A Fuzzy Approach to Robust Clusterwise Regression

Francesco Dotto[1], Alessio Farcomeni[2], Luis Angel García-Escudero[3], and Agustín Mayo-Iscar[3]

[1]Dipartimento di Scienze Statistiche, Sapienza Università di Roma

[2]Dipartimento di Sanità Pubblica e Malattie Infettive, Sapienza Università di Roma

[3]Departamento de Estadística e Investigación Operativa, Universidad de Valladolid. Valladolid, Spain

Abstract

A new robust fuzzy linear clustering method is proposed. We estimate coefficients of a linear regression model in each unknown cluster. Our method aims to achieve robustness by trimming a fixed proportion of observations. Assignments to clusters are fuzzy: observations contribute to estimates in more than one single cluster. We describe general criteria for tuning the method. The proposed method seems to be robust with respect to different types of contamination.

1 Introduction

Many clustering methods are based on estimate k groups of units according to the distance to suitably defined centroids. Oftentimes these centers are simply location parameters while in certain applications a different notion of cluster is of interest. For example, in linear clustering models, k groups of units forming a linear structure are taken in account, which 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 [18], that applied this type of procedure in economics, in [21] where these type of procedures are applied in marketing segmentation, and in [27], where the details about a feasible algorithm are provided. In [26] 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.

Our aim it to provide a fuzzy linear clustering method that is robust (see e.g. r51 for a detailed review of the main ideas in robust clustering). We focus on extending the TCLUST approach. This robust clustering method, that has been proposed in [10], allows the user to recover heteroscedastic structures and used trimming to avoid the deleterious effect of outlying observations. The extension of this procedure for linear clustering problems appeared in [11] and, following this idea, we are now interested in extending the robust linear clustering model proposed in [11] using the fuzzy approach. A review of robust regression can be found for instance in [16] and [6].

Fuzziness in clustering, that has been introduced in [26] and extended in [14], allows intermediate degrees of membership for each observation. It has several advantages in many applications. In some cases, e.g., [14] or [1], it is not possible to define meaningful hard partitions. Our robust fuzzy linear clustering model, based on trimming, also can be seen as an extension of the method in Hathaway and Bezdeck (see [15] for details). We use a likelihood based approach with trimming and constraints on the estimated residual variances.

The outline of the paper is as follows: in Section 2 we provide a motivation of the methodology and introduce the notation while in Section 3 we report the details of the algorithm and its justification. In Section 4 we illustrate how to choose tuning parameters based on examples. In Section 5 we report a simulation study. Finally Section 6 contains an application of our procedure to a real dataset and in Section 7 there are the concluding remarks and the proposal for further direction of research.

2 Methodology

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 the values taken by the p explanatory variables and $y_i \in \mathbb{R}$ is a continuous response variable for the individual i. Let us suppose to be interested in grouping them into k clusters in a fuzzy way and estimating a linear model in 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. Note that intermediate degrees of membership are allowed when $u_{ij} \in (0,1)$, that implies that an observation gives a contribution in the estimation of the parameters in each cluster which is proportional to its membership value u_{ij}. Secondly 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 off if $u_{ij} = 0$ for all $j = 1, ..., k$ and, thus, this observation has no membership contribution to any cluster. Let: $\alpha \in [0,1)$ be the fixed trimming proportion, $c \geq 1$ a fixed constant, $m > 1$ a fixed value of the fuzzifier parameter, $f(\cdot; \mu, \sigma^2)$ the p.d.f of a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean equal to μ and standard deviation equal to σ. A robust constrained fuzzy linear clustering problem can be defined through the maximization of the objective function

$$\sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^m \log\left(f(y_i; \boldsymbol{x_i'}\boldsymbol{b}_j + b_j^0, s_j^2)\right) \tag{2.1}$$

where 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 and $s_1^2, \ldots, s_k^2$ are the residual variances which obey to the constraint:

$$\frac{\max_{j=1}^k s_j^2}{\min_{j=1}^k s_j^2} \leq c. \tag{2.2}$$

Notice that $u_{i1} = ... = u_{ik} = 0$ for all $i \notin \mathcal{I}$, so these observations do not contribute to the summation in the target function (2.1) and these are the trimmed observations.

Using a maximum likelihood criterium like that one in (2.1) implies fixing a specific underlying linear model, which indeed allows us to better understand what the fuzzy clustering method is really aimed at. This maximum likelihood approaches has already been considered, among others, in [15], [20] and [17].

Another important issue in this approach is to consider, as already done in [11], the constraint given in (2.2). In fact, since the target function (2.1) is unbounded, the attempt of maximizing it without a constraint makes it a mathematically ill posed problem. For instance we can see that if we take an observation 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 (2.1) 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 usage of an objective function like that in (2.1), which does not take in account the effect of some additional clusters' weights p_j, lends the method a bias toward clusters with similar sizes (or more precisely toward a solution with similar values of $\sum_{j=1}^k u_{ij}^m$). Then equation (2.1) may be replaced by:

$$\sum_{i=1}^n \sum_{j=1}^k u_{ij}^m \log\left(p_j f(y_i; \boldsymbol{x_i'}\boldsymbol{b}_j + b_j^0, s_j^2)\right) \tag{2.3}$$

where $p_j \in [0, 1]$ and $\sum_{j=1}^k p_j = 1$ are some weights that the objective function also needs to be maximized on. Notice that, once the membership values are known, these weights are optimally determined as $p_j = \sum_{i=1}^n u_{ij}^m / \sum_{i=1}^n \sum_{j=1}^k u_{ij}^m$. Thus, this approach implies adding the term

$$\sum_{j=1}^k \left(\sum_{i=1}^n u_{ij}^m\right) \log\left(\sum_{i=1}^n u_{ij}^m \Big/ \sum_{i=1}^n \sum_{j=1}^k u_{ij}^m\right)$$

to the target function (2.1). This type of regularization is related to the “entropy regularization” and appeared in [23]. Note that our objective function (2.3) allow us to consider the same problem considered in the F-TCLUST approach [7] but in our case, instead of estimating robust clusters’ location and scatter parameters in a fuzzy way, we aim to estimate the parameters yielding a linear model in each fuzzy cluster. Moreover, setting $m = 1$, we can state that our procedure is equivalent to the one proposed in [11] without the second trimming step.

3 The algorithm

Maximization of (2.3) is of course not an easy task. In fact, an iterative procedure which is able to take in account all the imposed constraints is required. In the following we propose a computationally feasible algorithm which is supposed to be initialized with several random starting points and iterates within two alternating steps up to convergence or until a maximum value of iterations is reached. First, given the values of the parameter at each iteration, the membership values are obtained. Secondly, given the membership values computed in the previous step, the parameters are updated in order to maximize (2.3). Therefore we propose, in order to estimate the regression parameter in all the groups, the adaptation of an EM-type procedure like those one used in [3] that often appear when fitting finite mixture of observations [see, e.g., 22]. In any case we can see that the updating formulas for the membership values and for the parameters are similar to those applied in other fuzzy linear clustering algorithms like the one proposed in [15], [20] and [17]. Note that, like the whole procedure may be viewed as an adaptation of [7] for linear clustering, also the algorithm presents several analogies like, e.g., the application of the constraint (2.2) when the residual variances are computed.

3.1 The proposed algorithm

In this subsection the proposed algorithm is described and briefly commented, while a stepwise justification follows in the next section.

1 *Initialization:* The procedure is initialized several times by taking random starting values $p_1, \dots, p_k, b^0_1, \dots, b^0_k, \boldsymbol{b}_1, \dots, \boldsymbol{b}_k, s^2_1, \dots, s^2_k$. In order to provide these preliminaries estimation k subsets made of $p+1$ observation in general position are randomly selected and then, the linear systems for estimating the coefficients are resolved and the associated residual variances are computed. Not that, if required, the estimated residual variances $s^2_1, \dots, s^2_k$ are constraint in order to satisfy the relation (2.2)

2 *Iterative steps:* The following steps are executed until convergence or until a maximum number of iterations is reached.

2.1 *Update membership values:* Using the current parameter estimates we update the membership values using following criterium. If

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

then

$$u_{ij} = I\big\{p_j f(y_i; \boldsymbol{x_i'}\boldsymbol{b}_j + b^0_j, s^2_j) = \max_{q=1,\dots,k} p_q f(y_i; \boldsymbol{x_i'}\boldsymbol{b}_j + b^0_j, s^2_j)\big\} \text{ (hard assignment)}$$

with $I\{\cdot\}$ being a 0-1 indicator function which takes the value 1 if the expression within the brackets holds. If

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

then

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

2.2. *Trimmed observations:* Let

$$r_i = \sum_{j=1}^{k} u_{ij}^{m} \log(p_j f(y_i; \boldsymbol{x_i'}\boldsymbol{b}_j + b_j^0, s_j^2)) \tag{3.1}$$

and $r_{(1)} \leq r_{(2)} \leq ... \leq r_{(n)}$ be these values sorted. The observations to be trimmed are, at this stage of the algorithm, those with indexes $\{i : r_i < r_{(\lceil n\alpha \rceil)}\}$. The membership values for those observations are redefined as $u_{ij} = 0$, for every j, if $r_i < r_{([n\alpha])}$.

2.3 *Update parameters:* Given the membership values obtained in the previous steps, the cluster weights p_j are updated as:

$$p_j = \frac{\sum_{i=1}^{n} u_{ij}^{m}}{\sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^{m}}, \tag{3.2}$$

while for b_j^0 and $\boldsymbol{b}_j$, with $j = 1, 2, \ldots, k$, the usual (weighted) least square method is used. Indeed the closed forms for them are

$$\boldsymbol{b_j} = \frac{\frac{\sum_{i=1}^{n} u_{ij}^{m},\boldsymbol{x}'_i}{\sum_{i=1}^{n} u_{ij}^{m}} - \frac{\sum_{i=1}^{n} u_{ij} y_i}{\sum_{i=1}^{n} u_{ij}^{m}} \cdot \frac{\sum_{i=1}^{n} u_{ij}^{m} \boldsymbol{x}'_i}{\sum_{i=1}^{n} u_{ij}^{m}}}{\frac{\sum_{i=1}^{n} u_{ij}^{m} \boldsymbol{x}'_i \boldsymbol{x}_i}{\sum_{i=1}^{n} u_{ij}^{m}} - \left(\frac{\sum_{i=1}^{n} u_{ij}^{m} \boldsymbol{x}'_i}{\sum_{i=1}^{n} u_{ij}} \right)^2}$$

$$\text{and } b_j^0 = \frac{\sum_{i=1}^{n} u_{ij}^{m} y_i}{\sum_{i=1}^{n} u_{ij}^{m}} - \boldsymbol{b}'_j \frac{\sum_{i=1}^{n} u_{ij}^{m} \boldsymbol{x}_i}{\sum_{i=1}^{n} u_{ij}^{m}}. \tag{3.3}$$

Updating the s_j^2 parameters may be more complicate since the weighted sample residual variances used to update them, given by

$$d_j^2 = \frac{\sum_{i=1}^{n} u_{ij}^{m} (y_i - b_j^0 - \boldsymbol{x_i'}\boldsymbol{b_j})^2}{\sum_{i=1}^{n} u_{ij}^{m}}, \tag{3.4}$$

may not obey to the constraint given by (2.2). In that case, using the procedure used in [7] for constraining the eigenvalues of the scatter matrices of each cluster is needed. In order to do that consider the j-th residual variance component d_j and its truncated value which is given by

$$[d_j^2]_t = \begin{cases} d_j^2 & \text{if } d_j^2 \in [t, ct] \\ t & \text{if } d_j^2 < t \\ ct & \text{if } d_j^2 > ct \end{cases}, \tag{3.5}$$

with t being a threshold value (note that these truncated residual components do satisfy the required scatter constraint). To define the optimal residual variance component, an optimal threshold value t_{opt} is obtained by taking into account the aim of maximizing the target function (2.3). For instance it can be shown that t_{opt} is the value of t that minimizes the real-valued function:

$$t \mapsto \sum_{j=1}^{k} p_j \left(\log \left([d_j^2]_t \right) + \frac{d_j^2}{[d_j^2]_t} \right), \tag{3.6}$$

with p_j as defined in (3.2). A closed form to get the value t_{opt} exists: indeed (3.6) must be evaluated in the $2k+1$ points, which are properly defined and calculated in [8] and [7]. Once these optimal values are properly computed for each cluster, then the estimated weighted residual variance s_j^2 can be finally be updated computing $s_j^2 = [d_j^2]_{t_{opt}}$.

3 *Evaluate the objective function:* At the end of the iterative process (i.e. when convergence or a maximum number of iteration is reached) evaluate the value of the associated target function (2.3). The parameter set yielding the highest value of the objective function is returned as the output of the algorithm.

3.2 Justification of the algorithm

It can be shown that the two iterative steps of the algorithm increase the value of the target function (2.3). In one of them, given the regression parameters, we search the membership values that maximize the target function, while in other one, given the membership values we estimate the parameter regression and the constraint residual variance parameters that maximize (2.3). Through this iterative process the algorithm ends up in a local maximum of (2.3) and by the usage of many random starting points we aim to find the global maximum of the target function. The rest of this section aims to justify previous claims:

Membership values: Let us assume that the regression parameters and the $\boldsymbol{b}_j, b^0$ and s_j for $j =$

$1,\dots,k$ are known. Maximizing (2.3) is equivalent to minimize

$$\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^{m} D_{ij} \tag{3.7}$$

where $D_{ij} = -\log(p_j f(y_i; \boldsymbol{x}'_{\boldsymbol{i}}\boldsymbol{b}_j + b_j^0, s_j^2)) = \log\left[p_j^{-1}(2\pi s_j^2)^{1/2}\exp\left((y_i - \boldsymbol{x}'_{\boldsymbol{i}}\boldsymbol{b}_j - b_j^0)^2/(2s_j^2)\right)\right]$. If $p_j f(y_i; \boldsymbol{x}'_{\boldsymbol{i}}\boldsymbol{b}_j + b_j^0, s_j^2) < 1$ is satisfied for every observation i then the quantity $D_{ij} (> 0)$ can be seen as a measure of the distance of the observed value of the response variable y_i from its fitted value $\boldsymbol{x}'_{\boldsymbol{i}}\boldsymbol{b}_{\boldsymbol{j}} + b_j^0$. Thus, by the usage of the standard Lagrange multiplier, minimization of (3.7) with respect to the u_{ij}, yields the following optimal membership values

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

which coincide with the membership updates of the fuzzy assignments. On the other hand, if there exist a j such that $p_j f(y_i; \boldsymbol{x}'_{\boldsymbol{i}}\boldsymbol{b}_j + b_j^0, s_j^2) \geq 1$, that is equivalent to $\log(p_j f(y_i; \boldsymbol{x}'_{\boldsymbol{i}}\boldsymbol{b}_j + b_j^0, s_j^2)) \geq 0$ then, in order to maximize (2.3), the crisp assignment proposed is required. To show that, let us assume that

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

then the following holds

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

and thus we have to impose $u_{i1} = 1$ 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 $\lceil n\alpha \rceil$ observation with lowest values of the quantity r_i defined in (3.1) maximizes our target function (2.3), (indeed our target function (2.3) may be defined in an alternative way as $\sum_{i=1}^{n} r_i$). This approach to trimming is the basis of the "concentration step" that can be found in many other high breakdown point robust statistical procedures (see e.g. [25]).

Parameter estimation: In the parameter estimation step the membership values u_{ij} are supposed to be known from the previous "membership values" and the algorithm aims to maximize (2.3) with respect to parameters $p_j, b_j^0, \boldsymbol{b}_j$ and s_j^2.

- *Clusters weights:* Regarding in the p_j's values, it's straightforward to see that the quantity proposed in (3.2) is the one needed to maximize (2.3).

- *Coefficient's estimation:* In order to estimate the parameters b_j^0, and $\boldsymbol{b}_j$, the minimization of a weighted least squared function is required. Indeed it is straightforward to see that minimizing the weighted sum of squared residuals between the observed values of the response variable and the predicted values in each clusters is needed. The weights, that in this step are considered as known, are represented by the fuzzy assignments of each observation to each cluster. Formally speaking the required coefficients are the result of the following minimization:

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

  for which a closed form for the minimizers is given in (3.3).

- *Residual variance estimation:* Once that an estimation of $p_j, b_j^0, \boldsymbol{b}_j^0$ is provided by applying (3.2) and (3.3) then our aim is to solve the minimization:

  $$\min_{\sigma_1^2, \ldots, \sigma_k^2 > 0} \left[\sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^m \left(\frac{1}{2} \log(\sigma_j^2) + \frac{(y_i - b_j^0 - \boldsymbol{x_i'b_j})^2}{2\sigma_j^2} \right) \right] \tag{3.8}$$

  that easily translates into the following problem:

  $$\min_{\sigma_1^2, \ldots, \sigma_k^2 > 0} \left[\sum_{j=1}^{k} p_j \left(\log(\sigma_j^2) + \frac{d_j^2}{\sigma_j^2} \right) \right] \tag{3.9}$$

  where in d_j is the j-th weighted residual variance component defined in (3.2) and the values $\sigma_1^2 \ldots, \sigma_k^2$ must satisfy the constraint $\sigma_j^2 / \sigma_l^2 < c$ for every $j \neq l$.

4 Choosing parameters

The methodology described was designed to be as general possible, in order to be useful in many different applications. A consequence is that there are four choices that the user has to make: 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 clustering).

In this section we investigate the role and the importance of each choice, and give some guidelines on how to perform each decision in practice. We do so by example, through an illustration based on a simple simulated dataset. 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 $y_{1i} = 1 + 2x_{1i} + \varepsilon_{1i}$ 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: $y_{2i} = 10 - 1.5x_{2i} + \varepsilon_{2i}$ 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.

We now discuss each tuning parameter separately.

4.1 Fuzzifier Parameter

The fuzzifier parameter m in equation (2.3) takes values in the range $[1, +\infty)$ and it regulates the degree of fuzziness of the final clustering. Letting $m \longrightarrow \infty$ implies equal membership values $u_{ij} = 1/k$ for every i and for every j, 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. The optimal value of m depends, as intuitive, on the degree of overlap among clusters and on how much the researcher is prepared to accept fuzzy weights. Additionally, a suitable number of hard assignments (or approximately hard assignments) are needed to form a "cluster core", for obvious reasons of cluster identification and interpretation. Finally, the optimal value of m also depends on the scale of the response variable. This issue is a common problem that affects all the likelihood based fuzzy clustering algorithm like, e.g. the proposal that appeared in [28] and [24], and at our knowledge, has not been noted so far in the literature with the exception of [7]. This fact makes it quite difficult to find reasonable general criteria for choosing m. We illustrate and propose a heuristic procedure in the following.

In Figure 1 we applied our procedure to data 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: $m = 1.5$, that is a standard value, and $m = 1$ that implies no fuzzification in the model. Among the plots that appear in this Section, in order to graphically represent the fuzzification, we used a mixture of "red" and "green" colors with intensities proportional to the membership values of each observations. Additionally, through all the paper, points flagged as outlying under the model have been represented by "$*$" .

Figure 1 shows that when $m > 1$ the scale of the response variable leads to changes in the results, while when $m = 1$ results are scale independent.

Our proposal for choosing m in practice is to monitor, for each value of m in a pre-specified grid, two quantities: the proportion of hard assignments and the relative entropy of the fuzzy weights. The proportion of hard assignments is monitored in order to tune m in order to obtain a reasonable cluster core for each cluster. The relative entropy measures residual uncertainty in cluster assignments, and is computed as

$$\frac{\sum_{j=1}^{k} \sum_{i=1}^{n} u_{ij} \log u_{ij}}{[n(1-\alpha)] \log(k)}. \tag{4.1}$$

Figure 2 shows the proportion of hard assignments and the relative entropy of the fuzzy weights as a function of m in our simulated example. We did so by repeatedly applying our procedure for different values of m. We propose to choose m in order to balance these two quantities. Figure 2 is particularly useful in order to avoid extreme situations where there are zero hard assignments and explore the underlying degree of overlap.

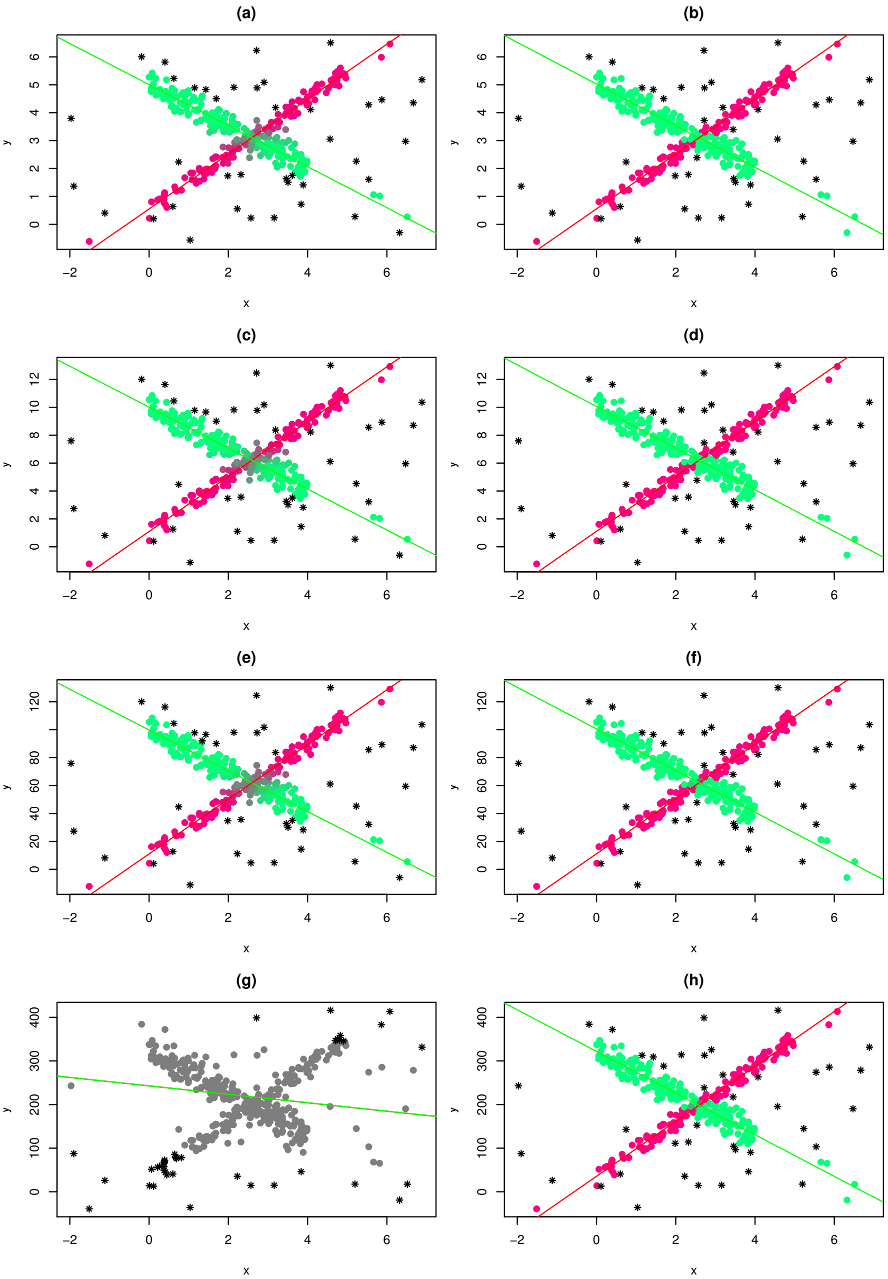

Figure 1: Different degrees of fuzzification obtained for different scale values $s \in \mathbb{R}^+$ (see the value on the y axis). (a) $m = 1.5$, $s = 0.5$. (b) $m = 1$, $s = 0.5$. (c) $m = 1.5$, $s = 1$. (d) $s = 1$, $s = 1$. (e) $m = 1.5$, $s = 10$. (f) $m = 1$, $s = 10$. (g) $m = 1.5$, $s = 32$. (h) $m = 1$, $s = 32$.

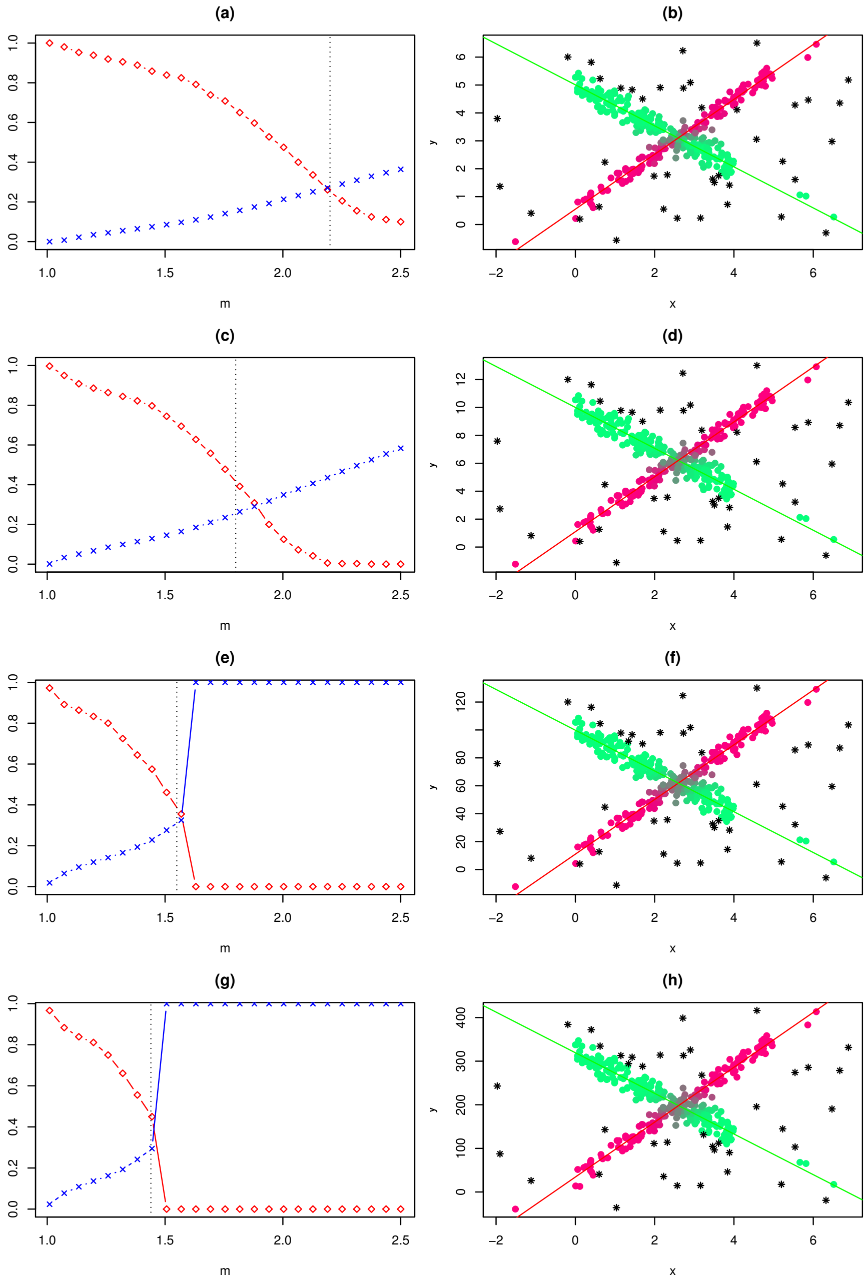

Figure 2: Left panels: relative entropy of the fuzzy weights, “×”, proportion of hard assignments, “◇”, as a function of scale; (a) $s = 0.5$. (c) $s = 1$ (e) $s = 10$. (g) $s = 32$. Right panels: clustering obtained for specific values of m through (b) $s = 0.5$, $m = 2.2$. (d) $s = 1$, $m = 1.8$. (f) $s = 10$, $m = 1.6$. (h) $s = 32$, $m = 1.4$.

4.2 Including clusters' weights

In equation (2.1) clusters' weights have been included. A possibility is to exclude these weights, with the consequence of shrinking assignments towards an equal number of observations within each cluster. This is also particularly relevant when the number of clusters is possibly misspecified.

In Figure 3 we represented a scenario where there are 40% of the clean observations in one cluster, 60% in the other one, and we estimated $k = 3$ clusters. We compare the performance of the model when the weights are kept in account in the likelihood function, like in equation (2.3), and when the weights do not appear in the objective function, like in (2.1).

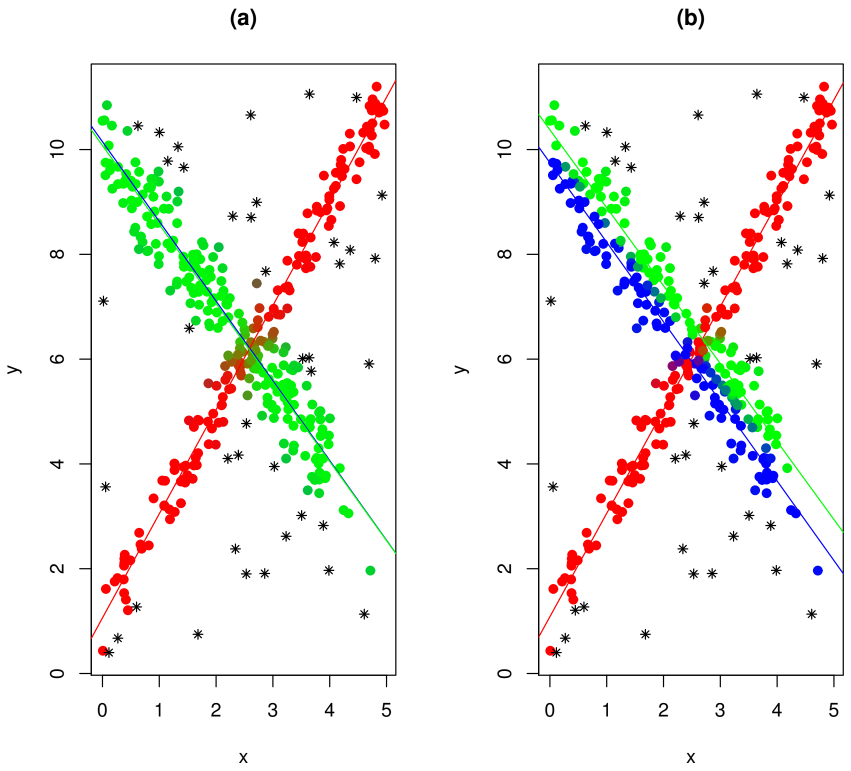

Figure 3: (a): estimated robust fuzzy clustering when $k = 3$ and p_j are used within the objective function. (b): estimated robust fuzzy clustering $k = 3$ and p_j are not use within the objective function.

When the weights p_j are taken into account our model is able to recover the real structure of the data, indeed, the estimated weights are equal to $0.01, 0.42, 0.57$ (note that one of them is really close to 0). Note also that two of the three estimated regression lines are overlapped. On the other hand, when weights are not used estimates of p_j correspond to $0.41, 0.30$ and 0.29, which are clearly biased towards an equal balanced clusters scenario and two almost parallel clusters are recovered.

4.3 Constraint in the residual variance

An important feature of the proposed algorithm is that no homoscedasticity assumption is made, or necessary, at all. This is a novel feature as in many switching regression models (e.g., [15] and [20]) where the residual variance is not kept into account, and implicitly or explicitly, clusters are assumed to be homoscedastic.

In order to give a brief illustration of how much bias might be obtained by assuming homoscedastic residuals when this assumption does not hold, we compare in Figure 4 estimates based on $c = 1$ and $c = 5$ in two different heteroscedastic scenarios. Recall that fixing $c = 1$ one forces the residual variances to be equal to each other.

In Figure 4, plot (a), where $c = 5$, variances are correctly estimated and classification is very good as only 18 observations out of 360 are wrongly classified. In the other hand, in 4, plot (b), we run the procedure again on the same data but after fixing $c = 1$. As could be expected, variance estimates are pooled. A consequence is that three additional observations are misclassified.

In panels 4, plot (c) and (d), we repeated this experiment, but with an increased difference in the underlying variances. In Figure 4, plot (c), where $c = 5$, we still have 18 misclassified observations. In Figure 4, plot (d), where $c = 1$, we have 32 misclassified observations.

One would be tempted to set a large value for the constraint limit c, but large values might be associated with spurious maximizers. In the following example we add 20 collinear outliers that form a spurious cluster and then implement our model with $\alpha = 0.05$ (that is, requiring that exactly 20 observations are trimmed). We do this twice: once constraining ratio of the

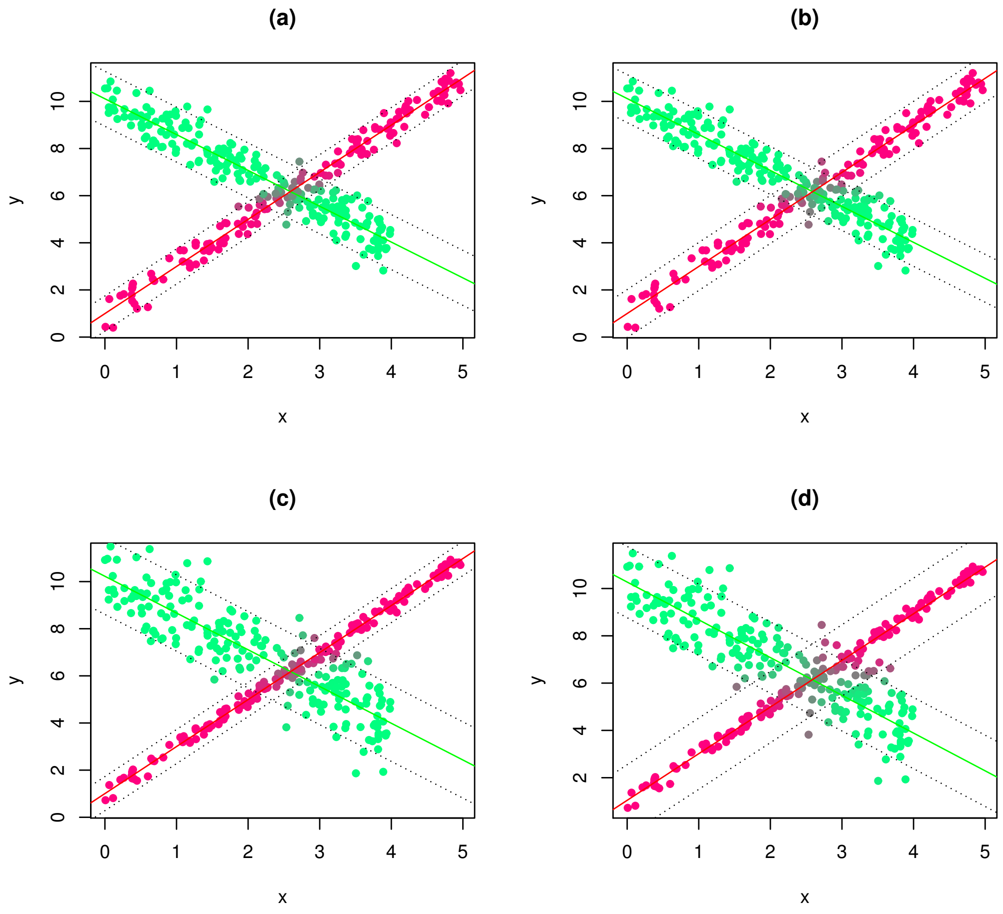

Figure 4: Estimated robust fuzzy clustering for different values of c, σ_1 and σ_2. (a) $c = 5$, $\sigma_1 = 0.4$ and $\sigma_2 = 0.6$. (b) $c = 1$, $\sigma_1 = 0.4$ and $\sigma_2 = 0.6$. (c) $c = 5$, $\sigma_1 = 0.2$ and $\sigma_2 = 1$. (d) $c = 1$, $\sigma_1 = 0.2$ and $\sigma_2 = 1$. The bands are obtained by adding $\pm 2\sigma_j$ to each obtained fitted regression line.

estimated residual variance to be lower than 5 and once with $c = 10^{10}$ (basically, without constraints).

We report the results in Figure 5, where in panel (b), with $c = 10^{10}$, the outliers are not discarded. Notably, small groups of collinear observations give an unbounded contribution to the likelihood if these pathologic solutions are not discarded in advance through appropriate constraints.

One could argue that the set of the 20 collinear points may be considered as a cluster and

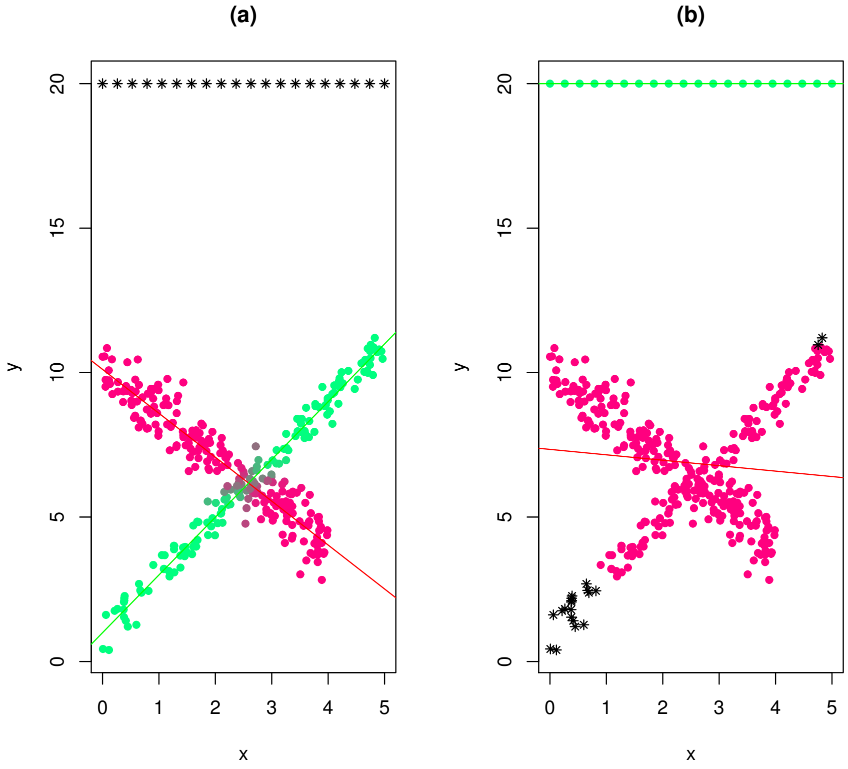

Figure 5: (a): FTCR with $c = 5$ and $\alpha = 0.05$. (b): FTCR with $c = 10^{10}$ and $\alpha = 0.05$.

thus these observation may not cause a problem in the estimation of the linear structure as k is properly increased. Nevertheless, since we are assuming that k is fixed in advance, and we are not aware of the presence of noisy observations, we are mainly interested in providing the best available solution with respect to the fixed number of groups k. Additionally we can say that observations having such perfect linear behavior are more likely to come from another data generating mechanism (i.e. an error in the data registration) and thus they should be excluded in order to provide meaningful results.

4.4 Setting the trimming level

One of the main innovative features of our approach is that a proportion of observations is trimmed and does not contribute to clustering or to parameter estimation. More precisely we perform trimming in the concentration step discarding a fixed proportion of observations, having the lowest contributions to the likelihood. The trimming level must be chosen in advance. There are several methods to do so (see [5] for a deep discussion on this point). In general, if α is too low there is an high risk of *masking*, with outliers used for estimation and loss of robustness. If α is too large, there is an high risk of *swamping*, with good observations discarded and loss of efficiency. A simple solution that we found effective for robust fuzzy linear clustering is plotting the average contribution to the likelihood r_i against each possible value of α. We show this in the right panel of Figure 6, and report in the left panel the estimated linear fuzzy clusters corresponding to the different values of α. Our proposal is to fix α as the lowest value after which the average contribution is stable. The idea is that after all outliers have been removed, the average contribution to the likelihood does not change much if adding or removing additional observations.

According to our proposal, the optimal value of α would be around 10%. An illustration of the rationale, based on the figure, follows. Figure 6, panel (a), clearly shows that when $\alpha = 0.2$ too many observations have been trimmed. This can be seen also in panel (b) as the the average contribution to the likelihood has stabilized for much smaller values of α. On the other hand panel (e) clearly shows that not enough observations have been trimmed. This can be also seen in panel (f) as the average contribution to the likelihood is much larger for larger values of α and much smaller for smaller values. Hence, there still are outliers available for trimming. Finally, $\alpha = 0.1$, the true underlying contamination level, is a fine choice since the likelihood looks almost stabilized.

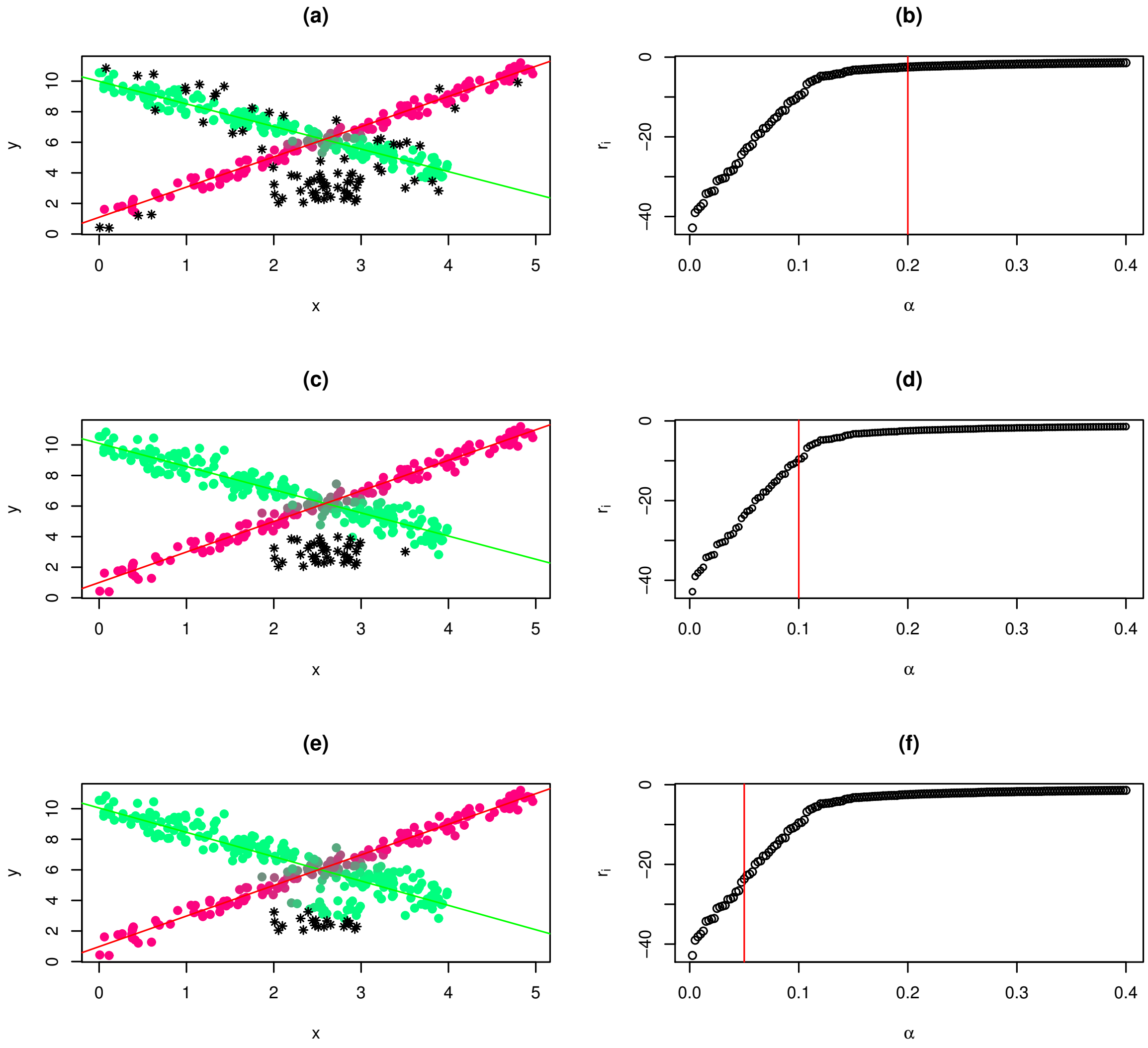

Figure 6: Left panels: Estimated linear clustering result for different trimming levels and $m = 1.5$. (a) $\alpha = 0.20$. (c) $\alpha = 0.10$. (e) $\alpha = 0.05$. Right panels: Average contribution to the likelihood for different values of α. A red line corresponds to the trimming level used on the corresponding left panel. (b): $\alpha = 0.20$. (d): $\alpha = 0.10$. (f): $\alpha = 0.05$.

As a concluding remark we would like to point out that a sensitivity analysis, obtained by varying the tuning parameters in a "proper" range, is always recommended. Additionally we would like to point out that these proposed heuristic tools can help the user to achieve sensible tuning of the procedure and that is not necessary to find the exact value for each tuning parameter since moving in a small range of the optimal value would yield, in most of the case, very similar results.

5 Simulation study

A simulation study was performed to illustrate and compare our procedure. We simulated 2 overlapped unbalanced linear clusters in two and three dimensions, that is to say that they are composed, respectively, by $p = 1$ and $p = 2$ explanatory variables. The two clusters were composed of $n_1 = 144$ and $n_2 = 216$ observations, and in the contaminated cases we added another 40 corrupted observations as described below. Both in two and three dimensions we had an homoscedastic and an heteroscedastic scenario. In two dimensions we generated a uniform covariate X_{1i} with support in $(0, 5)$ and a uniform covariate X_{2i} with support in $(0, 4)$. The underlying regression models were $y_i = 1 + 2x_{1i} + \varepsilon_{1i}$ for the observations within the first cluster and $y_i = 10 - 1.5x_{12} + \varepsilon_{2i}$ for the observations within the second cluster. In three dimensions we generated a uniform covariate X_{11i} with support in $(0, 5)$, another denoted with X_{12i} with support on $(5, 9)$, a third (X_{21i}) with support on $(0, 6)$ and, finally, X_{22} being an independent replica of X_{12i}. The underlying regression models were $y_i = 3 + 4x_{11i} - 2x_{12i} + \varepsilon_{1i}$ for the observations within the first cluster and $y_i = -2 - 2x_{12i} + 2x_{22} + \varepsilon_{2i}$. The errors ε_{1i} and ε_{2i} were 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.

We then considered four contamination's scenarios:

- Clean dataset

- 10% contaminated dataset with contaminating points generated by sampling from a uniform distribution in the range of the data

- 10% contaminated dataset with contaminating points generated by sampling from a uniform distribution defined on the range of the data ± 5.

- 10% contaminated dataset with pointwise contamination. For instance, in the case of $p = 1$ explanatory variable, we generated the contamination jittering around the point $(-1.5, 17)$, while, in the case of $p = 2$ explanatory variables, the contamination has been generated by jittering around the point $(1, -1.5, 18.5)$.

Therefore we have sixteen different data generating scenarios. Within the outlined scenarios we implemented the following different models:

1. The proposed fuzzy regression method based on the unconstrained EM algorithm and without the trimming step. This is the counterpart of our approach. It has been denoted in the plot with the acronym "EM" because it would ben the standard adaptation of the EM algorithm to the fuzzy framework.

2. Our proposed robust Fuzzy TCLUST Regression method, denoted in the plot with the acronym "FTCR".

3. The c-Regression model proposed in [15], which has been denoted in the plot with the acronym "cReg".

4. The Alternative switching Regression model proposed in [20], which has been denoted in the plot with the acronym "A-cReg".

It shall be noted that two of these procedures (EM and cReg) do not take into account robustness issues, while the other two (A-cReg and our proposed FTCR) do.

In our simulation study for each scenario we generated the data and estimated the four models above. We repeated this 500 times and report boxplots of the Mean Squared Error (MSE) for the slope and the intercept estimators in Figures 7 and 8 when $p = 1$ and $p = 2$, respectively (we stored the obtained estimations in a vector and compared it with the vector of the real values). Additionally, in Figures 9 and 10 we report boxplots of the misclassification rates when $p = 1$ and $p = 2$, respectively.

First of all, 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 MSEs (especially in scenarios (e) and (f), that are the pointwise contaminated scenarios) and high variability in performance. On the other hand, FTCR is mostly unaffected by contamination and it shows by far the best MSE and misclassification rates in presence of outliers.

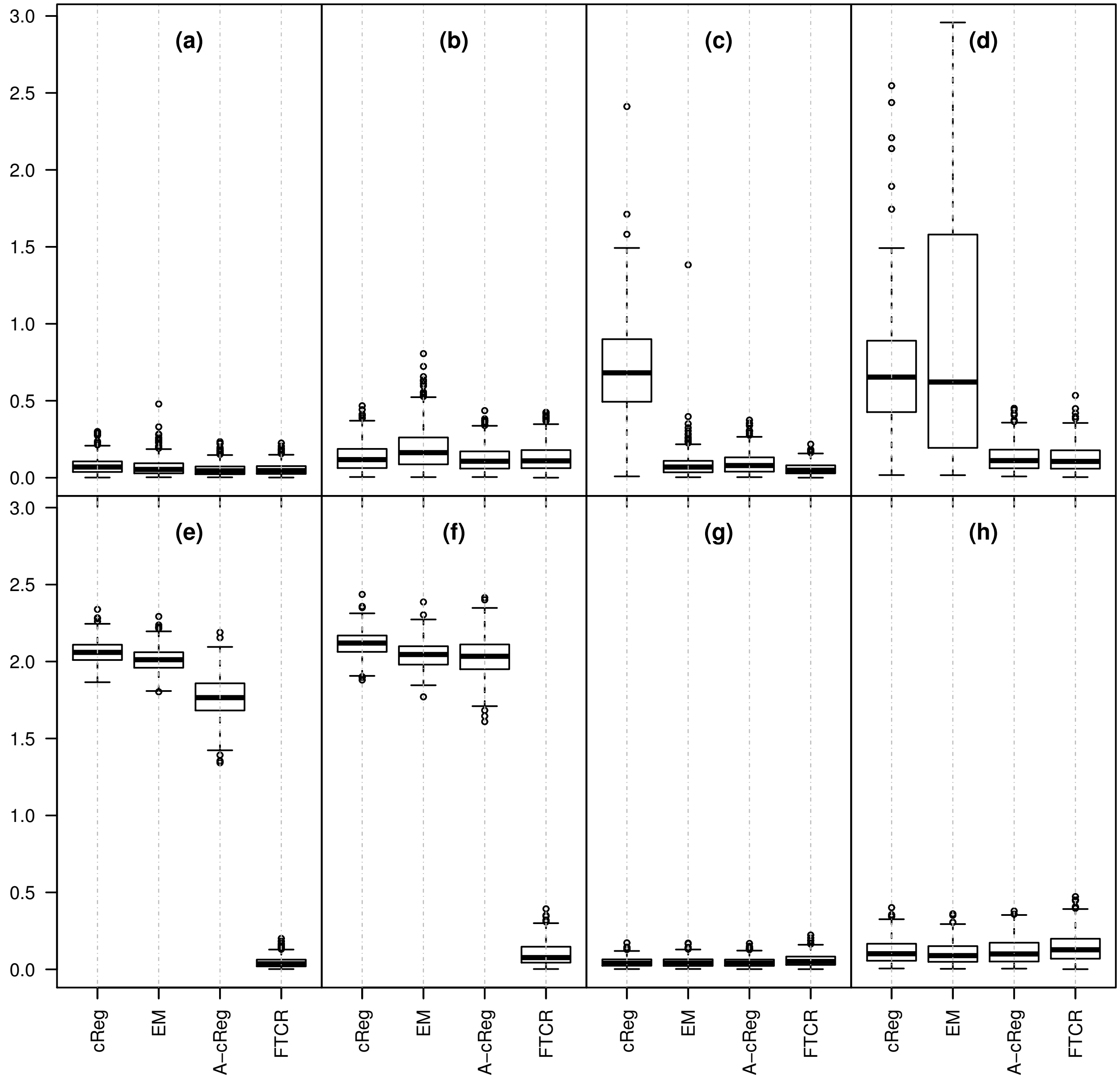

Figure 7: Simulation study: Boxplots representing the MSE of the estimation of $\boldsymbol{\beta}_j$ and b_j^0 when $p = 1$. The Homoscedastic clusters are in (a),(c),(e),(g). Heteroscedastic clusters are in (b), (d), (f), (h). Uniform contamination is in (a) and (b). Increased uniform contamination is in (c), and(d). Pointwise contamination in (e) and (f). Clean dataset is in (g) and (h)

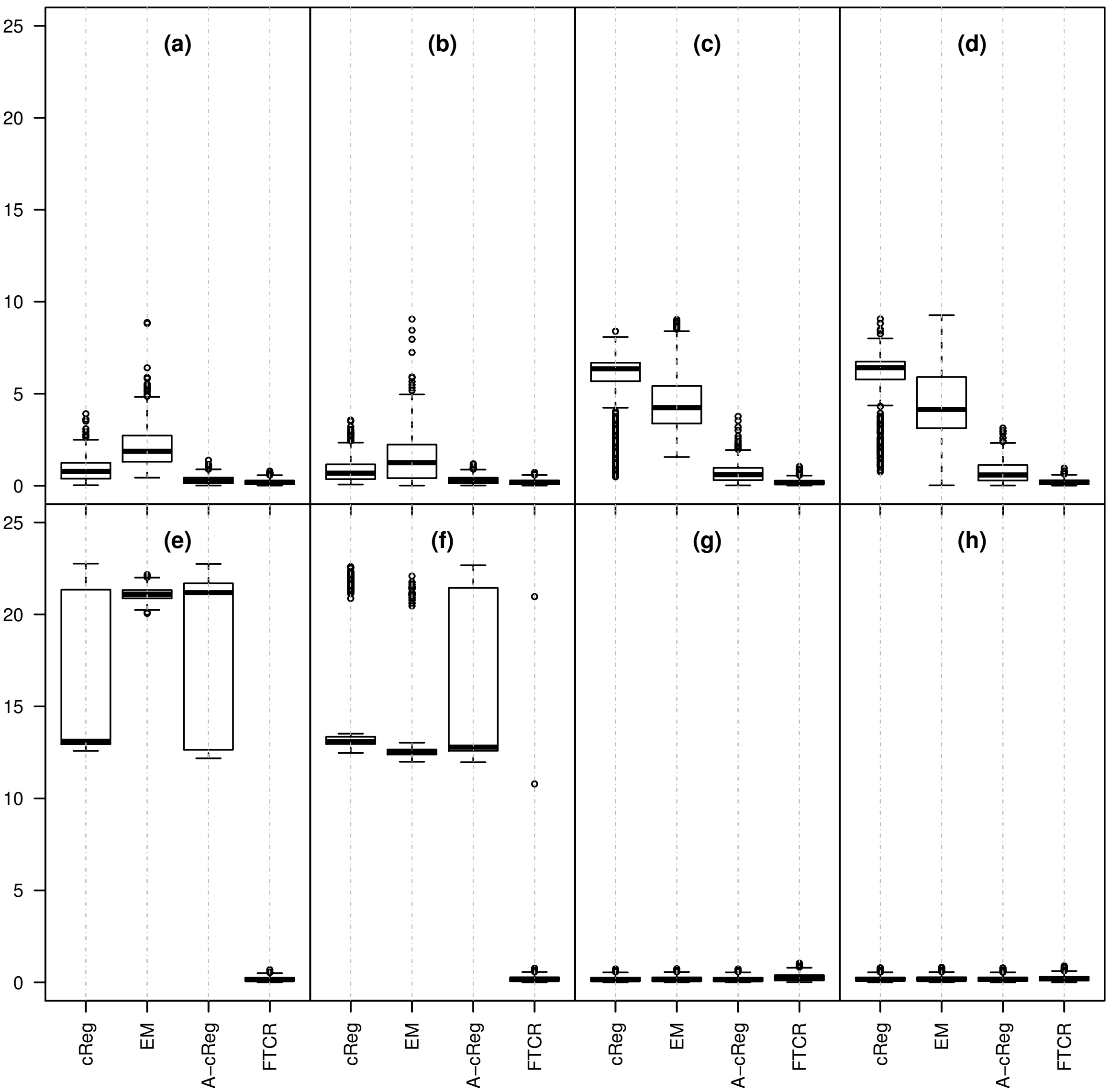

Figure 8: Simulation study: Boxplots representing the MSE of the estimation of $\boldsymbol{\beta}_j$ and b_j^0 when $p = 1$ within the same data scenarios outlined in Figure 7.

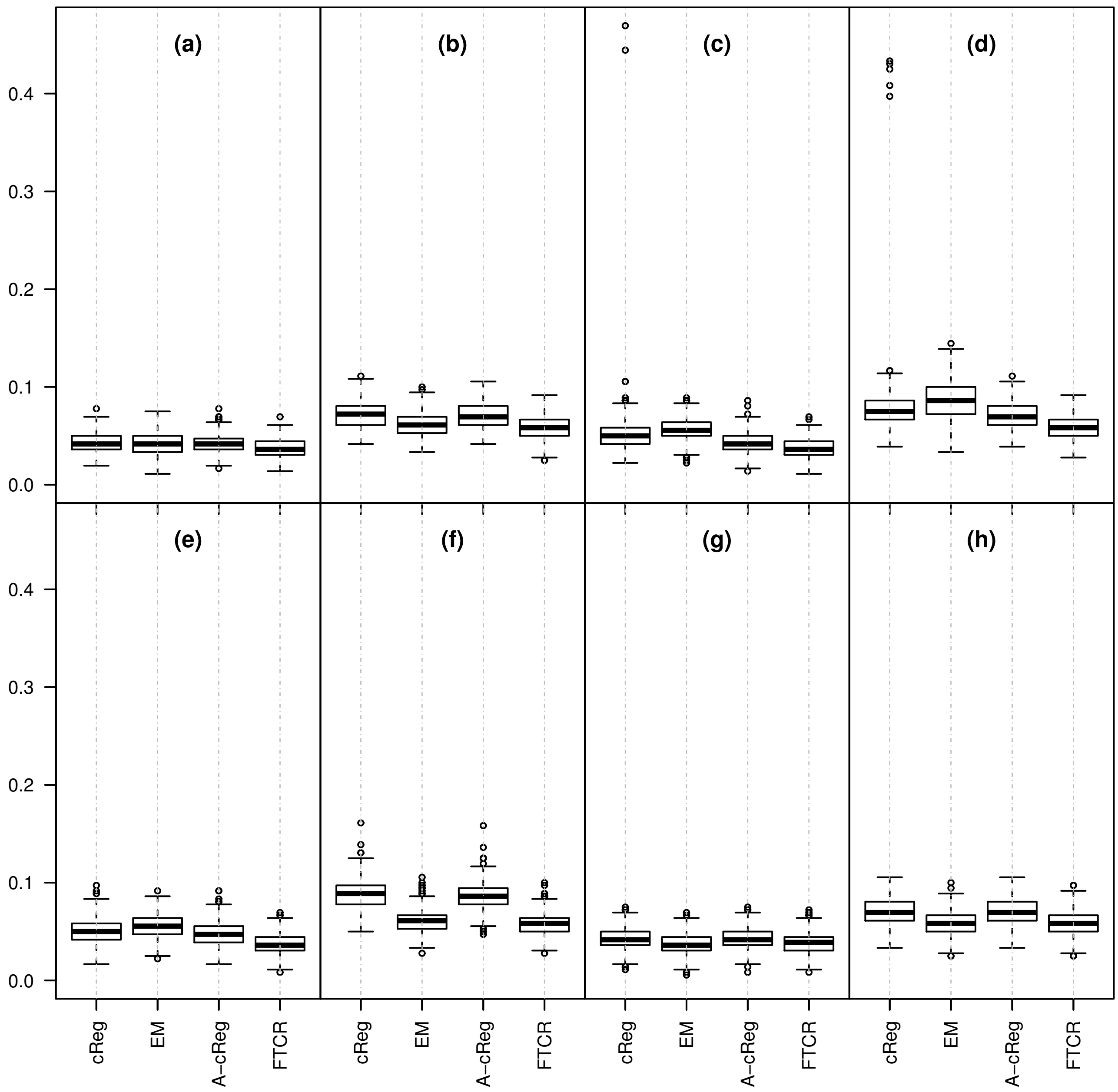

Figure 9: Simulation study. Misclassification error for $p = 1$ within the same data scenarios outlined in Figure 7.

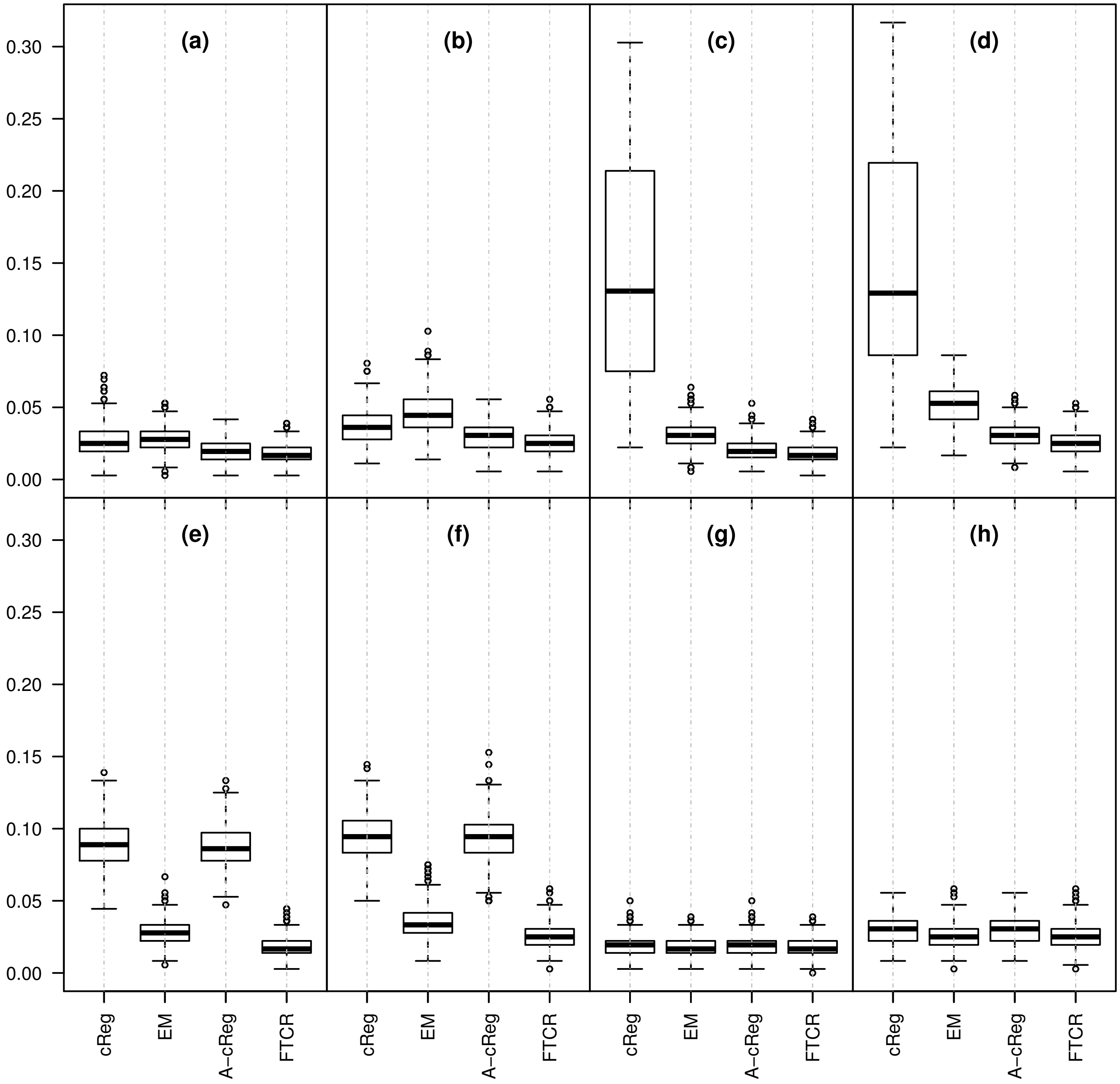

Figure 10: Misclassification rates when $p = 2$ within the same data scenarios outlined in Figure 7.

6 Real Data Example

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 [11], 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. The scatter plot in Figure 11 clearly shows that there should be three approximately linear groups, and an isolated group of outliers. This prompted us to fix $k = 3$. We applied our procedure twice: once without trimming Figure 11, (a), and once with a trimming level of 4% Figure 11, (b).

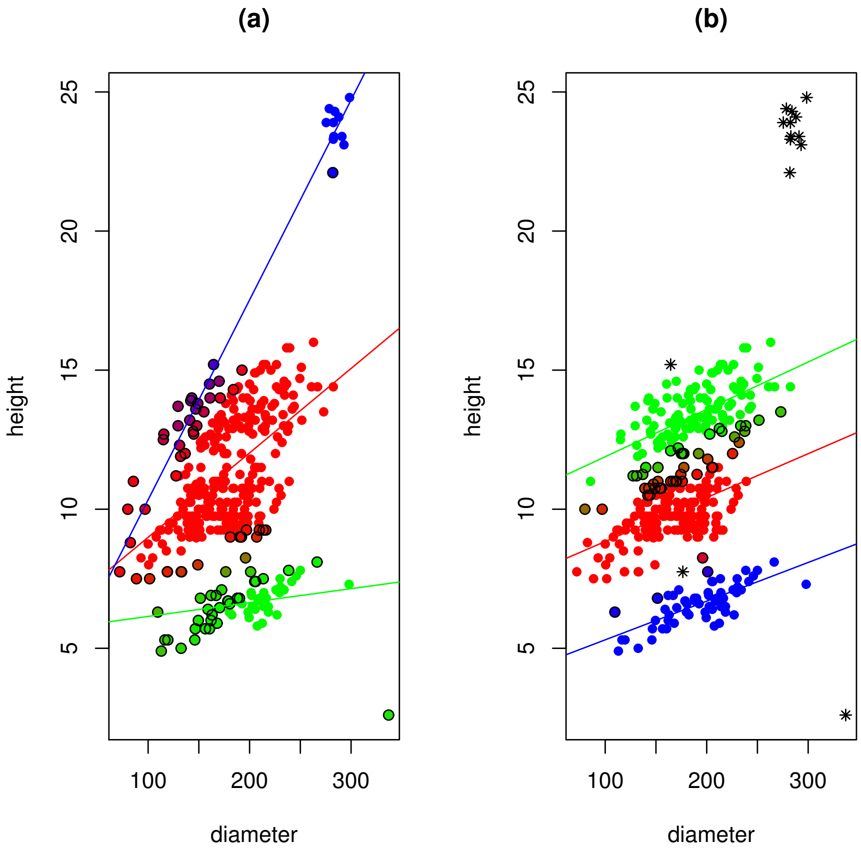

Figure 11: *Pinus Nigra Data example.* (a) scatter plot and results of the proposed procedure with no trimming. (b) scatter plot and FTCR with $\alpha = 0.04$.

It can be seen that the untrimmed procedure is not able to detect the most likely underlying linear relationships, 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. On the other hand, by trimming as few as 15 observations, we recover quite nicely the linear structures. The proportion of hard assignments is also much larger with FTCR procedure, 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 (quite arbitrarily) to one of the clusters. In order to stress this fact, the observation receiving more fuzziness(i.e. whose $\max_j u_{ij} \leq 0.95$) are plotted in Figure 11 with the symbol "$\circ$".

Regarding the obtained results a clear explanation for the detected three clusters is that pines are sampled in three different zones. It can be seen that three almost parallel lines are obtained, indicating 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: they are trees of a different species originally misclassified as *Pinus Nigra*.

We conclude this section by illustrating how we could have chosen the value for the trimming level α in this example. Although we used the same tuning that has been used in [11], that is to say that we imposed $k = 3$ groups and trimming 4% of the observations, a very similar choice for the trimming level α could have been done by using the heuristical tool proposed in the previous section. Indeed Figure 12, plot (a), shows how the average contribution to the likelihood is stabilized when α is set around the value 0.04. In fact even a slightly lower value could have been chosen and this choice would have avoided trimming the two clean observation that appear in Figure 11.

Additionally we also had to choose m and we set this parameter equal to 1.3. Indeed, as it clear from Figure 12, if $m > 1.3$ the proportion of hard assignments decreases sharply, leading to unclear clusters and improper estimation of the linear structures.

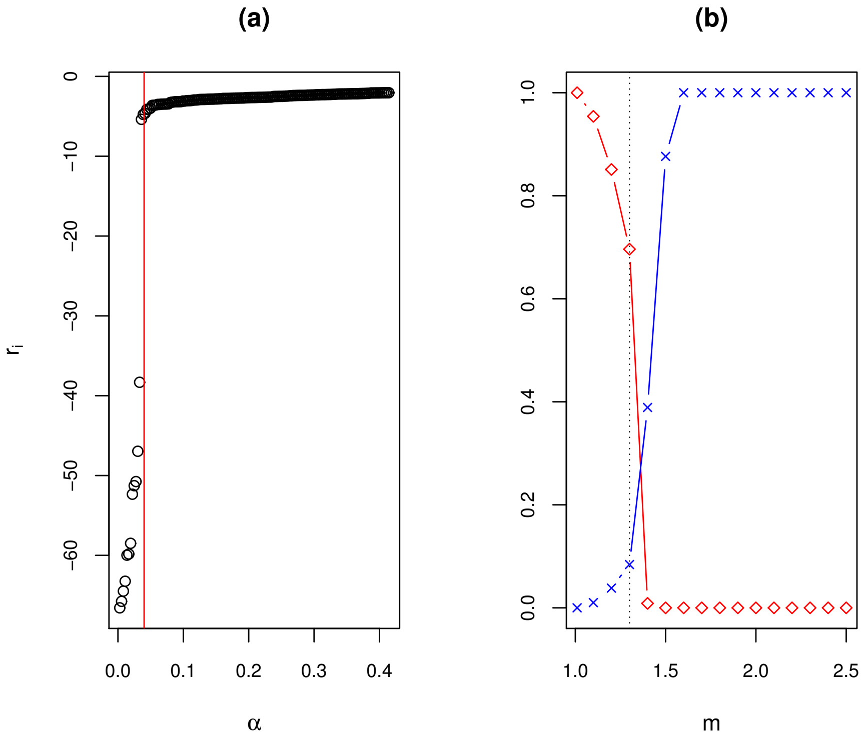

Figure 12: *Pinus Nigra example.* (a) average contribution to the likelihood as a function of α. (b) relative empty entropy and proportion of hard assignments as a function of m.

7 Conclusions and further directions

The proposed procedure is well preforming in terms of robustness, indeed it resists very well to all the type of contaminations and good efficiency is reached in parameter estimation. Nevertheless, as often happens in robust procedures, tuning is required in order to have sensible results. Automatic suitable methods to fix all the tuning parameters, required both within the robust clustering context and fuzzy clustering methods, are not available yet. This is not an easy task because all those parameters are interrelated (e.g. a larger α means that more observations may be discarded and a smaller k value is then needed). However we shown in Section 4 some heuristical tools that can be useful to help us in tuning all these parameters.

References

[1] Ali, Ameer M.; Karmakar, Gour C. nd Dooley, Laurence, S. (2008), "Review on Fuzzy Clustering Algorithm" *Journal of Advanced Computations*,2, 169-181.

[2] Bezdek, C.J., and Nikhil, R.P. (1995), "On Cluster Validityfor the fuzzy *c*-Mean Model" *IEEE transaction on fuzzy systems* 3, 370-379

[3] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, **39**, 18.

[4] DeSarbo, W.S., Cron, W.L., (1988), "A Maximum Likelihood Methodology for Clusterwise Linear Regression" *Journal of Classification*,**5**, 249-282

[5] Farcomeni, A., Greco, L., (2015) *Robust Methods for Data Reduction* Chapman and Hall/CRC

[6] Farcomeni, A., Ventura, L., (2012). "An overview of robust methods in medical research" *Statistical Methods in Medical Research* **21**, 111-133.

[7] Fritz, H., García-Escudero, L.A., and Mayo-Iscar, A.(2013). "Robust constrained fuzzy clustering" *Information Sciences* **245**, 38-52.

[8] Fritz, H., García-Escudero, L.A. and Mayo-Iscar, A. (2012), "A fast algorithm for robust constrained clustering" *Computational Statistics and Data Analysis* **61**, 124-136.

[9] Gath, I. and Geva, A.B. (1989), "Unsupervised optimal fuzzy clustering.", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 773-781.

[10] García-Escudero, L.A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2008), "A general trimming approach to robust cluster analysis", *Annals of Statistics*, **36**, 1324-1345.

[11] García-Escudero, L.A., Mayo-Iscar, A. and San Martín R. (2010), "Robust clusterwise linear regression through trimming", *Computational Statistics and Data Analysis*, **54**, 3057-3069.

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

[13] García-Escudero, L.A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2015), "Avoiding spurious maximizers in mixture modelling", *Stat Comput*, **25**, 619-633.

[14] Gustafson, E.E. and Kessel, W.C. (1979). "Fuzzy Clustering with a Fuzzy Covariance Matrix". *Proceedings of the IEEE lnternational Conference on Fuzzy Systems, San Diego, 1979*, 761-766.

[15] Hathaway, R.J. and Bezdek, J.C. (1993). "Switching regression models and fuzzy clustering". *IEEE Transactions on Fuzzy Systems*, **1**, 195-204.

[16] Heritier, S., Cantoni, E., Copt, S., and Victoria-Feser, M-P. (2009) *Robust methods in biostatistics* Chichester: Wiley

[17] Honda, K., Ohyama, T., Ichihashi, H., and Hotsu, A. (2008), "FCM-type switching regression with alternating least square method" *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ 2008)*,122-127.

[18] Hosmer, D.W. Jr. (1974), "Maximun Likelihood estimates of the parameters of a mixture of two regression lines." *Communications in Statstics* **3**, 995-1006.

[19] Kuo-Lung, W., (2012). "Analysis of parameter selections for fuzzy c-means " *Pattern Recognition*

[20] Kuo-Lung, W., Miin-Shen ,Y. and June-Nan, H. (2009). "Alternative fuzzy switch-

ing regression" *Proceedings of the International MultiConference of Engineers and Computer Scientist* **1**

[21] Lenstra, A.K., Lenstra J.K., Rinnoy Kan, A.H.G., Wansbeek, T.J. (1982) "Two lines least squares" *Annals of Discrete Mathematics* **16**, 201-211

[22] McLachlan, G. and Peel, D. (2000), *Finite Mixture Models,* John Wiley Sons, Ltd., New York.

[23] Miyamoto, S. and Mukaidono, M. (1997). "Fuzzy *c*-means as a regularization and maximum entropy approach" *Proceedings of the 7th International Fuzzy Systems Association World Congress (IFSA'97)*, **2**, 86-92.

[24] Rousseeuw, P.J., Kaufman, L. and Trauwaert, E. (1996). "Fuzzy clustering using scatter matrices" *Computational Statistics* & *Data Analysis*, **23**, 135-151

[25] Rousseeuw, P.J. and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, **41**, 212-223.

[26] Ruspini, E.H. (1969). "A New Approach to Clustering" *Information and Control*, **29**, 22-32.

[27] Späth, H. (1982), "A Fast Algorithm for Clusterwise Regression" *Computing* **29**, 175-181.

[28] Trauwaert, E., Kaufman, L. and Rousseeuw, P. (1991), "Fuzzy clustering algorithms based on the maximum likelihood principle" *Fuzzy Sets and Systems* **42**, 213-227.