

Algèbre Linéaire et **Analyse de Données**

Licence 2 - MIAHS

Guillaume Metzler

Université Lumière Lyon 2
Laboratoire ERIC, UR 3083, Lyon

guillaume.metzler@univ-lyon2.fr

Printemps 2022

Analyse de Données

Summary

1 Analyses de Données

- Généralités et Décomposition en Valeurs Singulières (SVD)
- Analyse en Composantes Principales (ACP)
- Généralisation des méthodes
- Analyse Factorielle des Correspondances (AFC)
- Analyse factorielle des Correspondances Multiples (ACM)

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).

$$X = \begin{matrix} & \mathbf{v}_1 & \cdots & \mathbf{v}_k & \cdots & \mathbf{v}_p \\ \mathbf{x}_1 & \left(\begin{array}{ccccc} x_{11} & \cdots & x_{1k} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ \mathbf{x}_i & x_{i1} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ \mathbf{x}_n & x_{n1} & \cdots & x_{nk} & \cdots & x_{np} \end{array} \right) \end{matrix},$$

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

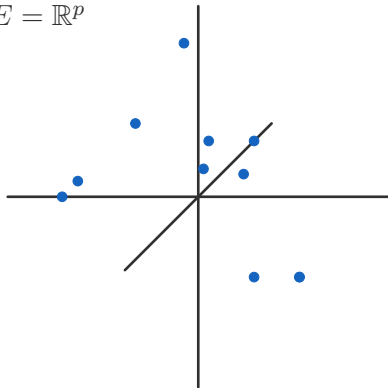
Introduction III

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

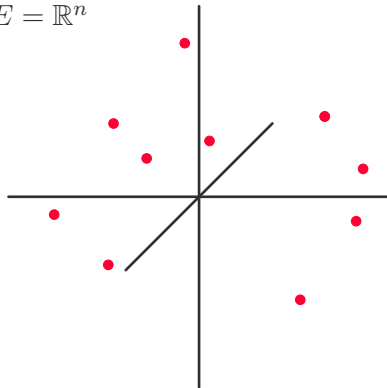
- On peut choisir de représenter les **individus** \mathbf{x}_i dans l'espaces des **variables** \mathbf{v}_k , une première représentation qui est sûrement la plus utilisée. Dans ce cas chaque point \mathbf{x}_i a pour coordonnées $(\mathbf{x}_{i1}, \dots, \mathbf{x}_{ip})$ dans l'espace \mathbb{R}^p .
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 $(\mathbf{v}_{1k}, \dots, \mathbf{v}_{nk})$ 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

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

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 , toute application d de $E \times E$ à valeurs dans \mathbb{R}^+ qui vérifient les propriétés suivantes :

- **symétrie** : $\forall \mathbf{x}, \mathbf{x}' \in E, d(\mathbf{x}, \mathbf{x}') = d(\mathbf{x}', \mathbf{x})$
- **séparation** : $\forall \mathbf{x}, \mathbf{x}' \in E, d(\mathbf{x}, \mathbf{x}') = 0 \iff \mathbf{x} = \mathbf{x}'$
- **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

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

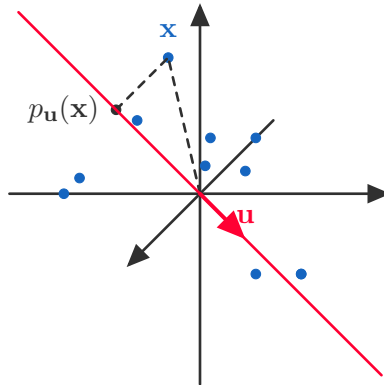
Ainsi on peut définir des distances pour tout entier $p > 1$ comme dans le cas des normes

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

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

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

↓ car \mathbf{u} est un vecteur unitaire

$$= \langle \mathbf{x}, \mathbf{u} \rangle^2,$$

↓ simple réécriture

$$= (\mathbf{x}^T \mathbf{u})^T (\mathbf{x}^T \mathbf{u}),$$

$$= \mathbf{u}^T \mathbf{x} \mathbf{x}^T \mathbf{u}.$$

Lien avec valeurs et vecteurs propres III

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

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

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

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

Lien avec valeurs et vecteurs propres IV

Définition 1.2: Matrice de Gram

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

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

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 $\mathbf{x}_1, \dots, \mathbf{x}_n$ des vecteurs de E et considérons la matrice G carrée d'ordre n définie par

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

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

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

↓ définition transposée et norme

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

↓ définition de norme

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

↓ \mathbf{u} 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 $G' = X^T X$.

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' ($\mathbf{u}_1, \dots, \mathbf{u}_{p'}$) plus grandes valeurs propres ($\lambda_1, \dots, \lambda_{p'}$) 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 $\mathcal{M}_p(\mathbb{R})$.

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 $\mathcal{M}_n(\mathbb{R})$.

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 :

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

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

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

où les σ_i ($\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq 0$) sont les 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 :

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

↓ définition de X^T

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

↓ orthogonalité de U

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

$$= U' \Sigma_p^2 U'^T,$$

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

↓ définition de X

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

↓ orthogonalité de U'

$$= U \Sigma \Sigma^T U^T,$$

$$= U \Sigma_n^2 U^T,$$

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

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

\implies les valeurs singulières de X sont les racines carrées des valeurs propres des matrices $X^T X$ ou $X X^T$.

Qualité de l'approximation I

Pour cela considérons notre matrice $X \in \mathcal{M}_{n,p}(\mathbb{R})$, où $\mathcal{M}_{n,p}(\mathbb{R})$ 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.

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

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 :

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

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 $\mathcal{M}_{n,p}(\mathbb{R})$ et considérons le problème d'optimisation suivant :

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

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 singulières ainsi que r vecteurs singuliers à gauche $(\mathbf{u}_1, \dots, \mathbf{u}_r)$ et à droite $(\mathbf{u}'_1, \dots, \mathbf{u}'_r)$ qui sont normés.

Ainsi la meilleure approximation de rang s est donnée par

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

où les valeurs singulières sont rangés par ordre décroissant :

$$\sigma_1 \geq \sigma_2 \geq \dots \sigma_s.$$

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

$$\tilde{X} = \sigma_1 \begin{array}{c} \mathbf{u}'_1 \\ \text{[red box]} \\ \mathbf{u}_1 \end{array} + \sigma_2 \begin{array}{c} \mathbf{u}'_2 \\ \text{[red box]} \\ \mathbf{u}_2 \end{array} + \dots + \sigma_s \begin{array}{c} \mathbf{u}'_s \\ \text{[red box]} \\ \mathbf{u}_s \end{array}$$

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

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

↓ en utilisant la SVD

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

↓ on arrange les termes

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

↓ les vecteurs \mathbf{u}_k et \mathbf{u}'_k 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 $\mathcal{M}_{3,2}(\mathbb{R})$. 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

$$X^T X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{pmatrix} \text{ en déterminant les racines du polynôme}$$

caractéristique $\mathcal{X}_{X^T X}(\lambda)$ défini par :

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

Exemple II

↓ on développe selon la ligne

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

↓ on développe les deux déterminants d'ordre 2 et on factorise
 $-\lambda(\lambda - 3)(\lambda - 1).$

Ainsi les racines de $\mathcal{X}_{X^T X}(\lambda)$ et donc les valeurs propres de $X^T X$ 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 $X^T X$. 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

$$X^T X - 3I_3 = \begin{pmatrix} -2 & 1 & 0 \\ 1 & -1 & 1 \\ 0 & 1 & -2 \end{pmatrix} \text{ dont la forme échelonnée réduite est}$$

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 donné 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 $X^T X - I_3 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$ 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 **quantitatifs réelles**, *i.e.* les variables prennent des valeurs réelles.

$$X = \begin{matrix} & \mathbf{v}_1 & \cdots & \mathbf{v}_k & \cdots & \mathbf{v}_p \\ \mathbf{x}_1 & \left(\begin{array}{ccccc} x_{11} & \cdots & x_{1k} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ \mathbf{x}_i & \begin{array}{ccccc} x_{i1} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ \mathbf{x}_n & \begin{array}{ccccc} x_{n1} & \cdots & x_{nk} & \cdots & x_{np} \end{array} \end{array} \right), \end{matrix}$$

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

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

ainsi que la variance s_k^2 de ces mêmes variables

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

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 à 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 $\mathbf{m} = (m_1, \dots, m_p)$ 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 jeu 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 $\hat{t}_1, \dots, \hat{t}_s$ 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 \hat{t}_1 est le sous-espace de dimension 1 qui maximise l'inertie du nuage du point après projection,
- l'axe défini par le vecteur \hat{t}_2 , *orthogonal* à \hat{t}_1 est le deuxième sous-espace de dimension 1 qui maximise l'inertie du nuage du point après projection,
- on continue avec \hat{t}_3 qui est *orthogonal* à \hat{t}_1 et \hat{t}_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 décorréler les variables, il faut donc faire intervenir cette matrices de corrélation. Pour rappel, il s'agit de la matrice $C \in \mathcal{M}_p(\mathbb{R})$ 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 :

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

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 $U' \in \mathcal{M}_p(\mathbb{R})$ et une matrice diagonale $\Sigma_p = \text{diag}(\lambda_1, \dots, \lambda_p)$ telle que $\lambda_1 \geq \dots \geq \lambda_p$. 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 $\hat{\mathbf{u}}$ 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 $\hat{\mathbf{m}}$:

Projection des individus VI

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n p_{\mathbf{u}'_m}(\mathbf{z}_i)^2 &= \frac{1}{n} \sum_{i=1}^n \langle \mathbf{z}_i, \mathbf{u}' \rangle^2 = \frac{1}{n} \sum_{i=1}^n \mathbf{u}'_m{}^T \mathbf{z}_i \mathbf{z}_i^T \mathbf{u}'_m \\ &= \frac{1}{n} \mathbf{u}'_m{}^T C \mathbf{u}'_m = \frac{1}{n} \lambda_m.\end{aligned}$$

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 :

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

$\langle \mathbf{v}'_j, \mathbf{v}'_j \rangle$ 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 $G \in \mathcal{M}_n(\mathbb{R})$ 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 $U \in \mathcal{M}_n(\mathbb{R})$ et une matrice diagonale $\Sigma_n = \text{diag}(\lambda_1, \dots, \lambda_n)$ telle que $\lambda_1 \geq \dots \geq \lambda_n$. 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} sont données par :

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

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