Statistical Methods & Applications (2021) 30:711–746
https://doi.org/10.1007/s10260-020-00540-8

ORIGINAL PAPER

Weighted likelihood latent class linear regression

Luca Greco[1] · Antonio Lucadamo[1] · Claudio Agostinelli[2]

Accepted: 5 July 2020 / Published online: 23 July 2020

Abstract
A weighted likelihood approach for robust fitting of a finite mixture of linear regression models is proposed. An EM type algorithm and its variant based on the classification likelihood have been developed. The proposed algorithm is characterized by an M-step that is enhanced by the computation of weights aimed at downweighting outliers. The weights are based on the Pearson residuals stemming from the assumption of normality for the error distribution. Formal rules for robust clustering and outlier detection are also defined based on the fitted mixture model. The behavior of the proposed methodologies has been investigated by numerical studies and real data examples in terms of both fitting and classification accuracy and outlier detection.

Keywords Classification · EM · Mixture · Outliers detection · Pearson residuals · Regression · Robustness · Weighted likelihood

1 Introduction

The problem of clustering around linear structures is particularly appealing and has received growing interest in the literature. Latent class regression has applications in many fields, including engineering, genetics, biology, econometrics, marketing, computer vision, pattern recognition, tomography, fault detection, among others. The reader is pointed to García-Escudero et al. (2009) for a large collection of references. This paper is motivated by the fact that noisy data frequently appear in

✉ Antonio Lucadamo
antonio.lucadamo@unisannio.it

Luca Greco
luca.greco@unisannio.it

Claudio Agostinelli
claudio.agostinelli@unitn.it

[1] DEMM Department, University of Sannio, Benevento, Italy

[2] Department of Mathematics, University of Trento, Trento, Italy

every field of application. When the sample data is contaminated by the occurrence of outliers, it is well known that maximum likelihood estimation (MLE) is likely to lead to unreliable results. In a mixture setting, the bias of at least one of the component parameters estimate can be arbitrarily large and the true underlying clustering structure might be hidden. Therefore, there is the need for a suitably robust procedure providing protection against outliers. The reader is pointed to the book by Farcomeni and Greco (2015a) for a gentle introduction to robustness issues.

The problem of robust fitting of a mixture of linear regressions has been already tackled in the literature. In general, the robust solutions are driven by a suitable modification of the EM algorithm for mixtures or the classification EM algorithm (CEM), concerning the M step, which is enhanced by some robust estimation approach in place of maximum likelihood. Some existing proposals are based on the idea of (hard) trimming: estimation is performed over a subset of the original data obtained after discarding those units with the lowest contributions to the likelihood function. According to such trimming strategies, potential outliers are discarded in the estimation process, that is observations are given crispy weights in $\{0, 1\}$. Neykov et al. (2007) introduced a mixture fitting approach based on the trimmed likelihood, García-Escudero et al. (2010) extended the TCLUST methodology, developed in García-Escudero et al. (2008) for mixtures of multivariate Gaussian distributions, exploiting the idea of impartial trimming in TCLUST-REG, a related proposal has been presented in García-Escudero et al. (2009) and an adaptive hard trimming procedure has been described in Riani et al. (2008) based on the Forward Search methodology. In particular, TCLUST-REG is characterized by group scatter constraints aimed at making the mixture fitting a well-posed problem and the addition of a second trimming step to mitigate the effect of outliers in the space of explanatory variables acting as leverage points. A very recent adaptive version of TCLUST-REG has been discussed in Torti et al. (2019). An alternative approach meant to automatically take into account leverage points has been considered by García-Escudero et al. (2017) where trimming and restrictions have been introduced to get a robust version of the cluster weighted model, named Trimmed Clustered Weighted Restricted Model (TCWRM). In this approach restrictions concern both the set of eigenvalues of the covariance matrix evaluated on the X-space and the variances of the regression error term. The reader is pointed to Torti et al. (2019) for a comparative analysis of TCLUST-REG and TCWRM under general settings. The benefits of trimming for robust regression clustering have been also investigated in Dotto et al. (2017) where a fuzzy approach has been developed.

In a different but complementary fashion, Bashir and Carter (2012) and Bai et al. (2012) modified the M step by resorting to soft rather than hard trimming procedures. Actually, they replaced the single component MLE problems by M- (and S-) estimation problems for linear regression (see also Campbell (1984) and Maronna et al. (2019)). In particular, in both papers the authors developed an EM-type algorithm featured by componentwise weights but this approach can be extend to obtain robust versions of the CEM algorithm based on M- and S-estimation, as well. According to a soft trimming strategy, observations are attached a weight lying in [0, 1] according to some measure of outlyingness. Potential outliers are expected to be heavily downweighted, whereas genuine observations receive a weight close to one.

It is worth to mention that there are different proposals aimed at robust latent class linear regression estimation that are not based on soft or hard trimming procedures in which the assumed model is embedded in a larger one to account for outliers. Yao et al. (2014) considered a mixtures of linear regression models with Student t error distributions; Punzo and McNicholas (2017) developed an approach based on the Contaminated Gaussian Cluster Weighted Model in which each mixture component has some parameters controlling the proportion of (different type of) outliers; Yu et al. (2017) proposed a case-specific and scale-dependent mean-shift mixture model and a penalized likelihood approach to induce sparsity among the mean-shift parameters.

Here, we propose the use of the weighted likelihood methodology (Markatou et al. 1998) as a valid alternative to the existing methods. Weighted likelihood is an appealing robust techniques for estimation and testing (Agostinelli and Markatou 2001). In particular, reliable statistical tools have been developed for linear regression (Agostinelli and Markatou 1998; Agostinelli 2002), generalized linear models (Alqallaf and Agostinelli 2016) and multivariate analysis (Agostinelli and Greco 2019). Recently, Greco and Agostinelli (2020) also introduced weighted likelihood estimation of mixtures of multivariate normal distributions. The authors explored the behavior of both EM and CEM type algorithms and found that weighted likelihood gives powerful devices for robust estimation, classification and outliers detection. Then, the same ideas can be extended to the context of mixtures of linear regressions.

Weighted likelihood belongs to the group of soft trimming techniques and the weighted likelihood estimator (WLE) can be thought as an M-estimator. The main differences are in the genesis of the weights and in their asymptotic behavior at the assumed model. Actually, weighted likelihood estimation can correspond to a minimum disparity estimation problem (Basu and Lindsay 1994). Then, the WLE is expected to be highly robust under contamination but, conversely to M-estimators, also asymptotically efficient at the assumed model. Some necessary preliminaries on weighted likelihood estimation are given in Sect. 2. The weighted EM and penalized CEM algorithms for robust fitting of mixtures of regressions are introduced in Sect. 3. Section 4 highlights some computational details, Sect. 5 is devoted to a very general result about consistency and aymptotic normality of the proposed estimator. Then, the outlier detection rule is described in Sect. 6 and some illustrative examples based on simulated data are presented in Sect. 7. In Sect. 8 some strategies to select the number of latent classes are presented. Section 9 gives some numerical studies whereas a real data example is discussed in Sect. 10. Concluding remarks end the paper.

2 Background

Let $y = (y_1, \cdots, y_n)^{\mathrm{T}}$ be a random sample of size n drawn from a r.v. Y with distribution function $M(y;\theta)$ and probability (density) function $m(y;\theta)$, which is an element of the parametric family of distributions $\mathcal{M} = \{M(y;\theta), \theta \in \Theta \subseteq \mathbb{R}^d, d \geq 1, y \in \mathcal{Y}\}$. Let $\hat{F}$ be the empirical distribution function. The WLE $\hat{\theta}^w$ is defined as the root of the Weighted Likelihood Estimating Equations (WLEE)

$$\sum_{i=1}^{n} w(y_i; \theta, \hat{F}) s(y_i; \theta) = 0, \tag{1}$$

where $s(y; \theta) = \sum_{i=1}^{n} s(y_i; \theta)$ is the score function. The WLEE in (1) is a modified version of the (system of) likelihood equations, since a data dependent weight, $w_i = w(y_i; \theta, \hat{F}) \in [0, 1]$, is attached to each individual score component. The weights are meant to be small for those data points that are in disagreement with the assumed sampling model. The degree of agreement between the data and the assumed model is measured by the Pearson residual function. Let

$$f^*(y) = \int_{\mathcal{Y}} k(y; t, h) d\hat{F}(t)$$

be a non parametric kernel density estimate and

$$m^*(y; \theta) = \int_{\mathcal{Y}} k(y; t, h) m(t; \theta) \ dt$$

a smoothed version of the model density obtained by using the same kernel function $k(y;\ t,\ h)$. Then, the Pearson residual is

$$\delta(y) = \frac{f^*(y) - m^*(y; \theta)}{m^*(y; \theta)},$$

with $\delta(y) \in [-1, +\infty)$. By smoothing the model, the Pearson residuals converge to zero with probability one for every y under the assumed model; the reader is pointed to Basu and Lindsay (1994), Markatou et al. (1998) and references therein. When the model is discrete, $f^*(y)$ is the empirical probability function and $m^*(y; \theta)$ simply reduces to $m(y; \theta)$. In this paper, we will make use of the Pearson residuals established in Agostinelli and Greco (2019). Actually, a valid WLEE can be also obtained by using Pearson residuals that are defined as

$$\delta(y) = \frac{f^*(\tilde{y}) - m^*(\tilde{y})}{m^*(\tilde{y})},$$

where $\tilde{y} = g(y; \theta)$ is a pivot at the assumed model whose (smoothed) distribution does not depend on the parameter value.

Large values of the Pearson residual function correspond to regions of the support of Y where the model fits the data poorly. According to this approach, outliers can be defined as *observations that are highly unlikely to occur under the assumed model* (Markatou et al. 1998), rather than from a geometric point of view as observation that are far from the model fitted to the bulk of the data, as in the classical theory of M-estimators.

The weight function is defined as

$$w(\delta(y)) = \frac{[A(\delta(y)) + 1]^+}{\delta(y) + 1}, \qquad (2)$$

where $[\cdot]^+$ denotes the positive part and $A(\delta)$ is the Residual Adjustment Function (RAF, Basu and Lindsay (1994)). The RAF plays the role to bound the effect of large Pearson residuals on the fitting procedure. By using a RAF such that $|A(\delta)| \leq |\delta|$ both outliers and inliers (whose nature will be described in the following) will be downweighted. The RAF function is connected to minimum disparity estimation problems. Actually, it is defined as $A(\delta) = (\delta + 1)G'(\delta) - G(0)$, with prime denoting differentiation, where $G(\cdot)$ is a strictly convex function over $[-1, +\infty)$ and thrice differentiable, which determines a disparity measure, that, in the continuous case, is defined as

$$\rho(f^*(y), m^*(y;\theta) = \int_y G(y) m^*(y;\theta)\ dy.$$

In principle, by following the approach developed in Markatou et al. (1998), it is possible to build a WLEE matching a minimum disparity objective function. One can consider the families of RAF stemming from the Symmetric Chi-Squared divergence, the family of Power divergence or Generalized Kullback–Leibler divergence measures. The resulting weight function is unimodal and decline smoothly to zero as $\delta(y) \to -1$ or $\delta(y) \to +\infty$. The weighting function corresponding to a Symmetric Chi-Squared divergence, which is driven by $G(\delta) = \frac{2\delta^2}{\delta+2}$, is given in Fig. 1.

Under the assumptions given in Markatou et al. (1998) and Agostinelli and Markatou (2001), that establish some regularity conditions on the model, the kernel and the weight function, at the assumed model, we have that:

1. $\hat{\theta}^w$ is a consistent and first order efficient estimator of θ, that is

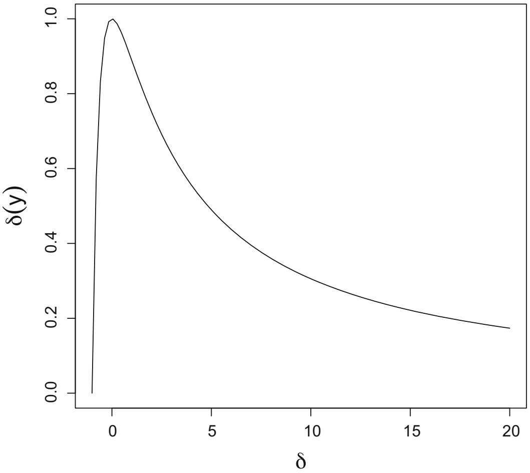

Fig. 1 Weighting function corresponding to a Symmetric Chi-Squared divergence

$$\sqrt{n}(\hat{\theta}^w - \theta)\xrightarrow{d}N(0, I_1^{-1}(\theta))$$

where $I_1(\theta) = E[u(Y;\theta)^2]$ is the expected Fisher information;

2. $\sup|w(y, \hat{\theta}^w, \hat{F}) - 1|\xrightarrow{a.s.}0$ (Agostinelli and Greco 2013);
3. the weighted versions of the likelihood ratio, Wald and score test all share the usual asymptotic behavior (Agostinelli and Markatou 2001).

For what concerns the robustness properties of the WLE, the reader is pointed to Lindsay (1994); Markatou et al. (1998). In particular, the the robust behaviour of the WLE in finite samples is characterized by the curvature parameter $A''(0)$ being negative, despite an unbounded influence function that is just that of the MLE. Furthermore, the WLE has a strong breakdown point equal to 0.5.

It is worth to claim that the shape of the kernel function has a very limited effect on weighted likelihood estimation. On the contrary, the smoothing parameter h allows to control the robustness/efficiency trade-off of the methodology in finite samples. Actually, large values of h lead to Pearson residuals all close to zero and weights all close to one and, hence, large efficiency, since the kernel density estimate is stochastically close to the postulated model. On the other hand, small values of h make the kernel density estimate more sensitive to the occurrence of outliers and the Pearson residuals become large for those data points that are in disagreement with the model.

2.1 Weighted likelihood for linear regression

Let us consider a linear regression model with normally distributed errors, i.e. $y = X\beta + \sigma\epsilon$, where y is the response, $X = [x_1, \ldots, x_p]$ is the $n \times p$ design matrix, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the vector of regression coefficients, σ is a scale parameter and $\epsilon \sim N(0, 1)$. In this setting, Pearson residuals and the weights can be evaluated over the scaled residuals $e = g(y; \beta, \sigma) = (y - X\beta)/\sigma$. An appealing strategy to compute Pearson residuals consists in using a normal kernel with bandwidth equal to h. In such a way, the smoothed model density is still normal with variance $(1 + h^2)$, that is

$$\delta(y) = \frac{f^*(e)}{\frac{1}{\sqrt{1+h^2}}\phi\left(\frac{e}{\sqrt{1+h^2}}\right)} - 1 \ , \qquad (3)$$

where $\phi(\cdot)$ denotes the standard normal density function. Then, the WLE of (β, σ) is obtained as the result of weighted least squares. Clearly, the computation of the WLE of (β, σ) yields an iterative procedure. At each iteration, based on the current parameter estimates, scaled residuals are obtained. Then, their non parametric density estimate is fitted based on the chosen kernel and Pearson residuals and weights are updated according to (3) and (2).

3 Robust fitting of a latent class linear regression model

Let us assume a latent class regression model featured by K components, where K is fixed in advance, with density function denoted by

$$m(y; x, \tau) = \sum_{k=1}^{K} \pi_k \phi(y; \mu_k, \sigma_k), \tag{4}$$

where $\mu_k = X\beta_k$, π_k is the prior probability of component k, (β_k, σ_k) are the component specific parameters and $\tau = (\pi_1, \ldots, \pi_K, \beta_1, \ldots, \beta_K, \sigma_1, \ldots, \sigma_K)^{\mathrm{T}}$ is the vector of all parameters.

The mixture loglikelihood function based on a sample of size n is

$$\ell(\tau) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \phi(y_i; \mu_{ik}, \sigma_k). \tag{5}$$

Maximum likelihood estimation is commonly performed by the EM algorithm, that works with the classification loglikelihood

$$\ell_c(\tau) = \sum_{i=1}^{n} \sum_{k=1}^{K} \log(\pi_k \phi(y_i; \mu_{ik}, \sigma_k)) u_{ik}, \tag{6}$$

where u_{ij} is an indicator of the ith unit belonging to the jth cluster. The EM algorithm iterates, over the index s, between the E step, in which posterior membership probabilities are evaluated as

$$u_{ik}^{(s)} = \frac{\pi_k^{(s-1)} \phi\left(y_i; \mu_{ik}^{(s-1)}, \sigma_k^{(s-1)}\right)}{\sum_{k=1}^{K} \pi_k^{(s-1)} \phi\left(y_i; \mu_{ik}^{(s-1)}, \sigma_k^{(s-1)}\right)}$$

and the M step, where parameters' estimates are updated as

$$\pi_k^{(s)} = \frac{\sum_{i=1}^{n} u_{ik}^{(s)}}{n}$$

$$\beta_k^{(s)} = (X^{\mathrm{T}} U_k^{(s)} X)^{-1} X^{{}^{\mathrm{T}} U_k^{(s)} y \sigma_k^{2(s)} = \frac{\left(y - \mu_k^{(s)}\right)^{\mathrm{T}} U_k^{(s)} \left(y - \mu_k^{(s)}\right)}{n}}$$

where $U_k^{(s)}$ is a diagonal matrix with elements u_{ik}.

At convergence, cluster assignments can be pursued according to a Maximum a Posteriori (MAP) rule: units are assigned to the most likely component. In the CEM algorithm, after the E step, a classification step is performed (together they form the CE step). Let $k_i = \mathrm{argmax}_k u_{ik}^{(s)}$, then $u_{ik_i}^{(s)} = 1$ and $u_{ik}^{(s)} = 0$ for $k \neq k_i$ and $U^{(s)}$ becomes a dummy matrix. Conversely to the EM algorithm, the CEM directly provides a classification of the units at convergence. Actually, the classification approach is aimed at maximizing the classification loglikelihood over both the mixture parameters and the individual components' labels.

Weighted versions of the above algorithms can be designed by introducing the computation of the weights defined in (2) before the M step at the current parameter value. In particular, the weighted EM (WEM) will require componentwise sets of weights, wheres in the weighted CEM (WCEM) weights will be computed conditionally on the current cluster assignments driven by the CE step. More in details, the WEM algorithm iterates between the classical E step and an M step in which the single components MLE problems are replaced by K one-step WLE problems. The single iteration is summarized in Algorithm 1. On the contrary, the WCEM algorithm iterates between the standard CE step and K one-step weighted likelihood based M-step in which weights are evaluated conditionally to the current cluster assignment and not for each component anymore, that is $u_{ik} = 1$ for $k = k_i$ and zero otherwise. Furthermore, the proposed algorithm can be successfully augmented by introducing scatter similarity restrictions as described by García-Escudero et al. (2010). These constraints are posed by fixing a constant c_σ such that

$$\frac{\max \sigma_k}{\min \sigma_k} \leq c_\sigma, \quad k = 1, 2, \ldots, K \tag{7}$$

and are needed to avoid spurious solutions and make the mixture fitting and classification well defined problems (see also Fritz et al. (2013); Garcia-Escudero et al. (2015); Greco and Agostinelli (2020)).

Algorithm 1 Computation of weights and the M step of the WEM algorithm

```
Weights
for $c = k, \ldots, K$ do
    $e_{ik}^{(s)} = \frac{1}{\hat{\sigma}^{(s)}}(y_i - \hat{\mu}_k^{(s)})$
    $\delta_{ik}^{(s)} = \frac{f^*\left(e_{ik}^{(s)}\right)}{\phi(e_{ik}^{(s)}; 0, \sqrt{(1+h^2)})} - 1$
    Obtain
    $w_{ik}^{(s)} = \frac{\left[A\left(\delta_{ik}^{(s)}\right) + 1\right]^+}{\delta_{ik}^{(s)} + 1}$
end for
M-step
for $c = k, \ldots, K$ do
    $\pi_k^{(s+1)} = \frac{\sum\limits_{i=1}^{n} \tilde{w}_{ik}^{(s)}}{\sum_{i=1}^{n} \sum_{k=1}^{K} \tilde{w}_{ik}^{(s)}}$
    $\beta_k^{(s+1)} = (X^{\mathrm{T}} \tilde{W}^{(s)} X)^{-1} X^{\mathrm{T}} \tilde{W}^{(s)} y$
    $\sigma_k^{(s+1)} = \frac{\sum\limits_{i=1}^{n} \left(y_i - \mu_k^{(s+1)}\right)^2 \tilde{w}_{ik}^{(s)}}{\sum_{i=1}^{n} w_{ik}^{(s)} u_{ik}^{(s)}}, \ \mu_k = X\beta_k$
    with $\tilde{w}_{ik} = u_{ik} w_{ik}$ and $\tilde{W}^{(s)} = diag\left[\tilde{w}_{ik}^{(s)}\right]$.
end for
```

Here, as one referee pointed out, the interest focuses on a general mixture model in which all parameters are class-dependent. However, in some applications a more parsimonious model may be needed, that is characterized by homogeneous slopes and is nested into the more general one. Maximum likelihood estimation can be performed by using a different set of estimating equations in the M-step and its weighted counterpart can be obtained as well. For instance, let us consider a mixture model whose components only differ in the intercept term, i.e $\beta = (\beta_{01}, \ldots, \beta_{0K}, \beta_{-0})$, where β_{-0} is the p_{-0} dimensional vector of common slopes with $p_{-0} \geq 1$. Then, the M-step in Algorithm 1 changes as follows:

$$\begin{aligned} \beta_{-0}^{(s+1)} &= \sum_{k=1}^{K} \left(X^T \tilde{W}^{(s)} X \right)^{-1} X^T \tilde{W}^{(s)} \left(y - \beta_{0k}^{(s)} \right) \\ \beta_{0k}^{(s+1)} &= \frac{\sum_{i=1}^{n} \left(y_i - x_i \beta_{-0}^{(s)} \right) \tilde{w}_{ik}}{\sum_{i=1}^{n} \tilde{w}_{ik}}, \quad k = 1, 2, \ldots, K. \end{aligned} \tag{8}$$

4 Computational details

One of the first issues to deal with the estimation of a mixture model by the EM or CEM algorithm and their robust counterparts is the choice of a suitable starting point. A solution is represented by subsampling (Markatou et al. 1998; Neykov and Müller 2003; Neykov et al. 2007; Torti et al. 2019). A subsample of size n^* is selected randomly from the data sample, then the model is fitted to these n^* observations by the classical EM (or CEM) algorithm to get a trial estimate. This approach shows some limitations since from the one hand n^* should be as small as possible in order to increase the chance of drawing at least one outlier free subsample, but from the other hand a larger trial sample size will avoid the algorithm to fail in finding a solution.

Here, in a different fashion, a deterministic initialization will be considered: first units are assigned to the different components by running TCLUST to the multivariate data (y, X), then cluster specific parameters' estimates are initialized by running a robust regression conditionally on clusters' assignments. In particular, weighted likelihood regression has been used but M-type regression could be used as well. This strategy is well justified since in García-Escudero et al. (2010) it is stated that TCLUST could serve as starting point for others approaches. The initial clustering depends on a couple of tuning constants that allow control of TCLUST: the level of trimming α and the eigen-ratio constraint factor (García-Escudero et al. 2008; Fritz et al. 2013), that will be denoted by c to avoid confusion with the scatter constraint factor c_σ in (7) characterizing the proposed WEM and WCEM algorithms.

An alternative deterministic initial solution may be obtained by computing the trimmed likelihood estimator of Neykov et al. (2007); other candidate initial solutions can be evaluated according to the approach discussed in Coretto and Hennig (2017) that is based on a combination of nearest neighbor denoising and agglomerative hierarchical clustering. Further starting points can be obtained by randomly perturbing the deterministic starting solution and/or the final one obtained from it (Farcomeni and Greco 2015b).

When using TCLUST, in order to avoid the algorithm to be dependent on the initial partition of the data and trapped in (spurious) local optima, a general advise is to run it (several times) for different values of (α, c) and then select the final fitted model according to one criterion. For instance, when the initial TCLUST is not properly robust with respect to the actual amount of contamination in the data, then the fitted model could exhibit lack of robustness as well. On the contrary, a large level of trimming can lead to solutions that are characterized by an excess of downweighting. As well, the choice of the eigen restriction factor could be crucial.

Formal solutions to the problem of root selection in weighted likelihood estimation have been provided in Markatou et al. (1998), Agostinelli (2006), Agostinelli and Greco (2019). Here, we decided to select the root leading to the minimum fitted approximate disparity as defined in Agostinelli and Greco (2019), that is

$$\tilde{\rho}(f^*, m^*) = \frac{1}{n}\sum_{i=1}^{n}\frac{G(\delta_i) + \delta_i}{\delta_i + 1} \tag{9}$$

where the Pearson residuals δ_i are evaluated conditionally on the final cluster assignments, that is $\delta_i = \delta_{ik_i}$, at convergence.

Another remarkable aspect is represented by the selection of the bandwidth parameter h. The tuning of the smoothing parameter h could be based on several quantities of interest stemming from the fitted mixture model: a safe selection can be achieved by monitoring the unit specific weights, residuals or the empirical downweighting level $(1 - \hat{\bar{\omega}})$ as h varies (Markatou et al. 1998; Greco 2017; Agostinelli and Greco 2018), with $\hat{\bar{\omega}} = n^{-1}\sum_{i=1}^{n}\hat{w}_i$ and the weights are evaluated conditionally on the final cluster assignments, that is $\hat{w}_i = \hat{w}_{ik_i}$.

An abrupt change in the monitored empirical downweighting level or in the residuals may indicate, for instance, the transition from a robust to a non robust fit (Agostinelli and Greco 2018) or to an extremely robust fit leading to the detection of an even intolerable number of false outliers, as shown in the bottom right panel of Fig. 10. Hence, monitoring can aid in the selection of a value of h that gives an appropriate compromise between efficiency and robustness at finite samples. A monitoring approach is commonly applied to select the trimming level in TCLUST, TCLUST-REG and TCWRM, for instance. The reader is pointed to Cerioli et al. (2018b) for a recent general account on the benefits and potentials of monitoring.

5 WEM and WCEM as special cases

The WEM and WCEM are obtained by replacing maximum likelihood by a different set of estimating equations, characterized by the introduction of weights aimed at bounding the effect of outliers on the fit. In a fashion similar to what stated in Bai et al. (2012), the proposed algorithms represent a special case of the algorithm first introduced by Elashoff and Ryan (2004), where an EM algorithm has been established for very general estimating equations. Here, in the M-step, it is suggested to solve a complete data estimating equation of the form

$$\Psi(y;X,\tau) = (\Psi_{\pi}(y;X,\tau), \Psi_{\beta}(y;X,\tau), \Psi_{\sigma}(y;X,\tau))^{\mathrm{T}=0} \tag{10}$$

with

$$\begin{aligned}
\Psi_{\pi}(y;X,\tau) &= (\Psi_{\pi_1}(y;X,\tau), \ldots, \Psi_{\pi_K}(y;X,\tau))^{\mathrm{T}},\\
\Psi_{\beta}(y;X,\tau) &= (\Psi_{\beta_1}(y;X,\tau), \ldots, \Psi_{\beta_K}(y;X,\tau))^{\mathrm{T}},\\
\Psi_{\sigma}(y;X,\tau) &= (\Psi_{\sigma_1}(y;X,\tau), \ldots, \Psi_{\sigma_K}(y;X,\tau))^{\mathrm{T}}
\end{aligned}$$

and

$$\begin{aligned}
\Psi_{\pi}(y;X,\tau) &= \sum_{i=1}^{n} \psi_{\pi_j}(y_i;\tau)u_{ij} = \sum_{i=1}^{n} w(y_i;\tau,\hat{F})s_{\pi_j}(y_i;\tau)u_{ij},\\
\Psi_{\beta}(y;X,\tau) &= \sum_{i=1}^{n} \psi_{\beta_j}(y_i;\tau)u_{ij} = \sum_{i=1}^{n} w(y_i;\tau,\hat{F})s_{\beta_j}(y_i;\tau)u_{ij},\\
\Psi_{\sigma}(y;X,\tau) &= \sum_{i=1}^{n} \psi_{\sigma_j}(y_i;\tau)u_{ij} = \sum_{i=1}^{n} w(y_i;\tau,\hat{F})s_{\sigma_j}(y_i;\tau)u_{ij}.
\end{aligned}$$

Very general conditions for consistency and asymptotic normality of the solution to (10) are given in Elashoff and Ryan (2004), whereas Bai et al. (2012) gives conditions in the case of M-estimators. The main requirements are that

1. ψ defines an unbiased estimating function, i.e. $E_{\tau}[\psi(Y;X,\tau)] = 0$;
2. $E_{\tau}[\Psi(Y;X,\tau)\Psi(Y;X,\tau)^{\mathrm{T}}]$ exists and is positive definite;
3. $E_{\tau}[\partial\Psi(Y;X,\tau)/\partial\tau]$ exists and is negative definite, $\forall\tau$.

This conditions are satisfied by the proposed WLEE, that are characterized by weighted score functions as in (10) (see also the Supplementary material in Agostinelli and Greco (2019)). Since the WLEE can be considered as M-type estimating equations and all the above requirements are fulfilled, one can state the following result, along the lines of Bai et al. (2012). Under the regularity conditions of Sect. 2, under the further identifiability conditions of the model (4) given in Hennig (2000), existence, consistency and asymptotic normality of the WLE $\hat{\tau}^{w}$ implicitly defined by equation (10) hold. In particular, the asymptotic covariance matrix of $\hat{\tau}^{w}$ can be obtained in the usual sandwich fashion. Consistency is defined conditionally on the true labels and concerns the case in which the WLEE admits a unique solution. In the presence of multiple solutions, the selection of the consistent root can be effectively pursued according to the strategies described in Sect. 4.

6 Outlier detection

The WEM and WCEM algorithms lead to classify all the sample units, both genuine and contaminated observations, meaning that also outliers are assigned to a cluster. Actually, we are not interested in classifying outliers and for purely clustering purposes outliers have to be discarded. Outlier detection should be based on the robust fitted model and performed separately by using formal rules. The key ingredients in outlier detection are the (scaled) residuals. For a fixed significance level α, an observation is flagged as an outlier when the corresponding residual in absolute value exceeds a fixed threshold, corresponding to the $(1-\alpha/2)$-level quantile of the reference standard normal distribution. In the case of finite mixtures, the main idea is that the outlyingness of each data point should be measured conditionally on the final assignment (Greco and Agostinelli 2020), i.e. an observation is flagged as outlying when

$$\frac{|y_i - \hat{\mu}_{ik_i}|}{\hat{\sigma}_{k_i}} > z_{1-\frac{\alpha}{2}} \tag{11}$$

with $\hat{\mu}_{ik_i} = x_i \hat{\beta}_{k_i}$. Popular choices are $\alpha = 0.05$ and $\alpha = 0.01$. The process of outlier detection may result in type-I and type-II errors. In the former case, a genuine observation is wrongly flagged as outlier (swamping), in the latter case, a true outlier is not identified (masking). Swamped genuine observations are false positives, whereas masked outliers are false negatives. A measure of the level of the test is provided by the rate of false positives, whereas the power of the testing procedure is given by the rate of true positives. A large number of flagged outliers is expected to lead to high power but, then, genuine observations are likely to be misclassified, that is swamping also increases. On the other side, with a low rate of correctly flagged true outliers, the power and the level are expected to both decrease. The outliers detection process could also be designed to take into account multiplicity arguments in the simultaneous testing of all the n data points. For instance, one could base the outlier detection rule on the False Discovery Rate (FDR, Cerioli and Farcomeni (2011)).

7 Illustrative examples with synthetic data

The overall behavior of WEM and WCEM is illustrated in the following examples based on simulated data. The proposed methodology has been tested on some data configurations that have been already used in the literature concerning robust fitting of mixtures of regression lines. The interest lies on both fitting and classification accuracy and in the outlier detection testing rule. The WLEE are based on a symmetric Chi-squared RAF. For each example, we display the data with their original clustering and the true regression lines superimposed and, in separated panels, the results stemming from WEM and WCEM. The outlier detection rule relies on the FDR at a 5% level. We use different symbols and colors for the clusters with a black + standing for the detected outliers (and the true outliers in the panel with the true assignments). In every situation the classical EM and CEM algorithms give unreliable results because of contamination in the sample at hand.

Example 1 Let us consider a mixture of three simple normal linear regressions. The regression lines were generated according to the models

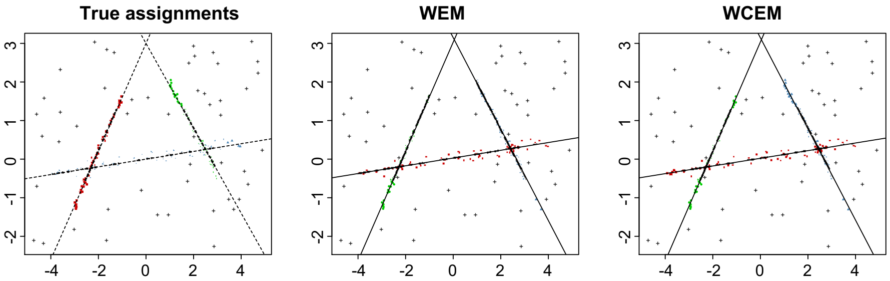

Fig. 2 Example 1. True assignments (left), WEM (middle), WCEM (right)

$$\begin{cases} y_1 = 3 + 1.4x + 0.1\epsilon \\ y_2 = 3 - 1.1x + 0.1\epsilon \\ y_3 = 0.2x + 0.1\epsilon \end{cases}$$

with $\epsilon \sim N(0,1)$ (Neykov et al. 2007). The clusters' sizes are 70, 70, 60, respectively. Then 50 outliers were added that are uniformly distributed in the rectangle that contains the genuine data points. Outliers are such that their distance from the true regression lines, as measured by the scaled residual in absolute value, is above the 0.95-level quantile of the standard normal distribution. The data, the fitted models and the final classification are displayed in Fig. 2: the left panel gives the true assignments and the true lines, the middle panel and the right panel display the results stemming from WEM and WCEM, respectively, with $c_\sigma = 2$. The weighted likelihood methodology provides quite satisfactory outcomes both in terms of fitting and classification accuracy.

To illustrate the problem concerning the initialization of the WEM and WCEM algorithms combined with the choice of c_σ and the selection of the best root, we consider different starting points obtained by varying the tuning parameters of TCLUST (α, c), for a fixed $h = 0.015$ and $c_\sigma = 2, 20$. The top row panels of Fig. 3 display the empirical downweighting level at convergence stemming from WEM. In each monotoring plot, three groups of solutions are apparent: in the central part we find the majority of solutions leading to a correct downweighting level (root 1), in the bottom left corner there are some solutions characterized by insufficient downweighting (root 2), whereas in the top right corner there are those solutions characterized by an excess of downweighting (root 3). Examples of root 2 and root 3 are given in the bottom row panels of Fig. 3, respectively. The root selection criterion based on the minimum fitted approximate disparity (9) leads to choose the right solution: we have $\tilde{\rho}(f^*, m^*) = 0.96$ for root 1, $\tilde{\rho}(f^*, m^*) = 1.31$ for root 2 and $\tilde{\rho}(f^*, m^*) = 1.58$ for root 3.

Example 2 Let us consider a mixture of two regression lines. Genuine data are drawn according to the model

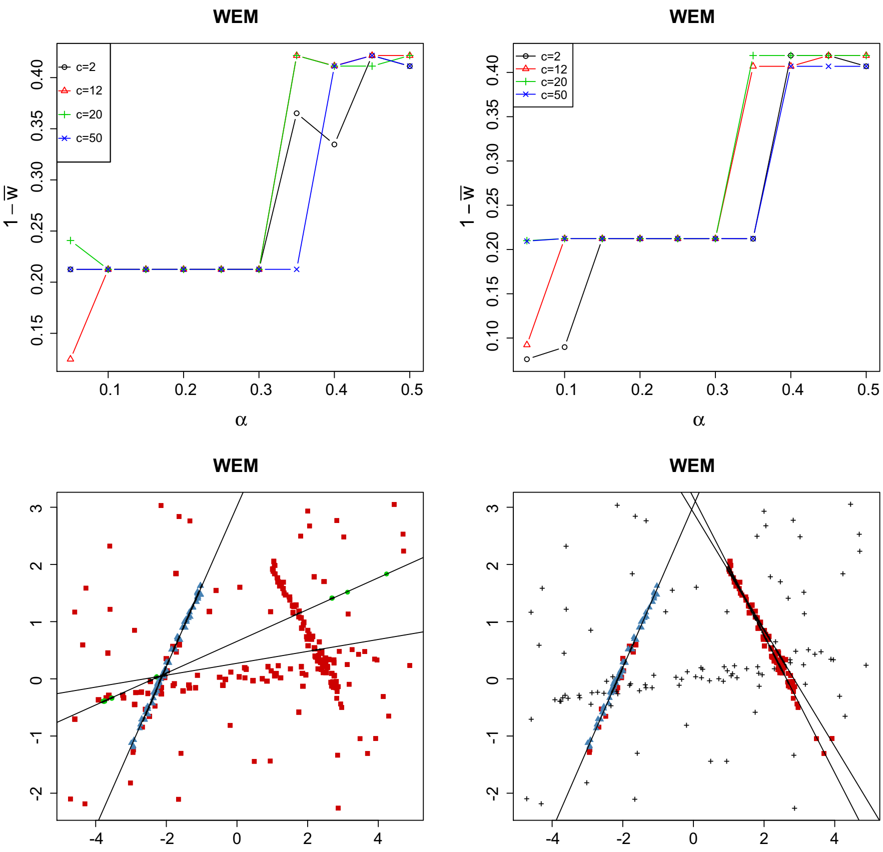

Fig. 3 Example 1. Top row: monitoring of $1-\bar{w}$ by varying (α, c) of the intial TCLUST for $c_\sigma = 2$ (left) and $c_\sigma = 20$. Bottom row: WEM root 2 (middle) and WEM root 3 (right)

$$\begin{cases} y_1 = 1 + x + \epsilon \\ y_2 = 3 + 5x + \epsilon \end{cases}$$

with $\epsilon \sim N(0,1)$ (Bai et al. 2012). Each group is composed by 100 points. By looking at the plots in Fig. 4, we notice that the two clusters are overlapped and the regression lines share the same sign of the slope. Then, 20 clustered bad leverage points are added in the top left corner that violate the patterns exhibited by the genuine points. In this scenario, both the classical EM and CEM lead to a fitted mixture in which one fitted component is wrongly rotated and attracted by the outliers, whereas the other is not able to fit neither of the two true linear structures. On the contrary, the behavior of the robust techniques is satisfactory. Here we set $(\alpha, c, c_\sigma) = (0.25, 2, 2)$.

Example 3 Let us consider a data constellation inspired by García-Escudero et al. (2009). We have a mixture of three linear models disposed according to a slanted π

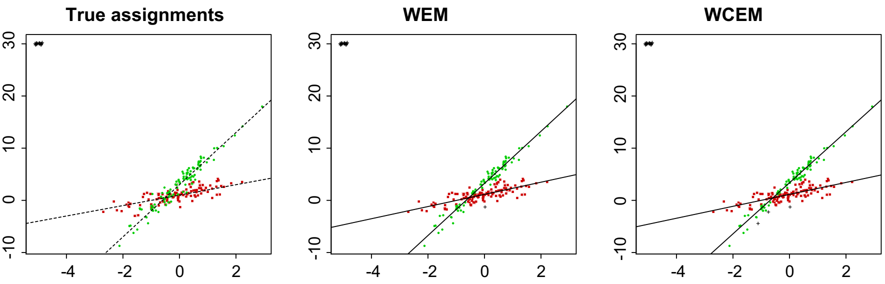

Fig. 4 Example 2. True assignments (left), WEM (middle), WCEM (right)

configuration. The sample size is 300, data are simulated according to equal membership probabilities. There are 50 outliers that are of two types: 25 are scattered in the rectangle that contains the genuine observations, 25 are inliers, since they lie between the linear patterns. Figure 5 displays the data and the results. The weighted likelihood methodology still provides accurate and satisfactory results. Here we set $(\alpha, c, c_\sigma) = (0.25, 50, 2)$.

Example 4 Here, we consider a data constellation similar to that analyzed in García-Escudero et al. (2017) (see their Fig. 6). The solutions displayed in the middle and right panel of Fig. 6 have been obtained for $c_\sigma = 5$. The initial TCLUST settings are $(\alpha, c) = (0.25, 10)$. The results are in strong agreement with those stemming from TCWRM, that, nevertheless, needs the specification of a further constraint on the eigenvalues in the covariates' space.

Example 5 This example has been taken from García-Escudero et al. (2010). In that paper, tha authors proposed TCLUST-REG allowing for a *second trimming* step to handle those data points acting as bad leverage points for the linear regressions. On the contrary, weighted likelihood regression is able to deal with outliers in the x-space and, according to our experience, there is not the need to introduce a second trimming. The data includes two linear regression clusters made up of 225 observations each from the model

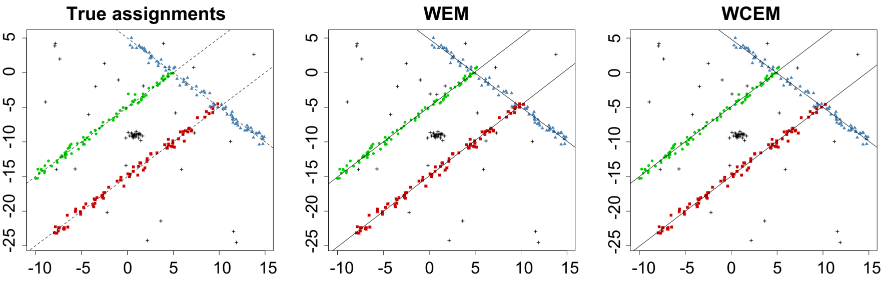

Fig. 5 Example 3. True assignments (left), WEM (middle), WCEM (right)

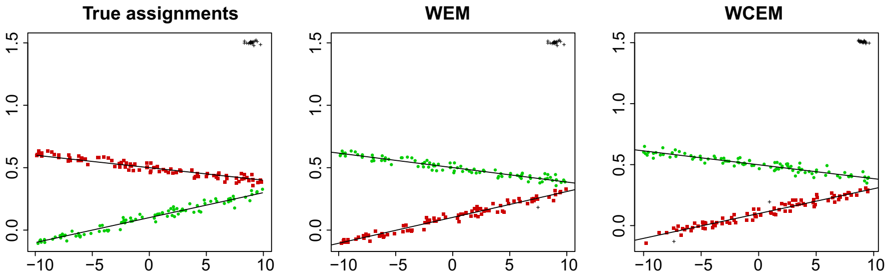

Fig. 6 Example 4. True assignments (left), WEM (middle), WEM (right)

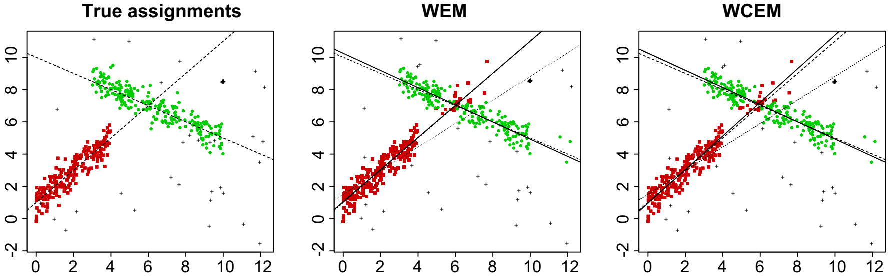

Fig. 7 Example 5. True assignments (left), WEM (middle), WCEM (right)

$$\begin{cases} y_1 = 1 + x + 0.5\epsilon \\ y_2 = 10 - 0.5x + 0.5\epsilon \end{cases}$$

with $\epsilon \sim N(0, 1)$. Then, 30 points are generated as a background noise and, finally, 20 more data points are concentrated around the point (10, 8.5), acting as bad leverage points in the estimation of one linear structure. This data configuration will be also considered in the numerical studies in Sect. 9 as a part of larger numerical study following the lines of García-Escudero et al. (2010). Figure 7 displays the true assignments with the true lines and the fitted models by WEM and WCEM, with $(\alpha, c, c_\sigma) = (0.25, 10, 2)$. In the middle and right panel, we superimposed both the true lines and the regression lines fitted by the trimmed likelihood, to better appreciate the nice behavior of WEM and WCEM in this scenario, since the trimmed likelihood approach of Neykov et al. (2007) is not able to take into account bad leverages. Actually, trimmed likelihood estimation suffers from the presence of the group of bad leverages, since one regression line is rotated towards their direction. On the contrary, the weighted likelihood technique still gives robust estimates, in a fashion similar to TCLUST-REG, but without any second trimming. It is worth noting that both WEM and WCEM wrongly classify some data points, even if characterized by large uncertainties. Actually, the misclassified points by WEM and WCEM are about those trimmed in the second step of TCLUST-REG.

8 Selecting the number of latent classes

The choice of the number of latent classes K is more than an open issue in latent class modeling. In the classical likelihood-based framework, a very general approach is to minimize over K some complexity-penalized version of the (negative) loglikelihood function, such as the well known BIC, AIC and ICL. Agostinelli (2002) and Greco and Agostinelli (2020) tackled the problem of model selection by introducing a weighted version of the AIC and BIC respectively, in which the genuine loglikelihood is replaced by its weighted counterpart evaluated at the WLE. Then, the proposed strategy is based on minimizing

$$Q^w(K) = -2q^w(y; \hat{\tau}) + m(K) \tag{12}$$

where $q^w(y; \hat{\tau}) = \sum_{i=1}^n \hat{w}_{ik_i} \ell(y_i; \hat{\tau})$ and the weights are those stemming from the largest fitted model (Agostinelli 2002). An alternative criterion can be built on the idea of minimum disparity by minimizing a penalized approximate disparity

$$D^w(K) = 2n\tilde{\rho}_k(f^*, m^*) + m(K), \tag{13}$$

where the subscript now is meant to stress the dependency on the number of latent classes. The penalty term $m(K)$ reflects model complexity depending on the number of free parameters. Since the larger the scatter similarity constraint the higher model complexity, Cerioli et al. (2018a) suggested a modified version of the penalty term $m(K, c_\sigma)$ that is also aimed at taking into account model complexity entailed by scatter similarity constraints.

For illustration purposes, let us consider the synthetic data used in Example 1. We tackle the problem of choosing between $K = 2$, $K = 3$ and $K = 4$ components. The results are shown in Table 1: both criteria (12) and (13) lead to choose the right number of latent classes (in bold).

It is worth to point out that criteria such those defined in (12) and (13) should be better used in conjunction with appropriate monitoring strategies, for instance by investigating their behavior as the smoothing parameter h varies (Agostinelli and Greco 2018; Farcomeni and Dotto 2018). In particular, at least in this example, the choice $K = 2$ always leads to a remarkable larger empirical downweighting level for every value of h.

9 Numerical studies

In this section the finite sample behavior of the proposed WEM and WCEM methodologies has been investigated by some numerical studies. We consider a mixture of two regression lines, i.e. with $p = 2$, according to the model described in

Table 1 Example 1: selection of the number of latent classes

K	$Q^w(K)$	$D^w(K)$
2	19908.88	866.27
3	**33.35**	**534.99**
4	39.66	543.19

Example 5. Here, by following the lines of García-Escudero et al. (2010), the covariates in the second groups are drawn from a uniform distribution $U(D, D+7)$, where the tuning parameter D controls the degree of overlapping by setting 3, 6, 12. Two different degrees of complexity have been taken into account: in the first we set equal clusters' proportions $\pi_1 = \pi_2$ and scales $\sigma_1 = \sigma_2 = 0.5$, whereas in the second we assumed unequal proportions and variances, with $\pi_1 = 0.4,\ \pi_2 = 0.6$, and $\sigma_1 = 0.4,\ \sigma_2 = 0.6$. The behavior of WEM and WCEM has been investigated both when any contamination does not occur ($\epsilon = 0$) and when outliers are present. For what concerns the contamination rates, we set $\epsilon = 10\%, 25\%$. Two types of outliers configurations have been considered. In the first scenario, outliers are generated as background noise (Cont.1), whereas in the second scenario we have both background noisy points and bad leverage points concentrated around a point mass (Cont.2). The considered sample size is $n = 500$. Table 2 summarizes the structure of the data for each combination of complexity, scenario and outliers' rate. Furthermore, we also considered the case with $p = 4$ by adding uninformative explanatory variables, that is the corresponding coefficients are set to zero. In summary, the numerical studies are composed by $2 \times 5 \times 3 \times 2 = 60$ separate simulations. The numerical studies are based on 500 Monte Carlo trials.

The weighted likelihood algorithms are based on a symmetric Chi-square RAF. The smoothing parameter h has been selected in such a way that the empirical downweighting level lies in the range (0.15, 0.20) for $\epsilon = 0.10$ and (0.35, 0.45) for $\epsilon = 0.25$, whereas it is about 0.10 when no outliers occur. We set $(\alpha, c, c_\sigma) = (0.25, 10, 2)$. The algorithm is assumed to reach convergence when $\max|\hat{\beta}^{(s+1)} - \hat{\beta}^{(s)}| < tol$, with a tolerance tol set to 10^{-4}, where $\hat{\beta}^{(s)}$ is the matrix of centroids estimates at the sth iteration and the differences are elementwise. The algorithms run on non-optimized R code.

Fitting accuracy has been evaluated according to the Mean Squared Error (MSE) for the mixture parameters, whereas classification accuracy has been measured by the Adjusted Rand Index (ARI) evaluated over true negatives, i.e. genuine observations that are not wrongly declared outliers. In order to detect outliers, we considered a testing rule with $\alpha = 0.01$, according to (11). In addition, we also

Table 2 Data configurations used in the numerical studies with $n = 500$, $p = 2, 4$

Complexity	Scenario	ϵ	Background outliers	Bad leverages
	No contamination	0	0	0
$\pi_1 = \pi_2$	Cont.1	0.10	50	0
$\sigma_1 = \sigma_2$		0.25	125	0
	Cont.2	0.10	30	20
		0.25	75	50
	No contamination	0	0	0
$\pi_1 \neq \pi_2$	Cont.1	0.10	50	0
$\sigma_1 \neq \sigma_2$		0.25	125	0
	Cont.2	0.10	30	20
		0.25	75	50

adopted a strategy based on the FDR for the same overall level, in order to take into account multiplicity effects. Then, we reported the empirical level and power of the test, measured as the swamping rate and the rate of true positives respectively, as explained in Sect. 6. Of course, when $\epsilon = 0$, swamping only is taken into account. The performance of the proposed WEM and WCEM has been compared with the classical EM and CEM (fitted by using the functions available from the R package `flexmix`), their M-type counterparts, MEM and MCEM, based on M-estimation at the M-step and TCLUST-REG. Here, we considered M-estimation based on the Tukey biweight function for an 85% efficiency level, whereas TCLUST-REG runs with the first trimming level set equal to 0.10 for $\epsilon = 0$ and the actual contamination rate under contamination and the second trimming level set to 0.15, as in García-Escudero et al. (2010). It is worth to point out that TCLUST-REG is built on a CEM-type algorithm. We implemented our own non optimized R code according to the details given in García-Escudero et al. (2010). In particular, we use the same starting values used for WEM and WCEM. In order to make a fair comparison across the different methodologies, we stress that, according to existing literature about it, we do not consider any testing strategy after TCLUST-REG but trimmed observations coincide with detected outliers.

As an overall result, we do appreciate the satisfactory behaviour of WEM and WCEM in terms of both classification and fitting accuracy. In particular, the superiority of WEM and WCEM with respect to the *oracle* TCLUST-REG make them quite valuable and promising. WEM and WCEM exhibit a satisfactory efficiency loss at the true model when contamination does not occur, whereas they provide stable results under contamination by dealing with outliers successfully. All Tables showing the detailed results of the numerical studies are given in the Appendix. The entries in Table 4 give the ARI evaluated over true negatives after that outliers have been discarded according to a testing rule based on a fixed 1% level or by controlling the overall level of the multiple testing procedure by using the FDR, when $p = 2$. The results are quite satisfactory both for the genuine and contaminated data. The classification accuracy clearly improves for increasing values of the tuning parameter D and there are no relevant differences when using a fixed level or multiplicity issues are taken into account. Table 8 give the results for the case $p = 4$. Table 5 gives the MSE corresponding to the fitted mixture parameters (β, σ, π) stemming from all the considered techniques, for $p = 2$, the entries in Table 9 give the results for the case $p = 4$. The overall behavior of WEM and WCEM is quite accurate.

Here, in order to present the results, the simulation setting has been divided in 9 macro scenarios by collapsing them with respect to $\epsilon = 0, 0.10, 0.25$ and $D = 3, 6, 12$. The empirical distribution of the ARI is displayed in Fig. 8. The results corresponding to the overlapping level $D = 12$ have not been reported since classification accuracy is almost always perfect for all macro-scenarios and procedures. Furthermore, the ARI for maximum likelihood under contamination is not given since it is well below those stemming from the robust techniques. The reader is pointed to the Tables given in the Appendix section. Classification accuracy provided by WEM and WCEM is quite satisfactory. We notice that WEM and WCEM improves over TCLUST-REG, in particular in the challenging case

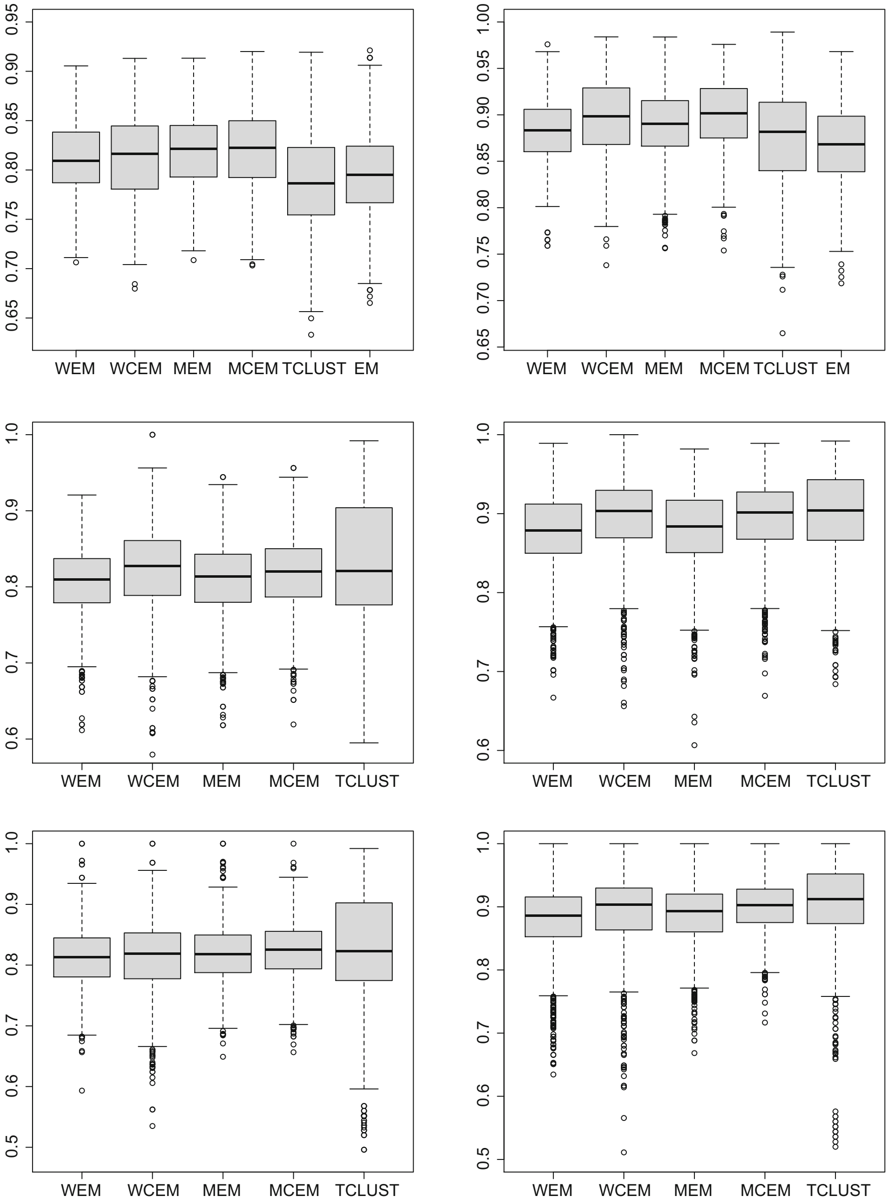

Fig. 8 ARI for WEM, WCEM, MEM, MCEM, TCLUST-REG, EM for $\epsilon = 0$ (top), $\epsilon = 0.10$ (middle) $\epsilon = 0.25$ (right) and $D = 3$ (left), $D = 6$ (right)

$D = 3$. Figure 9 gives the empirical distribution of the Sum of Squares for the regression coefficients. The overall behavior of WEM and WCEM is quite accurate in all macro scenarios. The loss of efficiency with respect to maximum likelihood is

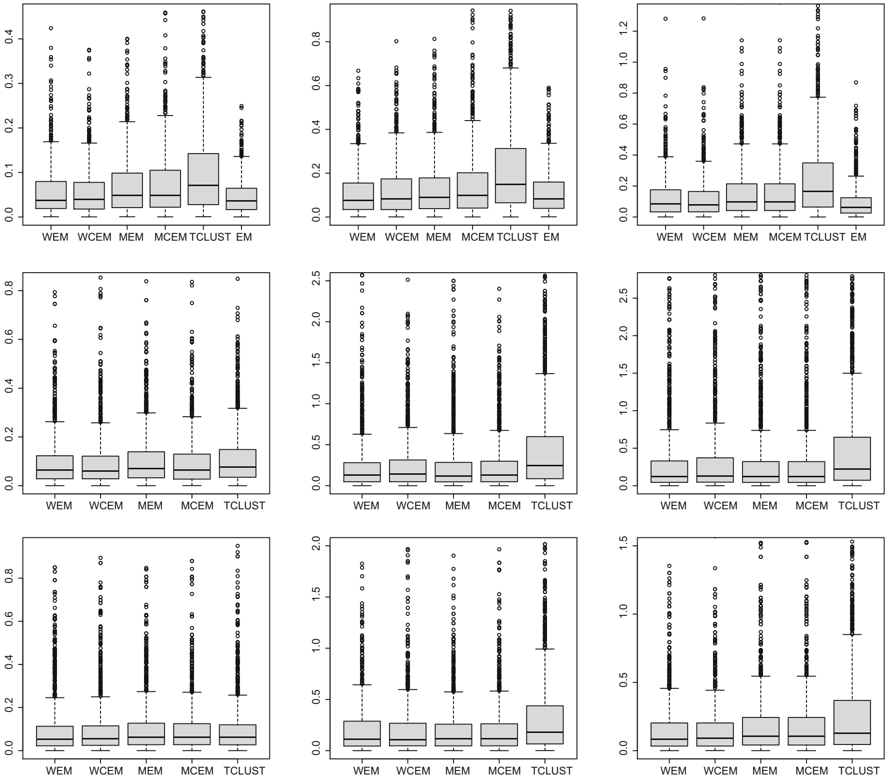

Fig. 9 MSE for WEM, WCEM, MEM, MCEM, TCLUST-REG, EM for $\epsilon = 0$ (top), $\epsilon = 0.10$ (middle) $\epsilon = 0.25$ (right) and $D = 3$ (left), $D = 6$ (middle), $D = 12$ (right)

negligible when no outliers occurs. The performance with respect to M-based techniques is often quite similar. On the other hand, WEM and WCEM improve over the *oracle* TCLUST-REG, in particular for $D = 6, 12$.

Swamping and power of the outliers tests are given in Tables 6 and 7, respectively, for $p = 2$, whereas Table 10 and Table 10 give the results for $p = 4$. It is worth to stress that the behavior of the tests depends on the actual robustness-efficiency trade-off of the procedure, that is on the value of the selected bandwidth parameter h for weighted likelihood estimation. In summary, the chosen values of h lead to an appreciable compromise between swamping and power. WEM and WCEM well compares with the results stemming from the other methods, in particular with TCLUST-REG. It is worth to notice that FDR leads to improved swamping but lower power than those resulting from the use of a fixed threshold (Cerioli et al. 2018a).

10 Pinus nigra data set

The data gives the height (in meters) and diameter (in millimeters) of $n = 362$ Pinus nigra trees located in the north of Palencia (Spain). The Diameter is considered as an explicative variables wheres Height is the response. The data are displayed in the left panel of Fig. 11. They exhibit the presence of three linear groups apart from a small group of trees forming its own cluster on the top right corner and one isolated point on the bottom right corner. The example has been taken from García-Escudero et al. (2010). We ran WEM and WCEM by setting $h = 0.01$, with $c_\sigma = 2$ and employing a symmetric chi square RAF. The starting solution stems from an unconstrained TCLUST (actually $c = 500$) wth $\alpha = 0.1$. The same solution has been obtained by using the denoising approach suggested by Coretto and Hennig (2017). The smoothing parameter has been selected according to a monitoring strategy. The left panel of Fig. 10 shows the empirical downweighting level as h varies on a fixed grid. We selected the value where the empirical downweighting level stabilizes. By monitoring the change in individual residuals (in absolute value) as h varies, we observe that the clustered outliers and the isolated outlier are clearly spotted during all the monitoring process and that the other data points have residuals below the threshold line for most of the monitoring. Here, the cut-off has been set equal to the square root of the qth quantile of the χ^2_1 distribution with $q = 1 - 0.99^{1/n}$, by using a Bonferroni adjustment to take into account multiplicity. The fitted models and detected outliers are shown in Fig. 11. The outlier detection rule is based on the FDR at 1% level. The results are in strong agreement with those stemming from TCLUST-REG. Evidence for the choice $K = 3$ has been confirmed by using the criteria (12) and (13), as given in Table 3.

The fitted model suggests that a more parsimonious mixture model characterized by homogeneous slopes may be fitted to the data at hand. To this end, we ran the

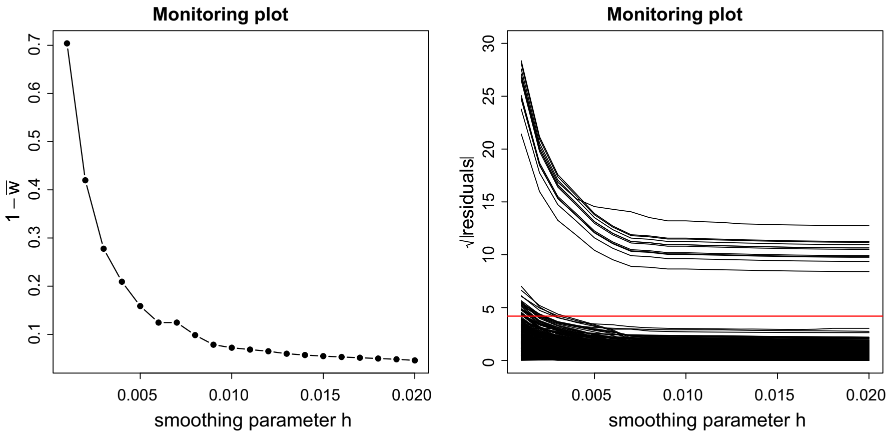

Fig. 10 Pinus nigra. Monitoring of the empirical downweighting level (left) and clusterwise residuals (right) from WEM

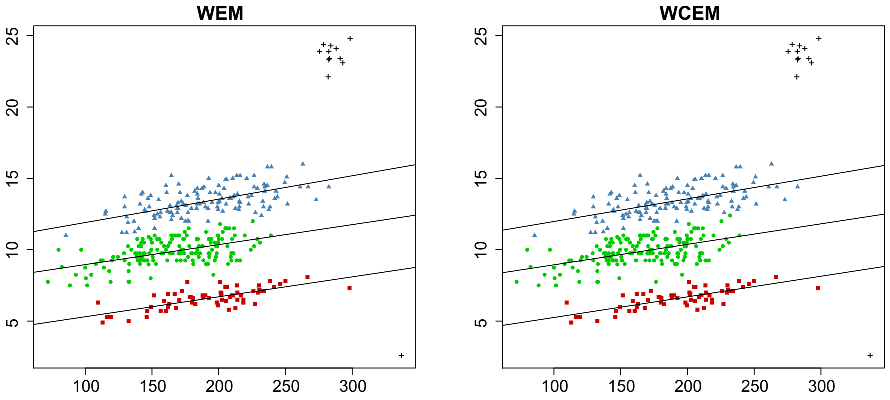

Fig. 11 Pinus nigra. Fitted mixtures by WEM (left) and WCEM (right). Clusters are denoted by different colors and symbols. Outliers are denoted by $+$

Table 3 Pinus nigra

K	$Q^w(K)$	$D^w(K)$
2	1514.70	621.29
3	**1468.09**	**585.66**
4	1489.23	617.21

Selection of the number of latent classes

WEM algorithm by assuming homogeneous slopes, that is using the estimating equations given in (8) for what concerns the estimation of the three class-dependent intercepts and the common slope. The fitted slope is $\hat{\beta}_{-0} = 0.015$, whereas the fitted intercepts are $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (3.75, 7.41, 10.44)$. The outlier detection rule (6) leads to identify the same data points. In order to test if there is evidence supporting the reduced model with homogeneous slopes, one could resort to the weighted likelihood ratio test (WLRT) developed in Agostinelli and Markatou (2001). The WLRT is obtained as

$$\Lambda^{oss} = 2\sum_{i=1}^{n} \hat{w}_i[\ell(y_i; \hat{\tau}) - \ell(y_i; \hat{\tau}_R)]$$

where $\hat{\tau}_R$ denotes the WLE under the reduced model. In this example we have $\Lambda^{oss} = 2.45$ with a p-value $Prob\left(\chi^2_2 > 2.45\right) = 0.29$, confirming evidence for the reduced model against the full model with class dependent slopes.

11 Concluding remarks

In this paper we developed weighted likelihood estimation for mixtures of linear structures in the presence of contamination in the data at hand. The proposed techniques behave satisfactory in all the considered scenario providing both fitting and classification accuracy. Furthermore, the suggested outliers detection rule exhibits reasonable level and power. The method inherits the main properties characterizing weighted likelihood estimation both in terms of efficiency at the assumed model and robustness in the presence of outliers. Both the WEM and WCEM compare satisfactory with existing methods. One of the main aspects concerns the selection of the smoothing parameter tuning the efficiency/robustness trade-off in finite samples. However, the same problem arises for the other robust techniques that all need some constant to be tuned. The researcher is advised to resort to a monitoring strategy for an effective tuning. Then, one clear advantage of the method is the availability of weighted counterparts of the likelihood ratio test and the information criteria with the standard asymptotic behavior.

Some possible further directions of research could concern initialization issues in order to reduce the number of initial partitions but also more challenging model selection problems that were out of the scope of the present paper, indeed. Moreover, the extension of the proposed methodology to mixtures of linear regressions with concomitant variables or to mixtures of generalized linear models seems feasible.

Acknowledgements The authors would like to thank two anonymous referees for their valuable comments and suggestions leading to an improved version of the paper.

Appendix

See Tables 4, 5, 6, 7, 8, 9, 10 and 11.

Table 4 Adjusted Rand Index evaluated over true negatives for WEM, WCEM, MEM, MCEM, TCLUST-REG, EM, for $p = 2$, different type of contamination, rate of contamination and degree of overlapping D among linear clusters

		WEM		WCEM		MEM		MCEM		TCLUST-REG	EM
	D	Fixed	FDR	Fixed	FDR	Fixed	FDR	Fixed	FDR		
		$\pi_1 = \pi_2$ and $\sigma_1 = \sigma_2$									
No Cont	3	0.811	0.813	0.807	0.810	0.812	0.817	0.813	0.818	0.788	0.793
	6	0.883	0.885	0.893	0.895	0.881	0.886	0.889	0.897	0.872	0.858
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4 continued

		WEM		WCEM		MEM		MCEM		TCLUST-REG	EM
	D	Fixed	FDR	Fixed	FDR	Fixed	FDR	Fixed	FDR		
Cont.1 10%	3	0.809	0.811	0.808	0.811	0.812	0.817	0.816	0.821	0.792	0.478
	6	0.874	0.876	0.892	0.893	0.878	0.879	0.894	0.895	0.876	0.236
	12	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	1.000	0.012
Cont.1 25%	3	0.792	0.794	0.782	0.783	0.792	0.795	0.799	0.802	0.781	0.392
	6	0.863	0.865	0.873	0.875	0.855	0.857	0.875	0.877	0.870	0.018
	12	0.997	0.998	0.996	0.997	0.996	0.997	0.997	0.998	0.998	0.003
Cont.2 10%	3	0.810	0.812	0.801	0.803	0.809	0.813	0.810	0.814	0.799	0.529
	6	0.871	0.873	0.882	0.884	0.881	0.884	0.894	0.896	0.887	0.328
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.099
Cont.2 25%	3	0.796	0.799	0.777	0.799	0.810	0.813	0.811	0.815	0.785	0.447
	6	0.849	0.862	0.871	0.873	0.869	0.872	0.896	0.897	0.873	0.283
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.006
		$\pi_1 \neq \pi_2$ and $\sigma_1 \neq \sigma_2$									
No cont	3	0.806	0.809	0.808	0.811	0.813	0.818	0.823	0.823	0.796	0.799
	6	0.885	0.886	0.898	0.900	0.888	0.890	0.895	0.897	0.887	0.881
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Cont.1 10%	3	0.807	0.810	0.821	0.824	0.814	0.818	0.821	0.825	0.804	0.571
	6	0.883	0.884	0.905	0.907	0.890	0.892	0.898	0.900	0.893	0.314
	12	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	0.029
Cont.1 25%	3	0.802	0.812	0.842	0.845	0.804	0.806	0.814	0.817	0.798	0.307
	6	0.883	0.890	0.918	0.920	0.883	0.884	0.899	0.900	0.885	0.000
	12	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	0.998	0.014
Cont.2 10%	3	0.814	0.816	0.824	0.826	0.817	0.820	0.824	0.827	0.805	0.565
	6	0.887	0.888	0.903	0.904	0.894	0.895	0.901	0.902	0.897	0.384
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.125
Cont.2 25%	3	0.819	0.823	0.851	0.854	0.817	0.820	0.825	0.828	0.806	0.453
	6	0.887	0.889	0.915	0.917	0.885	0.886	0.900	0.902	0.887	0.239
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.018

The outlier detection rule is based on a 0.01 level by using a fixed level and the False Discovery Rate

Table 5 Mean Squared Error for WEM, WCEM, MEM, MCEM, TCLUST-REG, EM, for $p = 2$, different type of contamination, rate of contamination and degree of overlapping D among linear clusters

	D	WEM			WCEM			MEM			MCEM			TCLUST-REG			EM		
		β	σ	π	β	σ	π	β	σ	π	β	σ	π	β	σ	π	β	σ	π
	$\pi_1 = \pi_2$ and $\sigma_1 = \sigma_2$																		
No cont.	3	0.031	0.021	0.006	0.031	0.021	0.009	0.036	0.052	0.007	0.034	0.057	0.007	0.046	0.033	0.024	0.031	0.002	0.008
	6	0.095	0.019	0.003	0.118	0.018	0.004	0.106	0.050	0.004	0.131	0.053	0.004	0.159	0.029	0.012	0.094	0.002	0.004
	12	0.097	0.018	0.002	0.094	0.015	0.002	0.135	0.048	0.002	0.135	0.048	0.002	0.127	0.025	0.002	0.082	0.001	0.001
Cont.1 10%	3	0.039	0.015	0.006	0.041	0.026	0.010	0.043	0.047	0.007	0.042	0.052	0.007	0.054	0.003	0.017	8.601	2.832	0.032
	6	0.107	0.015	0.003	0.126	0.024	0.004	0.099	0.046	0.003	0.118	0.049	0.003	0.261	0.002	0.007	43.128	2.450	0.075
	12	0.112	0.014	0.001	0.109	0.020	0.002	0.112	0.044	0.002	0.112	0.044	0.002	0.029	0.000	0.000	43.416	4.614	0.333
Cont.1 25%	3	0.092	0.058	0.009	0.092	0.071	0.019	0.100	0.048	0.008	0.083	0.095	0.008	0.138	0.006	0.021	25.674	10.238	0.023
	6	0.312	0.055	0.004	0.345	0.063	0.011	0.376	0.044	0.005	0.389	0.048	0.004	0.649	0.004	0.009	45.566	12.105	0.208
	12	0.592	0.051	0.002	0.639	0.055	0.003	0.736	0.042	0.001	0.679	0.042	0.001	0.984	0.003	0.001	47.330	11.251	0.205
Cont.2 10%	3	0.038	0.019	0.007	0.036	0.032	0.015	0.039	0.051	0.008	0.040	0.057	0.008	0.035	0.018	0.017	2.614	1.298	0.038
	6	0.115	0.018	0.004	0.129	0.024	0.007	0.121	0.053	0.006	0.131	0.053	0.004	0.175	0.015	0.008	38.110	1.149	0.028
	12	0.088	0.015	0.001	0.092	0.019	0.002	0.121	0.048	0.002	0.121	0.048	0.002	0.120	0.012	0.001	44.612	0.914	0.196
Cont.2 25%	3	0.101	0.036	0.008	0.097	0.062	0.015	0.101	0.065	0.012	0.083	0.065	0.010	0.107	0.009	0.018	23.982	5.952	0.004
	6	0.230	0.040	0.009	0.243	0.050	0.014	0.234	0.066	0.012	0.211	0.056	0.007	0.369	0.006	0.009	46.083	3.450	0.028
	12	0.231	0.052	0.002	0.253	0.055	0.005	0.259	0.047	0.004	0.259	0.047	0.004	0.424	0.002	0.002	45.546	9.149	0.156
		$\pi_1 \neq \pi_2$ and $\sigma_1 \neq \sigma_2$																	
No cont.	3	0.027	0.013	0.006	0.028	0.010	0.004	0.035	0.049	0.006	0.039	0.053	0.004	0.044	0.036	0.018	0.029	0.002	0.008
	6	0.123	0.021	0.004	0.131	0.018	0.003	0.141	0.054	0.004	0.161	0.056	0.004	0.164	0.032	0.008	0.095	0.001	0.004
	12	0.145	0.018	0.001	0.133	0.014	0.002	0.177	0.048	0.002	0.177	0.048	0.002	0.147	0.026	0.002	0.091	0.001	0.001
Cont.1 10%	3	0.041	0.017	0.007	0.039	0.022	0.004	0.046	0.050	0.006	0.044	0.055	0.006	0.040	0.018	0.014	8.431	2.401	0.022
	6	0.151	0.016	0.003	0.165	0.020	0.002	0.128	0.047	0.003	0.155	0.050	0.003	0.190	0.016	0.006	40.737	1.500	0.041
	12	0.123	0.014	0.001	0.134	0.016	0.003	0.111	0.043	0.001	0.111	0.043	0.001	0.316	0.000	0.001	39.665	4.297	0.165

Table 5 continued

	D	WEM			WCEM			MEM			MCEM			TCLUST-REG			EM		
		β	σ	π	β	σ	π	β	σ	π	β	σ	π	β	σ	π	β	σ	π
Cont.1 25%	3	0.071	0.061	0.007	0.067	0.060	0.011	0.018	0.051	0.007	0.070	0.056	0.007	0.122	0.005	0.013	22.809	9.932	0.038
	6	0.256	0.058	0.004	0.270	0.063	0.014	0.343	0.047	0.004	0.324	0.051	0.003	0.715	0.004	0.007	42.871	12.608	0.116
	12	0.521	0.054	0.019	0.885	0.044	0.047	0.845	0.045	0.038	0.707	0.045	0.039	1.013	0.003	0.002	46.845	12.007	0.098
Cont.2 10%	3	0.030	0.020	0.007	0.031	0.026	0.006	0.037	0.052	0.006	0.037	0.057	0.006	0.036	0.018	0.014	2.417	1.002	0.033
	6	0.106	0.019	0.004	0.123	0.023	0.003	0.119	0.051	0.004	0.133	0.053	0.003	0.179	0.016	0.006	31.062	0.908	0.016
	12	0.098	0.017	0.001	0.097	0.019	0.003	0.127	0.047	0.001	0.127	0.047	0.001	0.152	0.013	0.001	44.473	0.705	0.153
Cont.2 25%	3	0.071	0.064	0.008	0.060	0.069	0.013	0.059	0.053	0.008	0.059	0.056	0.006	0.083	0.005	0.013	20.032	4.868	0.009
	6	0.192	0.059	0.004	0.176	0.065	0.013	0.118	0.049	0.003	0.200	0.052	0.003	0.469	0.003	0.006	42.672	2.961	0.050
	12	0.275	0.055	0.002	0.301	0.060	0.020	0.331	0.046	0.002	0.331	0.046	0.002	0.618	0.002	0.001	43.350	9.931	0.065

Table 6 Swamping rate for WEM, WCEM, MEM, MCEM and TCLUST-REG, $p = 2$, for different type of contamination, rate of contamination and degree of overlapping D among linear clusters

		WEM		WCEM		MEM		MCEM		TCLUST-
	D	Fixed	FDR	Fixed	FDR	Fixed	FDR	Fixed	FDR	REG
		$\pi_1 = \pi_2$ and $\sigma_1 = \sigma_2$								
No Cont	3	0.035	0.003	0.034	0.003	0.070	0.012	0.078	0.014	0.147
	6	0.036	0.003	0.033	0.003	0.070	0.013	0.075	0.014	0.131
	12	0.037	0.003	0.033	0.003	0.073	0.013	0.073	0.013	0.100
Cont.1 10%	3	0.030	0.008	0.043	0.014	0.068	0.027	0.076	0.031	0.061
	6	0.032	0.007	0.041	0.011	0.068	0.026	0.073	0.029	0.045
	12	0.033	0.008	0.040	0.011	0.072	0.027	0.072	0.027	0.011
Cont.1 25%	3	0.052	0.050	0.075	0.063	0.067	0.038	0.076	0.044	0.075
	6	0.071	0.049	0.072	0.057	0.066	0.037	0.071	0.041	0.066
	12	0.078	0.047	0.069	0.052	0.066	0.037	0.066	0.037	0.027
Cont.2 10%	3	0.034	0.009	0.047	0.015	0.073	0.030	0.081	0.035	0.102
	6	0.036	0.009	0.041	0.011	0.082	0.039	0.079	0.035	0.086
	12	0.034	0.008	0.038	0.010	0.076	0.033	0.076	0.033	0.056
Cont.2 25%	3	0.058	0.031	0.087	0.056	0.101	0.071	0.095	0.063	0.072
	6	0.079	0.045	0.077	0.045	0.109	0.075	0.087	0.057	0.054
	12	0.081	0.047	0.085	0.051	0.075	0.044	0.087	0.057	0.016
		$\pi_1 \neq \pi_2$ and $\sigma_1 \neq \sigma_2$								
No cont	3	0.034	0.003	0.032	0.004	0.070	0.011	0.076	0.013	0.138
	6	0.037	0.003	0.033	0.003	0.074	0.012	0.077	0.014	0.127
	12	0.035	0.003	0.032	0.003	0.072	0.011	0.072	0.011	0.100
Cont.1 10%	3	0.031	0.007	0.038	0.011	0.069	0.027	0.076	0.032	0.093
	6	0.032	0.007	0.038	0.010	0.068	0.027	0.072	0.029	0.084
	12	0.030	0.007	0.034	0.009	0.067	0.024	0.068	0.024	0.011
Cont.1 25%	3	0.083	0.049	0.104	0.070	0.068	0.038	0.076	0.044	0.058
	6	0.085	0.051	0.103	0.070	0.070	0.040	0.074	0.042	0.060
	12	0.083	0.049	0.078	0.047	0.069	0.038	0.069	0.039	0.027
Cont.2 10%	3	0.033	0.009	0.041	0.013	0.070	0.027	0.077	0.032	0.093
	6	0.034	0.009	0.040	0.012	0.072	0.030	0.075	0.030	0.082
	12	0.034	0.009	0.039	0.011	0.071	0.028	0.072	0.028	0.056
Cont.2 25%	3	0.090	0.056	0.104	0.070	0.074	0.044	0.077	0.045	0.053
	6	0.084	0.050	0.102	0.067	0.068	0.038	0.072	0.041	0.050
	12	0.084	0.049	0.104	0.070	0.070	0.040	0.071	0.040	0.017

The outlier detection rule is based on a 0.01 level by using a fixed level and the False Discovery Rate

Table 7 Power of the outlier test for WEM, WCEM, MEM, MCEM and TCLUST, $p = 2$, for different type of contamination, rate of contamination and degree of overlapping D among linear clusters

		WEM		WCEM		MEM		MCEM		TCLUST
	D	Fixed	FDR	Fixed	FDR	Fixed	FDR	Fixed	FDR	
		$\pi_1 = \pi_2$ and $\sigma_1 = \sigma_2$								
Cont.1 10%	3	0.981	0.912	0.988	0.938	0.993	0.975	0.995	0.980	0.921
	6	0.976	0.919	0.988	0.941	0.990	0.970	0.994	0.976	0.926
	12	0.988	0.928	0.991	0.944	0.998	0.981	0.998	0.981	0.932
Cont.1 25%	3	0.968	0.956	0.966	0.953	0.971	0.958	0.982	0.971	0.945
	6	0.978	0.969	0.982	0.973	0.972	0.959	0.978	0.968	0.942
	12	0.986	0.982	0.985	0.979	0.983	0.974	0.985	0.975	0.951
Cont.2 10%	3	0.962	0.962	0.989	0.989	0.984	0.984	0.984	0.984	0.997
	6	0.918	0.958	0.962	0.962	0.957	0.957	0.978	0.978	0.995
	12	0.989	0.989	0.995	0.995	0.979	0.979	0.979	0.979	0.998
Cont.2 25%	3	0.927	0.913	0.950	0.943	0.944	0.939	0.963	0.959	0.923
	6	0.926	0.923	0.947	0.940	0.924	0.915	0.969	0.963	0.938
	12	0.997	0.993	0.996	0.993	0.980	0.980	0.986	0.980	0.968
		$\pi_1 \neq \pi_2$ and $\sigma_1 \neq \sigma_2$								
Cont.1 10%	3	0.978	0.915	0.985	0.930	0.992	0.972	0.995	0.978	0.993
	6	0.972	0.913	0.980	0.925	0.990	0.966	0.993	0.972	0.991
	12	0.988	0.923	0.987	0.923	0.998	0.979	0.998	0.979	0.932
Cont.1 25%	3	0.987	0.980	0.991	0.985	0.980	0.972	0.987	0.980	0.950
	6	0.987	0.979	0.987	0.979	0.977	0.967	0.983	0.976	0.944
	12	0.990	0.984	0.977	0.963	0.983	0.975	0.984	0.976	0.951
Cont.2 10%	3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997
	6	0.984	0.984	0.995	0.995	0.995	0.995	1.000	1.000	0.994
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998
Cont.2 25%	3	0.986	0.982	0.993	0.988	0.982	0.977	0.992	0.987	0.956
	6	0.992	0.987	0.994	0.989	0.992	0.985	0.994	0.989	0.959
	12	0.997	0.993	0.995	0.990	0.991	0.987	0.991	0.987	0.966

The outlier detection rule is based on a 0.01 level by using a fixed level and the False Discovery Rate

Table 8 Adjusted Rand Index evaluated over true negatives for WEM, WCEM, MEM, MCEM, TCLUST, EM, for $p = 4$, different type of contamination, rate of contamination and degree of overlapping among linear clusters

		WEM		WCEM		MEM		MCEM		TCLUST-	EM
	D	Fixed	FDR	Fixed	FDR	Fixed	FDR	Fixed	FDR	REG	
		$n_1 = n_2$ and $\sigma_1 = \sigma_2$									
No Cont	3	0.808	0.810	0.803	0.806	0.810	0.815	0.812	0.818	0.793	0.7928
	6	0.876	0.877	0.887	0.889	0.875	0.879	0.884	0.888	0.867	0.858
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.0009
Cont.1 10%	3	0.807	0.809	0.809	0.812	0.812	0.815	0.816	0.819	0.795	0.500
	6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.880	0.069
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.012
Cont.1 25%	3	0.789	0.791	0.778	0.779	0.793	0.794	0.801	0.804	0.785	0.385
	6	0.857	0.859	0.871	0.872	0.855	0.857	0.877	0.880	0.868	0.019
	12	0.997	0.997	0.997	0.996	0.994	0.994	0.998	0.998	0.998	0.003
Cont.2 10%	3	0.807	0.809	0.802	0.805	0.815	0.819	0.817	0.821	0.800	0.521
	6	0.860	0.862	0.878	0.880	0.883	0.886	0.895	0.897	0.881	0.273
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.036
Cont.2 25%	3	0.804	0.807	0.770	0.771	0.803	0.806	0.809	0.814	0.786	0.443
	6	0.870	0.873	0.863	0.865	0.883	0.885	0.899	0.900	0.878	0.285
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.006
		$n_1 \neq n_2$ and $\sigma_1 \neq \sigma_2$									
No cont	3	0.807	0.809	0.813	0.816	0.811	0.816	0.814	0.819	0.787	0.801
	6	0.885	0.887	0.899	0.901	0.889	0.891	0.894	0.896	0.881	0.880
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Cont.1 10%	3	0.803	0.806	0.824	0.826	0.812	0.817	0.818	0.823	0.805	0.517
	6	0.887	0.887	0.906	0.907	0.892	0.895	0.904	0.906	0.890	0.074
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.038
Cont.1 25%	3	0.803	0.806	0.838	0.841	0.804	0.808	0.814	0.818	0.801	0.282
	6	0.877	0.879	0.914	0.915	0.880	0.882	0.897	0.899	0.887	0.000
	12	0.993	0.993	0.988	0.988	0.993	0.993	0.993	0.993	0.998	0.015
Cont.2 10%	3	0.815	0.817	0.828	0.829	0.820	0.822	0.827	0.830	0.808	0.559
	6	0.886	0.887	0.907	0.908	0.896	0.898	0.903	0.905	0.898	0.292
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.019
Cont.2 25%	3	0.816	0.819	0.833	0.837	0.819	0.822	0.823	0.827	0.807	0.445
	6	0.890	0.892	0.907	0.909	0.892	0.894	0.898	0.900	0.889	0.253
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.018

The outlier detection rule is based on a 0.01 level by using a fixed level and the False Discovery Rate

Table 9 Mean Squared Error for WEM, WCEM, MEM, MCEM, TCLUST, EM, for $p = 4$, different type of contamination, rate of contamination D and degree of overlapping among linear clusters

	D	WEM			WCEM			MEM			MCEM			TCLUST-REG			EM		
		β	σ	π	β	σ	π	β	σ	π	β	σ	π	β	σ	π	β	σ	π
	$\pi_1 = \pi_2$ and $\sigma_1 = \sigma_2$																		
No cont.	3	0.101	0.022	0.007	0.099	0.023	0.010	0.125	0.057	0.008	0.130	0.063	0.008	0.111	0.034	0.016	0.064	0.002	0.008
	6	0.143	0.021	0.004	0.167	0.019	0.005	0.164	0.055	0.005	0.188	0.058	0.005	0.311	0.095	0.023	0.134	0.002	0.004
	12	0.151	0.018	0.001	0.147	0.015	0.002	0.189	0.050	0.002	0.189	0.050	0.002	0.380	0.087	0.005	0.113	0.001	0.001
Cont.1 10%	3	0.094	0.013	0.007	0.096	0.021	0.009	0.119	0.053	0.008	0.119	0.059	0.007	0.109	0.003	0.012	18.744	3.577	0.014
	6	0.245	0.011	0.001	0.231	0.015	0.002	0.190	0.048	0.001	0.190	0.048	0.001	0.302	0.002	0.006	53.500	4.582	0.284
	12	0.174	0.011	0.001	0.171	0.016	0.002	0.166	0.049	0.002	0.166	0.049	0.002	0.299	0.002	0.001	41.891	3.744	0.302
Cont.1 25%	3	0.203	0.045	0.009	0.202	0.061	0.027	0.200	0.053	0.010	0.185	0.061	0.009	0.190	0.006	0.015	26.620	10.245	0.021
	6	0.485	0.043	0.005	0.514	0.052	0.008	0.477	0.051	0.005	0.472	0.056	0.004	0.716	0.005	0.008	47.008	11.928	0.208
	12	0.775	0.035	0.002	0.825	0.044	0.006	0.959	0.047	0.002	0.773	0.047	0.002	1.005	0.003	0.001	49.180	11.226	0.205
Cont.2 10%	3	0.089	0.016	0.007	0.083	0.027	0.014	0.104	0.057	0.008	0.106	0.063	0.008	0.099	0.004	0.011	7.975	1.954	0.022
	6	0.188	0.015	0.006	0.190	0.023	0.008	0.181	0.061	0.007	0.200	0.063	0.006	0.270	0.003	0.007	43.775	1.438	0.021
	12	0.139	0.013	0.001	0.145	0.018	0.002	0.190	0.055	0.004	0.190	0.055	0.004	0.206	0.002	0.001	45.937	3.553	0.309
Cont.2 25%	3	0.230	0.065	0.013	0.242	0.066	0.029	0.223	0.072	0.013	0.203	0.067	0.009	0.175	0.011	0.016	24.754	5.928	0.005
	6	0.343	0.070	0.012	0.339	0.063	0.027	0.325	0.087	0.020	0.306	0.069	0.009	0.416	0.007	0.008	46.400	3.426	0.028
	12	0.346	0.042	0.002	0.345	0.047	0.005	0.377	0.057	0.007	0.377	0.057	0.007	0.464	0.002	0.002	45.177	9.092	0.155
	$\pi_1 \neq \pi_2$ and $\sigma_1 \neq \sigma_2$																		
No cont.	3	0.092	0.023	0.008	0.090	0.020	0.006	0.115	0.058	0.008	0.122	0.063	0.007	0.191	0.108	0.023	0.063	0.002	0.008
	6	0.171	0.022	0.004	0.182	0.019	0.003	0.198	0.056	0.004	0.219	0.059	0.004	0.343	0.100	0.014	0.135	0.001	0.004
	12	0.169	0.019	0.001	0.166	0.015	0.002	0.210	0.052	0.002	0.210	0.052	0.002	0.403	0.091	0.004	0.120	0.001	0.001
Cont.1 10%	3	0.098	0.013	0.007	0.099	0.020	0.005	0.117	0.054	0.007	0.115	0.059	0.007	0.096	0.003	0.008	10.515	2.593	0.008
	6	0.211	0.012	0.003	0.235	0.018	0.002	0.208	0.051	0.003	0.229	0.054	0.002	0.365	0.002	0.005	41.346	4.100	0.141
	12	0.190	0.012	0.001	0.197	0.016	0.003	0.206	0.050	0.002	0.206	0.050	0.002	0.343	0.002	0.001	43.977	4.378	0.169

Table 9 continued

	D	WEM			WCEM			MEM			MCEM			TCLUST-REG			EM		
		β	σ	π	β	σ	π	β	σ	π	β	σ	π	β	σ	π	β	σ	π
Cont.1 25%	3	0.172	0.049	0.007	0.177	0.058	0.008	0.177	0.058	0.007	0.171	0.064	0.006	0.182	0.005	0.010	23.823	9.860	0.037
	6	0.453	0.045	0.004	0.478	0.053	0.010	0.437	0.053	0.004	0.427	0.057	0.004	0.766	0.004	0.006	45.029	12.334	0.117
	12	0.657	0.043	0.002	0.975	0.052	0.022	0.675	0.051	0.002	0.671	0.051	0.002	1.089	0.003	0.002	48.623	11.830	0.097
Cont.2 10%	3	0.085	0.017	0.007	0.089	0.027	0.006	0.112	0.060	0.007	0.114	0.066	0.007	0.087	0.003	0.008	3.187	1.080	0.011
	6	0.168	0.015	0.004	0.176	0.023	0.003	0.200	0.059	0.005	0.210	0.060	0.004	0.265	0.002	0.004	35.369	0.950	0.005
	12	0.145	0.014	0.001	0.150	0.019	0.003	0.197	0.053	0.002	0.197	0.053	0.002	0.227	0.002	0.001	42.874	5.131	0.201
Cont.2 25%	3	0.111	0.050	0.007	0.120	0.058	0.007	0.109	0.056	0.007	0.112	0.061	0.006	0.161	0.005	0.010	21.033	4.948	0.011
	6	0.189	0.049	0.004	0.201	0.055	0.007	0.187	0.054	0.004	0.205	0.057	0.003	0.533	0.004	0.005	44.834	3.121	0.043
	12	0.225	0.045	0.001	0.217	0.048	0.008	0.229	0.053	0.002	0.225	0.052	0.002	0.742	0.003	0.001	43.262	9.856	0.063

Table 10 Swamping rate for WEM, WCEM, MEM, MCEM and TCLUST, $p = 4$, for different type of contamination, rate of contamination and degree of overlapping D among linear clusters

		WEM		WCEM		MEM		MCEM		TCLUST
	D	Fixed	FDR	Fixed	FDR	Fixed	FDR	Fixed	FDR	
		$\pi_1 = \pi_2$ and $\sigma_1 = \sigma_2$								
No Cont	3	0.037	0.004	0.037	0.004	0.077	0.015	0.086	0.019	0.133
	6	0.038	0.003	0.035	0.003	0.078	0.017	0.082	0.018	0.282
	12	0.037	0.003	0.033	0.003	0.077	0.014	0.077	0.014	0.258
Cont.1 10%	3	0.027	0.007	0.037	0.011	0.076	0.033	0.084	0.038	0.047
	6	0.029	0.007	0.034	0.009	0.076	0.032	0.076	0.032	0.050
	12	0.037	0.010	0.031	0.007	0.075	0.031	0.076	0.030	0.021
Cont.1 25%	3	0.066	0.038	0.084	0.054	0.075	0.045	0.087	0.054	0.060
	6	0.069	0.039	0.079	0.048	0.078	0.046	0.083	0.052	0.069
	12	0.066	0.036	0.073	0.042	0.076	0.044	0.076	0.044	0.036
Cont.2 10%	3	0.031	0.008	0.043	0.013	0.081	0.036	0.091	0.042	0.043
	6	0.035	0.008	0.042	0.012	0.096	0.050	0.097	0.051	0.047
	12	0.032	0.007	0.038	0.011	0.090	0.045	0.090	0.045	0.019
Cont.2 25%	3	0.107	0.077	0.098	0.062	0.116	0.083	0.102	0.067	0.061
	6	0.122	0.091	0.098	0.065	0.147	0.118	0.110	0.079	0.061
	12	0.070	0.039	0.077	0.045	0.094	0.063	0.094	0.063	0.026
		$\pi_1 \neq \pi_2$ and $\sigma_1 \neq \sigma_2$								
No cont	3	0.037	0.004	0.035	0.004	0.077	0.015	0.084	0.018	0.273
	6	0.038	0.004	0.034	0.004	0.079	0.017	0.083	0.018	0.278
	12	0.037	0.004	0.032	0.003	0.076	0.014	0.076	0.014	0.258
Cont.1 10%	3	0.027	0.006	0.037	0.010	0.076	0.033	0.085	0.039	0.036
	6	0.027	0.006	0.036	0.010	0.073	0.032	0.076	0.034	0.048
	12	0.028	0.007	0.036	0.010	0.077	0.032	0.077	0.032	0.021
Cont.1 25%	3	0.069	0.039	0.092	0.062	0.078	0.048	0.088	0.056	0.048
	6	0.070	0.039	0.090	0.059	0.079	0.047	0.084	0.050	0.063
	12	0.069	0.039	0.096	0.065	0.079	0.047	0.079	0.047	0.036
Cont.2 10%	3	0.030	0.008	0.044	0.015	0.084	0.039	0.091	0.045	0.034
	6	0.029	0.006	0.026	0.005	0.071	0.028	0.078	0.032	0.043
	12	0.027	0.006	0.025	0.005	0.072	0.027	0.073	0.027	0.018
Cont.2 25%	3	0.070	0.028	0.086	0.041	0.077	0.033	0.083	0.038	0.042
	6	0.075	0.033	0.085	0.040	0.085	0.035	0.078	0.035	0.054
	12	0.071	0.028	0.082	0.037	0.083	0.037	0.080	0.034	0.028

The outlier detection rule is based on a 0.01 level by using a fixed level and the False Discovery Rate

Table 11 Power of the outlier test for WEM, WCEM, ME, MCEM and TCLUS $p = 4$, for different type of contamination, rate of contamination and degree of overlapping D among linear clusters

		WEM		WCEM		MEM		MCEM		TCLUST-
	D	Fixed	FDR	Fixed	FDR	Fixed	FDR	Fixed	FDR	REG
		$\pi_1 = \pi_2$ and $\sigma_1 = \sigma_2$								
Cont.1 10%	3	0.973	0.900	0.980	0.927	0.993	0.975	0.995	0.980	0.920
	6	0.978	0.916	0.982	0.929	0.997	0.982	0.997	0.982	0.924
	12	0.988	0.927	0.981	0.909	0.997	0.978	0.997	0.999	0.930
Cont.1 25%	3	0.971	0.957	0.974	0.963	0.973	0.964	0.982	0.975	0.945
	6	0.966	0.952	0.970	0.959	0.969	0.959	0.977	0.967	0.941
	12	0.982	0.971	0.981	0.968	0.981	0.974	0.983	0.976	0.951
Cont.2 10%	3	0.956	0.956	1.000	1.000	1.000	1.000	1.000	1.000	0.935
	6	0.824	0.824	0.941	0.941	0.954	0.954	0.965	0.965	0.930
	12	0.968	0.968	0.979	0.979	0.980	0.980	0.980	0.980	0.948
Cont.2 25%	3	0.914	0.897	0.921	0.914	0.916	0.910	0.958	0.953	0.915
	6	0.968	0.954	0.974	0.962	0.970	0.960	0.978	0.968	0.933
	12	0.991	0.985	0.992	0.986	0.979	0.975	0.979	0.975	0.967
		$\pi_1 \neq \pi_2$ and $\sigma_1 \neq \sigma_2$								
Cont.1 10%	3	0.973	0.900	0.981	0.923	0.991	0.971	0.994	0.978	0.926
	6	0.968	0.894	0.976	0.916	0.988	0.968	0.992	0.971	0.922
	12	0.982	0.921	0.983	0.929	0.999	0.983	0.999	0.983	0.931
Cont.1 25%	3	0.975	0.972	0.979	0.978	0.977	0.974	0.982	0.982	0.951
	6	0.975	0.963	0.979	0.966	0.977	0.968	0.982	0.975	0.944
	12	0.981	0.970	0.973	0.960	0.981	0.974	0.981	0.974	0.948
Cont.2 10%	3	0.989	0.989	0.989	0.989	0.989	0.989	0.989	0.989	0.946
	6	0.946	0.894	0.954	0.901	0.988	0.963	0.988	0.965	0.941
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.950
Cont.2 25%	3	0.996	0.996	1.000	1.000	0.996	0.996	1.000	1.000	0.955
	6	0.986	0.985	0.996	0.996	0.995	0.995	1.000	1.000	0.958
	12	1.000	1.000	1.000	1.000	0.991	0.990	1.000	1.000	0.964

The outlier detection rule is based on a 0.01 level by using a fixed level and the False Discovery Rate

References

Agostinelli C (2002) Robust model selection in regression via weighted likelihood methodology. Stat Probab Lett 56(3):289–300

Agostinelli C (2006) Notes on Pearson residuals and weighted likelihood estimating equations. Stat Probab Lett 76(17):1930–1934

Agostinelli C, Greco L (2013) A weighted strategy to handle likelihood uncertainty in Bayesian inference. Comput Stat 28(1):319–339

Agostinelli C, Greco L (2018) Discussion on "The power of monitoring: how to make the most of a contaminated sample". Stat Methods Appl 27(4):609–619

Agostinelli C, Greco L (2019) Weighted likelihood estimation of multivariate location and scatter. Test 28(3):756–784

Agostinelli C, Markatou M (1998) A one-step robust estimator for regression based on the weighted likelihood reweighting scheme. Stat Probab Lett 37(4):341–350

Agostinelli C, Markatou M (2001) Test of hypotheses based on the weighted likelihood methodology. Stat Sin 11:499–514

Alqallaf F, Agostinelli C (2016) Robust inference in generalized linear models. Commun Stat Simul Comput 45(9):3053–3073

Bai X, Yao W, Boyer JE (2012) Robust fitting of mixture regression models. Comput Stat Data Anal 56(7):2347–2359

Bashir S, Carter E (2012) Robust mixture of linear regression models. Commun Stat Theory Methods 41(18):3371–3388

Basu A, Lindsay B (1994) Minimum disparity estimation for continuous models: efficiency, distributions and robustness. Ann Inst Stat Math 46(4):683–705

Campbell N (1984) Mixture models and atypical values. Math Geol 16(5):465–477

Cerioli A, Farcomeni A (2011) Error rates for multivariate outlier detection. Comput Stat Data Anal 55(1):544–553

Cerioli A, García-Escudero LA, Mayo-Iscar A, Riani M (2018a) Finding the number of normal groups in model-based clustering via constrained likelihoods. J Comput Gr Stat 27(2):404–416

Cerioli A, Riani M, Atkinson AC, Corbellini A (2018b) The power of monitoring: how to make the most of a contaminated multivariate sample. Stat Methods Appl 27:1–29

Coretto P, Hennig C (2017) Consistency, breakdown robustness, and algorithms for robust improper maximum likelihood clustering. J Mach Learn Res 18(1):5199–5237

Dotto F, Farcomeni A, García-Escudero LA, Mayo-Iscar A (2017) A fuzzy approach to robust regression clustering. Adv Data Anal Classif 11(4):691–710

Elashoff M, Ryan L (2004) An EM algorithm for estimating equations. J Comput Gr Stat 13(1):48–65

Farcomeni A, Dotto F (2018) The power of (extended) monitoring in robust clustering. Stat Methods Appl 27(4):651–660

Farcomeni A, Greco L (2015a) Robust methods for data reduction. CRC Press, Boca Raton

Farcomeni A, Greco L (2015b) S-estimation of hidden Markov models. Comput Stat 30(1):57–80

Fritz H, Garcia-Escudero L, Mayo-Iscar A (2013) A fast algorithm for robust constrained clustering. Comput Stat Data Anal 61:124–136

García-Escudero L, Gordaliza A, Matrán C, Mayo-Iscar A (2008) A general trimming approach to robust cluster analysis. Ann Stat 36(3):1324–1345

Garcia-Escudero L, Gordaliza A, Matran C, Mayo-Iscar A (2015) Avoiding spurious local maximizers in mixture modeling. Stat Comput 25(3):619–633

García-Escudero LA, Gordaliza A, San Martin R, Van Aelst S, Zamar R (2009) Robust linear clustering. J R Stat Soc Ser B (Statistical Methodology) 71(1):301–318

García-Escudero LA, Gordaliza A, Mayo-Iscar A, San Martín R (2010) Robust clusterwise linear regression through trimming. Comput Stat Data Anal 54(12):3057–3069

García-Escudero LA, Gordaliza A, Greselin F, Ingrassia S, Mayo-Íscar A (2017) Robust estimation of mixtures of regressions with random covariates, via trimming and constraints. Stat Comput 27(2):377–402

Greco L (2017) Weighted likelihood based inference for P (X< Y). Commun Stat Simul Comput 46(10):7777–7789

Greco L, Agostinelli C (2020) Weighted likelihood mixture modeling and model-based clustering. Stat Comput 30(2):255–277

Hennig C (2000) Identifiablity of models for clusterwise linear regression. J Classif 17(2):273–296

Lindsay BG (1994) Efficiency versus robustness: the case for minimum Hellinger distance and related methods. Ann Stat 22(2):1081–1114

Markatou M, Basu A, Lindsay BG (1998) Weighted likelihood equations with bootstrap root search. J Am Stat Assoc 93(442):740–750

Maronna R, Martin RD, Yohai V, Salibian-Barrera M (2019) Robust statistics: theory and methods (with R). Wiley, Hoboken

Neykov N, Filzmoser P, Dimova R, Neytchev P (2007) Robust fitting of mixtures using the trimmed likelihood estimator. Comput Stat Data Anal 52(1):299–308

Neykov NM, Müller CH (2003) Breakdown point and computation of trimmed likelihood estimators in generalized linear models. In: Dutter R, Filzmoser P, Gather U, Rousseeuw PJ (eds) Developments in robust statistics. Physica-Verlag, Heidelberg, pp 277–286

Punzo A, McNicholas PD (2017) Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. J Classif 34(2):249–293

Riani M, Cerioli A, Atkinson A, Perrotta D, Torti F (2008) Fitting mixtures of regression lines with the forward search. Min Massive Data Sets Secur Adv Data Min Search Soc Netw Text Min Appl Secur 19:271

Torti F, Perrotta D, Riani M, Cerioli A (2019) Assessing trimming methodologies for clustering linear regression data. Adv Data Anal Classif 13(1):227–257

Yao W, Wei Y, Yu C (2014) Robust mixture regression using the t-distribution. Comput Stat Data Anal 71:116–127

Yu C, Yao W, Chen K (2017) A new method for robust mixture regression. Can J Stat 45(1):77–94

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.