

# Statistique pour la Saé

## Introduction à l'analyse de données

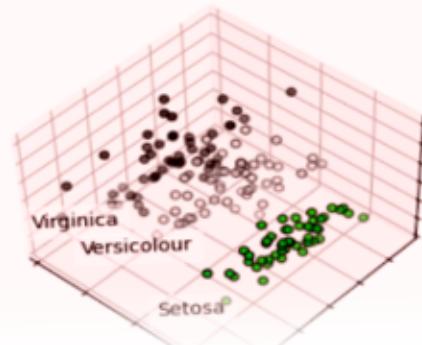
### SAé 2.04 – Exploitation d'une base de données

2023-2024

#### Dimensionality reduction

Reducing the number of random variables to consider.

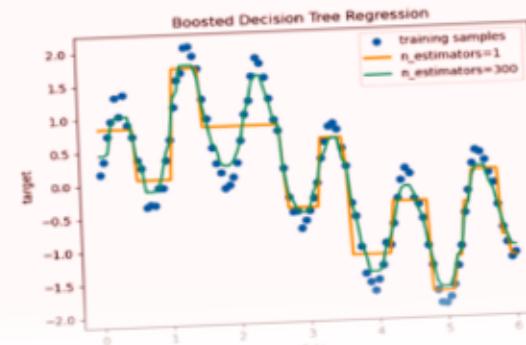
**Applications:** Visualization, Increased efficiency  
**Algorithms:** PCA, feature selection, non-negative matrix factorization, and more...



#### Regression

Predicting a continuous-valued attribute associated with an object.

**Applications:** Drug response, Stock prices.  
**Algorithms:** Gradient boosting, nearest neighbors, random forest, ridge, and more...

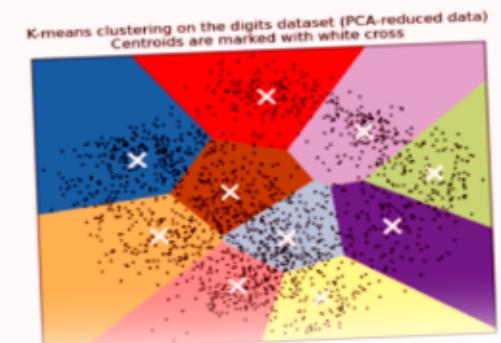


#### Clustering

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experiment outcomes

**Algorithms:** k-Means, HDBSCAN, hierarchical clustering, and more...



## • La SAé 2.04 : Exploitation d'une Base de Données

### Données de la SAé 2.04

Travail sur un jeu de données regroupant 4 fichiers sur les collèges en France en 2022, venant du site [data.gouv.fr](http://data.gouv.fr) :

- Effectifs d'élèves par section, niveau, genre, langues vivantes
- Etablissements de l'éducation prioritaire
- Indices de position sociale dans les collèges
- Indicateur de valeur ajoutée des collèges

### Étapes de la SAé 2.04

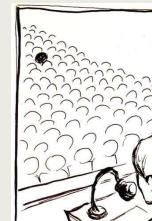
Le travail se fera en 3 parties successives, chacune avec un rendu à déposer sur Moodle :

- Partie 1 - Traduction et implantation d'un schéma relationnel
- Partie 2 - Peuplement et exploitation de la base de données
- **Partie 3 - Analyse statistique des données**

## • Les Statistiques dans la SAé 2.04

### Organisation des Statistiques pour la SAé 2.04

- Après les vacances, pendant 4 semaines, chaque semaine :
    - 2h TD sur PC (Python)  
avec T. Jézéquel (A,B,C,D) ou Yannick Favreau (E)
- ↪ apprentissage d'outils qui seront à utiliser sur la Partie 3 (Statistique) de la SAé
- Pendant la 4e semaine et la suivante
    - 1h de projet encadré (par T.Jézéquel ou Y. Favreau)
    - 2h de projet encadré (par enseignant.e de BD)
    - 3h en autonomie
- ↪ travail sur la Partie 3 (Statistique) de la SAé.



### Plan du cours

1. Analyse de données : 3 types de méthodes
2. Problème commun aux 3 types de méthodes
3. Préparation commune des données : centrer-réduire
4. Une méthode de régression : la régression linéaire multiple
5. Une méthode de clustering : l'algorithme des k-moyennes
6. Une méthode de réduction de dimension : l'ACP



### Évaluation de la Partie 3 (Statistique)

- Un rendu sous forme de rapport d'environ 8 à 10 pages.
- 1 note sur ce rapport, qui compte autant que celle d'un rendu de BD.

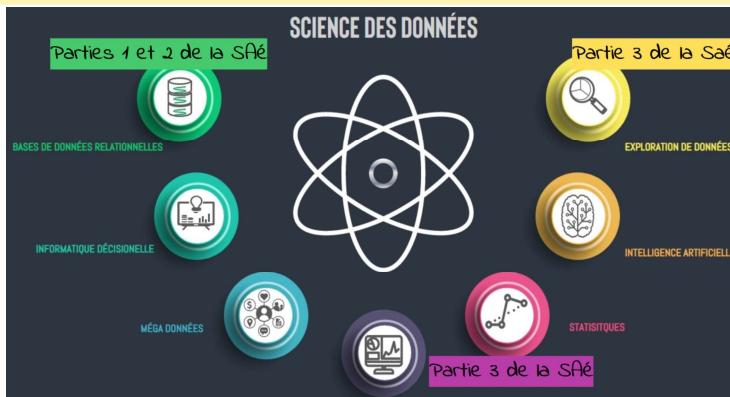


# 1. Analyse de données : 3 types de méthodes

Dans cette SAé : premiers pas en **Science des données**.

## Définition

La **science des données** (*data science* en anglais) est un domaine interdisciplinaire qui utilise les mathématiques, les statistiques, le calcul scientifique, les méthodes scientifiques, les process, les algorithmes et les systèmes informatiques automatisés pour extraire et extrapoler des connaissances à partir de grandes quantités de données brutes.



## • Analyse de données

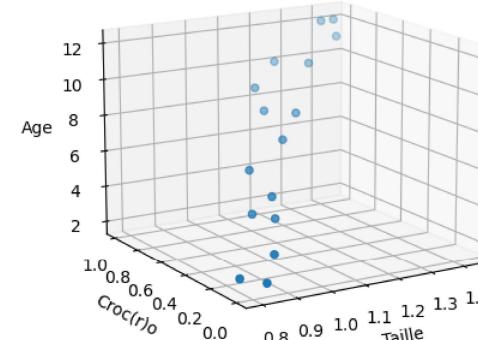
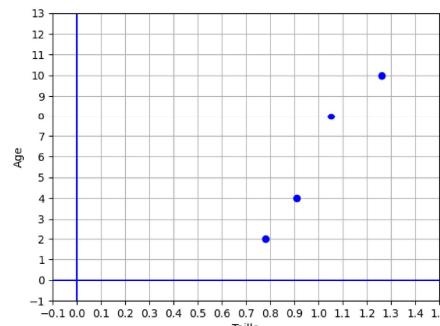
Dans la partie Statistique de la SAé, on sera plus particulièrement dans le domaine de l'**analyse de données**.

## Définition

L'**analyse des données** est une famille de méthodes statistiques dont les principales caractéristiques sont d'être multidimensionnelles et descriptives.

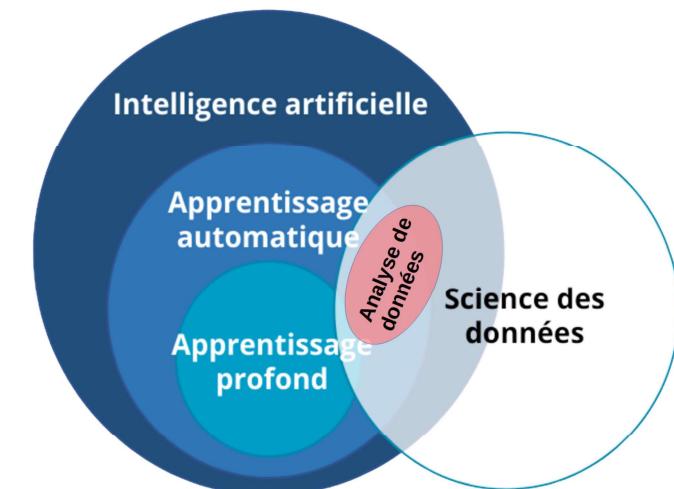
"multidimensionnelles" : dans le Cours 3 on a vu des représentations graphiques de 2 variables statistiques... donc 2 dimensions.

Ici, on va étudier des cas à plus de 2 variables !



## • Méthodes d'analyse de données

Il y a 3 principales catégories de méthodes d'analyse de données, qui se trouvent être aussi 3 des 4 catégories principales d'Apprentissage automatique (Machine Learning en anglais).



Ainsi, on va retrouver ces 3 grandes catégories de méthodes dans la bibliothèque Python de Machine Learning Scikit-learn

Scikit-learn homepage featuring:  
 - Logo and navigation links: Install, User Guide, API, Examples, Community, More.  
 - Main title: scikit-learn Machine Learning in Python  
 - Buttons: Getting Started, Release Highlights for 1.4, GitHub  
 - Description: Simple and efficient tools for predictive data analysis, Accessible to everybody, and reusable in various contexts, Built on NumPy, SciPy, and matplotlib, Open source, commercially usable - BSD license  
 - Sections:  
 - Dimensionality reduction: Visualizations of PCA and NMF.  
 - Regression: Visualizations of linear regression, decision trees, and boosted decision trees.  
 - Clustering: Visualizations of k-means clustering and hierarchical clustering.

### • Réduction de dimension

### • Régression

### • Clustering/Classification



## 2. Problème commun aux 3 méthodes

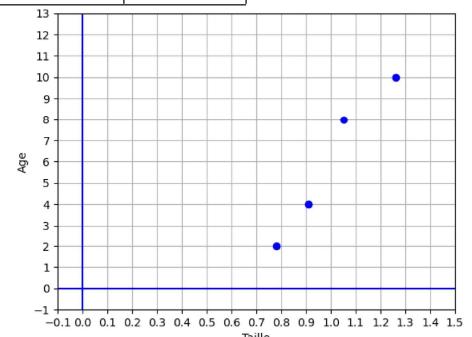
On travaille avec **plusieurs variables statistiques sur la même population** :

- 2 variables comme dans le Cours 3 sur les corrélations
- ... ou plus de 2.

On ne pourra donc pas forcément représenter le nuage de points !

**Exemple 1 : Âge et taille des enfants** (avec 2 variables statistiques)

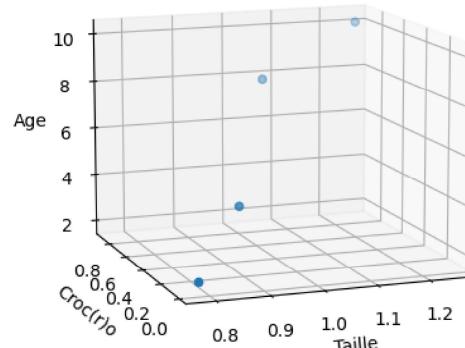
	Enfant A	Enfant B	Enfant C	Enfant D	Enfant E
Taille	0.84	0.98	1.14	1.32	1.44
Âge	2	4	8	10	12



**Exemple 1' : Âge et taille et prononciation des enfants** (3 variables statistiques)

On ajoute une donnée pour chaque enfant : la qualité de sa prononciation du mot croc(r)odile :

	Enfant A	Enfant B	Enfant C	Enfant D
Taille	0.78	0.91	1.05	1.26
Croc(r)o	0	0.3	0.8	0.9
Âge	2	4	8	10

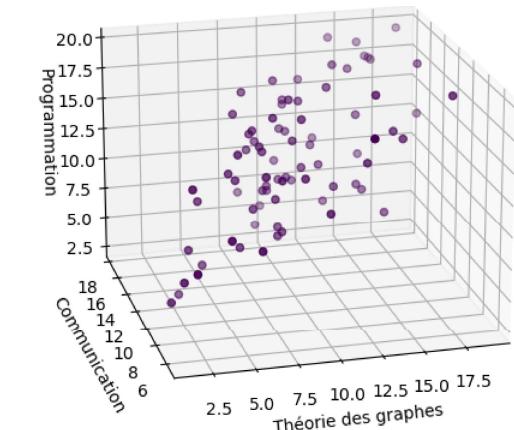
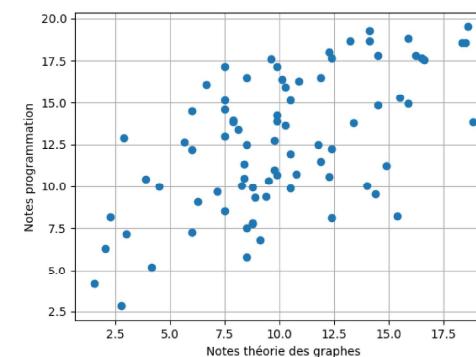


On a maintenant **3 variables statistiques sur la même population** :

- X : taille de l'enfant
- Y : prononciation de croc(r)odile
- Z : âge de l'enfant
- Population : enfants A,B,C,D et E

**Exemple 2 : Notes S2 2017** (avec 2 ou 3 variables statistiques)

- Population : étudiant.es de 1e année de 2017
- X : notes en programmation orientée objets
- Y : notes en théorie des graphes
- Z : notes en communication



Si on ajoute les notes de Statistique, ou d'Anglais... on ne peut plus faire de représentation graphique !

**Exemple 3**

Année (population)	nb abonnement internet haut débit	part des connexions internet supportant le 4k (en%)	Part (en%) d'individus ayant utilisé internet dans les 3 mois	Part (en%) d'individu jouant ou téléchargeant sur internet	revenus brut de la distribution de films	vente de ticket de cinéma
2003	23766	0	5	1	407980000	173500000
2004	23967	0	9	3	471200000	195800000
2005	23568	0.1	12	6	420410000	175600000
2006	23469	0.3	20	9	454500000	188800000
2007	23970	0.4	34	21	427500000	178500000
2008	23971	0.6	33	20	469530000	190300000
2009	23972	0.9	33	23	496410000	201600000
2010	23973	0.9	37	28	520580000	207100000
2011	23974	1.1	49	30	550590000	217200000
2012	23975	1.6	41	33	517120000	203600000
2013	24943	3.9	44	34	500300000	193700000
2014	25970	5.5	47	39	550410000	209100000
2015	26871	10.8	51	42	551970000	205400000
2016	27256	16.1	54	43	581760000	213100000
2017	27824	18.4	60	46	604650000	220500000

Population : .....

Nombre de séries statistiques : .....

Descripteurs : .....



### 3. Préparation commune des données : centrer-réduire

La première étape de préparation de ce type de données sera de **centrer-réduire** toutes les données, afin d'éviter que des échelles différentes ne donnent plus de poids à certaines variables qu'à d'autres.

#### Centrer-réduire

Pour chaque variable statistique  $X$ , on calcule la moyenne  $\bar{X}$  et l'écart-type  $\sigma(X)$  de la variable sur notre population, et pour chaque valeur  $X(i)$  :

- **on centre** en soustrayant la moyenne de la variable :  $X(i) - \bar{X}$ ,
- **on réduit** en divisant le résultat par l'écart-type :  $\frac{X(i) - \bar{X}}{\sigma(X)}$ .

9

#### Exemple : centrer-réduire 1 variable statistique

On prend une série de notes de 10 étudiant.es :

16, 11, 14, 10, 17, 12, 15, 9, 12, 14.

La moyenne sur ces 10 étudiant.es est 13, l'écart-type vaut 2.49.

- **on centre** : on soustraie la moyenne (13) à la note de chaque étudiant.e :

3, -2, 1, -3, 4, -1, 2, -4, -1, 1.

Par exemple pour le 1<sup>e</sup> étudiant.e, on a remplacé 16 par  $16 - 13 = 3$ .

- **on réduit** : on divise ce qu'on a obtenu par l'écart-type 2.49 :

1.2, -0.8, 0.4, -1.2, 1.6, -0.4, 0.8, -1.6, -0.4, 0.4.

(résultats arrondis à 1 décimale)

Par exemple pour le 1<sup>e</sup> étudiant.e, on a remplacé sa note centrée 3 par  $3 \div 2.49 \approx 1.2$ .

#### Exemple des âges, tailles et prononciation des enfants

	Enfant A	Enfant B	Enfant C	Enfant D
Taille	0.78	0.91	1.05	1.26
Croc(r)o	0	0.3	0.8	0.9
Âge	2	4	8	10

On calcule séparément la moyenne et l'écart-type pour chaque variable statistique :

- la taille X :  $\bar{X} = 1$ ,  $\sigma(X) \approx 0.18$
- la prononciation Y :  $\bar{Y} = \dots$ ,  $\sigma(Y) \approx 0.37$
- l'âge Z :  $\bar{Z} = \dots$ ,  $\sigma(Z) \approx 3.16$

On centre-réduit chaque variable/ligne séparément :

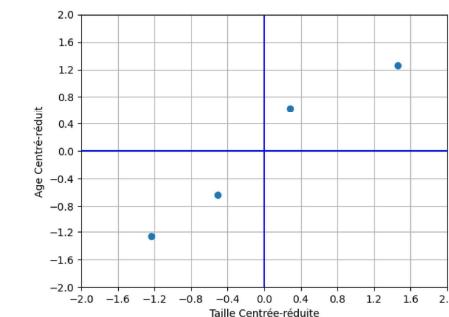
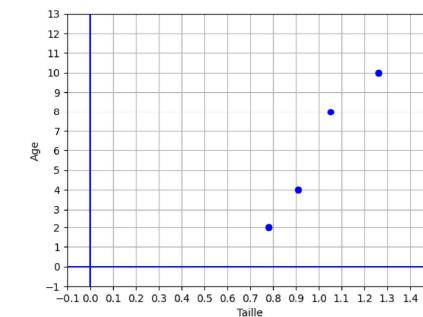
	Enfant A	Enfant B	Enfant C	Enfant D
Taille CR	$\frac{0.78-1}{0.18} \approx -1.22$	$\frac{0.91-1}{0.18} \approx -0.5$	$\frac{1.05-1}{0.18} \approx 0.28$	$\frac{1.26-1}{0.18} \approx 1.44$
Croc(r)o CR	$\frac{0-\dots}{0.37} \approx \dots$			
Âge CR				

10

11

#### Centrer-réduire vs nuage de points

Centrer-réduire ne change pas la forme du nuage de points.



#### Utilité de centrer-réduire

Centrer-réduire est utile surtout lorsqu'on a des données avec des échelles très différentes, comme dans l'exemple 3 ci-dessus par exemple.

Année (population)	nb abonnement internet haut débit	part des connexions internet supportant le 4k (en%)	Part (en%) d'individus ayant utilisé internet dans les 3 mois	Part (en%) d'individu jouant ou téléchargeant sur internet	revenus brut de la distribution de films	vente de ticket de cinéma
2003	23766	0	5	1	407980000	173500000
2004	23967	0	9	3	471200000	195600000
2005	23568	0.1	12	6	420410000	175600000
2006	23469	0.3	20	9	454500000	188800000

12



## 4. Une méthode de régression : la régression linéaire multiple

[Install](#) [User Guide](#) [API](#) [Examples](#) [Community](#) [More](#)  [Go](#)

**scikit-learn** Machine Learning in Python

[Getting Started](#) [Release Highlights for 1.4](#) [GitHub](#)

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

**Dimensionality reduction**  
Reducing the number of random variables to consider.

**Applications:** Visualization, Increased efficiency  
**Algorithms:** PCA, feature selection, non-negative matrix factorization, and more...

**Regression**  
Predicting a continuous-valued attribute associated with an object.

**Applications:** Drug response, Stock prices.  
**Algorithms:** Gradient boosting, nearest neighbors, random forest, ridge, and more...

**Clustering**  
Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experimental outcomes  
**Algorithms:** k-Means, HDBSCAN, hierarchical clustering, and more...

[Install](#) [User Guide](#) [API](#) [Examples](#) [Community](#) [More](#)  [Go](#)

13

### Régression linéaire "simple" : avec 2 variables statistiques

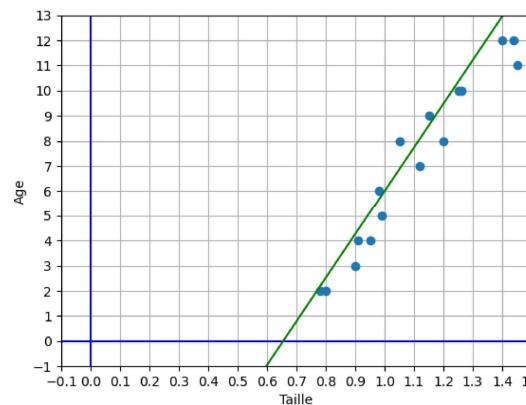
	Enfant A	Enfant B	Enfant C	Enfant D
Taille	0.78	0.91	1.05	1.26
Âge	2	4	8	10

La droite ci-contre a une équation de la forme  $y = ax + b$ , où  $a$  et  $b$  sont donnés par les formules du Cours 3 :

$$y \approx 17x - 11.$$

Ainsi, on peut trouver une approximation de l'âge d'un enfant à partir de sa taille :

$$\text{Age} \approx 17 \times \text{Taille} - 11$$



14

### Régression linéaire avec 3 variables statistiques

Pour avoir plus de précision sur l'âge lorsqu'il y a plus d'enfants, on peut souhaiter rajouter une variable statistique. Par exemple, la qualité de la prononciation du mot Croc(r)odile, sur une échelle allant de 0 et 1 :

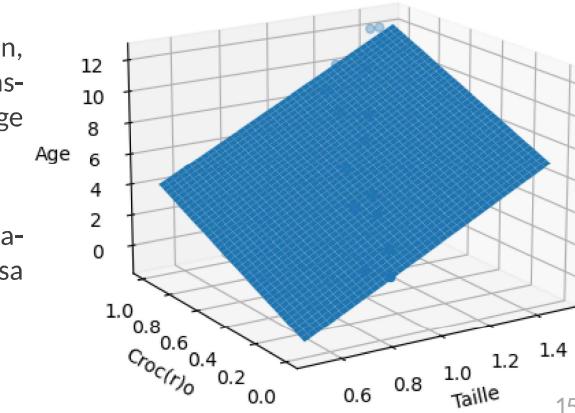
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Taille	0.84	0.98	1.14	1.32	1.44	0.8	0.95	1.2	1.25	1.4	0.9	0.99	0.98	1.12
Crocro	0	0.3	0.8	0.9	1	0.1	0.4	0.7	1	1	0.1	0.35	0.5	0.6
Âge	2	4	8	10	12	2	4	8	10	12	3	5	6	7

Dans un tel cas on peut trouver un plan, d'équation  $y = a_1x_1 + a_2x_2 + b$  passant de manière optimale par le nuage de points.

$$\text{Ici on trouvera : } y \approx 7x_1 + 5x_2 - 4.$$

Ainsi, on peut trouver une approximation de l'âge d'un enfant à partir de sa taille et de sa prononciation :

$$\text{Age} \approx 7 \times \text{Taille} + 5 \times \text{Crocro} - 4$$



15

### Régression linéaire avec 4 séries statistiques... ou plus ?

Pour avoir encore plus de précision sur l'âge des enfants, on peut encore rajouter des données. Par exemple, la capacité à se moucher (ou s'habiller) seul.e sur une échelle allant de 0 et 1 :

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Taille	0.84	0.98	1.14	1.32	1.44	0.8	0.95	1.2	1.25	1.4	0.9	0.99	0.98	1.12
Crocro	0	0.3	0.8	0.9	1	0.1	0.4	0.7	1	1	0.1	0.35	0.5	0.6
Moucher	0	0.2	0.9	0.8	1	0.1	0.3	0.8	0.9	1	0.1	0.4	0.4	0.5
Âge	2	4	8	10	12	2	4	8	10	12	3	5	6	7

On ne peut alors plus faire de représentation graphique...

Mais on peut toujours trouver une relation linéaire !

Dans un tel cas on peut trouver une équation  $y = a_1x_1 + a_2x_2 + a_3x_3 + b$  représentant de manière optimale la relation entre l'âge des enfants  $y$ , et les 3 autres variables  $x_1, x_2$  et  $x_3$  (la taille, la prononciation de Croc(r)odile, et le mouchage).

$$\text{Ici on trouvera : } y \approx 7x_1 + 3x_2 + x_3 - 4.$$

Ainsi, on peut trouver l'âge approximatif des enfants par la formule :

$$\text{Age} \approx 7 \times \text{Taille} + 3 \times \text{Crocro} + 1 \times \text{Moucher} - 4$$

16

## • Régression linéaire multiple

- avec 2 variables  $X$  et  $Y$ , équation de droite  $y = ax + b$
- avec 3 variables  $X_1, X_2$  et  $Y$ , équation de plan  $y = a_1x_1 + a_2x_2 + b$
- avec 4 variables  $X_1, X_2, X_3$  et  $Y$ , équation  $y = a_1x_1 + a_2x_2 + a_3x_3 + b$
- ...

### Définition

On suppose qu'on a  $n + 1$  variables statistiques :  $X_1, X_2, \dots, X_n$  et  $Y$  définies sur une même population.

Faire la **régression linéaire multiple** de  $Y$  en fonction de  $X_1, X_2, \dots, X_n$  consiste à trouver les coefficients  $a_1, a_2, \dots, a_n$  et  $b$  tels que l'équation

$$y = a_1x_1 + a_2x_2 + \dots + a_nx_n + b$$

soit la meilleure approximation possible de la relation entre  $Y$  et les variables  $X_1, X_2, \dots, X_n$ .

17

### Définitions

Les variables statistiques  $X_1, X_2, \dots, X_n$  sont appelées les **variables explicatives**.

La variable statistique  $Y$  est appelée la **variable endogène**.

Les coefficients  $a_1, a_2, \dots, a_n$  et  $b$  de l'équation trouvée sont appelés les **paramètres**.

Exemple des âges, tailles et prononciations des enfants :

- on a bien des séries statistiques définies sur une même population :  
.....
- la variable endogène est .....
- les variables explicatives sont .....
  
.....

18

## • Exemple de régression linéaire multiple

Gwendal Houssaye, Jean-Baptiste Le Chanu, Théo Leveque  
(projet 2018-2019)

Est-ce que l'évolution des dégâts causés par les sangliers à une raison ?



### 1) Présentation des données

- 1.1) Nombre de sanglier prélevé en France par année
- 1.2) Indemnisation en euro des dégâts causés par les sangliers en France
- 1.3) Consommation de viande de porc dans l'OCDE
- 1.4) Nombre de chasseur en France

19

On remarque qu'il y a une forte évolution pour la valeur «  $a_1$  » ( $\sim 1,0$ ). On en déduit que plus il y a de sangliers prélevés par les chasseurs plus les dégâts causés par des sangliers augmentent car ils se vengent en détruisant les cultures.

On remarque cependant qu'il y a une évolution plutôt moyenne des valeurs «  $a_2$  » ( $\sim -0,49$ ) et «  $a_3$  » ( $\sim 0,44$ ). On peut en déduire pour la valeur «  $a_2$  » que plus il y a de viande de porc consommée plus les sangliers sont calmes car les cochons et les sangliers ne font pas partie de la même famille.

Pour la valeur «  $a_3$  » on en déduit que moins il y a de chasseurs plus les dégâts des sangliers augmentent car les sangliers ont moins peur des chasseurs.

→ On étudiera cet exemple en TD sur Python !

20



## 5. Une méthode de clustering : l'algorithme des k-moyennes

Install User Guide API Examples Community More ▾

scikit-learn Machine Learning in Python

Getting Started Release Highlights for 1.4 GitHub

Simple and efficient tools for predictive data analysis  
 • Accessible to everybody, and reusable in various contexts  
 • Built on NumPy, SciPy, and matplotlib  
 • Open source, commercially usable - BSD license

**Dimensionality reduction**  
 Reducing the number of random variables to consider.  
**Applications:** Visualization, Increased efficiency  
**Algorithms:** PCA, feature selection, non-negative matrix factorization, and more...

**Regression**  
 Predicting a continuous-valued attribute associated with an object.  
**Applications:** Drug response, Stock prices.  
**Algorithms:** Gradient boosting, nearest neighbors, random forest, ridge, and more...

**Clustering**  
 Automatic grouping of similar objects into sets.  
**Applications:** Customer segmentation, Grouping experiment outcomes  
**Algorithms:** k-Means, HDBSCAN, hierarchical clustering, and more...

**But des méthodes de Clustering**

→ Regrouper les individus d'une population par groupes se ressemblant d'après les variables statistiques.

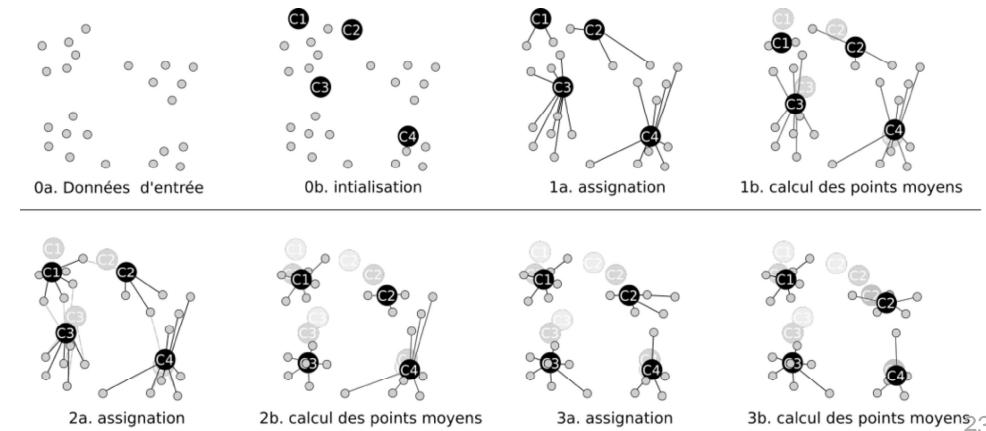
21

A partir de 4 variables statistiques, on ne peut plus faire de représentation graphique...

C'est bien pour ça qu'on va avoir besoin d'un algorithme ! Car sinon, l'œil aurait été aussi efficace.

On pourra alors interpréter a posteriori la répartition en groupes obtenue grâce aux moyennes des clusters dans chaque variable, comme pour l'exemple des notes du S2 2017.

Le principe de l'algorithme des k-moyennes est résumé ci-dessous :



22

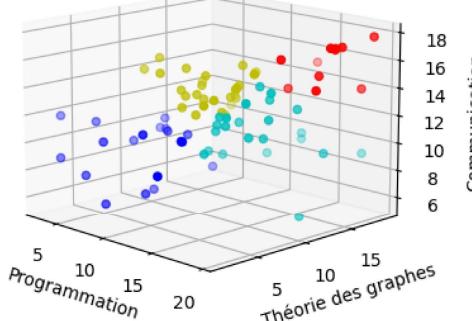
**Exemple : Notes S2 2017 (avec 3 variables statistiques)**

Population : étudiant.es de 1e année 2017

X : notes programmation orientée objets

Y : notes en théorie des graphes

Z : notes en communication



Ici par exemple, on a défini 4 clusters :

- le cluster rouge, où la moyenne en Prog est de 17.6, celle en Graphes de 15.8 et celle en Com de 16.
- le cluster bleu, où la moyenne en Prog est de 8.5, celle en Graphes de 5.4 et celle en Com de 9.9.
- le cluster jaune, où la moyenne en Prog est de 10.7, celle en Graphes de 10.5 et celle en Com de 13.8.
- le cluster turquoise, où la moyenne en Prog est de 15.7, celle en Graphes de 10.4 et celle en Com de 11.5.

Et on peut obtenir la liste des étudiant.es appartenant à chaque cluster.

22

24



## 6. Une méthode de réduction de dimension : l'Analyse en Composantes Principales

Tableau de notes :

Index	MATH	PHYS	FRAN	ANGL
jean	6	6	5	5.5
alan	8	8	8	8
anni	6	7	11	9.5
moni	14.5	14.5	15.5	15
didi	14	14	12	12.5
andr	11	10	5.5	7
pier	5.5	7	14	11.5
brig	13	12.5	8.5	9.5
evel	9	9.5	12.5	12

Tableau de notes Centré-Réduit:

Index	MATH	PHYS	FRAN	ANGL
jean	-1.0865	-1.28174	-1.50366	-1.6194
alan	-0.493865	-0.613006	-0.639857	-0.730706
anni	-1.0865	-0.947373	0.22395	-0.197488
moni	1.43221	1.56038	1.51966	1.75764
didi	1.28405	1.3932	0.511885	0.868948
andr	0.395092	0.0557278	-1.3597	-1.08618
pier	-1.23466	-0.947373	1.08776	0.513469
brig	0.98773	0.891645	-0.495889	-0.197488
evel	-0.197546	-0.111456	0.655853	0.691208

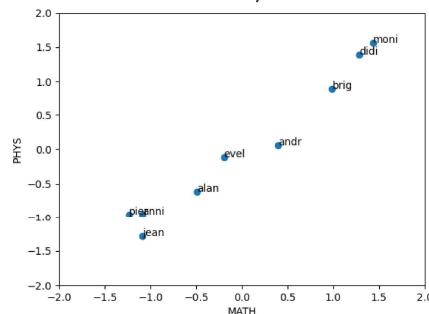


On a 4 notes pour chaque élève, donc si on voulait représenter chaque élève par un point, il faudrait pouvoir dessiner en dimension 4, ce qui est impossible...

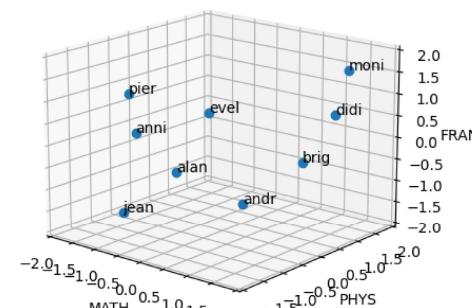
25

On peut représenter partiellement les notes des élèves :

Représentation de 2 notes (PHYS et MATH)



Représentation de 3 notes (PHYS, MATH et FRAN)

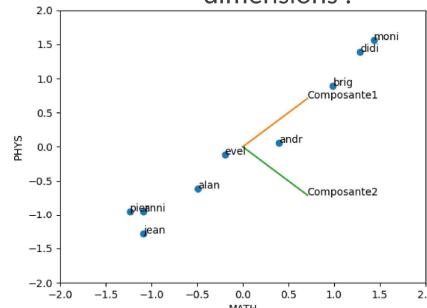


Mais ces représentations ne prennent pas en compte les autres notes...

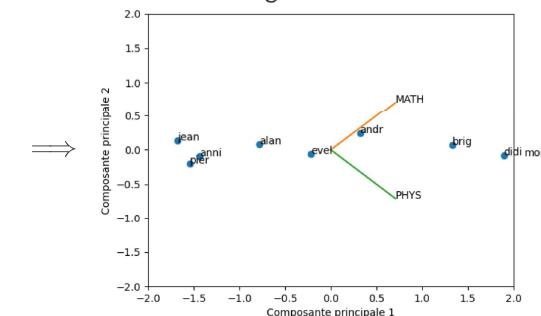
L'Analyse en Composantes Principales va permettre d'obtenir une représentation avec moins de dimensions, mais tenant compte de toutes les données.

### • Premier exemple : passer de 2 dimensions à 1 dimension

On part de cette représentation en 2 dimensions :



On fait un changement de coordonnées



On ne garde que la 1<sup>re</sup> coordonnée (Composante principale 1) :

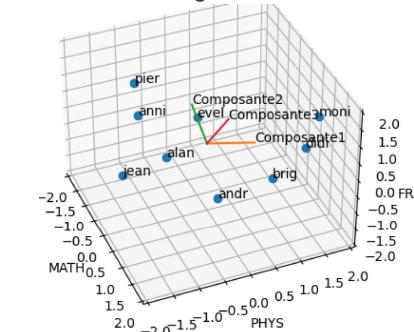


Cette représentation 1D tient bien compte des 2 notes : Pier et Anni ont la même note de PHYS, Jean et Anni ont la même note de MATH, mais en 1D ils ont 3 points distincts.

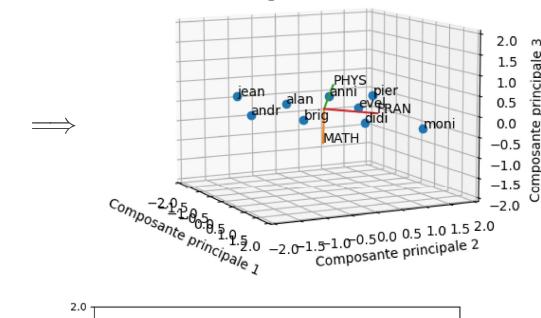
27

### • Deuxième exemple : passer de 3 dimensions à 2 dimensions

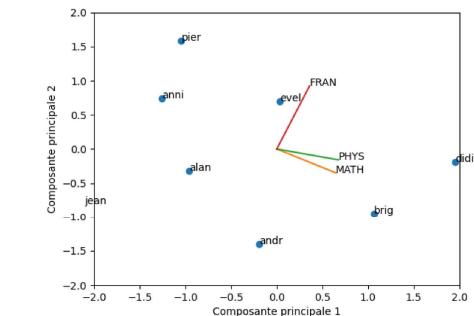
On part de cette représentation en 3 dimensions :



On fait un changement de coordonnées



On ne garde que les deux 1<sup>re</sup> coordonnées (Composantes principales 1 et 2) :



26

28

## • Deuxième exemple : de 3 à 2 dimensions sur les tableaux

On part des 3 premières colonnes du tableau de notes :

Index	MATH	PHYS	FRAN
jean	-1.0865	-1.28174	-1.50366
alan	-0.493865	-0.613006	-0.639857
anni	-1.0865	-0.947373	0.22395
moni	1.43221	1.56038	1.51966
didi	1.28405	1.3932	0.511885
andr	0.395092	0.0557278	-1.3597
pier	-1.23466	-0.947373	1.08776
brig	0.98773	0.891645	-0.495889
evel	-0.197546	-0.111456	0.655853

On fait un changement de coordonnées



	0	1	2
0	-2.1053	-0.807411	0.0127786
1	-0.961776	-0.321874	-0.0218886
2	-1.25951	0.735617	0.025543
3	2.52202	0.657582	-0.0488898
4	1.95179	-0.195353	0.0644815
5	-0.195679	-1.40261	-0.0485795
6	-1.04495	1.58526	0.0120363
7	1.06054	-0.943617	0.0373585
8	0.0328752	0.692409	-0.03284



	0	1
0	-2.1053	-0.807411
1	-0.961776	-0.321874
2	-1.25951	0.735617
3	2.52202	0.657582
4	1.95179	-0.195353
5	-0.195679	-1.40261
6	-1.04495	1.58526
7	1.06054	-0.943617
8	0.0328752	0.692409

On ne garde que les deux

1e colonnes

(Composantes principales 1 et 2) :

29

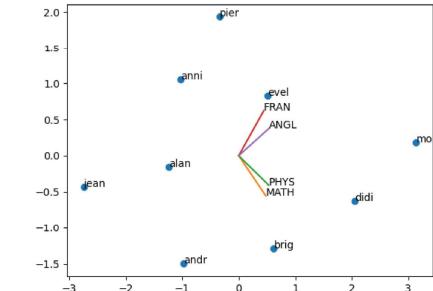
## • Troisième exemple : passer de 4 dimensions à 2 dimensions

On part du tableau complet à 4 dimensions/colonnes :

Index	MATH	PHYS	FRAN	ANGL
jean	-1.0865	-1.28174	-1.50366	-1.6194
alan	-0.493865	-0.613006	-0.639857	-0.730706
anni	-1.0865	-0.947373	0.22395	-0.197488
moni	1.43221	1.56038	1.51966	1.75764
didi	1.28405	1.3932	0.511885	0.868948
andr	0.395092	0.0557278	-1.3597	-1.08618
pier	-1.23466	-0.947373	1.08776	0.513469
brig	0.98773	0.891645	-0.495889	-0.197488
evel	-0.197546	-0.111456	0.655853	0.691208



...On fait un changement de coordonnées...



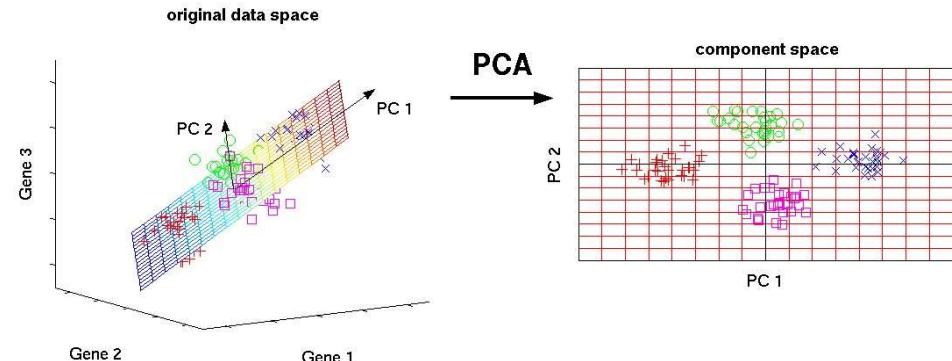
On ne garde que les deux

1e coordonnées/colonnes

(Composantes principales 1 et 2) :

30

**Intérêt / objectif :** Cette méthode permet d'avoir des informations du même type que celles de la Méthode des k-moyennes et de la Régression linéaire multiple réunies !

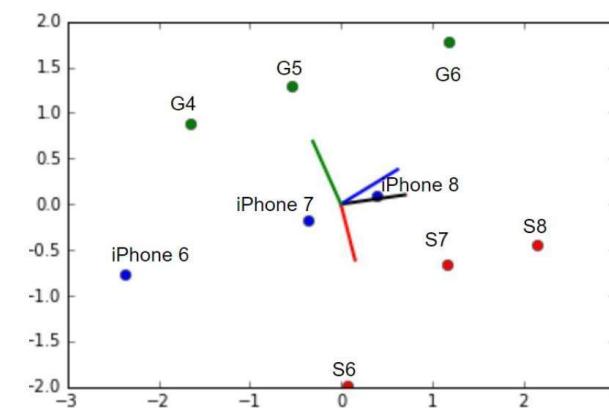


- la représentation finale permet de faire à l'oeil nu des regroupements du type de ceux faits par la Méthode des k-moyennes.
- le plan sur lequel on projette donne des informations du même type que les hyperplans de la Régression linéaire multiple.

31

## • Exemple

Glenn Anjubault, Tom Mallet (projet 2017-2018)



Légende :

-Points :

Bleu : Apple

Rouge : Samsung

Vert : LG

-Axes :

Bleu : Année de sortie

Rouge : Chiffre d'affaire de l'entreprise à l'année de sortie

Vert : Durée de vie

Noir : Autonomie de la batterie

32