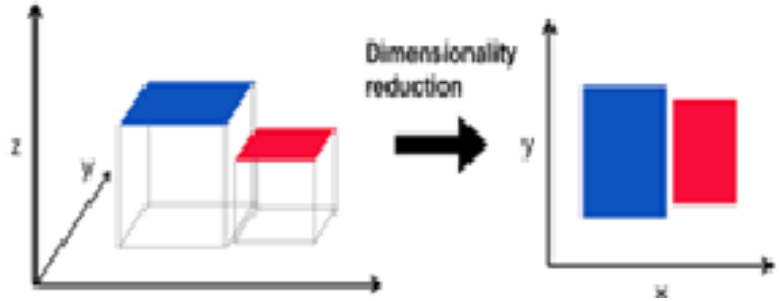
Réduction de la dimensionnalité de données

Dr. HADJADJ ABDELHALIM



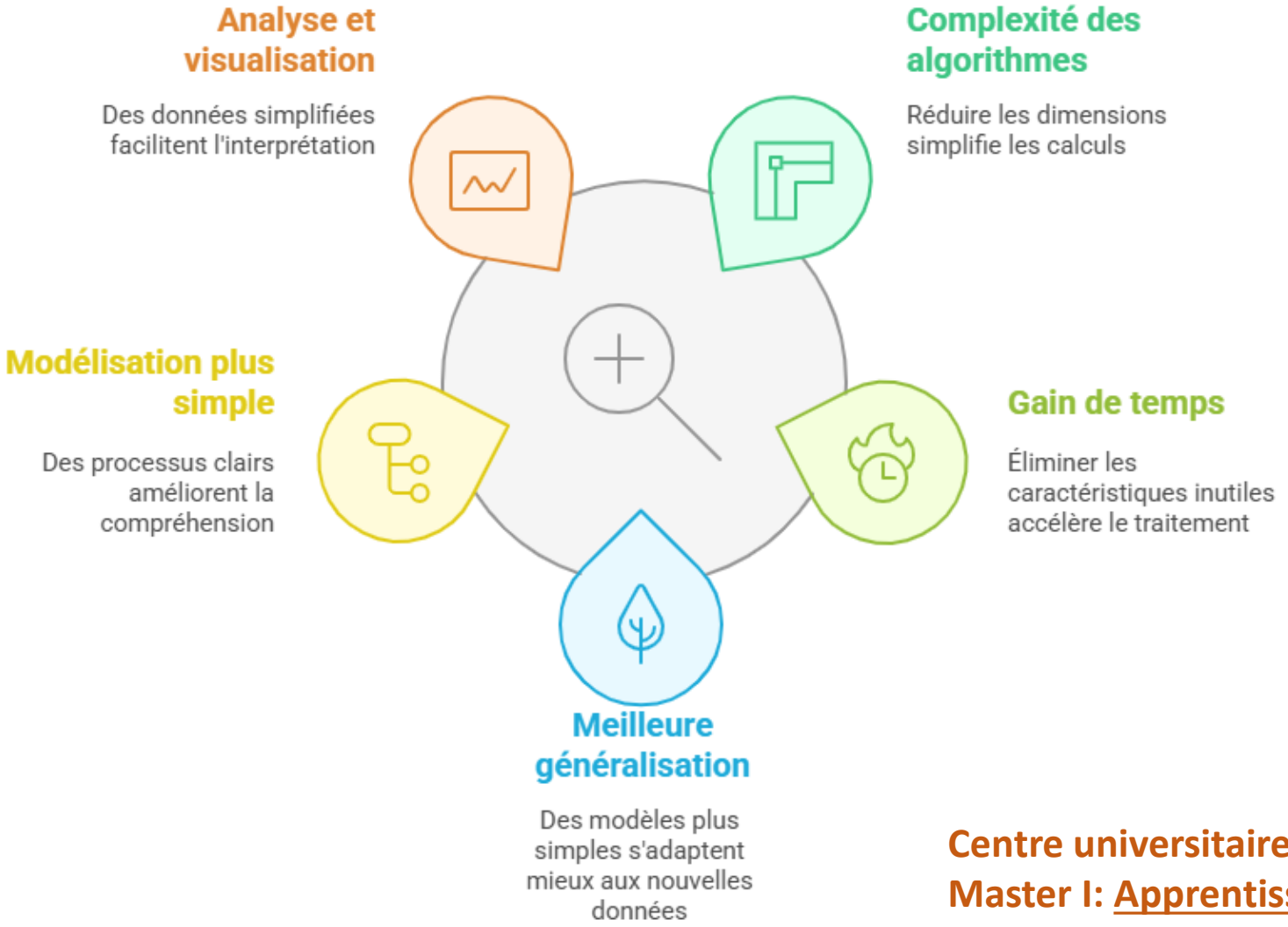
Centre universitaire de Mila
Master I: Apprentissage automatique

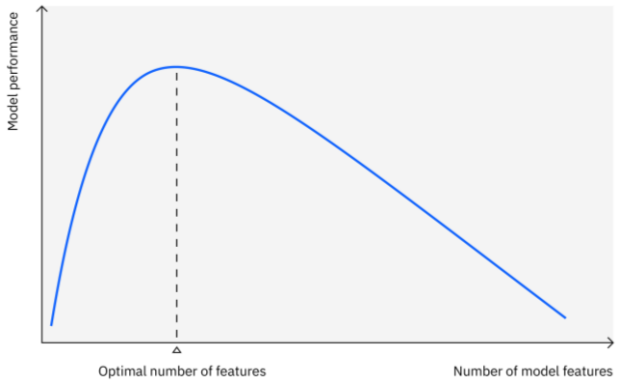
Réduction de la dimensionnalité de données

- 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.

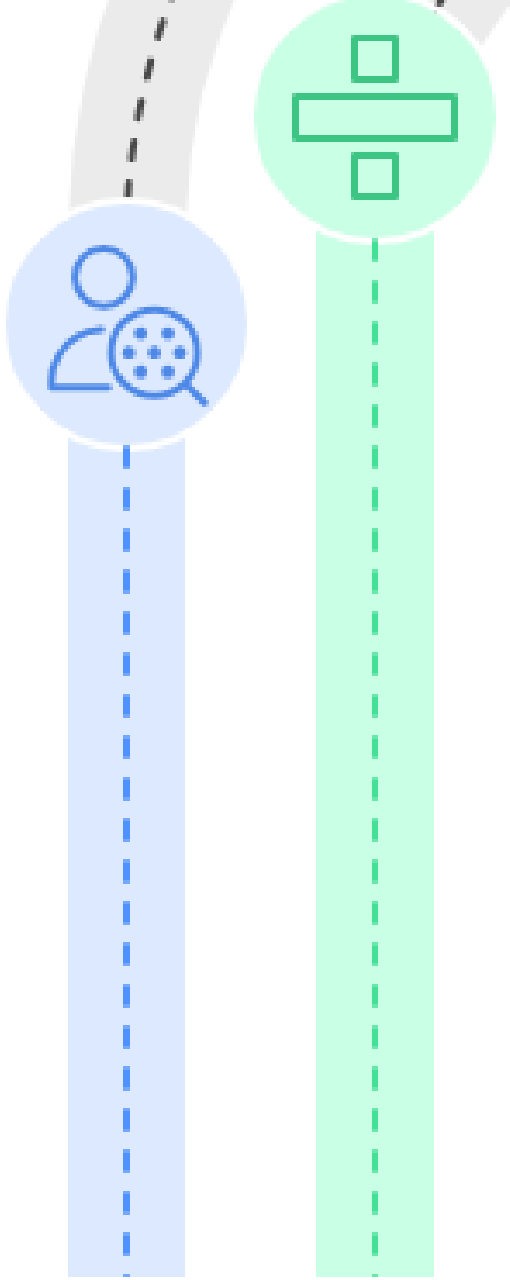
Pourquoi?

- 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 ?



Sélection d'attributs

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

$$\mathbf{x} = [x_1, x_2, \ldots, x_d]$$

$$\downarrow$$

$$\mathbf{z} = [x_{i_1}, x_{i_2}, \ldots, x_{i_k}] \qquad k \ll d$$

Extraction d'attributs

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

$$\mathbf{x} = [x_1, x_2, \ldots, x_d]$$

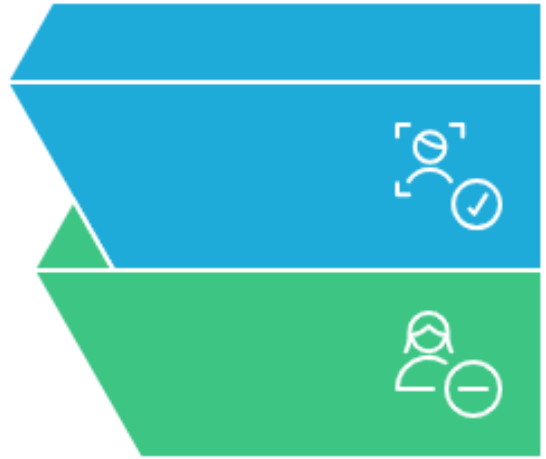
$$\downarrow \quad \mathbf{z} = f(\mathbf{x})$$

$$\mathbf{z} = [z_1, z_2, \ldots, z_k]$$

Principe de sélection d'attributs

- **Élimination des attributs inutiles**
  - Une fois qu'on a identifié ce sous-ensemble optimal,
    tous les autres attributs sont supprimés, car ils sont considérés comme non pertinents ou redondants pour la tâche.
- **Nombre de combinaisons possibles**
  - Si on a D attributs au départ, il existe exactement 2^{D-1} **sous-ensembles possibles** d'attributs.
    - Par exemple, si D=10, alors il y a 2^{10}=1024 sous-ensembles possibles (sauf l'ensemble vide, donc 1023 utilisables).
- **Problème de complexité**
  - Tester **toutes** les combinaisons d'attributs devient **très coûteux en temps** à mesure que D augmente.

Principe de sélection d’attributs



Principe de sélection d'attributs

- **1. Sélection en avant (par ajout progressif)**
    1. On commence **avec aucun attribut** (S = ∅).
    2. À chaque étape, on **ajoute l'attribut(**x_d**)** qui permet de **réduire le plus possible l'erreur de validation** du modèle $E(S \cup x_d)$.
    3. On continue à ajouter des attributs **tant que l'erreur diminue**.
    4. **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**.

$$x_j = argmin_d(E(S \cup x_d))$$

- ✓ 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.

Principe de sélection d’attributs

- **Sélection 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**.

Exemple 1

- 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.

Exemple 2

- S = {X1, X2, X3, X4}
- On entraîne un modèle avec tous les attributs disponibles :→ Erreur de validation = **0.25**

Attribut supprimé	Nouvel ensemble S'	Erreur de validation
X1	{X2, X3, X4}	**0.23** ✅ amélioration
X2	{X1, X3, X4}	0.27
X3	{X1, X2, X4}	0.24
X4	{X1, X2, X3}	0.26

- On choisit de **supprimer X1** (erreur = 0.23)

Extraction d’attributs: Analyse à composantes principales (ACP)

Centre universitaire de Mila
Master I: Apprentissage automatique

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 (ACP)**
  - L'**Analyse Discriminante (AD)**
  - L'**Analyse Factorielle (AF)**
  - L'**Analyse de Corrélation (AC)**

Analyse à composantes principales (ACP)

- L'ACP est une **méthode d'analyse non supervisée**, ce qui signifie qu'elle ne fait appel à aucune variable cible ou de sortie.
- Elle vise à **réduire la dimensionnalité** des données tout en **préservant un maximum de variance**, c'est-à-dire l'information la plus représentative du jeu de données.
- Le principe consiste à projeter les données sur de **nouvelles directions (appelées composantes principales)** qui capturent la plus grande variance possible.
  - Soit x un vecteur aléatoire représentant un point dans l'espace des données initial X.
  - On cherche une direction de projection $w^{(1)}$ telle que $\|w_{(1)}\|=1$ (norme unitaire).
  - La **première composante principale** est donnée par la projection de x sur cette direction :

$$z_1=(w_{(1)})^T x$$

où z_1 est une nouvelle variable représentant la donnée projetée.

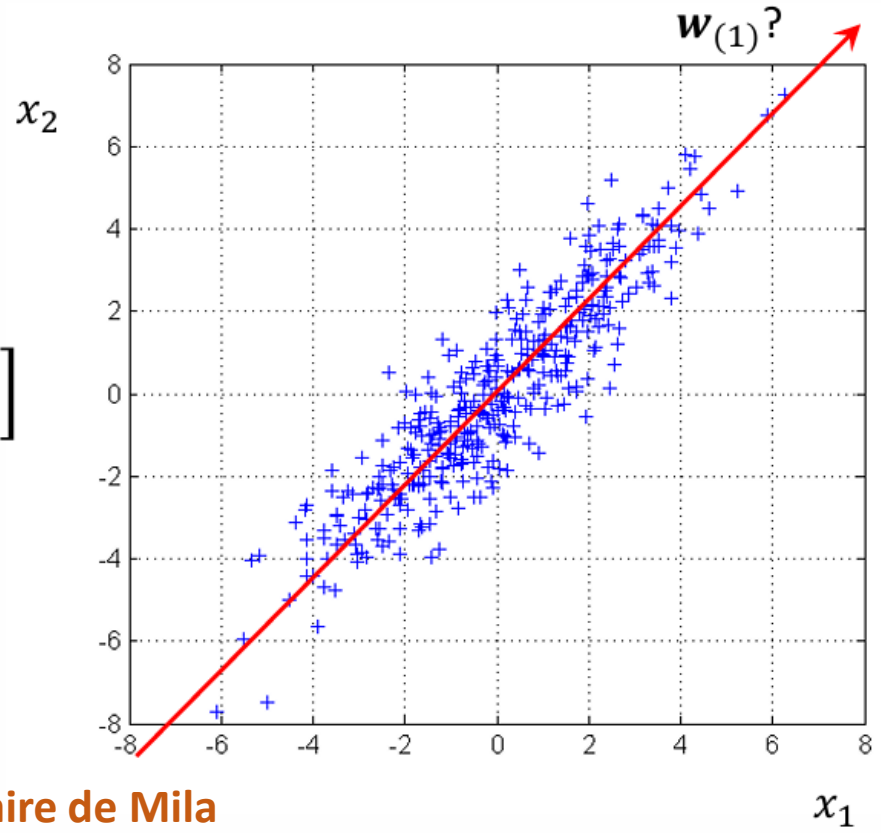
Analyse à composantes principales (ACP)

Exemple:

Quelle est la direction de projection qui maximise la variance?

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

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



Analyse à composantes principales (ACP)

- Soit :
  - M la **moyenne** du vecteur de données x,
  - Σ la **matrice de covariance** associée à x.
  - On peut démontrer que la **variance** de la première composante principale z_1 est donnée par :

$$\mathrm{Var}(z_1) = (\mathbf{w}^{(1)})^T \Sigma \mathbf{w}^{(1)}$$

- L'objectif est de trouver un vecteur $w^{(1)}$ qui :
  - **maximise** cette variance,
  - **satisfait la contrainte** : $\|w^{(1)}\|=1$ (c'est-à-dire une norme unitaire).
- 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 prenant égale à 0 la dérivée de la fonction par rapport à $\boldsymbol{w}_{(1)}$, on obtient:

$$2\boldsymbol{\Sigma}\,\boldsymbol{w}_{(1)} - 2\lambda\boldsymbol{w}_{(1)} = 0 \quad \Rightarrow \quad \boldsymbol{\Sigma}\,\boldsymbol{w}_{(1)} = \lambda\boldsymbol{w}_{(1)}$$

- On remarque que $\boldsymbol{w}_{(1)}$ correspond au vecteur propre de la matrice $\boldsymbol{\Sigma}$ et λ est une valeur propre de la même matrice.

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

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

Analyse à composantes principales (ACP)

•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.

• Pour déterminer la **deuxième direction principale** $w^{(2)}$, on cherche à :

•Maximiser la variance projetée :

$$\mathrm{Var}(z_2) = (\mathbf{w}^{(2)})^T \Sigma \mathbf{w}^{(2)}$$

- Sous les contraintes suivantes :
  - $\|w^{(2)}\|=1$ (norme unitaire),
  - $(w^{(1)})^T w^{(2)}=0$ (orthogonalité avec la première direction).
- Par un raisonnement similaire au premier cas, on peut démontrer que :
  - $w^{(2)}$ est le vecteur propre correspondant à la **deuxième plus grande valeur propre** de Σ.
- Cette procédure peut être **répétée** pour extraire les **M premières composantes principales**, chacune étant :
  - **Orthogonale** aux précédentes,
  - Associée à une **valeur propre décroissante**,
  - Et **maximisant la variance restante** à chaque étape.

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.

Analyse en Composantes Principales (ACP)

- Pour conserver environ **90 % de la variance totale**, on choisit le plus petit nombre M tel que :

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_M}{\sum_{d=1}^{D} \lambda_d} \approx 0{,}9$$

- Le **graphe de Scree** (ou éboulis) permet de visualiser :
  - Le **pourcentage de variance expliquée** par chaque composante principale,
  - En fonction du **nombre de composantes gardées**.
  - Il aide à choisir un MMM optimal en repérant un "coude" dans la courbe.
- Soit W une matrice dont les colonnes sont les **M premières composantes principales** $w^{(1)},\ldots,w^{(M)}$
- La transformation des données se fait par : $\mathbf{z} = \mathbf{W}^T(\mathbf{x} - \boldsymbol{\mu})$

  - x est une donnée d'origine (en dimension D),
  - μ est le **vecteur moyen** de l'ensemble des données,
  - Z est la représentation projetée dans un **espace de dimension réduite M**,
  - Cette transformation préserve au maximum la variance tout en réduisant la dimension.

Analyse en Composantes Principales (ACP)

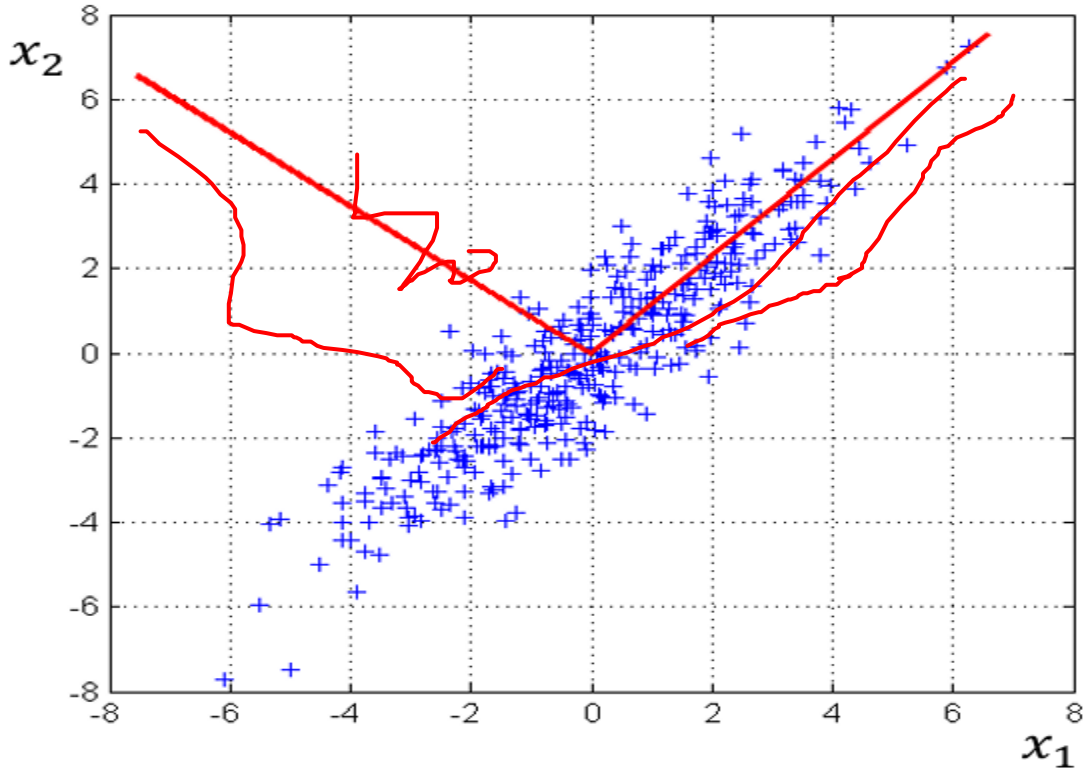
L'ACP donne les directions principales comme suit:

$$\mathbf{w} = \begin{bmatrix} 0.67 & -0.74 \\ 0.74 & 0.67 \end{bmatrix}$$

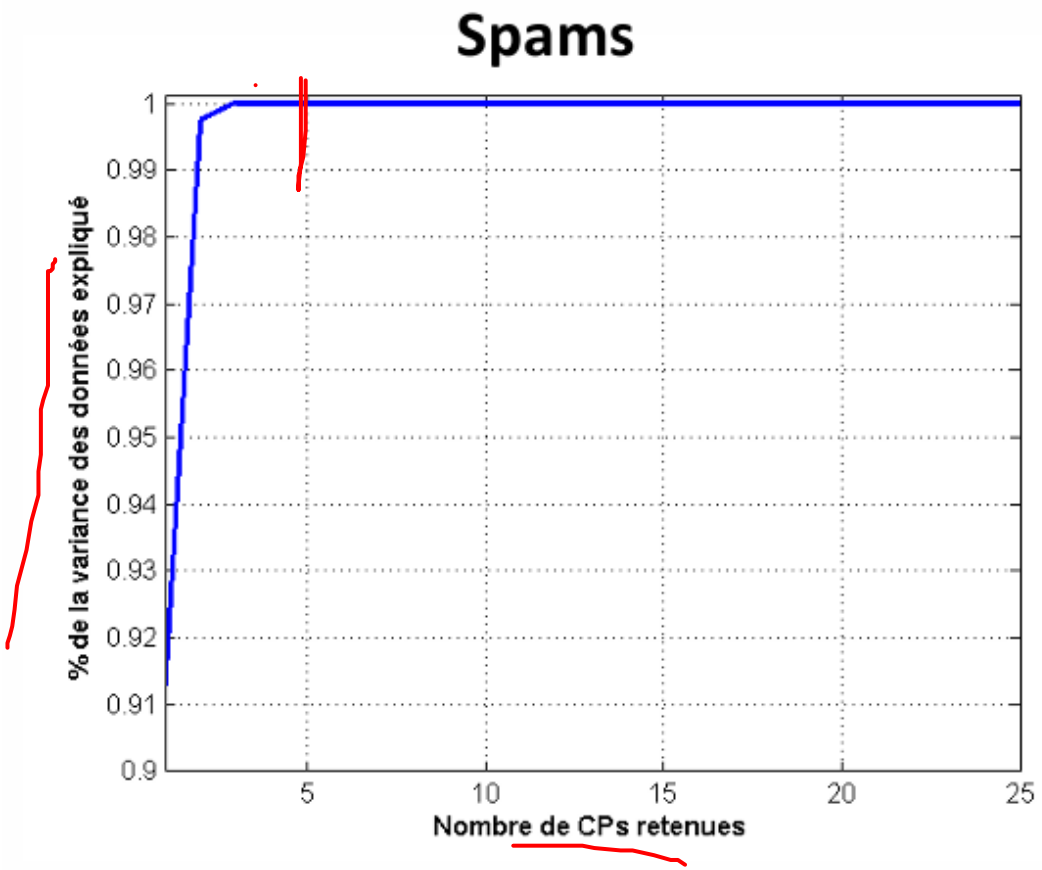
$w_{(1)}$ $w_{(2)}$

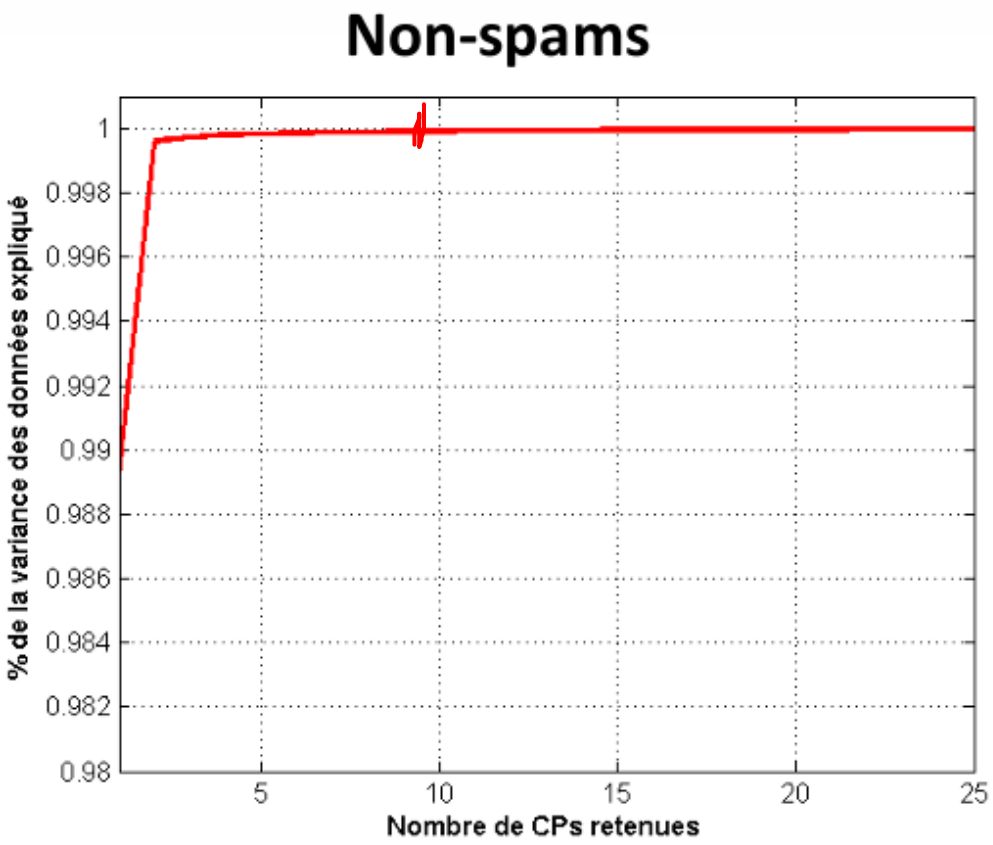
$\lambda_1 = 8.76.$

$\lambda_2 = 0.48$



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





Analyse discriminante linéaire (ADL)

Centre universitaire de Mila
Master I: Apprentissage automatique

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: $(D = 2)$

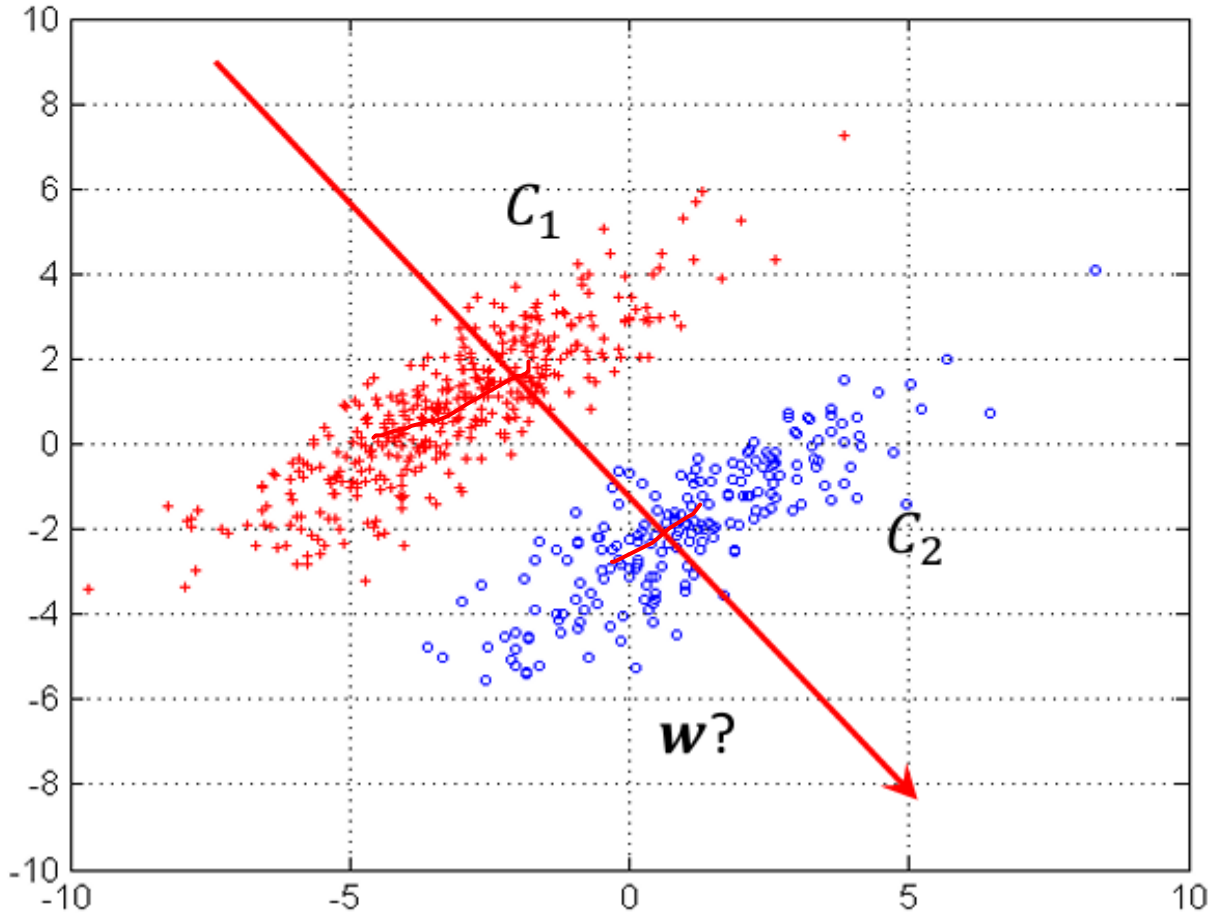
Quelle est la direction de projection $\boldsymbol{w}$ qui maximise la discrimination entre les deux classes C_1 et C_2 ?

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

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

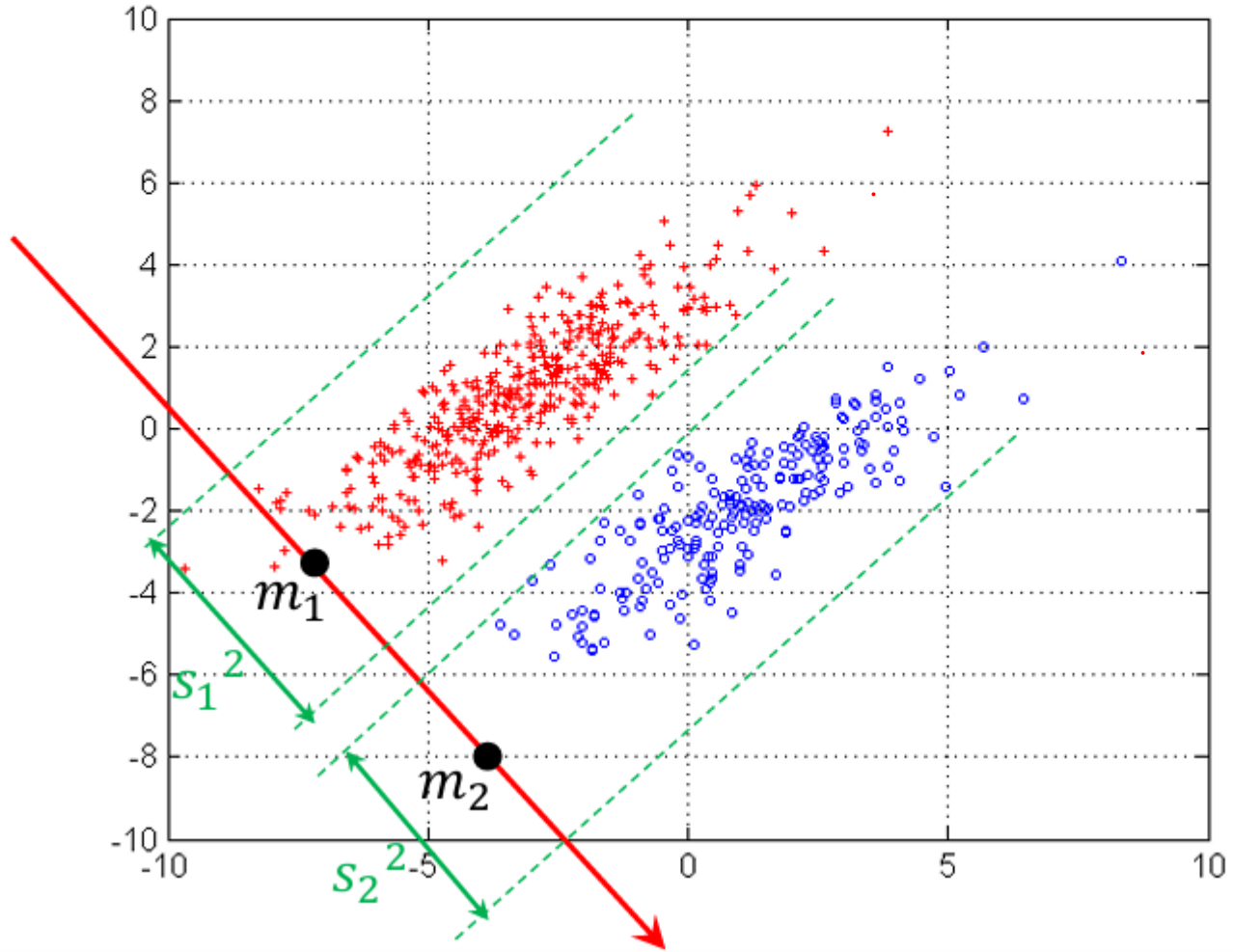
$$\boldsymbol{\Sigma}^{(1)} = \begin{bmatrix} 4 & 3 \\ 3 & 3 \end{bmatrix}$$

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



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.

$\boldsymbol{w}$?



Analyse Discriminante Linéaire (ADL)

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

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

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

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

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 .

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:

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

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

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

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

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

Analyse Discriminante Linéaire (ADL)

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