Classification en présence d'outliers (données aberrantes)
avec RMixmod (package de classification par modèles de mélanges)

F. Langrognet[a]

[a]Laboratoire de Mathématiques de Besançon
UMR 6623 - Université de Franche-Comté - CNRS
16, route de Gray - 25030 Besançon
florent.langrognet@univ-fcomte.fr

Mots clefs : classification, modèles de mélanges, outliers

Les modèles de mélanges offrent un cadre probabiliste flexible et efficace pour traiter des problématiques de classification supervisée ou non supervisée. L'objectif du projet MIXMOD est de diffuser un ensemble logiciel de classification des données par modèles de mélanges à un large spectre d'utilisateurs via plusieurs composants logiciels. La bibliothèque de calcul mixmodLib (C++) en est la pierre angulaire, résultat d'un travail de 15 ans sur la robustesse et la rapidité de calcul. Le package RMixmod, ensemble de fonctions pour R, interfacé avec mixmodLib (grâce à RCPP) est devenu un outil de référence pour la classification des données. Intégrant de nombreuses fonctionnalités (algorithmes de type EM, critères de sélection, modèles parcimonieux, stratégies d'initialisation, ...), cet ensemble logiciel permet de traiter des données quantitatives, qualitatives et mixtes, y compris dans des situations complexes.
Lorsque le jeu de données contient des individus parasites (c'est-à-dire ayant des valeurs aberrantes, encore appelés outliers) la classification devient alors particulièrement difficile (trouver le bon nombre de classes, affecter le bon label aux vrais individus, ...).

Comment traiter un jeu de données avec des outliers ?
En présence d'outliers, il peut être tentant d'appliquer un pré-traitement pour *nettoyer* le jeu de données avant de le soumettre à un logiciel de classification. Mais ces méthodes sont généralement peu efficaces.
A l'opposé, on peut considérer que la classification doit s'effectuer sur l'ensemble des individus avec une classe supplémentaire (celle des outliers).

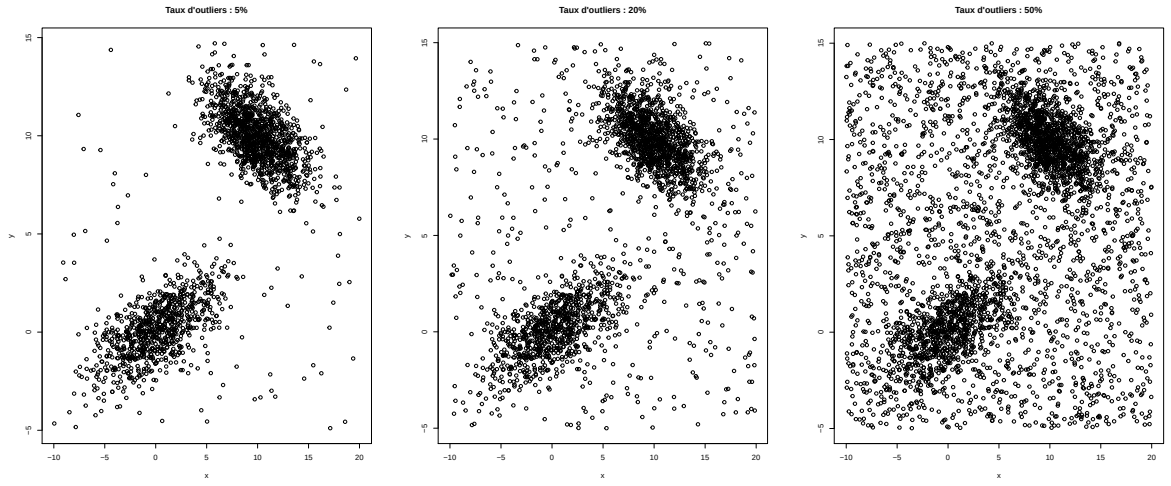
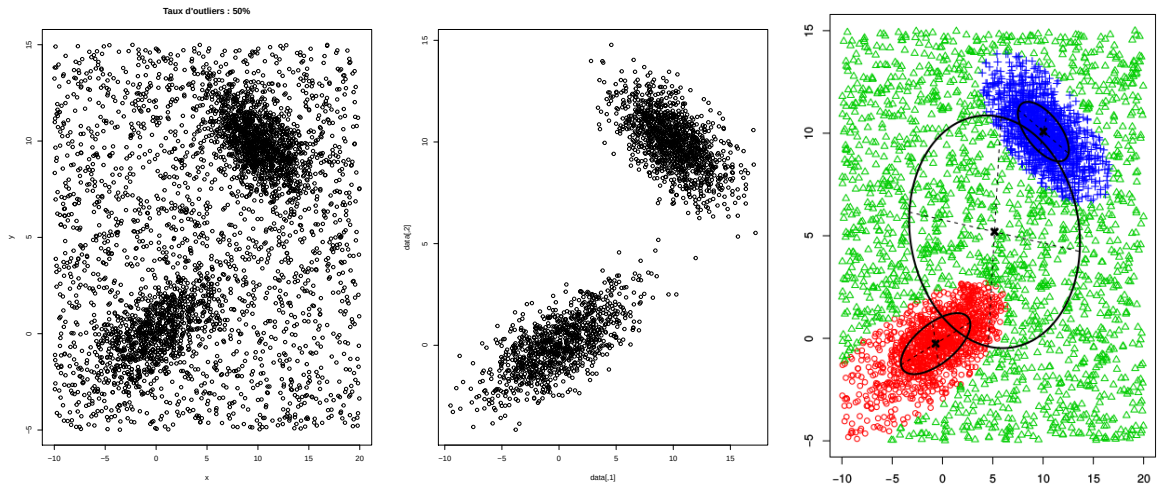


Figure 1: Jeux de données : 2 classes gaussiennes et des outliers répartis uniformément

Capacité de RMixmod à traiter les jeux de données avec des outliers
L'étude consiste à mettre à l'épreuve RMixmod sur ce type de problématique.
Bien évidemment, la difficulté est directement liée au pourcentage d'outliers et à leur répartition.

Lorsque la répartition des outliers suit une loi gaussienne, RMixmod la traite comme les autres classes sans difficulté. Mais que se passe t-il lorsque les outliers sont répartis selon d'autres lois ? Nous nous intéressons au cas où des outliers, répartis selon une loi uniforme, viennent s'ajouter à des individus issus de 2 lois gaussiennes (Figure 1).
La flexibilité des modèles de mélanges (ici gaussiens) permet non seulement de retrouver les classes d'origine (la classe rouge et la classe bleue de la figure 2c représentent bien les classes d'origine (figure 2b)) mais également de faire apparaître une classe contenant les outliers (individus en vert sur la figure 2c). Bien que de distribution uniforme, les outliers ont pu être modélisés efficacement par une gaussienne centrée et de variance suffisamment grande.

Jeu de données (avec outliers) Individus sans outlier Classification avec RMixmod



(a) (b) (c)

Figure 2: Classification en présence de 50% d'outliers

La quasi totalité des vrais individus est bien reclassée (voir tableau 1). En revanche, les outliers sont parfois considérés comme des vrais individus (dans 23% des cas). En effet, en simulant selon une loi uniforme, on peut obtenir des individus à l'intérieur des 2 vraies classes. Ce taux correspond d'ailleurs au rapport entre la surface des 2 vraies classes et la surface du support de la loi uniforme. Mais doit-on alors considérer ces individus comme des outliers ?

	Classe bleue	Classe rouge	Classe verte
Classe 1	97%	0%	3%
Classe 2	0%	98%	2%
Outliers	13%	10%	77%

Table 1: Reclassement des individus

Références

[1] Lebret R., Iovleff S., Langrognet F., C. Biernacki, G. Celeux, G. Govaert (2015). Rmixmod: The R Package of the Model-Based Unsupervised, Supervised and Semi-Supervised Classification Mixmod Library. *Journal of Statistical Software, 67(6), 1-29.*
[2] Biernacki C., Celeux G., Govaert G., Langrognet F., (2006). Model-Based Cluster and Discriminant Analysis with the MIXMOD Software. *Computational Statistics and Data Analysis, vol. 51/2, pp. 587-600.*
[3] Fraley C., Raftery A.E., (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97, 611-631.*