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 $\mathbf{v}_k$.

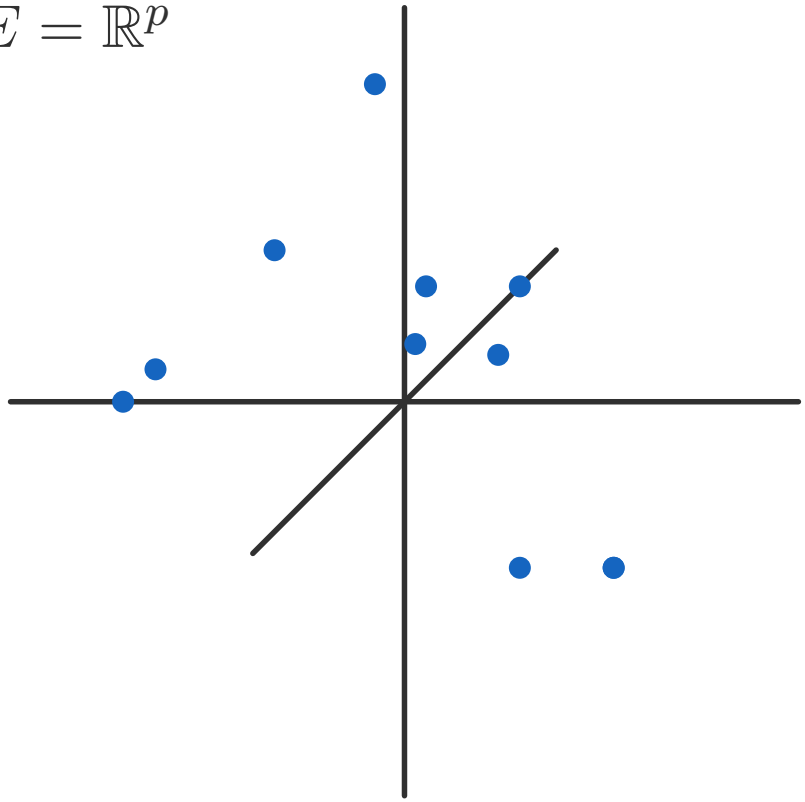
Introduction III

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

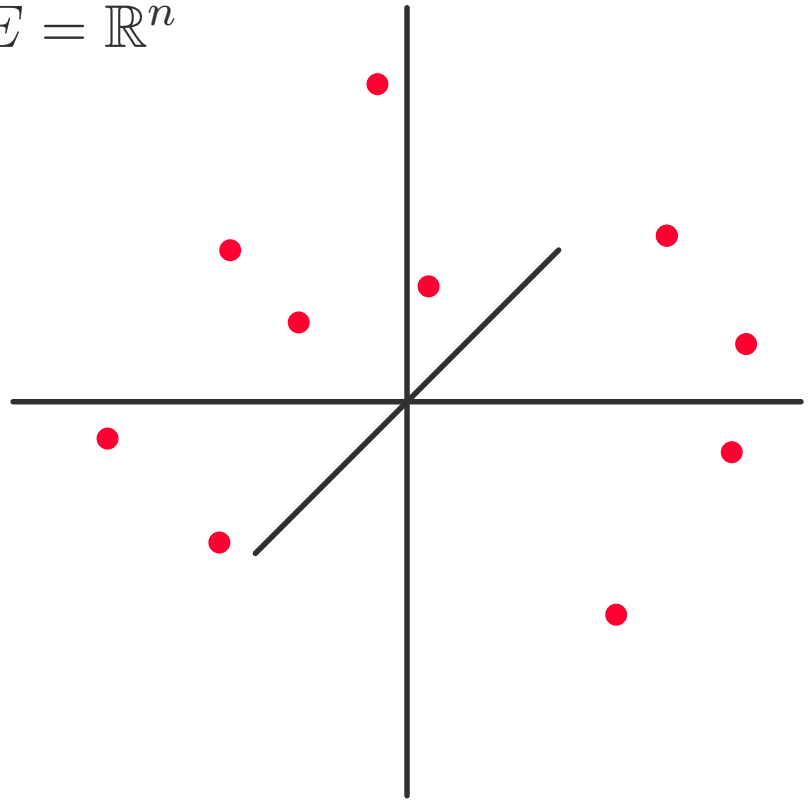
- On peut choisir de représenter les individus $\mathbf{x}_i$ dans l'espaces des variables $\mathbf{v}_k$, une première représentation qui est sûrement la plus utilisée. Dans ce cas chaque point $\mathbf{x}_i$ a pour coordonnées $(\mathbf{x}_{i1}, \cdots, \mathbf{x}_{ip})$ dans l'espace $\mathbb{R}^p$.
  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

$E = \mathbb{R}^p$



$E = \mathbb{R}^n$



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

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

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

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

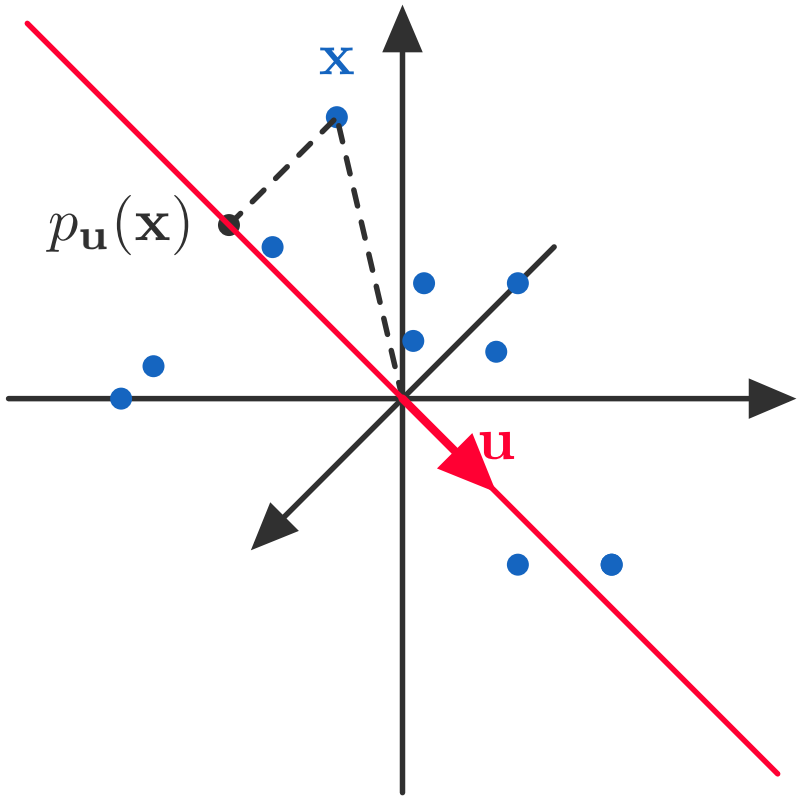
Ainsi on peut définir des distances pour tout entier $p > 1$ comme dans le cas des normes

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

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

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 projet 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 $X^T\mathbf{u}'_k = \mathbf{u}_k$

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

Nous avons également la relation inverse

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

Analyses de Données

Analyse en Composantes Principales (ACP)

Analyses de Données

Généralisation des méthodes

Analyses de Données

Analyse Factorielle des Correspondances (AFC)

Analyses de Données

Analyse factorielle des Correspondances Multiples (ACM)

The End