SY09 - TP3

Sarah Priour - Nina Martin

Printemps 2013

Table des matières

Chapitre 1

Classifieur Euclidien

1.1 Objectifs

L'objectif de cette section est d'étudier les performances d'un classifieur euclidien sur 2 échantillons issus de 2 lois normales de paramètres μ et Σ donnés.

Le classifieur euclidien consiste à déterminer la classe d'un individu à partir de sa distance euclidienne par rapport aux centres (ou moyennes) calculés pour chaque classe.

1.2 Simulation d'un échantillon

En pratique, il faudrait pour cette expérience disposer de données issues d'une réalisation précédente. Ici nous allons simuler ces données grâce à certaines fonctions de R.

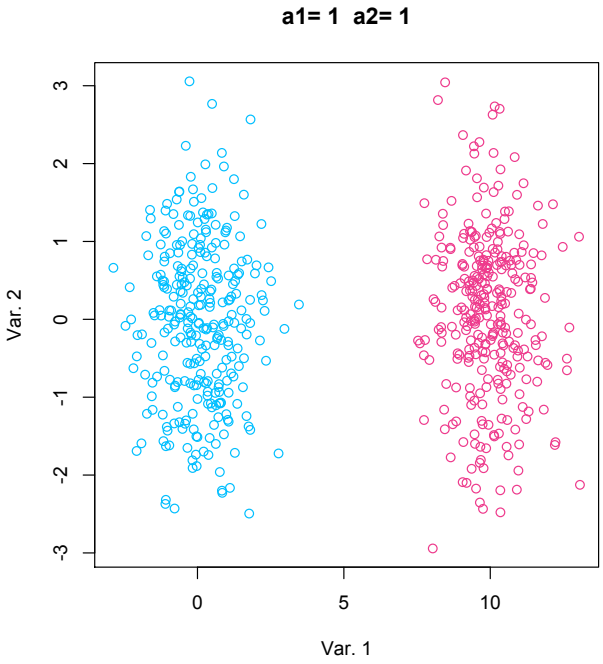
La première étape consiste à tirer la classe de l'individu en utilisant une loi de Bernoulli de paramètre π.

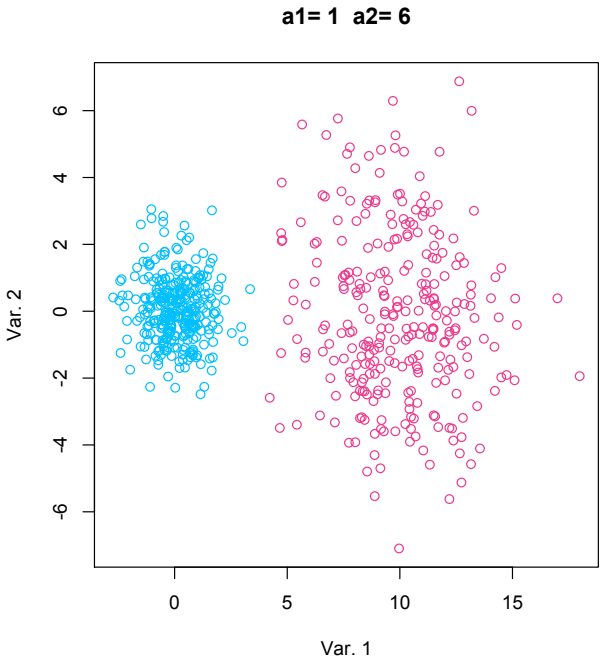
Ensuite il faut générer l'individu selon la classe déterminée.

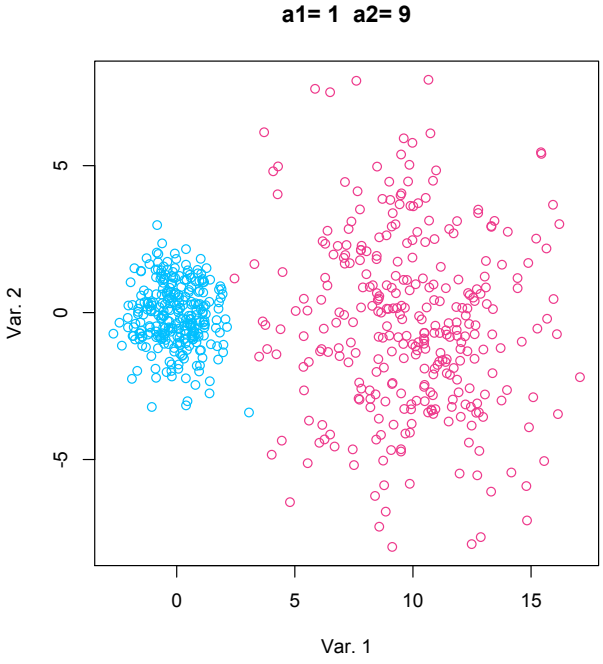
On peut remarquer que, plutôt que de réaliser n lois de Bernouilli de paramètre π renvoyant une valeur 0 (si l'individu appartient à la classe 2) ou 1 (si celui-ci appartient à la classe 1), on peut réaliser une seule loi binomiale de paramètres (n,π) renvoyant une valeur z comprise entre 0 et n représentant le nombre d'individus appartenant à la classe 1. Il ne reste alors plus qu'à générer z individus suivant la loi normale de la classe 1 en utilisant *mvrnorm(z,* μ_1*,* Σ_1*)* et $n - z$ individus suivant la loi normale de la classe 2 avec *mvrnorm(n - z,* μ_2*,* Σ_2*)*.

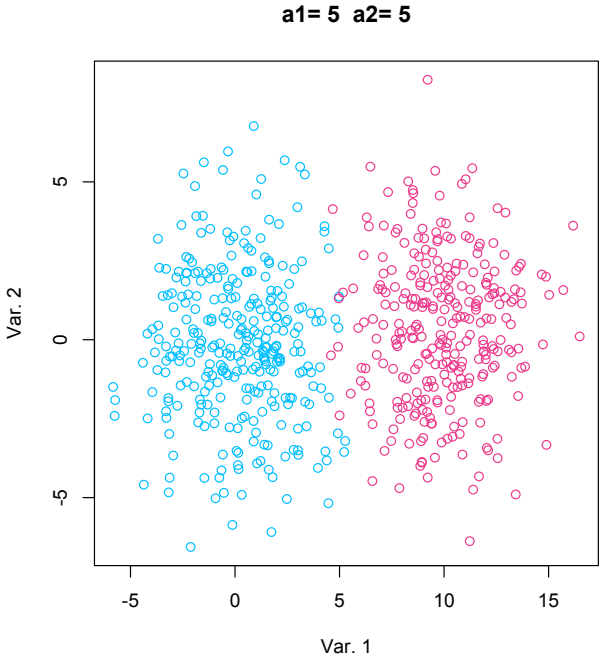
L'expérience est répétée avec 5 matrices de variances différentes. Ce sont essentiellement des matrices définies positives multiples de la matrice identité, ce qui donne un caractère sphérique aux classes.

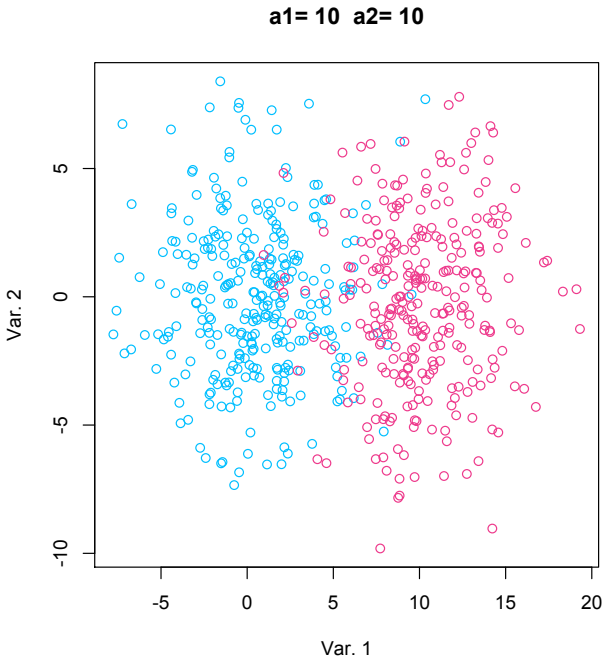
Les points peuvent ensuite être visualisés dans le plan.


a1= 1 a2= 1
Var. 2
Var. 1
3
2
1
0
-1
-2
-3
0
5
10


a1= 1 a2= 6
Var. 2
Var. 1
6
4
2
0
-2
-4
-6
0
5
10
15


a1= 1 a2= 9
Var. 2
Var. 1
5
0
-5
0
5
10
15


a1= 5 a2= 5
Var. 2
Var. 1
5
0
-5
-5
0
5
10
15



Comme on pouvait s'y attendre, les points des classes avec une variance élevée sont plus dispersés que pour les classes avec une variance faible.

1.3 Estimation de la probabilité d'erreur

On cherche maintenant à estimer la probabilité d'erreur associée au classifier euclidien dans les 5 situations.

Pour cela, il faut couper notre échantillon en 2 : un échantillon d'apprentissage à partir duquel nous pourrons déterminer les estimateurs μ_1 et μ_2 à l'aide de la moyenne empirique, et un échantillon de test sur lequel nous testerons le classifieur euclidien (qui permet d'associer un point à une classe en comparant les distances entre celui-ci et les centres des différentes classes).

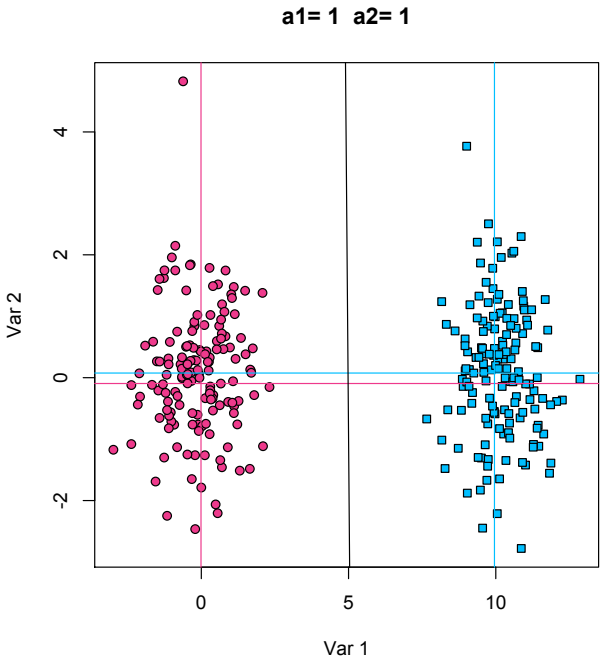
Il n'aurait pas été pertinent d'utiliser le même échantillon pour à la fois estimer les moyennes et tester le classifieur. En effet, les estimateurs auraient été calculés spécialement pour cet échantillon, et les résultats auraient été biaisés positivement.

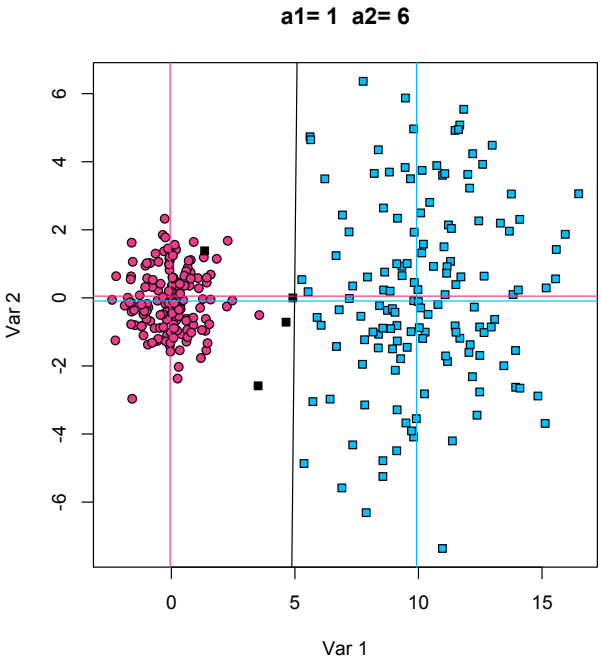
Pour l'élaboration de notre échantillon d'apprentissage, nous récupérons aléatoirement la moitié des individus de la classe 1 et la moitié des individus de la classe 2. L'objectif est d'obtenir le même ratio classe1/classe2 dans que dans l'échantillon de départ.

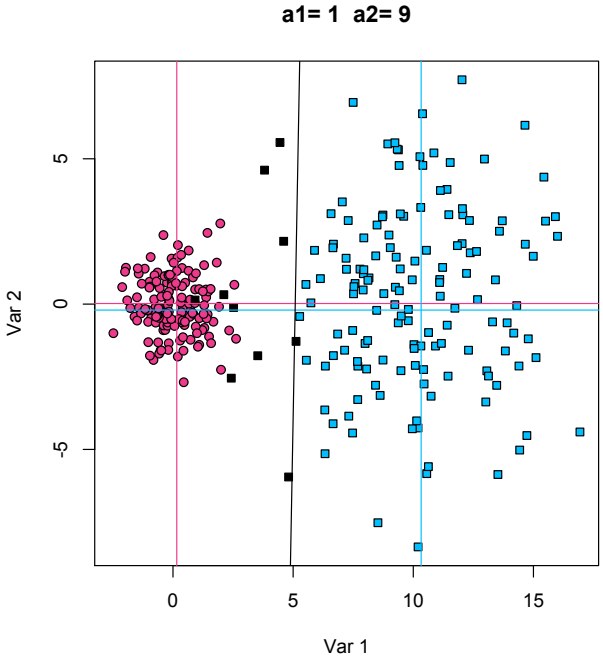
Pour l'échantillon de test, nous récupérons les individus qui n'ont pas été sélectionnés pour l'échantillon d'apprentissage.

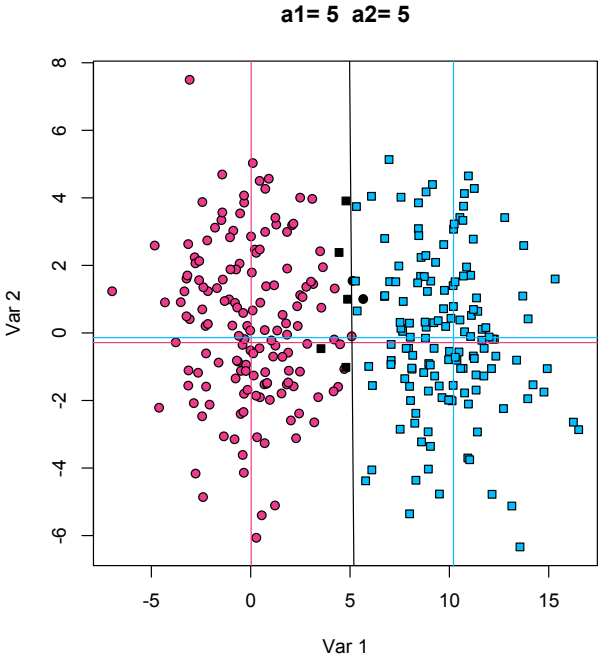
Une fonction *erreurEstimee* nous permet d'évaluer le taux d'erreur en relevant les individus qui ont été affectés à la mauvaise classe. Dans les graphes suivant, nous pouvons distinguer ces individus qui apparaissent en noir. La forme des points (ronds ou carrés) nous permet de connaitre la classe réelle de l'individu.

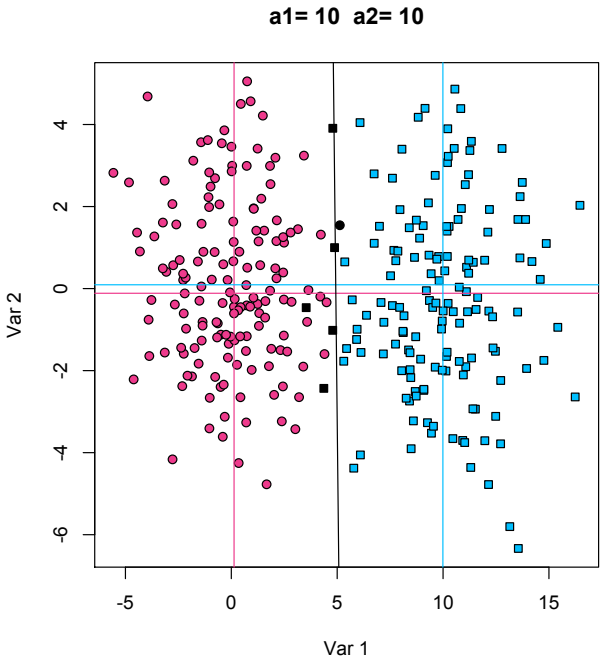
Sur les figures ci-dessous, on peut également observer les centres des classes, situés à l'intersection des droites roses pour la classe 1 et bleues pour la classe 2.


a1= 1 a2= 1
Var 2
Var 1
-2
0
2
4
0
5
10


a1= 1 a2= 6
Var 2
Var 1
-6
-4
-2
0
2
4
6
0
5
10
15


a1= 1 a2= 9
Var 2
Var 1
-5
0
5
0
5
10
15


a1= 5 a2= 5
Var 2
Var 1
-6
-4
-2
0
2
4
6
8
-5
0
5
10
15



A première vue on voit que le classifieur commet plus d'erreurs sur les classes avec une variance élevée. Cela était prévisible puisque les points de ces classes sont plus dispersés et donc plus sujet à se rapprocher des centres d'autres classes.

Il faut aussi rappeler que les deux classes sont relativement bien séparées au niveau de la 1$^{\text{ere}}$ variable (pour la 1$^{\text{ere}}$ loi $\mu_1[1] = 0$ et pour la 2$^{\text{ème}}$ loi $\mu_2[1] = 10$). Si les classes étaient plus proches, le classifieur commettrait plus d'erreurs et inversement.

Nous répétons l'expérience 10 fois afin d'obtenir une espérance et une variance sur le pourcentage d'erreur dans chaque situation. A chaque fois nous calculons un intervalle de confiance de niveau 5% sur l'espérance de la probabilité d'erreur en calculant le résultat suivant :

$$I_c = [\bar{x} - t_\alpha \frac{s}{\sqrt{n}}; \bar{x} + t_\alpha \frac{s}{\sqrt{n}}]$$

où $\bar{x}$ est la moyenne empirique, s est l'écart-type empirique, n la taille de l'échantillon (ici 10, car nous calculons 10 fois l'erreur), et t_α le niveau de confiance choisi (ici 0.05).

Nous obtenons les résultats suivants :

(a1,a2)	Moyenne	Variance	Intervalle de Confiance
(1,1)	0	0	[0 ;0]
(1,6)	0.0130	10^{-5}	[0.0129 ;0.0131]
(1,9)	0.0226	10^{-5}	[0.0226 ;0.0228]
(5,5)	0.0130	10^{-5}	[0.0129 ;0.0131]
(10,10)	0.0557	10^{-4}	[0.0554 ;0.0559]

TABLE 1.1 – Coordonnées des variables sur les composantes principales

On observe comme prévu une moyenne d'erreurs plus élevée lorsqu'on a des classes avec une variance élevée, les points étant plus dispersés dans le plan.

Chapitre 2

Règle de Bayes

2.1 Objectifs

Dans cette seconde partie, nous nous intéresserons à un problème de classification de cibles en deux classes ω_1 et ω_2 (missiles et avions) à partir de leur description par deux variables X_1 et X_2 issues de deux capteurs différents. Nous sommes donc une nouvelle fois confrontés à un problème de discrimination en deux classes.

Chaque variable suit dans chaque classe une loi normale avec les paramètres suivants :

$$f_{11}(x_1) \sim \mathcal{N}(-1, 1), \quad f_{21}(x_1) \sim \mathcal{N}(1, 1)$$

$$f_{12}(x_2) = f_{22}(x_2) \sim \mathcal{N}(0; 1)$$

On suppose l'indépendance conditionnelle de X_1 et X_2.

2.2 Expression de f_1 et f_2

X_1 et X_2 étant indépendantes, les densités conditionnelles du vecteur $X = (X_1, X_2)^\top$ sont donc $f_1(x) = f_{11}(x_1)f_{12}(x_2)$ dans la classe ω_1 et $f_2(x) = f_{21}(x_1)f_{22}(x_2)$ dans la classe ω_2.

La densité d'une variable aléatoire suivant une loi multidimensionnelle de paramètres μ et σ^2 s'exprime de la manière suivante :

$$f(x) = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Ainsi si nous faisons le produit des densités $f_{11}(x)$ et $f_{12}(x)$, de paramètres μ_{11}, ${\sigma_{11}}^2$ et μ_{12}, ${\sigma_{12}}^2$, nous obtenons :

$$f_1(x) = \frac{1}{2\pi\sigma_{11}\sigma_{12}} e^{-\frac{1}{2{\sigma_{11}}^2}(x_1 - \mu_{11})^2 - \frac{1}{2{\sigma_{12}}^2}(x_2 - \mu_{12})^2}$$

$$f_1(x) = \frac{1}{2\pi \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}$$

avec $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $\mu = \begin{pmatrix} \mu_{11} \\ \mu_{12} \end{pmatrix}$, $\Sigma = \begin{pmatrix} {\sigma_{11}}^2 & 0 \\ 0 & {\sigma_{12}}^2 \end{pmatrix}$ et par conséquent $\Sigma^{-1} = \begin{pmatrix} \frac{1}{{\sigma_{11}}^2} & 0 \\ 0 & \frac{1}{{\sigma_{12}}^2} \end{pmatrix}$.

Ceci est bien la densité d'une loi normale multidimensionnelle de paramètres μ et Σ.

Nous avons donc :

$$f_1(x) \sim \mathcal{N}(\mu_1, \Sigma_1)$$

avec $\mu_1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$ et $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

En appliquant le même raisonnement pour f_2, nous obtenons à nouveau une loi normale multidimensionnelle, avec pour paramètres $\mu_2 = \begin{pmatrix}\mu_{21}\\ \mu_{22}\end{pmatrix}$, $\Sigma_2 = \begin{pmatrix}{\sigma_{21}}^2 & 0\\ 0 & {\sigma_{22}}^2\end{pmatrix}$, soit :

$$f_2(x) \sim \mathcal{N}(\mu_2, \Sigma_2)$$

avec $\mu_2 = \begin{pmatrix}1\\ 0\end{pmatrix}$, $\Sigma_2 = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$.

2.3 Estimation des paramètres de f_1 et f_2

En utilisant la fonction *simul*, nous générons un échantillon de n réalisations issues des deux classes en proportions égales ($\pi_1 = \pi_2 = 0.5$). Nous estimons ensuite les paramètres de f_1 et f_2.

Cette expérience est réalisée pour différentes valeurs de n : 10, 100, 1000, 10000, 10000. Nous obtenons les résultats suivants :

Taille	$\hat{\mu_1}$	$\hat{\Sigma_1}$	$\hat{\mu_2}$	$\hat{\Sigma_2}$
$n = 10$	$\begin{pmatrix}-1.13\\ 0.87\end{pmatrix}$	$\begin{pmatrix}0.76 & 0\\ 0 & 0.35\end{pmatrix}$	$\begin{pmatrix}1.49\\ -0.03\end{pmatrix}$	$\begin{pmatrix}1.95 & 0\\ 0 & 0.70\end{pmatrix}$
$n = 100$	$\begin{pmatrix}-0.91\\ 0.07\end{pmatrix}$	$\begin{pmatrix}0.80 & 0\\ 0 & 1.12\end{pmatrix}$	$\begin{pmatrix}0.84\\ -0.25\end{pmatrix}$	$\begin{pmatrix}1.35 & 0\\ 0 & 0.93\end{pmatrix}$
$n = 1000$	$\begin{pmatrix}-1.05\\ -0.00\end{pmatrix}$	$\begin{pmatrix}1.01 & 0\\ 0 & 0.83\end{pmatrix}$	$\begin{pmatrix}0.95\\ -0.07\end{pmatrix}$	$\begin{pmatrix}1.09 & 0\\ 0 & 0.94\end{pmatrix}$
$n = 10000$	$\begin{pmatrix}-1.00\\ -0.01\end{pmatrix}$	$\begin{pmatrix}0.99 & 0\\ 0 & 0.98\end{pmatrix}$	$\begin{pmatrix}1.02\\ 0.00\end{pmatrix}$	$\begin{pmatrix}0.97 & 0\\ 0 & 0.96\end{pmatrix}$
$n = 100000$	$\begin{pmatrix}-1.00\\ -0.01\end{pmatrix}$	$\begin{pmatrix}0.99 & 0\\ 0 & 0.99\end{pmatrix}$	$\begin{pmatrix}0.99\\ 0.01\end{pmatrix}$	$\begin{pmatrix}1.01 & 0\\ 0 & 1.00\end{pmatrix}$

TABLE 2.1 – Estimation des paramètres de f_1 et f_2 sur des échantillons de n réalisations

Plus l'échantillon est grand, plus l'estimation des paramètres se rapproche de la valeur théorique, ce qui était prévisible puisque nous disposons d'un plus grand nombre de données.

2.4 Courbes d'isodensité de f_1 et f_2

Avec une représentation en 3 dimensions des échantillons générés précédemment, la fonction de densité serait représentée sur le troisième axe. Il peut alors être intéressant de représenter les courbes d'iso-densité (qui sont des courbes de niveau) dans le plan d'axes X_1 et X_2 pour se faire une idée de la représentation tridimensionnelle.

Pour calculer les courbes d'isodensité, on résout les équations $f_1(x) = K_1$ et $f_2(x) = K_2$ où K_1 et K_2 sont des constantes.

Dans notre cas, nous avons

$$f_1(x) = K_1$$

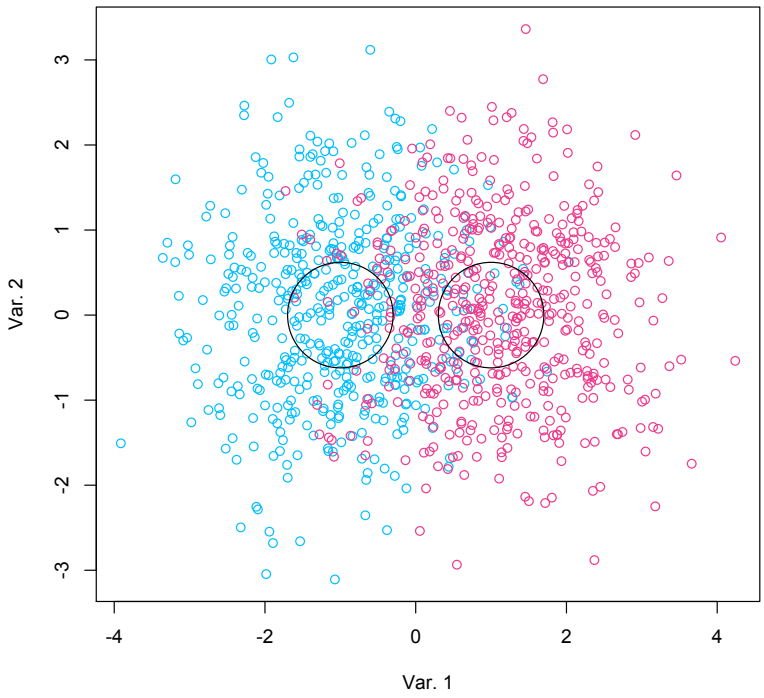
avec Σ la matrice identité et $\det(\Sigma) = 1$, soit :

$$\frac{1}{2}(x-\mu)^\top(x-\mu) = -\ln(2\pi K_1)$$
$$(x_1-\mu_1)^2 + (x_2-\mu_2)^2 = -2\ln(2\pi K_1)$$
$$(x_1+1)^2 + (x_2)^2 = -2\ln(2\pi K_1)$$

Ceci est l'équation d'un cercle de centre $\mu_1 = \begin{pmatrix}-1\\ 0\end{pmatrix}$ et de rayon $\sqrt{-2\ln(2\pi K_1)}$.

Pour f_2 nous obtenons un cercle de $\mu_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ et de rayon $\sqrt{-2\ln(2\pi K_2)}$.

En choisissant arbitrairement la valeur 0.20 pour K_1 et K_2 nous pouvons afficher les courbes d'isodensité de f_1 et f_2 dans le plan ($n = 1000$).



2.5 Règle de Bayes

La règle de Bayes est la règle qui vise à minimiser la probabilité d'erreur lors de la prise de décision. On cherchera donc la fonction de décision δ^* qui minimise le risque. On a :

$$\delta(x) = a_1 \text{ si } \frac{f_1(x)}{f_2(x)} > \frac{c_{12}-c_{22}}{c_{21}-c_{11}} \frac{\pi_1}{\pi_2}$$

$$\delta(x) = a_2 \text{ sinon.}$$

Les coûts sont définis par :

	ω_1	ω_2
a_1	c_{11}	c_{12}
a_{12}	c_{21}	c_{22}

où ω_1 et ω_2 représentent la classe réelle de l'individu considéré et a_1 et a_2 la décision prise pour cet individu (la classe à laquelle on a décidé de l'affecter).

Dans cet exercice, nous considérons que $c_{11} = c_{22} = 0$, autrement dit, les coûts engendrés par la décision a_l d'affecter un individu à sa classe réelle ω_k sont nuls, valeurs a priori assez intuitives.

En appliquant la règle de Bayes sur notre problème avec diverses valeurs pour c_{12}, c_{21}, π_1 et π_2, nous obtenons :

i. pour $c_{12} = c_{21} = 1$ et $\pi_1 = \pi_2$:

$$\delta^*(x) = a_1 \text{ si } f_1(x) > f_2(x)$$

$$\delta^*(x) = a_2 \text{ sinon}$$

ii. pour $c_{12} = 10$, $c_{21} = 1$ et $\pi_1 = \pi_2$:

$$\delta^*(x) = a_1 \text{ si } f_1(x) > 10 f_2(x)$$

$$\delta^*(x) = a_2 \text{ sinon}$$

iii. pour $c_{12} = c_{21} = 1$ et $\pi_2 = 10\pi_1$:

$$\delta^*(x) = a_1 \text{ si } f_1(x) > 10 f_2(x)$$

$$\delta^*(x) = a_2 \text{ sinon}$$

Dans la 1ère situation, les coûts associés aux choix de la mauvaise classe sont les mêmes dans chaque cas. Nous disposons par ailleurs de proportions égales. La règle de décision ne va donc pas favoriser une classe ou l'autre et se contenter de sélectionner celle dont la densité calculée pour l'individu est la plus élevée.

Dans la 2ème situation, le coût associé à la décision d'affecter à la casse ω_1 à un individu qui appartient en réalité à la classe ω_2 est très élevé comparée à l'erreur inverse. La règle de décision va alors d'une manière générale favoriser la classe ω_2 afin de minimiser le coût moyen.

Dans la 3ème situation, nous avons une proportion beaucoup plus élevée d'individus dans la classe ω_2. La règle de décision va donc favoriser cette classe, puisque les chances de tomber sur un individu de cette classe sont beaucoup plus élevées.

Pour le 2ème et le 3ème cas, nous nous retrouvons donc avec des règles de décision similaires, mais pour des raisons différentes (coût d'une décision élevé dans un cas, et proportion d'une classe élevée dans l'autre cas).

2.5.1 Frontières de décision

Pour le cas i.

$$f_1(x) = f_2(x)$$

$$(x+1)^2 = (x-1)^2$$

$$x = 0$$

La frontière de décision est la droite d'équation $x = 0$.

Pour les cas ii. et iii.

$$\frac{f_1(x)}{f_2(x)} < 10$$

$$\frac{e^{-\frac{1}{2}\,[(x_1+1)^2 + (x_2)^2]}}{e^{-\frac{1}{2}\,[(x_1-1)^2 + (x_2)^2]}} < 10$$

$$e^{-2(x)} < 10$$

$$x = -\frac{\ln(10)}{2}$$

La frontière de décision est la droite d'équation $x = -\frac{\ln(10)}{2}$ soit approximativement la droite d'équation $x = -1,15$.

Pour les représentations graphiques, nous avons choisi arbitrairement $n = 10000$.

Les points ronds correspondent aux individus dont la classe réelle est ω_1, les points carrés correspondent aux individus dont la classe réelle est ω_2. Les points roses correspondent aux individus auxquels on a affecté la classe ω_1, les points bleus correspondent aux individus auxquels on a affecté la classe ω_2.

Les points gris correspondent aux individus qui ont été affectés à la mauvaise classe après application de la règle de décision.

La droite verticale représente la frontière de décision. Les individus à gauche de cette droite sont affectés à la classe ω_1 et ceux à droite sont affectés à la classe ω_2. On peut d'ailleurs vérifier que nous avons uniquement des points à rose à gauche de la frontière et bleus à droite de la frontière.

Note : Sur les graphes ci-dessous, on peut avoir l'impression que la proportion de points gris est plus élevée qu'elle ne l'est en réalité. Cela est dû au fait que nous avons beaucoup de points qui se chevauchent et que nous avons choisi de faire apparaitre les individus affectés à la mauvaise classe en premier plan.

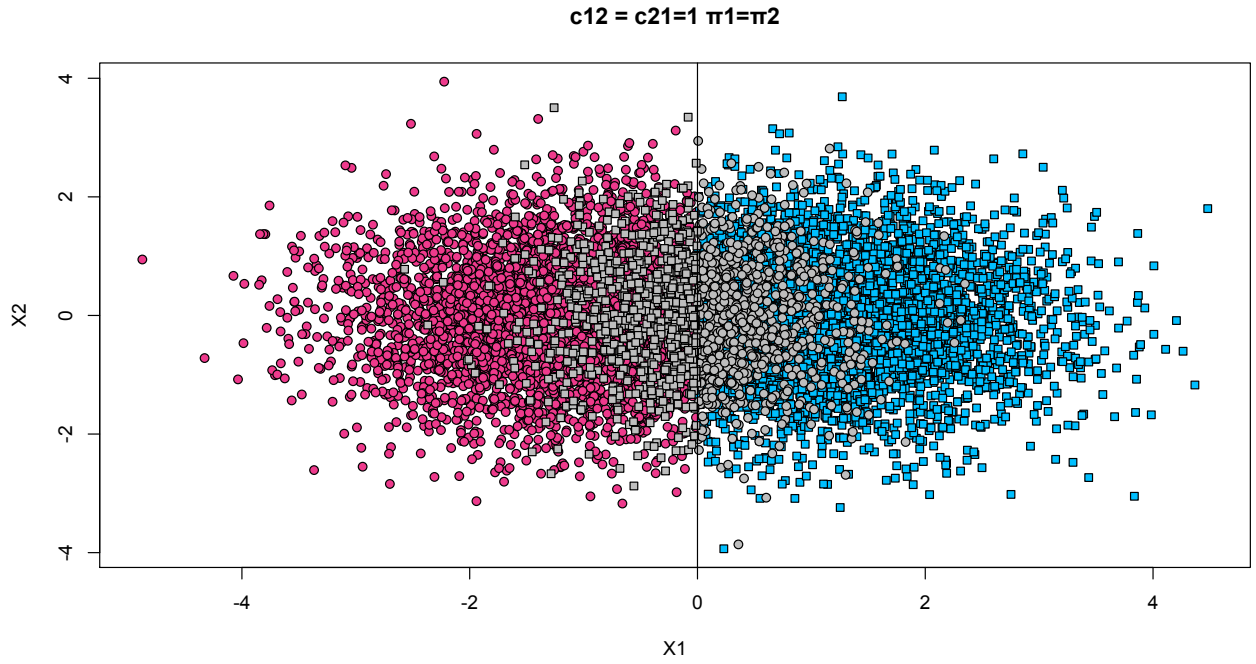


FIGURE 2.1 – Frontière de décision : $x = 0$

Sur la figure 2.1 on observe 2 classes de même proportion, avec des erreurs à proximité de la frontière de décision.

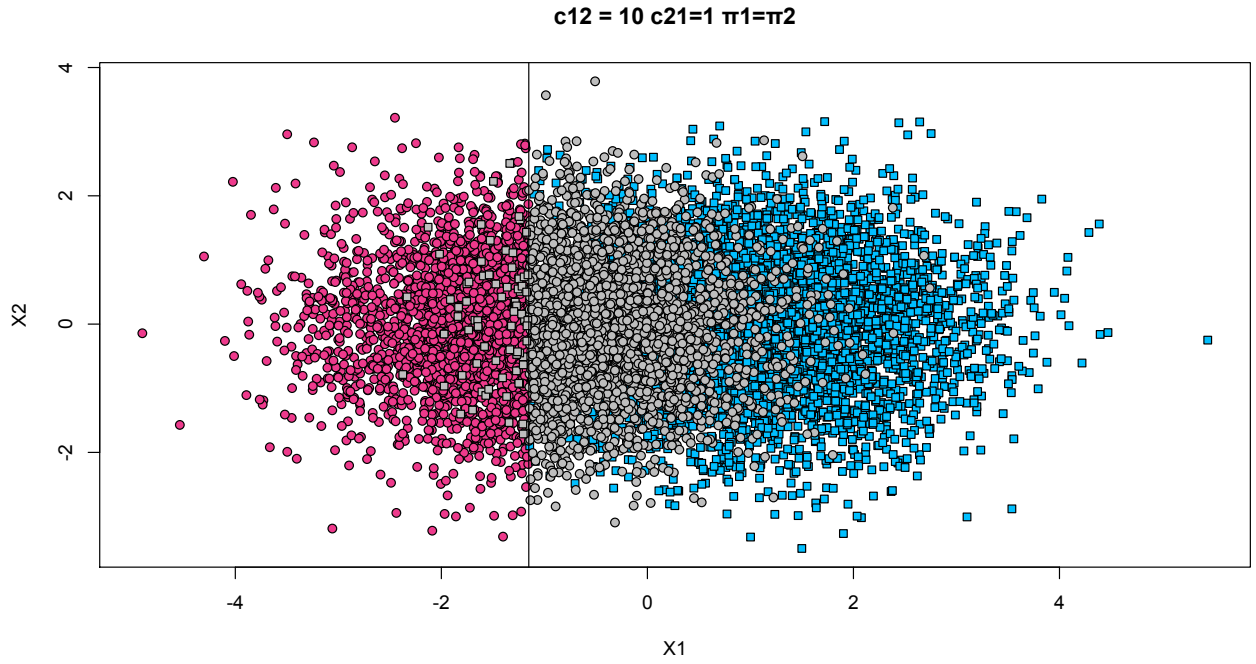


FIGURE 2.2 – Frontière de décision : $x = -\frac{\ln(10)}{2}$

Sur la figure 2.2 on observe à nouveau 2 classes de même proportion (une partie des carrés est cachée par les points gris erronés).

Cependant, on voit que la frontière de décision est décalée sur la gauche par rapport à la 1ère figure. Il en résulte que beaucoup de points de la classe ω_1 sont affectés à la classe ω_2 (points ronds et gris) et à l'inverse très peu de points de la classe ω_2 sont affectés à la classe ω_1 (points carrés et gris).

En effet, on cherche à minimiser le nombre de fois où la règle de décision choisit la classe ω_1 au lieu de la classe ω_2, le coût associé à cette erreur étant très élevé.

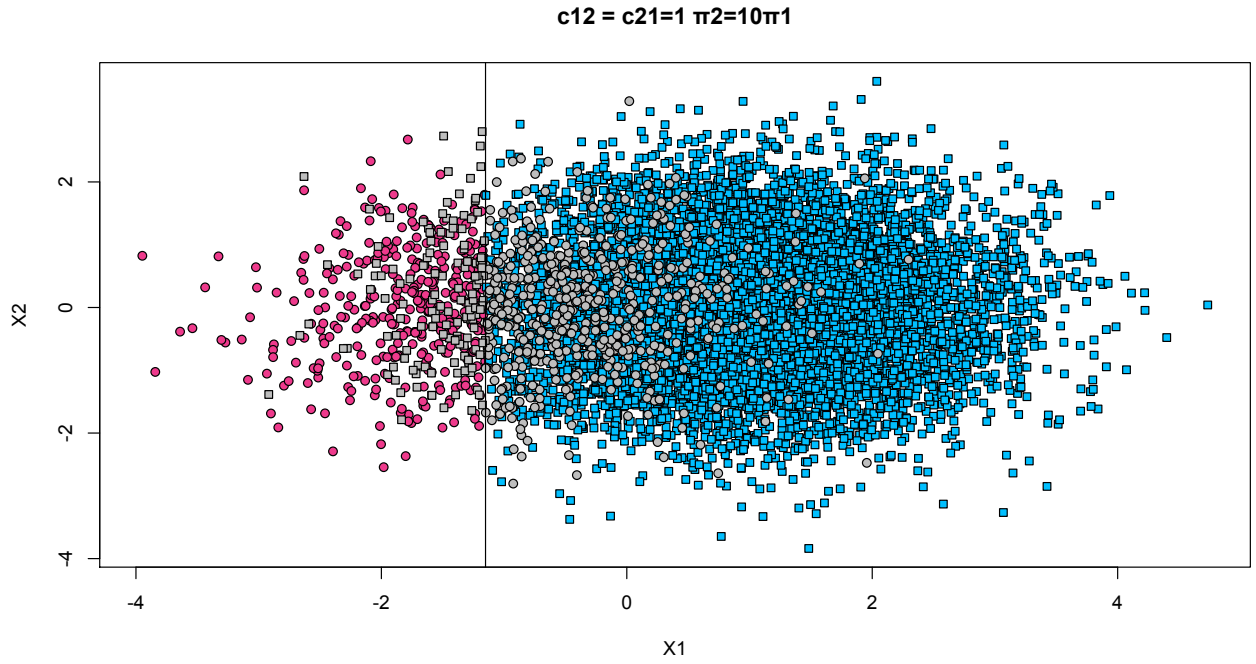


FIGURE 2.3 – Frontière de décision : $x = -\frac{\ln(10)}{2}$

Sur la figure 2.3, on a 2 classes en proportions inégales (plus de carrés (ω_2) que de ronds (ω_3)). La frontière de décision est à nouveau décalée sur la gauche. En effet, la classe ω_2 étant bien plus importante, nous avons intérêt à avoir une proportion d'erreurs aussi basse que possible pour cette classe.

2.5.2 Estimation des risques α et β

Nous cherchons les risques α et β associés à la règle de décision. Ces risques sont définis par $\alpha = P(\delta^*(x) = a_2|\omega_1)$ et $\beta = P(\delta^*(x) = a_1|\omega_2)$. On calcule les estimations de ces risques en calculant la proportion de points affectés à la classe ω_2 au sein de la classe ω_1 pour α et inversement pour β.

	α	β
$c_{12} = c_{21} = 1 \pi_1 = \pi_2$	0.17	0.16
$c_{12} = 10 c_{21} = 1$ et $\pi_1 = \pi_2$	0.56	0.01
$c_{12} = c_{21} = 1$ et $\pi_2 = 10\pi_1$	0.59	0.02

TABLE 2.2 – Risques α et β associés à la règle de Bayes

Dans le premier cas, les estimateurs des risques α et β sont très proches, en effet on ne cherche pas à minimiser l'un plus que l'autre.

Dans le 2ème et le 3ème cas, l'estimateur de β est sensiblement plus faible que celui de α. En effet dans un cas, nous cherchons à minimiser β car les coûts associés à ce risque sont très élevés, et dans l'autre cas on le minimise car la classe ω_2 étant plus importante, nous avons intérêt à minimiser la proportion de points de cette classe affectés à ω_1.

Conclusion

Au cours de ce TP, nous avons pu évaluer deux méthodes de classification supervisée. Nous avons d'abord étudié le classifieur euclidien en se basant sur un ensemble d'apprentissage pour déterminer les centres de classes. Celui-ci marche assez bien lorsque les classes sont bien séparées dans le plan et que les points sont relativement bien regroupés autour du centre (variance faible).

Nous avons ensuite étudié la règle de Bayes en supposant les distributions des classes connues. La règle de Bayes permet de classifier les individus en prenant en compte deux paramètres supplémentaires : le coût des différentes types d'erreurs et la proportion à priori de chaque classe. Cette règle permet de minimiser le risque plutôt que le nombre d'erreurs totales : on peut privilégier un type d'erreur par rapport à un autre.