélations** :

📌 **Définition :**

- Il s'agit d'une projection des variables initiales dans le **nouveau repère** formé par les axes principaux.
- Il permet d'analyser **l'association entre les variables** et leur contribution aux axes factoriels.

Chapitre 1 : Analyse en Composantes Principales (ACP)

Interprétation des coordonnées des variables :

- Plus une variable est proche du bord du cercle, plus elle est **bien représentée**.
- Une variable située près d'un axe contribue fortement à cet axe.
- Des variables proches l'une de l'autre sont **corrélées positivement**.
- Des variables opposées (≈180°) sont **corrélées négativement**.
- Des variables perpendiculaires sont **non corrélées**.

11.2. Contribution des Variables aux Axes Principaux

Chaque variable contribue différemment à la construction des axes. La **contribution d'une variable** à un axe est mesurée par son **cos²** et son **poids dans la variance expliquée**.

📌 Règles d'interprétation :

- **Cos² élevé (proche de 1)** → La variable est bien représentée par l'axe.
- **Cos² faible (proche de 0)** → La variable est peu expliquée par l'axe et peut être mieux représentée par un autre.
- **Valeurs propres élevées** → L'axe explique une part importante de l'information.

🔍 À vérifier après l'ACP :

- Sélection des variables qui contribuent fortement aux premiers axes.
- Éventuelle transformation ou suppression des variables mal représentées.

11.3. Exemple d'Implémentation en R

On peut visualiser le **cercle des corrélations** avec FactoMineR et factoextra :

```
# Charger les packages nécessaires
library(FactoMineR)
library(factoextra)
# Générer un jeu de données fictif
set.seed(123)
data <- data.frame(
Taille = rnorm(100, mean = 170, sd = 10),
Poids = rnorm(100, mean = 70, sd = 15),
Age = rnorm(100, mean = 40, sd = 12),
Score = rnorm(100, mean = 50, sd = 10)
)
# Effectuer l'ACP
res.pca <- PCA(data, scale.unit = TRUE, graph = FALSE)
# Cercle des corrélations
fviz_pca_var(res.pca, col.var = "cos2", repel = TRUE)
```

11.4. Applications et Cas d’Utilisation

- **Marketing** : Segmentation des clients en fonction de critères comportementaux.
- **Agriculture** : Classification de variétés de plantes en fonction de caractéristiques morphologiques.
- **Médecine** : Réduction de la dimensionnalité dans des bases de données médicales (ex : biomarqueurs).

12. Analyse en Composantes Principales (ACP) Normée : Procédure et Interprétation

L'**ACP normée** est une version de l'Analyse en Composantes Principales où les **variables sont standardisées** avant l'analyse. Cette standardisation permet d'éviter que certaines variables à grande variance dominent l'analyse.

12.1. Pourquoi une ACP Normée ?

- L'ACP classique utilise la **matrice de variance-covariance**. Si les variables ont des échelles très différentes, celles ayant une **grande variance** influenceront davantage les résultats.
- **Problème :** Une variable mesurée en **euros** et une autre en **centimètres** ne peuvent être directement comparées.
- **Solution : Standardisation des données** :

$$X'_{ij} = \frac{X_{ij} - \bar{X}_j}{\sigma_j}$$

où $\bar{X}_j$ est la moyenne et σ_j l'écart-type de la variable j.

Conséquence :

- Toutes les variables ont une **moyenne de 0** et un **écart-type de 1**.
- L'ACP est basée sur la **matrice des corrélations** au lieu de la matrice de variance-covariance.

12.2. Procédure de l'ACP Normée

Étapes principales :

- **Standardisation des données**

Transformation pour avoir une moyenne 0 et un écart-type 1.

- **Calcul de la matrice de corrélation**

Utilisée à la place de la matrice de variance-covariance.

- **Calcul des valeurs propres et des vecteurs propres**

- Les valeurs propres représentent la **variance expliquée** par chaque axe principal.
- Les vecteurs propres définissent les **composantes principales**.

- **Projection des individus et variables**

- Calcul des coordonnées des individus dans l'espace des nouvelles composantes.
- Construction du **cercle des corrélations** pour analyser les relations entre variables.

12.3. Nombre d'Axes à Retenir

Le choix du **nombre de composantes principales** est crucial pour ne pas perdre trop d'information ni en conserver inutilement.

- **Critères de sélection :**

Critère du Pourcentage de Variance

Retenir les axes qui expliquent **au moins 70-80% de la variance totale**.

Critère de Kaiser

Conserver les composantes ayant une **valeur propre supérieure à 1**.

Critère du Coude (Scree Plot)

Observer le graphique des valeurs propres et repérer le **point d'inflexion** (où la décroissance devient plus lente).

Contribution des Variables

Vérifier si les variables principales sont bien représentées sur les premiers axes.

12.4. Qualités et Défauts de l'ACP Normée

Avantages	Inconvénients
Permet de comparer des variables de nature différente	Perte d'information sur les variances initiales
Évite qu'une variable domine l'analyse	Résultats parfois moins interprétables
Applicable à des variables fortement corrélées	Ne fonctionne pas bien avec des données discrètes ou qualitatives
Facilite l'interprétation graphique (Cercle des corrélations, biplots, etc.)	Hypothèse de linéarité entre les variables (peu adapté aux relations complexes)

12.5. ETUDE DE CAS

Pour cette matrice, déterminer la matrice des données centrées et réduites (normées) Y.

- Déterminer la matrice des corrélations "tau".
- Diagonaliser la matrice "tau" et déterminer ses valeurs propres.
- Déterminer les vecteurs propres associés aux valeurs propres

$$X = \begin{bmatrix} 4 & 5 \\ 6 & 7 \\ 8 & 0 \end{bmatrix}$$

2. Centrage et Réduction : Matrice YYY des données normées

- **Centrage** : On soustrait la moyenne de chaque colonne.
- **Réduction** : On divise par l'écart-type de chaque colonne.
- **Calcul de la Moyenne et de l'Écart-Type**

$$\bar{X}_1 = \frac{4+6+8}{3} = 6, \quad \bar{X}_2 = \frac{5+7+0}{3} = 4$$

$$\sigma_1 = \sqrt{\frac{(4-6)^2+(6-6)^2+(8-6)^2}{3}} = \sqrt{\frac{4+0+4}{3}} = \sqrt{\frac{8}{3}} = \frac{2\sqrt{2}}{3}$$

$$\sigma_2 = \sqrt{\frac{(5-4)^2+(7-4)^2+(0-4)^2}{3}} = \sqrt{\frac{1+9+16}{3}} = \sqrt{\frac{26}{3}}$$

Matrice Normée Y

$$Y = \begin{bmatrix} \frac{4-6}{\sigma_1} & \frac{5-4}{\sigma_2} \\ \frac{6-6}{\sigma_1} & \frac{7-4}{\sigma_2} \\ \frac{8-6}{\sigma_1} & \frac{0-4}{\sigma_2} \end{bmatrix}$$

$$Y = \begin{bmatrix} \frac{-2}{\frac{2\sqrt{2}}{3}} & \frac{1}{\sqrt{26/3}} \\ 0 & \frac{3}{\sqrt{26/3}} \\ \frac{2}{\frac{2\sqrt{2}}{3}} & \frac{-4}{\sqrt{26/3}} \end{bmatrix}$$

$$Y = \begin{bmatrix} -\frac{3}{\sqrt{2}} & \frac{\sqrt{3}}{\sqrt{26}} \\ 0 & \frac{3\sqrt{3}}{\sqrt{26}} \\ \frac{3}{\sqrt{2}} & -\frac{4\sqrt{3}}{\sqrt{26}} \end{bmatrix}$$

3. Matrice de Corrélation τ

- La matrice de corrélation est définie par :

$$\tau = \frac{1}{n} Y^T Y$$

- Comme les donr ardisées, τ correspond directement à la **matrice de corrélation empirique**. En cal matriciel :

$$\tau = \begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix}$$

où $r_{12}=r_{21}$ est le coefficient de corrélation entre les deux variables

4. Diagonalisation de τ

- Nous devons résoudre l'équation caractéristique : $\det(\tau - \lambda I) = 0$

Les valeurs propres λ_1, λ_2 sont obtenues en résolvant :

$$\begin{vmatrix} 1-\lambda & r_{12} \\ r_{12} & 1-\lambda \end{vmatrix} = 0$$

Ce qui donne :

$$(1-\lambda)^2 - r_{12}^2 = 0$$

$$(1-\lambda) = \pm r_{12}$$

Donc les valeurs propres sont :

$$\lambda_1 = 1 + r_{12}, \quad \lambda_2 = 1 - r_{12}$$

Vecteurs propres associés

- Les vecteurs propres sont obtenus en résolvant :

$$(\tau - \lambda I)v = 0$$

- Pour chaque valeur propre λ_i, on trouve v_i.

Chapitre 1 : Analyse en Composantes Principales (ACP)

- **5. Implémentation en R**

```
# Définition de la matrice des données
X <- matrix(c(4, 5, 6, 7, 8, 0), nrow = 3, byrow = TRUE)
# Standardisation des données
X_norm <- scale(X)
# Calcul de la matrice de corrélation
tau <- cor(X_norm)
# Diagonalisation de tau
eig <- eigen(tau)
# Affichage des résultats
print("Matrice centrée et réduite Y :")
print(X_norm)
print("Matrice de corrélation tau :")
print(tau)
print("Valeurs propres de tau :")
print(eig$values)
print("Vecteurs propres de tau :")
print(eig$vectors)
```

6. Résolution analytique

- Nous allons effectuer **pas à pas** la résolution analytique du problème :
- Centrage et réduction de la matrice
- Calcul de la matrice de corrélation т
- Diagonalisation de т: détermination des valeurs propres
- Détermination des vecteurs propres

6.1. Matrice des Données et Standardisation

- La matrice initiale est :

$$X = \begin{bmatrix} 4 & 5 \\ 6 & 7 \\ 8 & 0 \end{bmatrix}$$

Chapitre 1 : Analyse en Composantes Principales (ACP)

Calcul des Moyennes des Colonnes

$$\bar{X}_1 = \frac{4+6+8}{3} = 6, \quad \bar{X}_2 = \frac{5+7+0}{3} = 4$$

Calcul des Écarts-types des Colonnes

$$\sigma_1 = \sqrt{\frac{(4-6)^2+(6-6)^2+(8-6)^2}{3}} = \sqrt{\frac{4+0+4}{3}} = \frac{2\sqrt{2}}{3}$$

$$\sigma_2 = \sqrt{\frac{(5-4)^2+(7-4)^2+(0-4)^2}{3}} = \sqrt{\frac{1+9+16}{3}} = \sqrt{\frac{26}{3}}$$

$$Y_{ij} = \frac{X_{ij} - \bar{X}_j}{\sigma_j}$$

Matrice Centrée et Réduite Y

- On applique la transformation :

$$Y = \begin{bmatrix} \frac{4-6}{\sigma_1} & \frac{5-4}{\sigma_2} \\ \frac{6-6}{\sigma_1} & \frac{7-4}{\sigma_2} \\ \frac{8-6}{\sigma_1} & \frac{0-4}{\sigma_2} \end{bmatrix}$$

$$Y = \begin{bmatrix} -\frac{3}{\sqrt{2}} & \frac{\sqrt{3}}{\sqrt{26}} \\ 0 & \frac{3\sqrt{3}}{\sqrt{26}} \\ \frac{3}{\sqrt{2}} & -\frac{4\sqrt{3}}{\sqrt{26}} \end{bmatrix}$$

2. Calcul de la Matrice de Corrélation τ

- La matrice de corrélation est donnée par :

$$\tau = \frac{1}{n} Y^T Y$$

Le produit $Y^T Y$ donne :

$$Y^T Y = \begin{bmatrix} Y_1^T Y_1 & Y_1^T Y_2 \\ Y_2^T Y_1 & Y_2^T Y_2 \end{bmatrix}$$

- Calculons chaque terme :

$$Y_1^T Y_1 = \left(-\frac{3}{\sqrt{2}}\right)^2 + 0^2 + \left(\frac{3}{\sqrt{2}}\right)^2 = \frac{9}{2} + 0 + \frac{9}{2} = 9$$

$$Y_2^T Y_2 = \left(\frac{\sqrt{3}}{\sqrt{26}}\right)^2 + \left(\frac{3\sqrt{3}}{\sqrt{26}}\right)^2 + \left(-\frac{4\sqrt{3}}{\sqrt{26}}\right)^2$$

$$= \frac{3}{26} + \frac{27}{26} + \frac{48}{26} = 1$$

$$Y_1^T Y_2 = \left(-\frac{3}{\sqrt{2}} \cdot \frac{\sqrt{3}}{\sqrt{26}}\right) + \left(0 \cdot \frac{3\sqrt{3}}{\sqrt{26}}\right) + \left(\frac{3}{\sqrt{2}} \cdot -\frac{4\sqrt{3}}{\sqrt{26}}\right)$$

$$= -\frac{3\sqrt{3}}{\sqrt{52}} - \frac{12\sqrt{3}}{\sqrt{52}}$$

$$= -\frac{3\sqrt{3}}{\sqrt{52}} - \frac{12\sqrt{3}}{\sqrt{52}}$$

$$= -\frac{15\sqrt{3}}{\sqrt{52}}$$

La matrice de corrélation τ est :

$$\tau = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$$

où $r_{12} = \frac{15\sqrt{3}}{\sqrt{52}}$.

3. Diagonalisation de τ : Calcul des Valeurs Propres

- On résout :

$$\det(\tau - \lambda I) = 0$$

$$\begin{vmatrix} 1-\lambda & r_{12} \\ r_{12} & 1-\lambda \end{vmatrix} = 0$$

Développement :

$$(1-\lambda)^2 - r_{12}^2 = 0$$

$$\lambda^2 - 2\lambda + (1 - r_{12}^2) = 0$$

Résolution :

$$\lambda = \frac{2 \pm \sqrt{4 - 4(1 - r_{12}^2)}}{2} = \frac{2 \pm \sqrt{4 - 4 + 4r_{12}^2}}{2}$$

$$= \frac{2 \pm 2r_{12}}{2} = 1 \pm r_{12}$$

4. Calcul des Vecteurs Propres

Pour chaque valeur propre λ_i, on résout :

$$(\tau - \lambda I)v = 0$$

Pour $\lambda_1 = 1 + r_{12}$:

$$\begin{bmatrix} 1-(1+r_{12}) & r_{12} \\ r_{12} & 1-(1+r_{12}) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0$$

$$\begin{bmatrix} -r_{12} & r_{12} \\ r_{12} & -r_{12} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0$$

En résolvant, on trouve v_1. De même pour $\lambda_2 = 1 - r_{12}$.