

Classification automatique : k -means

Séance 7 – Décomposer l'inertie

Christophe Ambroise 

UEVE, UMR CNRS 8071

4 septembre 2026

1

D'où vient k -means

2

Retour à la séance 1

Deux usages du même tableau, déjà vus : avec les étiquettes, prédire ; sans les étiquettes, **chercher des groupes**.

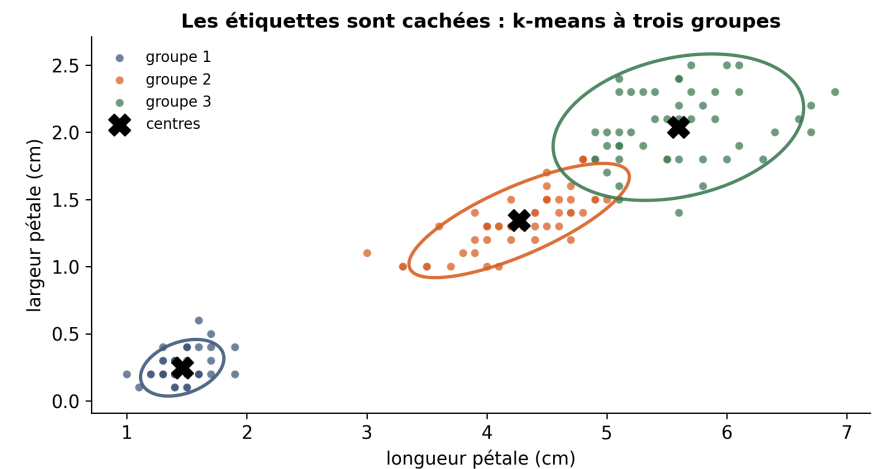


Figure 1

3

Apprendre sans étiquette

Rien à prédire (pas de variable réponse), pas de vérité terrain

Recherche de groupes . . .

Sans étiquette, le **critère devient le modèle** : changer de critère, c'est changer de définition du mot « groupe ».

4

Un modèle génératif

Supposons que chaque observation soit tirée en deux temps : d'abord un groupe $Z \in \{1, \dots, K\}$, puis une position autour du centre de ce groupe.

$$f(x) = \sum_{k=1}^K \pi_k \varphi(x; \mu_k, \Sigma_k), \quad \sum_k \pi_k = 1.$$

C'est un **mélange gaussien**. Les π_k, μ_k, Σ_k s'estiment par l'algorithme **EM**, en deux étapes alternées :

E — espérance

M — maximisation

$$\gamma_{ik} = \frac{\pi_k \varphi(x_i; \mu_k, \Sigma_k)}{\sum_{\ell} \pi_{\ell} \varphi(x_i; \mu_{\ell}, \Sigma_{\ell})} \quad \mu_k = \frac{\sum_i \gamma_{ik} x_i}{\sum_i \gamma_{ik}}$$

La probabilité que i vienne du groupe k .

Des moyennes **pondérées**.

5

k -means en est la limite

Prenons $\Sigma_k = \sigma^2 \mathbf{I}$ pour tout k , puis faisons $\sigma^2 \rightarrow 0$:

$$\gamma_{ik} = \frac{\pi_k \exp(-\|x_i - \mu_k\|^2 / 2\sigma^2)}{\sum_{\ell} \pi_{\ell} \exp(-\|x_i - \mu_{\ell}\|^2 / 2\sigma^2)} \xrightarrow{\sigma^2 \rightarrow 0} \mathbb{1}\{k = \arg \min_{\ell} \|x_i - \mu_{\ell}\|^2\}$$

Les *responsabilités* deviennent des **0 et des 1** : l'étape E se réduit à l'affectation au centre le plus proche, et l'étape M à la moyenne du groupe.

6

Vérification sur Iris

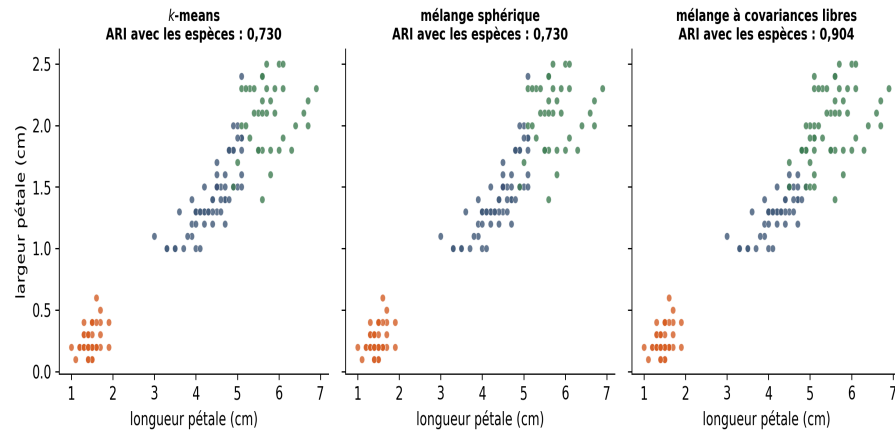


Figure 2

Le mélange sphérique, proportions égales, retrouve **exactement** la partition de k -means (indice de Rand ajusté de 1,000 entre les deux). Libérer les covariances change tout : le mélange général colle bien mieux aux espèces (0,904 contre 0,730).

Le problème

Une partition

$$\{1, \dots, n\} = C_1 \dot{\cup} C_2 \dot{\cup} \dots \dot{\cup} C_K, \quad n_k = |C_k|.$$

En notation matricielle, la **matrice d'appartenance** $\mathbf{C} \in \{0, 1\}^{n \times K}$ avec $C_{ik} = 1$ si $i \in C_k$. Les centres s'écrivent alors en une ligne :

$$\hat{\boldsymbol{\mu}}_k = \frac{\mathbf{C}^\top \mathbf{X}}{\text{diag}(\mathbf{C}^\top \mathbf{C})}.$$

Chercher exhaustivement ? Non

Le nombre de partitions de n objets est le **nombre de Bell** B_n :

$$B_{n+1} = \sum_{k=0}^n \binom{n}{k} B_k, \quad B_1 = 1, B_2 = 2, B_3 = 5, B_4 =$$

Pour les 150 fleurs d'Iris : $B_{150} \approx 10^{192}$.

Même en fixant $K = 3$, il reste environ 10^{70} partitions.

Le problème est **NP-difficile**. Tout algorithme praticable est heuristique — k -means en est un.

10

Le critère

$$I_W = \sum_{k=1}^K \sum_{i \in C_k} \|x_i - \hat{\mu}_k\|^2$$

L'**inertie intra-classe** : la somme des carrés des écarts de chaque point au centre de son groupe.

Petite quand les groupes sont **compacts**. C'est la quantité que k -means minimise, et rien d'autre.

11

L'inertie totale

$$I_T = \sum_{i=1}^n \|x_i - \hat{\mu}\|^2, \quad \hat{\mu} = \frac{1}{n} \sum_i x_i.$$

Elle ne dépend **pas** de la partition : c'est une propriété du nuage.

Et l'**inertie inter-classes**, qui mesure l'écartement des centres :

$$I_B = \sum_{k=1}^K n_k \|\hat{\mu}_k - \hat{\mu}\|^2.$$

12

Pourquoi la distance euclidienne

L'inertie est une somme de **carrés** euclidiens. Ce n'est pas neutre :

- le centre optimal d'un groupe est alors sa **moyenne**, calculable en une ligne — c'est ce qui rend l'algorithme praticable ;
- les carrés donnent un poids démesuré aux points éloignés.

Avec la distance ℓ_1 , le centre optimal devient la **médiane** : c'est k -médianes, plus robuste. Avec une distance quelconque et des centres pris parmi les observations : k -médoides.

13

La décomposition de l'inertie

14

Le théorème

Décomposition de l'inertie (Huygens)

Pour **toute** partition en K groupes,

$$\underbrace{\sum_i \|x_i - \hat{\mu}\|^2}_{I_T = \text{total}} = \underbrace{\sum_k \sum_{i \in C_k} \|x_i - \hat{\mu}_k\|^2}_{I_W = \text{intra}} + \underbrace{\sum_k n_k \|\hat{\mu}_k - \hat{\mu}\|^2}_{I_B = \text{inter}}$$

I_T est fixe. Donc **minimiser I_W , c'est exactement maximiser I_B .**

15

La preuve, en deux temps

Étape 1. En développant $\|x_i - \hat{\mu}_k\|^2$ et en utilisant $\sum_{i \in C_k} x_i = n_k \hat{\mu}_k$:

$$\sum_k \sum_{i \in C_k} \|x_i - \hat{\mu}_k\|^2 = \sum_i \|x_i\|^2 - \sum_k n_k \|\hat{\mu}_k\|^2.$$

Étape 2. De même, avec $\sum_k n_k \hat{\mu}_k = n \hat{\mu}$:

$$\sum_k n_k \|\hat{\mu}_k - \hat{\mu}\|^2 = \sum_k n_k \|\hat{\mu}_k\|^2 - n \|\hat{\mu}\|^2.$$

Conclusion. En sommant, les termes $\sum_k n_k \|\hat{\mu}_k\|^2$ s'annulent :

$$I_W + I_B = \sum_i \|x_i\|^2 - n \|\hat{\mu}\|^2 = I_T. \quad \square$$

16

En image

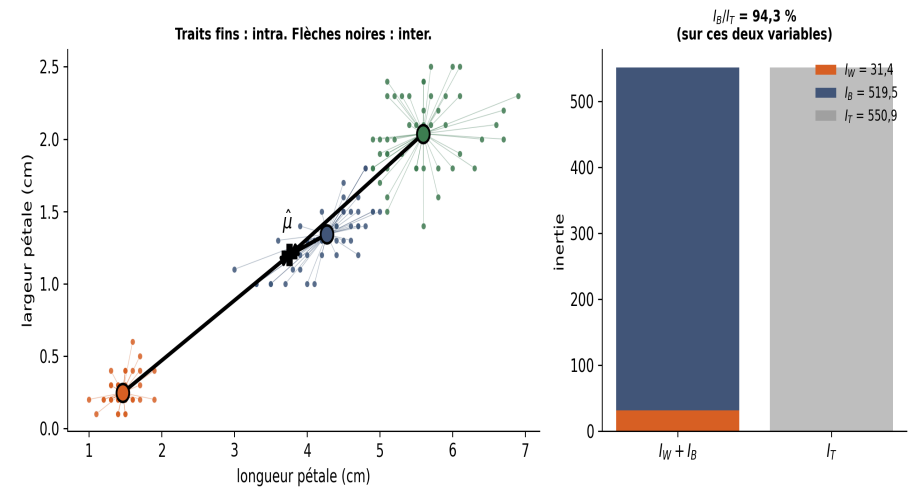


Figure 3

17

Théorème de la variance totale

En probabilité

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X | Z)] + \text{Var}(\mathbb{E}[X | Z])$$

Variance intra, puis variance des moyennes conditionnelles.

Sur un nuage de points

$$\frac{I_T}{n} = \frac{I_W}{n} + \frac{I_B}{n}$$

Z est l'étiquette de groupe, de loi $P(Z = k) = n_k/n$.

La même identité, lue deux fois : une décomposition de la variance selon un conditionnement. Ici, le conditionnement

18

Terme à terme

Probabilité	Nuage de points
$\text{Var}(X)$	I_T/n – inertie totale
$\mathbb{E}[\text{Var}(X Z)]$	I_W/n – inertie intra
$\text{Var}(\mathbb{E}[X Z])$	I_B/n – inertie inter
$P(Z = k)$	n_k/n
$\mathbb{E}[X Z = k]$	$\hat{\mu}_k$

Chercher une partition, c'est chercher **la variable latente Z qui explique le plus de variance**.

19

Une mesure de qualité

$$R^2 = \frac{I_B}{I_T} = 1 - \frac{I_W}{I_T} \in [0, 1].$$

C'est exactement le R^2 d'une analyse de variance à un facteur – la part de l'inertie **expliquée** par la partition.

Sur Iris (quatre variables), la partition de k -means en trois groupes atteint $R^2 = 88,4\%$.

20

Une conséquence immédiate

Soit une partition en K groupes, et la partition obtenue en **coupant** l'un des groupes en deux. Alors I_W ne peut que **diminuer** : les nouveaux centres sont au moins aussi proches de leurs points.

$K \mapsto I_W(K)$ est décroissante, de I_T en $K = 1$ à 0 en $K = n$.

Le critère ne peut donc **jamais** servir à choisir K : son minimum est toujours $K = n$.

21

Le même critère, chez l'ACP

La séance 6 décompose la **même** inertie totale, mais selon un sous-espace et non selon une partition :

$$I_T = \underbrace{\sum_{j \leq q} \lambda_j}_{\text{inertie projetée}} + \underbrace{\sum_{j > q} \lambda_j}_{\text{inertie résiduelle}} .$$

C'est le **théorème de Pythagore** : le vecteur $x_i - \hat{\mu}$ se décompose en sa projection sur le sous-espace et sa composante orthogonale.

22

Deux découpages du même total

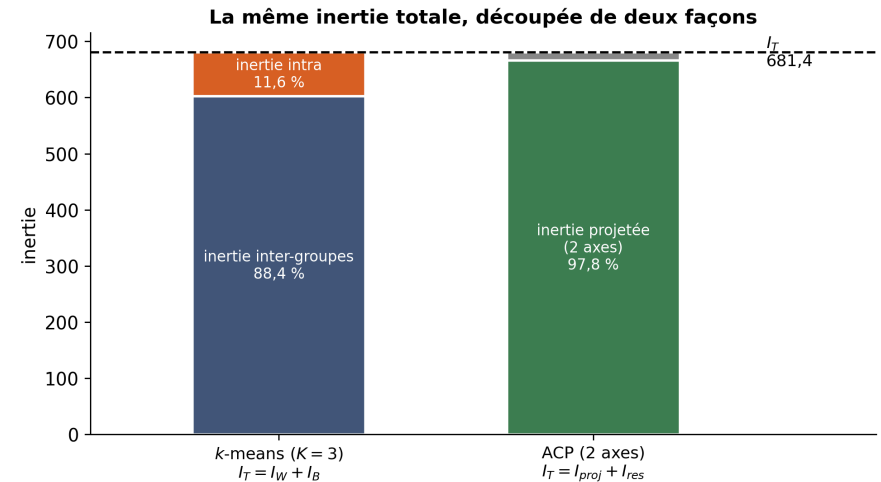


Figure 4

23

Ce que la figure dit

Sur les quatre variables d'Iris, $I_T = 681,4$ dans les deux cas.

Méthode	Ce qu'on choisit	Part de I_T retenue
k -means, $K = 3$	une partition	88,4 %
ACP, 2 axes	un sous-espace	97,8 %

Même quantité maximisée, deux familles de solutions. L'ACP **comprime** sans regrouper ; k -means **regroupe** sans comprimer. Elles se composent d'ailleurs très bien : projeter puis classer.

24

L'algorithme

25

Lloyd, en deux étapes

Partant de K centres, répéter jusqu'à stabilisation :

Affectation

$$z_i = \arg \min_k \|x_i - c_k\|^2$$

Chaque point rejoint le centre le plus proche.

Ce sont exactement les étapes E et M du mélange, en version dure.

Recentrage

$$c_k = \frac{1}{n_k} \sum_{i: z_i=k} x_i$$

Chaque centre devient la moyenne de son groupe.

26

Chaque demi-étape fait décroître I_W

Affectation. Chaque point passe à un centre au moins aussi proche : sa contribution ne peut pas augmenter.

Recentrage. Pour tout c , $\sum_{i \in C_k} \|x_i - c\|^2$ est minimisé en $c = \hat{\mu}_k$ – c'est la propriété de la moyenne.

I_W décroît à chaque demi-étape, et le nombre de partitions est **fini** : l'algorithme s'arrête en un nombre fini d'itérations.

27

La décroissance, mesurée

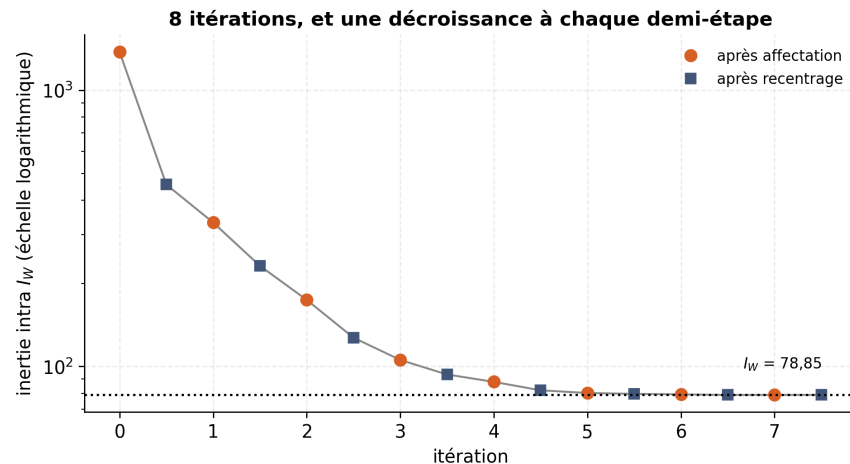


Figure 5

Huit itérations, de $I_W = 1374$ à $78,85$. Décroissance vérifiée à chaque demi-étape.

28

Mais l'optimum est local

Rien ne garantit que le minimum atteint soit **global**. Le résultat dépend de l'initialisation.

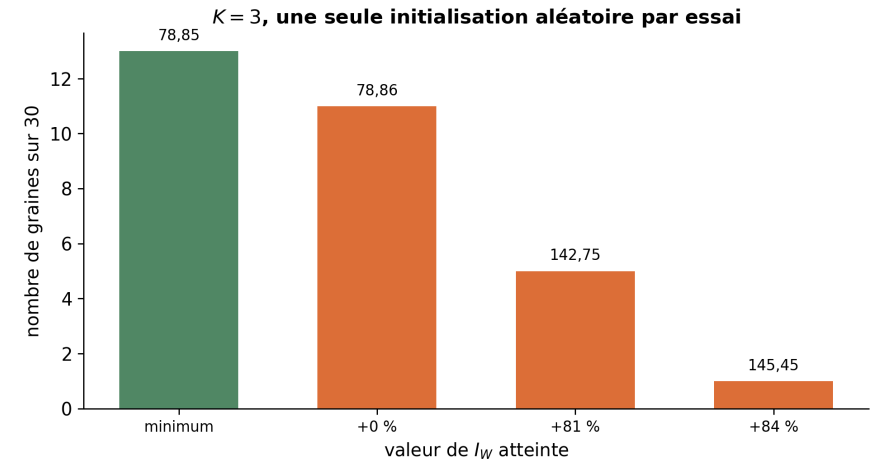


Figure 6

29

Ce que montrent les trente essais

Avec une seule initialisation aléatoire, le minimum global n'est atteint que **13 fois sur 30**. La pire solution est **84 % au-dessus**.

Toujours relancer : `n_init=50` en Python, `nstart=50` en R. Ce n'est pas une option de confort, c'est une condition de validité.

L'initialisation **k-means++** choisit les centres de départ éloignés les uns des autres, et réduit fortement le problème. C'est le défaut de scikit-learn.

30

k-means++

Plutôt que de tirer les centres au hasard uniforme :

1. tirer le premier centre uniformément parmi les observations ;
2. tirer chaque centre suivant avec une probabilité **proportionnelle à $D(x)^2$** , où $D(x)$ est la distance au centre déjà choisi le plus proche.

Les centres de départ sont ainsi écartés les uns des autres.

Garantie : l'espérance de I_W obtenu est au plus $8(\ln K + 2)$ fois l'optimum, **avant même** la première itération de Lloyd.

31

Coût

Une itération coûte $O(nKd)$: $n \times K$ distances en dimension d .

Avec `n_init` relances et T itérations : $O(\text{n_init} \times T \times nKd)$.

Linéaire en n : k -means passe à l'échelle là où une classification hiérarchique, en $O(n^2)$ de mémoire, ne passe plus.

32

Sur Iris

33

Les groupes ne sont pas les espèces

Table 1

groupe	setosa	versicolor	virginica
1	50	0	0
2	0	48	4
3	0	2	46

Setosa est retrouvée exactement. Entre *versicolor* et *virginica*, la frontière de k -means n'est pas celle de la biologie.

34

k -means bat les espèces sur son propre critère

Sur les quatre variables :

Partition	I_W	I_B/I_T
k-means, $K = 3$	78,85	88,4 %
les trois espèces	89,30	86,9 %

La partition en espèces est **moins bonne** que celle de k -means — pour le critère de k -means. Ce n'est pas un paradoxe : la vérité biologique n'a aucune raison de minimiser une somme de carrés.

35

En scikit-learn, en R

```
1 from sklearn.cluster import KMeans
2
3 km = KMeans(n_clusters=3, n_init=50, random_state=0).fit(X)
4 km.labels_      # la partition
5 km.cluster_centers_ # les centres
6 km.inertia_     # I_W, non normalisée

1 km <- kmeans(X, centers = 3, nstart = 50)
2 km$cluster; km$centers
3 km$tot.withinss # I_W
4 km$betweenss / km$totss # I_B / I_T
```

`n_init` / `nstart` **valent 1 par défaut** dans plusieurs versions : les fixer explicitement, toujours.

36

Faut-il standardiser ?

Iris est mesuré tout entier en centimètres, mais les étalements diffèrent d'un facteur quatre — et l'inertie est une somme de carrés.

Standardiser change la partition : **onze fleurs sur 150** changent de groupe (indice de Rand ajusté de 0,80 entre les deux).

Même règle qu'en séances 3 et 5 : sans unité commune, standardiser. Ici les unités **sont** communes, et le choix se discute — il faut le documenter, pas l'automatiser.

37

Choisir K : le coude

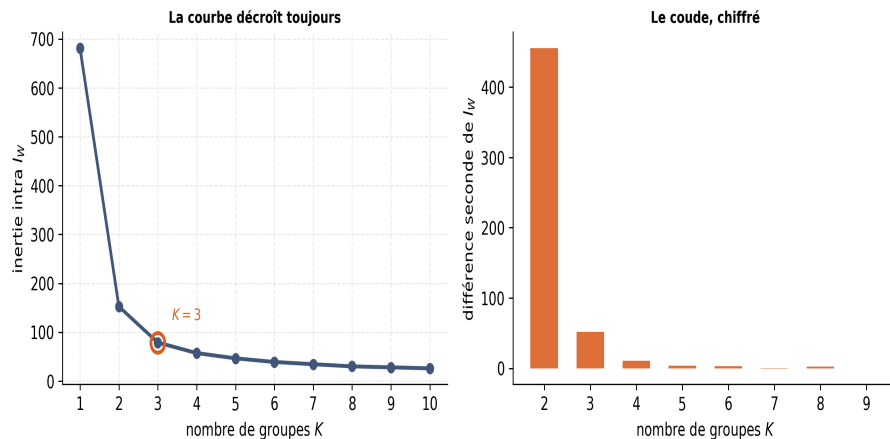


Figure 7

38

Ce que le coude ne dit pas

I_W **décroît toujours** avec K : à $K = n$, chaque point est son propre groupe et $I_W = 0$. Le minimum ne peut donc pas servir de critère.

Le coude cherche la rupture de pente. Sur Iris il désigne $K = 3$ — mais la différence seconde vaut 51,9 en $K = 3$ et 10,8 en $K = 4$: la lecture n'est pas sans appel.

Le coude **informe**, il ne décide pas. Silhouette, gap statistic ou critères de type BIC sur le mélange gaussien apportent d'autres avis — jamais une preuve.

39

Une autre voix : la silhouette

Pour chaque point, s_i compare sa distance moyenne à son propre groupe (a_i) et au groupe voisin le plus proche (b_i) :

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \in [-1, 1].$$

Sur Iris, la silhouette moyenne vaut **0,681 en $K = 2$** contre 0,553 en $K = 3$.

La silhouette dit **deux** groupes, le coude en dit trois, la biologie en compte trois. Deux critères géométriques, deux réponses

40

Projeter puis classer

L'ACP retenant 97,8 % de l'inertie sur deux axes, que se passe-t-il si l'on classe **dans le plan factoriel** plutôt que dans $\mathbb{R}^4$?

La partition est quasi identique : indice de Rand ajusté de **0,980**, et son inertie intra mesurée en dimension 4 vaut 78,86 contre 78,85.

Enchaîner ACP puis k -means ne coûte presque rien ici, et divise la dimension par deux. En grande dimension, c'est la pratique courante — pour la même raison qu'en séance 3 : les distances y redeviennent informatives.

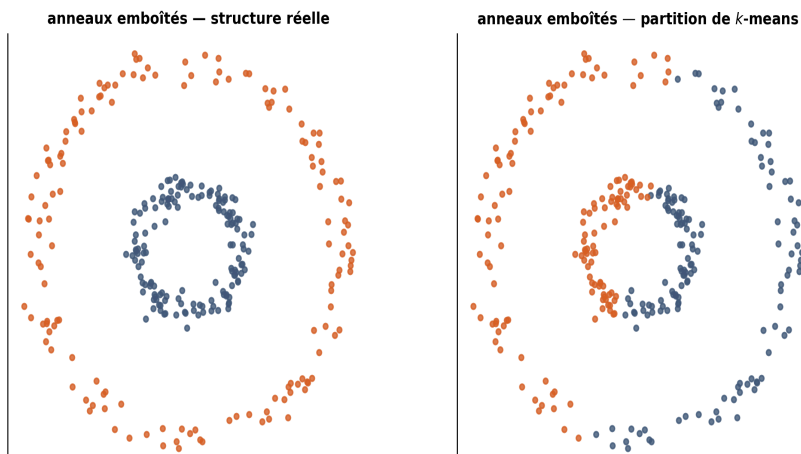
41

Limites

42

Des groupes sphériques, par construction

Minimiser $\sum \|x_i - c_k\|^2$ revient à supposer des groupes **isotropes et de même taille** — c'est l'hypothèse $\Sigma_k = \sigma^2 \mathbf{I}$ du mélange.



43

Deux échecs, deux remèdes

Nuages allongés : k -means coupe en travers. Remède — un mélange gaussien à **covariances libres**, celui de l'ouverture.

Anneaux emboîtés : aucun mélange gaussien n'y arrive. Remède — une méthode fondée sur la **connexité** (classification spectrale, DBSCAN), hors programme.

k -means trouve **toujours** K groupes, même quand il n'y en a aucun. Ce n'est pas un test d'existence de structure.

44

Sept séances, un même schéma

Séance	Méthode	Complexité réglée par
1	polynômes	degré
3	k -PPV	k
4	arbres	α
5	ridge	λ
6	ACP	q axes
7	k -means	K groupes

La différence : en supervisé, la validation croisée tranche. En non supervisé, il n'y a **pas d'étiquette** pour arbitrer — le choix

redevient une décision argumentée.

45

En bref

1. k -means est l'algorithme EM d'un mélange gaussien à covariances $\sigma^2 \mathbf{I}$, dans la limite $\sigma^2 \rightarrow 0$.
2. $I_T = I_W + I_B$ pour **toute** partition : c'est le théorème de la variance totale, et I_T ne dépend pas de la partition.
3. Minimiser l'intra, c'est maximiser l'inter ; I_B/I_T est un R^2 .
4. L'ACP décompose la **même** inertie, mais selon un sous-espace : Pythagore au lieu de Huygens.
5. Lloyd décroît à chaque étape vers un optimum **local** : relancer. Le coude informe sans décider.