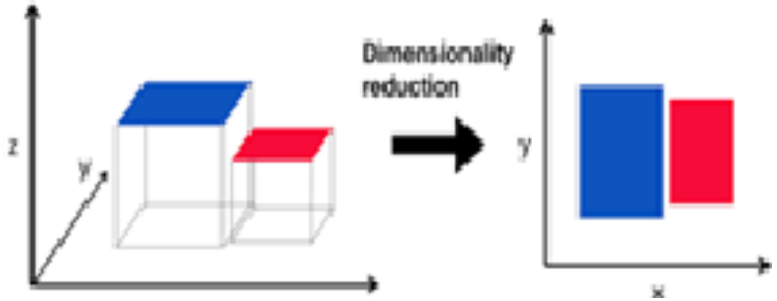
Réduction de la dimensionnalité de données

Dr. HADJADJ ABDELHALIM



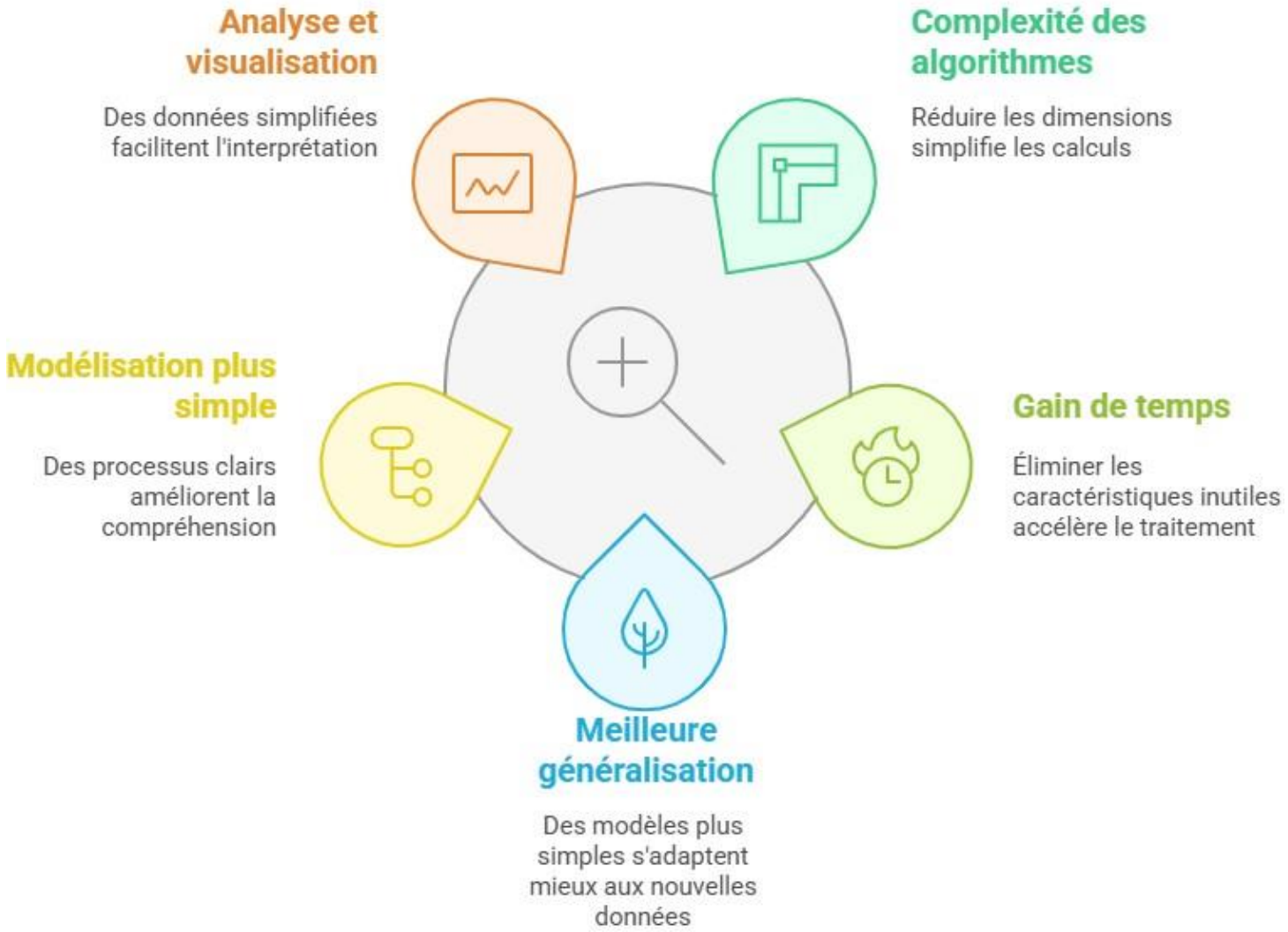
Université de Mila
Master I: Apprentissage automatique

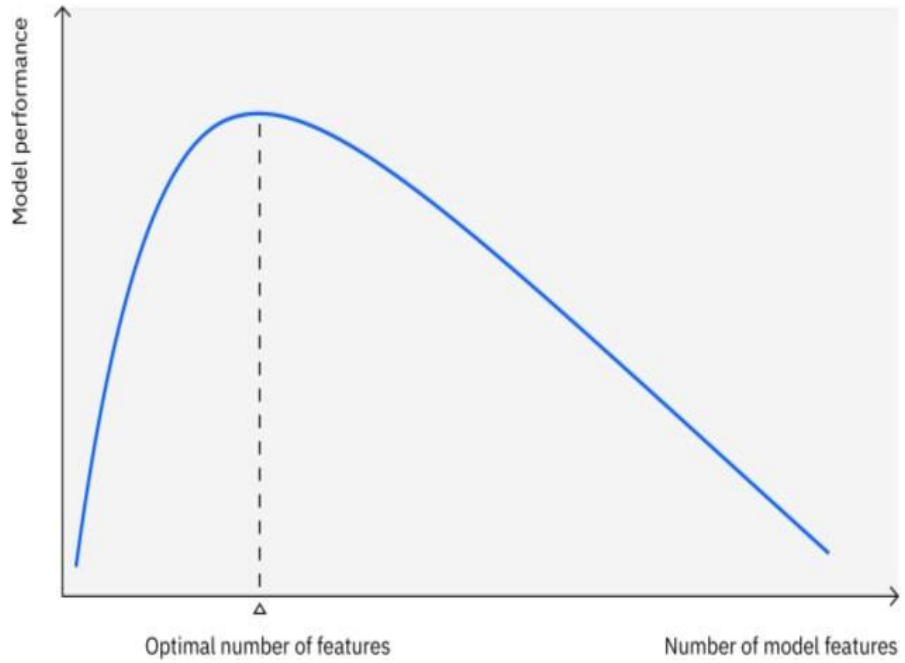
Réduction de la dimensionnalité de données

- La **réduction de dimensionnalité** est un ensemble de techniques qui visent à **réduire le nombre de variables** d'un jeu de données tout en **conservant l'essentiel de l'information**.
- Elle permet de simplifier les données, d'améliorer la vitesse des traitements, de faciliter la visualisation et souvent d'améliorer les performances des modèles prédictifs.

Pourquoi?

- En apprentissage automatique, les dimensions (ou caractéristiques) sont les variables d'entrée qui influencent la sortie d'un modèle.





Quelle méthode de réduction de dimension devrait être utilisée ?

1

Sélection d'attributs

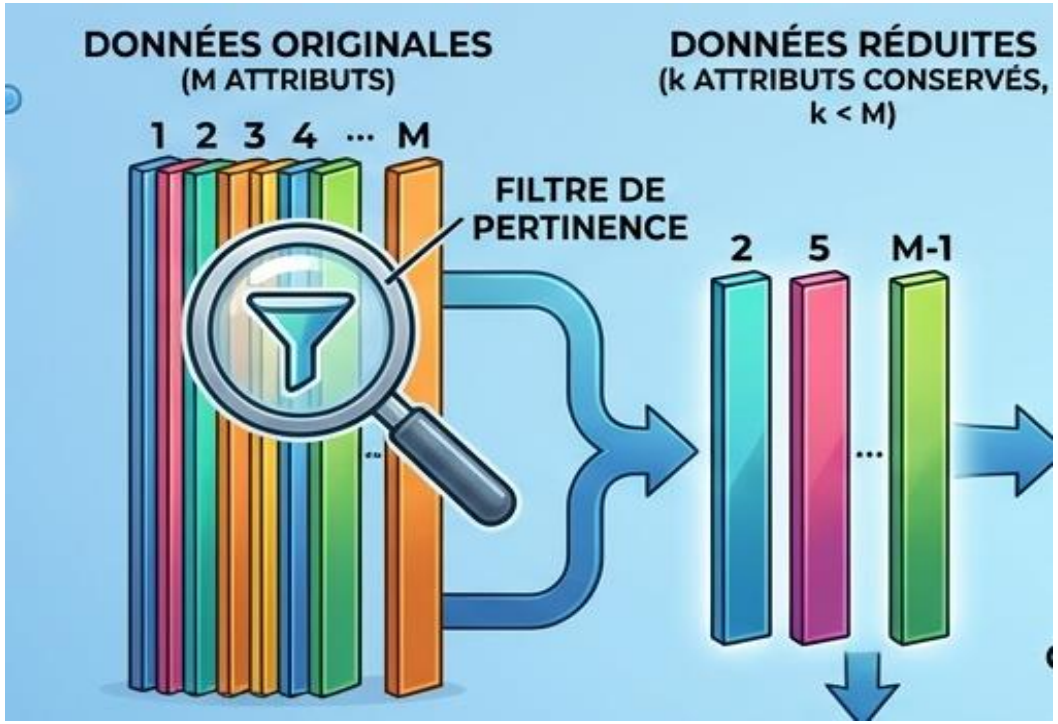
Conserve les attributs pertinents et élimine ceux qui sont non pertinents, réduisant la complexité.

2

Extraction d'attributs

Crée de nouveaux attributs tout en conservant l'information essentielle, peut êt supervisée ou non supervisée.

1. sélection d'attributs



Université de Mila
Master I: Apprentissage automatique

1. Principe de sélection d'attributs

- L'objectif est d'identifier le **plus petit sous-ensemble d'attributs** permettant de construire un modèle de **régression ou de classification performant**.
  - Une fois ce sous-ensemble sélectionné, les autres attributs sont **éliminés**, car ils sont jugés non pertinents ou redondants.
  - En ayant D attributs au départ, Il existera (2^{D}-1) sous ensembles possibles d'attributs qu'on peut former (**ex.** avec D = 10, on aura 2^{10} sous ensembles)

Il existe deux méthodes pour la sélection d'attributs:

1. Principe de sélection d'attributs

- **1.1 Sélection en avant (par ajout progressif)**
- On commence **avec aucun attribut** (S = ∅).
- À chaque étape, on **ajoute l'attribut(**x_d**)** qui permet de **réduire le plus possible l'erreur de validation** du modèle $E(S \cup x_d)$.
- On continue à ajouter des attributs **tant que l'erreur diminue**.
- **Arrêt** : Dès que l'ajout d'un nouvel attribut ne permet plus d'améliorer la performance (l'erreur devient stable ou augmente), on **arrête le processus**.

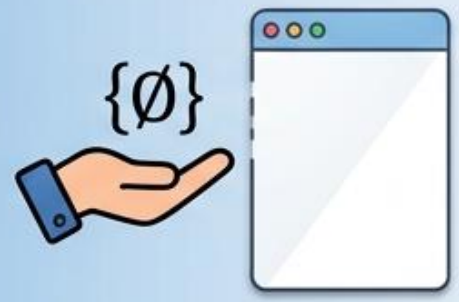
Cet algorithme de sélection d'attributs s'appelle: enroulement (wrapper), car le modèle de classification ou de régression est utilisé comme une routine pour la validation.

Sélection d'attributs par ajout progressif (Forward Selection)

ALGORITHME DE SÉLECTION D'ATTRIBUTS PAR ENROULEMENT (*WRAPPER*) : SÉLECTION EN AVANT

Le modèle (classification/régression) sert de routine de validation.

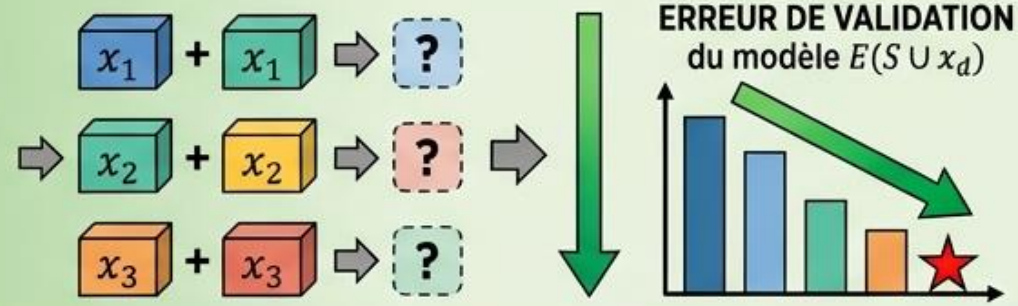
1. On commence avec aucun attribut ($S = \emptyset$).



SÉLECTION INITIALE (S)

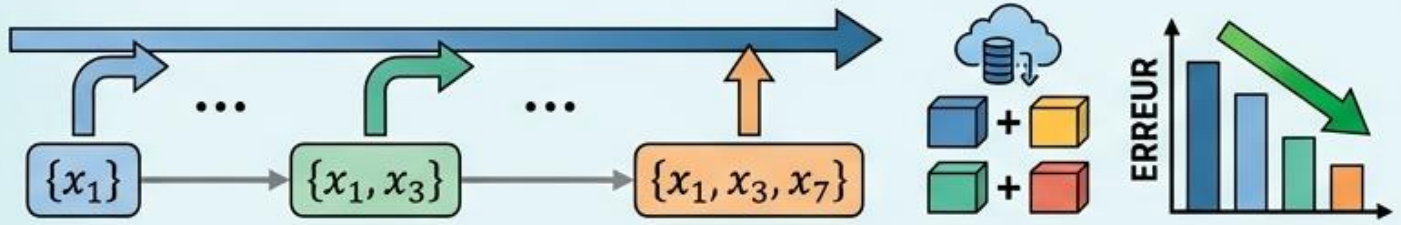
$\emptyset$

2. À chaque étape, on ajoute l'attribut (x_d) qui permet de réduire le plus possible l'erreur de validation.



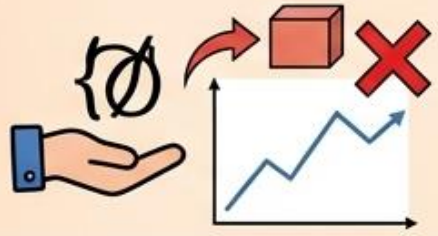
Classification de modèle

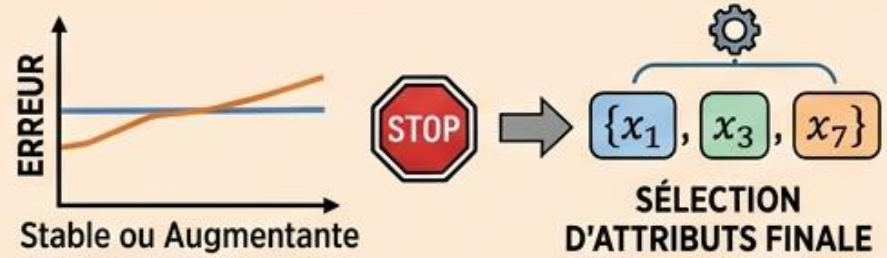
3. On continue à ajouter des attributs tant que l'erreur diminue.



Validation procedure

4. Arrêt : Dès que l'ajout d'un nouvel attribut ne permet plus d'améliorer la performance.





Sélection d'attributs par ajout progressif (Forward Selection)

Exemple

On a 4 attributs :

x_1 = âge

x_2 = revenu

x_3 = sexe

x_4 = nombre de visites

Étape 1 : départ

S = ∅

On teste chaque attribut seul :

- Modèle avec x_1 → erreur = 30%
- Modèle avec x_2 → erreur = 25% (meilleur)
- Modèle avec x_3 → erreur = 40%
- Modèle avec x_4 → erreur = 28%

On choisit **x_2 (revenu)**

Étape 2 : ajouter un attribut

On teste les combinaisons avec x_2 :

- $\{x_2, x_1\}$ → erreur = 20% (meilleur)
- $\{x_2, x_3\}$ → erreur = 24%
- $\{x_2, x_4\}$ → erreur = 22%

On ajoute **x_1 (âge),** S = $\{x_1, x_2\}$

Étape 3 : continuer

On teste :

$\{x_1, x_2, x_3\}$ → erreur = 21%

$\{x_1, x_2, x_4\}$ → erreur = 19% (meilleur)

On ajoute **x_4,** S = $\{x_1, x_2, x_4\}$

Étape 4 : arrêt

On teste le dernier attribut :

- $\{x_1, x_2, x_4, x_3\}$ → erreur = 19% (pas mieux)

L'erreur n'améliore plus → **on s'arrête**

Exemple

- Considérons le jeu de données **IRIS**, composé de N **= 150 observations** et de D **= 4 attributs** : x_1, x_2, x_3 et x_4.
  - on évalue la précision du modèle en utilisant **un seul attribut à la fois, les résultats pour** x_1 ,x_2 x_3 x_4 le suivant (**0.76**, **0.57**, **0.92**, **0.94**)
  - On choisit x_4 comme **premier attribut**, car il donne la meilleure précision.
- ensuite, on teste les combinaisons possibles de x_4 avec chacun des autres attributs restants (x_1, x_2, x_3) , les précisions obtenues, respectivement, 0.87, 0.92 et 0,92.
- On choisit x_3 comme **deuxième attribut**, car il améliore ou maintient la précision maximale.
- **Remarques**

  La sélection d'attributs est supervisée car les sorties de la régression/classification sont utilisées pour calculer l'erreur de validation.

1. Principe de sélection d'attributs

1.2 Selection en arrière (par élimination successive)

- On commence avec **tous les attributs** disponibles(S=A)
- À chaque étape, on **supprime l'attribut** dont l'élimination **diminue le plus l'erreur de validation**
- On continue à supprimer les attributs **jusqu'à ce que l'erreur ne s'améliore plus(**S−x_d. **)**.
- Lorsque l'élimination d'un attribut **n'améliore plus la performance** (l'erreur reste stable ou augmente), on **arrête le processus**.

1.2 SÉLECTION EN ARRIÈRE
(PAR ÉLIMINATION SUCCESSIVE)

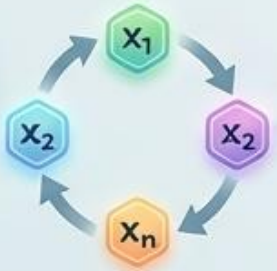
1. On commence avec **TOUS** les attributs disponibles (S=A)

$\{x1, x2, ..., x_n\}$

A = TOUS LES ATTRIBUTS

2. À chaque étape, on supprime l'attribut dont l'élimination **DIMINUE** le plus l'erreur de validation

3. On continue à supprimer les attributs **JUSQU'À CE QUE L'ERREUR NE S'AMÉLIORE PLUS**



4. Lorsque l'élimination d'un attribut **N'AMÉLIORE PLUS LA PERFORMANCE**, on arrête le processus

$S = \{x_1, x_3, ..., x_m\}$

Exemple

Étape 1 : départ

S = {x_1, x_2, x_3, x_4}

Erreur = 19%

On teste la suppression de chaque attribut :

•sans x_1 → erreur = 22%

•sans x_2 → erreur = 30%

•sans x_3 → erreur = 19%

•sans x_4 → erreur = 21%

On supprime **x_3** (aucun impact), S = {x_1, x_2, x_4}

Étape 2

On teste à nouveau :

•sans x_1 → erreur = 21%

•sans x_2 → erreur = 28%

•sans x_4 → erreur = 17%

On supprime **x_4,** S = {x_1, x_2}, Nouvelle erreur = **17%**

Étape 3 : arrêt

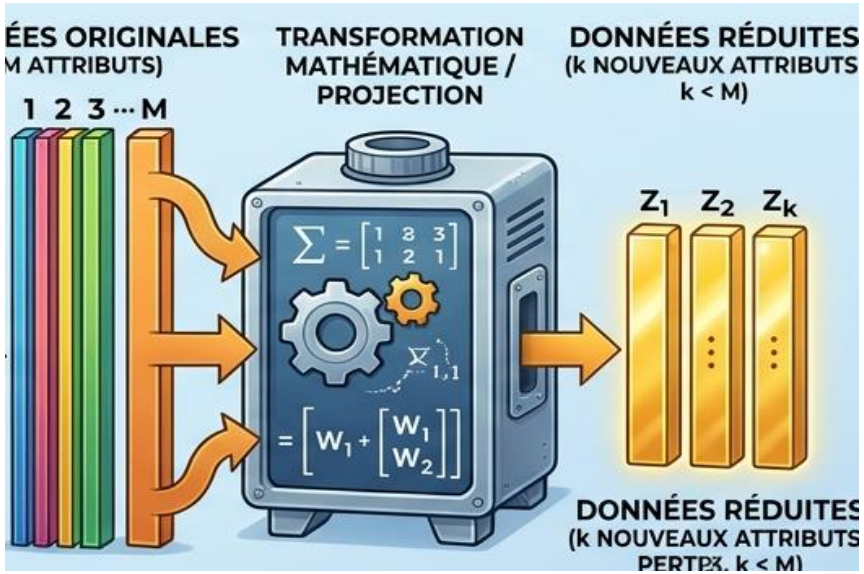
On teste :

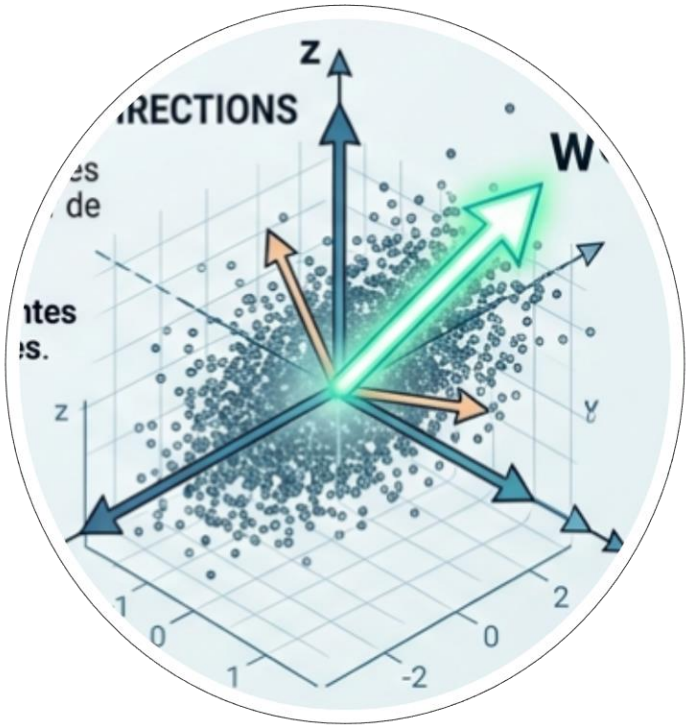
•sans x_1 → erreur = 25%

•sans x_2 → erreur = 35%

Toute suppression **augmente l'erreur**, **On s'arrête**

2. Extraction d'attributs: Analyse à composantes principales (ACP)





2. *Principe d'extraction d'attributs*

- **Diverses approches** permettent d'extraire de nouveaux attributs à partir d'un jeu de données.
- **Les méthodes de projection** jouent un rôle clé dans cette extraction.
- Leur objectif est de réduire la dimensionnalité : passer d'un espace à D **dimensions** à un sous-ensemble de M **attributs** (avec $M < D$), tout en **préservant au maximum** l'information structurelle des données.
- **Parmi les techniques les plus utilisées**, on trouve :
  - L'**Analyse en Composantes Principales**
  - L'**Analyse Discriminante** linéaire
  - L'**Analyse Factorielle**
  - L'**Analyse de Corrélation**

2. *Analyse à composantes principales (ACP)*

- ✓ L'ACP est une **méthode d'analyse non supervisée**,
- ✓ Elle conserve l'information importante sous forme de **variance maximale**.
- ✓ Elle projette les données sur de nouvelles directions appelées **composantes principales**.
- ✓ Chaque nouvelle variable est une **combinaison linéaire des variables initiales**.

Exemple:

- un vecteur x est projeté sur une direction $w^{(1)}$
- on obtient une nouvelle variable $z_1 = w^{(1)T} x$

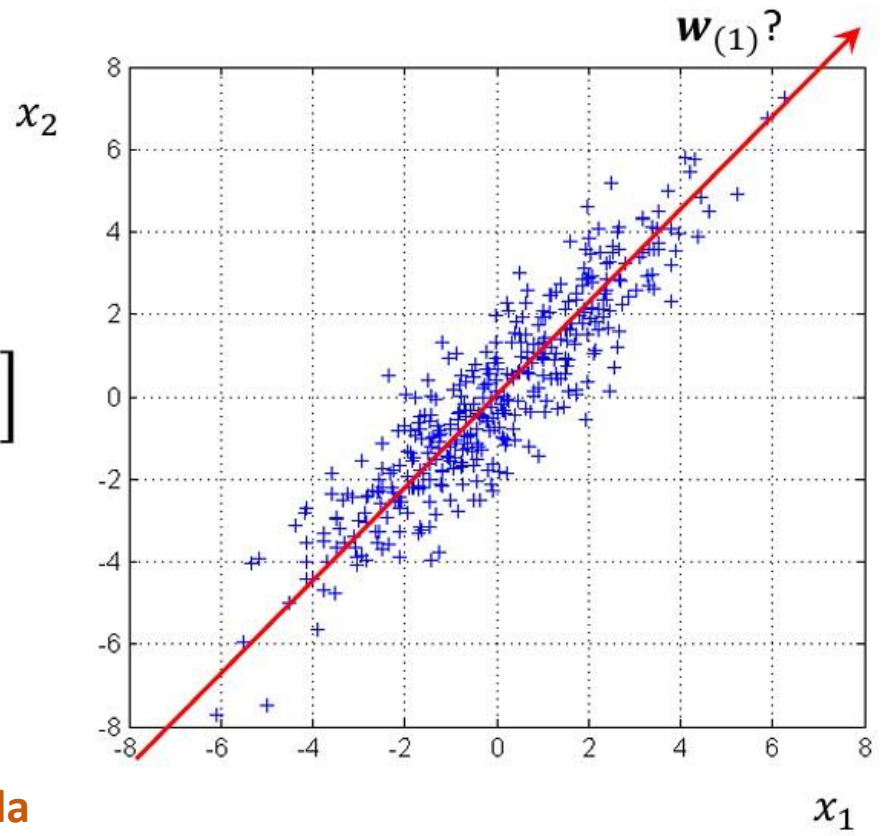
2. Analyse à composantes principales (ACP)

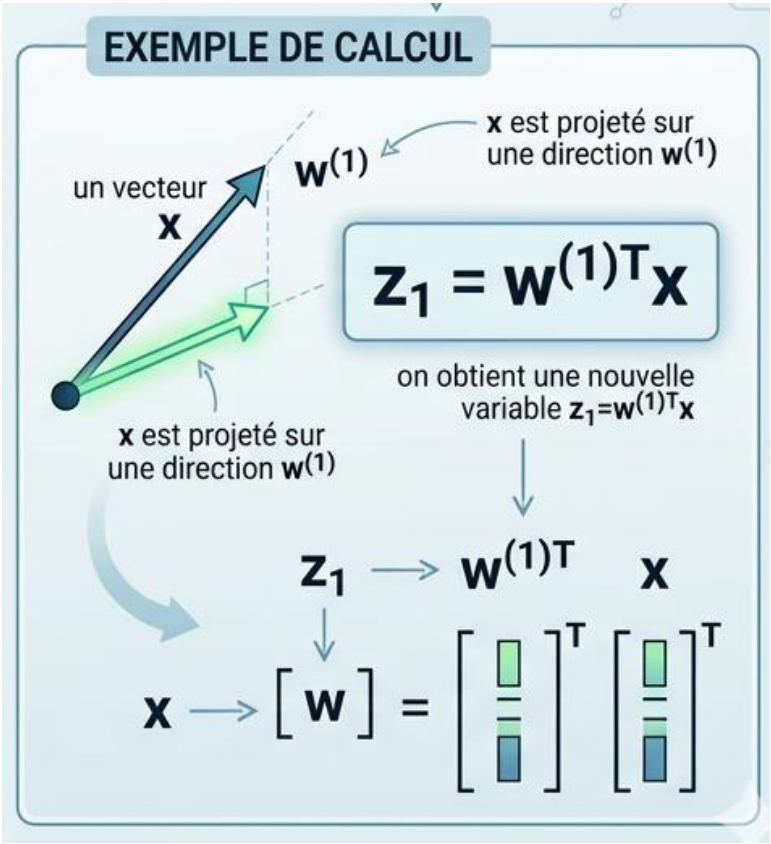
Exemple:

Quelle est la direction de projection qui maximise la variance?

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 4 & 4 \\ 4 & 5 \end{bmatrix}$$





2. *Analyse à composantes principales (ACP)*

- On considère un vecteur de données $\boldsymbol{x}$ avec une **moyenne $\boldsymbol{\mu}$** et une **matrice de covariance $\boldsymbol{\Sigma}$**.
- La variance d'une projection $z_1 = w^{(1)T}x$ est donnée par : $\mathrm{Var}(z_1) = w^{(1)T}\Sigma w^{(1)}$

Objectif de l'ACP:

- Trouver la direction $w^{(1)}$ qui **maximise la variance** des données projetées.
- Sous contrainte : $\| w^{(1)} \| = 1$

Pour cela, on résout le problème d'optimisation suivant, en utilisant la méthode des multiplicateurs de Lagrange : $\max_{\mathbf{w}^{(1)}} \left[(\mathbf{w}^{(1)})^T \Sigma \mathbf{w}^{(1)} - \lambda \left((\mathbf{w}^{(1)})^T \mathbf{w}^{(1)} - 1 \right) \right]$

- où : λ est le multiplicateur de Lagrange, il sert à maintenir la contrainte $\|w^{(1)}\|=1$ pendant l'optimisation.

Analyse à composantes principales (ACP)

On utilise les multiplicateurs de Lagrange :

$$\mathcal{L}\left(w^{(1)}\right) = w^{(1)T}\Sigma w^{(1)} - \lambda\left(w^{(1)T}w^{(1)} - 1\right)$$

Dérivation

On dérive par rapport à $w^{(1)}$et on égalise à zéro :

$$2\Sigma w^{(1)} - 2\lambda w^{(1)} = 0$$

Simplification

$$\Sigma w^{(1)} = \lambda w^{(1)}$$

2. Mathématiquement, cette équation signifie que :

$w^{(1)}$est un **vecteur propre** de la matrice de covariance Σ

λest la **valeur propre associée**

Puisque $\|\boldsymbol{w}_{(1)}\| = 1$, on remarque aussi que:

$$\boldsymbol{w}_{(1)}{}^{T}(\boldsymbol{\Sigma}\,\boldsymbol{w}_{(1)}) = \boldsymbol{w}_{(1)}{}^{T}\,(\lambda \boldsymbol{w}_{(1)}).$$

$$= \lambda(\boldsymbol{w}_{(1)}{}^{T}\,\boldsymbol{w}_{(1)})$$

$$= \lambda$$

- Donc, La **première direction de projection** qui maximise la variance correspond au **vecteur propre associé à la plus grande valeur propre** de la matrice de covariance Σ.

2. *Analyse à composantes principales (ACP)*

Soit $w^{(2)}$la deuxième direction de projection qui maximise la variance, définie par $\mathrm{Var}(z_2) = w^{(2)T}\Sigma w^{(2)}$ avec les contraintes $\| w^{(2)} \| = 1$et $w^{(1)T}w^{(2)} = 0$)orthogonalité avec la première direction).

En appliquant le même raisonnement que pour la première direction, on montre que $w^{(2)}$est un vecteur propre de Σ.
Ainsi, la deuxième direction principale $w^{(2)}$correspond au vecteur propre associé à la deuxième plus grande valeur propre de la matrice de covariance Σ.

De manière générale, en répétant ce processus, on peut extraire les M premières directions principales de projection, qui maximisent successivement la variance des données projetées tout en étant orthogonales entre elles.

2. *Analyse en Composantes Principales (ACP)*

- On note $w^{(1)}, w^{(2)}, \ldots, w^{(M)}$: les **M premières directions de projection** extraites par l'ACP. À chacune de ces directions est associée une **valeur propre** $\lambda_1, \lambda_2, \ldots, \lambda_M$.
- Ces directions sont appelées les **composantes principales (CP)** :
  - $w^{(1)}$: 1re composante principale,
  - $w^{(2)}$: 2e composante principale,
  - Et ainsi de suite.
- **Les premières composantes** sont généralement celles qui **capturent le plus de variance** dans les données.
- Exemple d'utilisation :
  - Si l'objectif est de **conserver 90 % de la variance totale**, on peut choisir le plus petit nombre *M* de composantes principales telles que :

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_M}{\lambda_1 + \lambda_2 + \cdots + \lambda_D} \geq 0{,}90$$

  - Cela permet de **réduire la dimension** tout en gardant une grande partie de l'information des données d'origine.

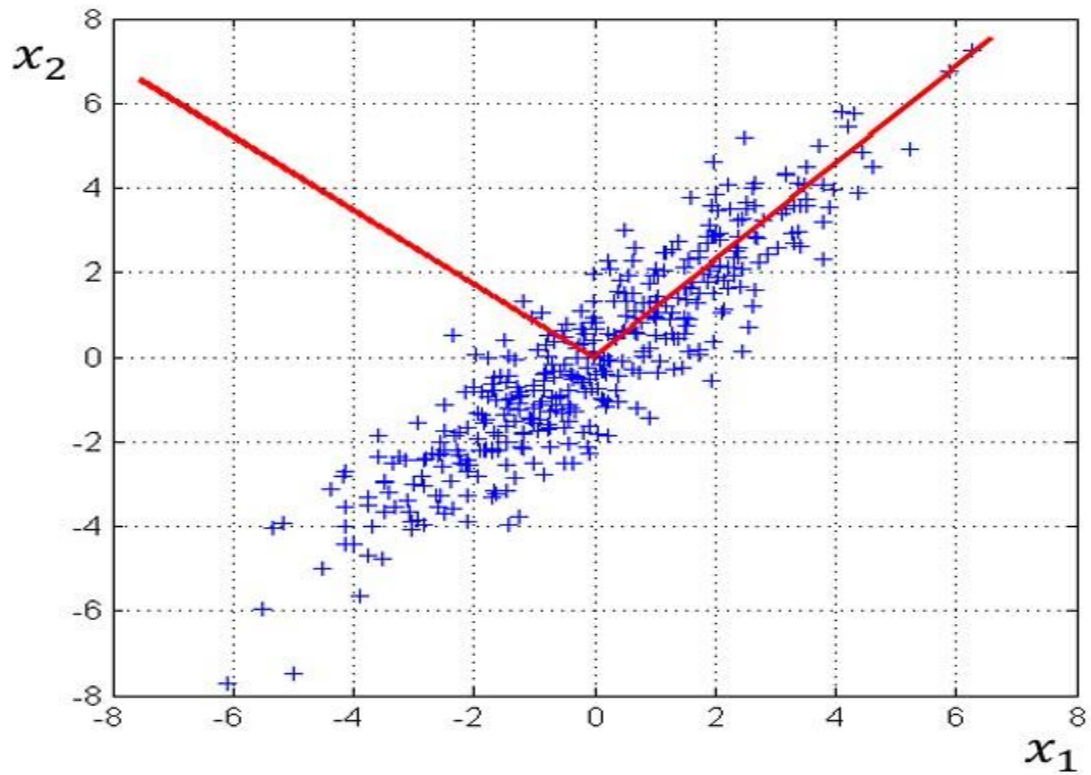
2. *Analyse en Composantes Principales (ACP)*

L'ACP donne les directions principales comme suit:

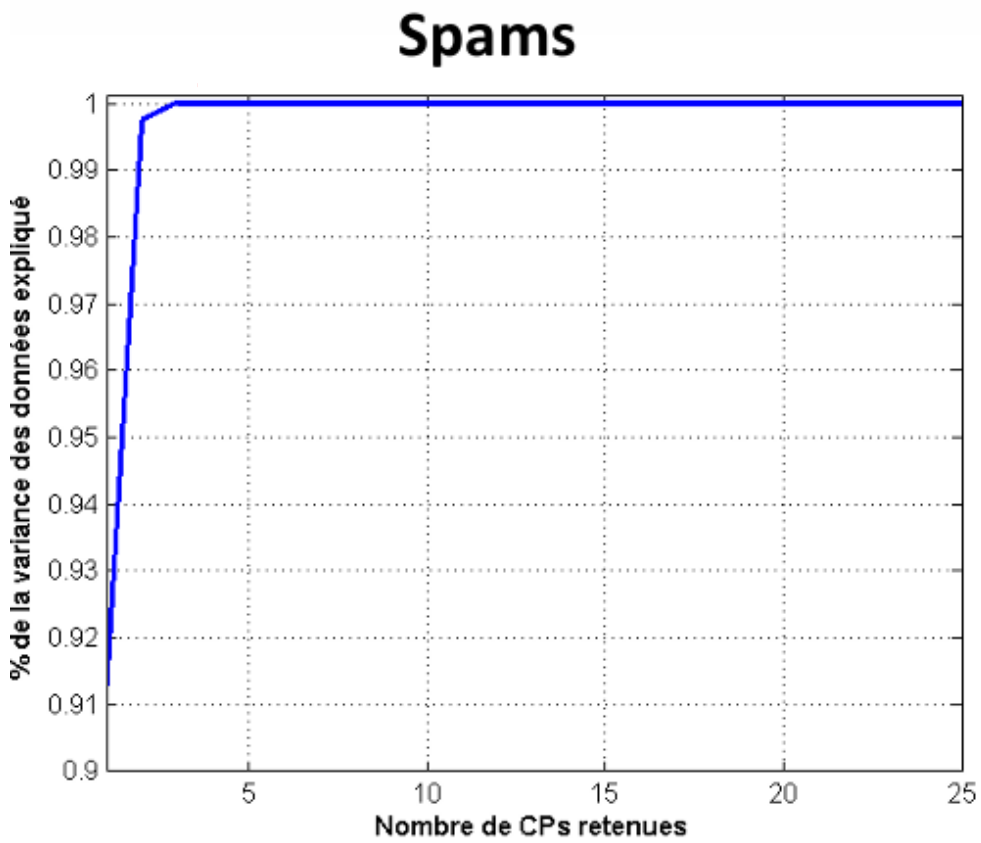
$$\mathbf{W} = \begin{bmatrix} \overset{\downarrow}{0.67} & \overset{\downarrow}{-0.74} \\ 0.74 & 0.67 \end{bmatrix} \quad \begin{matrix} \boldsymbol{w}_{(1)} \\ \\ \\ \end{matrix} \quad \begin{matrix} \boldsymbol{w}_{(2)} \\ \\ \\ \end{matrix}$$

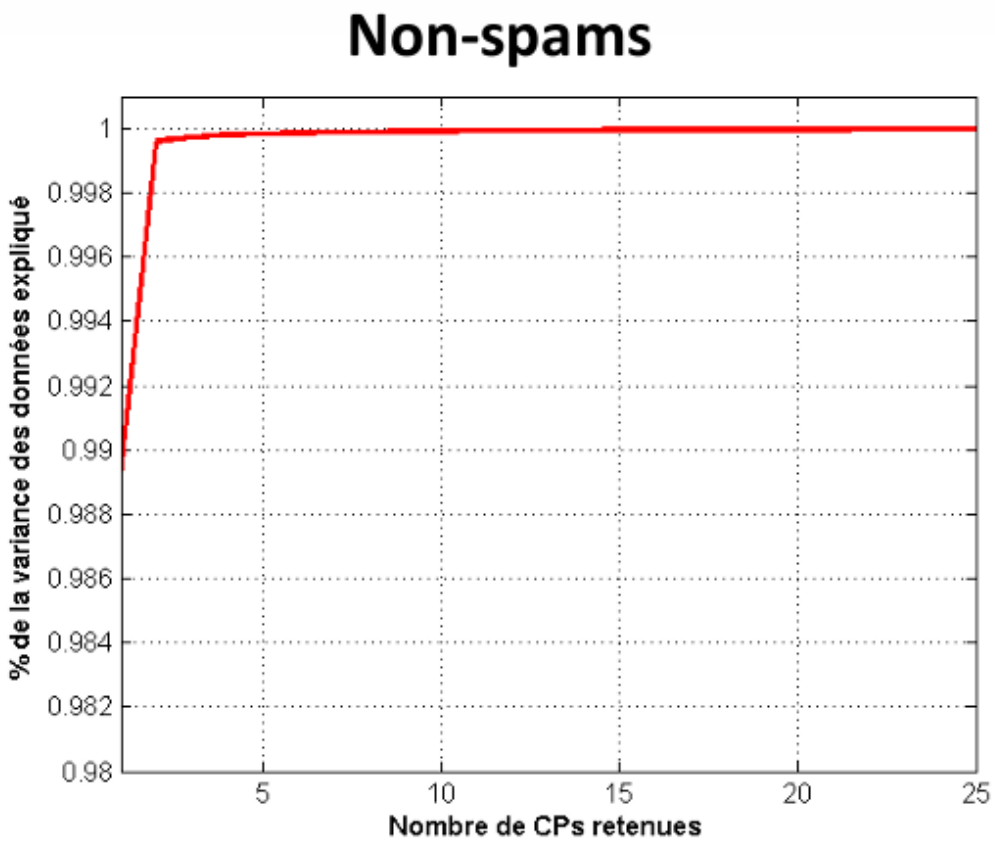
$$\lambda_1 = 8.76.$$

$$\lambda_2 = 0.48$$



Avec les données de spams de HP Labs, on obtient les graphes de Scree suivants avec les courriels spams et non-spams:





Exemple

1. Données initiales (2D)

Supposons 4 points :

$$x_1 = (2, 3), x_2 = (3, 5), x_3 = (4, 7), x_4 = (5, 8)$$

Ces points sont proches d'une droite (ils sont corrélés).

2. Étape 1 : moyenne

On calcule le centre :

$$\mu = \frac{1}{4}(x_1 + x_2 + x_3 + x_4) = (3.5, 5.75)$$

3. Étape 2 : centrage

On soustrait la moyenne :

$$x_i' = x_i - \mu$$

Exemple :

$$x_1' = (2, 3) - (3.5, 5.75) = (-, 1.5 - 2.75)$$

4. Étape 3 : matrice de covariance

On obtient une matrice du type :

$$\Sigma = \begin{pmatrix} 2.5 & 4.2 \\ 4.2 & 7.5 \end{pmatrix}$$

5. Étape 4 : vecteurs propres

On calcule :

$$\Sigma w = \lambda w$$

On obtient par exemple :

•1er vecteur propre :

$$w^{(1)} \approx (0.45, 0.89)$$

•2e vecteur propre :

$$w^{(2)} \approx (-, 0.890.45)$$

6. Interprétation

•$w^{(1)}$ =direction principale (max variance)

•$w^{(2)}$ =direction secondaire (peu de variance)

On garde souvent seulement $w^{(1)}$

7. Projection (réduction 2D → 1D)

On transforme chaque point :

$$z_i = w^{(1)T} x_i$$

Exemple :

$$z_1 = w^{(1)T} x_1$$

chaque point devient **un seul nombre**

3. Analyse discriminante linéaire (ADL)

Université de Mila
Master I: Apprentissage automatique

3. Analyse *Discriminante* Linéaire (ADL)

- **L'ACP** (Analyse en Composantes Principales) réduit la dimension en maximisant la **variance globale** des données, **sans tenir compte des classes**.
- Cependant, l'ACP **n'intègre pas d'information de classification**, ce qui peut être limitant pour des tâches supervisées (comme la classification).
- **L'ADL** (Analyse Discriminante Linéaire) est conçue pour résoudre ce problème
  - Elle **réduit la dimension** tout en **maximisant la séparation entre les classes**.
- **ADL est une méthode supervisée** :
  - Elle utilise les **étiquettes des classes** pour déterminer les directions les plus discriminantes.
- Objectif principal de l'ADL :
  - Trouver des axes de projection qui **maximisent la distance entre les moyennes des classes** tout en **minimisant la dispersion intra-classe**.

Exemple: $(D = 2)$

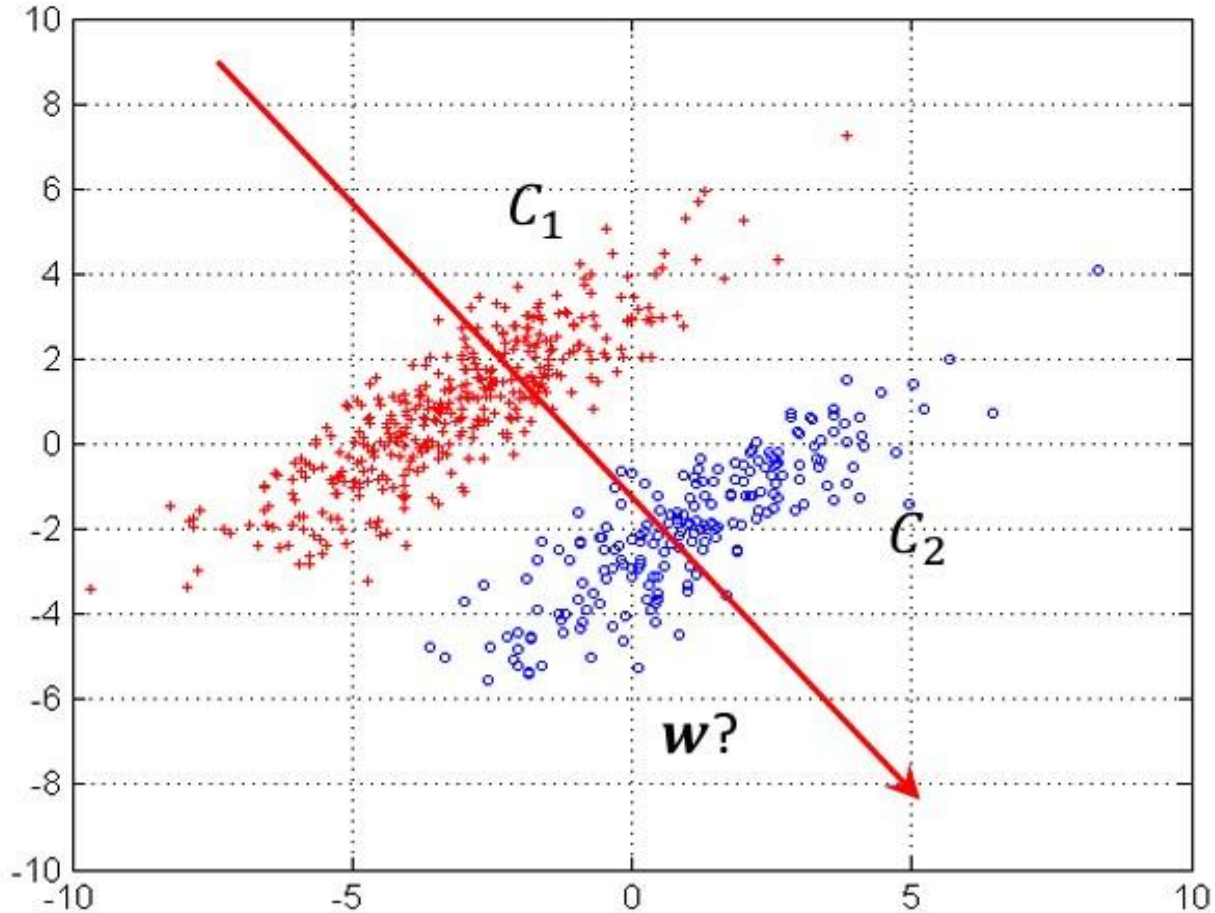
Quelle est la direction de projection $\boldsymbol{w}$ qui maximise la discrimination entre les deux classes C_1 et C_2 ?

$$\boldsymbol{\mu}^{(1)} = \begin{bmatrix} -3 \\ 1 \end{bmatrix}$$

$$\boldsymbol{\mu}^{(2)} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

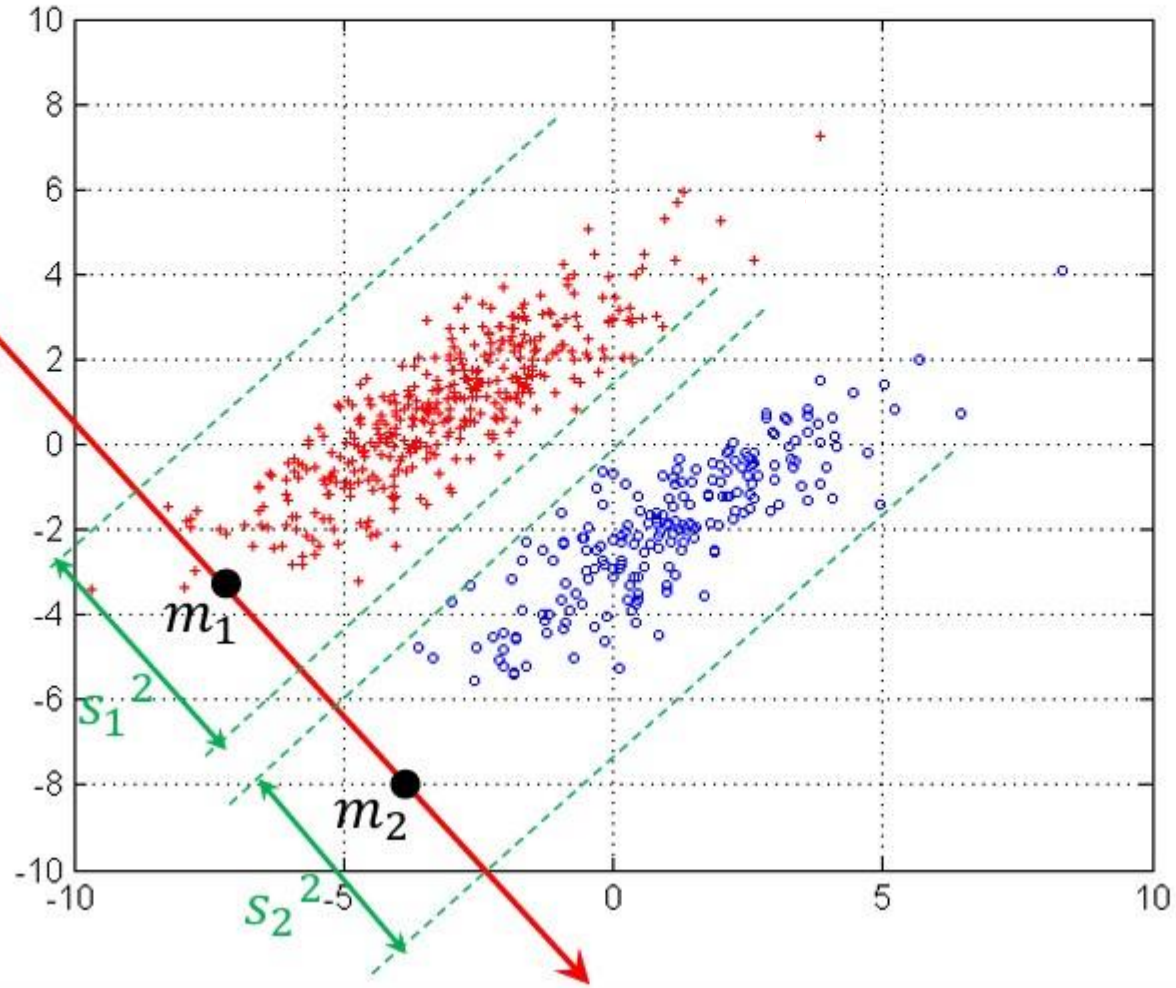
$$\boldsymbol{\Sigma}^{(1)} = \begin{bmatrix} 4 & 3 \\ 3 & 3 \end{bmatrix}$$

$$\boldsymbol{\Sigma}^{(2)} = \begin{bmatrix} 4 & 3 \\ 3 & 3 \end{bmatrix}$$



Soit m_1 et m_2 les moyennes des données de C_1 et C_2 après leurs projections sur $\boldsymbol{w}$ et s_1 et s_2 leurs variances.

$\boldsymbol{w}$?



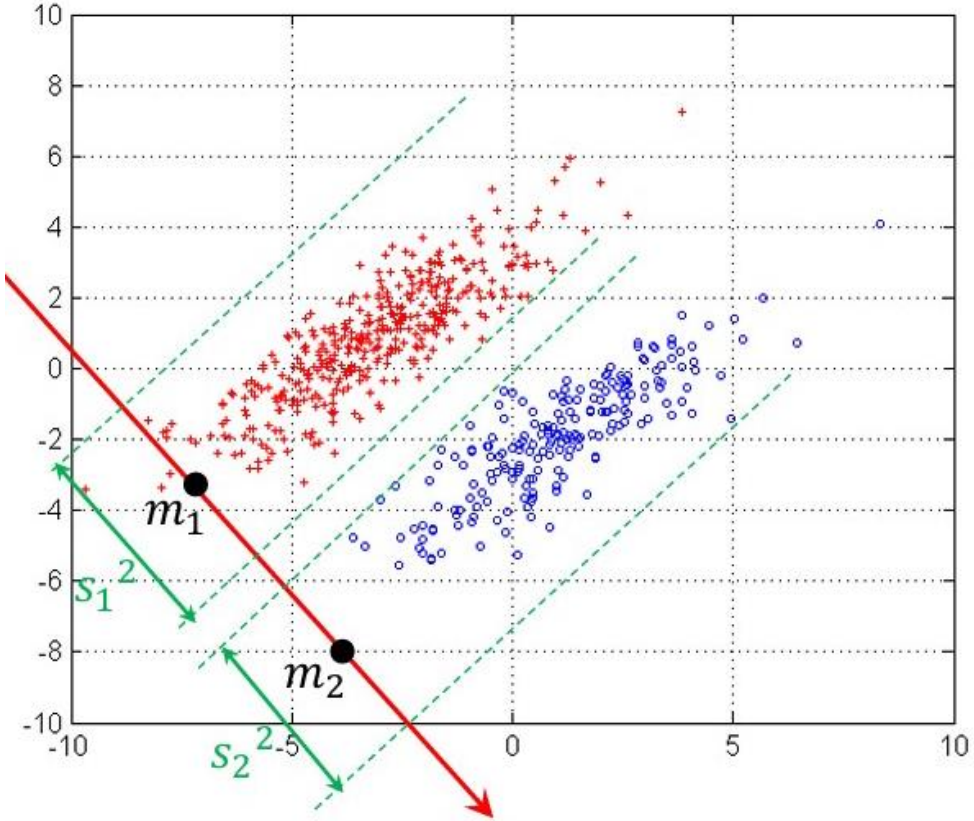
3. Analyse Discriminante Linéaire (ADL)

$$m_1 = \frac{\sum_{i=1}^{N} \boldsymbol{w}^T x^{(i)} y^{(i)}}{\sum_{i=1}^{N} y^{(i)}} = \boldsymbol{w}^T \boldsymbol{\mu}^{(1)}$$

$$m_2 = \frac{\sum_{i=1}^{N} \boldsymbol{w}^T x^{(i)} (1 - y^{(i)})}{\sum_{i=1}^{N} (1 - y^{(i)})} = \boldsymbol{w}^T \boldsymbol{\mu}^{(2)}$$

$$s_1{}^2 = \sum_{i=1}^{N} (\boldsymbol{w}^T x^{(i)} - m_1)^2 y^{(i)}$$

$$s_2{}^2 = \sum_{i=1}^{N} \left(\boldsymbol{w}^T x^{(i)} - m_2\right)^2 (1 - y^{(i)})$$



3. Analyse Discriminante Linéaire (ADL)

La direction $\boldsymbol{w}$ qui maximise la discrimination entre C_1 et C_2 est celle qui maximise la quantité:

$$J(\boldsymbol{w}) = \frac{(m_1 - m_2)^2}{{s_1}^2 + {s_2}^2}$$

On remarque que :

$$(m_1 - m_2)^2 = (\boldsymbol{w}^T \boldsymbol{\mu}^{(1)} - \boldsymbol{w}^T \boldsymbol{\mu}^{(2)})^2$$

$$= \boldsymbol{w}^T (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) \, (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})^T \boldsymbol{w}$$

$$= \boldsymbol{w}^T \boldsymbol{S}_{inter} \, \boldsymbol{w}$$

On appelle $\boldsymbol{S}_{inter}$ la matrice de variance interclasses.

Analyse Discriminante Linéaire (ADL)

Par ailleurs, on a:

$$s_1{}^2 = \sum_{i=1}^{N} (\boldsymbol{w}^T x^{(i)} - m_1)^2 y^{(i)}$$

$$= \sum_{i=1}^{N} (\boldsymbol{w}^T x^{(i)} - m_1)(\boldsymbol{w}^T x^{(i)} - m_1)^T y^{(i)}$$

$$= \sum_{i=1}^{N} \boldsymbol{w}^T (x^{(i)} - \boldsymbol{\mu^{(1)}})(x^{(i)} - \boldsymbol{\mu^{(1)}})^T \boldsymbol{w}\, y^{(i)}$$

$$= \boldsymbol{w}^T \boldsymbol{S_1} \boldsymbol{w}$$

On appelle $\boldsymbol{S_1}$ la matrice de variance intra-classe de C_1 .

3. Analyse Discriminante Linéaire (ADL)

De même, on définira la matrice intra-classe de C_2 : $\boldsymbol{S}_2$.

La variance totale intra classe après projection est:

$$s_1{}^2 + s_2{}^2 = \boldsymbol{w}^T \boldsymbol{S}_1 \boldsymbol{w} + \boldsymbol{w}^T \boldsymbol{S}_1 \boldsymbol{w}$$

$$= \boldsymbol{w}^T (\boldsymbol{S}_1 + \boldsymbol{S}_2) \boldsymbol{w}$$

$$= \boldsymbol{w}^T \boldsymbol{S}_{intra}\, \boldsymbol{w}$$

La fonction à maximiser sur $\boldsymbol{w}$ est alors donnée par:

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_{inter}\, \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_{intra}\, \boldsymbol{w}}$$

3. Analyse Discriminante Linéaire (ADL)

- Après la dérivation de $J(\boldsymbol{w})$ par rapport à $\boldsymbol{w}$, on obtient:

$$\boldsymbol{w} \approx \boldsymbol{S}_{intra}{}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})$$

Exemple:

- Pour notre exemple

$$\boldsymbol{S}_{intra} = \boldsymbol{S}_1 + \boldsymbol{S}_2 = 2\begin{bmatrix} 4 & 3 \\ 3 & 3 \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{S}_{intra}{}^{-1} \approx \begin{bmatrix} 0.5 & -0.50 \\ -0.5 & 0.67 \end{bmatrix}$$

$$\boldsymbol{w} = \boldsymbol{S}_{intra}{}^{-1}(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)}) = \begin{bmatrix} 0.5 & -0.50 \\ -0.5 & 0.67 \end{bmatrix}\begin{bmatrix} 4 \\ -3 \end{bmatrix}$$

$$= \begin{bmatrix} 3.5 \\ -4 \end{bmatrix}$$

3. Analyse Discriminante Linéaire (ADL)

Le vecteur $\boldsymbol{w}$ obtenu est en effet celui désiré.

$$w = S_{intra}^{-1}\left(\mu^{(2)} - \mu^{(1)}\right) = \begin{bmatrix} 0.5 & -0.50 \\ -0.5 & 0.67 \end{bmatrix} \begin{bmatrix} 4 \\ -3 \end{bmatrix}$$

$$= \begin{bmatrix} 3.5 \\ -4 \end{bmatrix}$$