Outils d'indexation et de classification de l'information

Tony GABOULAUD & Yanis MOUNSAMY

January 4, 2016

1 Introduction

Au cours de ce semestre, nous avons, en binôme, implémenté un ensemble de méthodes afin de discriminer des données reçus en entrée selon un ou plusieurs descripteurs (1D, 2D ou 3D). Pour cela, nous avons utilisé le langage MATLAB qui est parfaitement adapté aux calculs numériques notamment matricielles.

Le résultat pour chacune des méthodes est visualisé dans une fenêtre où l'on compare deux à deux les différentes classes (sur un total de 4).

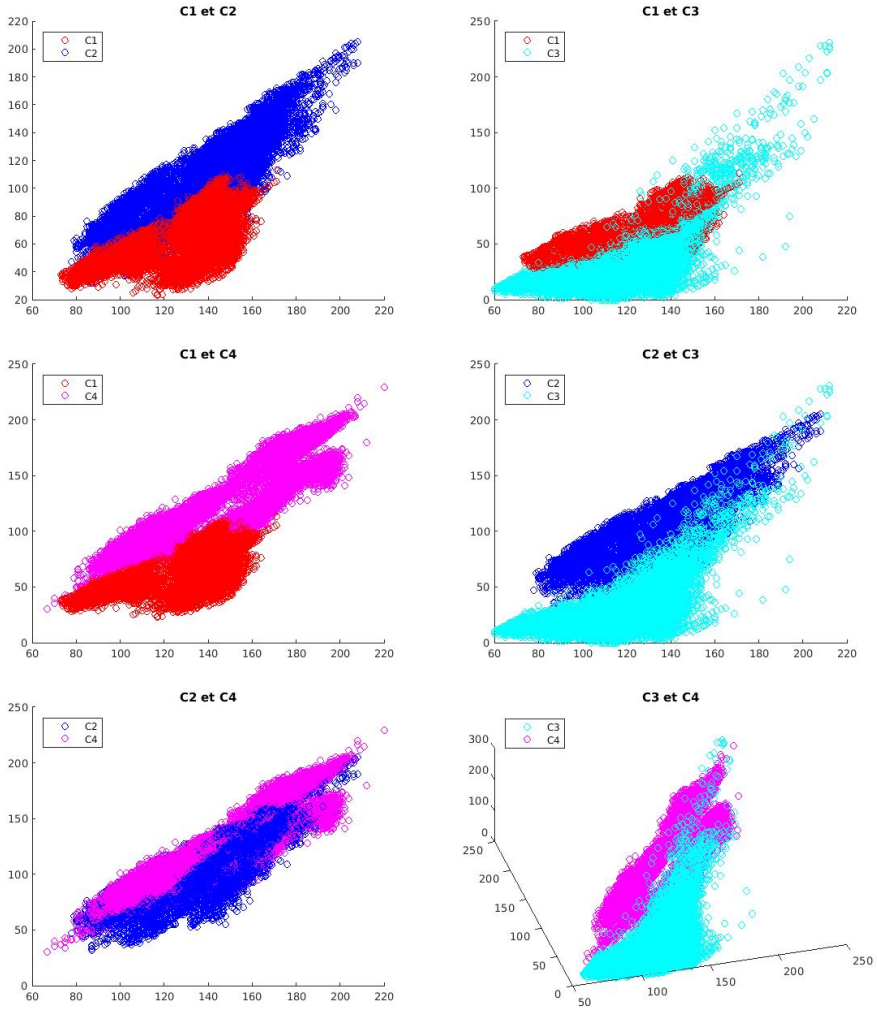

Figure 1: Visualisation des classes

2 Classification bayesienne

Pour deux classes C1 et C2 donnée on cherche à minimiser la propabilité d'erreur, on utilise un classifieur de maximum de vraissemblance :

```
function [ resC1, resC2 ] = classifieurMV( testC1,testC2, muC1,
                                                    sigmaC1, muC2, sigmaC2)

%C1
ptC1 = mvnpdf(testC1,muC1,sigmaC1);
ptC2 = mvnpdf(testC1,muC2,sigmaC2);

resC1=zeros(size(testC1,1),1);
resC1=resC1+(ptC2 > ptC1);
```

```
%C2
ptC1 = mvnpdf(testC2,muC1,sigmaC1);
ptC2 = mvnpdf(testC2,muC2,sigmaC2);

resC2=zeros(size(testC2,1),1);
resC2=resC2+(ptC1 > ptC2);
```

ptC1 est la propabilité d'appartenir à C1, ptC2 la propabilité d'appartenir à C2.
Dans resC1, les lignes ayant pour valeur 1 représentent les indices des données mal classé, c'est-à-dire des élements de la classe C1 mais dont la probabilité d'appartenir à C2 est plus élévé, il en est de même pour resC2 avec la classe C2. Ainsi pour connaitre le nombre d'élément mal classé dans C1, il suffit de faire `sum(resC1)`

On effectuant cette classification pour chacune des 4 classes de départ, nous obtenons les résultats suivant :
50,5 % d'erreur avec la classe C1
66 % d'erreur avec la classe C2
21,2 % d'erreur avec la classe C3
58 % d'erreur avec la classe C4

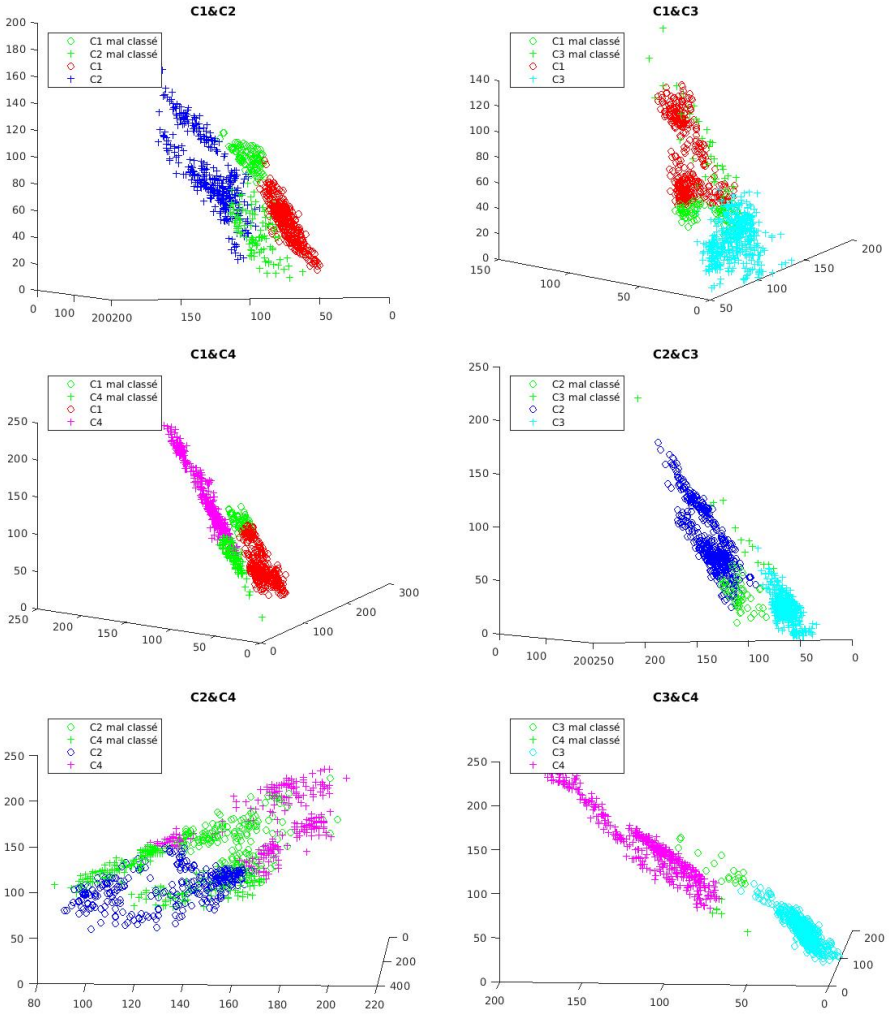

Figure 2: Visualisation des erreurs

On choisi alors par la suite de ne travailler qu'avec les classe C1, C3 et C4.
Pour C1&C3 on obtient 12% d'erreur
Pour C1&C3, 16% d'erreur
Pour C3&C4, 3% d'erreur

3 Classification ACP

L'analyse en composantes principales est une méthode d'analyse des données notamment de statistique multivariée, qui consiste à transformer des variables liées entre elles en de nouvelles variables indépendantes les unes des autres. Ces nouvelles variables sont appelées "composantes principales". Cela permet de réduire le nombre de variables et de rendre les données moins redondante.

De manière géométrique cela revient à faire le projeté des points dans un hyperplan. Lorsqu'on veut réduire en dimension un ensemble de N variables aléatoires, les n premiers axes de l'analyse en composantes principales représente le meilleur choix.

l'ACP nous donne une meilleure représentation des données pour une dimension réduite, mais ne nous permet pas d'obtenir une bonne classification.

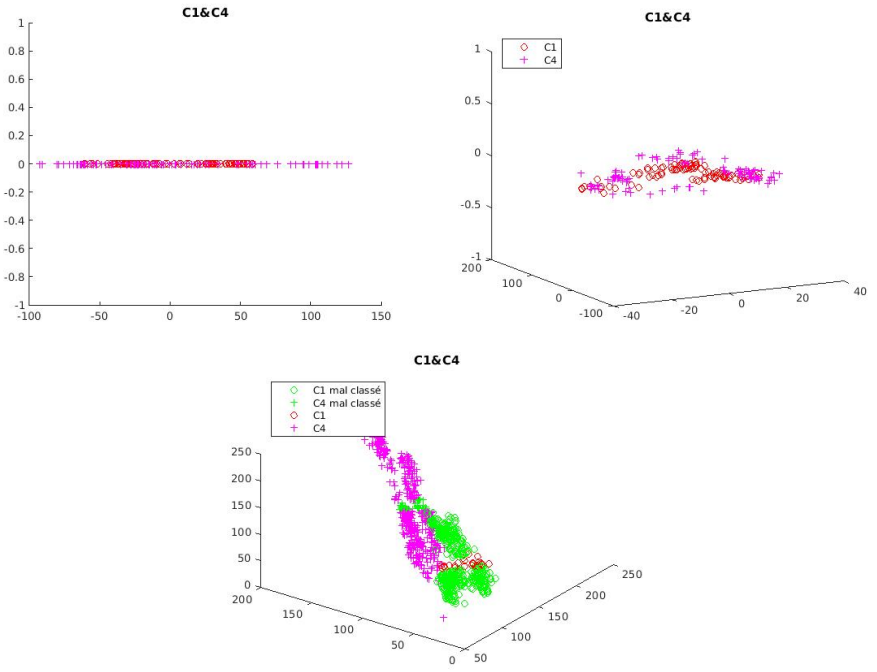

Figure 3: Résultat ACP : projection en 1D, projection en 2D, visualisation des mal classés trouvés

4 Classification Fisher

Avec fisher les données sont également projeté sur une droite, mais cette fois de façon à obtenir une bonne classification des données.

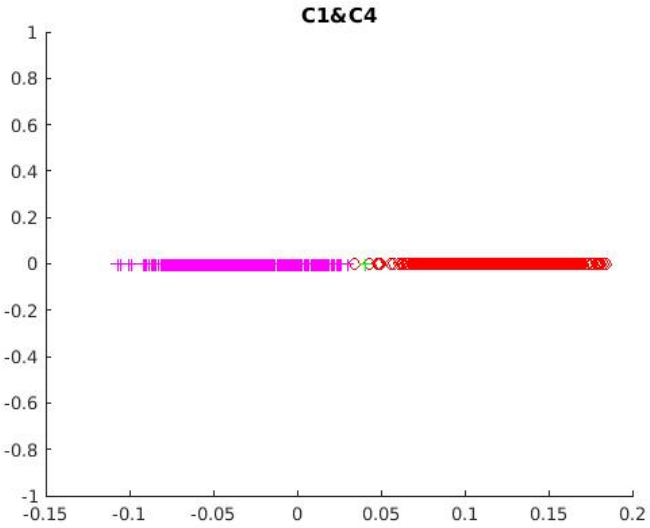

Figure 4: Résultat fisher, mal classés en vert

5 Classification Perceptron

Avec le perceptron nous utilisons une descente de gradient de manière itérative afin de déterminé les mal classés en cherchant s'il existe, une droite ou un plan séparant convenablement les données

6 Comparaison des résultats

	C1&C3	C1&C4	C3&C4
Bayesien	12.5 %	16.6 %	3.34 %
ACP 2D	48 %	47 %	45 %
ACP 1D	47 %	50 %	43 %
Fisher	32 %	0.15 %	0.32 %
Perceptron	8.6 %	2.6 %	2.5 %

Table 1: Pourcentage d'erreur avec différents classifieurs

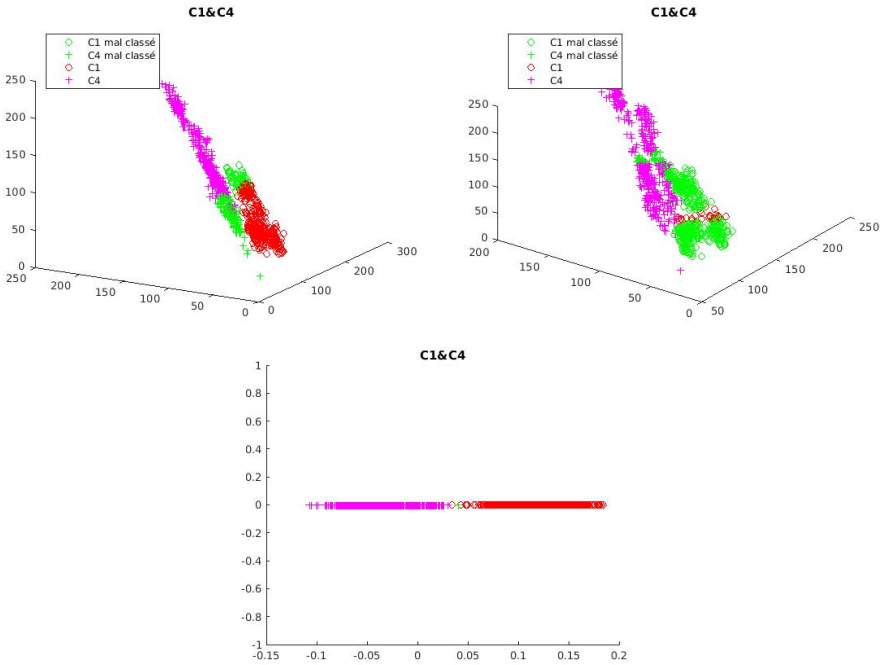

Figure 5: Visualisation des erreurs pour C1&C4 avec Bayesien, ACP et Fisher