Retour à l'analyse géométrique des données

→ **ACP : Analyse en Composantes Principales**

Tableau Individus × Variables numériques

→ **AC : Analyse des Correspondances simple**

Tableau de contingence (croisement de deux variables catégorielles)

→ **ACM : Analyse des Correspondances Multiples**

Tableau Individus × Variables catégorielles

Les principes/objectifs de l'Analyse géométrique des données

Soit un tableau de données avec n individus et p variables.
Il peut être représenté par la matrice :

$$X(n,p) = \begin{pmatrix} x_1^1 & \dots & \dots & \dots & x_1^p \\ x_i^1 & \dots & x_i^j & \dots & x_i^p \\ x_n^1 & \dots & \dots & \dots & x_n^p \end{pmatrix}$$

- Un individu peut être décrit par les p variables, donc dans p dimensions.
  Le nuage de points des individus intervient dans un espace de p dimensions.
  p est susceptible d'être élevé.

- Une variable peut être décrite par les n individus, donc dans n dimensions.
  Le nuage de points des variables intervient dans un espace de n dimensions.
  Le plus souvent en sciences sociales, n est élevé.

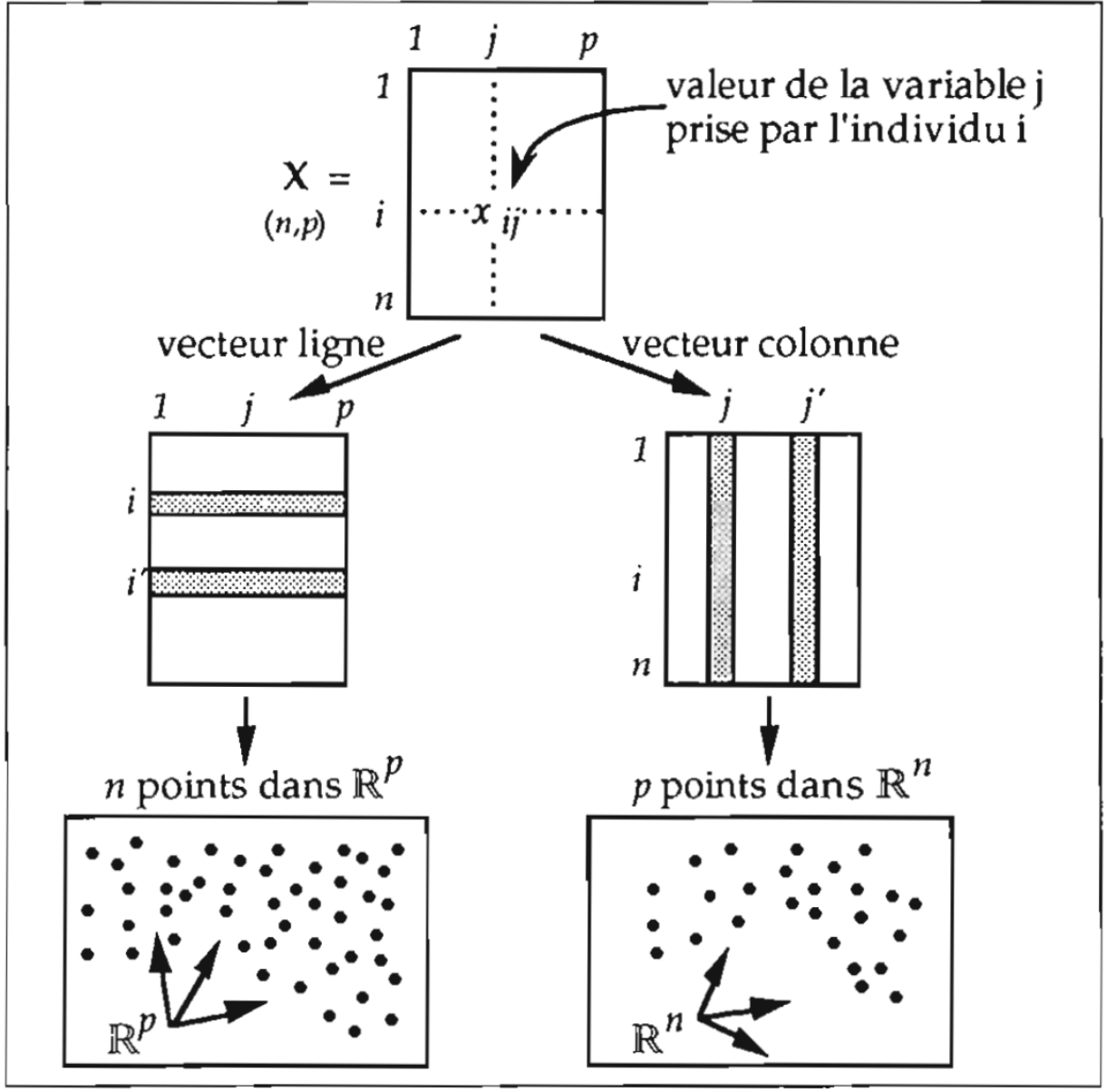


Figure 1
Principe de représentation géométrique

Source : Lebart, Morineau, Piron, Statistiques exploratoires multidimensionnelles, p. 24. https://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers11-10/010007837.pdf

Principe :

Calcul des distances entre les lignes

Calcul des distances entre les colonnes

Difficulté d'un tel tableau de données X(n,p) :

Dès que p>3, impossibilité de visualiser les relations entre l'ensemble des variables.

Une solution mise en œuvre par l'AGD :

Construire un sous-espace de dimension inférieure, donc « visualisable » & optimal, de façon à :

- Synthétiser l'information
- Identifier les liens les plus significatifs contenus dans le tableau
- Limiter la perte d'information (donc déformer le moins possible l'information), ou conserver un maximum de la variabilité du tableau de données initial

➔ **Changement de repère autour du nuage de points** qui soit le moins déformant possible.

Les nouveaux axes ainsi obtenus, **appelés axes factoriels**, doivent permettre de repérer les structures les plus prégnantes de sa population.

A la recherche de ces axes factoriels…

ie trouver le changement d'axes « optimal » (maximisant la variance) autour du nuage de points.

Arrière plan mathématique :

- Algèbre linéaire et calcul matriciel (transposition & diagonalisation de matrice, somme et produit de matrices…)
- Utilisation des propriétés algébriques et des opérateurs mathématiques qui permettent, à partir de la matrice représentant les données, de trouver les axes de projection optimaux.
- Ce que font les logiciels : diagonaliser la matrice pour obtenir les axes (valeurs propres, vecteurs propres…)

Les étapes de la démarche pour le chercheur :

Choix des variables pertinentes, codage, préparation des données, lecture statistique des résultats obtenus à partir du logiciel et INTERPRETATION (interprétation des axes, analyse de la structure du nuage de points sur un axe, dans un plan factoriel…)

A la recherche des axes factoriels…

1/ On recherche le meilleur axe, ie celui sur lequel le nuage se déforme le moins possible en projection (maximisation de la variance). On dit que ce nuage projeté est d'inertie maximale.

Cet axe, appelé axe principal, détermine une nouvelle direction dans le nuage de points qui suit l'axe d'allongement (d'étirement) maximal du nuage.

2/ On cherche le deuxième axe, orthogonal au premier, sur lequel le nuage se déforme le moins possible en projection une fois tenu compte de la variance prise en compte par le premier axe.

3/ …

A la recherche des axes factoriels…

Propriétés

1/ Si les axes sont appelés C_k, chacun d'eux est une combinaison linéaire des p variables d'origine.

$C_k = a_{1k}X_1 + a_{2k}X_2 + \ldots + a_{pk}X_p$

→ Coefficients a_{jk} à déterminer

2/ Les axes sont caractérisés par des valeurs propres (notées λ - variances du nuage projeté sur l'axe) et des vecteurs propres (directions d'allongement maximal du nuage)

3/ La première composante (le premier axe) est de variance maximale (λ_1 - valeur propre maximale). C'est la droite qui conserve au mieux la distance entre les points (après projection)

4/ La part de variance (%) dûe à un axe par rapport à la variance totale est appelée taux d'inertie de l'axe.

5/ On peut démontrer que le sous-espace optimal passe par le point G (point moyen du nuage).

A la recherche des axes factoriels…

On pourrait démontrer que l'orthogonalité des axes principaux successifs aboutit aux propriétés remarquables suivantes :

- ils sont 2 à 2 non corrélés (corrélation nulle). Il y a donc indépendance entre les dispersions projetées observées.
- ils permettent, ensemble, de reconstituer par additivité la variance totale du nuage.
- ils sont d'importance décroissante en termes de part de variance expliquée (taux d'inertie).

Remarques

1) L'orientation des axes est arbitraire. En effet, les vecteurs propres sont définis au signe près. La figure 1.1 - 5, concernant trois points, montre que toutes les images, obtenues suivant des orientations différentes des facteurs, respectent la forme du nuage c'est-à-dire les distances entre les points.

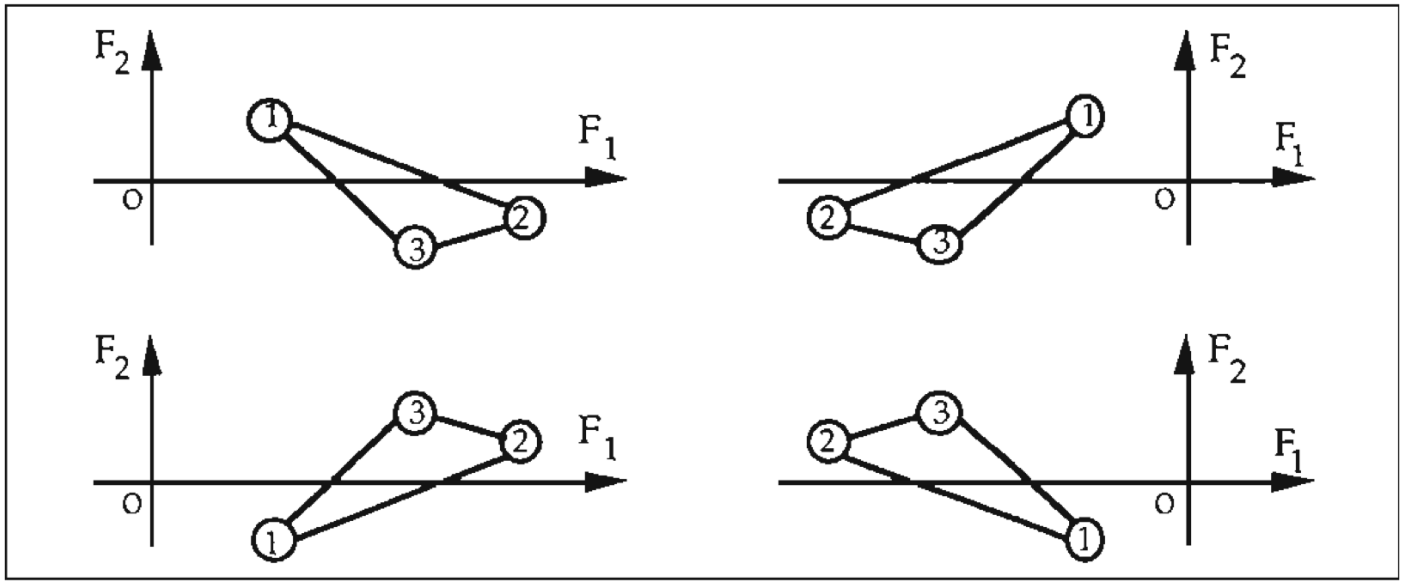


Figure 1.1 - 5
Orientation arbitraire des axes

Source : Lebart, Morineau, Piron, Statistiques exploratoires multidimensionnelles, p. 24.
https://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers11-10/010007837.pdf

Bibliographie théorique

- Benzécri J.P., Benzécri F., 1980, *Pratique de l'analyse des données: Analyse des correspondances, exposé élémentaire*, Paris, Dunod
- Lebart L., Morineau A., Piron M., 2000, *Statistique exploratoire multidimensionnelle*, Paris, Dunod
- Le Roux B. Rouanet H., 20004, *Geometric Data analysis: From correspodance analysis to structured data analysis*, Kluwer academic publishers
- Rouanet H., Le Roux B., 1993, *Analyse des données multidimensionnelles*, Paris, Dunod
- Saporta G., 1990, *Probabilités, analyse des données et statistiques*, technip

L'analyse en composantes principales (ACP)

$$X(n,p) = \begin{pmatrix} x_1^1 & \dots & \dots & \dots & x_1^p \\ x_i^1 & \dots & x_i^j & \dots & x_i^p \\ x_n^1 & \dots & \dots & \dots & x_n^p \end{pmatrix}$$

➔ **Les p variables sont continues**

L'ACP standard (normée)

➔**Les p variables sont standardisées (centrées-réduites) :**

variables de moyenne 0 et de variance 1 : $\frac{x_i - \bar{x}}{\sigma_x}$

Elimination des effets arbitraires liés à l'unité de mesure : ramène à la même échelle les variables. Chaque variable a le même poids dans l'analyse et donc la même contribution à la variance du nuage.

ACP Standard

- Obtenir les valeurs propres et les vecteurs propres des axes principaux :

  → Diagonalisation de la matrice des corrélations des p variables

- Chaque valeur propre mesure la part de variance expliquée par l'axe factoriel correspondant

- La somme de l'ensemble des valeurs propres correspond à la variance totale

- La variance totale est égale au nombre de variables actives de l'analyse :

$$\sigma^2 = \sum_{i=1}^{p} \lambda_i$$

Propriété de l'ACP Standard

- Variance du nuage (inertie totale du nuage)= somme des valeurs propres = nombre de variables

- **Nuage des variables** : le Cercle des corrélations

  la coordonnée d'une variable sur l'axe = sa corrélation à l'axe

  (il s'agit des coefficients de régression de la combinaison linéaire de l'axe avec les variables initiales)

  → Interprétation directe de l'axe
- Variables supplémentaires continues : corrélation à l'axe
- Variables supplémentaires qualitatives : corrélation des barycentres des m modalités à l'axe (projection)

- **Nuage des individus** :

$$x_l = \sum_{j=1}^{p} \left\| l_j \right\| \frac{x_{ij} - \bar{x}_j}{\sigma_j \sqrt{n}}$$

- Les individus supplémentaires : transformation du point supplémentaire (centré-réduit) et projection sur les axes

Propriété de l'ACP Standard

- Nuage des individus : qu'est-ce que 2 individus proches ?

  → Valeurs proches sur l'ensemble des variables

$$d(i,i') = \sum_{j=1}^{p} (x_i - x_{i'})^2 \qquad \text{(distance euclidienne)}$$

  Deux individus ayant exactement le même profil de réponse ont une distance nulle

- Nuage des variables : qu'est-ce que 2 variables proches ?

  → Valeurs proches sur l'ensemble des individus

- ATTENTION : les deux nuages ne se superposent pas.

Représentation simultanée impossible car 2 espaces différents.

Démarche

1/ Déterminer le nombre d'axes à conserver pour l'interprétation (diagramme des valeurs propres et règle du coude)

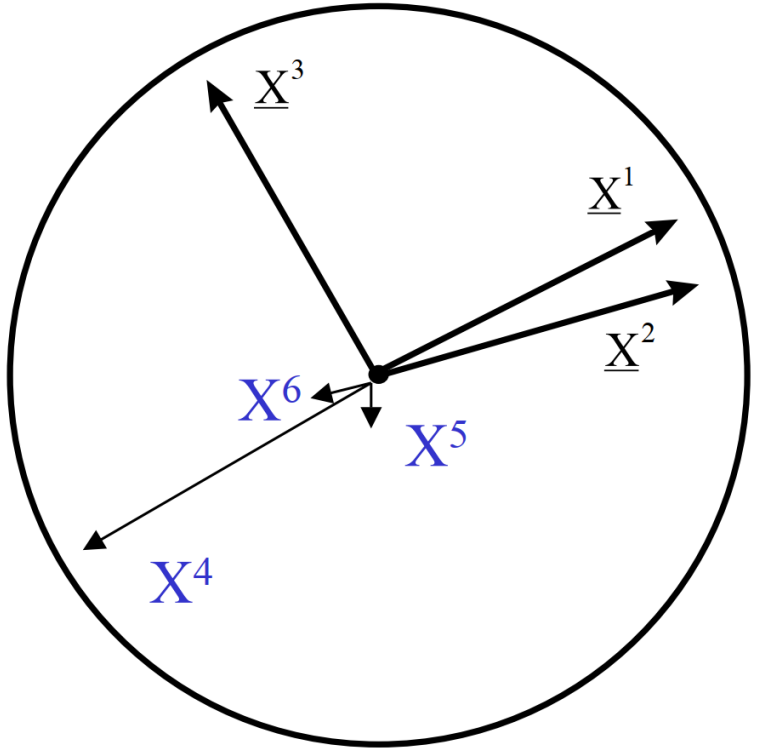
2/ Interprétation des axes 1 à 1 en ne tenant compte que des variables qui leur sont fortement corrélées / construction du cercle des corrélations

3/ Construction des nuages des individus et des variables illustratives

Aide à l'interprétation

X^1 et X^2 ont une corrélation proche de 1.

X^1 et X^3 ont une corrélation proche de 0.

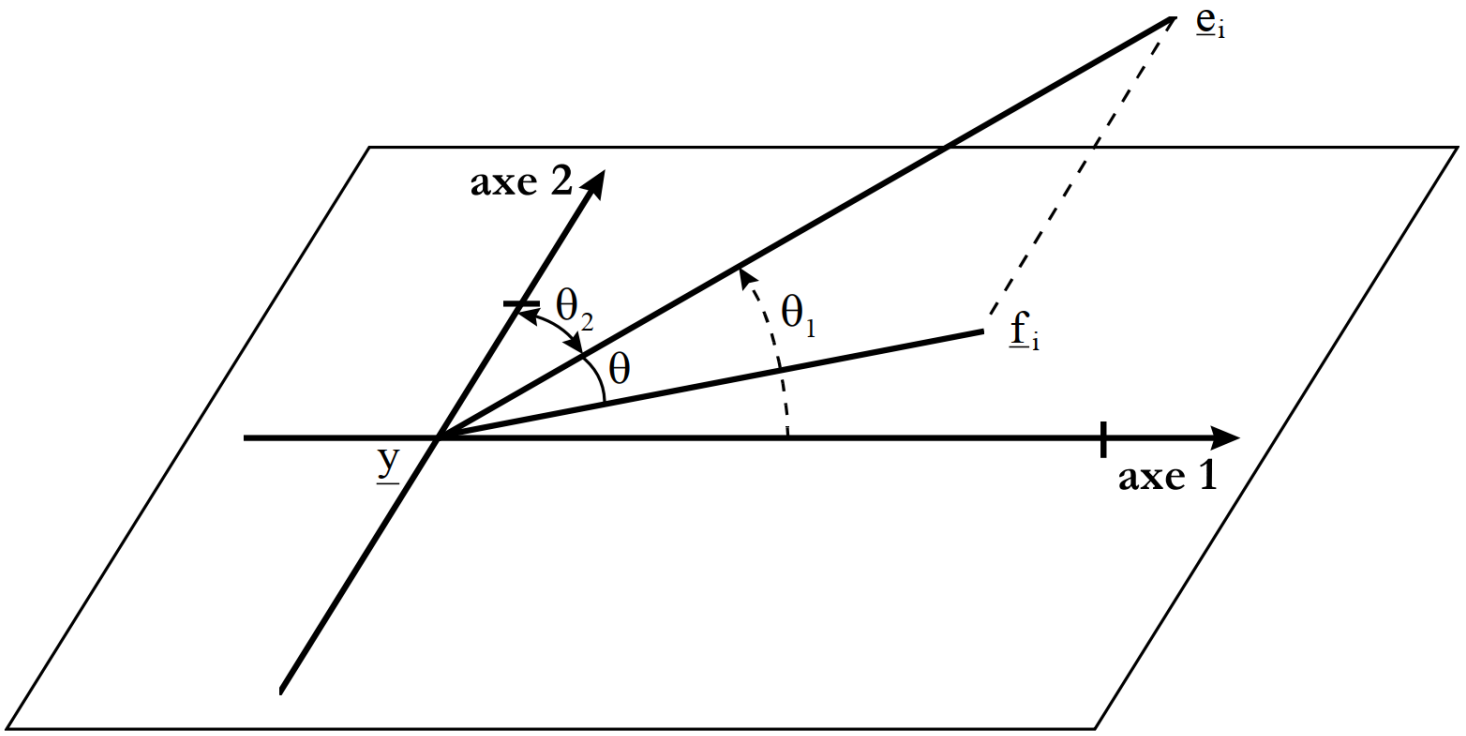


CERCLE DES CORRÉLATIONS

Le cosinus de l'angle formé par les variables X^i et X^j est le coefficient de corrélation linéaire de ces deux variables

Aide à l'interprétation

Cosinus carrés



$$\cos^2\theta = \cos^2\theta_1 + \cos^2\theta_2$$

Aide à l'interprétation

Valeurs-tests (VT)

Dans l'hypothèse d'un tirage au hasard, la VT d'une catégorie supplémentaire a 95% de chances d'être comprise dans l'intervalle (-1,96 ; 1,96).

(test loi Normale)

Une modalité dont la position est significativement différente de l'origine doit avoir une VT en-dehors de cet intervalle