**SY09 Printemps 2010 - TP 4**

**Analyses discriminantes quadratique et linéaire**

L'analyse discriminante étudie des données provenant de **groupes connus à priori**. Elle vise deux buts principaux:

* **Description:** Parmi les groupes connus, quelles sont les principales différences que l'on peut déterminer à l'aide des variables mesurées?
* **Classement:** Peut- on déterminer le groupe d'appartenance d'une nouvelle observation uniquement à partir des variables mesurées?

Dans ce TP nous allons utiliser L’analyses discriminantes pour

# **Règle de Bayes**

Dans cette premier exercice nous allons s’intéresser à la frontière de décision de la règle de Bayes avec deux classes gaussiennes de dimension 2. Les calcules vont se faire sur des paramètres différent. En effet nous allons varié les matrices de variance-covariance, les probabilités à priori et les centres de gravité des nuages de points.

## Calcule de la frontière avec la règle de Bayes

## On suppose que la population est répartie en deux classes, en proportions π1 et π2 = 1 − π1, issues des distributions gaussiennes bi variées N(μ1,Σ1) et N(μ2,Σ2).

Pour calculer les frontières de décisions, nous avons utilisé des fonctions discriminantes. La règle de Bayes s’écrit alors avec 



Les régions de décision sont séparées par des frontières d’équations 

1. , ,, 







La frontière est quadratique.

1. 

La frontière est quadratique.

1. 

La frontière est quadratique.

1. 

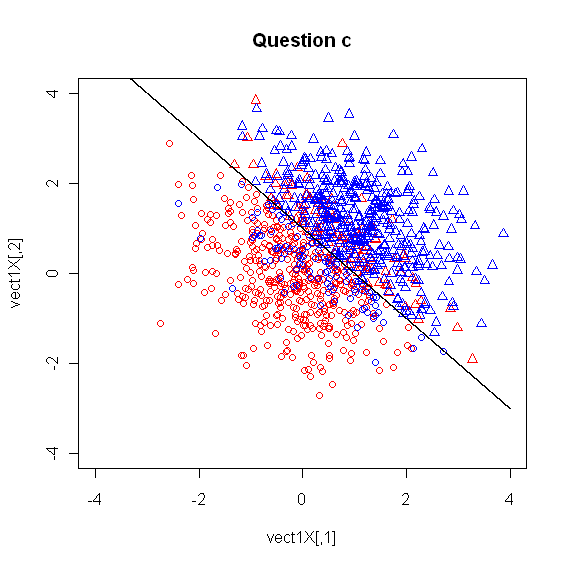
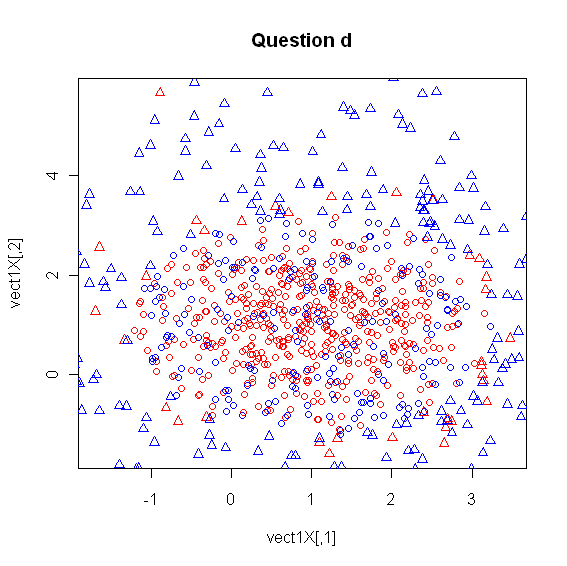
La frontière est un cercle.

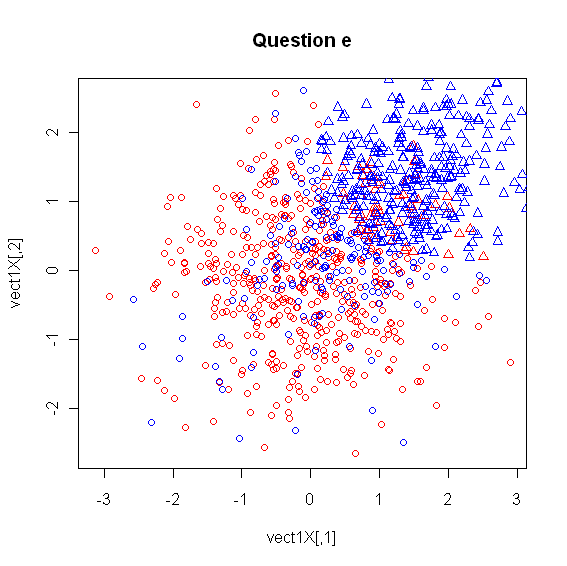


La frontière est une ellipse.

**Remarque** : Les frontières précédemment calculées respect les hypothèses selon lesquelles les frontières sont quadratiques, et plus particulièrement des coniques en dimension 2.

## Voici la représentation de Vect1X1 en fonction de Vect1X2, Vect2X1 en fonction de Vect2X2 ainsi que la frontière de décision calculée pour les trois premiers cas





## L’expression d’un estimateur de la probabilité d’erreur et les réalisations pour les différents cas.

L’erreur théorique dans le cadre de la **règle de Bayes** s’écrit sous la forme



Avec  : carré de la distance de Mahalanobis entre les 2 classes

**La borne de Batthacharyya** a pour expression

Avec 

|  |  |  |  |
| --- | --- | --- | --- |
| **Cas** | **Erreur estimé** | **Erreur théorique de Bayes** | **Borne de Batthacharyya** |
| **A** | **0.242** | **0.2397501** |  |
| **B** | **0.413** | **0.03055968** |  |
| **C** | **0.184** | **0.1990124** |  |
| **D** | **0.235** |  | **0.3651484** |
| **E** | **0.273** |  | **0.3855018** |

**Analyse des résultats:**

On remarque que la probabilité d’erreur théorique et celle estimée dans les cas a et c se rapprochent car on a deux échantillons de mêmeet. Les paramètres et les calculs sont donc cohérents et pertinents. Par contre, dans le cas b, on note un écart significatif. C’est écart s’explique par le fait qu’on l’on part de probabilité a priori égale à 0.1 pour la classe 1 alors qu’elle est en réalité de 0.5 (les cardinaux des classes 1 et 2 sont égaux tous les deux à 500).Dans les cas d et e, on se rapproche des cas a et c car avec une probabilité a priori de 0.6 pour la classe 1. Mais l’écart entre théorie et valeur estimée reste supérieur entre à celui que l’on a dans les cas a et c.

# **Règle de LDA, QDA sur les données Crabes et évaluation des performances**

## Explication ce que fait des instructions suivantes de langage R

* **contour :** est une fonction générique avec une méthode par défaut dans la base R. Il est de créer un contour de la parcelle, ou ajouter des lignes de contour d'une parcelle.
* **sample :** Cette fonction prend un échantillon avec la taille spécifiée dans l’ensemble des éléments en utilisant soit avec, soit sans remplacement.
* **lda :** effectue l’analyse discriminante linéaire
* **qda :** effectue l’analyse discriminante quadratique

1. *La différence entre predict et predict.lda*

* **preditct :** il s'agit d'une fonction générique pour les prédictions à partir des résultats des différents modèles montage des fonctions. La fonction invoque notamment les méthodes qui dépendent de la classe du premier argument.
* **predict.lda** : effectue une classification des observations multi variantes, et aussi projette les données sur les discriminants linéaires.

## Donner un estimateur de la probabilité d’erreur et calculer sa réalisation en se servant des données d’apprentissage utilisées pour effectuer la LDA et la QDA

Application numérique des paramètres :

,  ,

En appliquant les 2 méthodes LDA et QDA sur l’ensemble des données, nous avons reçu des différentes valeurs de probabilité d’erreur.

* **Pour la méthode LDA :** La densité des variables explicatives dans chaque groupe suit une loi multi-normale de même matrice de variances dans chacun des groupes où 

Dans ce cas, la probabilité théorique est calculée par la règle de Bayes



Ici, donc théoriquement = 0.2878168

La probabilité d’erreur réelle 

* **Pour la méthode QDA :** La densité des variables explicatives dans chaque groupe suit une loi multi-normale  où donc

et 

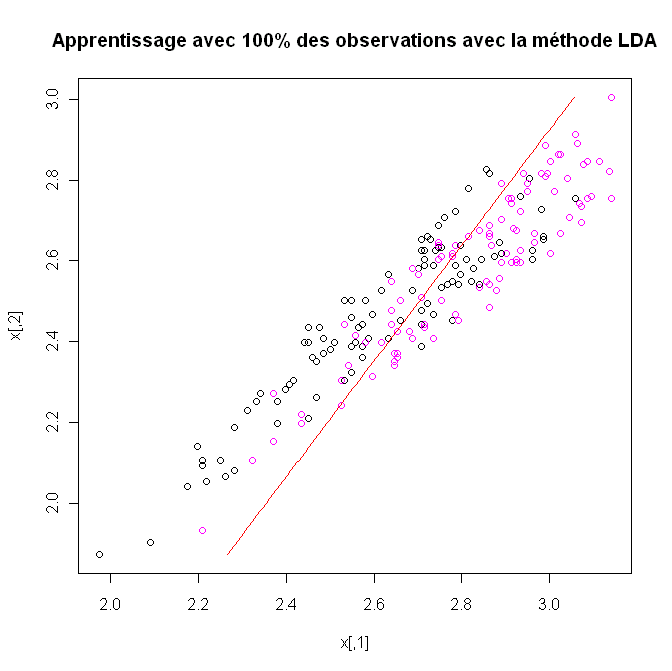
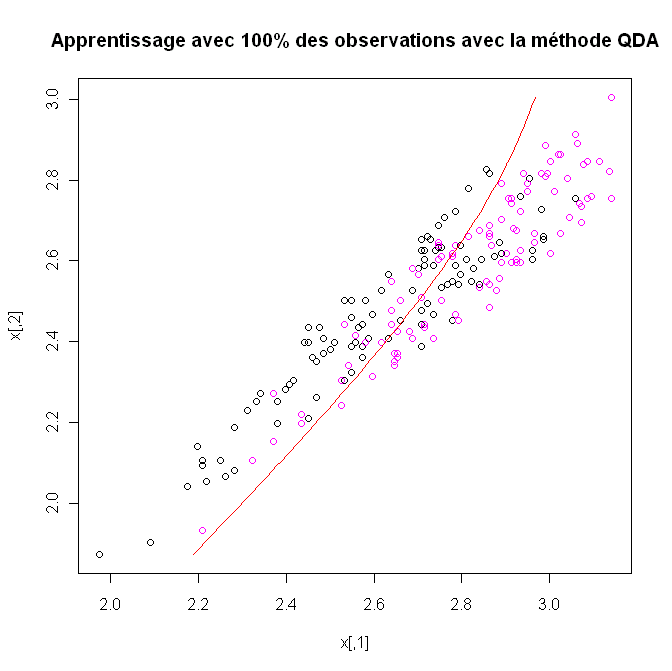
Dans ce cas, la probabilité théorique est calculée par la borne de Bhattacharyya

avec 

Théoriquement 

La probabilité d’erreur réelle 

Voilà la représentation des observations avec une visualisation des frontières de décision.



**Remarque :** Nous avons constaté que dans le cas de la méthode de QDA, .De plus, l’écart de la probabilité d’erreur théorique et réelle n’est plus important que celui du cas LDA. Donc, la méthode QDA est plus intéressante.

## Appliquer la méthode de l’échantillon de test avec 2/3 des observations pour l’ensemble d’apprentissage. Répéter la méthode de l’échantillon de test en choisissant différentes échantillons d’apprentissage et de test.

En réalité, en modifiant les probabilités de tirage aléatoire avec l’aide de fonction sample, nous ne pouvons connaitre exactement la classe de chacun des éléments de test. Nous avons modifié la taille de l’ensemble d’apprentissage qui nous permet de construire une frontière de décisions à l’aide de l’analyse quadratique ou linéaire, et on classe ensuite les éléments de l’échantillon de test en fonction de cette frontière.

Nous avons répété quelques fois l’expérience avec des différentes proportions de l’échantillon d’apprentissage (Appr) et de test (Test). En utilisant les 2 méthodes (LDA, QDA), nous avons retrouvé les probabilités d’erreur correspondantes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **2/3 Appr , 1/3 Test** | **1/2 Appr, 1/2 Test** | **1/3 Appr, 2/3 Test** | **1/10 Appr, 9/10 Test** |
|  | 0.3050847 | 0.3009709 | 0.3115942 | 0.3891892 |
|  | 0.2881356 | 0.2718447 | 0.2971014 | 0.372973 |

Nous avons constaté que dans tous les cas, la probabilité d’erreur de la méthode QDA est toujours moins importante que celle de la LDA. On peut en tirer le fait que : QDA est le meilleur modèle entre LDA et QDA