Réduction de dimension des données

Analyse en composantes principale

Manel sekma
sekma.manel@gmail.com

Note de cours Machine Learning avec Python
de Dr Manel SEKMA

Références

1. Hotelling, Harold. 1933. "Analysis of a Complex of Statistical Variables into Principal Components." *Journal of Educational Psychology* 24 (6): 417.
2. James, G., D. Witten, T. Hastie, and R. Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. Springer New York. https://books.google.ca/books?id=at1bmAEACAAJ.
3. Kuhn, M., and K. Johnson. 2013. *Applied Predictive Modeling*. SpringerLink : Bücher. Springer New York. https://books.google.ca/books?id=xYRDAAAAQBAJ.
4. Lê, Sébastien, Julie Josse, and François Husson. 2008. "FactoMineR: A Package for Multivariate Analysis." *Journal of Statistical Software* 25 (1): 1–18. https://doi.org/10.18637/jss.v025.i01.
5. Pearson, Karl. 1901. "LIII. On Lines and Planes of Closest Fit to Systems of Points in Space." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2 (11): 559–72.

Plan

(Apprentissage non supervisé, PCA)

1. Introduction
2. Sélection de variables ou extraction de variables
3. Analyse en composantes principale

Introduction

- Dans certaines applications , le nombre de variable utilisé pour représenter les données est très élevé.
  - Cas de traitement d'images haute-résolution (un pixel est représenter par ++variables)
  - Analyse de données génomiques (centaines de milliers de positions du génome peuvent être caractérisées)
  - ……
- Bien qu'une représentation des données contenant plus de variables soit intuitivement plus riche, il est plus difficile d'apprendre un modèle performant dans ces circonstances !

→Réduction de dimension des données

Introduction

Cas d'utilisation

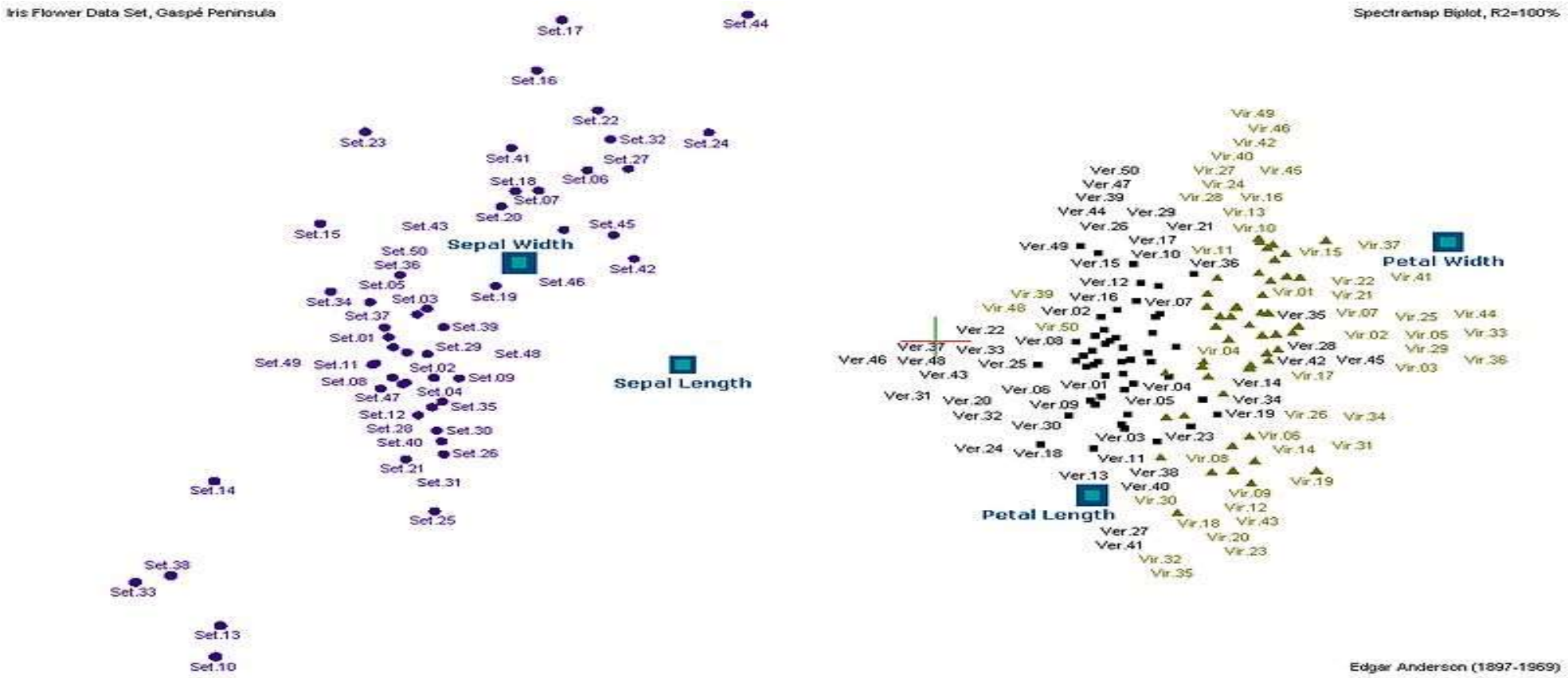
- Réduire les coûts algorithmiques: Dataset avec beaucoup de variables
  - réduire la dimension =>réduire l'espace en mémoire & temps de calcul
  - Si certaines variables sont inutiles, ou redondantes
  - …..

Customer (cust)	Job Configuration (conf)	First Job Year (year)	BHT (°F)	BHP (psi)	Treatment Pressure (psi)	Slurry Rate (bbl/min)	Acid Rate (bbl/min)	Proppant Concentration (lbm/gal)	Average JPT (hr)
1	Gel frac	2010	180	5,400	4,060.12	61.10	15.00	2.10	0.04
2	Gel frac	2012	175	4,500	4,824.50	96.66	10.15	2.07	0.23
3	N_2 foamed acid	2005	—	0	4,127.48	6.49	—	0.00	0.18
4	Gel frac	2011	315	13,046	10,333.36	61.94	15.93	3.51	0.14
5	Gel frac	2009	290	10,332	4,815.90	8.72	0.00	0.00	0.63
1	Gel frac	2010	290	10,500	8,536.23	55.43	9.41	2.02	1.38
3	Gel frac	2012	175	4,500	6,410.02	97.53	15.58	2.01	0.47
1	Gel frac	2007	160	5,373	5,438.87	51.72	20.00	1.61	0.29
1	Gel frac	2005	—	0	3,446.60	54.39	—	1.14	0.44
4	Gel frac	2010	180	5,175	2,572.72	46.14	8.00	2.63	0.21
5	Gel frac	2011	326	14,500	10,539.09	53.34	18.66	4.27	0.91
1	Gel frac	2007	140	3,262	2,444.42	45.27	20.00	1.29	2.74
2	Gel frac	2011	326	12,429	10,658.36	54.58	10.02	4.25	1.28
2	Gel frac	2011	334	13,713	10,765.09	99.92	15.51	4.26	1.01

Introduction

Cas d'utilisation

- **Visualiser** des données avec beaucoup de variables



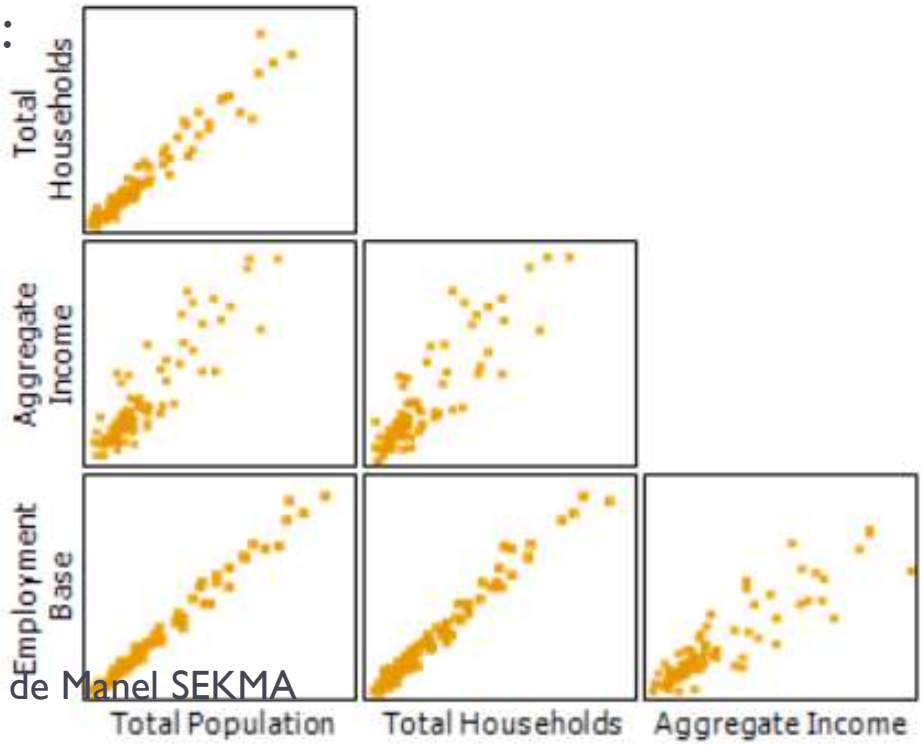
Cartographie spectrale des fleurs Iris de Fisher qui ont donné lieu à de nombreuses études en analyse des données.

Introduction

Cas d'utilisation

- Lorsque plusieurs variables sont très corrélées: multi colinéarité
  - l'existence de corrélations élevées entre les variables indépendantes (variables explicatives).
  - La multicolinéarité a pour conséquences :

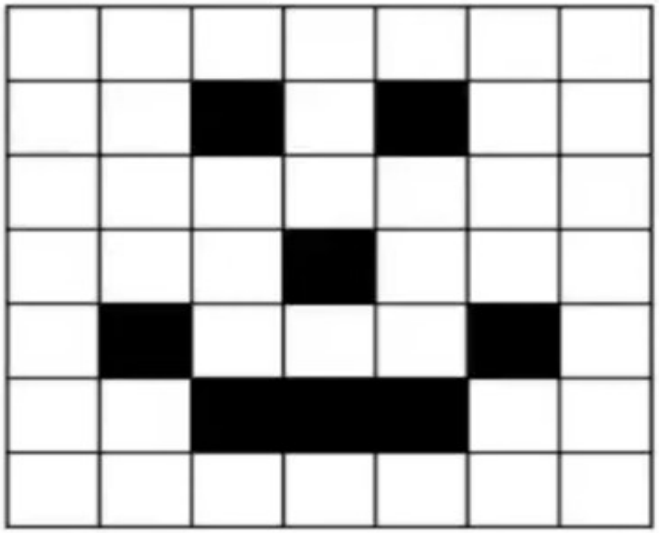
    - de fausser la précision de l'estimation
    - des coefficients de régression
    - de rendre sensible l'estimation
    - des coefficients à de petites
    - variations des données.



Introduction

Cas d'utilisation

- Débruitage et compression d'images
  - Image haute dimension ou vidéo



0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	0	0	0	0	0
0	0	0	1	0	0	0
0	1	0	0	0	1	0
0	0	1	1	1	0	0
0	0	0	0	0	0	0

Introduction

Représentation de données

- Le but de la réduction de dimension est de transformer une représentation $\mathbf{X} \in \mathbb{R}^{n \times p}$ des données

en une représentation :

$$\boldsymbol{X}^* \in \mathbb{R}^{n \times m} \textbf{ où } \boldsymbol{m \ll p}$$

Sélection de variables ou extraction de variables

- Deux possibilités pour réduire la dimension de nos données:
  - Sélection de variables: consiste à éliminer un nombre p-m de variables de nos données. Les variables sélectionnées gardent ainsi leur signification initiale, ce qui contribue à la lisibilité des modèles construits ultérieurement.
  - Extraction de variables: qui consiste à créer m nouvelles variables à partir des p variables initiales.

Sélection de variables ou extraction de variables

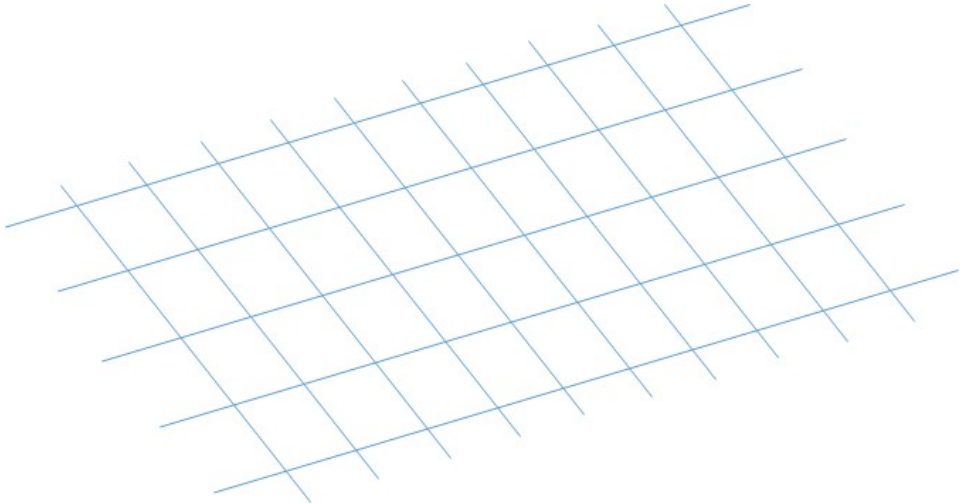
- Sélection de variables (*feature selection*)
  - Les méthodes de filtrage: basées sur des critères (par ex. minimisation de la redondance entre variables, maximisation de l'information mutuelle avec la classe à prédire) qui ne tiennent pas compte des résultats du modèle décisionnel ultérieur.
  - Les méthodes de conteneur (wrapper methods): basées sur des mesures des performances du modèle décisionnel qui emploie les variables sélectionnées.
  - Les méthodes embarquées/intégrées : l'opération de sélection est indissociable de la méthode de modélisation décisionnelle. (utilisé dans des réseaux de neurones)

Sélection de variables ou extraction de variables

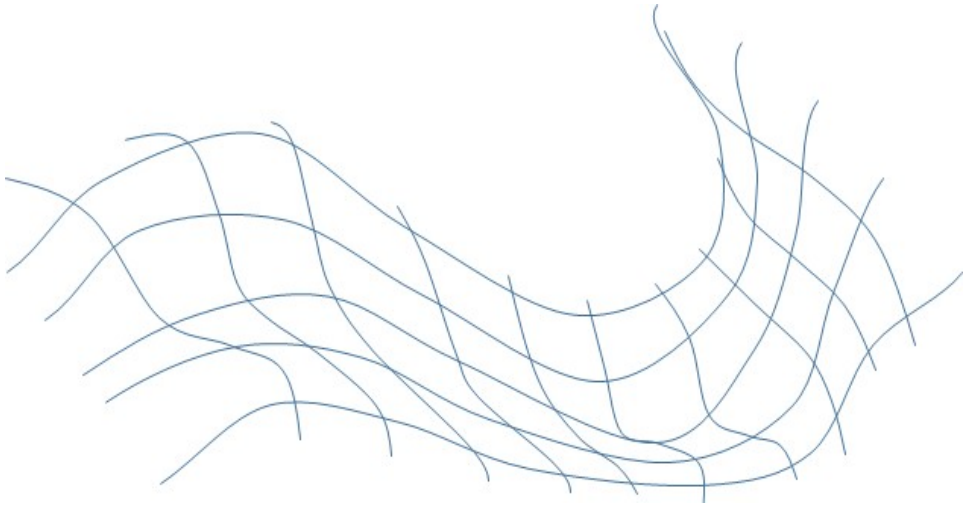
Extraction de variables (*feature extraction*)

Les nouvelles variables sont obtenues par des méthodes qui peuvent être

1. Linéaires : trouver un sous-espace linéaire de dimension k dans l'espace initial $\mathbb{R}^m$.
2. Non linéaires : trouver un sous-espace non linéaire de dimension k dans l'espace initial.



Sous-espace bidimensionnel linéaire dans l'espace tridimensionnel



Sous-espace bidimensionnel non linéaire dans l'espace tridimensionnel

Sélection de variables ou extraction de variables

Extraction de variables (*feature extraction*)

Méthodes factorielles linéaires:

- L'analyse en composantes principales (ACP), méthode à caractère exploratoire, adaptée à des données décrites par des variables quantitatives.
- L'analyse factorielle discriminante (AFD), méthode à caractère exploratoire et décisionnel, adaptée à des données décrites par des variables quantitatives et appartenant à plusieurs classes.
- L'analyse des correspondances multiples (ACM), méthode à caractère exploratoire, adaptée à des données décrites par des variables nominales.

Analyse en composantes principale

- Définition du PCA
- Variantes de PCA
- Comprendre l'intuition du PCA
- Mathématique de l'ACP

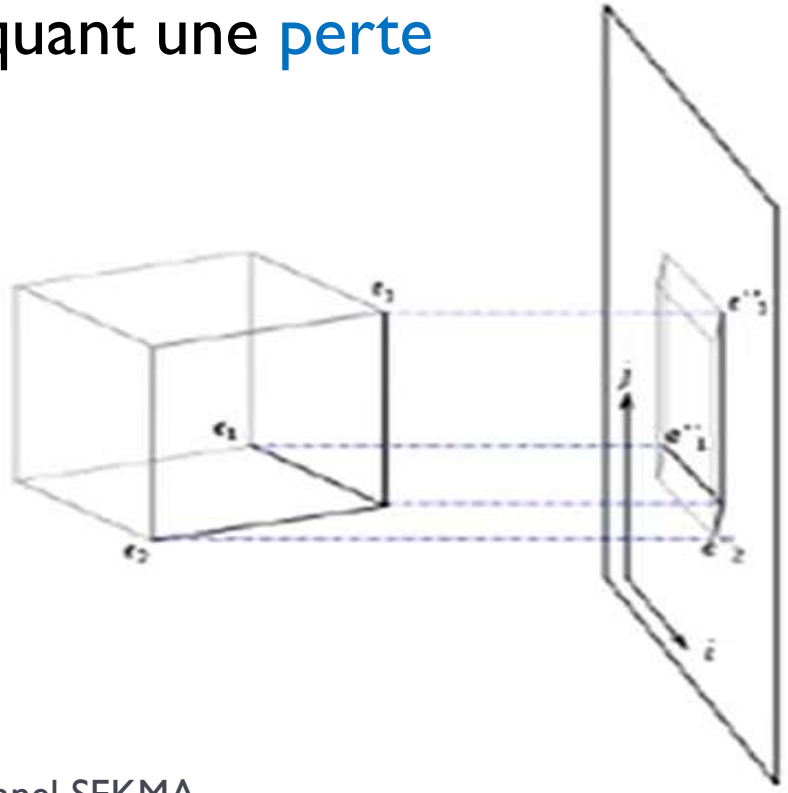
Définition

- Le PCA est un algorithme de réduction dimensionnelle non-supervisé capable d'identifier les corrélations et pattern dans un jeu de donnée et de le transformer en un ensemble de donnée avec un nombre réduit de variable en minimisant la perte d'information.

- Le PCA permet de mettre aussi en évidence la variabilité entre les différentes données qui composent la dataset, mais aussi la liaison entre les variables.

Comment ?

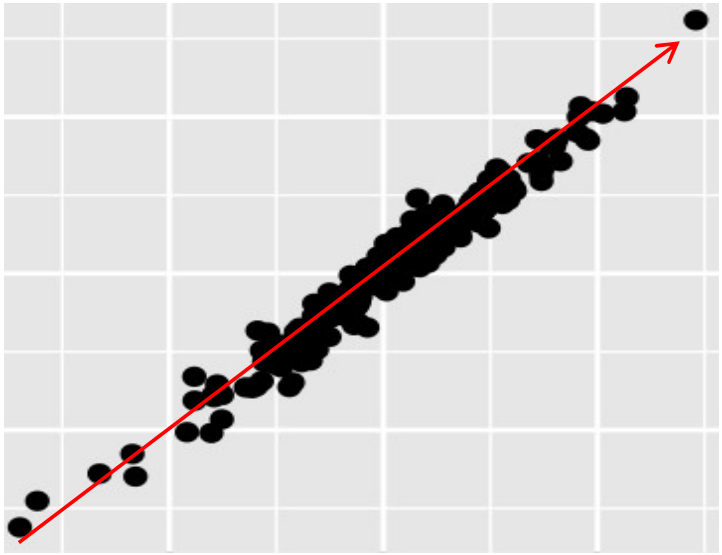
- Tout cela est réalisé par la projection de la Dataset initiale dans un espace réduit en utilisant les vecteurs propres.
- La projection, c'est la fonction qui permet de représenter des points dans un espace plus petit impliquant une perte d'information.



Comment ?

- Pour minimiser la perte :
  - Maximiser la variance de nos projections afin de pouvoir continuer à distinguer les exemples les uns des autres dans leur nouvelle représentation
  - Minimiser la distance entre nos données et nos projections

Maximisation de la variance: La variance des données est maximale selon l'axe indiqué par une flèche

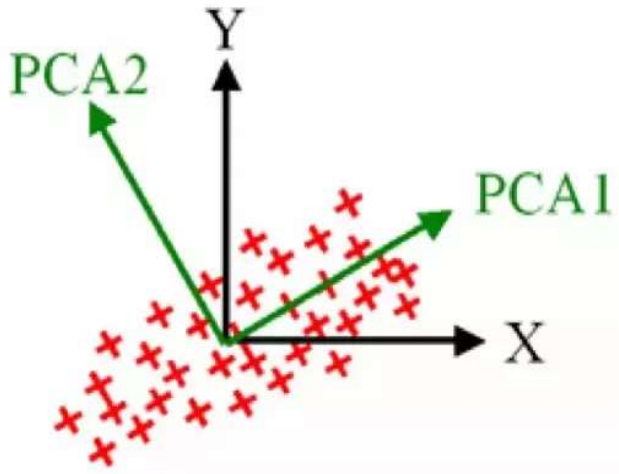


Formellement, une nouvelle représentation de X est définie par une base orthonormée sur laquelle projeter la matrice de données X

Définition

- Une ACP de la matrice $X \in \mathbb{R}^{n \times p}$ est une transformation linéaire orthogonale qui permet d'exprimer X dans une nouvelle base orthonormée, de sorte que :
  - la plus grande variance de X par projection s'aligne sur le premier axe de cette nouvelle base,
  - la seconde plus grande variance sur le deuxième axe,
  - et ainsi de suite…

→ Les axes de cette nouvelle base sont appelés composantes principales



Variantes de PCA

Dans la matrice X des données brutes chaque ligne correspond à une observation et chaque colonne à une variable initiale.

Customer (cust)	Job Configuration (conf)	First Job Year (year)	BHT (°F)	BHP (psi)	Treatment Pressure (psi)	Slurry Rate (bbl/min)	Acid Rate (bbl/min)	Proppant Concentration (lbm/gal)	Average JPT (hr)
1	Gel frac	2010	180	5,400	4,060.12	61.10	15.00	2.10	0.04
2	Gel frac	2012	175	4,500	4,824.50	96.66	10.15	2.07	0.23
3	N_2 foamed acid	2005	—	0	4,127.48	6.49	—	0.00	0.18
4	Gel frac	2011	315	13,046	10,333.36	61.94	15.93	3.51	0.14
5	Gel frac	2009	290	10,332	4,815.90	8.72	0.00	0.00	0.63
1	Gel frac	2010	290	10,500	8,536.23	55.43	9.41	2.02	1.38
3	Gel frac	2012	175	4,500	6,410.02	97.53	15.58	2.01	0.47
1	Gel frac	2007	160	5,373	5,438.87	51.72	20.00	1.61	0.29
1	Gel frac	2005	—	0	3,446.60	54.39	—	1.14	0.44
4	Gel frac	2010	180	5,175	2,572.72	46.14	8.00	2.63	0.21
5	Gel frac	2011	326	14,500	10,539.09	53.34	18.66	4.27	0.91
1	Gel frac	2007	140	3,262	2,444.42	45.27	20.00	1.29	2.74
2	Gel frac	2011	326	12,429	10,658.36	54.58	10.02	4.25	1.28
2	Gel frac	2011	334	13,713	10,765.09	99.92	15.51	4.26	1.01

Dataset

707	615	806	704	765		707	615	806	704	765	852
980	124	820	581	263		980	124	820	581	263	752
379	587	794	288	485		379	587	794	288	485	890
848	717	104	351	641		848	717	104	351	641	109
468	615	729	306	851		468	615	729	306	851	265
730	579	216	449	460		730	579	216	449	460	895
361	173	741	400	298		361	173	741	400	298	698
147	477	438	161	457		147	477	438	161	457	591

$X \in \mathbb{R}^{n \times p}$

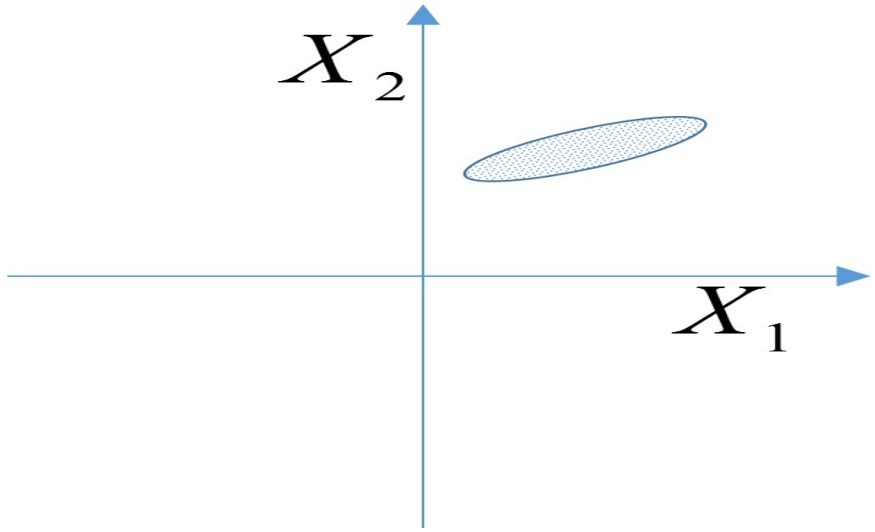
L'ACP connaît plusieurs variantes, selon le pré-traitement appliqué à la matrice X :

- **ACP générale**
- **ACP centrée**
- **ACP normée**

Variantes de PCA

I- ACP générale

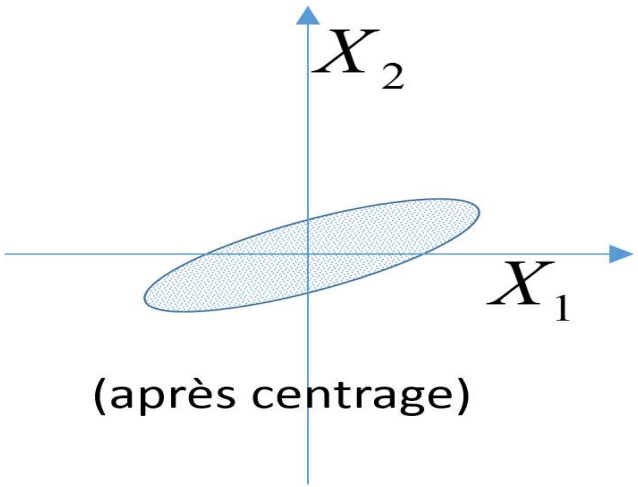
- appliquée directement sur la matrice X
- Interviennent dans l'analyse à la fois la *position* du nuage d'observations par rapport à l'origine et la *forme* du nuage.
- Cette variante est utilisée rarement, essentiellement pour tenir compte du zéro naturel de certaines variables.



Variantes de PCA

2- ACP centrée

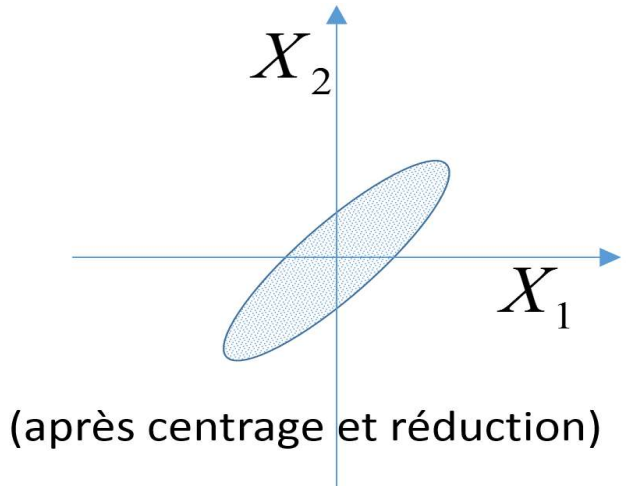
- centrage préalable des variables.
- La matrice analysée X', est obtenue en transformant X pour que chaque variable (chaque colonne) soit de moyenne nulle.
- Cela revient à s'intéresser à la *forme* du nuage d'individus par rapport à son centre de gravité.
- Cette variante est utilisée lorsque les variables initiales sont directement comparables (de même nature, intervalles de variation comparables).



Variantes de PCA

3- ACP normée

- réduction préalable des variables.
- La matrice analysée X', est obtenue en transformant X pour que chaque variable (chaque colonne) soit de moyenne nulle et d'écart-type unitaire.
- On s'intéresse donc à la forme du nuage d'individus après centrage et réduction des variables.
- Cette variante (la plus fréquemment rencontrée) est employée lorsque les variables (toutes quantitatives) sont de nature différente ou présentent des intervalles de variation très différents.



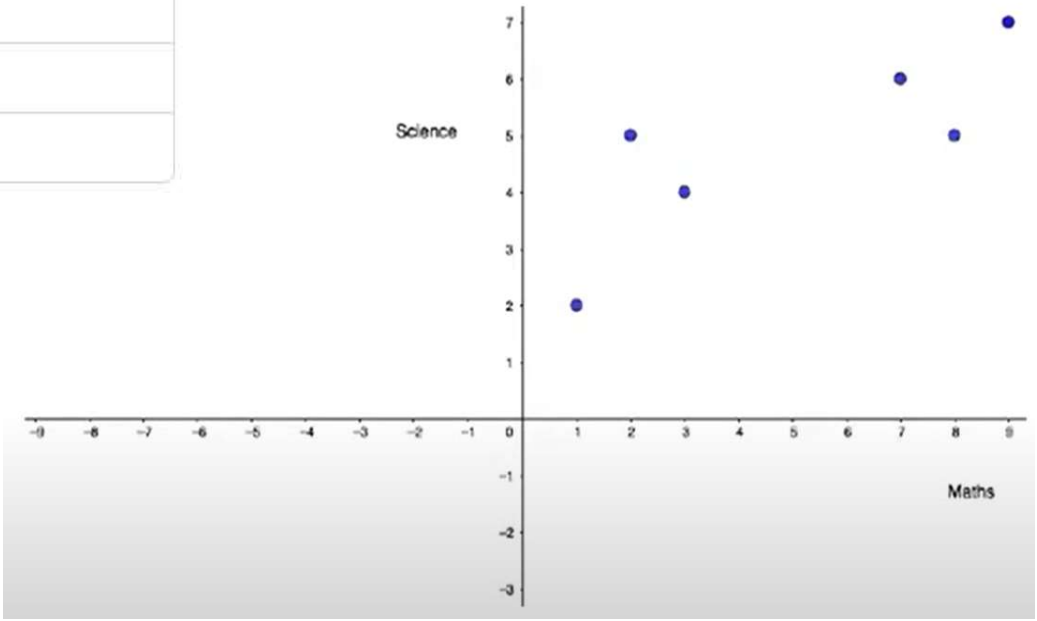
(après centrage et réduction)

Analyse en composantes principale

- Définition du PCA
- Variantes de PCA
- Comprendre l'intuition du PCA
- Mathématique de l'ACP

PCA à traves l'exemple

Étudiant	Note en Mathématiques (X_1)	Note en Français (X_2)
1	9	7
2	8	5
3	7	6
4	1	2
5	2	5
6	3	4

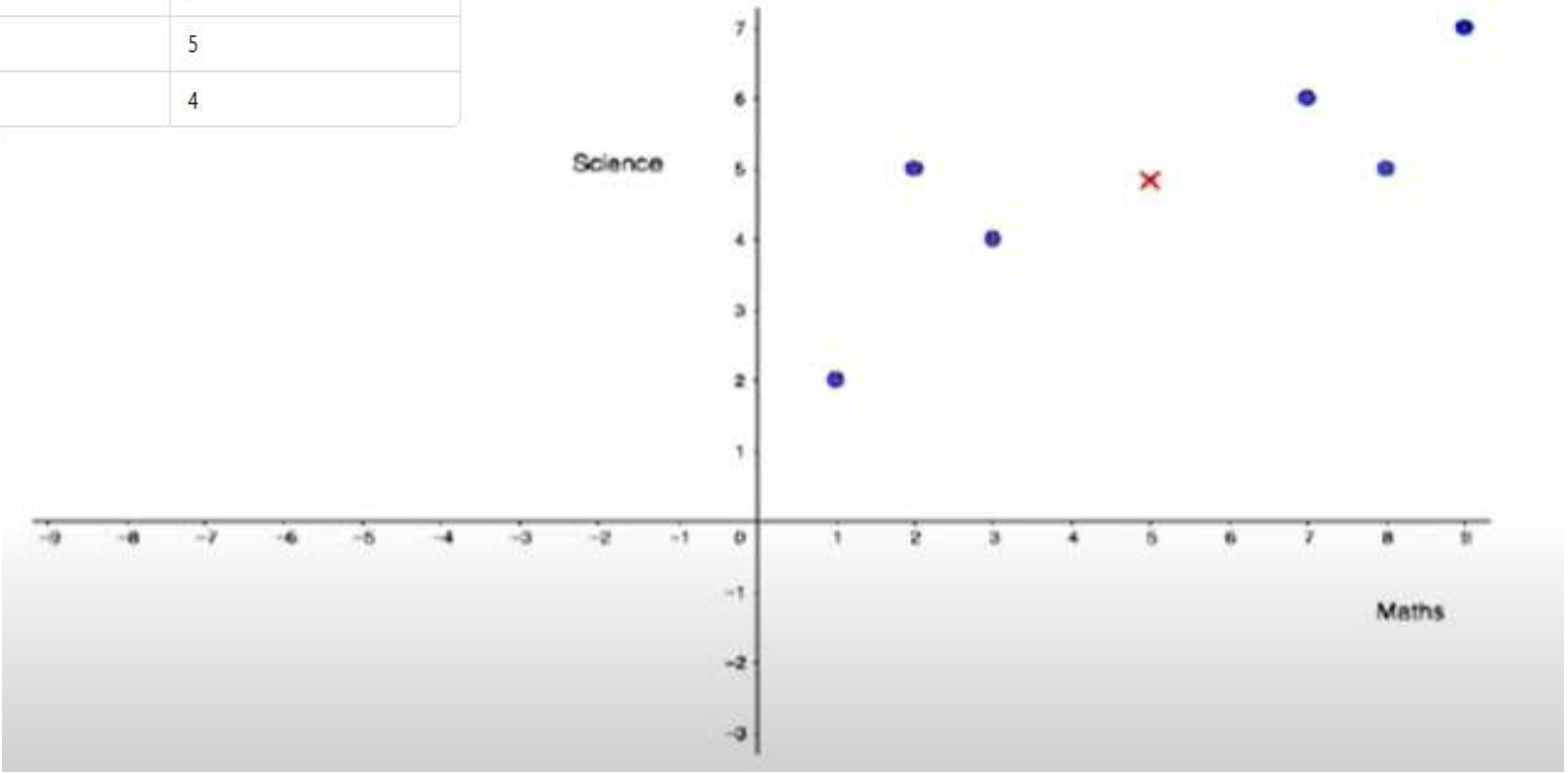


La représentation 2 D de données

PCA à traves l'exemple

- Calcule de la moyenne des données (la croix rouge)

Étudiant	Note en Mathématiques (X_1)	Note en Français (X_2)
1	9	7
2	8	5
3	7	6
4	1	2
5	2	5
6	3	4

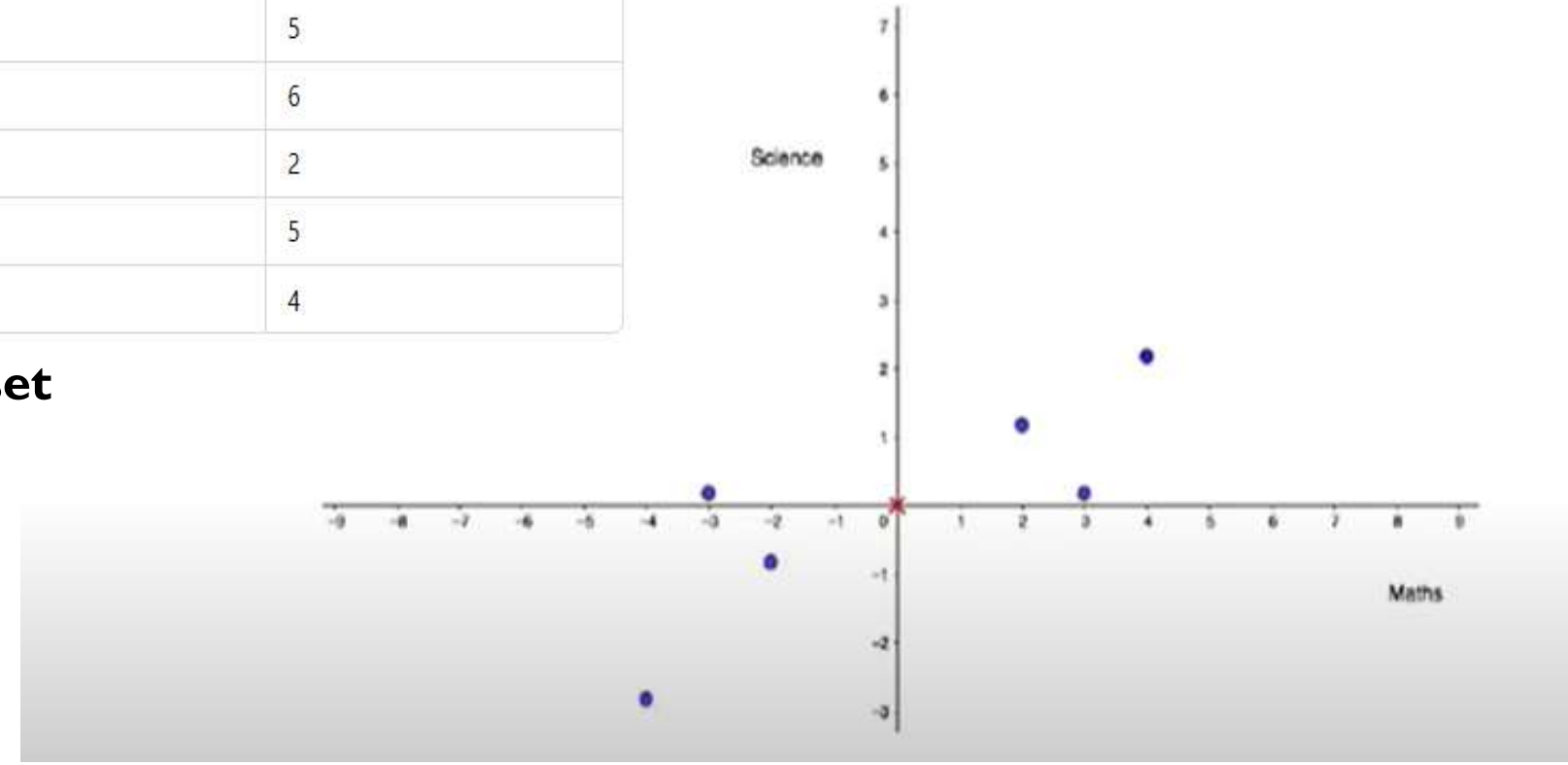


PCA à traves l'exemple

- Soustrait de la moyenne à tous les points => centrer les données pour ne pas impacter la variance de nos données

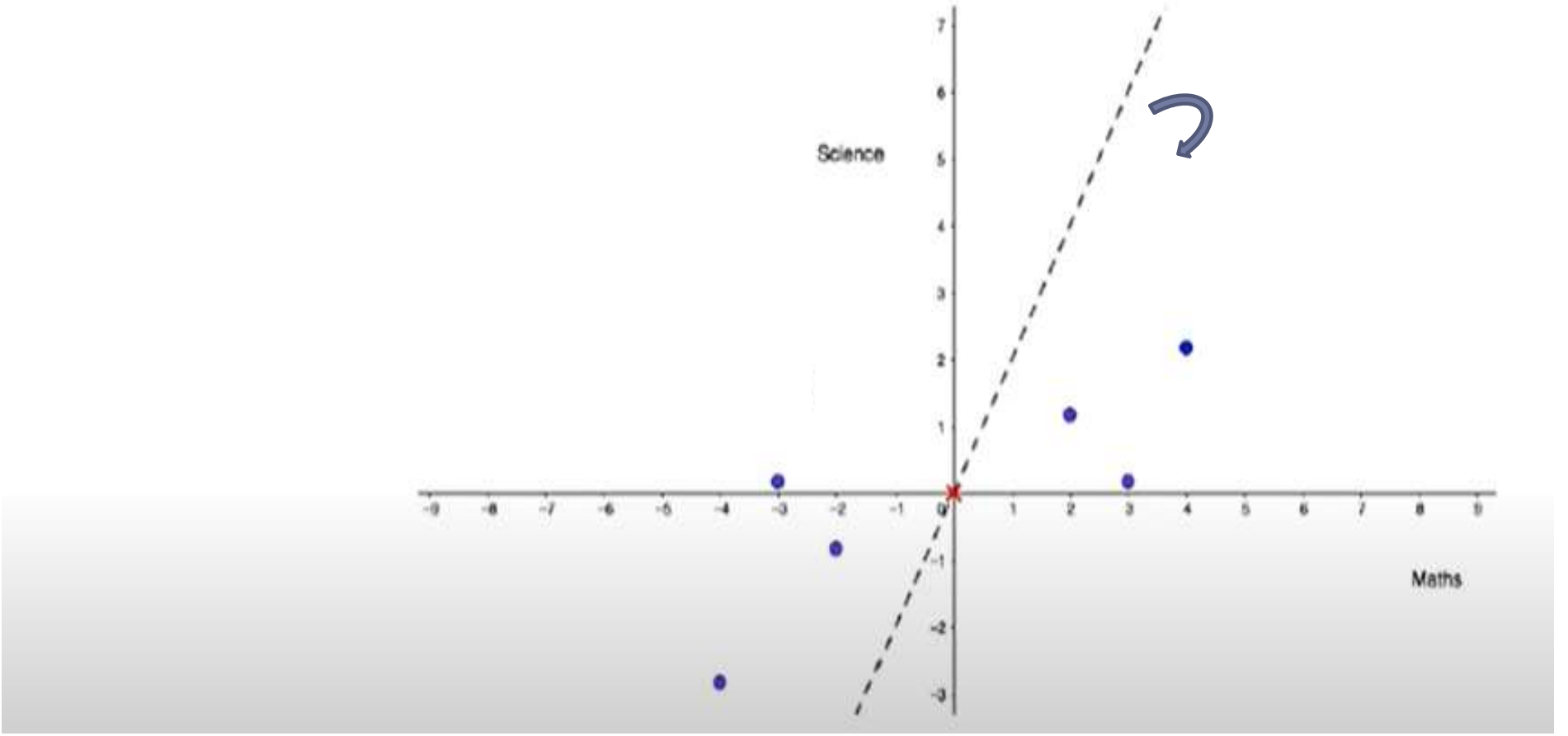
Étudiant	Note en Mathématiques (X_1)	Note en Français (X_2)
1	9	7
2	8	5
3	7	6
4	1	2
5	2	5
6	3	4

Dataset



PCA à traves l'exemple

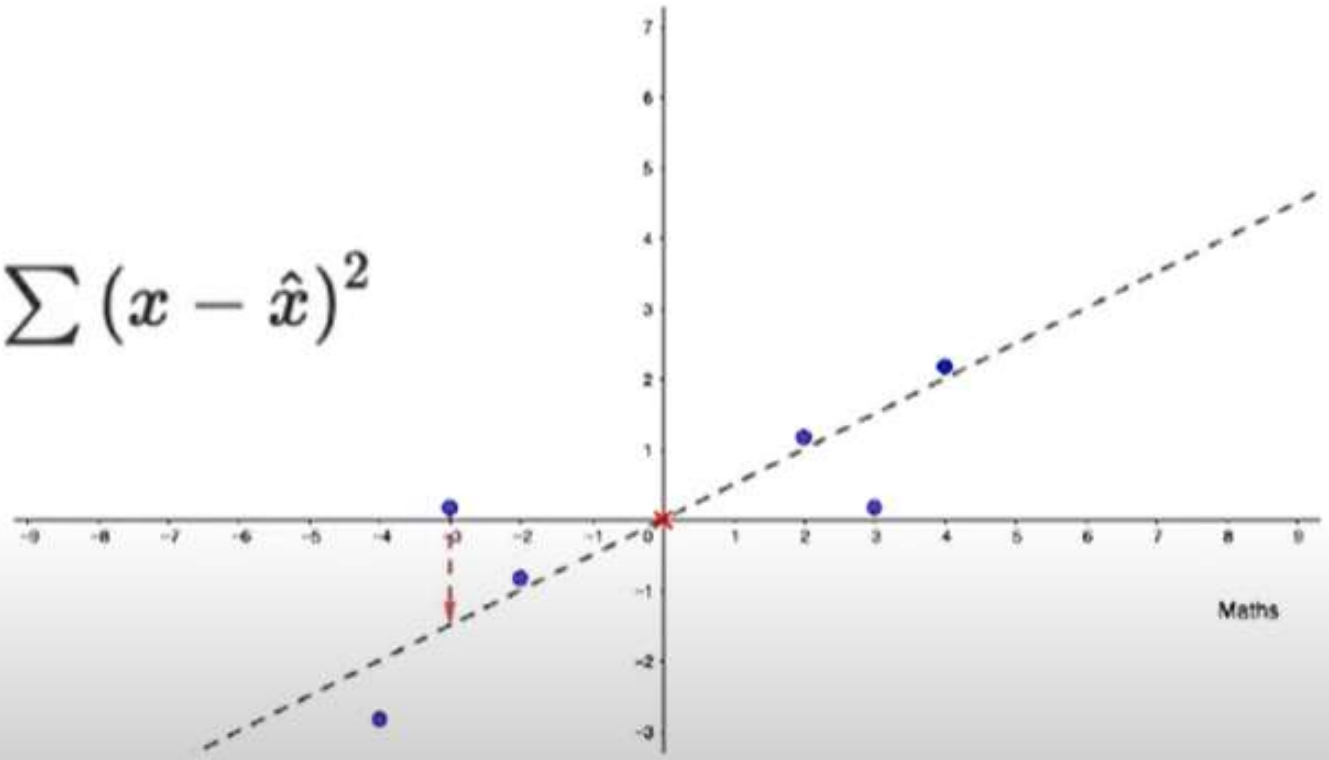
- Cherchant la projection qui minimise la distance avec les différents points



PCA à traves l'exemple

- Cette droite minimise le mieux la distance entre les points.
- Minimiser l'équation de min square erreur (par exemple MSE)
  - consiste à minimiser la moyenne de distance au carré entre les points projetés sur la droite

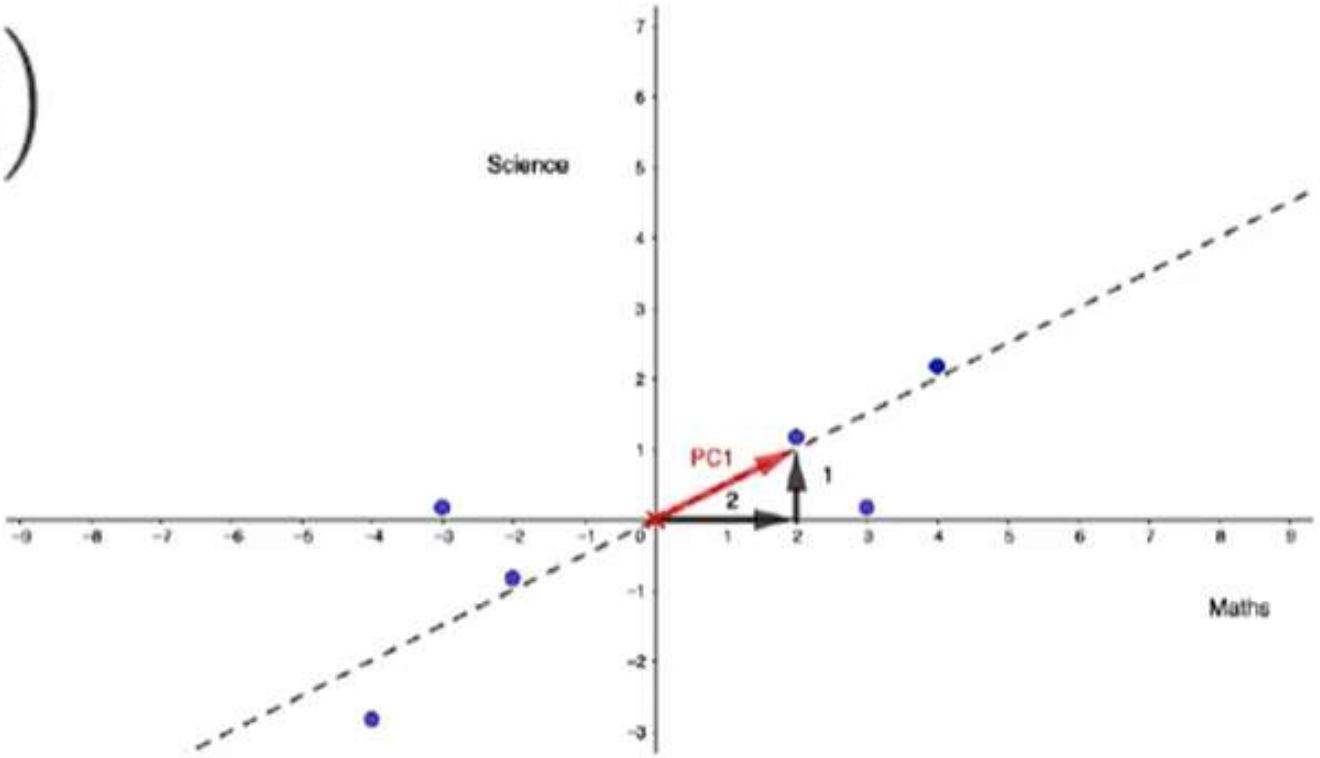
$$MSE = \frac{1}{n}\sum(x - \hat{x})^2$$



PCA à traves l'exemple

La première composante PC1

$$\overrightarrow{PC1} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$



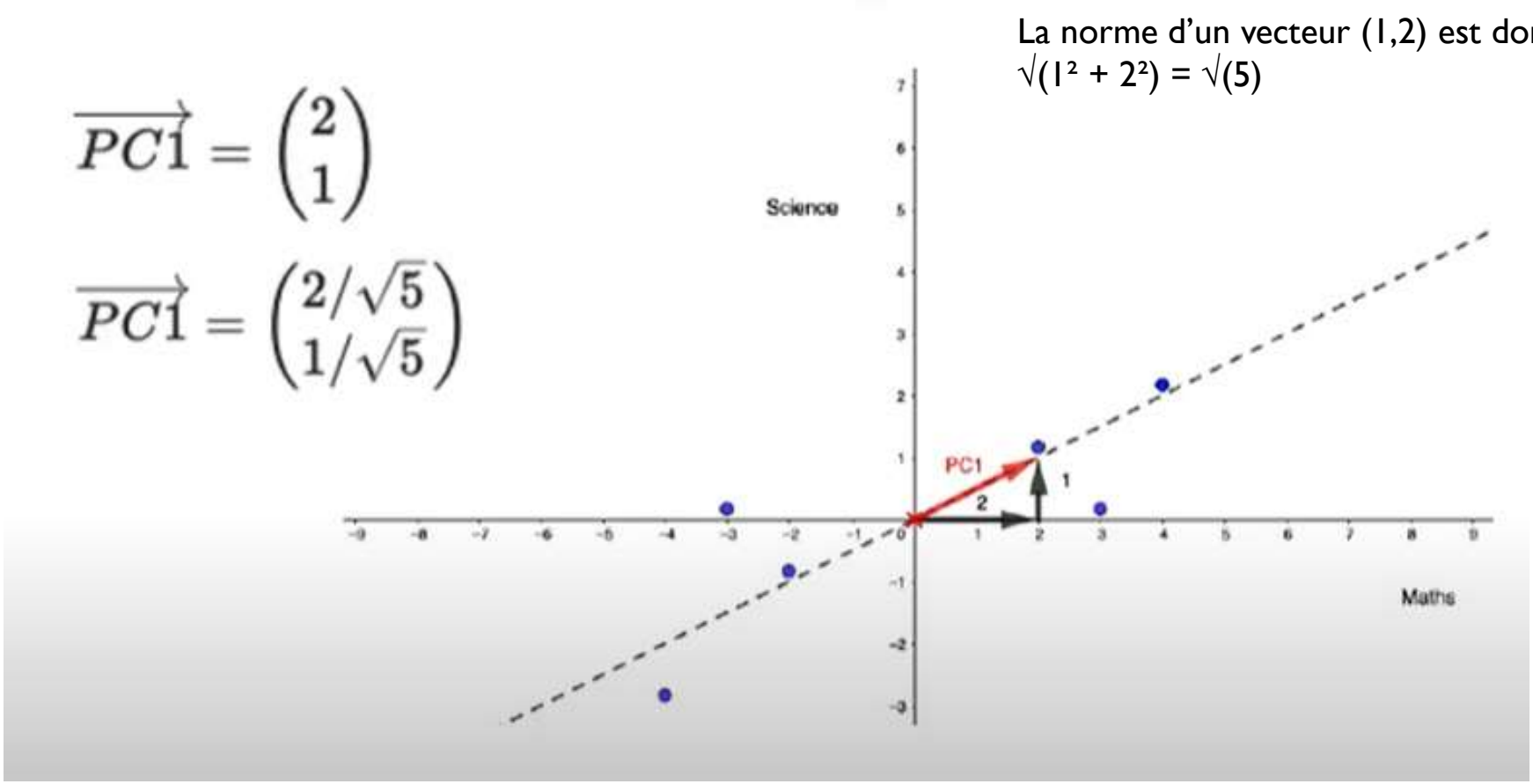
PCA à traves l'exemple

Normaliser le vecteur par sa norme ($\sqrt{(x^2 + y^2)}$)pour obtenir un vecteur unitaire car on cherche une base orthonormée

La norme d'un vecteur (1,2) est donnée par: $\sqrt{(1^2 + 2^2)} = \sqrt{(5)}$

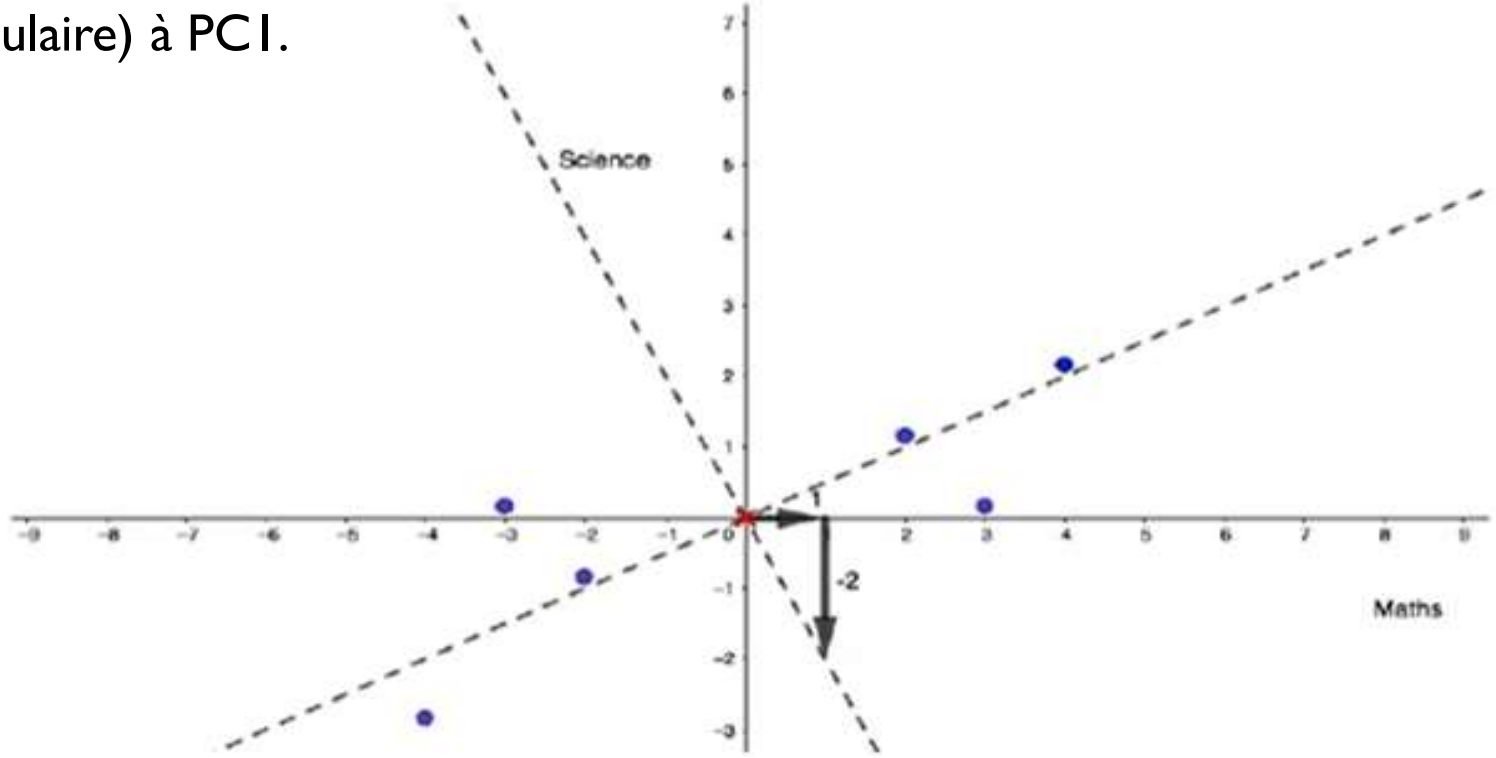
$$\overrightarrow{PC1} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

$$\overrightarrow{PC1} = \begin{pmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{pmatrix}$$



PCA à traves l'exemple

La deuxième composante, PC2 doit être **orthogonale** (perpendiculaire) à PC1.

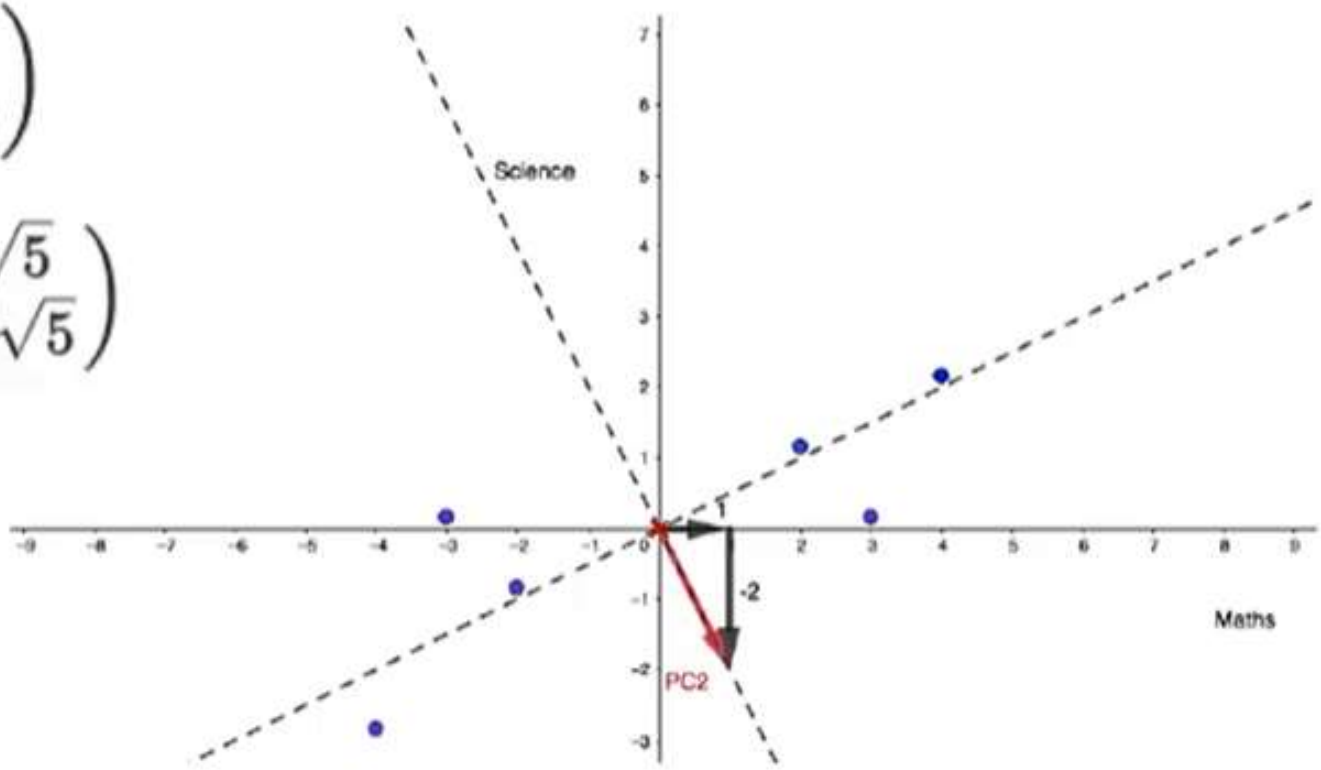


Un vecteur orthogonal à **(2,1)** est trouvé en inversant les coefficients et en changeant le signe d'un des termes soit: **(1, -2)**

PCA à traves l'exemple

$$\overrightarrow{PC2} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

$$\overrightarrow{PC2} = \begin{pmatrix} 1/\sqrt{5} \\ -2/\sqrt{5} \end{pmatrix}$$



PCA à traves l'exemple

Changer de base: **Transformer** les données en utilisant ces nouvelles directions P:(PC1, PC2).

$$X' = X \cdot P$$

Étudiant	Note en Mathématiques (X_1)	Note en Français (X_2)
1	9	7
2	8	5
3	7	6
4	1	2
5	2	5
6	3	4

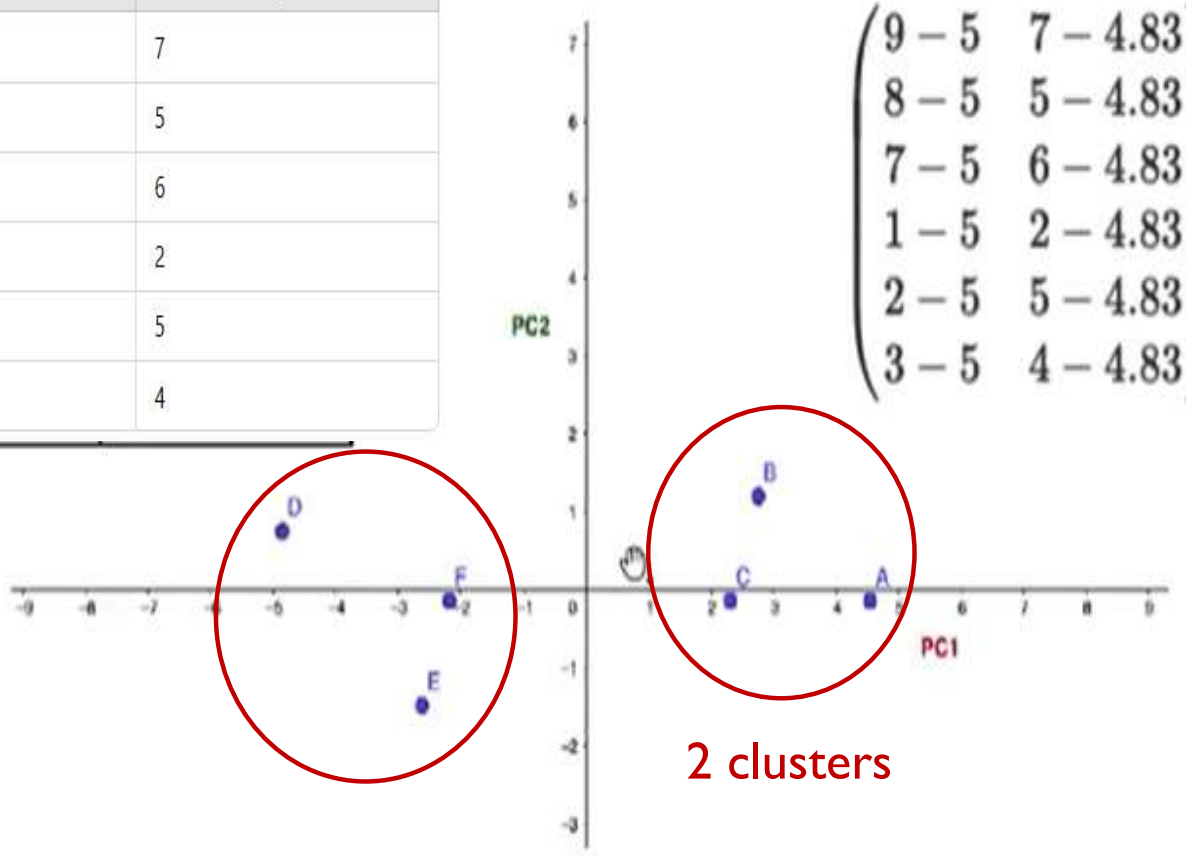
$$\begin{pmatrix} 9-5 & 7-4.83 \\ 8-5 & 5-4.83 \\ 7-5 & 6-4.83 \\ 1-5 & 2-4.83 \\ 2-5 & 5-4.83 \\ 3-5 & 4-4.83 \end{pmatrix} \begin{pmatrix} 2/\sqrt{5} & 1/\sqrt{5} \\ 1/\sqrt{5} & -2/\sqrt{5} \end{pmatrix}$$

- Calcule de produit matriciel entre les données centrées avec les composantes calculées au paravent → ce qui va donnée une nouvelle projection=> se sont les points a projeter sur les axes CP1, CP2.
- Les valeurs propres représentent **la quantité de variance** expliquée par chaque composante.

PCA à traves l'exemple

Étudiant	Note en Mathématiques (X_1)	Note en Français (X_2)
1	9	7
2	8	5
3	7	6
4	1	2
5	2	5
6	3	4

$$\begin{pmatrix} 9-5 & 7-4.83 \\ 8-5 & 5-4.83 \\ 7-5 & 6-4.83 \\ 1-5 & 2-4.83 \\ 2-5 & 5-4.83 \\ 3-5 & 4-4.83 \end{pmatrix} \begin{pmatrix} 2/\sqrt{5} & 1/\sqrt{5} \\ 1/\sqrt{5} & -2/\sqrt{5} \end{pmatrix}$$



2 clusters

La Matrice de Covariance

La **matrice de covariance** est calculée à partir des **données centrées**.

La **covariance** entre deux variables X_1 et X_2 est définie par la formule :

$$\mathrm{Cov}(X_1, X_2) = \frac{1}{n} \sum_{i=1}^{n} (X_{1,i} - \bar{X}_1)(X_{2,i} - \bar{X}_2)$$

où :

- n est le nombre d'observations
- $X_{1,i}$ et $X_{2,i}$ sont les valeurs des deux variables
- $\bar{X}_1$ et $\bar{X}_2$ sont les moyennes des variables

La Matrice de Covariance

La **matrice de covariance** est calculée à partir des **données centrées**.

Etape 1 : Calcul des moyennes

$$\bar{X}_1 = \frac{9+8+7+1+2+3}{6} = 5$$

$$\bar{X}_2 = \frac{7+5+6+2+5+4}{6} = 4.83$$

La Matrice de Covariance

Étape 2 : Centrage des données

On soustrait la moyenne de chaque colonne :

$$X_{\text{centré}} = \begin{bmatrix} 4 & 2.17 \\ 3 & 0.17 \\ 2 & 1.17 \\ -4 & -2.83 \\ -3 & 0.17 \\ -2 & -0.83 \end{bmatrix}$$

La Matrice de Covariance

Pour **une matrice X avec deux variables X_1** et X_2, la **matrice de covariance** est :

$$\Sigma = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) \end{bmatrix}$$

où :

- **Var**(X_1) est la variance de X_1
- **Var**(X_2) est la variance de X_2
- **Cov**(X_1, X_2) est la covariance entre X_1 et X_2

La Matrice de Covariance

Étape 3 : Calcul des variances et covariances

Formule de la variance :

$$\text{Var}(X_1) = \frac{1}{n}\sum(X_{1,i} - \bar{X}_1)^2$$

$$\text{Var}(X_2) = \frac{1}{n}\sum(X_{2,i} - \bar{X}_2)^2$$

$$\text{Cov}(X_1, X_2) = \frac{1}{n}\sum(X_{1,i} - \bar{X}_1)(X_{2,i} - \bar{X}_2)$$

AN: $\Sigma = \begin{bmatrix} 9.67 & 4.00 \\ 4.00 & 2.47 \end{bmatrix}$

- 9.67 : Variance de la première variable X_1, qui mesure la dispersion de la première colonne.
- 2.47 : Variance de la deuxième variable X_2, qui mesure la dispersion de la deuxième colonne.
- 4.00 : Covariance entre X_1 et X_2, qui indique dans quelle mesure les deux variables varient ensemble.

Calcul des valeurs propres

Les valeurs propres λ_1 et λ_2 sont obtenues en résolvant :

$$\det(\Sigma - \lambda I) = 0$$

Après calcul, les valeurs propres sont :

$$\lambda_1 = 11.31, \quad \lambda_2 = 0.83$$

Calcul du ratio de variance expliquée

Le **ratio de variance expliquée** par chaque composante principale est donné par :

$$\frac{\lambda_i}{\sum \lambda}$$

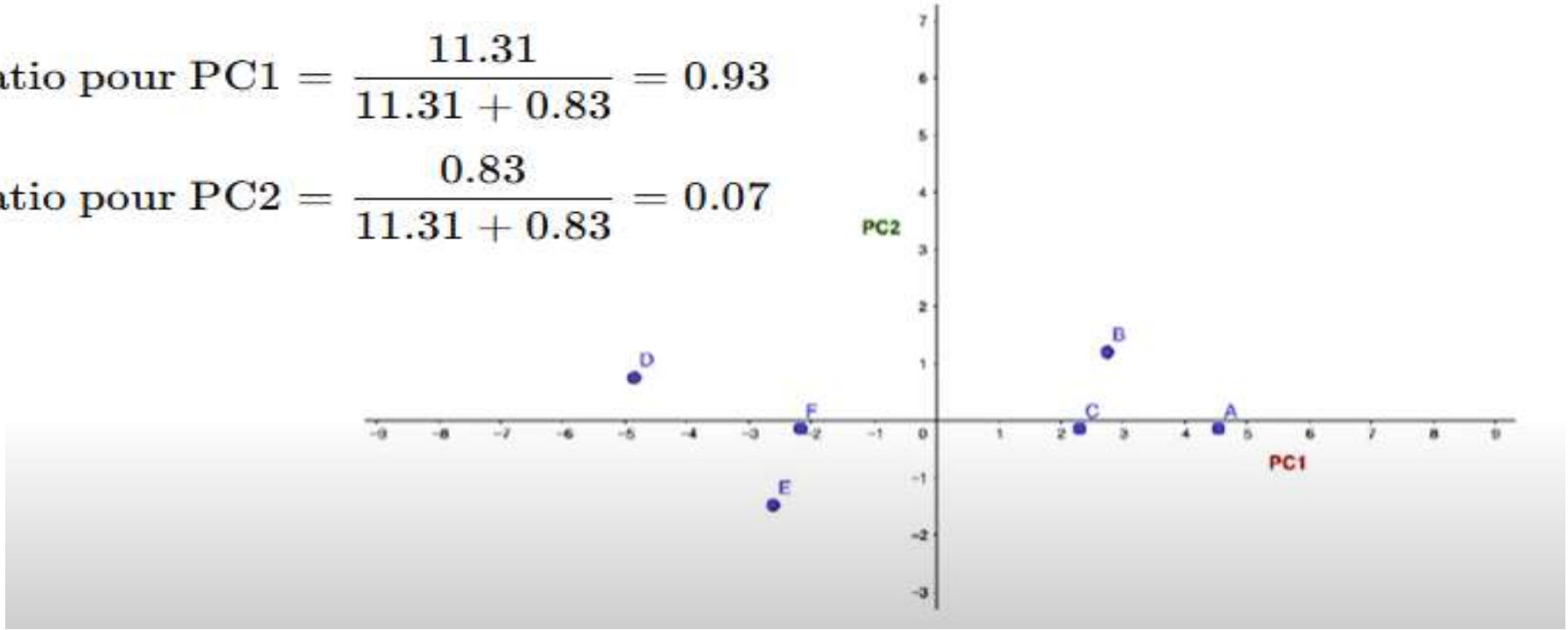
$$\text{Ratio pour PC1} = \frac{11.31}{11.31 + 0.83} = 0.93$$

$$\text{Ratio pour PC2} = \frac{0.83}{11.31 + 0.83} = 0.07$$

Interprétation des Résultats

$$\text{Ratio pour PC1} = \frac{11.31}{11.31 + 0.83} = 0.93$$

$$\text{Ratio pour PC2} = \frac{0.83}{11.31 + 0.83} = 0.07$$



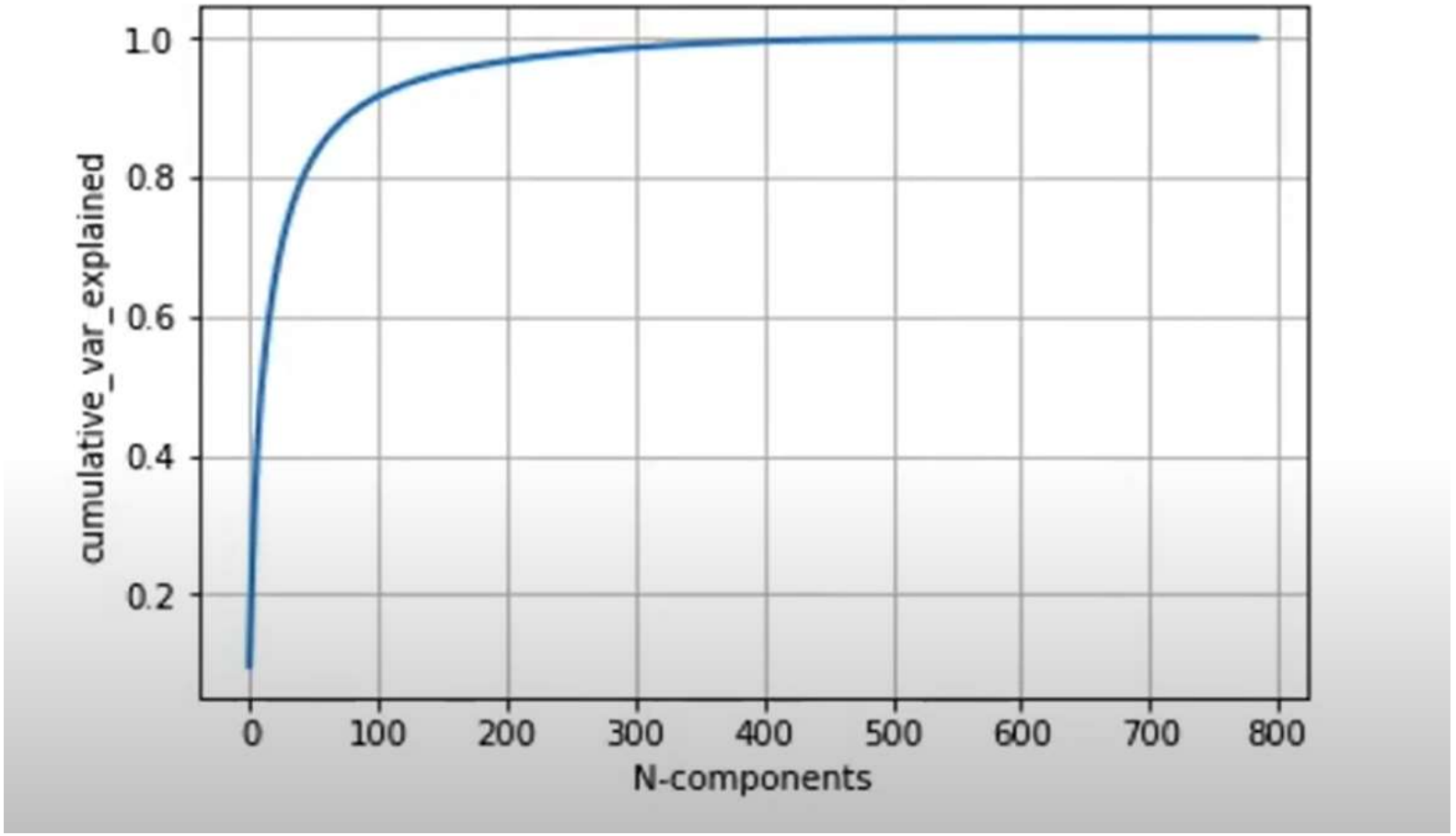
- **PC1 capture 93% de la variance totale** des données.
- **PC2 capture seulement 7% de la variance**, ce qui signifie que PC2 contient peu d'information.

Interprétation des Résultats

Projection sur l'axe avec
la variance la plus élevée

Choix de nombre de composantes

Variance cumulée



A 95% de l'information on peut choisir 100 composantes

Analyse en composantes principale

- Définition du PCA
- Variantes de PCA
- Comprendre l'intuition du PCA
- Code Python (TP1)