SAPIENZA
UNIVERSITÀ DI ROMA

Advances in Robust Clustering Methods with Applications

Scuola di Dottorato in Scienze Statistiche

Dottorato di Ricerca in Statistica Metodologica – XXIX Ciclo

Candidate

Francesco Dotto

ID number 1179428

Thesis Advisors

Prof. Alessio Farcomeni

Prof. Agustín Mayo-Iscar

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Statistics

October 2016

Thesis defended on 21 February 2017
in front of a Board of Examiners composed by:

Prof. Caterina Conigliani (chairman)

Prof. Maria Giovanna Ranalli

Dr. Pietro Coretto

Advances in Robust Clustering Methods with Applications
Ph.D. thesis. Sapienza – University of Rome

Author's email: francesco.dotto@uniroma1.it

To Emilio and Daniela...
my starting points

Contents

Contents iv

List of figures ix

List of tables xiii

Introduction 1

1 Robust Statistics: An overview 5

1.1 Contamination: some general notions 5

1.2 Contamination: models 6

1.2.1 Spurious outliers model 6

1.2.2 Tukey-Huber contaminated model 7

1.3 Multivariate Robust Statistics 7

1.3.1 Introduction 7

1.3.2 Multivariate outliers 8

1.3.3 MCD approach 9

1.3.4 Reweighted MCD approach 10

1.3.5 Alternative Approaches 11

1.4 Robust Statistics: some useful tools 12

1.4.1 The influence function and some related quantities 12

1.4.2 The breakdown point 13

2 Robust Clustering Methods 17

2.1 Introduction and state of art 17
2.1.1 Trimmed k-means 18
2.2 Heterogeneus robust clustering based on trimming 19
2.2.1 Formalization of the problem 19
2.2.2 A "naive" extension of the fast MCD algorithm 20
2.2.3 Spurious maximizers 21
2.2.4 Constraint based on the determinant 22
2.2.5 Hathaway-Dennis-Beale-Thompson constraints 23
2.3 The TCLUST methodology 24
2.3.1 Introduction 24
2.3.2 Mathematical formulation 24
2.3.3 The algorithm 25
2.3.4 Open Issues 26

3 Reweighting in Robust Clustering 29
3.1 Introduction 29
3.2 Methodology 30
3.3 The algorithm 33
3.4 Comments on the algorithm 34
3.5 Illustrative examples 35
3.6 Theoretical results 39
3.7 Simulation study 41

4 Extension of TCLUST to fuzzy linear clustering 49
4.1 Introduction 49
4.2 Methodology and algorithm 51
4.2.1 Defining the problem 51
4.2.2 Proposed algorithm 52
4.3 Interpretation and choice of the tuning parameters 55
4.3.1 Including clusters' weights 56

4.3.2 Number of clusters 56
4.3.3 Fuzzification Parameter 58
4.3.4 Constraints on the residual variances 60
4.3.5 Trimming level 63
4.3.6 Automatically choosing all parameters 63
4.4 Simulation study 65
4.4.1 Settings and methods 65
4.4.2 Automatic choice of the tuning parameters 73
4.4.3 Comments on the results of the simulation study 74

5 Real data examples **79**
5.1 Introduction 79
5.2 Applications of reweighted TCLUST 80
5.2.1 Swiss Bank Notes 80
5.2.2 Food Security Data 82
5.3 Applications of fuzzy linear clustering 85

6 Conclusions and further directions **89**
6.1 Concluding remarks on the reweighted TCLUST contribution 89
6.2 Concluding remarks on the TCLUST extension to fuzzy linear clustering models 90
6.3 Overall conclusions and further direction of research 91
6.3.1 Preliminary simulation results 93

Appendix A **97**
A.1 Proofs of the theoretical properties of the RTCLUST methodology . . 97
A.1.1 Introduction and notation 97
A.1.2 Proof of Theorem 1 98
A.1.3 Proof of Theorem 2 99
A.1.4 Proof of Theorem 3 99
A.1.5 Proof of Theorem 4 101

A.2 Justification of Algorithm 6 102

A.3 A proposal for standardizing the residual component in Algorithm 6 . 104

Bibliography **118**

List of Figures

2.1 Comparison between constrained and unconstrained clustering. 22

3.1 Two simulated data sets with their true assignments in (a.1) and (b.1). The result of TCLUST with $\alpha_0 = 0.33$ in (a.2) and (b.2). The final assignments obtained after applying the proposed methodology are given in (a.3) and (b.3). Noisy data and trimmed are denoted by $\circ$ in all graphs throughout the manuscript. 32

3.2 Evolution of $|\Sigma_j^l|$ in (a) and of π_j^l in (b) for different initial α_0 values ($\alpha_0 = .3$, .25, .2 and .15) for the data set shown in Figure 3.1 (b.1). The up-triangle symbols are the true parameters to be estimated. . . 36

3.3 Evolution of $|\Sigma_j^l|$ in (a) and of π_j^l in (b) for different initial c values ($c = 1$, 10 and 20 while the true c needed was 11.71) for the data set shown in Figure 3.1 (b.1). The up-triangle symbols are the true parameters to be estimated. 37

3.4 (a) The proposed iterative reweighting procedure when $k = 1$ started from $\alpha_0 = 0$ and $\alpha_L = 0.01$ (b) The (traditional) reweighted MCD started from $\alpha_0 = 0$ and $\alpha_L = 0.01$. Trimmed points are the black points. 38

3.5 Results when $\varepsilon = 0.05$. Every procedure is labeled as explained in the text. Values appearing in the Figure that are fixed in advance (e.g the trimming level for the `tclust` method) are plotted with the symbol "$\times$" while when the considered value exceeds the scale of the plot we used a "$\triangle$" 45

3.6 Results when $\varepsilon = 0.10$. Every procedure is labeled as explained in the text. 46

3.7 Simulation results study under no contamination ($\varepsilon = 0$). 47

4.1 (a) Robust fuzzy clustering results when $k = 3$ and p_j are used within the objective function. (b) Results when $k = 3$ and p_j are not used within the objective function. 57

4.2 (a) A simulated dataset with two overlapped linear clusters and 10% of contaminated points. (b) The associated "classification trimmed likelihood curves" when $c = 5$ and $m = 1.5$. 58

4.3 Different degrees of fuzzification obtained for different scale values s (y_i is replaced by $y_i \cdot s$). $m = 1.5$ and $s = 0.5$ in (a); $s = 1$ in (c); $s = 10$ in (e); $s = 32$ in (g). $m = 1$ (hard clustering) and $s = 0.5$ in (b); $s = 1$ in (d); $s = 10$ in (f); $s = 32$ in (h). 59

4.4 Left panels: relative entropy of the fuzzy weights, "×", proportion of hard assignments, "○", as a function of scale; (a) $s = 0.5$. (c) $s = 1$ (e) $s = 10$. (g) $s = 32$. Right panels: clustering obtained for specific values of m through (b) $s = 0.5$, $m = 2.2$. (d) $s = 1$, $m = 1.8$. (f) $s = 10$, $m = 1.6$. (h) $s = 32$, $m = 1.4$. 61

4.5 Estimated robust fuzzy clustering for different c values in two (less and more) heteroscedastic data sets. $c = 5$ is used in (a) and (d) and $c = 1$ in (b) and (d). The plotted bands are obtained by adding $\pm 2\widehat{s}_j$ to each fitted regression line. 62

4.6 (a) FTCR with $c = 5$ and $k = 2$. (b) FTCR with $c = 10^{10}$ and $k = 2$. (c) FTCR with $c = 10^{10}$ and $k = 3$. 62

4.7 Left panels: Estimated linear clustering result for different trimming levels and $m = 1.5$. (a) $\alpha = 0.20$. (c) $\alpha = 0.10$. (e) $\alpha = 0.05$. Right panels: Average contribution to the likelihood for different values of α. A red line corresponds to the trimming level used on the corresponding left panel. (b): $\alpha = 0.20$. (d): $\alpha = 0.10$. (f): $\alpha = 0.05$. 64

4.8 (a) The scatter plot of our dataset. (b) The results obtained using the tuning parameters chosen by cross-validation 65

4.9 Simulation study. Boxplots representing the MSE of $\boldsymbol{b}_j$ and b_j^0 for setting S1: $p = 1$, $k = 2$. The Homoscedastic clusters are in (a),(c),(e),(g). Heteroscedastic clusters are in (b), (d), (f), (h). Uniform contamination is in (a) and (b). Inflated uniform contamination is in (c), and(d). Pointwise contamination in (e) and (f). Clean dataset is in (g) and (h) 67

4.10 Simulation study. Misclassification error for setting S1: $p = 1$, $k = 2$. Legend as in Figure 4.9. 68

4.11 Simulation study. Boxplots representing the MSE of $\boldsymbol{b}_j$ and b_j^0 for setting S2: $p = 2$, $k = 2$. Same legend of Figure 4.9. 69

4.12 Simulation study. Misclassification error for setting S2: $p = 2$, $k = 2$. Legend as in Figure 4.9 70

4.13 Simulation study. Boxplots representing the MSE of $\boldsymbol{b}_j$ and b_j^0 for setting S3: $p = 4$, $k = 2$. Same legend of Figure 4.9 71

4.14 Simulation study. Misclassification error for setting S2: $p = 2$, $k = 2$. Legend as in Figure 4.9 72

4.15 Simulation study. Boxplots representing the MSE of $\boldsymbol{b}_j$ and b_j^0 for setting S4: $p = 1$, $k = 3$. Legend as in Figure 4.9 73

4.16 Simulation study. Misclassification error for setting S4: $p = 1$, $k = 3$. Legend as in Figure 4.9 74

4.17 Simulation study. Boxplots representing the MSE of $\boldsymbol{b}_j$ and b_j^0 for setting S5: $p = 2$, $k = 3$. Legend as in Figure 4.9 75

4.18 Simulation study. Misclassification error for setting S5: $p = 2$, $k = 3$. Legend as in Figure 4.9 75

4.19 Simulation study. Boxplots representing the MSE of $\boldsymbol{b}_j$ and b_j^0 for setting S6: $p = 4$, $k = 3$. Same legend of Figure 4.9 76

4.20 Simulation study. Misclassification error for setting S6: $p = 4$, $k = 3$. Legend as in Figure 4.9. 76

4.21 Boxplots with Mean Square Error for tuned and crossvalidated model, with competitors for comparison. C-val denotes FTCR with automatically chosen tuning. (a) Two Homoscedastic clusters uniformly contaminated, $p = 1$ covariate (Setting S1). (b) Two Heteroscedastic clusters with pointwise contamination, $p = 1$ covariate (Setting S1). . 77

5.1 Fourth against the sixth variable of the Swiss Bank Notes data set. (a) `G` stands for genuine bills, `F` for forged ones and 15 bills listed in Flury & Riedwyl (1988) as anomalous ones are surrounded by $\circ$ symbols. (b) The initial TCLUST solution with $\alpha_0 = 0.33$ (c) Final solution when applying the proposed iterative approach. Trimmed observations not coinciding with those in Flury and Riedwyl's list are surrounded by $\square$ symbols. 81

5.2 `ctlcurve` plot for the FIES data. 83

5.3 *Pinus Nigra example.* (a) *ctlcurve.* (b) average contribution to the likelihood as a function of α. (c) relative empty entropy and proportion of hard assignments as a function of m. 86

5.4 *Pinus Nigra example:* (a) Scatter plot and results of cReg method. (b) results of the "EM" method. (c) results of the A-cReg method. (d) results of the FTCR method. Circled observations are fuzzy assignments. 87

5.5 *Pinus Nigra Data example:* Results of the proposed procedure when searching for $k = 4$ clusters and no trimming imposed 88

A.1 Empirical Comparison of the impact of the scale of the data as before and after the standardization (A.7) 105

List of Tables

5.1 99% simultaneous confidence intervals for $\hat{\mu}$ provided by the TCLUST and the RTCLUST 81

5.2 Cluster profiles and measurements for the outlying countries. FIES: Food Insecurity Experience Scale. CE: Civic Engagement. St: Struggling. FS: Food Security. Co: Corruption index. YD: Youth Development. C-j: j-th cluster profile. 84

6.1 List of 10 different models obtained by imposing different constraint in decomposition (6.1) 92

6.2 Simulation results based on B=500 replicates: average MSE of the estimated vector mean in each cluster 95

Introduction

Robust methods in statistics are mainly concerned with deviations from model assumptions. As already pointed out in Huber (1981) and in Huber & Ronchetti (2009) "these assumptions are not exactly true since they are just a mathematically convenient rationalization of an often fuzzy knowledge or belief". For that reason "a minor error in the mathematical model should cause only a small error in the final conclusions". Nevertheless it is well known that many classical statistical procedures are "excessively sensitive to seemingly minor deviations from the assumptions".
All statistical methods based on the minimization of the average square loss may suffer of lack of robustness. Illustrative examples of how outliers' influence may completely alter the final results in regression analysis and linear model context are provided in Atkinson & Riani (2012). A presentation of classical multivariate tools' robust counterparts is provided in Farcomeni & Greco (2015).
The whole dissertation is focused on robust clustering models and the outline of the thesis is as follows.
Chapter 1 is focused on robust methods. Robust methods are aimed at increasing the efficiency when *contamination* appears in the sample. Thus a general definition of such (quite general) concept is required. To do so we give a brief account of some kinds of contamination we can encounter in real data applications. Secondly we introduce the "Spurious outliers model" (Gallegos & Ritter 2009*a*) which is the cornerstone of the robust model based clustering models. Such model is aimed at formalizing clustering problems when one has to deal with contaminated samples. The assumption standing behind the "Spurious outliers model" is that two different random mechanisms generate the data: one is assumed to generate the "clean" part while the another one generates the *contamination*. This idea is actually very common within robust models like the "Tukey-Huber model" which is introduced in Subsection 1.2.2. Outliers' recognition, especially in the multivariate case, plays a key role and is not straightforward as the dimensionality of the data increases. An overview of the most widely used (robust) methods for outliers detection is provided within Section 1.3. Finally, in Section 1.4, we provide a non technical review of the classical tools introduced in the Robust Statistics' literature aimed at evaluating the

robustness properties of a methodology.
Chapter 2 is focused on model based clustering methods and their robustness' properties.
Cluster analysis, "the art of finding groups in the data" (Kaufman & Rousseeuw 1990), is one of the most widely used tools within the unsupervised learning context. A very popular method is the k-means algorithm (MacQueen et al. 1967) which is based on minimizing the Euclidean distance of each observation from the estimated clusters' centroids and therefore it is affected by lack of robustness. Indeed even a single outlying observation may completely alter centroids' estimation and simultaneously provoke a bias in the standard errors' estimation. Cluster's contours may be inflated and the "real" underlying clusterwise structure might be completely hidden. A first attempt of robustifying the k- means algorithm appeared in Cuesta-Albertos et al. (1997), where a trimming step is inserted in the algorithm in order to avoid the outliers' exceeding influence.
It shall be noticed that k-means algorithm is efficient for detecting spherical homoscedastic clusters. Whenever more flexible shapes are desired the procedure becomes inefficient. In order to overcome this problem Gaussian model based clustering methods should be adopted instead of k-means algorithm. An example, among the other proposals described in Chapter 2, is the TCLUST methodology (García-Escudero et al. 2008), which is the cornerstone of the thesis. Such methodology is based on two main characteristics: trimming a *fixed* proportion of observations and imposing a constraint on the estimates of the scatter matrices. As it will be explained in Chapter 2, trimming is used to protect the results from outliers' influence while the constraint is involved as spurious maximizers may completely spoil the solution.
Chapter 3 and 4 are mainly focused on extending the TCLUST methodology.
In particular, in Chapter 3, we introduce a new contribution (compare Dotto et al. 2015 and Dotto et al. 2016*b*), based on the TCLUST approach, called reweighted TCLUST or RTCLUST for the sake of brevity. The idea standing behind such method is based on reweighting the observations initially flagged as outlying. This is helpful both to gain efficiency in the parameters' estimation process and to provide a reliable estimation of the true contamination level. Indeed, as the TCLUST is based on trimming a fixed proportion of observations, a proper choice of the trimming level is required. Such choice, especially in the applications, can be cumbersome. As it will be clarified later on, RTCLUST methodology allows the user to overcome such problem. Indeed, in the RTCLUST approach the user is only required to impose a high preventive trimming level. The procedure, by iterating through a sequence of decreasing trimming levels, is aimed at reinserting the discarded observations at each step and provides more precise estimation of the parameters and a

final estimation of the true contamination level $\hat{\alpha}$.
The theoretical properties of the methodology are studied in Section 3.6 and proved in Appendix A.1, while, Section 3.7, contains a simulation study aimed at evaluating the properties of the methodology and the advantages with respect to some other robust (reweigthed and single step procedures).
Chapter 4 contains an extension of the TCLUST method for fuzzy linear clustering (Dotto et al. 2016*a*). Such contribution can be viewed as the extension of Fritz et al. (2013*a*) for linear clustering problems, or, equivalently, as the extension of García-Escudero, Gordaliza, Mayo-Iscar & San Martín (2010) to the fuzzy clustering framework. Fuzzy clustering is also useful to deal with contamination. Fuzziness is introduced to deal with overlapping between clusters and the presence of *bridge points*, to be defined in Section 1.1. Indeed *bridge points* may arise in case of overlapping between clusters and may completely alter the estimated cluster's parameters (i.e. the coefficients of a linear model in each cluster). By introducing fuzziness such observations are suitably down weighted and the clusterwise structure can be correctly detected. On the other hand, robustness against *gross outliers*, as in the TCLUST methodology, is guaranteed by trimming a fixed proportion of observations. Additionally a simulation study, aimed at comparing the proposed methodology with other proposals (both robust and non robust) is also provided in Section 4.4.
Chapter 5 is entirely dedicated to real data applications of the proposed contributions. In particular, the RTCLUST method is applied to two different datasets. The first one is the "Swiss Bank Note" dataset, a well known benchmark dataset for clustering models, and to a dataset collected by Gallup Organization, which is, to our knowledge, an original dataset, on which no other existing proposals have been applied yet. Section 5.3 contains an application of our fuzzy linear clustering proposal to allometry data. In our opinion such dataset, already considered in the robust linear clustering proposal appeared in García-Escudero, Gordaliza, Mayo-Iscar & San Martín (2010), is particularly useful to show the advantages of our proposed methodology. Indeed allometric quantities are often linked by a linear relationship but, at the same time, there may be overlap between different groups and outliers may often appear due to errors in data registration.
Finally Chapter 6 contains the concluding remarks and the further directions of research. In particular we wish to mention an ongoing work (Dotto & Farcomeni, *In preparation*) in which we consider the possibility of implementing robust parsimonious Gaussian clustering models. Within the chapter, the algorithm is briefly described and some illustrative examples are also provided. The potential advantages of such proposals are the following. First of all, by considering the parsimonious models introduced in Celeux & Govaert (1995), the user is able to impose the

shape of the detected clusters, which often, in the applications, plays a key role. Secondly, by constraining the shape of the detected clusters, the constraint on the eigenvalue ratio can be avoided. This leads to the removal of a tuning parameter of the procedure and, at the same time, allows the user to obtain affine equivariant estimators. Finally, since the possibility of trimming a fixed proportion of observations is allowed, then the procedure is also formally robust.

Chapter 1

Robust Statistics: An overview

1.1 Contamination: some general notions

As briefly stated in the introduction, robust methods aim to provide methodologies which are resistant with respect to mild deviations from the assumed parametric model. This implies that in the observed sample there are points which do not follow the underlying distribution, that is to say, *contaminating points* or *outliers*. Contamination is a very general notion that may be defined in different ways depending on the context of application. Following Farcomeni & Greco (2015) we try here to give a non-exhaustive account of some types of contamination that are likely to be found in data analysis:

- *Extreme values* or *gross outliers*. Points unusually large (or small) with respect to one or more dimensions
- *Influential outliers* or *leverage points*. Points that do not follow the pattern shown by the majority of the data (i.e. points presenting negative correlation between two dimensions when data exhibit positive one)
- *Inliers*. Corrupted points that lie very close to the sample mean deflating the variance.
- *Bridge points*. Points lying very close to the boundaries of two clusters. These points play a key role in cluster analysis. Indeed bridge points may be very difficult to assign to one cluster and can be dangerous for parameters' estimation.

Generally speaking robust procedures are designed in order to be efficient in cases where *contamination* appears in the sample. As it will be clarified in the further

chapters, *impartial trimming* (García-Escudero et al. 2008) is a useful tool in order to deal with contaminating points.
It must be pointed out that trimming is supposed to discard the "farthest" values from the clusters' centers. In case of overlap between clusters, *bridge points* may appear in the sample and trimming may not work well. Thus a different framework will be considered in Chapter 4. In particular we will consider the case of *fuzzy* partitions, instead of *hard* partitions. Considering hard partitions in clustering is equivalent to assign a binary weight to the i-th observation $u_{ij} \in \{0, 1\}$ and $u_{ij} = 1$ if and only if observation i belongs to cluster j. On the other hand, in case of fuzzy partitions $u_{ij} \in [0, 1]$. Thus, within the fuzzy framework, each observation is simultaneously assigned to more than one cluster and the degree of membership of observation i to each cluster j is given by its fuzzy weight u_{ij}.

1.2 Contamination: models

Robust methods aim to contain the exceeding influence of the outlying points. To do so, the sample is generally supposed to come from two different probability density functions: one generating the "clean part" of the data, and the other one, generally called contaminating density, generating the contaminated part of the sample. Indeed, especially within the multivariate context, identifying the contaminated part of the data is necessary to be able to contain its influence by treating it in a different way (e.g by trimming or downweighting).

1.2.1 Spurious outliers model

One of the most widely used models suitable for cluster analysis is the "spurious outliers model", introduced in Gallegos & Ritter (2005). Such model keeps in account the presence of two different densities: one is a mixture made of k components, where each component of the mixture generates each cluster, and the other is a contaminating density, which generated the outlying component of the data. A more detailed definition follows.

Definition 1. Let $x_i \in \mathbb{R}^p$ be a sample point, $f(\cdot)$ the multivariate normal density, μ_j and Σ_j be location and scatter parameters, respectively, of the j-th group. Additionally let $g_{\psi_i}(\cdot)$ be the contaminating density and K the number of groups. Then the likelihood function associated to the spurious outliers model is given by:

$$\left[\prod_{j=1}^{K}\prod_{i\in R_j} f(x_i;\mu_j;\Sigma_j)\right]\left[\prod_{i\notin R_j} g_{\psi_i}(x_i)\right] \tag{1.1}$$

Additionally it must be pointed out that, in equation (1.1), $R = \bigcup_{j=1}^{K} R_j$ represents the set of the clean observation and is such that $\#R = \lceil n(1-\alpha) \rceil$. As it will be clarified in the further sections, only the clean data give a contribution to the likelihood function, while, outliers give no contribution to function (1.1).

1.2.2 Tukey-Huber contaminated model

Spurious outliers model is an adaptation of the Tukey-Huber contamination1 model (Tukey 1962 and Huber et al. 1964), which is defined as follows.

Definition 2. Let F be the model generating the data, generally assumed to be Gaussian throughout the whole dissertation, G the contaminating model and ε the proportion of observation arising from the contaminating model. The ε-neighborhood or Tukey-Huber contaminated model is defined as:

$$\mathcal{P}(F, \varepsilon) = \{F_\varepsilon | F_\varepsilon = (1-\varepsilon)F(X;\theta) + \varepsilon G(X), \theta \in \Theta, X \in \mathcal{X}\} \tag{1.2}$$

Generally speaking the more a statistics (output of a procedure) $T(F)$ is resistant to contamination, the more is considered as robust.

Definitions 1 and 2 provide a very general formalization of the concept of contaminated models. Throughout this dissertation, whenever we refer to contaminated data, we implicitly refer to a data generating mechanism outlined either in Definition 1 or in Definition 2.

1.3 Multivariate Robust Statistics

1.3.1 Introduction

Given a multivariate sample $X = (x_1, \ldots, x_n)$ with $x_i \in \mathbb{R}^p$ and $p \geq 1$, the sample mean vector, $\hat{\mu}$, and the sample covariance matrix, $\hat{\Sigma}$, are standard tools for describing location, variability and pairwise dependence in the data. Usage of such quantities is also motivated by the fact that these are the MLE estimators of the location and scale parameters at the multivariate Gaussian model. It shall be noticed that the multinormal distributional assumption of the data is pretty common although it may be too restrictive in some cases.
As in the univariate case, such quantities suffer of lack of robustness since even one single observation may completely spoil the yielding estimates. Illustrative examples may be encountered, among the others, in García-Escudero et al. (2012). Thus

the influence of outlying points needs to be controlled although, as the dimensionality of the data increases, outliers' identification becomes an hard task. Indeed, as visualization's tools can not be applied, alternative methods are required.

1.3.2 Multivariate outliers

Let us suppose, that $x_i \sim \mathcal{N}(\mu, \Sigma)$ where $x_i \in \mathbb{R}^p$ and $\mathcal{N}(\mu, \Sigma)$ stands for the multivariate normal distribution with location parameter μ and scale parameter Σ. Generally speaking outliers are observation placed "far" from the bulk of the data. For that reason suitable methods for defining the distance from the bulk of the data are required. Clearly the usage of the simple Euclidean distance from a suitably defined center of the data is not enough since such value is affected by the scale. The most commonly used distance measure in multivariate statistics is the Mahalanobis distance:

$$d_\Sigma(x_i, \mu) = \sqrt{(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)} \tag{1.3}$$

where in equation (1.3) $x_i \in \mathbb{R}^p$ is a sample point, $\mu \in \mathbb{R}^p$ is a location parameter and $\Sigma \in PD(\mathcal{M}^{pxp})$ is a scale parameter. It easy to note that if $\Sigma = I_{pxp}$ then the Mahalanobis distance is equivalent to the Euclidean distance while, on the other hand, as $p = 1$ the Mahalanobis distance reduces to the well known z-score. Mahalanobis distance evaluation plays a key role in multivariate outliers' detection: every observation exceeding a pre-fixed value of the Mahalanobis distance may be flagged as outlying. In order to fix this value one may refer to the asymptotic distribution of the Mahalanobis distance which can be approximated:

$$d^2_\Sigma(x_i; \mu) \sim \chi^2_p \tag{1.4}$$

Clearly formula (1.4) provides a very "naive" approximation and the true parameters of the distribution are supposed to be known. In order to improve (1.4), Gnanadesikan & Kettenring (1972) proposed an exact distribution for the Mahalanobis distance as the MLE is plugged in instead of the true parameters' values:

$$d^2_{\bar{S}}(x_i; \bar{x}) \sim \frac{(n-1)^2}{n} Beta\left(\frac{p}{2}, \frac{n-p-1}{2}\right) \tag{1.5}$$

Typical choices for cut off values are .975 or .99 that correspond to flagging as outlying observation out of the boundaries of the 97.5% or 99% of the tolerance ellipsoid. Clearly neither approximation (1.4) nor approximation (1.5) are efficient under contamination. Indeed even one single outlier may completely alter the estimation of $\bar{x}$ and, simultaneously, inflate, or deflate as well, $|\hat{\Sigma}|$. Possible consequences are

swamping or *masking* effects. *Swamping* occurs when clean observations are flagged as outlying. Such undesired effect may be caused by a deflation of $|\hat{\Sigma}|$ which may lead to wrongly consider many observations "too far" from the center of the data and thus, flagging them as outlying. On the contrary, as $|\hat{\Sigma}|$ is overestimated, outliers may not be recognized and then *masking* occurs.
For these reasons, some robust counterparts of the classical methods are required.

1.3.3 MCD approach

In order to protect the estimators from the influence of the "farthest" points, Rousseeuw (1985) proposed to estimate the parameters using a subset containing only the bulk of the data. The bulk of the data can be recognized as the subsample containing the $n(1-\alpha)$ data points which yield the covariance matrix having the minimum determinant. Once the subset containing the data with the minimum determinant of the covariance matrix is identified, then the population mean is estimated straightforwardly by using the sample mean of these points; while the covariance matrix is estimated by multiplying the sample variance of these points for a constant which guarantees the consistency of the estimator. More formally, let α be the fixed proportion discarded and let z_i be a binary vector such that $\sum z_i = n \cdot (1-\alpha)$. The MCD estimators are defined as:

$$\hat{\mu}_{MCD} = \frac{1}{\sum_i z_i} \sum_{i=1}^{n} z_i x_i \tag{1.6}$$

$$\hat{\Sigma}_{MCD} = \frac{c(p,\alpha)}{\sum_i z_i - 1} \sum_{i=1}^{n} (x_i - \hat{\mu}_{MCD})(x_i - \hat{\mu}_{MCD})^T z_i \tag{1.7}$$

where the constant term in equation (1.7) is a factor which makes the MCD consistent at the Normal model by inflating the estimated covariance matrix. Its explicit formula is the following:

$$c(p,\alpha) = \frac{1-\alpha}{F_{\chi^2_{p+2}}(q_{p,1-\alpha})} \tag{1.8}$$

More theoretical details on this argument can be found in Liu et al. (1999).
The most popular algorithm for the MCD is the FASTMCD algorithm, proposed in Rousseeuw & van Driessen (1999), which iterates the following steps:

Algorithm 1.

1. Let $\hat{\theta}_0 = (\hat{\mu}_0, \hat{\Sigma}_0)$ an initial estimate of the parameters obtained by sampling randomly a subset of observations having size $n \cdot (1-\alpha)$

2. Calculate the robust distances $d_{0i} = d(x_i, \hat{\theta}_0)$
3. Sort the distances in non increasing order and take the subset of size $n \cdot (1-\alpha)$ having the lowest values
4. Update the estimate of the parameter $\hat{\theta}_1 = (\hat{\mu}_1, \hat{\Sigma}_1)$
5. Iterate steps 2-4 up to convergence

It is straightforward to see that at each iteration of the algorithm the determinant of the estimated scatter matrix decreases, since, at each step, the observations "closest" to each other are inserted in the subsample. By initializing the algorithm from different starting points the global optimum for the objective function (the determinant of the scatter matrix) is more likely to be reached. Usually the algorithm is implemented by initializing it 500 times and typical values for α are either 0.25 or 0.50.
Despite computational issues, the MCD algorithm is one of the most widely used approaches to provide robust estimates in a multivariate context; additionally there are interesting asymptotic properties (Butler et al. 1993, in Croux & Haesbroeck 1999 and Cator et al. 2012). From the robustness' point of view it can be shown that the asymptotic breakdown point (to be better defined in the further sections) is often equal to the chosen trimming rate. As it will be clarified later on, the maximum value for the breakdown point that can be reached by an estimator is 0.5 which implies that, if $\alpha = 0.5$, then the MCD estimator is the affine equivariant estimator having the highest possible value for the asymptotic breakdown point.

1.3.4 Reweighted MCD approach

Robustness may cause loss of efficiency since part of the observations are generally discarded. Indeed as α is fixed too high, then too many observations are discarded, provoking a loss of efficiency in the parameter estimation. For that reason, the MCD estimator may be reweighted to increase of efficiency. This leads to a new estimator: the reweighted MCD, usually called RMCD. The reweighting process works as follows:

Algorithm 2.

1. For each $i = 1, \ldots, n$ compute the distances $d_i = d_{\hat{\Sigma}_{MCD}}(x_i, \hat{\mu}_{MCD})$ from the MCD estimators defined in (1.6) and (1.7)
2. Set $z_i = 1$ if the value of d_i is below a fixed threshold and $z_i = 0$ otherwise.

3. Update the estimation using formulas (1.6) and (1.7) to update the estimates.

A common choice for fixing the threshold to be used in the step 1 of Algorithm 2 is the α' quantile of the χ^2_p distribution. Alternatively a better approximation is provided in Hardin & Rocke (2004) where a scaled F distribution is proposed:

$$d^2_{\hat{\Sigma}_{MCD}}(x_i; \hat{\mu}_{MCD}) \sim \frac{pm}{(m-p-1)} F_{p,m-p+1} \tag{1.9}$$

In equation (1.9) m is a constant whose expression can be found in Hardin & Rocke (2005). It shall be pointed out that approximation (1.9) is optimal for MCD estimators, while an suitable approximation for RMCD estimators, provided in Cerioli (2010), is given by:

$$d^2_{\hat{\Sigma}_{RMCD}}(x_i; \hat{\mu}_{RMCD}) \sim \frac{(\sum z_i - 1)^2}{\sum z_i} Beta\left(\frac{p}{2}, \frac{\sum z_i - p - 1}{2}\right) \tag{1.10}$$

1.3.5 Alternative Approaches

Another popular robust estimator has been proposed in Rousseeuw (1984) and in Rousseeuw (1985) where the Minimum Volume Ellipsoid (MVE) estimator has been introduced. Operatively speaking, the MVE estimator looks for the ellipsoid of the minimum volume that contains $n(1-\alpha)$ observations. It shall be noticed that the idea of this estimator is pretty similar to the idea standing behind the MCD estimators. An algorithmic method for computing the minimum volume ellipsoid has been proposed Van Aelst & Rousseeuw (2009). Nevertheless, due to the efficiency of the fast MCD algorithm (Algorithm 1), this last estimator has become much more popular than the MVE estimator so far. MCD and MVE are based on hard rejection rules, following the "impartial trimming principle", explained in García-Escudero et al. (2008) and in Cuesta-Albertos et al. (2008*a*). Another method based on trimming is the forward search approach, Atkinson et al. (2004) and Atkinson et al. (2004), which is based on the idea of starting from a subset of clean observations and iteratively looking for the best sets of increasing size based on the estimates at the previous steps.
Alternative robust approaches are mainly based on underweighting outlying observations instead of trimming them. Among the others, we recall methodologies in Donoho (1982) and Stahel (1981), where the idea is to assign a weight to each observation depending on its "outlyingness" measured by using a univariate projection of each observation. To our knowledge, an application of these methodologies for clustering has not been proposed yet.

1.4 Robust Statistics: some useful tools

A brief review of how to evaluate the robustness of a procedure follows. It must be pointed out that some tools require technical arguments and a direct usage within the robust clustering context is not straightforward. These concepts will be briefly mentioned within this chapter and recalled, as required, along the whole thesis.

1.4.1 The influence function and some related quantities

Influence function: definition

In order to describe the effect of the departure from the assumed model F within a neighborhood Hampel (1974) and Huber (1981) introduced the idea of *influence function*. From a mathematical point of view the influence function is defined as the Gateaux derivative of the functional $T(F)$, with $F \in \mathcal{F}$ along the direction of x. More precisely:

Definition 3. Let $F_\varepsilon = (1-\varepsilon)F + \varepsilon\delta_x$ where δ_x is Dirac delta random variable degenerate in x. The influence function for an infinitesimal point mass contamination ε, at location x, at the model F, is given by:

$$IF(x;T,F) = \lim_{\varepsilon \to 0} \frac{T(F_\varepsilon) - T(F)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} T(F_\varepsilon)|_{\varepsilon=0} \tag{1.11}$$

The influence function provides a global overview of the robustness properties of an estimator as Gateaux derivative's computation at location x aims at measuring the effect that a contaminating point x may have on an estimator $T(F)$. Whenever such effect is bounded we its yields that its influence function is bounded as well and the estimator can be considered robust. As an illustrative example let us consider the univariate standard Gaussian model. Let us also consider the sample mean and the sample median as estimators of the location parameter. Their influence function is given, respectively, by:

$$IF(x;\hat{\mu},\Phi) = x \tag{1.12}$$

$$IF(x;Med,\Phi) = \sqrt{\frac{\pi}{2}} sign(x) \tag{1.13}$$

It is clear from equations (1.12) and (1.13) that the influence function associated to the sample mean is unbounded while it becomes bounded in the case of the median. This reflects the well known properties of such estimators. Indeed the influence that a single point may have on the sample mean estimator is unbounded, while, in the case of the median, the estimator moves in the direction of the outlier in a bounded way.

The gross error sensitivity

As briefly stated in the previous paragraph, evaluation of the boundedness of the influence function of an estimator is pretty important in order to assess its robustness properties. Mathematically speaking, the boundedness of the influence function is evaluated by computing the *gross error sensitivity*, defined as the upper bound of the influence function:

$$\gamma^* = \sup_{x \in \mathbb{R}^p} ||IF(x; T, F)|| \tag{1.14}$$

Equation (1.14) allows to measure the highest influence that a fixed size contamination may have on the value of an estimator. A bounded value of γ^* implies that an estimator is influenced in a "bounded" way by any type of contamination. It is indeed a very important property from the robustness point of view and, as a result, estimators having bounded values of γ^* are formally robust. They are usually referred to B (bias) - robust estimators.

Local shift sensitivity

Data are often slightly changed (due to inaccuracies in data registration or to operations like rounding), and in terms of robustness it is interesting to measure the effect that such changes may have on the chosen estimator. We recall here the definition of *local shift sensitivity* given by:

$$\lambda^* = \sup_{x \neq y; x, y, \in \mathbb{R}^p} \frac{||IF(x; T, F) - IF(y; T, F)||}{||x - y||} \tag{1.15}$$

Equation (1.15) describes the effect of shifting an observation x to a close point y. It is straightforward to see that as $y = x + \varepsilon$ with $\varepsilon \to 0$ we go back to equation (1.11).

1.4.2 The breakdown point

Introduction

The notion of breakdown point is related with the proportion of observations that can be arbitrarily replaced until an estimator (an output of a procedure) breaks down. Roughly speaking, the higher is the breakdown point, the higher is the robustness of the given procedure. Formal definition of such concept depends strictly on the application of interest. There follow some very general definitions of breakdown point and the formalization proposed in Gallegos & Ritter (2009*a*) which is suitable to evaluate the robustness of a clustering method.

Finite sample breakdown point and its generalizations

The finite sample breakdown point, also known as *individual breakdown point* (Ruwet et al. 2013), provides a data dependent definition of the concept of the breakdown point associated to a given dataset.

Definition 4. Let $X_r \in \mathcal{X}_r$ where $\mathcal{X}_r$ is the collection of all datasets X_r of size having $(n-r)$ elements in common with the original data X_n. The *finite sample breakdown point* is defined as:

$$\varepsilon^i = \max\left\{\frac{r}{n} : \sup_{\mathcal{X}_r} ||T(X) - T(X_r)|| \in K\right\} \tag{1.16}$$

where a K is a bounded and closed set that does not contain the boundary points of the parameter space.

In order to overcome the dependency from the data Donoho & Huber (1983) introduced the notion of *universal breakdown point*.

Definition 5. Let $\mathcal{D}$ be the set of all datasets $X_n \in \mathbb{R}^p$ in general position. The universal breakdown point is defined as:

$$\varepsilon^{(u)} = \max_{X_n \in \mathcal{D}} \varepsilon^{(i)} \tag{1.17}$$

It shall be noticed that Definition 5 generalizes Definition 4 since the class containing all the dataset $X_n \in \mathbb{R}^p$ is considered in computing the breakdown point instead of considering a single dataset X_n.

Restricted breakdown point

As noticed in Ruwet et al. (2013), according to Definitions 4 and 5, a set of estimators obtained by a clustering model may have 0 value for the breakdown point despite their robustness. Indeed "some datasets can hardly be clustered in k clusters simply because the do not come from a k-component model and this makes any clustering method have a 0 value for the universal breakdown point" (Ruwet et al. 2013). For that reason, Gallegos & Ritter (2005) introduced the notion of *restricted breakdown point* with respect to some subclass $\mathcal{K} \subset \mathcal{D}$ of *admissible datasets*. In particular, within cluster analysis, the condition of "well clustered" datasets, proposed in the reference above, is kept in account. The *restricted breakdown point* is defined as follows.

Definition 6. Let $\mathcal{D}$ be the set of all datasets $X_n \in \mathbb{R}^p$ in general position and let $\mathcal{K}$ be a subset of $\mathcal{D}$ containing the datasets for which condition of "well clustered" data holds. Then the *restricted breakdown point* is defined as

$$\varepsilon^{(r)} = \max_{X_n \in \mathcal{K}} \varepsilon^{(i)} \tag{1.18}$$

Definition 6 is generally the one adopted to assess the robustness of a clustering model. Compare as an example Ruwet et al. (2013) where an explicit computation of the `tclust` method is provided.

More details may be encountered, besides the reference provided so far, in Huber (1981), in Huber & Ronchetti (2009), where the dissertation on the breakdown point has been hugely extended, and in Ruckdeschel & Horbenko (2012).

Chapter 2

Robust Clustering Methods

2.1 Introduction and state of art

Generally speaking "clusters may be thought as regions of high density separated from other such regions by regions of low density"(Hartigan 1975). Cluster analysis aims to identify a prefixed number of clusters within a given dataset. To do so, observations are usually grouped around suitably defined centroids following the aim of maximizing the heterogeneity between the groups and minimizing the homogeneity within the groups. Clusters' centroids are either observations or quantities, computed on clusters' observations, somehow representative of the whole cluster. Detailed reviews on clustering methods are provided, among the others, in Atkinson & Riani (2012) and in Hennig et al. (2015), or in Farcomeni & Greco (2015) and in Ritter (2014) for robust methods. Additionally alternative approaches based on grouping around different types of structures have been proposed so far, as a different notion of cluster may be of interest. In particular, one may be interested in clustering around linear structures or other type of manifolds as in García-Escudero et al. (2009), in García-Escudero, Gordaliza, Mayo-Iscar & San Martín (2010) and in Hennig (2003).

Among the different approaches to cluster analysis we may distinguish between *distance based* methods and *model based* methods. In the latter approach the assumption of an underlying population model is needed. In our opinion the latter approach has many advantages for mainly two reasons: first more flexible clusters' shapes are allowed. Secondly, as we are referring to a specified (and flexible as possible) statistical model, further inferential properties of the clustering method can be studied and assessed. Finally it must be pointed out that many *distance based* methods aim at optimizing fixed quantities which are directly connected with prob-

abilistic assumptions. As an example consider the k-means algorithm (MacQueen et al. 1967). This clustering method aims to minimize the Euclidean distance from k centroids which corresponds, from a probabilistic point of view, to assume that the data arise from k spherical homoscedastic multinormal populations. As in the case of the k means, oftentimes probabilistic assumptions are implicitly done even in cases where an underlying model is not properly formalized.
The outline of the chapter is as follows. Firstly we introduce the most relevant contributions related with robust clustering models. We start from the k-means' robust counterpart, the trimmed k-means (Cuesta-Albertos et al. 1997), and then, in section 2.2, we introduce more sophisticated methods which are able to deal with data divided in heterogeneous clusters. In section 2.3, we present the `tclust` method (García-Escudero et al. 2008) and the open issues related with this methodology.

2.1.1 Trimmed k-means

One of the most widely adopted approach for cluster analysis is the k-means algorithm. Given a sample $\{x_1, \ldots, x_n\}$ with $x_i \in \mathbb{R}^p$, k-means algorithm aims to minimize the following quantity:

$$\inf_{m_1,\ldots,m_k \in \mathbb{R}^p} \sum_{i=1}^{n} \min_{j=1,\ldots,k} ||x_i - m_j||^2 \tag{2.1}$$

It shall be noticed that the optimization problem introduced in formula (2.1) is based on minimizing a least square criterium and therefore, every solution to (2.1) may be affected by lack of robustness (García-Escudero, Gordaliza, Matrán & Mayo-Iscar 2010). A naive way to robustify the solution of optimization (2.1) is to replace the sample mean with the central observation of each cluster. This is, indeed, the strategy that led to formalize the PAM (partitioning around medoids) algorithm. Nevertheless, as it is proved in García-Escudero & Gordaliza (1999), PAM algorithm only provide a mild robustification. It must be pointed out that the influence function of the estimators of the centers is bounded, which implies that a single observation has a bounded influence on the centers' estimation, but, on the other hand, the associated break down point is equal to 0 even in the cases where the condition of "well clustered datasets" given in Ruwet et al. (2013) holds. This fact implies that, although the influence of a single observation is bounded, even one single observation placed very far can completely spoil the solution. Indeed, as a very far observation is inserted within the sample, the centers' estimation moves in the direction of such observation, and thus, cluster's contours may be inflated in an uncontrolled way. As a consequence the true underlying clusterwise structure may be completely hidden and the procedure completely breaks.

Following the ideas behind the MCD estimators, Cuesta-Albertos et al. (1997) proposed an embedded trimming step within the k-means algorithm in order to reach a break down point equal to α, the prefixed trimming level. The methodology aims to find the set of centroids optimizing the following minimization problem:

$$\inf_{\boldsymbol{Y}} \inf_{m_1,\ldots,m_k} \sum_{x_I \in \boldsymbol{Y}}^{n} \min_{j=1,2\ldots,k} ||x_i - m_j||^2 \tag{2.2}$$

It shall be noticed that the squared distance from the estimated centroids is calculated only for the observations included in $\boldsymbol{Y}$, where $\boldsymbol{Y}$ is a subset of the sample having size equal to $\lceil n \cdot (1-\alpha) \rceil$ and α is the proportion of observations to be trimmed off.
Despite its good properties, trimmed k-means has serious drawbacks when the assumption of homoscedasticity and sphericity of the clusters does not hold. Minimization of (2.2) is equivalent to maximizing the loglikelihood function associated with a trimmed mixture of k spherical multinormal population with common unit variances.

2.2 Heterogeneus robust clustering based on trimming

2.2.1 Formalization of the problem

As briefly stated in the previous section trimmed k-means algorithm is optimal whenever spherical groups are supposed. On the other hand, as data strongly depart from this assumption, the method potentially fails and yields wrong classification results. The adaptation of trimmed k-means for heterogeneous groups detection leads to the formulation of the "spurious outliers model", introduced in Gallegos (2002) and in Gallegos & Ritter (2005) and briefly outlined in Chapter 1 , Definition 2, and whose likelihood function is given in formula (1.1) and recalled in formula (2.9). Spurious outliers can be viewed as an extension of the Tukey-Huber contaminating model within the clustering context, while, on the other hand, the resulting estimators are an adaptation of the MCD philosophy for clustering models. The last term of equation (1.1) is the likelihood function associated to the noise component of the dataset. The maximum likelihood estimator of (1.1) exists if and only if the following condition on the contaminating density holds:

$$\arg\max_{\mathcal{R}} \max_{\mu_j,\Sigma_j} \prod_{j=1}^{k} \prod_{i \in R_j} f(x_i;\mu_j,\Sigma_j) \subseteq \arg\max_{\mathcal{R}} \prod_{i \notin \cup_{j=1}^{k} R_j} g_\psi(x_i) \tag{2.3}$$

As pointed out in Farcomeni (2014*a*), condition (2.3) states that identification of clean observations by maximization of the right hand term of (2.3) identifies the same observations as would identification of contaminated observations by maximizing the part of the likelihood corresponding to the noise. Thus, once clean observations are identified by maximizing the right hand term of (2.3), then the contaminated entries are optimally identified.
Additionally, if the condition (2.3) holds, the MLE of the likelihood function (1.1) has a simple representation and, its maximization reduces to the maximization of:

$$\sum_{j=1}^{n} \sum_{i \in R_j} \log f(x_i; \mu_j, \Sigma_j) \tag{2.4}$$

keeping the constraint $\# \cup_{j=1}^{k} R_j = \lceil n(1-\alpha) \rceil$. The MLE of $g_\psi(x_i)$ is in fact the Dirac's delta.
Additionally it shall be noticed that minimizing (2.4) in the case $k = 1$ is equivalent to perform the minimization that leads to the MCD estimators. In order to maximize (2.4) an iterative procedure, which will be described in the further subsection, is required.

2.2.2 A "naive" extension of the fast MCD algorithm

As pointed out in García-Escudero, Gordaliza, Matrán & Mayo-Iscar (2010) maximization of (2.4) requires an algorithm that is a "naive" extension of the fast MCD algorithm outlined in Chapter 1. The algorithm iterates the following steps:

Algorithm 3.

1. *Initialization*: Initialize randomly k initial centers $m_1, \ldots, m_k$ and k covariance matrices $\Sigma_1, \ldots, \Sigma_k$

2. *Concentration steps*:

   2.1 Keep the set H containing the $\lceil n(1-\alpha) \rceil$ observations closest (w.r.t the Mahalanobis distance) to the estimated centroids $m_1, \ldots, m_k$.

   2.2 For each $i = 1 \ldots n$ obtain the clusters' assignments by computing the minimization $\inf_j d^2_{\Sigma_j}(x_i; m_j)$.

   2.3 Update the estimates of the clusters' centers $m_1, \ldots, m_k$ and $\Sigma_1, \ldots, \Sigma_k$.

3. Repeat steps 2.1 - 2.3 until there are no improvements in equation (2.4).

The iterative procedure is an EM-type algorithm whose convergence to a local maximum has been proved in Dempster et al. (1977). To be more precise, it is a Classification-EM algorithm (Celeux & Govaert 1992). Indeed, in the EM algorithm the *a posteriori* probabilities of each observation to belong to each cluster are kept in account. Such estimated probabilities play the role of weighting the yielding parameters' estimations. This is a very common algorithm within the mixture modelling context. As in clustering context one may be interested in *fully* assigning an observation to each group, *crispy* weights, computed by the *a posteriori* probabilities are generally considered.

As a final remark it must be pointed out that equation (2.4) is unbounded. Thus there can be spurious maximizers of the objective function which can completely spoil the solution, as we now detail.

2.2.3 Spurious maximizers

Maximization of (2.4) is a mathematically ill-posed problem since the objective function is unbounded. Such problem, noticed in many contributions (Maronna & Jacovkis (1974) among the others), still remains an open issue within the mixture modelling and model based clustering literature. Nowadays, in order to avoid problems related with the unboundedness of the objective function (2.4), the optimization is performed under proper constraints. These generally involve the estimated scatter matrices. Depending on the context of application, there are different constraints that have been proposed so far. Some examples, to be better defined in the further sections are the eigenvalue ratio constraint or the the Hathaway-Dennis-Beale-Thompson constraint.
Generally speaking, spurious maximizers can be defined as a set of points either too close among each other or lying in a lower-dimensional space. As a consequence, the covariance matrix associated to these points is almost singular, its determinant is very close to 0 and thus the objective function tends to infinity. As a consequence on the final results the clusterwise structure of the data is hidden and the set of such observations is identified in the final output as one of the detected groups. Figure 2.1 reports the classification's results of a clustering procedure when $k = 2$. Panel (a) shows the results as no constraint has been imposed. As a consequence, a set of collinear points, which yield variance equal to 0 for one component, is recognized as a cluster and the real underlying clusterwise structure is not properly recognized by the model. On the other hand, in panel (b), we plotted the results obtained by keeping constraining the estimated clusters' variances: the two clusters that clearly appear in the data are correctly recognized by the procedure and the observations

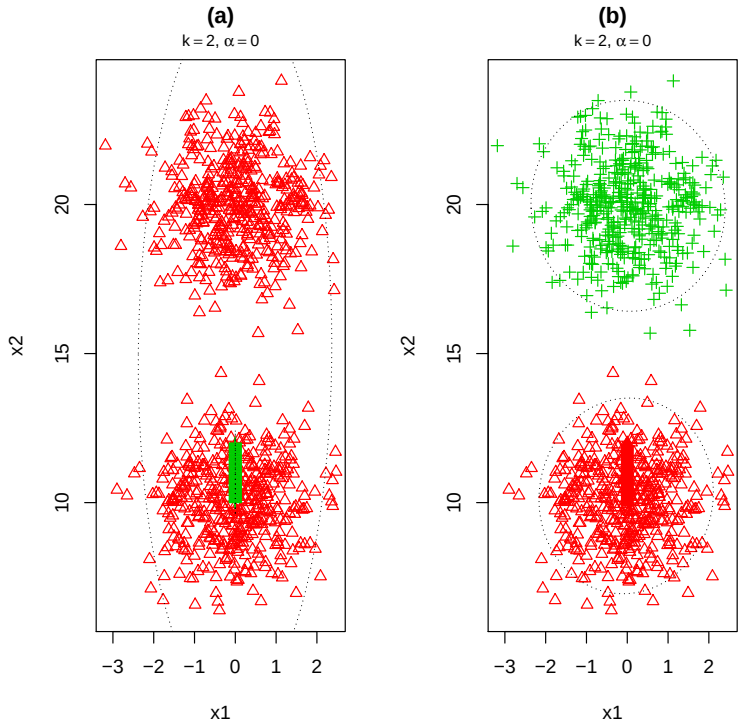

Figure 2.1: Comparison between constrained and unconstrained clustering.

look well classified.

2.2.4 Constraint based on the determinant

Unit determinant covariance matrix

In order to avoid that any determinant of the estimated scatter matrices potentially goes to 0, Gallegos (2002) proposed to factorize the group covariance matrix Σ_j as $\Sigma_j = \sigma_j U_j$ where $U_j = \Sigma_j/|\Sigma_j|^{1/p}$. In this factorization U_j is the group "shape" matrix and it is such that its determinant is equal to 1, while σ_j is the group scale parameter. The resulting clustering algorithm iterates within the same steps of Algorithm 3 but a modified computation of the Mahalanobis distance is used in the *Concentration step*. The modified Mahalanobis distance is given by:

$$\widetilde{d_{\Sigma_j}}(x, m_j) = (x - m_j)^T (U_j)^{-1} (x - m_j) \tag{2.5}$$

This clustering method of course is able to avoid the undesired effects of the spurious maximizers but, on the other hand, solutions containing groups with equal scales are favored.

Homogeneity

Another proposal can be found in Gallegos & Ritter (2005) where the same unknown scatter matrix is imposed as the covariance matrix of each group. As in the previously presented cases, we refer to and adaptation of the spurious outliers model, whose likelihood function is given by:

$$\left[\prod_{j=1}^{K}\prod_{i\in R_j} f(x_i;\mu_j,\Sigma)\right]\left[\prod_{i\notin R_j} g_{\psi_i}(x_i)\right] \tag{2.6}$$

It shall be noticed that in equation (2.6) the scatter component does not depend on the estimated clusters. This method avoids the effect of the spurious maximizers but, clearly, fails in presence of high heterogeneity between groups.

2.2.5 Hathaway-Dennis-Beale-Thompson constraints

A further proposal appeared in Gallegos & Ritter (2009*b*). This is based on constraining the Hathaway-Dennis-Beale-Thompson (HDBT) ratio of the k estimated covariance matrices. This is an adaptation of the constraint proposed in Hathaway (1985) for the multivariate case and is defined as follows.

Definition 7. Given a set of estimated covariance matrices $\Sigma_1 \ldots, \Sigma_k$ the HDBT ratio is defined as the maximum value c such that the following holds:

$$\Sigma_j \succeq c\Sigma_l \text{ for } j \leq 1, l \leq k \tag{2.7}$$

where the operator $\succeq$ in equation (2.7) recalls the Löwner ordering on the space of the symmetric matrices.

It shall be noticed that, as c is imposed to be equal to 1, then homoscedasticity is imposed, while, as c tends to 0 more degrees of freedom in the scatter estimation are allowed. Furthermore Gallegos & Ritter (2009*b*) showed that an explicit computation of the HDBT ratio is given by:

$$\min_{j,k,l} \lambda_k\left(\Sigma_l^{-1/2}\Sigma_j\Sigma_l^{-1/2}\right) \tag{2.8}$$

where $\lambda_k(\Sigma_l)$ the k-th eigenvalue of the matrix Σ_l. Operatively speaking the resulting algorithm, which iterates exactly the same steps of Algorithm 3, does not yield constrained solutions. Indeed, as pointed out in Fritz et al. (2013*b*) "the authors propose to obtain all possible local maxima of the trimmed likelihood and, afterwards, the ratio in (2.7) and the value of the trimmed likelihood for these local

maxima are monitored in order to choose sensible candidate clustering solutions". Usage of such approach has the advantage that affine equivariance of the estimators is preserved. On the other hand finding an optimal combination of local maxima of the objective function and a "suitable" value for the HDBT ratio is not straightforward even considering the heuristics proposed in Gallegos & Ritter (2009*b*).

2.3 The TCLUST methodology

2.3.1 Introduction

The `tclust` methodology (García-Escudero et al. 2008) is a robust model based clustering method designed with the aim of fitting clusters with different scatters and different weights. The robustness of the method is guaranteed by the fact that a fixed proportion of observations α is trimmed. Additionally, the methodology is designed to deal with collinear points that may arise in a given sample. Indeed, the effect of the spurious maximizers is avoided by constraining the ratio between the highest and the lowest eigenvalues of the estimated scatter matrices. Usage of such constraint (ER, eigenvalue ratio) guarantees the consistency to the population parameters. A formal proof of this statement is provided in García-Escudero et al. (2008).
This methodology has been implemented in the open source software R, within the `tclust` package (Fritz et al. 2012*a*). A formal study of its robustness properties can be found in Ruwet et al. (2012) and in Ruwet et al. (2013). Nowadays many extensions of such method have been proposed. In particular, the method has been extended for linear clustering problems in García-Escudero et al. (2009) and in García-Escudero, Gordaliza, Mayo-Iscar & San Martín (2010), for fuzzy methods in Fritz et al. (2013*a*), for achieving robustness against entry-wise outliers in Farcomeni (2014*a*), and for double clustering methods in Farcomeni (2009).

2.3.2 Mathematical formulation

The objective function of the `tclust` is given by:

$$\left[\prod_{j=1}^{K}\prod_{i\in R_j}\pi_j f(x_i;\mu_j;\Sigma_j)\right]\left[\prod_{i\notin R_j} g_{\psi_i}(x_i)\right] \tag{2.9}$$

An equivalent formulation of such objective function that can be used whenever condition (2.3) holds is given by:

$$\sum_{j=1}^{n} \sum_{i \in R_j} \log(\pi_j f(x_i; \mu_j, \Sigma_j)) \tag{2.10}$$

It shall be noticed that difference between (2.4) and (2.9) is that the latter include clusters weights π_j, and thus a bias toward equal sized clusters is avoided.
Additionally the maximization is performed under the so called eigenvalue ratio (ER) constraint defined as:

$$\frac{M_n}{m_n} = \frac{\max_{j=1,2,\ldots,K} \max_{l=1,2,\ldots,p} \lambda_l(\Sigma_j)}{\min_{j=1,2,\ldots,K} \min_{l=1,2,\ldots,p} \lambda_l(\Sigma_j)} \tag{2.11}$$

where, in formula (2.11), $\lambda_l(\Sigma_j)$ are the eigenvalues of the scatter matrix Σ_j for $j = 1, 2, \ldots, K$ and for $l = 1, 2, \ldots, p$ and c is a fixed constant ≥ 1. Usage of the constraint defined in (2.11) has two main advantages: a feasible algorithmic implementation is available (compare García-Escudero et al. (2015) for details) and it has an easy geometric interpretation as well. Indeed, as $c = 1$ spherical clusters are imposed, while, as c increases, more differently shaped clusters are allowed in the final output of the procedure. Although the estimators obtained under constraint (2.11) are not affine equivariant, imposing high values for the constant c allows to obtain "almost" affine equivariant estimators. Finally ER constraint has strong relationship with the HBDT constraint. Indeed it can proved that (Ruwet et al. 2012), if ER holds, then also HDBT holds but the converse is not true. Additionally, in the afore mentioned reference is proved, in terms of influence function, that TCLUST method is robust under more general conditions, which can be veiwed, as authors commented in Ruwet et al. (2013) as "compensation for the loss of affine equivariance".

2.3.3 The algorithm

Clearly maximization of (2.9) cannot be performed analytically, and thus an iterative procedure is required. In particular the `tclust` algorithm is given by the following steps:

Algorithm 4.

1. *Initialization*: Initialize randomly k initial centers $m_1^0, \ldots, m_k^0$, k covariance matrices $\Sigma_1^0, \ldots, \Sigma_k^0$ and k values $p_1^0, \ldots, p_k^0$ or the clusters' weights.

2. *Concentration steps*:

2.1 Keep the set H containing the $\lceil n(1-\alpha)\rceil$ observations closest (w.r.t the Mahalanobis distance) to the estimated centroids $m_1, \ldots, m_k$.

2.2 For each $i = 1 \ldots n$ obtain the clusters' assignments by computing the minimization $\min_j d^2_{\Sigma_j}(x_i; m_j)$.

2.3 Update the estimates of the clusters' centers $m_1, \ldots, m_k$, clusters' scatter matrices $\Sigma_1, \ldots, \Sigma_k$, and clusters' weights $p_1, \ldots, p_k$. In the scatter matrices's estimation apply the algorithm proposed in García-Escudero et al. (2015) to obtain variances obeying constraint (2.11).

3. Repeat Steps 2.1 - 2.3 until there are no improvements in equation (2.9).

4. Draw several different random starting values and recompute the values of the objective function. Keep the configuration yielding the maximal value of (2.9) as the final output of the algorithm.

It shall be noticed that, as we are referring to an impartial trimming based method, only the fixed proportion of $\lceil n \cdot (1-\alpha)\rceil$ observations contribute to the parameters'estimation, while the remaining are discarded.

2.3.4 Open Issues

Simulations and theoretical results have shown that the TCLUST method is robust and gives efficient estimations both in terms of the clusters' parameters and in terms of classification's results. Nevertheless, as often times happens for robust methods, tuning of the procedure is required and is not automatic. Reasonable values" for the trimming level α and for the constraint on the eigenvalues c are required.

Fixing the trimming level

All trimming based methods, including the TCLUST, require to fix in advance α, the proportion of observations to be discarded. The loss in fixing a trimming level α is not symmetric: if it is too low, outliers can completely spoil the solution. If it is too high, a loss of efficiency (which is usually less problematic than the first scenario) is incurred. We now outline some heuristic proposals to fix such tuning parameter. Fritz et al. (2012*a*) introduced the `ctl curves`, that will be used, within this thesis in Chapter 4. *ctl curves* are helpful for the user to find the number of underlying groups have an idea of the amount of contamination. Operatively speaking, by looking at the plot of the *ctl curves* the user is able to monitor the evolution of the

objective function as the imposed trimming level and the number of clusters are increased. The idea is that once the outlying component of the data is trimmed, then the objective function shows a more stable trend for increasing values of α. Useful guidelines on the interpretation of this plot are also provided in García-Escudero et al. (2015). Another heuristic method for fixing reasonable values of α is proposed in Farcomeni & Greco (2015) where the G-statistics has been introduced.
Our proposal, to be better explained within Chapter 3, is to use an iterative method based on reweighting. The idea is to fix a high initial trimming level α_0. Then reweighting is based on flagging as outlying observations whose Mahalanobis distance is above the opportune quantile of the χ^2 distribution, and updating the parameters. The procedure is stopped as soon as the trimming level and the proportion of observations discarded by the outlier test coincide. As it can be seen from the simulation study, such method does not need much tuning, can resist to high proportion of outliers and is efficient even with little or no contamination.

Fixing reasonable value for the ER constraint

Fixing a proper constraint c is also cumbersome. There are indeed two important facts that should be kept in mind:

1. Whenever too "restrictive" values are imposed, one may incur in solution biased towards spherical clusters. On the contrary, as too "high values" are imposed, the risk of considering spurious solutions increases.

2. Imposition of constraint (2.11) leads to the loss of the affine equivariance of the estimators (although this problem may be overcome by imposing "very" high values)

In García-Escudero et al. (2015) appeared a contribution mainly focused on fixing reasonable values for the constant c. In the afore mentioned reference the authors propose to monitor the evolution of the objective function obtained as different values of c are imposed. Indeed, is not necessary to be really precise in fixing such constant. A huge range of values of c is suitable for avoiding the effect of the spurious maximizers and simultaneously contain the bias in the scatters' estimations.
The idea of using the geometric constraints outlined in Celeux & Govaert (1995), instead of the ER, is an ongoing work that will be briefely outlined in the section containing the further direction of research. Depending on the imposed constraint different properties in terms of the affine equivariance of the estimators can be obtained. Additionally these constraints have a direct geometric interpretability.

Other clustering methods

All the robust clustering methods mentioned so far are based on trimming. However there are several interesting robust proposals which are not based on trimming.
As usual, let us assume that the dataset is divided in k groups. Non trimming approaches are based on fitting a mixture of k Gaussian components and accommodating the "noisy" part of the data in a component generated by a different probability distribution.
Banfield & Raftery (1993) and Fraley & Raftery (1998) propose to fit a mixtures of k Gaussian distributions for the set of the "clean" data and a uniform distribution defined on the convex hull of the data for the noisy component of the dataset. Later on Coretto & Hennig (2013) provided a more robust approach. This is based on the idea of classifying the data as noisy whenever they have the density, for all Gaussian components, with values smaller than a fixed constant c. Such method is robust and several theoretical properties have been proved in Coretto & Hennig (2013), but tuning is cumbersome. Indeed fixing and interpreting the values of the afore mentioned threshold for the "contaminating" Gaussian density is not straightforward. Some guidelines are provided in Coretto & Hennig (in press 2016). Additionally robust clustering models can be implemented by adapting the "forward search" to the clustering context. Indeed, the plots outlined in Atkinson et al. (2004) provide some heuristic useful to both determine the number of underlying groups in the dataset and recognize the farthest observations.

Chapter 3

Reweighting in Robust Clustering

3.1 Introduction

Within this chapter we propose an iterative method targeted at simultaneously increasing the efficiency and estimating the proportion of contamination in a dataset. The main part of the contents of this Chapter can be found in Dotto et al. (2015) and in Dotto et al. (2016*b*). The outline of the chapter is as follows. Firstly we briefly introduce the problem. In section 3.2 we formally introduce the methodology. In Section 3.3 we outline the algorithm while a detailed explanation of each step is reported in section 3.4. Section 3.5 contains some illustrative examples, in Section 3.6 we study the theoretical properties of the methodology while in Section 3.7 we report the simulation study. Finally, in Chapter 5 we apply the proposed methodology to two different dataset and report the results obtained, while the proofs of the theoretical statements are stored in Appendix A.1.

Generally speaking robust clustering models aim to provide robustness by considering outlier-free subsamples extracted from the data and by discarding observations outside these subsamples. To do so trimming is generally used. In Chapter 2 the problem of fixing a proper value for the trimming level α, compulsory for applying robust methods based on trimming, has been generally introduced. We now wish to recall that the loss in fixing a trimming level α is not symmetric: if it is too low, outliers can completely spoil the solution. If it is too high, a loss of efficiency (which is usually less problematic than the first scenario) is incurred. For this reason, a preventive (higher than needed) trimming level is often considered. This could result in a high number of non-outlying observations which are wrongly trimmed, and loss of efficiency in subsequent statistical analyses. Carefully tuning the trimming level

may be cumbersome in several applications, and the final results may be dependent on a subjective choice of this tuning parameter. Additionally a high number of wrongly trimmed observations (due to the consideration of high initial preventive α_0 trimming levels) could be a major problem as researchers usually would like to assign as many observations as possible to a cluster. Failure to assign a clean observation to a cluster might be associated with practical consequences. For instance in marketing research not assigning a potential buyer to a his/her appropriate cluster is associated to loss of the revenue associated with the future transaction. For that reason we now aim to reduce as much as possible, in a data driven fashion, the trimming proportion.

A popular solution in robust statistics is to resort to reweighting methodologies. Reweighting of each observation x_i is usually based on the Mahalanobis distance through $w_i = v(d_i)$, with $v(\cdot)$ being a non-increasing function. The weights w_i allow us to compute (one-step) reweighed location and scatter estimators which have good robustness performance and better efficiency behavior just by considering weighted sample means and weighted sample covariances. See Lopuhaa (1999) for a detailed discussion on the properties of reweighted estimators. The approach could be then iterated (e.g., Cerioli 2010).

A very simple and widely applied approach is to use binary weights. Given initial T and S (robust) location and scatter matrices estimators and their associated Mahalanobis distances $d_i = d_S(x_i, T)$, we can simply use

$$w_i = 1 \text{ if } d_i \leq \sqrt{\chi^2_{p,\alpha_L}} \text{ and } w_i = 0 \text{ otherwise.} \tag{3.1}$$

We use the notation $\chi^2_{p,\beta}$ for a $1-\beta$ quantile of the χ^2_p distribution and α_L is taken as a positive value close to 0. This allows to recover some of the wrongly trimmed observations, which could have not been taken into account when computing T and S, by assuming a normal distribution for the non-outlying part of data. In our proposal, as we now detail, we fix a sequence of decreasing trimming level and for each trimming level we update parameter's estimation. At each step we aim to reinsert observations initially flagged as outlying comparing their Mahalanobis distance with the threshold given in (3.1) in order to increase the efficiency and reduce the number of discarded observations.

3.2 Methodology

Let us assume that the number of clusters k is known in advance but the proportion of observations π_j in each cluster is unknown and the true contamination level π_0

is also unknown. Non-outlying observations come from a mixture of k normally distributed components, and contamination might be present in our data. We also loosely make the assumption that the components are not too much overlapping. The proposed methodology is initialized with a large trimming level α_0 which -hopefully- guarantees the detection of a proportion $1-\alpha_0$ of outlier-free observations in the most central regions of each cluster. These observations can be seen somehow as the *cores* of the clusters. Starting from the cores we will consider a sequence of decreasing trimming levels $\alpha_0 > \alpha_1 > ... > \alpha_L$ with α_0 being an initial preventive (i.e., surely higher than needed) trimming level and α_L is a value close to 0 that can be interpreted in a similar fashion as parameter α_L in (3.1).

In this iterative process better estimates of the cluster centroids, scatter matrices, cluster proportions, and the contamination level are consecutively obtained. Providing efficient estimates of these parameters is helpful to detect the outliers and, consequently, avoid their insertion in the final set of the clustered data eventually stopped prior to reaching the small trimming levels that would include outliers in estimation sets. Our proposal, to be better detailed below, can be seen as an extension of the procedure presented in García-Escudero & Gordaliza (2007) where the final trimming level had to be determined manually.

Figure 3.1 shows the result of applying the proposed methodology to two simulated datasets. The first one shown in panel (a.1) is the result of simulating a mixture of two normal components with no contamination. In panel (b.1) 10% of the observations are replaced by outlying data points. A more detailed description of the simulation scheme will be given in Section 3.7. Panels (a.2) and (b.2) show the results of TCLUST (García-Escudero et al. 2008) with $\alpha_0 = 0.33$ trimming. Several wrongly trimmed observations can be seen, but also that the TCLUST procedure successfully identifies cluster cores. Finally, panels (a.3) and (b.3) show the results of the proposed methodology, which we name RTCLUST, which in both cases adapts well to the true underlying contamination.

The underlying idea is that using an initial very robust estimator would make the procedure resistant to a very high proportion of outliers (i.e., have a breakdown point of α_0). On the other hand, iteratively decreasing the trimming level would make the procedure almost as efficient as the non-robust counterparts. A similar idea but with a different rationale was proposed in Hardin & Rocke (2004), where an initial solution is improved based on a scaled F approximation to the distribution of Mahalanobis distances (see also Hardin & Rocke 2005). We will compare in simulations below.

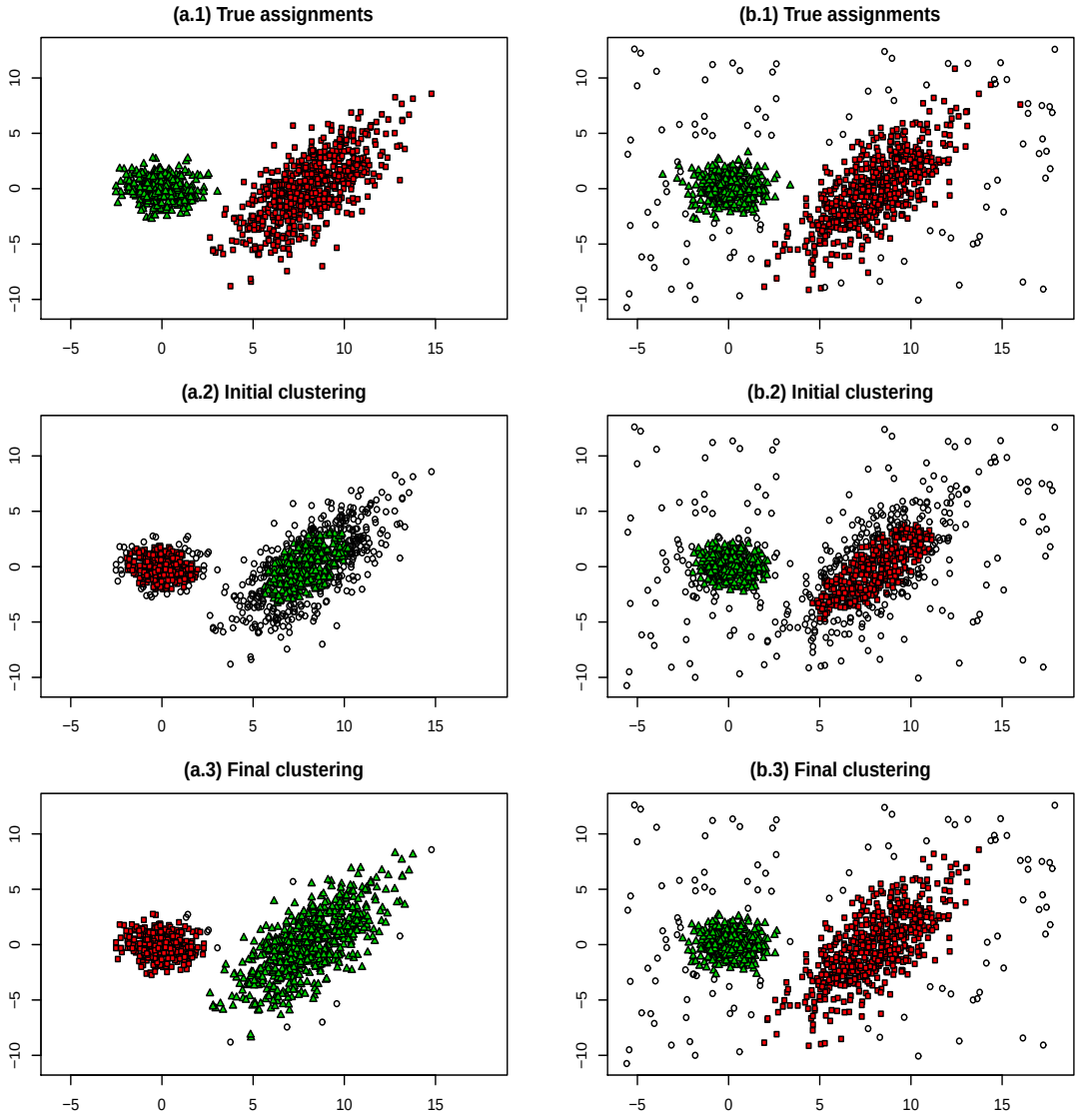

Figure 3.1: Two simulated data sets with their true assignments in (a.1) and (b.1). The result of TCLUST with $\alpha_0 = 0.33$ in (a.2) and (b.2). The final assignments obtained after applying the proposed methodology are given in (a.3) and (b.3). Noisy data and trimmed are denoted by $\circ$ in all graphs throughout the manuscript.

It is important to stress that while we will estimate the contamination level, and evaluate masking and swamping, what we are proposing is *not* a method to simultaneously perform robust clustering and outlier detection. We aim at obtaining robust and efficient estimates of partitions and model parameters. Outlier detection should then be based on robust estimators, but should be performed separately based on formal rules (see e.g. Cerioli & Farcomeni (2011) for a general discussion on this point).

3.3 The algorithm

There it follows a brief description of the algorithm which iterates the following steps:

Algorithm 5.

1. *Initialization:* Set the initial parameters' set $\pi_1^0, ..., \pi_k^0, \pi_{k+1}^0$, $\mu_1^0, ..., \mu_k^0$ and $\Sigma_1^0, ..., \Sigma_k^0$ obtained by applying the TCLUST with a high trimming level α_0.

2. *Reweighting process:* Consider $\alpha_l = \alpha_0 - l \cdot \varepsilon$ with $\varepsilon = (\alpha_L - \alpha_0)/L$ for $l = 1, ..., L$

   2.1 *Fill the clusters:* Given $\pi_1^{l-1}, ..., \pi_k^{l-1}, \pi_{k+1}^{l-1}$, $\mu_1^{l-1}, ..., \mu_k^{l-1}$ and $\Sigma_1^{l-1}, ..., \Sigma_k^{l-1}$ from the previous step, let us consider

   $$D_i = \min_{1 \leq j \leq k} d^2_{\Sigma_j^{l-1}}(x_i, \mu_j^{l-1}) \tag{3.2}$$

   and sort these values as $D_{(1)} \leq ... \leq D_{(n)}$. Take the sets

   $$A = \{x_i : D_i \leq D_{([n(1-\alpha_l)])}\} \text{ and } B = \{x_i : D_i \leq \chi^2_{p,\alpha_L}\}$$

   Now, use the distances in (3.2) to obtain a partition $A \cap B = \{H_1, ..., H_k\}$ with

   $$H_j = \left\{ x_i \in A \cap B \ : \ d_{\Sigma_j^{l-1}}(x_i, \mu_j^{l-1}) = \min_{q=1,...,k} d_{\Sigma_q^{l-1}}(x_i, \mu_q^{l-1}) \right\}.$$

   2.2 *Update cluster weights* The proportion of contamination is estimated by computing

   $$\pi_{k+1}^l = 1 - \frac{\#B}{n}.$$

   Given $n_j = \#H_j$ and $n_0 = n_1 + ... + n_k$ the cluster weights are estimated by computing:

   $$\pi_j^l = \frac{n_j}{n_0}\big(1 - \pi_{k+1}^l\big). \tag{3.3}$$

   2.2 *Update locations and scatters:* Update the cluster centers by taking μ_j^l equal the sample mean of the observations in H_j and the scatter by computing the sample covariance matrix of the observations in H_j multiplied by its consistency factor.

3. *Output of the algorithm:* $\mu_1^L, ..., \mu_k^L$ and $\Sigma_1^L, ..., \Sigma_k^L$ are the final parameters estimates for the normal components. From them, final assignments are done by computing

   $$D_i = \min_{1 \leq j \leq k} d^2_{\Sigma_j^L}(x_i, \mu_j^L),$$

for $i = 1, ..., n$. Observations assigned to cluster j are those in H_j with

$$H_j = \left\{ x_i : d_{\Sigma_j^L}(x_i, \mu_j^L) = \min_{q=1,...,k} d_{\Sigma_q^L}(x_i, \mu_q^L) \text{ and } D_i \leq \chi^2_{p,\alpha_L} \right\}$$

and the trimmed observations are observations not assigned to any of these H_j sets (i.e., those observations with $D_i > \chi^2_{p,\alpha_L}$).

3.4 Comments on the algorithm

Initialization

As briefly stated in the previous section we initialized the algorithm with the output of the TCLUST methodology on which a high trimming level has been imposed. Nevertheless we wish to point out that other robust proposals may also be used as initialization of the `rtclust` algorithm. For instance, methods derived from the maximization of (2.10) with different constraints on the Σ_j matrices and/or removing π_j weights can be used. See Cuesta-Albertos et al. (1997), Hennig (2003), Gallegos & Ritter (2005) or Neykov et al. (2007) among others. Whenever an initialization that does not keep in account the π_j weights is used, then one may consider $\pi_1^0 = ... = \pi_k^0 = 1/k$ to initialize the procedure.

Filling the clusters

Step 2.1 is targeted at keeping outliers outside $A \cap B$, while increasing the trimming size in a controlled fashion. Indeed at each step of the reweighting process only a prefixed number of closest observations, given by $\lceil n(1-\alpha_l) \rceil$, where α_l represents the current trimming level, are inserted in the set of the clean observations. Alongside, better parameter estimates are obtained by increasing the active sample size.

Estimating clusters weights

It shall be noticed that clusters weights are not estimated by directly computing the estimated cluster proportions. Indeed, in order to provide better estimations of location and scatter parameters, each cluster is filled, in the reweighting process, by considering the subset of the most central observations. On the other hand, we aim to provide precise estimations of the cluster weights and thus, we estimate them by applying formula (3.3).

Estimating covariance matrices

In the step 2.2 we use well-known correction factors (see, e.g. Liu et al. 1999) to inflate the covariance matrix estimates based on trimmed data. These guarantee consistency at the normal model. At each stage the fraction of observations in the central region of group j is $n_j/n\pi_j^l = n_0/(n(1-\pi_{k+1}^l))$, where $n\pi_j^l$ is an estimate of the total number of observations in group j. Additionally, covariance estimates need to be corrected by considering correcting factor defined as

$$c_j = \left(\eta_{\frac{n_0}{n(1-\pi_{k+1}^l)}}\right)^{-1} \text{ if } \frac{n_0}{n(1-\pi_{k+1}^l)} < 1$$

and

$$c_j = 1 \quad \text{if } \frac{n_0}{n(1-\pi_{k+1}^l)} \geq 1$$

where $\eta_\beta = P\left(\chi^2_{p+2} \leq \chi^2_{p,\beta}\right)/\beta$ and $\beta = \#H_j/n\pi_j$.
We finally update the scatter matrices as

$$\Sigma_j^l = S_j^l \cdot c_j.$$

Remark 1. More sophisticated rules for discarding outliers, for instance, based on using the Beta distribution or multiple testing corrections could have been tried (Cerioli 2010, Cerioli & Farcomeni 2011). However, for sake of clarity of presentation, we have preferred the simpler use of a rule just based on χ^2_{p,α_L}. There is still room for improvement regarding better detection of outlying observations.

Remark 2. Sometimes, we could be interested in forcing some "a priori" constraints like those in (2.11) to the final estimated clusters scatter matrices. In this case, constraints can be forced by truncating the scatter matrices eigenvalues in the updating step 2.2, as done in Fritz et al. (2013*b*).

3.5 Illustrative examples

The two component normal mixture shown in panels (b.1) of Figure 3.1 account for 36% and 54% of the observations, respectively, while a 10% of not "very overlapped" contamination is added. The scatter matrix for the first component is Σ_1 equal to the identity matrix and Σ_2 is a scatter matrix with $|\Sigma_2| = 20$ and eigenvalues equal to 11.708 and 1.708. This means that the "true" eigenvalue ratio for these two scatter matrices is equal to 11.708. A more detailed description of the process generating this data set will be given in Section 3.7. We will use this data set in order to illustrate the lack of dependence of the final solution on the initializing trimming

level α_0 and on the initial value of the restriction factor c when TCLUST is used as initializing procedure. Figure 3.2 shows the evolution of the determinants of the scatter matrices, i.e. $\{|\Sigma_j^l|\}_{l=0}^L$ for $j = 1, 2$ in panel (a), and the evolution of the estimated contamination level and estimated cluster sizes, i.e. $\{\pi_j^l\}_{l=0}^L$ for $j = 0, 1, 2$ in (b). These evolutions are studied for different values of $\alpha_0 = 0.3$, 0.25, 0.2 and 0.15 and it is always considered the same (wrong) eigenvalue ratio constraint value $c = 5$ for the TCLUST method as initializing procedure. We can see that the final output is not very dependent on the initializing trimming level and that the output estimated parameters are very close to the true ones we want to estimate (i.e., $|\Sigma_1| = 1$ and $|\Sigma_2| = 20$ for the cluster scatter matrices determinants and $\pi_0 = 0.1$, $\pi_1 = 0.36$ and $\pi_2 = 0.54$ for the contamination level and cluster sizes).

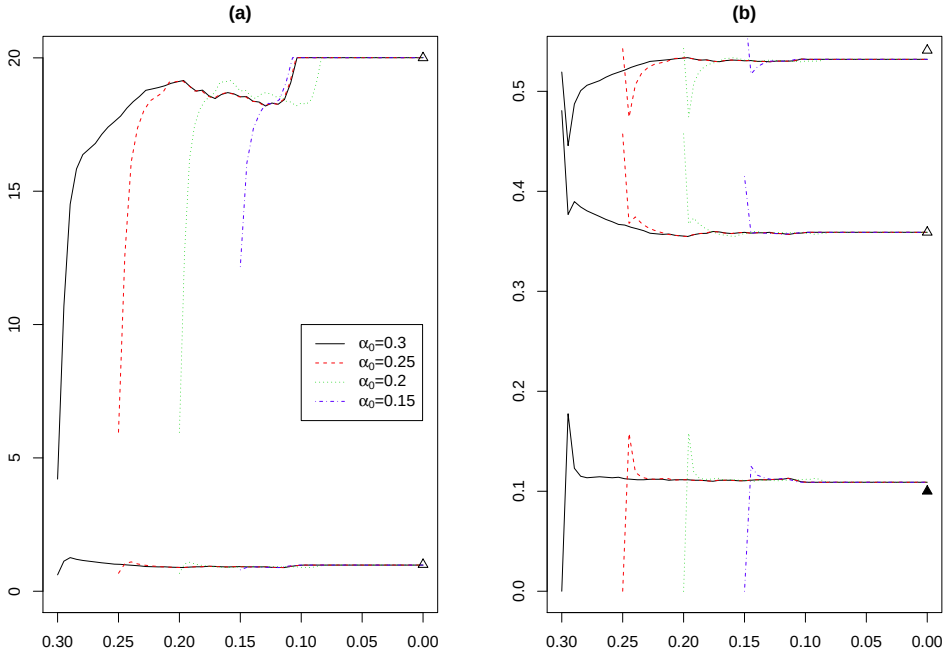

Figure 3.2: Evolution of $|\Sigma_j^l|$ in (a) and of π_j^l in (b) for different initial α_0 values ($\alpha_0 = .3$, .25, .2 and .15) for the data set shown in Figure 3.1 (b.1). The up-triangle symbols are the true parameters to be estimated.

Analogously, the same type of study was made to analyze the possible dependence on the initializing choice of c. The results are shown in Figure 3.3 where c values equal to 1, 10 and 20 were chosen. Recall that the true eigenvalue ratio for the considered scatter matrices was exactly equal to 11.708 (which is not equal to any of the c initializing values tried). We can see again that the obtained results are accurate and that they are not very dependent on the initial c value.

It is also important to note, in Figure 3.2 and Figure 3.3, that no great changes are noticeable in the estimated parameters when the procedure approximately reaches

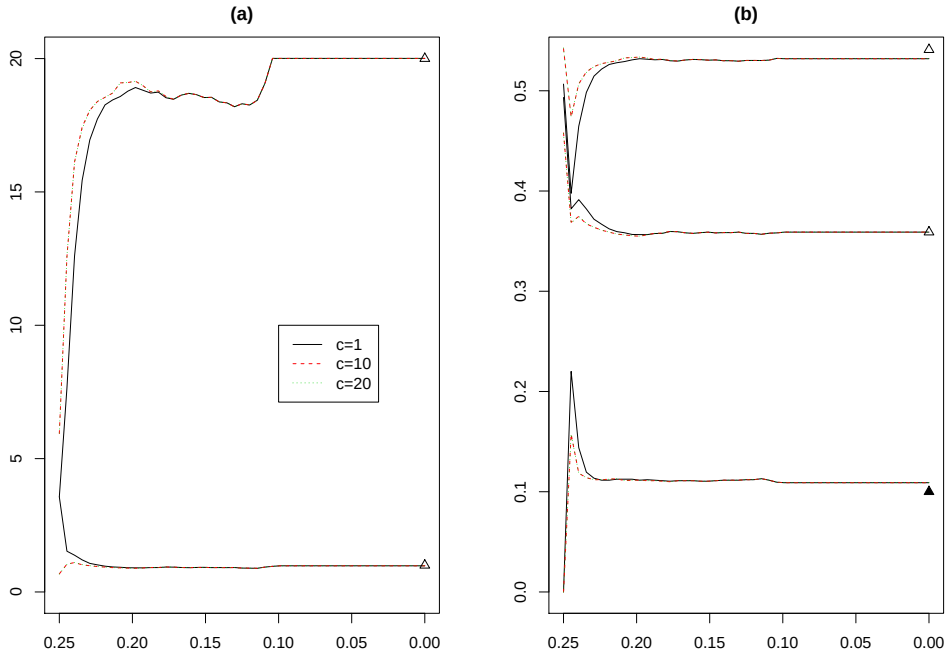

Figure 3.3: Evolution of $|\Sigma_j^l|$ in (a) and of π_j^l in (b) for different initial c values ($c = 1$, 10 and 20 while the true c needed was 11.71) for the data set shown in Figure 3.1 (b.1). The up-triangle symbols are the true parameters to be estimated.

the true contamination level. This is because, we count on quite accurate estimators of the parameters of the normal distributions components throughout μ_j^l and Σ_j^l when $\alpha_l \approx 0.1$. Due to their effect the set $A \cap B$ defined in Step 2.1 remains essentially the same and equal to the set having all the regular (non-noisy) observations already included. On the other hand, one-step procedures only take into account the information from truncated sub-samples corresponding to central regions in the normal components. From this central regions, it is not so easy to have very accurate parameters estimations for the normal components parameters.

To reinforce our previous claims, we will illustrate the advantages of the proposed iterative trimming procedure with respect to one-step reweighting approaches even in the $k = 1$ case. When $k = 1$, the reweighted MCD is clearly one of the most popular robust location and scatter estimator. After considering an initial large trimming level α_0, reweighting is done to increase efficiency as described in Section 3.1.

Figure 3.4 is based in a simulated data set of size $n = 1000$ generated from a bivariate normal distribution accounting for 73% of the data (the bulk of data), a 24% amount of pointwise contamination placed at $(4, 8)$ (labeled with an "arrow" symbol) and 3% of background contamination. Figure 3.4,(b) shows the result of applying the reweighted MCD approach in Section 3.1 by using the "robustbase"

package in R available in the CRAN repository with the default initial trimming level $\alpha_0 \simeq 0.5$ and $\alpha_L = 0.01$ and the function "tolEllipsePlot" (from "robustbase") to plot the 0.99 tolerance ellipses (the classical and the MCD-based robust ones). Despite there exists a "good" initial sub-population including more than half of the observations, the final estimation is very distorted by the added pointwise contamination as can be seen in 3.4,(b). On the other hand, Figure 3.4,(a) shows how the proposed iterative trimming resists very well this pointwise contamination.

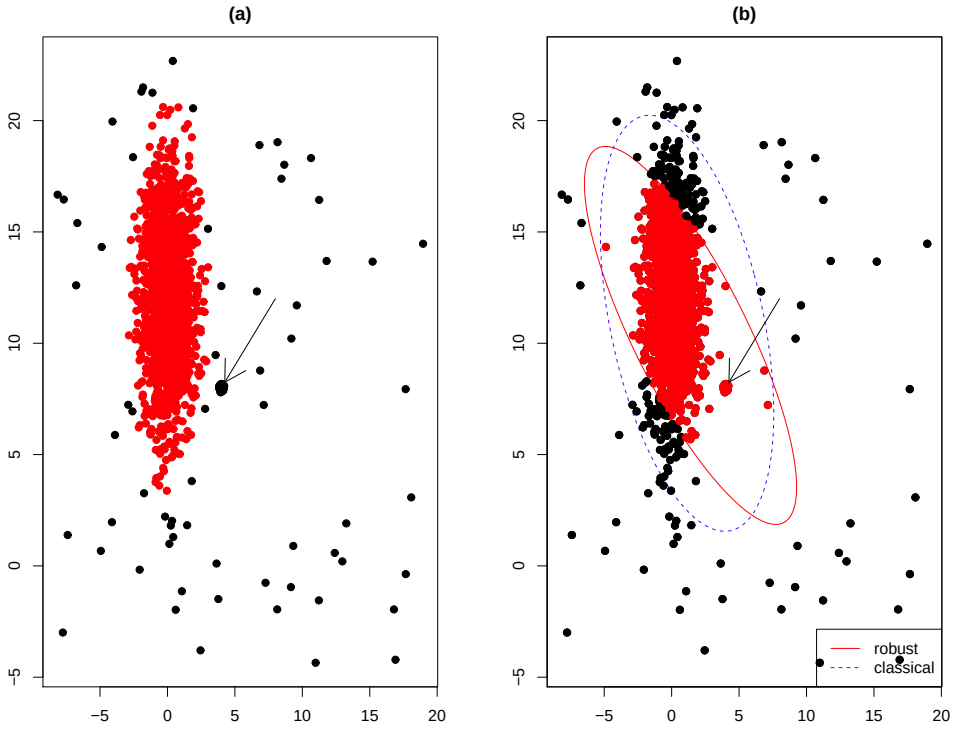

Figure 3.4: (a) The proposed iterative reweighting procedure when $k = 1$ started from $\alpha_0 = 0$ and $\alpha_L = 0.01$ (b) The (traditional) reweighted MCD started from $\alpha_0 = 0$ and $\alpha_L = 0.01$. Trimmed points are the black points.

Finally, an additional important parameter for the proposed methodology is α_L. In all the shown illustrative examples, the same $\alpha_L = 0.01$ has been take. The α_L parameter has to do with the quantile in the χ^2_p distribution and it plays the same role as in all analogous reweighting methods. For instance, $\alpha_L = 0.01$ means that around 1% of the observations are wrongly discarded when we have normal components without contamination. The smaller the α_L the lesser is the number of proportion of wrongly trimmed observations but higher is the risk of incorporating near outlying observations.

3.6 Theoretical results

The algorithm proposed in the previous section admits a population counterpart for a theoretical underlying probability P. Let us denote by

$$\theta_P^0 = (\pi_{1P}^0, \ldots, \pi_{kP}^0, 0, \mu_{1P}^0, \ldots, \mu_{kP}^0, \Sigma_{1P}^0, \ldots, \Sigma_{kP}^0)$$

the population parameters obtained when applying the TCLUST methodology to distribution P, for a fixed α_0 and c. This solution does exist under very mild assumptions (see Proposition 2 in García-Escudero et al. 2008). Note also that we are setting $\pi_{k+1P}^0 = 0$ and $\sum_{j=1}^{k} \pi_{1P}^j = 1$ given that we do not dispose of reliable estimators for the contamination level at this initial $l = 0$ stage. In a similar fashion, we use the notation

$$\theta_P^l = (\pi_{1P}^l, \ldots, \pi_{kP}^l, \pi_{k+1P}^l, \mu_{1P}^l, \ldots, \mu_{kP}^l, \Sigma_{1P}^l, \ldots, \Sigma_{kP}^l)$$

for the population values of the parameters obtained after applying l steps of the proposed algorithm when considering as fixed and known the underlying distribution P. A more formal definition of these population θ_P^l parameters is given in the Appendix A.1.

Let $\{x_1, ..., x_n\}$ be a realization of an independent identically distributed sample from distribution P and let P_n denote its associated empirical measure. We have that the $\theta_{P_n}^l$ parameters exactly coincide with those obtained from the algorithm presented in Section 3.3.

Next result shows that the parameters are bounded when considering a finite number of steps L under mild assumptions on P. In Theorem 1, the assumption concerning the noncoincidence of population centers at any iteration serves to exclude certain pathological cases in clustering problems. The proof of this result and other in this section appears in Appendix A.1.

Theorem 1. Assume P is an absolutely continuous distribution with a strictly positive density function. Additionally assume that $\mu_{j_1P}^l \neq \mu_{j_2P}^l$ for every $j_1 \neq j_2$ and every $l = 0, 1, ..., L$. We have that

$$\max_{j=1,\ldots,k;l=0,1,\ldots,L} \|\mu_{jP}^l\| < \infty$$

and

$$0 < \min_{j=1,\ldots,k;l=0,1,\ldots,L;q=1,\ldots,p} \lambda_q(\Sigma_{jP}^l) \leq \max_{j=1,\ldots,k;l=0,1,\ldots,L;q=1,\ldots,p} \lambda_q(\Sigma_{jP}^l) < \infty$$

where $\{\lambda_q(S)\}_{q=1}^p$ is the set of eigenvalues of matrix S.

As shown in Appendix A.1, the proof of the previous result relies on the fact that the optimal set (i.e., the set including all the non-trimmed regions in $\mathbb{R}^p$) can be represented as a union of k ellipsoids having all of them non-null P probability mass.

The following result needs the same assumptions as in Theorem 1 but notice that these assumptions only concern the underlying distribution P.

Theorem 2. Under the same assumptions of Theorem 1, we have that there exists a compact set K and $n_0 \in \mathbb{N}$ such that $\theta^l_{P_n} \in K$ for $n > n_0$ with probability 1.

Now, we can state a consistency result for the parameters obtained from random samples of size n from P toward those obtained from the population problem when n increases.

Theorem 3. Under the same assumptions of Theorem 1, we have

$$\theta^l_{P_n} \to \theta^l_P, \ P\text{-almost surely},$$

for every $l = 0, 1, ..., L$ when $n \to \infty$.

Another important issue is to analyze if this reweighting approach is able to retain the robustness properties of the TCLUST initializing method. In order to do that, we resort to the "addition r-components" breakdown-point (BP) notion, as given in Cuesta-Albertos et al. (2008*b*). This notion is a multivariate adaptation of a univariate proposal by Hennig (2004). It is easy to see that classical BP notions in clustering are sample-dependent and, then, they cannot be directly applied. The considered BP notion is based on the assumption that measuring the BP of a clustering procedure should require "well-clusterized" data set prior to contamination (compare Definition 6 in Chapter 1 and Ritter 2014). With this idea in mind, a sequence of data sets composed by groups with bounded "intra-group" variability and with "between-groups" distance going to infinity are considered for studying the change in the estimated parameters caused by the addition of r outliers. The "separation" of this r outlier should converge to infinity (a more precise formulation of these two concepts can be encountered in the aforementioned references). Under this ideal setting, the BP of a clustering procedure corresponds to the minimum r required in order that one of the estimated parameters breaks down (in the classical sense) of the estimation of any parameter in the model.

Theorem 4. Let $\mathcal{X}_n = \{x_1, ..., x_n\}$, $n \in \mathbb{N}$, be sequence of well-clustered data sets in $k \geq 2$ clusters and let $\mathcal{Y}_n = \{y_{1,n}, ..., y_{r,n}\}$, $n \in \mathbb{N}$, be sequence of r outliers with separations converging to infinity (see Cuesta-Albertos et al. (2008*b*) for a more precise

statement). If $r \leq [\alpha_0 n]$, then the obtained parameters from TCLUST with trimming level α_0 and the L subsequent parameters obtained throughout the described reweighting procedure do not break down by the addition of these r outliers.

It is important to note that, even under this very "ideal" clustering setting (well-clustered), Hennig (2004) proves that maximum likelihood estimator of a normal mixture model breaks down with the addition of only $r = 1$ outlier, as well as other robust proposals like maximum likelihood estimators of t-mixture models or the addition to the normal mixture of a uniformly distributed component in the convex hull defined by the data.

3.7 Simulation study

We now study the performance of the previously described procedure when applied to several (contaminated) mixtures of Gaussian distributions. Additionally, we detail how the data sets used in previous sections have been simulated in the illustrative examples.

The non-outlying part of the dataset comes from a mixture of two p-variate normal distributions $\pi_1 N(\mu_1, \Sigma_1) + \pi_2 N(\mu_2, \Sigma_2)$ with centers $\mu_1 = (0, 0, 0, ..., 0)'$ and $\mu_2 = (8, 0, ..., 0)'$ and covariance matrices

$$\Sigma_1 = I_p \text{ and } \Sigma_2 = \sqrt[p]{\lambda} \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 2 & 2 & \cdots & 2 \\ 1 & 2 & 3 & 3 & \cdots & 3 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & 3 & 4 & \cdots & p \end{pmatrix}.$$

This means that $|\Sigma_1| = 1$ and $|\Sigma_2| = \lambda$.

To generate outliers we fix an hypercube where each dimension includes the range of the non-contaminated data. Outlying observations are generated uniformly within this hypercube, but outliers with squared Mahalanobis distances from μ_1 and μ_2 (using Σ_1 and Σ_2) smaller than $\chi^2_{p,\nu}$ are discarded. The operation is repeated until the desired proportion of ε outliers have been obtained. The parameter ν controls how far away contaminated data points are.

We generate data sets of size $n = 1000$ under all possible combinations of the following scenarios:

- Three data dimensions: $p = 2, 4$ and 6

- Three contamination levels $\varepsilon = 0.10$, 0.05, and zero.
- Two scales $\lambda = 1$ and 5
- Balanced clusters $\pi_j = 0.5$ for $j = 1, 2$ and unbalanced clusters $\pi_1 = 0.4$ and $\pi_2 = 0.6$
- Two ν values, $\nu = 0.01$ and $\nu = 0.005$
- Two types of contamination: a symmetric one obtained sampling from a uniform distribution in the hypercube defined by the range of the non-contaminated part of the data and an asymmetric one obtained by sampling from a uniform distribution defined on $[-3, 0] \times [-7, -2] \times [-2, 2]^{p-2}$, which is closer to the second cluster than to the first.

The case $\varepsilon = 0$ is used to evaluate efficiency of the proposed methodology when applied to clean data.

Regarding the illustrative examples in Figure 3.1 we generated two datasets once from a bivariate normal distribution, fixing $\lambda = 20$, $\pi_1 = 0.4$ $\pi_2 = 0.6$, with symmetric contamination and $\nu = 0.01$. A contamination level $\varepsilon = 0$ was used in (a.1) and $\varepsilon = 0.10$ in (b.1).

We compare the performance of the following robust clustering proposals:

- `rtclust33` and `rtclust20`: The proposed iterative reweighting approach started from TCLUST with initial trimming levels $\alpha_0 = 0.33$ and $\alpha_0 = 0.2$
- `HR33` and `HR20`: a one-step version of the procedure by Hardin & Rocke (2004) started from TCLUST with initial trimming levels $\alpha_0 = 0.33$ and $\alpha_0 = 0.2$
- `HR-it33` and `HR-it20`: the iterated and adapted version of Hardin & Rocke (2004) started from TCLUST with initial trimming levels $\alpha_0 = 0.33$ and $\alpha_0 = 0.2$
- `tclust33`, `tclust20`, `tclust10` and `tclust05`: TCLUST with fixed trimming levels $\alpha_0 = 0.33, 0.2, 0.1$ and 0.05

The same value $\alpha_L = 0.01$ was used for RTCLUST and Hardin and Rocke's methods. For iterative procedures we fixed $L = 20$. The TCLUST procedure was included with with trimming levels which could be higher or the correct one. The same eigenvalue restriction factor $c = 12$ is always applied when using TCLUST (in the initialization of RTCLUST and in the direct application of TCLUST). Note

that $c = 12$ could be smaller or larger than the true eigenvalue ratio, depending on p and λ.

The Hardin and Rocke's methods are clustering algorithms based on the MCD philosophy. These methods are going to be initialized in this simulation study with exactly the same TCLUST robust clustering initial solution used for RTCLUST. Indeed Hardin & Rocke (2004) commented in their work that "any" robust clustering solution can be used and we have seen that TCLUST always provides quite sensible initial solutions for all the considered data sets in the simulation study. In fact, TCLUST always removes all noisy observations (together with others wrongly trimmed ones) with these high trimming levels ($\alpha_0 = 0.33$ and 0.2). Let $\mu_1^0, ..., \mu_k^0$, $\Sigma_1^0, ..., \Sigma_k^0$ and $H_0^0, H_1^0, ..., H_k^0$ being the solution obtained by applying the TCLUST method. The Hardin and Rocke's approach proposes cut-off values to declare outliers based on the approximation

$$\frac{k_j(m_j - p + 1)}{pm_j} d^2_{\Sigma_j^0}(x_i, \mu_j^0) \sim F_{p,m_j-p+1}, \tag{3.4}$$

where $k_j = \eta_{\beta_j}$ is a correction factor (as that used in Section 3.3) with

$$\beta_j = \tilde{h}_j/n_j \text{ for } \tilde{h}_j = \#H_j^0$$

and

$$n_j = \#\{x_i : d_{\Sigma_j^l}(x_i, m_j^l) = \min_{q=1,...,k} d_{\Sigma_q^l}(x_i, m_q^l)\}$$

and m_j is the approximated degrees of freedom for the associated Wishart distribution (see details in Hardin & Rocke (2004) and Hardin & Rocke (2005)). "HR33" and "HR20" apply directly the cut-off values in (3.4) to the observations in the H_j^0 sets while "HR-it33" and "HR-it20" refine these H_j^0 sets until stabilization by applying the iterative steps described in Section 3.3 of Hardin & Rocke (2004).

For all the 96 different data scenarios, we generated the data 500 times and evaluated the performance of the methods in terms of:

- Mean Square Error for estimation of the mean vectors μ_1 and μ_2, indicated in the plot with MSE_μ.
- Mean Square Errors associated to the logarithm of the eigenvalue ratio, indicated in the plots with MSE_Σ. We decided to report the error associated to this quantity since this ratio is forced in the initialization step to be smaller than a fixed constant $c = 12$ to avoid spurious maximizers. Nevertheless, as already commented, this is not necessarily the true eigenvalue ratio and we want to see how far the final estimated ratio is with respect the true one given

that the proper estimation of the cluster scales play a key role in the detection of outliers.

- The estimated contamination level $\hat{\varepsilon}$.
- *Swamping:* the proportion of non outlying observations that are trimmed
- *Masking:* the proportion of outliers that are not trimmed

Figures 3.5 and 3.6 summarize the simulation results obtained when $\varepsilon = 0.05$ and 0.1, respectively. Figures are separated in five row panels, one for each performance measure, and three column panels, one for each data dimensionality p. Given that there are several settings, in order to summarize the results in a concise way we do not distinguish among them further and just report the average performance measures all together. Note that some values exceed the scale of the plots, as identified by the upward triangle symbols.

The iterative reweighing procedure efficiently estimates the mean vector and the covariance matrix in every data scenario. In all cases we see small MSE values, and not much variability, meaning that results do not depend on the simulation setting considered. The MSE values are smaller than those obtained when applying TCLUST with large trimming values as 0.20 and 0.33. Moreover, MSE is even slightly better than what obtained with an oracle TCLUST whose trimming level is exactly equal to the true contamination level ε. This happens for two reasons. The first is that reweighting can adapt well to the positioning of the outliers, therefore flexibly trimming more or less as needed within each replicate. The second is that TCLUST is based on a sometimes wrong eigenvalue ratio constraint value $c = 12$. RTCLUST does not have further constraints and therefore can exceed this value when needed.

As far as estimation of the contamination level $\hat{\varepsilon}$ is concerned, RTCLUST provides very stable results in all simulation scenarios, with a systematic slight overestimation of ε. On the other hand, the procedures based on Hardin and Rocke approach may underestimate contamination levels in a remarkable way. The swamping proportion is small for all reweighting approaches but masking proportions can be very high in some scenarios with Hardin and Rocke's proposals. Underestimation of the contamination level is clearly more harmful than overestimation, as outliers included in the estimation set might break down the estimates. We believe that the problem with the Hardin and Rocke's approach is within the correction factor, which exploits an estimator of the fraction of observations in each cluster. The latter might not be resistant to outliers in our experience.

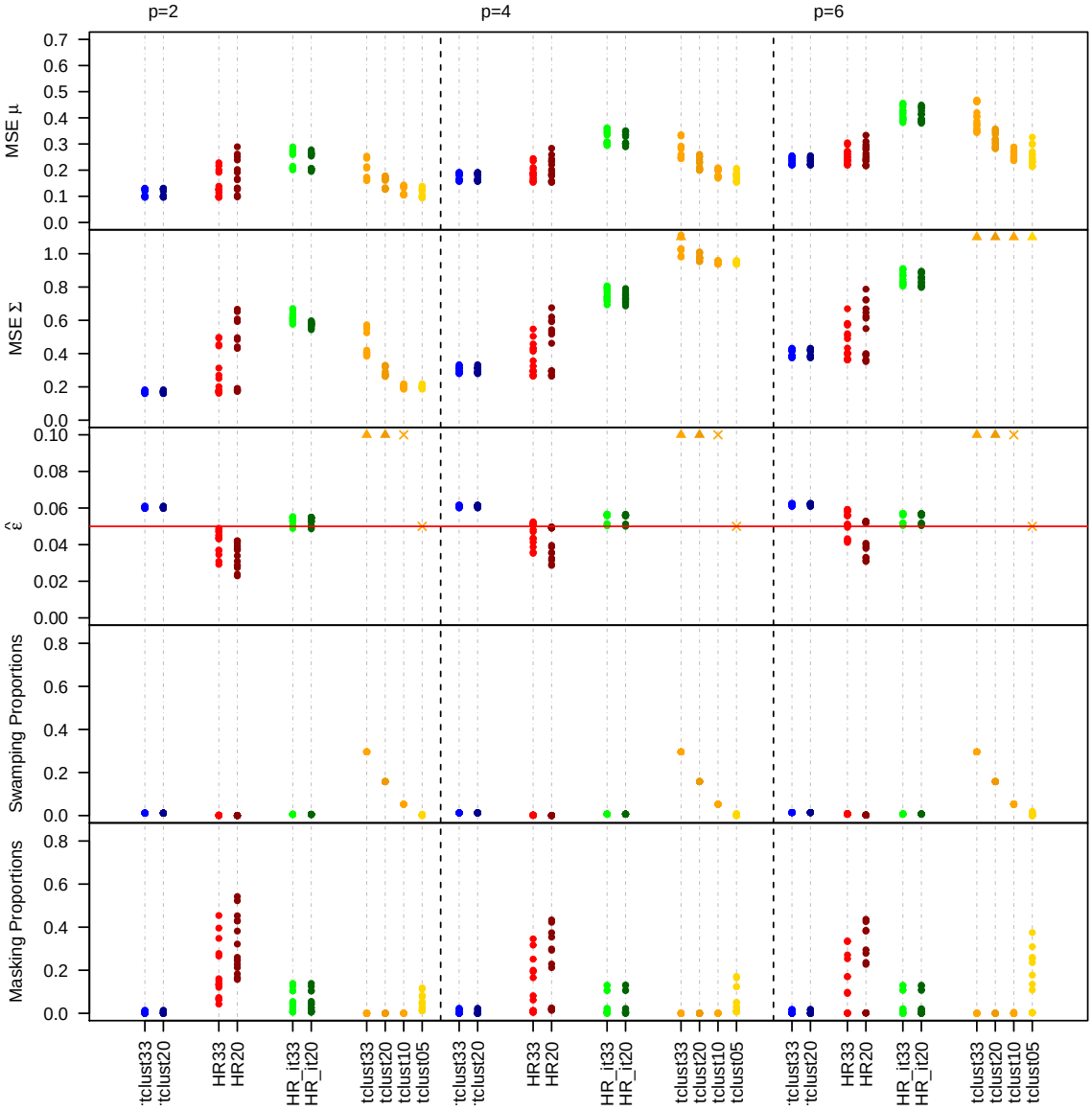

Figure 3.5: Results when $\varepsilon = 0.05$. Every procedure is labeled as explained in the text. Values appearing in the Figure that are fixed in advance (e.g the trimming level for the `tclust` method) are plotted with the symbol "$\times$" while when the considered value exceeds the scale of the plot we used a "$\triangle$"

We end this section by comparing the performance of these methods in the non contaminated $\varepsilon = 0$ case. This is reported in Figure 3.7.

We can see that the iteratively reweighting approach exhibits a very good performance in terms of providing small MSE values. We can also see that the (non-iterated) Hardin and Rocke's approaches are very competitive in this non-contaminated $\varepsilon = 0$ case. RTCLUST wrongly discards a limited proportion of observations, about 1%. This is not so surprising as $\alpha_L = 0.01$ in this section.

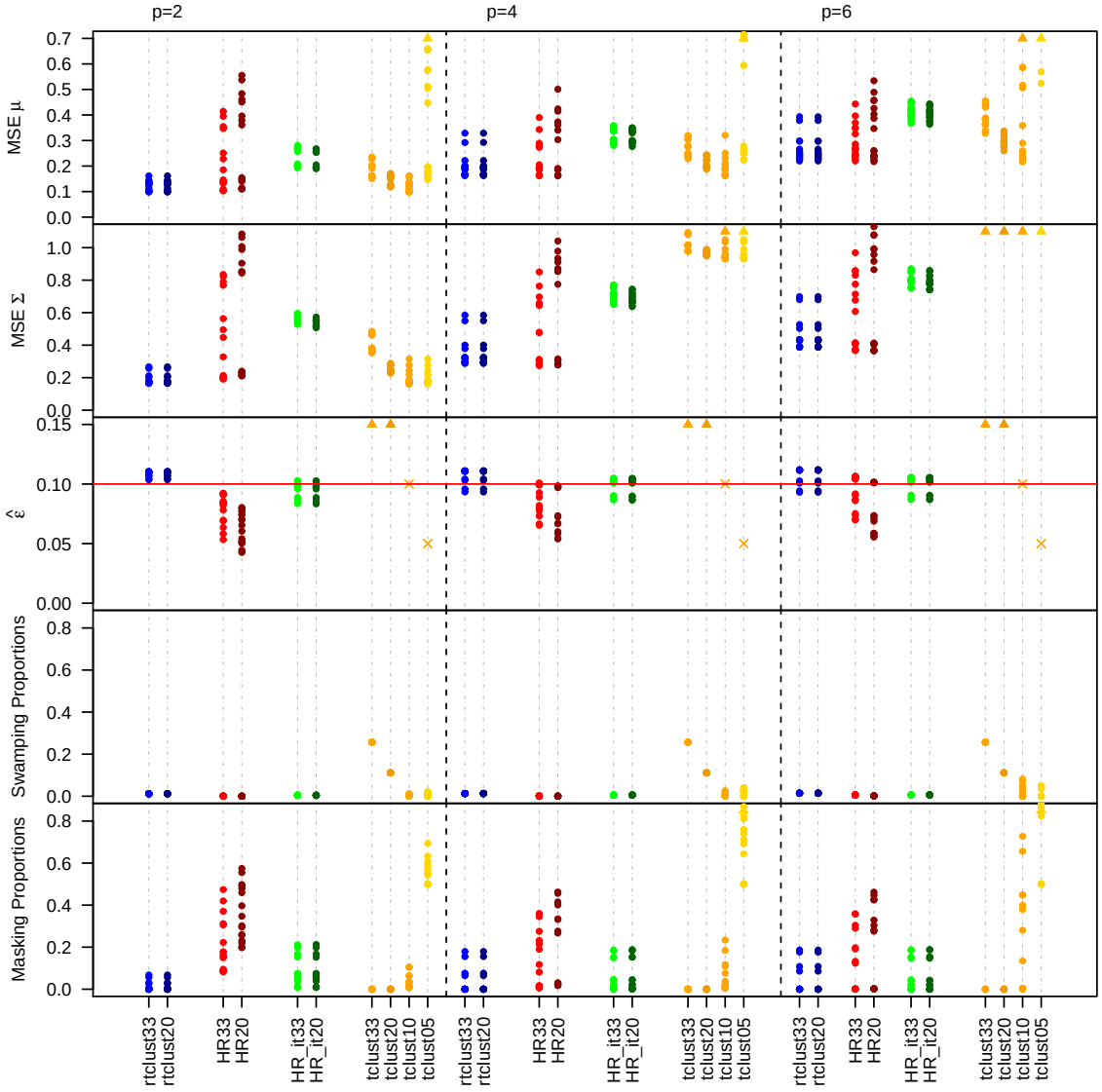

Figure 3.6: Results when $\varepsilon = 0.10$. Every procedure is labeled as explained in the text.

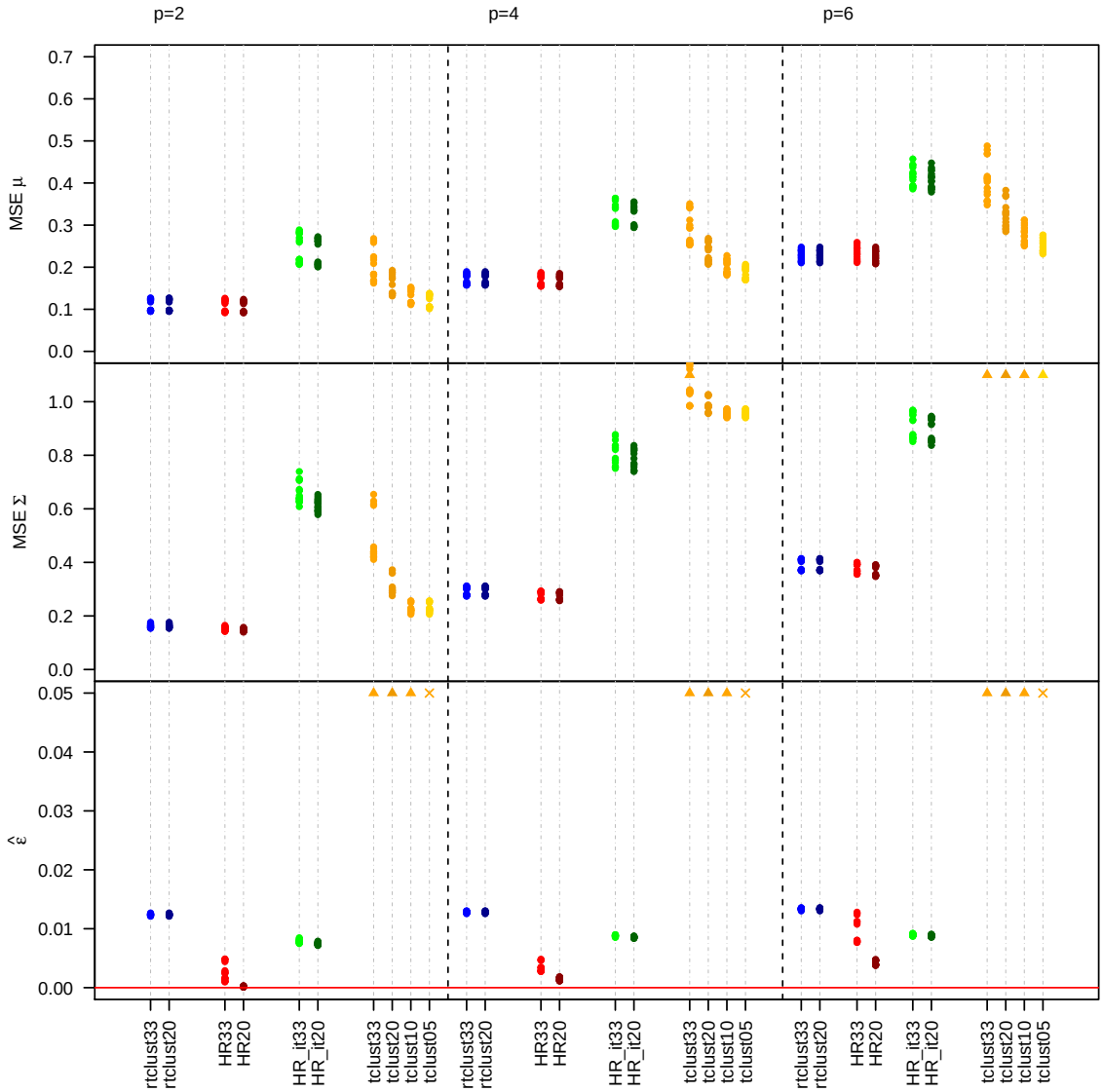

Figure 3.7: Simulation results study under no contamination ($\varepsilon = 0$).

Chapter 4

Extension of TCLUST to fuzzy linear clustering

4.1 Introduction

In this chapter we report the contribution proposed in Dotto et al. (2016*a*). As we now detail, the proposed method is a robust fuzzy linear clustering model. It shall be noticed that these type of models appear in the literature with different names (e.g switching regression models, linear clustering models or regression clustering models). Throughout this thesis we refer to such models using the term "linear clustering" in order to stress the fact that we aim to cluster our dataset around a linear structure.

Linear clustering models are based on searching k groups of units forming a linear structure. This implies that each unit is assigned to the group minimizing the regression error (i.e. its squared residuals from the estimated regression line). First attempts to fit $k = 2$ regression lines can be found in David & David (1974), that applied this type of procedure in economics, in Lenstra et al. (1982) where this type of procedures are applied in marketing segmentation, and in Späth (1982), where the details about a feasible algorithm are provided. In DeSarbo & Cron (1988) the EM algorithm has been used in this context and the methodology has been extended to the multidimensional case and for $k > 2$. The general linear clustering method could be then applied in many different research fields like medicine, psychology, biology, image reconstruction, and many others. See also Spiliopoulou et al. (2006) and Van Aelst et al. (2006) where such type of methodology has been extended, respectively, in a hierarchical clustering framework and for orthogonal regression clustering.

Our aim is to provide a fuzzy linear clustering method that is robust. To be more precise, we focus on extending the TCLUST approach. The TCLUST approach was adapted to be applied in fuzzy clustering problems in Fritz et al. (2013*a*). Another contribution where trimming is applied in fuzzy clustering to reach robustness can be found in D'Urso et al. (2015) and Kim et al. (1996). However, these proposals were not aimed at dealing with linear clusters. An extension of the TCLUST methodology for linear clustering problems appeared in García-Escudero, Gordaliza, Mayo-Iscar & San Martín (2010) and, following this idea, we are now interested in extending that methodology for performing robust linear fuzzy clustering.

Since we focus on model based clustering we will a formal maximum likelihood approach, as in Hathaway & Bezdek (1993), Wu et al. (2009) and Honda et al. (2008). A review of robust regression can be found for instance in Heritier et al. (2009) and Farcomeni & Ventura (2012). Robust methods for linear regression appeared for instance in Bai (2012). Methods for robustly estimating several unknown regression lines have appeared for instance in Ingrassia et al. (2014), McLachlan & Peel (2004), Yao & Li (2014). It shall be noted that fuzzy modeling is a framework which might seem somehow related to mixture modeling, but it is instead different in principles. In mixture models of regressions (e.g., DeSarbo & Cron 1988) a true underlying cluster label is always assumed to exist for each observation, and posterior probabilities summarize the researcher's uncertainty for this label. In fuzzy modeling, nonnegative membership values are assumed which may generate overlapping clusters where subjects may be shared among all clusters. Moreover, as we will see later, the proposed methodology in this work allows a kind of transition between "hard/crisp" and "fuzzy" clustering partition. The method can return a "core" of observations with 0-1 membership values and the remaining observations may belong to more than one cluster (i.e., membership values within the $(0, 1)$ interval). The degree of "fuzziness" is controlled throughout a tuning parameter $m \geq 1$, while it would make no sense to tune posterior probabilities. Fuzziness in clustering, that has been introduced in Ruspini (1969) and extended in Bezdek (2013), has several advantages in many applications. In some cases, e.g., Gustafson & Kessel (1978) or Ali et al. (2008), it is not possible to define meaningful hard partitions. See also D'Urso et al. (2011).

Our robust fuzzy linear clustering model can also be seen as an extension of the methodology introduced in Hathaway & Bezdek (1993). This last method is an adaptation of the fuzzy c-means algorithm (Bezdek 2013) for linear clustering problems and is based on minimization of the sum of the weighted distance of each point from the estimated regression line. The weights of the residual distance are given by the fuzzy membership values of each point to each cluster. This proposal is

not robust with respect to contamination and additionally can not take into account varying cluster weights. An alternative robust approach has been proposed in Wu et al. (2009) where an alternative measure of the distance of the residuals has been proposed. This method is indeed robust but it is seen in simulation to resist only to certain types of contamination.

The outline of the chapter is as follows. We provide the methodology in Section 4.2. We discuss in Section 4.3 the interpretation of the tuning parameters and, additionally, we give heuristics and an automatic method for choosing them. In Section 4.4 we report a simulation study.

4.2 Methodology and algorithm

4.2.1 Defining the problem

Let $\{(y_i, \mathbf{x}_i')\}_{i=1}^n \subset \mathbb{R}^{p+1}$ be a dataset where $\mathbf{x}_i \in \mathbb{R}^p$ are p explanatory variables and $y_i \in \mathbb{R}$ is a continuous response for the individual i. We are interested in grouping them into k clusters in a fuzzy way, and estimating a linear model within each group. Therefore, our aim is twofold: first of all we estimate a set of membership values $u_{ij} \in [0,1]$ for all $i = 1, ..., n$ and $j = 1, ..., k$, where a membership value 1 indicates that object i belongs at all to cluster j and conversely a membership value 0 indicates that object i does not belong to cluster j. Intermediate degrees of membership are obtained when $u_{ij} \in (0,1)$. We estimate the regression coefficients and the intercept parameters $\boldsymbol{b}_j \in \mathbb{R}^p$ and $b_j^0 \in \mathbb{R}$. Additionally we consider that an observation is fully trimmed if $u_{ij} = 0$ for all $j = 1, ..., k$ and, thus, this observation has no membership to any of the clusters. This is in contrast with the alternative robust approach in Wu et al. (2009), which sets $u_{ij} = 1/k$ for outliers.

Let $\alpha \in [0,1)$ be a fixed trimming proportion, $c \geq 1$ a fixed constant controlling ratio of cluster residual variances, $m \geq 1$ a fixed value of the fuzzifier parameter. A robust constrained fuzzy linear clustering problem can be defined as the task of maximizing the objective function

$$\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^m \log\left(f(y_i; \boldsymbol{x_i'b}_j + b_j^0, s_j^2)\right) \tag{4.1}$$

where $f(\cdot; \mu, \sigma^2)$ is the p.d.f of a normal distribution with mean μ and standard deviation σ, $f(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2}\exp(-(x-\mu)^2/(2\sigma^2))$. The membership values

$u_{ij} \geq 0$ are assumed to satisfy

$$\sum_{j=1}^{k} u_{ij} = 1 \text{ if } i \in \mathcal{I} \text{ and } \sum_{j=1}^{k} u_{ij} = 0 \text{ if } i \notin \mathcal{I},$$

for a subset

$$\mathcal{I} \subset \{1, 2, ..., n\} \text{ with } \#\mathcal{I} = [n(1-\alpha)],$$

and $s_1^2, \ldots, s_k^2$ are the residual variances which satisfy the constraint

$$\frac{\max_{j=1}^{k} s_j^2}{\min_{j=1}^{k} s_j^2} \leq c. \tag{4.2}$$

Note that $u_{i1} = ... = u_{ik} = 0$ for all $i \notin \mathcal{I}$, so these "trimmed" observations do not contribute to the objective function (4.1).

Constraint in (4.2) is needed as the target function (4.1) is unbounded otherwise. For instance, if we pick any x_i and take b_1^0 and $\boldsymbol{b}_1$ such that $y_i = b_1^0 + \boldsymbol{x}_i'\boldsymbol{b}_1$ then (4.1) tends to infinity whenever $u_{i1} = 1$ and $u_{l1} = 0$ for every $l \neq i$, just by taking $s_1^2 \to 0$.

The maximization of (4.1) assumes "a priori" that clusters have equal size and, thus, biases the procedure towards the detection of clusters with similar sizes (where the "size" of cluster j is seen in fuzzy clustering as $\sum_{i=1}^{n} u_{ij}$). To remove this assumption we might include clusters' weights p_j and replace (4.1) with this new target function

$$\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^m \log\left(p_j f(y_i; \boldsymbol{x_i'b}_j + b_j^0, s_j^2)\right), \tag{4.3}$$

where $p_j \in [0, 1]$ and $\sum_{j=1}^{k} p_j = 1$ are additional parameters. Conditionally on the membership values, these weights are optimally determined as

$$\hat{p}_j = \sum_{i=1}^{n} u_{ij}^m / \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^m. \tag{4.4}$$

This extra term so implies adding an "entropy regularization" term to the target function (4.1). This type of "entropy regularization" was discussed in Sadaaki & Masao (1997), see also Farcomeni (2014*a*).

4.2.2 Proposed algorithm

Maximization of (4.3) shall be performed through a constrained iterative procedure. We suggest to repeatedly initialize from several random starting points, and iterate two steps up to convergence or until a maximum number of iterations is reached. The

two updating steps are as follows: first, conditionally on current parameter values, membership values are obtained. Secondly, conditionally on current membership values, parameters are updated in order to maximize (4.3). Therefore we propose, in order to estimate the regression parameter in all the groups, the adaptation of an EM-type procedure (Dempster et al. (1977), McLachlan & Peel (2004)). Updating formulas are similar to those used in Hathaway & Bezdek (1993), Wu et al. (2009) and Honda et al. (2008). As the model might be viewed as an adaptation of Fritz et al. (2013*a*) for linear clustering, also the algorithm presents several analogies. algorithm iterates the following steps::

Algorithm 6.

1. Initialize randomly k initial regression parameters $\boldsymbol{b}_j$ and b_j^0 and k values $p_1^0, \ldots, p_k^0$ for the clusters' weights.

2. Compute the unconstrained residual variances

$$d_j^2 = \frac{\sum_{i=1}^n u_{ij}^m (y_i - b_j^0 - \boldsymbol{x_i'b_j})^2}{\sum_{i=1}^n u_{ij}^m} \tag{4.5}$$

   Apply the algorithm for constraining the variances, if required, to the quantities obtained in formula (4.5) to obtain the estimated residual variances s_j.

3. *Update membership values:* Using the current parameter estimates we update the membership values. If

$$\max_{q=1,\ldots,k} \{p_q f(y_i; \boldsymbol{x_i'b}_q + b_q^0, s_q^2)\} \geq 1,$$

   then we define "hard" assignments as

$$u_{ij} = I\big\{p_j f(y_i; \boldsymbol{x_i'b}_j + b_j^0, s_j^2) = \max_{q=1,\ldots,k} p_q f(y_i; \boldsymbol{x_i'b}_q + b_q^0, s_q^2)\big\}$$

   with $I\{\cdot\}$ being a 0-1 indicator function. If

$$\max_{q=1,\ldots,k} \big\{p_q f(y_i; \boldsymbol{x_i'b}_q + b_q^0, s_q^2)\big\} < 1,$$

   then we define "fuzzy" assignments as

$$u_{ij} = \left(\sum_{q=1}^k \left(\frac{\log(p_j f(y_i; \boldsymbol{x_i'b}_j + b_j^0, s_j^2)}{\log(p_j f(y_i; \boldsymbol{x_i'b}_q + b_q^0, s_q^2)} \right)^{\frac{1}{m-1}} \right)^{-1}.$$

4. *Trimming:* Compute

$$r_i = \sum_{j=1}^k u_{ij}^m \log(p_j f(y_i; \boldsymbol{x_i'b}_j + b_j^0, s_j^2)) \tag{4.6}$$

and sort them as $r_{(1)} \leq r_{(2)} \leq ... \leq r_{(n)}$. The membership values for the observations x_i with $r_i < r_{([n\alpha])}$ are fixed as $u_{ij} = 0$ and, thus, these observations are discarded at this stage of the algorithm.

5. *Update parameters:*

   - The cluster weights (if included) p_j are updated as (4.4)
   - For b_j^0 and $\boldsymbol{b}_j$, with $j = 1, 2, \ldots, k$, the usual (weighted) least square method is used. Closed forms are available as

$$\begin{aligned} \boldsymbol{b_j} &= \left(\frac{\sum_{i=1}^n u_{ij}^m \boldsymbol{x}_i \boldsymbol{x}_i'}{\sum_{i=1}^n u_{ij}^m} - \frac{\sum_{i=1}^n u_{ij}^m \boldsymbol{x}_i}{\sum_{i=1}^n u_{ij}^m} \cdot \frac{\sum_{i=1}^n u_{ij}^m \boldsymbol{x}_i'}{\sum_{i=1}^n u_{ij}^m} \right)^{-1} \\ &\quad \cdot \left(\frac{\sum_{i=1}^n u_{ij}^m y_i \boldsymbol{x}_i}{\sum_{i=1}^n u_{ij}^m} - \frac{\sum_{i=1}^n u_{ij}^m y_i}{\sum_{i=1}^n u_{ij}^m} \cdot \frac{\sum_{i=1}^n u_{ij}^m \boldsymbol{x}_i}{\sum_{i=1}^n u_{ij}^m} \right), \\ b_j^0 &= \frac{\sum_{i=1}^n u_{ij}^m y_i}{\sum_{i=1}^n u_{ij}^m} - \boldsymbol{b}_j' \frac{\sum_{i=1}^n u_{ij}^m \boldsymbol{x}_i}{\sum_{i=1}^n u_{ij}^m}. \end{aligned} \tag{4.7}$$

   - s_j^2, is updated by computing the initial unconstrained estimation d_j^2

$$d_j^2 = \frac{\sum_{i=1}^n u_{ij}^m (y_i - b_j^0 - \boldsymbol{x_i'} \boldsymbol{b_j})^2}{\sum_{i=1}^n u_{ij}^m} \tag{4.8}$$

     and then applying, if required, a suitable algorithm to impose the required constraint and obtain s_j^2, as we now detail.

Whenever the weighted sample residual variances d_j does not obey to the desired constraint (4.2) a similar procedure to that used in Fritz et al. (2013*a*) is then needed. In that case, let us consider the j-th residual variance component d_j^2 and its truncated value given by

$$[d_j^2]_t = \begin{cases} d_j^2 & \text{if } d_j^2 \in [t, ct] \\ t & \text{if } d_j^2 < t \\ ct & \text{if } d_j^2 > ct \end{cases}, \tag{4.9}$$

with t being a threshold value. These truncated residual components do satisfy the required scatter constraint. An optimal threshold value t_{opt} is obtained by taking into account the aim of maximizing the target function (4.3). It can be seen (see details in the Appendix) that t_{opt} is the value of t minimizing the real-valued function

$$t \mapsto \sum_{j=1}^k p_j \left(\log\left([d_j^2]_t\right) + \frac{d_j^2}{[d_j^2]_t} \right), \tag{4.10}$$

with p_j as given in (4.4). A closed way to get t_{opt} exists by evaluating (4.10) in $2k+1$ points (see, again, the Appendix section for details). Once this optimal threshold value is determined, the s_j^2 parameters are finally updated as $s_j^2 = [d_j^2]_{\text{opt}}$.

At each step of the algorithm the objective function (4.3) is increased. A more detailed justification of each step is provided in the Appendix.

4.3 Interpretation and choice of the tuning parameters

The methodology described was designed to be as general as possible. A drawback of this is that there are five choices to be done in advance: the number of clusters, k, the fuzzifier parameter m, the trimming level α, the bound on the ratios of residual variances c, and whether or not including cluster weights (that is, whether or not to shrink towards approximately balanced clusters).

Although "clustering" is clearly an "unsupervised learning" method, there is an increasing global consensus about the fact that it may not be a fully automatized task. For instance, the user is supposed to play an active role by specifying the type of clusters that he/she is particularly interested in (see, e.g., Hennig & Liao (2013)). As different choices of parameters yield very different clustering results, this choice must be guided by the final purpose of the analysis. Discussions on the role of tuning in robust clustering can be found in Coretto & Hennig (in press 2016).

Note also that parameters are clearly interrelated. For instance, a high trimming level α could delete smaller clusters and, thus, a smaller number of groups k may be needed. If we allow for higher cluster variabilities, then some observations, which may be otherwise considered as noise, may be included within the main clusters and a smaller k is so needed. Therefore, we do not think that a fully automatized way to fix simultaneously all these parameters is to be expected. However, we consider that some practical guidelines and helpful heuristic tools may be given in order to help the user to make this choice. Note also that fixing some of these parameters may be seen as a way to specify the type of clusters the user of the clustering method is actually interested in. This section is aimed at presenting the importance of each parameter choice and give some guidelines on how to make each decision in practice.

In order to do that, we resort to an illustration based on simple simulated datasets. We simulate two overlapped two-dimensional linear clusters. The first cluster is made of 144 observations. The explanatory variable X_{1i} is generated as a uniform distribution in the range $[0, 5]$, while the response variable is generated, for each $i = 1, \ldots, n$, as $y_i = 1 + 2x_{1i} + \varepsilon_{1i}$ where $\varepsilon_{1i} \sim \mathcal{N}(0, \sigma_1^2)$ and $\sigma_1 = 0.4$. The second cluster is made of 216 observations. The independent variable X_{2i} is generated from a uniform in $[0, 4]$ and the response variable as: $y_i = 10 - 1.5x_{2i} + \varepsilon_{2i}$

where $\varepsilon_{2i} \sim \mathcal{N}(0, \sigma_2^2)$ and $\sigma_2 = 0.6$. Additional noisy observations are added to our data set as needed.

It must be pointed out that information criteria (like i.e BIC or AIC) may be not applicable since these are monotone with respect to some of the tuning parameters (e.g. m and α). We nevertheless provide a method for providing a kind of automatic choice of the five tuning parameters simultaneously, or a subset of them, at the end of this section by resorting to a pseudo "cross-validation" criterium.

4.3.1 Including clusters' weights

Although this is properly not a tuning parameter, choosing to maximize likelihood (4.1) or (4.3) is an important decision to be made. As was commented in Section 4.2.1, the maximization of (4.1) assumes that clusters have equal size and, thus, biases the procedure towards the detection of clusters with similar 'sizes. In Figure 4.1 we represented a scenario where there are 40% of the clean observations in one cluster, 60% in the other one, and let us suppose that we search for $k = 3$ clusters. We compare the performance of the proposed procedure when the weights are kept into account in the likelihood function, like in equation (4.3), and when the weights do not appear in the objective function, like in (4.1). When the weights p_j are taken into account, we are able to recover the real structure of the data even though we wrongly set $k = 3$. Indeed, the cluster sizes obtained ($\sum_{i=1}^{n} u_{ij}$) are equal to 0.01, 0.42 and 0.57 (one of them is really close to 0) and two of the three estimated regression lines are overlapped. On the other hand, the cluster sizes obtained are 0.41, 0.30 and 0.29 when p_j weights are not used in the likelihood, which are clearly biased towards an equal balanced clusters scenario and two almost parallel clusters are recovered.

Therefore, if there are no particular interest in the detection of clusters with similar sizes, it might be useful to maximize (4.3) given that clusters with weights p_j close to 0 are obtained if a higher k than needed was wrongly chosen. In fact, this is the rationale behind the "classification trimmed likelihood curves" that will be presented in the following subsection.

4.3.2 Number of clusters

Parameter k obviously has to do with the number of clusters that the procedure initially looks for. The choice of the number of groups is one of the more difficult problems in cluster analysis. An underlying issue is the definition of what a cluster

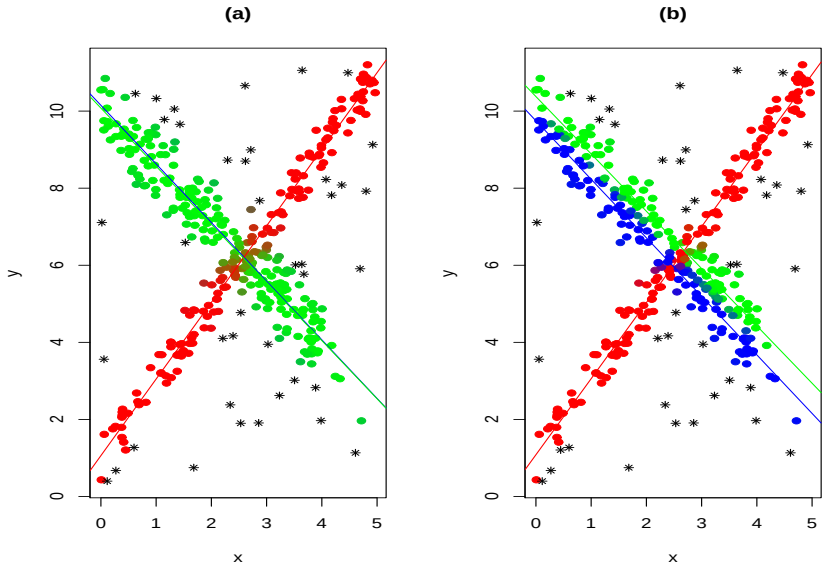

Figure 4.1: (a) Robust fuzzy clustering results when $k = 3$ and p_j are used within the objective function. (b) Results when $k = 3$ and p_j are not used within the objective function.

is. In this section we extend a heuristic tool based on monitoring the objective function (4.3) depending on the number of clusters k and the trimming proportion α, which has been proposed in García-Escudero et al. (2011) for non-fuzzy robust clustering. The "classification trimmed likelihood curves" (*ctlcurve* method) can be used to fix k and α, simultaneously once that parameters c and m are fixed in advance by the user.

In Figure 4.2,(b), we plot the objective function (4.3) with respect to different values of the trimming level α and number of groups k when $c = 5$ and $m = 1.5$. It is to easy to see that when 10% of observations are trimmed the objective function moderately increases as the number of clusters is increased. This plot leads to set $k = 2$ (which is the minimum number of clusters at which there is curve convergence with the one above), that is, in fact, the real number of linear components set in the data generation process. These curves also suggest a trimming level α around 0.1 (recall that the true contamination level was 10%) given that there are not noticeable improvements in (4.3) when increasing k at this point. Note that a higher k value is needed when, for instance, $\alpha = 0$ but there is no need to increase k when considering a trimming level higher than the "true" 10% contamination level.

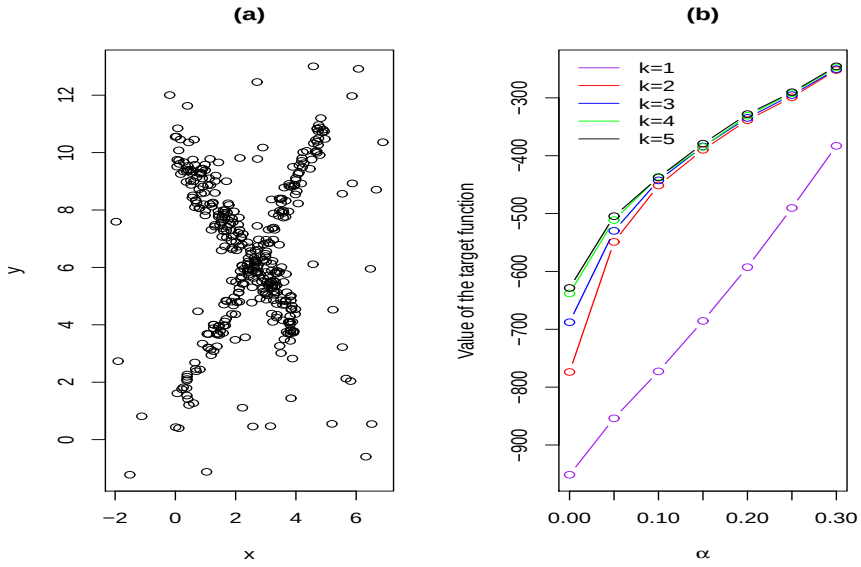

Figure 4.2: (a) A simulated dataset with two overlapped linear clusters and 10% of contaminated points. (b) The associated "classification trimmed likelihood curves" when $c = 5$ and $m = 1.5$.

4.3.3 Fuzzification Parameter

The fuzzifier parameter m in equations (4.1) and (4.3) takes values in the range $[1, +\infty)$ and it regulates the degree of fuzziness of the final clustering. Letting $m \to \infty$ implies equal membership values $u_{ij} = 1/k$ regardless of the data; while when $m = 1$ crispy weights $\{0, 1\}$ are always obtained and all observations are hard assigned to one and only one cluster. In fact, if we fix $m = 1$, our procedure reduces to the method in García-Escudero, Gordaliza, Mayo-Iscar & San Martín (2010), after the removal of the proposed second trimming step. Thus, the optimal value of m depends on the degree of overlap among clusters and on how much the researcher is prepared to accept and use fuzzy membership values.

It is also very important to take into account that the effect of a fuzzifier parameter $m > 1$, for the proposed methodology, is dependent on the measurement scale used for the response variable. In order to see that, we applied in Figure 4.3 our procedure to a simulated data set and having different scales for the response variable. We did so by multiplying the response variable by $s \in \mathbb{R}^+$ (i.e. y_i is replaced by $y_i \cdot s$) and for each scenario we chose two different values for the fuzzifier parameter: $m = 1.5$ (a quite standard choice for the fuzzifier parameter) and $m = 1$ which implies no fuzzification at all. In order to graphically represent the degree of fuzzification, we used a mixture of "red" and "green" colors with intensities proportional to the membership values of each observation. Additionally, throughout the paper,

points flagged as outlying under the model have been represented by "$*$". Figure 4.3 shows that the scale of the response variable leads to changes in the results when $m > 1$, while when $m = 1$ (hard clustering) results are scale independent.

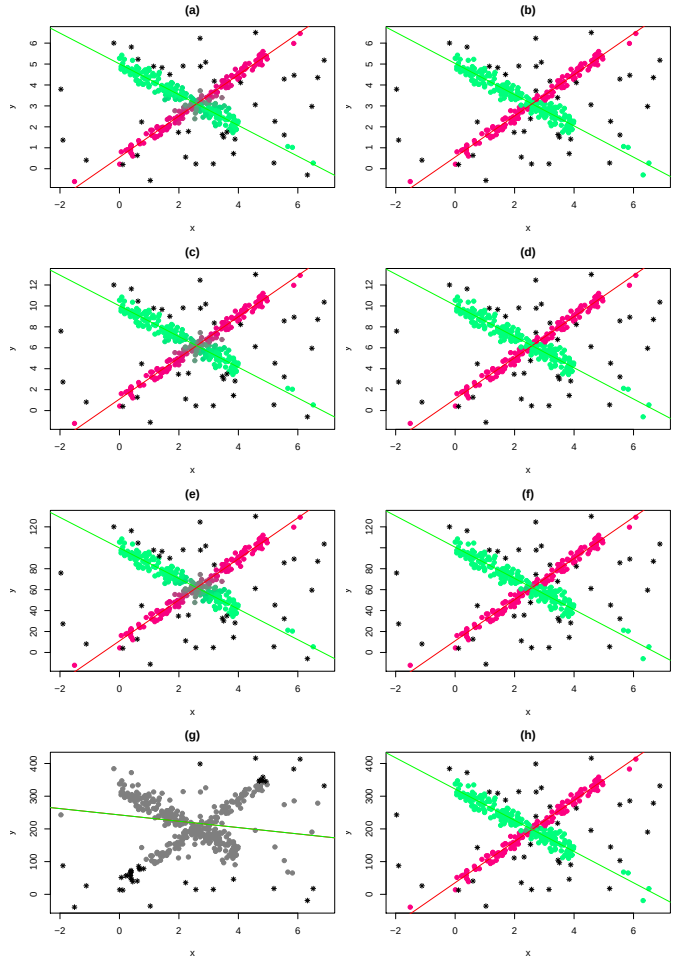

Figure 4.3: Different degrees of fuzzification obtained for different scale values s (y_i is replaced by $y_i \cdot s$). $m = 1.5$ and $s = 0.5$ in (a); $s = 1$ in (c); $s = 10$ in (e); $s = 32$ in (g). $m = 1$ (hard clustering) and $s = 0.5$ in (b); $s = 1$ in (d); $s = 10$ in (f); $s = 32$ in (h).

It is not difficult to see that this issue is a shared problem with others likelihood-based fuzzy clustering algorithms as Trauwaert et al. (1991), Rousseeuw et al. (1996) and Gath & Geva (1989). To our knowledge, this has been noted so far only in Fritz et al. (2013*a*). Note also that no additional problems appear due to the residual variance terms constraint as long as (4.2) is equivariant with respect to scale changes in the response variable.

Our proposal for choosing the m parameter is to monitor simultaneously the following two quantities: the proportion of hard assignments and the relative entropy of the fuzzy weights. The proportion of hard assignments (or approximately hard

assignments) has to do with the size of the cluster "cores" (i.e., the proportion of observations undoubtedly assigned to clusters). For certain applications, it is clearly interesting to have as higher as possible percentage of observations within these cores. The relative entropy measures residual uncertainty in cluster assignments, and it is proposed to be computed as

$$\sum_{j=1}^{k}\sum_{i=1}^{n} u_{ij}\log u_{ij} \Big/ [n(1-\alpha)]\log(k). \tag{4.11}$$

There is a clear trade-off between the proportion of hard assignments and the relative entropy. These two criteria cannot be simultaneously controlled by moving parameter m but the user can set m by considering a kind of "compromise" between these two opposite goals. The user may also change the response variable measurement scale in cases in which the proportion of hard assignments is basically constant with respect to m.

Figure 4.4 shows the proportion of hard assignments and the relative entropy as a function of m in our simulated example. We did so by repeatedly applying our procedure for different values of m. These plots suggest interesting m parameter values as those shown in Figure 4.3. These type of plots are useful to avoid extreme situations (very large degrees of fuzziness or zero proportions of hard assignments) and to explore the underlying degree of overlap.

4.3.4 Constraints on the residual variances

An important feature is that no cluster homoscedasticity assumption is made when $c > 1$. This is a novel feature as in many fuzzy switching regression models (see, e.g., Hathaway & Bezdek 1993, Wu et al. 2009) the residual variances are not kept into account and clusters are (implicitly or explicitly) assumed to be homoscedastic. In the "crisp" switching regression literature heteroscedasticity has already been considered in DeSarbo & Cron (1988) and Leisch (2006).

In order to give a brief illustration of how much clustering results might change by allowing for homoscedastic residuals, we compare in Figure 4.5 estimates based on $c = 1$ and $c = 5$ in two different heteroscedastic scenarios. To be more precise, we use the simulation scheme as before but with $\sigma_1 = 0.4$ and $\sigma_2 = 0.6$ in (a) and (b) and $\sigma_1 = 0.2$ and $\sigma_2 = 1$ in (c) and (d). In Figure 4.5,(a), the residual variances are correctly estimated when $c = 5$ and classification is very good as only 18 observations out of 360 are wrongly classified. On the other hand, in Figure 4.5,(b), we run the procedure on the same data set but after fixing $c = 1$. The estimated residual variances are now forced to be equal and 3 additional observations

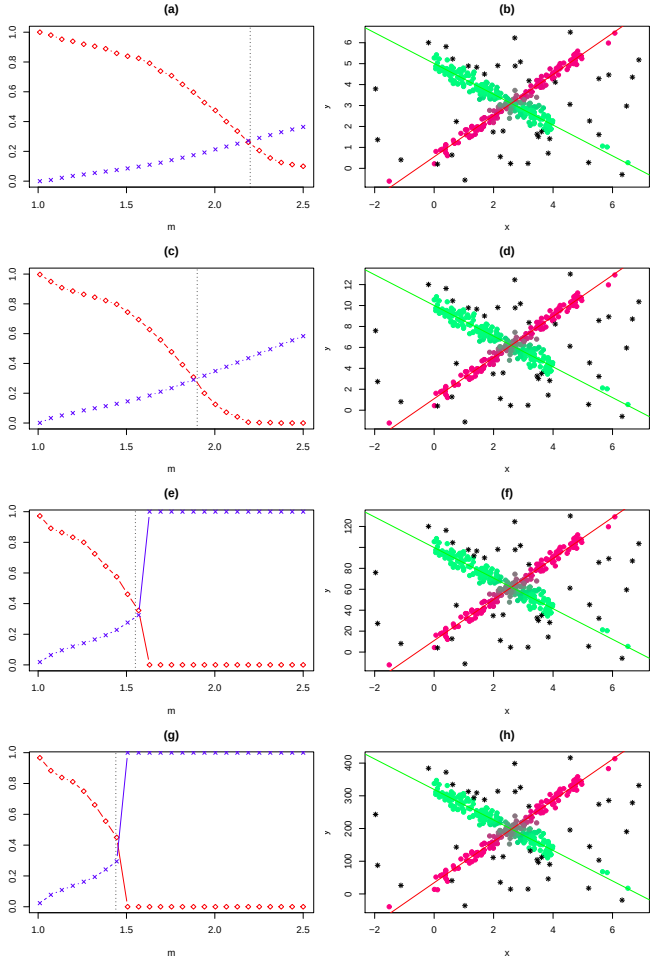

Figure 4.4: Left panels: relative entropy of the fuzzy weights, "×", proportion of hard assignments, "○", as a function of scale; (a) $s = 0.5$. (c) $s = 1$ (e) $s = 10$. (g) $s = 32$. Right panels: clustering obtained for specific values of m through (b) $s = 0.5$, $m = 2.2$. (d) $s = 1$, $m = 1.8$. (f) $s = 10$, $m = 1.6$. (h) $s = 32$, $m = 1.4$.

are misclassified. In panels (c) and (d), we repeated this experiment, but with an increased difference in the underlying variances. In Figure 4.5,(c) we still have 18 misclassified observations when $c = 5$ but in (d), where $c = 1$, we have 32 misclassified observations.

One would be tempted to set a large value for the constraint limit c, but too large values might be associated with spurious maximizers. Compare Chapter 2 for further details. In the following example, a set of spurious maximizers, we add a small proportion of almost collinear points(points that approximately lie on the same hyperplane) that form a spurious cluster (McLachlan & Peel 2000) . We have applied the proposed methodology with $k = 2$ and constraints on the residual variances, with $k = 2$ without constraints (almost unconstrained with $c = 10^{10}$) and with $k = 3$ and constraints. The results obtained can be seen in Figure 4.6.

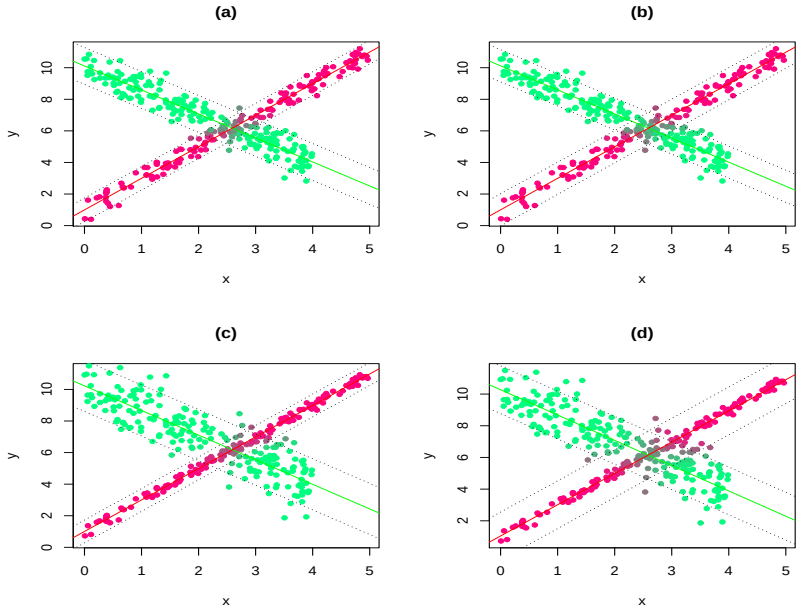

Figure 4.5: Estimated robust fuzzy clustering for different c values in two (less and more) heteroscedastic data sets. $c = 5$ is used in (a) and (d) and $c = 1$ in (b) and (d). The plotted bands are obtained by adding $\pm 2\widehat{s}_j$ to each fitted regression line.

We can see that these few almost collinear points are detected as a new additional cluster in Figure 4.6,(c). However, when $k = 2$, we clearly can see that the cluster partition shown in (a) is surely a more sensible one that the one shown in (b) where no constraints to avoid spurious solutions have been incorporated.

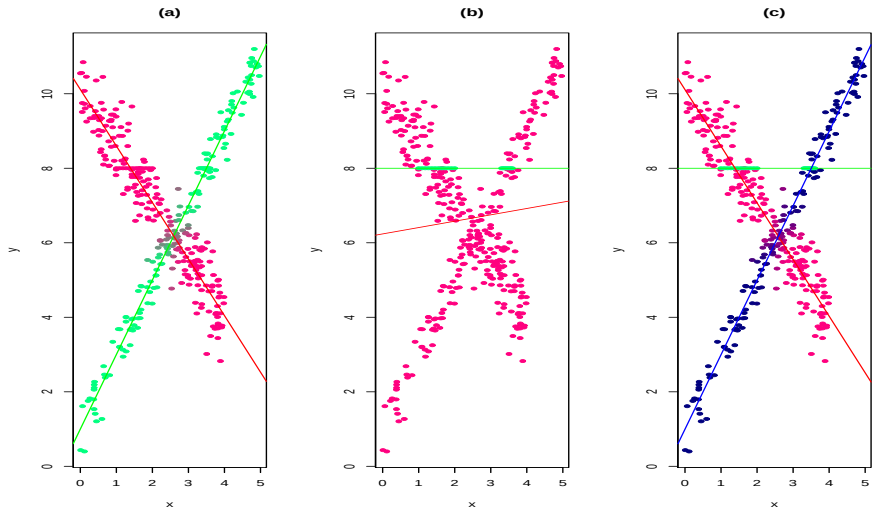

Figure 4.6: (a) FTCR with $c = 5$ and $k = 2$. (b) FTCR with $c = 10^{10}$ and $k = 2$. (c) FTCR with $c = 10^{10}$ and $k = 3$.

4.3.5 Trimming level

The issue of fixing a proper trimming level has been hugely discussed in Chapter 2, while in Chapter 3 we outlined a method based on reweighting which basically overcomes this problem. Since we now focus on a different problem we wish to outline some simple heuristic methods. The usage of *ctlcurves* as outlined above has proved to be intuitive and effective. An alternative heuristic approach is based on plotting $r_{[n(1-\alpha)]}$ against the trimming level α, where $r_{(1)} \leq r_{(2)} \leq ... \leq r_{(n)}$ are the sorted r_i values that were introduced in (4.6). These r_i values are the individual contributions of our n observations to (4.3). Given a candidate trimming level α_0, we apply the proposed methodology to obtain optimal regression and optimal membership values. We then plot the associated $\{\alpha, r_{[n(1-\alpha)]}\}$ curve. Our proposal is to consider α_0 as a sensible choice for the trimming level if this curve increases quickly when $\alpha < \alpha_0$ but slowly when $\alpha \geq \alpha_0$. The idea is that after all outliers have been removed, the individual contributions to the likelihood $r_{[n(1-\alpha)]}$ for two observations must be essentially the same when considering similar α values. We show these curves different candidate α_0 values ($\alpha = 0.2$, 0.1 and 0.05) in the left. Figure 4.7,(a), clearly shows that when $\alpha_0 = 0.2$ too many observations have been trimmed. This can be seen in panel (b) as the individual contributions to the likelihood have stabilized for much smaller values of α than $\alpha_0 = 0.2$. On the other hand, panel (e) clearly shows that not enough observations have been trimmed with $\alpha_0 = 0.05$. This can be also seen in panel (f) as the individual contributions are still increasing quickly when $\alpha = \alpha_0 = 0.05$ and, hence, there still are outliers available for trimming. Finally, $\alpha = 0.1$, the true underlying contamination level, is a fine choice.

4.3.6 Automatically choosing all parameters

In the proposed method many tuning parameters are involved, and the choices are interwined (that is, the optimal α depends on the chosen k, and so on). A user might not be willing to spend time exploring the data, or prior information might not be strong enough to guide the choice. We propose here a pseudo cross-validation method which we have found promising for choosing a good set of parameters. This will be demonstrated for some simulated data sets in this Section and in Section 4.4.

Cross-validation in unsupervised learning is slightly more complicated than the supervised framework, as no measurement of the target outcome is available. For a general discussion see Perry (2009) and references therein. In our framework, the user has no choice other than checking stability. We proceed as follows: we fix a

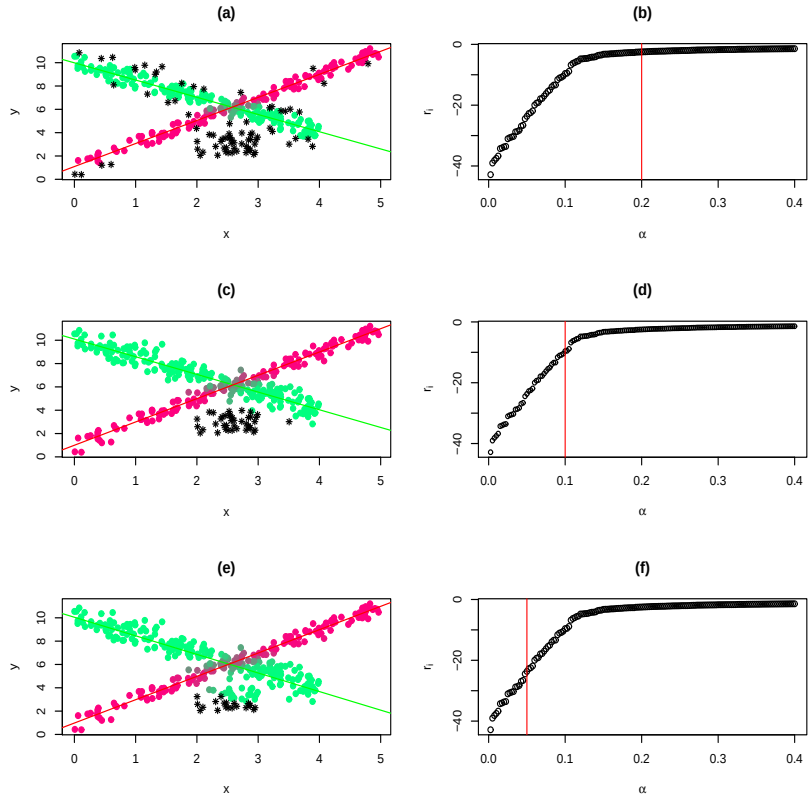

Figure 4.7: Left panels: Estimated linear clustering result for different trimming levels and $m = 1.5$. (a) $\alpha = 0.20$. (c) $\alpha = 0.10$. (e) $\alpha = 0.05$. Right panels: Average contribution to the likelihood for different values of α. A red line corresponds to the trimming level used on the corresponding left panel. (b): $\alpha = 0.20$. (d): $\alpha = 0.10$. (f): $\alpha = 0.05$.

grid of candidate tuning parameters. For each combination of tuning parameters, we randomly select a subset of observations (e.g., 50 or 75% of them), fit the model, and record the Euclidean distance between the estimated intercepts and slopes and the estimates based on the full data. We then repeat several (e.g., 100) times and use the mean or median Euclidean distance as a scoring rule. The Adjusted Rand Index (Hubert & Arabie 1985) can be also applied as stability measure. The "optimal" combination of parameters is that based on the minimal scoring rule.

To illustrate, consider the example in Figure 4.8. It can be seen that pseudo cross validation in this case was able to recover the true structure.

As was commented, different users may be interested in different clustering partitions depending on their final data-analysis purposes. We believe that we can surely find the best partition, depending on our clustering purposes, among a reduced list of partitions having the highest pseudo cross-validation indexes. I.e., it is recom-

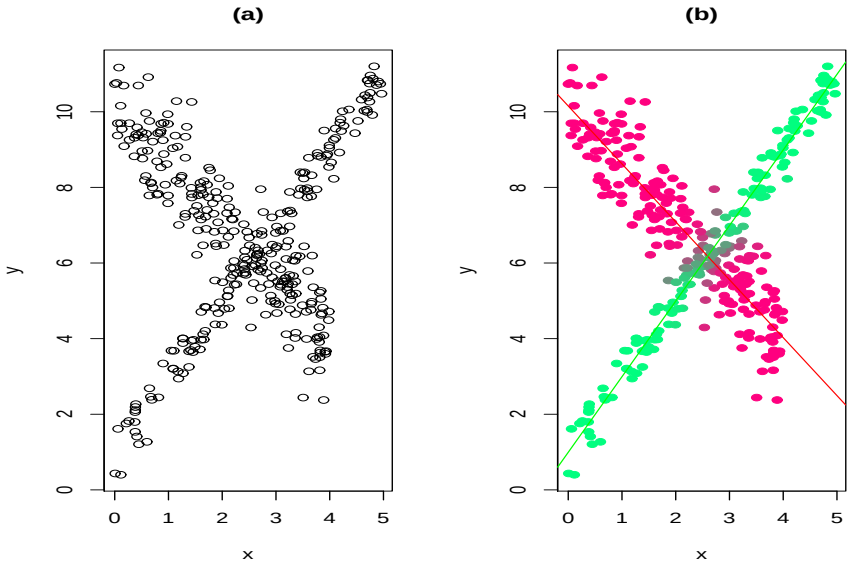

Figure 4.8: (a) The scatter plot of our dataset. (b) The results obtained using the tuning parameters chosen by cross-validation

mended to have a careful look at other stable partitions (not only the "optimal" one) by using the heuristic tools already presented in this Section. Finally, as a concluding remark, we would like to point out that a sensitivity analysis, obtained by varying the tuning parameters in a reasonable range, is always recommended regardless of how tuning parameters are chosen.

4.4 Simulation study

4.4.1 Settings and methods

In order to validate the proposed approach, a simulation study has been performed. We compared the proposed method (tagged FTCR throughout) with (i) the proposed method with no trimming, that is, $\alpha = 0$ with $c = 10^{10}$ and $m = 1.5$ (tagged EM throughout), (ii) the c-regression model of Hathaway & Bezdek (1993) (tagged cReg) and (iii) the alternative switching regression model of Wu et al. (2009), tagged A-cReg throughout. Two of these procedures (EM and cReg) are not designed to resist to contamination, while A-cReg and FTCR are formally robust. It is also important to note that the EM approach without constraints ($c = 10^{10}$) often provides very poor results, specially in higher p cases, because it might return partitions including clusters made up of few almost collinear observations (i.e., "spurious" clusters).

We generated data based on k =2 and 3 linear clusters and p =1, 2 and 3 covariates. For each setting we have four possible contamination schemes: (i) no contamination, (ii) uniform contamination, (iii) uniformly distributed background noise contamination and (iv) pointwise contamination. Contamination scheme (ii) corresponds to generating outliers from a uniform distribution with support within the range of the data (response and explanatory variables). Contamination scheme (iii) corresponds to the same, but with each dimension of the support brought farther from zero by two units when $p = 1, 2$ and five units when $p = 4$. Contamination scheme (iv) corresponds to generating outliers from a Gaussian distribution centered in a point $(\mathbf{x}, y)$, specified below, with standard deviation 0.1. This creates very concentrated "spherical" cluster of outliers, which do not follow a linear structure but might be influential. For each scenario we moreover compare homoscedastic and heteroscedastic underlying clusters.

The total number of settings is therefore 48. For each setting we generate data, estimate parameters based on the four procedures, and evaluate the Mean Squared Error (MSE) for slope and regression parameters (after matching through an increasing order for the slopes) and misclassification rate of observations. The results are averaged over $B = 500$ replicates.

We now outline for each combination of k and p how data was generated in more detail, and the results. Throughout tuning parameters are fixed at reasonable values given the data generating distribution, for all of the four procedures.For instance the trimming level α has been fixed equal to 0.10 while $m = 1.5$ and $c = 5$. It shall be noticed that in case of uncontaminated data we run our model overestimating the proportion of outliers. Nevertheless simulation's results have shown that the overestimation of the true contamination level lead us to moderate loss in terms of efficiency in the parameter estimation process. At the end of the section we very briefly discuss the performance of our automatic method for choosing tuning parameters with FTCR.

Setting S1: $p = 1$ and $k = 2$ where the data is generated as follows:

1. The first cluster is made of $n_1 = 144$ observations; the explanatory variable X_{11i} is distributed according to a uniform distribution with support in $(0, 5)$ and the regression model is $y_i = 1 + 2x_{1i} + \varepsilon_{1i}$.

2. The second cluster is made of $n_2 = 216$ observations; the explanatory variable X_{21i} is distributed according to a uniform distribution in $(0, 4)$ and the regression model is $y_i = 10 - 1.5x_{2i} + \varepsilon_{2i}$

3. 40 (that is to say an amount of 10%) contaminating points have been added as described previously. Pointwise contamination has been obtained from a Gaussian centered in $(x, y) = (-1.5, 17)$.

The errors ε_{1i} and ε_{2i} are zero-centered normals with standard deviation σ_1 and σ_2, respectively, where $\sigma_1 = 0.4$, and $\sigma_2 = 0.8$ in the heteroscedastic case, and $\sigma_1 = \sigma_2 = 0.4$ in the homoscedastic case.

In Figures 4.9 and 4.10 we report the MSE for slopes and intercepts, and misclassification rates, respectively.

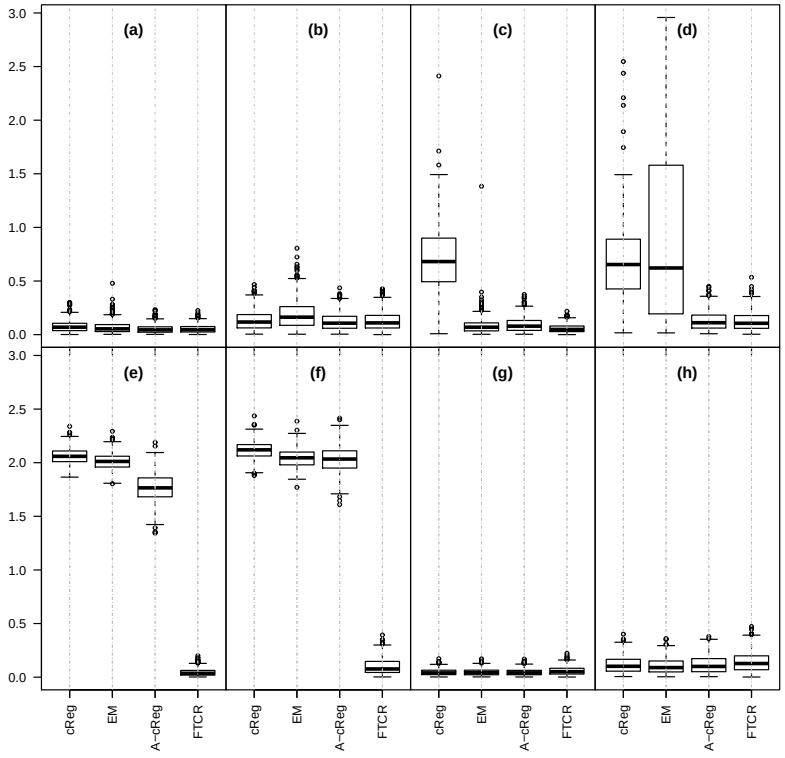

Figure 4.9: Simulation study. Boxplots representing the MSE of $\boldsymbol{b}_j$ and b_j^0 for setting S1: $p = 1$, $k = 2$. The Homoscedastic clusters are in (a),(c),(e),(g). Heteroscedastic clusters are in (b), (d), (f), (h). Uniform contamination is in (a) and (b). Inflated uniform contamination is in (c), and(d). Pointwise contamination in (e) and (f). Clean dataset is in (g) and (h)

Setting S2: $p = 2$ and $k = 2$ where the data is generated, with two covariates, as follows:

1. The first cluster is made of $n_1 = 144$.. The two covariates X_{11i} and X_{12i} are uniformly distributed in the range $(0, 5)$ and $(5, 9)$, respectively. The underlying regression model is $y_i = 3 + 4x_{11i} - 2x_{12i} + \varepsilon_{1i}$.

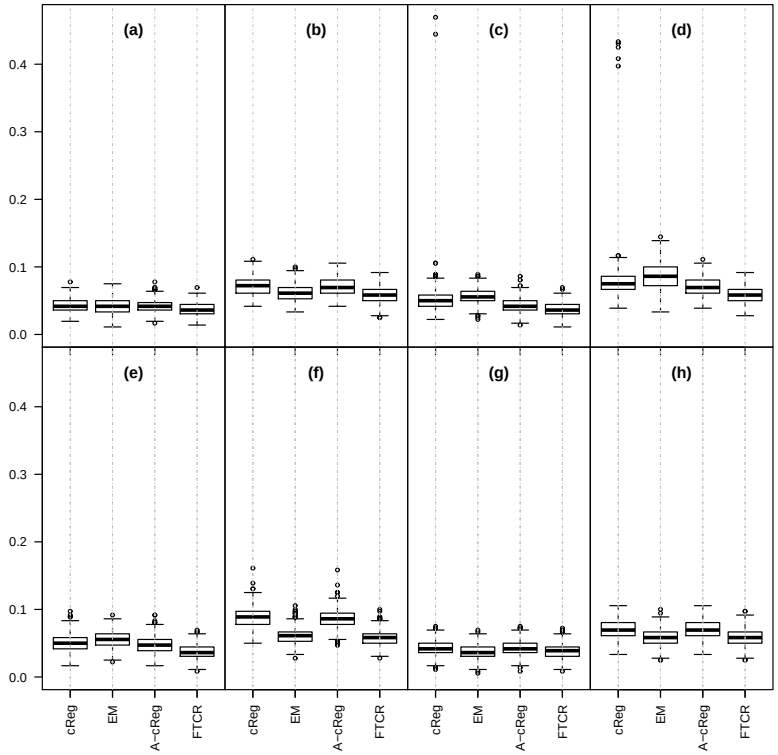

Figure 4.10: Simulation study. Misclassification error for setting S1: $p = 1$, $k = 2$. Legend as in Figure 4.9.

2. The second cluster is made of $n_2 = 216$ observations. The two covariates X_{21i} and X_{22i} are uniformly distributed in the range $(0, 6)$ and $(6, 9)$, respectively. The underlying regression model is $y_i = -2 - 2x_{21i} + 2x_{22i} + \varepsilon_{2i}$.

3. 40 contaminating points have been added as described previously. Pointwise contamination has been obtained by centering the Gaussian distribution on $(x_1, x_2, y) = (1, -1.5, 18.5)$.

The errors ε_{1i} and ε_{2i} are zero-centered normals with standard deviation σ_1 and σ_2, respectively, where $\sigma_1 = 0.4$, and $\sigma_2 = 0.8$ in the heteroscedastic case, and $\sigma_1 = \sigma_2 = 0.4$ in the homoscedastic case.

Results are reported in Figures 4.11 and 4.12.

Setting S3: $p = 4$ and $k = 2$ where the data is generated as follows:

1. The first cluster is made of $n_1 = 144$ observations. The four covariates $X_{11i}, X_{12i}, X_{13i}, X_{14i}$ are uniformly distributed in the range $(0, 5)$, $(5, 9)$, $(2, 7)$ and $(3, 8)$, respectively. The underlying linear model is $y_i = 3 + 2x_{11i} - 0.5x_{12i} + 2x_{13i} + 4x_{13i} + \varepsilon_{1i}$

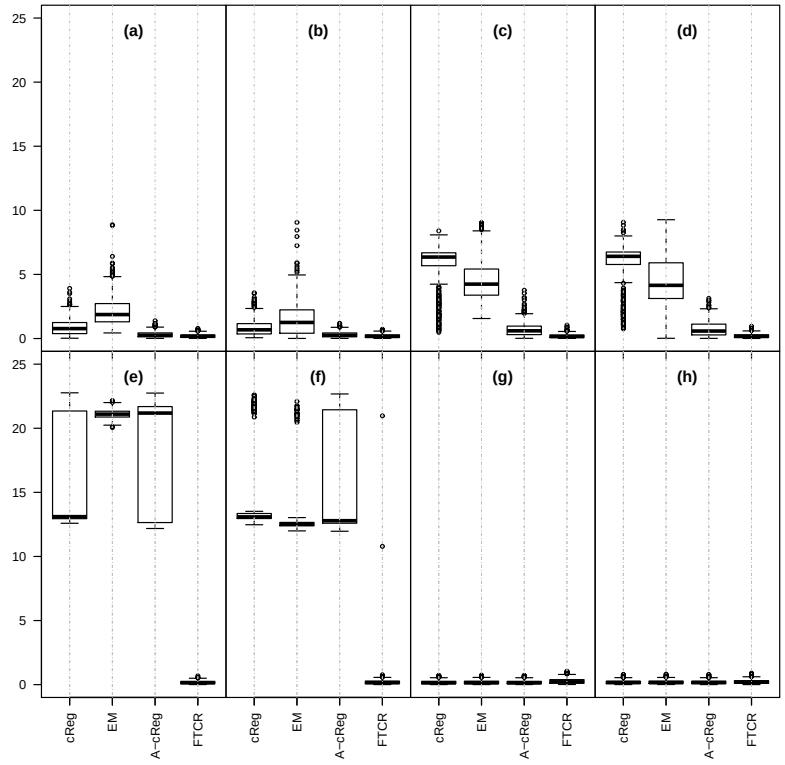

Figure 4.11: Simulation study. Boxplots representing the MSE of $\boldsymbol{b}_j$ and b_j^0 for setting S2: $p = 2$, $k = 2$. Same legend of Figure 4.9.

2. The second cluster is made of $n_2 = 216$ observations. The four covariates $X_{21i}, X_{22i}, X_{23i}, X_{24i}$ are uniformly distributed in the range $(0, 6)$, $(4, 12)$, $(0, 8)$ and $(1, 5)$, respectively. The underlying linear model is $y_i = 6 - 1.5x_{21i} - 0.1x_{22i} + 3x_{23i} + 6x_{24i} + \varepsilon_{2i}$.

3. 40 contaminated points are added following the schemes described in the previous section. Pointwise contamination is generated by centering the Gaussian distribution on $(x_1, x_2, x_3, x_4, y) = (3, 8, 4, 2.5, 9)$.

The error terms ε_i are zero-centered normal variables having standard deviation σ_i equal to 0.4 in case of homoscedastic clusters and to $\sigma_1 = 0.4$ and $\sigma_2 = 0.8$ in the heteroscedastic case.

Results are reported in Figures 4.13 and 4.14.

Setting S4: $p = 1$ and $k = 3$ where the data is generated as follows:

1. $n_1 = 150$ observations from the first cluster. A covariate X_{11i} is uniformly distributed in the range $(0, 5)$ and the underlying linear model is given by $y_i = 3 + 2x_{11i} + \varepsilon_{1i}$

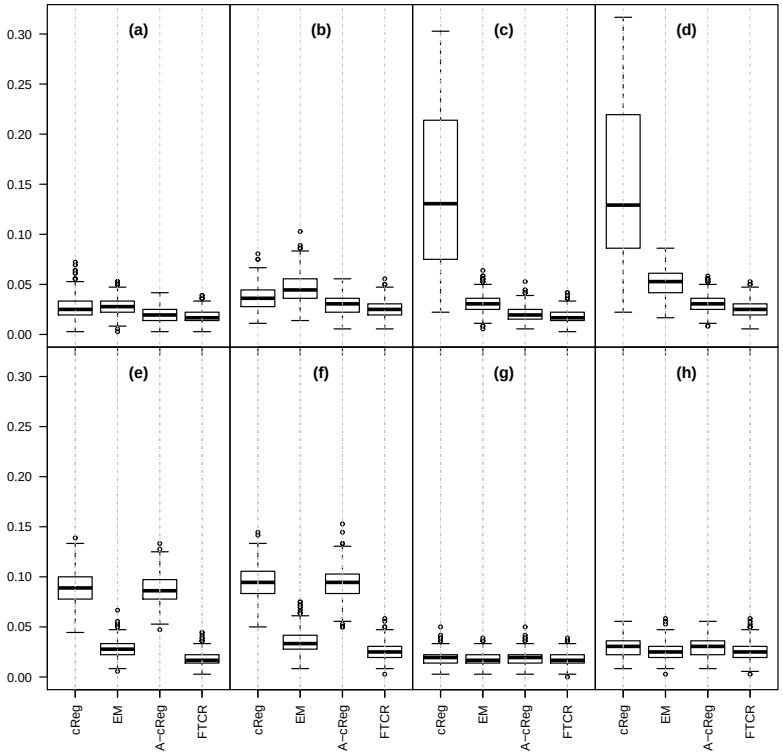

Figure 4.12: Simulation study. Misclassification error for setting S2: $p = 2$, $k = 2$. Legend as in Figure 4.9

2. The second cluster is made of $n_2 = 160$ observations and the covariate X_{21i} is uniformly distributed in the range $(0, 6)$. The underlying linear model is given by $y_i = 6 - 2x_{21i} + \varepsilon_{2i}$

3. The third cluster is made of $n_3 = 140$ observations and the covariate X_{31i} is uniformly distributed in the range $(0, 5)$. The underlying linear model is given by: $y_i = 5 + 4x_{31i} + \varepsilon_{3i}$

4. 50 contaminated points are added following the schemes described in the previous section. In the case of pointwise contamination we center the Gaussian distribution on $(x_1, y) = (-1.5, 20)$.

The errors ε_{1i}, ε_{2i} and ε_{3i} are zero-centered normals with standard deviations σ_1, σ_2, and σ_3, respectively; where $\sigma_1 = 0.5$, and $\sigma_2 = 0.6$ and $\sigma_3 = 0.4$ in the heteroscedastic case, and $\sigma_1 = \sigma_2 = \sigma_3 = 0.4$ in the homoscedastic case.

Results are reported in Figures 4.15 and 4.16.

Setting S5: $p = 2$ and $k = 3$ where the data is generated as follows:

1. The first cluster is made of $n_1 = 150$ observations. The two covariates

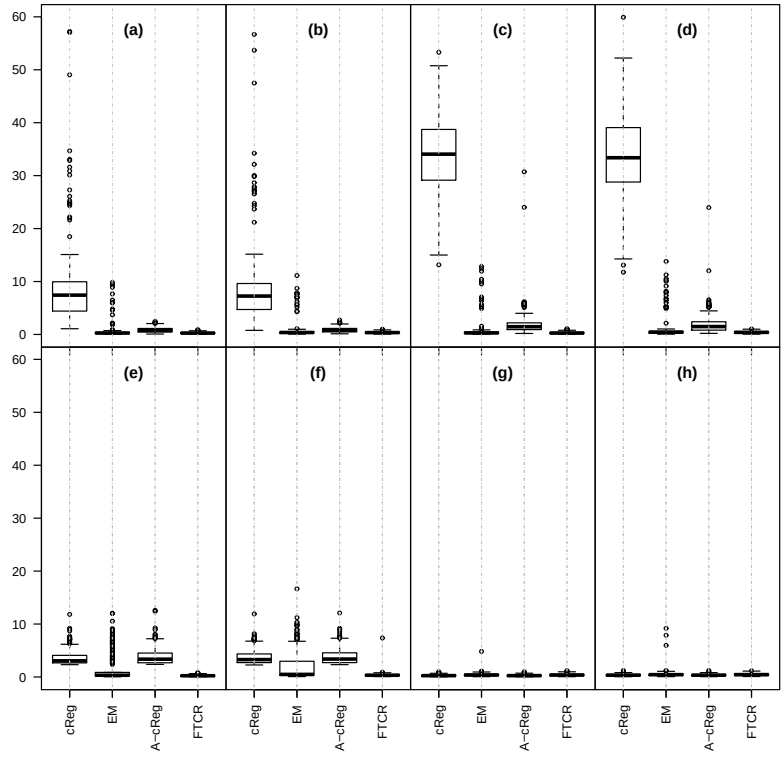

Figure 4.13: Simulation study. Boxplots representing the MSE of $\boldsymbol{b}_j$ and b_j^0 for setting S3: $p = 4$, $k = 2$. Same legend of Figure 4.9

X_{11i} and X_{12i} are uniformly distributed in the range $(0, 5)$ and $(5, 9)$, respectively. The underlying linear model is given by $y_i = 3 + 2x_{11i} - 0.5x_{21i} + \varepsilon_{1i}$

2. The second cluster is made of $n_2 = 160$ observations. The two covariates X_{21i} and X_{22i} are uniformly distributed in the range $(0, 6)$ and $(4, 12)$, respectively. The underlying linear model is given by $y_i = 6 - 2x_{21i} - 0.1x_{22i} + \varepsilon_{2i}$

3. The third cluster is made of $n_3 = 140$ observations and the two covariates X_{31i} and X_{32i} are identically distributed to X_{11i} and X_{12i}, respectively. The underlying linear model is given by: $y_i = 5 + 4x_{31i} - 2.5x_{32i} + \varepsilon_{3i}$

4. 50 contaminated points are added following the schemes described in the previous section. In the case of pointwise contamination the Gaussian is centered on $(x_1, x_2, y) = (-1, 1, 20)$.

The errors ε_{1i}, ε_{2i} and ε_{3i} are zero-centered normals with standard deviation σ_1 and σ_2, respectively, where $\sigma_1 = 0.5$, and $\sigma_2 = 0.6$ and $\sigma_3 = 0.6$ in the heteroscedastic case, and $\sigma_1 = \sigma_2 = \sigma_3 = 0.4$ in the homoscedastic case.

Results are reported in Figures 4.17 and 4.18.

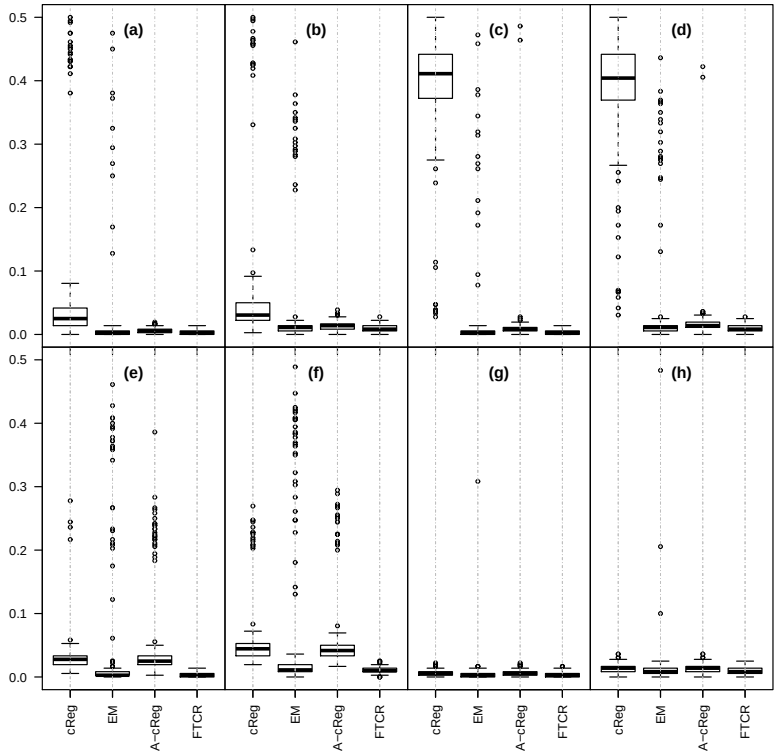

Figure 4.14: Simulation study. Misclassification error for setting S2: $p = 2$, $k = 2$. Legend as in Figure 4.9

Setting S6: $p = 4$ and $k = 3$ where the data is generated as follows:

1. The first cluster is made of $n_1 = 150$ observations. The four covariates $X_{11i}, X_{12i}, X_{13i}$ and X_{14i} are uniformly distributed in the range $(0, 5)$, $(5, 9)$, $(1, 7)$ and $(1, 5)$, respectively. The underlying linear model is $y_i = 3 + 2x_{11i} - 1x_{12i} + 2x_{13i} + x_{14i} + \varepsilon_{1i}$

2. The second cluster is made of $n_2 = 160$ observations. The covariates $X_{21i}, X_{22i}, X_{23i}$ and X_{24i} are uniformly distributed in the range $(0, 6)$, $(4, 9)$, $(0, 7)$ and $(1, 5)$, respectively. The underlying linear model is $y_i = 6 - 2x_{21i} + x_{22i} + x_{23i} + 6x_{24i} + \varepsilon_{2i}$

3. The third cluster is made of $n_3 = 140$ observations and the covariates $X_{31i}, X_{32i}, X_{33i}$ and X_{34i} are uniformly distributed in the range $(0, 6)$, $(5, 9)$, $(0, 8)$ and $(1, 5)$, respectively. The underlying linear model is given by $y_i = 8 + 4x_{31i} - 2.5x_{32i} + 2x_{33i} + 3x_{34i} + \varepsilon_{3i}$

4. 50 contaminated points are added following the schemes described in the previous section. Pointwise contamination is generated centered on $(x_1, x_2, x_3, x_4, y) = (1, 5, 2, 2, 18.5)$.

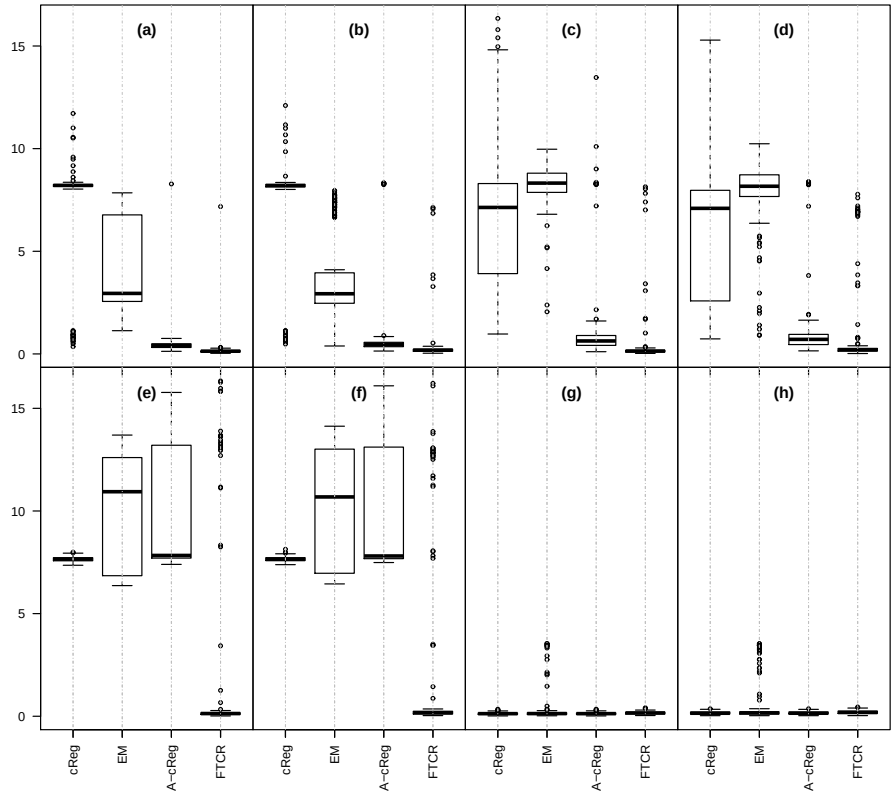

Figure 4.15: Simulation study. Boxplots representing the MSE of $\boldsymbol{b}_j$ and b_j^0 for setting S4: $p = 1$, $k = 3$. Legend as in Figure 4.9

The errors ε_{1i}, ε_{2i} and ε_{3i} are zero-centered normals with standard deviation σ_1 and σ_2, respectively, where $\sigma_1 = 0.4$, and $\sigma_2 = 0.5$ and $\sigma_2 = 0.3$ in the heteroscedastic case, and $\sigma_1 = \sigma_2 = \sigma_3 = 0.4$ in the homoscedastic case.

Results are reported in Figures 4.19 and 4.20.

4.4.2 Automatic choice of the tuning parameters

We give in this section a brief evaluation of our proposed resampling method for automatic choice of the tuning parameters. In Figure 4.21 we report the MSE for slopes and intercepts in two scenarios extrapolated from scenario S1. For comparison we report also the same for "oracle" FTCR with fixed tuning parameters, and the three other methods. It can be seen that the MSE with automatically chosen tuning is slightly definitely comparable with that of the oracle FTCR. We therefore deem pseudo cross-validation as promising.

Quite similar results (not reported for reasons of space) are reported also in other scenarios.

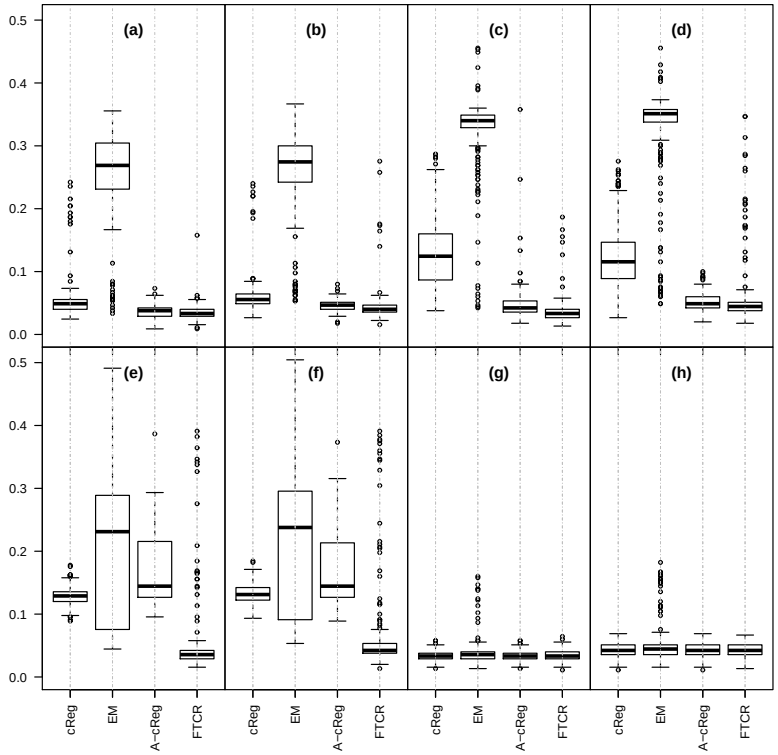

Figure 4.16: Simulation study. Misclassification error for setting S4: $p = 1$, $k = 3$. Legend as in Figure 4.9

4.4.3 Comments on the results of the simulation study

In all scenarios it can be seen that the procedures have more or less the same performance under no contamination, with the EM being only slightly better than the other three and FTCR being only slightly worse than the other ones. This is the loss of efficiency which is expected for any robust procedure, and it is in our opinion very reasonable. On the other hand, in contaminated scenarios non-robust procedures break down, showing very large MSE values (especially in scenarios (e) and (f), that are the pointwise contaminated scenarios) and high variability in performance. We speculate the superior performance of FTCR in pointwise contaminated scenarios is due to the fact that jittered clustered outliers might be wrongly detected as a spurious cluster by other methods, or might not be sufficiently downweighted. Trimming is able to completely remove the influence of outliers.

As a matter of fact, FTCR is mostly unaffected by contamination and it shows by far the best MSE and misclassification rates in presence of outliers. In setting *S6* with pointwise contamination (panels (e) and (f) of Figures 4.19 and 4.20) not much difference is seen among the procedures due to the curse of dimensionality and lack of clear true underlying structures. Nevertheless, FTCR seems to be left-skewed,

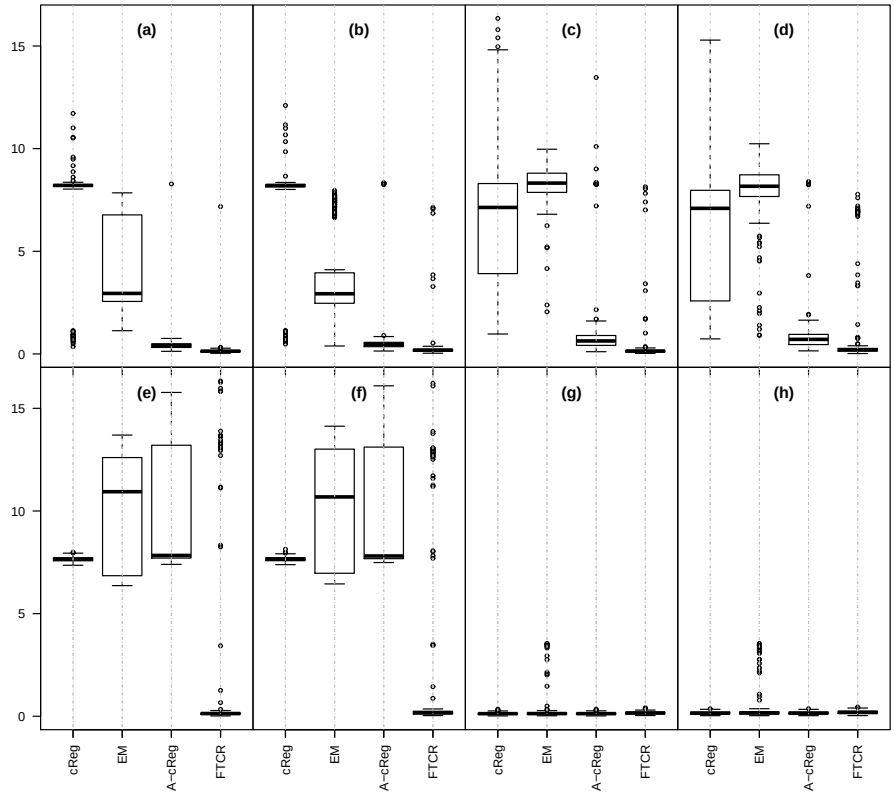

Figure 4.17: Simulation study. Boxplots representing the MSE of $\boldsymbol{b}_j$ and b_j^0 for setting S5: $p = 2$, $k = 3$. Legend as in Figure 4.9

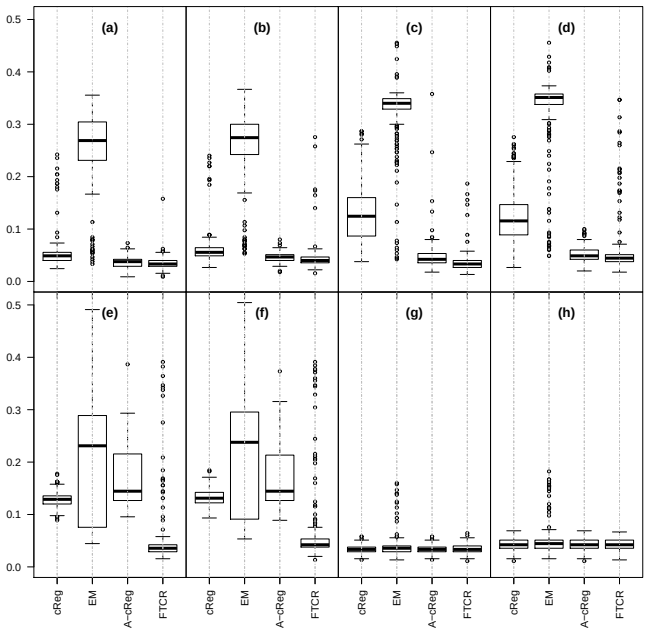

Figure 4.18: Simulation study. Misclassification error for setting S5: $p = 2$, $k = 3$. Legend as in Figure 4.9 .

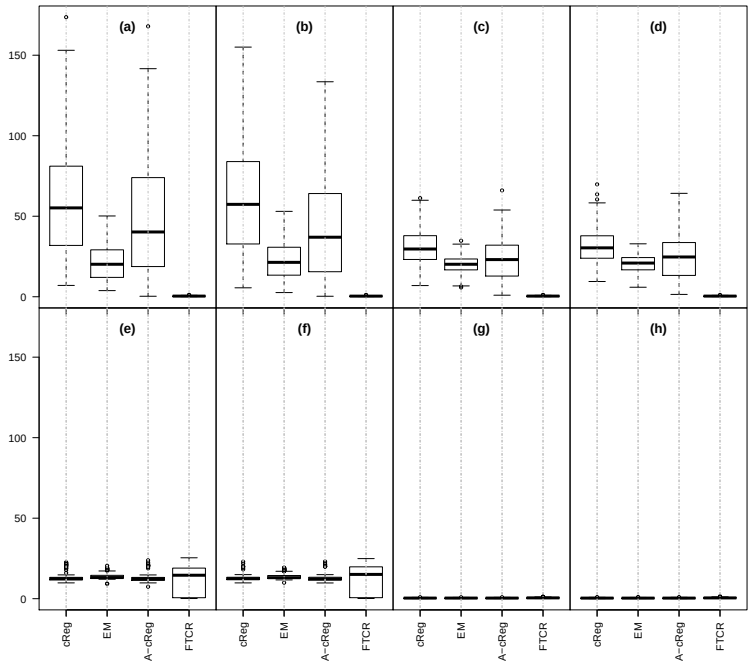

Figure 4.19: Simulation study. Boxplots representing the MSE of $\boldsymbol{b}_j$ and b_j^0 for setting S6: $p = 4$, $k = 3$. Same legend of Figure 4.9

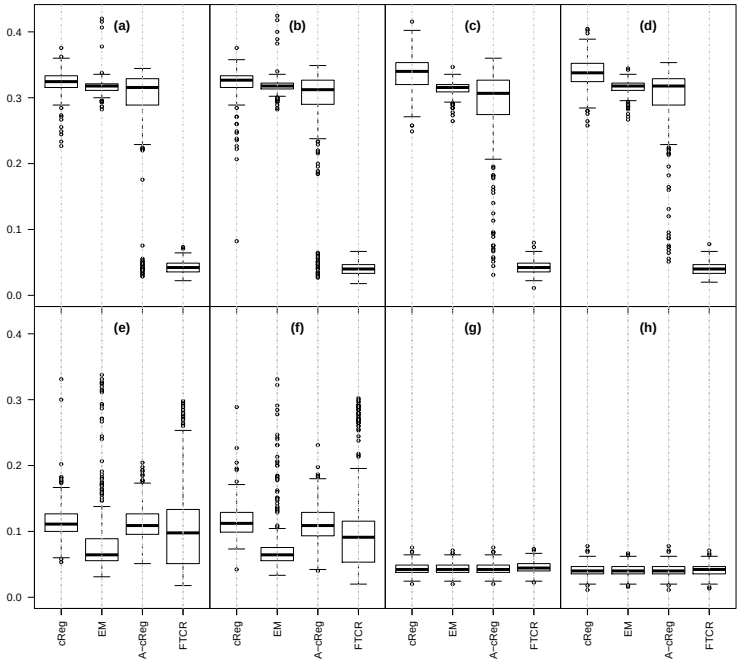

Figure 4.20: Simulation study. Misclassification error for setting S6: $p = 4$, $k = 3$. Legend as in Figure 4.9.

therefore still dominating the other procedures in many of the simulated scenarios.

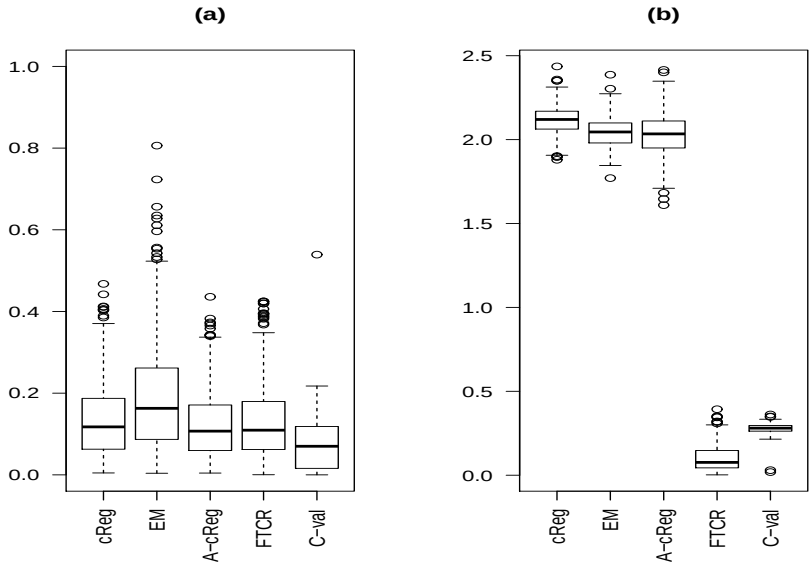

Figure 4.21: Boxplots with Mean Square Error for tuned and crossvalidated model, with competitors for comparison. C-val denotes FTCR with automatically chosen tuning. (a) Two Homoscedastic clusters uniformly contaminated, $p = 1$ covariate (Setting S1). (b) Two Heteroscedastic clusters with pointwise contamination, $p = 1$ covariate (Setting S1).

Chapter 5

Real data examples

5.1 Introduction

Within this chapter we apply the proposed contributions to real datasets. The outline of the chapter is as follows.
In section 5.2.1 we report the application of the RTCLUST methodology to the "Swiss Bank Note" dataset as already done in Dotto et al. (2016*b*). Such dataset, provided within many different R packages (`tclust` by García-Escudero et al. 2008 and `mclust` by Fraley & Raftery (2012) among the others), has been widely used to illustrate other clustering proposals (both robust and non robust). For the sake of brevity we just report the results obtained applying the TCLUST method, which also served as initialization for applying the RTCLUST methodology, and the results obtained after the reweighting process
In Section 5.2.2 we report the analysis appeared in Dotto et al. (2016*b*) on a dataset provided by the GALLUP Organization. This is a novel dataset ans to our knowledge no applications have appeared yet.
In Section 5.3 we report the application of the proposed fuzzy linear clustering model to allometry data (Dotto et al. 2016*a*). As a comparison we also applied all the methods used in the simulation study and compare the results.

5.2 Applications of reweighted TCLUST

5.2.1 Swiss Bank Notes

In this Section we apply the proposed iterative reweighting approach to the 6-dimensional "Swiss Bank Notes" data set presented in Flury & Riedwyl (1988) which describes certain features of 200 Swiss 1000-franc bank notes divided in two groups: 100 genuine and 100 counterfeit notes. This is a well known benchmark data set. In Flury & Riedwyl (1988), it is pointed out that the group of forged bills is not homogeneous since 15 observations arise from a different pattern and are, for that reason, outliers. Figure 5.1,(a) shows a scatterplot of the fourth ("Distance of the inner frame to lower border") against the sixth variable ("Length of the diagonal") with the classification of bills given in Flury & Riedwyl (1988) by using symbols "`G`" for the genuine bills and "`F`" for the forged ones. The previously commented 15 "anomalous" forged bills are surrounded by circles in this graph. Figure 5.1,(b) shows the results of applying TCLUST with a high trimming level $\alpha_0 = 0.33$ and $c = 12$. We can see that the 15 outlying points are successfully discarded and observations in the "cores" of the genuine and forged bills groups are correctly found. However, due to the use of this high trimming level, many observations are also discarded apart from the 15 clear outliers. We have surrounded these "probably wrongly" trimmed observations by square symbols. Finally, Figure 5.1,(c) shows the results of applying the proposed iterative trimming approach starting from the TCLUST's solution in (b) with $\alpha_L = 0.001$. We can see that the proportion of "probably wrongly" trimmed observations reduces to 4 (also surrounded by square symbols). One of these 4 observations is a genuine bill which clearly exhibits certain anomalous behavior in these two plotted variables and we could also see that the other 3 (wrongly) trimmed observations analogously seems to exhibit slight deviations in some of the (non-plotted) variables.

We have used a smaller $\alpha_L = 0.001$ value in this real data example. If $\alpha_L = 0.01$ then 7 wrongly trimmed observations (instead of 4) are obtained. As stated in the introduction, RTCLUST is not an outlier detection method. Estimates of the clusters location and scatter matrices do not change notably with the choice of α_L, which makes RTCLUST a good choice for robust clustering and parameters estimation for this data set. Formal rules for outlier detection could be then based on RTCLUST robustly estimated parameters.
As a final comment to the analysis we also report the table containing the confidence intervals for the estimated clusters' centroids obtained by the TCLUST with $\alpha_0 =$ .33 and for the estimation provided by the reweighted TCLUST. It can be seen that

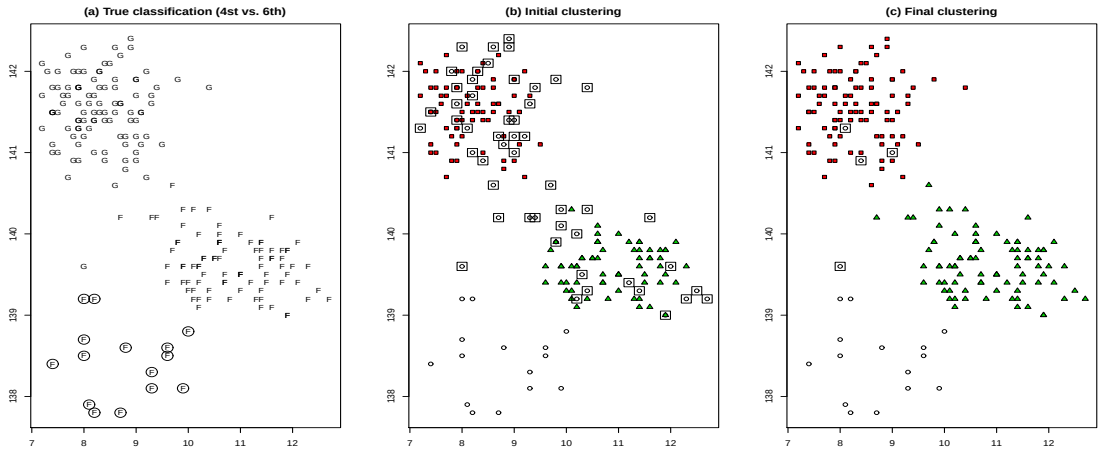

Figure 5.1: Fourth against the sixth variable of the Swiss Bank Notes data set. (a) `G` stands for genuine bills, `F` for forged ones and 15 bills listed in Flury & Riedwyl (1988) as anomalous ones are surrounded by ∘ symbols. (b) The initial TCLUST solution with $\alpha_0 = 0.33$ (c) Final solution when applying the proposed iterative approach. Trimmed observations not coinciding with those in Flury and Riedwyl's list are surrounded by □ symbols.

in most of the cases the confidence intervals associated to the RTCLUST estimation are narrower which mean that more precise estimations are available.

Table 5.1: 99% simultaneous confidence intervals for $\hat{\mu}$ provided by the TCLUST and the RTCLUST

	TCLUST				RTCLUST			
	C-1		C-2		C-1		C-2	
Variable	lower	upper	lower	upper	lower	upper	lower	upper
Length	214.78	215.15	214.64	214.90	214.83	215.15	214.65	214.91
Ht Left	129.76	130.09	130.13	130.38	129.77	130.08	130.15	130.39
Ht Right	129.53	129.83	130.05	130.27	129.57	129.85	130.04	130.33
IF Lower	7.87	8.51	10.49	11.30	8.01	8.59	10.44	11.25
IF Upper	9.98	10.50	10.82	11.46	9.87	10.44	10.80	11.40
Diagonal	141.31	141.71	139.44	139.75	141.37	141.73	139.47	139.79

We conclude with an analysis based on $k = 1$. As half of the bank notes are genuine ones, one could think that setting $k = 1$ and trimming 50% of the observations would identify them. Use of TCLUST with $k = 1$, $\alpha_0 = 0.5$ and $c = 12$ (which is the default value of c fixed in the `tclust` package in Fritz et al. 2012*b*) successfully identifies 96 genuine bills (out of the 100 non-trimmed observations). The standard application of RTCLUST, started from this TCLUST solution with $\alpha_0 = .5$

and $\alpha_L = 0.001$, returns a final set with 102 notes which includes 98 genuine bills. Therefore, RTCLUST is well-suited to discover, in an automatized way, the genuine observations. On the other hand, use of MCD through the well-known `robustbase` package with $\alpha = 0.5$ returns 103 bills (i.e., the largest integer less than or equal to $(n+p+1)/2$ as the "best" subset found and used for computing the raw estimates. Surprisingly, only 42 out of these 103 observations are genuine ones. Additionally, things become even worse when applying the default consistency correction factor for the covariance matrix estimation and the use of (3.1) with $\alpha_L = 0.025$, as this finally leads to 176 notes used for robust estimation.

5.2.2 Food Security Data

In this section we apply the proposed procedure to an original and very recent data set on an investigation of the status of food insecurity in the world in 2014. Food security is defined by the Committee on World Food Security of United Nations as when people

> *at all times, have physical, social and economic access to sufficient safe and nutritious food that meets their dietary needs and food preferences for an active and healthy life.*

For reviews see Godfray et al. (2010) and Jones et al. (2013).

In 2014, the Gallup Organization conducted a World Poll based on a questionnaire given to a representative sample of about 1000 adults from each of several areas in the world. Areas mostly correspond to countries, while in some cases countries have been split in different areas (e.g., Congo has been split in two, Brazzaville and Kinshasa areas). The Gallup World Poll (GWP) answers are then routinely summarized by Gallup into thematic indeces, which are evaluated for each polled subject and could be used to make comparisons across countries. A detailed description of the GWP can be found at `http://www.gallupworldpoll.com/content/24046/About.aspx`. In 2014 the usual GWP questionnaire has been augmented with eight questions, in partnership with the Voices of the Hungry (VoH) project of the Food and Agriculture Organization (FAO) of the United Nations. These questions were aimed at evaluating specifically a new index, the Food Insecurity Experience Scale (FIES). A very challenging issue that has been tackled by the VoH team is the standardization of the FIES score over different cultures and languages. Details on how this was performed are given in Cafiero et al. (2016). A more general discussion is provided in Ballard et al. (2013), Cafiero et al. (2014).

We have obtained the individual standardized FIES scores, in addition to the rest of GWP data for 2014. Data have been aggregated at country level, taking sampling weights into account. Our aim is to cluster and identify outlying countries, and secondly to evaluate the discriminating power of FIES after taking into account information collected by the other indices. Our final data set, aggregated over subjects, is therefore made of $n = 127$ countries and $p = 6$ indeces. These are Food Insecurity Experience Scale, Civic Engagement Index, Struggling Index, Food Security Index, Corruption Index, Youth Development Index. The aim of each index is rather self-explanatory from its name. Details can be found in Gallup (2015) and on the GWP website.

In order to explore the number of groups we use the `ctlcurves` of García-Escudero et al. (2011), which for different values of k show the log-likelihood at convergence of TCLUST, as a function of α and k. They can be used to determine both the number of groups and the optimal trimming level. The `ctlcurve` for the FIES data is reported in Figure 5.2.

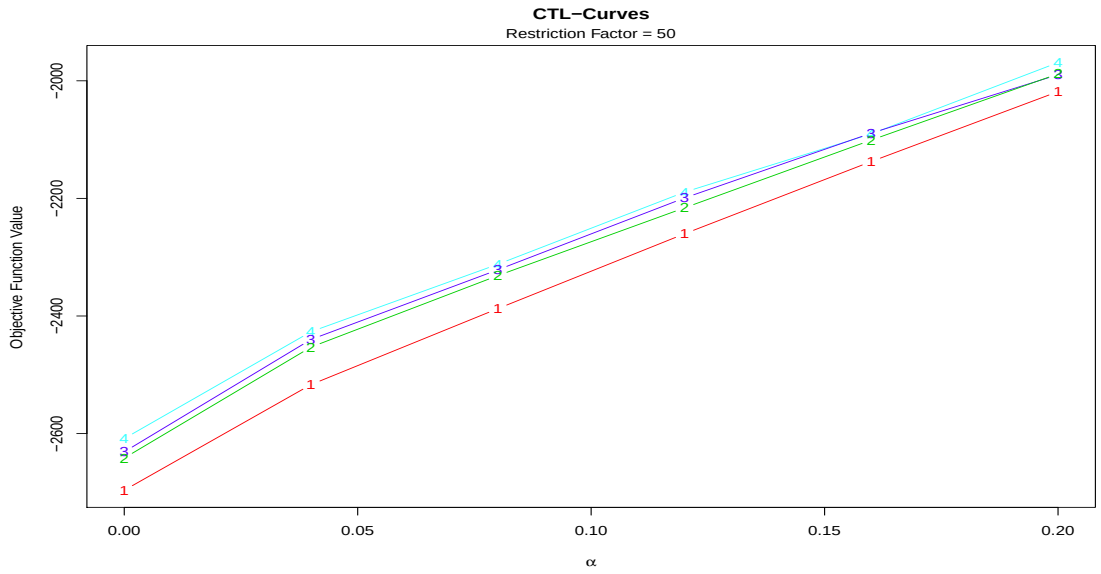

Figure 5.2: `ctlcurve` plot for the FIES data.

As sometimes happens, Figure 5.2 clearly indicates that there should be $k > 2$ groups, but it is unclear as with respect to the choice between $k = 3$ and $k = 4$. Additionally, it is definitely not conclusive with respect to the optimal trimming level α, which here is a parameter of interest as it is connected with the number (and identity) of outlying nations. The final estimates depend on the choice of α. In this example, RTCLUST can be seen as an automatic way of choosing the optimal trimming level, as the one balancing between robustness and efficiency. For the

proposed methodology we do not need to specify α. We have applied our method both based on $k = 3$ and $k = 4$. As with $k = 4$ two groups are not very separated, we prefer $k = 3$ and report only those results for reasons of space. We run `rtclust` with $k = 3$, initial trimming level $\alpha_0 = 0.2$, $\alpha_L = 0.001$.The results are remarkably stable with respect to the tuning parameters. Nine countries (7.1%) are flagged as outlying, 13 are classified in group 1, 95 in group 2, and 10 in group 3. The cluster profiles (cluster means) and raw measurements for the outlying countries are reported in Table 5.2. It shall be noted that groups 1 and 3 are of similar size as the group of outliers. Countries in groups 1 and 3 are very similar though and close to the reported profiles, while outliers are provably scattered, or have extremal values at least in one of the dimensions considered.

Table 5.2: Cluster profiles and measurements for the outlying countries. FIES: Food Insecurity Experience Scale. CE: Civic Engagement. St: Struggling. FS: Food Security. Co: Corruption index. YD: Youth Development. C-j: j-th cluster profile.

	FIES	CE	St	FS	Co	YD
C-1	-0.34	44.64	55.19	69.66	45.53	75.79
C-2	0.13	31.42	63.24	53.99	74.33	58.61
C-3	0.41	22.55	63.94	52.51	67.97	44.90
Myanmar	-0.95	66.84	85.80	13.21	53.33	85.48
Sweden	-0.64	43.22	48.08	76.68	37.39	59.67
Georgia	-0.43	21.24	60.49	41.31	30.85	67.56
New Zealand	-0.12	57.98	55.52	67.02	40.50	66.94
Paraguay	0.06	17.43	81.05	88.26	66.31	40.64
Rwanda	0.27	13.12	69.74	61.54	9.29	84.48
Cambodia	0.90	26.93	62.30	20.61	73.53	86.70
South Sudan	3.81	35.17	51.53	35.22	58.29	49.97
Haiti	5.04	35.33	51.47	43.66	57.24	32.07

It can be seen that the three clusters are well separated in terms of all of the items considered. The first cluster is characterized by the lowest food insecurity (and largest food security), corruption and struggling, and by the largest civic engagement and youth development. Sadly, only a minority of countries are assigned to cluster 1. The third cluster is characterized by largest food insecurity, lowest civic engagement and youth development. No differences are seen in terms of struggling and FS index between clusters 2 and 3. Finally, not surprisingly the corruption index is higher in the slightly more developed countries belonging to cluster 2 than in those in cluster 3. The outliers are easily explained, as for instance Haiti and South Sudan have an

extremely high FIES. Sweden might belong to cluster 1, but its corruption is so low and its food security (however measured) is so large that it is outlying. All other outliers have at least one measurement in complete disagreement with the three clusters. A special note regards Myanmar, where there might have been problems with the questionnaire and with the sampling, and whose measurements therefore might not be completely reliable.

It shall be noted that the new FIES score is able to separate very well the three clusters, while Gallup's FS score only discriminates between the first and the other two. Other evidence in favor of the added value of FIES is that if we remove it and repeat the analysis the average silhouette width decreases by about 4%.

5.3 Applications of fuzzy linear clustering

Allometry studies the relationships between biometric measurements in humans, animals, and plants. Clusterwise regression is particularly useful for allometric studies since relations between biometric measurements are often linear or close to linear, possibly after transformation, and additionally there might be different relationships according to other variables which might not even be measured. For instance, the relationship between head circumference and height in humans is different at different age classes. In our experience, groups are seldom perfectly separated and overlapping may hinder the true relationships if not properly taken into account (e.g., through fuzzy weights). Additionally, outlying biometric measurements are often present.

We illustrate based on an example already considered in García-Escudero, Gordaliza, Mayo-Iscar & San Martín (2010), where sharp clusterwise regression was implemented. Here we implement fuzzy clusterwise regression, showing that use of fuzzy weights leads to better clustering and better understanding of bridge points between clusters. Data is made of 362 measurements of height and diameter of *Pinus Nigra* trees located in the north of Palencia (Spain). We aim to explore the linear relationship between these two quantities. Our explanatory variable is "diameter", while the outcome is "height". This is justified by the fact that roughly measuring the diameter of a tree is extremely simple, while measuring the height is expensive as the tree must be climbed or at least measured with more complex tools and by a team of operators. Hence a reliable way of predicting the height from the diameter would be cost-effective. The diameter and height can be used to roughly estimate the age and the volume (that is, the amount of wood) of each tree.

The scatter plot in Figure 5.4 clearly shows that there should be three approxi-

mately linear groups, and an isolated group of outliers.

To confirm this we use both our heuristic and automatic methods for choosing the tuning parameters. In Figure 5.3,(a) gives the *ctlcurve*, which indicates that we should fix $k = 3$ and $\alpha = 0.04$. This is confirmed in panel (b), where we show the average contribution to the likelihood. Finally, from panel (c) we see that for $m > 1.3$ the proportion of hard assignments decreases sharply. This makes us set $m = 1.3$.

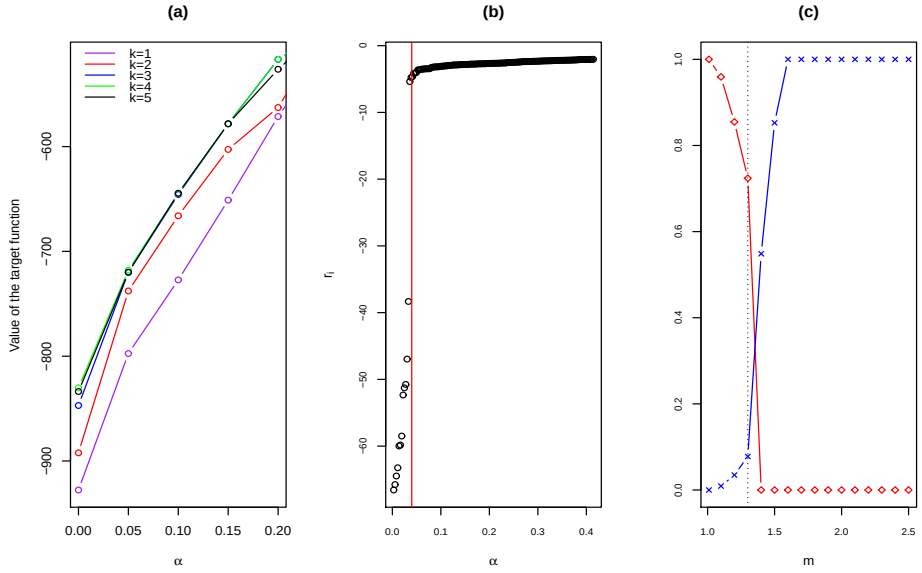

Figure 5.3: *Pinus Nigra example.* (a) *ctlcurve.* (b) average contribution to the likelihood as a function of α. (c) relative empty entropy and proportion of hard assignments as a function of m.

Finally, it shall be noted that the tuning parameters chosen with our proposed automatic method based on pseudo cross-validation are $k = 3$, $m = 1.3$, $\alpha = 0.04$ and $c = 25$. There is substantial agreement between heuristics and automatic tuning, and these are the parameters we use in the following.

We apply four procedures: the fuzzy c-means regression method, Figure 5.4,(a), our method without trimming in (b), the A-cReg method in (c), and, our proposal in (d).

It can be seen that the untrimmed procedures are not able to detect the most likely underlying linear relationships even if one additional cluster is used, as done in Figure 5.5, as the small isolated groups of observations have a direct influence on one of the clusters, and an indirect one on the other two. A very large coefficient is estimated for the group including the isolated outliers, while another group includes too many observations with many fuzzy memberships.

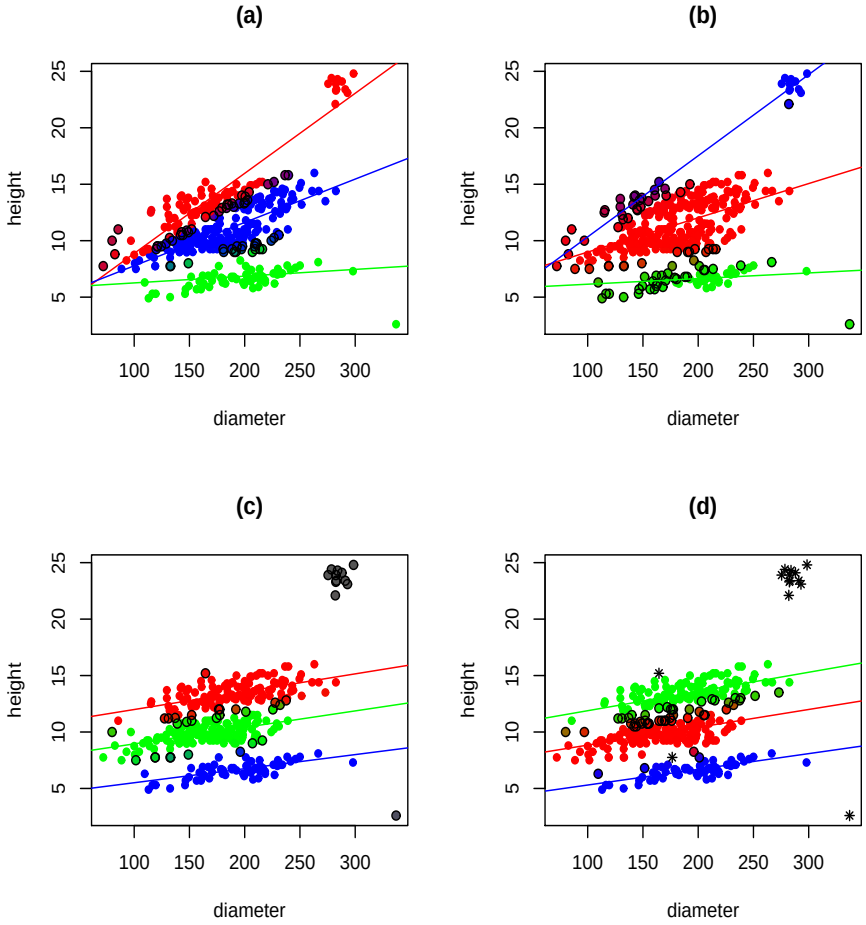

Figure 5.4: *Pinus Nigra example:* (a) Scatter plot and results of cReg method. (b) results of the "EM" method. (c) results of the A-cReg method. (d) results of the FTCR method. Circled observations are fuzzy assignments.

On the other hand, by trimming as few as 15 observations, we recover quite nicely the linear structures. A similar result is obtained with A-cReg, but at the price of a longer computational time. The FTCR procedure (and similarly A-cReg) give a good proportion of hard assignments, indicating that the estimated clusters are well separated. There is also a fair proportion of fuzzy cluster assignments, which might mislead interpretation if hard assigned to one of the clusters. The FTCR procedures formally detects outliers while A-cReg gives $u_{ij} = 1/k$ membership values. More distant outliers, as seen in simulations, might lead the two procedures to behave differently. In Figure 5.4, less fuzzy observations (i.e. whose $\max_j u_{ij} \leq 0.95$) are plot with the symbol "○".

A clear interpretation for the three clusters is that pines are sampled in three different zones. It can be seen that three almost parallel lines are obtained, indi-

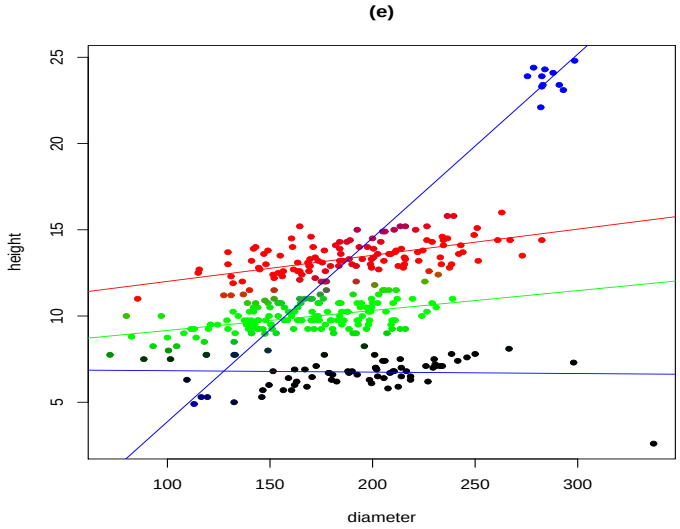

Figure 5.5: *Pinus Nigra Data example:* Results of the proposed procedure when searching for $k = 4$ clusters and no trimming imposed

cating a similar relationship between diameter and height within the three zones. We can therefore speculate that environmental conditions (e.g., quota, rainfall, sun exposure) are similar in the three zones, but that immigration of the species has occurred in different times; where in the "green" zone trees are older (and therefore bigger) and the most recent colonization (with younger and smaller trees) has occurred in the "blue" zone. Additionally, outliers can be easily justified since they are trees of a different species which seems to be misclassified as "Pinus Nigra".

Chapter 6

Conclusions and further directions

6.1 Concluding remarks on the reweighted TCLUST contribution

In Chapter 3 we have presented an iteratively reweighed approach that can recover wrongly trimmed observations when applying robust clustering procedures based on high (preventive) trimming levels. This approach also makes easier the use of the TCLUST robust clustering method by eliminating the need to calculate the initial trimming level and the eigenvalue constraint. RTCLUST has two advantages over TCLUST: first, a sometimes not easily chosen tuning parameter, the trimming level, does not need to be perfectly specified in advance and the same happens for the eigenvalue ratio constraint value c. Secondly, it conjugates high robustness (as it can resist to an α_0 proportion of outliers) with high efficiency (as under no or little contamination the proportion of discarded observations will be much lower than α_0). The simulation study and the real data example also show how this methodology could be useful in practical applications. There is still room for further work. Formal theoretical properties could be explored. As commented in Remark 1, the outlier labeling process at each iteration could also be refined. We have applied very simple thresholds based on the χ^2 approximation for the Mahalanobis distances. More accurate procedures could be obtained, for instance, by considering small sample approximations or correcting for the multiple testing when labeling outliers (see, e.g., Cerioli 2010, Cerioli & Farcomeni 2011). The multiple testing approach to reweighting might be tweaked to yield a simultaneous robust estimation and outlier detection method. The proposed methodology assumes that the number of groups k is known in advance. Estimating a correct k value is an important, but difficult too, problem. In fact, this is an ill-posed problem because the total number of groups

depends on the type of clusters we are searching for or on what we understand as noise. For instance, a set made up with several disperse observations can be seen as a proper group with a large scatter or it can also be seen as background noise. Therefore, searching for the proper number of groups k would require making some subjective choices specifying all these aspects somehow. Another interesting open research line has to do with the extension of this iteratively reweighing approach for mixture modeling. This could be useful in order to address severe overlaps among groups.

6.2 Concluding remarks on the TCLUST extension to fuzzy linear clustering models

In Chapter 4 we have proposed a procedure for robust fuzzy linear clustering. The procedure can resist to different types of contamination, still being efficient in parameter estimation. When there is overlap between groups (and m is well calibrated), some observations might receive fractional membership values. On the other hand, observations that are well separated are hard assigned to a cluster and form the cluster core. Finally, observations far from any cluster are trimmed.

The updating algorithm is based on several closed form expressions. This avoids us the use of numerical maximization routines, with obvious advantages in terms of computational complexity.

As often happens with robust procedures, tuning is required in order to obtain sensible results. We have described some heuristical tools that we found useful for satisfactory tuning. We have also briefly outlined a new method, based on pseudo cross-validation. We have provided some initial evidence that the method might provide reasonable results, but a full exploration of its properties and performance is beyond the scope of this thesis. We leave it for further work.

As another further direction for research, the robustness' properties of the procedure could be extended in order to guarantee robustness against the effect of entry-wise outliers. In order to perform this improvement, instead of applying trimming to a fixed proportion of observations, snipping techniques, like the one proposed in Farcomeni (2014*b*) and Farcomeni (2014*a*), may be applied. Additionally assessing the goodness of fit of the procedure may be necessary and meaningful results might be obtained for instance by extending one of the robust tests proposed in Cerioli & Farcomeni (2011) and Cerioli et al. (2013).

6.3 Overall conclusions and further direction of research

The two contributions proposed within this thesis are based on the `tclust` method. In particular we focused on making the tuning of the `tclust` easier and extending it to the linear clustering models. Besides the possible extensions of the such contributions there is still room for extending the `tclust` approach to different statistical frameworks.
As an example we wish to introduce here an ongoing work (Dotto & Farcomeni, *In preparation*) aimed at introducing geometric constraints within the `tclust` algorithm. We consider the parameterizations of the covariance matrix of each group outlined in Celeux & Govaert (1995) and implemented within the `mclust` R package in Fraley & Raftery (2002). Let us consider the eigenvalue decomposition given by:

$$\boldsymbol{\Sigma_k} = \lambda_k D_k A_k D_k^T \tag{6.1}$$

where $\lambda_k = |\boldsymbol{\Sigma_k}|^{1/d}$ is a measure of the volume of the k-th cluster, A_k is an orthogonal matrix with the eigenvalues of Σ_k on the diagonal and it describes the shape of each cluster, and D_k is a matrix whose columns are given by the eigenvectors of Σ_k and it determines the direction of each cluster. Combining all possible assumptions regarding scale, volume and orientation, Celeux & Govaert (1995) described 14 different models. A shorter list can be found in Fraley & Raftery (2007), and we summarize it in Table 6.1. In such table we report the model name as commonly referred (and used also in `R` library `mclust`), the final parameterization of Σ_k, cluster shapes and properties of invariance of the solutions. The simplest model, EII, involves spherical clusters and its solution corresponds to homogeneous model-based clustering and k-means. The solution is invariant only with respect to isometric transformations (that is, preserving distances). Model VVV, on the other hand, corresponds to the unconstrained case where Σ_k is arbitrary.

As stated in Table 6.1 different parameterizations of the covariance matrix imply different properties in terms of equivariance. A notable drawback of (2.11), which is recalled in the following formula,

$$\frac{M_n}{m_n} = \frac{\max_{j=1,2,\ldots,K} \max_{l=1,2,\ldots,p} \lambda_l(\Sigma_j)}{\min_{j=1,2,\ldots,K} \min_{l=1,2,\ldots,p} \lambda_l(\Sigma_j)} \tag{6.2}$$

is that all properties of affine invariance are lost, as any affine transformation (except for traslations) lead to different eigenvalue ratios. We note here that spurious solutions need not arise under some formulations, basically those pooling scale or volume across clusters. A full account on whether (2.11) is needed to avoid spurious

Table 6.1: List of 10 different models obtained by imposing different constraint in decomposition (6.1)

Model Name	Parametrization	ER	Invariance
EII	λI	Not required	Isometric transformations
VII	$\lambda_k I$	Not required	Isometric transformations
EEI	λA	Not required	Scaling
VEI	$\lambda_k A$	Not required	Scaling
EVI	λA_k	Not required	Traslation
VVI	$\lambda_k A_k$	Required	Traslation
EEE	$\lambda D A D^T$	Not required	Linear transformations
EEV	$\lambda D_k A D_k^T$	Not required	Linear transformations
VEV	$\lambda_k D_k A D_k^T$	Not required	Linear transformations
VVV	$\lambda_k D_k A_k D_k^T$	Required	Traslation

solutions is given in the ER column of Table 6.1. All the models above are not resistant to contamination and for that reason, in our proposal, we robustly estimate model parameters by developing a CEM algorithm augmented with an impartial trimming step. Our task is to maximize the classification log-likelihood

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} z_{ij} \left[\ln \sum_{k=1}^{K} \pi_k \phi(\boldsymbol{x_i} | \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \right], \tag{6.3}$$

where z_{ij} is a binary indicator that the i-th observation belongs to the j-th cluster, with $\sum_j z_{ij} \leq 1$ and $\sum_{ij} z_{ij} = n(1-\alpha)$, where α is the trimming level. Consequently, $n\alpha$ observations are not classified into any cluster, and do not contribute to the objective function (6.3). In particular the algorithm aimed at maximizing the (6.3) iterates the following steps:

Algorithm 7.

1. *Initialization*: Initialize randomly k initial centers $m_1^0, \ldots, m_k^0$, k covariance matrices $\Sigma_1^0, \ldots, \Sigma_k^0$ and k values $p_1^0, \ldots, p_k^0$ or the clusters' weights.

2. *Concentration steps*:

   2.1 Keep the set H containing the $\lceil n(1-\alpha) \rceil$ observations closest (w.r.t the Mahalanobis distance) to the estimated centroids $m_1, \ldots, m_k$.

   2.2 For each $i = 1 \ldots n$ obtain the clusters' assignments by computing the minimization $\min_j d^2_{\Sigma_j}(x_i; m_j)$.

   2.3 Update the estimates of the clusters' centers $m_1, \ldots, m_k$, clusters' scatter matrices $\Sigma_1, \ldots, \Sigma_k$, and clusters' weights $p_1, \ldots, p_k$. Depending on

the chosen model different procedures are needed in order to update the scatter matrices. In particular Celeux & Govaert (1995) analyzed all the possible cases outlined in Table 6.1 and provided closed forms, whenever these are available, and the required iterative procedure, whenever a closed form for the given estimator does not exist.

3. Repeat Steps 2.1 - 2.3 until there are no improvements in equation (6.3).

4. Draw several different random starting values and recompute the values of the objective function. Keep the configuration yielding the maximal value of (2.9) as the final output of the algorithm.

Preliminary simulation study showed very good results both in terms of robustness and efficiency. Clearly optimal choice of a proper model plays a key role. Besides the heuristics we are now also working on an automatic method of choosing the parametrization of the scatter matrix which suits best the underlying structure of the data. Pretty good results have been reached by using the tests provided in Vuong (1989) and Clarke (2003). Within these contributions the authors developed parametric and non parametric tests for model selection both in case of nested and overlapping models.
In synthesis, in this further work, we will propose the MTCLUST methodology, that is, TCLUST restricted to parametrizations of the kind (6.1). The advantages are that in certain cases the eigenvalue ratio constraint is not needed, and hence affine equivariance is retained; and that robust and parsimonious clustering will be made available to the interested audience.

6.3.1 Preliminary simulation results

Within this subsection we provide the preliminary simulation's results obtained comparing the new outlined methodology with the TCLUST methodology. In particular we generate the data following the different schemes outlined within table (6.1) and add a fixed proportion (α) of contaminating points. Then we compare the performance of the new proposed methodology and of the TCLUST methodology imposing two different trimming levels (i.e 5% and 10%) in terms of mean square error of the estimated vector mean of each cluster. Additionally, in the last column we reported the mean square error of the model automatically chosen by performing the parametric test introduced in Vuong (1989). The procedures involved in this preliminary simulation study are labelled as follows:

- `mtclust.10` Our poposed method on which we imposed a trimming level equal

to 10%. The constraint applied is either the eigenvalue ratio, when required, or a constraint which obeys the true data generation mechanism.

- `tclust.10` The `tclust` method on which we imposed a trimming level equal to 10%
- `mtclust.05` Our proposed method on which we imposed a trimming level equal to 5%.The constraint applied is either the eigenvalue ratio, when required, or a constraint which obeys the true data generation mechanism.
- `tclust.05` The `tclust` method on which we imposed a trimming level equal to 5%
- The model automatically chosen by applying the Vuoung Test (Vuong 1989) for selecting nested and overlapping models on which we imposed a trimming level equal to 10%.

It shall be noticed that the performance obtained are, in most cases, pretty similar to the one obtained by applying the `tclust`. Moderate improvements are obtained as trimming level is underestimated. We believe that this happens because of the effect of the geometric constraint imposed in the different cases. Indeed by constraining the shape of each cluster we include "less dangerous" outliers when the true trimming level is underestimated. Additionally substantial improvements are obtained in the cases in which the eigenvalue ratio is unconstrained. Indeed, within this simulation setting the imposed bound for the eigenvalue ratio is equal to 12. This of course leads to a great loss of efficiency whenever the true eigenvalue ratio is above this value. Nevertheless by choosing constrained models which do not require the usage of the eigenvalue ratio this loss of efficiency can be avoided. As an example if we consider settings 41 or 54 of Table 6.2 we can easily see that by choosing a different constrained model a significant improvement of efficiency can be obtained, and similar consideration con be done by reading settings 45 and 48.

Table 6.2: Simulation results based on B=500 replicates: average MSE of the estimated vector mean in each cluster

Setting	p	α	Data Genearation	mtclust.10	tclust.10	mtclust.05	tclust.05	Chosen Model
1	2	0.1	EII	0.128	0.128	0.362	0.544	0.128
2	4	0.1	EII	0.182	0.182	0.377	0.606	0.182
3	6	0.1	EII	0.229	0.229	0.415	0.751	0.229
4	2	0.05	EII	0.137	0.137	0.123	0.123	0.123
5	4	0.05	EII	0.192	0.193	0.178	0.178	0.178
6	6	0.05	EII	0.236	0.236	0.219	0.219	0.219
7	2	0.1	VII	0.209	0.208	0.570	0.615	0.208
8	4	0.1	VII	0.288	0.288	0.681	0.737	0.288
9	6	0.1	VII	0.359	0.358	0.78	0.833	0.358
10	2	0.05	VII	0.221	0.221	0.195	0.196	0.196
11	4	0.05	VII	0.310	0.308	0.278	0.278	0.278
12	6	0.05	VII	0.383	0.383	0.344	0.344	0.344
13	2	0.1	EEI	0.156	0.157	0.369	0.471	0.157
14	4	0.1	EEI	0.294	0.294	0.447	0.536	0.296
15	6	0.1	EEI	0.417	0.418	0.567	0.660	0.420
16	2	0.05	EEI	0.166	0.166	0.152	0.151	0.152
17	4	0.05	EEI	0.302	0.302	0.283	0.283	0.284
18	6	0.05	EEI	0.439	0.440	0.413	0.413	0.414
19	2	0.1	VEI	0.196	0.196	0.470	0.525	0.196
20	4	0.1	VEI	0.354	0.354	0.588	0.681	0.354
21	6	0.1	VEI	0.524	0.524	0.802	0.896	0.524
22	2	0.05	VEI	0.204	0.203	0.185	0.185	0.185
23	4	0.05	VEI	0.375	0.375	0.349	0.348	0.348
24	6	0.05	VEI	0.541	0.540	0.498	0.498	0.498
25	2	0.1	EVI	0.267	0.267	0.384	0.409	0.266
26	4	0.1	EVI	0.414	0.414	0.516	0.542	0.414
27	6	0.1	EVI	0.549	0.549	0.664	0.699	0.549
28	2	0.05	EVI	0.272	0.272	0.249	0.249	0.249
29	4	0.05	EVI	0.434	0.434	0.405	0.405	0.405
30	6	0.05	EVI	0.569	0.569	0.536	0.536	0.536
31	2	0.1	VVI	0.266	0.266	0.546	0.553	0.267
32	4	0.1	VVI	0.395	0.395	0.705	0.713	0.395
33	6	0.1	VVI	0.492	0.492	0.826	0.838	0.492
34	2	0.05	VVI	0.301	0.300	0.266	0.266	0.266
35	4	0.05	VVI	0.416	0.415	0.374	0.373	0.374
36	6	0.05	VVI	0.523	0.523	0.477	0.477	0.478
37	2	0.1	EEV	0.356	0.365	0.552	0.502	0.356
38	4	0.1	EEV	1.289	1.344	1.493	1.483	1.289
39	6	0.1	EEV	2.710	2.774	2.727	4.845	2.710
40	2	0.05	EEV	0.371	0.396	0.344	0.351	0.344
41	4	0.05	EEV	1.339	1.435	1.264	1.274	1.264
42	6	0.05	EEV	2.875	3.169	2.716	2.759	2.716
43	2	0.1	VEV	0.396	0.410	1.208	0.490	0.396
44	4	0.1	VEV	1.410	1.480	1.787	1.673	1.410
45	6	0.1	VEV	2.981	35.193	11.427	63.947	2.981
46	2	0.05	VEV	0.452	0.494	0.397	0.412	0.397
47	4	0.05	VEV	1.504	1.675	1.346	1.414	1.346
48	6	0.05	VEV	3.284	10.449	2.982	35.849	2.982
49	2	0.1	VVV	0.593	0.594	0.656	0.653	0.548
50	4	0.1	VVV	1.987	1.987	2.228	2.243	1.872
51	6	0.1	VVV	83.816	83.875	85.751	94.9	3.930
52	2	0.05	VVV	0.613	0.613	0.540	0.539	0.508
53	4	0.05	VVV	2.355	2.352	1.915	1.912	1.807
54	6	0.05	VVV	74.705	74.590	83.400	83.636	3.878

Appendix A

A.1 Proofs of the theoretical properties of the RTCLUST methodology

A.1.1 Introduction and notation

In this section some modifications to the notation used in Chapter 3 are introduced in order to better outline the proofs of the theoretical results stated within the chapter.

Let $\theta = (\pi_1, \ldots, \pi_k, \pi_{k+1}, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k) \in \Theta$, where Θ is the considered parametric space. We define

$$D_\theta(x) = \min_{1 \leq j \leq k} D_\theta^j(x)$$

where $D_\theta^j(x) = d^2_{\Sigma_j}(x, \mu_j)$ is the Mahalanobis distance from the center μ_j and the scatter matrix Σ_j.

Given a fixed probability measure P, let us consider $G_\theta^P(u) = P[D_\theta(\cdot) \leq u]$ and its β quantile $D_\theta^{P,\beta} = \inf_u\{G_\theta^P(u) \geq \beta\}$. If θ_P^{l-1} with

$$\theta_P^{l-1} = (\pi_{1P}^{l-1}, \ldots, \pi_{kP}^{l-1}, \pi_{k+1P}^{l-1}, \mu_{1P}^{l-1}, \ldots, \mu_{kP}^{l-1}, \Sigma_{1P}^{l-1}, \ldots, \Sigma_{kP}^{l-1}),$$

are the values of the parameters at stage $l-1$, we consider the sets

$$A_{\theta_P^{l-1}}^{P,1-\alpha_l} = \{x | D_{\theta_P^{l-1}}(x) < D_{\theta_P^{l-1}}^{P,1-\alpha_l}\},$$

$$B_{\theta_P^{l-1}} = \{x | D_{\theta_P^{l-1}}(x) \leq \chi^2_{p,1-\alpha_L}\}$$

and

$$H_{j\theta_P^{l-1}}^{P,1-\alpha_l} = \left\{x | D_{\theta_P^{l-1}}^j(x) = D_{\theta_P^{l-1}}(x)\right\} \cap A_{\theta_P^{l-1}}^{P,1-\alpha_l} \cap B_{\theta_P^{l-1}}.$$

The consistency factors for the scatter matrices are obtained as

$$\left(c_{\theta_P^{l-1}}^{P,1-\alpha_l}\right)^{-1} = \eta_{P\left(A_{\theta_P^{l-1}}^{P,1-\alpha_l} \cap B_{\theta_P^{l-1}}\right) \Big/ P\left(B_{\theta_P^{l-1}}\right)}$$

if $P\left(A_{\theta_P^{l-1}}^{P,1-\alpha_l} \cap B_{\theta_P^{l-1}}\right)/P(B_{\theta_P^{l-1}}) < 1$ and equal to 1 otherwise. As done in Chapter 3, $\eta_\beta = P\left(\chi^2_{p+2} \leq \chi^2_{p,\beta}\right)/\beta$. Then, by using this notation and $I_A(\cdot)$ as the indicator function of set A, we have updated parameters:

$$\pi_{jP}^l = P\left(H_{j\theta_P^{l-1}}^{P,1-\alpha_l}\right),$$

$$\pi_{k+1P}^l = 1 - P\left(B_{\theta_P^{l-1}}\right),$$

$$\mu_{jP}^l = \int x I_{H_{j\theta_P^{l-1}}^{P,1-\alpha_l}}(x) dP(x)$$

and

$$\Sigma_{jP}^l = \left(\int xx' I_{H_{j\theta_P^{l-1}}^{P,1-\alpha_l}}(x) dP(x) - \mu_{jP}^l (\mu_{jP}^l)'\right) c_{\theta_P^{l-1}}^{P,1-\alpha_l}.$$

Given $\{x_1, ..., x_n\}$ being a realization of an independent identically distributed (i.i.d.) sample from distribution P, let P_n denote its associated empirical measure. When replacing the (unknown) P by P_n in previous expressions, we obtain $\theta_{P_n}^l$ exactly as the parameters appearing in Algorithm 5.

A.1.2 Proof of Theorem 1

The required bounds for the parameters have already been proved when $l = 0$ in García-Escudero et al. (2008). Notice that assuming an absolutely continuous distribution P automatically guarantees the PR condition in García-Escudero et al. (2008).

Let us also assume that the solution of that TCLUST population problem satisfy $\pi_{jP}^0 > 0$ for $1 \leq j \leq k$ (otherwise it is clear that k should have been decreased for clustering purposes). In order to apply an inductive reasoning, let us suppose that the parameters in θ_P^{l-1} do satisfy the boundedness condition in the statement of Theorem 1. Given that P has a strictly positive density function, if $\mu_{j_1P}^{l-1} \neq \mu_{j_2P}^{l-1}$ for every $j_1 \neq j_2$, then it is trivial to prove that each $H_{j\theta_P^{l-1}}^{P,1-\alpha_l}$ contains a non empty open ball and consequently $\pi_{jP}^l > 0$ for $1 \leq j \leq k$. This also implies that the eigenvalues $\{\lambda_q(\Sigma_{jP}^l)\}_{q=1}^p$ can be uniformly bounded from below by a strictly positive constant. The other bounds follow from the boundedness of the $H_{j\theta_P^{l-1}}^{P,1-\alpha_l}$ sets, which is a consequence of the previously assumed bounds for the θ_P^{l-1} parameters.

A.1.3 Proof of Theorem 2

As commented before, we recover the parameters in Algorithm 5 in Chapter 3 when the unknown probability measure P is replaced by the empirical measure P_n. Therefore, we use the notation $\theta^l_{P_n}$ for those parameters obtained from an i.i.d. random sample $\{x_1, ..., x_n\}$ from P.

Lemma A.4 and Lemma A.5 in García-Escudero et al. (2008) guarantee that there exists a compact set K satisfying $\theta^0_{P_n} \in K$ for $n > n_0$ with probability 1. An inductive reasoning, similar to that applied in the proof of Theorem 1, would show that the $H^{P_n,1-\alpha_l}_{j\theta^{l-1}_{P_n}}$ is uniformly bounded with probability 1. It may happen that one of these sets would have 0 probability mass under P_n. In that case, we just need to take $\mu^l_{jP_n} = \mu^{l-1}_{jP_n}$ (recall that $\mu^{l-1}_{jP_n}$ was bounded because of the inductive reasoning applied) and take $\Sigma^l_{jP_n}$ equal to the zero matrix.

A.1.4 Proof of Theorem 3

In this proof, we will apply results from of Empirical Processes theory (see, e.g., van der Vaart & Wellner 1997) and the inductive reasoning again to prove consistency of the sample parameters toward the population ones. Some technical lemmas will be also needed in this proof.

From García-Escudero et al. (2008) we know that the sample solution of the TCLUST is consistent to the population one. I.e., we have that

$$\theta^0_{P_n} \to \theta^0_P, \ P\text{-almost surely.}$$

By assuming the consistency in the $(l-1)$-th iteration, i.e.

$$\theta^{l-1}_{P_n} \to \theta^{l-1}_P, \ P\text{-almost surely,} \tag{A.1}$$

we now have to prove the consistency for the l-th iteration.

By the notation introduced in as A.1 we see that:

Lemma 1. For a probability distribution Q in $\mathbb{R}^p$, for each $\theta \in \Theta$, $a \in \mathbb{R}$, and $1 \leq j \leq k$, the sets A^Q_θ, B_θ are contained in a Vapnik-Chervonenkis (VC) classes of sets Ξ, $A^{Q,1-\alpha}_\theta \cap B_\theta$ are contained in a VC class Λ and $H^{Q,1-\alpha}_{j\theta}$ are contained in a VC classes of sets Ψ. These classes are given by

$$\Xi = \{U_{\theta a} | \theta \in \Theta; a \in \mathbb{R}\}$$

$$\Lambda = \{U_{\theta a} \cap U_{\theta b} | \theta \in \Theta; a, b \in \mathbb{R}\}$$

and

$$\Psi = \{V_{\theta j} \cap U_{\theta a} \cap U_{\theta b} | \theta \in \Theta; 1 \leq j \leq k; a, b \in \mathbb{R}\},$$

where

$$U_{\theta a} = \{x | D_\theta(x) \leq a\} \text{ and } V_{\theta j} = \{x | D_\theta(x) = D_\theta^j(x)\}.$$

Proof. Since $D_\theta(x)$, for $\theta \in \Theta$, is the minimum of k functions belonging to a finite dimensional subspace of functions, then $\{D_\theta(x) | \theta \in \Theta\}$ is a VC class by lemmas 2.6.15 and 2.6.18 in van der Vaart & Wellner (1997). Analogously. Ξ, Λ and Ψ are VC classes of sets by application of lemmas 2.6.15, 2.6.17 and 2.6.18 in van der Vaart & Wellner (1997). Sets $A_\theta^{Q,1-\alpha}$ and B_θ are contained in Ξ for $\theta \in \Theta$. Their intersection $A_\theta^{Q,1-\alpha} \cap B_\theta$ are contained in Λ and, for $j = 1, \ldots, k$, $H_{j\theta}^{Q,1-\alpha}$ are contained in Ψ. □

Lemma 2. Under the assumptions of Theorem 3 and assuming (A.1), we have

$$D_{\theta_{P_n}^{l-1}}^{P_n,1-\alpha_l} \to D_{\theta_P^{l-1}}^{P,1-\alpha_l}, \ P\text{-almost surely.}$$

Proof. In order to prove this lemma we need to show

$$\sup_{\theta \in K} \left| D_\theta^{P_n,1-\alpha_l} - D_\theta^{P,1-\alpha_l} \right| \to 0, \ P\text{-almost surely,}$$

in a compact set $K \subseteq \Theta$. This follows exactly as in Lemma A.7 in García-Escudero et al. (2008), given the assumed convergence (A.1). □

Lemma 3. Under the assumptions of Theorem 3 and assuming (A.1), the following convergences hold

$$\pi_{jP_n}^l \to \pi_{jP}^l, \text{ for } j = 1, \ldots, k, k+1, \ \mu_{jP_n}^l \to \mu_{jP}^l \text{ and } \Sigma_{jPn}^l \to \Sigma_{jP}^l.$$

Proof. Due to the Glivenko-Cantelli property of the classes Ψ, Λ and Ξ together with (A.1) and the consistency results for the quantiles in Lemma 2, we can state that

$$P_n\left(H_{j\theta_{P_n}^{l-1}}^{P_n,1-\alpha_l}\right) \to P\left(H_{j\theta_P^{l-1}}^{P,1-\alpha_l}\right),$$

and, consequently $\pi_{jP_n}^l \to \pi_{jP}^l$ for $1 \leq j \leq k$. Analogously, the consistency $\pi_{k+1P_n}^l \to \pi_{k+1P}^l$ follows from the convergence

$$P_n\big(B_{\theta_{P_n}^{l-1}}\big) \to P\big(B_{\theta_P^{l-1}}\big), \tag{A.2}$$

that it is obtained in a similar fashion.

Due to Glivenko-Cantelli property for the class $\{xI_H(x)|H \in \Psi\}$ and Lemma 2, we have that $\mu^l_{jPn} \to \mu^l_{jP}$, P-almost surely.

Additionally, we have consistency for the consistency factors as

$$c^{P_n,1-\alpha_l}_{\theta^{l-1}_{P_n}} \to c^{P,1-\alpha_l}_{\theta^{l-1}_{P}}, \ P\text{-almost surely.}$$

This last consistency is trivial given that

$$P_n\left(A^{P_n,1-\alpha_l}_{\theta^{l-1}_{P_n}} \cap B_{\theta^{l-1}_{P_n}}\right) \to P\left(A^{P,1-\alpha_l}_{\theta^{l-1}_{P}} \cap B_{\theta^{l-1}_{P}}\right),$$

together with the convergence (A.2) and the fact that $\eta_\beta = P\big(\chi^2_{p+2} \leq \chi^2_{p,\beta}\big)/\beta$ (seen as a function on β) is a continuous function for $\beta \in (0,1)$.

Therefore, given that the class $\{xx'I_H(x)|H \in \Psi\}$ is also a Glivenko-Cantelli class and the consistency of those $c^{P_n,1-\alpha_l}_{\theta^{l-1}_{P_n}}$ factors, we see that $\Sigma^l_{jP_n} \to \Sigma_{jP}$ P-almost surely for $1 \leq j \leq k$. □

The combination of all the above lemmas then allow us to argue in favor of consistency at each iteration $l = 1, \ldots, L$, by applying the inductive reasoning.

A.1.5 Proof of Theorem 4

It can be easily proven, by using the same arguments as in Cuesta-Albertos et al. (2008*b*) and in Hennig (2004), that the TCLUST with a trimming level α_0 does not break down with the addition of less than $[\alpha_0 n]$ outliers for these well-clusterized dataset and this type of contamination scheme. This level of resistance to outliers, given by $[\alpha_0 n]$, cannot be deteriorated throughout the proposed reweighing approach in a finite number of iterations L by applying a straightforward inductive reasoning again.

A.2 Justification of Algorithm 6

At each step of the algorithm an increase of the target function (4.3) is obtained, therefore the algorithm must converge at least to a local maximum. By using several random starting points we increase the likelihood of finding the global maximum of the target function. The rest of this section aims at justification of these claims:

Membership values: Conditionally on $\boldsymbol{b}_j, b^0$ and s_j for $j = 1, \ldots, k$, maximizing (4.3) is equivalent to minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^m D_{ij} \tag{A.3}$$

where $D_{ij} = -\log(p_j f(y_i; \boldsymbol{x_i'b}_j + b_j^0, s_j^2)) = \log\left[p_j^{-1}(2\pi s_j^2)^{1/2}\exp\left((y_i - \boldsymbol{x_i'b}_j - b_j^0)^2/(2s_j^2)\right)\right]$. If $p_j f(y_i; \boldsymbol{x_i'b}_j + b_j^0, s_j^2) < 1$ for all j then $D_{ij} (> 0)$ can be seen as a measure of the distance between y_i and its fitted value $\boldsymbol{x_i'b_j} + b_j^0$. Thus minimization of (A.3) with respect to the u_{ij} yields

$$u_{ij} = \left(\sum_{q=1}^{k}\left(\frac{D_{ij}}{D_{iq}}\right)^{\frac{1}{m-1}}\right)^{-1}.$$

If there exists j such that $\log(p_j f(y_i; \boldsymbol{x_i'b}_j + b_j^0, s_j^2)) \geq 0$, in order to maximize (4.3) a crisp assignment is required. To see that assume without loss of generality that

$$\log(p_1 f(y_i; \boldsymbol{x_i'b}_1 + b_1^0, s_1^2)) = \max_{j=1,2\ldots,k} \log(p_j f(y_i; \boldsymbol{x_i'b}_j + b_j^0, s_j^2)) > 0,$$

then the following holds

$$\sum_{j=1}^{k} u_{ij}^m \log\left(p_j f(y_i; \boldsymbol{x_i'b}_j + b_j^0, s_j^2))\right) \leq \log\left(p_1 f(y_i; \boldsymbol{x_i'b}_1 + b_1^0, s_1^2))\right) \sum_{j=1}^{k} u_{ij}^m$$

$$\leq \log\left(p_1 f(y_i; \boldsymbol{x_i'b}_1 + b_1^0, s_1^2))\right) \sum_{j=1}^{k} u_{ij} = \log\left(p_1 f(y_i; \boldsymbol{x_i'b}_1 + b_1^0, s_1^2)\right).$$

and thus the optimal solution is $u_{i1} = 1$ and $u_{ij} = 0$ for every $j \neq 1$.

Trimmed observations: Within our algorithm a fixed proportion α of observations is allowed to be discarded. It is straightforward to see that discarding the $\lceil n\alpha \rceil$ observations with lowest values of the quantity r_i defined in (4.6) maximizes our target function (4.3).

Parameter estimation: Conditionally on u_{ij} we then maximize (4.3) with respect to $p_j, b_j^0, \boldsymbol{b}_j$ and s_j^2.

It is straightforward to see that (4.4) is the optimal solution of (4.3) in terms of p_j.

To estimate b_j^0, and $\boldsymbol{b}_j$ a weighted least squares approach is required. The weights are the current u_{ij} values. Formally speaking the following minimization problem arises:

$$\min_{\beta_j^0 \in \mathbb{R},\, \boldsymbol{\beta}_j \in \mathbb{R}^p} \left[\sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^m (y_i - \beta_j^0 - \boldsymbol{x}_i' \boldsymbol{\beta}_j)^2 \right]$$

for which a closed form solution is given in (4.7).

Finally, to estimate the residual variances we need to solve

$$\min_{\sigma_1^2,\ldots,\sigma_k^2>0} \left[\sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^m \left(\frac{1}{2} \log(\sigma_j^2) + \frac{(y_i - b_j^0 - \boldsymbol{x}_i' \boldsymbol{b}_j)^2}{2\sigma_j^2} \right) \right] \tag{A.4}$$

that easily translates into the following problem:

$$\min_{\sigma_1^2,\ldots,\sigma_k^2>0} \left[\sum_{j=1}^{k} p_j \left(\log(\sigma_j^2) + \frac{d_j^2}{\sigma_j^2} \right) \right] \tag{A.5}$$

where in d_j is the j-th weighted residual variance component defined in (4.8) and the values $\sigma_1^2 \ldots, \sigma_k^2$ must satisfy the constraint $\sigma_j^2/\sigma_l^2 < c$ for every $j \neq l$. From (A.5), it is easy to see that the use of truncated residual variance components, as done in Section 4.2.1, is the optimal way of updating the s_j^2 parameters.

For sake of self-containedness, we show that the optimal threshold value can be obtained by evaluating $2k+1$ times function (4.10). This can be done by considering $e_1 \leq e_2 \leq \ldots \leq e_{2k}$ obtained after ordering $d_1^2, d_2^2, \ldots, d_k^2, d_1^2/c, d_2^2/c, \ldots, d_k^2/c$. Then, let us consider any $2k+1$ values $f_1, \ldots, f_{2k+1}$ satisfying $f_1 < e_1 \leq f_2 \leq e_2 \leq \ldots \leq f_{2k} \leq e_{2k} < f_{2k+1}$. The critical points of the auxiliary target function (4.10) are

$$t_i = \frac{\sum_{j=1}^{k} p_j \big(d_j^2 I\{d_j^2 < f_i\} + \frac{1}{c} d_j^2 I\{d_j^2 > c f_i\} \big)}{\sum_{j=1}^{k} p_j \big(I\{d_j^2 < f_i\} + I\{d_j^2 > c f_i\} \big)},$$

and, thus, these are the only $2k+1$ points that need to be evaluated.

A.3 A proposal for standardizing the residual component in Algorithm 6

It shall be noticed that calibrating the fuzzification parameter m may be cumbersome due to its sensibility w.r.t. the scale of the data. As outlined in Section 4.3.3 the user may be forced to change the measurement scale of the response variable in case the proportion of hard assignments is basically constant with respect to m. We now outline a standardization step which may be embedded within the algorithm and moderates the effect of the scale of the response variable on the computation of the fuzzy weights. Let us recall that the update is given by:

$$u_{ij} = \left(\sum_{q=1}^{k} \left(\frac{\log(p_j f(y_i; \boldsymbol{x_i'} \boldsymbol{b}_j + b_j^0, s_j^2)}{\log(p_j f(y_i; \boldsymbol{x_i'} \boldsymbol{b}_q + b_q^0, s_q^2)} \right)^{\frac{1}{m-1}} \right)^{-1}. \tag{A.6}$$

where $f(\cdot; \boldsymbol{b}_j + b_j^0, s_j^2)$ stands for the p.d.f of the Gaussian random variable centered at $\boldsymbol{x_i'} \boldsymbol{b}_j + b_j^0$ with variance equal to s_j^2.Let assume, without loss of generality, that $y \sim \mathcal{N}(\mu, \sigma^2)$, then if we multiply y by a fixed positive constant a it yields that $a \cdot y \sim \mathcal{N}(a\mu, a^2\sigma^2)$. It easy to see that the increase in the variance component is not linear as the increase in the mean of the variable. Thus, as a increases then the resulting density becomes closer and closer to a uniform density. As a result the required ratio needed for computing of the fuzzy weights of each cluster, formula (A.6), tends to be equal to $1/k$ for each observation. In such case the clusterwise struture of the data is completely hidden and the method is not able to estimate the different cluster centroids (which in our case are represented by the coefficients of a linear model). In order to overcome this issue we briefly outline here a standardization of the residual component s_j which moderates such undesired effect and is given by:

$$u_{ij} = \left(\sum_{q=1}^{k} \left(\frac{\log(p_j f(y_i/s_j^2; (\boldsymbol{x_i'} \boldsymbol{b}_j + b_j^0, 1)/s_j^2, 1)}{\log(p_j f(y_i/s_q^2; (\boldsymbol{x_i'} \boldsymbol{b}_q + b_q^0)/s_q^2, 1)} \right)^{\frac{1}{m-1}} \right)^{-1} \tag{A.7}$$

It shall be noticed that in formula (A.7) the same quantities of (A.6) are involved. The only difference is in the residual component since formula (A.7) involves studentized residuals which allow to contain the effect of the scale of the explanatory variable. Figure A.1 contains an application of the proposed standardization on a simulated dataset, as described in Section 4.3. As in Subsection 4.3.3, once we generated the data, we multiplied the values of the response variable by two fixed constants (i.e. 1000 and 1/1000) and reported the results. Figure A.1 clearly shows how the impact of the scale can be fully contained by standardizing the variables in the fuzzy weigths' computation. In particular, as the data are displayed on "high"

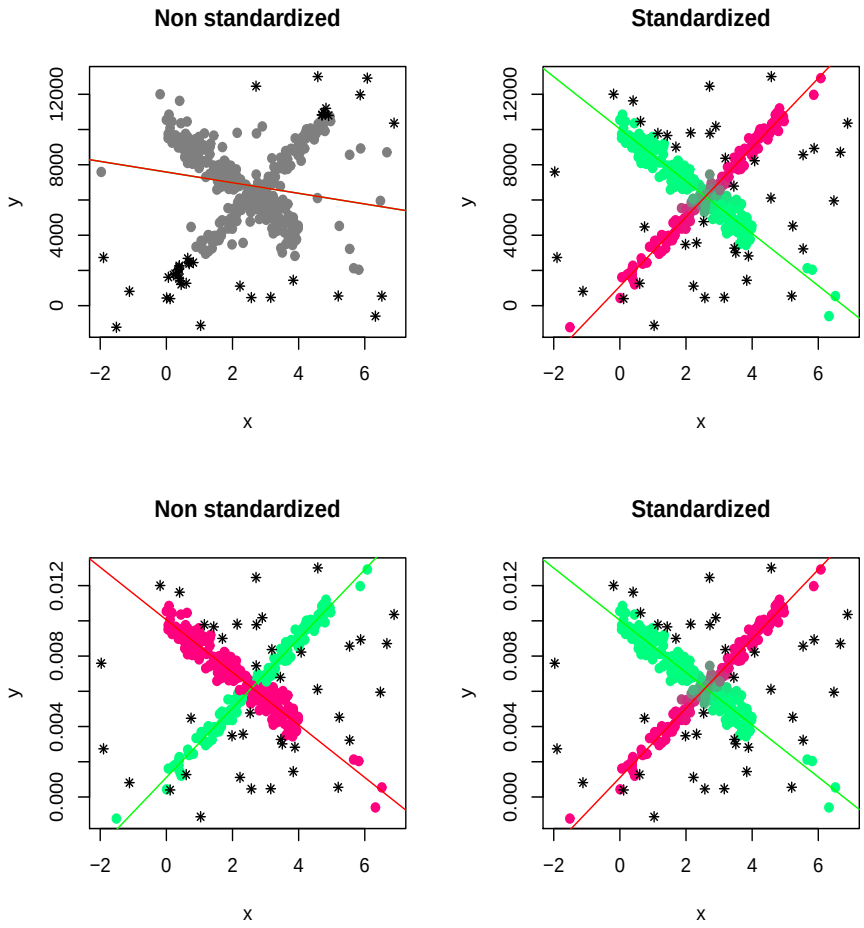

Figure A.1: Empirical Comparison of the impact of the scale of the data as before and after the standardization (A.7)

values for the scale parameters, then a complete fuzzification is obtained and no linear model can be corrected estimated. On the other hand, as the scale is very low, then, regardless to m, no fuzzification is reached. On the contrary, the plots on the right side show how such problems are completely overcome with studentized residuals.

Acknowledgements

La sezione dei ringraziamenti delle tesi ha generalmente assunto una connotazione del tutto particolare. Capita spesso di imbattersi in ringraziamenti dai toni altisonanti, a tratti elegiaci, dei propri direttori di tesi e dei loro "assistenti". Altrimenti non è raro imbattersi in sterminati elenchi di nomi con annessi aneddotti, talvolta imbarazzanti, che a me personalmente, hanno sempre ricordato, a tratti, la sezione dei saluti de "La Posta di Sonia" (i nati a fine anni '80 coglieranno la citazione), ed a tratti discorsi degni del ritiro di un premio Oscar.

Nel tentativo di non (s)cadere in nessuna di queste vorrei, come prima cosa, dire grazie ad Alessio, per due motivi. Il primo, forse scontato, è quello di avermi insegnato con pazienza inesauribile a fare ricerca, spedendomi, nel momento migliore, in quel di Valladolid. Il secondo risale a esattamente dieci anni fa: durante i suoi tutoraggi del corso di Statistica di Base introdusse il problema della robustezza statistica. Anche in una sua prima formalizzazione elementare il problema mi ha in qualche modo affascinato e ad oggi, grazie lui, posso provare a dare un mio (modesto) contributo a questa tanto dibattuta problematica di ricerca.

Mi sento poi in dovere di ringraziare diverse persone incontrate nei due dipartimenti in cui ho soggiornato come studente di dottorato, il Departamento de Estadistica e I.O de la Universidad de Valladolid ed il dipartimento di Scienze Statistiche di Roma, in particolare:

- Il Prof. Marco Perone Pacifico in primis, per avermi portato ad un primo piccolo traguardo e perchè non è facile trovare tanta competenza e tanta disponibilità in una persona sola.

- I due diversi coordinatori di dottorato che ho avuto: il Prof. Fulvio de Santis e il Prof. Pierluigi Conti, per aver cercato, sempre e comunque, di aiutare i dottorandi a muoversi nel difficile mondo della ricerca, talvolta, ulteriormente complicato, da interminabili scartoffie burocratiche.

- Il prof. Pierpaolo Brutti ed il suo inesauribile estro.
- Los profesóres Agustín Mayo-Iscar y Luis Angel García-Escudero porqué siempre me han dado confianza a pesar de las dificultades que se encontran en el trabajo y han hecho mi estadía en Valladolid interesante, cada día más.
- Il prof. Marco Alfó: un riferimento sereno e sicuro per me e per tanti altri studenti.

Sarebbe infine un sacrilegio non menzionare gli amici, i piú recenti e quelli di una vita, Alberto, Guido, Rostro, Macho, Ciuffo, Timo e casa de *i marci* in generale, Carota, Filippo, Dega, Marta; i colleghi, quelli delle stanze 40 e 41, e quelli conosciuti in dipartimento e naturalmente diventati degli amici, Luca, Manuel, Rosa, Giulia, Adele, Francesca (Rosa), Maria Francesca Marino, Francesca (Matano), sfuggita alle mie angherie in quel di Pittsburgh, Carla e suoi cornetti, e tra i giovani "veterani" del dipartimento Stefania e Maria Brigida. Grazie per aver reso lo studio, prima, ed il lavoro, poi, piú piacevole. Un affettuoso abbraccio non puó che andare alla grande famiglia di Catania e di Valeggio e ai miei genitori, a cui la tesi è dedicata.

Bibliography

Ali, A. M., Karmakar, G. C. & Dooley, L. S. (2008), 'Review on fuzzy clustering algorithms', *Journal of Advanced Computations* **2**, 169–181.

Atkinson, A. & Riani, M. (2012), *Robust diagnostic regression analysis*, Springer Science & Business Media.

Atkinson, A., Riani, M. & Cerioli, A. (2004), *Exploring multivariate data with the forward search*, Springer Series in Statistics, Springer, New York.

Bai, X. (2012), Robust linear regression, PhD thesis, Kansas State University.

Ballard, T., Kepple, A. & Cafiero, C. (2013), The food insecurity experience scale: developing a global standard for monitoring hunger worldwide, Technical report, Food and Agriculture Organization of the United Nations, Rome.

Banfield, J. & Raftery, A. (1993), 'Model-based Gaussian and non-Gaussian clustering', *Biometrics* **49**, 803–821.

Bezdek, J. C. (2013), *Pattern recognition with fuzzy objective function algorithms*, Springer Science & Business Media.

Butler, R., Davies, P. & Jhun, M. (1993), 'Asymptotics for the minimum covariance determinant estimator', *The Annals of Statistics* **21**, 1385–1400.

Cafiero, C., Melgar-Quinonez, H. R., Ballard, T. J. & Kepple, A. W. (2014), 'Validity and reliability of food security measures', *Annals of the New York academy of sciences* **1331**, 230–248.

Cafiero, C., Nord, M., Viviani, S., del Grossi, M. E., Ballard, T. J., Kepple, A. W., Miller, M. & Nwosu, C. (2016), Methods for estimating comparable rates of food insecurity experienced by adults throughout the world, Technical report, Food and Agriculture Organization of the United Nations, Rome.

Cator, E. A., Lopuhaä, H. P. et al. (2012), 'Central limit theorem and influence function for the mcd estimators at general multivariate distributions', *Bernoulli* **18**, 520–551.

Celeux, G. & Govaert, A. (1992), 'A classification EM algorithm for clustering and two stochastic versions', *Computational Statistics and Data Analysis* **14**, 315–332.

Celeux, G. & Govaert, G. (1995), 'Gaussian parsimonious clustering models', *Pattern recognition* **28**, 781–793.

Cerioli, A. (2010), 'Multivariate outlier detection with high-breakdown estimators', *Journal of the American Statistical Association* **105**, 147–156.

Cerioli, A. & Farcomeni, A. (2011), 'Error rates for multivariate outlier detection', *Computational Statistics and Data Analysis* **55**, 544–553.

Cerioli, A., Farcomeni, A. & Riani, M. (2013), 'Robust distances for outlier-free goodness-of-fit testing', *Computational Statistics & Data Analysis* **65**, 29–45.

Clarke, K. A. (2003), 'Nonparametric model discrimination in international relations', *Journal of Conflict Resolution* **47**(1), 72–93.

Coretto, P. & Hennig, C. (2013), 'A consistent and breakdown robust model-based clustering method', *arXiv preprint arXiv:1309.6895* .

Coretto, P. & Hennig, C. (in press 2016), 'Robust improper maximum likelihood: tuning, computation, and a comparison with other methods for robust gaussian clustering', *Journal of the American Statistical Association* .

Croux, C. & Haesbroeck, G. (1999), 'Influence function and efficiency of the minimum covariance determinant scatter matrix estimator', *Journal of Multivariate Analysis* **71**, 161–190.

Cuesta-Albertos, J., Gordaliza, A. & Matrán, C. (1997), 'Trimmed k-means: an attempt to robustify quantizers', *Annals of Statistics* **25**, 553–576.

Cuesta-Albertos, J., Matrán, C. & Mayo-Iscar, A. (2008*a*), 'Robust estimation in the normal mixture model based on robust clustering', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 779–802.

Cuesta-Albertos, J., Matran, C. & Mayo-Iscar, A. (2008*b*), 'stimation in the normal mixture model based on robust clustering', *J Roy Stat Soc, Ser. B* **70**, 779–802.

David, J. & David, W. (1974), 'Maximum likelihood estimates of the parameters of a mixture of two regression lines', *Communications in Statistics-Theory and Methods* **3**, 995–1006.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **39**, 1–38.

DeSarbo, W. S. & Cron, W. L. (1988), 'A maximum likelihood methodology for clusterwise linear regression', *Journal of classification* **5**, 249–282.

Donoho, D. & Huber, P. (1983), The notion of breakdown point, *in* 'A Festschrift for Erich L. Lehmann', Wadsworth Statist./Probab. Ser., Wadsworth, pp. 157–184.

Donoho, D. L. (1982), Breakdown properties of multivariate location estimators, Technical report, Harvard University, Boston.
http://statweb.stanford.edu/~donoho/Reports/Oldies/BPMLE.pdf.

Dotto, F. & Farcomeni, A. (n.d.), Robust inference for constrained model-based clustering. In Preparation.

Dotto, F., Farcomeni, A., García-Escudero, L. A. & Mayo-Iscar, A. (2015), The rtclust algorithm for robust clustering, *in* '10^{th} Scientific Meeting of the Classification and Data Analysis Group of the Italian Statstical Sociey. Book of Astract', CUEC.

Dotto, F., Farcomeni, A., García-Escudero, L. A. & Mayo-Iscar, A. (2016*a*), 'A fuzzy approach to robust regression clustering', *Advances in Data Analysis and Classification. (to appear)* .
http://link.springer.com/article/10.1007/s11634-016-0271-9.

Dotto, F., Farcomeni, A., García-Escudero, L. & Mayo-Iscar, A. (2016*b*), 'A reweighting approach to robust clustering', *Submitted* .

D'Urso, P., De Giovanni, L. & Massari, R. (2015), 'Trimmed fuzzy clustering for interval valued data', *Advances in Data Analysis and Classification* **9**, 21–40.

D'Urso, P., Massari, R. & Santoro, A. (2011), 'Robust fuzzy regression analysis', *Information Sciences* **181**, 4154–4174.

Farcomeni, A. (2009), 'Robust double clustering: a method based on alternating concentration steps', *Journal of classification* **26**, 77–101.

Farcomeni, A. (2014*a*), 'Robust constrained clustering in presence of entry-wise outliers', *Technometrics* **56**, 102–111.

Farcomeni, A. (2014*b*), 'Snipping for robust k-means clustering under component-wise contamination', *Statistics and Computing* **24**, 907–919.

Farcomeni, A. & Greco, L. (2015), *Robust Methods for Data Reduction*, CRC Press.

Farcomeni, A. & Ventura, L. (2012), 'An overview of robust methods in medical research', *Statistical Methods in Medical Research* **21**, 111–133.

Flury, B. & Riedwyl, H. (1988), *Multivariate Statistics. A Practical Approach*, Chapman and Hall, London.

Fraley, C. & Raftery, A. (1998), 'How many clusters?which clustering method? Answers via model-based cluster analysis', *The Computer Journal* **41**, 578–588.

Fraley, C. & Raftery, A. (2002), 'Model based clustering, discriminant analysis and density estimation', *Journal of the American Statistical Association* **97**, 611–631.

Fraley, C. & Raftery, A. (2012), *mclust: Model-Based Clustering / Normal Mixture Modeling*. R package version 3.4.11.
URL: *http://CRAN.R-project.org/package=mclust*

Fraley, C. & Raftery, A. E. (2007), 'Model based methods of classification: Using the `mclust` software in chemometrics', *Journal of Statistical Software* **6**.

Fritz, H., Garcıa-Escudero, L. A. & Mayo-Iscar, A. (2012*a*), 'tclust: An R package for a trimming approach to cluster analysis', *Journal of Statistical Software* **47**, 1–26.

Fritz, H., García-Escudero, L. A. & Mayo-Iscar, A. (2013*a*), 'Robust constrained fuzzy clustering', *Information Sciences* **245**, 38–52.

Fritz, H., García-Escudero, L. & Mayo-Iscar, A. (2012*b*), 'tclust: An R package for a trimming approach to cluster analysis', *J Stat Softw* **47**.
URL: *http://www.jstatsoft.org/v47/i12*

Fritz, H., García-Escudero, L. & Mayo-Iscar, A. (2013*b*), 'A fast algorithm for robust constrained clustering', *Computational Statistics and Data Analysis* **61**, 124–136.

Gallegos, M. (2002), Maximum likelihood clustering with outliers, *in* K. Jajuga, A. Sokolowski & H. Bock, eds, 'Classification, Clustering and Data Analysis: Recent advances and applications', Springer-Verlag, pp. 247–255.

Gallegos, M. & Ritter, G. (2005), 'A robust method for cluster analysis', *Annals of Statistics* **33**, 347–380.

Gallegos, M. & Ritter, G. (2009*a*), 'Trimmed ML estimation of contaminated mixtures', *Sankhya* **71**, 164–220.

Gallegos, M. & Ritter, G. (2009*b*), 'Trimming algorithms for clustering contaminated grouped data and their robustness', *Advances in Data Analysis and Classification* **3**, 135–167.

Gallup (2015), *Worldwide Research Methodology and Codebook*, Gallup, Inc., Washington, D.C.

García-Escudero, L. A. & Gordaliza, A. (1999), 'Robustness properties of k-means and trimmed k-means', *Journal of the American Statistical Association* **94**, 956–969.

García-Escudero, L. A., Gordaliza, A. & Matrán, C. (2012), 'Trimming tools in exploratory data analysis', *Journal of Computational and Graphical Statistics* **12**, 434–449.

García-Escudero, L. A., Gordaliza, A., Matrán, C. & Mayo-Iscar, A. (2015), 'Avoiding spurious local maximizers in mixture modeling', *Statistics and Computing* **25**, 619–633.

García-Escudero, L. A., Gordaliza, A., Mayo-Iscar, A. & San Martín, R. (2010), 'Robust clusterwise linear regression through trimming', *Computational Statistics and Data Analysis* **54**, 3057–3069.

García-Escudero, L. & Gordaliza, A. (2007), 'The importance of the scales in heterogeneous robust clustering', *Computational Statistics and Data Analysis* **51**, 4403–4412.

García-Escudero, L., Gordaliza, A., Matrán, C. & Mayo-Iscar, A. (2008), 'A general trimming approach to robust cluster analysis', *Annals of Statistics* **36**, 1324–1345.

García-Escudero, L., Gordaliza, A., Matrán, C. & Mayo-Iscar, A. (2010), 'A review of robust clustering methods', *Advances in Data Analysis and Classification* **4**, 89–109.

García-Escudero, L., Gordaliza, A., Matrán, C. & Mayo-Iscar, A. (2011), 'Exploring the number of groups in robust model-based clustering', *Statistics and Computing* **21**, 585–599.

García-Escudero, L., Gordaliza, A., San Martín, R., Van Aelst, S. & Zamar, R. (2009), 'Robust linear clustering', *Journal of the Royal Statical Society: Series B (Statstical Methdology)* **71**, 301–318.

Gath, I. & Geva, A. B. (1989), 'Unsupervised optimal fuzzy clustering', *IEEE Transactions on pattern analysis and machine intelligence* **11**, 773–780.

Gnanadesikan, R. & Kettenring, J. R. (1972), 'Robust estimates, residuals, and outlier detection with multiresponse data', *Biometrics* **28**, 81–124.

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, K., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M. & Toulmin, C. (2010), 'Food security: the challenge of feeding 9 billion people', *Science* **327**, 812–818.

Gustafson, D. & Kessel, W. (1978), 'Fuzzy clustering with a fuzzy covariance matrix." scientific systems', *Inc., Cambridge, MA* .

Hampel, F. R. (1974), 'The influence curve and its role in robust estimation', *Journal of the American Statistical Association* **69**, 383–393.

Hardin, J. & Rocke, D. (2004), 'Outlier detection in the multiple cluster setting using the Minimum Covariance Determinant estimator', *Computional Statistics and Data Analysis* **44**, 625–638.

Hardin, J. & Rocke, D. (2005), 'The distribution of robust distances', *Journal of Computational and Graphical Statistics* **14**, 928–946.

Hartigan, J. A. (1975), *Clustering algorithms*, Wiley Series in Probability and Mathematical Statistics.

Hathaway, R. J. (1985), 'A constrained formulation of maximum-likelihood estimation for normal mixture distributions', *The Annals of Statistics* **13**, 795–800.

Hathaway, R. J. & Bezdek, J. C. (1993), 'Switching regression models and fuzzy clustering', *IEEE Transactions on fuzzy systems* **1**, 195–204.

Hennig, C. (2003), 'Clusters, outliers and regression: fixed point clusters', *Journal of Multivariate Analysis* **83**, 183–212.

Hennig, C. (2004), 'Breakdown points for maximum likelihood-estimators of location-scale mixtures', *Ann Stat* **32**, 1313–1340.

Hennig, C. & Liao, T. F. (2013), 'How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **62**, 309–369.

Hennig, C., Meila, M., Murtagh, F. & Rocci, R. (2015), *Handbook of cluster analysis*, CRC Press.

Heritier, S., Cantoni, E., Copt, S. & Victoria-Feser, M.-P. (2009), *Robust methods in Biostatistics*, Vol. 825, John Wiley & Sons.

Honda, K., Ohyama, T., Ichihashi, H. & Notsu, A. (2008), Fcm-type switching regression with alternating least squares method, *in* 'Fuzzy Systems, 2008. FUZZ-IEEE 2008.(IEEE World Congress on Computational Intelligence). IEEE International Conference on', IEEE, pp. 122–127.

Huber, P. (1981), *Robust statistics.*, Wiley series in probability and mathematical statistics: probability and mathematical statistics, New York.

Huber, P. J. et al. (1964), 'Robust estimation of a location parameter', *Annals of Mathematical Statistics* **35**, 73–101.

Huber, P. & Ronchetti, E. (2009), *Robust statistics, second edition*, Wiley series in Probability and Statistics, New York.

Hubert, L. & Arabie, P. (1985), 'Comparing partitions', *Journal of classification* **2**, 193–218.

Ingrassia, S., Minotti, S. C. & Punzo, A. (2014), 'Model-based clustering via linear cluster-weighted models', *Computational Statistics & Data Analysis* **71**, 159–182.

Jones, A., Ngure, F., Pelto, G. & Young, S. (2013), 'What are we assessing when we measure food security? A compendium and review of current metrics', *Advances in Nutrition* **4**, 481–505.

Kaufman, L. & Rousseeuw, P. (1990), *Finding groups in data: An introduction to cluster analysis*, 9th edn, Wiley-Interscience.

Kim, J., Krishnapuram, R. & Davé, R. (1996), 'Application of the least trimmed squares technique to prototype-based clustering', *Pattern Recognition Letters* **17**, 633–641.

Leisch, F. (2006), 'A toolbox for k-centroids cluster analysis', *Computational statistics & data analysis* **51**, 526–544.

Lenstra, A. K., Lenstra, J., Kan, A. R. & Wansbeek, T. (1982), 'Two lines least squares', *North-Holland Mathematics Studies* **66**, 201–211.

Liu, R. Y., Parelius, J. M. & Singh, K. (1999), 'Multivariate analysis by data depth: descriptive statistics, graphics and inference', *The Annals of Statistics* **27**, 783–858.

Lopuhaa, H. P. (1999), 'Asymptotics of reweighted estimators of multivariate location and scatter', *The Annals of Statistics* **27**, 1638–1665.

MacQueen, J. et al. (1967), 'Some methods for classification and analysis of multivariate observations', *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* **1**, 281–297.

Maronna, R. & Jacovkis, P. M. (1974), 'Multivariate clustering procedures with variable metrics', *Biometrics* **30**, 499–505.

McLachlan, G. & Peel, D. (2000), *Finite mixture models*, Wiley Series in Probability and Statistics, New York.

McLachlan, G. & Peel, D. (2004), *Finite mixture models*, John Wiley & Sons.

Neykov, N., Filzmoser, P., Dimova, R. & Neytchev, P. (2007), 'Robust fitting of mixtures using the trimmed likelihood estimator', *Computational Statics and Data Analysis* **52**, 299–308.

Perry, P. O. (2009), 'Cross-validation for unsupervised learning', *arXiv preprint arXiv:0909.3052* .

Ritter, G. (2014), *Robust cluster analysis and variable selection*, CRC Press.

Rousseeuw, P. J. (1984), 'Least median of squares regression', *Journal of the American statistical association* **79**, 871–880.

Rousseeuw, P. J. (1985), 'Multivariate estimation with high breakdown point', *Mathematical Statistics and applications* **8**, 283–297.

Rousseeuw, P. J. & van Driessen, K. (1999), 'A fast algorithm for the minimum covariance determinant estimator', *Technometrics* **41**, 212–223.

Rousseeuw, P., Kaufman, L. & Trauwaert, E. (1996), 'Fuzzy clustering using scatter matrices', *Computational Statistics & Data Analysis* **23**, 135–151.

Ruckdeschel, P. & Horbenko, N. (2012), 'Yet another breakdown point notion: Efsbp', *Metrika* **75**, 1025–1047.

Ruspini, E. H. (1969), 'A new approach to clustering', *Information and control* **15**, 22–32.

Ruwet, C., García-Escudero, L., Gordaliza, A. & Mayo-Iscar, A. (2012), 'The influence function of the TCLUST robust clustering procedure', *Advances in Data Analysis and Classification* **6**, 107–130.

Ruwet, C., García-Escudero, L., Gordaliza, A. & Mayo-Iscar, A. (2013), 'On the breakdown behavior of robust constrained clustering procedures', *TEST* **22**, 466–487.

Sadaaki, M. & Masao, M. (1997), Fuzzy c-means as a regularization and maximum entropy approach, *in* 'Proceedings of the 7^{th} International Fuzzy Systems Association World Congress (IFSA 1997)'.

Späth, H. (1982), 'A fast algorithm for clusterwise linear regression', *Computing* **29**, 175–181.

Spiliopoulou, M., Kruse, R., Borgelt, C., Nürnberger, A. & Gaul, W. (2006), *From data and information analysis to knowledge engineering: proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation eV, University of Magdeburg, March 9-11, 2005*, Springer Science & Business Media.

Stahel, W. A. (1981), *Robuste schätzungen: infinitesimale optimalität und schätzungen von kovarianzmatrizen*, Eidgenössische Technische Hochschule [ETH] Zürich.

Trauwaert, E., Kaufman, L. & Rousseeuw, P. (1991), 'Fuzzy clustering algorithms based on the maximum likelihood priciple', *Fuzzy Sets and Systems* **42**, 213–227.

Tukey, J. W. (1962), 'The future of data analysis', *Annals of Mathematical Statistics* **33**, 1–67.

Van Aelst, S. & Rousseeuw, P. (2009), 'Minimum volume ellipsoid', *Wiley Interdisciplinary Reviews: Computational Statistics* **1**, 71–82.

Van Aelst, S., Wang, X. S., Zamar, R. H. & Zhu, R. (2006), 'Linear grouping using orthogonal regression', *Computational Statistics & Data Analysis* **50**.

van der Vaart, A., V. & Wellner, J., A. (1997), *Weak Convergence and Empirical Processes*, Springer, New York.

Vuong, Q. H. (1989), 'Likelihood ratio tests for model selection and non-nested hypotheses', *Econometrica: Journal of the Econometric Society* pp. 307–333.

Wu, K.-L., Yang, M.-S. & Hsieh, J.-N. (2009), Alternative fuzzy switching regression, *in* 'International Multi-Conference of Engineers and Computer Scientists (IMECS 2009) Vol. I'.

Yao, W. & Li, L. (2014), 'A new regression model: modal linear regression', *Scandinavian Journal of Statistics* **41**, 656–671.