Algèbre Linéaire
et
Analyse de Données
Licence 2 - MIASHS

Guillaume Metzler

Université Lumière Lyon 2
Laboratoire ERIC, UR 3083, Lyon

guillaume.metzler@univ-lyon2.fr

Printemps 2022

Analyse de Données

Summary

Introduction I

Objectif : réduction de dimension en partant d'espaces de dimensions n ou p.

Synthétiser l'information en adoptant une représentation des données dans un espace de dimension 2 voire 3, permettant de **visualiser** les informations contenues dans les données.

Nous utilisons surtout les notions suivantes :

- la notion de distances entre des points, les projections orthogonales et la notion de métrique,
- la recherche de valeurs propres d'un endomorphisme et ses vecteurs propres.

Introduction II

Dans ce qui suit : n désignera **le nombre d'individus** dans notre échantillon (ou le nombre d'exemples) et p **le nombre de descripteurs** pour un exemple donné (*i.e.* le nombre de variables).

$$X = \begin{array}{c} \\ \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_i \\ \vdots \\ \mathbf{x}_n \end{array} \begin{array}{c} \begin{array}{cccccc} \mathbf{v}_1 & \cdots & \mathbf{v}_k & \cdots & \mathbf{v}_p \end{array} \\ \begin{pmatrix} x_{11} & \cdots & x_{1k} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} & \cdots & x_{np} \end{pmatrix} \end{array},$$

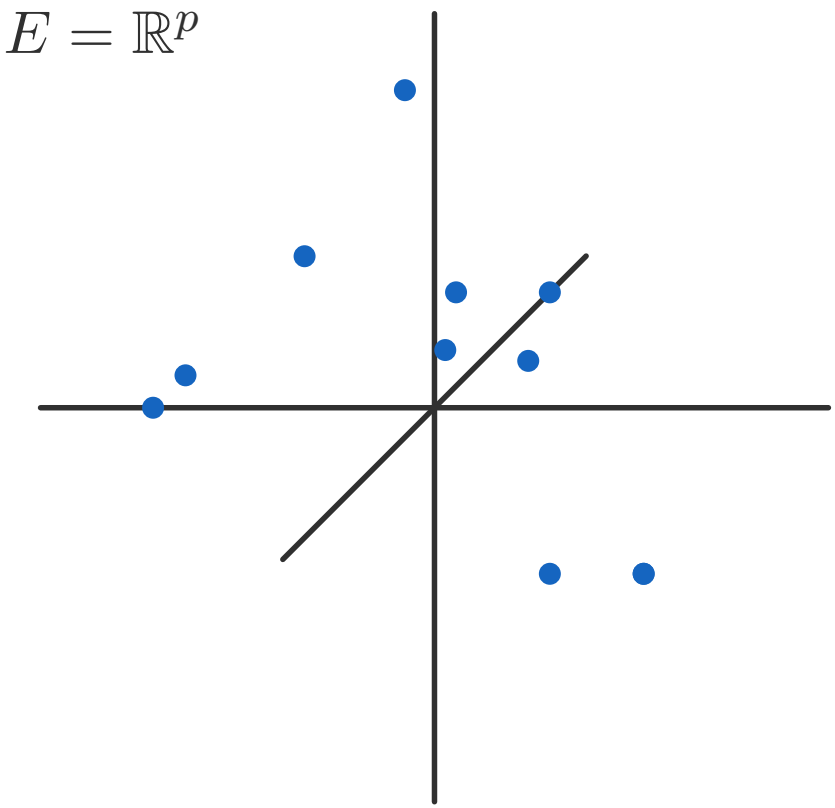
où $\mathbf{x}_i$ représente l'individu i avec les valeurs $\mathbf{x}_{ik}$ prises par les différents descripteurs $\mathbf{v}_k$.

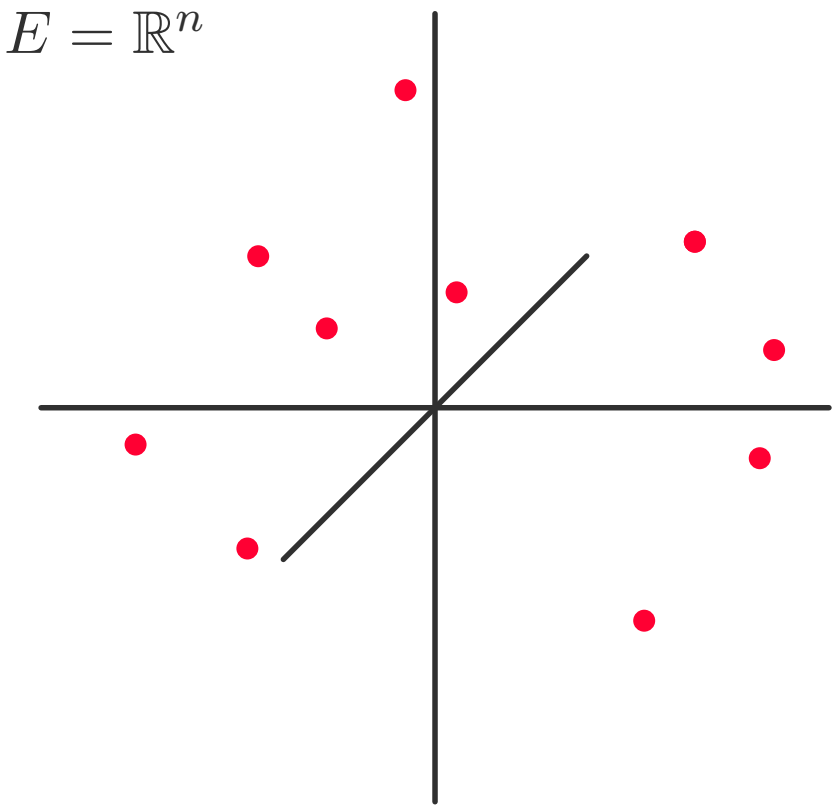
Introduction III

A partir de ce simple tableaux de données, il est possible d'adopter deux représentations

- On peut choisir de représenter les individus $\mathbf{x}_i$ dans l'espaces des variables $\mathbf{v}_k$, une première représentation qui est sûrement la plus utilisée. Dans ce cas chaque point $\mathbf{x}_i$ a pour coordonnées $(\mathbf{x}_{i1}, \cdots, \mathbf{x}_{ip})$ dans l'espace $\mathbb{R}^p$.
  On obtient un premier nuage de points que l'on notera **nuage des individus**.
- On peut également faire le choix de représenter les variables dans l'espace des individus. Dans ce cas chaque point $\mathbf{v}_k$ a pour coordonnées $(\mathbf{v}_{1k}, \cdots, \mathbf{v}_{nk})$ dans l'espace $\mathbb{R}^n$.
  Ce deuxième nuage de points est appelé **nuage des variables**.

Introduction IV





Introduction V

On montrera qu'il existe un lien très fort entre ces deux représentations et que ce dernier repose sur la décomposition en valeurs singulières de cette matrice de données.
L'objectif de cette partie est de fournir des réponses à des questions comme

- Est-ce que des variables sont corrélées entre elles ?
- Quelles directions de l'espace permettent d'expliquer au mieux la variabilité observée au sein des données ?
- Est-ce qu'il est possible d'obtenir une représentation fiable de nos données dans un espace de dimension faible afin de visualiser les informations ? Quel serait le sens de cette nouvelle présentation ?
- Est-ce qu'il existe des groupes d'individus dont le comportement pourrait être expliqué par des variables particulières ?

Introduction VI

Bien évidemment, ces questions de représentations vont se limiter aux espaces de dimension 2 et 3 pour les aspects visualisation. Nous verrons aussi comment mesurer la perte de l'information lors de cette étape de projection.

Les techniques de réduction de dimension étudiées dans cette partie sont :

- l'Analyse en Composantes Principales (ACP),
- l'Analyse Factorielle des Correspondance (AFC),
- l'Analyse factorielle des Correspondances Multiples (ACM).

Analyses de Données

Généralités et Décomposition en Valeurs Singulières (SVD)

Généralités I

Lorsque l'on étudie des données, nous sommes amenés, le plus souvent, à nous intéresser à deux choses :

- l'analyse des corrélations entre les variables
- l'analyse des distances entre les individus

Ces deux critères recherches nous permettent de voir si notre jeu de données est riche en information. Pour quantifier cette information dans un nuage de points composé de n individus dans un espace de dimension p, on va mesurer la **variance**, notée Var_{tot} qui se trouve dans ce nuage (encore appelée **inertie totale**).

Généralités II

Cette *variance totale* ou *inertie totale* est définie par

$$Var_{tot} = \frac{1}{n^2} \sum_{\mathbf{x} \in X} \sum_{\mathbf{x}' \in X} d^2(\mathbf{x}, \mathbf{x}') = \frac{1}{n} \sum_{i=1}^{n} d^2(\mathbf{x}_i, \bar{\mathbf{x}}),$$

où $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$ est appelé **barycentre du nuage de points**.
Il représente un individu *moyen* qui représente le nuage de points.

La notion de *variance* employée ici fait appel à la notion de distance que nous n'avons pas encore définie.

Généralités III

Définition 1.1: Distance

Soit E un ensemble (par exemple un espace vectoriel, mais ce n'est pas une obligation). On appelle **distance** sur l'espace E, tout application d de $E \times E$ à valeurs dans $\mathbb{R}^+$ qui vérifient les propriétés suivantes :

- **symétrie :** $\forall \mathbf{x}, \mathbf{x}' \in E,\ d(\mathbf{x}, \mathbf{x}') = d(\mathbf{x}', \mathbf{x})$
- **sépération :** $\forall \mathbf{x}, \mathbf{x}' \in E,\ d(\mathbf{x}, \mathbf{x}') = 0 \iff \mathbf{x} = \mathbf{x}'$
- **inégalité triangulaire :**
  $\forall \mathbf{x}, \mathbf{x}', \mathbf{x}'' \in E,\ d(\mathbf{x}, \mathbf{x}'') \leq d(\mathbf{x}, \mathbf{x}') + d(\mathbf{x}', \mathbf{x}'').$

Généralités IV

Nous avons déjà rencontré des distances dans la première partie de ce document. Ce sont les distances induites par les normes, lorsque l'ensemble E est un espace vectoriel. On a alors

$$\forall \mathbf{x}, \mathbf{x}' \in E, \quad d(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|.$$

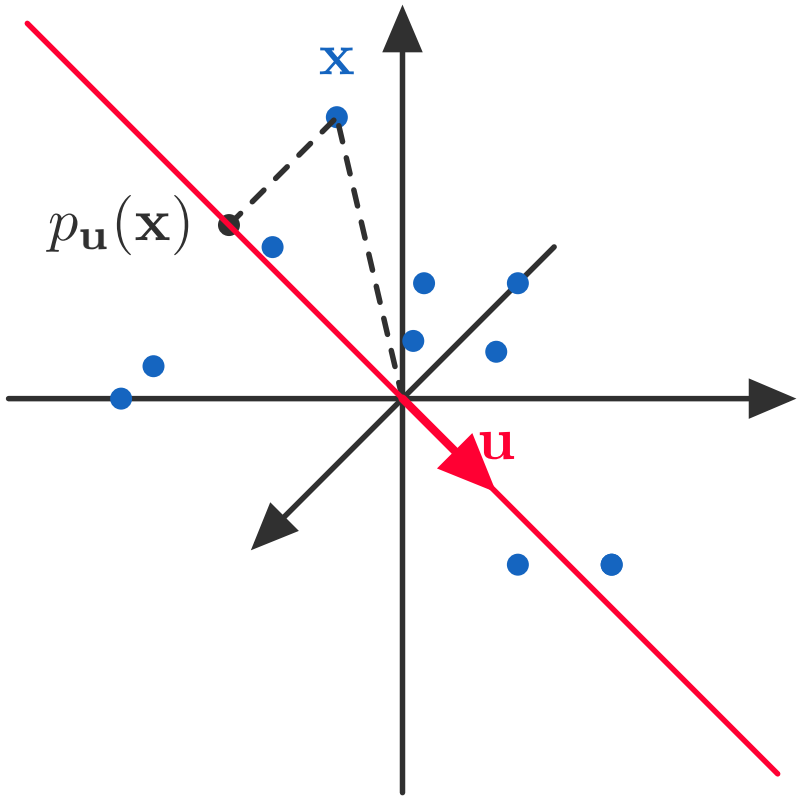
Ainsi on peut définir des distances pour tout entier $p > 1$ comme dans le cas des normes

$$\text{Distance de Manhattan} \quad d(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{n} |\mathbf{x} - \mathbf{x}'|.$$

$$\text{Distance Euclidienne} \quad d(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{i=1}^{n} (\mathbf{x} - \mathbf{x}')^2}.$$

Généralités V

Un des premier objectif des méthodes que nous étudierons est de pouvoir fournir une représentation fidèle des données dans un espace de dimension inférieure.



Généralités VI

On va donc chercher la droite $\mathbf{u}$ qui maximise la variance de cette nouvelle représentation Cela conduit à résoudre un problème d'optimisation suivant :

$$\min_{\mathbf{u}\in\mathbb{R}^p} \sum_{i=1}^{n} \|\mathbf{x} - p_{\mathbf{u}}(\mathbf{x})\|^2,$$

où $p_{\mathbf{u}}(\mathbf{x})$ désigne le projeté orthogonal de $\mathbf{x}$ sur la droite vectorielle engendrée par $\mathbf{u}$.
En utilisant les propriétés d'orthogonalité :

$$\|\mathbf{x} - p_{\mathbf{u}}(\mathbf{x})\|^2 = \|\mathbf{x}\|^2 - \|p_{\mathbf{u}}(\mathbf{x})\|^2.$$

Ainsi, notre problème de minimisation initial est équivalent à

$$\max_{\mathbf{u}\in\mathbb{R}^p} \sum_{i=1}^{n} \|p_{\mathbf{u}}(\mathbf{x})\|^2.$$

Lien avec valeurs et vecteurs propres I

On se rappelle que le projeté $p_{\mathbf{u}}(\mathbf{x})$ d'un vecteur $\mathbf{x}$ sur la droite vectorielle engendrée par $\mathbf{u}$ est donnée par

$$p_{\mathbf{u}}(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{u} \rangle}{\|\mathbf{u}\|} \mathbf{u}.$$

Dans la suite on supposera, sans perte de généralité, que le vecteur $\mathbf{u}$ est un vecteur de norme égale à 1. Nous aurons alors : $p_{\mathbf{u}}(\mathbf{x}) = \langle \mathbf{x}, \mathbf{u} \rangle \mathbf{u}$. Dans ce cas

$$\begin{aligned} \|p_{\mathbf{u}}(\mathbf{x})\|^2 &= \langle p_{\mathbf{u}}(\mathbf{x}), p_{\mathbf{u}}(\mathbf{x}) \rangle, \\ &= \langle \langle \mathbf{x}, \mathbf{u} \rangle \mathbf{u}, \langle \mathbf{x}, \mathbf{u} \rangle \mathbf{u} \rangle, \\ &\quad \downarrow \text{ linéarité du produit scalaire} \\ &= \langle \mathbf{x}, \mathbf{u} \rangle^2 \langle \mathbf{u}, \mathbf{u} \rangle, \end{aligned}$$

Lien avec valeurs et vecteurs propres II

$$
\begin{aligned}
&\downarrow \text{ car } \mathbf{u} \text{ est un vecteur unitaire} \\
&= \langle \mathbf{x}, \mathbf{u} \rangle^2, \\
&\downarrow \text{ simple réécriture} \\
&= (\mathbf{x}^T \mathbf{u})^T (\mathbf{x}^T \mathbf{u}), \\
&= \mathbf{u}^T \mathbf{x} \mathbf{x}^T \mathbf{u}.
\end{aligned}
$$

Lien avec valeurs et vecteurs propres III

Notons maintenant que $X = (\mathbf{x}_1^T, \cdots, \mathbf{x}_n^T)$ est une matrice à n lignes et p colonnes dont les lignes sont formées par les $\mathbf{x}_i$. On peut alors écrire :

$$\sum_{i=1}^{n} \|p_{\mathbf{u}}(\mathbf{x}_i)\|^2 = \sum_{i=1}^{n} \mathbf{u}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{u} = \sum_{i=1}^{n} \mathbf{u}^T \left(\mathbf{x}_i \mathbf{x}_i^T\right) \mathbf{u} = \mathbf{u}^T \left(X^T X\right) \mathbf{u}.$$

Dans cette relation, la matrice $X^T X$ est une matrice de $\mathscr{M}_p(\mathbb{R})$.Nous avons alors l'équivalence suivante :

$$\max_{\mathbf{u} \in \mathbb{R}^p} \sum_{i=1}^{n} \|p_{\mathbf{u}}(\mathbf{x})\|^2 = \max_{\mathbf{u} \in \mathbb{R}^p,\, \|\mathbf{u}\|^2 = 1} \mathbf{u}^T X^T X \mathbf{u}.$$

Lien avec valeurs et vecteurs propres IV

Définition 1.2: Matrice de Gram

Soit E un espace euclidien de dimension p et $\mathbf{x}_1, \cdots, \mathbf{x}_n$ des vecteurs de E. On appelle **matrice de Gram**, notée G, la matrice carrée des produits scalaires scalaires entre les individus, dont la matrice d'ordre n telle que :

$$G_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle, \ \forall i, j = 1, \cdots, n.$$

Nous pourrions définir une matrice similaire pour définir le produit scalaire entre les variables.
Il s'agit également d'une matrice de Gram.

Lien avec valeurs et vecteurs propres V

Proposition 1.1: Valeurs propres de la matrice de Gram

Soit E un espace euclidien de dimension p et $\mathbf{x}_1, \cdots, \mathbf{x}_n$ des vecteurs de E et considérons la matrice G carrée d'ordre n définie par

$$G_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle, \ \forall i, j = 1, \cdots, n.$$

Alors G est symétrique et semi-définie positive, *i.e.* elle admet n valeurs propres positives ou nulles.

Preuve : Rappelons que si $\mathbf{u}$ est un vecteur propre (non nul !) G associée à la valeur propre λ, alors

$$G\mathbf{u} = \lambda \mathbf{u}.$$

Lien avec valeurs et vecteurs propres VI

Or $G = XX^T$, donc $G\mathbf{u} = XX^T\mathbf{u} = \lambda\mathbf{u}$. Si on pré-multiplie chaque membre de l'égalité par $\mathbf{u}^T$, nous avons

$$\mathbf{u}^T XX^T\mathbf{u} = \lambda\mathbf{u}^T\mathbf{u},$$

$$\downarrow \text{ définition transposée et norme}$$

$$\left(X^T\mathbf{u}\right)^T\left(X^T\mathbf{u}\right) = \lambda\|\mathbf{u}\|^2,$$

$$\downarrow \text{ définition de norme}$$

$$\|X^T\mathbf{u}\|^2 = \lambda\|\mathbf{u}\|^2,$$

$$\downarrow \mathbf{u} \text{ est un vecteur non nul}$$

$$\lambda = \frac{\|X^T\mathbf{u}\|^2}{\|\mathbf{u}\|^2} \geq 0.$$

On peut rédiger une démonstration analogue dans le cas où $G' = X^TX$.

Lien avec valeurs et vecteurs propres VII

Supposons que l'on souhaite maintenant projeter les données sur un espace de dimension $1 < p' < p$. Cela se fera toujours en étudiant les valeurs propres et vecteurs propres de la matrice G'.

Vu que cette matrice est symétrique, on rappelle que ses vecteurs propres forment une base orthonormée de l'espace de départ.

Ainsi, si l'on cherche la représentation dans l'espace de dimension p' qui maximise la variance, il suffit de déterminer les p' vecteurs propres associés aux p' $(\mathbf{u}_1, \cdots, \mathbf{u}_{p'})$ plus grandes valeurs propres $(\lambda_1, \cdots, \lambda_{p'})$ associées à la matrice G'.

Lien avec valeurs et vecteurs propres VIII

Les coordonnées d'une donnée $\mathbf{x}_i$ sur la droite vectorielle engendrée par $\mathbf{u}_s$ est donnée par :

$$\langle \mathbf{x}_i, \mathbf{u}_s \rangle = \mathbf{x}_i^T \mathbf{u}_s.$$

Pour l'ensemble des vecteurs $\mathbf{x}_i$, les coordonnées sur le droite vectorielle engendrée par $\mathbf{u}_s$ sont données par le vecteur :

$$\langle X^T, \mathbf{u}_s \rangle = X \mathbf{u}_s.$$

On peut déterminer les coordonnées de cette façon pour l'ensemble des vecteurs $\mathbf{x}_i$ dans la base des vecteurs propres $\mathbf{u}_s$. Ces coordonnées sont données par la matrice

$$XU_{p'},$$

où $U_{p'}$ est une matrice de dimension $p \times p'$.

Dualité des représentations I

Nous venons de voir que chercher à obtenir une représentation d'un ensemble d'individus revient à diagonaliser la matrice $X^T X$ de $\mathscr{M}_p(\mathbb{R})$.

De même, si on souhaite projeter les variables, représentés dans l'espace des individus, dans un espace de dimension inférieure, nous devons diagonaliser la matrice XX^T de $\mathscr{M}_n(\mathbb{R})$.

Il y a cependant un fait plutôt marquant entre ces deux matrices ... toutes les valeurs propres non nulles sont égales !

En effet, notons λ_k la k-ème plus grande valeur propre de la matrice XX^T et $\mathbf{u}_k$ le vecteur propre associé. Par définition, nous avons alors :

$$XX^T\mathbf{u}_k = \lambda_k\mathbf{u}_k \quad \text{d'où} \quad X^TXX^T\mathbf{u}_k = \lambda_k X^T\mathbf{u}_k.$$

Dualité des représentations II

Ainsi si $\mathbf{u}_k$ est le vecteur propre de la matrice XX^T associé à la valeur propre λ_k, alors le vecteur $X^T\mathbf{u}_k$ est un vecteur propre de la matrice X^TX associé à la même valeur propre λ_k.

On en déduit les relations suivantes entre les deux vecteurs propres $\mathbf{u}'_k$ et $X\mathbf{u}'_k = \mathbf{u}_k$

$$\mathbf{u}'_k = \frac{X^T\mathbf{u}_k}{\|X^T\mathbf{u}_k\|} = \frac{1}{\sqrt{\lambda_k}}X^T\mathbf{u}_k.$$

Nous avons également la relation inverse

$$\mathbf{u}_k = \frac{X\mathbf{u}'_k}{\|X\mathbf{u}'_k\|} = \frac{1}{\sqrt{\lambda_k}}X\mathbf{u}'_k.$$

Décomposition en valeurs singulières I

Proposition 1.2: Décomposition en valeurs singulières

Soit X une matrice réelle d'ordre $n \times p$ et de rang $r \leq \min(n, p)$. Alors la matrice X peut être factorisée de la façon suivante

$$X = U \Sigma U'^T,$$

où $U \in \mathscr{M}_n(\mathbb{R})$ et $U' \in \mathscr{M}_p(\mathbb{R})$ sont des matrices orthogonales et $\Sigma \in \mathscr{M}_{n,p}(\mathbb{R})$ est une matrice remplie de 0 sauf sur la diagonale principale où, pour tout $i = 1, \cdots, r$, on a :

$$\Sigma_{ii} = \sigma_i,$$

où les σ_i $(\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0)$ sont liées aux valeurs propres non nulles indifféremment de la matrice XX^T ou X^TX.

Décomposition en valeurs singulières II

Dans cette proposition :

- les valeurs σ_i sont appelée *valeurs singulières* de la matrice X.
- la matrice U est composé des *vecteurs singuliers* à gauche de la matrice X. Ils correspondent aux vecteurs propres de la matrice XX^T.
- la matrice U' est composé des *vecteurs singuliers à droite* de la matrice X. Ils correspondent aux vecteurs propres de la matrice X^TX.

De plus, les colonnes des matrices U et U' forment une base orthonormée des espaces $\mathbb{R}^n$ et $\mathbb{R}^p$ respectivement.

Décomposition en valeurs singulières III

Pour cela remarquons que si $X = U\Sigma U'^T$ alors $X^T = U'\Sigma^T U^T$ ce qui implique :

$$
\begin{aligned}
X^T X &= X^T U\Sigma U'^T, \\
&\quad \downarrow \text{ définition de } X^T \\
&= U'\Sigma^T U^T U\Sigma U'^T, \\
&\quad \downarrow \text{ orthogonalité de } U \\
&= U'\Sigma^T\Sigma U'^T, \\
&= U'\Sigma_p^2 U'^T,
\end{aligned}
$$

où $\Sigma^T\Sigma = \Sigma_p^2 \in \mathscr{M}_p(\mathbb{R})$.

$$
\begin{aligned}
XX^T &= XU'\Sigma^T U^T, \\
&\quad \downarrow \text{ définition de } X \\
&= U\Sigma U'^T U'\Sigma^T U^T, \\
&\quad \downarrow \text{ orthogonalité de } U' \\
&= U\Sigma\Sigma^T U^T, \\
&= U\Sigma_n^2 U^T,
\end{aligned}
$$

où $\Sigma\Sigma^T = \Sigma_n^2 \in \mathscr{M}_n(\mathbb{R})$.

$\Longrightarrow$ les valeurs singulières de X sont les racines carrés des valeurs propres des matrices $X^T X$ ou XX^T.

Qualité de l'approximation I

Pour cela considérons notre matrice $X \in \mathscr{M}_{n,p}(\mathbb{R})$, où $\mathscr{M}_{n,p}(\mathbb{R})$ est l'espace vectoriel des matrices muni du produit scalaire de Frobenius $\langle \cdot, \cdot \rangle_F$ et de la norme induite $\| \cdot \|_F$ qui nous servira à définir la distance entre deux matrices.

$$\langle A, B \rangle = Tr(A^T B) = \sum_{i=1}^{n} \sum_{j=1}^{p} a_{ij} b_{ij} \quad \text{et} \quad \|A - B\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} (a_{ij} - b_{ij})^2.$$

La norme de Frobenius de la différence de deux matrices peut ainsi se voir comme la distance euclidienne entre deux matrices.

Qualité de l'approximation II

On souhaite maintenant approximer, au sens de la norme de Frobenius, une matrice X par une matrice $\tilde{X}$ telle que $\tilde{X}$ soit de rang inférieur à un rang donné s. On souhaite donc résoudre le problème suivant :

$$\min_{\tilde{X} \in \mathscr{M}_{n,p}(\mathbb{R}),\, rg(\tilde{X}) \leq s} \|X - \tilde{X}\|_F^2.$$

La solution à ce problème d'optimisation est donnée par le *Théorème d'Eckart-Young*.

Qualité de l'approximation III

Théorème 1.1: Théorème d'Eckart-Young

Soit X une matrice de $\mathscr{M}_{n,p}(\mathbb{R})$ et considérons le problème d'optimisation suivant :

$$\min_{\tilde{X}\in\mathscr{M}_{n,p}(\mathbb{R}),\, rg(\tilde{X})\leq s} \|X-\tilde{X}\|_F^2.$$

Alors la solution de ce problème d'optimisation est donnée par la décomposition en valeurs singulières de la matrice X.

Qualité de l'approximation IV

Quelques remarques à propos de ce théorème :

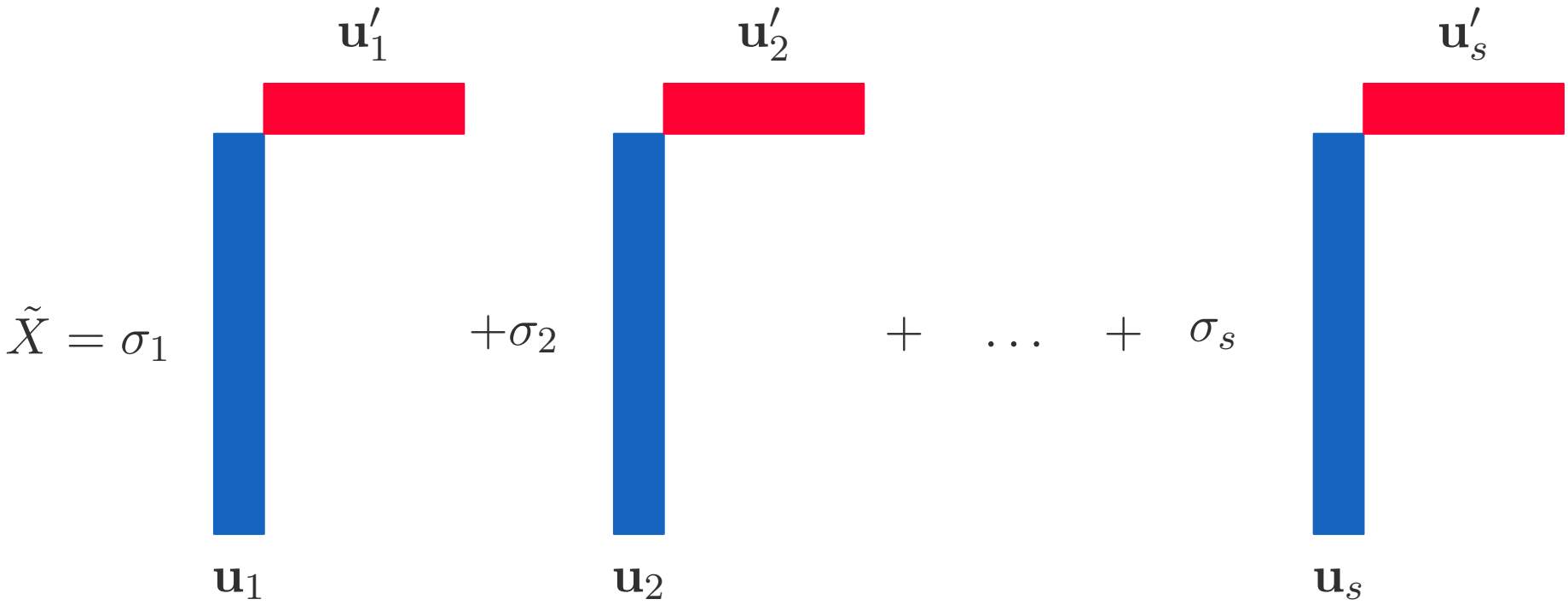
- si on cherche une approximation de rang s et que la matrice de design (autre nom donnée à la matrice des données) X est de rang $< s$, on a alors $X = \tilde{X}$.
- si la matrice de design est de rang $r > s$, alors elle possède r valeurs singuliers ainsi que r vecteurs singuliers à gauche $(\mathbf{u}_1, \cdots, \mathbf{u}_r)$ et à droite $(\mathbf{u}'_1, \cdots, \mathbf{u}'_r)$ qui sont normés.
  Ainsi la meilleure approximation de rang s est donnée par

$$\tilde{X} = \sigma_1 \mathbf{u}_1 (\mathbf{u}'_1)^T + \cdots + \sigma_s \mathbf{u}_s (\mathbf{u}'_s)^T,$$

  où les valeurs singulières sont rangés par ordre décroissant :
  $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_s$.

Qualité de l'approximation V

Chaque élément de cette somme représente des matrices de rang 1 mais comme pour tout $i \neq j$, $\mathbf{u}_i \perp \mathbf{u}_j$ et $\mathbf{u}'_i \perp \mathbf{u}'_j$, alors la somme des matrices $\mathbf{u}_i(\mathbf{u}'_i)^T$ avec $\mathbf{u}_j(\mathbf{u}'_j)^T$ donnent bien une matrice de rang 2. Ainsi notre approximation $\tilde{X}$ peut se représenter comme



Qualité de l'approximation VI

- enfin, si l'on souhaite mesurer la qualité τ de l'information, on peut regarder un premier indicateur naïf qu'est le quotient de la somme des valeurs singulières associées à l'approximation sur la somme des valeurs singulières de la matrice X :

$$\tau = \frac{\sum_{k=1}^{s} \sigma_k(X)}{\sum_{k=1}^{r} \sigma_k(X)}.$$

  Plus la valeur de s est grande, meilleure sera l'approximation.
  Cette valeur est un bon indicateur pour savoir si l'information contenue dans la matrice de design peut être synthétisée dans un espace de dimension faible.
  En revanche, ce critère ne repose pas sur la même norme que celle employée dans le problème d'optimisation, il faudrait donc introduire un moyen de mesurer qualité de l'approximation qui se fonde sur la norme de Frobenius.

Qualité de l'approximation VII

Pour cela on mesure plutôt :

$$\tau = \frac{\|\tilde{X}\|_F^2}{\|X\|_F^2},$$

où la norme de Frobenius d'une matrice X est donnée par

$$\|X\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2.$$

Mais plutôt que de calculer cette somme là, on va voir que l'on peut à nouveau faire intervenir les valeurs singulières de la matrice X (ou les valeurs propres de la matrice XX^T !). Pour cela, on se rappelle que si X est une matrice de rang r, alors $X = \sum_{k=1}^{r} \sigma_k \mathbf{u}_k (\mathbf{u}'_k)^T$

Qualité de l'approximation VIII

$$\|X\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2,$$

$$\downarrow \text{ en utilisant la SVD}$$

$$= \sum_{k=1}^{r} \sigma_k^2 \sum_{i=1}^{n}\sum_{j=1}^{p} (\mathbf{u}_k)_i^2 (\mathbf{u}'_k)_j^2,$$

$$\downarrow \text{ on arrange les termes}$$

$$= \sum_{k=1}^{r} \sigma_k^2 \left(\sum_{i=1}^{n} (\mathbf{u}_k)_i^2\right) \left(\sum_{j=1}^{p} (\mathbf{u}'_k)_j^2\right),$$

$$\downarrow \text{ les vecteurs } \mathbf{u}_k \text{ et } \mathbf{u}'_k \text{ sont normés}$$

Qualité de l'approximation IX

$$= \sum_{k=1}^{r} \sigma_k^2.$$

Ainsi

$$\tau = \frac{\sum_{k=1}^{s} \sigma_k^2(X)}{\sum_{k=1}^{r} \sigma_k^2(X)} = \frac{\sum_{k=1}^{s} \lambda_k(X)}{\sum_{k=1}^{r} \lambda_k(X)}.$$

Mettons cela en oeuvre sur un exemple.

Exemple I

Soit $X = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$ une matrice de $\mathscr{M}_{3,2}(\mathbb{R})$. On va essayer de déterminer la décomposition en valeurs singulières de cette matrice, ses vecteurs singuliers contenus dans les matrices U et U' ainsi que la qualité d'une approximation de rang 1.
On commence par déterminer les valeurs propres de la matrice $X^TX = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{pmatrix}$ en déterminant les racines du polynôme caractéristique $\mathcal{X}_{X^TX}(\lambda)$ défini par :

$$\begin{aligned} \mathcal{X}_{X^TX}(\lambda) &= det(X^TX - \lambda I_3) \\ &= \begin{vmatrix} 1-\lambda & 1 & 0 \\ 1 & 2-\lambda & 1 \\ 0 & 1 & 1-\lambda \end{vmatrix}, \end{aligned}$$

Exemple II

$$\downarrow \text{ on développe selon la ligne}$$

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

$$\downarrow \text{ on développe les deux déterminants d'ordre 2 et on factorise}$$

$$-\lambda(\lambda-3)(\lambda-1).$$

Ainsi les racines de $\mathcal{X}_{X^TX}(\lambda)$ et donc les valeurs propres de X^TX sont $3, 1$ et 0. Les valeurs singulières de X sont donc $\sigma_1 = \sqrt{3}$ et $\sigma_2 = 1$.

On cherche maintenant les vecteurs propres associés à chaque valeur propre de la matrice X^TX. Ces vecteurs propres vont définir les *vecteurs singuliers à droite* de notre matrice X (*i.e.* ils vont définir la matrice U'.)

Exemple III

- **Vecteur propre associé à $\lambda = 3$:** on a $X^TX - 3I_3 = \begin{pmatrix} -2 & 1 & 0 \\ 1 & -1 & 1 \\ 0 & 1 & -2 \end{pmatrix}$ dont la forme échelonnée réduite est donnée par la matrice $\begin{pmatrix} 1 & 0 & -1 \\ 0 & -1 & 2 \\ 0 & 0 & 0 \end{pmatrix}$. Un élément du noyau de cette matrice est donnée par le vecteur $(1, 2, 1)$, donc une base orthonormée de ce sous-espace propre est donné par le vecteur $\mathbf{u}'_1 = \frac{1}{\sqrt{6}}(1, 2, 1)$. On en déduit de suite $\mathbf{u}_1 = \frac{X\mathbf{u}'_1}{\sigma_1} = \frac{1}{\sqrt{2}}(1, 1)$.

Exemple IV

- **Vecteur propre associé à $\lambda = 1$:** on effectue le même processus que précédemment. On a $X^TX - I_3 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$ dont la forme échelonnée réduite est donnée par la matrice $\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$. Un élément du noyau de cette matrice est donnée par le vecteur $(1, 0, -1)$, donc une base orthonormée de ce sous-espace propre est donné par le vecteur $\mathbf{u}'_2 = \frac{1}{\sqrt{2}}(1, 0, -1)$.

  On en déduit de suite $\mathbf{u}_2 = \frac{X\mathbf{u}'_1}{\sigma_2} = \frac{1}{\sqrt{2}}(1, -1)$.

Exemple V

- **Vecteur propre associé à $\lambda = 0$:** On a $X^T X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{pmatrix}$ dont la forme échelonnée réduite est donnée par la matrice $\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$.
  Un élément du noyau de cette matrice est donnée par le vecteur $(1, -1, 1)$, donc une base orthonormée de ce sous-espace propre est donné par le vecteur $\mathbf{u}'_3 = \frac{1}{\sqrt{3}}(1, -1, 1)$.

Exemple VI

Finalement, en posant $U = (\mathbf{u}_1, \mathbf{u}_2)$, $\Sigma = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ et $U' = (\mathbf{u}'_1, \mathbf{u}'_2, \mathbf{u}'_3)$, on obtient la décomposition en valeurs singulières de la matrice X.
Regardons maintenant la qualité de l'approximation de rang 1. La matrice $\tilde{X}$ associée est définie par :

$$\tilde{X} = \sigma_1 \mathbf{u}_1 (\mathbf{u}'_1)^T = \frac{\sqrt{3}}{2} \begin{pmatrix} 1 & 2 & 1 \\ 1 & 2 & 1 \end{pmatrix}.$$

La qualité de cette approximation de rang 1 est donnée par

$$\tau = \frac{\|\tilde{X}\|_F^2}{\|X\|_F^2} = \frac{3}{3+1} = \frac{3}{4}.$$

Ainsi, l'approximation de rang 1 permet de restituer 75% de l'information initialement présente.

Pour finir

Nous avons tous les outils nécessaires et à appliquer pour présenter les techniques d'analyse de données.

Toutes les techniques présentées dans ce cours se basent sur les outils montrés précédemment. Les seules différences viendront de la nature des variables (*quantitatives* ou *qualitatives*) et donc de la préparation de ces dernières pour effectuer leur analyse.

On commence par l'analyse des variables *quantitatives*.

Analyses de Données

Analyse en Composantes Principales (ACP)

Principe I

L'analyse en composantes principale est une méthode réduction de la dimension qui s'emploie lorsque les variables étudiées sont **quantitatives réelles**, *i.e.* les variables prennent des valeurs réelles.

$$X = \begin{array}{c} \\ \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_i \\ \vdots \\ \mathbf{x}_n \end{array} \begin{array}{c} \begin{array}{cccccc} \mathbf{v}_1 & \cdots & \mathbf{v}_k & \cdots & \mathbf{v}_p \end{array} \\ \begin{pmatrix} x_{11} & \cdots & x_{1k} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} & \cdots & x_{np} \end{pmatrix} \end{array},$$

Principe II

Individus : comme nous l'avons mentionné dans la section précédente, la mesure que nous utiliserons pour mesurer la distance entre les individus est la distance euclidienne :

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} ((\mathbf{x}_i)_k - \mathbf{x}_j)_k)^2.$$

Cette distance sera donc la base de l'étude du nuage des individus.

Variable : nous n'utiliserons pas de notion de distance entre les variables mais nous allons plutôt chercher à mesurer la **liaison** entre ces dernières en étudiant le **coefficient de corrélation** noté $r(\mathbf{v}_k, \mathbf{v}_l)$

Principe III

Pour cela, nous aurons besoin d'étudier les grandeurs suivantes :

La moyenne m_k d'une variable $\mathbf{v}_k$

$$m_k = \frac{1}{n}\sum_{l=1}^{n}(\mathbf{v}_k)_i,$$

ainsi que la variance s_k^2 de ces mêmes variables

$$s_k^2 = \frac{1}{n}\sum_{l=1}^{n}\left((\mathbf{v}_k)_i - m_k\right)^2.$$

Ces quantités nous serviront à étudier notre matrice de design X avec notre ACP.

Objectif ACP

l'ACP va transformer des variables **corrélées** en un nouvel ensemble de variables **décorrélées** qui vont se présenter comme une combinaison linéaire des anciennes variables. Ces nouvelles variables seront appelées **axes principaux**.
Ces axes correspondent a des directions de l'espace selon lesquelles la variance est maximale.

En tant que technique cherchant à préserver au maximum la variance dans les données, l'ACP se présente comme un problème aux valeurs propres sur une certaine matrice de variance : *la matrice de variance-covariance* des données X. On pourra également (et c'est le choix que l'on fera par la suite) de travailler sur *la matrice de corrélation* des variables.

Transformation des données I

La première étape consiste à **centrer** notre jeu de données, *i.e.* faire en sorte que les moyennes de chaque variables $\mathbf{v}_k$ soient égales à 0. On obtient alors une nouvelle matrice $X_{centrée}$ définie par

$$X_{centrée} = \begin{pmatrix} x_{11} - m_1 & \cdots & x_{1k} - m_k & \cdots & x_{1p} - m_p \\ \vdots & & \vdots & & \vdots \\ x_{i1} - m_1 & \cdots & x_{ik} - m_k & \cdots & x_{ip} - m_p \\ \vdots & & \vdots & & \vdots \\ x_{n1} - m_1 & \cdots & x_{nk} - m_k & \cdots & x_{np} - m_p \end{pmatrix} = X - \mathbf{1}\mathbf{m}^T,$$

où $\mathbf{m} = (m_1, \cdots, m_p)$ est le vecteur des moyennes des variables. C'est aussi le barycentre du nuage de points et $\mathbf{1}$ est un vecteur colonne de taille n ne comprenant que des 1.

Transformation des données II

On peut ensuite réduire notre jeux de données en divisant chaque variable par son écart-type, c'est le choix que nous ferons ici mais il n'est pas obligatoire. On obtient alors une nouvelle matrice $X_{cen-red}$ définie par

$$X_{cen-red} = \begin{pmatrix} \frac{x_{11}-m_1}{s_1} & \dots & \frac{x_{1k}-m_k}{s_k} & \dots & \frac{x_{1p}-m_p}{s_p} \\ \vdots & & \vdots & & \vdots \\ \frac{x_{i1}-m_1}{s_1} & \dots & \frac{x_{ik}-m_k}{s_k} & \dots & \frac{x_{ip}-m_p}{s_p} \\ \vdots & & \vdots & & \vdots \\ \frac{x_{n1}-m_1}{s_1} & \dots & \frac{x_{nk}-m_k}{s_k} & \dots & \frac{x_{np}-m_p}{s_p} \end{pmatrix}.$$

Transformation des données III

Réduire ses données n'est pas sans conséquence sur l'ACP qui repose sur la variance des données. En effet en faisant ce choix :

- toutes les variables vont avoir la même variance (égale à 1) ce qui va éviter de tirer l'ACP vers les variables dont la variance est élevée simplement parce que les valeurs prises par cette dernière sont plus grandes.
- en revanche, si les données associées à une variable présentent un bruit important (mauvaise collecte des données, problème avec l'outil de mesure, ...) alors cette dernière aura une variance semblable à une variable qui serait elle plus informative.

Transformation des données IV

Remarque importante : **il y a deux écoles pour réaliser l'ACP : l'ACP classique et une ACP dite normée. Cette deuxième version nécessite de diviser l'ensemble des colonnes de notre matrice $X_{cen-red}$ par $\sqrt{n}$.**

On suppose maintenant que nos données sont centrées et réduites et les valeurs sont divisées par $\sqrt{n}$ comme décrit précédemment et on cherche maintenant à projeter nos individus dans un espace de dimension plus faible.

Projection des individus I

On va procéder de façon semblable à la SVD. Si on cherche à projeter les données sur un sous-espace de dimension s, on va commencer par chercher des directions $\mathbf{u}'_1, \cdots, \mathbf{u}'_s$ sur lesquelles on va maximiser la variance (ou l'inertie) du nuage de points. Ces directions seront appelées **axes principaux** avec cette même convention que pour la SVD :

- l'axe défini par le vecteur $\mathbf{u}'_1$ est le sous-espace de dimension 1 qui maximise l'inertie du nuage du point après projection,
- l'axe défini par le vecteur $\mathbf{u}'_2$, *orthogonal à* $\mathbf{u}_1$ est le deuxième sous-espace de dimension 1 qui maximise l'inertie du nuage du point après projection,
- on continue avec $\mathbf{u}'_3$ qui est *orthogonal* à $\mathbf{u}'_1$ et $\mathbf{u}'_2$, et ainsi de suite.

Projection des individus II

Nous allons faire la même chose ici, mais nous ne travaillerons pas directement sur la matrice de design X mais plutôt sur sa version centrée-réduite normée $X_{cen-red}$.
En effet, on rappelle que le but de l'ACP est aussi de décorrélée les variables, il faut donc faire intervenir cette matrices de corrélation. Pour rappel, il s'agit de la matrice $C \in \mathscr{M}_p(\mathbb{R})$ définie par

$$C = \frac{1}{n} X_{cen-red}^T X_{cen-red},$$

où les termes c_{ij} sont données par $c_{ij} = \frac{1}{n} \sum_{l=1}^{n} \left(\frac{x_{li} - m_i}{s_i} \right) \left(\frac{x_{lj} - m_j}{s_j} \right)$ et sont tous compris dans l'intervalle $[-1, 1]$.

Projection des individus III

Dans la suite, nous noterons z_{ij} les éléments de la matrice $X_{cen-red}/\sqrt{n}$ qui correspondent aux données centrées réduites normée de la matrice de design X afin d'éviter toute confusion. Ainsi, on définit :

$$Z = \begin{array}{c} \\ \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_k \\ \vdots \\ \mathbf{z}_n \end{array} \begin{array}{c} \begin{array}{cccccc} \mathbf{v}_1' & \cdots & \mathbf{v}_k' & \cdots & \mathbf{v}_p' \end{array} \\ \begin{pmatrix} \frac{x_{11}-m_1}{s_1\sqrt{n}} & \cdots & \frac{x_{1k}-m_k}{s_k\sqrt{n}} & \cdots & \frac{x_{1p}-m_p}{s_p\sqrt{n}} \\ \vdots & & \vdots & & \vdots \\ \frac{x_{i1}-m_1}{s_1\sqrt{n}} & \cdots & \frac{x_{ik}-m_k}{s_k\sqrt{n}} & \cdots & \frac{x_{ip}-m_p}{s_p\sqrt{n}} \\ \vdots & & \vdots & & \vdots \\ \frac{x_{n1}-m_1}{s_1\sqrt{n}} & \cdots & \frac{x_{nk}-m_k}{s_k\sqrt{n}} & \cdots & \frac{x_{np}-m_p}{s_p\sqrt{n}} \end{pmatrix} \end{array}.$$

Projection des individus IV

En définissant $C = Z^T Z$, on définit une matrice symétrique réelle, elle est donc orthogonalement semblable à une matrice diagonale, *i.e.* il existe donc une matrice orthogonale $U' \in \mathscr{M}_p(\mathbb{R})$ et une matrice diagonale $\Sigma_p = diag(\lambda_1, \cdots, \lambda_p)$ telle que $\lambda_1 \geq \cdots \geq \lambda_p$. On a

$$C = U'\Sigma_p U'^T \quad \text{et pour tout } m \leq p, \quad C\mathbf{u}'_m = \lambda_m \mathbf{u}'_m.$$

où U' est formée des vecteurs propres de C. Donc le vecteur propre $\mathbf{u}'_m$ est associé à la valeur propre λ_m qui est la $m-$ème plus grande valeur propre.

Projection des individus V

Ainsi, les nouvelles coordonnées d'un individu $\mathbf{x}_i$ sur un axe principal $\mathbf{u}'$ sont données par :

$$p_{\mathbf{u}'}(\mathbf{z}_i) = \langle \mathbf{z}_i, \mathbf{u}' \rangle = \sum_{k=1}^{p} z_{ik} u'_k.$$

On peut montrer que le nuage projeté sur un axe principal $\mathbf{u}$ est également centré, *i.e.*

$$\frac{1}{n}\sum_{i=1}^{n} p_{\mathbf{u}'}(\mathbf{z}_i) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{p} z_{ik} u'_k = \sum_{k=1}^{p} u'_k \underbrace{\frac{1}{n}\sum_{i=1}^{n} z_{ik}}_{=0} = 0.$$

Enfin, on peut également calculer la variance associée au m-ème axe principal $\mathbf{u}_m$:

Projection des individus VI

$$\frac{1}{n}\sum_{i=1}^{n} p_{\mathbf{u}'_m}(\mathbf{z}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{z}_i, \mathbf{u}'\rangle^2 = \frac{1}{n}\sum_{i=1}^{n} {\mathbf{u}'_m}^T \mathbf{z}_i \mathbf{z}_i^T \mathbf{u}'_m$$

$$= \frac{1}{n}{\mathbf{u}'_m}^T C \mathbf{u}'_m = \frac{1}{n}\lambda_m.$$

Cette dernière égalité nous indique que la variance du projeté sur l'axe engendré par le vecteur $\mathbf{u}'_m$ est égal, au facteur $1/n$ près, à la valeur propre associée au vecteur $\mathbf{u}'_m$.
Plus généralement, on obtient alors les composantes principales des n individus sur l'axe factoriel $\mathbf{u}'_m$ en déterminant

$$\mathbf{p}_{\mathbf{u}'_m} = \mathbf{p}_{\mathbf{u}'_m}(Z) = Z\mathbf{u}'_m.$$

Projection des variables I

On peut également procéder au même type d'analyse en travaillant sur le nuage des variables En revanche, la distance utilisée sera un peu différente vu que l'on ne s'intéresse plus aux individus mais plutôt aux variables.
La distance euclidienne employée fera alors intervenir le coefficient de corrélation :

$$d(\mathbf{v}'_k, \mathbf{v}'_l)^2 = \|\mathbf{v}'_k - \mathbf{v}'_l\|^2 = \langle \mathbf{v}'_k, \mathbf{v}'_k \rangle + \langle \mathbf{v}'_l, \mathbf{v}'_l \rangle - 2\langle \mathbf{v}'_k, \mathbf{v}'_l \rangle = 2\left(1 - c_{kl}\right),$$

$\langle \mathbf{v}'_j, \mathbf{v}'_j \rangle$ désigne la variance du vecteur $\mathbf{v}'_j$ qui est égale à 1 par définition et c_{kl} est le coefficient de corrélation linéaire entre les variables $\mathbf{v}'_k$ et $\mathbf{v}'_l$.

Projection des variables II

En revanche, on va cette fois-ci s'intéresser à la matrice des produits scalaires entre les individus. On retrouve donc notre matrice de Gram $G \in \mathscr{M}_n(\mathbb{R})$ sur les individus

$$G = ZZ^T.$$

Or G est une matrice symétrique réelle, elle est donc orthogonalement semblable à une matrice diagonale, *i.e.* il existe donc une matrice orthogonale $U \in \mathscr{M}_n(\mathbb{R})$ et une matrice diagonale $\Sigma_n = diag(\lambda_1, \cdots, \lambda_n)$ telle que $\lambda_1 \geq \cdots \geq \lambda_n$. On a

$$G = U\Sigma_n U^T \quad \text{et pour tout } m \leq n, \quad G\mathbf{u}_m = \lambda_m \mathbf{u}_m.$$

où U est formée des vecteurs propres de G. Donc le vecteur propre $\mathbf{u}_m$ est associé à la valeur propre λ_m qui est la m-ème plus grande valeur propre.

Projection des variables III

Ainsi, les nouvelles coordonnées d'une variable $\mathbf{v}'_j$ sur un axe principal $\mathbf{u} \in \mathbb{R}^n$ sont données par :

$$p_{\mathbf{u}}(\mathbf{v}'_j) = \langle \mathbf{v}'_j, \mathbf{u} \rangle = \sum_{i=1}^{n} z_{ij} u_i.$$

Regardons maintenant d'un peu plus près les propriétés du vecteur $\mathbf{p_u} = \mathbf{p_u}(Z^T) \in \mathbb{R}^p$ (on regarde les composantes des p variables sur $\mathbf{u}$) et remarquons le lien suivant entre les vecteurs variables et les coordonnées des projections

$$\mathbf{u}_m = \frac{1}{\sqrt{\lambda_m}} Z \mathbf{u}'_m = \frac{1}{\sqrt{\lambda_m}} \mathbf{p}_{\mathbf{u}'_m}(Z).$$

Projection des variables IV

Ce dernier élément correspond à la projection des n individus sur l'axe factoriel $\mathbf{u}'_m$. Les composantes de l'axe factoriel $\mathbf{u}_m$ sont donc proportionnelles aux composantes principales des individus sur l'axe factoriel $\mathbf{u}'_m$.

On peut également montrer que $p_{\mathbf{u}_m}(\mathbf{v}'_j)$ est égal au coefficient de corrélation linéaire entre les vecteurs $\mathbf{u}_m$ et $\mathbf{v}'_j$ (les deux étant des vecteurs de norme égale à 1) , c'est donc une valeur comprise entre -1 et 1. Mais c'est aussi égal au coefficient de corrélation linéaire entre les vecteurs $\mathbf{v}'_j$ et $\mathbf{p}_{\mathbf{u}'_m}(Z)$.

Dualité I

Regardons maintenant les liens qui existent l'espace des individus et l'espaces des variables.

En projetant les nuages de points sur les axes successifs, on décompose la variance, donc la variance totale n'est rien d'autre que la somme des valeurs propres de notre problème

$$Var = \sum_{k=1}^{p} \lambda_k = p.$$

Comme nous l'avons vu dans le cas de la SVD, il est également possible de faire le lien entre les deux espaces de représentations à l'aide des relations suivantes :

$$\mathbf{u}'_m = \frac{1}{\sqrt{\lambda_m}} \mathbf{p}_{\mathbf{u}_m}(Z) = \frac{1}{\sqrt{\lambda_m}} Z\mathbf{u}_m.$$

Dualité II

déjà présentée précédemment, ainsi que

$$\mathbf{u}_m = \frac{1}{\sqrt{\lambda_m}} \mathbf{p}_{\mathbf{u}'_m}(Z) = \frac{1}{\sqrt{\lambda_m}} Z^T \mathbf{u}'_m.$$

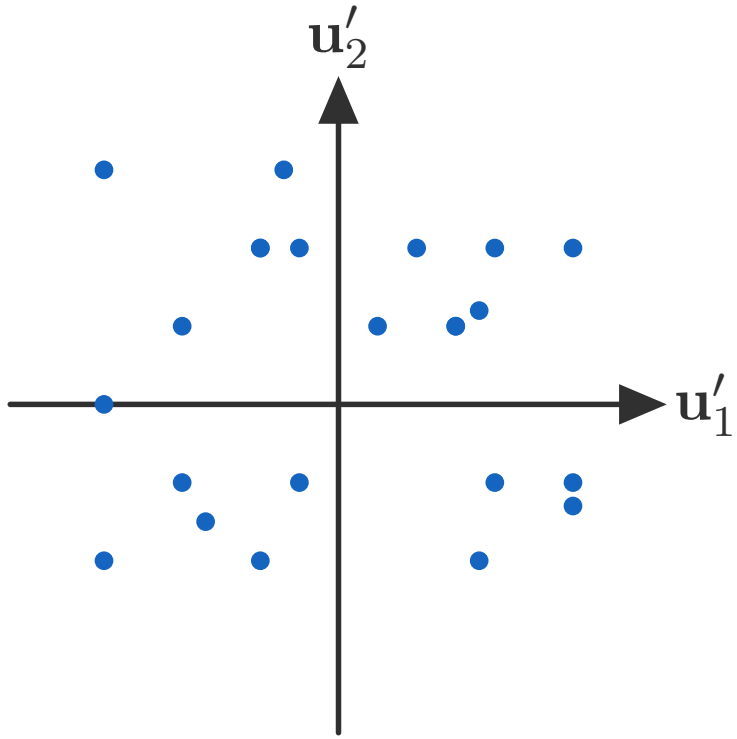
Ces relations montrent qu'il est suffisant de procéder à une seule diagonalisation (par exemple celle de la matrice des corrélations linéaires) afin d'obtenir toutes les informations nécessaires à la synthèse des données dans les deux espaces de représentations.

Analyse du nuage des individus I

Si on cherche à représenter les individus dans un espace de dimension $s < p$, on va considérer un nouveau repère affine qui sera centré en le barycentre des données $\mathbf{m} = (m_1, \cdots, m_p) \in \mathbb{R}^p$ et on prendra comme repère les axes définis par les vecteurs propres de la matrice de corrélations linéaires C.

Dans ce nouvel espace, les individus auront alors pour coordonnées les composantes principales des axes $\mathbf{p}_{\mathbf{u}'_1}(Z), \mathbf{p}_{\mathbf{u}'_2}(Z), \cdots, \mathbf{p}_{\mathbf{u}'_s}(Z)$.

Analyse du nuage des individus II



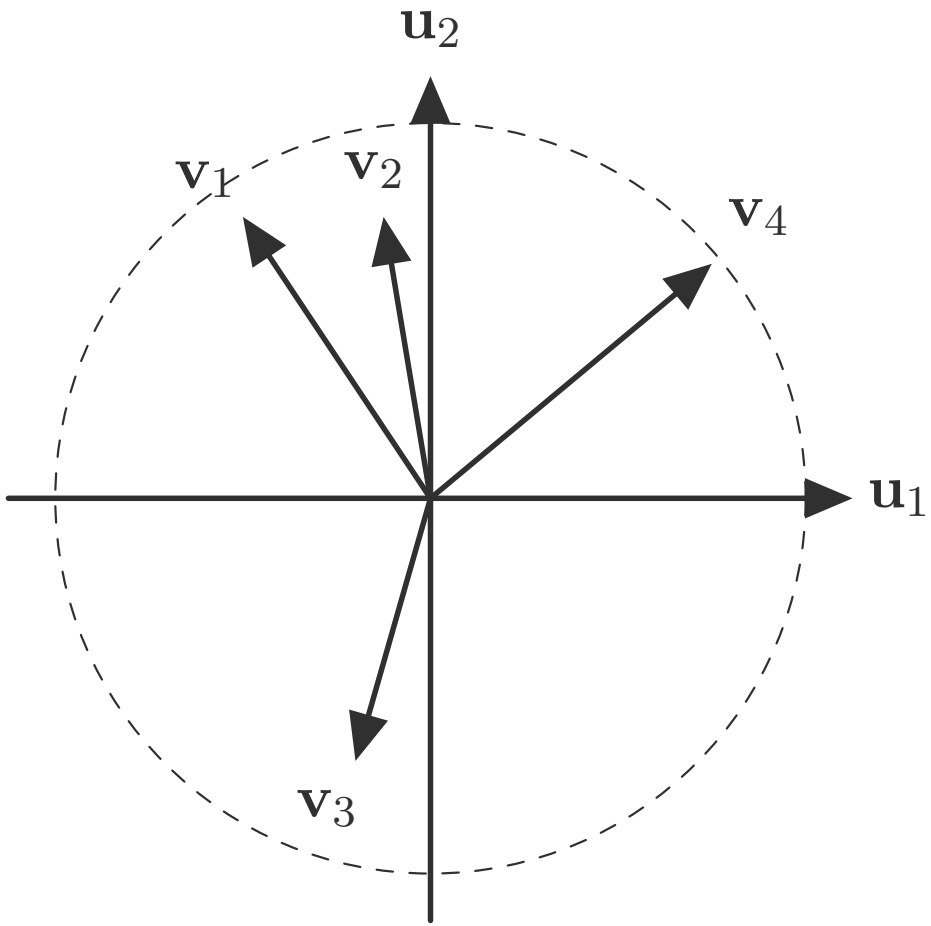
Projection des individus dans l'espace engendré par $(\mathbf{u}'_1, \mathbf{u}'_2)$.

Analyse du nuage des variables I

En ce qui concerne le variable, on considère le repère affine centré en l'origine de l'espace $\mathbb{R}^n$ (et non plus centré en le barycentre !) dont une base est formée par les axes principaux $\mathbf{u}_1, \ldots, \mathbf{u}_s$ qui forment à nouveau une base orthonormale. On rappelle que, dans cet espace, les points ont, sur chaque axe principal, comme coordonnées les valeurs $\mathbf{p}_{\mathbf{u}_1}(Z^T), \mathbf{p}_{\mathbf{u}_2}(Z^T), \ldots, \mathbf{p}_{\mathbf{u}_s}(Z^T)$.

Pour rappel, ces coordonnées ne sont rien d'autres que les coefficients de corrélations linéaires entre les variables $\mathbf{v}_k$ et l'axe principale $\mathbf{u}_k$. Les coordonnées sont donc toutes comprises dans l'intervalle $[-1, 1]$

Analyse du nuage des variables II

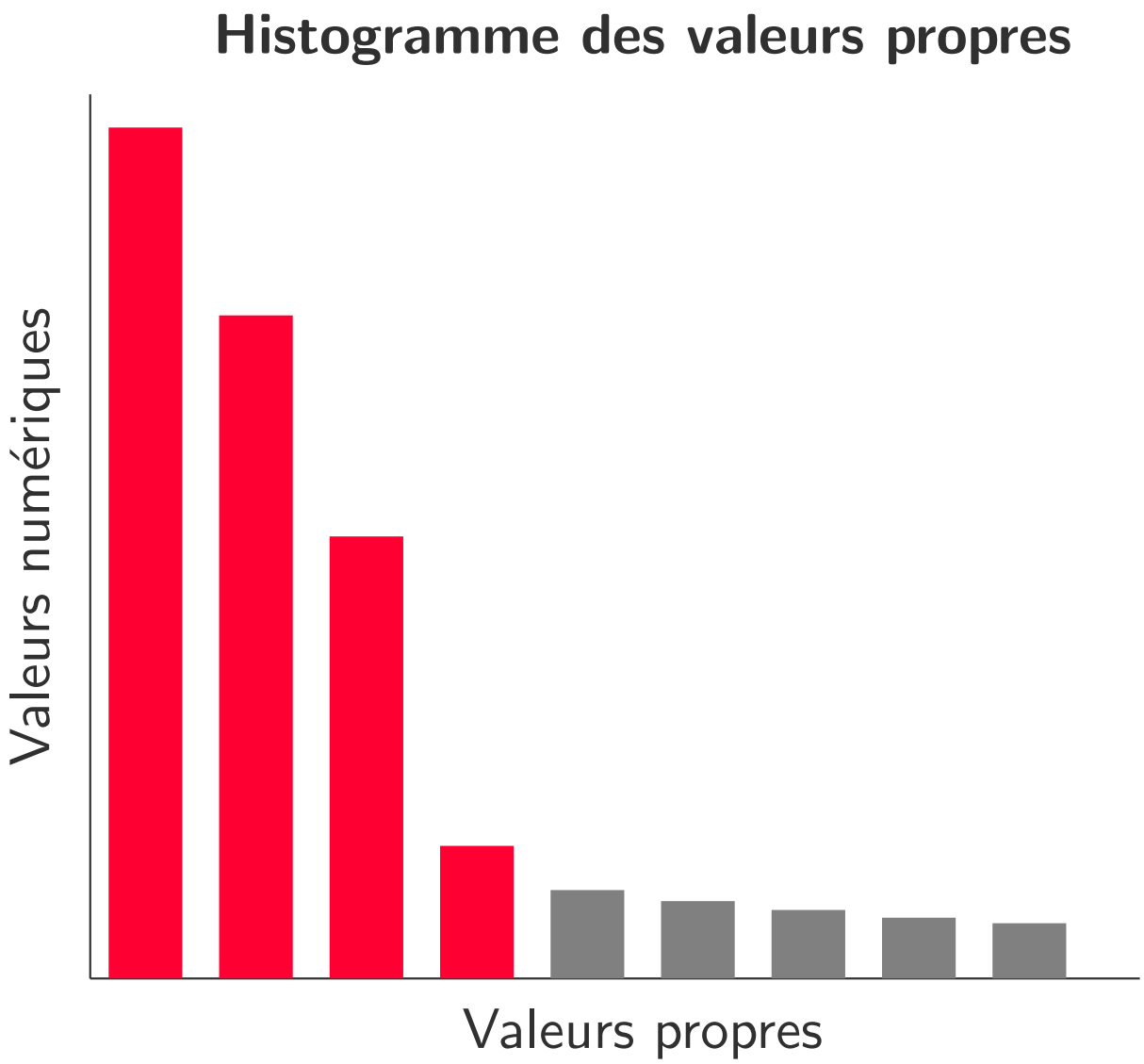


Cercle des corrélations dans l'espace engendré par $(\mathbf{u}_1, \mathbf{u}_2)$.

Choix du nombre d'axes I

- **La règle de Cattell :** cette première représentation se base sur l'utilisation d'un histogramme sur lequel on représente les valeurs propres par ordre décroissant.
  Le but est de chercher une rupture dans la courbe de décroissance des valeurs propres et de ne conserver que les espaces propres associés aux valeurs propres se trouvant avant cette rupture.
- **La règle de Kayser :** cette règle est un peu plus simple que la précédente, elle consiste simplement à considérer les espaces propres dont les valeurs propres associées sont supérieures à 1.
  En effet, la moyenne des valeurs propres est égale à 1 Cela revient donc à conserver les axes dont la variance est supérieure à la moyenne.

Choix du nombre d'axes II



Qualité de la représentation I

On pourra cependant donner quelques critères quantitatifs pour juger de la qualité de la représentation que ce soit **à l'échelle des individus** ou encore **à l'échelle des variables**.

- On peut évaluer la qualité de la représentation de l'individu i **sur l'axe** $\mathbf{u}'_m$ en évaluant l'inertie de la projection de cet individu sur l'axe $\mathbf{u}_m$ et en divisant cela par l'inertie totale de l'individu i.
  Cette valeur est égale au $\cos^2$ de l'angle formé entre le vecteur de la projection de $\mathbf{z}_i$ sur $\mathbf{u}'_m$ et $\mathbf{z}_i$:

$$\frac{\left(p_{\mathbf{u}'_m}(\mathbf{z}_i)\right)^2}{\sum_{k=1}^{p}(z_{ik})^2}.$$

  On peut également mesurer la qualité de la représentation du nuage complet sur l'axe $\mathbf{u}'_m$ par la quantité suivante

Qualité de la représentation II

$$\frac{\lambda_m}{p}.$$

Enfin, la contribution de l'individu i **à l'axe** $\mathbf{u}_m$ est l'inertie de la projection de i sur $\mathbf{u}'_m$ mais divisée cette fois-ci par l'inertie totale du nuage projeté sur ce même axe

$$\frac{\left(p_{\mathbf{u}'_m}(\mathbf{z}_i)\right)^2}{\lambda_m}.$$

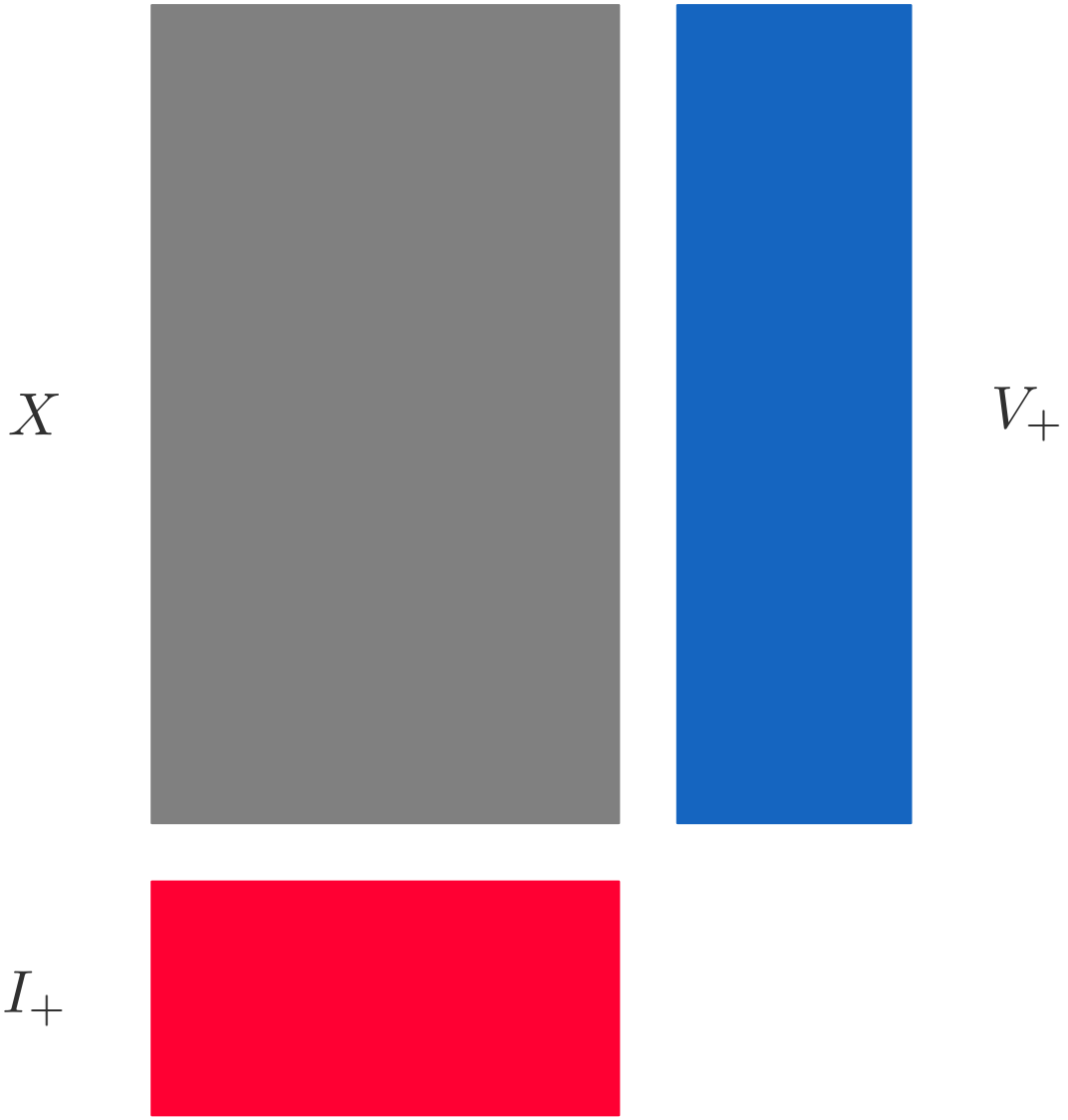
Qualité de la représentation III

- Nous pouvons faire la même chose en ce qui concerne les variables avec les éléments présentés dans le sections précédentes. On rappelle que le vecteur $\mathbf{p}_{\mathbf{u}_m}(Z^T)$ contient les coordonnées des individus $\mathbf{z}_i$ sur l'axe principal $\mathbf{u}_m$ qui sont des coefficients de corrélation. D'où

$$(\mathbf{p}_{\mathbf{u}'_m}(Z))_k = cor(\mathbf{v}'_k, \mathbf{u}_m).$$

Les variables fortement corrélées à un axe sont donc celles qui contribuent le plus à la définition de cet axe. Etant donnée un axe $\mathbf{u}_m$, on s'intéresse donc aux variables ayant les plus fortes coordonnées sur cet axe.
Sachant que $\mathbf{p}_{\mathbf{u}'_m}(Z) = \sqrt{\lambda_m}\mathbf{u}_m$ mais aussi que $(\mathbf{p}_{\mathbf{u}_m}(Z^T))_k = cor(\mathbf{v}'_k, \mathbf{u}_m)$, on pourra interpréter l'axe principal $\mathbf{p}_{\mathbf{u}'_m}(Z)$ en fonction des groupements des variables ayant une coordonnées forte sur $\mathbf{u}_m$.

Ajout de nouvelles informations I



Ajout de nouvelles informations II

A l'échelle des individus : de multiples raisons font que de nouveaux individus ne sont pas utilisés ou pris en compte pour la réalisation de l'ACP et ne sont utilisés qu'a posteriori : *données tests en Machine Learning - individus qui auraient un impact trop important sur les résultats ou sur lesquels planent une incertitude sur l'extraction des caractéristiques et que l'on préfère écarter pour ne pas fausser l'analyse, ...*

Nous avions procédé au **centrage** (on retranchait le vecteur moyenne $\mathbf{m} = (m_1, \cdots, m_p)$ et à la **réduction** (multiplication par $diag(s_1^{-1}, \cdots, s_p^{-1})$) puis à la normalisation des données initiales X avant d'effectuer l'ACP. Pour ces nouvelles données I_+ on effectue exactement la même chose et on obtient une nouvelle matrice Z_+ dont les éléments sont définis par

$$(Z_+)_{ij} = \frac{(I_+)_{ij} - m_j}{s_j\sqrt{n}}.$$

Ajout de nouvelles informations III

On fera attention au fait que les moyennes m_j et écart-types s_j employés sont bien ceux calculés sur les données initiales X et non sur la concaténation des anciens individus avec les individus supplémentaires.

La projection de ces nouveaux individus sur un axe principal $\mathbf{u}'_m$ se déduit alors directement de ce qui précède :

$$\mathbf{p}_{\mathbf{u}'_m}(Z_+) = Z_+\mathbf{u}'_m.$$

Ajout de nouvelles informations IV

A l'échelle des variables ? Le fonctionnement est un peu différent car ce sont des caractéristiques que l'on n'a pas encore rencontré.
On va cette fois-ci obtenir notre matrice transformée en utilisant comme moyennes et écart-types ceux **des nouvelles variables !**. On définit alors

$$m_{+,j} = \frac{1}{n}\sum_{i=1}^{n}(V_+)_{ij} \text{et} \quad s_{+,j} = \frac{1}{n}\sum_{i=1}^{n}\left((V_+)_{ij} - m_{+,j}\right)^2.$$

On peut alors définir notre matrice Y_+ dont le terme général en (i,j) est donné par

$$(Y_+)_{ij} = \frac{(V_+)_{ij} - m_{+,j}}{s_{+,j}\sqrt{n}}.$$

La projection des nouvelles variables sur le sous-espace engendré par $\mathbf{u}_m$ se déduit alors directement de ce qui précède : $\mathbf{p}_{\mathbf{u}'_m}(Y_+) = Y_+^T\mathbf{u}'_m$.

Pour finir I

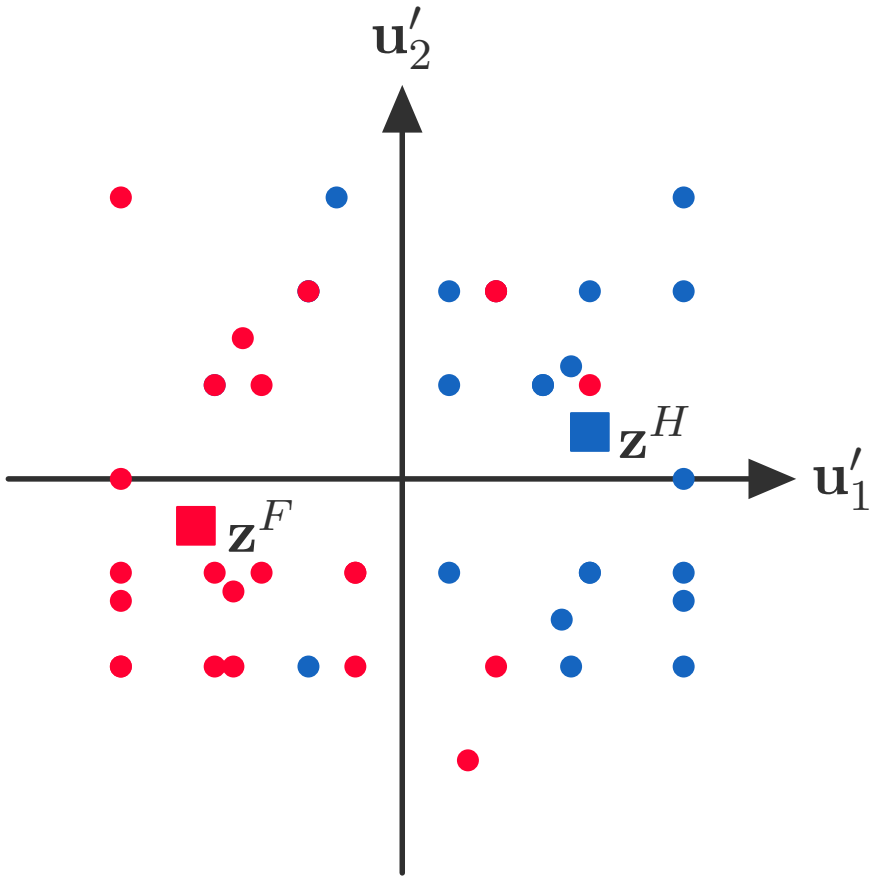
Considérons une variable nominale $\mathbf{q}$ possédant q modalités, prenons l'exemple d'une variable avec deux modalités pour simplifier la présentation (Homme H et Femme F). On peut calculer les barycentres $\mathbf{z}^H$ et $\mathbf{z}^F$ des groupes :

$$\mathbf{z}^H = \frac{1}{|H|} \sum_{i:q_i=H} z_i \quad \text{et} \quad \mathbf{z}^F = \frac{1}{|F|} \sum_{i:q_i=F} z_i.$$

Ces individus moyens peuvent ainsi être projetés sur le sous-espace engendré par un axe principal $\mathbf{u}_m$ comme nous l'avons plus tôt pour la projection de nouveaux individus, *i.e.* à l'aide de la relation :

$$p_{\mathbf{u}'_m}(\mathbf{z}^H) = (\mathbf{z}^H)^T \mathbf{u}'_m = \langle \mathbf{z}^H, \mathbf{u}'_m \rangle \quad \text{et} \quad p_{\mathbf{u}'_m}(\mathbf{z}^F) = (\mathbf{z}^F)^T \mathbf{u}'_m = \langle \mathbf{z}^H, \mathbf{u}'_m \rangle.$$

Pour finir II



Analyses de Données

Généralisation des méthodes

Consignes

Je vous invite à lire la Section 9 du polycopié de cours afin de comprendre le principe.

L'idée de cette section est de vous présenter la définition des axes principaux et les projections dans le cas où l'on attribue des poids non uniformes aux différents individus et lorsque la métrique employée n'est pas euclidienne mais quelconque.

La partie n'est pas très longue à lire, il vous est donc demandé de l'étudier chez vous et je répondrai à vos questions lors de la prochaine séance. Elle est cependant importante pour la présentation de l'AFC et de l'ACM.

Analyses de Données

Analyse Factorielle des Correspondances (AFC)

Contexte I

L'AFC consiste à analyser des tables de contingence , tables qui croisent deux variables qualitatives, disons P et Q, ayant chacune un certain nombre de modalités p et q respectivement. Ainsi, une table de contingence, notée N, sera un tableau de taille $p \times q$ dont chaque chaque entrée n_{ij} sera égal au nombre d'individus possédant **à la fois la caractéristique p_i et la caractéristique q_j**.

$$N = \begin{array}{c} \\ \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_i \\ \vdots \\ \mathbf{p}_n \end{array} \begin{array}{c} \begin{array}{cccccc} \mathbf{q}_1 & \cdots & \mathbf{q}_j & \cdots & \mathbf{q}_p \end{array} \\ \begin{pmatrix} n_{11} & \cdots & n_{1j} & \cdots & n_{1q} \\ \vdots & & \vdots & & \vdots \\ n_{i1} & \cdots & n_{ij} & \cdots & n_{iq} \\ \vdots & & \vdots & & \vdots \\ n_{p1} & \cdots & n_{pj} & \cdots & n_{pq} \end{pmatrix} \end{array},$$

Contexte II

Un exemple (purement fictif) de matrice de contingence est donné en considérant deux modalités :

- **animal de compagnie** (P) dont les modalités sont *chat* - *chien* - *souris* - *lézard*,
- **couleur des yeux** (Q) dont les modalités sont *bleu* - *vert* - *marron*

Pour cela on dispose d'un échantillon de 248 personnes.

$$N = \begin{array}{r} \\ \text{chat} \\ \text{chien} \\ \text{souris} \\ \text{lézard} \end{array} \begin{array}{c} \begin{array}{ccc} \text{bleu} & \text{vert} & \text{marron} \end{array} \\ \begin{pmatrix} 45 & 34 & 12 \\ 3 & 21 & 90 \\ 2 & 6 & 8 \\ 12 & 8 & 7 \end{pmatrix} \end{array},$$

Analyses Préliminaires I

L'objectif de cette partie est de commencer par quelques analyses simples de la table de contingence.

$$N = \begin{array}{c} \\ \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_i \\ \vdots \\ \mathbf{p}_n \\ \\ \end{array} \begin{array}{c} \begin{array}{cccccc} \mathbf{q}_1 & \cdots & \mathbf{q}_j & \cdots & \mathbf{q}_p & \end{array} \\ \begin{pmatrix} n_{11} & \cdots & n_{1j} & \cdots & n_{1q} \\ \vdots & & \vdots & & \vdots \\ n_{i1} & \cdots & n_{ij} & \cdots & n_{iq} \\ \vdots & & \vdots & & \vdots \\ n_{p1} & \cdots & n_{pj} & \cdots & n_{pq} \end{pmatrix} \\ \begin{array}{cccccc} n_{.1} & \cdots & n_{.j} & \cdots & n_{.q} & \end{array} \end{array} \begin{array}{c} \\ n_{1.} \\ \vdots \\ n_{i.} \\ \vdots \\ n_{p.} \\ n \end{array}$$

Analyses Préliminaires II

On peut d'abord définir ce que l'on appelle des **marges** relatives à cette table :

- $n_{i\cdot} = \sum_{j:q_j \in Q} n_{ij}$ qui correspond au nombre d'individus ayant la modalité p_i pour la variable P.
- $n_{\cdot j} = \sum_{i:p_i \in P} n_{ij}$ qui correspond au nombre d'individus ayant la modalité q_j pour la variable Q.

On peut alors avoir accès au nombre total d'individus n par les relations

$$n = \sum_{j:q_j \in Q} n_{\cdot j} = \sum_{i:p_i \in P} n_{i\cdot} = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}.$$

Analyses Préliminaires III

Reprenons notre exemple précédent avec nos 248 personnes et faisons cette fois-ci figurer les marges :

$$N = \begin{array}{r} \\ \text{chat} \\ \text{chien} \\ \text{souris} \\ \text{lézard} \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} \text{bleu} & \text{vert} & \text{marron} \end{array} \\ \begin{pmatrix} 45 & 34 & 12 \\ 3 & 21 & 90 \\ 2 & 6 & 8 \\ 12 & 8 & 7 \end{pmatrix} \\ \begin{array}{ccc} 62 & 69 & 117 \end{array} \end{array} \begin{array}{l} \\ 91 \\ 114 \\ 16 \\ 27 \\ 248 \end{array} ,$$

Analyses Préliminaires IV

Cette première table N est appelée **table des effectifs bruts**, mais il est parfois d'usage de travailler non pas avec les effectifs mais plutôt avec les fréquences. Pour cela on définit une table des fréquences F ayant la même dimension que F et de terme général

$$f_{ij} = \frac{n_{ij}}{n},$$

i.e. on regarde simplement la proportion d'individus qui possèdent à la fois les caractéristiques p_i et q_j parmi les n individus considérés.

$$F = \begin{array}{c} \\ \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_i \\ \vdots \\ \mathbf{p}_n \end{array} \begin{array}{c} \begin{array}{cccccc} \mathbf{q}_1 & \cdots & \mathbf{q}_j & \cdots & \mathbf{q}_p \end{array} \\ \begin{pmatrix} f_{11} & \cdots & f_{1j} & \cdots & f_{1q} \\ \vdots & & \vdots & & \vdots \\ f_{i1} & \cdots & f_{ij} & \cdots & f_{iq} \\ \vdots & & \vdots & & \vdots \\ f_{p1} & \cdots & f_{pj} & \cdots & f_{pq} \end{pmatrix} \end{array} \begin{array}{c} \\ f_{1\cdot} \\ \vdots \\ f_{i\cdot} \\ \vdots \\ f_{p\cdot} \end{array}$$

Analyses Préliminaires V

De la même façon on pourra également définir des **marges** sur la matrice de fréquences

- $f_{i\cdot} = \sum_{j:q_j \in Q} f_{ij}$ qui correspond à la proportion d'individus ayant la modalité p_i.
- $f_{\cdot j} = \sum_{i:p_i \in P} f_{ij}$ qui correspond à la proportion d'individus ayant la modalité q_j.

Analyses Préliminaires VI

La somme totale des éléments de F (qui doit valoir 1 !) peut se définir par

$$1 = \sum_{j:q_j \in Q} f_{\cdot j} = \sum_{i:p_i \in P} f_{i\cdot} = \sum_{i=1}^{p} \sum_{j=1}^{q} f_{ij}.$$

Pour la suite de la présentation, nous noterons $\mathbf{p}_i$ **le vecteur des fréquences des modalités de** Q **des individus ayant la modalité** p_i **et** $\mathbf{q}_j$ **le le vecteur des fréquences des modalités de** P **des individus ayant la modalité** q_j.

On ne résonnera donc qu'avec les fréquence à partir de maintenant.

Analyses Préliminaires VII

On reprend à nouveau notre exemple précédent avec notre table de contingence N et on calcule la table des fréquences correspondante

$$F = \begin{array}{c} \\ \text{chat} \\ \text{chien} \\ \text{souris} \\ \text{lézard} \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} \text{bleu} & \text{vert} & \text{marron} & \end{array} \\ \begin{pmatrix} 0.18 & 0.14 & 0.05 \\ 0.01 & 0.09 & 0.36 \\ 0.01 & 0.02 & 0.03 \\ 0.05 & 0.03 & 0.03 \end{pmatrix} \\ \begin{array}{ccc} 0.25 & 0.28 & 0.47 \end{array} \end{array} \begin{array}{c} \\ 0.37 \\ 0.46 \\ 0.06 \\ 0.11 \\ 1 \end{array}$$

Cette table des fréquences va nous permettre d'étudier les liaisons ou corrélations entre les variables P et Q

Analyses Préliminaires VIII

Plus précisément, on va plutôt chercher à savoir si **des variables sont indépendantes ou non**. D'un point de vue statistique, deux variables P et Q sont indépendantes, notée $P \perp Q$ si les fréquences observées pour une variable donnée ne dépendent pas des fréquences des modalités observées pour l'autre variable. On peut aussi traduire cela par :

$$\forall (i,j) \ : \ (p_i, q_j) \in P \times Q \ f_{ij} = f_{i\cdot} f_{\cdot j}.$$

A travers cette relation on peut aussi dire que la probabilité qu'un individu possède en même temps les attributs p_i et q_j est égale à la probabilité qu'un individu possède l'attribut p_i, $f_{i\cdot}$ multipliée par la probabilité qu'il possède l'attribut q_j, $f_{\cdot j}$. On dit aussi que **la probabilité jointe est égale au produit des probabilités marginales**.
Ce que l'on va chercher à faire ici c'est de mesurer l'écart à cette indépendance en comparant les valeurs de f_{ij} aux produits $f_{i\cdot} f_{\cdot j}$.

Analyses Préliminaires IX

Une telle comparaison offre deux cas possibles :

- $f_{ij} > f_{i\cdot}f_{\cdot j}$, la probabilité jointe est supérieure au produit des probabilités marginales, cela signifie que les modalités p_i et q_j **s'attirent**,
- au contraire si $f_{ij} < f_{i\cdot}f_{\cdot j}$, la probabilité jointe est inférieure au produit des probabilités marginales, cela signifie que les modalités p_i et q_j **se repoussent**.

Pour cela on va comparer une table des fréquences dite empirique , qui contient les fréquences observées,*i.e.* il s'agit de notre table F, et on va construire une deuxième table F' qui contiendra les effectifs théoriques dans le cas où les deux variables étudiées sont indépendantes.

Analyses Préliminaires X

Reprenons la table des fréquences observées F de notre exemple

$$F = \begin{array}{c} \\ \text{chat} \\ \text{chien} \\ \text{souris} \\ \text{lézard} \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} \text{bleu} & \text{vert} & \text{marron} & \\ \end{array} \\ \begin{pmatrix} 0.18 & 0.14 & 0.05 \\ 0.01 & 0.09 & 0.36 \\ 0.01 & 0.02 & 0.03 \\ 0.05 & 0.03 & 0.03 \end{pmatrix} \\ \begin{array}{ccc} 0.25 & 0.28 & 0.47 \end{array} \end{array} \begin{array}{c} \\ 0.37 \\ 0.46 \\ 0.06 \\ 0.11 \\ 1 \end{array}.$$

et construisons la table des fréquences théoriques F' dans le cas où nos deux variables sont indépendantes, pour rappel, il suffit simplement de faire le produit des marginales qui sont données ci-dessus

Analyses Préliminaires XI

$$F = \begin{array}{r} \\ \text{chat} \\ \text{chien} \\ \text{souris} \\ \text{lézard} \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} \text{bleu} & \text{vert} & \text{marron} \end{array} \\ \begin{pmatrix} 0.10 & 0.10 & 0.17 \\ 0.11 & 0.13 & 0.22 \\ 0.01 & 0.02 & 0.03 \\ 0.03 & 0.03 & 0.05 \end{pmatrix} \\ \begin{array}{ccc} 0.25 & 0.28 & 0.47 \end{array} \end{array} \begin{array}{c} \\ 0.37 \\ 0.46 \\ 0.06 \\ 0.11 \\ 1 \end{array}$$

Les entrées mises en avant en $\text{tr}\{rouges\}représententdescasoùlesmodalitésontendanceà$**s'attirer**, $c'est-à-direque\text{f}_{ij} > f_{i\cdot}f_{\cdot j}$. A l'inverses, celles en bleues représentent des cas où les modalités ont plutôt tendance à **se repousser** c'est-à-dire que $f_{ij} < f_{i\cdot}f_{\cdot j}$.

Analyses Préliminaires XII

En cas d'indépendances entre les deux variables, nous avons les propriétés suivantes sur les lignes et les colonnes

- les lignes i de terme général $\frac{f_{ij}}{f_{i\cdot}}$ sont **proportionnelles à la marge** :

$$f_{ij} = f_{i\cdot} f_{\cdot j} \iff \frac{f_{ij}}{f_{i\cdot}} = f_{\cdot j}$$

- les colonnes j de terme général $\frac{f_{ij}}{f_{\cdot j}}$ sont **proportionnelles à la marge** :

$$f_{ij} = f_{i\cdot} f_{\cdot j} \iff \frac{f_{ij}}{f_{\cdot j}} = f_{i\cdot}$$

Analyse Factorielle des correspondances

Contrairement à l'ACP, on ne parlera pas ici d'espace des *variables* ou des *individus*, la structure des données étant très différente. On va plutôt parler de profils :

- **les profils lignes :** qui peuvent être vus comme des vecteurs de l'espace $\mathbb{R}^q$. Ils correspondent aux *individus* dans le cadre de l'ACP.
- **les profils colonnes :** qui peuvent être vus comme des vecteurs de l'espace $\mathbb{R}^p$. Ils correspondent aux *variables* dans le cadre de l'ACP.

L'objectif reste le même, obtenir une représentation des données dans un espace de dimension inférieure à l'espace de départ.

Applications aux profils lignes I

il s'agit d'une matrice L de taille $p \times q$ dont chaque entrée l_{ij} est définie par

$$l_{ij} = \frac{f_{ij}}{f_{i\cdot}}.$$

Ainsi

$$L = \begin{array}{c} \\ \mathbf{l}_1 \\ \vdots \\ \mathbf{l}_i \\ \vdots \\ \mathbf{l}_n \end{array} \begin{pmatrix} \frac{f_{11}}{f_{1\cdot}} & \cdots & \frac{f_{1j}}{f_{1\cdot}} & \cdots & \frac{f_{1q}}{f_{1\cdot}} \\ \vdots & & \vdots & & \vdots \\ \frac{f_{i1}}{f_{i\cdot}} & \cdots & \frac{f_{ij}}{f_{i\cdot}} & \cdots & \frac{f_{iq}}{f_{i\cdot}} \\ \vdots & & \vdots & & \vdots \\ \frac{f_{p1}}{f_{p\cdot}} & \cdots & \frac{f_{pj}}{f_{p\cdot}} & \cdots & \frac{f_{pq}}{f_{p\cdot}} \end{pmatrix} \begin{array}{c} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{array},$$

Applications aux profils colonnes I

Analyses de Données

Analyse factorielle des Correspondances Multiples (ACM)

The End