Robust fitting of mixtures of GLMs by weighted likelihood

Luca Greco¹

Received: 12 July 2020 / Accepted: 25 April 2021 / Published online: 8 May 2021
© Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Finite mixtures of generalized linear models are commonly fitted by maximum likelihood and the EM algorithm. The estimation process and subsequent inferential and classification procedures can be badly affected by the occurrence of outliers. Actually, contamination in the sample at hand may lead to severely biased fitted components and poor classification accuracy. In order to take into account the potential presence of outliers, a robust fitting strategy is proposed that is based on the weighted likelihood methodology. The technique exhibits a satisfactory behavior in terms of both fitting and classification accuracy, as confirmed by some numerical studies and real data examples.

Keywords Classification · EM · GLM · Mixture · Outliers · Weighted likelihood

Mathematics Subject Classification MSC 62F35 · MSC 62G35 · MSC 62H25 · MSC 62H30

1 Introduction

Mixtures of generalized linear models (GLMs) represent a general and flexible approach to model different type of data when the sample at hand exhibits unobserved heterogeneity or overdispersion issues. Here, we consider finite mixtures of regression models of the form

$$q(y; x, \tau) = \sum_{k=1}^K \pi_k m(y; \mu_k, \theta_k), \quad (1)$$

where K denotes the number of components, $m(y; \mu, \theta)$ is a probability (density) function that is a member of an exponential family of distributions, μ_k are component-specific expected values such that $g(\mu_k) = \eta_k = x\beta_k$, for a certain link function

✉ Luca Greco
luca.greco@unisannio.it

¹ DEMM Department, University of Sannio, Benevento, Italy

$g(\cdot)$, $\beta_k \in \mathbb{R}^p$, $p > 1$ are regression coefficients, x is a p -vector of covariates, π_k are prior membership probabilities, such that $\pi_k \in (0, 1)$ and $\sum_{k=1}^K \pi_k = 1$, $\theta_k > 0$ are dispersion parameters and $\tau = (\pi_1, \dots, \pi_K, \beta_1, \dots, \beta_K, \theta_1, \dots, \theta_K)$ denotes the vector of all parameters. The reader is pointed to McLachlan and Peel (2004) and references therein for a general account on the topic.

It is well known that inference for mixture models based on the likelihood suffers from lack of robustness (Farcomeni and Greco 2015a; Greco and Agostinelli 2020). Actually, the occurrence of outliers can lead to arbitrarily biased component parameters estimates that are not able to unveil the underlying clustering structure in the data. Furthermore, classification rules stemming from the fitted model, as those corresponding to a maximum a posteriori approach, becomes unreliable as long as they depend on inaccurate estimates. There is a large literature about the problem of robust estimation of mixtures, even if the emphasis relies on the case of multivariate Gaussian or linear regression components. Robust estimation is commonly performed by resorting to a suitable modification of the EM (or the Classification EM, CEM hereafter) algorithm obtained by replacing the standard likelihood equations characterizing the M-step. For instance, robust estimation of mixtures can be achieved by using soft-trimming procedures based on M-type estimators (Bai et al. 2012; Bashir and Carter 2012; Farcomeni and Greco 2015b) or on weighted likelihood estimating equations (Markatou 2000; Greco and Agostinelli 2020; Greco et al. 2020). According to this approach, the M-step is enhanced by the computation of data-dependent weights laying in $[0, 1]$, aimed to bound the effect of anomalous observations. In a different but complementary fashion, one could resort to impartial trimming and the `tclust` methodology: a trimming step precedes the M-step, aiming at discarding a fixed rate of observations exhibiting the lowest contributions to the likelihood, based on the current parameters' values. The main results about `tclust` in the framework of mixtures of multivariate Gaussian components are given in García-Escudero et al. (2008) and Fritz et al. (2013); further developments can be found in Dotto et al. (2018), Dotto and Farcomeni (2019), whereas the problem of mixtures of linear regressions has been tackled in García-Escudero et al. (2009, 2010). A related approach has been developed by Neykov and Müller (2003) and Neykov et al. (2007) based on the so-called *trimmed likelihood*.

Despite the interest risen by the problem of robust fitting of mixtures, there are still few contributions in the framework of mixtures of GLMs. Exceptions are represented by Karlis and Xekalaki (2001), Lu et al. (2003) and Neykov et al. (2007), in which the attention is focused on robust estimation of mixtures of Poisson regression models. In general, we can state that the existing techniques still do not reflect the generality of the available theoretical results for mixtures of GLMs, as well as in the standard likelihood framework [the reader is also pointed to Grün and Leisch (2007)]. In this paper, a general procedure for robust fitting of mixtures of GLMs has been developed that is based on the weighted likelihood methodology. The technique makes use of the results from Alqallaf and Agostinelli (2016) concerning weighted likelihood estimation and inference for GLMs and represents an extension of the methodology introduced by Greco and Agostinelli (2020), Greco et al. (2020) in the framework of mixtures of multivariate Gaussian components and linear regressions, respectively.

The paper is organized as follows: Section 2 gives necessary background about weighted likelihood estimation with particular emphasis on GLM; the proposed methodology for robust fitting of the mixture model given in (1) is described in Sect. 3; Sect. 4 gives some rules for outlier detection after robust fitting; Sect. 5 focuses on model selection issues; illustrative examples based on synthetic data are given in Sect. 6; Sect. 7 gives some numerical studies; real data examples are discussed in Sect. 8. Concluding remarks end the paper in Sect. 9.

2 Background

The weighted likelihood methodology (Markatou et al. 1998; Agostinelli and Markatou 2001) provides robust inferential tools under contamination, while keeping full efficiency at the assumed model. This task is achieved by modifying the classical likelihood equations through a set of data-dependent weights whose role is that of bounding the effect on inference of those observations that are unlikely to occur under the postulated model. At the same time, the weights are constructed in such a way to converge uniformly to unity at the assumed model, i.e., when the sample is not corrupted by outliers (Agostinelli and Greco 2013), hence recovering full efficiency.

Let $y = (y_1, y_2, \dots, y_n)^\top$ be a sample of size n from a r.v. Y with distribution function $F(y)$ and probability (density) function $f(y)$. Let $\mathcal{Q} = \{Q(y; \tau), \tau \in T \subseteq \mathbb{R}^d, d \geq 1, y \in \mathcal{Y}\}$ be the assumed parametric model, with corresponding density $q(y; \tau)$, and $\hat{F}_n$ the empirical distribution function. Assume that the support of $Q(y; \tau)$ is the same as that of $F(y)$ and independent of τ . The detection of those data points not in agreement with the assumed model comes from the inspection of Pearson residuals, that are defined as

$$\delta(y) = \delta(y; \tau, F) = \frac{f(y)}{q(y; \tau)} - 1,$$

with $\delta(y) \in [-1, +\infty)$. The finite sample counterpart of (2) can be obtained as

$$\delta_n(y) = \delta(y; \tau, \hat{F}_n) = \frac{\hat{f}_n(y)}{q(y; \tau)} - 1,$$

where $\hat{f}_n(y)$ is a consistent estimate of the true density $f(y)$. In discrete families of distributions, $\hat{f}_n(y)$ can be driven by the observed relative frequencies, whereas in continuous models one could consider a nonparametric density estimate based on the kernel function $k(y; t, h)$, that is

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

Moreover, with continuous data, a smoothed model density is commonly involved in the denominator, that is

$$q^*(y; \tau) = \int_{\mathcal{Y}} k(y; t, h) dQ(t; \tau).$$

Large values of the Pearson residual function correspond to regions of the support $\mathcal{Y}$ where the model fits the data poorly, meaning that the observation is unlikely to occur under the assumed model. Then, outliers are expected to show large Pearson residuals (Basu and Lindsay 1994; Markatou et al. 1998; Agostinelli and Greco 2019). In an opposite fashion, values of $\delta_n(y)$ near minus one tell that there are *holes* in the data with respect to the predicted model (Park et al. 2002). Here, we also want to downweight such inliers. Therefore, in order to limit the effect of those observations not in agreement with the assumed model in the fitting process, a weight function is introduced, defined as

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

where $A(\delta)$ is the residual adjustment function (RAF, Basu and Lindsay 1994) and $[\cdot]^+$ denotes the positive part. The RAF is such that $A(\delta) \leq \delta$ for $\delta \geq 0$ and acts on residuals in a fashion similar to the classical Huber or Tukey function in an M-estimation framework (Maronna et al. 2019). Here, we consider the families of RAF based on the power divergence measure and the generalized Kullback–Leibler divergence (Park et al. 2002; Kuchibhotla and Basu 2015). Then, the weighted likelihood estimate (WLE) is defined as the root of the weighted likelihood estimating equation (WLEE)

$$\sum_{i=1}^n w(y_i; \tau, \hat{F}_n) s(y_i; \tau) = 0, \quad (3)$$

where $s(y_i; \tau)$ is the individual term of the score function. Maximum likelihood is recovered for $A(\delta) = \delta$. Even if other functions could have been considered to bound the effect of large Pearson residuals, we consider the RAF in light of the connections between weighted likelihood estimation and minimum disparity estimation. Actually, the RAF arises naturally from a minimum disparity estimation problem, although the construction of the WLEE does not depend on the availability of an objective function (Markatou et al. 1998; Agostinelli and Greco 2019).

2.1 Weighted likelihood estimation in GLMs

In the specific setting of GLMs, with data of the form (y_i, x_i) , with $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$, let $r(y_i; x_i, \beta, \theta)$ be the i th scaled residual: the deviance or Anscombe residuals are feasible alternative choices. At the assumed model, the distribution of the scaled residuals is expected to be standard normal. Generally speaking, the standardized residuals $r(y_i; x_i, \beta, \theta) / \sqrt{1 - h_{ii}}$, where the h_{ii} are the generalized leverages, tend to be preferable because they are more symmetric, but both are common and can be used. According to the results established in

Alqallaf and Agostinelli (2016) and Agostinelli and Greco (2019), Pearson residuals can be obtained as

$$\delta = \delta(r(y;x, \beta, \theta)) = \frac{\hat{f}_n(r(y;x, \beta, \theta))}{m^*(r)} - 1, \quad (4)$$

where, now, $\hat{f}_n(r)$ is a nonparametric kernel density estimate evaluated over the residuals, rather than on the original data, based on the kernel function $k(r; t, h)$ with bandwidth h and

$$m^*(r) = \int k(r; t, h) \phi(t) dt$$

is a smoothed version of the residuals (asymptotic) model density obtained by using the same kernel function. Here, $\phi(\cdot)$ denotes the standard normal density function. In particular, when a Gaussian kernel is involved, then

$$m^*(r) = \frac{1}{\sqrt{1+h^2}} \phi\left(\frac{r}{\sqrt{1+h^2}}\right).$$

The WLEE for the regression vector coefficient β is

$$(X^\top W \Omega X) \beta = X^\top W \Omega [\eta + g'(\mu)(y - \mu)] \quad (5)$$

where X is the $n \times p$ design matrix, $\mu = g^{-1}(\eta)$, $\eta = X\beta$, $V(\mu)$ is the variance function, $W = \text{diag}(1/[V(\mu_i)g'(\mu_i)^2])$, with prime denoting differentiation, and $\Omega = \text{diag}(w_i)$, $w_i = w(\delta_i)$ and $\delta_i = \delta(r(y_i; x_i, \beta))$. The solution $\hat{\beta}$ stemming from (5) can be found by applying a standard iterative weighted least squares algorithm enhanced by the computation of the weight matrix Ω in each step at the current parameters' values (Alqallaf and Agostinelli 2016).

Consistency and asymptotic normality of the WLE $\hat{\beta}$ are driven by the asymptotic behavior of the Pearson residuals in (4), under some regularity conditions on the model, the kernel and the weight function given in Markatou et al. (1998), Agostinelli and Markatou (2001), Agostinelli and Greco (2019). A sufficient condition to establish consistency and asymptotic normality of the WLE is that the Pearson residuals evaluated from the Anscombe or deviance residuals converge uniformly to zero at the assumed model. As a consequence, at the assumed model the weights in (2) converge uniformly to one and the WLEE coincides with the standard likelihood equations, hence recovering full efficiency. It is worth remarking that the use of deviance residuals in weighted likelihood estimation in a GLM framework is a novel approach, not considered in Alqallaf and Agostinelli (2016).

When the assumed GLM is characterized by an unknown dispersion parameter θ , a WLEE for θ can be driven from the (weighted) method of moments as

$$\sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2 \hat{w}_i}{V(\hat{\mu}_i)} - (n^w - p)\theta = 0 \quad (6)$$

where $n^w = \sum_{i=1}^n \hat{w}_i$, $\hat{\mu}_i = g^{-1}(x_i \hat{\beta})$ and $\hat{w}_i = w(\delta(y_i; x_i, \hat{\beta}))$. The variance of $\hat{\beta}$ can be consistently estimated by the inverse of the weighted Fisher information matrix, evaluated at the WLE, that is $\widehat{\text{Var}}(\hat{\beta}) = \hat{\theta}(X^\top \hat{W} \hat{\Omega} X)^{-1}$, with $\hat{W} = \text{diag}(\hat{w}_i)$.

3 A weighted likelihood EM-type algorithm

Maximum likelihood estimation of mixtures model is commonly achieved by using the EM (Dempster et al. 1977) or the classification EM (CEM) (Celeux and Govaert 1993) algorithm. The EM algorithm works with the complete loglikelihood function

$$\ell_c(\tau) = \sum_{i=1}^n \sum_{k=1}^K u_{ik} \log[\pi_k m(y_i; x_i, \mu_k, \theta_k)],$$

where u_{ik} is an indicator of the i th observation belonging to the k th mixture component. First, at the E-step, based on current parameters' values, at the s th iteration, one obtains posterior probabilities

$$u_{ik}^{(s)} \propto \pi_k m(y_i; x_i, \beta_k^{(s-1)}, \theta_k^{(s-1)}),$$

$i = 1, 2, \dots, n$, $k = 1, 2, \dots, K$. Then, parameter estimates are updated at the M-step by solving K distinct weighted least squares problems with weights $u_{ik}^{(s)}$.

By exploiting the methodology developed in Greco and Agostinelli (2020) and Greco et al. (2020), a weighted version of the classical EM algorithm can be designed introducing a weighting step, characterized by the computation of component-wise weights $w_{ik}^{(s)} = w(\delta_{ik})$, with $\delta_{ik} = \delta_n(r_{ik}^{(s-1)})$ and $r_{ik} = r(y_i; x_i, \beta_k, \theta_k)$. Then, the weighted EM algorithm (WEM) iterates between the classical E step and an M step in which the weighted least square problems also depend on the robustness weights $w_{ik}^{(s)}$. Then, the M step is composed by K complete-data WLEE of the form (5), that is

$$(X^\top W_k U_k \Omega_k X) \beta = X^\top W_k U_k \Omega_k [\eta_k + g'(\mu_k)(y - \mu_k)] \quad (7)$$

now with $W_k = \text{diag}\left(1/\left[V\left(\mu_{ik}^{(s)}\right)g'\left(\mu_{ik}^{(s)}\right)^2\right]\right)$, $\Omega_k = \text{diag}\left(w_{ik}^{(s)}\right)$ and $U_k^{(s)} = \text{diag}\left(u_{ik}^{(s)}\right)$. The s^{th} iteration of the WEM algorithm is summarized in Algorithm 1. In particular, the WEM requires that K kernel density estimates are obtained in each step to compute component-dependent Pearson residuals. In a similar fashion, a weighted CEM (WCEM) algorithm can be developed: let $k_i = \arg\max_k u_{ik}^{(s)}$, then $u_{ik_i}^{(s)} = 1$ and $u_{ik}^{(s)} = 0$ for $k \neq k_i$ and $U_k^{(s)}$ becomes a dummy matrix. Therefore, weights are computed conditionally on the current cluster assignments driven by the classification step.

3.1 Properties

The proposed algorithm represents a special case of the EM-type algorithm first introduced by Elashoff and Ryan (2004), that has been established for general estimating equations. The complete data WLEE (7) to be solved in the M-step is of the type (3), with

$$\Psi_{\tau} = \sum_{i=1}^n w_i s(y_i; x_i, \tau) u_{ik} = \mathbf{0}$$

where, now, $s(y; x, \tau) = \partial \ell_c(y; x, \tau) / \partial \tau^{\top}$ and $\Psi_{\tau} = (\Psi_{\pi}, \Psi_{\beta}, \Psi_{\theta})^{\top}$. The main requirements to establish consistency and asymptotic normality of the proposed WLE for a finite mixture of GLMs are that

1. Ψ_{τ} defines an unbiased estimating function at the assumed model, i.e.,

$$E_{\tau}[\Psi_{\tau}(Y; x, \tau)] = 0$$

2. $V(\tau) = E_{\tau}[\Psi_{\tau}(Y; x, \tau) \Psi_{\tau}(Y; x, \tau)^{\top}]$ exists and is positive definite
3. $B(\tau) = E_{\tau}[\partial \Psi_{\tau}(Y; x, \tau) / \partial \tau]$ exists and is negative definite, $\forall \tau$.

Assume the following:

- (A1) $A''(\delta)(\delta + 1)^{\alpha}$ is bounded for some fixed α and for all $\delta \geq -1$.
- (A2) The supports of $m(\cdot)$ and $f(\cdot)$ are independent of τ .
- (A3) The density $f(\cdot)$ is twice differentiable, with bounded first and second derivatives.
- (A4) The kernel $k(\cdot; \cdot)$ is of bounded variation, with $\lim_{|z| \rightarrow \infty} k(z) = 0$, $h \rightarrow 0$, $nh \rightarrow \infty$, as $n \rightarrow \infty$.
- (A5) The random vectors $(A(\delta(Y)) + 1)m(Y)s(Y; x, \tau)/f(Y; x, \tau)$, $A'(\delta(Y))s(Y; x, \tau)$ and $m^{\alpha-1}(Y)s(Y; x, \tau)/f^{\alpha-1}(Y; x, \tau)$ have component-wise finite moments of some order strictly greater than 2.
- (A6) There exists a compact subset T_0 of T such that

$$\mathbb{E} \sup_{\tau \in T_0} [m^{\alpha-1}(Y)s(Y; x, \tau)s^{\top}(Y; x, \tau)/f^{\alpha-1}(Y; x, \tau)]$$

is finite.

- (A7) $\int |A'(\delta(y))| f(y; x, \tau) s(y; x, \theta) s^{\top}(y; x, \tau) dy$ is finite for $\tau \in T_0$.

Unbiasedness can be claimed by using the fact that $\sup_r |\hat{f}_n(r) - m^*(r)| \xrightarrow{P} 0$, as $n \rightarrow \infty$, at the true model $f(y) = q(y; x, \tau_0)$, for some $\tau_0 \in T$ (Markatou et al. 1998; Agostinelli and Greco 2013, 2019). Then, we have that

$$\sup_y |w(\delta(y))| \xrightarrow{P} 1.$$

Let τ_f be such that that $q(y; x, \tau_f)$ is *close* to the true distribution $f(y)$, that is, τ_f is implicitly defined by

$$\Psi^0(y; x, \tau) = \Psi^0 = \int_{\mathcal{Y}} w(y; \tau, F) s(y; x, \tau) dF(y) = \mathbf{0}.$$

Under the assumptions A(1)–A(7), we have that (Kuchibhotla and Basu 2015; Agostinelli and Greco 2019)

$$\begin{aligned} & - \sqrt{n}(\Psi_{\tau} - \Psi_{\tau}^0) \xrightarrow{d} N(0, V(\tau_f)) \\ & - \hat{\tau} \xrightarrow{\tau_f} \tau_f \\ & - \sqrt{n}(\hat{\tau} - \tau_f) \xrightarrow{d} N(0, B^{-1}(\tau_f) V(\tau_f) B^{-1}(\tau_f)), \end{aligned}$$

where B and V have been defined above. Under the true model, $B^{-1}(\tau_f) V(\tau_f) B^{-1}(\tau_f)$ reduces to the inverse of the expected Fisher Information, and the WLE is asymptotically first order efficient.

Algorithm 1 Single iteration from the WEM algorithm.

E step

for $k = 1, \dots, K$ **do**

$$u_{ik}^{(s)} = \frac{\pi_k^{(s-1)} m(y_i; \mu_k^{(s-1)}, \theta_k^{(s-1)})}{\sum_{k=1}^K \pi_k^{(s-1)} m(y_i; \mu_k^{(s-1)}, \theta_k^{(s-1)})}$$

end for

Weighting step

for $k = 1, \dots, K$ **do**

$$\delta_{ik}^{(s)} = \frac{f^*(r_{ik}^{(s)})}{m^*(r_{ik}^{(s)})} - 1$$

$$w_{ik}^{(s)} = \frac{[A(\delta_{ik}^{(s)}) + 1]^+}{\delta_{ik}^{(s)} + 1}$$

end for

M-step

for $k = 1, \dots, K$ **do**

$$\begin{aligned} \pi_k^{(s+1)} &= \frac{\sum_{i=1}^n u_{ik}^{(s)} w_{ik}^{(s)}}{\sum_{i=1}^n \sum_{k=1}^K u_{ik}^{(s)} w_{ik}^{(s)}} \\ \beta_k^{(s+1)} &= \left(X^\top W_k^{(s)} U_k^{(s)} \Omega_k^{(s)} X \right)^{-1} X^\top W_k^{(s)} U_k^{(s)} \Omega_k^{(s)} \left[\eta_k^{(s)} + g'(\mu_k^{(s)}) (y - \mu_k^{(s)}) \right] \\ \theta_k^{(s+1)} &= \frac{1}{n_k^{(s)} - p} \sum_{i=1}^n \frac{(y_i - \mu_{ik}^{(s+1)})^2}{V(\mu_{ik}^{(s+1)})} w_{ik}^{(s)} u_{ik}^{(s)} \end{aligned}$$

$$\text{with } n_k^{(s)} = \sum_{i=1}^n w_{ik}^{(s)} u_{ik}^{(s)}.$$

end for

3.2 Computational details

WEM and WCEM can be initialized by subsampling (Markatou et al. 1998; Neykov and Müller 2003; Neykov et al. 2007; Torti et al. 2019). A subsample of size $n^* \geq K(p + 1)$ 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. On the one hand, n^* should be as small as possible in order to increase the chance of drawing an outliers free subsample, but on the other hand a larger trial sample size can be necessary to find an initial solution.

In this paper, we consider a *deterministic* initialization. A starting outliers free partition of the data is obtained by running `tclust` on the $(x, g(y))$ data. Actually, we prefer to look for the initial partition on the link scale rather than on the response scale. This approach is the natural development of the strategy already successfully adopted by Greco et al. (2020) for mixtures of linear regressions. Then, initial parameters' values are obtained by fitting a GLM in each element of the partition. In an alternative fashion, one could evaluate a trial MLE by fitting a mixture model on the initial outlier free partition, for instance, by using the available functions from the R package `flexmix` (Grun and Leisch 2008).

In order to avoid the algorithm to be dependent on initial values and increase the chance to find a global rather than a local solution, a simple and common strategy is to consider several initializations. For instance, when using the deterministic initialization, different partitions can be obtained by suitably varying the trimming level or the restriction factor of `tclust`. Here, it is suggested to select the solution leading to the minimum fitted approximate disparity as defined in Agostinelli and Greco (2019), that is

$$\tilde{\rho}(f^*, m^*) = \frac{1}{n} \sum_{i=1}^n \frac{G(\hat{\delta}_{ik_i}) + \hat{\delta}_{ik_i}}{\hat{\delta}_{ik_i} + 1}, \quad (8)$$

where $\hat{\delta}_{ik_i} = \delta(r(y_i; x_i, \hat{\beta}_{k_i}, \hat{\theta}_{k_i}))$. Actually, the RAF function is connected to minimum disparity estimation problems. In principle, by following the approach developed in Markatou et al. (1998), it is possible to build a WLEE matching a minimum disparity objective driven by some function $G(\delta)$, where $G(\cdot)$ is a strictly convex function over $[-1, \infty)$ and thrice differentiable, such that $A(\delta) = (\delta + 1)G'(\delta) - G(0)$. For instance, the RAF corresponding to the Hellinger distance (HD) disparity is driven by $G(\delta) = 2 \left[\sqrt{\delta + 1} - 1 \right]^2$, that corresponding to the symmetric Chi-squared (SCH2) disparity is $G(\delta) = \delta^2 / (\delta + 2)$.

For what concerns the convergence criterion, the WEM (or WCEM) stops iterating when the maximum difference in absolute value among subsequent estimates is small enough, that is, below a certain value tol ,

$$\max |\hat{\beta}^{(s+1)} - \hat{\beta}^{(s)}| < \text{tol}.$$

3.3 WLE settings

Weighted likelihood estimation requires some choices to be made about the type of RAF, the kernel and its bandwidth. According to authors' experience, the shape of the kernel function or the selected RAF has a limited effect on weighted likelihood estimation. On the contrary, the choice of the smoothing parameter h is a crucial task. By varying h , one is allowed 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. The sum of the weights at convergence can be considered as a rough indicator of the actual level of contamination in the data. Guidelines for a safe choice of h are given in Agostinelli and Greco (2018): the authors recommend to monitor the WLE analyses as h varies by looking, for instance, at the trajectories corresponding to unit-specific residuals or weights. An abrupt change in such trajectories usually unveils a transition from a robust to a non-robust fit, hence leading to an adaptive selection of the smoothing parameter that corresponds to an appropriate compromise between robustness and efficiency.

4 Classification and outliers detection

For the purpose of classification, when using the WEM algorithm, cluster assignments can be pursued according to a maximum-a-posteriori (MAP) rule: units are assigned to the most likely component, at convergence. Conversely, the WCEM directly provides a classification of the units at convergence. In both cases, the classification step allows cluster assignments' for all data points, regardless of their degree of outlyingness. Some information about the agreement between the data and the assumed model can be driven from cluster-specific weights at convergence $\hat{w}_{ik_i} = w(\delta(\hat{r}_{ik_i}))$, $\hat{r}_{ik_i} = r(y_i; x_i, \hat{\beta}_{k_i}, \hat{\theta}_{k_i})$. However, outliers should not be considered for purely clustering tasks, that is, they are not supposed to be classified into any of the formed clusters but appropriately identified. Therefore, outliers' detection should be performed separately and based on formal rules driven by the robust fit.

The suggested outlier detection rule parallels the one outlined in Greco and Agostinelli (2020) in a multivariate framework and Greco et al. (2020) in a linear regression setting. For a fixed significance level α , conditionally on the final clusters assignment, 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, i.e.,

$$|r(y_i; x_i, \hat{\beta}_{k_i}, \hat{\theta}_{k_i})| > z_{1-\frac{\alpha}{2}}. \quad (9)$$

Popular choices are $\alpha = 0.05$ and $\alpha = 0.01$. The problem of testing for outliers is clearly affected by multiplicity issues because of the simultaneous testing of all the n data points. Here, we consider the approach presented by Cerioli and Farcomeni (2011), where the overall level of the test is controlled by means of the false discovery rate. This strategy leads to larger thresholds and avoids an intolerable number of false positives, that is genuine observation wrongly flagged as outliers (swamping effect). At the same time, the outlier detection rule is expected to guarantee a large power, that is to detect a large number of true outliers and suffer a tolerable rate of false negatives (masking effect).

5 Model selection

In the classical likelihood-based framework, a very general approach to select the number of clusters is to minimize over $k \in \{1, 2, \dots, k, \dots, K\}$ some complexity-penalized versions of the (negative) loglikelihood function, such as the well-known BIC, AIC and ICL. By following Greco et al. (2020), two alternative approaches aimed at dealing with the problem of the selection of K are suggested. One could resort to a weighted version of the BIC

$$BIC^w(k) = -2 \sum_{i=1}^n \hat{w}_{ik_i} \ell(y_i; \hat{\tau}) + m(k)$$

where the weights $\hat{w}_{ik_i} = w(\delta(r(y_i; x_i, \hat{\beta}_k)))$ are those stemming from the largest (in terms of model complexity) fitted model (Agostinelli 2002) or consider a disparity-based penalized criterion of the form

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

where the subscript now is meant to stress the dependency on the number of latent classes. As usual, the penalty term $m(k)$ reflects model complexity depending on the number of free parameters.

6 Illustrative examples

6.1 Example1: Poisson regression mixture

Let us consider a mixture of two Poisson regression models $Y_k \sim \text{Pois}(\mu_k)$, $k = 1, 2$, generated according to

$$\begin{cases} \log \mu_1 = 3 + x \\ \log \mu_2 = 3 - x \end{cases}$$

with $x \sim U(-2, 2)$. Each group consists of a hundred points. Then, 50 outliers were added that are uniformly distributed in the rectangle that contains the genuine data.

Outliers are such that their single-term deviance values evaluated with respect to the true models are above the 0.99-level quantile of the χ_1^2 distribution. We set $h = 0.05$ and used a RAF based on the HD disparity. The data, the fitted model by WEM and the final classification are displayed in Fig. 1. The weights are based on the deviance residuals. The outlier detection rule has significance level $\alpha = 0.01$ controlled by the FDR. The weighted likelihood methodology provides quite satisfactory outcomes both in terms of fitting and in terms of classification accuracy. The WCEM gives very similar results and is not reported here. Actually, the fitted mixture model (solid line in the right panel) is *close* to the true model (solid line in the left panel) and able to recover the underlying structure of the data: genuine observations have been accurately classified and well separated from the outliers. On the contrary, maximum likelihood leads to a completely unreliable fit (dashed lines in the right panel) that completely miss the underlying structure of the sample at hand.

6.2 Example 2: Gamma regression mixture

Let us consider a mixture of two Gamma regression models $Y_k \sim \text{Gamma}(\mu_k, \theta_k)$, $k = 1, 2$, generated according to

$$\begin{cases} \mu_1 = \frac{x}{3+x/16} = \left(\frac{3}{x} + \frac{1}{16}\right)^{-1} \\ \mu_2 = \frac{x}{0.5+x/2} = \left(\frac{1}{2x} + \frac{1}{2}\right)^{-1} \end{cases}$$

with $x \sim U(0, 50)$ and $(\theta_1, \theta_2) = (1/50, 1/30)$. The assumed model for each component is the Michaelis–Menten model that is particularly useful in those situations where the mean response is bounded. Each group consists of 500 points. Then, 100

Fig. 1 Example 1. True assignments (left), fitted model by WEM (right). Clusters stemming from WEM are denoted by different colors and symbols. The empty black circles denote true (left) or detected (right) outliers. The solid lines give the true (left) and fitted by WEM (right) model. The dashed lines in the left panel correspond to maximum likelihood

outliers were added that are clustered themselves to form two separate groups of bad leverage points of equal size. We set $h = 0.02$ and used a RAF based on the SCHI2 disparity. The data, the fitted model by WEM and classical EM and the final classification stemming from WEM and an outlier detection rule at an overall level $\alpha = 0.01$ controlled by the FDR are displayed in Fig. 2. The WEM (and the WCEM as well) is effective in recovering the underlying model and successfully leads to flag outliers. The fitted components (solid lines in the right panel) resemble the true model displayed in the left panel. An accurate classification is provided for the genuine data, and outliers are successfully detected. On the contrary, maximum likelihood estimation (dashed lines in the right panel) is badly affected from the occurrence of outliers indeed, turning in fitted components that do not represent any of the true underlying data structures.

7 Numerical studies

The finite sample behavior of the proposed WEM and WCEM has been investigated by some numerical studies. As well as in the illustrative examples, we focus on mixtures of Poisson and Gamma regressions. The numerical studies are based on 500 Monte Carlo trials; the WEM and WCEM are assumed to reach convergence for a tolerance $\text{tol} = 10^{-5}$. We used both deviance (WEM1 and WCEM1) and Anscombe (WEM2 and WCEM2) residuals. The algorithms run on non-optimized R code, but computational time always lies in a feasible range. For what concerns the contamination rates, we set $\epsilon = 0\%, 10\%, 20\%$. The scenario without contamination in the data allows to investigate the efficiency loss of the proposed robust procedures WEM and WCEM with respect to maximum likelihood estimation at the assumed

Fig. 2 Example 2. True assignments (left), fitted model (right) by WEM (solid line) and EM (dotted line). Clusters stemming from WEM are denoted by different colors and symbols. The empty black circles denote true (left) or detected (right) outliers. The solid lines give the true (left) and fitted by WEM (right) model. The dashed lines in the left panel correspond to maximum likelihood

model. Maximum likelihood estimation has been performed by using the classical EM algorithm. In the case of Poisson data, the behavior of the weighted likelihood techniques under contamination has been also compared with the trimmed likelihood methodology, which has been developed both under an EM (TEM) and a CEM algorithm (TCEM). The trimmed likelihood did not work fine in the Gamma case, and results have not been shown. The trimmed EM algorithms have been also implemented by non-optimized R-code and are based on the same initialization strategy used for both WEM and WCEM. Here, we consider *oracle* trimmed algorithms with trimming level equal to the contamination rate ϵ . The numerical studies have been carried out for $p = 2$ and $p = 5$. In the latter case, uninformative explanatory variables have been added, that is whose corresponding coefficients are set to zero.

Fitting accuracy has been measured by storing in each replication the sum of squared errors (SSEs) for the mixture parameters (π, β, θ) . Classification accuracy has been evaluated by the adjusted Rand index (ARI) computed over true negatives, i.e., genuine observations that are not wrongly declared outliers. The empirical distribution for the MLE has been displayed only in the non-contaminated case, since the observed bias is very large in the presence of outliers, whereas the comparison between WEM and WCEM with the oracle trimmed likelihood algorithm takes place only under contamination.

Outlier detection has been performed at an overall $\alpha = 0.01$ significance level based on (9) and the FDR, in order to take into account multiplicity issues. The behavior of the testing strategy has been summarized by the swamping rate and the power of the test. It is worth noticing that any outliers detection rule is not available and currently used after TEM or TCCEM, but outliers are commonly identified with trimmed observations. For this reason, the comparison with WEM and WCEM in terms of swamping and power is not properly fair (Greco and Agostinelli 2020; Greco et al. 2020).

7.1 Poisson regression mixture

Let us assume a mixture of $K = 2$ Poisson regression models, according to the scheme described in Example 1. Figure 3 displays the violin plots corresponding to the distribution of the SSE for the regression coefficients from MLE, WEM, WCEM, TEM and TCCEM. The weighted algorithms are based on a HD RAF. Under the true model, the weighted algorithms show a close agreement with maximum likelihood. In a similar fashion, we do not observe remarkable differences with the results stemming from the trimmed likelihood approach with oracle trimming level under contamination. These features confirm the satisfactory accuracy of both WEM and WCEM. Furthermore, the choice of the type of residuals, deviance or Anscombe, does not seem to be relevant. The results concerning classification accuracy are given in Fig. 4. The WEM and WCEM still exhibit a satisfactory behavior well comparable with that stemming from maximum likelihood or the oracle trimming strategies. For what concerns the task of outliers detection, the entries in Table 1 give mean swamping and power of the robust procedures in the presence of outliers. By keeping in mind the unfairness of the comparison, as explained above, the

weighted methodologies exhibit null swamping for the chosen outliers configuration, whereas the oracle trimmed procedure shows slight larger power. These findings confirm the satisfactory behavior of WEM and WCEM for the task of outlier detection.

7.2 Gamma regression mixture

Let us assume a mixture of $K = 2$ Gamma regression models. We assume a logarithmic link function and expected values as those used in the Poisson numerical studies. As well, we consider scattered outliers whose distance from the true model is larger than the 0.99-level quantile of the χ_1^2 distribution. The WEM and WCEM are based on a SCH2 RAF function. The behavior of WEM and WCEM is displayed in Figs. 5 and 6 for what concerns fitting accuracy, and in Fig. 7 for the task of classification. Both fitting and classification accuracy of WEM and WCEM are satisfactory in all considered scenarios. The capability to provide an accurate estimate of the dispersion parameter is particularly remarkable. The entries in Table 2 give swamping and power. The proposed strategy is able to correctly identify outliers with a reasonable large power. As well in the Poisson case, we do not observe any difference between the case with $p = 2$ and $p = 5$.

8 Vitamin D level and Covid-19 mortality

Low-serum concentration of 25-hydroxy vitamin D is expected to be related to susceptibility to acute respiratory tract infections. In Ilie et al. (2020), the authors look for evidence supporting the hypothesis that vitamin D may play a protective role for COVID-19, since vitamin D has been claimed to support the immune system in reacting to an infection. Actually, they investigated the crude association between the mean levels of vitamin D in twenty European countries and the number of deaths (over 100 thousand inhabitants) caused by COVID-19 by April 8th. By fitting a regression line, they found a not convincingly significant association (with a p value 0.054) and a very unsatisfactory goodness of fit as measured by $R^2 = 0.192$ (see Maruotti et al. 2020 for more comments). These findings also depend on the fact that the data suggest the presence of (at least) two subgroups, and their fitted straight line does not pass through any of the groups. Here, we stress the need for a mixture model and consider Poisson regression for count data, in a more appropriate fashion, rather than least squares. In particular, an overdispersed Poisson regression model has been considered, such that $\text{Var}(Y_{ik}) = \theta_k \mu_{ik}$, $\theta_k > 0$, to take into account possible other sources of uncertainty. Figure 8 displays the fitted mixture model by WEM and EM, with cluster assignments and outliers stemming from WEM. Portugal has been identified as a moderate outlier (with a p value 0.014). The negative crude association between vitamin D level and deaths due to Covid-19 is confirmed but with a different strength in the two groups. The fitted slopes correspond to an expected reduction in the number of deaths for a unit increase in vitamin D level equal to 0.12 in cluster1 and to 0.05 in cluster 2. A 95% level confidence interval for the slope in

Fig. 3 Mixture of Poisson regressions: SSE for the regression coefficients β for MLE, WEM, WCEM, TEM, TCEM, when $p = 2$ (left), $p = 5$ (right), no contamination (top), $\epsilon = 10\%$ (middle), $\epsilon = 20\%$ (bottom)

Fig. 4 Mixture of Poisson regressions: ARI stemming from MLE, WEM, WCEM, TEM, TCEM, when $p = 2$ (left), $p = 5$ (right), no contamination (top), $\epsilon = 10\%$ (middle), $\epsilon = 20\%$ (bottom)

Table 1 Mixture of Poisson regressions: swamping and power for WEM, WCEM, TEM, TCEM, when $p = 2, 5$ and $\epsilon = 10\%, 20\%$

	$\epsilon = 10\%$		$\epsilon = 20\%$	
	Swamping	Power	Swamping	Power
$p = 2$				
WEM1	0.00	0.90	0.00	0.93
WEM2	0.00	0.90	0.00	0.93
WCEM1	0.00	0.90	0.00	0.93
WCEM2	0.00	0.90	0.00	0.93
TEM	0.02	0.97	0.02	0.98
TCEM	0.02	0.97	0.02	0.98
$p = 5$				
WEM1	0.00	0.90	0.00	0.91
WEM2	0.00	0.90	0.00	0.91
WCEM1	0.00	0.90	0.00	0.91
WCEM2	0.00	0.90	0.00	0.91
TEM	0.04	0.96	0.03	0.96
TCEM	0.04	0.96	0.03	0.96

Table 2 Mixture of Gamma regressions: swamping and power for WEM, WCEM, when $p = 2, 5$ and $\epsilon = 10\%, 20\%$

	$\epsilon = 10\%$		$\epsilon = 20\%$	
	swamping	power	swamping	power
$p = 2$				
WEM1	0.00	0.88	0.00	0.88
WEM2	0.00	0.89	0.00	0.89
WCEM1	0.00	0.87	0.00	0.87
WCEM2	0.00	0.88	0.00	0.88
$p = 5$				
WEM1	0.00	0.82	0.00	0.82
WEM2	0.00	0.86	0.00	0.86
WCEM1	0.00	0.80	0.00	0.80
WCEM2	0.00	0.84	0.00	0.84

cluster 1 is $(-0.080, -0.032)$ and $(-0.202, -0.055)$ in cluster 2. Here, inferences should be considered conditionally on the final cluster assignments $\hat{u}_{ik_i}$ and based on the fixed weights $\hat{w}_{ik_i}$, i.e., by paralleling the classical approach, the variance-covariance matrix of $\hat{\beta}_k$ is estimated by $\hat{\theta}_k(X^\top \hat{W}_k \hat{U}_k \hat{\Omega}_k X)^{-1}$, with $\hat{U}_k = \text{diag}(u_{ik_i})$.

The detected outlier has the clear effect of attracting the second regression component fitted by maximum likelihood and the EM algorithm. An effective diagnostic tool that clearly spots the outlier is given in Fig. 9 in the form of a normal quantile-quantile plot evaluated over component-wise robust residuals: Portugal corresponds to the point in the bottom left corner. The outlying nature of Portugal could be explained since it exhibits the lowest Vitamin D level across the considered countries

Fig. 5 Mixture of Gamma regressions: SSE for the regression coefficients β for MLE, WEM, WCEM, when $p = 2$ (left), $p = 5$ (right), no contamination (top), $\epsilon = 10\%$ (middle), $\epsilon = 20\%$ (bottom)

Fig. 6 Mixture of Gamma regressions: SSE for the dispersion parameters θ for MLE, WEM, WCEM, when $p = 2$ (left), $p = 5$ (right), no contamination (top), $\epsilon = 10\%$ (middle), $\epsilon = 20\%$ (bottom)

Fig. 7 Mixture of Gamma regressions: ARI stemming from MLE, WEM, WCEM, when $p = 2$ (left), $p = 5$ (right), no contamination (top), $\epsilon = 10\%$ (middle), $\epsilon = 20\%$ (bottom)

Fig. 8 Vitamin D level and Covid-19 mortality: fitted WEM and EM with cluster assignments and outliers stemming from WEM

Fig. 9 Vitamin D level and Covid-19 mortality: normal quantile–quantile plot of WEM residuals

associated with an unexpected low number of registered deaths. The same analysis has been run with more recent mortality data collected at <https://www.worldometers.info/coronavirus/> on May 21st. The new analysis did not unveil Portugal as an outlier anymore, but now it is properly classified in cluster 2. The composition of clusters is not altered. The main change concerns the fitted slope in cluster 1 that increases from -0.128 to -0.023 (with a 0.95%-level confidence interval $-0.044, -0.001$)).

9 Conclusions

In this paper, a robust method to fit a mixture of GLMs has been presented and discussed. The technique is based on weighted likelihood estimation. It mainly relies on the proposal from Alqallaf and Agostinelli (2016) and the theoretical results in

Agostinelli and Greco (2019) and extends and generalizes the procedure in Greco et al. (2020) from mixtures of linear regressions to the more general and challenging framework of GLMs. Actually, the methodology allows to model data clustered around nonlinear structures, defined through a specified link function, satisfactory and provides both accurate estimates and model-based classification. Moreover, a general and effective strategy to detect outliers can be applied stemming from the robust fit. The nice behavior of the method has been investigated through synthetic examples, some numerical studies and a real data example.

Acknowledgements The author wish to thank the associate editor and two anonymous referees for their comments and suggestions that helped to improve the paper.

References

- Agostinelli, C.: Robust model selection in regression via weighted likelihood methodology. *Stat. Prob. Lett.* **56**(3), 289–300 (2002)
- Agostinelli, C., Greco, L.: A weighted strategy to handle likelihood uncertainty in Bayesian inference. *Comput. Stat.* **28**(1), 319–339 (2013)
- Agostinelli, C., Greco, L.: Discussion on: the power of monitoring: how to make the most of a contaminated sample. *Stat. Methods Appl.* **27**(4), 609–619 (2018)
- Agostinelli, C., Greco, L.: Weighted likelihood estimation of multivariate location and scatter. *Test* **28**(3), 756–784 (2019)
- Agostinelli, C., Markatou, M.: Test of hypotheses based on the weighted likelihood methodology. *Stat. Sin.* **11**, 499–514 (2001)
- Alqallaf, F., Agostinelli, C.: Robust inference in generalized linear models. *Commun. Stat. Simul. Comput.* **45**(9), 3053–3073 (2016)
- Bai, X., Yao, W., Boyer, J.E.: Robust fitting of mixture regression models. *Comput. Stat. Data Anal.* **56**(7), 2347–2359 (2012)
- Bashir, S., Carter, E.: Robust mixture of linear regression models. *Commun. Stat. Theory Methods* **41**(18), 3371–3388 (2012)
- Basu, A., Lindsay, B.: Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Ann. Inst. Stat. Math.* **46**(4), 683–705 (1994)
- Celeux, G., Govaert, G.: Comparison of the mixture and the classification maximum likelihood in cluster analysis. *J. Stat. Comput. Simul.* **47**(3–4), 127–146 (1993)
- Cerioni, A., Farcomeni, A.: Error rates for multivariate outlier detection. *Comput. Stat. Data Anal.* **55**(1), 544–553 (2011)
- Dempster, A., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **39**, 1–38 (1977)
- Dotto, F., Farcomeni, A.: Robust inference for parsimonious model-based clustering. *J. Stat. Comput. Simul.* **89**(3), 414–442 (2019)
- Dotto, F., Farcomeni, A., García-Escudero, L., Mayo-Iscar, A.: A reweighting approach to robust clustering. *Stat. Comput.* **28**(2), 477–493 (2018)
- Elashoff, M., Ryan, L.: An EM algorithm for estimating equations. *J. Comput. Graph. Stat.* **13**(1), 48–65 (2004)
- Farcomeni, A., Greco, L.: *Robust Methods for Data Reduction*. CRC Press, Boca Raton (2015a)
- Farcomeni, A., Greco, L.: S-estimation of hidden Markov models. *Comput. Stat.* **30**(1), 57–80 (2015b)
- Fritz, H., García-Escudero, L., Mayo-Iscar, A.: A fast algorithm for robust constrained clustering. *Comput. Stat. Data Anal.* **61**, 124–136 (2013)
- García-Escudero, L.A., Gordaliza, A., Matrán, C., Mayo-Iscar, A.: A general trimming approach to robust cluster analysis. *Ann. Stat.* **36**(3), 1324–1345 (2008)
- García-Escudero, L., Gordaliza, A., San Martín, R., Van Aelst, S., Zamar, R.: Robust linear clustering. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **71**(1), 301–318 (2009)

- García-Escudero, L.A., Gordaliza, A., Mayo-Isar, A., San Martín, R.: Robust clusterwise linear regression through trimming. *Comput. Stat. Data Anal.* **54**(12), 3057–3069 (2010)
- Greco, L., Agostinelli, C.: Weighted likelihood mixture modeling and model-based clustering. *Stat. Comput.* **30**(2), 255–277 (2020)
- Greco, L., Lucadamo, A., Agostinelli, C.: Weighted likelihood latent class linear regression. *Stat. Methods Appl.* (2020). <https://doi.org/10.1007/s10260-020-00540-8>
- Grün, B., Leisch, F.: Fitting finite mixtures of generalized linear regressions in R. *Comput. Stat. Data Anal.* **51**(11), 5247–5252 (2007)
- Grun, B., Leisch, F.: Flexmix version 2: finite mixtures with concomitant variables and varying and constant parameters. *J. Stat. Softw.* **28**(4), 1–35 (2008)
- Ilie, P.C., Stefanescu, S., Smith, L.: The role of vitamin D in the prevention of coronavirus disease 2019 infection and mortality. *Aging Clin. Exp. Res.* (2020). <https://doi.org/10.1007/s40520-020-01570-8>
- Karlis, D., Xekalaki, E.: Robust inference for finite Poisson mixtures. *J. Stat. Plan. Inference* **93**(1–2), 93–115 (2001)
- Kuchibhotla, A., Basu, A.: A general set up for minimum disparity estimation. *Stat. Prob. Lett.* **96**, 68–74 (2015)
- Lu, Z., Hui, Y., Lee, A.: Minimum hellinger distance estimation for finite mixtures of Poisson regression models and its applications. *Biometrics* **59**(4), 1016–1026 (2003)
- Markatou, M.: Mixture models, robustness, and the weighted likelihood methodology. *Biometrics* **56**(2), 483–486 (2000)
- Markatou, M., Basu, A., Lindsay, B.G.: Weighted likelihood equations with bootstrap root search. *J. Am. Stat. Assoc.* **93**(442), 740–750 (1998)
- Maronna, R., Martin, R.D., Yohai, V., Salibián-Barrera, M.: *Robust Statistics: Theory and Methods* (with R). Wiley, Hoboken (2019)
- Maruotti, A., Belloc, F., Nicita, A.: Comments on: The role of vitamin d in the prevention of coronavirus disease 2019 infection and mortality. *Aging Clin. Exp. Res.* **32**(8), 1621–1623 (2020)
- McLachlan, G., Peel, D.: *Finite Mixture Models*. Wiley, Hoboken (2004)
- Neykov, N.M., Müller, C.H.: Breakdown point and computation of trimmed likelihood estimators in generalized linear models. In: *Developments in Robust Statistics*, pp 277–286. Springer (2003)
- Neykov, N., Filzmoser, P., Dimova, R., Neytchev, P.: Robust fitting of mixtures using the trimmed likelihood estimator. *Comput. Stat. Data Anal.* **52**(1), 299–308 (2007)
- Park, C., Basu, A., Lindsay, B.: The residual adjustment function and weighted likelihood: a graphical interpretation of robustness of minimum disparity estimators. *Comput. Stat. Data Anal.* **39**(1), 21–33 (2002)
- Torti, F., Perrotta, D., Riani, M., Cerioli, A.: Assessing trimming methodologies for clustering linear regression data. *Adv. Data Anal. Classif.* **13**(1), 227–257 (2019)

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