METRON (2021) 79:225–240
https://doi.org/10.1007/s40300-021-00214-9

Robust estimation for multivariate wrapped models

Giovanni Saraceno[1] · **Claudio Agostinelli**[1] · **Luca Greco**[2]

Received: 15 October 2020 / Accepted: 8 June 2021 / Published online: 26 June 2021

Abstract
A weighted likelihood technique for robust estimation of multivariate Wrapped distributions of data points scattered on a $p-$dimensional torus is proposed. The occurrence of outliers in the sample at hand can badly compromise inference for standard techniques such as maximum likelihood method. Therefore, there is the need to handle such model inadequacies in the fitting process by a robust technique and an effective downweighting of observations not following the assumed model. Furthermore, the employ of a robust method could help in situations of hidden and unexpected substructures in the data. Here, it is suggested to build a set of data-dependent weights based on the Pearson residuals and solve the corresponding weighted likelihood estimating equations. In particular, robust estimation is carried out by using a Classification EM algorithm whose M-step is enhanced by the computation of weights based on current parameters' values. The finite sample behavior of the proposed method has been investigated by a Monte Carlo numerical study and real data examples.

Keyword CEM algorithm · Multivariate wrapped distributions · Pearson residuals · Robust estimators · Torus · Weighted likelihood

1 Introduction

Multivariate circular observations arise commonly in all those fields where a quantity of interest is measured as a direction or when instruments such as compasses, protractors, weather vanes, sextants or theodolites are used [24]. Circular (or directional) data can be

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40300-021-00214-9.

✉ Giovanni Saraceno
giovanni.saraceno@unitn.it

Claudio Agostinelli
claudio.agostinelli@unitn.it

Luca Greco
luca.greco@unisannio.it

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

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

seen as points on the unit circle and represented by angles, provided that an initial direction and orientation of the circle have been chosen.

These data might be successfully modeled by using appropriate wrapped distributions such, e.g, the Wrapped Normal or the Wrapped Cauchy on the unit circle. The reader is pointed to [9,19,25] for modeling and inferential issues on circular data. Wrapping can be explained as the geometric translation of a distribution with support on $\mathbb{R}$ to a space defined on a *circular* object, e.g., a unit circle [25].

When data come in a multivariate setting, we might extend the univariate wrapping around the circle by using a component-wise wrapping of multivariate distributions around a $p-$dimensional torus. Let

$$\mathcal{M} = \left\{ m(\boldsymbol{x}; \Omega) = c_p |\Sigma|^{-\frac{1}{2}} h(d(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)), \Omega = (\boldsymbol{\mu}, \Sigma), \boldsymbol{\mu} \in \mathbb{R}^p, \Sigma \in PDS(p) \right\}$$

be the elliptically symmetric family of distributions where $PDS(p)$ is the set of all positive-definite symmetric $p \times p$ matrices, c_p is a normalization constant depending on $p > 1$, $h(\cdot)$ is a non-negative scalar function, called density generating function, and $d(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \left[(\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right]^{1/2}$ is the Mahalanobis distance. For example, the multivariate Normal distribution and the multivariate Student t_ν distribution belong to this family choosing $h(d) = \exp(-d^2/2)$ and $h(d) = (1 + d^2/\nu)^{-(p+\nu)/2}$, respectively, as density generating function. As particular case, the multivariate Cauchy distribution can be obtained for $\nu = 1$. Let $\boldsymbol{X}$ be a multivariate random variable whose distribution belongs to the family of elliptically symmetric distributions. Then, the distribution of $\boldsymbol{Y} = \boldsymbol{X} \bmod 2\pi$ is

$$M^\circ(\boldsymbol{y}) = \sum_{\boldsymbol{j} \in \mathbb{Z}^p} [M(\boldsymbol{y} + 2\pi \boldsymbol{j}; \Omega) - M(2\pi \boldsymbol{j}; \Omega)],$$

with density function

$$m^\circ(\boldsymbol{y}) = \sum_{\boldsymbol{j} \in \mathbb{Z}^p} m(\boldsymbol{y} + 2\pi \boldsymbol{j}; \Omega),$$

$\boldsymbol{y} \in (0, 2\pi]^p$, $\Omega = (\boldsymbol{\mu}, \Sigma)$, where $M(\cdot)$ and $m(\cdot)$ are the distribution and density function of $\boldsymbol{X}$, respectively, and the modulus operator mod is applied component-wise. As a special case, let $\boldsymbol{X}$ be multivariate Normal, i.e. $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$. Then, the distribution of $\boldsymbol{Y} = \boldsymbol{X} \bmod 2\pi$ is Wrapped Normal and denoted as $WN_p(\boldsymbol{\mu}, \Sigma)$. An appealing property of the Normal distribution that carries over to the Wrapped Normal is its closure with respect to convolution [7,19]. This property will be particularly relevant in the implementation of our methodology.

Given an i.i.d. sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ of size n from $\boldsymbol{Y}$ on the p-torus, likelihood based inference about the parameters of Wrapped distributions can be trapped in numerical and computational hindrances since the log-likelihood function

$$\ell(\Omega) = \sum_{i=1}^{n} \log \left[\sum_{\boldsymbol{j} \in \mathbb{Z}^p} m(\boldsymbol{y}_i + 2\pi \boldsymbol{j}; \Omega) \right],$$

involves the evaluation of an infinite series. [2] proposed an Iterative Reweighted Maximum Likelihood Estimating Equations algorithm in the univariate setting, that is available in the `R` package `circular` [4]. Algorithms based on the Expectation-Maximization (EM) method have been used by [15] for parameter estimation in autoregressive models of Wrapped Normal distributions and by [10], [32] and [14] in a Bayesian framework according to a data augmentation approach to estimate the missing unobserved wrapping coefficients. An innovative estimation strategy based on EM and Classification EM algorithms has been discussed

in [28]. In order to perform maximum likelihood estimation, the wrapping coefficients are treated as latent variables.

We can think of $\boldsymbol{y}_i = \boldsymbol{x}_i \mod 2\pi$ where $\boldsymbol{x}_i$ is a sample from a random variable whose distribution belongs to the elliptically symmetric family of distributions. The EM algorithm works with the complete log-likelihood function given by

$$\ell_C(\Omega) = \sum_{i=1}^{n} \sum_{\boldsymbol{j} \in \mathbb{Z}^p} v_{i\boldsymbol{j}} \log m(\boldsymbol{y}_i + 2\pi \boldsymbol{j}; \Omega), \tag{1}$$

that is characterized by the missing unobserved wrapping coefficients $\boldsymbol{j}$ and $v_{i\boldsymbol{j}}$ is an indicator of the ith unit having the $\boldsymbol{j}$ vector as wrapping coefficients. The EM algorithm iterates between an Expectation (E) step and a Maximization (M) step. In the E-step, the conditional expectation of (1) is obtained by estimating the $v_{i\boldsymbol{j}}$ with the posterior probability that $\boldsymbol{y}_i$ has $\boldsymbol{j}$ as wrapping coefficient based on current parameters' values, i.e.

$$v_{i\boldsymbol{j}} = \frac{m(\boldsymbol{y}_i + 2\pi \boldsymbol{j}; \Omega)}{\sum_{\boldsymbol{b} \in \mathbb{Z}^p} m(\boldsymbol{y}_i + 2\pi \boldsymbol{b}; \Omega)}, \qquad \boldsymbol{j} \in \mathbb{Z}^p, \quad i = 1, \ldots, n.$$

In the M-step, the conditional expectation of (1) is maximized with respect to Ω. The reader is pointed to [28] for computational details about such maximization problem for the multivariate Wrapped Normal distribution.

An alternative estimation strategy is based on the CEM-type algorithm. The substantial difference is that the E-step is followed by a C-step (where C stands for classification) in which $v_{i\boldsymbol{j}}$ is estimated as either 0 or 1 and so that each observation $\boldsymbol{y}_i$ is associated to the most likely wrapping coefficients $\boldsymbol{j}_i$ with $\boldsymbol{j}_i = \arg\max_{\boldsymbol{b} \in \mathbb{Z}^p} v_{i\boldsymbol{b}}$.

When the sample data is contaminated by the occurrence of outliers, it is well known that maximum likelihood estimation, also achieved through the implementation of the EM or CEM algorithm, is likely to lead to unreliable results [13]. Then, there is the need for a suitable robust procedure providing protection against those unexpected anomalous values. There have been few attempts to deal with outliers in circular data analysis for univariate distributions, mainly focused on the Von Mises distribution [2,20,21,33]. On the contrary, the robust technique proposed here is based on multivariate Wrapped distributions and, to the best of our knowledge, there are no competing techniques of robust estimation for multivariate models.

An attractive solution to develop a robust estimation algorithm for multivariate wrapped distributions would be to modify the likelihood equations in the M-step. Such a modification could be achieved by the introduction of a set of weights aimed to bound the effect of those observations deviating from the assumed model. Here, it is suggested to evaluate weights according to the weighted likelihood methodology ([26]). Weighted likelihood is an appealing robust technique for estimation and testing [5]. The methodology leads to a robust fit and gives the chance to detect outliers and possible substructures in the data. Furthermore, the weighted likelihood methodology works in a very satisfactory fashion when combined with the EM and CEM algorithms, as in the case of mixture models [17,18].

The remainder of the paper is organized as follows. Section 2 gives brief but necessary preliminaries on weighted likelihood. The weighted CEM algorithm for robust fitting of multivariate Wrapped models on data on a $p-$dimensional torus is described in Sect. 3, while some theoretical properties are discussed in Sect. 3.1. Section 4 reports the results of some numerical studies, whereas a real data example is discussed in Sect. 5. Concluding remarks end the paper.

2 Preliminaries on weighted likelihood

Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be a random sample of size n drawn from a r.v. $\boldsymbol{Y}$ with distribution function F and probability (density) function f. Let $\mathcal{M} = \{M(\boldsymbol{y}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d, d \geq 1, \boldsymbol{y} \in \mathcal{Y}\}$ be the assumed parametric model, with corresponding density $m(\boldsymbol{y}; \boldsymbol{\theta})$, and $\hat{F}_n$ the empirical distribution function. Assume that the support of M is the same as that of F and independent of $\boldsymbol{\theta}$. A measure of the agreement between the *true* and assumed model is provided by the Pearson residual function $\delta(\boldsymbol{y})$, with $\delta(\boldsymbol{y}) \in [-1, +\infty)$, [23,26], defined as

$$\delta(\boldsymbol{y}) = \delta(\boldsymbol{y}; \boldsymbol{\theta}, F) = \frac{f(\boldsymbol{y})}{m(\boldsymbol{y}; \boldsymbol{\theta})} - 1 \,. \tag{2}$$

The finite sample counterpart of (2) can be obtained as

$$\delta_n(\boldsymbol{y}) = \delta(\boldsymbol{y}; \boldsymbol{\theta}, \hat{F}_n) = \frac{\hat{f}_n(\boldsymbol{y})}{m(\boldsymbol{y}; \boldsymbol{\theta})} - 1 \,, \tag{3}$$

where $\hat{f}_n(\boldsymbol{y})$ is a consistent estimate of the true density $f(\boldsymbol{y})$. In discrete families of distributions, $\hat{f}_n(\boldsymbol{y})$ can be driven by the observed relative frequencies [23], whereas in continuous models one could consider a non parametric density estimate based on the kernel function $k(\boldsymbol{y}; \boldsymbol{t}, h)$, that is

$$\hat{f}_n(\boldsymbol{y}) = \int_{\mathcal{Y}} k(\boldsymbol{y}; \boldsymbol{t}, h) d\hat{F}_n(\boldsymbol{t}) \,. \tag{4}$$

Moreover, in the continuous case, the model density in (3) can be replaced by a smoothed model density, obtained by using the same kernel involved in non-parametric density estimation [8,26], that is

$$\hat{m}(\boldsymbol{y}; \boldsymbol{\theta}) = \int_{\mathcal{Y}} k(\boldsymbol{y}; \boldsymbol{t}, h) m(\boldsymbol{t}; \boldsymbol{\theta}) \, d\boldsymbol{t}$$

leading to

$$\delta_n(\boldsymbol{y}) = \delta(\boldsymbol{y}; \boldsymbol{\theta}, \hat{F}_n) = \frac{\hat{f}_n(\boldsymbol{y})}{\hat{m}(\boldsymbol{y}; \boldsymbol{\theta})} - 1 \,. \tag{5}$$

By smoothing the model, the Pearson residuals in (5) converge to zero with probability one for every $\boldsymbol{y}$ under the assumed model and it is not required that the kernel bandwidth h goes to zero as the sample size n increases. Large values of the Pearson residual function correspond to regions of the support $\mathcal{Y}$ where the model fits the data poorly, meaning that the observation is unlikely to occur under the assumed model. The reader is pointed to [3,8,26] and references therein for more details.

Observations leading to large Pearson residuals in (5) are supposed to be down-weighted. Then, a weight in the interval [0, 1] is attached to each data point, that is computed accordingly to the following weight function

$$w(\delta(\boldsymbol{y})) = \min\left\{1, \frac{[A(\delta(\boldsymbol{y})) + 1]^+}{\delta(\boldsymbol{y}) + 1}\right\} \,, \tag{6}$$

where $[\cdot]^+$ denotes the positive part and $A(\delta)$ is the Residual Adjustment Function (RAF, [8,23,29]). The weights $w(\delta_n(\boldsymbol{y}))$ are meant to be small for those data points that are in disagreement with the assumed model. Actually, the RAF plays the role to bound the effect of large Pearson residuals on the fitting procedure. $A(\cdot)$ is an increasing, twice differentiable, function in $[-1, \infty)$, such that $A(0) = 0$ and $A'(0) = 1$.

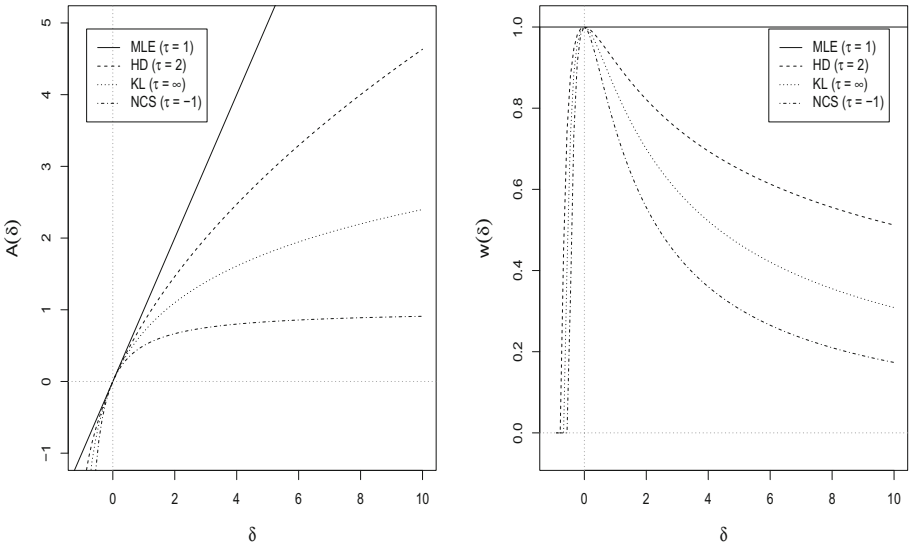

Fig. 1 RAF from Power Divergence Measure (left) and corresponding weight function (right) for different values of τ

The weight function (6) might be based on the families of RAF stemming from the Symmetric Chi-squared divergence [26], the Generalized Kullback-Leibler divergence [30]

$$A_{gkl}(\delta, \tau) = \frac{\log(\tau\delta + 1)}{\tau}, \ 0 \leq \tau \leq 1; \tag{7}$$

or the Power Divergence Measure [11,12]

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

In the latter case, special cases are maximum likelihood (ML, $\tau = 1$, as the weights become all equal to one), Hellinger distance (HD, $\tau = 2$), Kullback–Leibler divergence (KL, $\tau =\!\!\rightarrow \infty$) and Neyma-s Chi-Square (NCS, $\tau = -1$). The RAF stemming from the Power Divergence Measure are illustrated in the left panel of Fig. 1. The resulting weight function (6) is unimodal and declines smoothly to zero as $\delta(\boldsymbol{y}) \to -1$ or $\delta(\boldsymbol{y}) \to \infty$, as displayed in the right panel of Fig. 1. See also [29] for further ways of defining RAFs.

According to the chosen RAF, robust estimation can be based on a Weighted Likelihood Estimating Equation (WLEE), defined as

$$\sum_{i=1}^{n} w(\delta_n(\boldsymbol{y}_i); \boldsymbol{\theta}, \hat{F}_n) s(\boldsymbol{y}_i; \boldsymbol{\theta}) = 0 \,, \tag{8}$$

where $s(\boldsymbol{y}_i; \boldsymbol{\theta})$ is the individual contribution to the score function. Therefore, weighted likelihood estimation can be thought as a root solving problem. Finding the solution of (8) requires an iterative weighting algorithm.

Remark 1 Several functions could be defined to bound the effect of large Pearson residuals rather than using the RAF. However, the use of the RAF is strictly connected to weighted likelihood estimation. First, this choice is motivated by historical reasons, in the spirit of the

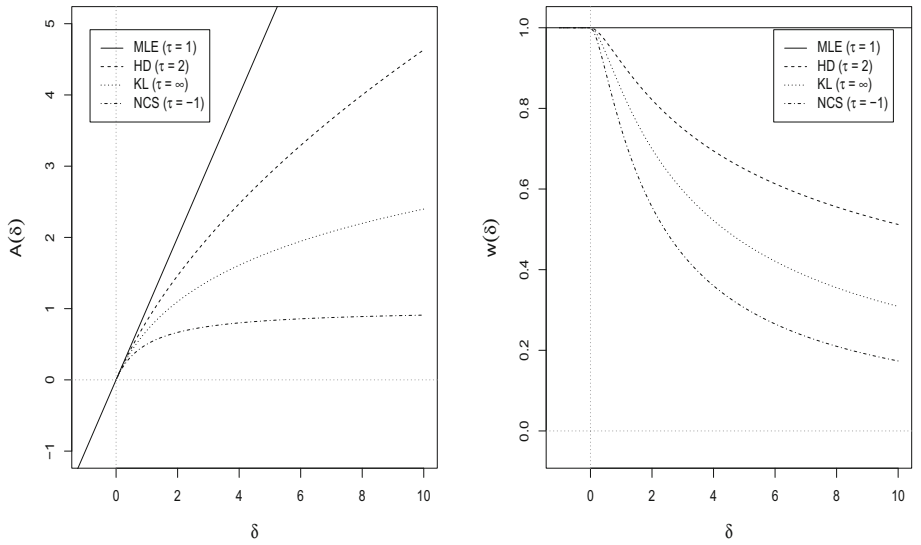

Fig. 2 Modified RAF from Power Divergence Measure (left) and corresponding weight function (right) for different values of τ

work by [23,26], among others. Then, the special role played by the RAF is justified in light of the connection between weighted likelihood estimation and minimum disparity estimation. Actually, the RAF arises naturally from a minimum disparity estimation problem, although the construction of the WLEE does not depend on the availability of an objective function [3].

Remark 2 As pointed out in [29], values of the Pearson Residuals in the interval $(0, \infty)$ are related to outliers, while values in $(-1, 0)$ to inliers. RAFs can act on this last interval in opposite ways. For instance, the RAF related to the HD leads to downweighting while the Negative Exponential Disparity (see [23]) leads to an upweigthing of the observations. Since inliers represent a minor issue for data in p–dimensional torus, we decided to modify our RAFs in the interval $(-1, 0)$ setting them equal to the identity function. The plots of the modified RAFs together with the corresponding weights are reported in Fig. 2. We used these RAFs in our simulations and examples.

The corresponding weighted likelihood estimator $\hat{\boldsymbol{\theta}}^{w}$ (WLE) is consistent, asymptotically normal and fully efficient at the assumed model, under some general regularity conditions pertaining the model, the kernel and the weight function [3,5,26]. Its robustness properties have been established in [23] in connection with minimum disparity problems. It is worth remarking that under very standard conditions, one can build a simple WLEE matching a minimum disparity objective function, hence inheriting its robustness properties.

In finite samples, the robustness/efficiency trade-off of weighted likelihood estimation can be tuned by varying the smoothing parameter h in Eq. (4). Large values of h lead to Pearson residuals all close to zero and weights all close to one and, hence, large efficiency, since $\hat{f}_n(\boldsymbol{y})$ is stochastically close to the postulated model. On the other hand, small values of h make $\hat{f}_n(\boldsymbol{y})$ more sensitive to the occurrence of outliers and the Pearson residuals become large for those data points that are in disagreement with the model. On the contrary, the shape of the kernel function $k(\boldsymbol{y}; \boldsymbol{t}, h)$ has a very limited effect.

For what concerns the tasks of testing and setting confidence regions, a weighted likelihood counterpart of the classical likelihood ratio test, and its asymptotically equivalent Wald and Score versions, can be established. Note that, all share the standard asymptotic distribution at the true model, according to the results stated in [5], that is

$$\Lambda(\boldsymbol{\theta}) = 2\sum_{i=1}^{n} w_i \left[\ell(\hat{\boldsymbol{\theta}}^w; \boldsymbol{y}_i) - \ell(\boldsymbol{\theta}; \boldsymbol{y}_i)\right] \xrightarrow{p} \chi^2_p \,,$$

with $w_i = w(\delta_n(\boldsymbol{y}_i); \hat{\boldsymbol{\theta}}^w, \hat{F}_n)$. Profile tests can be obtained as well.

3 A weighted CEM algorithm

As previously stated in the "Introduction" [28] provided effective iterative algorithms to fit a multivariate Wrapped distribution on the p-dimensional torus. Here, robust estimation is achieved by a suitable modification of their CEM algorithm, consisting in a weighting step before performing the M-step, in which data-dependent weights are evaluated according to (6) yielding a WLEE (8) to be solved in the M-step.

In the special case of the multivariate Wrapped Normal distribution, the construction of Pearson residuals in (5) involves a multivariate Wrapped Normal kernel with covariance matrix $h\Lambda$. Since the family of multivariate Wrapped Normal is closed under convolution, then the smoothed model density is still Wrapped Normal with covariance matrix $\Sigma + h\Lambda$. Here, we set $\Lambda = \Sigma$ so that h can be a constant independent of the variance-covariance structure of the data. The problem becomes more challenging if other elliptically symmetric distributions are considered, since smoothed densities require numerical evaluations.

The weighted CEM algorithm is structured as follows:

0 *Initialization.* Starting values can be obtained by maximum likelihood estimation evaluated over a randomly chosen subset. The subsample size is expected to be as small as possible in order to increase the probability to get an outliers' free initial subset but large enough to guarantee estimation of the unknown parameters. A starting solution for $\boldsymbol{\mu}$ can be obtained by the circular mean, whereas the diagonal entries of Σ can be initialized as $-2\log(\hat{\rho}_r)$, where $\hat{\rho}_r$ is the sample mean resultant length and the off-diagonal elements by $\rho_c(\boldsymbol{y}_r, \boldsymbol{y}_s)\sigma^{(0)}_{rr}\sigma^{(0)}_{ss}$ $(r \neq s)$, where $\rho_c(\boldsymbol{y}_r, \boldsymbol{y}_s)$ is the circular correlation coefficient, $r = 1, 2, \ldots, p$ and $s = 1, 2, \ldots, p$, see [19] pag.176,equation8.2.2. In order to avoid the algorithm to be dependent on initial values, a simple and common strategy is to run the algorithm from a number of starting values using the bootstrap root searching approach as in [26]. A criterion to choose among different solutions will be illustrated in Sect. 5.

1. *E-step*. Based on current parameters' values, first evaluate posterior probabilities

$$v_{i\boldsymbol{j}} = \frac{m(\boldsymbol{y}_i + 2\pi\boldsymbol{j}; \Omega)}{\sum_{\boldsymbol{b}\in\mathbb{Z}^p} m(\boldsymbol{y}_i + 2\pi\boldsymbol{b}; \Omega)} \,, \qquad \boldsymbol{j} \in \mathbb{Z}^p, \quad i = 1, \ldots, n \,,$$

2. *C-step*. Set $\boldsymbol{j}_i = \arg\max_{\boldsymbol{b}\in\mathbb{Z}^p} v_{i\boldsymbol{b}}$ and $v_{i\boldsymbol{j}} = 1$ for $\boldsymbol{j} = \boldsymbol{j}_i$, and $v_{i\boldsymbol{j}} = 0$ otherwise. Note that, at each iteration the classification algorithm provides also an estimate of the original unobserved sample obtained as $\hat{\boldsymbol{x}}_i = \boldsymbol{y}_i + 2\pi\boldsymbol{j}_i$, $i = 1, \ldots, n$.
3. *W-step* (weighting step). Based on current parameters' values, compute Pearson residuals according to (5) and evaluate the weights as

$$w_i = w(\delta_n(\boldsymbol{y}_i), \Omega, \hat{F}_n).$$

4. *M-step*. Update parameters' values by solving the WLEE

$$\sum_{i=1}^{n} w_i s(\boldsymbol{y}_i + 2\pi \boldsymbol{j}_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} w_i s(\hat{\boldsymbol{x}}_i; \boldsymbol{\theta}) = \mathbf{0}\,,$$

conditionally on $\boldsymbol{j}_i$ $(i = 1, \ldots, n)$, with $s(\boldsymbol{x}; \boldsymbol{\theta}) = \partial \log m(\boldsymbol{x}; \boldsymbol{\theta})/\partial\theta^\top$. In the Normal case, the WLEE returns weighted mean and variance-covariance matrix with weights w_i, given by

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{i=1}^{n} w_i \hat{\boldsymbol{x}}_i}{\sum_{i=1}^{n} w_i},$$
$$\hat{\Sigma} = \frac{\sum_{i=1}^{n} w_i (\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{\mu}}_i)(\hat{\boldsymbol{x}}_j - \hat{\boldsymbol{\mu}}_j)^\top}{\sum_{i=1}^{n} w_i}.$$

3.1 Properties

The WLEE to be solved in the M-step is of the type (8). Let denote it by $\Psi_n = \mathbf{0}$. Let θ_f be such that $f(\boldsymbol{y})$ is *close* to $m^\circ(\boldsymbol{y}; \boldsymbol{\theta}_f)$, that is $\boldsymbol{\theta}_f$ is implicitly defined by

$$\Psi = \int w(\delta(\boldsymbol{y})) s(\boldsymbol{y} + 2\pi \boldsymbol{j}; \boldsymbol{\theta}_f)\, dF(\boldsymbol{y}) = \mathbf{0},$$

given $\boldsymbol{j}$. We have the following results:

(i)
$$\sqrt{n}\,(\Psi_n - \Psi) \xrightarrow{d} N(0, V(\theta))$$

(ii)
$$\hat{\theta}^w \xrightarrow{a.s.} \theta_f$$

(iii)
$$\sqrt{n}\left(\hat{\theta}^w - \theta_f\right) \xrightarrow{d} N(0, B^{-1}(\theta_f) V(\theta_f) B^{-1}(\theta_f))$$

with

$$V(\theta) = \lim_{n\to\infty} \mathbb{V}ar\left[\int k((y - Y)/h) A'(\delta(y)) s(y; \theta)\, dy\right]$$
$$= \mathbb{V}ar\left[A'(\delta(Y)) s(Y; \theta)\right]$$

and

$$B(\theta) = \int A(\delta(y)) \nabla_2 m(y; \theta)\, dy - \int A'(\delta(y))(\delta(y) + 1) s(Y; \theta) s^\top(Y; \theta) m(y; \theta)\, dy\,,$$

where $V(\theta)$ is finite and positive definite and $B(\theta)$ is non-zero for $\theta = \theta_f$. At the true model, $B^{-1}(\theta_f) V(\theta_f) B^{-1}(\theta_f)$ coincides with the inverse of the expected Fisher information matrix and the WLE recovers full efficiency. **Details about the assumptions and proofs can be found in** [3,22].

In particular, one can also relax the mathematical device of evaluating integrals and their approximations given by sums on a trimmed set to avoid numerical instabilities due the occurrence of small (almost null) densities in the tails that would affect the denominator of Pearson residuals. As stated in [22], trimming is not necessary and could not be considered, especially in those models where the tails decay exponentially.

4 Numerical studies

The finite sample behavior of the proposed weighted CEM has been investigated by some numerical studies based on 500 Monte Carlo trials each, in the Normal case, with data drawn from a $WN_p(\boldsymbol{\mu}, \Sigma)$. We set $\boldsymbol{\mu} = 0$, whereas in order to account for the lack of affine equivariance of the Wrapped Normal model [28], we considered different covariance structures Σ as in [6]. In particular, for fixed condition number $CN = 20$, we obtained a random correlation matrix R. Then, the correlation matrix R has been converted into the covariance matrix $\Sigma = D^{1/2} R D^{1/2}$, with $D = \text{diag}(\sigma^2 \mathbf{1}_p)$, where σ is a chosen constant and $\mathbf{1}_p$ is a p-dimensional vector of ones. Outliers have been generated by shifting a proportion ϵ of randomly chosen data points by an amount k_ϵ in the direction of the smallest eigenvalue of Σ. We considered sample sizes $n = 50, 100, 500$, dimensions $p = 2, 5$, contamination level $\epsilon = 0, 5\%, 10\%, 20\%$, contamination size $k_\epsilon = \pi/4, \pi/2, \pi$ and $\sigma = \pi/8, \pi/4, \pi/2$.

For each combination of the simulation parameters, we compare the performance of CEM and weighted CEM algorithms. The weights used in the W-step are computed using the Generalized Kullback–Leibler RAF in Eq. (7) with $\tau = 0.1$. According to the strategy described in [5], the bandwidth h has been selected by setting $\Lambda = \Sigma$, so that h is a constant independent of the scale of the model. Here, h is obtained so that any outlying observation located at least three standard deviations away from the mean in a component-wise fashion, is attached a weight not larger than 0.12 when the rate of contamination in the data has been fixed equal to 20%. The algorithm has been initialized according to the root search approach described in [26] based on 15 subsamples of size 10. It is worth remarking here that there are not other robust proposals to be compared with our method, to the best of our knowledge.

The weighted CEM is assumed to have reached convergence when at the $(k+1)$–th iteration

$$\max\left(\sqrt{2(1-\cos(\hat{\boldsymbol{\mu}}^{(k)} - \hat{\boldsymbol{\mu}}^{(k+1)}))}, \max|\hat{\Sigma}^{(k)} - \hat{\Sigma}^{(k+1)}|\right) < 10^{-6}$$

where differences are element-wise and $\max|\hat{\Sigma}^{(k)} - \hat{\Sigma}^{(k+1)}|$ **denotes the maximum absolute difference in any of the components of the matrix** $\hat{\Sigma}^{(k)} - \hat{\Sigma}^{(k+1)}$. The algorithm has been implemented so that $\mathbb{Z}^p$ is replaced by the Cartesian product $\times_{s=1}^{p} \mathcal{J}$ where $\mathcal{J} = (-J, -J+1, \ldots, 0, \ldots, J-1, J)$ for some J providing a good approximation. Here we set $J = 3$. The algorithm runs on `R` code [31] available from the authors upon request.

Fitting accuracy has been evaluated according to

(i) the average angle separation ([9])

$$\text{AS}(\hat{\boldsymbol{\mu}}) = \frac{1}{p}\sum_{i=1}^{p}(1 - \cos(\hat{\mu}_i - \mu_i)) \,,$$

which ranges in $[0, 2]$, for the mean vector;

(ii) the divergence

$$\Delta(\hat{\Sigma}) = \text{trace}(\hat{\Sigma}\Sigma^{-1}) - \log(\det(\hat{\Sigma}\Sigma^{-1})) - p \,,$$

for the variance-covariance matrix. Here, we only report the results stemming from the challenging situation with $n = 100$ and $p = 5$.

Figure 3 displays the average angle separation whereas Fig. 4 gives the divergence to measure the accuracy in estimating the variance-covariance matrix for the weighted CEM (in dark grey) and CEM (in light grey). The weighted CEM exhibits a fairly satisfactory fitting accuracy both under the assumed model (i.e. when the sample at hand is not corrupted by the

occurrence of outliers) and under contamination. The robust method outperforms the CEM method, especially in the estimation of the variance–covariance components. The algorithm results in biased estimates for both the mean vector and the variance–covariance matrix only for the large contamination rate $\epsilon = 20\%$, with small contamination size and a large σ. Actually, in this data constellation outliers are not well separated from the group of genuine observations. A similar behavior has been observed for the other sample sizes. Complete results are made available in the "Supplementary Material".

4.1 Monitoring the smoothing parameter

As pointed out in Sect. 2, in finite samples the robustness/efficiency trade-off of weighted likelihood estimation can be tuned by varying the smoothing parameter h used in kernel density estimation. In the numerical studies above, h has been selected according to an objective criterion (see Section 4.1 in [26] for the details). However, practitioners are advised to monitor the behavior of weighted likelihood estimation as h varies in a reasonable range [16]. Here, the procedure is illustrated over a sample of size $n = 100$ from the previous numerical studies with $\sigma = \frac{\pi}{4}, \epsilon = 10\%, k_\epsilon = \frac{\pi}{2}$.

Figure 5 shows the trajectories of the weights at convergence corresponding to different values of h in the range $[0.001, 0.25]$. In particular, the weights relative to the generated outliers are in dark grey, whereas those for the genuine observations are displayed in light grey. Outliers are correctly downweighted for several values of h and, as expected, beyond a certain value the analysis becomes not robust. On the other side, weights corresponding to genuine observations rapidly goes to unity for increasing h. The (red) dashed line indicates the value of h used in the simulation study. Such a value correctly downweights the outlying observations.

5 Real data example: protein data

The data under consideration [27] contain bivariate information about 63 protein domains that were randomly selected from three remote Protein classes in the Structural Classification of Proteins (SCOP). In the following, we consider the data set corresponding to the 39th protein domain. A bivariate Wrapped Normal has been fitted to the data at hand by using the weighted CEM algorithm, based on a Generalized Kullback-Leibler RAF with $\tau = 0.25$ and $J = 6$. The tasks of bandwidth selection and initialization have been resolved according to the same strategy described above in Sect. 4.

The inspection of the data suggests the presence of at least a couple of clusters that make the data non homogeneous.

Figure 6 displays the data on a flat torus together with fitted means and 95% confidence regions corresponding to three different roots of the WLEE (that are illustrated by different colors): one root gives location estimate $\boldsymbol{\mu}_1 = (1.85, 2.34)$ and a positive correlation $\rho_1 = 0.79$; the second root gives location estimate $\boldsymbol{\mu}_2 = (1.85, 5.86)$ and a negative correlation $\rho_2 = -0.80$; the third root gives location estimate $\boldsymbol{\mu}_3 = (1.61, 0.88)$ and correlation $\rho_3 = -0.46$. The first and second roots are very close to maximum likelihood estimates obtained in different directions when unwrapping the data: this is evident from the shift in the second coordinate of the mean vector and the change in the sign of the correlation. In both cases the data exhibit weights larger than 0.5, except in few cases, corresponding to the most extreme observations, as displayed in the first two panels of Fig. 7. In none of the two cases the bulk

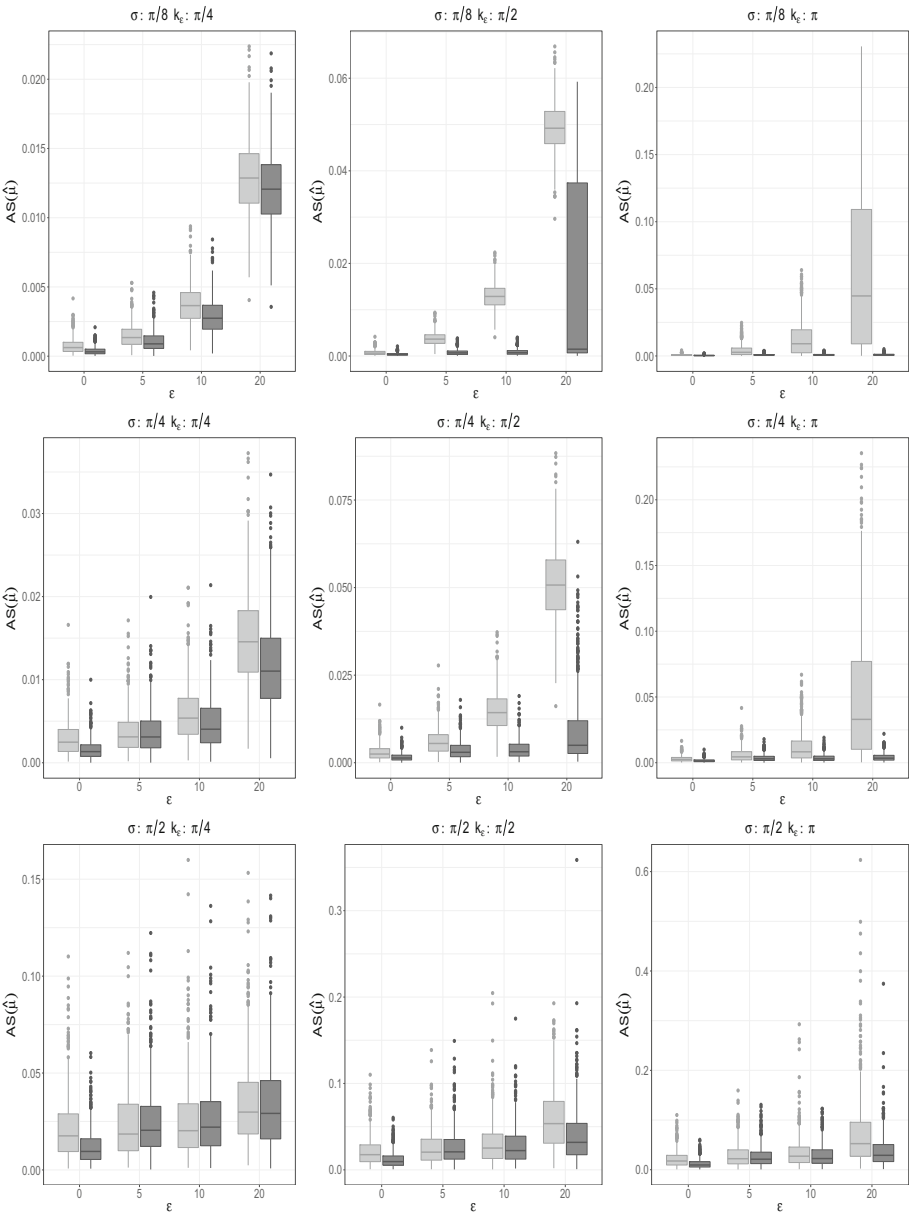

Fig. 3 Distribution of average angle separation for $n = 100$ and $p = 5$ using weighted CEM (in dark grey) and the CEM (in light grey). The contamination rate ϵ is given on the horizontal axis. Increasing contamination size k_ϵ from left to right, increasing σ from top to bottom

of the data corresponds to an homogeneous sub-group. On the contrary, the third root is able to detect an homogeneous substructure in the sample, corresponding to the most dense region in the data configuration. A weight close to zero is attached to almost half of the data points, as shown in the third panel of Fig. 7. These findings still confirm the ability of the

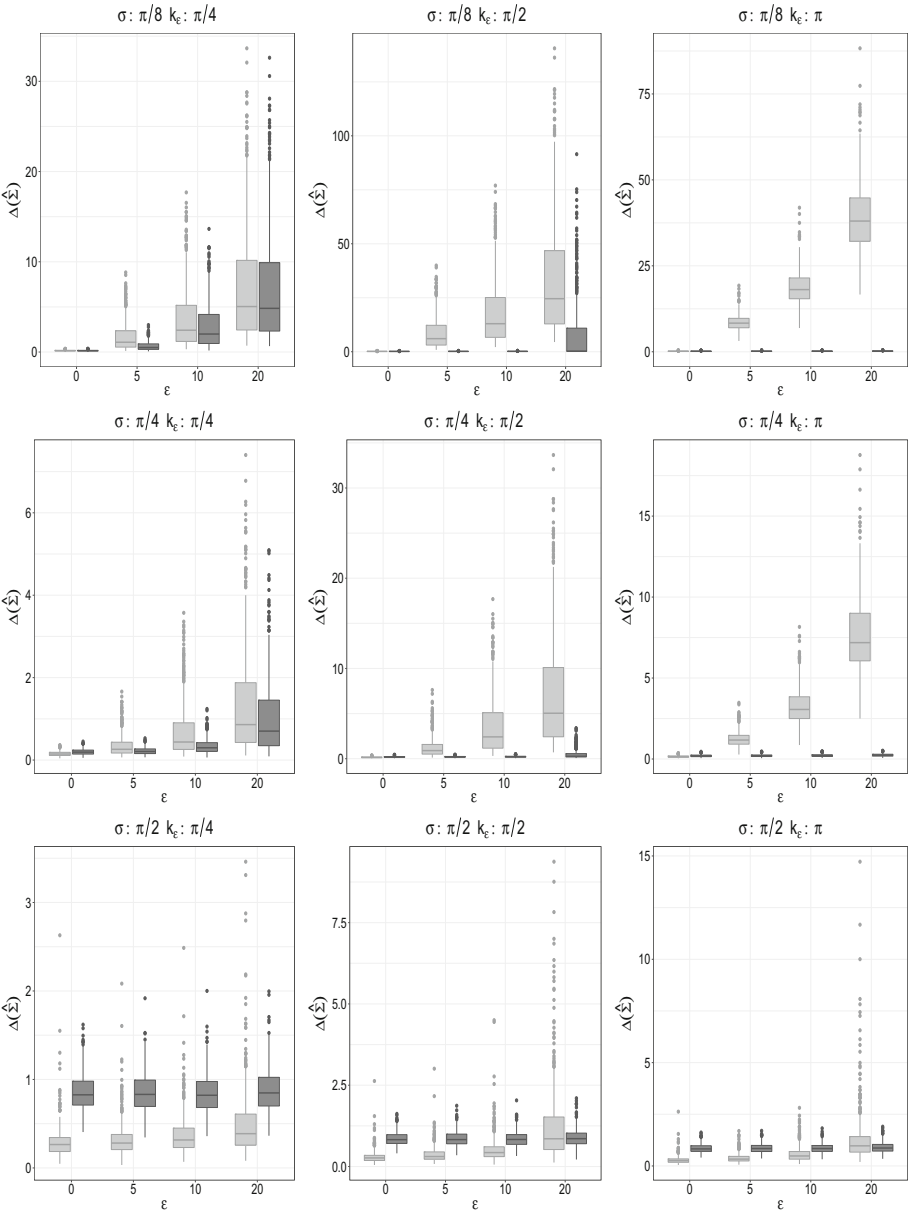

Fig. 4 Distribution of the divergence measure for $n = 100$ and $p = 5$ using the weighted CEM (in dark grey) and the CEM (in light grey). The contamination rate ϵ is given on the horizontal axis. Increasing contamination size k_ϵ from left to right, increasing σ from top to bottom

weighted likelihood methodology to tackle such uneven patterns as a diagnostic of hidden substructures in the data. In order to select one of the three roots we have found, we consider the strategy discussed in [1], that is, we select the root leading to the lowest fitted probability

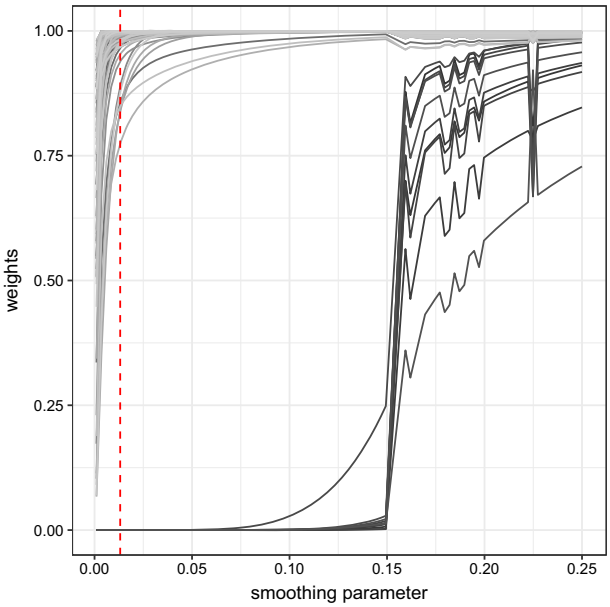

Fig. 5 Final weights of contaminated observations (dark grey) and uncontaminated observations (light grey) computed with respect to the smoothing parameter h. The dashed (red) line indicates the value used in the simulation study

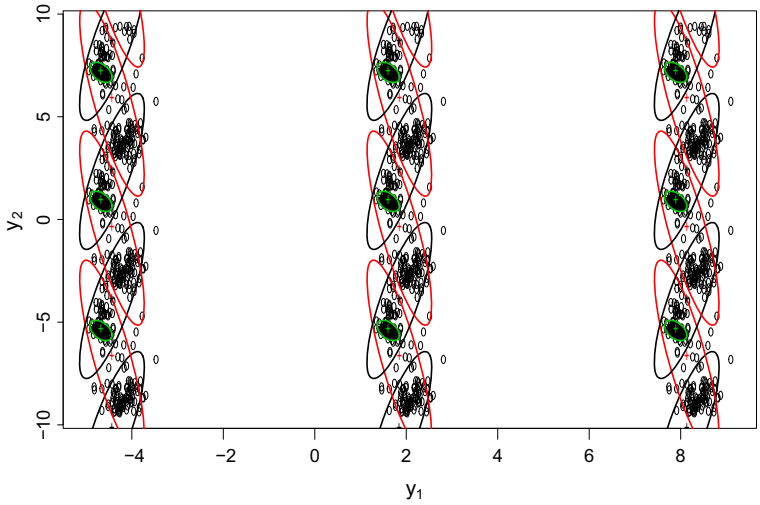

Fig. 6 Protein data. Fitted means (+) and 95% confidence regions corresponding to three different roots from weighted CEM ($J = 6$)

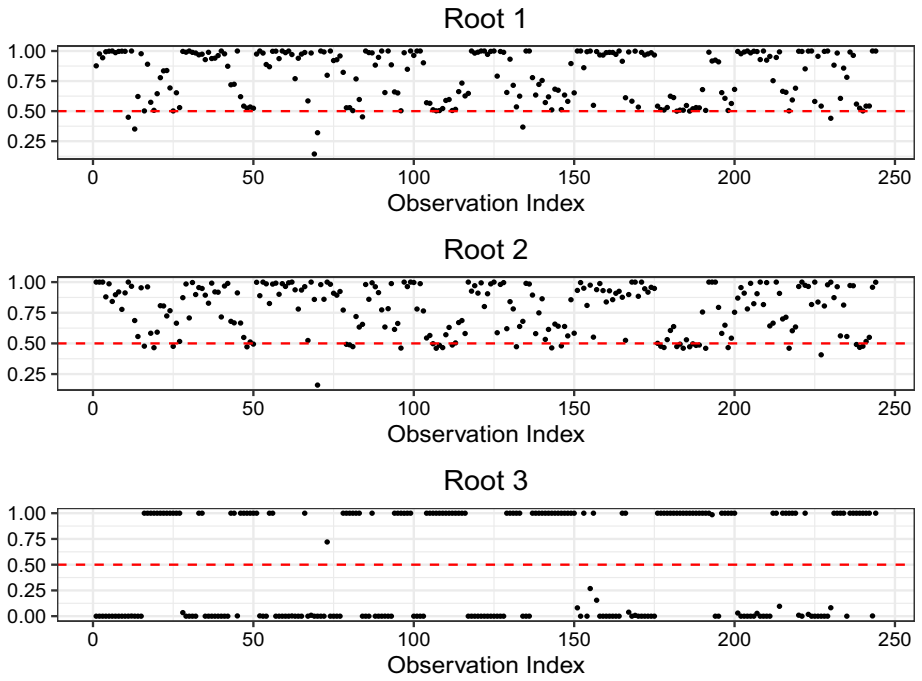

Fig. 7 Protein data. Weights corresponding to three different roots from weighted CEM

$$\text{Prob}_{\hat{\Omega}}\left(\delta_n(\mathbf{y};\hat{\Omega},\hat{F}_n) < -0.95\right).$$

This probability has been obtained by drawing 5000 samples from the fitted bivariate Wrapped Normal distribution for each of the three roots. The criterion correctly leads to choose the third root, for which an almost null probability is obtained, wheres the fitted probabilities for the first and second root are 0.204 and 0.280, respectively.

6 Conclusions

In this paper an effective strategy for robust estimation of multivariate Wrapped models on a $p-$dimensional torus has been presented. The method inherits the good computational properties of the CEM algorithm developed in [28] jointly with the robustness properties stemming from the employ of Pearson residuals and the weighted likelihood methodology. In this respect, it is particularly appealing the opportunity to work with a family of distribution that is close under convolution and allows to parallel the procedure one would have developed on the real line by using the multivariate normal distribution. The proposed weighted CEM works satisfactory at least in small to moderate dimensions, both on synthetic and real data. It is worth stressing that the method can be easily extended to other multivariate wrapped models.

Funding Open access funding provided by Università degli Studi di Trento within the CRUI-CARE Agreement.

References

1. Agostinelli, C.: Notes on Pearson residuals and weighted likelihood estimating equations. Stat. Probab. Lett. **76**(17), 1930–1934 (2006)
2. Agostinelli, C.: Robust estimation for circular data. Comput. Stat. Data Anal. **51**(12), 5867–5875 (2007)
3. Agostinelli, C., Greco, L.: Weighted likelihood estimation of multivariate location and scatter. Test **28**(3), 756–784 (2019)
4. Agostinelli, C., Lund U.: R package circular: circular statistics (version 0.4-93). https://r-forge.r-project.org/projects/circular/ (2017)
5. Agostinelli, C., Markatou, M.: Test of hypotheses based on the weighted likelihood methodology. Stat. Sin. 499–514 (2001)
6. Agostinelli, C., Leung, A., Yohai, V.J., Zamar, R.H.: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. TEST **24**(3), 441–461 (2015)
7. Baba, Y.: Statistics of angular data: wrapped normal distribution model. Proc. Inst. Stat. Math. **28**, 41–54 (1981). (in Japanese)
8. Basu, A., Lindsay, B.G.: Minimum disparity estimation for continuous models: efficiency, distributions and robustness. Ann. Inst. Stat. Math. **46**(4), 683–705 (1994)
9. Batschelet, E.: Circular Statistics in Biology. Academic Press, NewYork (1981)
10. Coles, S.: Inference for circular distributions and processes. Stat. Comput. **8**, 105–113 (1998)
11. Cressie, N., Read, T.R.C.: Multinomial goodness-of-fit tests. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **46**, 440–464 (1984)
12. Cressie, N., Read, T.R.C.: Statistic, cressie-read. In: Kotz, S., Johnson, N.L. (eds.) Encyclopedia of Statistical Sciences, supplementary volume, pp. 37–39. Wiley (1988)
13. Farcomeni, A., Greco, L.: Robust Methods for Data Reduction. CRC Press, New York (2016)
14. Ferrari, C.: The Wrapping Approach for Circular Data Bayesian Modeling. PhD Thesis, Alma Mater Studiorum University di Bologna. Dottorato di Ricerca in Metodologia Statistica per la Ricerca Scientifica (2009)
15. Fisher, N.I., Lee, A.J.: Time series analysis of circular data. J. R. Stat. Soc. Ser. B **56**, 327–339 (1994)
16. Greco, L., Agostinelli, C.: 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. Stat. Methods Appl. **27(4)**, 609–619 (2018)
17. Greco, L., Agostinelli, C.: Weighted likelihood mixture modeling and model-based clustering. Stat. Comput. **30**(2), 255–277 (2020)
18. Greco, L., Lucadamo, A., Agostinelli, C.: Weighted likelihood latent class linear regression. Stat. Methods Appl. (2020). https://doi.org/10.1007/s10260-020-00540-8
19. Jammalamadaka, S.R., SenGupta, A.: Topics in Circular Statistics. Multivariate Analysis, vol. 5. World Scientific, Singapore (2001)
20. Kato, S., Eguchi, S.: Robust estimation of location and concentration parameters for the von Mises-Fisher distribution. Stat. Pap. **57**(1), 205–234 (2016)
21. Ko, D.J., Chang, T.: Robust M-estimators on spheres. J. Multivar. Anal. **45**(1), 104–136 (1993)
22. Kuchibhotla, A.K., Basu, A.: Ayanendranath A minimum distance weighted likelihood method of estimation. In:Technical Report, Interdisciplinary Statistical Research Unit (ISRU), Indian Statistical Institute, Kolkata, India. https://faculty.wharton.upenn.edu/wp-content/uploads/2018/02/attemptv4p1.pdf (2018)
23. Lindsay, B.G.: Efficiency versus robustness: the case for minimum Hellinger distance and related methods. Ann. Stat. **22**, 1018–1114 (1994)
24. Mardia, K.V.: Statistics of Directional Data. Academic Press, London (1972)
25. Mardia, K.V., Jupp, P.E.: Directional Statistics. Wiley, New York (2000)
26. Markatou, M., Basu, A., Lindsay, B.G.: Weighted likelihood equations with bootstrap root search. J. Am. Stat. Assoc. **93**(442), 740–750 (1998)

27. Najibi, S.M., Maadooliat, M., Zhou, L., Huang, J.Z., Gao, X.: Protein structure classification and loop modeling using multiple Ramachandran distributions. Comput. Struct. Biotechnol. J. **15**, 243–254 (2017). https://doi.org/10.1016/j.csbj.2017.01.011
28. Nodehi, A., Golalizadeh, M., Maadooliat, M., Agostinelli, C.: Estimation of parameters in multivariate wrapped models for data on a p-torus. Comput. Stat. https://doi.org/10.1007/s00180-020-01006-x (2020)
29. Park, C., Basu, A., Lindsay, B.G.: The residual adjustment function and weighted likelihood: a graphical interpretation of robustness of minimum disparity estimators. Comput. Stat. Data Anal. **39**(1), 21–33 (2002)
30. Park, C., Basu, A.: The generalized Kullback-Leibler divergence and robust inference. J. Stat. Comput. Simul. **73**(5), 311–332 (2003)
31. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/ (2020)
32. Ravindran, P., Ghosh, S.K.: Bayesian analysis of circular data using wrapped distributions. J. Stat. Theory Pract. **5**, 547–561 (2011)
33. Sau, M.F., Rodriguez, D.: Minimum distance method for directional data and outlier detection. Adv. Data Anal. Classif. **12**(3), 587–603 (2018)