

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

Introduction

Motivations

- Apprendre des techniques permettant de dégager les informations présentes dans un jeu de données → en dégager la substantifique moelle !
- Partir de la représentation la plus classique d'un jeu de données, *i.e.* sous forme d'un tableau, et en apprendre une représentation synthétique sous forme de graphiques.
- Nous verrons aussi quelles techniques adopter en fonction de la nature de nos données.
- On va donc voir comment synthétiser l'information à des fins d'**interprétations** mais aussi de **visualisation**.

Motivations

- On va s'éloigner de la simple statistique descriptive pour se plonger dans l'étude de variables dites **multidimensionnelles**.
- Pourquoi multidimensionnelles ? Parce qu'un objet ou un individu pourra être représenté par une multitude d'informations ou **caractéristiques**.
- Toutes ces informations sont structurées dans des **tableaux/matrices** dont les éléments peuvent à la fois être des **chiffres** ou du **texte**.

Motivations

Les applications sont nombreuses dans le domaine des sciences sociales :

- **Marketing** : déterminer des profils clients dans un panel - identifier les cibles prioritaires et les meilleurs arguments dans des campagnes publicitaires
- **Sociologie** : identifier les profils d'individus les plus sensibles aux fake news dans la population - détecter des communauté d'individus dans les réseaux sociaux
- **Economie** : identification des facteurs clefs pour la prise de décision dans un but commercial
- ...

On retrouve également des applications dans le domaine médical, biologique, de détection de fraudes, ...

Comment faire cela ?

Nos données sont représentées sous une forme structurée, *i.e.* des tableaux ou des matrices de taille $n \times p$.

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

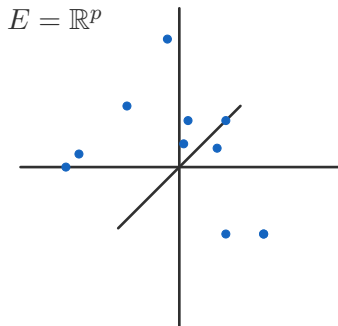
Chaque ligne i représentera un **individu** qui sera décrit par un ensemble de p **caractéristiques** ou **attributs** \mathbf{v}_j que l'on appellera également **variables** (ou **features**).

Individu \rightarrow un vecteur de taille p , *i.e.* $\mathbf{x}_i \in \mathbb{R}^p$.

Variables \rightarrow un vecteur de taille n , *i.e.* $\mathbf{v}_j \in \mathbb{R}^n$.

Deux visualisation possibles

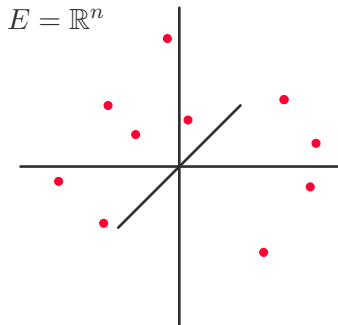
La tableau précédent suggère deux visualisation possibles de notre jeu de données



Une première représentation dans l'**espace des variables**, ainsi chaque individu est représenté par un vecteur de \mathbb{R}^p .

Deux visualisation possibles

On peut également faire le choix de représenter les **variables** dans l'espace des **individus**.



Ainsi chaque **variable** est représenté par un vecteur de \mathbb{R}^n .

Motivations

Nous verrons que les deux représentations permettent de dégager des informations différentes dans notre jeu de données : on pourra, par exemple, détecter d'éventuelles **corrélations** entre des variables ou des **groupes d'individus**.

D'un point de vue technique nous verrons également que ces deux représentations là sont très liées et qu'en étudiant simplement l'une de ces représentations nous sommes capables d'en déduire des informations sur l'autre représentation.

Objectifs

- Il est difficile d'interpréter et de synthétiser des informations se trouvant dans des espaces à **grande dimension**. Notre vision humaine nous limite à la visualisation en 3 dimension.
- Il va donc falloir développer des techniques qui permettant la synthèse et la visualisation de ces résultats dans des espaces de dimension faible (2 ou 3 en général) tout en préservant l'information qui était présente dans la représentation initiale.
- La quantité d'informations sera représentée par les notions de **variance** (quand on étudiera les variables) ou encore par la notion de **distances entre les individus** (si on étudie les individus et non les variables).

Un exemple pour motiver ce cours I

Nous demandons aux étudiants qui suivent ce cours s'ils sont satisfaits, à travers 5 critères sur lesquels ils doivent se positionner sur une échelle de 1 à 5, 1 correspondant à "très insatisfait" et 5 à "très satisfait". Voici ces 5 critères :

1. Clarté du cours écrit
2. Fluidité de la lecture du cours écrit
3. Facilité à comprendre les exemples du cours
4. Clarté des vidéos
5. Satisfaction vis à vis de l'enseignant dans la vidéo

Un exemple pour motiver ce cours II

	Q1	Q2	Q3	Q4	Q5
Individu 1	3	3	3	5	5
Individu 2	2	3	1	5	4
Individu 3	2	3	3	4	5
Individu 4	1	1	1	1	1
Individu 5	5	5	4	3	3
Individu 6	4	5	5	2	3
Individu 7	5	5	5	3	3
Individu 8	1	1	1	1	1

TABLE – Résultats du questionnaire sur un ensemble de 8 individus.

Un exemple pour motiver ce cours III

Les outils statistiques de statistiques descriptives nous permettent simplement de décrire un individu *moyen*.

Il laisse à penser que l'échantillon est globalement moyennement satisfait vis à vis de l'ensemble des critères étudiés.

	Q1	Q2	Q3	Q4	Q5
Moyenne	2.875	3.25	3	3	3.125

TABLE – Réponse moyenne sur les différents critères vis à vis des réponses fournies au questionnaire et présentées dans la Table 1.

En réalité, différents profils peuvent être extraits de cette table...

Un exemple pour motiver ce cours IV

	Q1	Q2	Q3	Q4	Q5
Individu 1	3	3	3	5	5
Individu 2	2	3	1	5	4
Individu 3	2	3	3	4	5
Individu 4	1	1	1	1	1
Individu 5	5	5	4	4	3
Individu 6	4	5	5	2	3
Individu 7	5	5	5	3	3
Individu 8	1	1	1	1	1

TABLE – Création de groupes en fonction des réponses des individus. On identifie trois profils différents en fonction de la nature des réponses, identifiés en marron, gris et blanc.

Un exemple pour motiver ce cours V

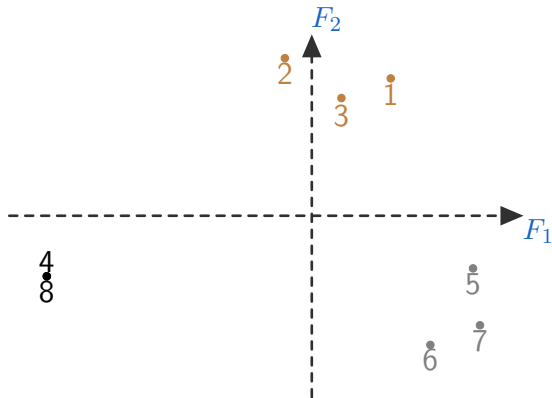


FIGURE – Représentation des individus, *i.e.* des réponses fournies au questionnaire, dans un espace à deux dimensions obtenu par l'ACP.

Quels outils ?

Pour cela on aura besoin de techniques basées sur l'**Algèbre Linéaire** et nous allons, entre autre, mobiliser ce que vous avez pu voir sur les matrices :

- Propriétés des matrices
- Représentation des vecteurs dans une base
- Réduction des endomorphismes

Mais aussi d'outils géométriques pour mieux appréhender et utiliser les notions de distances :

- Espace euclidien
- Projections
- Produit scalaire - norme

Quelles techniques ?

Les techniques d'**analyse de données** que nous allons étudier vont mobiliser les outils d'algèbre linéaire qui seront présentés.

D'un point de vue technique ou d'un point de vue mathématique, il n'y aura rien de nouveau, il faudra simplement comprendre et interpréter les résultats obtenus avec nos outils mathématiques.


Enfin, nous verrons bien sûr quels outils employés et dans quelles circonstances ainsi que l'interprétation des résultats en fonction du type de données.

Présentations des enseignants

- **Chargé de CM :** Guillaume METZLER - MCF Informatique - Enseignant à l'ICOM et rattaché au Laboratoire ERIC (Batiment K - Bureau K073)
- **Intervenants en TD :**
 - Eliz Peyraud - Doctorante en Maths Appliquées - ICOM
 - Zied Selmi - ATER en Informatique (Computer Vision)
 - Moi-même x)

Déroulement du cours

Programme :

- 11 séances de CM portant sur l'Algèbre Linéaire et l'Analyse de Données.
Une séance est également réservée pour effectuer un partiel de 1h45 sur la première partie du cours, consacrée à l'Algèbre Linéaire
- 7 séances de TD sur la partie Algèbre Linéaire puis 4 séances de TIC/TP sur la partie Analyses de Données (sur machine avec ).
La dernière séances de TIC/TP sera évaluée à nouveau.

Au total, vous aurez donc deux évaluations de 2h00.

Summary

1 Introduction

2 Algèbre Linéaire

- Espaces vectoriels et applications linéaires
- Espaces vectoriels de dimension finie
- Matrices et calcul matriciel
- Systèmes linéaires
- Réduction des endomorphismes
- Formes quadratiques et espaces euclidiens

3 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)

Cours et Supports

Le cours se compose :

- des planches pour la présentation des cours magistraux
- d'un polycopié qui reprend de façon détaillée le contenu des slides
- d'un petit poly avec des exercices

L'ensemble de ces supports sont disponibles à l'adresse suivante (les slides seront régulièrement mis à jour) :

https://guillaumemetzler.github.io/aladd_lyon2.html

Attention : présence aux TD obligatoires ! Aucune correction ne sera mise en ligne.

C'est parti !