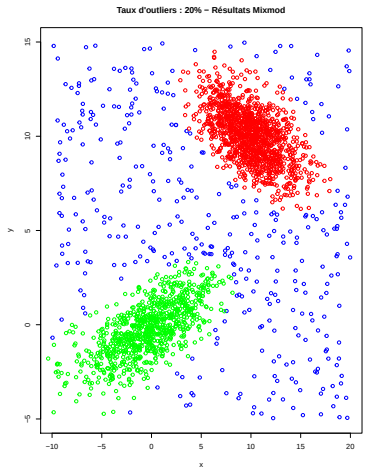
Classification en présence d'outliers (données aberrantes) avec RMixmod

Package de classification par modèles de mélanges



F. Langrognet

laboratoiredemathématiquesdebesançon
UNIVERSITÉ DE FRANCHE-COMTÉ • CNRS • UMR 6623

Rmixmod

Etude de la robustesse de RMixmod

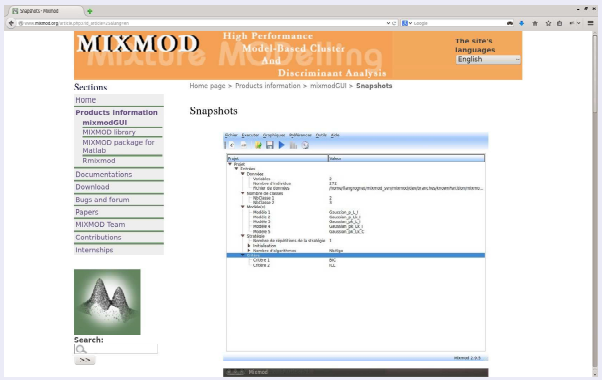
Classification des données en présence d'outliers

MIXMOD

Fiche d'identité

Diffuser auprès d'un large public un ensemble logiciel de classification par modèles de mélanges

- Projet initié en 2001 et très actif depuis 15 ans
- Soutenu par 5 organismes
- Licence GNU GPL
- www.mixmod.org
- Plusieurs milliers de téléchargements par an
- Domaines variés (finance, climatologie, génome, ...)

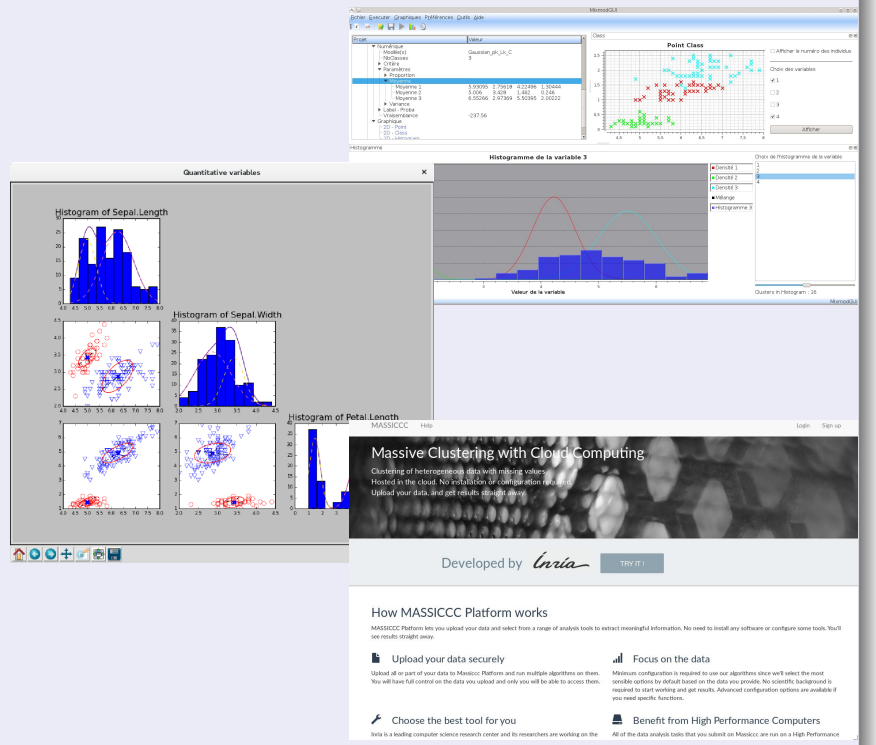


MIXMOD

Ensemble logiciel

Proposer des composants logiciels
adaptés aux demandes des utilisateurs

- 2002...2011 : composant pour Scilab
- 2002...(2015) : mixmodForMatlab
- Depuis 2011 : mixmodGUI
- Depuis 2012 : RMixmod
- 2016 : PyMixmod
- 2016 : MASSICCC



mixmodLib : Bibliothèque de calcul (C++)

Classification des données en présence d'outliers

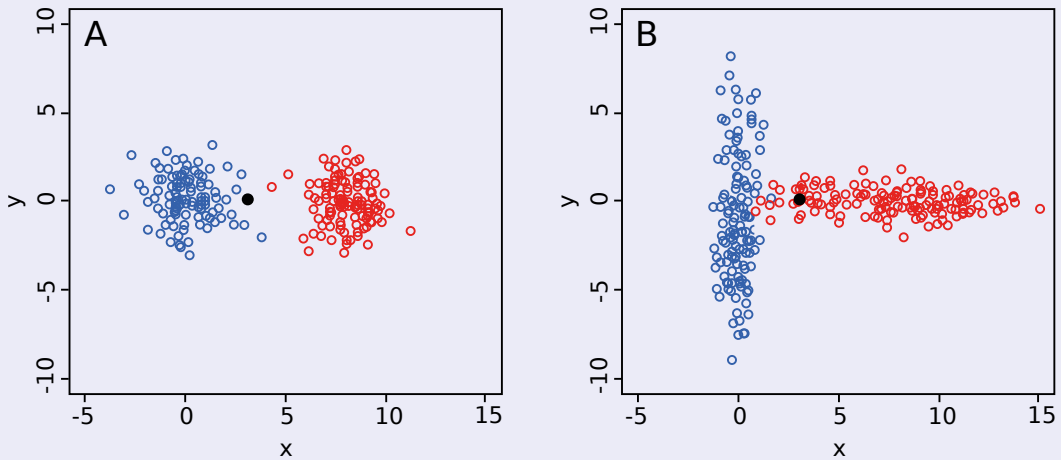
Fonctionnalités (1)

Problématiques traitées

- Classification non supervisée
- Classification supervisée

Modèles de mélange

- Outils souples pour modéliser un large spectre de situations
- Calcul des paramètres du modèle sous-jacent - Caractérisation des classes (proportion, moyenne, dispersion)
- Classification des individus avec des métriques adaptées



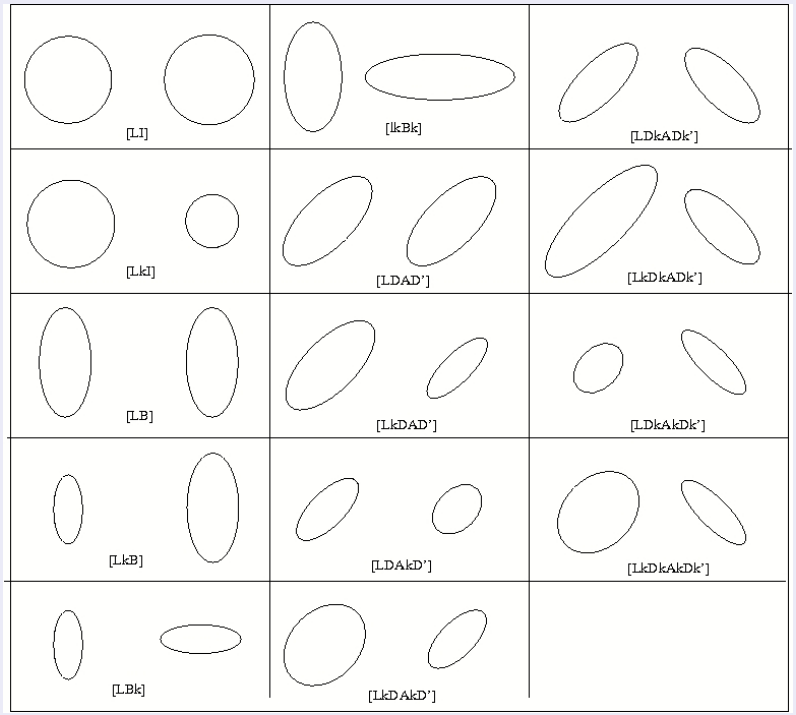
Fonctionnalités (2)

Modèles et métriques

Données quantitatives

14 modèles gaussiens

basés sur la décomposition en valeur sigulière de la matrice de variance



Données quantitatives en grande dimension

8 modèles spécifiques pour la grande dimension

Données qualitatives

5 modèles multinomiaux

basés sur une reparamétrisation de la distribution Multinomiale

Données mixtes

20 modèles hétérogènes

pour les données quantitatives/qualitatives

Fonctionnalités (3)

Algorithmes

Maximisation de la vraisemblance (ou vraisemblance complétée)

- EM (Expectation Maximisation)
- SEM (Stochatitic EM)
- CEM (Classification EM)

Critères

- BIC (Bayesian Information Criterion)
- ICL (Integrated Completed Likelihood)
- NEC (Normalized Entropy Criterion)
- CV (Cross Validation)

Initialisations et Stratégies

- 6 initialisations
  Ex : 'random', 'short runs of EM',...
- Algorithmes chaînés
  Ex: 100 iterations de **SEM** puis 50 iterations de **EM**

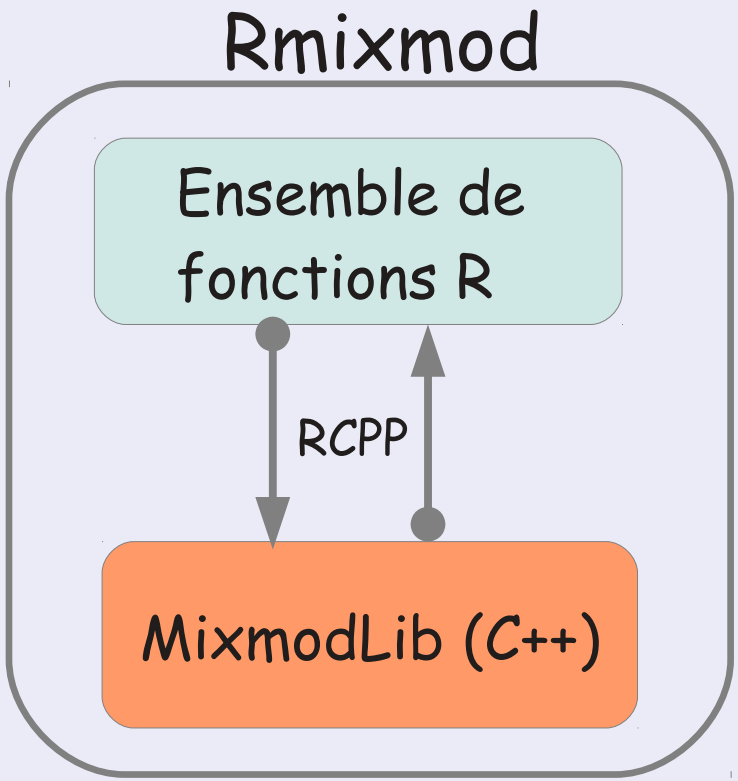
Et aussi...

- Connaissance partielle des labels des individus (semi-supervisé)
- Individus pondérés

Classification des données en présence d'outliers

Rmixmod : le package R de Mixmod

Architecture



Avantages

- Atouts de R
  - Environnement “familier” de l’utilisateur
  - Interface avec d’autres packages
  - Outils de visualisation
- Atouts de mixmodLib (C++)
  - Développée depuis 2001
  - Largement diffusée, et utilisée
  - Eprouvée, robuste, rapide

Classes - Fonctions

Classes Rmixmod

Classes (S4)

Mixmod
MixmodCluster [<-Mixmod]
MixmodLearn [<-Mixmod]
MixmodPredict
MixmodResults
MixmodDAResults [<-MixmodResults]
Model
MultinomialModel [<-Model]
GaussianModel [<-Model]
Parameter
GaussianParameter [<-Parameter]
MultinomialParameter [<-Parameter]
Strategy

Fonctions Rmixmod

Fonctions

mixmodCluster
mixmodLearn
mixmodPredict

mixmodStrategy
mixmodGaussianModel
mixmodMultinomialModel
sortByCriterion
nbFactorFromData

summary
print
hist
histCluster
plot
PlotCluster
barplot
barplotCluster

Etude de la robustesse de RMixmod

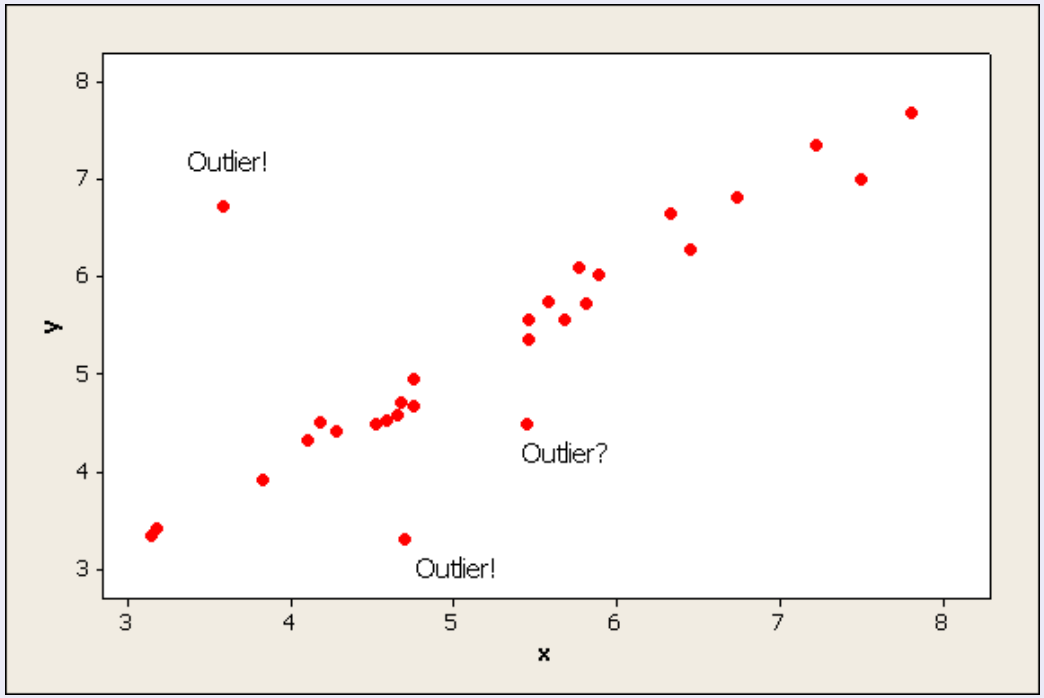
Qu'est ce qu'un outlier ?

Donnée aberrante

Outlier

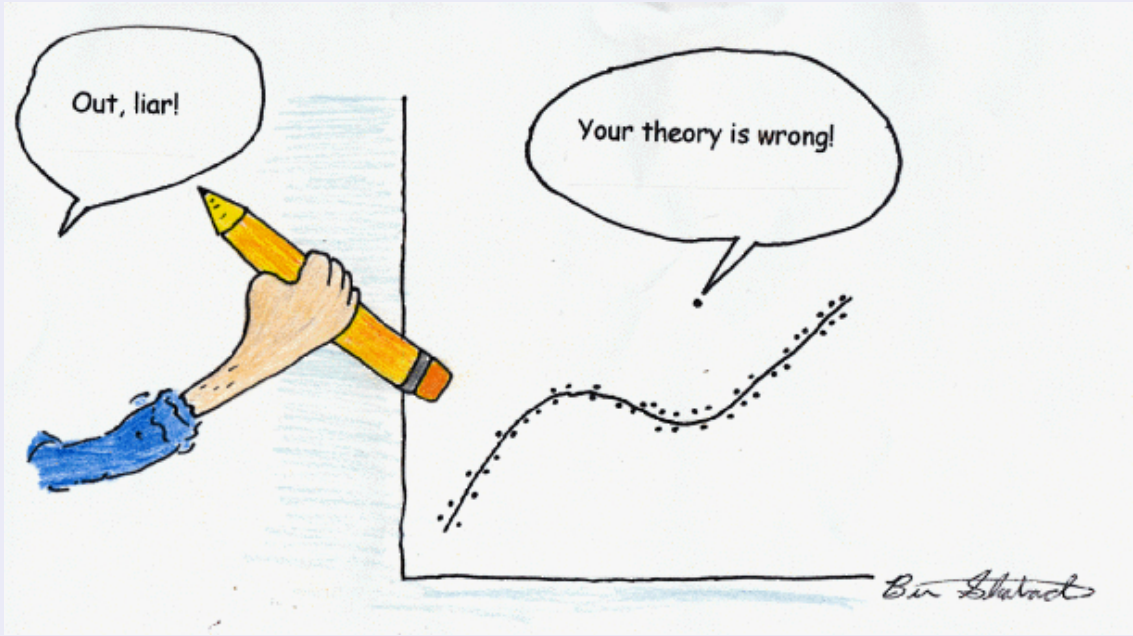
Une donnée qui ne ressemble pas aux autres

- Tout le monde “comprend” ce qu’est une donnée aberrante
- Mais donner une définition est complexe et subjectif
  Selon quel critère (de ressemblance) ?
- Sans critère objectif, il est tentant de considérer certains individus comme étant des outliers



Un outlier ce n'est pas...

Un individu qui ne "rentre" pas dans notre modèle



C'est le modèle qui doit s'adapter aux données et non l'inverse !

Les bonnes questions à se poser

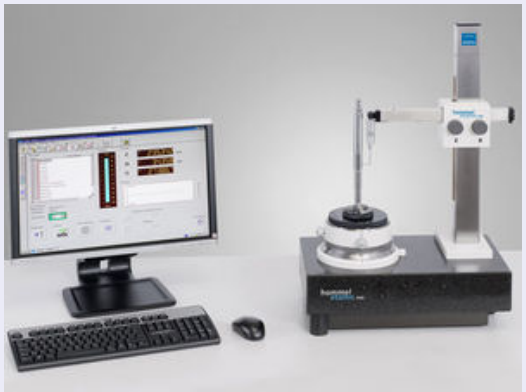
Si un individu ne rentre pas dans le modèle

- D'où vient cet individu ? Est-ce un individu de la population étudiée ?

Population étudiée VS outliers

Un outlier est un individu qui n'appartient pas à la population étudiée

En général, ce sont des erreurs expérimentales



- Erreurs de mesure

- Erreurs de recopie

On passe de donnée aberrante à donnée qui ne fait pas partie de la population étudiée
Un outlier peut alors 'ressembler" à un individu de la population étudiée...

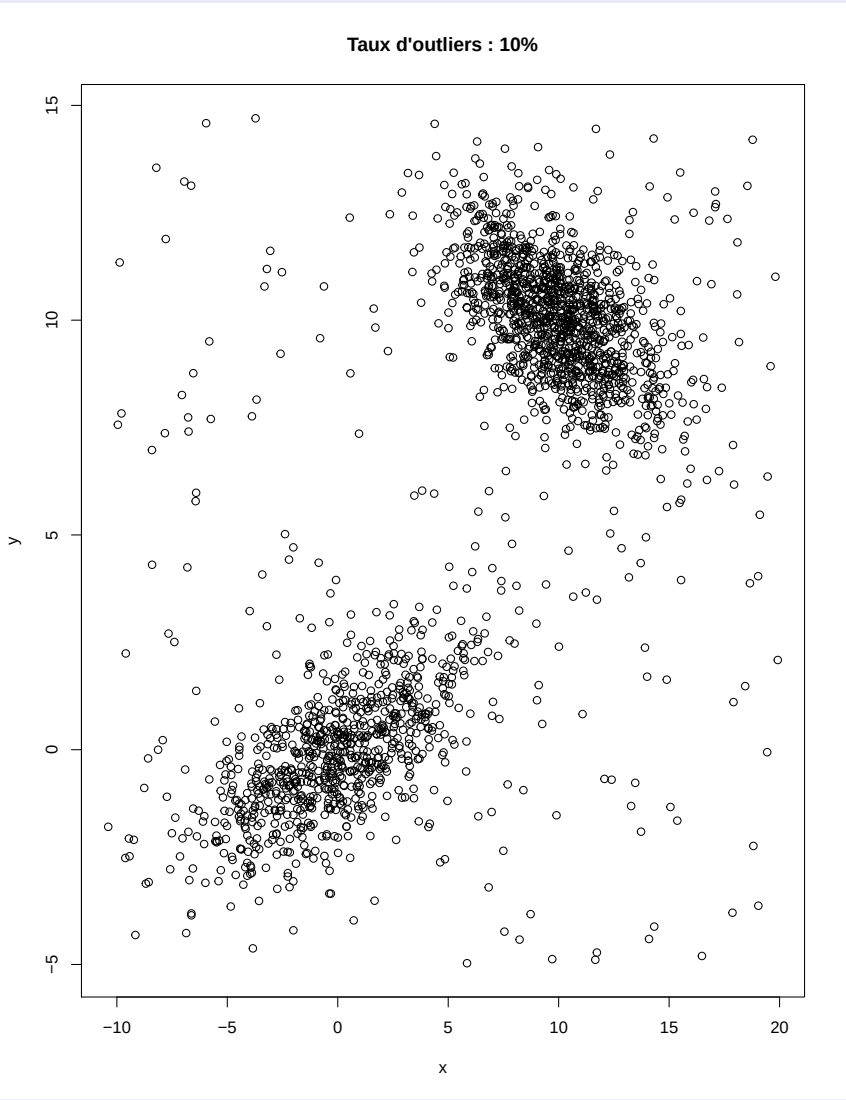
Etude de la robustesse de RMixmod

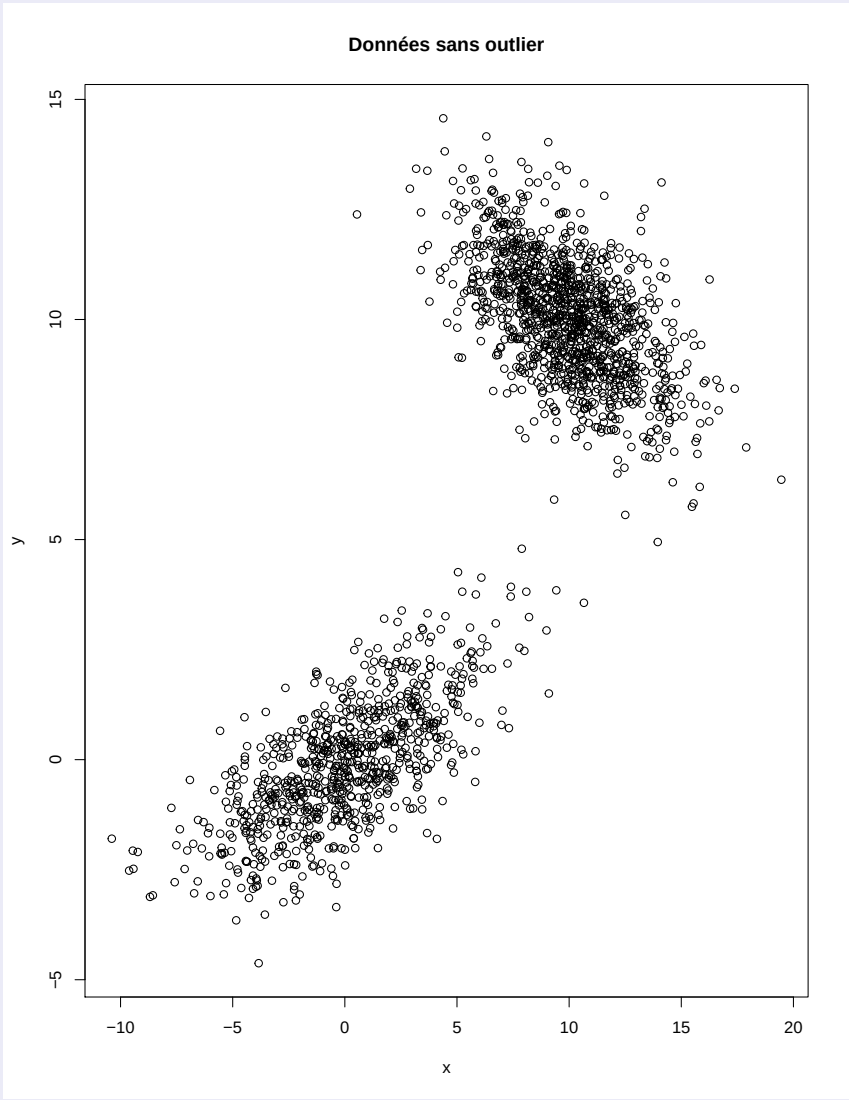
Classification des données en présence d'outliers

Nettoyage de la population

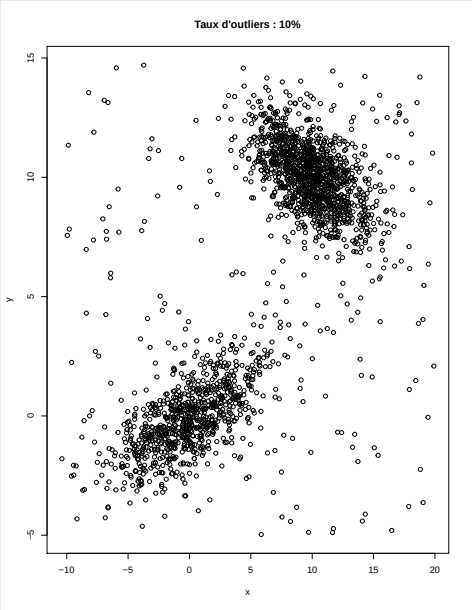
Nettoyage de la population

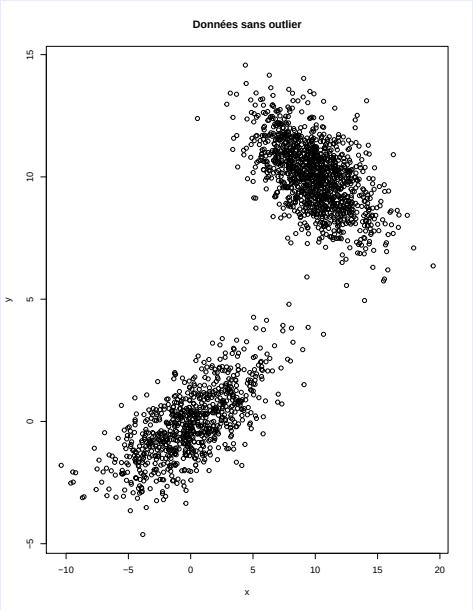
Enlever les outliers





Comment nettoyer les données ?



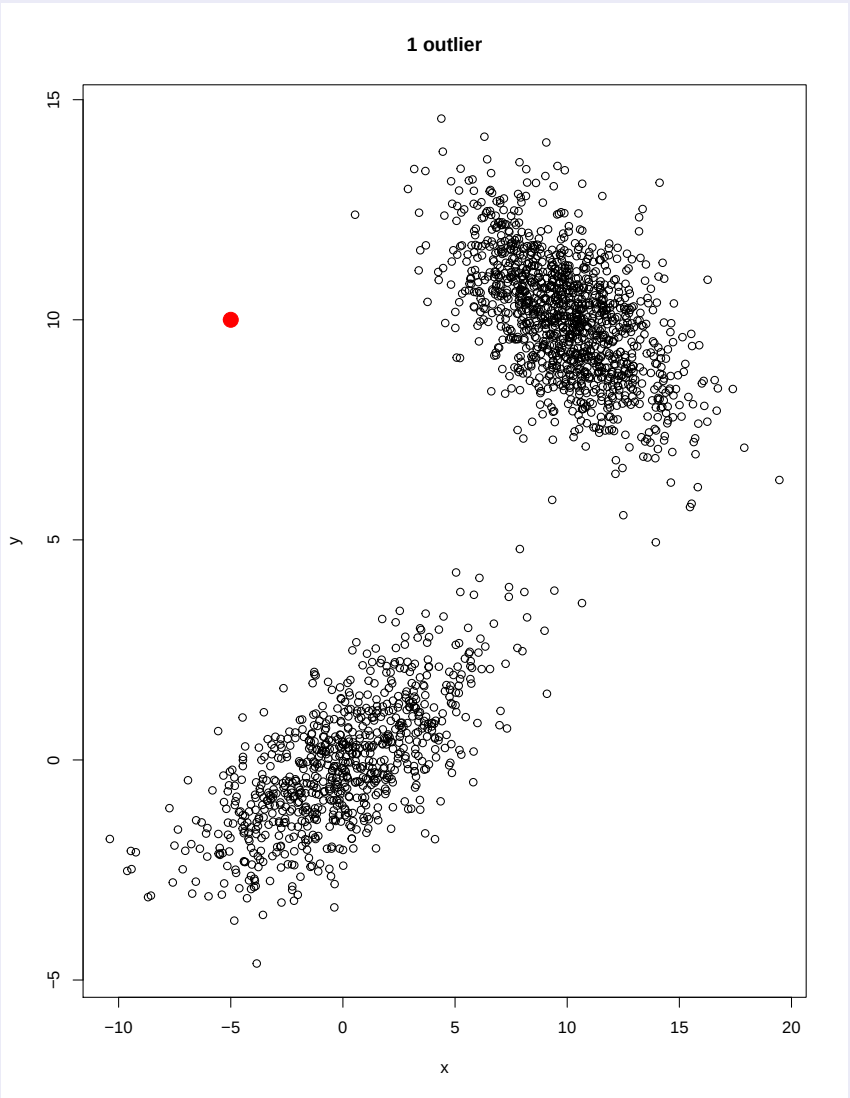


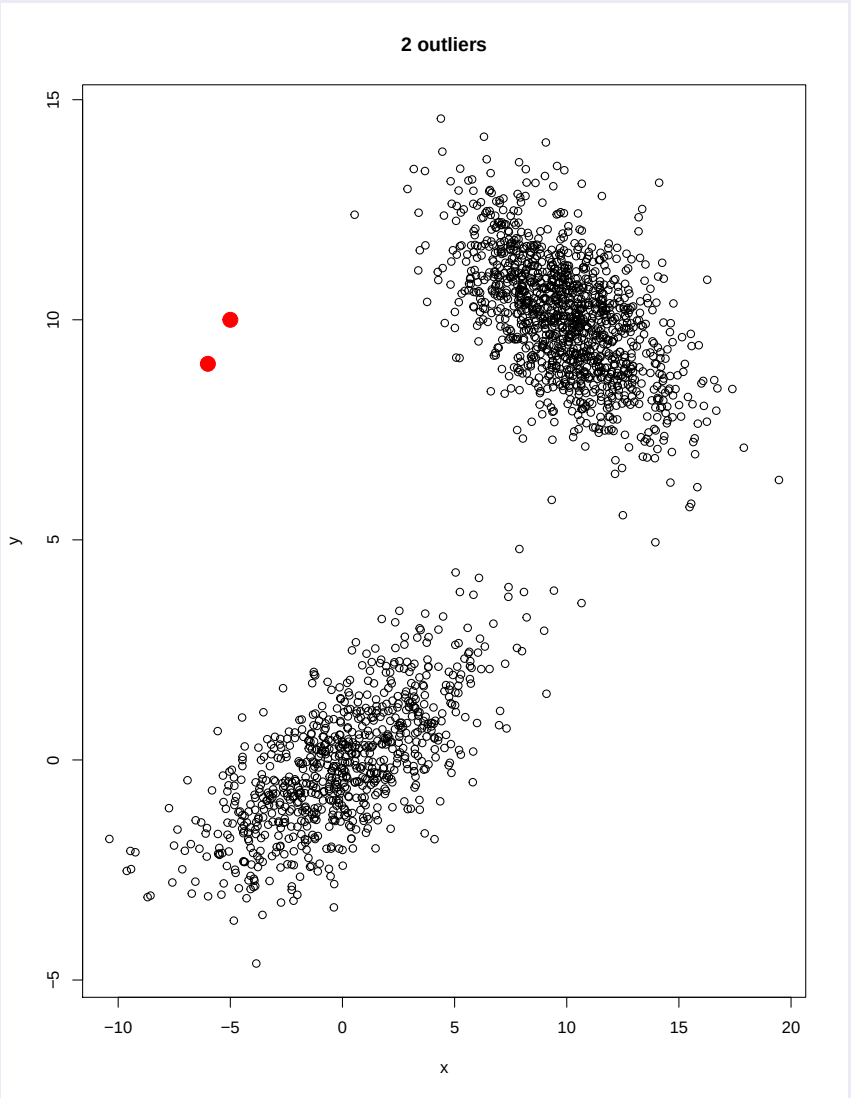
- Comment enlever des individus sans connaître le modèle ?
- Risques
  - Conserver à tord des outliers
  - Retirer à tord de ”vrais individus”

=> Perte d'information et biais sur la classification

Outliers ou classe supplémentaire ?

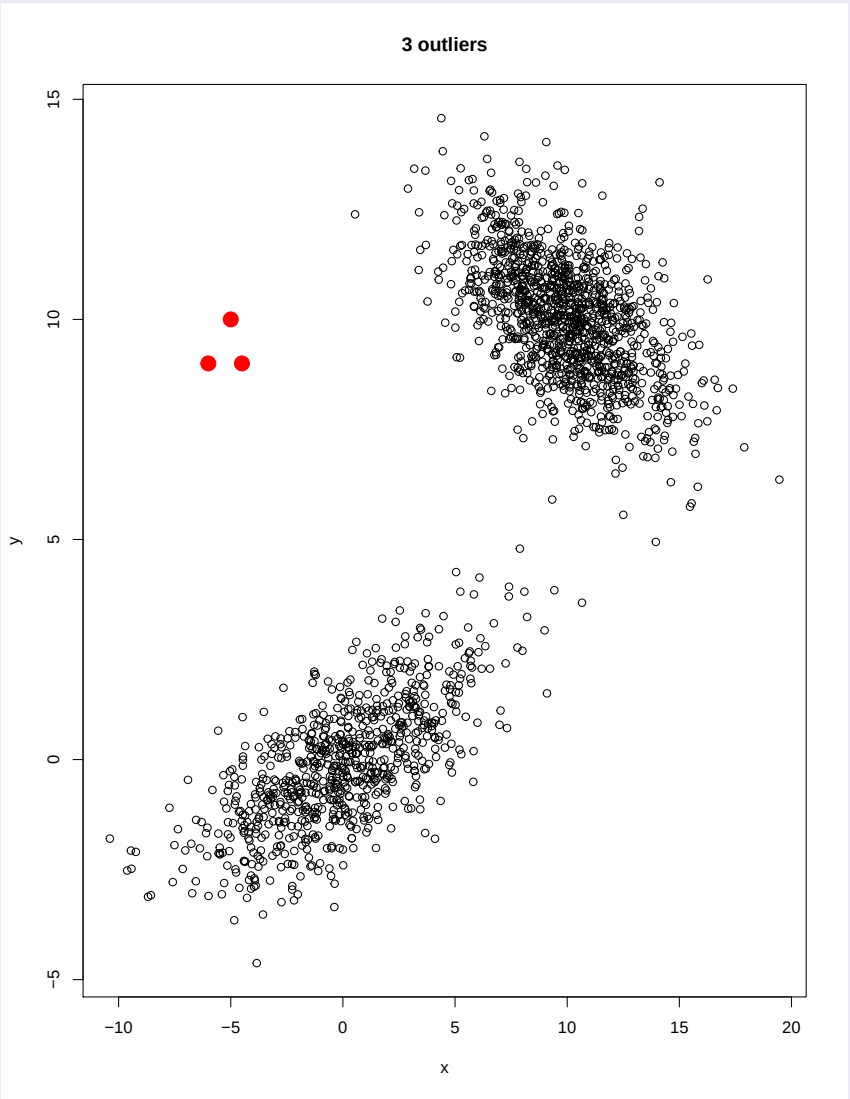
1 ou 2 outlier(s)

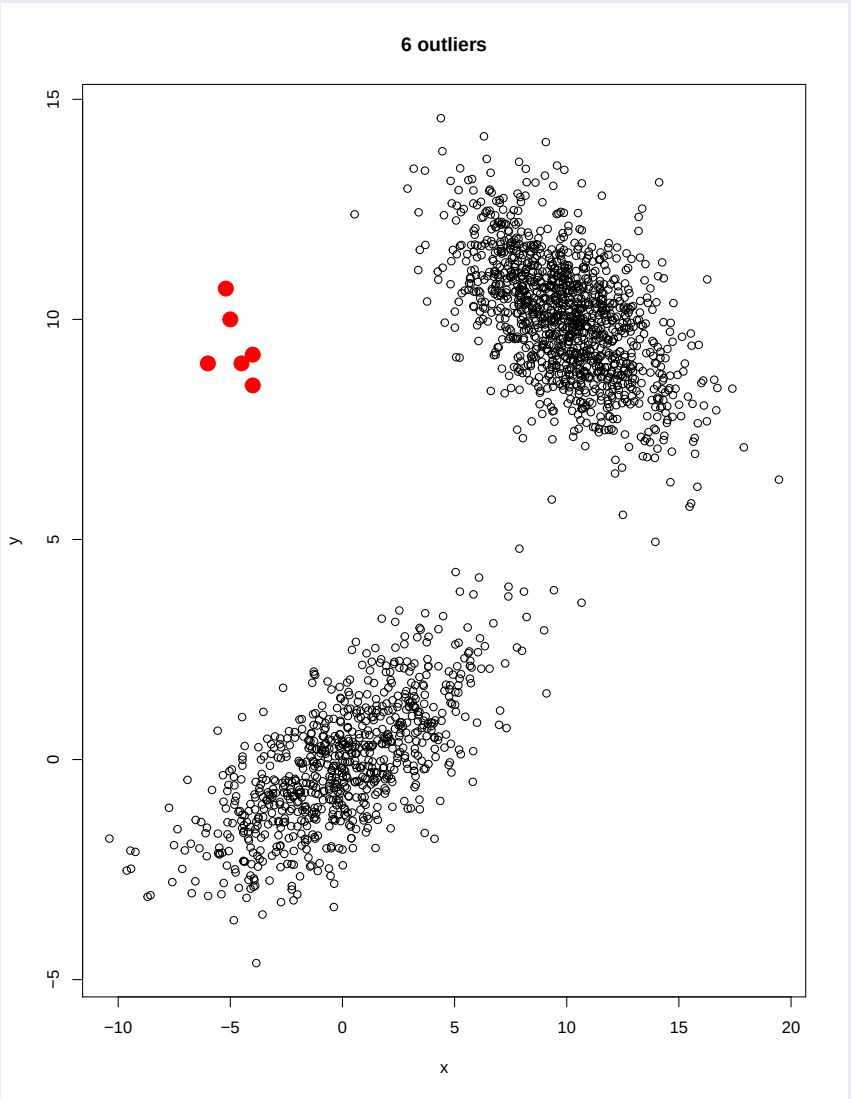




Outliers ou classe supplémentaire ?

3, 6 outliers





Sans nettoyage de la population

Sans nettoyage de la population

Sans nettoyage

Le nettoyage est risqué, subjectif et difficile (sans connaissance du modèle sous-jacent)

Et si l'on considérait une classe de bruit ?

Classification sur l'ensemble de la population en $k + 1$ clusters
k *vraies classes et une classe de bruit*

Classification des données en présence d'outliers

RMixmod et traitement des outliers

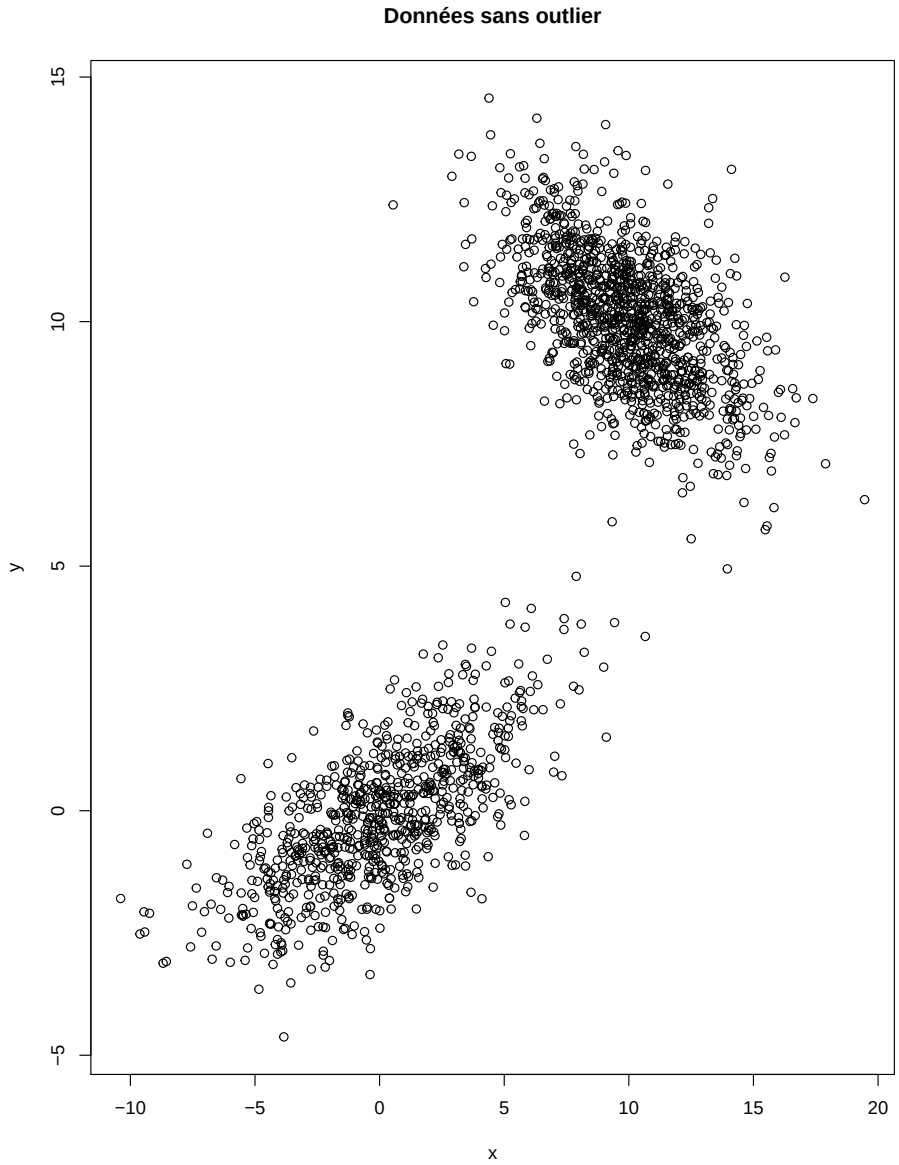
Jeux de donnnées

- Jeu de données simulées (on connaît les paramètres)
  Mélange de 2 gaussiennes

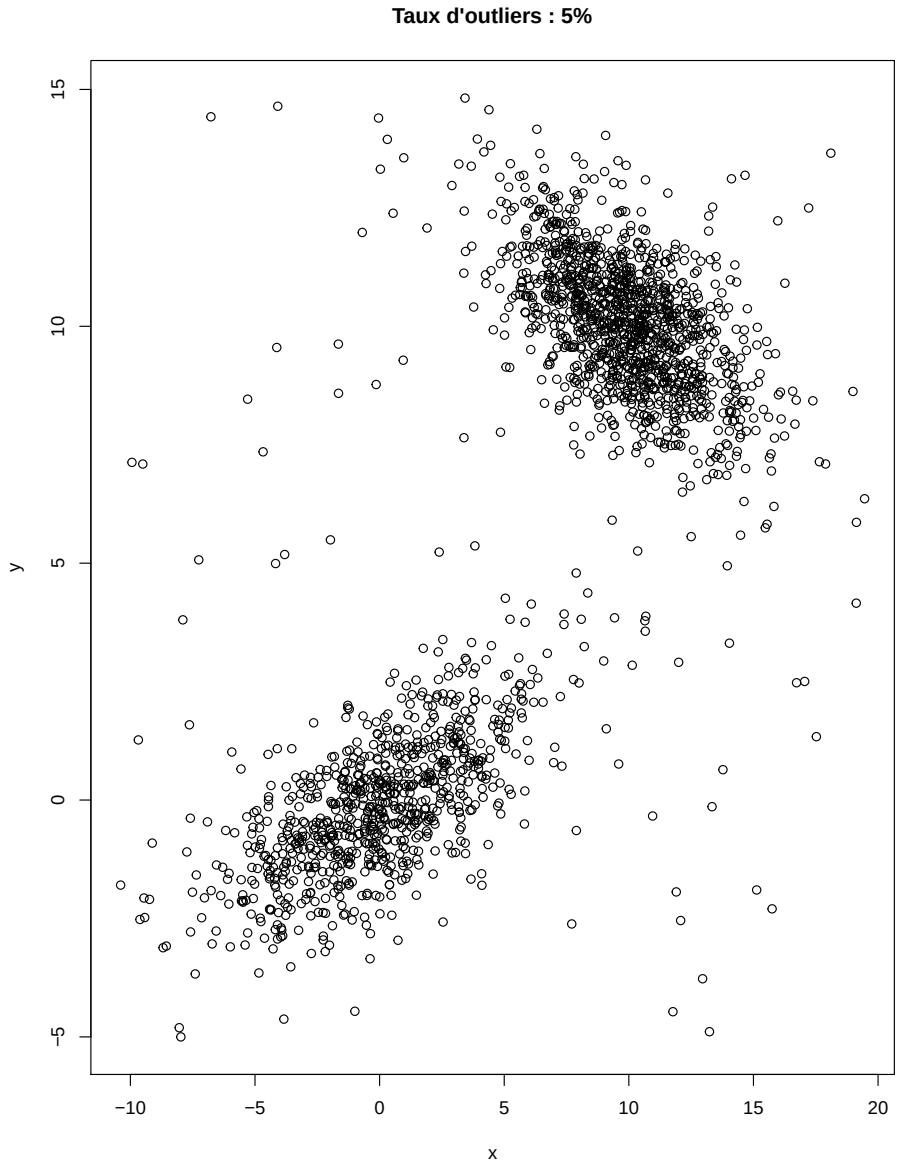
p	0.4	0.6
μ	(0,0)	(10,10)
Σ	$\begin{pmatrix} 10 & 3 \\ 3 & 2 \end{pmatrix}$	$\begin{pmatrix} 6 & -2 \\ -2 & 2 \end{pmatrix}$

- On ajoute des outliers
  - Selon une loi gaussienne
    Pas de difficulté (mélange gaussien)
  - Selon une loi uniforme
    Comment RMixmod peut traiter ces données ?

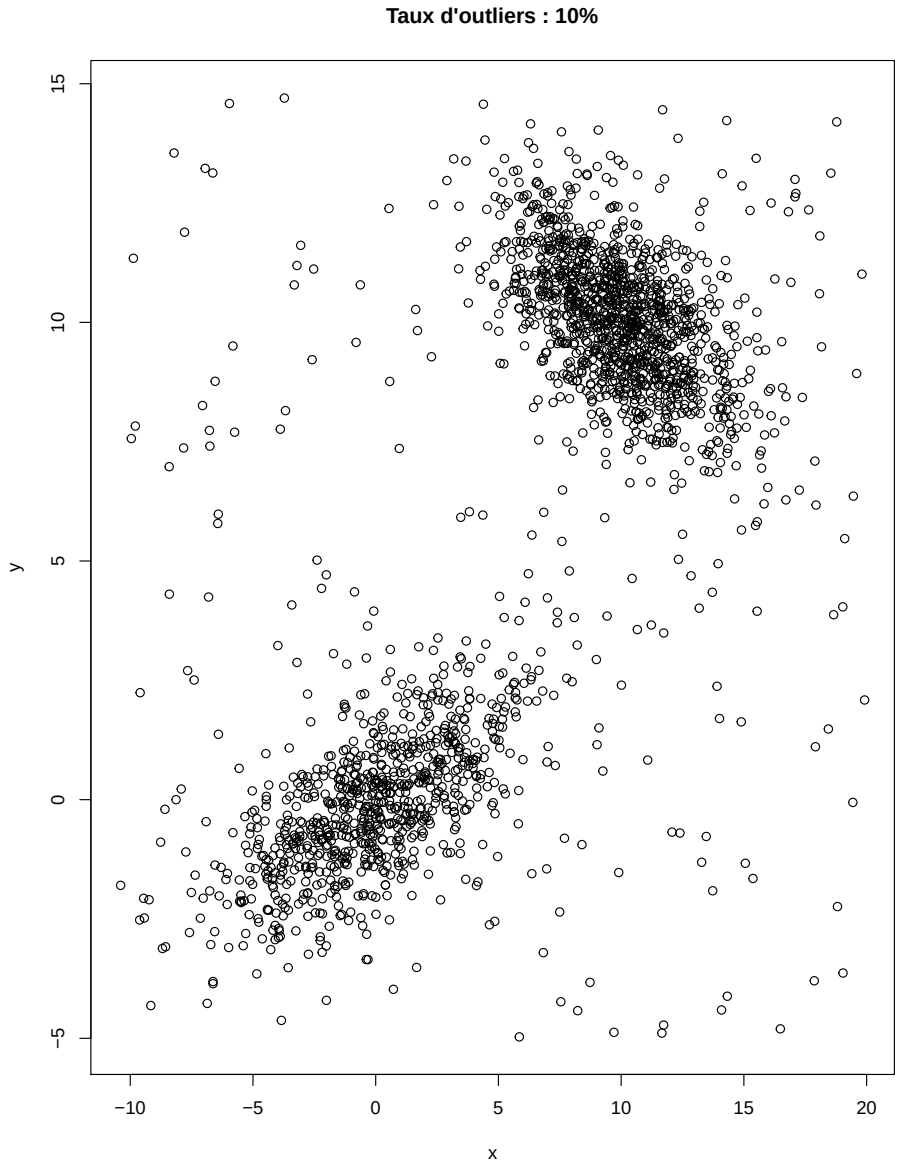
Jeu de données



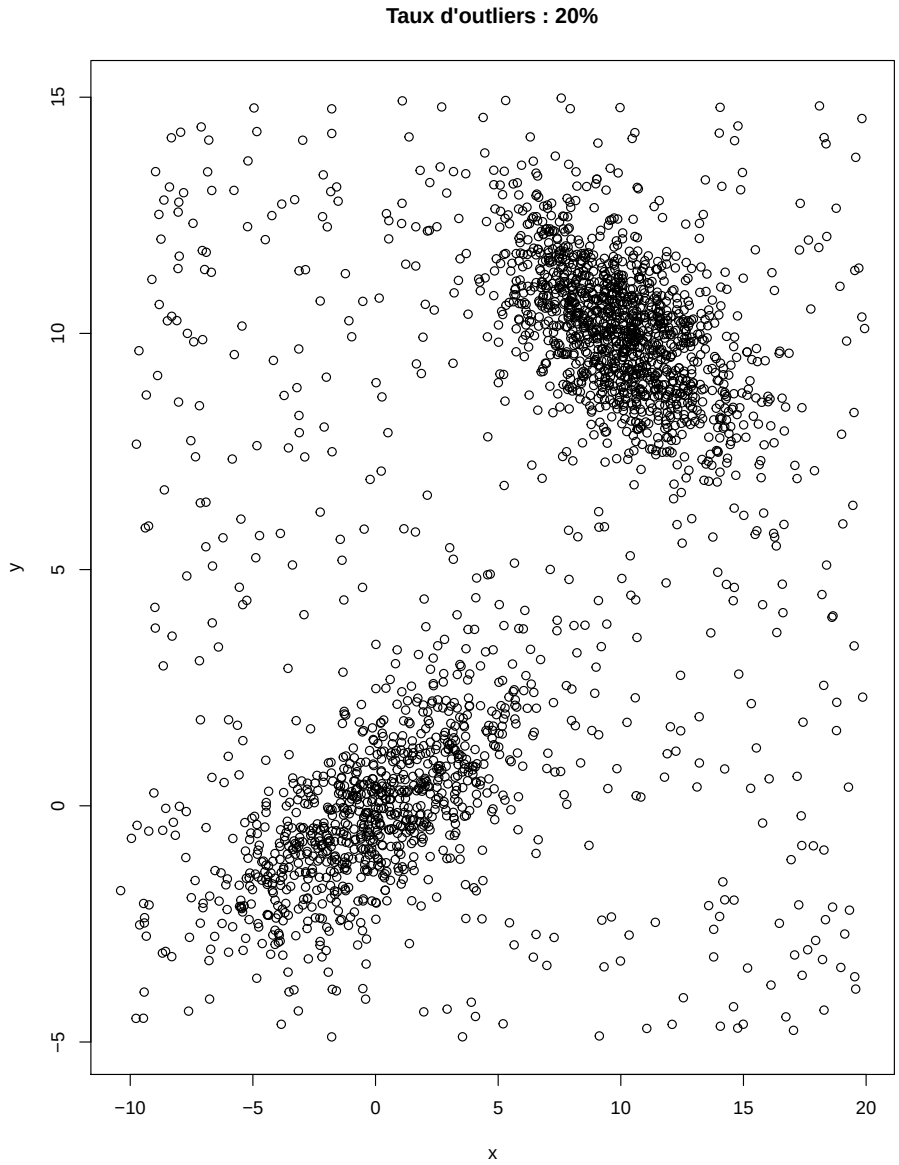
Jeu de données + 5% d'outliers



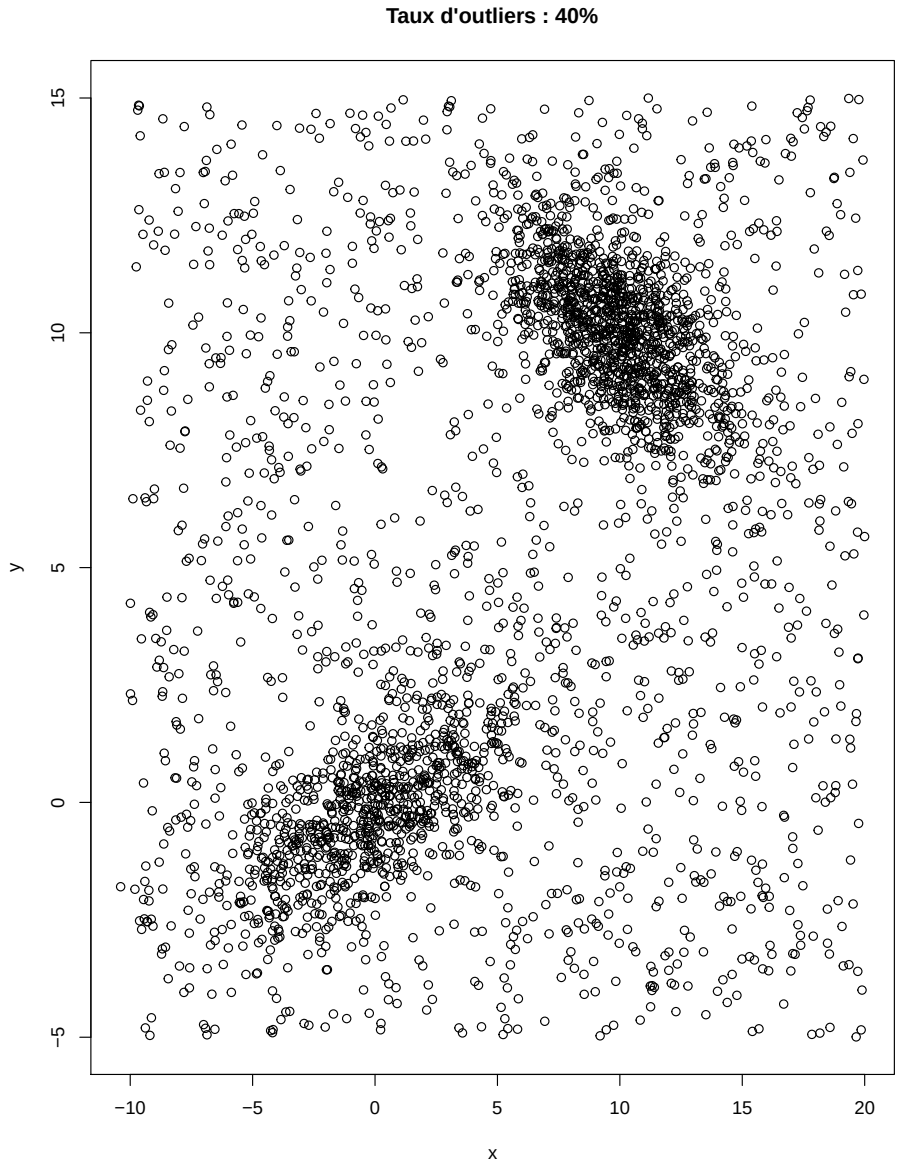
Jeu de données + 10% d'outliers



Jeu de données + 20% d'outliers



Jeu de données + 40% d'outliers



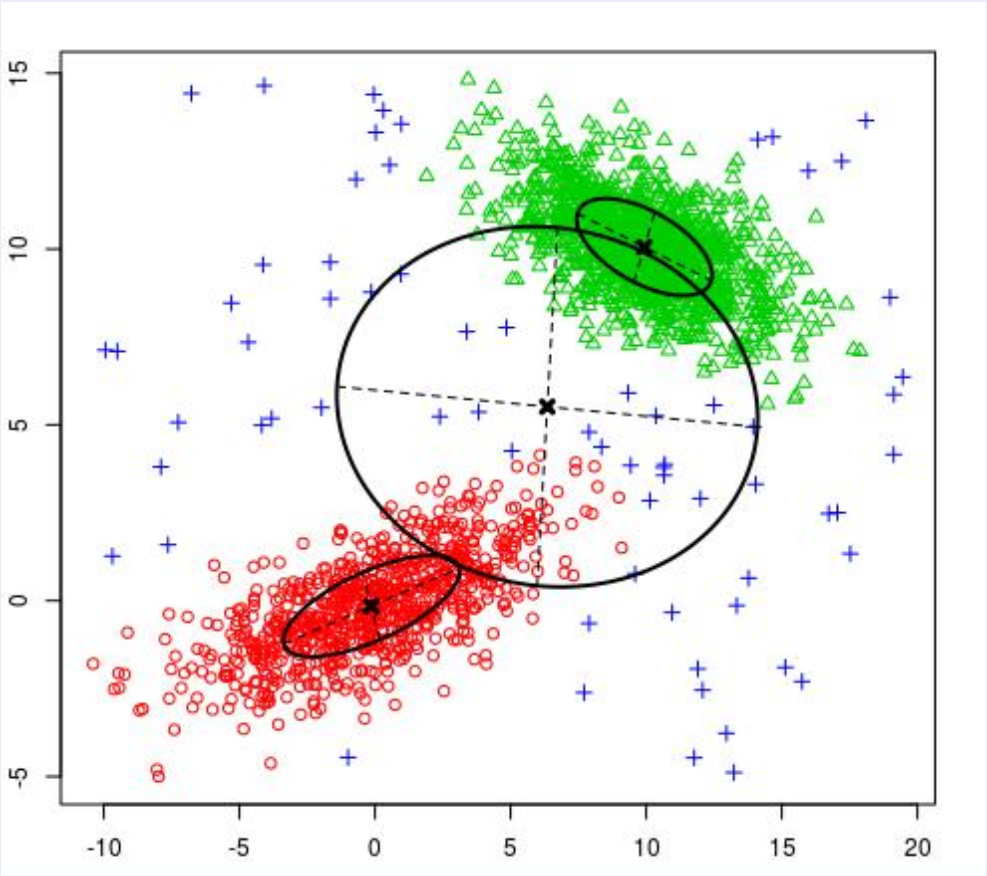
Rmixmod

Commandes R

```
library(Rmixmod)

model<-mixmodGaussianModel(listModels=
  c("Gaussian_pk_Lk_Ck"))

out<-mixmodCluster(data_outliers, nbCluster=3,
  models=model)

plot(out)

summary(out)
```

Jeu de données + 5% d'outliers

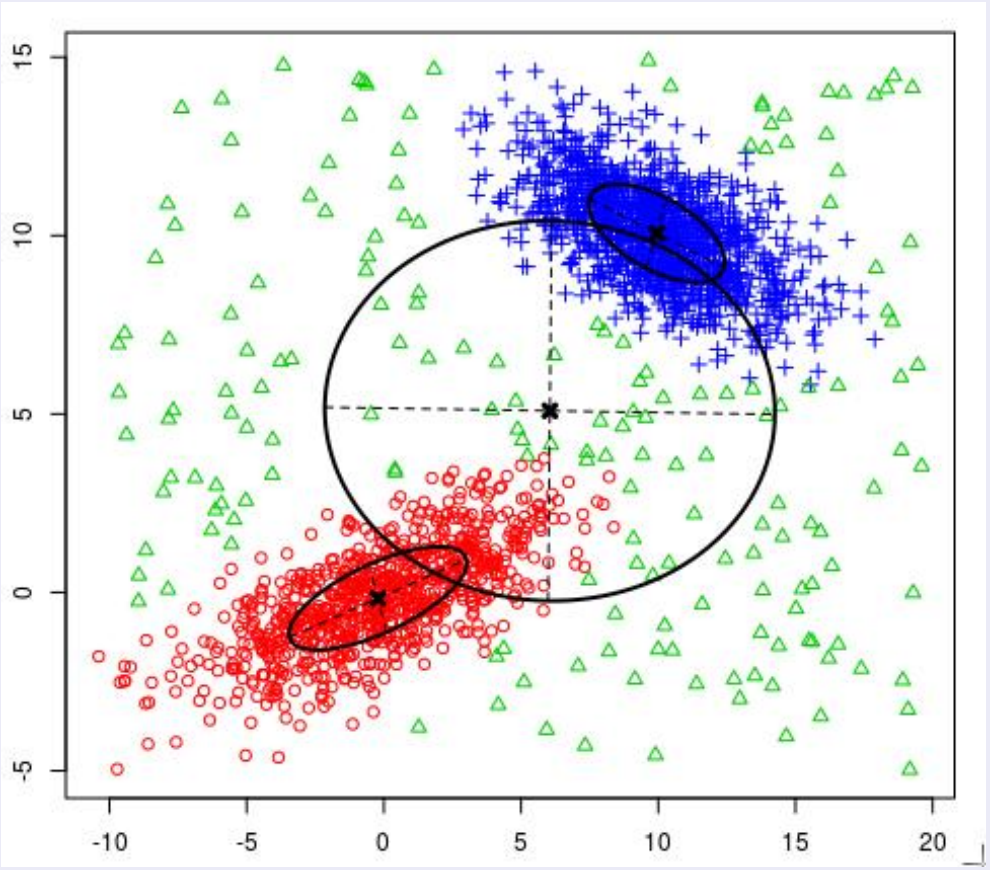
Résultats




```
**************************************************************
* Number of samples     =  2105
* Problem dimension     =  2
**************************************************************
*       Number of cluster =  3
*              Model Type =  Gaussian_pk_Lk_Ck
*               Criterion =  BIC(20448.4361)
*              Parameters =  list by cluster
*                  Cluster  1 :
                         Proportion =  0.5697
                              Means =  9.9582 10.0515
                          Variances = |     6.2942    -1.9716 |
                                      |    -1.9716     1.8706 |
*                  Cluster  2 :
                         Proportion =  0.3852
                              Means =  -0.1527 -0.1408
                          Variances = |    10.4635     3.1552 |
                                      |     3.1552     2.0714 |
*                  Cluster  3 :
                         Proportion =  0.0451
                              Means =  5.8006 5.8317
                          Variances = |    67.3496    -8.0327 |
                                      |    -8.0327    27.3253 |
*          Log-likelihood =  -10159.1754
**************************************************************
```

Jeu de données + 10% d'outliers

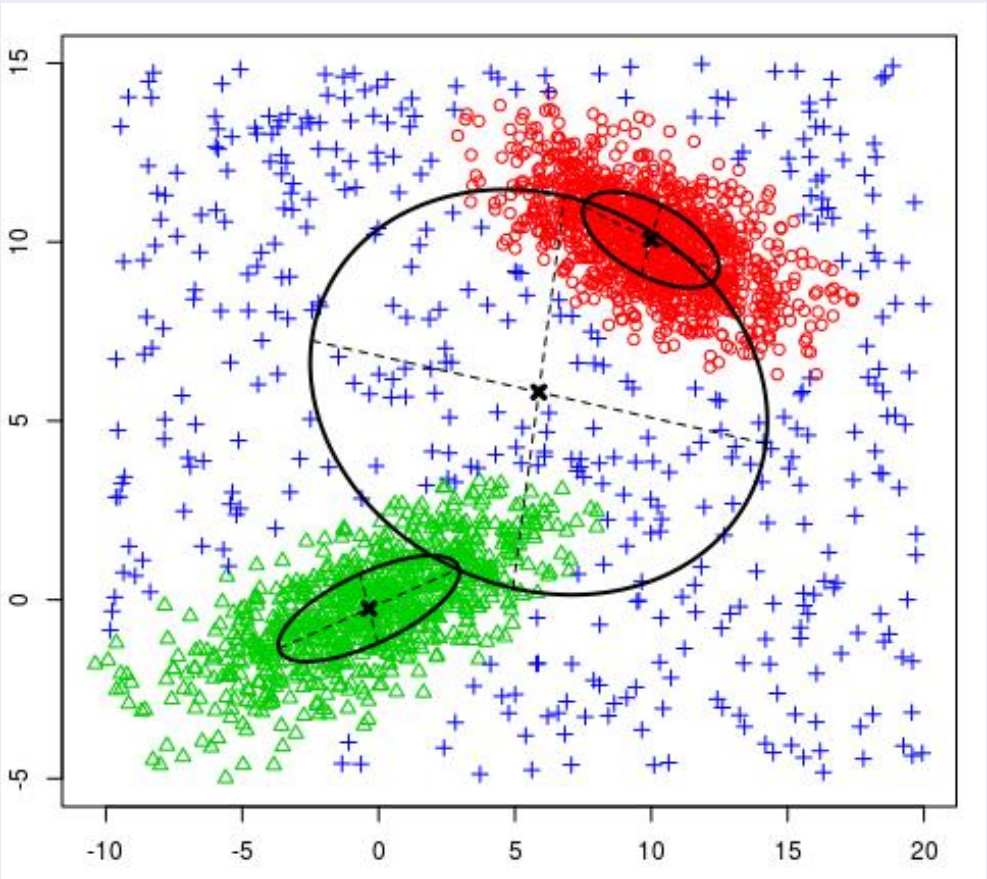
Résultats




```
************************************************************
* Number of samples     =  2222
* Problem dimension     =  2
************************************************************
*       Number of cluster =  3
*              Model Type =  Gaussian_pk_Lk_Ck
*               Criterion =  BIC(22482.1398)
*              Parameters =  list by cluster
*                  Cluster  1 :
                         Proportion =  0.3586
                              Means =  -0.2188 -0.1673
                          Variances =  |    10.5094     3.2179 |
                                       |     3.2179     2.0876 |
*                  Cluster  2 :
                         Proportion =  0.1056
                              Means =  6.0493 5.0887
                          Variances =  |    67.2919    -0.4733 |
                                       |    -0.4733    28.3726 |
*                  Cluster  3 :
                         Proportion =  0.5358
                              Means =  9.9469 10.0639
                          Variances =  |     6.1423    -1.9812 |
                                       |    -1.9812     1.8866 |
*          Log-likelihood =  -11175.5675
************************************************************
```

Jeu de données + 20% d'outliers

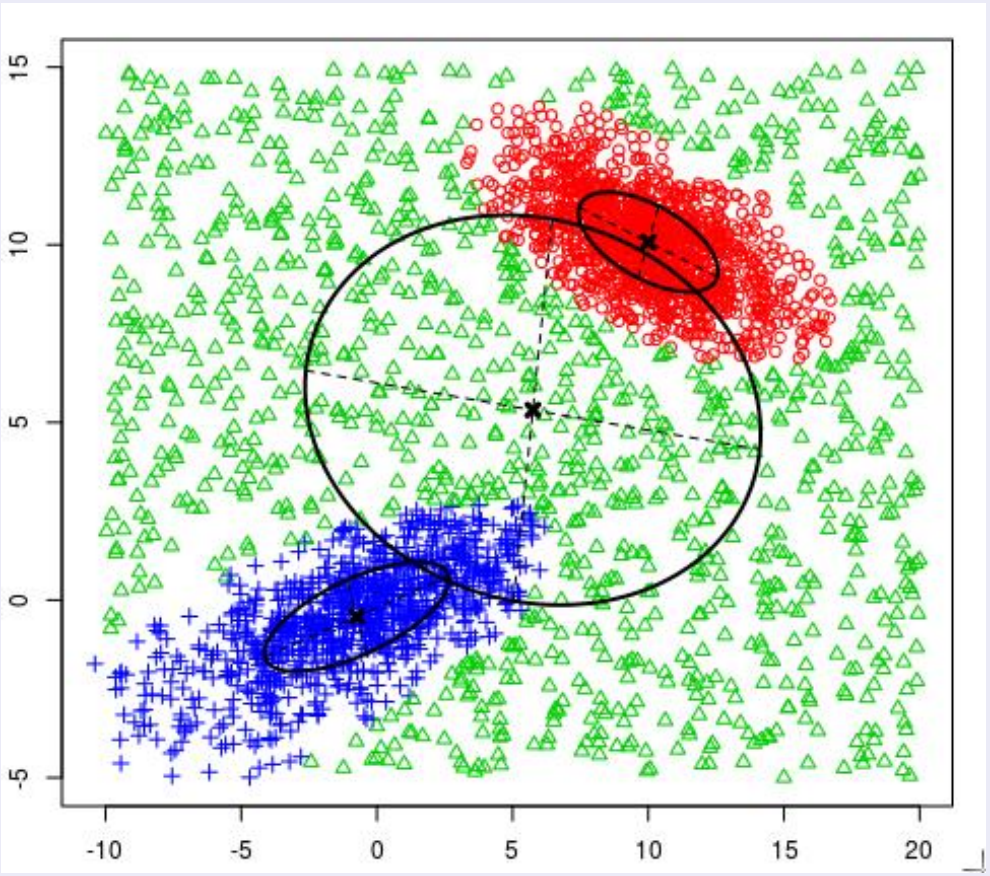
Résultats




```
************************************************************
* Number of samples     =  2500
* Problem dimension     =  2
************************************************************
*       Number of cluster =  3
*              Model Type =  Gaussian_pk_Lk_Ck
*               Criterion =  BIC(26985.3772)
*              Parameters =  list by cluster
*                  Cluster  1 :
                         Proportion =  0.4687
                              Means =  9.9897 10.0608
                          Variances = |     6.2357    -1.9286 |
                                      |    -1.9286     1.8030 |
*                  Cluster  2 :
                         Proportion =  0.3265
                              Means =  -0.3556 -0.2485
                          Variances = |    11.0408     3.3911 |
                                      |     3.3911     2.1923 |
*                  Cluster  3 :
                         Proportion =  0.2048
                              Means =  5.8745 5.8028
                          Variances = |    70.2530    -6.8008 |
                                      |    -6.8008    32.0964 |
*          Log-likelihood =  -13426.1842
************************************************************
```

Jeu de données + 40% d'outliers

Résultats




```
****************************************************************
* Number of samples     =  3333
* Problem dimension     =  2
****************************************************************
*       Number of cluster =  3
*              Model Type =  Gaussian_pk_Lk_Ck
*               Criterion =  BIC(39110.3155)
*              Parameters =  list by cluster
*                  Cluster  1 :
                         Proportion =  0.3646
                              Means =  10.0035 10.0770
                          Variances = |     6.6177    -2.0875 |
                                      |    -2.0875     1.9659 |
*                  Cluster  2 :
                         Proportion =  0.3776
                              Means =  5.7464 5.3491
                          Variances = |    70.4542    -5.4527 |
                                      |    -5.4527    30.1093 |
*                  Cluster  3 :
                         Proportion =  0.2579
                              Means =  -0.7390 -0.4682
                          Variances = |    11.5557     3.4871 |
                                      |     3.4871     2.3748 |
*          Log-likelihood =  -19486.2089
****************************************************************
```

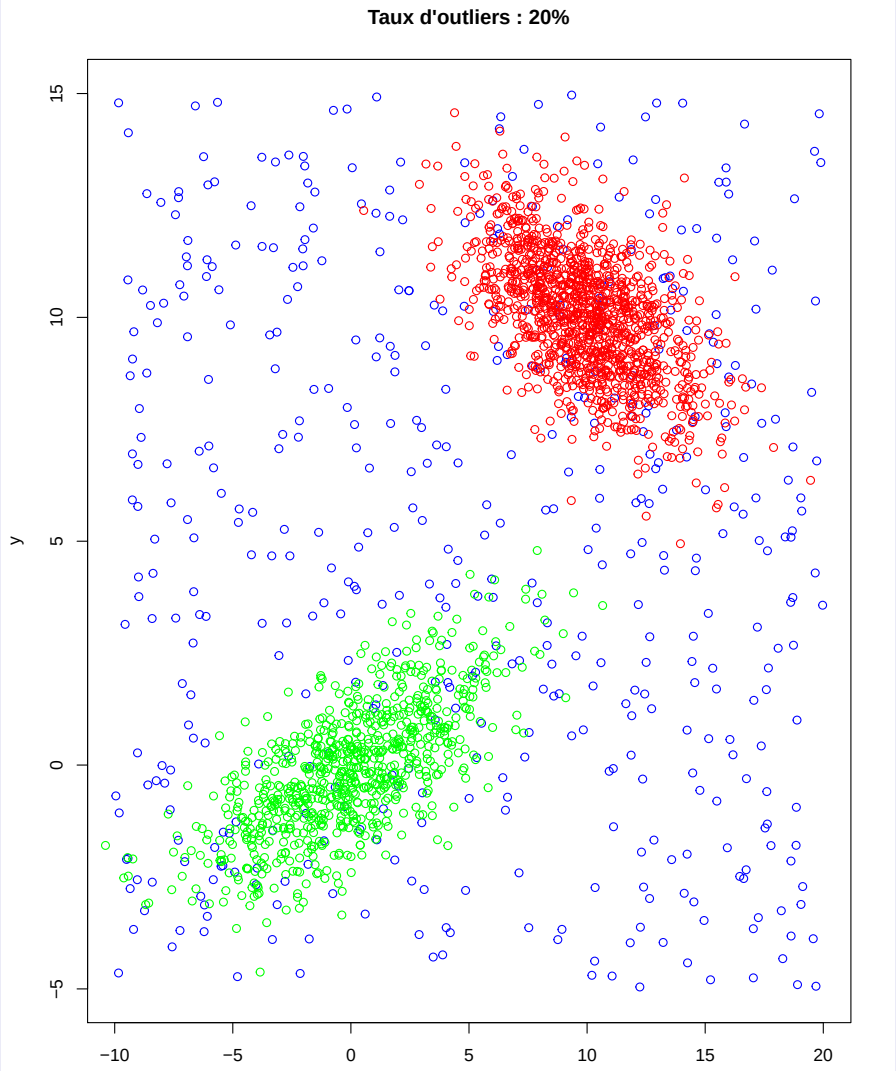
Analyse des résultats

20% d'outliers

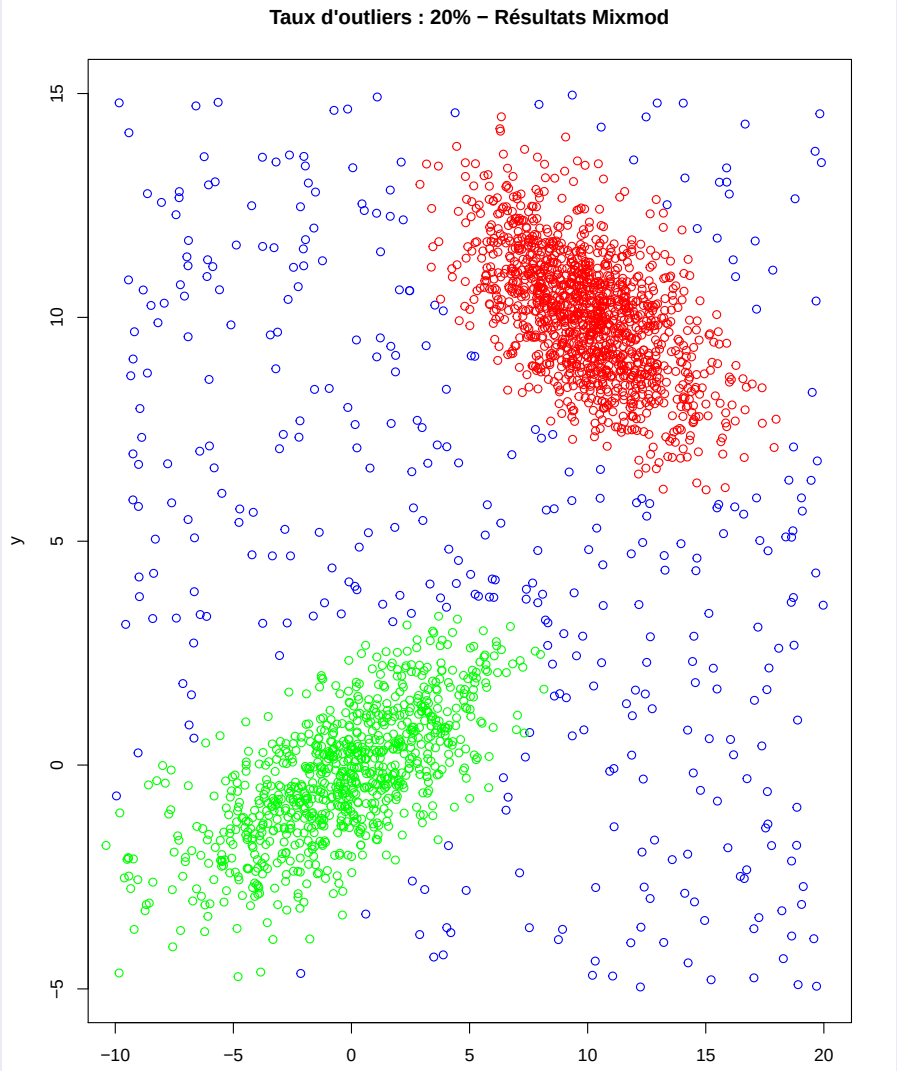
Analyse des résultats

Labels

Données initiales



Résultats RMixmod



Analyse des résultats

Qualité du classement (labels)

	classe 1	classe 2	outliers
classe 1	98%	0%	2%
classe 2	0%	98.7%	1.3%
outliers	16%	17%	67%

- Classes 1 et 2 : Excellent taux de reclassement
- Outliers : Très bon taux de reclassement
  - Certains individus (tirés au hasard selon la loi unforme) se trouvent dans le ”domaine” des classes 1 ou 2
  - Ils sont donc considérés comme des individus de ces classes
  - => Pas de biais sur l'estimation des paramètres

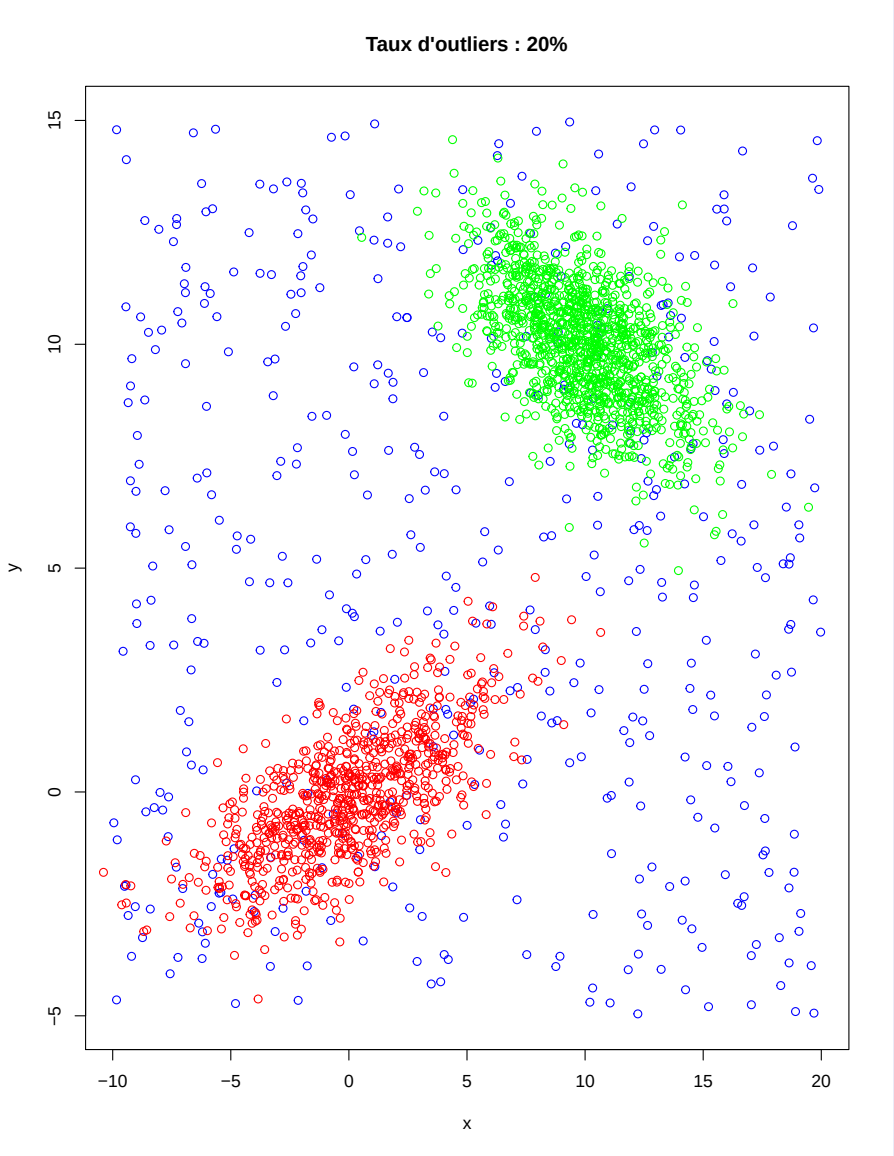
Analyse des résultats

Paramètres

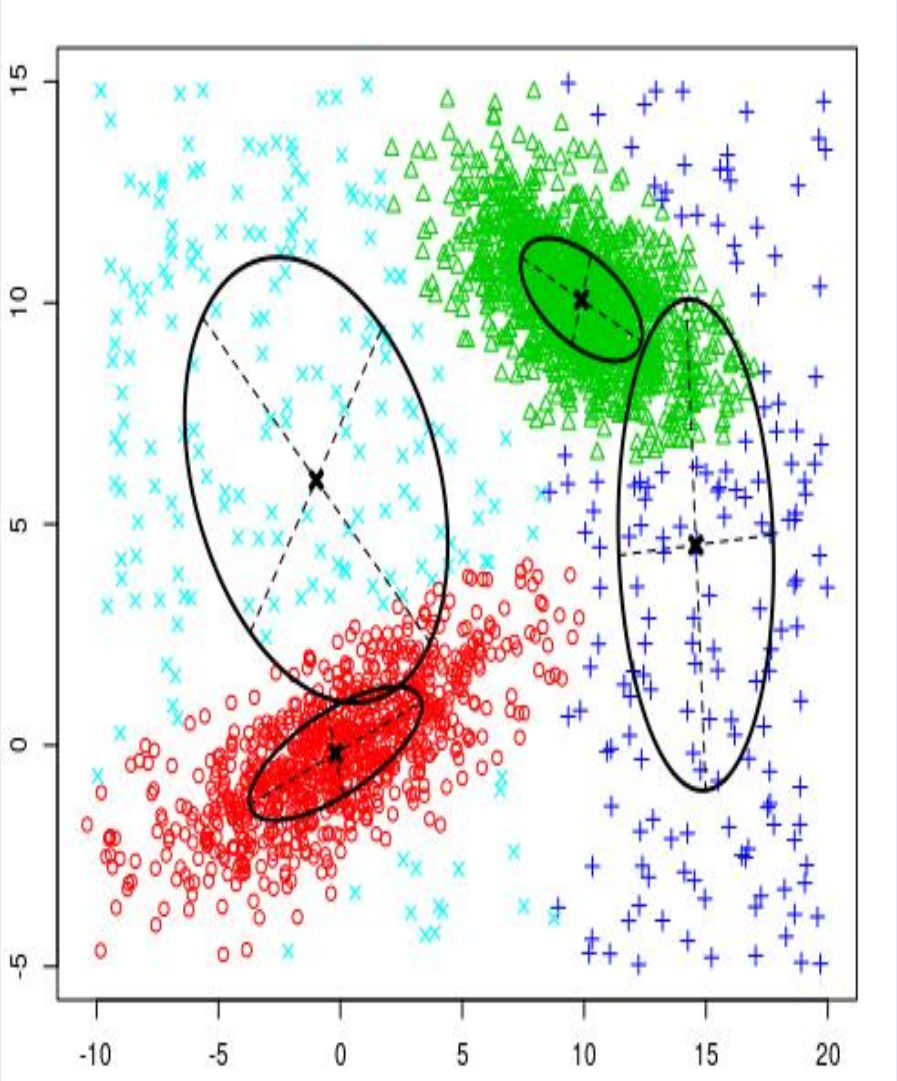
```
*****************************************************************
* Number of samples     =  2500
* Problem dimension     =  2
*****************************************************************
*       Number of cluster =  3
*              Model Type =  Gaussian_pk_Lk_Ck
*               Criterion =  BIC(26985.3772)
*              Parameters =  list by cluster
*                  Cluster  1 :
                         Proportion =  0.4687
                              Means =  9.9897 10.0608
                          Variances = |     6.2357    -1.9286 |
                                      |    -1.9286     1.8030 |
*                  Cluster  2 :
                         Proportion =  0.3265
                              Means =  -0.3556 -0.2485
                          Variances = |    11.0408     3.3911 |
                                      |     3.3911     2.1923 |
*                  Cluster  3 :
                         Proportion =  0.2048
                              Means =  5.8745 5.8028
                          Variances = |    70.2530    -6.8008 |
                                      |    -6.8008    32.0964 |
*          Log-likelihood =  -13426.1842
*****************************************************************
```

Et si on augmentait le nombre de classes ?

4 classes ?



Résultats RMixmod



4 classes - Analyse des résultats

Labels

	classe 1	classe 2	outliers (classe 3 et 4)
classe 1	97.8%	0%	2.2%
classe 2	0%	98.9%	1.1%
outliers	15.6%	15.8%	68.6%

Qualité de classification équivalente qu'avec 3 classes

4 classes - Analyse des résultats

Paramètres

```
*************************************************************
* Number of samples     =  2500
* Problem dimension     =  2
*************************************************************
*       Number of cluster =  4
*              Model Type =  Gaussian_pk_Lk_Ck
*               Criterion =  BIC(26821.5970)
*              Parameters =  list by cluster
*                  Cluster  1 :
                         Proportion =  0.1348
                              Means =  2.4304 4.8879
                          Variances = |    44.5556   -13.4250 |
                                      |   -13.4250    27.5719 |
*                  Cluster  2 :
                         Proportion =  0.4882
                              Means =  9.8424 10.1053
                          Variances = |     5.8952    -1.8743 |
                                      |    -1.8743     1.9807 |
*                  Cluster  3 :
                         Proportion =  0.0525
                              Means =  16.2754 5.4183
                          Variances = |     5.3842    -1.9707 |
                                      |    -1.9707    32.5546 |
*                  Cluster  4 :
                         Proportion =  0.3245
                              Means =  -0.3508 -0.2418
                          Variances = |    11.5645     3.5182 |
                                      |     3.5182     2.1714 |
*          Log-likelihood =  -13320.8220
*************************************************************
```

En augmentant le nombre de classes, on modélise mieux les outliers (meilleure vraisemblance)

Etude de la robustesse de RMixmod

Conclusion

Classification en présence d'outliers avec RMixmod

- En considérant une classe pour les outliers, RMixmod permet
  - d'obtenir une excellente qualité de classification (labels)
  - d'estimer avec précision les paramètres des "vraies" classes sans biais (les outliers ayant été affectés dans leur classe)

- Les modèles de mélanges de lois gaussiennes montrent leur puissance et leur souplesse
  Solution efficace et robuste y compris lorsque les individus qui ne suivent pas une loi gaussienne

Classification en présence d'outliers avec RMixmod

FIN

Merci de votre attention

- Site web : http://www.mixmod.org
- contact :
  - contact@mixmod.org
  - florent.langrognet@univ-fcomte.fr