JDSSV

Journal of Data Science, Statistics, and Visualisation

October 2022, Volume II, Issue VI. *doi: 10.52933/jdssv.v2i6.47*

Robust Model-Based Clustering

Juan D. González
Acoustic Propagation Department
UNIDEF-CONICET

Ricardo A. Maronna
Universidad de Buenos Aires

Victor J. Yohai
Universidad de Buenos Aires and
CONICET

Ruben H. Zamar
University of British Columbia

Abstract

We propose a class of Fisher-consistent robust estimators for mixture models. These estimators are then used to build a *robust model-based clustering* procedure. We study in detail the case of multivariate Gaussian mixtures and propose an algorithm, similar to the EM algorithm, to compute the proposed estimators and build the robust clusters. An extensive Monte Carlo simulation study shows that our proposal outperforms other robust and non-robust, state of the art, model-based clustering procedures. We apply our proposal to a real data set and show that again it outperforms alternative procedures.

Keywords: mixture models, EM–algorithm, scatter S–estimators.

1. Introduction

Let $f(\mathbf{x},\boldsymbol{\theta})$ be a p-dimensional density function indexed by a q-dimensional parameter $\boldsymbol{\theta}$, and let $F_{\boldsymbol{\theta}}(\mathbf{x})$ be the corresponding distribution function. We consider the mixture model

$$h(\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\Theta})=\sum_{k=1}^{K}\alpha_k f(\mathbf{x},\boldsymbol{\theta}_k), \tag{1}$$

where $K > 0$ is some given integer, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K) \in [0,1]^K$, $\sum_{k=1}^K \alpha_k = 1$, and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K) \in \mathbb{R}^{q \times K}$. We assume that we have n independent observations from model (1). The important case of Gaussian mixtures is obtained when the *kernel density* $f(\mathbf{x}, \boldsymbol{\theta})$ is a multivariate normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ (that is, $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$).

The seminal work by Dempster et al. (1977) introduced the EM algorithm to compute the maximum likelihood estimators (MLE) for the parameters of a Gaussian mixture with K components. The MLE is efficient when applied to clean Gaussian data but performs poorly in the presence of *cluster outliers*, that is, data points that are far away from all the clusters (see García-Escudero et al. (2010)). Several authors addressed the problem of robust estimation of the parameters of a Gaussian mixture. García-Escudero et al. (2008) proposed the maximization of the likelihood of a multivariate normal mixture after trimming a given fraction, α, of the data. This procedure has a very good performance when the fraction α is well specified. However, α is often unknown and difficult to estimate. Another approach, that builds on previous work by Banfield and Raftery (1993) was proposed by Coretto and Hennig (2016) and Coretto and Hennig (2017). This approach consists of the addition of an improper uniform distribution with level δ to account for possible outliers. In the first implementation of this procedure called RIMLE, the level δ is a fixed input parameter. In the current implementation, called OTRIMLE, δ is estimated from the data.

We present a new estimating procedure with some desirable properties: (i) the estimators of the mixture model parameters are Fisher-consistent and (ii) our method does not require prior knowledge of the fraction of outliers in the data.

The rest of the paper is organized as follows. In Section 2, we present a general framework for the robust estimation of the parameters of a mixture model. In Section 3, this general framework is applied to the case of multivariate Gaussian mixtures. In Section 4, we give a computing algorithm. In Section 5, we discuss several practical issues including the allocation of observations to clusters and the flagging of outliers. In Section 6, we present the results of a simulation study. In Section 7, we apply our clustering procedure to a real dataset and compare the results with those of alternative clustering procedures. In Section 8, we provide some concluding remarks. Mathematical proofs and further details are given in the Appendix.

2. Robust Estimation of Mixture Models

We consider the problem of robust estimation of the parameters of the mixture model (1), $(\boldsymbol{\alpha}, \boldsymbol{\Theta})$, using a random sample $\mathbf{x}_1, ...\mathbf{x}_n$ from this model.

First, we give some general background and context for our proposal. We can think of model (1) as the marginal density of an observation, $\mathbf{X}$, from a random experiment with outcome $(\mathbf{U}, \mathbf{X})$, where the conditional density of $\mathbf{X}$ given $\mathbf{U} = \mathbf{u}$ is $p(\mathbf{x}, \boldsymbol{\Theta} | \mathbf{U} = \mathbf{u}) = \prod_{j=1}^K [f(\mathbf{x}, \boldsymbol{\theta}_j)]^{u_j}$ and the label vector $\mathbf{U}$ has a multinomial distribution Mult(K,$\boldsymbol{\alpha}$). Therefore, the joint density of $(\mathbf{U}, \mathbf{X})$ is $p(\mathbf{u}, \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\Theta}) = \prod_{j=1}^K [\alpha_j f(\mathbf{x}, \boldsymbol{\theta}_j)]^{u_j}$.

As in the classical EM algorithm, a key building block in the proposed robust estimation framework is the conditional probability that an observation $\mathbf{X}$ comes from the k^{th}

population given that $\mathbf{X} = \mathbf{x}$:

$$\widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\Theta}) = \frac{\alpha_k f(\mathbf{x}, \boldsymbol{\theta}_k)}{\sum_{j=1}^{K} \alpha_j f(\mathbf{x}, \boldsymbol{\theta}_j)}. \tag{2}$$

Another key building block is the robust base estimator discussed below. Finally, given the robust estimators $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\Theta}})$ produced by our proposal, observation $\mathbf{x}_i$, $i = 1, ..., n$, is assigned to cluster G_k iff $\widetilde{\alpha}_k(\mathbf{x}_i, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\Theta}}) = \max_{1 \le j \le K} \widetilde{\alpha}_j(\mathbf{x}_i, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\Theta}})$.

2.1. The Base Robust Estimator

We assume that given a random sample $\mathbf{x}_1, .., \mathbf{x}_n$ from the kernel density $f(\mathbf{x}, \boldsymbol{\theta})$, the parameter $\boldsymbol{\theta}$ has a robust estimator $\widehat{\boldsymbol{\theta}}$, which can be expressed as a function of h sample averages and satisfies a fixed point equation. More precisely, there exists a function $\mathbf{g} : \mathbb{R}^h \to \mathbb{R}^q$ and h real valued functions $\eta_j(\mathbf{x}_i, \boldsymbol{\theta})$, $1 \le j \le h$, such that

$$\widehat{\boldsymbol{\theta}} = \mathbf{g}\left(\frac{1}{n}\sum_{i=1}^{n} \eta_1\left(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}\right), ..., \frac{1}{n}\sum_{i=1}^{n} \eta_h\left(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}\right)\right). \tag{3}$$

In this case, the corresponding asymptotic functional $\boldsymbol{\theta}(F)$ for $\widehat{\boldsymbol{\theta}}$ when the underlying distribution is F satisfies the fixed point equation

$$\boldsymbol{\theta}(F) {=} \mathbf{g}\left(E_F\left\{\eta_1\left(\mathbf{x}, \boldsymbol{\theta}(F)\right)\right\}, ..., E_F\left\{\eta_h\left(\mathbf{x}, \boldsymbol{\theta}(F)\right)\right\}\right). \tag{4}$$

Many robust estimators satisfy this requirement.

Example: For simplicity's sake, let us consider a univariate location M-estimator $\widehat{\theta}$ implicitly defined by the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \psi(x_i - \widehat{\theta}) = 0.$$

To express $\widehat{\theta}$ as in (3) we write

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\psi(x_i - \widehat{\theta})}{x_i - \widehat{\theta}}\left(x_i - \widehat{\theta}\right) = 0.$$

Setting $W(x) = \psi(x)/x$ (defined by $\lim_{x \to 0} \psi(x)/x$ when $x = 0$) we have

$$\frac{1}{n}\sum_{i=1}^{n} W(x_i - \widehat{\theta})\left(x_i - \widehat{\theta}\right) = 0$$

or equivalently

$$\widehat{\theta} = \frac{\sum_{i=1}^{n} W(x_i - \widehat{\theta}) x_i}{\sum_{i=1}^{n} W(x_i - \widehat{\theta})}.$$

This satisfies (3) with $\eta_1(x, \theta) = W(x - \theta)x$, $\eta_2(x, \theta) = W(x - \theta)$ and $g(u, v) = u/v$. Similarly, the (more realistic) case of simultaneous location and scale M-estimators Huber (1964) can also be written as (3). In fact, many robust estimators satisfy (3) and (4). In particular, we show in Section 3 that Davies (1987) S estimators of multivariate location and scatter satisfy these conditions and therefore can be used for the robust estimation of the parameters of a multivariate Gaussian mixture.

2.2. The Mixture Model Estimator

Suppose now that we have a robust base estimator $\widehat{\boldsymbol{\theta}}$ satisfying (3) and (4). Then, given a random sample $\mathbf{x}_1, .., \mathbf{x}_n$, from model (1) we define the estimators

$$\left(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\Theta}}\right), \ \widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, ..., \widehat{\alpha}_K), \ \widehat{\boldsymbol{\Theta}} = (\widehat{\boldsymbol{\theta}}_1, ..., \widehat{\boldsymbol{\theta}}_K)$$

for the mixture model parameters $\boldsymbol{\alpha}= (\alpha_1, ..., \alpha_K)$ and $\boldsymbol{\Theta}= (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ as follows. Given $\mathbf{X} = \mathbf{x}$, let $\widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\Theta})$ be the conditional probability that this observation comes from the k^{th} sub-population, as given in equation (2). Then, $\widehat{\alpha}_k$ and $\widehat{\boldsymbol{\theta}}_k$ satisfy the fixed point equations:

$$\widehat{\alpha}_k = \frac{1}{n} \sum_{i=1}^{n} \widetilde{\alpha}_k(\mathbf{x}_i, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\Theta}}), \ 1 \leq k \leq K, \tag{5}$$

$$\widehat{\boldsymbol{\theta}}_k = g\left(\sum_{i=1}^{n} \frac{\widetilde{\alpha}_k(\mathbf{x}_i, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\Theta}})}{\widehat{\alpha}_k} \eta_1(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}_k), ..., \sum_{i=1}^{n} \frac{\widetilde{\alpha}_k(\mathbf{x}_i, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\Theta}})}{\widehat{\alpha}_k} \eta_h(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}_k)\right), \ 1 \leq k \leq K, \tag{6}$$

respectively.

Notice that $\widehat{\boldsymbol{\theta}}_k$ is the base estimator defined in (3) (still using the n observations) but with simple averages replaced by weighted averages. The i^{th} observation $\mathbf{x}_i$ has a weight proportional to the conditional probability, $\widetilde{\alpha}_k(\mathbf{x}_i, \boldsymbol{\alpha}, \boldsymbol{\Theta})$, that $\mathbf{x}_i$ belongs to the k^{th} sub-population.

Given the mixed model distribution H, we denote by $\mathbf{T}(H) = (\boldsymbol{\alpha}(H)\,\mathbf{,}\,\boldsymbol{\Theta}(H))$, the corresponding asymptotic functional of the robust estimators. The K components of $\boldsymbol{\alpha}(H)$ and $\boldsymbol{\Theta}(H)$ satisfy the fixed point equations

$$\alpha_k = \ E_H\left(\widetilde{\alpha}_k(\mathbf{x}_i, \boldsymbol{\alpha}\mathbf{,} \boldsymbol{\Theta})\right), \ 1 \leq k \leq K, \tag{7}$$

$$\boldsymbol{\theta}_k \ = g\left(E_H\left(\frac{\widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}\mathbf{,} \boldsymbol{\Theta})}{\alpha_k} \eta_1(\mathbf{x}, \boldsymbol{\theta}_k)\right), ..., E_H\left(\frac{\widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}\mathbf{,} \boldsymbol{\Theta})}{\alpha_k} \eta_h(\mathbf{x}, \boldsymbol{\theta}_k)\right)\right), \ 1 \leq k \leq K. \tag{8}$$

The theorem below shows that if the robust base estimator $\widehat{\boldsymbol{\theta}}$ is Fisher consistent, that is, if the corresponding asymptotic functional $\boldsymbol{\theta}(F)$ satisfies the equation

$$\boldsymbol{\theta}\mathbf{=g}\left(E_{F_{\boldsymbol{\theta}}}\left\{\eta_1\left(\mathbf{x}, \boldsymbol{\theta}\left(F_{\boldsymbol{\theta}}\right)\right)\right\}, ..., E_{F_{\boldsymbol{\theta}}}\left\{\eta_h\left(\mathbf{x}, \boldsymbol{\theta}(F_{\boldsymbol{\theta}})\right)\right\}\right), \quad \text{for all } \boldsymbol{\theta}\mathbf{,} \tag{9}$$

where $F_{\boldsymbol{\theta}}$ is the distribution function corresponding to the density $f(\mathbf{x},\boldsymbol{\theta}\mathbf{)}$, then the estimators for the mixture distribution parameters proposed above are also Fisher consistent.

Theorem 1. *Suppose that* $\mathbf{\Theta}_0 = (\boldsymbol{\theta}_{01}, ..., \boldsymbol{\theta}_{0K})$ *and* $\boldsymbol{\alpha}_0 = (\alpha_{01}, ..., \alpha_{0K})$ *are the true values of* $\mathbf{\Theta}$ *and* $\boldsymbol{\alpha}$*, respectively. Let* H_0 *be the corresponding true mixture distribution with density*

$$h_0(\mathbf{x},\boldsymbol{\alpha}_0, \mathbf{\Theta}_0) = \sum_{k=1}^{K} \alpha_{0k} f(\mathbf{x},\boldsymbol{\theta}_{0k}).$$

Suppose that the base estimator $\widehat{\boldsymbol{\theta}}$ *is Fisher consistent, then* $\left(\widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{\Theta}}\right)$ *is also Fisher consistent. That is,*

$$\mathbf{T}(H_0) = (\boldsymbol{\alpha}(H_0), \mathbf{\Theta}(H_0)) = (\boldsymbol{\alpha}_0, \mathbf{\Theta}_0)\,, \quad \textit{for all } (\boldsymbol{\alpha}_0, \mathbf{\Theta}_0).$$

2.3. Computing Strategy

Let $\mathbf{x}_1, ..., \mathbf{x}_n$ be a random sample from the mixture model (1) and let H_n be the corresponding empirical distribution function. We compute estimators $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}}) = \mathbf{T}(H_n)$ using an iterative approach. Suppose that, at step m, the current values of the estimators are $\boldsymbol{\alpha}^m = (\alpha_1^m, ..., \alpha_K^m)$ and $\mathbf{\Theta}^m = (\boldsymbol{\theta}_1^m, ..., \boldsymbol{\theta}_K^m)$. Then, for $1 \leq k \leq K$, we set

$$\alpha_k^{m+1} = E_{H_n}(\widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}^m, \mathbf{\Theta}^m)),\ 1 \leq k \leq K,$$

and

$$\boldsymbol{\theta}_k^{m+1} = \boldsymbol{g}\left(E_{H_n}\left(\frac{\widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}^m, \mathbf{\Theta}^m)}{\alpha_k^{m+1}}\eta_1(\mathbf{x}, \boldsymbol{\theta}_k^m)\right), ..., E_{H_n}\left(\frac{\widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}^m, \mathbf{\Theta}^m)}{\alpha_k^{m+1}}\eta_h(\mathbf{x}, \boldsymbol{\theta}_k^m)\right)\right).$$

Observe that if $(\boldsymbol{\alpha}^m, \mathbf{\Theta}^m) \to (\boldsymbol{\alpha}, \mathbf{\Theta})$, then $(\boldsymbol{\alpha}, \mathbf{\Theta})$ satisfies the fixed point equations (5) and (6).

Remark 1. Unlike other high breakdown point estimators, our procedure is not defined as the optimizer of an objective function, but as the solution of the fixed point problem defined in equations (5)-(6). In particular, although the algorithm described above bears some similarity with the EM algorithm, in that it alternates between two types of steps, none of them being a maximization step.

Initial estimators One way to define the initial estimators $\boldsymbol{\alpha}^0$ and $\mathbf{\Theta}^0$ for a multivariate normal mixture is given in Section 4.

Stopping rule. For each m, let H^m be the mixture model distribution with $(\boldsymbol{\alpha}, \mathbf{\Theta}) = (\boldsymbol{\alpha}^m, \mathbf{\Theta}^m)$. We stop the iterations when H^m and H^{m+1} are close enough. See Section 4 for further details for the case of multivariate normal mixtures.

3. Robust Estimation of Normal Mixtures

In this section, we propose a robust estimator for the parameters of a multivariate normal mixture model, based on the estimators defined in Section 2.2. In this case, the kernel density (1) is a multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$ and the chosen robust base estimator is the S estimator for multivariate location and scatter matrix (Davies (1987)), defined as follows. Given a p-dimensional vector $\boldsymbol{\mu}$, a $p \times p$ symmetric and positive definite matrix $\boldsymbol{\Sigma}$, and a distribution F on $\mathbb{R}^p$, the asymptotic scale functional $\sigma(F, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is implicitly defined by the equation

$$E_F \left(\rho_c \left(\frac{d(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\sigma(F, \boldsymbol{\mu}, \boldsymbol{\Sigma})} \right) \right) = b, \tag{10}$$

with

$$d^2(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}), \tag{11}$$

where $0.5 \leq b \leq 1$ and $\rho_c(d) = \rho(d/c)$, for a non-negative and non-decreasing function ρ such that $\rho(0)$=0 and $\sup \rho(d) = 1$. The tuning constant $c > 0$ is chosen so that

$$E(\rho_c(Y^{1/2})) = b, \qquad Y \sim \chi^2_{(p)}. \tag{12}$$

Then, if F is $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have $\sigma(F,\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 1$. The value of b determines the breakdown point of the estimator which is equal to $\min(b, 1-b)$. Finally, the S estimator functional of multivariate location and scatter is defined by

$$(\boldsymbol{\mu}(F), \boldsymbol{\Sigma}(F)) = \arg \min_{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})=1} |\boldsymbol{\Sigma}|, \tag{13}$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

Given a sample $\mathbf{x}_1, ..., \mathbf{x}_n$ in $\mathbb{R}^p$, the S estimator of multivariate location and scatter is obtained by replacing F by the empirical distribution F_n. That is,

$$(\boldsymbol{\mu}(F_n), \boldsymbol{\Sigma}(F_n)) = \arg \min_{\sigma(F_n,\boldsymbol{\mu},\boldsymbol{\Sigma})=1} |\boldsymbol{\Sigma}|, \tag{14}$$

with $\sigma(F_n, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ given by the equation

$$\frac{1}{n} \sum_{i=1}^{n} \rho_c \left(\frac{d(\mathbf{x}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\sigma(F_n, \boldsymbol{\mu}, \boldsymbol{\Sigma})} \right) = b.$$

3.1. S-estimators Fit the General Framework of Section 2.1

To write the asymptotic S functional as a fixed point of a function of means we need to introduce the auxiliary parameters $\boldsymbol{\Sigma}^*$ and s^*. The fixed point equation satisfied by the augmented S functional $(\boldsymbol{\mu}(F), \boldsymbol{\Sigma}(F), s^*(F), \boldsymbol{\Sigma}^*(F))$ is given in the following theorem.

Theorem 2. *Let $\psi = \rho'$ and $W(d) = \psi(d)/d$. Let $(\boldsymbol{\mu}(F), \boldsymbol{\Sigma}(F))$ be the S functional, then there exist a $p \times p$ symmetric and positive definite matrix $\boldsymbol{\Sigma}^*(F)$ and a scalar $s^*(F)$ such that $(\boldsymbol{\mu}(F), \boldsymbol{\Sigma}(F), s^*(F), \boldsymbol{\Sigma}^*(F))$ satisfies the following fixed point equations*

$$\begin{aligned}
\boldsymbol{\mu}(F) &= \frac{E_F\left(W\left(d(\mathbf{x}, \boldsymbol{\mu}(F), \boldsymbol{\Sigma}(F))\right)\mathbf{x}\right)}{E_F\left(W\left(d(\mathbf{x}, \boldsymbol{\mu}(F), \boldsymbol{\Sigma}(F))\right)\right)}, \\
\boldsymbol{\Sigma}^*(F) &= \frac{E_F\left(W\left(d(\mathbf{x}, \boldsymbol{\mu}(F), \boldsymbol{\Sigma}(F))\right)(\mathbf{x} - \boldsymbol{\mu}(F))(\mathbf{x} - \boldsymbol{\mu}(F))^T\right)}{E_F(W\left(d(\mathbf{x}, \boldsymbol{\mu}(F), \boldsymbol{\Sigma}(F))\right))}, \\
s^*(F) &= E_F\left((1/b)s^*(F)\rho(d(\mathbf{x}, \boldsymbol{\mu}\left(F\right), \boldsymbol{\Sigma}^*(F))/s^*(F))\right), \\
\boldsymbol{\Sigma}(F) &= s^*(F)^2 \boldsymbol{\Sigma}^*(F).
\end{aligned}$$

Theorem 2 shows that the augmented S functional $(\boldsymbol{\mu}(F), \boldsymbol{\Sigma}(F), s^*(F), \boldsymbol{\Sigma}^*(F))$ satisfies the requirements specified for the base estimating functional given in Section 2 with

$$\begin{aligned}
\eta_1(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^*, s^*, \boldsymbol{\Sigma}) &= W\left(d(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right)\mathbf{x}, \\
\eta_2(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^*, s^*, \boldsymbol{\Sigma}) &= W\left(d(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right), \\
\eta_3(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^*, s^*, \boldsymbol{\Sigma}) &= W\left(d(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right)(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}, \\
\eta_4(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^*, s^*, \boldsymbol{\Sigma}) &= (1/b)s^*\rho\left(d\left(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^*\right)/s^*\right),
\end{aligned} \tag{15}$$

and

$$\mathbf{g}(z_1, z_2, z_3, z_4) = \left(z_1/z_2, z_3/z_2, z_4, z_4^2 z_3/z_2\right).$$

and fixed point equations

$$\begin{aligned}
\boldsymbol{\mu} &= E(\eta_1(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^*, s^*, \boldsymbol{\Sigma}))/E(\eta_2(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^*, s^*, \boldsymbol{\Sigma})), \\
\boldsymbol{\Sigma}^* &= E(\eta_3(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^*, s^*, \boldsymbol{\Sigma}))/E(\eta_2(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^*, s^*)), \\
s^* &= E(\eta_4(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^*)), \\
\boldsymbol{\Sigma} &= E(\eta_4(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^*, s^*, \boldsymbol{\Sigma}))^2 E(\eta_3(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^*, s^*, \boldsymbol{\Sigma}))/E(\eta_2(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^*, s^*, \boldsymbol{\Sigma})).
\end{aligned} \tag{16}$$

3.2. The Loss Function

In this paper, we use the loss function

$$\rho(t) = \begin{cases} 1.38t^2 & \text{if} \quad 0 \le t < 2/3 \\ 0.55 - 2.69t^2 + 10.76t^4 - 11.66t^6 + 4.04t^8 & \text{if} \quad 2/3 \le |t| \le 1 \\ 1 & \text{if} \quad |t| > 1. \end{cases} \tag{17}$$

This is a simplified version of the optimal ρ function obtained by Yohai and Zamar (1997) for robust regression. Simulation studies showed that the S estimators for multivariate location and scatter based on this type of ρ functions have better performance than those based on the more traditional Tukey bisquare loss function (see Maronna and Yohai (2017)). For the simulation study, we set $b = 0.5$, which is the value maximizing the breakdown point. To simplify the notation, in the following we write ρ instead of ρ_c.

The values of c that satisfy equation (12) with $b = 0.5$ for ρ functions given in (17) can be found in Table 1 for $1 \le p \le 20$. An approximation (good for p in the range $1 \le p \le 400$) is given by

$$\hat{c}(p) = -\frac{0.1642}{p} + 0.5546\sqrt{p}.$$

The maximum error of this approximation is 0.015. That is, $|\hat{c}(p) - c(p)| \le 0.015$ for $1 \le p \le 400$.

Table 1: Value of the tuning constants satisfying equation (12) for different values of p.

p	1	2	3	4	5	6	7	8	9	10
c	1.21	2.08	2.70	3.19	3.61	3.99	4.33	4.65	4.94	5.22

p	11	12	13	14	15	16	17	18	19	20
c	5.48	5.73	5.97	6.20	6.42	6.64	6.84	7.04	7.24	7.43

4. Computing Algorithm

We now apply the computing strategy described in Section 2.3 to the case of a mixture of Gaussian distributions with fixed point equations (16).

Remark 2. It must be noted that the fixed point equation may have multiple solutions, and therefore the solution actually obtained depends on the initial estimator. The satisfactory behavior of our estimator is based on the fact that we use the K *Tau centers estimator* given in Gonzalez et al. (2019) as initial estimator of the algorithm. This estimator has been shown empirically to be sufficiently close to the "good" solution of the fixed point equations.

Initialization. We will assume that the number of clusters, K, is given. The initial values $\boldsymbol{\mu}^0 = (\boldsymbol{\mu}_1^0, ..., \boldsymbol{\mu}_K^0)$, $\boldsymbol{\Sigma}^0 = \left(\boldsymbol{\Sigma}_1^0, ..., \boldsymbol{\Sigma}_K^0\right)$, $\boldsymbol{\alpha}^0 = (\alpha_1^0, ..., \alpha_K^0)$ and s^{*0} can be obtained as follows:

Initial estimator for $\boldsymbol{\mu}_k$, $1 \le k \le K$: we use the K Tau centers estimator for $\boldsymbol{\mu}^0$.

Initial estimator for $\boldsymbol{\alpha}$: we first make an initial assignment of the data points to sub-populations by minimizing their Euclidean distances to the initial cluster centers $\boldsymbol{\mu}_k^0$. The initial values for the α_k are then taken equal to the relative frequency of each sub-population.

Initial estimator for $\boldsymbol{\Sigma}_k$, $1 \le k \le K$: we use the points assigned to each sub-population to compute the robust estimator of scatter proposed by Davies (1987).

Iteration. Let $\boldsymbol{\alpha}^m$, $\boldsymbol{\mu}^m$ and $\boldsymbol{\Sigma}^m$, be the current values for the mixture parameters. Then, $\boldsymbol{\alpha}^{m+1}$ and $\boldsymbol{\mu}^{m+1}, \boldsymbol{\Sigma}^{m+1}$ are computed as follows.

(a) Compute $\widetilde{\alpha}_{ki}$, $1 \le i \le n, 1 \le k \le K$, the probability that $\mathbf{x}_i$ belongs to the k^{th} sub-population when the mixture model parameters are $\boldsymbol{\alpha}^m$, $\boldsymbol{\mu}^m$ and $\boldsymbol{\Sigma}^m$

$$\widetilde{\alpha}_{ki} = \frac{f\left(\mathbf{x}_i, \boldsymbol{\mu}_k^m, \boldsymbol{\Sigma}_k^m\right)\alpha_k^m}{\sum_{l=1}^K f\left(\mathbf{x}_i, \boldsymbol{\mu}_l^m, \boldsymbol{\Sigma}_l^m\right)\alpha_l^m}, \tag{18}$$

where

$$f\left(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}.$$

(b) Update $\alpha_k, 1 \le k \le K$,

$$\alpha_k^{m+1} = \frac{\sum_{i=1}^n \widetilde{\alpha}_{ki}}{n}. \tag{19}$$

(c) Update $\boldsymbol{\mu}_k$, $1 \leq k \leq K$. First, we compute $\widetilde{d}_{ik} = d(\mathbf{x}_i,\boldsymbol{\mu}_k^m, \boldsymbol{\Sigma}_k^m)$, $1 \leq i \leq n$, $1 \leq k \leq K$. Then, $\boldsymbol{\mu}_k^{m+1}$ is the expectation of $\mathbf{x}$ when $\mathbf{x}_i$, $1 \leq i \leq n$ has probability $\widetilde{\alpha}_{ki} W(\widetilde{d}_{ik}) / \sum_{i=1}^n \widetilde{\alpha}_{ki}\ W(\widetilde{d}_{ik})$, that is,

$$\boldsymbol{\mu}_k^{m+1} = \frac{\sum_{i=1}^n \widetilde{\alpha}_{ki}\ W(\widetilde{d}_{ik}) \mathbf{x}_i}{\sum_{i=1}^n \widetilde{\alpha}_{ki}\ W(\widetilde{d}_{ik})}.$$

(d) Update $\boldsymbol{\Sigma}_k^*$, $1 \leq k \leq K$. Here $(\boldsymbol{\Sigma}_k^*)^{m+1}$ is the expectation of $(\mathbf{x} - \boldsymbol{\mu}_k^{m+1})(\mathbf{x} - \boldsymbol{\mu}_k^{m+1})^{\mathrm{T}}$ when $\mathbf{x}_i$, $1 \leq i \leq n$, has probability $\widetilde{\alpha}_{ki} W(\widetilde{d}_{ik}) / \sum_{i=1}^n \widetilde{\alpha}_{ki}\ W(\widetilde{d}_{ik})$, then

$$(\boldsymbol{\Sigma}_k^*)^{m+1} = \frac{\sum_{i=1}^n \widetilde{\alpha}_{ki} W(\widetilde{d}_{ik}) (\mathbf{x}_i - \boldsymbol{\mu}^{m+1})(\mathbf{x}_i - \boldsymbol{\mu}^{m+1})^{\mathrm{T}}}{\sum_{i=1}^n \widetilde{\alpha}_{ki}\ W(\widetilde{d}_{ik})}.$$

(e) Update $s_k^*, 1 \leq k \leq K$. First, we recompute $\widetilde{d}_{ik} = d(\mathbf{x}_i,\boldsymbol{\mu}_k^{m+1}, (\boldsymbol{\Sigma}_k^*)^{m+1})$, $1 \leq i \leq n$, $1 \leq k \leq K$. Then

$$(s_k^*)^{m+1} = (1/b) s_k^{*m} \frac{1}{n} \sum_{i=1}^n \frac{\widetilde{\alpha}_{ki}}{\alpha_k^{m+1}} \rho\left(\widetilde{d}_{ik} / s_k^{*m}\right).$$

(f) Update $\boldsymbol{\Sigma}_k, 1 \leq k \leq K$

$$\boldsymbol{\Sigma}_k^{m+1} = \left[(s_k^*)^{m+1}\right]^2 (\boldsymbol{\Sigma}_k^*)^{m+1}.$$

Stopping Rule. The iterations stop when

$$\left\| \boldsymbol{\alpha}^{m+1} - \boldsymbol{\alpha}^m \right\| < \delta$$

and

$$\sum_{k=1}^K d_{KL}(F_k^{m+1}, F_k^m) < \delta,$$

where $\delta > 0$ is the desired precision and $d_{KL}(F_k^{m+1}, F_k^m)$ are the Kullback–Leibler divergences between the distributions of the k^{th} components obtained at iterations m and $m+1$, respectively. This divergence is computed as in Hershey and Olsen (2007).

Remark 3. It may happen on some occasions that the matrix inversion required to compute the distance d defined in (11) is not feasible because of collinearity. In this case, we notice that these distances are only needed to compute the probabilities $\widetilde{\alpha}_{ki}$ in (18), in the weights $W(\widetilde{d}_{ik})$ and also in step (e) when updating s_k^*. Fortunately, this problem can be resolved using the general procedure to compute Mahalanobis distances described in Section 6.3.1 of Maronna et al. (2019).

We now make a conceptual comparison between our algorithm and the EM algorithm for the case of multivariate normal mixtures. The update of the mixture weights $\boldsymbol{\alpha}$, steps (a) and (b) of the iteration, is exactly the same in both algorithms. The updates of $\boldsymbol{\mu}_k$, step (c), are quite similar in both algorithms. In both cases the updating formulas are weighted means of the observations $\mathbf{x}_i$. However, while the weights used in the

EM algorithm are proportional to $\widetilde{\alpha}_{ki}$, the probability that $\mathbf{x}_i$ belongs to the k^{th} subpopulation, the weights used in our robust algorithm are proportional to the products $\widetilde{\alpha}_{ki} W(\widetilde{d}_{ik})$. The extra factor $W(\widetilde{d}_{ik})$ decreases with the distance of $\mathbf{x}_i$ to the center $\boldsymbol{\mu}_k$ of the k^{th} mixture component, ensuring that outliers that are far away from all the cluster centers have small - even zero - weight, and therefore have little influence on the value of the updated estimators of $\boldsymbol{\mu}_k$. A similar comment applies to the update of the matrix $\mathbf{\Sigma}_k^*$, step (d) in both algorithms. Our robust algorithm has two extra steps, steps (e) and (f), which are needed for a technical reason related to the use of S estimators; the matrix $\mathbf{\Sigma}_k^*$ is slightly biased as an estimator of $\mathbf{\Sigma}_k$ and requires a scalar correction factor $(s_k^*)^2$, which is calculated in step (e) and used in step (f). These steps are not needed in the case of the EM algorithm.

Notice that if $\rho(d) = ad^2$ for some constant $a > 0$, then $W(\widetilde{d}_{ik}) = 2a$ for all d_{ik}, and our algorithm reduces to the EM algorithm. Moreover, in the case that $\rho(d_{ik}/s^*) = \rho_c(d_{ik}) = \rho(d_{ik}/c)$, with ρ given by (17), if c is sufficiently large (as is our recommended default) and there are no outliers, then $d_{ik} \leq (2/3)c$ for all i and $\rho_c(d_{ik}) = (1.38/c)d_{ik}^2$. Therefore, when the data have no outliers, the estimators produced by the robust algorithm and the classical EM algorithm are very similar. However, when there are outliers, these outliers may severely affect the EM-algorithm but not much the robust algorithm because the outliers will be assigned small or even zero weights.

5. Robust Clustering

We can use the robust estimators $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_K)$, $\hat{\mathbf{\Sigma}} = \left(\hat{\mathbf{\Sigma}}_1, ..., \hat{\mathbf{\Sigma}}_K\right)$ and $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, ..., \hat{\alpha}_K)$ to define robust clusters. This approach is called *robust model-based clustering* (RMBC) .

Let

$$\widehat{P}(\mathbf{x}_i \in G_k) = \frac{f\left(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_k, \hat{\mathbf{\Sigma}}_k\right)\hat{\alpha}_k}{\sum_{l=1}^K f\left(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_l, \hat{\mathbf{\Sigma}}_l\right)\hat{\alpha}_l}. \tag{20}$$

Observation $\mathbf{x}_i$ is assigned to cluster G_j if

$$\widehat{P}(\mathbf{x}_i \in G_j) > \widehat{P}(\mathbf{x}_i \in G_k), \quad \text{for all } k \neq j.$$

Since the denominator in (20) does not depend on k we just need to compare the numerators in the log scale:

$$\delta_k(\mathbf{x}) = \log \hat{\alpha}_k - \frac{1}{2}\log|\hat{\mathbf{\Sigma}}_k| - \frac{1}{2}d^2(\mathbf{x}, \hat{\boldsymbol{\mu}}_k, \hat{\mathbf{\Sigma}}_k).$$

Remark 4. Flagging outliers: Let $\mathcal{E}_k = \{\mathbf{x} \in \mathbb{R}^p : d^2(\mathbf{x}, \hat{\boldsymbol{\mu}}_k, \hat{\mathbf{\Sigma}}_k) \leq \chi^2_{p,1-\beta}\}$, where $\chi_{p,\gamma}$ is the γ percentile of the χ^2 distribution with p degrees of freedom and $\beta = 10^{-3}$, and set $\mathcal{E}^K = \cup_{k=1}^K \mathcal{E}_k$. Observation $\mathbf{x}_i$ is flagged as an outlier if it falls outside $\mathcal{E}^K$. For the sake of consistency in the comparison, this definition of outliers will be used for all the methods compared in this work.

6. Simulation Study

We conduct a large simulation study to compare our proposal with several other robust and non robust clustering procedures. We considered several simulation scenarios and different performance metrics. The details are given below.

6.1. Scenarios Used in the Simulation Study

We generate 500 replications from six different simulation scenarios. In the first four scenarios, the data have a contaminated mixture distribution with K components. The density has the following form

$$(1-\varepsilon)\sum_{j=1}^{K}\alpha_k f(\mathbf{x},\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)+\varepsilon f_c(x), \tag{21}$$

where ε is the fraction of contamination and f_c is the contamination density. The first three scenarios, taken from Coretto and Hennig (2016), have fixed covariance matrices and are named SunSpot5, SideNoise2 and SideNoise2H (as in the given reference). We also simulated another scenario, SideNoise3, with fixed covariance matrix and two scenarios called RandScatterMatrix and RandScatterMatrixH, where the covariance matrices are generated at random for each replication. In the random covariance scenarios, the outliers are generated in a different appropriate way for each replication.

SunSpot5: In this case, we have $K = 5$ clusters, with weights $\boldsymbol{\alpha} = (0.15, 0.30, 0.10, 0.15, 0.30)$, in $\mathbb{R}^2$. The kernel distribution is normal with

$$\boldsymbol{\mu}_1 = (0,3) \quad \boldsymbol{\mu}_2 = (7,1) \quad \boldsymbol{\mu}_3 = (5,9) \quad \boldsymbol{\mu}_4 = (-13,5) \quad \boldsymbol{\mu}_5 = (-9,5),$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & -1.5 \\ -1.5 & 2 \end{pmatrix} \quad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 2 & 1.3 \\ 1.3 & 2 \end{pmatrix},$$

$\boldsymbol{\Sigma}_4 = 0.5\boldsymbol{I}_2$ and $\boldsymbol{\Sigma}_5 = 2.5\boldsymbol{I}_2$. In general, $\boldsymbol{I}_p$ denotes the identity matrix of dimension p. The fraction of contamination is $\varepsilon = 0.005$ with uniform distribution in the rectangle $[30, 40] \times [30, 40]$. The sample size in this case is $n = 1000$. This scenario generates a few isolated outliers far away from the bulk of data.

SideNoise2: In this case, we consider $K = 2$ clusters with weights $\boldsymbol{\alpha} = (0.75, 0.25)$ in $\mathbb{R}^2$. The kernel distribution is normal,

$$\boldsymbol{\mu}_1 = (-10,5),\ \boldsymbol{\mu}_2 = (3,13),\ \boldsymbol{\Sigma}_1 = 0.4I_2,\ \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1.5 & -1.1 \\ -1.1 & 1.5 \end{pmatrix}.$$

In this case, we take $\varepsilon = 0.10$ and the outliers are generated with uniform distribution in the square $[-50, 5] \times [-50, 5]$. The sample size for this scenario is $n = 1000$.

SideNoise2H: In this case, we also consider $K = 2$ clusters with weights $\boldsymbol{\alpha} = (0.75, 0.25)$ but this time in $\mathbb{R}^{20}$. The generating process for the first two coordinates is as in the previous case, including the addition of outliers (only the first two coordinates are contaminated). The remaining eighteen coordinates are independent standard normal random variables. The sample size for this scenario is $n = 2000$.

SideNoise3: In this case, we consider $K = 3$ clusters, with weights $\boldsymbol{\alpha}= (0.28, 0.33, 0.39)$ in $\mathbb{R}^2$. The kernel distribution has multivariate distribution with

$$\boldsymbol{\mu}_1 = (-2, -2), \quad \boldsymbol{\mu}_2 = (7, 1). \quad \boldsymbol{\mu}_3 = (15, 19),$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & -1.5 \\ -1.5 & 2 \end{pmatrix}, \quad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 2 & 1.3 \\ 1.3 & 2 \end{pmatrix}.$$

In this case, the fraction of contamination is $\varepsilon = 0.10$ with uniform distribution in the rectangle $[-20, 15] \times [-50, 5]$. The sample size for this scenario is $n = 1000$.

RandomScatter: In this case, we have $K = 6$ clusters with weights

$$\boldsymbol{\alpha} = (1/11, 2/11, 2/11, 2/11, 2/11, 2/11).$$

in $\mathbb{R}^2$. The kernel distribution is normal and $\boldsymbol{\mu}_k = 3(k-3)(1, 1), 1 \leq k \leq 6$. For each replication, $\boldsymbol{\Sigma}_k = \boldsymbol{U}_k \boldsymbol{U}_k^{\mathrm{T}}$, where $\boldsymbol{U}_k$ is a 2×2 random matrix, whose elements are independent uniform random variables on $[-1, 1]$. The fraction of contamination is $\varepsilon = 0.05$ generated from a uniform distribution on a region obtained as follows. The outliers are generated with uniform distribution in a box obtained by expanding by a factor of two the smallest box that contains the clean data. The sample size for this scenario is $n = 1200$.

RandomScatterH: The observations are generated as in the previous case but now with $p = 10$, and $\boldsymbol{\mu}_k = 3(k-3)\mathbf{1}, 1 \leq k \leq 6$, where $\mathbf{1}$ is a vector of 10 ones. Moreover, the $\boldsymbol{U}_k$s are of dimension 10×10. The sample size is $n = 1200$.

Clean Gaussian data: several clean Gaussian data scenarios are defined by choosing $\varepsilon = 0$ in the various settings described above.

Small samples scenarios: In addition to the sample sizes mentioned above for each scenario, we also compared the considered clustering procedures in a small sample setting. We did so for all the model scenarios in our simulation experiment that do not include a large number of variables. Moreover, to avoid the possible occurrence of clusters with very few observations we simulated clusters with balanced sizes, taking $\boldsymbol{\alpha} = (1/K)(1, 1...1)$. The number of observations is $n = 50 \times K$, where K is the number of clusters. In this way the expected number of good observations of each cluster is $50(1-\varepsilon)$, where ε is the fraction of outliers. In Figure 1 we show plots of these two–dimensional small sample scenarios.

Remark 5. For all the contaminated scenarios, if some generated outliers fall inside the 99% probability ellipsoids of the distributions used to generate the clean clusters, then these data points are removed and replaced by new ones generated in the same way until they all fall outside the 99% probability ellipsoids.

6.2. Estimators Compared in the Simulation Study

RMBC: This is the clustering procedure that we propose based on the estimators described in Sections 4 and 5. This procedure is applied using the function RMBC in the R package **RMBC** with all the default parameters.

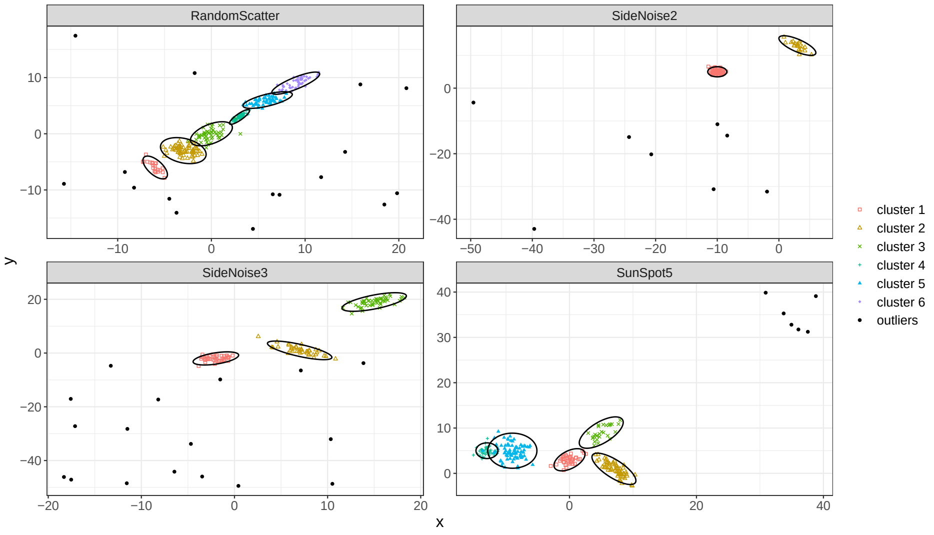

Figure 1: Plots of the small sample size two dimensional scenarios

Otrimle: This approach was proposed by Coretto and Hennig (2016). The outliers are identified by adding a cluster with an improper uniform density with level parameter δ,

$$g_\delta\left(\mathbf{x},\boldsymbol{\theta}\right) = \alpha_0\delta + \sum_{j=1}^{K}\alpha_j f\left(\mathbf{x},\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right), \qquad \sum_{j=0}^{k}\alpha_j = 1.$$

The first term in the mixture, δ, can be interpreted as an outlier generating improper density. The estimator $\hat{\boldsymbol{\theta}}_\delta$ maximizes the pseudo likelihood of the sample, that is,

$$\hat{\boldsymbol{\theta}}_\delta = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} g_\delta\left(\mathbf{x}_i,\boldsymbol{\theta}\right).$$

The pseudo maximum likelihood estimator is computed using an approach similar to the EM algorithm. Once the estimators, $(\widehat{\alpha}_j,\widehat{\boldsymbol{\mu}}_j,\widehat{\boldsymbol{\Sigma}}_j), 1 \leq j \leq K$ and $\widehat{\delta}$ are computed, the probability that each observation belongs to a given cluster is obtained as if we were dealing with all true densities. Each observation $\mathbf{x}_i$ is assigned to the cluster with largest posterior probability. An observation is called an outlier if it is assigned to the pseudo uniform distribution. The R package `otrimle` (based on an algorithm proposed in Coretto and Hennig (2017)) estimates all the parameters including δ.

Mclust: Fraley and Raftery (2002) proposed the maximum likelihood estimator for the Gaussian mixture model with the maximization carried out by the EM algorithm, introduced by Dempster et al. (1977). In our simulation study and example, we use the function Mclust in the R package mclust described in Scrucca et al. (2016).

Tclust: García-Escudero et al. (2008) proposed the *α-trimmed maximum likelihood estimators*, $0 < \alpha < 1$, which maximizes the function

$$\prod_{j=1}^{K}\prod_{i\in C_j}\alpha_j f(\mathbf{x}_i,\boldsymbol{\mu}_j\boldsymbol{\Sigma}_j),$$

where $C_1, ...C_K$ are disjoint subsets of $\{1, ..., n\}$ such that if $C_0 = \{1, ..., n\} - \cup_{j=1}^{K} C_j$ then $\#C_0 = \alpha n$. The main idea is that αn data points are collected in C_0 and labeled as potential outliers, while the remaining $\mathbf{x}_i \in C_j$ with $j > 0$ are regular observations. This idea was previously explored by Gallegos and Ritter (2005), under the assumption that the α_i, $1 \leq i \leq K$ are equal and all the covariance determinants $|\mathbf{\Sigma}_i|$ $1 \leq i \leq K$ are also equal. García-Escudero et al. (2008) study this estimator under a more general constraint $\Gamma \leq \delta$, where $\Gamma = \lambda_{max}/\lambda_{min}$ and λ_{max} and λ_{min} are the maximum and minimum eigenvalues of all the matrices $\mathbf{\Sigma}_j$, $j = 1, \ldots, K$. This procedure is implemented in the function tclust in the package **tclust** built by Fritz et al. (2012). In our simulation study, we consider two versions of this procedure: tclust with $\alpha = 0.05$ which is the default value for α in the function tclust, and tclustOracle, where α matches the contamination fraction in the given scenario.

Remark 6. The function tclust assigns the trimmed observations to a separate cluster called G_0. For a fair evaluation of the cluster results we reassign these observations to the clusters with largest estimated probability (see equation (20)) calculated using the parameter estimates reported by the function tclust. Moreover, the outliers are also flagged using the confidence 99.9% ellipsoids described at the end of Section 5, again using the reported parameter estimates.

All the considered procedures are run using the default values for their tuning parameters, except for tclustOracle, where the trimming parameter is set equal to the true contamination level in the given scenario.

6.3. Performance Measures

Misclassification Rate (MCR): This measure focuses on the identification of the true clusters. Suppose we have n observations known to belong to K clusters labeled $1, 2, ..., K$. Suppose that we run a clustering algorithm to estimate these K clusters. We compare each estimated cluster with each one of the true clusters and search for the matching that produces the minimum number, m, of misclassified observations. Then, the MCR is defined as $MCR = m/n$. An observation is considered as misclassified if and only if it is generated as a member of one of true clusters and assigned by the clustering procedure to another one, or if it is flagged as an outlier (see Remark 4 for the definition of outlier). Therefore, generated outliers assigned to a cluster are not considered as misclassified.

Extended Misclassification Rate (EMCR): It is similar to MCR but adding a cluster containing the outliers. Then, both the outliers that are assigned to one of the original clusters as well as the good observations flagged as outliers are considered as misclassified.

Kullback–Leibler divergence (KLD): This divergence is computed between the estimated and true mixture densities, using the procedure described in Hershey and Olsen (2007).

Sensitivity: This measure evaluates the procedure's ability to flag the true outliers. An observation is flagged as outlier if it falls outside of all the 0.999 confidence ellipsoids calculated using the estimated location and scatter parameters. Then, the sensitivity is defined as the proportion of actual outliers that are flagged as such.

Specificity Rate(SR): It is the fraction of good observation (no outliers) classified in one of the clusters. In our tables we show $1 - SR$, that is, the fraction of good observations classified as outliers (false outliers) .

6.4. Simulation Results

For each procedure, scenario and replication, we compute the performance measures described above. Table 2 gives the MCR, KL-divergence and 1-SR performances for the case of clean Gaussian data and large samples. The EMCR is not included because for clean data the EMCR is equal to the MCR. Table 3 gives the same results plus the EMCR and the sensitivity average performances for the case of contaminated Gaussian data. Tables 4 and 5 report the same results than as those of Tables 2 and 3 for the case of the small sample case settings described later on in this section.

Table 2: Large sample simulation results for the six scenarios and the different clustering procedures without outliers. We do not report the EMCR results here, because for clean data these are equal to the MCR results.

Method	SunSpot5	SideNoise2	SideNoise2H	SideNoise3	R.Scatter	R.ScatterH
			Misclassification Rate (%)			
RMBC	2.15 (0.028)	0.13 (0.005)	0.13 (0.004)	0.16 (0.006)	0.83 (0.038)	0.34 (0.008)
otrimle	2.56 (0.134)	1.69 (0.329)	0.1 (0.004)	0.15 (0.015)	10.64 (0.489)	1.69 (0.138)
tclust	5.33 (0.275)	0.48 (0.011)	0.33 (0.006)	0.36 (0.008)	7.05 (0.265)	1.06 (0.014)
tclustoracle	4.53 (0.177)	0.09 (0.003)	0.08 (0.002)	0.09 (0.003)	5.6 (0.208)	0.39 (0.006)
mclust	6.63 (0.435)	0.09 (0.004)	0.1 (0.003)	0.09 (0.004)	3.85 (0.275)	7.36 (0.357)
			Kullback-Leibler Divergence			
RMBC	0.04 (0.001)	0.01 (0.0004)	0.12 (0.001)	0.02 (0.001)	0.04 (0.001)	0.21 (0.001)
otrimle	0.12 (0.021)	1.87 (0.507)	0.12 (0.001)	0.01 (0.002)	3.53 (0.340)	2.11 (0.062)
tclust	0.08 (0.002)	0.05 (0.001)	0.17 (0.001)	0.03 (0.001)	0.66 (0.022)	2.21 (0.023)
tclustoracle	0.03 (0.001)	0.01 (0.001)	0.12 (0.001)	0.01 (0.001)	0.49 (0.011)	2.15 (0.016)
mclust	0.03 (0.002)	0 (0.000)	0.12 (0.001)	0.01 (0.001)	0.09 (0.007)	0.91 (0.033)
			(1-Specificity) (%)			
RMBC	0.14 (0.006)	0.13 (0.005)	0.13 (0.004)	0.16 (0.006)	0.17 (0.006)	0.32 (0.007)
otrimle	0.52 (0.084)	1.25 (0.208)	0.1 (0.004)	0.15 (0.015)	8.52 (0.393)	1.48 (0.105)
tclust	0.5 (0.013)	0.48 (0.011)	0.33 (0.006)	0.36 (0.008)	1.85 (0.079)	1.02 (0.014)
tclustoracle	0.08 (0.003)	0.09 (0.003)	0.08 (0.002)	0.09 (0.003)	0.24 (0.007)	0.35 (0.006)
mclust	0.07 (0.004)	0.09 (0.004)	0.1 (0.003)	0.09 (0.004)	0.09 (0.006)	0.31 (0.027)

Results for clean Gaussian data for large samples: The simulation results for clean Gaussian data reveal the superiority of RMBC, which outperforms the other robust procedures in terms of both, MCR and KL-divergence. In this case, mclust and tclustOracle are designed for dealing with clean Gaussian data. Therefore, it is surprising that RMBC outperforms these two procedures in three scenarios (SunSpot5, RandomScatter and RandomScatterH). This may be due to the fact that RMBC uses a very good initial estimator.

Results for Contaminated Gaussian Data for large samples: Notice that in general, as expected, tclustOracle outperforms tclust, for all the metrics, whenever the

Table 3: Large sample simulation results for the six scenarios and the different clustering procedures with outliers.

Method	SunSpot5	SideNoise2	SideNoise2H	SideNoise3	R.Scatter	R.ScatterH
Misclassification Rate (%)						
RMBC	2.19 (0.057)	0.05 (0.003)	0.04 (0.002)	0.08 (0.004)	1.29 (0.119)	0.63 (0.107)
otrimle	2.43 (0.121)	1.14 (0.250)	0.11 (0.004)	0.18 (0.016)	8.83 (0.449)	1.48 (0.116)
tclust	5.35 (0.280)	5.75 (0.474)	9.43 (0.539)	12.46 (0.631)	7.49 (0.308)	0.37 (0.019)
tclustoracle	7.92 (0.314)	0.09 (0.005)	0.13 (0.004)	0.12 (0.006)	7.49 (0.308)	0.37 (0.019)
mclust	13.92 (0.223)	0.07 (0.004)	24.05 (0.053)	0.53 (0.012)	11.04 (0.227)	10.79 (0.231)
Extended Misclassification Rate (%)						
RMBC	2.18 (0.057)	0.08 (0.004)	0.1 (0.003)	0.4 (0.009)	1.4 (0.117)	0.6 (0.102)
otrimle	2.42 (0.121)	1.06 (0.226)	0.15 (0.004)	0.3 (0.015)	8.48 (0.431)	1.4 (0.110)
tclust	5.32 (0.278)	7.32 (0.515)	11.19 (0.574)	14.94 (0.654)	7.29 (0.295)	0.35 (0.018)
tclustoracle	8 (0.319)	0.14 (0.005)	0.21 (0.005)	0.31 (0.009)	7.29 (0.295)	0.35 (0.018)
mclust	14.31 (0.228)	9.67 (0.049)	31.67 (0.052)	9.55 (0.055)	15.56 (0.216)	15.32 (0.219)
Kullback-Leibler Divergence						
RMBC	0.04 (0.001)	0.03 (0.001)	0.2 (0.001)	0.06 (0.001)	0.07 (0.010)	0.37 (0.058)
otrimle	0.08 (0.016)	1.21 (0.404)	0.14 (0.001)	0.02 (0.001)	2.36 (0.215)	2.01 (0.046)
tclust	0.07 (0.002)	1.62 (0.017)	1.74 (0.034)	0.63 (0.007)	0.55 (0.017)	2.19 (0.023)
tclustoracle	0.06 (0.002)	0.1 (0.007)	0.19 (0.004)	0.02 (0.002)	0.55 (0.017)	2.19 (0.023)
mclust	0.12 (0.001)	1.12 (0.003)	2.11 (0.001)	1.06 (0.003)	0.41 (0.011)	1.51 (0.021)
Sensitivity (%)						
RMBC	100 (0e+00)	99.61 (0.029)	99.33 (0.027)	96.72 (0.080)	96.61 (0.227)	100 (0e+00)
otrimle	100 (0e+00)	99.69 (0.026)	99.44 (0.024)	98.56 (0.057)	98.01 (0.340)	100 (0e+00)
tclust	100 (0e+00)	79.26 (0.831)	73.39 (0.854)	63.14 (0.822)	96.33 (0.123)	100 (0.003)
tclustoracle	84.08 (1.139)	99.45 (0.035)	99.12 (0.042)	98.1 (0.077)	96.33 (0.123)	100 (0.003)
mclust	16.98 (1.676)	4.48 (0.103)	0.05 (0.007)	9.22 (0.203)	0.11 (0.025)	0.01 (0.007)
(1-Specificity) (%)						
RMBC	0.13 (0.008)	0.05 (0.003)	0.04 (0.002)	0.08 (0.004)	0.2 (0.039)	0.41 (0.057)
otrimle	0.39 (0.065)	0.89 (0.165)	0.11 (0.004)	0.18 (0.016)	6.58 (0.335)	1.34 (0.092)
tclust	0.43 (0.012)	0.01 (0.001)	0.09 (0.005)	0.05 (0.004)	0.26 (0.011)	0.3 (0.008)
tclustoracle	0.10 (0.004)	0.09 (0.005)	0.13 (0.004)	0.12 (0.006)	0.26 (0.011)	0.3 (0.008)
mclust	0.09 (0.004)	0.00 (0.00)	0.09 (0.003)	0 (0.001)	0.00 (0.00)	0.00 (0.00)

tclust default trimming parameter ($\alpha = 0.05$) falls below the fraction of contamination in the simulated data. On the other hand, tclust does relatively well when the percentage of outliers is below 5%. Unfortunately, the true fraction of contamination in the data is seldom known and difficult to estimate from the data.

Regarding MCR, RMBC has the best performance in all the considered scenarios except for RandomScatterH where tclust has the best performance. Even in this case, RMBC is a relatively close runner up. Regarding KLD, RMBC overall has a good performance. It comes first under all the considered scenarios except for SideNoise2H and SideNoise3 where otrimle and tclustOracle perform slightly better. Regarding sensitivity, RMBC in general performs very well, detecting always at least 96% of the true outliers. As expected, mclust fails to identify the greatest majority of the true outliers. Also notice that tclust has poor sensitivity (fails to detect a large fraction of true outliers) whenever

Table 4: Small sample simulation results for the four low dimensional scenarios and the different clustering procedures without outliers.

Method	SunSpot5	SideNoise2	SideNoise3	RandomScatter
		Misclassification Rate (%)		
RMBC	3.53 (0.074)	0.85 (0.033)	0.83 (0.028)	1.79 (0.054)
otrimle	5.00 (0.197)	3.60 (0.226)	2.51 (0.167)	13.20 (0.376)
tclust	6.53 (0.186)	0.84 (0.028)	0.92 (0.026)	11.23 (0.234)
tclustoracle	5.58 (0.132)	0.05 (0.005)	0.04 (0.004)	10.01 (0.171)
mclust	5.46 (0.256)	0.06 (0.008)	0.10 (0.008)	4.3 (0.232)
		Kullback-Leibler Divergence		
RMBC	0.19 (0.003)	0.19 (0.004)	0.19 (0.003)	0.20 (0.002)
otrimle	0.42 (0.029)	0.83 (0.106)	0.45 (0.091)	4.67 (0.266)
tclust	0.20 (0.003)	0.13 (0.002)	0.14 (0.002)	0.75 (0.013)
tclustoracle	0.10 (0.001)	0.06 (0.001)	0.06 (0.001)	0.64 (0.009)
mclust	0.09 (0.002)	0.06 (0.001)	0.07 (0.001)	0.15 (0.006)
		(1-Specificity) (%)		
RMBC	0.57 (0.019)	0.85 (0.033)	0.83 (0.028)	0.73 (0.019)
otrimle	2.22 (0.155)	3.60 (0.226)	2.50 (0.165)	11.99 (0.364)
tclust	1.12 (0.024)	0.84 (0.028)	0.92 (0.026)	1.33 (0.031)
tclustoracle	0.05 (0.003)	0.05 (0.005)	0.04 (0.004)	0.16 (0.006)
mclust	0.05 (0.004)	0.07 (0.008)	0.10 (0.008)	0.07 (0.007)

its default trimming parameter ($\alpha = 0.05$) falls below the percentage of contamination in the data.

Results for small samples: The results reported in Tables 4 and 5 show that, as already seen before in the large sample setting, RMBC also exhibits a good relative performance in the considered small sample setting. In the case of no outliers, Table 4 shows that mclust has the lowest MCR in two of the four scenarios, while RMBC does the same in the other two. In the presence of outliers Table 5 shows that RMBC has the best MCR in three of the four scenarios, and the best EMCR in all four of them.

7. Application to Real Data

Phytoplankton, being a primary food source for a wide range of sea creatures, plays a fundamental role in the marine ecosystem. Furthermore, there are some phytoplankton species that can be used as biological indicators of pollution in oceanic areas, and others that produce massive algal blooms that affect activities carried out by man. So estimating phytoplankton abundance is an important ecological problem.

The acoustic monitoring of phytoplankton is a potentially useful technique for estimating the abundance of these organisms in real time. Therefore, in the last decade, ultrasound techniques have been developed to obtain information about these organisms. See for example, Blanc et al. (2004), Bok et al. (2010) and Blanc et al. (2017).

In particular, we will work with data from Cinquini et al. (2016), obtained by taking laboratory measurements of ultrasonic acoustic signals; a pulse is emitted by a trans-

Table 5: Small sample simulation results for the four low dimensional scenarios and the different clustering procedures with outliers.

Method	SunSpot5	SideNoise2	SideNoise3	RandomScatter
Misclassification Rate (%)				
RMBC	3.95 (0.105)	0.56 (0.031)	0.57 (0.029)	2.22 (0.102)
otrimle	4.63 (0.182)	3.10 (0.182)	1.94 (0.120)	11.31 (0.345)
tclust	8.36 (0.225)	16.89 (0.702)	13.52 (0.483)	12.52 (0.234)
tclustoracle	11.38 (0.264)	1.31 (0.214)	0.64 (0.108)	12.52 (0.234)
mclust	18.99 (0.063)	0.49 (0.124)	0.66 (0.052)	15.43 (0.065)
Extended Misclassification Rate (%)				
RMBC	3.83 (0.102)	0.54 (0.028)	0.82 (0.031)	2.27 (0.098)
otrimle	4.51 (0.178)	2.81 (0.164)	1.88 (0.108)	10.80 (0.327)
tclust	8.15 (0.220)	18.18 (0.719)	15.57 (0.508)	12.18 (0.225)
tclustoracle	11.64 (0.276)	1.52 (0.215)	1.10 (0.112)	12.18 (0.225)
mclust	21.43 (0.068)	8.38 (0.162)	9.75 (0.103)	19.94 (0.061)
Kullback-Leibler Divergence				
RMBC	0.19 (0.003)	0.19 (0.004)	0.20 (0.004)	0.19 (0.003)
otrimle	0.34 (0.023)	0.61 (0.072)	0.3 (0.020)	3.95 (0.245)
tclust	0.17 (0.002)	1.49 (0.023)	0.59 (0.010)	0.71 (0.012)
tclustoracle	0.16 (0.002)	0.41 (0.016)	0.16 (0.004)	0.71 (0.012)
mclust	0.23 (0.001)	2.04 (0.009)	1.12 (0.004)	0.48 (0.008)
Sensitivity (%)				
RMBC	100 (0.000)	99.69 (0.055)	97.15 (0.142)	96.77 (0.153)
otrimle	99.40 (0.244)	99.76 (0.049)	98.64 (0.102)	98.12 (0.217)
tclust	99.00 (0.199)	72.56 (0.851)	68.18 (0.715)	93.76 (0.183)
tclustoracle	85.15 (0.768)	97.37 (0.255)	95.68 (0.217)	93.76 (0.183)
mclust	0.03 (0.022)	25.44 (0.591)	10.47 (0.341)	0.01 (0.009)
(1-Specificity) (%)				
RMBC	0.52 (0.019)	0.56 (0.031)	0.55 (0.025)	0.50 (0.017)
otrimle	1.88 (0.138)	3.10 (0.182)	1.94 (0.120)	9.90 (0.327)
tclust	0.61 (0.017)	0.02 (0.004)	0.04 (0.005)	0.15 (0.008)
tclustoracle	0.31 (0.011)	0.28 (0.019)	0.27 (0.016)	0.15 (0.008)
mclust	0.16 (0.007)	0.001 (0.001)	0.001 (0.001)	0.00 (0.000)

ducer, this pulse interacts with phytoplankton suspended in the water and produces an acoustic dispersion (scattering), which is recorded by an electronic acquisition device.

Description of the dataset

A filtering process of the signal is performed in a first stage. Portions of the signal belong to one of the two main cases:

- (a) Signals corresponding to the acoustic response of phytoplankton.
- (b) Signals corresponding to spurious dispersers, such as bubbles or particles in suspension, whose intensity is greater than in case (a).

To classify a signal in one of these two groups, biologists create a vector (X_1, X_2) defined as follows:

$$X_1 = \text{ratio of filtered to non-filtered signal power},$$

$$X_2 = \text{filtered signal power expressed in dB}.$$

The available data consists of 375 such measurements (see Figure 2). These data are particularly useful to compare robust procedures because 20% of these measurements are known to be outliers produced by a communication failure between the electronic device (digital oscilloscope) and the software for acquiring the acoustic signal. This failure occurs once every 5 microseconds, which allows the scientists to identify the outliers. The region $X_1 < 0.75$ and $X_2 > 20$ contains 93.3% of the outliers.

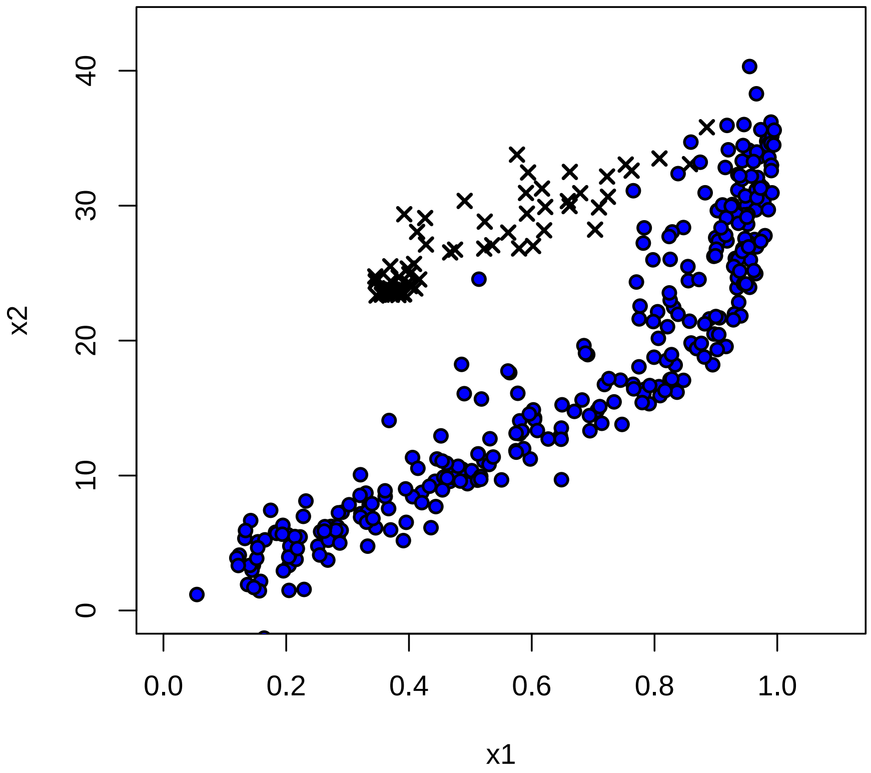

Figure 2: Original Data ($n = 375$). Circles and crosses correspond to regular observations and outliers, respectively.

Clustering analysis

Now we apply the four model-based clustering procedures compared in our simulation study to assess their ability to separate the observations of type (a) and (b). The performance of the estimators is evaluated by the MCR, the Kullback-Leibler divergence and the sensitivity.

Since in this example the outliers are known, we can remove these outliers and define the "true groups" as the partition produced by MCLUST (the classical procedure) applied to the clean data. We call this "the reference partition". Then, we apply the four clustering procedures to the whole data set including the outliers.

In the first panel of Figure 3, we show the clean data obtained after the true outliers identified by the biologists are removed. In this panel, we also show the allocation of the observations to the two clusters. By scientific prior knowledge, we know that measurements of the type (a), tend to have larger values of X_1 and X_2. Therefore, in the partition of the clean measurements produced by MCLUST, we identify the data points represented by triangles as measurements of type (a), and those represented by circles as measurements of type (b).

Table 6 shows the performance measures for the four considered procedures. Overall, RMBC has the best performance. Otrimle and TCLUST with oracle tuning parameter equal to 0.2 come second, except for KLD where they exhibit the worst performance. MCLUST, TCLUST and Otrimle have zero sensitivity because they fail to flag the true outliers. To reproduce Figure 2 and Table 6 of this example link to https://github.com/jdgonzalezwork/RMBC_Reproducibility.

Table 6: Performance of the compared model-based clustering procedures applied to the phytoplankton data. The reported values for TCLUST correspond to the choice $\alpha = 0.20$, the actual fraction of outliers in the data (ORACLE).

	RMBC	TCLUST	MCLUST	Otrimle
Misclassification Rate %	6.33	7.67	18.67	7.00
Kullback-Leibler Divergence	0.29	2.69	0.56	2.09
Sensitivity %	74.67	0.00	0.00	0.00

8. Conclusions

We present a general framework for the robust estimation of the parameters of a mixture model and show how this can be used to perform robust model-based clustering. Our proposal has some desirable features:

- The procedure is Fisher consistent under mild regularity assumptions.
- The procedure compares favorably with other robust and non-robust model-based clustering proposals in an extensive simulation study and a real data application.
- The procedure can be applied using an efficient computing algorithm implemented in the R package **RMBC**.
- The procedure's tuning parameters do not depend on the (usually unknown) fraction of outliers in the data.

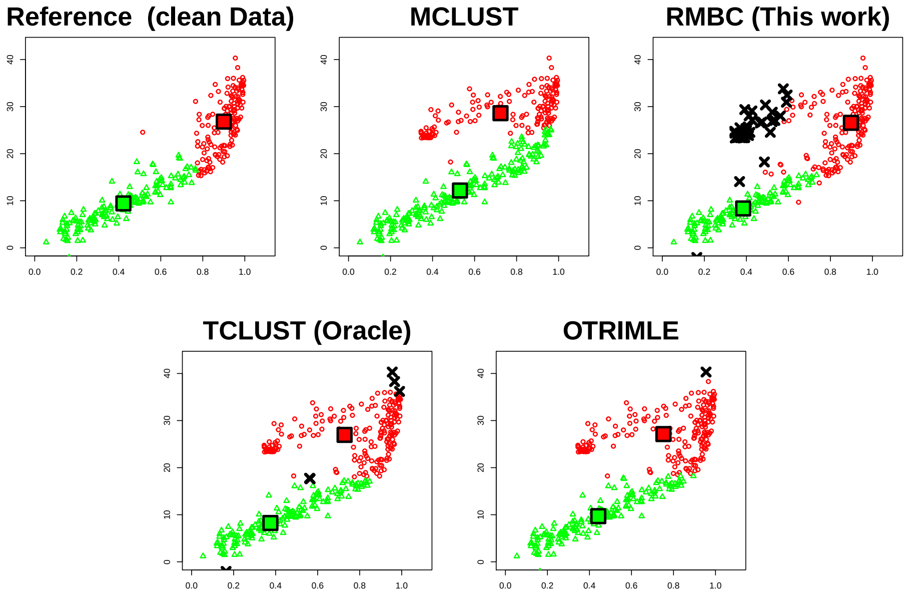

Figure 3: Results from the classic and robust model-based clustering procedures. The cluster of triangles and the cluster of circles correspond to observations of type (a) and (b), respectively. Crosses and squares represent detected outliers and cluster estimated centers, respectively.

Acknowledgements

Juan D. González was partially supported by Grant PIDDEF 02-20, Program of the Argentinian Ministry of Defense, Victor J. Yohai by grants 20020170100330BA from the Universidad de Buenos Aires and PICT-201-0377 from ANPYCT, Argentina, and Ruben H. Zamar by the Natural Sciences and Engineering Research Council of Canada. We would like to thank the anonymous reviewers for their careful review of our paper and their thoughtful comments that contributed to introduce important improvements in it.

References

Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, 49(3):803–821, ISSN: `0006341X, 15410420`, `http://www.jstor.org/stable/2532201`.

Blanc, S., Mosto, P., de Milou, M. E., and BenÃtez, C. (2004). AN ALTERNATIVE PROPOSAL: ACOUSTIC TECHNIQUES TO ASSES DETECTION AND MONITORING OF TOXIC ALGAL BLOOMS. *Acoustics Letters*, 68:54 – 59, ISSN: `0717-6538`, DOI: `10.4067/S0717-65382004000200009`, `http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0717-65382004000200009&nrm=iso`.

Blanc, S., Prario, I., Cinquini, M., Bos, P., and Tolivia, A. (2017). Ultrasonic scattering responses from phytoplankton: Measurements and modelling. In *2017 IEEE/OES*

Acoustics in Underwater Geosciences Symposium (RIO Acoustics), pages 1–9. DOI: 10.1109/RIOAcoustics.2017.8349702.

Bok, T. H., Paeng, D.-G., Kim, E., Na, J., and Kang, D. (2010). Ultrasound backscattered power from Cochlodinium polykrikoides, the main red tide species in the Southern Sea of Korea. *Journal of Plankton Research*, 32(4):503–514, ISSN: 0142-7873, DOI: 10.1093/plankt/fbq001, https://doi.org/10.1093/plankt/fbq001.

Cinquini, M., Bos, P., Prario, I., and Blanc, S. (2016). Advances on modelling, simulation and signal processing of ultrasonic scattering responses from phytoplankton cultures. *Proceedings of Meetings on Acoustics*, 28(1):070002, DOI: 10.1121/2.0000366, https://asa.scitation.org/doi/abs/10.1121/2.0000366.

Coretto, P. and Hennig, C. (2016). Robust improper maximum likelihood: Tuning, computation, and a comparison with other methods for robust gaussian clustering. *Journal of the American Statistical Association*, 111(516):1648–1659, DOI: 10.1080/01621459.2015.1100996, https://doi.org/10.1080/01621459.2015.1100996.

Coretto, P. and Hennig, C. (2017). Consistency, breakdown robustness, and algorithms for robust improper maximum likelihood clustering. *Journal of Machine Learning Research*, 18(142):1–39, http://jmlr.org/papers/v18/16-382.html.

Davies, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15(3):1269–1292, ISSN: 00905364, http://www.jstor.org/stable/2241828.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, ISSN: 00359246, http://www.jstor.org/stable/2984875.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, ISSN: 01621459, http://www.jstor.org/stable/3085676.

Fritz, H., García-Escudero, L. A., and Mayo-Iscar, A. (2012). tclust: An R package for a trimming approach to cluster analysis. *Journal of Statistical Software*, 47(12):1–26, DOI: 10.18637/jss.v047.i12, https://www.jstatsoft.org/index.php/jss/article/view/v047i12.

Gallegos, M. T. and Ritter, G. (2005). A robust method for cluster analysis. *The Annals of Statistics*, 33(1):347–380, ISSN: 00905364, http://www.jstor.org/stable/3448666.

García-Escudero, L., Gordaliza, A., Matrán, C., and Mayo, A. (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification*, 4:89–109, DOI: 10.1007/s11634-010-0064-5.

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, 36(3):1324–1345, ISSN: 00905364, http://www.jstor.org/stable/25464669.

Gonzalez, J. D., Yohai, V. J., and Zamar, R. H. (2019). Robust clustering using tau-scales. https://arxiv.org/abs/1906.08198.

Hershey, J. and Olsen, P. (2007). Approximating the Kullback Leibler divergence between gaussian mixture models. volume 4, pages IV–317. ISBN: 1-4244-0728-1, DOI: 10.1109/ICASSP.2007.366913.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, ISSN: 00034851, http://www.jstor.org/stable/2238020.

Maronna, R., Martin, R., Yohai, V., and Salibián-Barrera, M. (2019). *Robust Statistics: Theory and Methods (with R)*. Wiley Series in Probability and Statistics. Wiley, ISBN: 9781119214687, https://books.google.be/books?id=K5RxDwAAQBAJ.

Maronna, R. A. and Yohai, V. J. (2017). Robust and efficient estimation of multivariate scatter and location. *Computational Statistics & Data Analysis*, 109:64–75, ISSN: 0167-9473, DOI: https://doi.org/10.1016/j.csda.2016.11.006, https://www.sciencedirect.com/science/article/pii/S0167947316302705.

Scrucca, L., Fop, M., Murphy, T., and Raftery, A. (2016). mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *The R Journal*, 8:205–233, DOI: 10.32614/RJ-2016-021.

Yohai, V. J. and Zamar, R. H. (1997). Optimal locally robust M-estimates of regression. *Journal of Statistical Planning and Inference*, 64(2):309–323, ISSN: 0378-3758, DOI: https://doi.org/10.1016/S0378-3758(97)00040-2, https://www.sciencedirect.com/science/article/pii/S0378375897000402.

A. Proofs of Theorems

Proof of Theorem 1

We must show that for $1 \leq k \leq K$,

$$\alpha_{0k} = E_{H_0}(\widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}_0, \boldsymbol{\Theta}_0)) \tag{22}$$

and

$$\boldsymbol{\theta}_{0k} \ = \mathbf{g}\left(\frac{E_{H_0}\left(\widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}_0, \boldsymbol{\Theta}_0)\eta_r(\mathbf{x}, \boldsymbol{\theta}_{0k})\right)}{E_{H_0}(\widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}_0, \boldsymbol{\Theta}_0))}, ..., \frac{E_{H_0}\left(\widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}_0, \boldsymbol{\Theta}_0)\eta_h(\mathbf{x}, \boldsymbol{\theta}_{0k})\right)}{E_{H_0}(\widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}_0, \boldsymbol{\Theta}_0))}\right). \tag{23}$$

To prove (22) we write

$$\begin{aligned}
&E_{H_0}(\widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}_0, \boldsymbol{\Theta}_0)) \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}_0, \boldsymbol{\Theta}_0) \sum_{l=1}^{K} \alpha_{0l} f\ (\mathbf{x}, \boldsymbol{\theta}_{0l}) d\mathbf{x} \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{f\left(\mathbf{x}, \boldsymbol{\theta}_{0k}\right) \alpha_{0k}}{\sum_{l=1}^{K} \alpha_{0l} f\ (\mathbf{x}, \boldsymbol{\theta}_{0l})} \sum_{l=1}^{K} \alpha_{0l} f\ (\mathbf{x}, \boldsymbol{\theta}_{0l}) d\mathbf{x} \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f\left(\mathbf{x}, \boldsymbol{\theta}_{0k}\right) \alpha_{0k} d\mathbf{x} = \alpha_{0k}.
\end{aligned} \tag{24}$$

To prove (23), by (6), it is enough, to show that fixing $1 \leq r \leq h$ and $1 \leq k \leq K$ we have

$$\frac{E_{H_0}\left(\widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}_0, \boldsymbol{\Theta}_0)\eta_r(\mathbf{x}, \boldsymbol{\theta}_{0k})\right)}{E_{H_0}(\widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}_0, \boldsymbol{\Theta}_0))} = E_{F_{\boldsymbol{\theta}_{0k}}}\left(\eta_r(\mathbf{x}, \boldsymbol{\theta}_{0k})\right). \tag{25}$$

By (2) and (24) we get

$$\begin{aligned}
\frac{E_{H_0}\left(\widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}_0, \boldsymbol{\Theta}_0)\eta_r(\mathbf{x}, \boldsymbol{\theta}_{0k})\right)}{E_{H_0}(\widetilde{\alpha}_k(\mathbf{x}, \boldsymbol{\alpha}_0, \boldsymbol{\Theta}_0))} &= \frac{1}{\alpha_{0k}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{f\left(\mathbf{x}, \boldsymbol{\theta}_{0k}\right) \alpha_{0k} \eta_r(\mathbf{x}, \boldsymbol{\theta}_{0k}) \sum_{l=1}^{K} \alpha_{0l} f\ (\mathbf{x}, \boldsymbol{\theta}_{0l})}{\sum_{l=1}^{K} \alpha_{0l} f\ (\mathbf{x}, \boldsymbol{\theta}_{0l})} d\mathbf{x} \\
&= \frac{1}{\alpha_{0k}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f\left(\mathbf{x}, \boldsymbol{\theta}_{0k}\right) \alpha_{0k} \eta_r(\mathbf{x}, \boldsymbol{\theta}_{0k}) d\mathbf{x} \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f\left(\mathbf{x}, \boldsymbol{\theta}_{0k}\right) \eta_r(\mathbf{x}, \boldsymbol{\theta}_{0k}) d\mathbf{x} \\
&= E_{F_{\boldsymbol{\theta}_{0k}}}\left(\eta_r(\mathbf{x}, \boldsymbol{\theta}_{0k})\right),
\end{aligned}$$

proving (25). □

To show that the S estimator functional fits the general framework outlined in Section 2.1, we must show that this functional satisfies a system of fixed point equations. To obtain the estimating equations of the S functional, we consider a minimization problem which is equivalent to (13) but free of side constraints. We introduce the auxiliary functional A defined as

$$A(F, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{1/(2p)} \sigma(F, \boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

The following lemmas establish the relationship between the functionals S and A.

Lemma 1. *For all $\lambda > 0$,*

$$\sigma(F, \boldsymbol{\mu}, \lambda\boldsymbol{\Sigma}) = \sigma(F, \boldsymbol{\mu}, \boldsymbol{\Sigma})/\sqrt{\lambda} \tag{26}$$

and so

$$A(F, \boldsymbol{\mu}, \lambda\boldsymbol{\Sigma}) = A(F, \boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{27}$$

In other words, the functional $A(F, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ depends on the shape but not on the size of $\boldsymbol{\Sigma}$.

Proof

Note that for any $\lambda > 0$, $\sigma(F,\boldsymbol{\mu}, \lambda\boldsymbol{\Sigma})$ satisfies the equation

$$E_F\left(\rho\left(\frac{d(\mathbf{x}, \boldsymbol{\mu}, \lambda\boldsymbol{\Sigma})}{\sigma(F, \boldsymbol{\mu}, \lambda\boldsymbol{\Sigma})}\right)\right) = b.$$

Since $d(\mathbf{x}, \boldsymbol{\mu}, \lambda\boldsymbol{\Sigma}) = d(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})/\lambda^{1/2}$ we get

$$E_F\left(\rho\left(\frac{d(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\sqrt{\lambda}\sigma(F, \boldsymbol{\mu}, \lambda\boldsymbol{\Sigma})}\right)\right) = b.$$

Then, (26) is proved. Now we will show (27):

$$\begin{aligned} A(F, \boldsymbol{\mu}, \lambda\boldsymbol{\Sigma}) &= |\lambda\boldsymbol{\Sigma}|^{1/(2p)}\sigma(F, \boldsymbol{\mu}, \lambda\boldsymbol{\Sigma}) \\ &= (\lambda^p)^{1/(2p)}|\boldsymbol{\Sigma}|^{1/(2p)}\frac{1}{\sqrt{\lambda}}\sigma(F, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= |\boldsymbol{\Sigma}|^{1/(2p)}\sigma(F, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= A(F, \boldsymbol{\mu}, \boldsymbol{\Sigma}).\square \end{aligned} \tag{28}$$

Lemma 2. *Let $(\boldsymbol{\mu}(F), \boldsymbol{\Sigma}(F)) = \arg\min_{\boldsymbol{\mu},\boldsymbol{\Sigma}>0,\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})=1} |\boldsymbol{\Sigma}|$, where $\boldsymbol{\Sigma} > 0$ stands for $\boldsymbol{\Sigma}$ is positive definite. Then*

$$A(F, \boldsymbol{\mu}(F), \boldsymbol{\Sigma}(F)) = \min_{\boldsymbol{\mu},\boldsymbol{\Sigma}>0} A(F, \boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{29}$$

Moreover, let $(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ such that $A(F, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) = \min_{\boldsymbol{\mu},\boldsymbol{\Sigma}>0} A(F, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{\mu}(F)=\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}(F) = \sigma(F,\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)^2\boldsymbol{\Sigma}^*$.

Proof

We shall first show that the S functional $(\boldsymbol{\mu}(F), \boldsymbol{\Sigma}(F))$ also minimizes $A(F,\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with the constraint $\sigma(F, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = 1$. Since to minimize $|\boldsymbol{\Sigma}|$ is equivalent to minimizing $|\boldsymbol{\Sigma}|^{1/2p}$, we have

$$\begin{aligned} A(F, \boldsymbol{\mu}(F), \boldsymbol{\Sigma}(F)) &= \min_{\boldsymbol{\mu},\boldsymbol{\Sigma}>0,\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})=1} |\boldsymbol{\Sigma}|^{1/(2p)} = \min_{\boldsymbol{\mu},\boldsymbol{\Sigma}>0,\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})=1} |\boldsymbol{\Sigma}|^{1/(2p)}\sigma(F, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \min_{\boldsymbol{\mu},\boldsymbol{\Sigma}>0,\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})=1} A(F, \boldsymbol{\mu}, \boldsymbol{\Sigma}), \end{aligned} \tag{30}$$

Let now $(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ be such that $A(F, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) = \min_{\boldsymbol{\mu},\boldsymbol{\Sigma}>\mathbf{0}} A(F, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. By Lemma 1, we have $\sigma^2(F, \boldsymbol{\mu}^*, \sigma^2(F, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)\boldsymbol{\Sigma}^*) = 1$ and $A(F, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) = A(F, \boldsymbol{\mu}^*, \ \sigma^2(F, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)\boldsymbol{\Sigma}^*)$, and then by (30)

$$\begin{aligned} A(F, \boldsymbol{\mu}^*, \ \sigma^2(F, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)\boldsymbol{\Sigma}^*) &= \min_{\boldsymbol{\mu}.\boldsymbol{\Sigma}>\mathbf{0}} A(F, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \min_{\boldsymbol{\mu},\boldsymbol{\Sigma}>\mathbf{0},\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})=1} A(F, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \min_{\boldsymbol{\mu},\boldsymbol{\Sigma}>0,\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})=1} |\boldsymbol{\Sigma}|^{1/(2p)}. \end{aligned}$$

Then $(\boldsymbol{\mu}(F), \boldsymbol{\Sigma}(F)) = (\boldsymbol{\mu}^*, \sigma^2(F, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)\boldsymbol{\Sigma}^*)$, and this proves the Lemma. □

Proof of Theorem 2

We start by obtaining fixed point equations that the critical points $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ of $A(F,\boldsymbol{\mu}, \boldsymbol{\Sigma})$ satisfy. Note that the equation

$$\frac{\partial A(F, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = 0$$

is equivalent to

$$\frac{\partial \sigma(F, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = 0.$$

Since

$$\frac{\partial d^2(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = -2\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}),$$

it follows that

$$\frac{\partial d(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = \frac{-\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})}{d(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}.$$

Then

$$\frac{\partial (d(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})/\sigma(F, \boldsymbol{\mu}, \boldsymbol{\Sigma}))}{\partial \boldsymbol{\mu}} = \frac{\frac{-\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}\sigma(F, \boldsymbol{\mu}, \boldsymbol{\Sigma}) - d(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\frac{\partial \sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}}}{\sigma^2(F, \boldsymbol{\mu}, \boldsymbol{\Sigma})}.$$

Implicit differentiation of $\sigma(F,\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\mu}$ gives

$$E_F\left(\psi\left(\frac{d(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\sigma(F, \boldsymbol{\mu}, \boldsymbol{\Sigma})}\right)\frac{\frac{-\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}\sigma(F, \boldsymbol{\mu}, \boldsymbol{\Sigma}) - d(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\frac{\partial \sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}}}{\sigma^2(F, \boldsymbol{\mu}, \boldsymbol{\Sigma})}\right) = 0,$$

where $\psi = \rho'$. Putting $\partial\sigma(F,\boldsymbol{\mu}, \boldsymbol{\Sigma})/\partial\boldsymbol{\mu}= 0$ and multiplying both sides by $-\sigma(F,\boldsymbol{\mu}, \boldsymbol{\Sigma})^2\boldsymbol{\Sigma}$ we get

$$E_F\left(\psi\left(\frac{d(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\sigma(F, \boldsymbol{\mu}, \boldsymbol{\Sigma})}\right)\frac{(\mathbf{x} - \boldsymbol{\mu})}{d(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}\sigma(F, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right) = 0,$$

and

$$E_F\left(\frac{\psi\left(\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)}{\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}}(\mathbf{x} - \boldsymbol{\mu})\right) = 0.$$

Setting $W(t) = \psi(t)/t$ we get

$$E_F\left(W\left(\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)(\mathbf{x}-\boldsymbol{\mu})\right)=0,$$

or equivalently

$$\boldsymbol{\mu}=\frac{E_F\left(W\left(\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)\mathbf{x}\right)}{E_F\left(W\left(\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)\right)}. \tag{31}$$

We now differentiate $A(F,\boldsymbol{\mu},\boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}$. We will use the following results

$$\frac{\partial|\boldsymbol{\Sigma}|}{\partial\boldsymbol{\Sigma}}=|\boldsymbol{\Sigma}|\boldsymbol{\Sigma}^{-1} \tag{32}$$

and

$$\frac{\partial(\mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{b})}{\partial\boldsymbol{\Sigma}}=-\boldsymbol{\Sigma}^{-1}\mathbf{a}\mathbf{b}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}. \tag{33}$$

Then,

$$\frac{\partial d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial\boldsymbol{\Sigma}}=\frac{-\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}}{2d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}.$$

Differentiating $\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}$ we get

$$E_F\left(\psi\left(\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)\frac{\frac{-\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}}{2d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})-d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})\frac{\partial\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial\boldsymbol{\Sigma}}}{\sigma^2(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)=0. \tag{34}$$

Besides, differentiating $\log A(F,\boldsymbol{\mu},\boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}$ we get

$$\frac{\frac{\partial\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial\boldsymbol{\Sigma}}}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}+\frac{\boldsymbol{\Sigma}^{-1}}{2p}=0,$$

and therefore

$$\frac{\partial\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial\boldsymbol{\Sigma}}=-\frac{1}{2p}\boldsymbol{\Sigma}^{-1}\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma}). \tag{35}$$

Therefore replacing $\partial\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})/\partial\boldsymbol{\Sigma}$ in (34) we get

$$E_F\left(\psi\left(\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)\frac{\frac{-\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}}{2d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})+d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})\frac{1}{2p}\boldsymbol{\Sigma}^{-1}\ \sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma^2(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)=0$$

and

$$E_F\left(\psi\left(\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)\left(\frac{-\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}}{2d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}+\frac{1}{2p}\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\boldsymbol{\Sigma}^{-1}\ \right)\right)=0.$$

Pre and post multiplying by $\boldsymbol{\Sigma}$ we obtain

$$E_F\left(\psi\left(\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)\left(\frac{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}}{2d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}-\frac{1}{2p}\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\boldsymbol{\Sigma}\right)\ \right)=0,$$

$$E_F\left(\psi\left(\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)\frac{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}}{2d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)=E_F\left(\psi\left(\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)\frac{1}{2p}\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)\boldsymbol{\Sigma}$$

$$\frac{p}{\sigma^2(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}E_F\left(\psi\left(\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)\frac{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}}{\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}}\right)=E_F\left(\psi\left(\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)\boldsymbol{\Sigma}.$$

Then, putting

$$c_0=\frac{p}{\sigma^2(F,\boldsymbol{\mu},\boldsymbol{\Sigma})E_F\left(\psi\left(\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)}$$

we get

$$\boldsymbol{\Sigma}=c_0E_F\left(W\left(\frac{d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})}\right)(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\right). \tag{36}$$

Since by Lemma 1, $A(F,\boldsymbol{\mu},c\boldsymbol{\Sigma})=A(F,\boldsymbol{\mu},\boldsymbol{\Sigma})$, c_0 may be changed by any other scalar. Then, we can put in (36)

$$c_0=\frac{1}{E_F\left(W\left(\frac{d(\mathbf{x},\boldsymbol{\mu}^*(F),\boldsymbol{\Sigma}^*(F))}{\sigma(F,\boldsymbol{\mu}^*(F),\boldsymbol{\Sigma}^*(F))}\right)\right)},$$

and $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ will still be a critical point. Then, (31) and (36) imply that there exists $(\boldsymbol{\mu}^*(F),\boldsymbol{\Sigma}^*(F))$ which minimizes $A(F,\boldsymbol{\mu},\boldsymbol{\Sigma})$ and satisfies

$$\boldsymbol{\mu}^*(F)=\frac{E_F\left(W\left(\frac{d(\mathbf{x},\boldsymbol{\mu}^*(F),\boldsymbol{\Sigma}^*(F))}{\sigma(F,\boldsymbol{\mu}^*(F),\boldsymbol{\Sigma}^*(F))}\right)\mathbf{x}\right)}{E_F\left(W\left(\frac{d(\mathbf{x},\boldsymbol{\mu}^*(F),\boldsymbol{\Sigma}^*(F))}{\sigma(F,\boldsymbol{\mu}^*(F),\boldsymbol{\Sigma}^*(F))}\right)\right)}, \tag{37}$$

$$\boldsymbol{\Sigma}^*(F)=\frac{E_F\left(W\left(\frac{d(\mathbf{x},\boldsymbol{\mu}^*(F),\boldsymbol{\Sigma}^*(F))}{\sigma(F,\boldsymbol{\mu}^*(F),\boldsymbol{\Sigma}^*(F))}\right)(\mathbf{x}-\boldsymbol{\mu}^*(F))(\mathbf{x}-\boldsymbol{\mu}^*(F))^{\mathrm{T}}\right)}{E_F\left(W\left(\frac{d(\mathbf{x},\boldsymbol{\mu}^*(F),\boldsymbol{\Sigma}^*(F))}{\sigma(F,\boldsymbol{\mu}^*(F),\boldsymbol{\Sigma}^*(F))}\right)\right)}. \tag{38}$$

Then, by Lemma 2, the values $(\boldsymbol{\mu}(F),\boldsymbol{\Sigma}(F))$ defined as

$$\boldsymbol{\mu}(\boldsymbol{F})=\boldsymbol{\mu}^*(F) \tag{39}$$

$$\boldsymbol{\Sigma}(F)=\sigma(F,\boldsymbol{\mu}^*(F),\boldsymbol{\Sigma}^*(F))^2\boldsymbol{\Sigma}^*(F) \tag{40}$$

minimize $|\boldsymbol{\Sigma}|$ subject to $\sigma(F,\boldsymbol{\mu},\boldsymbol{\Sigma})=1$. By Lemma 1

$$\frac{d(\mathbf{x},\boldsymbol{\mu}^*(F),\boldsymbol{\Sigma}^*(F))}{\sigma(F,\boldsymbol{\mu}^*(F),\boldsymbol{\Sigma}^*(F))}=d(\mathbf{x},\boldsymbol{\mu}(F),\boldsymbol{\Sigma}(F)).$$

and therefore (37) and (38) can be written as

$$\boldsymbol{\mu}(\boldsymbol{F})=\frac{E_F\left(W\left(d(\mathbf{x},\boldsymbol{\mu}(\boldsymbol{F}),\boldsymbol{\Sigma}(F))\right)\mathbf{x}\right)}{E_F\left(W\left(d(\mathbf{x},\boldsymbol{\mu}(\boldsymbol{F}),\boldsymbol{\Sigma}(F))\right)\right)}, \tag{41}$$

$$\boldsymbol{\Sigma}^*(F)=\frac{E_F\left(W\left(d(\mathbf{x},\boldsymbol{\mu}(\boldsymbol{F}),\boldsymbol{\Sigma}(F)\right)(\mathbf{x}-\boldsymbol{\mu}(F))(\mathbf{x}-\boldsymbol{\mu}(F))^{\mathrm{T}}\right)}{E_F\left(W\left(d(\mathbf{x},\boldsymbol{\mu}(F),\boldsymbol{\Sigma}(F))\right)\right)}. \tag{42}$$

Define $s^*(F)=\sigma(F,\boldsymbol{\mu}(F),\boldsymbol{\Sigma}^*(F))$, then by (10)

$$E_F(\rho(d(\mathbf{x},\boldsymbol{\mu}(F),\boldsymbol{\Sigma}^*(F))/s^*(F)))=b,$$

and multiplying both sides by $2s^*(F)$

$$s^*(F) = E_F((1/b)s^*(F)\rho(d(\mathbf{x}, \boldsymbol{\mu}(F), \boldsymbol{\Sigma}^*(F))/s^*(F))). \tag{43}$$

Finally, by (40)

$$\boldsymbol{\Sigma}(F) = s^*(F)^2\boldsymbol{\Sigma}^*(F). \tag{44}$$

Then, (41)-(44) prove Theorem 2.□

Affiliation:

Ruben H. Zamar
University of British Columbia
Department of Statistics
Earth Sciences Building, Room 3182
2207 Main Mall
Vancouver, BC Canada V6T 1Z4
E-mail: ruben@stat.ubc.ca

Journal of Data Science, Statistics, and Visualisation https://jdssv.org/
published by the International Association for Statistical Computing
http://iasc-isi.org/

ISSN 2773-0689
October 2022, Volume II, Issue VI *Submitted:* 2021-06-20
doi:10.52933/jdssv.v2i6.47 *Accepted:* 2022-04-06