Intelligence Artificielle et Données

Réduction de dimension, Partie 1 : ACP

Akka Zemmari

LaBRI, Université de Bordeaux

2024 - 2025

Malédiction de la dimensionnalité

Malédiction de la dimensionnalité = "Curse of Dimensionality" = CD

[1] https://medium.com/@paritosh_30025/curse-of-dimensionality-f4edb3efa6ec

Malédiction de la dimensionnalité

Malédiction de la dimensionnalité = "Curse of Dimensionality" = CD En gros[1] :

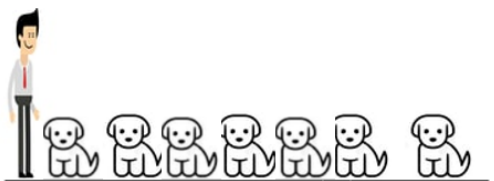


Figure 1 - One-dimension scenario

Malédiction de la dimensionnalité

Malédiction de la dimensionnalité = "Curse of Dimensionality" = CD En gros[1] :

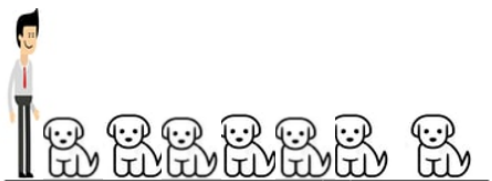


Figure 1 - One-dimension scenario

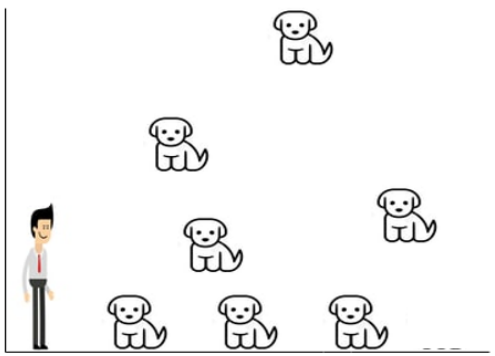


Figure 2- Two-dimension Scenario

Malédiction de la dimensionnalité

Malédiction de la dimensionnalité = "Curse of Dimensionality" = CD En gros[1] :

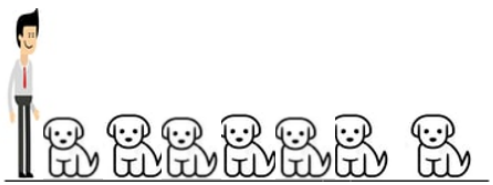


Figure 1 - One-dimension scenario

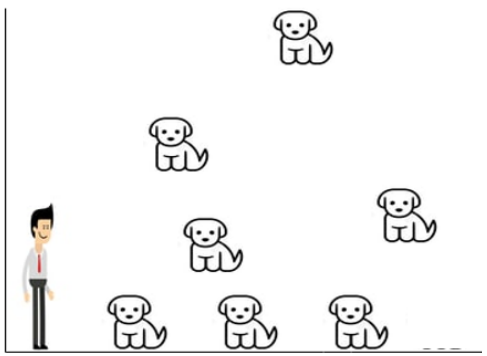


Figure 2- Two-dimension Scenario

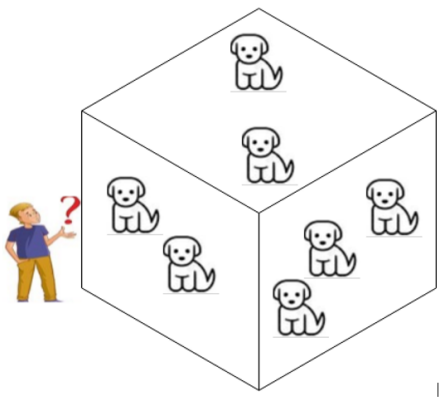


Figure 3 – Three-dimension scenario

[1] https://medium.com/@paritosh_30025/curse-of-dimensionality-f4edb3efa6ec

Malédiction de la dimensionnalité

Malédiction de la dimensionnalité = "Curse of Dimensionality" = CD En gros[1] :

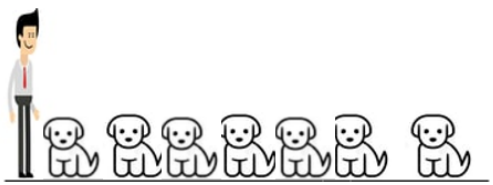


Figure 1 - One-dimension scenario

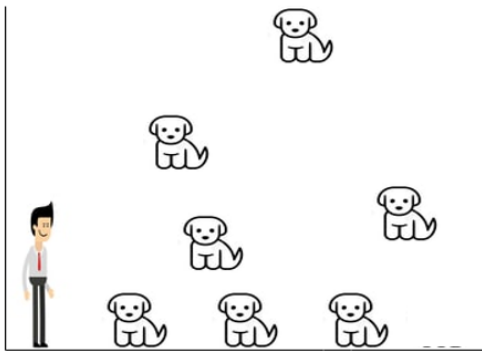


Figure 2- Two-dimension Scenario

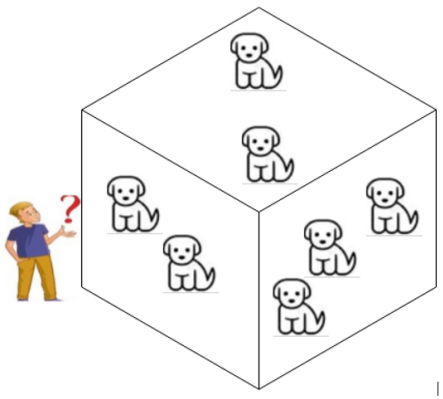


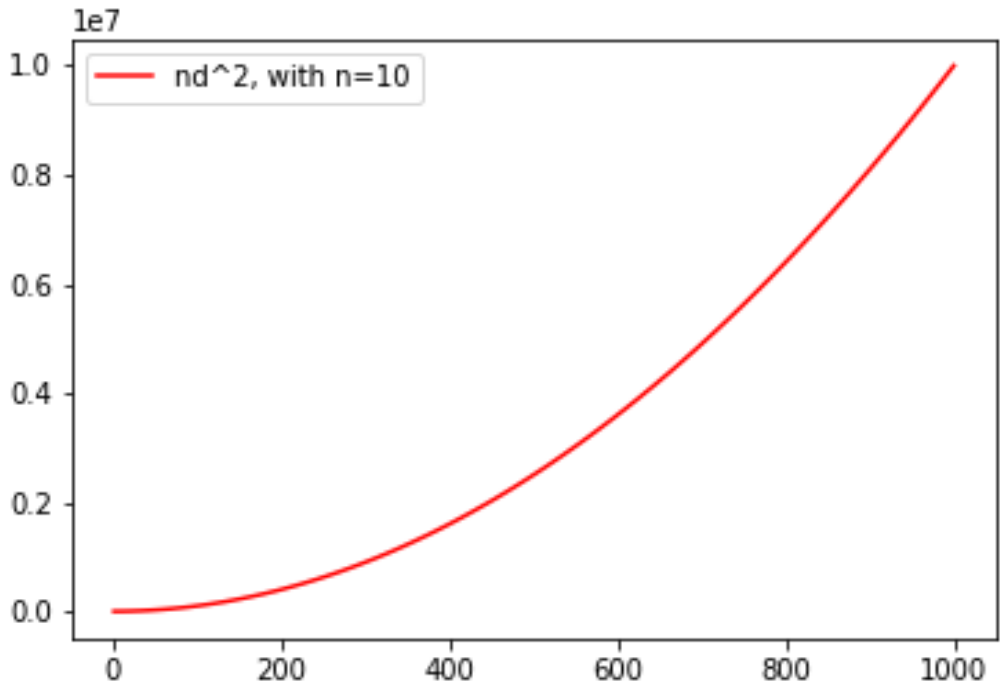
Figure 3 – Three-dimension scenario

Dans la suite : Soit $\mathcal{S}$ un dataset avec d variables (dimension d) et n exemples

[1] https://medium.com/@paritosh_30025/curse-of-dimensionality-f4edb3efa6ec

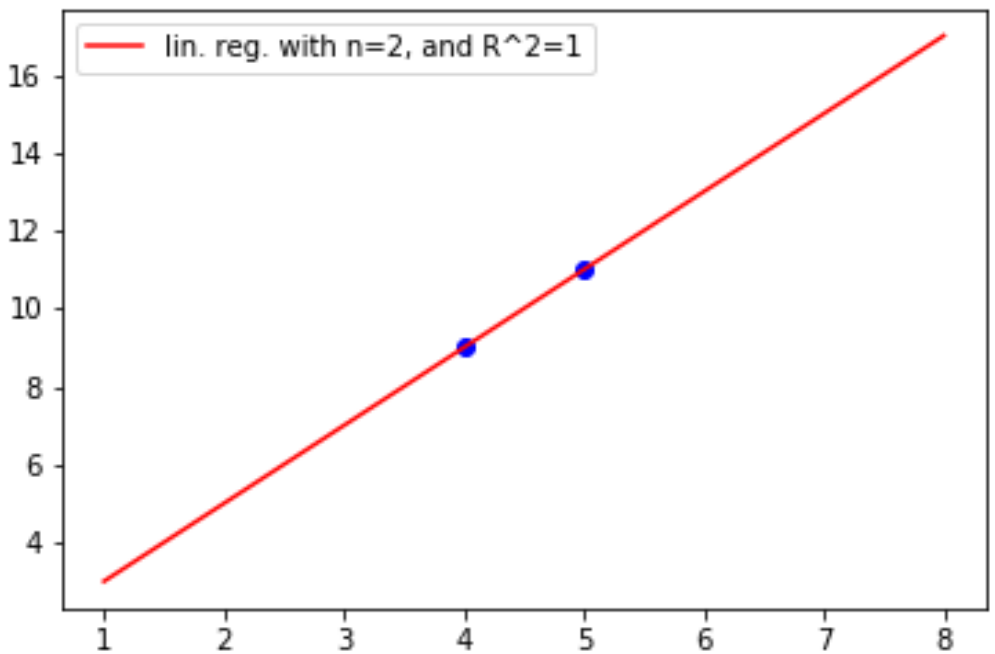
CD et complexité

- La complexité augmente en fonction de la dimension d
- Exemple : on veut estimer la matrice de covariance, la complexité en temps est alors de $O\left(nd^2\right)$



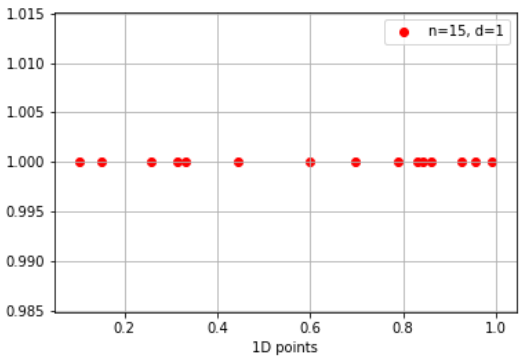
CD et sur-apprentissage (overfitting)

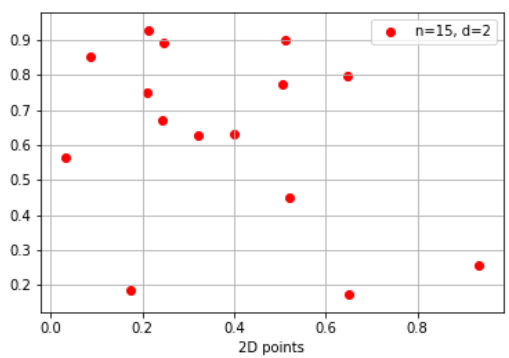
- Si d est grand, n peut être trop petit pour une bonne estimation des paramètres d'un modèle.
- Exemple : retour sur la régression linéaire avec un $R^2 = 1$ mais ....

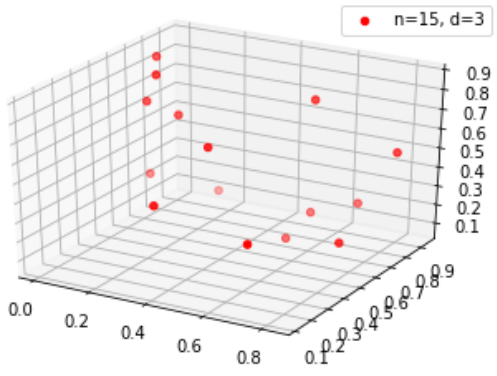


CD et nombre d'exemples

Soit les trois figures suivantes :

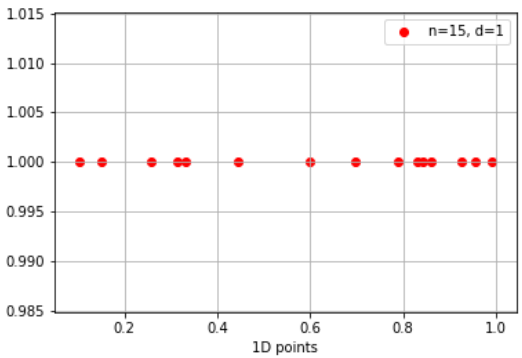


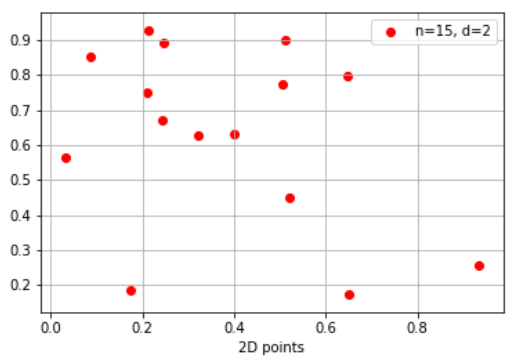


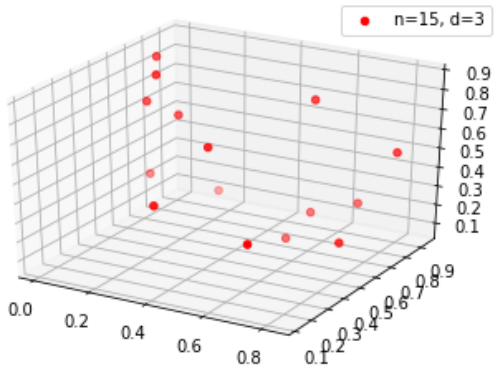


CD et nombre d'exemples

Soit les trois figures suivantes :



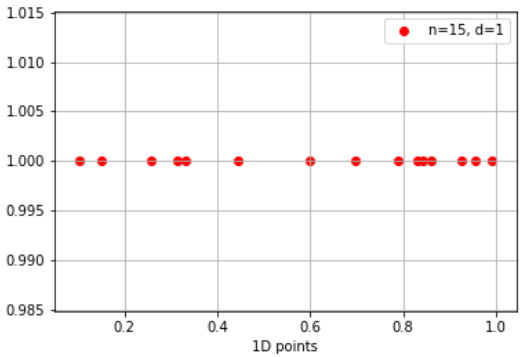


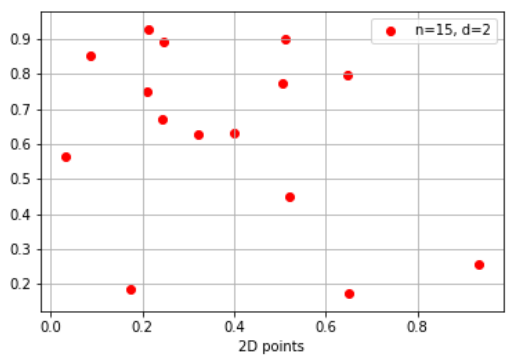


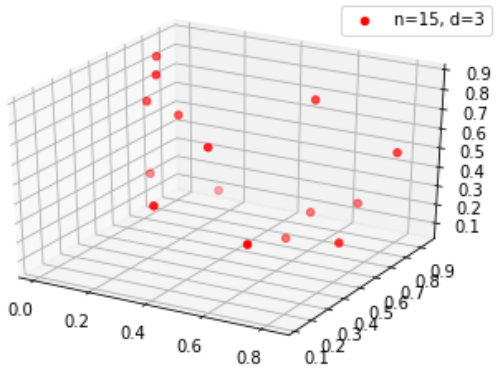
- Que pouvez-vous dire sur la densité des points dans chaque figure ?

CD et nombre d'exemples

Soit les trois figures suivantes :



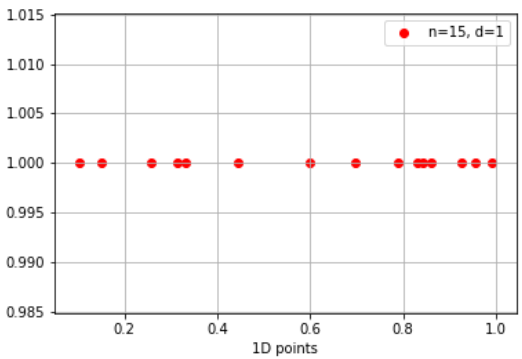


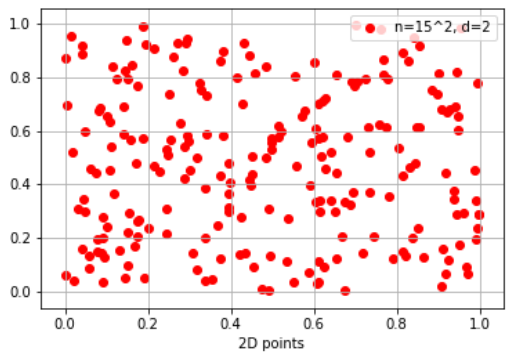


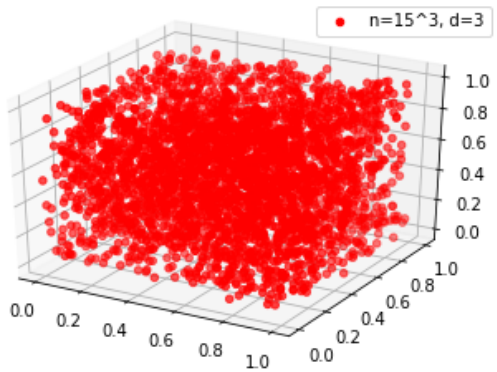
- Que pouvez-vous dire sur la densité des points dans chaque figure ?
- Si vous deviez utiliser un k-nn avec par exemple $k = 1$, quelle dimension auriez-vous choisi ? Pourquoi ?

CD et nombre d'exemples

Si on veut retrouver la même densité que pour $d = 1$:

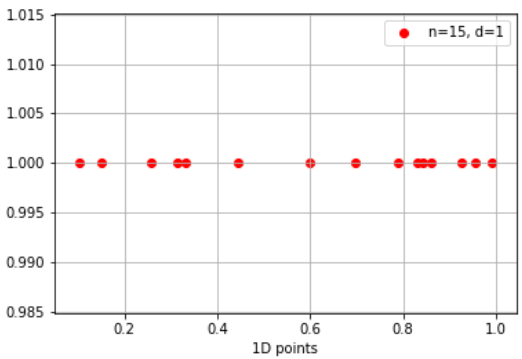


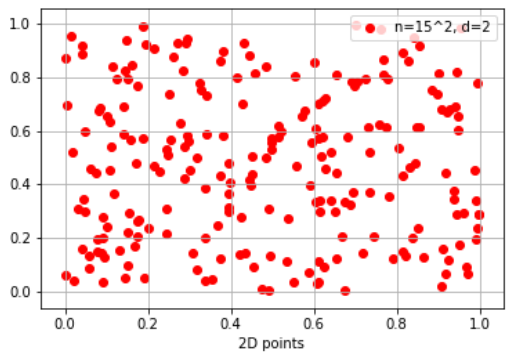


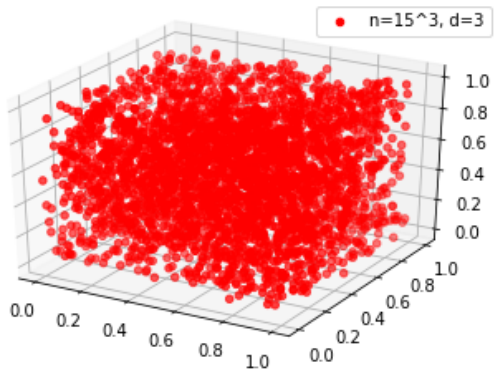


CD et nombre d'exemples

Si on veut retrouver la même densité que pour $d = 1$:



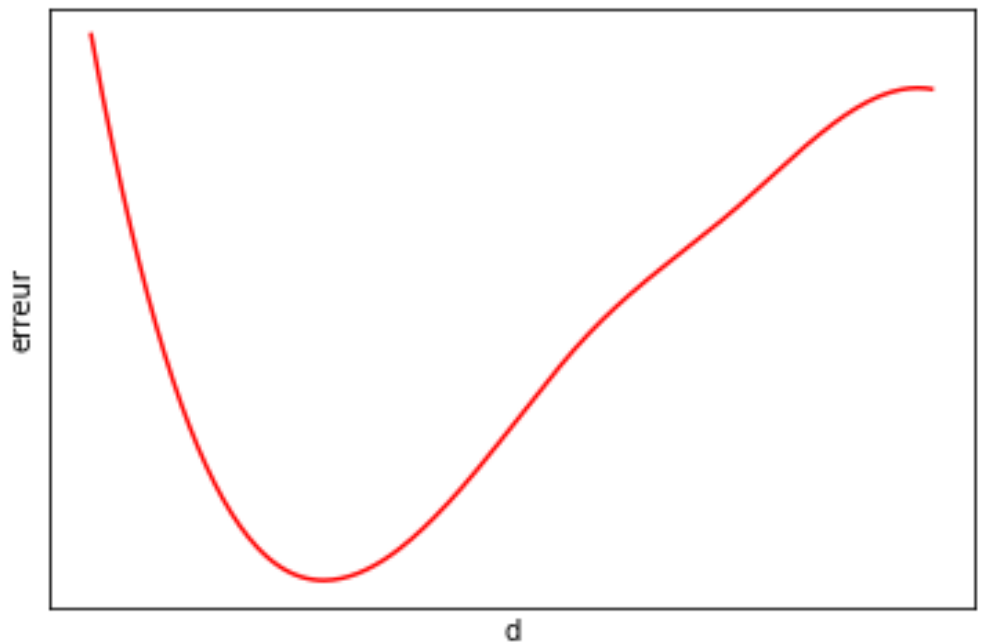




- Il faudrait donc disposer de $n' = n^d$ exemples pour garantir la même densité ...

CD et nombre d'exemples

En fait, si on fixe n (le nombre d'exemples) et qu'on augmente d, la courbe des erreurs (de classification) se comportera comme suit :



Malédiction de la dimensionnalité

Pour résumer ...

- Il faut éviter d'avoir beaucoup de descipteurs, ...
- Dans la pratique : on n'a pas vraiment le choix !

Malédiction de la dimensionnalité

Exemple :

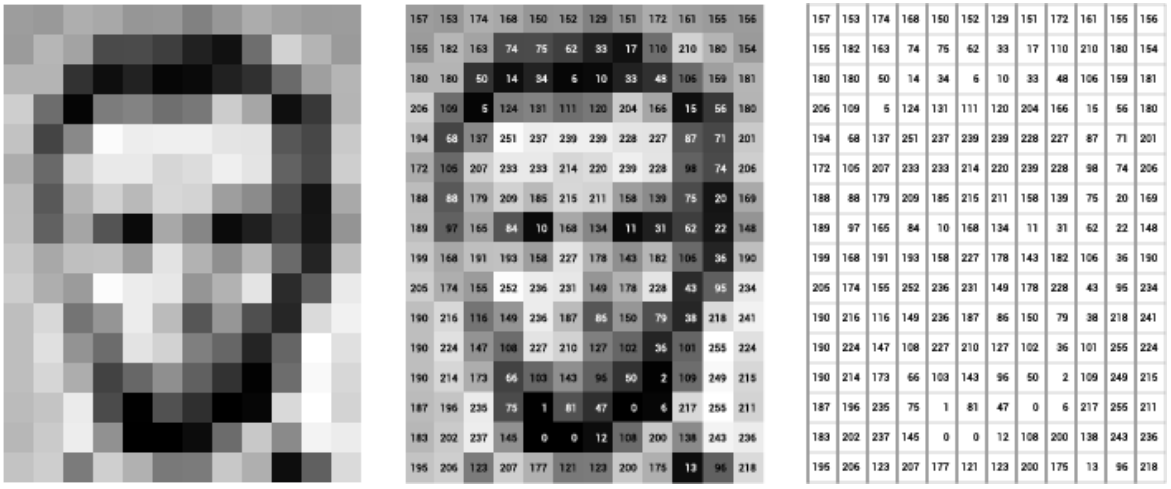


Figure: Une image = un tableau à $m \times n$ pixels

- Avec un pixel = un descipteur, l'image est de dimension $m \times n$. $\Rightarrow$ Dans le cas de la figure : on a $d = 16 \times 12 = 192$
- Ainsi, si un échantillon de taille 10 est suffisant pour $d = 1$, il nous faudrait (juste !!) 10^{192} images.

Malédiction de la dimensionnalité

Remarques :

- Tous les pixels ne sont pas significatifs
- $\rightarrow$ on n'a pas vraiment $m \times n$ descripteurs
- Avec du "feature engineering", on peut réduire la dimension

Réduction de dimension

Pas de "Feature engineering" ici mais plutôt de la combinaison de descipteurs (Feature Combination) :

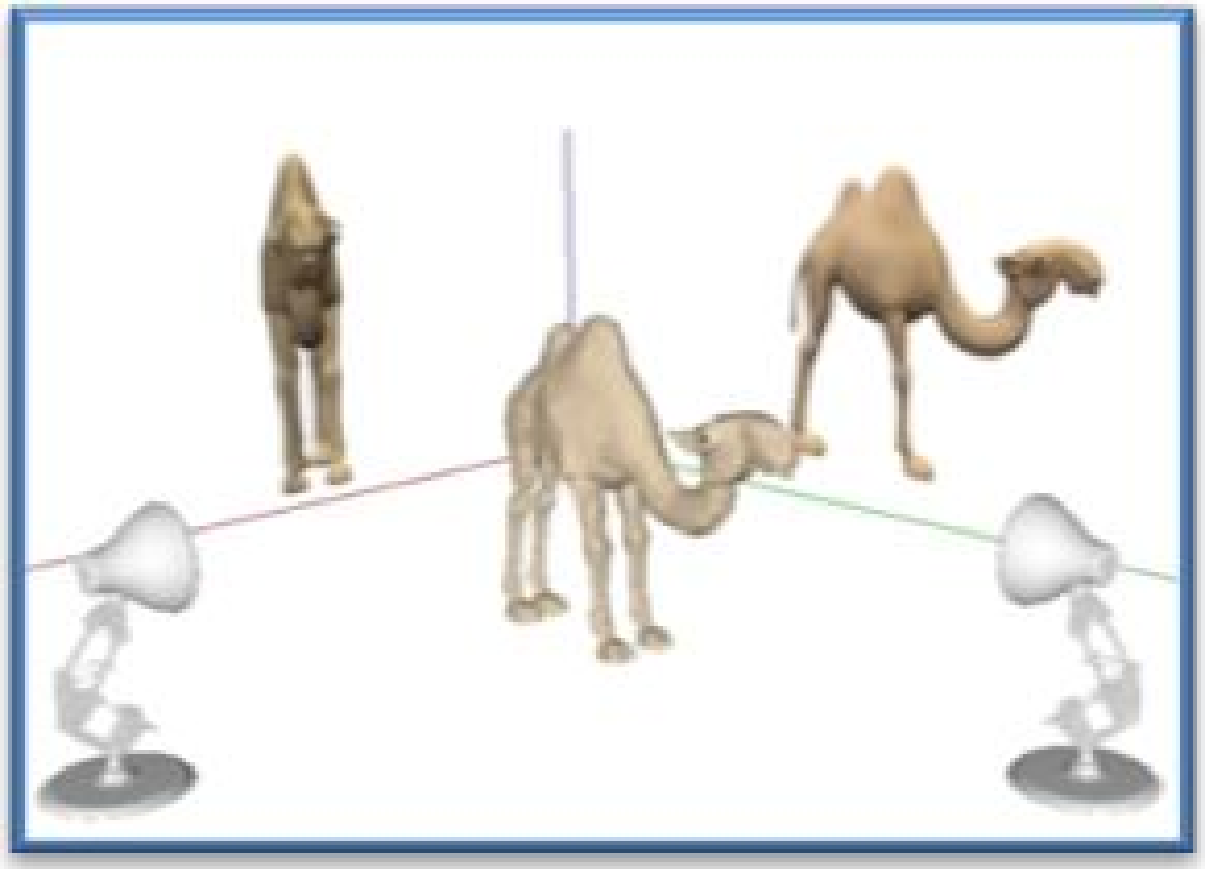
- Analyse en Composantes Principale (ACP)
- Analyse Linéaire Discriminante (LDA).

ACP : idées de base

- Transformer des variables très corrélées en nouvelles variables décorrélées les unes des autres.
- Il s'agit en fait de résumer l'information qui est contenue dans un dataset en un certain nombre de variables synthétiques : les Composantes principales.

ACP : idées de base

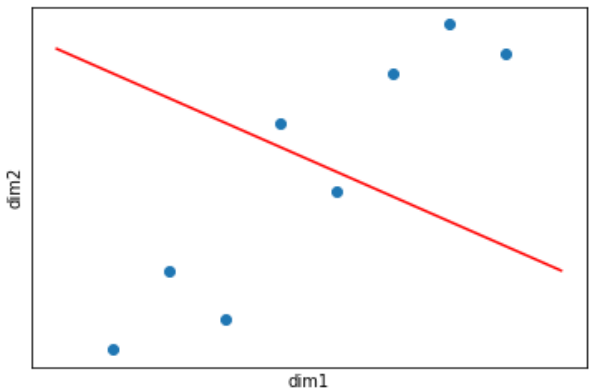
Exemple 1.

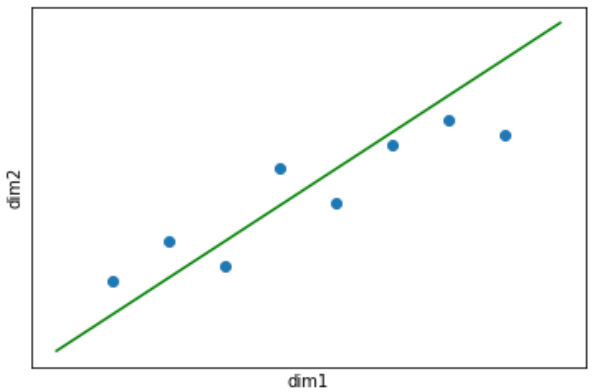


ACP : idées de base

Exemple 2.

Projeter des données 2-D sur une ligne tout en minimisant l'erreur de projection.
Faites votre choix :

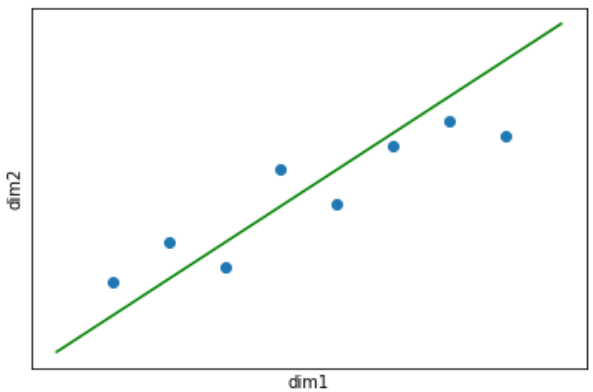




Quelle est la meilleure droite (1-D) ? Expliquer.

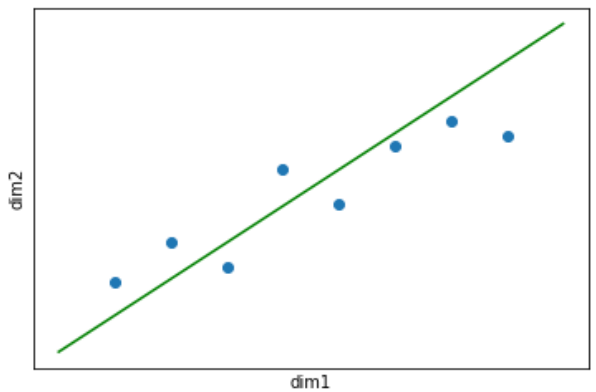
ACP : idées de base

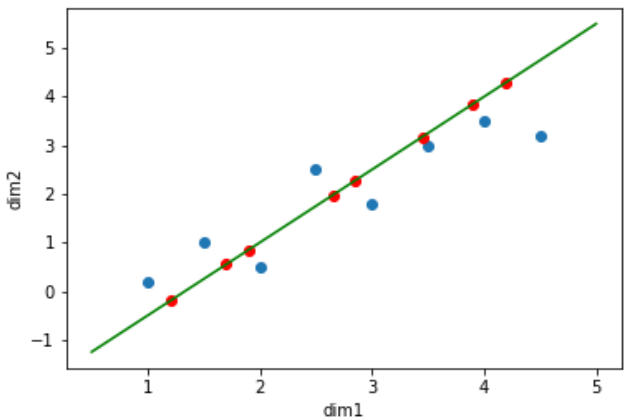
On voit clairement que la meilleure droite est la droite verte. La projection des points sur cette droite conserve plus l'information contenue dans le nuage.



ACP : idées de base

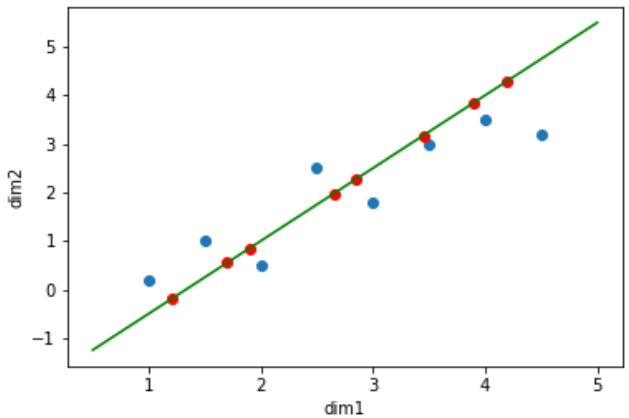
On voit clairement que la meilleure droite est la droite verte. La projection des points sur cette droite conserve plus l'information contenue dans le nuage.

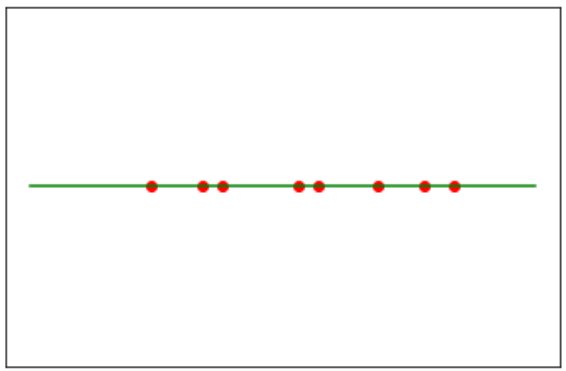




ACP : idées de base

- Une fois les points projetés, il faut transformer le système de coordonnées pour obtenir une représentation 1-D.





- les nouvelles données (1-D) ont la même variance que les données initiales dans la direction de la ligne verte.
- Justement, l'ACP préserve la plus grande variance des données. (On y reviendra).

ACP : utilisation

- Compression de données,
- Visualisation de données,
- Accélération de l'apprentissage,
- ...

ACP : Objectif

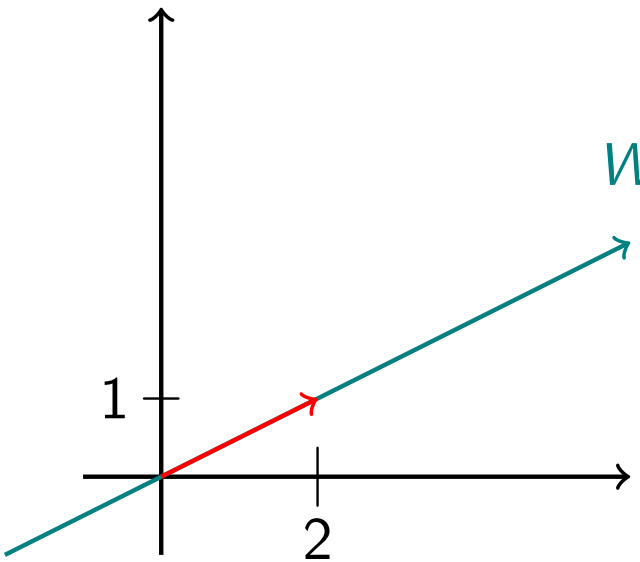
- Données d'origine : $X = (X_1, X_2, \cdots, X_p)$.
  Les variables X_i sont fortement corrélées.
- Ce que l'on cherche : $Y = (Y_1, Y_2, \cdots, Y_q)$, tel que :
  - $Y_j = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$,
  - les variables Y_j sont indépendantes,
  - avec une perte d'information la plus petite possible ...

ACP : Rappels

- V un espace de dimension d, W un sous espace de V de dimension k.
- On peut toujours trouver k vecteurs de dimension d, $\{e_1, e_2, \cdots, e_k\}$ formant une base orthonormée de W, i.e., $< e_i, e_j >= 0$ pour tout $i \neq j$ et $< e_i, e_i >= 1$.
- $\Rightarrow$ tout vecteur u de W peut s'écrire $u = \sum_{i=1}^{k} \alpha_i e_i$.

ACP : Rappels

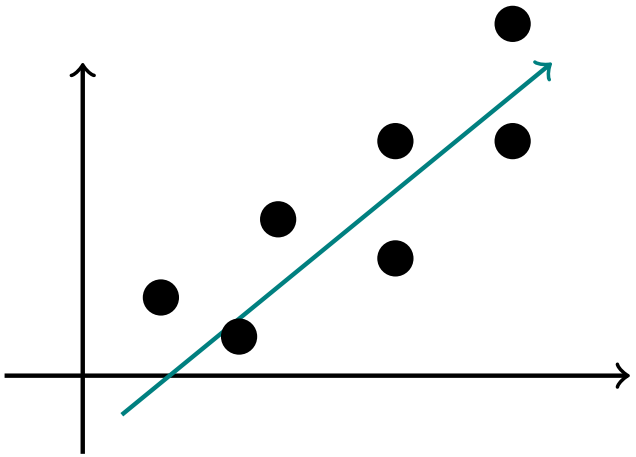
- Exemple.

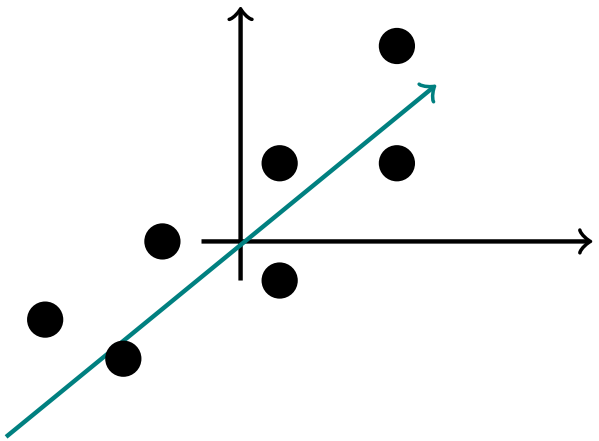


- $V = \mathbb{R}^2$,
- W la droite d'équation $x - 2y = 0$,
- $e = \left(\frac{1}{\sqrt{5}}, \frac{-2}{\sqrt{5}}\right)^t$ est une base orthonormée de W

ACP

L'origine du (des) nouvel (nouveaux) axe(s) doit être au centre du nuage de points :





Ce qui revient à centrer les données :

$$z = x - \frac{1}{n}\sum_{i=1}^{n} x_i = x - \overline{x},$$

Attention : prenez le temps de comprendre ce que représentent les x, x_i, et z, des scalaires ? des vecteurs ? des matrices ?

ACP

Rappelons ce que l'on cherche à faire...

- On cherche à trouver une représentation assez fidèle d'un dataset $\mathcal{S} = \{x_1, x_2, \cdots, x_n\}$ dans un sous espace W de dimension $k < d$
- On va plutôt travailler avec les données centrées, i.e., $\{z_1, z_2, \cdots, z_n\}$.

ACP

- Soit $\{e_1, e_2, \cdots, e_k\}$ la base orthonormée de W. Chaque vecteur u dans W peut être écrit en fonction de la base :

$$\sum_{i=1}^{k} \alpha_i e_i$$

- z_j sera représenté par un vecteur dans W :

$$z_j = \sum_{i=1}^{k} \alpha_{ji} e_i$$

ACP

- l'erreur, pour z_j, peut alors être représentée par :

$$erreur_j = \left\| z_j - \sum_{i=1}^{k} \alpha_{ji} e_i \right\|^2$$

- L'erreur totale pour tout le dataset $\mathcal{S}$:

$$J(e_1, e_2, \cdots, e_k, \alpha_{11}, \alpha_{12}, \cdots, \alpha_{nk}) = \sum_{j=1}^{n} \left\| z_j - \sum_{i=1}^{k} \alpha_{ji} e_i \right\|^2$$

Objectif de l'ACP : trouver les bons paramètres $e_1, e_2, \cdots, e_k$ et $\alpha_{11}, \alpha_{12}, \cdots, \alpha_{nk}$ tels que J soit minimale.

ACP

Théorème de l'ACP

J est minimale si on prend comme base, pour W, les k vecteurs propres associés aux k plus grandes valeurs propre de de la "scatter matrix" $S = (n-1)\hat{\Sigma}$. $\hat{\Sigma}$ étant la matrice de covariance.

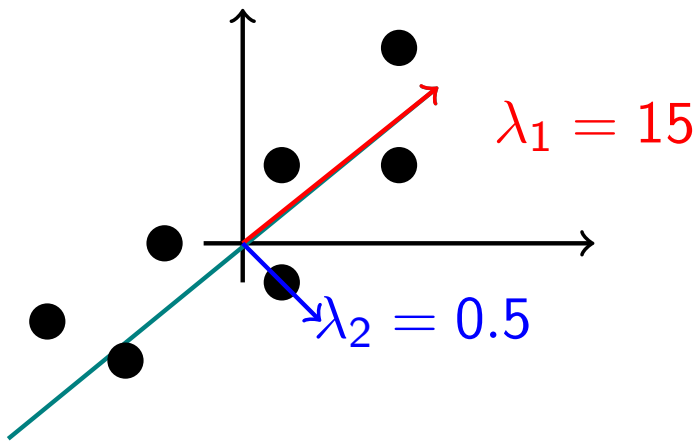
Preuve :

Voir le tableau.

ACP

Interprétation

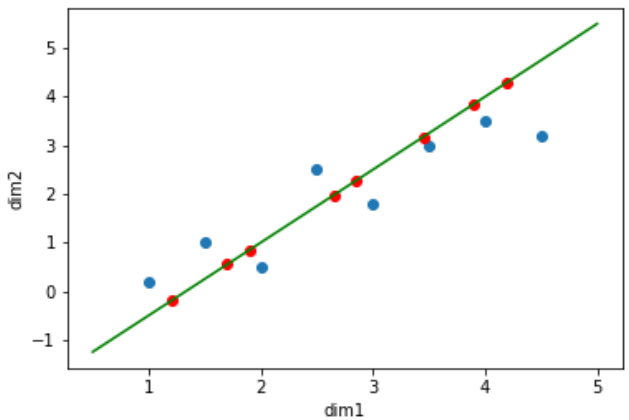
- Chaque valeur propre représente la variance dans la direction de son vecteur propre associé.

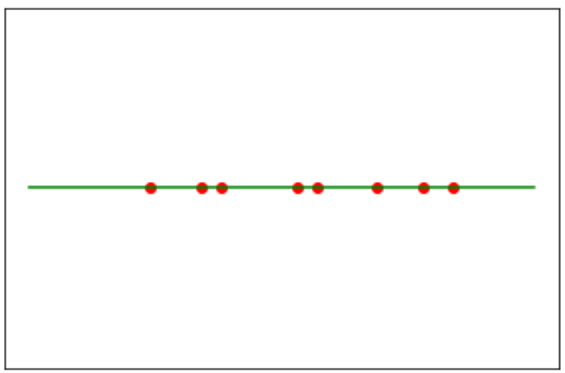


- En fait, l'ACP ne fait rien d'autre qu'une rotation jusqu'à ce que les directions conservant un maximum de variance soit trouvées ...

ACP : une dernière étape

- Comment changer les coordonnées pour obtenir le vecteur y de dimension k :





- Soit E la matrice $E = (e_1, e_2, \cdots, e_k)$. Le nouveau vecteur s'obtient comme suit :

$$y = E^t x$$

- Exercice : les vecteurs propres forment la base standard sous E^t.

L'ACP en une page

Données : $\mathcal{S} = \{x_1, x_2, \cdots, x_n\}$.
Chaque x_i est un vecteur de dimension d;
Objectif : réduire la dimension de d à k

1. Calculer la moyenne de l'échantillon $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. Centrer les données : $z_j = x_j - \overline{x}$
3. Calculer la "scatter matrix" : $S = (n-1)\sum_{j=1}^{n} z_j z_j^t$
4. Calculer les vecteurs propres $e_1, e_2, \cdots, e_k$ associés aux k plus grandes valeurs propres de S
5. Soit $E = (e_1, e_2, \cdots, e_k)$
6. Le vecteur recherché est

$$y = E^t z$$

L'ACP en pratique

Voir le Jupyter notebook...

ACP : Nombre de facteurs

On récapitule :

- On cherche à projeter le nuage de points dans des espaces géométriques de dimension 1 (un axe), 2 (un plan), 3 (un espace "ordinaire"), ... de manière à déformer le moins possible les distances entre les individus, et donc la variabilité observée.
- Le meilleur axe est le premier axe factoriel ;
- Le meilleur plan est défini par l'axe factoriel précédent et un deuxième axe orthogonal au précédent. ...
- **La variance** de la composante principale Y_k est égale à la k-ième **valeur propre**.

ACP : Nombre de facteurs

Question :

Quel est le nombre d'axes à garder ? 1 ?, 2 ?, ...

Réponse : cela dépond ...

- On peut vouloir (absolument) garder une variabilité supérieure à un seuil $\rightarrow$ la somme des plus grande valeurs propres doit être supérieure à la variance souhaitée;
- On peut vouloir garder une certaine qualité des données (exemple : ACP pour la compression)
- de manière générale : utiliser un critère, par exemple le critère du "coude".

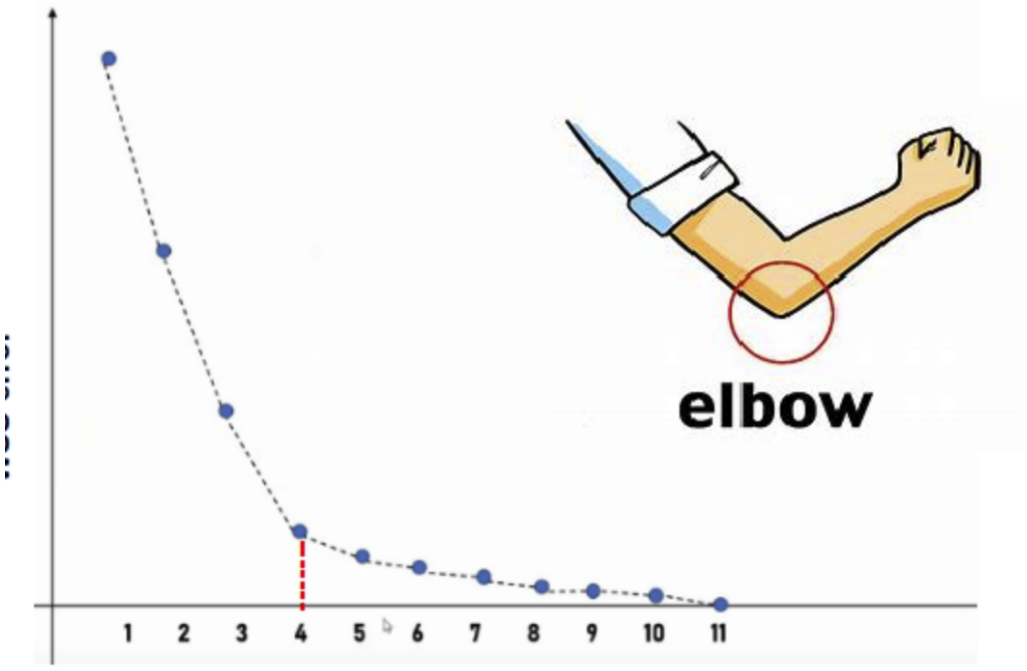
ACP : Nombre de facteurs

Question :

Quel est le nombre d'axes à garder ? 1 ?, 2 ?, ...

Réponse : cela dépond ...

- de manière générale : utiliser un critère, par exemple le critère du "coude" :



ACP : Nombre de facteurs

Question :

Quel est le nombre d'axes à garder ? 1 ?, 2 ?, ...

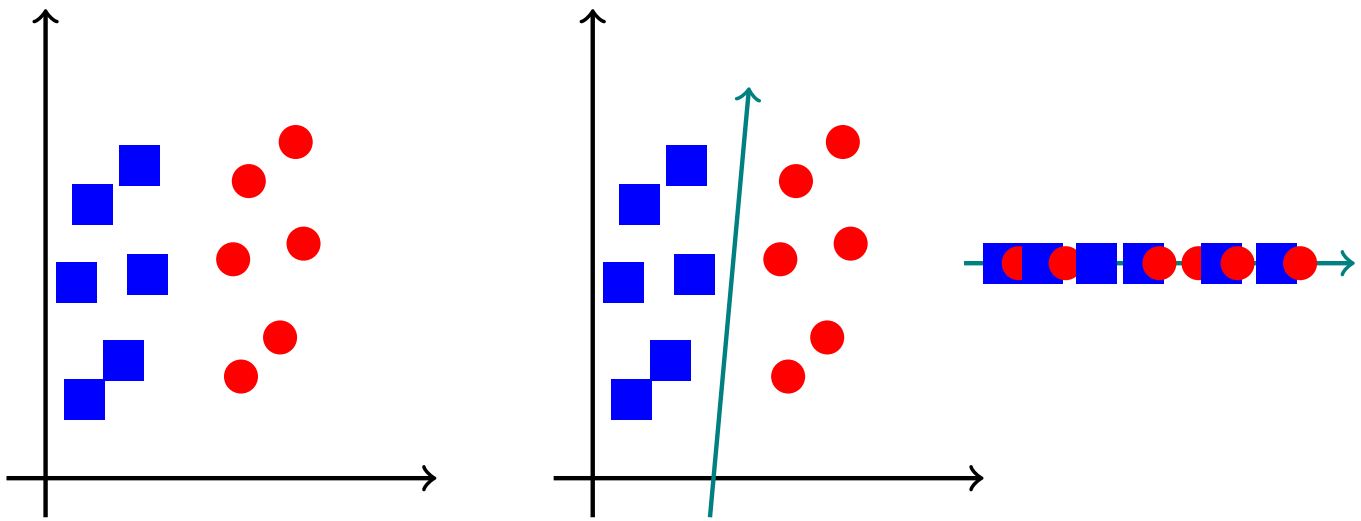
Réponse : cela dépond ...

- de manière générale : utiliser un critère, par exemple le critère du "coude" :

Voir le Jupyter notebook

ACP : limites

- L'ACP est conçue pour représenter des données, et non pour les classifier,
  - elle préserve la variance autant que possible,
  - si les directions à plus grandes variances coincident avec les bons choix pour la classification alors elle peut servir
- En général, la direction conservant la plus grande variance peut ne pas être avantageuse pour la classification :



ACP : limites

Solution : Faire une LDA