|  |
| --- |
| Université Lille 1 – Master 1 |
| RDF – TP n°07 |
| Réduction de la dimension par analyse en composantes principales et analyse factorielle discriminante |

|  |
| --- |
| BARCHID Sami – SLIMANI Anthony  10/03/2019 |

# Introduction

L’objectif de ce TP est d’apprendre la réduction de l’espace de représentation en identifiant un axe de projection optimal. Pour ce faire, nous avons vu deux méthodes dans ce TP : l’analyse en composantes principales (non supervisé) et l’analyse factorielle discriminante (supervisé).

Ce rapport est divisé en 3 parties :

* **Analyse des ensembles de données** : présentation et analyse des données multi-variées utilisées pour ce TP.
* **Analyse en composantes principales :** expérimentation et interprétations de l’espace de représentation obtenu par analyse en composantes principales puis sur la classification résultante grâce à l’analyse linéaire discriminante.
* **Analyse factorielle discriminante :** expérimentation et interprétations de l’espace de représentation obtenu par analyse factorielle discriminante puis sur la classification résultante grâce à l’analyse linéaire discriminante.

# Analyse des ensembles de données

Durant ce TP, nous avons eu à disposition deux ensembles :

* Un ensemble d’apprentissage présenté en figure 1.
* Un ensemble de test présent en figure 2.

Le but pour la suite du TP sera de réduire l’espace de représentation selon un axe optimale selon une des deux techniques présentées aux point suivants. Ensuite, nous appliquerons une classification par analyse linéaire discriminante sur les données d’apprentissage projetées dans le nouvel espace de représentation trouvé. Et enfin, nous appliquerons la classification identifiée sur les données de test projetées dans le nouvel espace de représentation calculé.

|  |  |
| --- | --- |
| Figure – DONNÉES D’APPRENTISSAGE | Figure – DONNÉES DE TEST |

Les données sont multivariées où nous avons D, le nombre d’attributs, qui vaut 2. L’objectif ici est donc de réduire D en un d=1 (donc passer d’un espace de représentation en deux dimensions à un espace à une seule dimension).

Nous avons les classes suivantes :

* Classe 1 en rouge
* Classe 2 en vert
* Classe 3 en bleu

Nous avons ensuite analysé les relations entre les deux attributs étudiés en calculant la matrice de covariance sur les données d’apprentissage :

Cette matrice de covariance indique deux choses :

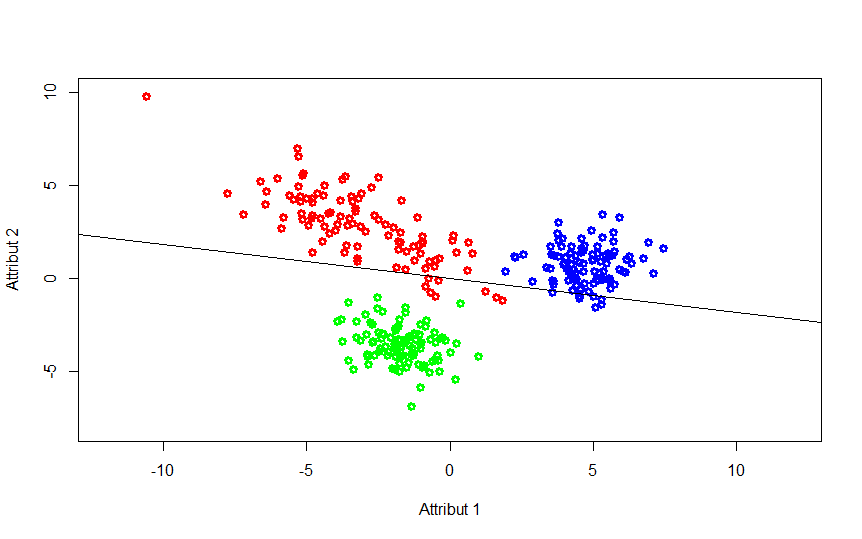
* Aucun attribut est indépendant (rien n’est nul dans la matrice de covariance)
* = (car une matrice de covariance est symétrique) et ces deux valeurs sont négatives. Ce qui veut dire que, si l’attribut 1 augmente, l’attribut 2 va diminuer et vice-versa.

# Analyse en composantes principales

Pour cette partie du TP, nous sommes dans le cadre d’un apprentissage non-supervisé. Nous allons donc réduire l’espace de représentation grâce à l’analyse en composantes principales (ACP).

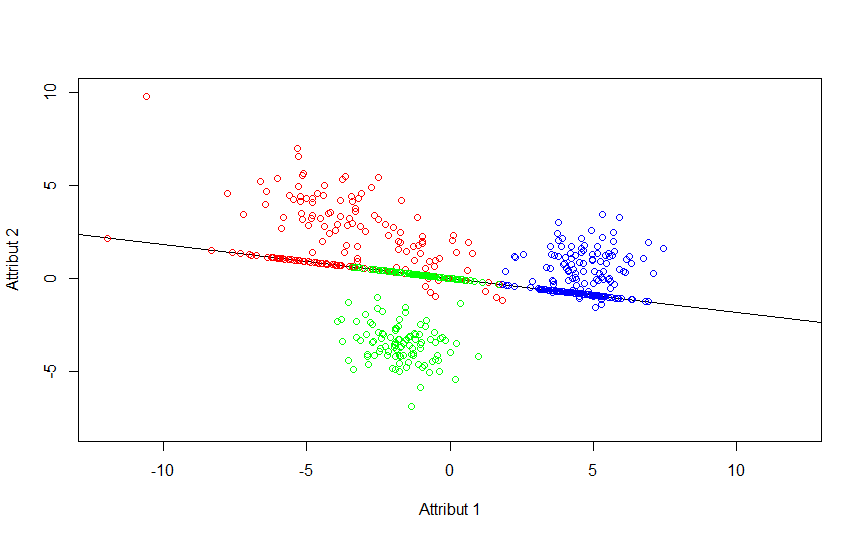
### ACP sur les données d’apprentissage

Nous avons commencé par appliqué l’ACP (analyse en composantes principales) sur les données d’apprentissage pour fixer un axe principal le plus discriminant possible. Le résultat obtenu est montré sur l’image suivante :

 Figure 3 – DONNÉES D’APPRENTISSAGE AVEC AXE PRINCIPALE

La droite vue ici est l’axe principale (qui est la droite dont le vecteur directeur est le vecteur propre principale de la matrice de covariance pour les données d’apprentissage). Cette droite sera notre nouvel espace de représentation obtenu par analyse de composantes principales.

En projetant les données d’apprentissage sur cet axe, nous obtenu l’image suivante :

 Figure 4 – DONNÉES D’APPRENTISSAGE PROJETÉES SUR L’AXE PRINCIPALE

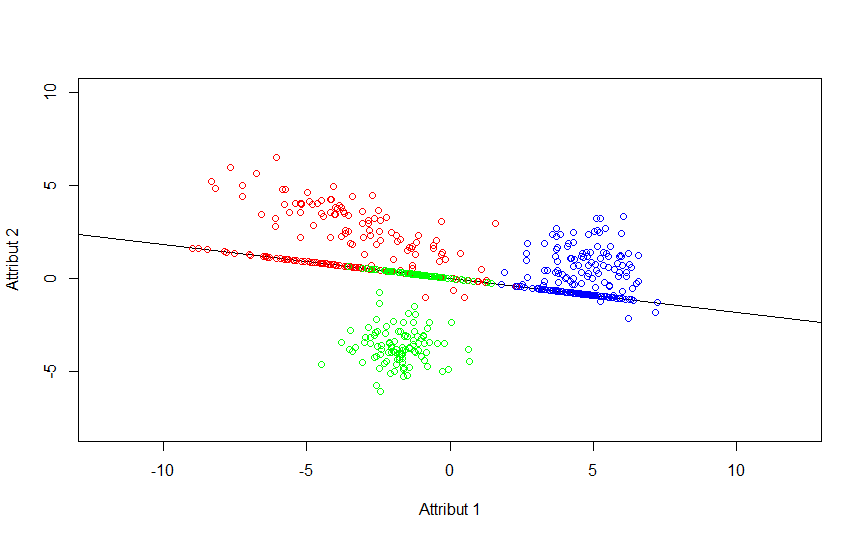
En observant la projection des données d’apprentissage sur l’axe, nous observons que les données peuvent être, en majorité, discriminées convenablement. En effet, nous observons bien des regroupements des observations projetées pour chaque classe sur le nouvel espace de représentation (l’axe).

Cependant, nous remarquons déjà qu’il y aura une portion d’erreurs puisque, certaines observations d’une classe seront projetées sur une autre classe (par exemple, les points rouges au centre seront projetés dans la classe 2 en vert).

### Classification par ALD des données de test projetées

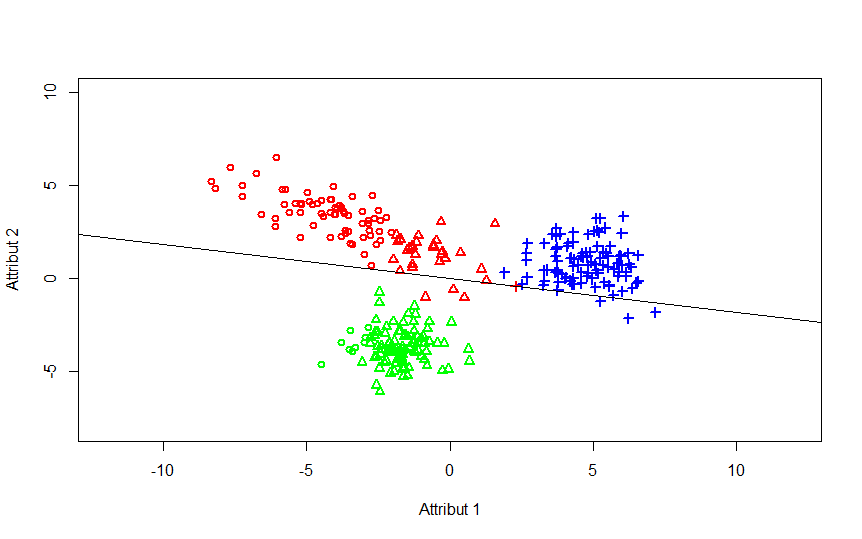
Nous allons appliquer la règle de décision de l’ALD apprise avec les données d’apprentissage sur les données de test projetées sur l’axe principale calculé auparavant.

Tout d’abord, nous projetons les données de tests sur l’espace de représentation obtenu sur les données d’apprentissage, ce qui nous donne l’image suivante :

 Figure 5 – DONNÉES DE TEST PROJETÉES SUR L’AXE PRINCIPALE

Nous remarquons que, comme pour les données d’apprentissage, les données de test pourront être discriminées convenablement.

Nous appliquons, donc, l’analyse linéaire discriminante sur les données de test projetées, ce qui donne la figure suivante :

 Figure 6 – ANALYSE LINÉAIRE DISCRIMINANTE DES DONNÉES DE TEST PROJETÉES SUR L’AXE PRINCIPALE

Dans la figure 6, les classes originales des observations sont données par la couleur tel que (rien ne change par rapport à avant) :

* Classe 1 en rouge
* Classe 2 en vert
* Classe 3 en bleu

Et les classes trouvées par classification suivant l’ALD sont données de la manière suivante :

* Classe 1 donnée par des cercles
* Classe 2 donnée par des triangles
* Classe 3 donnée par des croix (+)

Ainsi, nous observons plusieurs choses :

* Des observations de la classe 1 (en rouge) ont été identifiées comme des éléments de la classe 2 (triangle) et inversement. Ces points sont ceux qui étaient projetés (dans la figure 5) dans les regroupements de classes autres
* Les observations de la classe 3 (en bleu) ont tous été bien classifiés. On remarque aussi qu’une observation de la classe rouge (très éloignée du reste de sa classe) a été identifiée comme appartenant à la classe 3.

Nous pouvons, ensuite, calculer les taux de bonne classification des données de test. Le taux de bonne classification est de 0.866666 0.87 donc **87%**.

Et pour chaque classe, le taux de bonne classification est :

* Classe 1 : 69%
* Classe 2 : 91%
* Classe 3 : 100%

Nous pouvons donc confirmer que les données ont pu être correctement discriminées avec un taux de bonne classification correcte (mais pas exceptionnel non plus, 13% d’erreur ça fait beaucoup). Cependant, la classe 1 a été celle avec le plus d’erreurs et donne un taux de bonne classification qui, lui, n’est pas satisfaisant.

# Analyse factorielle discriminante

Ici, nous sommes dans le contexte d’un apprentissage supervisé (contrairement à la partie précédente). Nous avons donc cherché à réduire l’espace de représentation en utilisant l’analyse factorielle discriminante, qui maximise le critère de Fisher.