

Statistique et apprentissage

Chargés de cours (ordre alphabétique) :

Julien Bect, Ziad Kobeissi, Gilles Faÿ, Laurent Le Brusquet,
Vincent Lescarret, Arshak Minasyan, Arthur Tenenhaus[†] & Xujia Zhu

[†] Coordinateur du cours

1/53

Cours 10/9

Apprentissage non supervisé : deux exemples

Objectifs du cours 10

- ▶ Comprendre les enjeux de l'apprentissage non-supervisé à travers deux exemples de tâches non supervisées.
- ▶ Savoir réduire la dimension d'un jeu de données grâce à l'**analyse en composantes principales**.
- ▶ Savoir réaliser un partitionnement de données (*clustering*) par l'**algorithme des K-moyennes**.

2/53

Plan du cours

- 1 – Introduction à l'apprentissage non supervisé
- 2 – Analyse en composantes principales
- 3 – Clustering
- 4 – Un avant-goût de quelques méthodes plus avancées
- 5 – Annexes

3/53

Plan du cours

- 1 – Introduction à l'apprentissage non supervisé
- 2 – Analyse en composantes principales
- 3 – Clustering
- 4 – Un avant-goût de quelques méthodes plus avancées
- 5 – Annexes

Rappel : apprentissage supervisé

- ▶ On observe des **couples** (X_i, Y_i) :

$$(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{\text{iid}}{\sim} P^{X,Y},$$

avec $X_i \in \mathcal{X}$: **exemple** et $Y_i \in \mathcal{Y}$: **étiquette**.

- ▶ On cherche à approcher le **prédicteur optimal**

$$h^* = \operatorname{argmin}_h \mathbb{E}(L(Y, h(X))),$$

qui est une **propriété de la loi conditionnelle** $P^{Y|X}$:

$$\begin{aligned} h^*(x) &= \operatorname{argmin}_{\tilde{y} \in \mathcal{Y}} \mathbb{E}(L(Y, \tilde{y}) \mid X = x) \\ &= \operatorname{argmin}_{\tilde{y} \in \mathcal{Y}} \int L(y, \tilde{y}) P^{Y|X=x}(dy). \end{aligned}$$

4/53

Apprentissage non supervisé

Apprentissage sans l'aide d'un « enseignant » :

- ▶ on observe **seulement les exemples**,

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} P^X,$$

et on s'intéresse à la loi **P^X** .

Supposons que $\mathcal{X} \subset \mathbb{R}^p$ et que P^X admet une ddp f^X .

Problème : fléau de la dimension

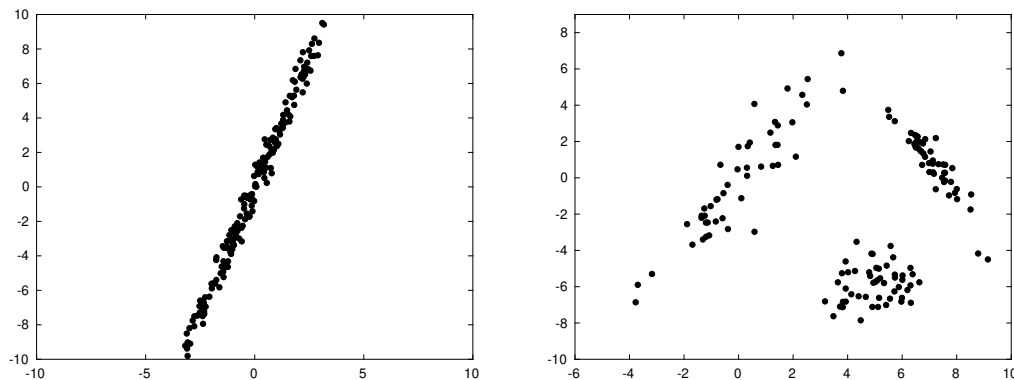
Estimer une densité f^X « quelconque » a un coût (taille d'échantillon nécessaire pour atteindre une précision donnée) qui **augmente exponentiellement avec la dimension p** .[†]

[†] La *statistique non paramétrique*, une branche de la statistique qui s'intéresse notamment à l'estimation de densité sous des hypothèses très générales, fournit des énoncés rigoureux (hors programme) qui permettent d'étayer cette affirmation.

5/53

Objectifs de l'apprentissage non-supervisé

- 1 Idéalement, **estimer la densité** f^X de la loi des données.
 - ▢ à moins que p ne soit assez petit (disons $p \lesssim 5$, rare en apprentissage), c'est en général un **problème trop difficile**[†].
- 2 **Révéler des « structures »** dans la loi des données
(sans passer explicitement par l'estimation de la densité)



[†] En petite dimension, on peut utiliser par exemple un *estimateur à noyau* (hors programme).

6/53

Plan du cours

1 – Introduction à l'apprentissage non supervisé

2 – Analyse en composantes principales

2.1 – Approximation de faible rang

2.2 – Recherche du sous-espace optimal : la SVD

2.3 – Variances et covariances empiriques des composantes

3 – Clustering

4 – Un avant-goût de quelques méthodes plus avancées

5 – Annexes

Objectif : réduction de dimension

On cherche une transformation

$$\begin{aligned} T : \mathcal{X} &\rightarrow \mathcal{Z} \subset \mathbb{R}^q && \text{avec } q \ll p \\ x &\mapsto z = T(x) \end{aligned}$$

ainsi qu'une procédure de **reconstruction**

$$\begin{aligned} \tilde{T} : \mathcal{Z} &\rightarrow \mathcal{X} \\ z &\mapsto \hat{x} = \tilde{T}(z) \end{aligned}$$

telles que

$$\frac{1}{n} \sum_{i=1}^n L(x_i, \hat{x}_i) = \frac{1}{n} \sum_{i=1}^n L\left(x_i, \underbrace{\tilde{T}(T(x_i))}_{z_i}\right)$$

soit le plus petit possible (avec $L(x, \hat{x})$ une fonction de perte).

Remarque : plus généralement, \mathcal{Z} pourrait être une *variété* de dimension q , généralisation abstraite des notions de courbes ($q = 1$) et de surface ($q = 2$) ; voir géométrie différentielle.

7/53

Plan du cours

1 – Introduction à l'apprentissage non supervisé

2 – Analyse en composantes principales

2.1 – Approximation de faible rang

2.2 – Recherche du sous-espace optimal : la SVD

2.3 – Variances et covariances empiriques des composantes

3 – Clustering

4 – Un avant-goût de quelques méthodes plus avancées

5 – Annexes

Réduction « linéaire » de dimension

Soit $x_1, \dots, x_n \in \mathbb{R}^p$ un échantillon observé. Soit $q < p$.

Définition : sous-espace affine

$\mathcal{A}_q \subset \mathbb{R}^p$ est un **sous-espace affine** de **dimension q** s'il existe

- ▶ $\mu \in \mathbb{R}^p$,
- ▶ A matrice de taille $p \times q$ de **rang q** ,

tels que $\mathcal{A}_q = \text{Aff}_{\mu, A} = \{y \in \mathbb{R}^p \text{ tel que } y = \mu + Az, z \in \mathbb{R}^q\}$.

Définition : analyse en composantes principales (ACP)

L'ACP consiste à trouver la **meilleure approximation** des données, pour la **perte quadratique**, par un **sous-espace affine \mathcal{A}_q** .

La dimension q est soit donnée, soit déterminée automatiquement.

8/53

Réduction « linéaire » de dimension (suite)

On cherche donc $\mathcal{A}_q = \text{Aff}_{\mu, A}$ et (z_i) tels que

$$\mu, A, (z_i) \in \operatorname{argmin} \sum_{i=1}^n \|x_i - (\mu + Az_i)\|^2. \quad (*)$$



Il n'y a **pas unicité** de la solution.

- ⇒ Si \tilde{A} a la même image que A , alors il existe des \tilde{z}_i tels que $Az_i = \tilde{A}\tilde{z}_i$ pour tout i .
- ⇒ On supposera sans perte de généralité que les colonnes de la matrice A sont **orthonormées** :

$$A^T A = \text{Id}_q.$$

Remarque : l'hypothèse d'orthonormalité ne lève pas complètement l'indétermination de A ...

9/53

Réduction « linéaire » de dimension (suite)

⇒ Fixons μ , A et (z_i) , et posons $\tilde{z}_i = z_i - \bar{z}$. Alors

$$\begin{aligned}\mu + Az_i &= \mu + A(\tilde{z}_i + \bar{z}) \\ &= \underbrace{\mu + A\bar{z}}_{\tilde{\mu}} + A\tilde{z}_i.\end{aligned}$$

⇒ Sans perte de généralité, on cherchera les z_i t.q. $\bar{z} = 0$.

10/53

Résultat partiel

Proposition

La minimisation du critère pour une matrice A donnée conduit à :

$$\begin{aligned}\mu &= \bar{x}, \\ z_i &= A^\top (x_i - \bar{x}),\end{aligned}$$

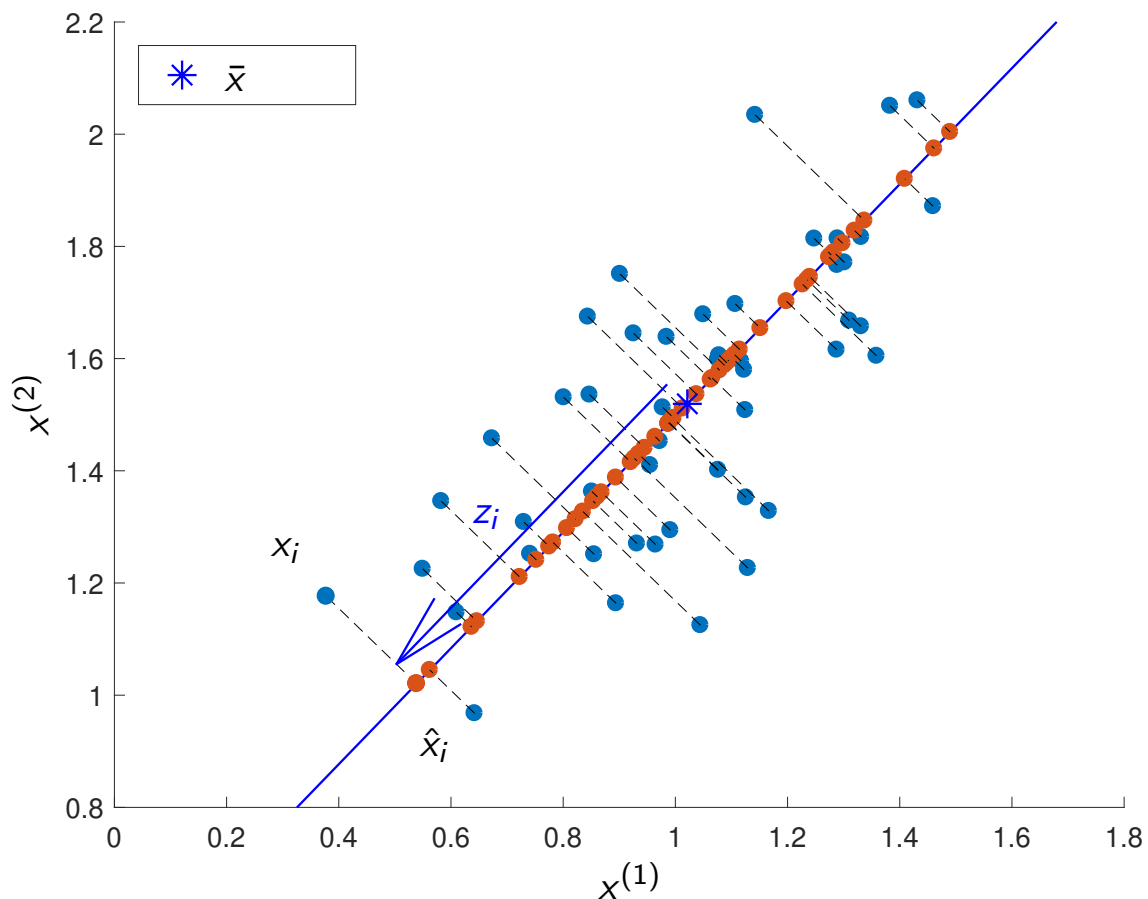
et on a l'interprétation géométrique :

⇒ $\hat{x}_i = \mu + Az_i$ est la **projection orthogonale** de x_i sur $\text{Aff}_{\mu,A}$.

Conséquence. En insérant ce résultat dans (\star) , il vient

$$A = \operatorname{argmin} \sum_{i=1}^n \left\| \left(\operatorname{Id}_p - AA^\top \right) (x_i - \bar{x}) \right\|^2.$$

11/53



Résultat partiel : démonstration

Fixons A et (z_i) , avec $\bar{z} = 0$, et posons $v_i = x_i - Az_i$. Alors

$$\begin{aligned} \sum_i \|x_i - (\mu + Az_i)\|^2 &= \sum_i \|v_i - \mu\|^2 \\ &= n \left\| \mu - \frac{1}{n} \sum_i v_i \right\|^2 + c \end{aligned}$$

où c ne dépend pas de μ . Ainsi, le μ optimal est

$$\mu = \frac{1}{n} \sum_i v_i = \bar{x} - A\bar{z} = \bar{x}.$$

On fixe donc $\mu = \bar{x}$, et on procède de façon similaire pour déterminer chaque z_i . Pour tout i le minimum est atteint (exercice) en

$$z_i = A^\top (x_i - \bar{x}),$$

et on vérifie que $\bar{z} = \frac{1}{n} \sum_i z_i = A^\top (\bar{x} - \bar{x}) = 0$. □

Remarque : on peut aussi retrouver les expressions rapidement en calculant le gradient du critère.

Plan du cours

1 – Introduction à l'apprentissage non supervisé

2 – Analyse en composantes principales

2.1 – Approximation de faible rang

2.2 – Recherche du sous-espace optimal : la SVD

2.3 – Variances et covariances empiriques des composantes

3 – Clustering

4 – Un avant-goût de quelques méthodes plus avancées

5 – Annexes

Notations

Soit X la matrice des observations :

$$X = \begin{pmatrix} (x_1)^\top \\ \vdots \\ (x_n)^\top \end{pmatrix} = \begin{pmatrix} x_1^{(1)} & \dots & x_1^{(p)} \\ \vdots & & \vdots \\ x_n^{(1)} & \dots & x_n^{(p)} \end{pmatrix}$$

Dans la suite on supposera, sans perte de généralité, $\bar{x} = 0$.

On cherche A telle que

$$\begin{aligned} A &= \operatorname{argmin} \sum_{i=1}^n \left\| \left(\operatorname{Id}_p - AA^\top \right) x_i \right\|^2 \\ &= \operatorname{argmin} \left\| \left(\operatorname{Id}_p - AA^\top \right) X^\top \right\|_F^2 \end{aligned}$$

avec $\|\cdot\|_F$ la **norme de Frobenius** :

$$\|M\|_F^2 = \sum_{i,j} M_{ij}^2 = \operatorname{tr}(M^\top M) = \operatorname{tr}(MM^\top).$$

Décomposition en valeurs singulières (SVD)

Théorème

Soit M une matrice réelle de taille $n \times p$. Il existe des matrices :

- ▶ U orthogonale de taille $n \times n$ ($U^\top U = \text{Id}_n$),
- ▶ V orthogonale de taille $p \times p$ ($V^\top V = \text{Id}_p$),
- ▶ $D = \text{diag}(d_1, \dots, d_r, 0, \dots, 0)$ de taille $n \times p$,
avec $d_1 \geq d_2 \geq \dots \geq d_r > 0$

telles que :

$$M = UDV^\top,$$

et r est le rang des matrices D et M .

Les scalaires $d_1, \dots, d_r, 0, \dots, 0$ sont les **valeurs singulières de M** .

- ▶ d_1^2, \dots, d_r^2 sont les valeurs propres $\neq 0$ de MM^\top et $M^\top M$.

Démonstration. Voir TD 8, exercice bonus.



15/53

Solution du problème d'optimisation

Soient U , D et V les matrices issues de la SVD de X :

$$X = UDV^\top.$$

Théorème fondamental de l'ACP

Soient

- ▶ v_1, v_2, \dots, v_p les colonnes de V ,
- ▶ $V_q = (v_1 \mid \dots \mid v_q)$ la sous-matrice des q premières colonnes.

Alors

$$V_q \in \operatorname{argmin}_A \left\| \left(\text{Id}_p - AA^\top \right) X^\top \right\|_F^2,$$

où A parcourt l'ensemble des matrices $p \times q$ de rang q .

L'ACP en résumé

Algorithme : Analyse en Composantes Principales (ACP)

Réaliser l'ACP de l'échantillon (x_1, \dots, x_n) consiste à :

- 1 Calculer la moyenne \bar{x} et **centrer les observations** : $x_i \leftarrow x_i - \bar{x}$.
- 2 Former la matrice X des observations centrées.
- 3 Calculer la matrice V par **SVD de X**
(les valeurs singulières sont utiles aussi, cf. section suivante)
- 4 Réaliser la **réduction de dimension** : $z_i = V_q^\top x_i$.

Reconstruction. $\hat{x}_i = \bar{x} + V_q z_i$.

Vocabulaire.

- ▶ v_1, \dots, v_q (colonnes de V_q) : **axes principaux**.
- ▶ $z_i^{(1)}, \dots, z_i^{(q)}$: **composantes principales**.

17/53

Exemple « Code postal » (pas MNIST, un autre!)

Données : $n = 658$ images 16×16 du chiffre « 3 » $\rightarrow p = 256$

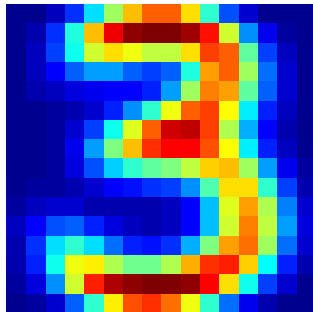


Source : The Elements of Statistical Learning, Springer

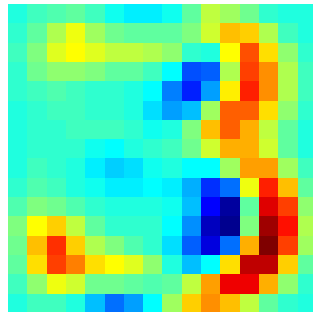
18/53

Exemple « Code postal » (suite)

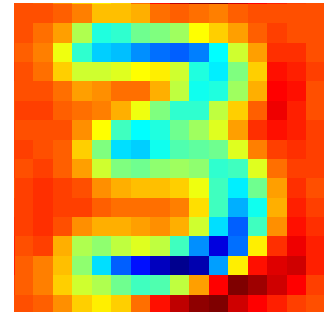
Visualisation des 2 premiers axes principaux



moyenne \bar{x}



axe principal v_1

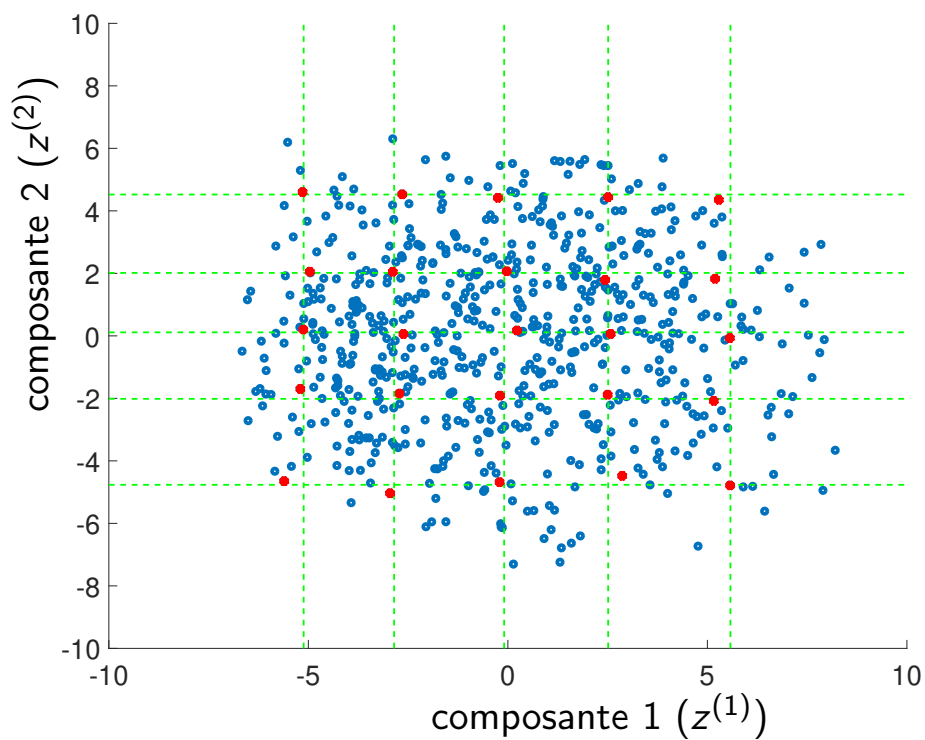


axe principal v_2

$$\forall i, \hat{x}_i = \bar{x} + z_i^{(1)} v_1 + z_i^{(2)} v_2$$

19/53

Plan principal $(z^{(1)}, z^{(2)})$

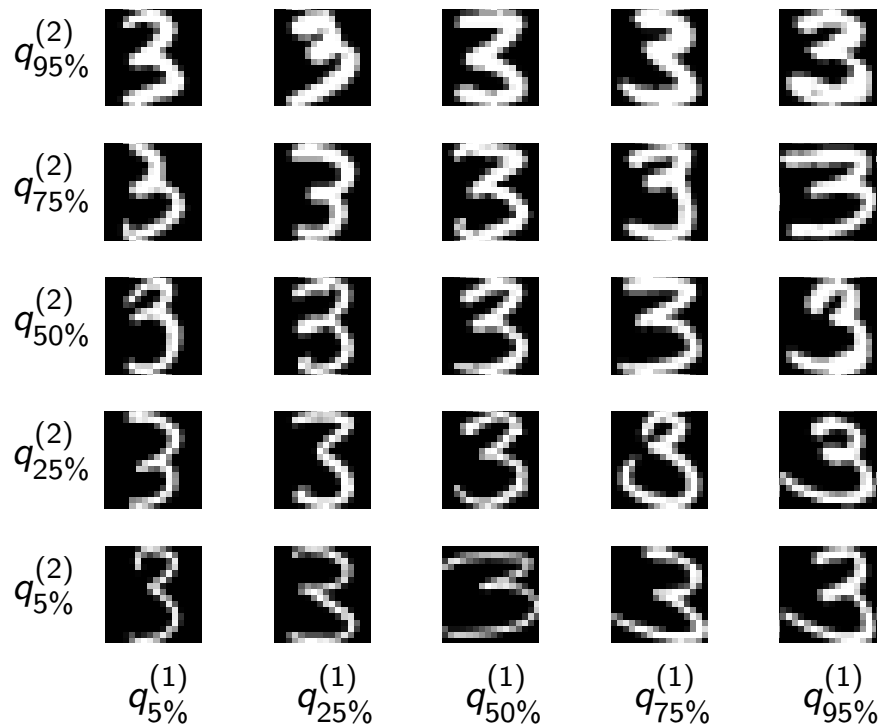


Lignes en pointillés : quantiles à 5%, 25%, 50%, 75%, 95%.

Points rouges : exemples montrés sur le slide suivant.

20/53

Interprétation des composantes ($z^{(1)}, z^{(2)}$) à l'aide de 25 observations localisées sur une grille du plan principal.



21/53

Plan du cours

1 – Introduction à l'apprentissage non supervisé

2 – Analyse en composantes principales

2.1 – Approximation de faible rang

2.2 – Recherche du sous-espace optimal : la SVD

2.3 – Variances et covariances empiriques des composantes

3 – Clustering

4 – Un avant-goût de quelques méthodes plus avancées

5 – Annexes

Matrice de covariance des composantes

Soit $\hat{\Sigma}_Z$ la matrice de covariance empirique des q composantes

$$\begin{aligned}\hat{\Sigma}_Z &= \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})(z_i - \bar{z})^\top \\ &= \frac{1}{n} \sum_{i=1}^n z_i z_i^\top \quad (\text{car } \bar{z} = \frac{1}{n} \sum_{i=1}^n z_i = 0) \\ &= \frac{1}{n} \mathbf{Z}^\top \mathbf{Z}\end{aligned}$$

avec $\mathbf{Z} = \begin{pmatrix} z_1^\top \\ \vdots \\ z_n^\top \end{pmatrix}$. Rappel : $z_i = V_q^\top x_i$, donc $\mathbf{Z} = \mathbf{X} V_q$.

En utilisant $\mathbf{X} = \mathbf{U} \mathbf{D} \mathbf{V}^\top$, il vient

$$\begin{aligned}\hat{\Sigma}_Z &= \frac{1}{n} V_q^\top V D^\top D V^\top V_q \\ &= \frac{1}{n} \text{diag}(d_1^2, \dots, d_q^2).\end{aligned}$$

22/53

Matrice de covariance des composantes (suite)

Conclusions.

- ▶ La **variance (empirique)** de la composante $z^{(j)}$ est $\frac{d_j^2}{n}$.
 - ▮▮▮ Composantes rangées par ordre de variance décroissante.
- ▶ Les **covariances (empiriques)** sont nulles.
 - ▮▮▮ Les composantes sont **décorrélées**.

23/53

Variance totale d'un échantillon

Définition / Proposition

La **variance totale** de l'échantillon p -varié (x_1, \dots, x_n) est

$$VT(x_1, \dots, x_n) = \sum_{j=1}^p \text{var} \left(x_1^{(j)}, \dots, x_n^{(j)} \right).$$

Les x_i étant centrés, on a

$$VT(x_1, \dots, x_n) = \frac{1}{n} \text{tr}(X^\top X) = \frac{1}{n} \sum_{j=1}^r d_j^2.$$

Démonstration. En effet, en utilisant que les x_i sont centrés, on a

$$VT(x_1, \dots, x_n) = \sum_{j=1}^p \left(\frac{1}{n} \sum_{i=1}^n \left(x_i^{(j)} \right)^2 \right) = \frac{1}{n} \|X\|_F^2 = \frac{1}{n} \text{tr}(X^\top X).$$

En utilisant ensuite $X = UDV^\top$, avec $U^\top U = \text{Id}_n$ et $V^\top V = \text{Id}_p$, il vient

$$VT(x_1, \dots, x_n) = \frac{1}{n} \text{tr}(D^\top D) = \frac{1}{n} \sum_{j=1}^r d_j^2.$$

24/53

Proportion de variance expliquée

Variance totale de l'échantillon reconstruit $(\hat{x}_1, \dots, \hat{x}_n)$:

$$VT(\hat{x}_1, \dots, \hat{x}_n) = \frac{1}{n} \text{tr}(\hat{X}^\top \hat{X}) = ?.$$

En utilisant $\hat{X} = ZV_q^\top$, il vient :

$$VT(\hat{x}_1, \dots, \hat{x}_n) = \text{tr}(V_q \hat{\Sigma}_Z V_q^\top) = \frac{1}{n} \sum_{j=1}^q d_j^2.$$

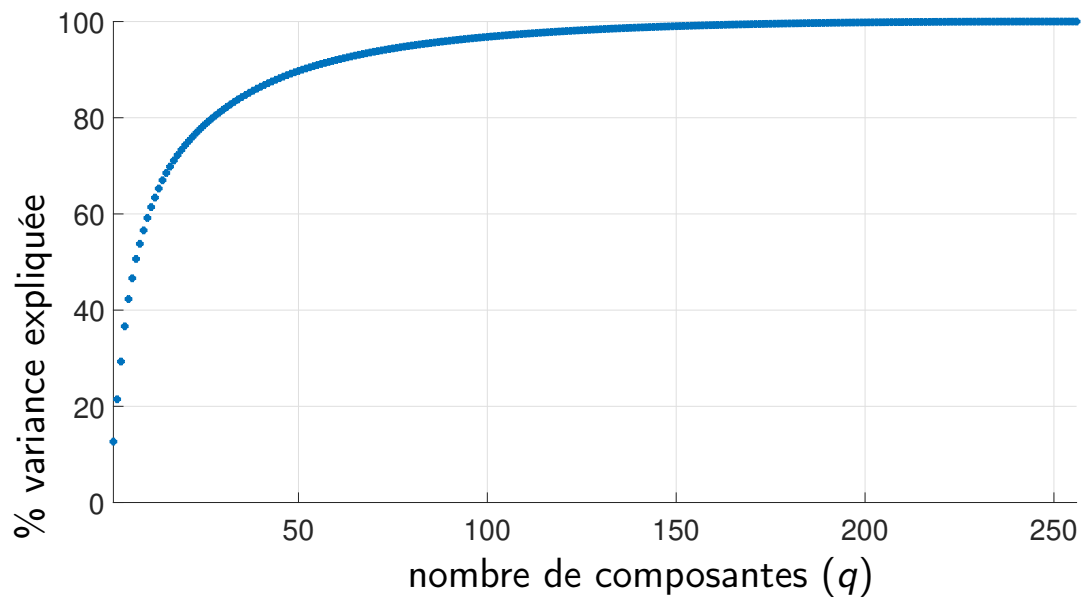
Proportion de variance expliquée

On appelle **proportion de variance expliquée** la quantité

$$\frac{VT(\hat{x}_1, \dots, \hat{x}_n)}{VT(x_1, \dots, x_n)} = \frac{\sum_{j=1}^q d_j^2}{\sum_{j=1}^r d_j^2}.$$

25/53

Exemple « Code postal » (MNIST, $p = 28^2 = 784$)



Remarque : similarité avec le coefficient de détermination (R^2) en régression.

26/53

Plan du cours

1 – Introduction à l'apprentissage non supervisé

2 – Analyse en composantes principales

3 – Clustering

3.1 – Dissimilarité

3.2 – Algorithme K -means

3.3 – Choix du nombre de clusters

4 – Un avant-goût de quelques méthodes plus avancées

5 – Annexes

Définition : clustering, clusters

Soit $E = \{x_1, \dots, x_n\}$ un échantillon de n observations $x_i \in \mathcal{X}$.

- ▶ On suppose $\mathcal{X} \subset \mathbb{R}^p$, donc $E \subset \mathbb{R}^p$.

Définitions

Le clustering[†] consiste à **partitionner** l'ensemble E en K **parties non vides** $E_k \subset E$, $1 \leq k \leq K$, regroupant des données « similaires ».

Le nombre K peut être fixé ou déterminé automatiquement.

Les ensembles E_k sont appelés **groupes** ou grappes (**clusters**).

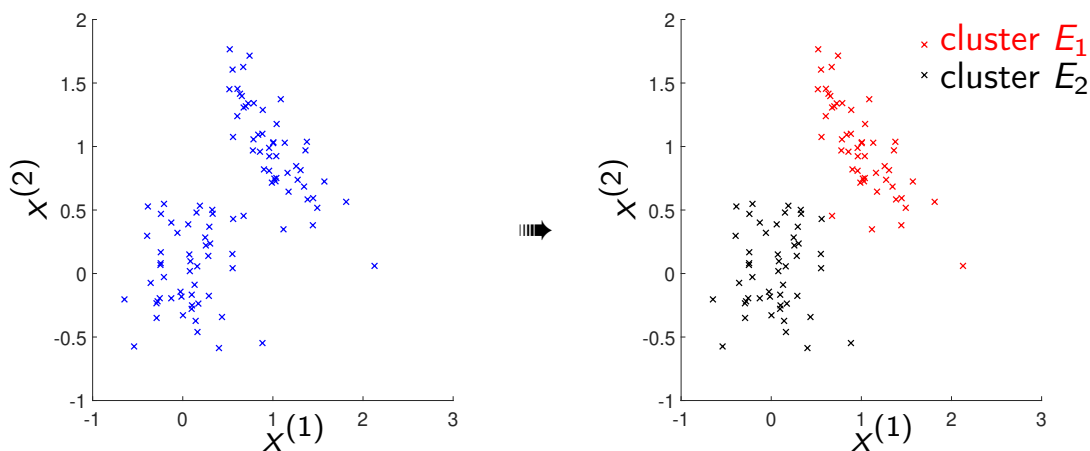
Notations.

- ▶ On note $\pi_k = \{i \leq n \mid x_i \in E_k\}$ les indices du cluster E_k .
- ▶ $\Pi = \{\pi_1, \dots, \pi_K\}$ est une partition de $\{1, \dots, n\}$.

[†] ou « partitionnement des données », en bon français.

Un exemple de résultat

Exemple avec $p = 2$ et $K = 2$



Plan du cours

1 – Introduction à l'apprentissage non supervisé

2 – Analyse en composantes principales

3 – Clustering

3.1 – Dissimilarité

3.2 – Algorithme K -means

3.3 – Choix du nombre de clusters

4 – Un avant-goût de quelques méthodes plus avancées

5 – Annexes

Dissimilarité : définition

On cherche une partition telle que, pour tout k ,

- ▶ les exemples[†] du cluster E_k sont « similaires » entre eux,
- ▶ et aussi dissimilaires que possibles de ceux des autres clusters.

Définition

Dans les algorithmes de clustering, on appelle **dissimilarité** la fonction $D : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ utilisée pour mesurer la « distance » entre les exemples.

Remarque : **pas toujours une distance** mais vérifie en général

- ▶ la propriété de **symétrie** : $D(x, y) = D(y, x)$,
- ▶ la propriété de **positivité** : $D(x, y) \geq 0$.

[†] ou « observations », « données », « individus » ...

Dissimilarité : exemples

- ▶ Forme générale : $D(x_i, x_{i'}) = \sum_{j=1}^p d_j(x_i^{(j)}, x_{i'}^{(j)})$
- ▶ Variable quantitative : $d_j(x_i^{(j)}, x_{i'}^{(j)}) = f(|x_i^{(j)} - x_{i'}^{(j)}|)$.
Exemple : $d_j(x_i^{(j)}, x_{i'}^{(j)}) = (x_i^{(j)} - x_{i'}^{(j)})^2$.
Remarque : il est souvent utile de normaliser les variables :
 $x_i^{(j)} \rightarrow \frac{x_i^{(j)}}{s_j}$, (choix courant pour s_j : écart-type empirique)
- ▶ Variable qualitative : $d_j(x_i^{(j)}, x_{i'}^{(j)}) = \text{cste}$ si $x_i^{(j)} \neq x_{i'}^{(j)}$ (0 sinon)

30/53

Inertie intra-cluster & inertie inter-cluster

Notons $d_{ii'} = D(x_i, x_{i'})$.

Inertie intra-cluster

On appelle inertie **intra**-cluster la quantité :

$$W(\Pi) = \frac{1}{2} \sum_{k=1}^K \sum_{i, i' \in \pi_k} d_{ii'}.$$

(W=Within)

Inertie inter-cluster

On appelle inertie **inter**-cluster la quantité :

$$B(\Pi) = \frac{1}{2} \sum_{k, k' \neq k} \sum_{i \in \pi_k} \sum_{i' \in \pi_{k'}} d_{ii'}.$$

(B=Between)

31/53

Inertie intra-cluster & inertie inter-cluster (suite)

Propriété

$$W(\Pi) + B(\Pi) = \frac{1}{2} \sum_{i,i'} d_{ii'}$$

Définition

$T = \frac{1}{2} \sum_{i,i'} d_{ii'}$ est l'**inertie totale**.

- Ne dépend pas de la partition.

Démonstration de la propriété :

$$\begin{aligned} T &= \frac{1}{2} \sum_{i,i'} d_{ii'} = \frac{1}{2} \sum_{k,k'} \sum_{i \in \pi_k} \sum_{i' \in \pi_{k'}} d_{ii'} \\ &= \underbrace{\frac{1}{2} \sum_k \sum_{i,i' \in \pi_k} d_{ii'}}_{W(\Pi)} + \underbrace{\frac{1}{2} \sum_{k,k' \neq k} \sum_{i \in \pi_k} \sum_{i' \in \pi_{k'}} d_{ii'}}_{B(\Pi)} \end{aligned}$$

32/53

Partition optimale

On aimerait déterminer la **partition optimale** :

$$\Pi_{\star} = \arg \min_{\Pi} W(\Pi)$$

Remarque : comme $W(\Pi) + B(\Pi) = T$, $\Pi_{\star} = \arg \max_{\Pi} B(\Pi)$.

Difficulté : Il s'agit d'un problème d'optimisation **combinatoire**

- 34105 partitions pour $n = 10$ et $K = 4$,
- $\approx 7.5 \cdot 10^{11}$ partitions pour $n = 20$ et $K = 5$.

Solution : Recherche d'une partition sous-optimale

- ➡ algorithme K -means

33/53

Plan du cours

1 – Introduction à l'apprentissage non supervisé

2 – Analyse en composantes principales

3 – Clustering

3.1 – Dissimilarité

3.2 – Algorithme K -means

3.3 – Choix du nombre de clusters

4 – Un avant-goût de quelques méthodes plus avancées

5 – Annexes

Dissimilarité considérée : $d_{ij'} = \|x_i - x_{i'}\|^2$.

Avec cette dissimilarité ([démonstration](#)) :

$$W(\Pi) = \sum_{k=1}^K n_k \sum_{i \in \pi_k} \|x_i - \bar{x}_k\|^2$$

où $\bar{x}_k = \frac{1}{n_k} \sum_{i \in \pi_k} x_i$ est le barycentre du cluster, $n_k = |\pi_k|$.

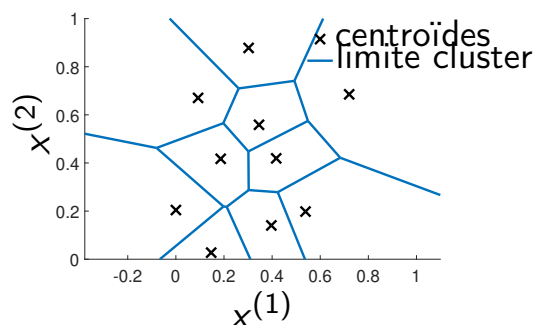
⇒ \bar{x}_k s'appelle le **centroïde** du cluster k .

Principe de l'algorithme K -means

Itérativement,

- ▶ Partant d'une partition Π , calculer les centroïdes \bar{x}_k .
- ▶ Modifier Π de sorte que chaque x_i soit associé au cluster π_k dont le centroïde (courant) \bar{x}_k est le plus proche.

⇒ diagramme de Voronoï



Algorithme K –means

Require: $K > 0$

{nombre de clusters}

Require: $(\bar{x}_{1,0}, \dots, \bar{x}_{K,0})$

{initialisation des centroïdes}

$t \leftarrow 0$

repeat

Step 1

{construction de Π_t à partir des centroïdes}

for all k **do**

$\pi_{k,t} = \{i \text{ t.q. } k = \arg \min_{k'} \|x_i - \bar{x}_{k',t}\|\}$

end for

Step 2

{mise à jour des centroïdes}

for all k **do**

$\bar{x}_{k,t} = \frac{1}{|\pi_{k,t}|} \sum_{i \in \pi_{k,t}} x_i$

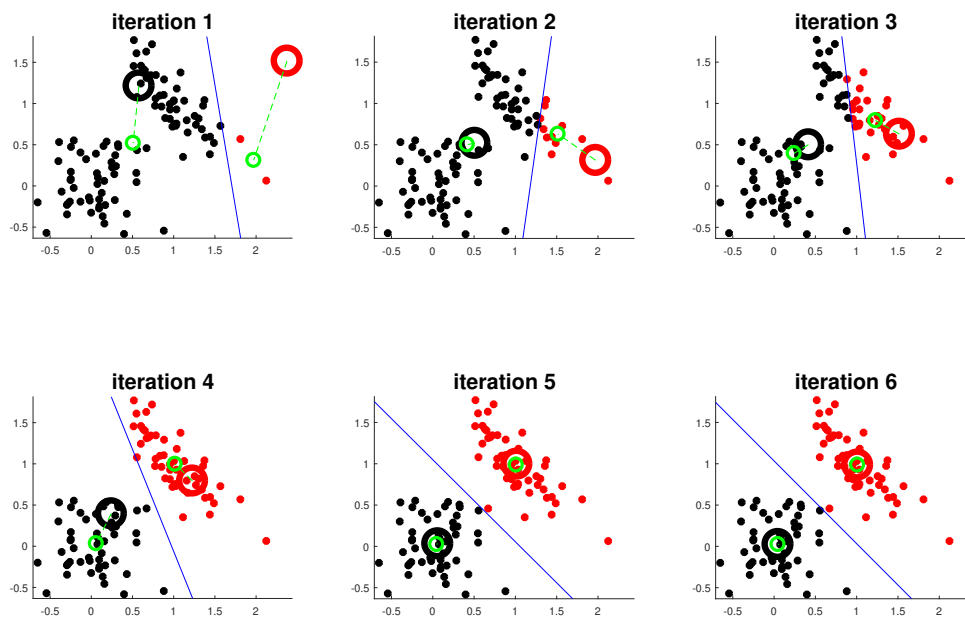
end for

$t \leftarrow t + 1$

until $W(\Pi_{t-1}) = W(\Pi_{t-2})$

return Π_{t-1}

35/53



36/53

Propriétés de l'algorithme K –means

Proposition

Soit $(\Pi_t)_{t \geq 0}$ la suite des partitions construites par l'algorithme.

Alors, il existe T tel que :

- ① $\forall t \leq T, W(\Pi_t) < W(\Pi_{t-1}),$
- ② $W(\Pi_{T+1}) = W(\Pi_T).$



L'algorithme se termine en un nombre fini d'itérations, mais

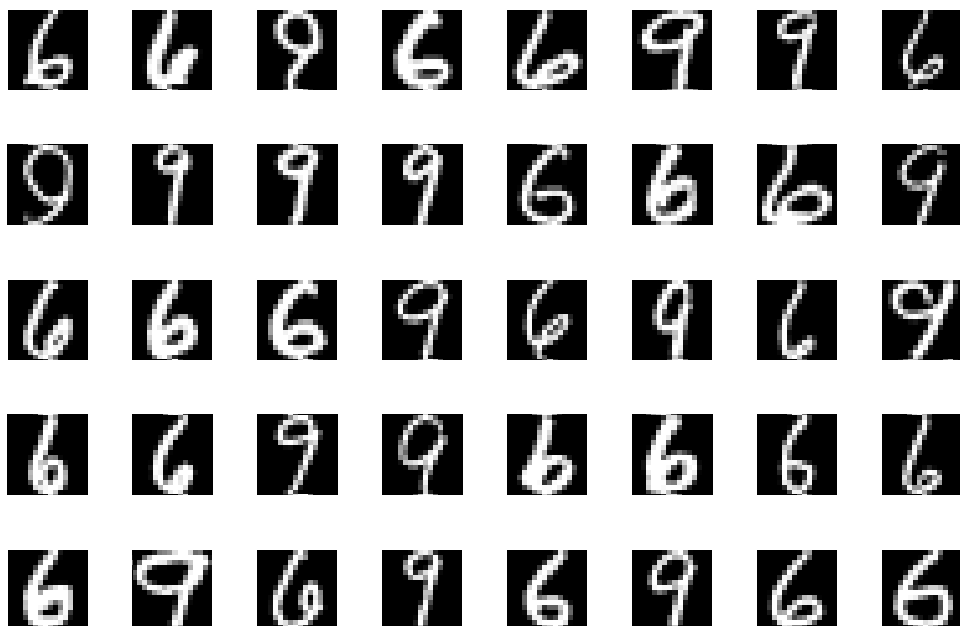
- ▶ la partition Π_T n'est **pas, en général, la partition optimale** ;
- ▶ elle dépend du point de départ $(\bar{x}_{1,0}, \dots, \bar{x}_{K,0})$.

⇒ Recommandé : essais avec différentes initialisations aléatoires.

37/53

Retour sur l'exemple « Code postal »

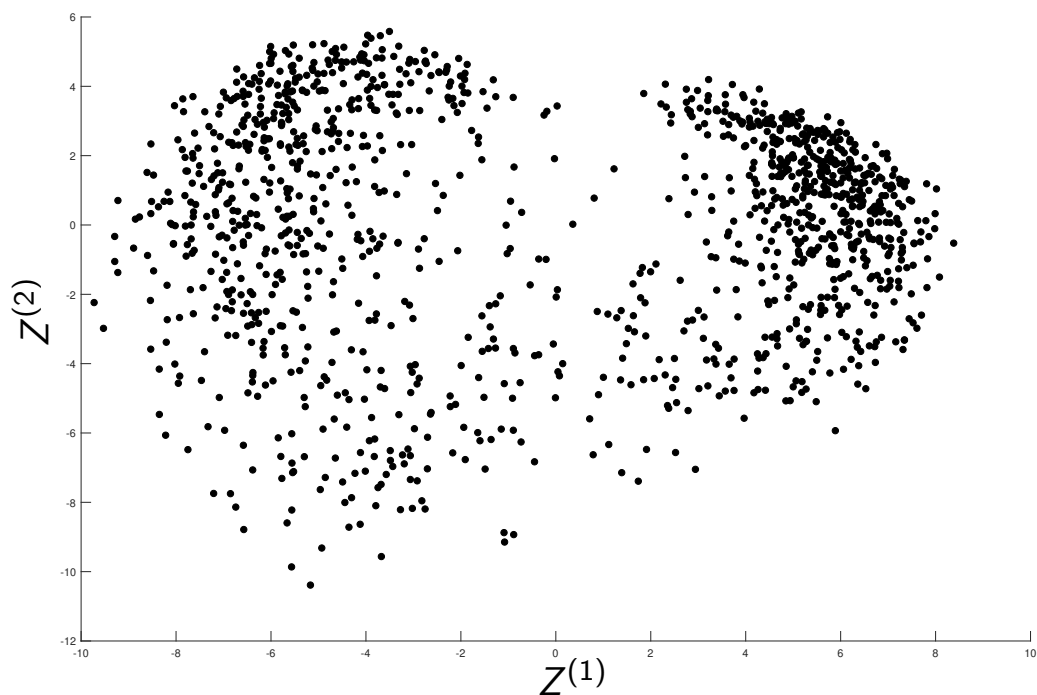
Considérons les chiffres « 6 » (664 images) et « 9 » (644 images)



38/53

Exemple « Code postal »

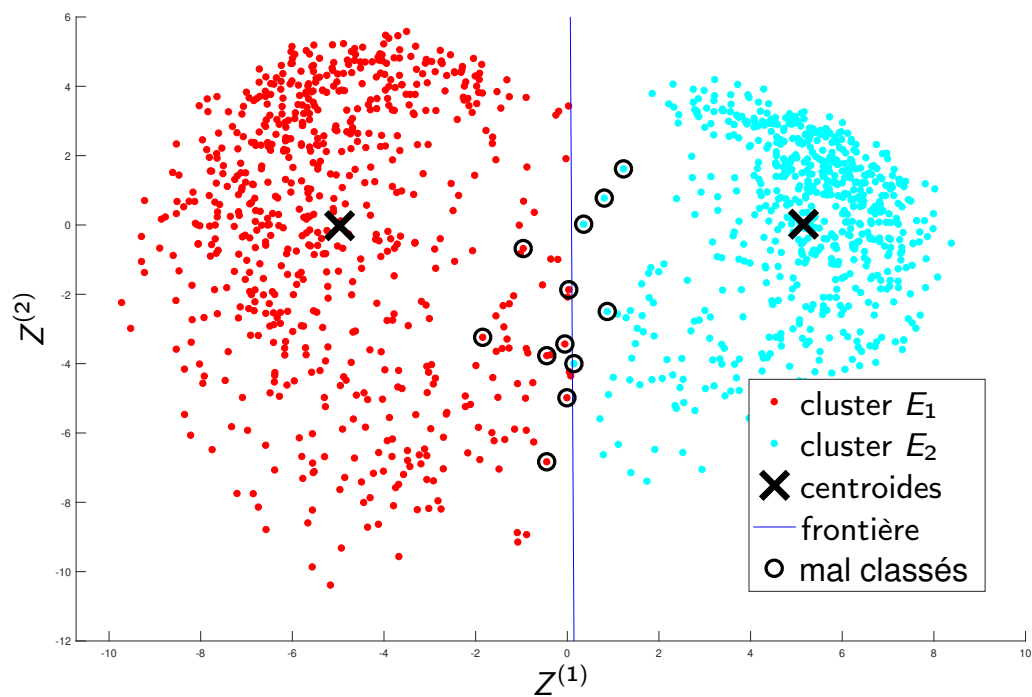
Appliquons l'algorithme K -means dans le plan principal (ACP).



39/53

Exemple « Code postal »

taux mal classés = 0.92%



Note : on utilise ici les étiquettes, supposées indisponibles dans un cadre non-supervisé, à la seule fin d'évaluer la qualité de la partition obtenue.

40/53

Plan du cours

1 – Introduction à l'apprentissage non supervisé

2 – Analyse en composantes principales

3 – Clustering

3.1 – Dissimilarité

3.2 – Algorithme K -means

3.3 – Choix du nombre de clusters

4 – Un avant-goût de quelques méthodes plus avancées

5 – Annexes

Homogénéité / dispersion

Rappel. On cherche une partition telle que, pour tout k ,

- ▶ les exemples[†] du cluster E_k sont « **similaires** » entre eux,
- ▶ et aussi **dissimilaires** que possibles de ceux **des autres clusters**.

Définition : dispersion

On mesure (souvent) la **dispersion** du cluster E_k par

$$S_k = \left(\frac{1}{|\pi_k|} \sum_{i \in \pi_k} \|x_i - \bar{x}_k\|^q \right)^{\frac{1}{q}},$$

avec q un réel positif à choisir[†].

Interprétation. Plus S_k est petit, plus le cluster est **homogène**.

[†] Le polycopié de P.-H. Cournède et scikit-learn utilisent $q = 1$.

Indice de Davies-Bouldin

Définition : similarité des clusters E_k et $E_{k'}$

$$R_{k,k'} = \frac{S_k + S_{k'}}{\|\bar{x}_k - \bar{x}_{k'}\|}, \quad 1 \leq k, k' \leq K, \quad k \neq k'.$$

Interprétation. Les clusters sont d'autant plus dissimilaires que leurs dispersions sont petites comparées à la distance entre les centroïdes.

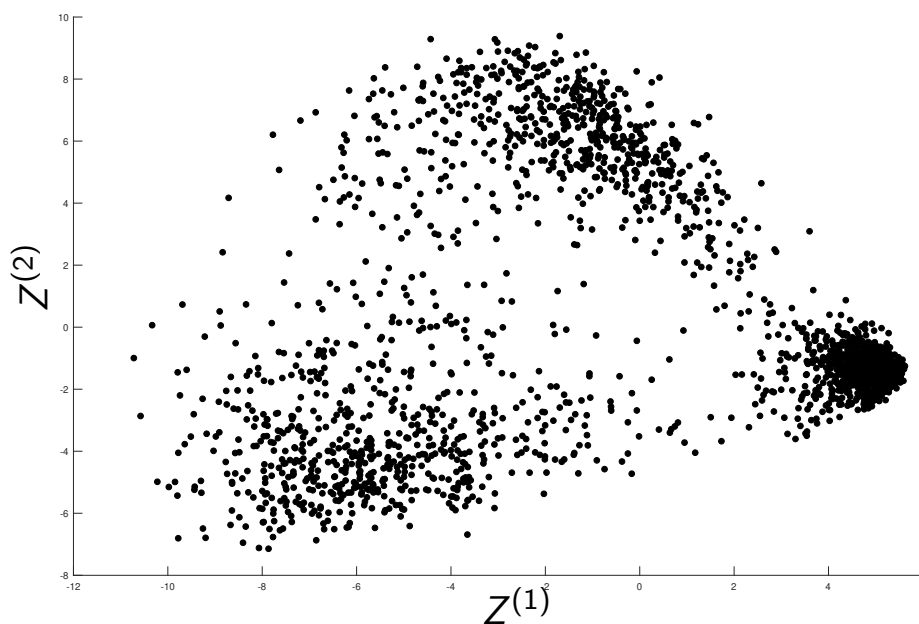
Définition : indice de Davies-Bouldin d'une partition

$$DB = \frac{1}{K} \sum_{k=1}^K \max_{k' \neq k} R_{k,k'}$$

Utilisation : on choisit K de façon à **minimiser DB**.

42/53

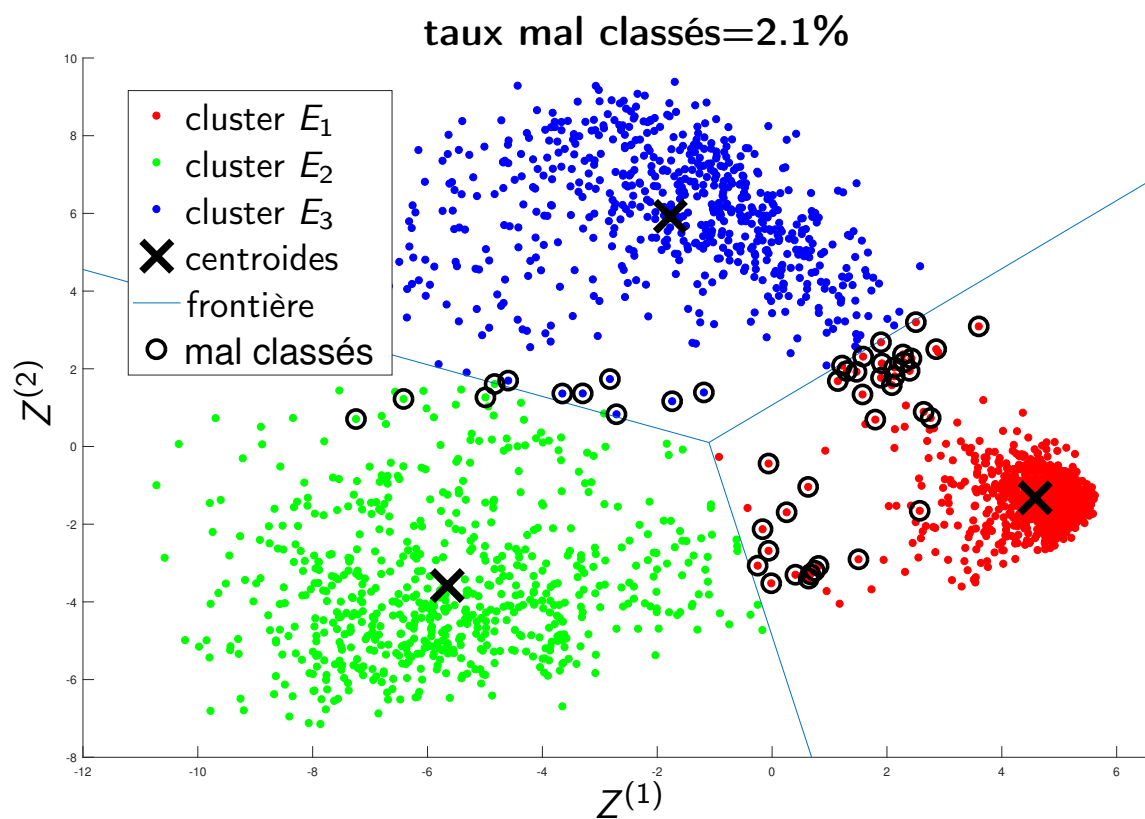
Exemple « Code postal » avec les chiffres 1, 6 et 9



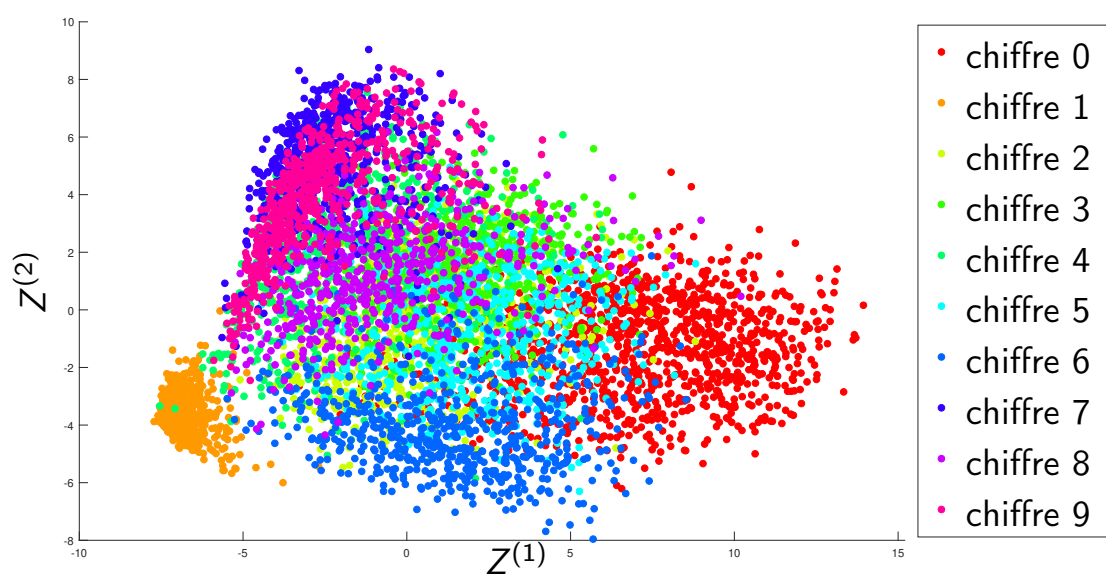
K	2	3	4	5	6	7	8
$DB(K)$	0.76	0.42	0.77	0.89	0.76	0.77	0.79

43/53

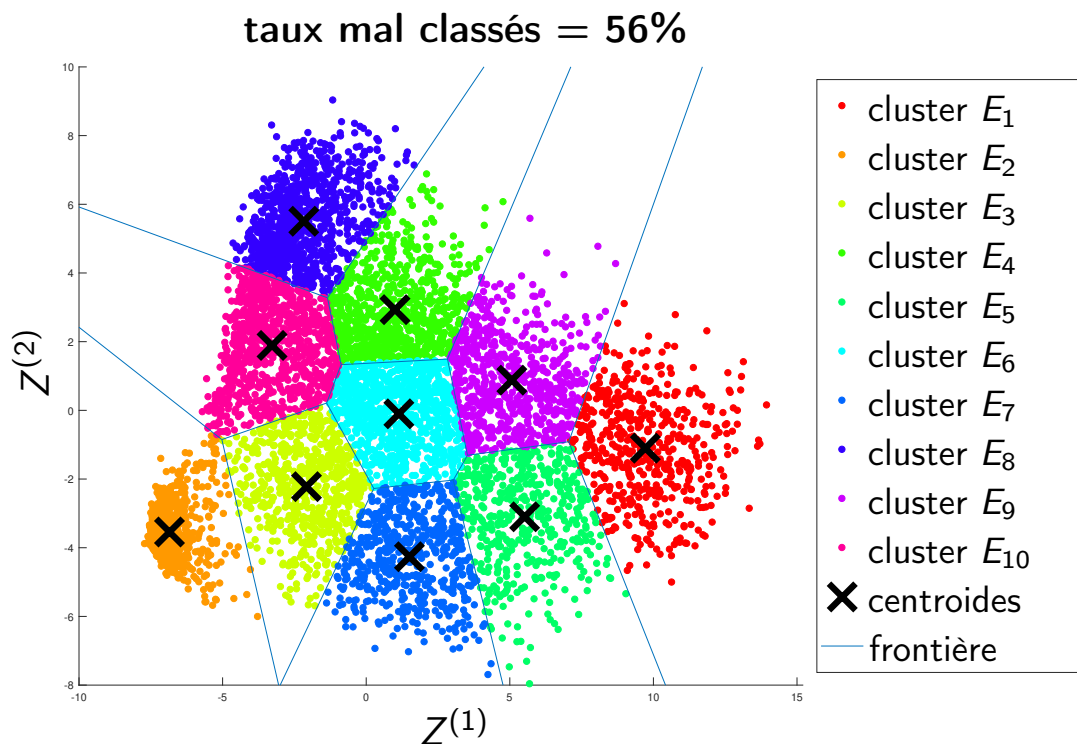
Exemple « Code postal » avec les chiffres 1, 6 et 9



Exemple « Code postal » avec les chiffres de 0 à 9



Exemple « Code postal » avec les chiffres de 0 à 9



46/53

Exemple « Code postal » avec les chiffres de 0 à 9

	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	total
« 0 »	498	0	22	6	260	82	64	0	262	0	1194
« 1 »	0	1000	4	0	0	0	0	0	0	1	1005
« 2 »	3	1	234	122	12	202	54	3	60	40	731
« 3 »	1	0	29	230	4	211	5	5	131	42	658
« 4 »	0	21	70	112	2	42	3	144	19	239	652
« 5 »	2	0	61	37	66	171	88	1	119	11	556
« 6 »	3	6	135	0	128	43	335	0	10	4	664
« 7 »	0	2	2	49	0	6	0	458	1	127	645
« 8 »	2	7	82	138	1	93	1	17	41	160	542
« 9 »	0	10	0	64	0	3	0	303	7	257	644
total	509	1047	639	758	473	853	550	931	650	881	7291

Peu concluant ➡ besoin d'une meilleure mesure de dissimilarité !
(et, en particulier, d'une *meilleure représentation*)

47/53

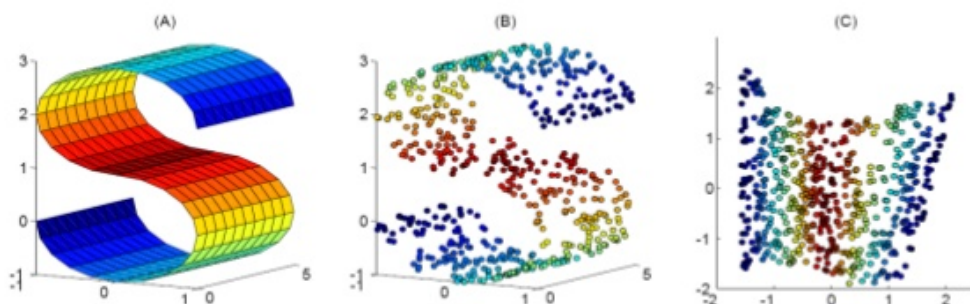
Plan du cours

- 1 – Introduction à l'apprentissage non supervisé
- 2 – Analyse en composantes principales
- 3 – Clustering
- 4 – Un avant-goût de quelques méthodes plus avancées
- 5 – Annexes

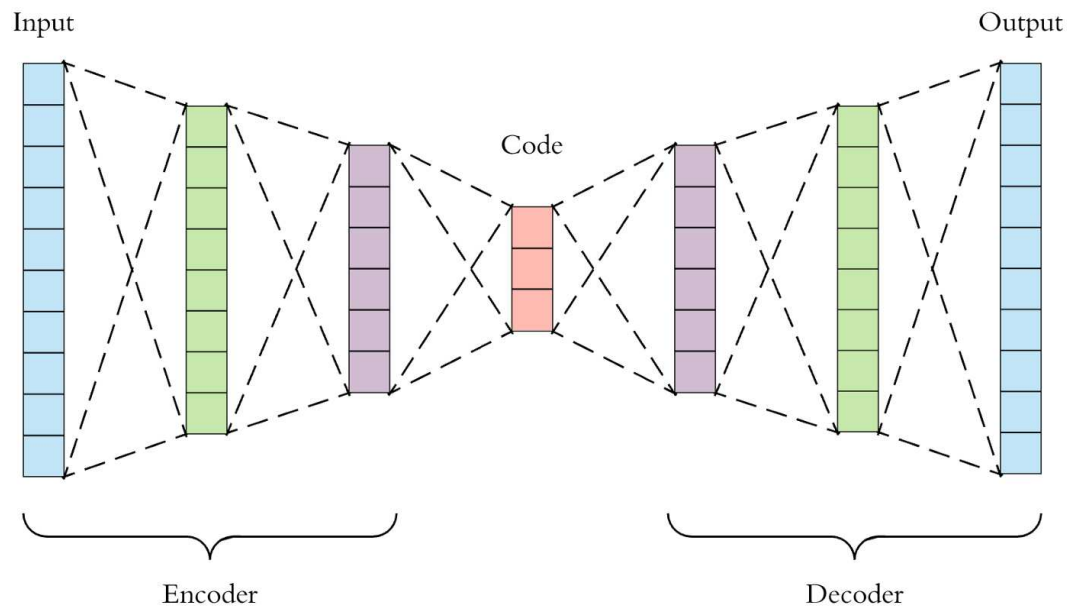
Réduction de dimension non-linéaire

Nonlinear Dimensionality Reduction

- Many data sets contain essential nonlinear structures that invisible to PCA.



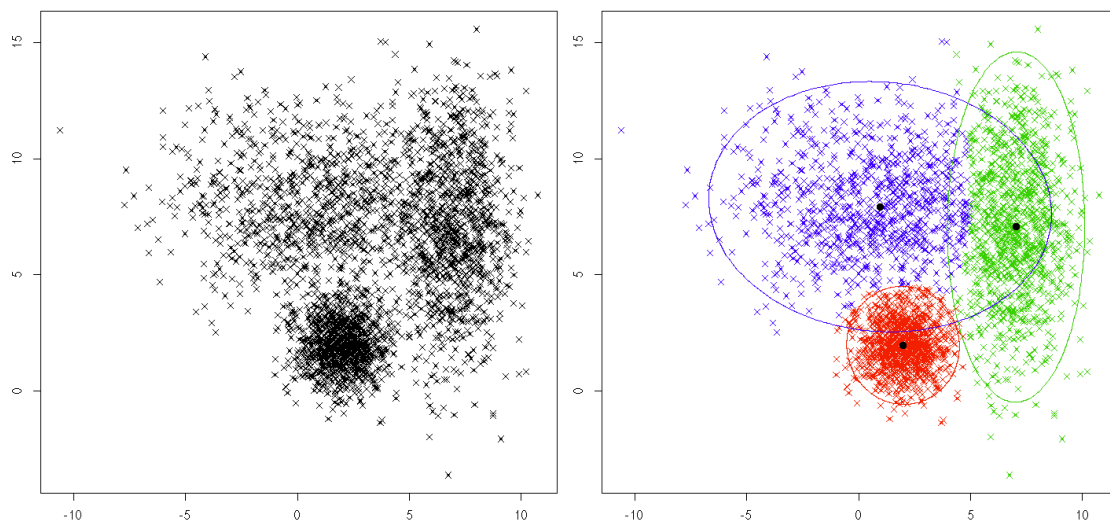
Exemple : auto-encodeur



source : <https://towardsdatascience.com, Applied Data Deep Learning Part 3>

49/53

Clustering fondé sur les modèles de mélange



source : bioinfo-fr.net

50/53

Plan du cours

1 – Introduction à l'apprentissage non supervisé

2 – Analyse en composantes principales

3 – Clustering

4 – Un avant-goût de quelques méthodes plus avancées

5 – Annexes

5.1 – Démonstration du théorème fondamental de l'ACP

5.2 – Expressions de T et $W(\Pi)$ pour $d_{ii'} = \|x_i - x_{i'}\|^2$

5.3 – Coefficient de silhouette d'une partition

Plan du cours

1 – Introduction à l'apprentissage non supervisé

2 – Analyse en composantes principales

3 – Clustering

4 – Un avant-goût de quelques méthodes plus avancées

5 – Annexes

5.1 – Démonstration du théorème fondamental de l'ACP

5.2 – Expressions de T et $W(\Pi)$ pour $d_{ii'} = \|x_i - x_{i'}\|^2$

5.3 – Coefficient de silhouette d'une partition

Démonstration du théorème fondamental de l'ACP

$$\|(\text{Id}_p - AA^\top) X^\top\|_F^2 = \|VD^\top U^\top - AA^\top VD^\top U^\top\|_F^2$$

Propriété de la norme de Frobenius : si U et V sont orthogonales,

$$\|VMU^\top\|_F^2 = \|M\|_F^2.$$

$$\text{D'où : } \|(\text{Id}_p - AA^\top) X^\top\|_F^2 = \|D^\top - V^\top AA^\top VD^\top\|_F^2.$$

Soit $\mathcal{M}_{n,p,q}$ l'ensemble des matrices de taille $n \times p$ et de rang q . Alors

$$D_q = \text{diag}(d_1, \dots, d_q, 0, \dots, 0) \in \operatorname{argmin}_{M \in \mathcal{M}_{n,p,q}} \|D^\top - M^\top\|_F^2$$

(matrice diagonale des q plus grandes valeurs singulières).

On obtient le résultat en vérifiant que $V^\top V_q V_q^\top VD^\top = D_q^\top$. □

[retour au slide 16](#)

51/53

Plan du cours

1 – Introduction à l'apprentissage non supervisé

2 – Analyse en composantes principales

3 – Clustering

4 – Un avant-goût de quelques méthodes plus avancées

5 – Annexes

5.1 – Démonstration du théorème fondamental de l'ACP


5.2 – Expressions de T et $W(\Pi)$ pour $d_{ii'} = \|x_i - x_{i'}\|^2$

5.3 – Coefficient de silhouette d'une partition

Expressions de T et $W(\Pi)$ pour $d_{ij'} = \|x_i - x_{i'}\|^2$

$$\begin{aligned} T &= \frac{1}{2} \sum_{i,i'} \|x_i - x_{i'}\|^2 \\ &= \frac{1}{2} \sum_{i,i'} \|(x_i - \bar{x}) - (x_{i'} - \bar{x})\|^2 \\ &= n \sum_i \|x_i - \bar{x}\|^2 - \sum_{i,i'} (x_i - \bar{x})^\top (x_{i'} - \bar{x}) \\ &= n \sum_i \|x_i - \bar{x}\|^2 \end{aligned}$$

$$\begin{aligned} W(\Pi) &= \frac{1}{2} \sum_k \sum_{i,i' \in \pi_k} \|x_i - x_{i'}\|^2 \\ &= \frac{1}{2} \sum_k \sum_{i,i' \in \pi_k} \|(x_i - \bar{x}_k) - (x_{i'} - \bar{x}_k)\|^2 \\ &= \sum_k n_k \sum_{i \in \pi_k} \|x_i - \bar{x}_k\|^2 \end{aligned}$$

 retour au slide 34

52/53

Plan du cours

1 – Introduction à l'apprentissage non supervisé

2 – Analyse en composantes principales

3 – Clustering

4 – Un avant-goût de quelques méthodes plus avancées

5 – Annexes

5.1 – Démonstration du théorème fondamental de l'ACP

5.2 – Expressions de T et $W(\Pi)$ pour $d_{ij'} = \|x_i - x_{i'}\|^2$

5.3 – Coefficient de silhouette d'une partition

Coefficient de silhouette d'une partition

Autre indicateur de la cohérence globale d'une partition Π .

Soit $i \in \pi_k$. Pour le point x_i , on définit :

- ▶ $a(x_i)$: moyenne des distances aux autres points du cluster π_k
- ▶ $b(x_i)$: minimum de cette moyenne si x_i appartenait à un autre cluster

$$a(x_i) = \frac{1}{|\pi_k|} \sum_{i' \in \pi_k} \|x_{i'} - x_i\|$$
$$b(x_i) = \min_{k' \neq k} \left(\frac{1}{|\pi_{k'}|} \sum_{i' \in \pi_{k'}} \|x_{i'} - x_i\| \right)$$

Interprétation : $a(x_i) \ll b(x_i)$ si les clusters sont homogènes et bien séparés.

Coefficient de silhouette de la partition Π

$$s(\Pi) = \frac{1}{n} \sum_{i=1}^n \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))}$$

Choix du nombre K de clusters :

$\forall \Pi, s(\Pi) \leq 1$ et on choisit la partition tel que $s(\Pi)$ est maximum.