Analyse, Classification et Indexation des données (ACID)

Réduction de dimension, Partie 2 : LDA

Akka Zemmari

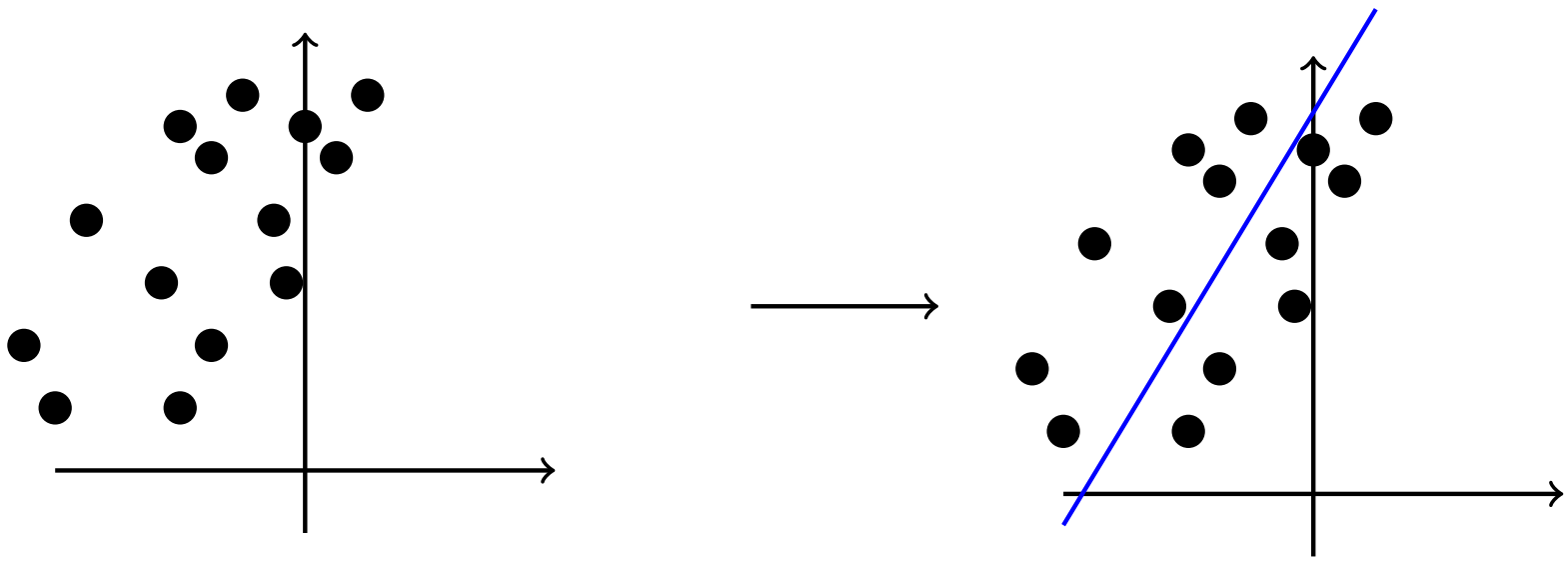
LaBRI, Université de Bordeaux

2024 - 2025

Motivation

Limites de l'ACP

- L'ACP est conçue pour représenter des données, et non pour les classifier,
  - elle préserve la variance autant que possible :

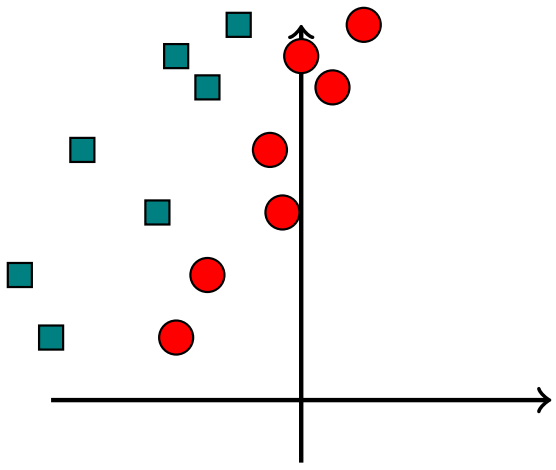


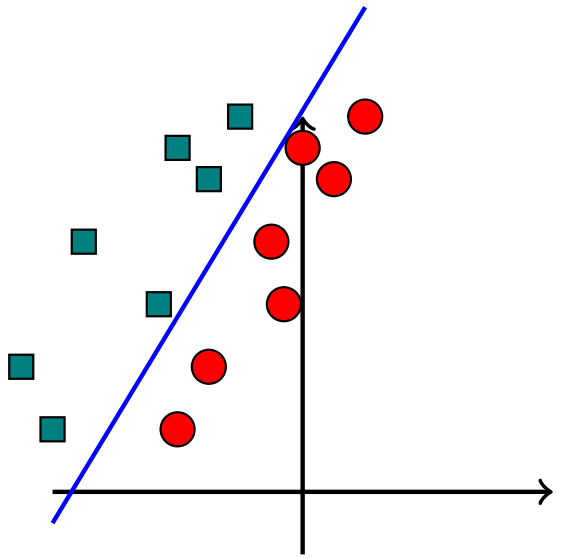
  - si les directions à plus grandes variances coïncident avec les bons choix pour la classification alors elle peut servir.

Motivation

Limites de l'ACP

- L'ACP est conçue pour représenter des données, et non pour les classifier,
- En général, la direction conservant la plus grande variance peut ne pas être avantageuse pour la classification :

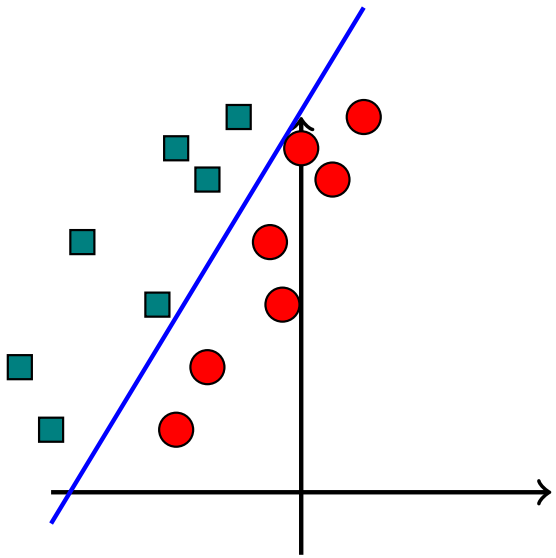


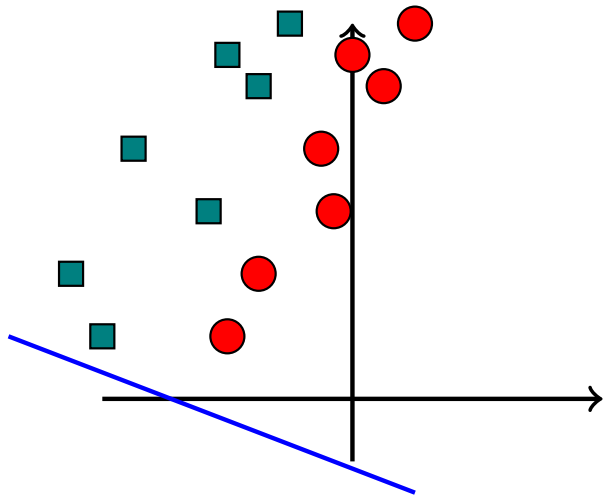


Motivation

Idée de base

- Préférer une projection sur une ligne concervant la séparation entre les classes :



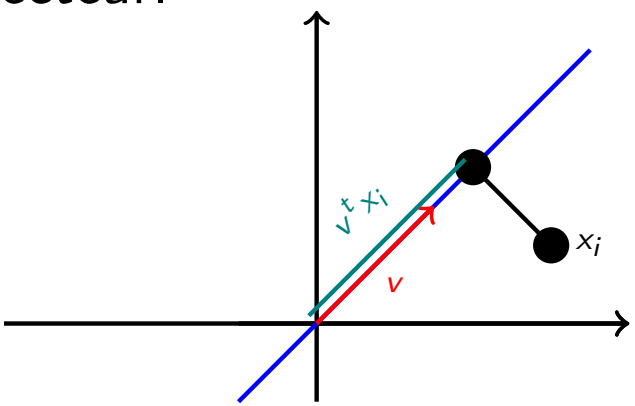


Formalisation

- Supposons que l'on dispose d'un corpus avec n enregistrements $\{x_1, x_2, \cdots, x_n\}$, chacun de dimension d.
- Chaque élément x_i appartient à une classe.
- Le nombre de classes est 2 (classification binaire).
  - n_1 éléments appartiennent à la première classe,
  - n_2 éléments appartiennent à la seconde classe.

Formalisation

- Supposons que l'on projette chaque élément x_i du corpus sur une droite passant par l'origine et ayant le vecteur unitaire v comme vecteur directeur.



  - Le réel $v^t x_i$ représente la distance entre la projection de x_i et l'origine,
  - $v^t x_i$ est la projection de x_i sur un sous-espace de dimension 1.

Que cherche-t-on à faire ?

Questions

- Comment mesurer la qualité de la projection en terme de séparation des classes ?
  - Soit μ_1 et μ_2 les moyennes des éléments classés 1 et 2 resp.
  - Soit $\widetilde{\mu_1}$ et $\widetilde{\mu_2}$ leurs projections sur le nouvel axe :

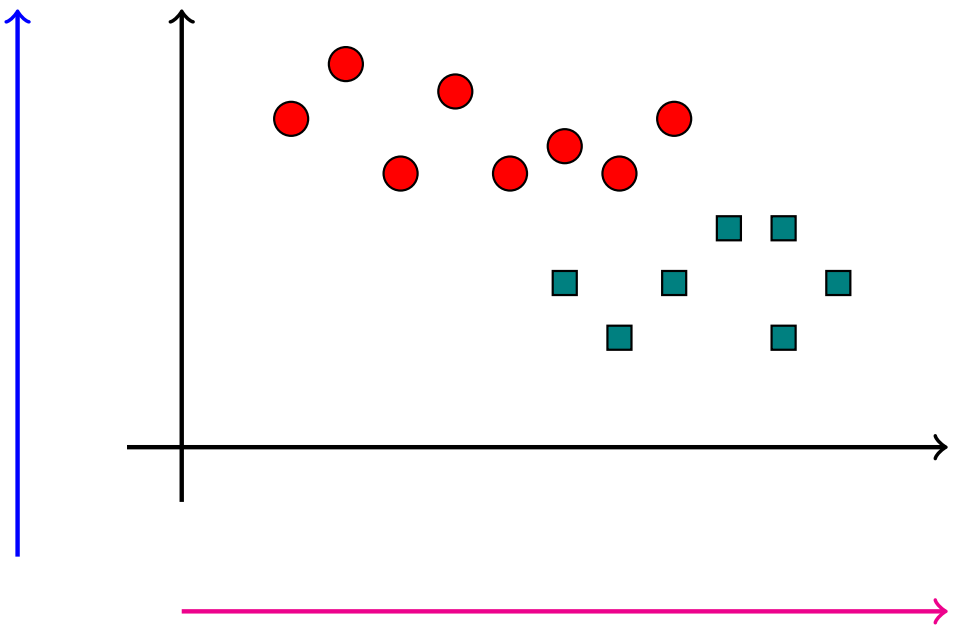
$$\widetilde{\mu_i} = \frac{1}{n_i} \sum_{x_j \in C_i} v^t x_j = v^t \left(\frac{1}{n_i} \sum_{x_j \in C_i} x_j \right) = v^t \mu_i$$

- Intuitivement, on a envie de choisir un axe tel que $\mid \widetilde{\mu_1} - \widetilde{\mu_2} \mid$ soit la plus grande possible (pourquoi ?)

A la recherche de la séparation ...

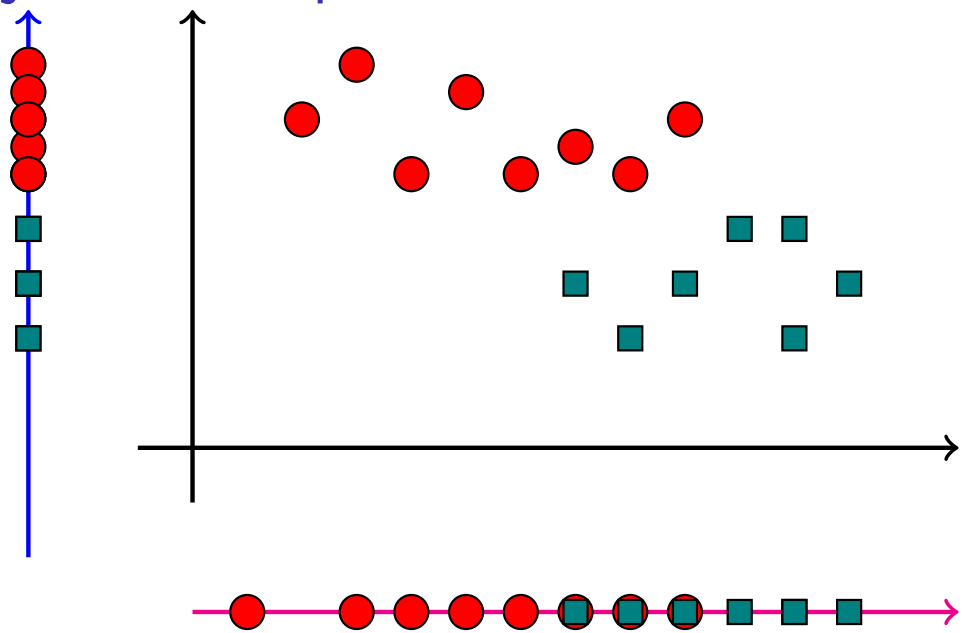
Est-ce toujours vrai ?

- Entre les deux axes suivants, lequel vous choisissez ? (D_1) ou (D_2) ? Pourquoi ?



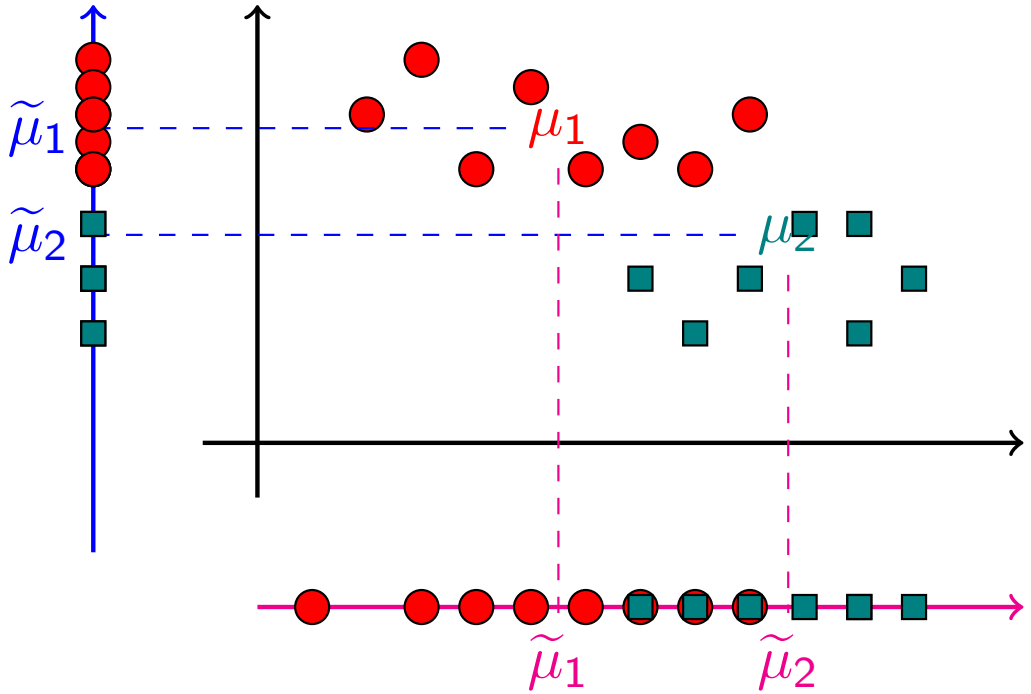
A la recherche de la séparation ...

Voyons la projection des points ...



A la recherche de la séparation ...

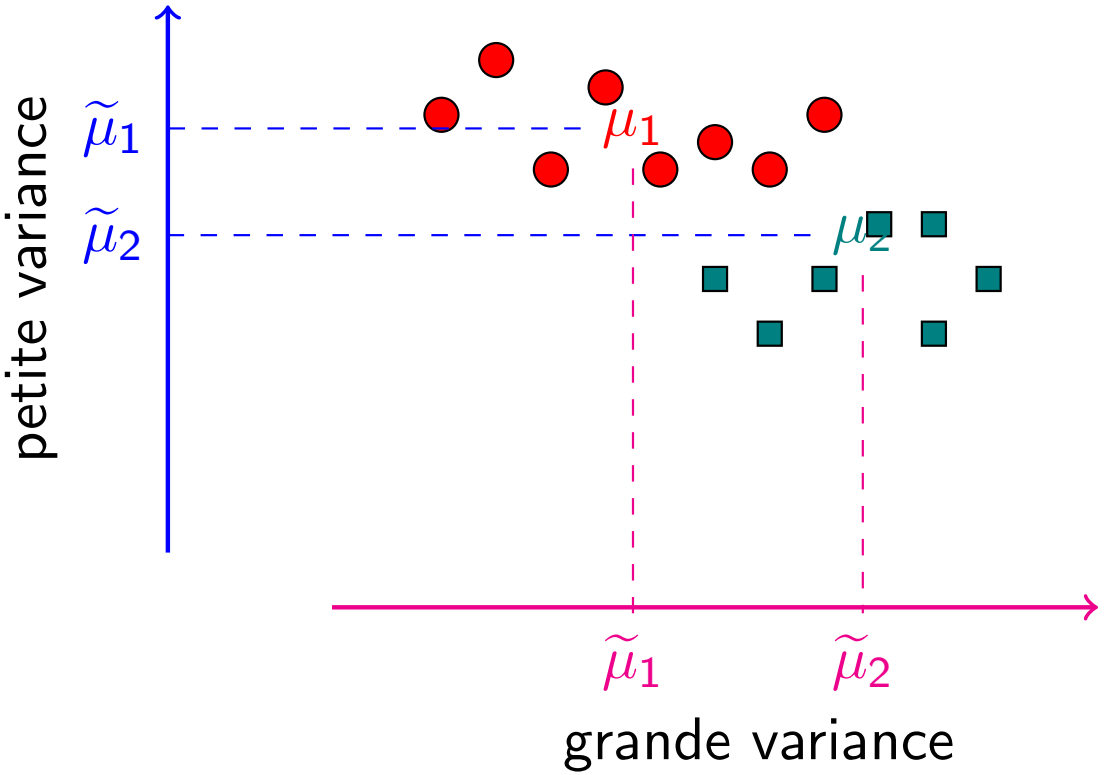
Voyons la projection des points, des moyennes et les distances :



L'axe bleu semble meilleur que l'axe violet. Il préserve mieux la séparation des classes ... alors que les distances entre les projections des moyennes ne donnent pas cette impression.
Question : Qu'est-ce qu'on n'a pas pris en compte ?

A la recherche de la séparation ...

Observons un peu plus le nuage de points



On doit normaliser (diviser) le terme $\mid \widetilde{\mu}_1 - \widetilde{\mu}_2 \mid$ par un terme proportionnel à la variance

Solution de Fisher

- On doit normaliser (diviser) le terme $|\ \widetilde{\mu}_1 - \widetilde{\mu}_2\ |$ par un terme proportionnel à la variance,
- Solution de Fisher : normaliser $|\ \widetilde{\mu}_1 - \widetilde{\mu}_2\ |$ en la divisant par la matrice de dispersion (scatter) :
  - Soit $y_j = v^t x_j$ les projections des éléments du corpus,
  - la matrice de dispersion des projections des exemples de la classe C_i est donné par :

$$\widetilde{s}_i^2 = \sum_{y_j \in C_i} (y_j - \widetilde{\mu}_i)^2 .$$

Discriminant de Fisher

- On doit normaliser par la matrice de dispersion des deux classes
- La solution de Fisher est de projeter les éléments du corpus sur la ligne dans la direction v qui maximise :

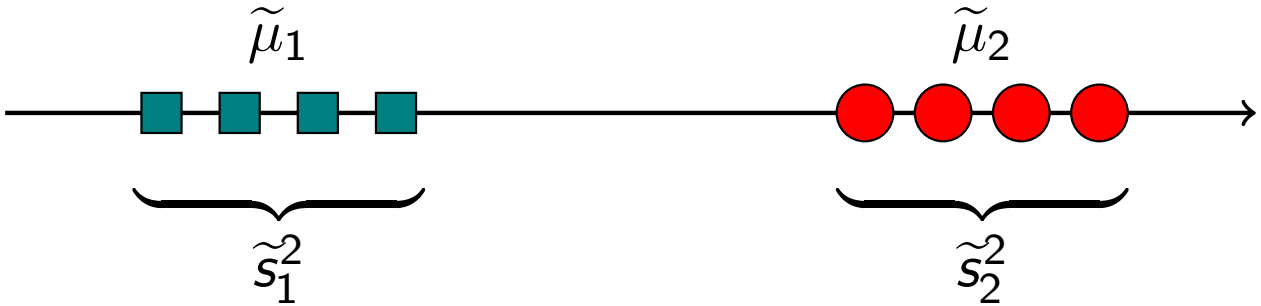
$$J(v) = \frac{(\widetilde{\mu}_1 - \widetilde{\mu}_2)^2}{\widetilde{s}_1^2 + \widetilde{s}_2^2}$$

- Question : quelle est l'intuition ?

Discriminant de Fisher

$$J(v) = \frac{(\widetilde{\mu}_1 - \widetilde{\mu}_2)^2}{\widetilde{s}_1^2 + \widetilde{s}_2^2}$$

Sur une figure :



A la recherche de la meilleur direction

$$J(v) = \frac{(\widetilde{\mu}_1 - \widetilde{\mu}_2)^2}{\widetilde{s}_1^2 + \widetilde{s}_2^2}$$

- On doit exprimer J en fonction de v et chercher à la maximiser,
- On définit les matrices de dispersion de chacune des classes :

$$S_1 = \sum_{x_j \in C_1} (x_j - \mu_1)(x_j - \mu_1)^t, \quad S_2 = \sum_{x_j \in C_2} (x_j - \mu_2)(x_j - \mu_2)^t,$$

A la recherche de la meilleur direction

$$J(v) = \frac{(\widetilde{\mu}_1 - \widetilde{\mu}_2)^2}{\widetilde{s}_1^2 + \widetilde{s}_2^2}$$

- On définit la matrice de dispersion intra-classes :

$$S_w = S_1 + S2$$

- on montre que :

$$\widetilde{s}_1^2 = v^t S_1 v \quad \text{et} \quad \widetilde{s}_2^2 = v^t S_2 v$$

Ainsi :

$$\widetilde{s}_1^2 + \widetilde{s}_2^2 = v^t S_w v$$

A la recherche de la meilleur direction

$$J(v) = \frac{(\widetilde{\mu}_1 - \widetilde{\mu}_2)^2}{\widetilde{s}_1^2 + \widetilde{s}_2^2}$$

- On définit à présent la matrice de dispersion inter-classes :

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^t$$

S_B mesure alors la séparation entre les deux moyennes des classes (avant projection).

- on montre que :

$$(\widetilde{\mu}_1 - \widetilde{\mu}_2)^2 = v^t S_B v.$$

Ainsi :

$$J(v) = \frac{v^t S_B v}{v^t S_w v}$$

A la recherche de la meilleur direction

$$J(v) = \frac{v^t S_B v}{v^t S_w v}$$

- Il faut donc résoudre l'équation

$$\frac{d}{dv} J(v) = 0$$

- Comme pour l'ACP, on résout l'équation et on trouve que la meilleur direction est celle donnée par le vecteur :

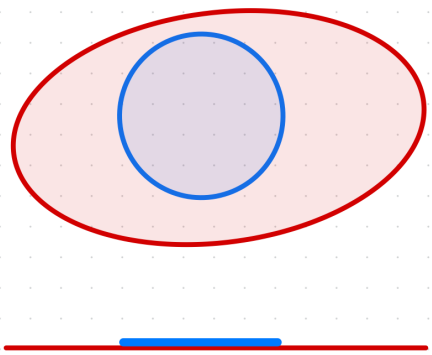
$$v = S_w^{-1} (\mu_1 - \mu_2)$$

Multiple Discriminant Analysis (MDA)

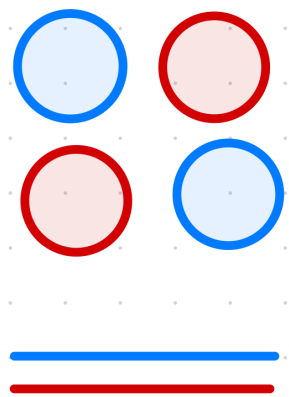
- Généralise la LDA à plusieurs classes
- Si on a k classes, on peut réduire la dimension à $1, 2, \cdots, k-1$ dimensions

LDA (et MDA) vs ACP

- la LDA-MDA réduit la dimension à $k-1$ au maximum (pour k classes), ce qui n'est pas le cas de l'ACP.
- la LDA-MDA peut échouer si $J(v)=0$



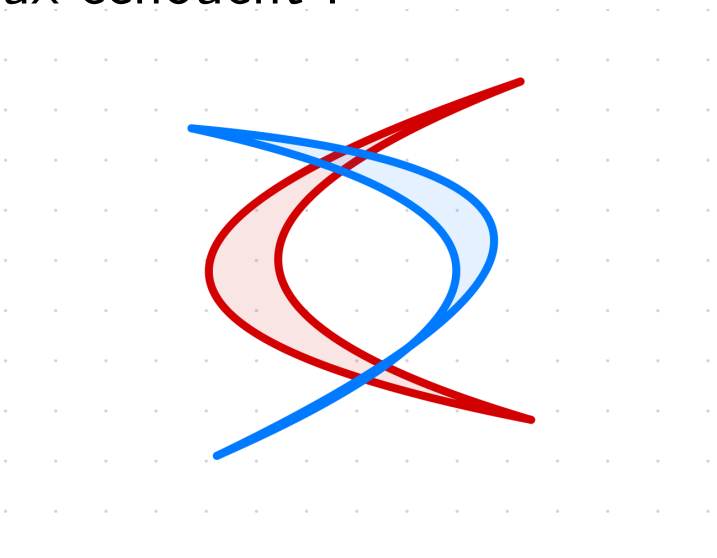
ACP est meilleure



Les deux n'y arrivent pas

LDA (MDA) vs ACP

Autre cas où les deux échouent :



Les classent s'entremêlent, $J(v)$ est trop grande.

Exemple

Voir le jupyter notebook