Noname manuscript No.
(will be inserted by the editor)

Weighted likelihood estimation of multivariate location and scatter

Claudio Agostinelli · Luca Greco

July 3, 2018

Abstract A novel approach to obtain weighted likelihood estimates of multivariate location and scatter is discussed. A weighting scheme is proposed that is based on the univariate distribution of the Mahalanobis distances rather than the multivariate distribution of the data at the assumed model. This strategy allows to avoid the curse of dimensionality affecting multivariate non-parametric density estimation, that is involved in the construction of the weights through the Pearson residuals. Asymptotic properties of the proposed weighted likelihood estimator are also discussed. Then, weighted likelihood based outlier detection rules and robust dimensionality reduction techniques are developed. The effectiveness of the methodology is illustrated through some numerical studies and real data examples.

Keywords Dimensionality Reduction · Discriminant Analysis · Mahalanobis distance · Multivariate Normal · Outlier detection · Pearson residuals · PCA · Robustness · Weighted Likelihood

Mathematics Subject Classification (2000) MSC 62F35 · MSC 62G35 · MSC 62H25 · MSC 62H30

1 Introduction

Several multivariate techniques are based on the assumption of multivariate normality and the use of the sample mean vector and covariance matrix. Actually, they lead to a simple description of the overall shape of the multivariate

C. Agostinelli
Department of Mathematics,University of Trento, Italy
E-mail: claudio.agostinelli@unitn.it

L. Greco
DEMM Department, University of Sannio, Italy
E-mail: luca.greco@unisannio.it

data at hand through the related ellipsoid (Huber and Ronchetti, 2009). It is well known that small departures from normality may invalidate multivariate estimation of location and scatter and have dramatic effects on all those techniques based on them, such as Principal Component Analysis or Discriminant Analysis, for instance (Maronna et al, 2006; Huber and Ronchetti, 2009). Such departures result in data inadequacies that are typically observed in the form of several outliers. Outliers are observations that exhibit patterns not shared by the remaining genuine part of the data. They can be defined as observations that are highly unlikely to occur under the assumed model (Markatou et al, 1998). In other words, outliers *contaminate* the data with respect to (w.r.t.) the postulated model. Outliers may be generated by an unexpected and unknown overlapping random mechanism but can also occur because of a wrong specification, when, for instance, the true underlying model is characterized by longer tails and/or an asymmetric shape. However, we should keep in mind that any model is only an approximation to reality.

On the contrary, by supplying robust estimates of multivariate location and scatter, one could rely on techniques that are resistant to contamination (Hubert et al, 2008). Furthermore, the appropriate use of robust estimators may also lead to detect outliers, find unexpected structures in the data and explore the types of occurred departures. There is a growing literature on robust multivariate estimation. The reader is pointed to the book by Farcomeni and Greco (2016) for a recent account on robustness in a multivariate setting.

Robust estimates of multivariate location and covariance are obtained by attaching a weight to each data point in order to bound the effect of possible outliers on the resulting fit. Weights are determined according to an outlyingness measure, that is a measure of the distance of the multivariate data point from the robust fit. In summary, we can consider three main classes of estimators.

1. Estimators based on hard trimming: weights are 0-1 and outliers are trimmed. The final estimate is based on a subset of the original data points, whose size is tuned by the user. The Minimum Covariance Determinant (MCD) is undoubtedly one of the most popular techniques (Rousseeuw, 1985; Croux and Haesbroeck, 1999).
2. Estimators based on adaptive hard trimming: outliers are trimmed but the final sample is determined adaptively by the data. The main tool is represented by the Forward Search (FS, see Riani et al, 2009; Atkinson and Riani, 2012, for a recent account).
3. Estimators based on soft trimming: outliers are down-weighted, with weights varying in $[0, 1]$, and the final estimate consists of a weighted mean and weighted covariance matrix. This feature characterizes M-estimators and related methods such as S-estimators (Lopuhaa, 1989) and MM-estimators (Salibian-Barrera et al, 2006), but also the weighted likelihood estimator (WLE, Markatou et al, 1998; Kuchibhotla and Basu, 2015, 2018a). In this class we also include those methods stemming from projection of multi-

variate data onto univariate directions, as the Stahel-Donoho estimator, for instance.

The weighting strategy that characterizes the MCD, the FS and M-estimation is based on the inspection of the Mahalanobis distances

$$d(y;\mu,\Sigma) = \sqrt{(y-\mu)^\top \Sigma^{-1}(y-\mu)}\ , \tag{1}$$

where y denotes a p-variate observation, $p > 1$, sampled from a multivariate normal model, $Y \sim N_p(\mu, \Sigma)$ with mean vector $\mu = (\mu_1, \mu_2, \ldots, \mu_p)^\top$ and $p \times p$ covariance matrix Σ. Let $(\hat{\mu}, \hat{\Sigma})$ be a robust estimate of location and scatter, then data points are discarded or down-weighted according to their distance $d(y; \hat{\mu}, \hat{\Sigma})$ from the robust fit: the larger the robust distance the closer to zero the weight and more likely the point will be treated as an outlier.

In a different fashion, the computation of the WLE is not based on such robust distances, but outlyingness is measured according to the agreement between the data, summarized by a non parametric density estimate, and the assumed multivariate normal model. Actually, the weighting scheme based on the computation of a multivariate density estimate becomes troublesome for large dimensions, because of the curse of dimensionality (Huber, 1985; Scott and Wand, 1991). With growing dimensions the data are more sparse and kernel density estimation may become unfeasible. The reader is pointed to Deng and Wickham (2011) for a comparison of several density estimation methods available from the statistical environment `R` (R Core Team, 2018). This feature represents a serious limitation of the weighted likelihood methodology in a multivariate framework. Such a restriction is much more annoying since all the other multivariate estimators that we have mentioned so far are well behaved in large dimensions. It is worth to stress here that we only consider the case where the sample size n is larger than the dimension p.

In this paper, a novel approach to overcome this hindrance is presented. We introduce a weighting algorithm that is still based on non parametric density estimation, but now a univariate density estimate is evaluated over (robust) distances. We obtain multivariate estimators of location and covariance that are consistent and expected to be highly efficient at the assumed multivariate normal model, as there is negligible downweighting. On the contrary, the proposed multivariate WLE is robust w.r.t. the presence of outliers, since the weights will penalize the contribution of those data points exhibiting large distances from the fitted model. A similar goal to build efficient and robust estimators has been pursued by Gervini and Yohai (2002) in a regression framework, for instance.

In the following, the emphasis is on those situations in which a multivariate normal sample is contaminated by some anomalous values following a different random mechanism, that is the data come from the ϵ-contaminated model

$$G(y) = (1-\epsilon)N(y;\mu,\Sigma) + \epsilon H(y), \tag{2}$$

where $H(y)$ is an arbitrary distribution generating outliers and $\epsilon < 0.50$. Nevertheless, we also expect that the proposed robust method works satisfactory

under some type of misspecifications. Let

$$\mathcal{Q} = \left\{ q(y;\mu,\Sigma) = \mathrm{cost}(p)|\Sigma|^{-\frac{1}{2}} b[d(y;\mu,\Sigma)], \mu \in \mathbb{R}^p, \Sigma \in PDS(p), p > 1 \right\}$$

be an elliptically symmetric family of distributions where $PDS(p)$ is the set of all positive-definite symmetric $p \times p$ matrices and $\mathrm{cost}(p)$ is a normalization constant depending on p. Misspecification issues may arise since the shape of $b(\cdot)$ is only approximately known (Maronna, 1976). The multivariate normal family corresponds to $b(d) = \exp(-d^2/2)$. Another example is represented by the multivariate Student t_ν distribution, with $b(d) = (1 + d^2/\nu)^{-(p+\nu)/2}$.

Nevertheless, the proposed approach is meant to be much more general. Therefore, hereafter $G(y)$ will denote the *true* underlying model and $Q(y;\theta)$ the *specified* parametric model. The rest of the paper is structured as follows. Some background on the weighted likelihood methodology is given in Section 2. The new weighting algorithm is introduced in Section 3 and then detailed in the case of the multivariate normal model in Section 3.1. Asymptotic properties are discussed in Section 4. Outlier detection rules are illustrated in Section 5. Some numerical studies are given in Section 6 and real data examples concerning estimation, outlier detection, principal component analysis and discriminant analysis are presented in Section 7.

2 Background

Beran (1977) studied the minimum Hellinger distance estimation in case of continuous parametric models with asymptotic first order efficiency and very interesting robust properties. Since then, minimum disparity estimation has become a popular and attractive technique from both the robustness and efficiency perspectives. Lindsay (1994) established first order efficiency under fairly general conditions on the class of disparities for discrete models. Park and Basu (2004) discussed a general framework for continuous models, even if their approach is somewhat restrictive since it excluded some common disparities such as the Pearson's chi-square, the Hellinger distance and the likelihood disparity. A review has been provided in Basu et al (2011), in which several applications concerning different statistical problems were illustrated.

Minimum disparity estimators are obtained by minimizing the corresponding disparity function but can be also defined as the root of an estimating equation defined as the first derivative of the disparity function. In most cases, the estimating function is obtained as the product of the usual score function with another function that can be thought as a weighting factor. The estimation procedure requires an integral evaluation over the whole support of the data in both approaches. In a different fashion, in Markatou et al (1998) integrals were naturally replaced by summations, leading to the weighted likelihood estimating equations (WLEE), whose solution can be found by one readily available iterative reweighting algorithm. However, the corresponding estimator does not minimize any proper objective function. Then, in summary, minimum disparity estimation describes a minimization problem whereas weighted

likelihood estimation equations is a root solving problem. These two methods have been dealt with separately in the literature until the work by Kuchibhotla and Basu (2015) and Kuchibhotla and Basu (2018a,b), who provided a unified approach meant to reconcile them. In the papers mentioned above, it is shown that, under very standard conditions, one can build a simple WLEE matching a minimum disparity objective function. In the following, we briefly review this approach.

Let $\mathbf{y} = (y_1, \cdots, y_n)$ be a random sample from a p-random vector Y with unknown distribution function G and corresponding density function g. The assumed model for Y is

$$\mathcal{Q} = \{q(y;\theta); \theta \in \Theta \subset \mathbb{R}^p, p \geq 1\}$$

where $q(y;\theta)$ is a probability density function and $Q = Q(y;\theta)$ is the corresponding distribution function. Furthermore, it is assumed that the support of G is the same as that of Q and independent of θ. Let $\theta_g \in \Theta$ be such that g is *close* to $q(y;\theta_g)$ in some appropriate sense. Let $\hat{G}_n$ denote the empirical distribution function. The Pearson residual function $\delta(y)$ (Lindsay, 1994; Markatou et al, 1998) is defined by comparing the true density by the model density as

$$\delta(y) = \delta(y;\theta,G) = \frac{g(y)}{q(y;\theta)} - 1 \ .$$

It is worth to notice that $\int \delta(y)\ dQ(y;\theta) = 0$ for all θ and G. Let C be a thrice differentiable convex function defined on $[-1,\infty)$, satisfying $C(0) = 0$, and $k \in \mathbb{R}$. Consider the class of disparities defined by

$$\begin{aligned}\rho_C(g, q_\theta) &= \int C\left(\delta(y)\right)\ dQ(y;\theta) \\ &= \int C\left(\delta(y)\right) + k\delta(y)\ dQ(y;\theta) \\ &= \int \frac{C\left(\delta(y)\right) + k\delta(y)}{\delta(y)+1}\ dG(y)\ ,\end{aligned}$$

where q_θ stands for $q(y;\theta)$, and let $\theta_g = \arg\min_{\theta\in\Theta} \rho_C(g, q_\theta)$ be the best fitting parameter according to the disparity measure ρ_C. Let $\{A_n\}_{n=1}^{\infty}$ be a sequence such that $A_n \uparrow \mathbb{R}^p$ as $n \to \infty$, $\mathbb{1}_{A_n}$ be the indicator function of the set A_n, and consider the following approximation of $\rho_C(g, q_\theta)$

$$\tilde{\rho}_C(g, q_\theta) = \int \mathbb{1}_{A_n}(y) \frac{C\left(\delta(y)\right) + k\delta(y)}{\delta(y)+1}\ dG(y)\ .$$

A natural estimate of $\tilde{\rho}_{C,n}(g, q_\theta)$ is given by

$$\begin{aligned}\tilde{\rho}_{C,n}(\hat{g}_n, q_\theta) &= \int \mathbb{1}_{A_n}(y) \frac{C\left(\delta_n(y)\right) + k\delta_n(y)}{\delta_n(y)+1}\ d\hat{G}_n(y) \\ &= \frac{1}{n}\sum_{i=1}^{n} \kappa_{n,i} \frac{C\left(\delta_n(y_i)\right) + k\delta_n(y_i)}{\delta_n(y_i)+1}\ , \qquad (3)\end{aligned}$$

where, $\kappa_{n,i} = \mathbb{1}_{A_n}(y_i)$ and

$$\delta_n(y) = \delta(y;\theta,\hat{G}_n) = \frac{\hat{g}_n(y)}{q(y;\theta)} - 1 \ ,$$

is the finite sample Pearson residual, which compares $\hat{g}_n(y)$, that is an estimate of $g(y)$, with the model density $q(y;\theta)$. In discrete families of distributions, $\hat{g}_n(y)$ can be driven by the observed relative frequencies (Lindsay, 1994). In continuous models, a non parametric density estimate

$$\hat{g}_n(y) = \int k(y;t,h)\ d\hat{G}_n(t)$$

based on the kernel $k(\cdot;t,h)$ with bandwidth h is often used (Basu and Lindsay, 1994; Markatou et al, 1998). Furthermore, in this situation, the model density is replaced by a smoothed model density $q^*(y;\theta)$

$$q^*(y;\theta) = \int k(y;t,h)\ dQ(t;\theta) \ .$$

The key idea is that by smoothing the model, the convergence of $\hat{g}_n(y)$ towards $q^*(y;\theta)$ does not require the bandwidth h to go to zero, as n increases. In the proposed approach, we will not use the smoothed model and we consider $A_n = \{y : \hat{g}_n(y) > \gamma_n/2\}$ for some $\gamma_n \downarrow 0$ as $n \to \infty$ at some rate as specified in Kuchibhotla and Basu (2018a). The estimating equation stemming from the objective function (3) is given by

$$\frac{1}{n}\sum_{i=1}^{n} \kappa_{n,i} \frac{A\left(\delta_n(y_i)\right) + k}{\delta_n(y_i) + 1}\ s(y_i;\theta) = 0 \ , \tag{4}$$

where $s(y_i;\theta) = \nabla \log(q(y_i;\theta))$ denotes the i-th contribution to the score function. The trimming sequence $\kappa_{n,i}$ is meant to avoid numerical instabilities due the occurrence of small (almost null) densities in the denominator for y in the tails. As stated in Kuchibhotla and Basu (2018a), trimming is not necessary and could not be considered, especially in those models where the tails decay exponentially.

Hereafter we set $k = 1$ so that we can formulate (4), apart from the trimming function $\kappa_{n,i}$ (whenever introduced) and the smoothed model, as the WLEE defined in Markatou et al (1998)

$$\sum_{i=1}^{n} w(y_i) s(y_i;\theta) = 0 \ , \tag{5}$$

with

$$w(y) = w(\delta_n(y)) = \kappa_{n,i}\ \frac{A(\delta_n(y)) + 1}{\delta_n(y) + 1} \ , \tag{6}$$

where $A(\cdot)$ is the Residual Adjustment Function (RAF, Lindsay, 1994; Basu and Lindsay, 1994; Markatou et al, 1998; Park et al, 2002). The RAF plays

the role to bound the effect of large Pearson residuals on the fitting procedure, as well as the Huber and Tukey-bisquare function bound large distances in M-type estimation. By using a RAF such that $|A(\delta)| \leq |\delta|$ both outliers and inliers will be downweighted. Here, we consider the families of RAF based on the Power Divergence Measure

$$A_{pdm}(\delta, \tau) = \begin{cases} \tau\left((\delta+1)^{1/\tau} - 1\right) & \tau < \infty \\ \log(\delta+1) & \tau \to \infty \ . \end{cases}$$

Special cases are maximum likelihood ($\tau = 1$, as the weights become all equal to one), Hellinger distance ($\tau = 2$), Kullback–Leibler divergence ($\tau \to \infty$) and Neyman's Chi–Square ($\tau = -1$). An alternative is represented by the families of RAF based on the Generalized Kullback-Leibler divergence

$$A_{gkl}(\delta, \tau) = \frac{\log(\tau\delta + 1)}{\tau}, \ 0 \leq \tau \leq 1;$$

maximum likelihood is a special case when $\tau \to 0$ and Kullback–Leibler divergence is obtained for $\tau = 1$ (see Cressie and Read, 1984, 1988; Park and Basu, 2003, and references therein). As one referee pointed out, one could define other functions to bound the effect of large Pearson residuals. However, in this paper, we still focused on the residual adjustment function. First, this choice is motivated by historical reasons, in the spirit of the work by Lindsay (1994) and Markatou et al (1998), among others. Then, despite the construction of the WLEE does not depend on the availability of an objective function, the RAF still arises naturally from a minimum disparity estimation problem. Therefore, the special role played by the RAF is justified in light of the connections between weighted likelihood estimation and minimum disparity estimation.

When the model is correctly specified, the Pearson residual function evaluated at the true parameter value converges almost surely to zero, whereas, otherwise, for each value of the parameters, large Pearson residuals detect regions where the observation is unlikely to occur under the assumed model. Hence, those observations lying in such regions are attached a weight that decreases with increasing Pearson residual. Large Pearson residuals and small weights will correspond to data points that are likely to be outliers.

Under the assumptions (A1)-(A9) stated in Section **??** of the Supplementary Material, Kuchibhotla and Basu (2018a, Theorem 3.4) proves that there exists a zero of the WLEE (5), say $\hat{\theta}_n$, which converges almost surely to θ_g and

$$n^{1/2}(\hat{\theta}_n - \theta_g) \xrightarrow{d} N(0, B^{-1}(\theta_g)V(\theta_g)B^{-1}(\theta_g)) \ .$$

with $V(\theta)$ and $B(\theta)$ defined as in (A9). Furthermore, it is proved that under the model there exists a $\theta_0 \in \Theta$ such that $g = q(y; \theta_0)$, $\theta_g \equiv \theta_0$ and $B^{-1}(\theta_0)V(\theta_0)B^{-1}(\theta_0) = I^{-1}(\theta_0)$ where $I(\theta)$ represents the expected Fisher Information matrix. The last statement means that the WLE is asymptotically first order efficient at the model.

Consider a contamination model such as in (2),

$$G(y) = Q_{\epsilon,n}(y) = (1-\epsilon)Q(y;\theta) + \epsilon H_n(y)$$

where the contaminant component H might depend on the sample size n, that is we assume a sequence $\{H_n\}_{n=1}^{\infty}$. Let $\hat{\theta}(G) = \arg\min_{\theta\in\Theta} \rho_C(g, q_\theta)$. Following Simpson (1987), it is possible to state that breakdown occurs for the functional $\hat{\theta}(G)$ at a contamination level ϵ if there exists a sequence $\{Q_{\epsilon,n}\}$ such that $||\hat{\theta}(Q_{\epsilon,n})-\theta_0|| \to \infty$ as $n \to \infty$. Under assumptions (B1)-(B4) stated in section **??** of the Supplementary Material, Kuchibhotla and Basu (2018a, Theorem 4.2) proves that the asymptotic breakdown point of the WLE is at least 0.5.

3 A new type of weighted likelihood estimator

Let us consider a measurable function h and the distribution function of $h(Y)$ both under the true model $G(y)$ and the postulated model $Q(y;\theta)$. We denote them by $F(y) = F(h(y))$ and $M(y;\theta) = M(h(y);\theta)$, respectively, whereas $f(y)$ and $m(y;\theta)$ are the corresponding densities. The newly established WLE is defined as the root of a WLEE that is obtained by combining the densities $f(y)$ and $m(y;\theta)$ in the construction of the Pearson residuals and the weights, while the score function is the one stemming from the assumed model $Q(y;\theta)$. In details, the Pearson residual function is now defined as

$$\delta(y) = \delta(y;\theta, f) = \frac{f(y)}{m(y;\theta)} - 1 \ .$$

In a similar fashion, the finite sample Pearson residual is given by

$$\delta_n(y) = \delta_n(y;\theta, \hat{f}_n) = \frac{\hat{f}_n(y)}{m(y;\theta)} - 1$$

where $\hat{f}_n(y) = \hat{f}_n(h(y))$ is a non parametric kernel estimate evaluated over the transformed data set $(h(y_1), \ldots, h(y_n))$. The WLEE is the same as in expression (4) but, now, the weights are driven from the distribution of the transformed data. We note that, obviously, when h is the identity function, the WLEE proposed by Kuchibhotla and Basu (2018a) is obtained.

This strategy looks particularly promising when the dimension of Y is large, while the dimension of $h(Y)$ is small, since the evaluation of the Pearson residuals would require only a non parametric density estimate in few dimensions. In particular, the interest lies in those situations in which the distribution of $h(Y)$ is univariate. An appealing situation stems from the use of pivots such that $m(y;\theta) = m(y)$. This is the approach that will be pursued in the development of weighted likelihood estimates of multivariate location and scatter.

In general, the use of transformations $h(Y)$ also turns to be useful and lead to improved kernel density estimates when the support of Y is bounded, e.g. the support is the non negative real line. Actually, in those cases, the kernel procedure should be adapted to account for the bounded support of the distribution. Then, appropriate transformations can be used so that the support of $h(Y)$ is the whole real line. These issues will be discussed below.

3.1 Weighted likelihood estimation based on robust distances

Here, we develop weighted likelihood estimation of the parameters of the multivariate normal model by using the aforementioned idea of calculating weights on transformed data. In order to define a set of weights whose computation does not need the evaluation of a multivariate kernel density estimate and does not suffer from problems due to large dimensionality, we suggest to focus on the transformation

$$d^2 = d^2(y, \theta) = (y - \mu)^\top \Sigma^{-1}(y - \mu) \ ,$$

that defines the squared Mahalanobis distance. At the multivariate normal model, the squared Mahalanobis distance satisfies

$$d^2(Y, \theta) \sim \chi^2_p \ , \qquad \theta = (\mu, \Sigma)$$

at the true parameter values. The main feature of this approach is that $M(y; \theta)$ does not depend on θ, whereas the transformed values do.

The weighting scheme will be based on Pearson residuals aiming at measuring the degree of agreement between a univariate kernel density estimate based on the vector of squared distances $d^2 = d^2(y; \theta)$ and their underlying χ^2_p distribution at the assumed multivariate normal model, that are defined as follows

$$\delta_n(d_i^2; \theta, \hat{F}_n) = \frac{\hat{f}_n(d_i^2)}{m_{\chi^2_p}(d_i^2)} - 1 \ . \tag{7}$$

Here, $\hat{f}_n(d_i^2) = \hat{f}_n(d^2(y_i; \theta))$ is a non parametric kernel density estimate based on the set $\left\{d_1^2, \ldots, d_n^2\right\}$ and $m_{\chi^2_p}$ is the density function of a χ^2_p variate. This strategy will lead, in most situations, to downweight those observations that exhibit a large distance from the robust fit.

The behavior of the Pearson residual function in (7) and the resulting weight function are exemplified in Figure 1. The true underlying model for the squared distances is assumed to be an ϵ-contaminated model of the form $f(x) = (1 - \epsilon)\chi^2_p(x) + \epsilon\chi^2_p(x, c)$, where the perturbing component is a non-central χ^2_p distribution with non centrality parameter c. This is equivalent to assume that the model in (2) is a mean shift model where $H(y)$ is a $N(y; \tau, \Sigma)$ and $c = 0.5(\mu - \tau)^\top(\mu - \tau)$ (Cerioli et al, 2013). Here we set $p = 2, c = 5, \epsilon = 0.05$. Large squared distances are likely to occur under the contaminating component and are expected to be downweighted at the χ^2_2 distribution. The left panel displays the Pearson residual function (7), that takes large values at large distances and, hence, detect a region where outlying distances are likely to occur. The weight function based on the Hellinger distance RAF is given in the right panel and clearly decreases at large distances. The vertical dashed line in the third panel gives the 0.975-level quantile of the χ^2_2 distribution: this is the quantile commonly used to declare a large distance and detect outliers in robust multivariate estimation.

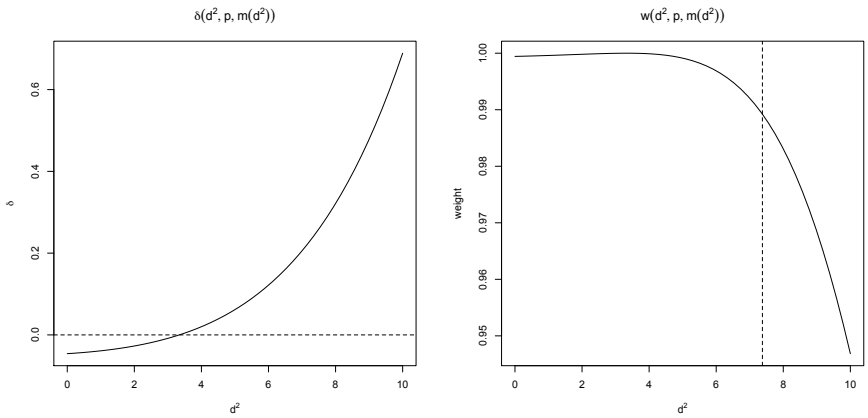

Fig. 1 Pearson Residual function (left panel). Weight function (right panel) based on the Hellinger distance RAF.

As stated in the Introduction, we are mainly interested in the multivariate normal model, but the same approach can be used for any elliptically symmetric family of distributions by using the distribution of $d^2(y;\mu,\Sigma)$ at the specified model. For instance, at the Student t_ν model, the distribution of squared distances is a scaled Fisher distribution.

The WLE of multivariate location and scatter $(\hat{\mu},\hat{\Sigma})$ is obtained as a weighted mean and weighted covariance matrix with data dependent weights. It is a common practice to consider an unbiased weighted likelihood estimates of the covariance matrix, that can be defined as

$$\hat{\Sigma}_u = \frac{\sum_{i=1}^n (y_i-\hat{\mu})(y_i-\hat{\mu})^\top \hat{w}_i}{\gamma \sum_{i=1}^n \hat{w}_i}\ , \qquad \gamma = 1 - \frac{\sum_{i=1}^n \hat{w}_i^2}{(\sum_{i=1}^n \hat{w}_i)^2}\ ,$$

where $\hat{w}_i = w(\delta_n(\hat{d}_i^2,\hat{\theta},\hat{F}_n))$, $\hat{\theta} = (\hat{\mu},\hat{\Sigma}_u)$, $\hat{d}_i = d(y_i,\hat{\mu},\hat{\Sigma}_u)$, $i = 1,2,\ldots,n$. Here, unbiasedness is meant in analogy with the multivariate normal model since when all the weights are equal to one, then $(n-1)$ appears in the denominator. Actually, this is the approach currently implemented in the `R` function `cov.wt` to get an unbiased estimate of scatter. It is worth to mention that the WLE of scatter does not require any consistency adjustment.

The computation of $(\hat{\mu},\hat{\Sigma}_u)$ yields an iterative procedure, as illustrated in Algorithm 1. At each iteration, based on the current values $(\hat{\mu},\hat{\Sigma}_u)$ robust distances are obtained. Then, their non parametric density estimate is fitted based on the chosen kernel and Pearson residuals and weights are updated. Algorithm 1 shares the main features of the iterative procedure developed to obtain weighted likelihood estimates in linear regression (Agostinelli and Markatou, 1998) and generalized linear models (Alqallaf and Agostinelli, 2016). Actually, at each iteration squared distances and their non-parametric density estimate are updated, whereas the model is held fixed.

Algorithm 1 WLE based on the Mahalanobis distance

Initialize $(\hat{\mu}, \hat{\Sigma}_u)$
Calculate squared distances

$$\hat{d}_i^2 = d^2(y_i, \hat{\mu}, \hat{\Sigma}_u)$$

Evaluate a nonparametric density estimate

$$\hat{f}_n(d_i^2) = n^{-1} \sum_{j=1}^{n} k(d_i^2; \hat{d}_j^2, h) \qquad i = 1, \ldots, n$$

Compute Pearson residuals

$$\delta_n(\hat{d}_i^2; \hat{\mu}, \hat{\Sigma}_u, \hat{F}_n) = \frac{\hat{f}_n(\hat{d}_i^2)}{m_{\chi_p^2}(\hat{d}_i^2)} - 1$$

Compute weights

$$\hat{w}_i = \frac{A(\delta_n(\hat{d}_i^2, \hat{\mu}, \hat{\Sigma}_u, \hat{F}_n)) + 1}{\delta_n(\hat{d}_i^2, \hat{\mu}, \hat{\Sigma}_u, \hat{F}_n) + 1}$$

Update $(\hat{\mu}, \hat{\Sigma}_u)$

Algorithm 1 may be initialized by drawing a large number of random subsets of fixed dimension $(p + 1)$. The sample mean and covariance matrix are evaluated over each subsample and used as starting values (Markatou et al, 1998). A deterministic solution to set initial values can be also implemented, stemming from that described in Hubert et al (2012). The fixed-point Algorithm 1 may generate multiple roots because of its dependence upon the different starting values. According to the results stated in Agostinelli (2006), we implemented a strategy that select the solution leading the lowest fitted probability

$$\Pr_{\hat{\theta}} \left[\delta(\hat{d}^2; \hat{\theta}, \hat{G}_n) < -0.95 \right] \ . \tag{8}$$

An alternative strategy would consist in minimizing the approximate disparity (3), meant as an approximate objective function. It is worth to note that Pearson residuals involved in (8) and (3) have to be evaluated at the fitted parameter value based on the original multivariate data for purely selection purposes. Actually, the sum of the weights provides a guidance for root selection, as well: when $\sum_{i=1}^{n} \hat{w}_i \approx 1$ than the WLE is close to the MLE, whereas when $\sum_{i=1}^{n} \hat{w}_i$ is too small, than the corresponding WLE is a degenerate solution, indicating that it only represents a small subset of the data. The reliability of the suggested root selection criteria has been illustrated on a real data example considered in Subsection 7.2. The results are given in the Supplementary material.

3.2 Computational issues

The kernel density estimate $\hat{m}_n(d_i^2)$ is expected to allocate all its probability mass over $[0, +\infty)$ because of the non-negativeness of the d_i^2. On the contrary, it would be biased at the boundary (Karunamuni and Alberts, 2005) and the comparison between the squared distances and the χ_p^2 model unfair. In the development of Algorithm 1, four methods are suggested to come through this issue. The first three are designed to obtain an unbiased at the boundary kernel density estimate, whereas the fourth is based on the distribution of log-transformed squared distances, moving the problem over the whole real line.

1. The reflection technique (Silverman, 1986) is based on data augmentation by adding the reflections of all the points in the boundary. Then, it is possible to implement any method originally designed for the whole real line. A reflection kernel can be defines as follows
$$k(y; t, h) = \frac{1}{h} k\left(\frac{y-t}{h}\right) + \frac{1}{h} k\left(\frac{y+t}{h}\right)$$
where $k(\cdot)$ is a symmetric and differentiable probability density: the reflection of a normal density leads to a folded normal kernel.
2. A kernel density estimate over $(0, \infty)$ can be also obtained by first log-transforming the squared distances, fitting a non parametric density estimate over the whole real line, i.e. $\hat{m}_n(\log d^2)$, and then back-transforming the fitted density to $(0, \infty)$, i.e. $\hat{m}_n(d^2) = \frac{1}{d^2}\hat{m}_n(e^{\log d^2})$, (Bowman and Azzalini, 1997). In this paper, we make use of the code available from the `R`-package `sm`.
3. The Gamma kernel (Chen, 2000)
$$k(y; t, h) = \Gamma(t; y/h + 1, h)$$
along with its version obtained by swopping the role of y and t (Jones and Henderson, 2007), where $\Gamma(t; a, b)$ denotes the probability density function of a Gamma variate with shape parameter a and scale parameter b evaluated at the point t. This is an appealing alternative that does not involve any transformation.
4. Log-transformed squared distances are distributed according to a $\log \chi_p^2$ model whose probability density function is
$$p(x; p) = \frac{1}{2^{p/2}\Gamma(p/2)} \exp\left[\frac{1}{2}(px - \exp(x))\right], \; x \in \mathbb{R} \; .$$
Then, Pearson residuals and weights can be evaluated on this new scale by comparing the fitted kernel density based on log-transformed squared distances with the $\log \chi_p^2$ distribution. Notice that this approach is fully compatible with our new general approach in the sense that the transformation $h(y) = \log(d^2(y; \theta))$ is used here.

The choice of the kernel is not crucial with regard to the properties of the method, both in terms of efficiency loss at the model and robustness, as confirmed by numerical studies. On the contrary, the smoothing parameter h indexing the kernel function $k(\cdot;\cdot)$ plays an important role in regulating the robustness/efficiency trade-off of the weighted likelihood methodology. 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. In other words, in finite samples more smoothing will lead to higher efficiency but larger bias under contamination.

The selection of h can be tuned by monitoring the empirical downweighting level $(1-\bar{\omega})$ as h varies, with $\bar{\omega} = n^{-1}\sum_{i=1}^{n}\hat{w}_i$, as already suggested by Markatou et al (1998). The idea is that $(1-\bar{\omega})$ gives a rough idea of the rate of contamination. This approach has been used in Greco (2017). As a toy example, the left panel of Figure 2 displays a perturbed sample of $n = 1000$ bivariate points: 700 are generated from a $N(0, \mathrm{I}_2)$ model, whereas 300 come from a $N(3, \mathrm{I}_2)$ distribution. We fit a bivariate normal model to the data at hand by using the proposed WLE, based on a reflection kernel and Symmetric chi-square RAF. Tolerance ellipses of level 0.95 have been over-imposed, that correspond to two different values of the smoothing parameter h (ellipses are based on a χ^2_2 distribution). The larger h clearly leads to a biased fit. The right panel gives $(1-\bar{\omega})$ as h varies. An abrupt change is visible before $h = 0.005$, that indicates the transition from a robust to a non robust fit. This adaptive procedure is reinforced in light of the nice features of the monitoring approach outlined in Cerioli et al (2017) and applied to the multivariate WLE framework in Agostinelli and Greco (2017). In particular, in the latter paper the WLE analyses are also monitored by looking at the behavior of individual robust distances as h varies.

We end this section on computational issues by remarking that we set all $k_{n,i}$ values equal to one. Actually, trimming was not necessary indeed.

4 Asymptotic properties

Asymptotic properties of the proposed weighted likelihood estimators are studied following the approach in Kuchibhotla and Basu (2018a,b). Let us consider the following estimating equation

$$T_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\kappa_i \frac{A(\delta_n(Y_i))+1}{\delta_n(Y_i)+1} s(Y_i;\theta) = 0$$

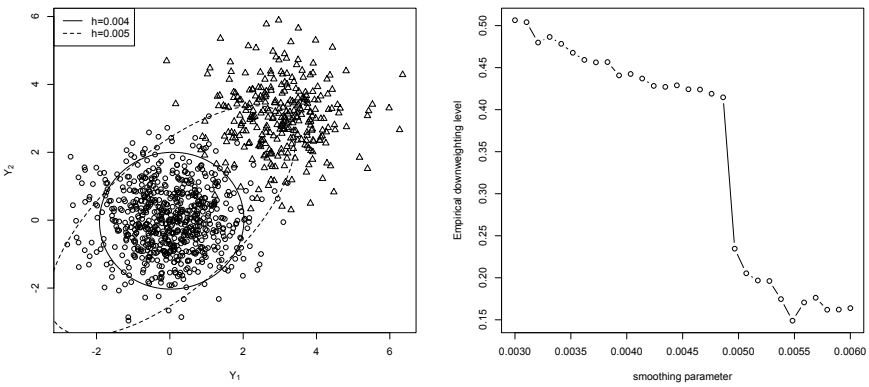

Fig. 2 Simulated data: fitted 95% tolerance ellipse based on the WLE (left) and empirical downweighting level for varying h (right). Outliers are denoted by triangles.

where $\kappa_i = \mathbb{1}_{D_n}$ is a trimming function, $D_n = \{y : f(y) \geq \gamma_n\}$, and δ, δ_n are as defined in Section 3. Let $\Psi_n(\theta)$ be the proposed WLEE

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \kappa_{n,i} \frac{A(\delta_n(Y_i)) + 1}{\delta_n(Y_i) + 1} s(Y_i;\theta) = 0$$

where now $\kappa_{n,i} = \mathbb{1}_{A_n}$ and $A_n = \{y : \hat{f}(y) > \gamma_n/2\}$. Under the set of assumptions (C) stated in Section **??** of the Supplementary Material we have that $T_n(\theta) - \Psi_n(\theta) = o_p(n^{-1/2})$ (see Theorem **??**) and that (see Corollary **??**)

$$n^{1/2}\left(\Psi_n(\theta) - \int \frac{A(\delta(y)) + 1}{\delta(y) + 1} s(y;\theta) g(y)\ dy\right) \xrightarrow{D} N(0, V(\theta))$$

where $V(\theta) = \mathbb{V}ar_g\left[A'(\delta(Y))s(Y;\theta)\right]$ as defined in assumption (C9). Using this previous result and the uniform convergence of the derivatives of the estimating function $\Psi_n(\theta)$ (see Section **??** for details) we have that there exists a zero of $\Psi_n(\theta)$, $\hat{\theta}_n$, which converges almost surely to θ_f and

$$n^{1/2}(\hat{\theta}_n - \theta_f) \xrightarrow{D} N(0, B^{-1}(\theta_f)V(\theta_f)B^{-1}(\theta_f))\ .$$

with $B(\theta)$ defined in assumption (C9), see Section **??** for details. Under the true model $g(y) = q(y;\theta_0)$, then it is easy to see that $B^{-1}(\theta_f)V(\theta_f)B^{-1}(\theta_f)$ reduced to the inverse of the expected Fisher Information $I(\theta_0)$ and the WLE is asymptotically first order efficient (see Corollay **??**). These results also provide the Influence Function of the proposed estimator (see Theorem **??** and Corollay **??**). In particular, it coincides with that of the maximum likelihood estimator under the true model.

5 Outlier detection

The availability of robust estimates of location and scatter $(\hat{\mu}, \hat{\Sigma})$ allows to activate some procedures designed to identify multivariate outliers. Actually, outliers in the sample are revealed by their large distances $d(y; \hat{\mu}, \hat{\Sigma}_u)$ from the robust fit. The use of robust estimates in place of the sample vector mean and covariance matrix avoids *masking* and *swamping* effects in outlier detection: there is masking whenever an outlier is not detected, swamping when a genuine observation is flagged as an outlier.

The problem of outlier detection consists in testing the n null hypotheses that each data point is a realization of a multivariate normal distribution, i.e. $H_{0i} : y_i \sim N_p(\mu, \Sigma)$. The detection rule will depend on the (asymptotic) distribution of the squared robust distances. A common approach to define cut-off values to flag outliers is based on the χ^2_p distribution to approximate the distribution of squared robust distances. A more accurate distributional result in finite samples may be used after the computation of the reweighted MCD estimator (Cerioli, 2010), but not in the case of M-type estimation. A rule of thumb is based on the 0.975 or 0.99-level quantile of the reference distribution. The outliers detection process could also be designed to take into account multiplicity arguments in the simultaneous testing of all the n data points, that is by considering the size of the test of the intersection hypothesis $\cap_i H_{0i}$. For instance, cut-off values can be based on a $(1-\alpha_{mult})$-level quantile such that the simultaneous testing of all the data points corresponds to a global nominal level α, with $\alpha_{mult} = 1 - (1-\alpha)^{1/n}$ (Cerioli, 2010). An alternative strategy is obtained by controlling the overall level of the simultaneous testing procedure by the False Discovery Rate (Cerioli and Farcomeni, 2011).

The asymptotic distribution of squared robust distances at the postulated multivariate normal model based on the WLE is still χ^2_p, because of its consistency. However, we argue that in finite samples their null distribution can be better approximated by using a result that resembles the classical one concerning the Mahalanobis distance evaluated at the unbiased MLE (Gnanadesikan and Kettenring, 1972), that is

$$d^2(Y_i, \hat{\mu}, \hat{\Sigma}_u) \approx \frac{(n-1)^2}{n} \text{Beta}\left(\frac{p}{2}, \frac{n-p-1}{2}\right) . \tag{9}$$

The reader is pointed to Ververidis and Kotropoulos (2008) to revise the classical proof based on the unbiased MLE. Our claim stems from the consideration that the weights are expected to tend to unity under the assumed model. The same result does not hold for M-type estimation since Huber and Tukey's bisquare weights do not share the asymptotic behavior of the weights in (6) at the postulated model. A close result has been established in the case of the reweighted MCD (Cerioli, 2010). The use of the scaled Beta distribution (9) in the process of outlier detection will result in smaller thresholds than those based on the χ^2_p.

6 Numerical studies

In this section we investigate the finite sample behavior of the newly proposed WLE of multivariate location and scatter and its performance for the goal of outlier detection through some numerical studies.

We first consider the accuracy of the multivariate WLE. The strategies outlined in Section 3 to compute Pearson residuals and the corresponding weights are all considered: folded normal kernel (WLEa), log and back transform (WLEb), log transform with $\log \chi^2_p$ (WLEc), gamma kernel (WLEd). The multivariate WLE has been also compared with the deterministic reweighted MCD (with 50% breakdown point and based on six initial solutions (Hubert et al, 2012)) and the S-estimator (with Rocke type weights and an asymptotic rejection point set equal to the running contamination rate, that has been designed to work properly for large dimensions), evaluated by using the functions from the `R` package `rrcov`. The WLE also runs on the deterministic algorithm set for the MCD.

Several combinations of (n, p) have been taken into account. Data have been generated according to the model (2), that is genuine data are sampled from a multivariate normal distribution with uncorrelated components and unit variance, i.e $\Sigma = \mathrm{I_p}$, whereas a percentage ϵ of outliers comes from a different perturbing stochastic mechanism. We consider the following scenarios:

- point mass contamination, $H(y) = \Phi_p(y; ka, \delta \mathrm{I_p})$, $\delta = 0.01$;
- location shift model, $H(y) = \Phi_p(y; ka, \mathrm{I_p})$, leading to clustered outliers;
- inflated scale model, $H(y) = \Phi_p(y; 0, k\mathrm{I_p})$, leading to radial outliers,

with $k = 1, 2, 3, \ldots, K$. The case $k = 0$ corresponds to the uncontaminated setting. When $p \leq 10$, contamination is designed to affect all dimensions, $a = (1, 1, \ldots, 1)^\top$, wheres when $p > 10$, outliers only contaminate the first five dimensions $a = (1, 1, 1, 1, 1, 0, \ldots, 0)^\top$. We show results corresponding to a contamination level $\epsilon = 20\%$ and two data configurations: in the first $n = 100$ and $p = 10$, in the second $n = 500$ and $p = 50$. The Hellinger RAF has been used for the scenario with point mass contamination, whereas the Symmetric Chi-square RAF has been used for the location shift model and the GKL ($\tau = 0.9$) RAF in the case of radial outliers. The same smoothing parameter has been used for each value of k, calibrated so that, on average, the empirical downweigting level was always in the range $[0.65 - 0.75]$. The numerical studies are based on 1000 Monte Carlo trials. All Figures and Tables concerning the numerical studies have been moved to Section **??** of the Supplementary Material for reason of space.

The following performance measures were considered:

1. $||\hat{\mu}||$
2. $\log \frac{\mathrm{trace}(\hat{\Sigma}_u)}{p}$
3. $\log_{10} \mathrm{cond}(\hat{\Sigma}_u)$
4. $||\hat{\Sigma}_{u_{jj}} - 1||,\ j = 1, 2, \ldots, p$
5. $||\hat{\Sigma}_{u_{jh}}||,\ j, h = 1, 2, \ldots, p, j \neq h$

6. computational time (in seconds on a 3,4 GHz Intel Core i5).

They are expected to be as close as possible to zero.

Figures **??** and **??** display the average performance measures for $n = 100, p = 10$ and $n = 500, p = 50$, respectively, under point mass contamination. The WLE provides very accurate results and an appealing behavior in that it compares remarkably well with the S-estimator and the MCD, whatever the chosen weighting scheme. In the first case, the WLE outperforms both the MCD and the S-estimator, in particular at $a = 1, 2, 3$. It is worth noting that the S-estimator is still badly affected by the point mass contamination at $a = 3$. In the second case, the WLE still performs satisfactory: in all the panels its performance measures exhibit the desired pattern even if at larger distances than the MCD. On the contrary, the S-estimator seems to fail. The behavior of the WLE under the location shift model is illustrated in Figures **??** and **??**. When $n = 100, p = 10$, the WLE performs not dissimilarly from the S-estimator and the MCD. As well as before, the S-estimator does not show robust features for $n = 500, p = 50$, whereas the WLE behaves in a fashion similar to the MCD for all values of k but at $k = 4$. Figures **??** and **??** gives the results in the case of radial outliers. The WLE still compares satisfactory both with the S-estimator and the MCD when $n = 100, p = 10$ and its behavior is particularly appealing when $n = 500, p = 50$, when also the MCD shows some lack of accuracy in estimating the elements of the covariance matrix. The fact that the WLE performs remarkably better than the S-estimator is a noticeable result, in that both of them fall in the general category of soft-trimming estimators, as stated in the Introduction.

A *redescending* pattern is observed very often. In the case of overlapping between the *genuine* multivariate normal model and the contaminating component $H(y)$, i.e. when outliers are not located at large distances, the robust methods all provide less accurate estimates since they are not completely able to identify outliers as such. On the contrary, at larger distances outliers are downweighted correctly. The redescending patterns of all the considered performance measures give evidence to the fact that the WLE exhibits an outlier stability property, in that as the distance of the outliers from the bulk of the genuine data increases, than the estimator behaves as if outliers were simply deleted from the sample at hand.

The sixth panels in all figures show the computational time. One needs to keep in mind that the comparison with the MCD and S- estimators is unfair, since the WLE is still based on an unoptimized `R` code, that will be soon available from the `R` package `wle`. Actually, computational time for the WLE remains in a feasible range, even when $p = 50$. It seems that the use of folded normal kernel leads to save computational time w.r.t. the other kernels but all of them provides very similar performances.

We now explore the finite sample behavior of of the outlier detection rules based on the robust distances stemming from the proposed multivariate WLE. First, we want to investigate the swamping error under the correct model. Here, the rate of swamping is a measure of the level α_{ind} of the individual testing

procedure about $H_{0i} : y_i \sim N(0, \mathrm{I_p}), i = 1, 2, \ldots, n$. A gamma kernel and Hellinger RAF have been used. The entries in Table **??** give the swamping error for $n = 50, 100$, p=5, 10 and several choices of the smoothing parameter h, both according to the χ^2_p and scaled Beta distribution in (9), for nominal levels $\alpha = 0.010, 0.025$. The last column gives the average empirical downweighting level. The results seem to confirm that the scaled Beta distribution provides an accurate approximation to the distribution of robust distances at the correct model in finite samples. Actually, the approximation improves as h increases and the weights all tend to unity.

Then, it is of interest to investigate both swamping and masking effects under contamination. The masking error corresponds to the type-II error of the test. Here, we assume that outliers are generated according to the location shift model. The reader is pointed to the paper by Cerioli et al (2013) for some regularity conditions on the perturbing component $H(y)$, that are necessary to control the level of the testing procedure. Actually, in the case of severe overlapping, the test is expected to show large masking effects. Both swamping and masking errors are given in Tables **??** and **??** for $\epsilon = 0.10$ and in Tables **??** and **??** for $\epsilon = 0.20$, according to different level of separations indexed by $k = 2, 2.5, 3$, different h values, a Symmetric-Chi-square RAF and $n = 50, 100$. Cut-off points have been set equal to the $(1 - \alpha_{ind}) = 0.975$ and $(1 - \alpha_{mult}) = 1 - 0.975^{1/n}$ level quantile of the reference distribution. The inspection of Tables **??** and **??** says that the actual level of the test is still acceptable, even with $p = 10$, both for the individual and intersection null hypothesis. On the contrary, in the presence of overlapping between the two components in (2), masking may be relevant, as we observe for $k = 2$ and $p = 5$ in Table **??** and Table **??**. When outliers are located far from the genuine part of the data, the power of the testing procedure increases in a desirable fashion. We also notice that the procedure becomes more powerful with growing dimensionality. A similar phenomenon is described in Cerioli et al (2013). By looking at the entires in Tables **??** and **??**, we note that by increasing the contamination rate both swamping and masking errors are larger than before. In particular, when $n = 50, p = 10$ the testing procedures are too liberal leading to large rates of false discoveries, whereas the size of the tests are again acceptable with $n = 100$. In a similar fashion, masking can still be non negligible also at $k = 2.5$. Only when $n = 100, p = 10$ the *blessing of dimensionality* makes the tests more powerful. We also give the empirical downweighting level corresponding to each scenario for the given h value. We note that $(1 - \bar{w})$ slightly increases as outliers are far apart. The empirical downweighting level exhibits more stability for $p = 10$. As a increases, h should become larger to keep $(1 - \bar{w})$ constant. As well, h need to be tuned with both n and p. Furthermore, we can appreciate how the selection of the smoothing parameter affects the two errors: swamping decreases with increasing h, whereas the opposite is true for masking, that is the more robust the procedure, the larger is the type-I error and the smaller is the type-II error. Actually, the numerical studies suggest that the smoothing parameter is to be set by monitoring the WLE analyses as h varies, but also by taking

into account different choices for the RAF, in order to improve the accuracy of both the fitting and the outlier detection procedure.

7 Real data examples

In this section we provide some real data examples concerning multivariate estimation of location and scatter, outlier detection, principal component analysis and discriminant analysis. The proposed weighted likelihood methodology is also compared with the WLE based on a multivariate kernel density estimate, in order to highlight its advantages, and other popular robust multivariate tools, to better assess its reliability. In all the examples, the multivariate normal model is considered as the central model, i.e. the model for the bulk of the data, and this assumption is expected to be tenable for the data at hand. Actually, we are interested in providing a simple description of the overall shape of the p-dimensional data through the vector mean, scatter and related ellipsoids, that is a an important aspect in many multivariate techniques.

7.1 StarsCYG data

The StarsCYG data give the effective temperature at the surface and the light intensity, both on a log scale, of 47 stars in the star cluster CYG OB1. Five stars are clear outliers: the points 11, 20, 30, 34 correspond to giant stars that do not lie on the main sequence and the point 7 also does not share the correlation structure of the remaining 43 stars. Figure 3 displays 0.975-level tolerance ellipses stemming from the proposed WLE, the WLE based on a bivariate kernel density estimate (WLEmulti), the reweighted MCD (with 50% breakdown point), the MM-estimator (with 50% breakdown point and 95% shape efficiency) and the MLE. All the methods we outlined to compute the weights for the WLE gave very similar results and hence, only the one based on log and back transform of distances is shown. A Hellinger RAF has been used. The weighted likelihood contours are based on the distributional approximation given in (9), whereas that stemming from the reweighted MCD is based on the result given in Cerioli (2010) (for those data in the non trimmed set) and that derived from the MM-estimator is based on the χ^2_2 distribution. The tolerance ellipse stemming from maximum likelihood is also based on the scaled Beta distribution. The fitted robust ellipses do not exhibit any significant difference and are all able to catch the correlation structure in the main sequence of stars: the multivariate WLE gives a correlation of 0.701, the proposed WLE gives 0.681, the MCD gives 0.655 and the MM gives 0.691. On the contrary, the MLE leads to inflated variability and negative correlation. The newly proposed WLE behaves not dissimilarly from the WLE based on the bivariate kernel. The left panel of Figure 4 displays the monitoring of $\sum_{i=1}^{n} \hat{w}_i$ as the smoothing parameter h varies. Here, h has been selected so that $1 - \bar{w} = 0.15$. It is worth to note that a different h is selected for each

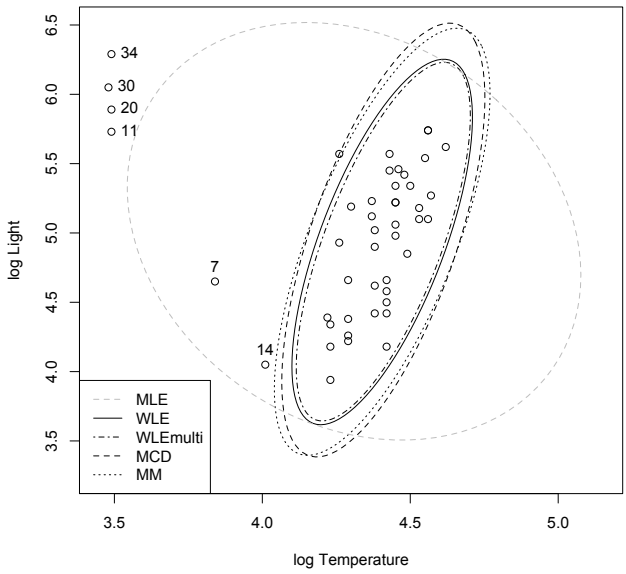

Fig. 3 StarsCYG data: fitted 97.5% tolerance ellipse based on the WLE, WLEmulti, reweighted MCD and MM-estimator.

weighting scheme. Moreover, it is worth to remark that the sum of the weights remains stable for growing h. This behavior proves the robustness of the WLE that does not deteriorate by increasing h. The right panel of Figure 4 gives the final weights corresponding to the WLE. All the outliers are given a weight that is almost null, wheres the genuine observations receive a weight that is equal or very close to one but observation 14, whose weight is about 0.48.

7.2 Auto data

The Auto data give information on technical and insurance characteristics of $n = 195$ cars collected in 1985 by the Insurance Institute for Highway Safety, for a total of $p = 15$ variables. The car are of two types: running on a gasoline or diesel engine. The are only 20 cars running on diesel that may be identified as outliers w.r.t. the others. In particular, the cars equipped with a diesel engine exhibit larger values of `compression-ratio` than those with a gasoline engine. Several outliers of different nature may also arise corresponding, for instance, to cars with peculiar technical features or deserving specific insurance conditions (Farcomeni and Greco, 2016). In the following we assume that the p-variate Normal model provides a simple but valid model for the data at hand.

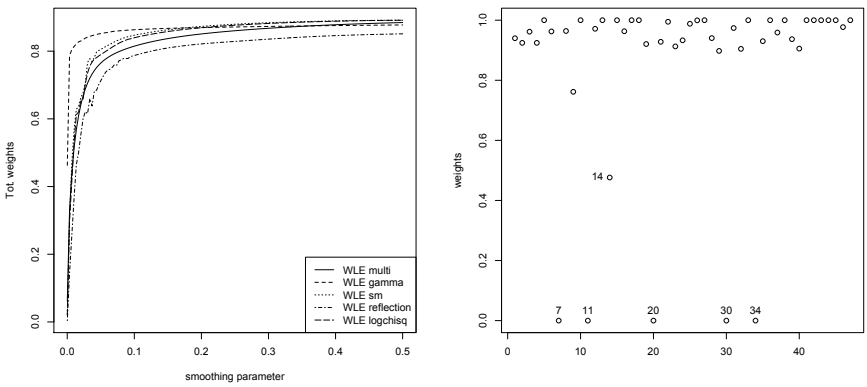

Fig. 4 StarsCYG data: monitoring of the sum of the weights for all the considered WLEs (left) and final weights from WLE based on `sm.density`.

First, we show the limitations of the original WLE based on a multivariate kernel and the advantages of the proposed methodology. Let us consider a subset of $n = 100$ cars, composed by eighty cars running on gasoline sampled at random from the original set and all the twenty cars with a diesel engine. We also take into account only $p = 5$ variables, `compression-ratio` being among them. Figure 5 shows the empirical downweigting level corresponding to the multivariate-kernel based WLE (left panel), implemented in the `R` function `wle.norm.multi` from the `wle` package, and the proposed multivariate WLE (right panel), based on a gamma kernel and the Hellinger RAF, as h varies. By looking at the left panel, we observe an abrupt change at $h = 0.53$. Actually, a robust solution is obtained only for $0.42 < h < 0.53$. In this range, the sum of the weights is always below 0.4, which means that the robust solution has been found at the cost of an excess of downweighting. For smaller h values the algorithm is not able to find a solution, whereas for larger values the solution is not robust. On the contrary, the proposed multivariate WLE leads to a robust fit for any value of h and at the cost of a smaller empirical downweigting level. Actually, the excess of downweighting characterizing the old fashioned WLE leads to an intolerable rate of swamping, as illustrated in Figure 6, that displays robust distances from both techniques, and too narrow tolerance ellipses, as given in Figure 7 for a couple of bivariate marginals. Here cut-off values and tolerance ellipses correspond to the 0.975 quantile of the scaled Beta distribution in (9). The dashed line in Figure 6 gives the cut-off value at the $1 - \alpha_{mult} = (0.975)^{(1/n)}$ level quantile, to take into account multiplicity. In general, the multivariate kernel based WLE does not lead to any solution for large p (w.r.t. n) and when a robust solution is available, it only happens at the cost of excessive downweighting, that implies intolerable swamping. In the case of the auto data, the old fashioned WLE does not lead to a robust solution for $p > 10$, when using the first p variables.

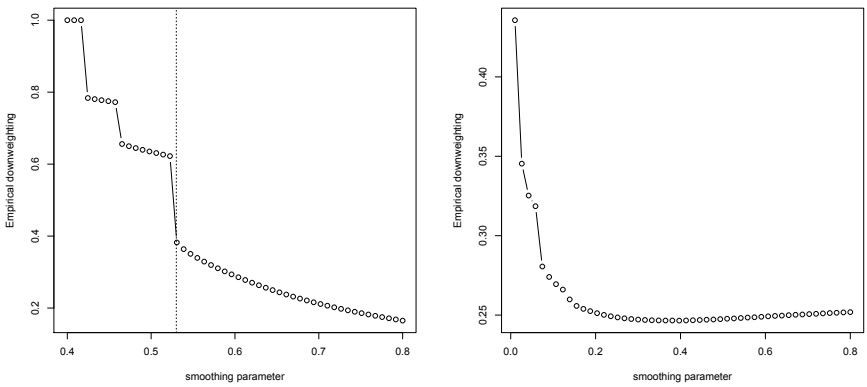

Fig. 5 Auto data subsample: monitoring the empirical downweighting level of the multivariate kernel (left) and univariate kernel (right) based WLE.

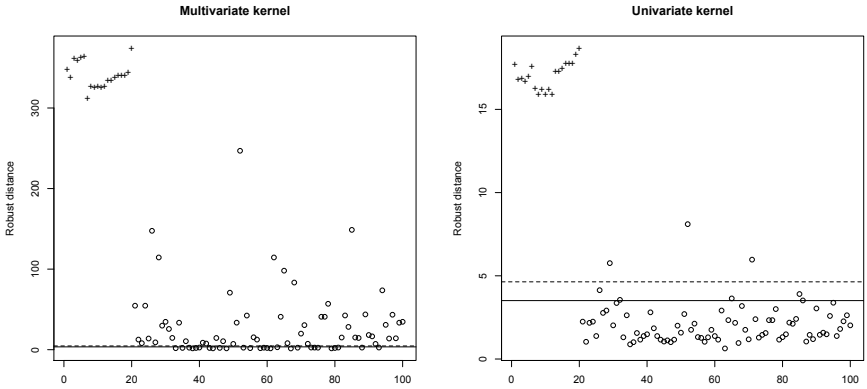

Fig. 6 Auto data subsample: robust distances from the multivariate kernel based WLE (left) and the proposed WLE (right) based on a gamma kernel and the Hellinger distance. Cars running on a diesel engine are denoted by a +. The solid line gives the 0.975-level cut-off value, the dashed line gives $(1-\gamma)$-level cut-off value to take into account multiplicity.

Let us consider the all data, now. Figure 8 gives the robust distances corresponding to each car stemming from the proposed WLE, with Hellinger RAF and gamma kernel. The fitted model corresponds to $h = 0.05$ and $1-\bar{w} \approx 70\%$. It is worth to remark that by increasing h the empirical downweighting level remains stable and the robustness of the procedure does not vanish. The group of cars running on diesel is clearly characterized by the largest distances and is well separated from the remaining cars. The inspection of the left panel of Figure 8 also unveils some other outlying cars that may exhibit peculiar characteristics. The group of cars running on diesel and the other outliers are clearly spotted by the QQ-plot in the left panel. The different nature of the

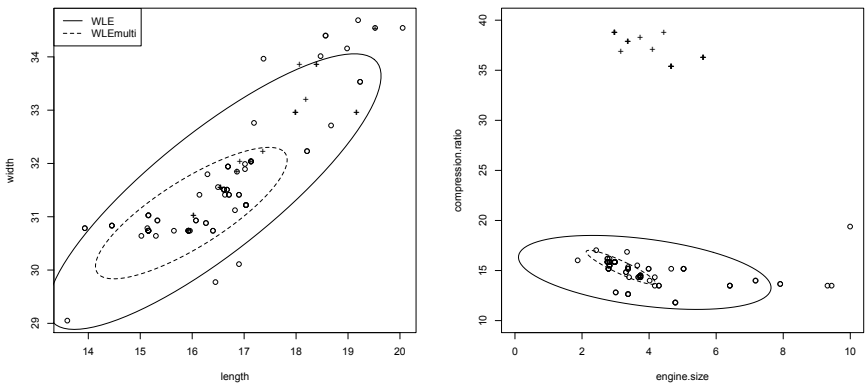

Fig. 7 Auto data subsample: 0.975-level tolerance ellipses from the multivariate kernel based WLE (left) and the proposed WLE (right) based on a gamma kernel and the Hellinger distance, for two bivariate marginals. Cars running on a diesel engine are denoted by a +.

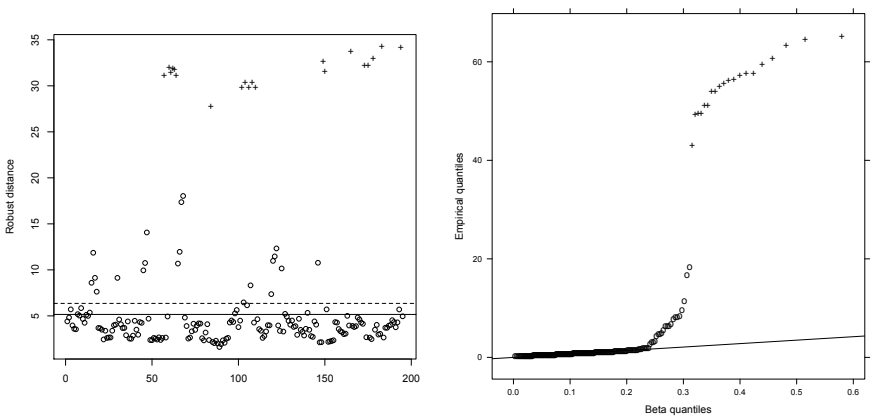

Fig. 8 Auto data: robust distances based on WLE (left) and scaled Beta QQ-plot. Cars running on a diesel engine are denoted by a +. The solid line gives the 0.975-level cut-off value, the dashed line gives $(1-\gamma)$-level cut-off value to take into account multiplicity.

several outliers that we have identified can be investigated further by exploring the distance-distance plot in Figure 9. The robust distances based on the WLE are compared with the classical distances based on the MLE. An important feature of such plot is that the cut-off values are determined according to the same scaled Beta distribution, hence being the same on both axes. It is worth noting that the group of cars running on a diesel engine would not have been detected by looking at the classical distances based on the MLE. The results driven by the use of the MCD, S and MM estimators are very similar.

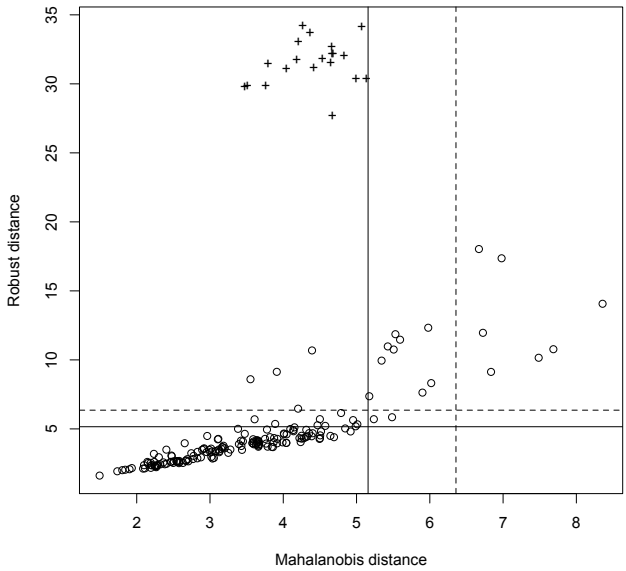

Fig. 9 Auto data: distance-distance plot based on the WLE. Cars running on a diesel engine are denoted by a +. The solid lines give the 0.975-level cut-off values, the dashed lines give $(1-\gamma)$-level cut-off value to take into account multiplicity.

7.3 Principal Component Analysis

Principal Component Analysis (PCA) is undoubtedly the most popular technique for dimension reduction. The data are projected onto a lower dimensional sub-space so that they are as spread out as possible. This feature allows to express the covariance structure of the data by means of a small number of new variables (the principal components). These new variables are obtained as linear combination of the original set of variables and are orthogonal each other. The coefficients of the linear combinations are given by the eigenvectors of the covariance matrix and each component accounts for an amount of total variability proportional to the corresponding eigenvalue. PCA is clearly sensitive to the occurrence of outliers, that, in particular, may inflate the variability accounted for by the first components hence leading to wrongly rotated loadings. One approach to robust PCA is based on the eigen-decomposition of a robust estimate of covariance. Here, we employ the WLE to perform a robust PCA on the Auto data. The same example has been discussed in Farcomeni and Greco (2016). Let us consider the first $k = 3$ components. The percentage of explained variance from standard PCA is 76.5% whereas the robust analysis gives a smaller value of 73.6%. In order to better explain the deleterious effect of outliers on standard PCA and the effectiveness of our weighted approach,

Figure 10 displays the pairwise score-plots based on the first three components. The group of cars running on diesel is clearly spotted by the robust components in the left panels, whereas this does not happens in the right panel. A typical effect due to the presence of outliers can be seen in the last panel: the second and third component from standard PCA still show a linear trend and only the effect of the outlying cars leads to a null correlation.

Robust PCA is an effective tool in outlier detection when the dimensionality is not of a manageable size. The usual tool is an outlier map, displayed in Figure 11, that is obtained by plotting for each data point its score and orthogonal distance: the group of outlying cars is clearly separated from the rest but also other outlying points are visible. Guidelines to find the cut-off values are given in Hubert et al (2005).

We only mention here, that the WLE of multivariate location and scatter could be used in the technique developed by Greco and Farcomeni (2016) to obtain sparse and robust PCA.

7.4 Discriminant Analysis

Discriminant analysis is concerned with the problem of assigning data to one of several groups. The observations within each group are assumed to arise from a multivariate normal distribution. In linear discriminant analysis (LDA) it is assumed homogeneity of the covariance matrices, whereas in quadratic discriminant analysis (QDA) the groups are allowed to have different scatters. Let $\pi_j, j = 1, 2, \ldots, k$ denote the prior probabilities. The linear discriminant rule classifies observations by maximizing

$$\log \hat{\pi}_j - \frac{1}{2} d^2(y, \hat{\mu}_j, \hat{\Sigma}_p)$$

and the quadratic discriminant rule classifies observations by maximizing

$$\log \hat{\pi}_j - \frac{1}{2} d^2(y, \hat{\mu}_j, \hat{\Sigma}_j) - \frac{1}{2} \log |\hat{\Sigma}_j|$$

where $\hat{\pi}_j = \frac{n_j}{\sum_{j=1}^{k} n_j}$ is an estimate of prior probabilities to be used when prior information is not available, $\hat{\mu}_j$ is an estimate of the group vector mean, $\hat{\Sigma}_p$ is a pooled estimate of the common scatter and, $\hat{\Sigma}_j$ is an estimate of the group scatter matrix. Actually, an appealing approach to define a discriminant function that is not prone to contamination in the data is based on robust estimates of location and common covariance matrix (Hubert and Van Driessen, 2004; He and Fung, 2000). Here, we apply weighted likelihood to perform robust LDA and QDA. In particular, we consider two different strategies to obtain a robust pooled estimate of the covariance matrix, in a fashion similar to what happens when using the MCD estimator (Todorov and Pires, 2007). By paralleling the standard technique, the first estimate (WLEA) is obtained by averaging the

Table 1 Diabetes data: misclassification rates for different rules based on all the data and on leave-one-out cross validation.

		ALL	CV
LDA	MLE	0.131	0.131
	WLEA	0.124	0.131
	WLEB	0.076	0.103
	MCDA	0.124	0.131
	MCDB	0.083	0.117
QDA	MLE	0.076	0.110
	WLE	0.076	0.103
	MCD	0.083	0.103

unbiased estimates from each group as follows:

$$\hat{\Sigma}_p^A = \frac{\sum_{j=1}^k \gamma_j \omega_j \hat{\Sigma}_{uj}}{\sum_{j=1}^k \gamma_j \omega_j}, \ \omega_j = \sum_{i=1}^{n_j} \hat{w}_{ij}$$

The second estimate (WLEB) can be obtained after the following steps. First center the data from each group by a robust estimate of location $\hat{\mu}_{j0}$; one could use the L1 (spatial) median, for instance. Then, evaluate the WLE $(\hat{\mu}_p, \hat{\Sigma}_p^B)$ from all the centered data and update the group vector means as $\hat{\mu}_j^B = \hat{\mu}_{j0} + \hat{\mu}_p$. The latter approach needs only one robust estimate of covariance rather than one for each group as in the former one. Nevertheless, an alternative, even if slightly more demanding, still consists in running Algorithm 1 for each group and centering the data by using the WLE of location from each of them in the first step.

Let us apply weighted LDA and QDA to the Diabetes data. These data consists of three measurement of plasma, `glucose`, `insuline` and `sspg`, made on 145 non-obese adult patients classified into three groups: normal subjects, chemical diabetes and overt diabetes. The weights are based on the Hellinger RAF and the kernel density stemming from log and back transform. The data, along with the fitted groups according to LDA based on WLEB and QDA stemming from group-wise WLEs, are displayed in Figure 12. The fitted groups appears as 0.975-level tolerance ellipses. It is worth noting the differences among the two techniques concerning, in particular, the peculiar nature of the overt diabetes group. Actually, the nature of correlation between `glucose` and `sspg` and `insuline` and `sspg` in the third group is different from what happens in the other two groups. The entries in Table 1 give the estimates of the misclassification rate based on all the data (ALL) and on leave-one-out cross validation (CV) based on the WLE, MLE and MCD, for LDA and QDA. The use of robust estimates of multivariate location and scatter improves classification accuracy over the standard approach based on the MLE and the WLE performs satisfactory compared to the MCD. In particular, both LDA based on WLEB and QDA stemming from the group-wise WLEs lead to the same results.

Aknowledgements

The authors wish to thank the Associate Editor and two anonymous referees for their comments and suggestions that lead to an improved version of the paper. The authors also wish to thank Kuchibhotla A. and Basu A. who kindly shared their technical report.

References

Agostinelli C (2006) Notes on pearson residuals and weighted likelihood estimating equations. Statistics & Probability Letters 76(17):1930–1934

Agostinelli C, Greco L (2017) Discussion of "the power of monitoring: how to make the most of a contaminated multivariate sample" by andrea cerioli, marco riani, anthony c. atkinson and aldo corbellini. Statistical Methods & Applications, https://doiorg/101007/s10260-017-0416-9

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

Alqallaf F, Agostinelli C (2016) Robust inference in generalized linear models. Communications in Statistics-Simulation and Computation 45(9):3053–3073

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

Basu A, Lindsay BG (1994) Minimum disparity estimation for continuous models: efficiency, distributions and robustness. Annals of the Institute of Statistical Mathematics 46(4):683–705

Basu A, Shioya H, Park C (2011) Statistical Inference: The minimum distance approach, Monograph on Statistics and Applied Probability, vol 120. CRC Press, Boca Raton, FL

Beran R (1977) Minimum Hellinger distance estimates for parametric models. The Annals of Statistics 5(3):445–463

Bowman A, Azzalini A (1997) Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations, vol 18. OUP Oxford

Cerioli A (2010) Multivariate outlier detection with high-breakdown estimators. Journal of the American Statistical Association 105(489):147–156

Cerioli A, Farcomeni A (2011) Error rates for multivariate outlier detection. Computational Statistics & Data Analysis 55(1):544–553

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

Cerioli A, Riani M, Atkinson AC, Corbellini A (2017) The power of monitoring: how to make the most of a contaminated multivariate sample. Statistical Methods & Applications, https://doiorg/101007/s10260-017-0409-8

Chen S (2000) Probability density function estimation using gamma kernels. Annals of the Institute of Statistical Mathematics 52(3):471–480

Cressie N, Read T (1984) Multinomial goodness-of-fit tests. Journal of the Royal Statistical Society, Series B (statistical methodology) 46:440–464

Cressie N, Read T (1988) Cressie–Read Statistic, Wiley, pp 37–39. In: Encyclopedia of Statistical Sciences, Supplementary Volume, edited by S. Kotz and N.L. Johnson

Croux C, Haesbroeck G (1999) Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. Journal of Multivariate Analysis 71(2):161–190

Deng H, Wickham H (2011) Density estimation in r. Electronic publication

Farcomeni A, Greco L (2016) Robust methods for data reduction. CRC press

Gervini D, Yohai VJ (2002) A class of robust and fully efficient regression estimators. Annals of Statistics pp 583–616

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

Greco L (2017) Weighted likelihood based inference for $p(x < y)$. Communications in Statistics-Simulation and Computation 46(10):7777–7789

Greco L, Farcomeni A (2016) A plug-in approach to sparse and robust principal component analysis. Test 25(3):449–481

He X, Fung W (2000) High breakdown estimation for multiple populations with applications to discriminant analysis. Journal of Multivariate Analysis 72(2):151–162

Huber P (1985) Projection pursuit. The Annals of Statistics pp 435–475

Huber P, Ronchetti E (2009) Robust statistics. Hoboken: John Wiley & Sons

Hubert M, Van Driessen K (2004) Fast and robust discriminant analysis. Computational Statistics & Data Analysis 45(2):301–320

Hubert M, Rousseeuw P, Vanden Branden K (2005) Robpca: a new approach to robust principal component analysis. Technometrics 47(1):64–79

Hubert M, Rousseeuw P, Van Aelst S (2008) High-breakdown robust multivariate methods. Statistical Science pp 92–119

Hubert M, Rousseeuw P, Verdonck T (2012) A deterministic algorithm for robust location and scatter. Journal of Computational and Graphical Statistics 21(3):618–637

Jones M, Henderson D (2007) Kernel-type density estimation on the unit interval. Biometrika pp 977–984

Karunamuni R, Alberts T (2005) On boundary correction in kernel density estimation. Statistical Methodology 2(3):191–212

Kuchibhotla A, Basu A (2015) A general set up for minimum disparity estimation. Statistics and Probability Letters 96:68–74

Kuchibhotla A, Basu A (2018a) A minimum distance weighted likelihood method of estimation. Tech. rep., Interdisciplinary Statistical Research Unit (ISRU), Indian Statistical Institute, Kolkata, India, URL https://faculty.wharton.upenn.edu/wp-content/uploads/2018/02/attemptv4p1.pdf

Kuchibhotla A, Basu A (2018b) Supplement to: "a minimum distance weighted likelihood method of estimation". Tech. rep., Interdisciplinary Statistical Research Unit (ISRU), Indian Statistical Institute, Kolkata, India, URL https://faculty.wharton.upenn.edu/wp-content/uploads/2018/02/attemptv4p2.pdf

Lindsay B (1994) Efficiency versus robustness: The case for minimum hellinger distance and related methods. The Annals of Statistics 22:1018–1114

Lopuhaa H (1989) On the relation between s-estimators and m-estimators of multivariate location and covariance. The Annals of Statistics pp 1662–1683

Markatou M, Basu A, Lindsay BG (1998) Weighted likelihood equations with bootstrap root search. Journal of the American Statistical Association 93(442):740–750

Maronna R (1976) Robust estimation of multivariate location and scatter. The Annals of Statistics 4(1):51–67

Maronna R, Martin R, Yohai V (2006) Robust statistics. John Wiley & Sons, Chichester

Park C, Basu A (2003) The generalized kullback-leibler divergence and robust inference. Journal of Statistical Computation and Simulation 73(5):311–332

Park C, Basu A (2004) Minimum disparity estimation: asymptotic normality and breakdown point results. Bull Inform Cybernet 36:19–33, special Issue in Honor of Professor Takashi Yanagawa

Park C, Basu A, Lindsay B (2002) The residual adjustment function and weighted likelihood: a graphical interpretation of robustness of minimum disparity estimators. Computational Statistics & Data Analysis 39(1):21–33

R Core Team (2018) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/

Riani M, Atkinson A, Cerioli A (2009) Finding an unknown number of multivariate outliers. Journal of the Royal Statistical Society: Series B (statistical methodology) 71(2):447–466

Rousseeuw P (1985) Multivariate estimation with high breakdown point. Mathematical statistics and applications 8:283–297

Salibian-Barrera M, Van Aelst S, Willems G (2006) Principal components analysis based on multivariate mm estimators with fast and robust bootstrap. Journal of the American Statistical Association 101(475):1198–1211

Scott DW, Wand M (1991) Feasibility of multivariate density estimates. Biometrika pp 197–205

Silverman B (1986) Density estimation for statistics and data analysis, vol 26. CRC press

Simpson D (1987) Minimum Hellinger distance estimation for the analysis of count data. Journal of the American Statistical Association 82(399):802–807

Todorov V, Pires A (2007) Comparative performance of several robust linear discriminant analysis methods. REVSTAT Statistical Journal 5:63–83

Ververidis D, Kotropoulos C (2008) Gaussian mixture modeling by exploiting the mahalanobis distance. IEEE transactions on signal processing 56(7):2797–2811

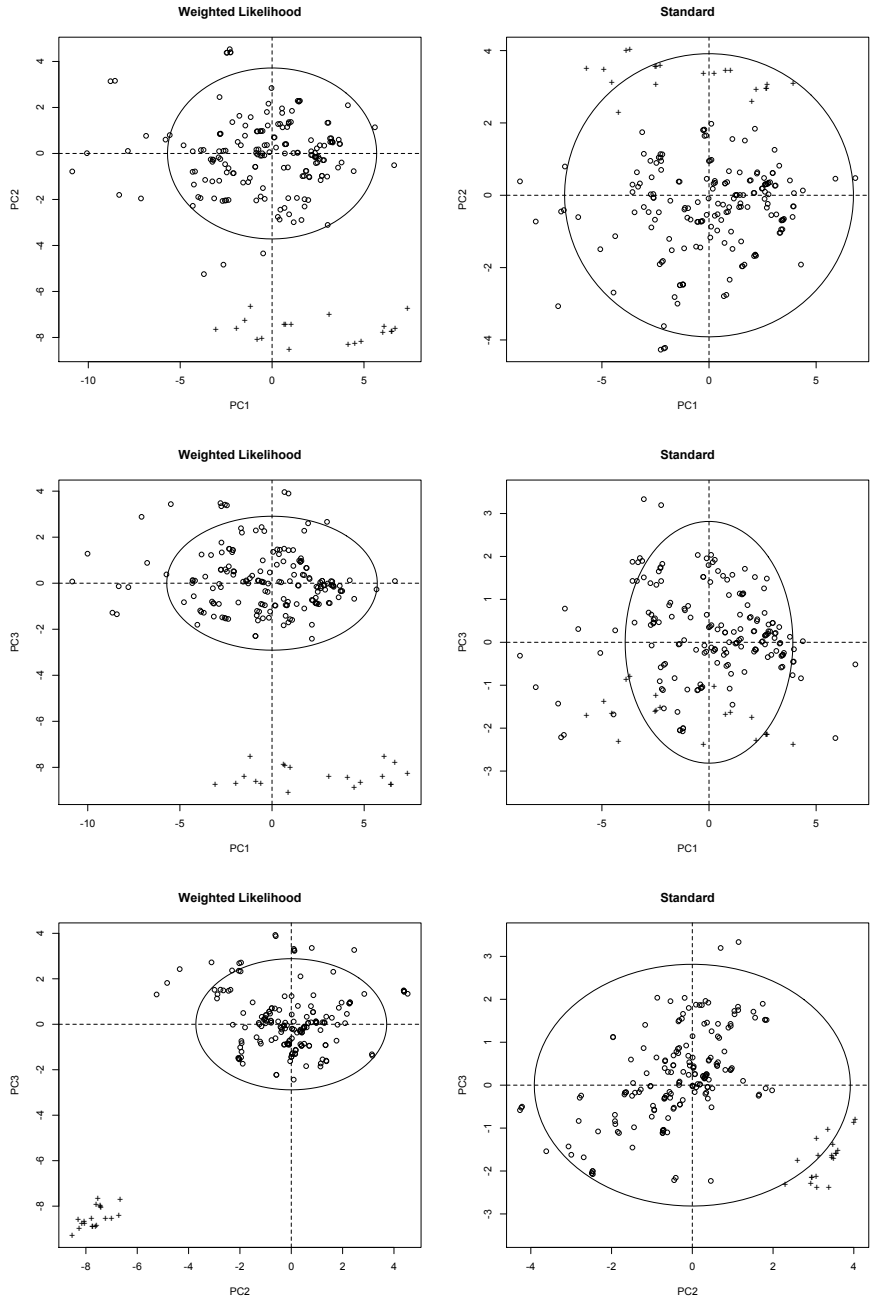

Fig. 10 Auto data: pairwise score-plots based on the first three components, robust (left) and standard (right). Cars running on a diesel engine are denoted by a +.

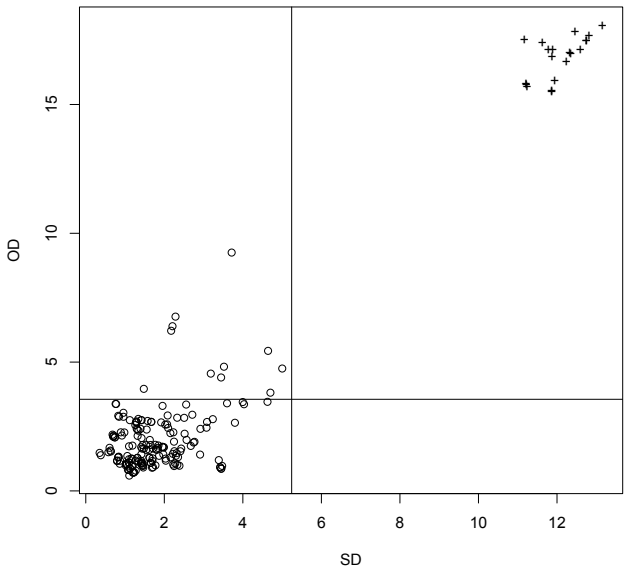

Fig. 11 Auto data: outlier map based on WLE with $k = 3$ components. Cars running on a diesel engine are denoted by a $+$.

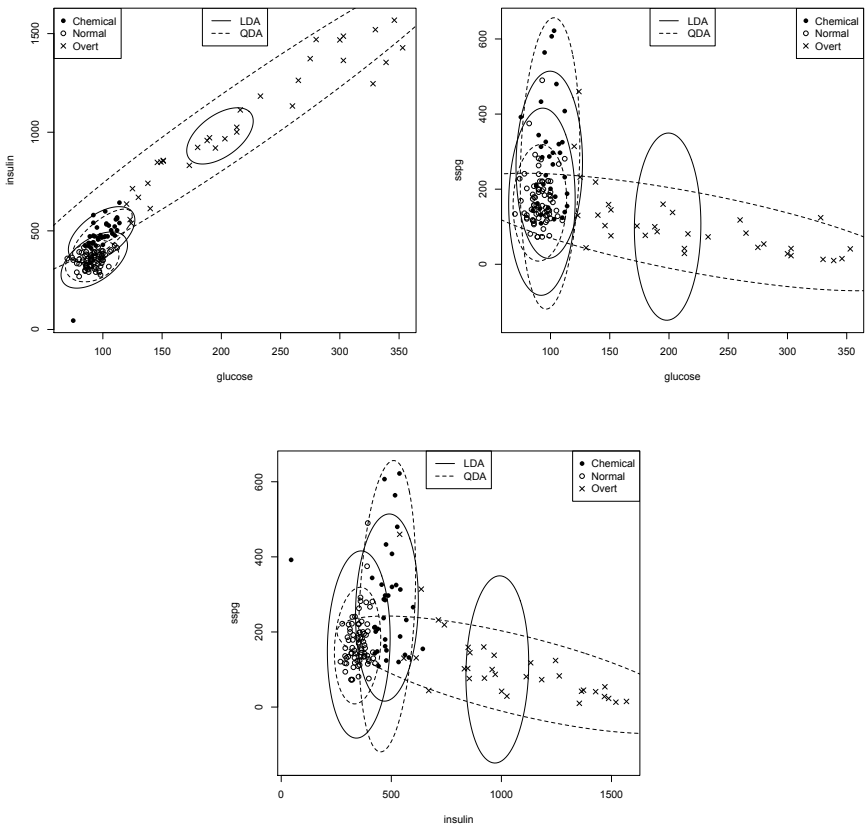

Fig. 12 Diabetes data: pairwise scatter-plots with 0.975-level tolerance ellipses over-imposed for each group, driven by LDA based on WLEB and QDA based on group-wise WLEs.