AStA Advances in Statistical Analysis
https://doi.org/10.1007/s10182-024-00494-2

ORIGINAL PAPER

Weighted likelihood methods for robust fitting of wrapped models for *p*-torus data

Claudio Agostinelli[1] · Luca Greco[2] · Giovanni Saraceno[3]

Received: 28 February 2023 / Accepted: 12 February 2024

Abstract

We consider, robust estimation of wrapped models to multivariate circular data that are points on the surface of a p-torus based on the weighted likelihood methodology. Robust model fitting is achieved by a set of weighted likelihood estimating equations, based on the computation of data dependent weights aimed to down-weight anomalous values, such as unexpected directions that do not share the main pattern of the bulk of the data. Weighted likelihood estimating equations with weights evaluated on the torus or obtained after unwrapping the data onto the Euclidean space are proposed and compared. Asymptotic properties and robustness features of the estimators under study have been studied, whereas their finite sample behavior has been investigated by Monte Carlo numerical experiment and real data examples.

Keywords Circular data · Expectation-maximization algorithm · Outliers · Pearson residual · Ramachandran plot

Mathematics Subject Classification 62H11 · 62F35

✉ Luca Greco
l.greco@unifortunato.eu

Claudio Agostinelli
claudio.agostinelli@unitn.it

Giovanni Saraceno
gsaracen@buffalo.edu

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

[2] University Giustino Fortunato, Benevento, Italy

[3] Department of Biostatistics, University of Buffalo, New York, NY, USA

Published online: 11 March 2024

1 Introduction

Multivariate circular data arise commonly in many different fields, including the analysis of wind directions (Lund 1999; Agostinelli 2007), animal movements (Ranalli and Maruotti 2020; Rivest et al. 2016), handwriting recognition (Bahlmann 2006), people orientation (Baltieri et al. 2012), cognitive and experimental psychology (Warren et al. 2017), human motor resonance (Cremers and Klugkist 2018), neuronal activity (Rutishauser et al. 2010) and protein bioinformatics (Mardia et al. 2007, 2012; Eltzner et al. 2018). The reader is pointed to Mardia and Jupp (2000a); Jammalamadaka and SenGupta (2001); Pewsey et al. (2013) for a general review. The data can be thought as points on the surface of a p-torus, embedded in a $(p+1)$ -dimensional space, whose surface is obtained by revolving the unit circle in a $p-$ dimensional manifold. A p-torus is topologically equivalent to a product of a circle p times by itself, written $\mathbb{T}^p$, $p \geq 1$ (Munkres 2018). The peculiarity of torus data is periodicity, which reflects in the boundedness of the sample space and often of the parametric space.

In order to illustrate the nature of torus data, let us consider a bivariate example, concerning $n = 490$ backbone torsion angle pairs (ϕ, ψ) for the protein 8TIM. Data are available from the R package BAMBI (Chakraborty and Wong 2021) and are extracted from the vast Protein Data Bank (Bourne 2000). The protein is an example of a TIM barrel folded into eight α-helices and eight parallel β-strands, alternating along the protein tertiary structure. It gets its name from the enzyme triose-phosphate isomerase, a conserved metabolic enzyme (Chang et al. 1993). The data are shown in Fig. 1 according to the Ramachandran plot of the angles over $[0, 2\pi) \times [0, 2\pi)$, in the right panel, or $[-\pi, \pi) \times [-\pi, \pi)$, in the left panel. Clearly, this type of graphical display is not unique and depends on how the angles are represented. Actually, the Ramachandran plot does not allow to show the intrinsic periodicity of the angles. In order to account for such wraparound nature of the data, one should topologically glue both pairs of opposite edges together with no twists. Then,

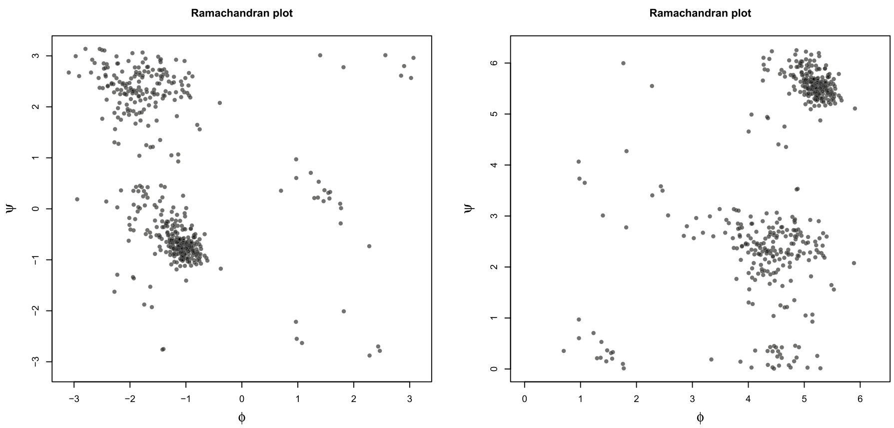

Fig. 1 8TIM protein data. Ramachandran plot over $[0, 2\pi) \times [0, 2\pi)$ (right) and over $[-\pi, \pi) \times [-\pi, \pi)$

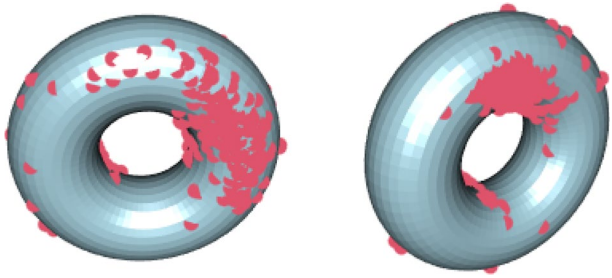

Fig. 2 8TIM protein data. Bivariate angles as points on the surface of a torus from two different perspectives

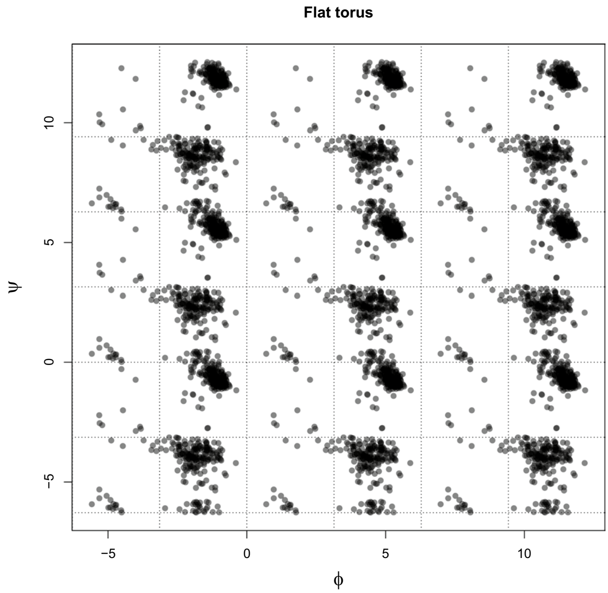

Fig. 3 8TIM protein data. Flat torus plot. The dotted lines give multiples of $\mp\pi$

the resulting surface is that of a torus with one hole (say, of genus one) in three dimensions. The data on the torus are shown in Fig. 2 from two different perspectives. The limitations of the Ramachandran plot in the two dimensional space can be circumvented by *unwrapping* the data on a flat torus, that is the angles are revolved around the unit circle a fixed number of times in each dimension and transformed into linear data, according to $\boldsymbol{x} = \boldsymbol{y} + 2\pi\boldsymbol{j}$, for a given $\boldsymbol{j} \in \mathbb{Z}^2$. This representation is shown in Fig. 3 where the data are given for different choices of $\boldsymbol{j} \in \mathbb{Z}^2$: then, the same data structure repeats itself to reflect the periodic nature of the data. Dotted lines give multiples of π.

The problem of modeling circular data has been tackled through suitable distributions, such as the von Mises (Mardia 1972). In a different fashion, in this paper, we focus our attention on the family of wrapped distributions (Mardia and Jupp 2000a). Wrapping is a popular method to define distributions for torus data. Let $\boldsymbol{X} = (X_1, X_2, \dots, X_p)$ be a *linear* random vector with distribution function $M(\boldsymbol{x}; \boldsymbol{\theta})$ and corresponding probability density function $m(\boldsymbol{x}; \boldsymbol{\theta})$, with $\boldsymbol{x} \in \mathbb{R}^p$ and $\boldsymbol{\theta} \in \Theta$. Assume that each component is wrapped around the unit circle, i.e., $Y_d = X_d \bmod 2\pi$, $d = 1, 2, \dots, p$, where mod denotes the modulus operator. Then, the distribution of $\boldsymbol{Y} = \boldsymbol{X} \bmod 2\pi$ is a p−variate wrapped distribution with distribution function

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

and probability density function

$$m^{\circ}(\boldsymbol{y}; \boldsymbol{\theta}) = \sum_{\boldsymbol{j} \in \mathbb{Z}^p} m(\boldsymbol{y} + 2\pi \boldsymbol{j}; \boldsymbol{\theta}) \tag{1}$$

$\boldsymbol{y} = (y_1, y_2, \dots, y_p) \in [0, 2\pi)^p$, $\boldsymbol{j} = (j_1, j_2, \dots, j_p) \in \mathbb{Z}^p$. The p-dimensional vector $\boldsymbol{j}$ is the vector of wrapping coefficients, that, if it was known, would describe how many times each component of the p-toroidal data point was wrapped. In other words, if we knew $\boldsymbol{j}$ along with $\boldsymbol{y}$, we would obtain the unwrapped data $\boldsymbol{x} = (x_1, x_2, \dots, x_p)$ as $\boldsymbol{x} = \boldsymbol{y} + 2\pi \boldsymbol{j}$. Hereafter, we concentrate on unimodal and elliptically symmetric densities of the form

$$m(\boldsymbol{x}; \boldsymbol{\theta}) \propto |\Sigma|^{-1/2} h\left((\boldsymbol{x} - \boldsymbol{\mu})^{\top} \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right), \tag{2}$$

where $h(\cdot)$ is a strictly decreasing and nonnegative function, $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma)$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_p)$ is a location vector and Σ is a $p \times p$ positive definite scatter matrix. When $h(t) = \exp(-t/2)$, the multivariate normal distribution is recovered as a special case. Applying the component-wise wrapping of a p-variate normal distribution $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ onto a p-dimensional torus, one obtains the multivariate wrapped normal (WN), $\boldsymbol{Y} \sim WN_p(\boldsymbol{\mu}, \Sigma)$, with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix Σ. Without loss of generality, we let $\boldsymbol{\mu} \in [0, 2\pi)^p$ to ensure identifiability.

Torus data are not immune to the occurrence of outliers, which are unexpected values, such as angles or directions, that do not share the main pattern of the bulk of the data. The key to understanding circular outliers lies in the intrinsic periodic nature of the data. In particular, outliers in the circular setting differ from those in the linear case, in that angular distributions have bounded support. For classical *linear* data in a Euclidean space, one single outliers can lead the mean to minus or plus infinity. In contrasts, breakdown occurs in directional data when contamination causes the mean direction to change by at most π (Davies and Gather 2005, 2006). Marginally, the occurrence and subsequent detection of anomalous circular data points clearly depends on the concentration of the data around some main direction. The lower the concentration, the more outliers are unlikely to occur and have a little effect on estimates of location or spread. Furthermore, in a multivariate framework, outliers can violate the main correlation structures of the data and lead to misleading associations. Therefore, when outliers do contaminate the torus data at hand, they can very badly affect likelihood based estimation, leading to unreliable inferences. The problem of robust fitting for directional data has been addressed, since the works of Lenth (1981); Ko and Guttorp (1988); He and Simpson (1992); Agostinelli (2007), mainly for univariate problems. A very first attempt to develop a robust parametric technique well suited for p-torus data and wrapped models can be found in Saraceno et al. (2021). A second approach has been discussed in Greco et al. (2021). They are both based on a set of weighted data-augmented estimating equations that are solved using a classification expectation-maximization (CEM) algorithm, whose M-step is enhanced by the computation of a set of data dependent weights aimed to down-weight outliers.

The main contributions of this paper can be summarized as follows. We generalize, the approach in Saraceno et al. (2021) building a set of weighted likelihood estimating equations (WLEE, Markatou et al. 1998) as weighted counterparts of the likelihood equations. The technique is developed in a very general framework for unimodal and elliptically symmetric distributions and not limited to the WN model. The resulting weighted likelihood estimator (WLE) can be evaluated according to different weighting schemes. We shed new light on the nature, definition and treatment of torus outliers. In details, it is shown how the different approaches to evaluate weights can be justified in light of the current definition of outliers in use. We present and discuss a new strategy to obtain weights for robust fitting based on the unwrapped data, after imputing the vector of wrapping coefficients $\boldsymbol{j}$. It is shown that the estimating equations based on the unwrapped data can be properly used for sufficiently enough concentrated distributions on the torus. Furthermore, this work is meant to be a step forward the existing literature also because it is accompanied by formal theoretical results about the asymptotic behavior and the robustness properties of the proposed estimators.

The remainder of the paper is organized according to the following structure. Some background on maximum likelihood estimation of wrapped models is given in Sect. 2. The concept of outlyingness for torus data is shown in Sect. 3. Methods for weighted likelihood fitting are shown in Sect. 4. Theoretical properties are shown in Sect. 5. Numerical studies are shown in Sect. 6. Real data examples are given in Sect. 7. `R` (R Core Team 2021) code to run the proposed algorithms and replicate the real examples is available as supplementary material.

2 Maximum likelihood estimation

Given, an i.i.d sample $(\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_n)$ from $\boldsymbol{Y} \sim m^\circ(\boldsymbol{y}; \boldsymbol{\theta})$, the maximum likelihood estimate (MLE) is obtained by maximizing the log-likelihood function

$$\ell^\circ(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log m^\circ(\boldsymbol{y}_i; \boldsymbol{\theta}) \tag{3}$$

or solving the corresponding set of estimating equations $\sum_{i=1}^{n} u^\circ(\boldsymbol{y}_i; \boldsymbol{\theta}) = \boldsymbol{0}$, where

$$u^{\circ}(\boldsymbol{y};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log m^{\circ}(\boldsymbol{y};\boldsymbol{\theta}) = \frac{\nabla_{\boldsymbol{\theta}} m^{\circ}(\boldsymbol{y};\boldsymbol{\theta})}{m^{\circ}(\boldsymbol{y};\boldsymbol{\theta})}$$

is the score function. For a wrapped unimodal elliptically symmetric model, i.e., given by wrapping (2) onto the p-torus, let

$$v_{ij} = v_{ij}(\boldsymbol{\mu}, \Sigma) = \frac{h'(\boldsymbol{y}_i + 2\pi \boldsymbol{j}; \boldsymbol{\mu}, \Sigma)}{\sum_{\boldsymbol{k}\in\mathbb{Z}^p} h(\boldsymbol{y}_i + 2\pi \boldsymbol{k}; \boldsymbol{\mu}, \Sigma)}. \tag{4}$$

Then, the MLE is the solution to the following fixed point equations

$$\begin{aligned} \boldsymbol{\mu} &= \frac{\sum_{i=1}^n \sum_{\boldsymbol{j}\in\mathbb{Z}^p} v_{ij}(\boldsymbol{y}_i + 2\pi\boldsymbol{j})}{\sum_{i=1}^n \sum_{\boldsymbol{k}\in\mathbb{Z}^p} v_{ik}} \\ \Sigma &= -\frac{2}{n} \sum_{i=1}^n \sum_{\boldsymbol{j}\in\mathbb{Z}^p} v_{ij}(\boldsymbol{y}_i + 2\pi\boldsymbol{j} - \boldsymbol{\mu})(\boldsymbol{y}_i + 2\pi\boldsymbol{j} - \boldsymbol{\mu})^{\top} . \end{aligned} \tag{5}$$

The reader is pointed to Appendix A for details. Finding the MLE requires an iterative procedure alternating between the computation of (4) based on current parameters values and finding the (updated) solution to (5). An approximate MLE can be obtained using crispy assignments after the computation of (4), that is we let

$$\hat{\boldsymbol{j}}_i = \mathrm{argmax}_{\boldsymbol{j}\in\mathbb{Z}^p} v_{ij} \tag{6}$$

and solve the estimating equation

$$\sum_{i=1}^n u(\hat{\boldsymbol{x}}_i; \boldsymbol{\theta}) = \boldsymbol{0} \tag{7}$$

based on the *unwrapped* (fitted) linear data $\hat{\boldsymbol{x}}_i = \boldsymbol{y}_i + 2\pi\hat{\boldsymbol{j}}_i$.

In the special situation given by the WN, the derivation of the MLE through the fixed point equations in (5) coincides with that obtained from an expectation-maximization (EM) algorithm based on a data augmentation procedure (Fisher and Lee 1994; Coles 1998; Jona Lasinio et al. 2012; Nodehi et al. 2021). In a similar fashion, the approximate MLE can be obtained from a classification EM (CEM) algorithm (Nodehi et al. 2021). See Appendix B.

Remark 1 The infinite sum over $\mathbb{Z}^p$ makes likelihood inference challenging and hence, it is common to replace it by a sum over the Cartesian product $\mathcal{C}_J = \mathcal{J}^p$, where $\mathcal{J} = (-J, -J+1, \dots, 0, \dots, J-1, J)$ for some J providing a good approximation, since the summands of the series converge to zero. The approximation based on the truncated series works when

$$\Pr\{(\boldsymbol{Y} - \boldsymbol{\mu}) \in (-2\pi J, 2\pi J]^p\} \leq \sum_{d=1}^p \Pr\left\{(Y_k - \mu_k) \in (-2\pi J, 2\pi J]\right\}$$

is negligible; this is the case when $\left(\mu_d - 4\Sigma_{dd}^{1/2}, \mu_d + 4\Sigma_{dd}^{1/2}\right) \subseteq (-2\pi J, 2\pi J]$, for $d = 1, 2, \dots, p$ (see also Kurz et al. 2014). Actually, in case of the wrapped elliptically symmetric family, the density in (1) tends to that of a uniform distribution as concentration decreases (see also Mardia and Jupp 2000b, for the WN case).

As noticed in Nodehi et al. (2021), the MLE for location is equivariant under affine transformation of the data in the original (unwrapped) linear space. On the contrary, this is not the case for the scatter matrix estimates. Furthermore, it is worth to remark that solving (7) does not lead to consistent estimates for Σ , since the $\hat{\boldsymbol{j}}_i$ cannot be a consistent estimates of the unknown wrapping coefficients. Therefore, there is lack of consistency for $\hat{\boldsymbol{x}}_i$, as well. The population estimating equation

$$\int_{\mathbb{T}^p} u^\circ(\boldsymbol{y}; \boldsymbol{\mu}, \Sigma) m^\circ(\boldsymbol{y}; \boldsymbol{\mu_0}, \Sigma_0)\, d\boldsymbol{y} = \boldsymbol{0} \tag{8}$$

is solved by the true values $(\boldsymbol{\mu_0}, \Sigma_0)$, hence making the MLE estimator Fisher consistent. In contrasts, the estimating equation (8) is not the population estimating equations corresponding to (7). Actually, we can always re-express our observations so that $z_i = \boldsymbol{y_i} - \boldsymbol{\mu} \in (-\pi, \pi]^p$. It is not difficult to see that $\hat{\boldsymbol{x}}_i = z_i$. Then, the distribution from which the $\hat{\boldsymbol{x}}_i$s are sampled is not $m(\boldsymbol{x}; \boldsymbol{\mu_0}, \Sigma_0)$. However, the distribution is still elliptically symmetric around $\boldsymbol{\mu}$ and its support is any hyper-cube of length 2π and in particular we can take $T(\boldsymbol{\mu}) = \times_{k=1}^p (\mu_k - \pi, \mu_k + \pi]$. We call this distribution the unwrapped model and we denote it by

$$m^u(\boldsymbol{x}; \boldsymbol{\mu_0}, \Sigma_0) = m^\circ(\boldsymbol{x}; \boldsymbol{\mu_0}, \Sigma_0)\mathbb{I}(\boldsymbol{x} \in T(\boldsymbol{\mu}_0)).$$

Now, we can define Σ_0^u as the solution to the CEM population estimating equation

$$\int_{\mathbb{R}^p} u(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) m^u(\boldsymbol{x}; \boldsymbol{\mu_0}, \Sigma_0)\, d\boldsymbol{x} = \boldsymbol{0}. \tag{9}$$

For illustrative purposes, let us consider the following univariate examples. In Fig. 4, we compare the unwrapped normal density $m^u(x; 0, \sigma_0^2)$ with the original normal density $m(x; 0, \sigma_0^2)$, for $\sigma_0 = 3\pi/8 \approx 1.178$ (left panel) and $\sigma_0 = \pi/2 \approx 1.571$

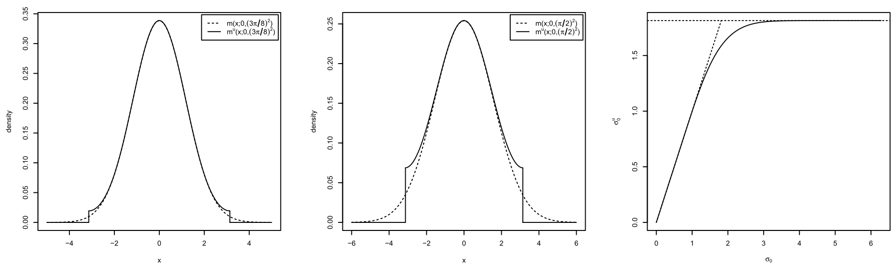

Fig. 4 Unwrapped normal density $m^u(x; 0, \sigma_0^2)$ compared with the original normal density $m(x; 0, \sigma_0^2)$, $\sigma_0 = 3\pi/8$, (left panel), $\sigma_0 = \pi/2$ (middle panel); σ_0^u versus σ_0 (right panel)

(middle panel). We find that $\sigma_0^u \approx 1.163$ and $\sigma_0^u \approx 1.460$, respectively. For small values of σ_0 the two densities are very similar apart from the truncation of the tails in the range $(-\pi, \pi]$. On the opposite, the difference becomes marked for large values of σ_0 the relation between σ_0 and σ_0^u is displayed in the right panel of Fig. 4. It follows that (7) can be safely used for $\sigma \leq \pi/2$. However, in most practical cases, distributions characterized by large concentrations are not of interest and the identification of outliers become unfeasible, as already shown in Sect. 1.

3 Outlyingness of torus data

We distinguish at least two approaches in the definition of outliers. The probabilistic approach is based on the idea that outliers are values *that are highly unlikely to occur under the assumed model* (Markatou et al. 1998; Agostinelli 2007). Under this perspective, outlyingness can be measured according to the degree of agreement between the data and the assumed model, as provided by the Pearson residual (Lindsay 1994). In contrasts, according to the geometric approach, outliers are observations *which deviate from the pattern set by the majority of the data* (Huber and Ronchetti 2009; Rousseeuw et al. 2011) with respect to a geometric distance. However, it is not straightforward to define and measure geometric distances on the torus (Mardia and Frellsen 2012). This makes the probabilistic point of view very appealing in this framework.

A simple but effective way to introduce outliers on the torus is that of considering the classical gross error model (Huber and Ronchetti 2009) on the unwrapped linear space. Let $0 \leq \epsilon < 0.5$ and $g(\boldsymbol{x})$ be an arbitrary density function. Then, the *true* density on the Euclidean space is

$$f(\boldsymbol{x}) = (1-\epsilon)m(\boldsymbol{x};\boldsymbol{\theta}) + \epsilon g(\boldsymbol{x}) \tag{10}$$

whereas, on the torus, we have that

$$\begin{aligned} f^\circ(\boldsymbol{y}) &= \sum_{\boldsymbol{j}\in\mathbb{Z}^p} f(\boldsymbol{y} + 2\pi\boldsymbol{j}) \\ &= (1-\epsilon)\sum_{\boldsymbol{j}\in\mathbb{Z}^p} m(\boldsymbol{y} + 2\pi\boldsymbol{j};\boldsymbol{\theta}) + \epsilon \sum_{\boldsymbol{j}\in\mathbb{Z}^p} g(\boldsymbol{y} + 2\pi\boldsymbol{j}) \\ &= (1-\epsilon)m^\circ(\boldsymbol{y};\boldsymbol{\theta}) + \epsilon g^\circ(\boldsymbol{y}). \end{aligned} \tag{11}$$

A measure of the agreement between the true and assumed model on the probabilistic ground is provided by the Pearson residual function (Lindsay 1994; Basu and Lindsay 1994; Markatou et al. 1998). Let $K_H(\boldsymbol{y})$ be a smooth family of (circular) kernel functions with bandwidth matrix H. Let $\hat{f}^\circ(\boldsymbol{y})$ and $\hat{m}^\circ(\boldsymbol{y};\boldsymbol{\theta})$ be smoothed densities, obtained by convolution between $K_H(\boldsymbol{y})$ and $f^\circ(\boldsymbol{y})$ and $m^\circ(\boldsymbol{y};\boldsymbol{\theta})$, respectively. In Saraceno et al. (2021) it has been suggested to measure the outlyingness of torus data based on (11) and using the Pearson residual function defined on $\boldsymbol{y} \in \mathbb{T}^p$ as

$$\delta^\circ(\boldsymbol{y};\boldsymbol{\theta}) = \frac{\hat{f}^\circ(\boldsymbol{y})}{\hat{m}^\circ(\boldsymbol{y};\boldsymbol{\theta})} - 1 \tag{12}$$

with $\delta^{\circ}(\boldsymbol{y};\boldsymbol{\theta}) \in [-1,+\infty)$, see also (Agostinelli 2007). The same probabilistic definition of outliers can be applied on the unwrapped linear space rather than on the torus, in a dual fashion. Therefore, in a CEM-based framework, one can define outlyingness on the unwrapped rather than circular data, based on (10). Actually, for a given $\boldsymbol{x} \in \mathbb{R}^p$, one can define the Pearson residual function

$$\delta(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\hat{f}(\boldsymbol{x})}{\hat{m}(\boldsymbol{x};\boldsymbol{\theta})} - 1, \tag{13}$$

where $\hat{f}(\boldsymbol{x})$ and $\hat{m}(\boldsymbol{x};\boldsymbol{\theta})$ are linear smoothed model densities. However, according to the results shown in Sect. 2, the use of a C-step does not lead to observe data directly from $m(\boldsymbol{x};\boldsymbol{\theta})$ but from the wrapped-unwrapped mechanism $m^u(\boldsymbol{x};\boldsymbol{\theta})$. Then, it would be correct to consider the Pearson residual function

$$\delta^u(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\hat{f}^u(\boldsymbol{x})}{\hat{m}^u(\boldsymbol{x};\boldsymbol{\theta})} - 1 \tag{14}$$

instead, with $\delta^u(\boldsymbol{x};\boldsymbol{\theta}) \in [-1,+\infty)$.

Large Pearson residuals detect points in disagreement with the model. This points are supposed to be down-weighted in the estimation process using a proper weighting function. The evaluation of a proper set of weights requires measuring the outlyingness of each data point with respect to a given (robust) fit of the postulated model. Based on the weighted likelihood methodology (Markatou et al. 1998), the weights are obtained from the finite sample counterparts of the Pearson residuals shown in (12) or (14). In the former case, we have

$$\delta_n^{\circ}(\boldsymbol{y};\boldsymbol{\theta}) = \frac{\hat{f}_n^{\circ}(\boldsymbol{y})}{\hat{m}^{\circ}(\boldsymbol{y};\boldsymbol{\theta})} - 1, \tag{15}$$

where $\hat{f}_n^{\circ}(\boldsymbol{y})$ is a circular kernel density estimate on the torus. As well, in the case of unwrapped data, we have that

$$\delta_n^u(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\hat{f}_n^u(\boldsymbol{x})}{\hat{m}^u(\boldsymbol{x};\boldsymbol{\theta})} - 1, \tag{16}$$

where $\hat{f}_n(\boldsymbol{x})$ is a kernel density estimate evaluated on the hyperplane over the fitted unwrapped (complete) data $(\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_n)$. In practice, for concentrated circular distributions, the Pearson residuals in (16) can be approximated by

$$\delta_n(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\hat{f}_n^u(\boldsymbol{x})}{\hat{m}(\boldsymbol{x};\boldsymbol{\theta})} - 1. \tag{17}$$

Smoothing the model makes the Pearson residuals converge to zero with probability one under the assumed model and it is not required that the kernel bandwidth goes to zero as the sample size increases (Markatou et al. 1998). In general, the choice of the kernel is not crucial.

Remark 2 When the model is the multivariate WN distribution, we can use a multivariate WN kernel with covariance matrix $H = \mathrm{diag}(h^2)$, since the smoothed model density is still an element of the WN family with covariance matrix $\Sigma + H$.

Remark 3 In practice, under the WN model, the distribution of the unwrapped data can be approximated by a multivariate normal variate for *concentrated* distributions, that is whenever all the variances are sufficiently *small*. In this case, using a multivariate normal kernel with bandwidth matrix $H = \mathrm{diag}(h^2)$ returns a smoothed model that is still normal with variance-covariance matrix $\Sigma + H$. It is worth to stress that the WN distribution inherits this property of closure with respect to convolution from the normal model. The closure to convolution property makes the use of the Gaussian kernel very appealing.

Remark 4 The family of elliptical distributions is not closed under convolution. e.g., see Sec 5.3.4 of (Prestele 2007). However, some subfamilies of elliptical distributions are closed under convolution; for example, the class of elliptical stable distributions are closed under convolutions.

Despite several weight functions could be used, in the weighted likelihood methodology it is common to consider

$$w(\delta) = \min\left\{1, \frac{[A(\delta)+1]^+}{\delta+1}\right\}, \tag{18}$$

where $w(\delta) \in [0, 1]$, $[\cdot]^+$ denotes the positive part and $A(\delta)$ is the residual adjustment function (RAF, Lindsay 1994; Basu and Lindsay 1994; Park et al. 2002), whose special role is related to the connections between weighted likelihood estimation and minimum disparity estimation. In practice, the RAF acts by bounding the effect of those points leading to large Pearson residuals. The function $A(\cdot)$ is assumed to be increasing and twice differentiable in $[-1, +\infty)$, with $A(0) = 0$ and $A'(0) = 1$. The weights decline smoothly to zero as $\delta \to \infty$ (outliers) and depending on the RAF also as $\delta \to -1$ (inliers). In particular, the weight function (18) can involve a RAF based on the Symmetric Chi-squared divergence (Markatou et al. 1998), the family of Power divergences (Lindsay 1994) or the Generalized Kullback–Leibler divergence (Park and Basu 2003) (see Saraceno et al. 2021, for details).

3.1 The geometric approach

The probabilistic approach allows to identify outliers either on the torus or after unwrapping the data, in a purely dual fashion. On the other hand, the geometric approach can be used only in the latter situation, as described in Greco et al. (2021). By exploiting the methodology developed in Agostinelli and Greco (2019), under the elliptically symmetric model in (3) and for a known wrapping coefficient vector $\boldsymbol{j}$, Pearson residuals and weights can be based on the squared Mahalanobis distance $d^2 = d^2(\boldsymbol{x}; \boldsymbol{\theta}) = [(\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})]$. In particular, finite sample Pearson residuals are defined as

$$\delta_n^{du}(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\hat{f}_n^u(d^2)}{\chi_u^2(d^2;p)} - 1, \tag{19}$$

where $\hat{f}_n(d^2)$ is a (unbounded at the boundary) kernel density estimate evaluated over squared Mahalanobis distances $d^2(\hat{\boldsymbol{x}};\hat{\boldsymbol{\theta}})$ and $\chi_u^2(d^2;p)$ is the density of the Mahalanobis distance evaluated under the wrapped-unwrapped model $m^u(\cdot;\boldsymbol{\theta})$. For concentrated circular distributions, the Pearson residual in (19) can be approximated by

$$\delta_n^{d}(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\hat{f}_n(d^2)}{\chi^2(d^2;p)} - 1, \tag{20}$$

where $\chi^2(\cdot;p)$ denotes the (asymptotic) distribution of Mahalanobis distances for the original linear data. Figure 5 shows two examples of $\chi_u^2(d^2;p)$ for $p = 6$ when $\sigma_0 = 3\pi/8$ (left panel) and $\sigma_0 = \pi/2$ (right panel). In the first case the support of the distribution is the interval $[0, 42.\bar{6})$ while in the second case is the interval $[0, 24)$.

4 Robust fitting based on WLEE

Robust fitting of a multivariate wrapped unimodal elliptically symmetric model to torus data can be achieved according to a weighted version of the population estimating equations (8), i.e.,

$$\int_{\mathbb{T}^p} w^\circ(\boldsymbol{y}) u^\circ(\boldsymbol{y};\boldsymbol{\mu},\Sigma) m^\circ(\boldsymbol{y};\boldsymbol{\mu_0},\Sigma_0)\, d\boldsymbol{y} = \boldsymbol{0}, \tag{21}$$

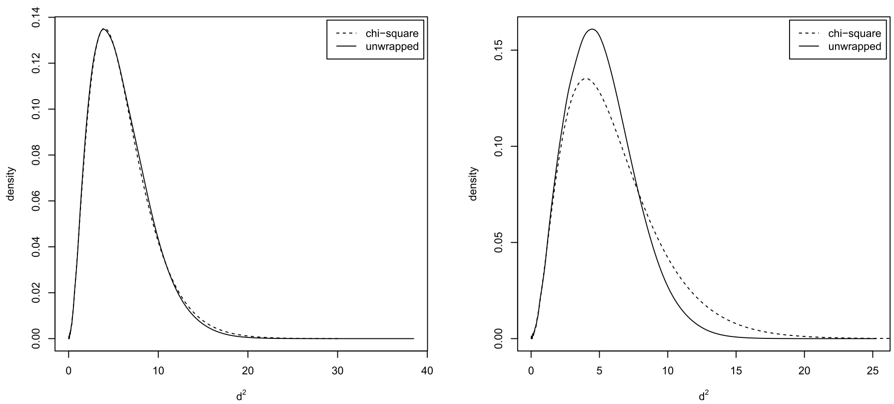

Fig. 5 Distribution of the squared Mahalanobis distance for the unwrapped observations from a wrapped normal model with $\sigma_0 = 3\pi/8$, (left panel) and $\sigma_0 = \pi/2$ (right panel)

where the weight function is given by $w^\circ(\boldsymbol{y}) = w(\delta^\circ(\boldsymbol{y};\boldsymbol{\theta}))$. We notice that $w^\circ(\boldsymbol{y})$ is a periodic function, i.e., $w^\circ(\boldsymbol{y}) = w^\circ(\boldsymbol{y} + 2\pi\boldsymbol{j})$, $\boldsymbol{j} \in \mathbb{Z}^p$. The sample version of (21), that is

$$\sum_{i=1}^{n} w(\delta_n^\circ(\boldsymbol{y}_i;\boldsymbol{\theta}))u^\circ(\boldsymbol{y}_i;\boldsymbol{\theta}) = \boldsymbol{0}$$

specializes to the following WLEE for unimodal elliptically symmetric distributions

$$\begin{aligned}\boldsymbol{\mu} &= \frac{\sum_{i=1}^{n} w(\delta_n^\circ(\boldsymbol{y}_i)) \sum_{\boldsymbol{j}\in\mathbb{Z}^p} v_{i\boldsymbol{j}}(\boldsymbol{y}_i + 2\pi\boldsymbol{j})}{\sum_{i=1}^{n} w(\delta_n^\circ(\boldsymbol{y}_i)) \sum_{\boldsymbol{k}\in\mathbb{Z}^p} v_{i\boldsymbol{k}}} \\ \Sigma &= -\frac{2}{\sum_{i=1}^{n} w(\delta_n^\circ(\boldsymbol{y}_i))} \sum_{i=1}^{n} w(\delta_n^\circ(\boldsymbol{y}_i)) \sum_{\boldsymbol{j}\in\mathbb{Z}^p} v_{i\boldsymbol{j}}(\boldsymbol{y}_i + 2\pi\boldsymbol{j} - \boldsymbol{\mu})(\boldsymbol{y}_i + 2\pi\boldsymbol{j} - \boldsymbol{\mu})^\top.\end{aligned} \tag{22}$$

with $w(\delta_n^\circ(\boldsymbol{y}_i)) = w(\delta_n^\circ(\boldsymbol{y}_i;\boldsymbol{\mu},\Sigma))$. The WLEE can be solved by a suitable modification of the iterative procedure shown in Sect. 2 to find the MLE. At iteration (s), based on current $v_{i\boldsymbol{j}}^{(s)}$ obtained as in (4), a set of data dependent weights $w_i^{(s)} = w\left(\delta_n^\circ(\boldsymbol{y}_i;\boldsymbol{\mu}^{(s)},\Sigma^{(s)})\right)$ is computed, whose effect is that of down-weighting the contribution of those points with large Pearson residuals based on the current fit. Then, updated estimates from iteration (s) to $(s+1)$ can be obtained by solving the WLEE in (22). In practice, the summation over $\mathbb{Z}^p$ is replaced by a summation over C_J.

According to a similar reasoning, we can consider a weighted counterpart of the population estimating equation (9), that is

$$\int_{\mathbb{R}^p} w(\boldsymbol{x})u(\boldsymbol{x};\boldsymbol{\mu},\Sigma)m^u(\boldsymbol{x};\boldsymbol{\mu_0},\Sigma_0)\,d\boldsymbol{x} = \boldsymbol{0}. \tag{23}$$

We notice that, in this situation, the use of (12) or (13) leads to the same estimator. Hence, one can build a WLEE based on the fitted unwrapped linear data $\hat{\boldsymbol{x}}_i$, with

weights whose evaluation can be now based on (15), (16) or (19). At iteration (s), estimates are updated according to

$$\begin{aligned}\hat{\boldsymbol{\mu}}^{(s+1)} &= \frac{\sum_{i=1}^{n} w_i^{(s)} \dot{h}_i^{(s)} \hat{\boldsymbol{x}}_i^{(s)}}{\sum_{i=1}^{n} w_i^{(s)} \dot{h}_i^{(s)}} \\ \hat{\Sigma}^{(s+1)} &= -\frac{2}{\sum_{i=1}^{n} w_i^{(s)}} \sum_{i=1}^{n} w_i^{(s)} \dot{h}_i^{(s)} \left(\hat{\boldsymbol{x}}_i^{(s)} - \hat{\boldsymbol{\mu}}^{(s+1)}\right)\left(\hat{\boldsymbol{x}}_i^{(s)} - \hat{\boldsymbol{\mu}}^{(s+1)}\right)^\top,\end{aligned} \tag{24}$$

where $\dot{h}_i^{(s)} = h'(d(\hat{\boldsymbol{x}}_i^{(s)};\hat{\boldsymbol{\mu}}^{(s)},\hat{\Sigma}^{(s)}))/h(d(\hat{\boldsymbol{x}}_i^{(s)};\hat{\boldsymbol{\mu}}^{(s)},\hat{\Sigma}^{(s)}))$.

We stress that the derivation of the WLEE for torus data generalizes the approach introduced in Saraceno et al. (2021), that was confined to a data augmentation perspective rather than on genuine maximum likelihood estimation. Therefore, here it is possible to derive a WLE that is the weighted counterpart of the MLE (and of its approximated version) and we are not limited to a CEM-type algorithm.

Remark 5 For a fixed bandwidth matrix H, the newly established weighting approach based on (16) requires that a multivariate kernel density estimate is computed at each iteration. The same is also true when using the weights in (19). In contrasts, the procedure based on (15) requires the evaluation of a more demanding torus kernel density estimate only once. However, computing a new kernel density estimate for linear data at each iteration adds no computational burden.

4.1 Bandwidth selection

The finite sample robustness of the WLE depends on the selection of the smoothing parameter h, whatever the type of Pearson residuals among those listed above. Large values of h lead to smooth kernel density estimates that are stochastically close to the postulated model. As a result, Pearson residuals are all close to zero, weights all close to one, and the WLE gains efficiency at the model but is less robust. On the opposite, small values of h make the kernel estimate more sensitive to the occurrence of outliers. Then, Pearson residuals become large where the data are in disagreement with the model and such points are properly down-weighted: the WLE loses efficiency at the model but recover robustness to outliers contamination.

The selection of h is still an open issue in weighted likelihood estimation. From a practical point of view, selecting a too small value for h can lead to an undue excess of down-weighting and hide relevant features in the data. In contrasts, a too large value could provide an insufficient down-weighting and misleading estimates, as well as the MLE. One strategy relies on a monitoring approach (Agostinelli and Greco 2018; Greco and Agostinelli 2020; Greco et al. 2020) in the selection of the bandwidth. It is suggested to run the procedure for different values of the smoothing parameter h and monitor the behavior of estimates and/or weights as h varies in a reasonable range. Monitoring the weights as h varies is expected to describe a transition from a robust to a nonrobust fit, since for increasing values of h all the weights approach one and the methodology does not allow to discriminate between the genuine part of the data and the outliers, anymore. As well, one can monitor a summary of the weights, such as the empirical down-weighting level $1 - \bar{w}$, where $\bar{w}$ denotes the average of the weights. It can be considered as a rough estimate of the amount of down-weighting. The approach of monitoring unveils patterns and substructures otherwise hidden that can aid the comprehension of the phenomenon under study and the sources of contamination.

4.2 Initialization

The iterative algorithm to solve the WLEE in (22) or (24) can be initialized using subsampling. 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. The initial value for the mean vector

$\boldsymbol{\mu}$ is set equal to the circular sample mean. Initial diagonal elements of Σ can be obtained as $\Sigma_{rr}^{(0)} = -2\log(\hat{\rho}_r)$, where $\hat{\rho}_r$ is the sample mean resultant length, whereas its off-diagonal elements are given by $\Sigma_{rs}^{(0)} = \rho_c(\boldsymbol{y}_r, \boldsymbol{y}_s)\sigma_{rr}^{(0)}\sigma_{ss}^{(0)}$ $(r \neq s)$, where $\rho_c(\boldsymbol{y}_r, \boldsymbol{y}_s)$ is the circular correlation coefficient, $r, 2 = 1, 2, \dots, p$ (Jammalamadaka and SenGupta 2001). It is suggested to run the algorithm from several starting points. The *best* solution can be selected by minimizing the probability to observe a small Pearson residual (Agostinelli and Greco 2019; Saraceno et al. 2021). According to the experience of the authors, a small number of subsamples is sufficient and very often they led to the same solution.

4.3 Outliers detection

The objective of a robust analysis is twofold: from the one hand we protect model fitting from the adverse effect of anomalous values, from the other hand it is of interest to provide effective tools to identify outliers based on formal rules and the robust fit. The process of outliers detection allows to investigate deeply their source and nature and unveil hidden and unexpected substructures in the data that are worth studying and may not have been considered otherwise (Farcomeni and Greco 2016). The inspection of weights provides a first approach for the task of outliers detection: points whose weight is below a fixed, and opportunely low, threshold (see also Greco and Agostinelli 2020 in a different framework) could be declared as outlying. However, it would be desirable to base outliers detection on an appropriate statistic to test outlyingness of each data point. In this respect, at least when robust fitting relies on (24), it is suggested to build a decision rule based on the fitted unwrapped linear data at convergence, treating them as a proper sample from a multivariate *linear* variate with density function as in (2). This approximation is supposed to work as long as torus data show a sufficiently high concentrated distribution. Therefore, one can pursue outliers detection looking at the squared robust distances $d^2(\hat{\boldsymbol{x}}_i; \hat{\theta})$. Outlying data are those whose distance exceeds a fixed threshold corresponding to the $(1-\alpha)$-level quantile of a chi-square distribution with p degrees of freedom (Greco et al. 2021).

5 Properties

Here, the asymptotic behavior of the proposed estimators and their robustness properties are investigated. The reader is pointed to Agostinelli and Greco (2019) for details on the asymptotic behavior of the WLE in a general setting. Hereafter, we assume broad regularity conditions for consistency and asymptotic normality of the MLE to hold.

5.1 Asymptotic distribution under the model

The following Lemma give the conditions to ensure the required asymptotic behavior of the Pearson residuals in (15), (16) and (19) and the corresponding weights at the assumed model. Henceforth, $\hat{f}(\boldsymbol{y}) = \hat{m}(\boldsymbol{y}, \boldsymbol{\theta}_0)$ (a.s.) and

$$\delta(\boldsymbol{y}; \boldsymbol{\theta}) = \frac{\hat{m}(\boldsymbol{y}, \boldsymbol{\theta}_0)}{\hat{m}(\boldsymbol{y}, \boldsymbol{\theta})} - 1,$$

where $\hat{m}(\boldsymbol{y}; \boldsymbol{\theta}) = \int k(\boldsymbol{y} - \boldsymbol{t}) m^*(\boldsymbol{t}; \boldsymbol{\theta})\, d\boldsymbol{t}$ is the smoothed model is involved in the definition of Pearson residuals in use, i.e., $m^*(\boldsymbol{y})$ can be $m^\circ(\boldsymbol{y})$, $m^u(\boldsymbol{x})$ or $\chi^2_u(d^2)$, respectively. Moreover, let δ_n be the Pearson residuals defined as either in (15), (16) or (19), and $\hat{f}_n$ be a kernel density estimator with kernel $K_H(\cdot)$ and bandwidth matrix H, corresponding to $\hat{f}^\circ_n$, $\hat{f}^u_n$ or $\hat{f}^{du}_n$, respectively, according to the definition of δ_n in use.

Lemma 1 *Assume that: (i) the kernel $K_H(\cdot)$ is of bounded variation; (ii) the model is correctly specified, that is, there exists $\boldsymbol{\theta}_0 \in \Theta$ such that $f^\circ(\boldsymbol{y}) = m^\circ(\boldsymbol{y}; \boldsymbol{\theta}_0)$ (a.s.); (iii) the model density is positive over the support $\mathcal{Y}$, that is, there exists $K > 0$ such that $\sup_{\boldsymbol{y}\in\mathcal{Y},\boldsymbol{\theta}\in\Theta} m^\circ(\boldsymbol{y}; \boldsymbol{\theta}) \geq K$; (iv) $A(0) = 0$, $A'(0) = 1$ and $A''(\delta)$ is a bounded and continuous function w.r.t. δ. Then,*

$$\sup_{\boldsymbol{y}\in\mathcal{Y},\boldsymbol{\theta}\in\Theta} |\delta_n(\boldsymbol{y}; \boldsymbol{\theta}) - \delta(\boldsymbol{y}; \boldsymbol{\theta})| \xrightarrow{a.s.} 0$$

$$\sup_{\boldsymbol{y}\in\mathcal{Y},\boldsymbol{\theta}\in\Theta} |w(\delta_n(\boldsymbol{y}; \boldsymbol{\theta})) - w(\delta(\boldsymbol{y}; \boldsymbol{\theta}))| \xrightarrow{a.s.} 0$$

$$\sup_{\boldsymbol{y}\in\mathcal{Y},\boldsymbol{\theta}\in\Theta} |w'(\delta_n(\boldsymbol{y}; \boldsymbol{\theta})) - w'(\delta(\boldsymbol{y}; \boldsymbol{\theta}))| \xrightarrow{a.s.} 0 \,.$$

Proof Under assumptions (i) and (ii) we have that $\hat{f}_n(\boldsymbol{y}) \xrightarrow{a.s.} \hat{m}(\boldsymbol{y}; \boldsymbol{\theta}_0)$ uniformly w.r.t. $\boldsymbol{y}$ as a result of the Glivenko–Cantelli theorem (Rao 2014). Under (iii) we obtain

$$\begin{aligned}\sup_{\boldsymbol{y}\in\mathcal{Y},\boldsymbol{\theta}\in\Theta} \left|\delta_n(\boldsymbol{y}; \boldsymbol{\theta}) - \delta(\boldsymbol{y}; \boldsymbol{\theta})\right| &= \sup_{\boldsymbol{y}\in\mathcal{Y},\boldsymbol{\theta}\in\Theta} \left|\frac{\hat{f}(\boldsymbol{y}) - \hat{m}(\boldsymbol{y}; \boldsymbol{\theta}_0)}{\hat{m}(\boldsymbol{y}; \boldsymbol{\theta})}\right| \\ &\leq \frac{\sup_{\boldsymbol{y}\in\mathcal{Y},\boldsymbol{\theta}\in\Theta} \left|\hat{f}(\boldsymbol{y}) - \hat{m}(\boldsymbol{y}; \boldsymbol{\theta}_0)\right|}{K} \\ &\xrightarrow{a.s.} 0 \,.\end{aligned}$$

the second and third statements follows from equation (18), assumption (iv) and the continuous mapping theorem. □

Remark 6 Assumption (iii) in Lemma 1 is plausible in the case of toroidal densities. It allows to relax the mathematical device of evaluating the supremum of the Pearson residuals, since it avoids the occurrence of small (almost null) densities in the tails that would affect the denominator of Pearson residuals (Agostinelli and Greco

2019). It is satisfied for wrapped models obtained from (2) under e.g., the assumption that $h(\cdot)$ is strictly positive in the hyper-cube $\times_{i=1}^{p}(\mu_i - \pi, \mu_i + \pi]$ and Σ is positive definite.

Lemma 2 *Assume that for all* $\boldsymbol{y}$ *and* $\boldsymbol{\theta}$, $\Psi(\boldsymbol{y};\boldsymbol{\theta}) = w(\delta(\boldsymbol{y};\boldsymbol{\theta}))u(\boldsymbol{y};\boldsymbol{\theta})$ *is differentiable and the matrix* $\dot{\Psi}(\boldsymbol{y};\boldsymbol{\theta})$ *with elements* i, j *be* $\partial\Psi_i/\partial\theta_j$ *is positive definite and* $\mathbb{E}_{\theta_0}(\dot{\Psi}(\boldsymbol{Y};\boldsymbol{\theta}))$ *is finite, then*

- *i. for every* n, *if there exists a solution* $\check{\boldsymbol{\theta}}_n$ *of* $\sum_{i=1}^{n}\Psi(\boldsymbol{Y}_i;\boldsymbol{\theta}) = \mathbf{0}$ *this solution is unique*;
- *ii. let* $\check{\boldsymbol{\theta}}_n$ *be the sequence of solutions, then* $\check{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \boldsymbol{\theta}_0$ *as* $n \to \infty$.

Proof Part *i.* is an application of Theorem 10.9 in Maronna et al. (2019). For part *ii.* notice that $\Psi(\boldsymbol{y};\boldsymbol{\theta}_0) = u(\boldsymbol{y};\boldsymbol{\theta}_0)$ and by a first order Taylor expansion around $\boldsymbol{\theta}_0$ of $\Psi(\boldsymbol{y};\boldsymbol{\theta})$ we have

$$0 = \sum_{i=1}^{n}\Psi(\boldsymbol{Y}_i;\check{\boldsymbol{\theta}}_n) = \sum_{i=1}^{n}u(\boldsymbol{Y}_i;\boldsymbol{\theta}_0) + \sum_{i=1}^{n}\dot{\Psi}(\boldsymbol{Y}_i;\boldsymbol{\theta}_i)(\check{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

, hence

$$\check{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = \left[\frac{1}{n}\sum_{i=1}^{n}\dot{\Psi}(\boldsymbol{Y}_i;\boldsymbol{\theta}_i)\right]^{-1}\frac{1}{n}\sum_{i=1}^{n}u(\boldsymbol{Y}_i;\boldsymbol{\theta}_0)$$

On the right hand side, the first term is bounded almost surely, while the second term goes to zero almost surely by the strong law of large numbers for i.i.d. random variables. Hence, $\check{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \boldsymbol{\theta}_0$ as $n \to \infty$. □

Theorem 1 (Consistency) *Under the assumptions of Lemmas 1 and 2. Assume* $\Psi_n(\boldsymbol{y};\boldsymbol{\theta}) = w(\delta_n(\boldsymbol{y};\boldsymbol{\theta}))u(\boldsymbol{y};\boldsymbol{\theta})$ *is differentiable and the matrix* $\dot{\Psi}_n(\boldsymbol{y};\boldsymbol{\theta})$ *with elements* i, j *be* $\partial\Psi_{n,i}/\partial\theta_j$ *is positive definite, then*

- *i. for every* n, *if there exists a solution* $\hat{\boldsymbol{\theta}}_n$ *of* $\sum_{i=1}^{n}\Psi_n(\boldsymbol{Y}_i;\boldsymbol{\theta}) = \mathbf{0}$ *this solution is unique*;
- *ii. let* $\hat{\boldsymbol{\theta}}_n$ *be the sequence of solutions, then* $\hat{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \boldsymbol{\theta}_0$ *as* $n \to \infty$.

Proof For each n consider a first order Taylor expansion around $\check{\boldsymbol{\theta}}_n$ of $\Psi_n(\boldsymbol{Y}_i;\boldsymbol{\theta})$ and hence,

$$\sum_{i=1}^{n}\left(\Psi_n(\boldsymbol{Y}_i;\hat{\boldsymbol{\theta}}_n) - \Psi_n(\boldsymbol{Y}_i;\check{\boldsymbol{\theta}}_n)\right) = \sum_{i=1}^{n}\dot{\Psi}_n(\boldsymbol{Y}_i;\boldsymbol{\theta}_{n,i})(\hat{\boldsymbol{\theta}}_n - \check{\boldsymbol{\theta}}_n)$$

and, since

$$\begin{aligned}\mathbf{0} = \sum_{i=1}^{n} \Psi_n(\boldsymbol{Y}_i;\hat{\boldsymbol{\theta}}_n) &= \sum_{i=1}^{n} \left(\Psi_n(\boldsymbol{Y}_i;\hat{\boldsymbol{\theta}}_n) - \Psi_n(\boldsymbol{Y}_i;\check{\boldsymbol{\theta}}_n)\right) + \sum_{i=1}^{n} \left(\Psi_n(\boldsymbol{Y}_i;\check{\boldsymbol{\theta}}_n) - \Psi(\boldsymbol{Y}_i;\check{\boldsymbol{\theta}}_n)\right) \\ &= \sum_{i=1}^{n} \dot{\Psi}_n(\boldsymbol{Y}_i;\boldsymbol{\theta}_{n,i})(\hat{\boldsymbol{\theta}}_n - \check{\boldsymbol{\theta}}_n) + \sum_{i=1}^{n} \left(\Psi_n(\boldsymbol{Y}_i;\check{\boldsymbol{\theta}}_n) - \Psi(\boldsymbol{Y}_i;\check{\boldsymbol{\theta}}_n)\right),\end{aligned}$$

we have

$$\hat{\boldsymbol{\theta}}_n - \check{\boldsymbol{\theta}}_n = -\left[\frac{1}{n}\sum_{i=1}^{n} \dot{\Psi}_n(\boldsymbol{Y}_i;\boldsymbol{\theta}_{n,i})\right]^{-1} \frac{1}{n}\sum_{i=1}^{n} \left(\Psi_n(\boldsymbol{Y}_i;\check{\boldsymbol{\theta}}_n) - \Psi(\boldsymbol{Y}_i;\check{\boldsymbol{\theta}}_n)\right).$$

the first term is bounded almost surely, while for the second term, we notice that

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n} \left(\Psi_n(\boldsymbol{Y}_i;\check{\boldsymbol{\theta}}_n) - \Psi(\boldsymbol{Y}_i;\check{\boldsymbol{\theta}}_n)\right)\right\| &= \left\|\frac{1}{n}\sum_{i=1}^{n} \left(w(\delta_n(\boldsymbol{Y}_i;\check{\boldsymbol{\theta}}_n)) - w(\delta(\boldsymbol{Y}_i;\check{\boldsymbol{\theta}}_n))\right)u(\boldsymbol{Y}_i;\check{\boldsymbol{\theta}}_n)\right\| \\ &\leq \sup_{y\in\mathcal{Y},\theta\in\Theta} \left|w(\delta_n(\boldsymbol{Y}_i;\check{\boldsymbol{\theta}}_n)) - w(\delta(\boldsymbol{Y}_i;\check{\boldsymbol{\theta}}_n))\right| \\ &\quad \times \frac{1}{n}\sum_{i=1}^{n} \left\|u(\boldsymbol{Y}_i;\check{\boldsymbol{\theta}}_n)\right\|\end{aligned}$$

the first term goes to zero almost surely by Lemma 1, while the second term is bounded almost surely by assumption on the second moment of the score function. Hence, $\hat{\boldsymbol{\theta}}_n - \check{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \mathbf{0}$ on the other hand, by Lemma 2 we have $\check{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \xrightarrow{a.s.} \mathbf{0}$ and this concludes the proof. □

Remark 7 We stress again that the WLE is consistent for $\boldsymbol{\theta}_0 = (\boldsymbol{\mu}_0, \Sigma_0^u)$ in the case of (23), but the differences between the solutions to the population estimating equations (21) and (23) are negligible for concentrated circular distributions, as well as for (8) and (9).

Theorem 2 (*Asymptotic distribution*) *Under the assumptions of Theorem* 1. *Assume, for each n*, Ψ_n *be twice differentiable with respect to* $\boldsymbol{\theta}$ *with bounded derivatives*; *let* $\dot{\Psi}_{n,jk} = \partial\Psi_{n,j}/\partial\theta_k$ *assume, for all* $\boldsymbol{y}, \boldsymbol{\theta}$ $|\dot{\Psi}_{n,jk}| \leq K(\boldsymbol{y})$ *with* $\mathbb{E}(K(\boldsymbol{Y})) < \infty$. *Then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, A^{-1}),$$

where $A = \mathbb{E}_{\theta_0}(\Psi(\boldsymbol{y};\boldsymbol{\theta}_0)\Psi(\boldsymbol{y};\boldsymbol{\theta}_0)^\top)$.

Proof The proof is similar to Theorem 10.11 of Maronna et al. (2019). Let $\epsilon(\boldsymbol{\theta}) = \mathbb{E}_{\theta_0}\Psi(\boldsymbol{Y};\boldsymbol{\theta})$ and B the matrix of derivatives with elements $\partial\epsilon_j/\partial\theta_k|_{\theta=\theta_0}$.

For each n and j call $\ddot{\Psi}_{n,j}$ be the matrix with elements $\partial\Psi_{n,j}/\partial\theta_k\partial\theta_h$. Let $A_n = \frac{1}{n}\sum_{i=1}^n \Psi_n(\boldsymbol{Y}_i;\boldsymbol{\theta}_0)$, $B_n = \frac{1}{n}\sum_{i=1}^n \dot{\Psi}_n(\boldsymbol{Y}_i;\boldsymbol{\theta}_0)$ and C_n is the matrix with its jth row equals to $(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \frac{1}{2n}\sum_{i=1}^n \ddot{\Psi}_{n,j}(\boldsymbol{Y}_i;\boldsymbol{\theta}_i)$. We notice that $\frac{1}{n}\sum_{i=1}^n \ddot{\Psi}_{n,j}(\boldsymbol{Y}_i;\boldsymbol{\theta}_i)$ is bounded and since $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \xrightarrow{a.s.} \mathbf{0}$ by Theorem 1, this implies that $C_n \xrightarrow{a.s.} \mathbf{0}$. From a second order Taylor expansion around $\boldsymbol{\theta}_0$ of $\Psi_n(\boldsymbol{y};\boldsymbol{\theta})$ it is easy to see that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -(B_n + C_n)^{-1}\sqrt{n}A_n.$$

From the proof of Theorem 1, we have $A_n - \frac{1}{n}\sum_{i=1}^n u(\boldsymbol{Y_i};\boldsymbol{\theta}_0) \xrightarrow{a.s.} \mathbf{0}$. In a similar way, using Lemma 1 we have $B_n - B \xrightarrow{a.s.} \mathbf{0}$. Since $u(\boldsymbol{Y_i};\boldsymbol{\theta}_0)$ are i.i.d and finite second moments, by multivariate central limit theorem we have $\frac{1}{\sqrt{n}}\sum_{i=1}^n u(\boldsymbol{Y_i};\boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, A)$ and hence, $\sqrt{n}A_n$ has the same limit. We notice that B coincides with the second derivatives of the log-likelihood and we had assume it positive definite. So, by the multivariate Slutsky's lemma, see, e.g., Maronna et al. (2019, Theorem 10.10) we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, B^{-1}AB^{-1\top})$$

on the other hand, under Bartlett's assumption we have $A = B$ and the result holds.

□

In the next corollary we provide a set of assumptions so that the previous results can be applied to wrapped unimodal elliptical symmetric models.

Corollary 1 *Consider a wrapped unimodal elliptically symmetric model as in (2). Let $\boldsymbol{\theta}_0 = (\boldsymbol{\mu}_0, \Sigma_0)$ be the true values with Σ_0 be a nonsingular covariance matrix, i.e., the sample $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ is i.i.d. from $m(\cdot;\boldsymbol{\theta}_0)$. Let $h(\cdot)$ be a strictly decreasing, non-negative function with uniformly bounded third derivatives and $h(\cdot)$ is positive in the region $T(\boldsymbol{\mu}_0)$. Assumptions in Lemma 1 hold. Then,*

i. the sequence $\hat{\theta}_n$ solutions of (22) is strongly consistent for $\boldsymbol{\theta}_0$ and

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, I(\boldsymbol{\theta}_0)^{-1}),$$

where $I = \mathbb{E}_{\boldsymbol{\theta}_0}(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^\top} m^\circ(\boldsymbol{Y};\boldsymbol{\theta}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ is the expected Fisher information matrix.

ii. the sequence $\tilde{\theta}_n$ solutions of (24) is strongly consistent for $\boldsymbol{\theta}_0^u = (\boldsymbol{\mu}_0, \Sigma_0^u)$ and

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0^u) \xrightarrow{d} N(\mathbf{0}, I^u(\boldsymbol{\theta}_0^u)^{-1}),$$

where $I^u = \mathbb{E}_{\boldsymbol{\theta}_0}(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^\top} m^u(\boldsymbol{Z};\boldsymbol{\theta}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^u}$ is the expected Fisher information matrix based on $m^u(\cdot;\boldsymbol{\theta}_0)$.

5.2 Influence function

The influence function (IF) plays a very important role in the evaluation of local robust properties of estimators in a classic robust framework (Huber and Ronchetti 2009). For a class of minimum distance estimators and weighted likelihood estimators (Beran 1977; Lindsay 1994), under broad regularity conditions and the assumed model, the IF coincides with that of the MLE. This feature suggests their high efficiency from one side, but a lack of local robustness on the other. The IF was used to investigate the robustness of some estimators for the circular mean direction in Him and Simpson (1992), but its use was unsatisfactory. Here, we discuss the IF of the proposed WLE in a more general setting.

Given a distribution function F, let $\boldsymbol{T} : F \mapsto \boldsymbol{T}(F) \in \Theta$ be a statistical functional that admits a von Mises expansion (Serfling 2009). Given, the gross error neighborhood $F_\epsilon(\boldsymbol{x}) = (1-\epsilon)F(\boldsymbol{x}) + \epsilon \mathbb{1}_z(\boldsymbol{x})$ we define the influence function of $\boldsymbol{T}$ at z as

$$\mathrm{IF}(z; \boldsymbol{T}, F) = \lim_{\epsilon \downarrow 0} \frac{\boldsymbol{T}(F_\epsilon) - \boldsymbol{T}(F)}{\epsilon} = \frac{\partial}{\partial \epsilon} \boldsymbol{T}(F_\epsilon)\big|_{\epsilon=0}.$$

Let $M_{\boldsymbol{\theta}} = M(\boldsymbol{x}; \boldsymbol{\theta})$ be the assumed model and $u(\boldsymbol{x}; \boldsymbol{\theta})$ the corresponding score function. Let $\boldsymbol{T}_F = \boldsymbol{T}(F)$ be the statistical functional solution of the weighted likelihood estimating equations

$$\int w(\boldsymbol{x}; \boldsymbol{T}(F), F) u(\boldsymbol{x}; \boldsymbol{T}(F))\, dF(\boldsymbol{x}) = \boldsymbol{0},$$

where we have $\boldsymbol{T}(M_{\boldsymbol{\theta}}) = \boldsymbol{\theta}$. The derivation of the IF for such functional is similar to the case of M-estimators (Huber and Ronchetti 2009). We have that

$$\frac{\partial}{\partial \delta} w(\delta) = \Big(\frac{\partial}{\partial \delta} A(\delta) - w(\delta) \Big) (\delta + 1)^{-1}$$

and

$$\frac{\partial}{\partial \epsilon} \delta(\boldsymbol{x}; \boldsymbol{T}(F_\epsilon), F_\epsilon)|_{\epsilon=0} = -\frac{k(\boldsymbol{x}; z, H)}{\hat{m}(\boldsymbol{x}; \boldsymbol{T}(F))} + (\delta(\boldsymbol{x}; \boldsymbol{T}(F), F) + 1)(1 - \hat{u}(\boldsymbol{x}; \boldsymbol{T}(F))\mathrm{IF}(z; \boldsymbol{T}, F)),$$

where $\hat{m}(\boldsymbol{x}; \boldsymbol{\theta})$ is the smoothed model and $\hat{u}(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log \hat{m}(\boldsymbol{x}; \boldsymbol{\theta})$. Then, we obtain

$$IF(z; \boldsymbol{T}, F) = D(F)^{-1} N(z, F),$$

where

$$
\begin{aligned}
N(\boldsymbol{z},F) &= w(\boldsymbol{z};\boldsymbol{T}(F),F)u(\boldsymbol{z};\boldsymbol{T}(F)) \\
&\quad + \int w'(\delta(\boldsymbol{x};\boldsymbol{T}(F),F))\frac{k(\boldsymbol{x};\boldsymbol{z},H)}{\hat{m}(\boldsymbol{x};\boldsymbol{T}(F))}u(\boldsymbol{x};\boldsymbol{T}(F))\,dF(\boldsymbol{x}) \\
&\quad - \int w'(\delta(\boldsymbol{x};\boldsymbol{T}(F),F))(\delta(\boldsymbol{x};\boldsymbol{T}(F),F)+1)u(\boldsymbol{x};\boldsymbol{T}(F))\,dF(\boldsymbol{x}) \\
&= w(\boldsymbol{z};\boldsymbol{T}(F),F)u(\boldsymbol{z};\boldsymbol{T}(F)) \\
&\quad + \int (A'(\delta(\boldsymbol{x};\boldsymbol{T}(F),F)) - w(\delta(\boldsymbol{x};\boldsymbol{T}(F),F))) \\
&\quad \times \left(\frac{k(\boldsymbol{x};\boldsymbol{z},H)}{\hat{f}(\boldsymbol{x})} - 1\right)u(\boldsymbol{x};\boldsymbol{T}(F))\,dF(\boldsymbol{x})
\end{aligned}
$$

and

$$
\begin{aligned}
D(F) &= \int w'(\delta(\boldsymbol{x};\boldsymbol{T}(F),F))(\delta(\boldsymbol{x};\boldsymbol{T}(F),F)+1)\hat{u}(\boldsymbol{x};\boldsymbol{T}(F))u(\boldsymbol{x};\boldsymbol{T}(F))^\top\,dF(\boldsymbol{x}) \\
&\quad - \int w(\boldsymbol{x};\boldsymbol{T}(F),F)u'(\boldsymbol{x};\boldsymbol{T}(F))\,dF(\boldsymbol{x}) \\
&= \int (A'(\delta(\boldsymbol{x};\boldsymbol{T}(F),F)) - w(\delta(\boldsymbol{x};\boldsymbol{T}(F),F)))\hat{u}(\boldsymbol{x};\boldsymbol{T}(F))u(\boldsymbol{x};\boldsymbol{T}(F))^\top\,dF(\boldsymbol{x}) \\
&\quad - \int w(\delta(\boldsymbol{x};\boldsymbol{T}(F),F))u'(\boldsymbol{x};\boldsymbol{T}(F))\,dF(\boldsymbol{x}),
\end{aligned}
$$

where $u'(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\log m(\boldsymbol{x};\boldsymbol{\theta})$. Under the model, we obtain the classical IF, that for the WLE corresponds to that of the MLE, i.e.

$$
\mathrm{IF}(\boldsymbol{z};\boldsymbol{T},M_{\boldsymbol{\theta}_0}) = I(\boldsymbol{\theta}_0)^{-1}\,u(\boldsymbol{z};\boldsymbol{\theta}_0),
$$

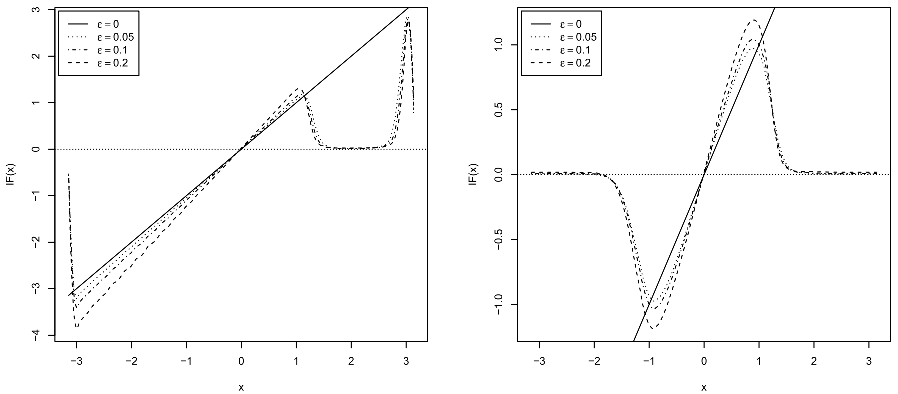

Fig. 6 WEM. Influence function for the location functional $\mu(F)$ with $f^\circ(\boldsymbol{y}) = (1-\varepsilon)m^\circ(\boldsymbol{y};0,\sigma_0^2) + \varepsilon m^\circ(\boldsymbol{y};\pi/2,(\pi/16)^2)$, for $\varepsilon = 0, 0.05, 0.10, 0.20$ and $\sigma_0 = \pi/8$ (left panel) and $\sigma_0 = \pi/4$ (right panel)

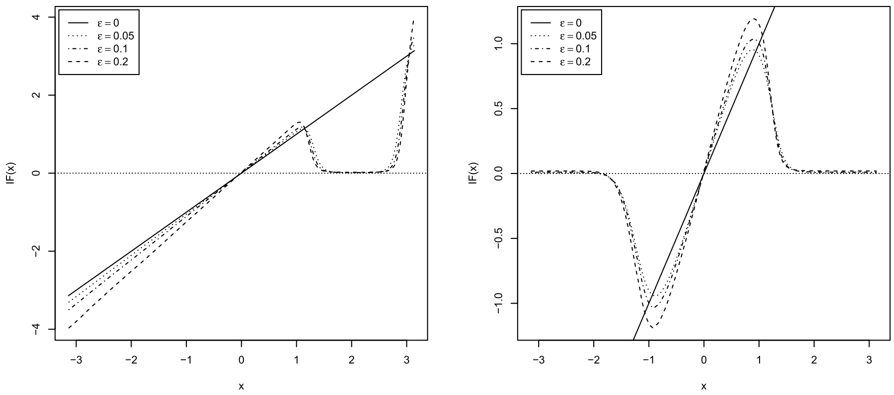

Fig. 7 WCEM using Pearson residuals as in (15) or (16). Influence function for the location functional $\mu(F)$ with $f^{\circ}(\boldsymbol{y}) = (1-\varepsilon)m^{\circ}(\boldsymbol{y}; 0, \sigma_0^2) + \varepsilon m^{\circ}(\boldsymbol{y}; \pi/2, (\pi/16)^2)$, for $\varepsilon = 0, 0.05, 0.10, 0.20$ and $\sigma_0 = \pi/8$ (left panel) and $\sigma_0 = \pi/4$ (right panel)

where $I(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}(u'(\boldsymbol{x}; \boldsymbol{\theta}))$ is the expected Fisher information matrix. However, the behavior of the IF under a distribution other than the postulated model is very different. As an example let us consider a simple setting in which $m^{\circ}(\boldsymbol{y}; \mu, \sigma^2)$ is the univariate WN and we are interested in evaluating the IF for the location functional when the data are from a two components mixture $f^{\circ}(\boldsymbol{y}) = (1-\varepsilon)m^{\circ}(\boldsymbol{y}; 0, \sigma_0^2) + \varepsilon m^{\circ}(\boldsymbol{y}; \pi/2, (\pi/16)^2)$. In Fig. 6, we show the IF of the location functional $\mu(F^{\circ})$ defined as the solution to the estimating equation (21) for $\sigma_0 = \pi/8$ (left panel) and $\sigma_0 = \pi/4$ (right panel). In this setting, the IF is a periodic function and in a region of high probability for the contaminating distribution the influence of a point is almost null. On the opposite, the behavior of the IF outside that region is similar to that of the maximum likelihood functional. We also notice

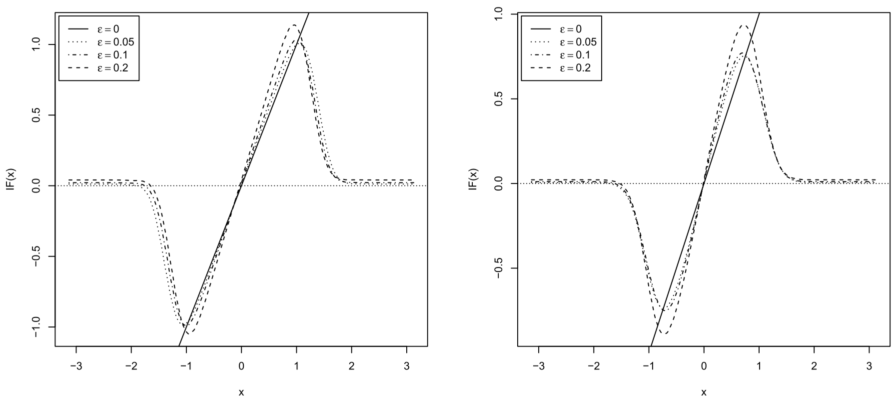

Fig. 8 WCEM using Pearson residuals as in (19). Influence function for the location functional $\mu(F)$ with $f^{\circ}(\boldsymbol{y}) = (1-\varepsilon)m^{\circ}(\boldsymbol{y}; 0, \sigma_0^2) + \varepsilon m^{\circ}(\boldsymbol{y}; \pi/2, (\pi/16)^2)$, for $\varepsilon = 0, 0.05, 0.10, 0.20$ and $\sigma_0 = \pi/8$ (left panel) and $\sigma_0 = \pi/4$ (right panel)

the change in sign at the antimode ($\pm\pi$). When we consider the location functional $\mu(F)$ associated to the WLE defined by (23) with Pearson residuals as in (12) or (13), the IF is not periodic and it is zero outside the interval $(\mu - \pi, \mu + \pi)$. Inside the interval, the behavior of the IF is similar to that of $\mu(F^\circ)$, as it is shown in Fig. 7. in contrasts, the IF of $\mu(F)$ with Pearson residuals built according to the geometric approach is symmetric, since only the magnitude of the outliers plays a role in the Mahalanobis distance, as shown in Fig. 8.

6 Numerical studies

In this section, we investigate the finite sample behavior of the proposed WLEs given by the WLEE in (22) and (24), for the different weighting schemes considered. The numerical studies are limited to the WN case. Since solving the WLEE in this case is equivalent to consider a weighted counterpart of the EM or CEM algorithms, in order to make it easier to read the results, we denote the WLE solution to (22) as WEM and the approximate WLE solution to (24) as WCEM-torus, WCEM-unwrap and WCEM-dist, depending on whether weights are based on residuals in (15), (17) or (20), respectively. The MLE and its approximated version have been also taken into account and are denoted by EM and CEM, respectively. We consider numerical studies based on $N = 500$ Monte Carlo trials. Data are sampled from a $p-$variate WN with null mean vector and variance-covariance matrix $\Sigma = D^{1/2}RD^{1/2}$, where R is a random correlation matrix with condition number set equal to 20 and $D = \sigma \mathrm{I}_p$. Contamination has been added by replacing a proportion ϵ of randomly selected data points. Those observations are shifted by an amount k_ϵ in the direction of the smallest eigenvector of Σ and perturbed by adding some noise from a $p-$variate wrapped normal with independent components and marginal scale σ_ϵ. We considered a sample size $n = 250$, number of dimensions $p = 2, 5$, $\sigma = \pi/8, \pi/4$, $\epsilon = 0, 0.10, 0.20$, $k_\epsilon = \pi/2, \pi$, $\sigma_\epsilon = 0.05$, $J = 2$. The case $\epsilon = 0$ concerns the situation without contamination and allows to investigate the behavior of the proposed robust methods at the true model. When $p = 5$, contamination only affects the first two dimensions. The bandwidths have been chosen so that all the WLEs return an empirical downweighting level close to the nominal contamination size to make a fair comparison. The weights are based on a GKL RAF. Initialization is based on subsampling with twenty subsamples of size $p + p(p+1)/2 + 5$. This choice did not represent an issue. Moreover, very often the different starting values led to the same solution. All the algorithms are assumed to reach convergence when

$$\max\left(g(\hat{\boldsymbol{\mu}}^{(s+1)} - \hat{\boldsymbol{\mu}}^{(s)}), \|\hat{\Sigma}^{(s+1)} - \hat{\Sigma}^{(s)}\|\right) < 10^{-6},$$

where $g(\boldsymbol{\mu}) = \sqrt{2(1-\cos(\boldsymbol{\mu}))}$. Fitting accuracy is evaluated according to

(i) the square root average angle separation

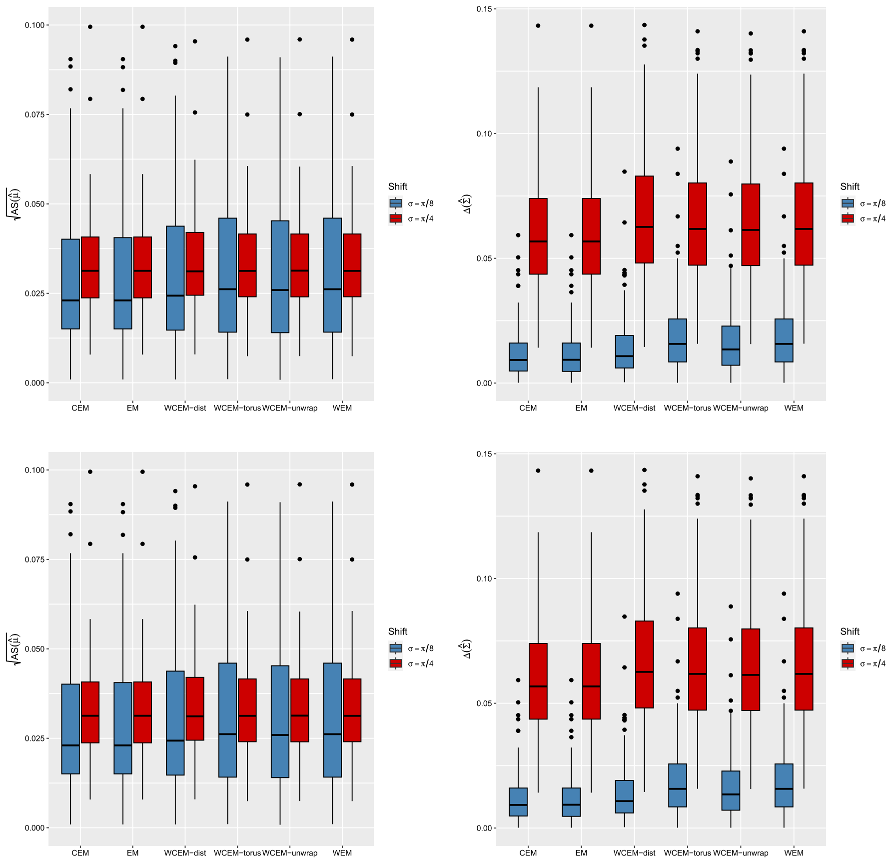

Fig. 9 Box-plots for $\sqrt{AS(\hat{\boldsymbol{\mu}})}$ (left) and $\Delta(\hat{\Sigma})$ (right) for $p = 2$ (top) and $p = 5$ (bottom), $\sigma = \pi/8, \pi/4$ when $\epsilon = 0$

$$\sqrt{AS(\hat{\boldsymbol{\mu}})} = \sqrt{\frac{1}{p}\sum_{j=1}^{p}(1 - \cos(\hat{\boldsymbol{\mu}}_j)),}$$

(ii) the divergence:

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

The effectiveness of the outliers detection rules shown in Sect. 4 is assessed in terms of swamping and power, that is evaluating the rate of genuine observation wrongly declared outliers and that of outliers correctly detected, respectively, for an overall significance level $\alpha = 1\%$. Both univariate and multivariate kernel density estimation involved in the computation of Pearson residuals in (17) and (20), respectively, has been performed using the functions available from package `pdfCluster` (Azzalini and Menardi 2014). The numerical studies are based on nonoptimized `R`

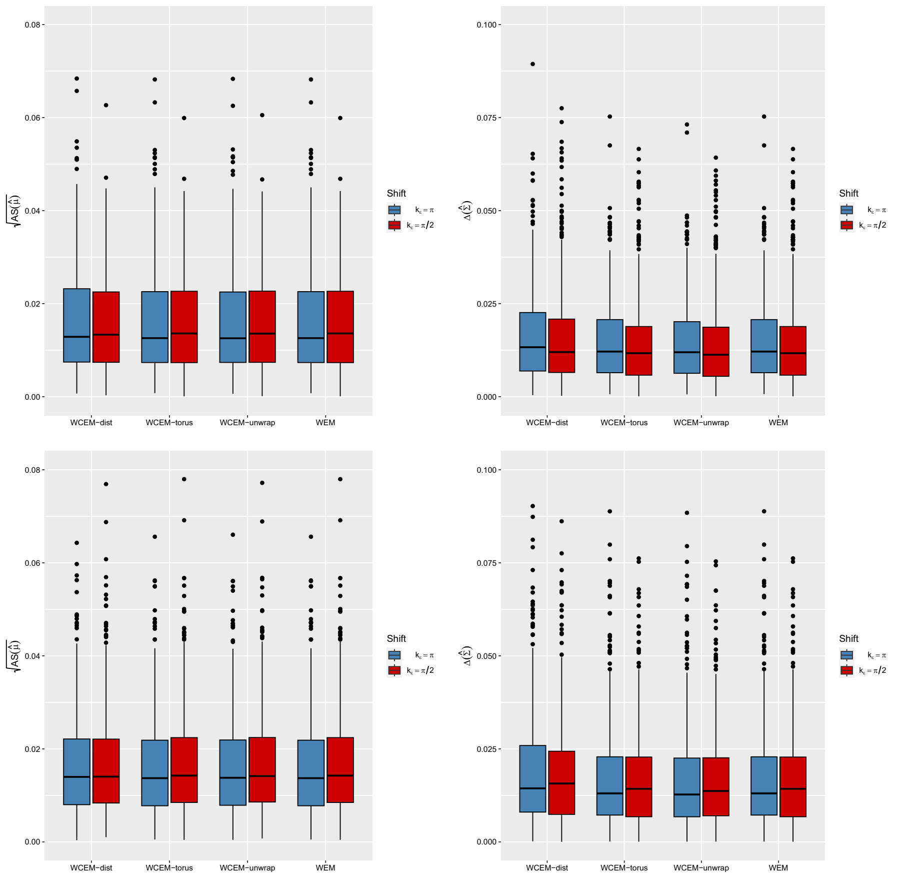

Fig. 10 Box-plots for $\sqrt{AS(\hat{\boldsymbol{\mu}})}$ (left) and $\Delta(\hat{\Sigma})$ (right) for $p = 2$, $\sigma = \pi/8$, $k_\epsilon = \pi/2, \pi$ when $\epsilon = 10\%$ (top) and $\epsilon = 20\%$ (bottom)

code and have been run on a 3.4 GHz Intel Core i5 quad-core. Codes are available as supplementary material.

Figure 9 displays the results under the true model, for $p = 2, 5$: the robust methods all provide accurate results in this scenario and the observed differences with respect to the MLE are tolerable. Figures 10 and 11 give the empirical distributions of the four WLEs in the presence of contamination when $p = 2$ and $\sigma = \pi/8$ or $\sigma = \pi/4$, respectively. As well, Figs. 12 and 13 concern the case with $p = 5$. The MLE becomes unreliable and it is not shown. In contrasts, the robust techniques always provide resistant estimates, as expected. We do not observe relevant differences among the robust proposals in terms of fitting accuracy. For what concerns the task of outliers detection, all the suggested WLEs return an average rate of swamping close to the nominal level and a power almost always equal to one, for all considered scenario and they do not exhibit different performances.

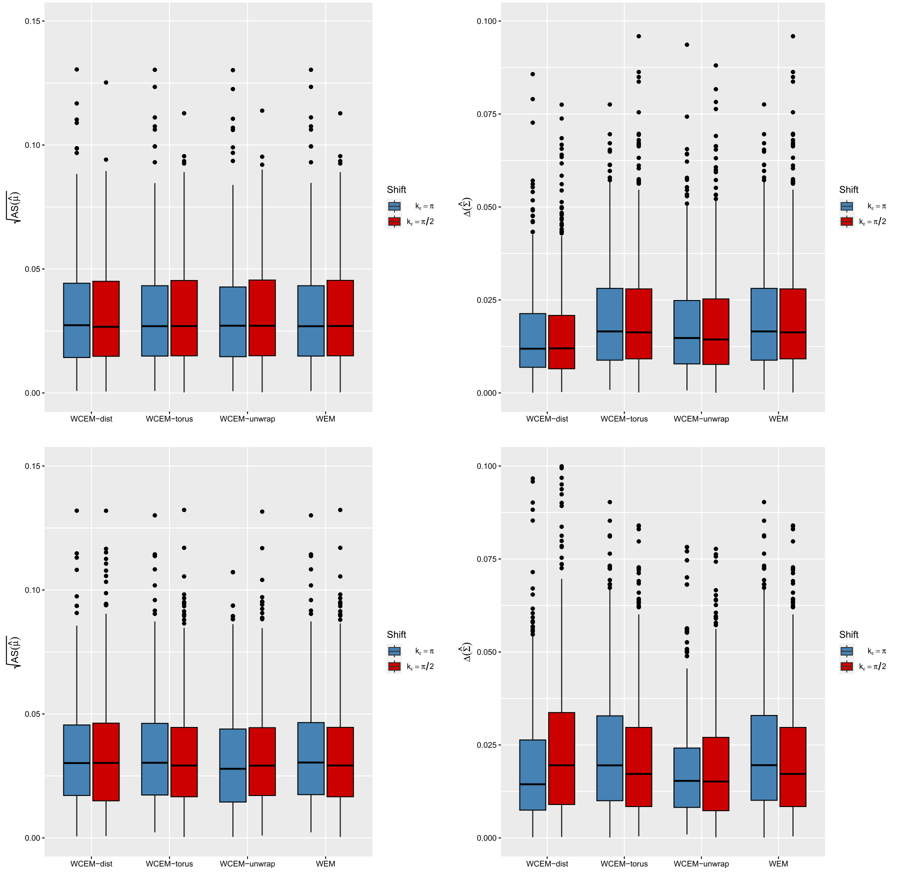

Fig. 11 Box-plots for $\sqrt{AS(\hat{\boldsymbol{\mu}})}$ (left) and $\Delta(\hat{\Sigma})$ (right) for $p = 2$, $\sigma = \pi/4$, $k_\epsilon = \pi/2, \pi$ when $\epsilon = 10\%$ (top) and $\epsilon = 20\%$ (bottom)

Computational time was always in a feasible range. However, based on the current codes, there is a remarkable time saving from the use of WCEM-unwrap or WCEM-dist with respect to WCEM-torus and WEM. One main reason could be the use of the functions from `pdfCluster` in the former two methods. For instance, when $p = 2$, $\sigma = \pi/4$, $\epsilon = 20\%$ the median elapsed time was about 12 s for the WEM and the WCEM-torus, but only 1.3 s for the WCEM-unwrap and slightly larger (still less than two) for the WCEM-dist. The advantage of using the WCEM combined with Pearson residuals in (17) was overwhelming for $p = 5$: with $\sigma = \pi/4$ and $\epsilon = 20\%$ the WEM and WCEM-torus took a median time of about 75 and 80 s, respectively for $k_\epsilon = \pi/2$, whereas the WCEM-unwrap took about 9 s and the WCEM-dist about 35 s. The case with $k_\epsilon = \pi$ was less computationally demanding but still the differences were noticeable: about 55 s for the WEM and WCEM-torus, about 27 s for the WCEM-dist and only about 4 s for the WCEM-unwrap. The ability to evaluate weights on the unwrapped data rather than on the torus reduced the computational time, indeed.

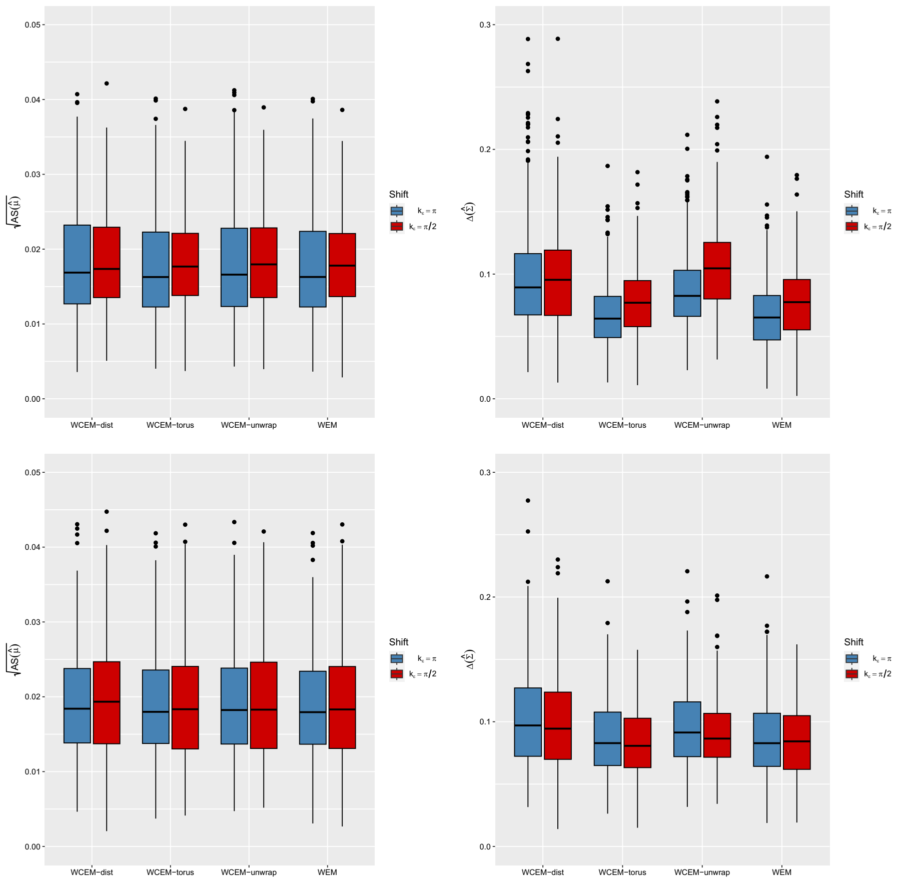

Fig. 12 Box-plots for $\sqrt{AS(\hat{\boldsymbol{\mu}})}$ (left) and $\Delta(\hat{\Sigma})$ (right) for $p = 5$, $\sigma = \pi/8$, $k_\epsilon = \pi/2, \pi$ when $\epsilon = 10\%$ (top) and $\epsilon = 20\%$ (bottom)

7 Real data examples

7.1 8TIM protein data

Let us consider the 8TIM protein data shown in Sect. 1. We compare the results from maximum likelihood estimation and its robust counterparts based on weighted likelihood estimation under the WN model assumption. We use the same notation shown in Sect. 6 to denote the different estimates. The data and the fitted models given by the EM and WCEM-unwrap based on (16) are shown in Fig. 14: the Ramachandran plot of the angles over $[0, 2\pi) \times [0, 2\pi)$ is given in the left panel, whereas data are displayed on a flat torus in the right panel, to account for their cyclic topology. The results from the WEM or WCEM-torus are indistinguishable. In both panels, the fitted models are represented through tolerance ellipses based on the 0.99–level quantile of the χ^2_2 distribution. The data clearly show a multi-modal clustered pattern.

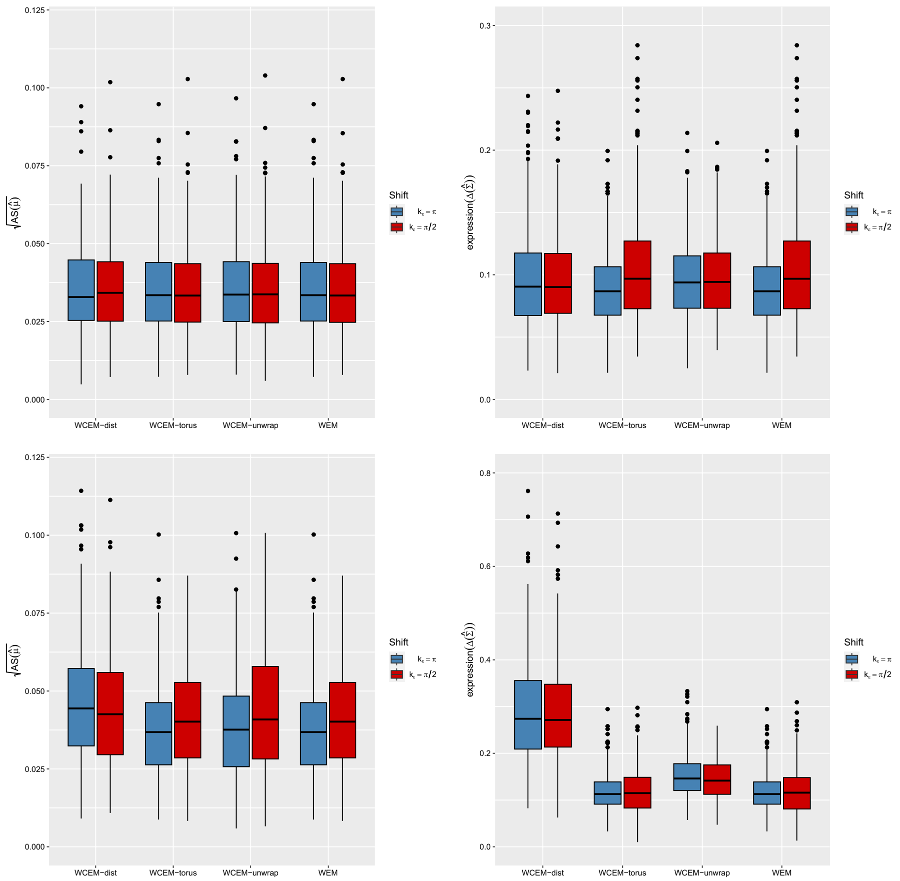

Fig. 13 Box-plots for $\sqrt{AS(\hat{\boldsymbol{\mu}})}$ (left) and $\Delta(\hat{\Sigma})$ (right) for $p = 5$, $\sigma = \pi/4$, $k_\epsilon = \pi/2, \pi$ when $\epsilon = 10\%$ (top) and $\epsilon = 20\%$ (bottom)

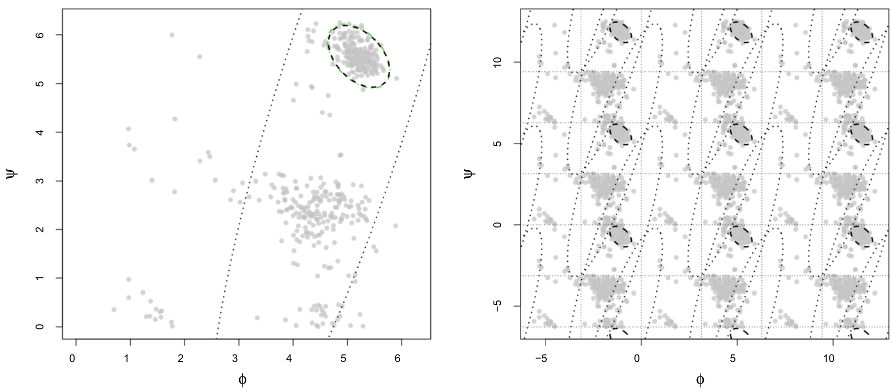

Fig. 14 8TIM protein data. Left panel: Ramachandran plot. Right panel: unwrapped data on a flat torus. 99% Tolerance ellipses over imposed: robust fit (dashed line), maximum likelihood estimation (dotted line)

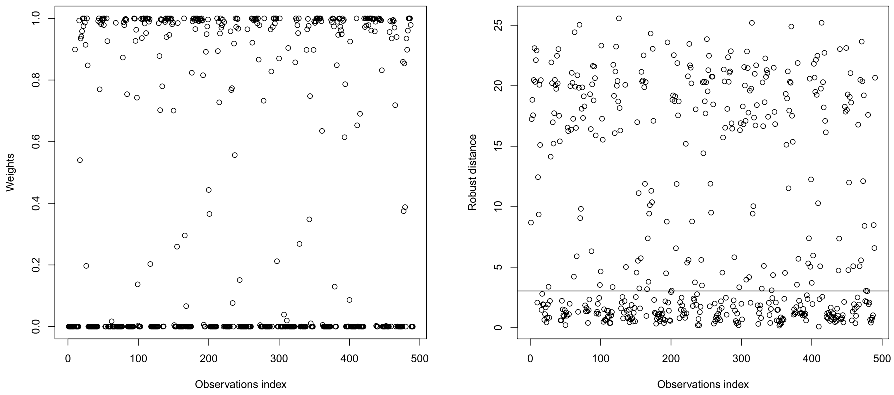

Fig. 15 8TIM protein data. Left panel: weights. Right panel: robust distances. The horizontal line gives the square root of the 0.99-level quantile of the χ^2_2 distribution

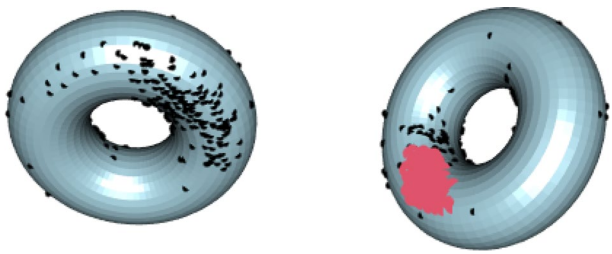

Fig. 16 8TIM protein data. Bivariate angles as points on the surface of a torus from two different perspectives: genuine observations correspond to (red) larger dots, the remaining are outliers (color figure online)

Actually, the robust analyses give strong indication of the presence of several clusters: they all disclose the presence of different structures, otherwise undetectable by maximum likelihood estimation. The tolerance ellipses corresponding to the robustly fitted WN distribution enclose those points in the most dense area, whereas the others are severely down-weighted. There is strong agreement with the findings from the analysis in Chakraborty and Wong (2021). In the left panel of Fig. 15 we displayed the weights from the WCEM-unwrap algorithm. According to an outliers detection testing rule performed at a significance level $\alpha = 0.01$, the actual rate of contamination is about 46%. The right panel of Fig. 15 shows the corresponding distance plot based on robust distances. The horizontal line gives the (square root) $\chi^2_{0.99,2}$ cut-off. Figure 16 shows genuine points and outliers on the torus.

The clustered structure of the data suggested by the outcome of the robust analyses can be further explored using a monitoring plot of the weights as the bandwidth h varies on a chosen grid of values. In this example, the bandwidth matrix is $H = \text{diag}(h^2)$. The vertical line gives the bandwidth actually used. The dark trajectories in Fig. 17 correspond to those points receiving a large weight

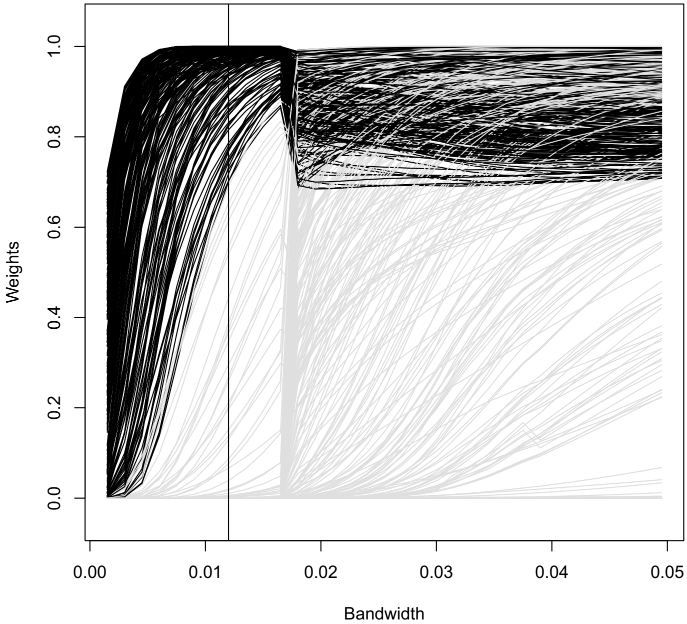

Fig. 17 8TIM protein data. Monitoring plot of weights from the robust fit. The vertical line gives the selected bandwidth value

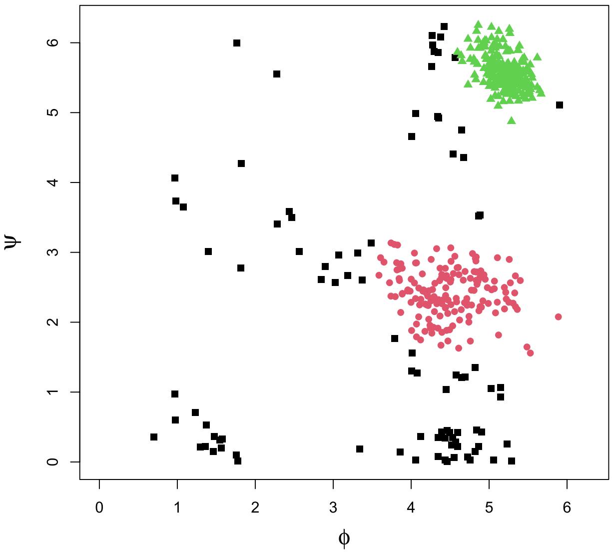

Fig. 18 8TIM protein data. Model based clustering

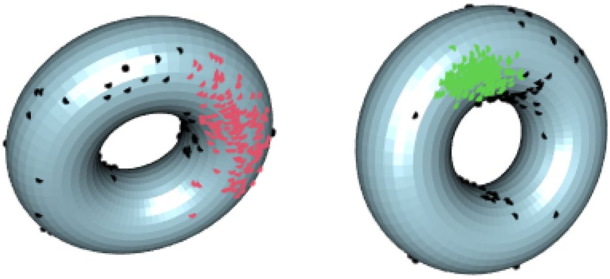

Fig. 19 8TIM protein data. Model based clustering on the torus

in the robust analysis, whereas the gray lines refer to the other points. For small values of the bandwidth h, at least two groups can be detected. As h increases, we notice a transition from the robust to a nonrobust fit since, many other observations are attached large weights and the size of global down-weighting reduces. In particular, some data points exhibit very steep trajectories, as they are no more down-weighted from some point ahead. This behavior suggest the presence of a second group of observations. A closer look at Fig. 17 also unveils a third group, which is composed by those points whose weight is still low for large values of the bandwidth on the right end part of the plot. These points highlight features that are not assimilable to the previous groups. Hence, the robust analysis indicates at least three groups. This finding is confirmed by the results stemming from a proper model based clustering of the torus data at hand (Greco et al. 2022), whose cluster assignments are shown in Figs. 18 and 19.

7.2 RNA data

RNA is assembled as a chain of nucleotides that constitutes a single strand folded onto itself. A nucleotide contains the five-carbon sugar deoxyribose, a nucleobase, that is a nitrogenous base, and one phosphate group. Then, each nucleotide in RNA molecules presents seven torsion angles: six dihedral angles and one angle for the base. Data have been taken from the large RNA data set (Wadley et al. 2007). Here, we consider a subsample of size $n = 260$, obtained after joining data from two distinct clusters, whose sizes are 232 and 28, respectively, and we neglect the information about group labels in the fitting process. Since, the sizes of two clusters are very unbalanced, a feasible robust method is expected to fit the majority of the data belonging to the larger cluster and to lead to detect the data from the smaller cluster as outliers, as they share a different pattern. Figure 20 gives the distance plot from WCEM-torus, WCEM-unwrap and WCEM-dist, under the WN model. We do not appreciate noticeable differences among the results. Each technique leads to detect the smaller group, denoted by black dots. Actually, in this case, the outcome from the robust analysis allows to cope with an unsupervised classification problem and to discriminate between the two groups, with a satisfactory balance between swamping and power.

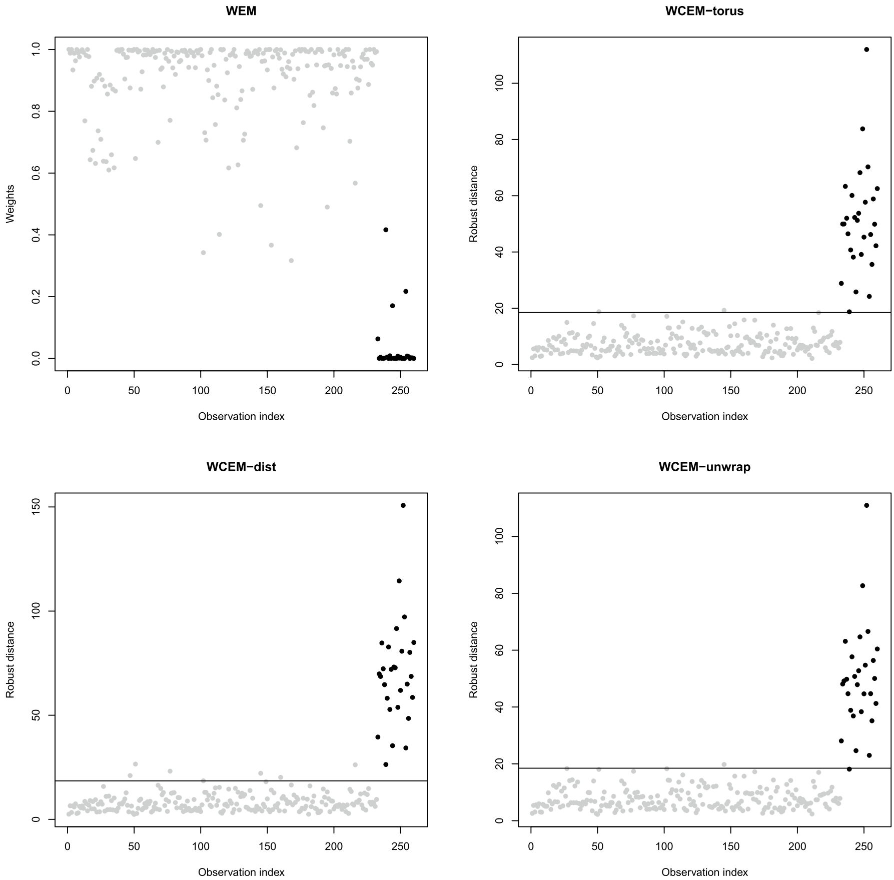

Fig. 20 RNA data. Weights returned by the WEM (left-right panle). Squared distance plots for the different weighting scheme as given by the WCEM-torus, WCEM-unwrap and WCEM-dist (clockwise in the other panels). Black dots give points from the smaller *outlying* cluster. The horizontal line gives the 0.99-level quantile of the χ^2_7 distribution

Appendix A: MLE for wrapped unimodal elliptically symmetric distributions

Let us consider the circular model

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

where

$$m(\boldsymbol{x}; \boldsymbol{\theta}) \propto |\Sigma|^{-1/2} h\big((\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\big)$$

is a unimodal elliptically symmetric distribution. The log-likelihood function based on an i.i.d. sample $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ is

$$
\begin{aligned}
\ell^\circ(\boldsymbol{\mu}, \Sigma) &= \sum_{i=1}^n \log m^\circ(\boldsymbol{y}_i; \boldsymbol{\mu}, \Sigma) \\
&= \sum_{i=1}^n \log \sum_{\boldsymbol{j}\in\mathbb{Z}^p} m(\boldsymbol{y}_i + 2\pi\boldsymbol{j}; \boldsymbol{\mu}, \Sigma) \\
&\propto \sum_{i=1}^n \log \sum_{\boldsymbol{j}\in\mathbb{Z}^p} |\Sigma|^{-\frac{1}{2}} h\big[(\boldsymbol{y}_i + 2\pi\boldsymbol{j} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{y}_i + 2\pi\boldsymbol{j} - \boldsymbol{\mu})\big] \\
&= \frac{n}{2} \log|\Sigma^{-1}| + \sum_{i=1}^n \log \sum_{\boldsymbol{j}\in\mathbb{Z}^p} h\big[\mathrm{tr}\big((\boldsymbol{y}_i + 2\pi\boldsymbol{j} - \boldsymbol{\mu})(\boldsymbol{y}_i + 2\pi\boldsymbol{j} - \boldsymbol{\mu})^\top \Sigma^{-1}\big)\big]
\end{aligned}
$$

Recall that for given square matrices A and B, both symmetric and positive definite we have that

1. $\nabla_A \mathrm{tr}(BA) = B^\top$,
2. $\nabla_A \log(|A|) = \left(A^{-1}\right)^\top$,
3. $\nabla_{\boldsymbol{x}}(\boldsymbol{x}^\top A \boldsymbol{x}) = 2A\boldsymbol{x}$.

Let $d_{ij}(\boldsymbol{\mu}, \Sigma) = (\boldsymbol{y}_i + 2\pi\boldsymbol{j} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{y}_i + 2\pi\boldsymbol{j} - \boldsymbol{\mu})$. Taking the derivatives w.r.t. $\boldsymbol{\mu}$ and Σ^{-1}, the likelihood equations are

$$
\begin{aligned}
\nabla_{\boldsymbol{\mu}} \ell^\circ(\boldsymbol{\mu}, \Sigma) &= \sum_{i=1}^n \nabla_{\boldsymbol{\mu}} \log \sum_{\boldsymbol{j}\in\mathbb{Z}^p} h(d_{ij}(\boldsymbol{\mu}, \Sigma)) \\
&= \sum_{i=1}^n \frac{\sum_{\boldsymbol{j}\in\mathbb{Z}^p} \nabla_{\boldsymbol{\mu}} h(d_{ij}(\boldsymbol{\mu}, \Sigma))}{\sum_{\boldsymbol{k}\in\mathbb{Z}^p} h(d_{ik}(\boldsymbol{\mu}, \Sigma))} \\
&= 2\sum_{i=1}^n \frac{\sum_{\boldsymbol{j}\in\mathbb{Z}^p} h'(d_{ij}(\boldsymbol{\mu}, \Sigma)) \Sigma^{-1} (\boldsymbol{y}_i + 2\pi\boldsymbol{j} - \boldsymbol{\mu})}{\sum_{\boldsymbol{k}\in\mathbb{Z}^p} h(d_{ik}(\boldsymbol{\mu}, \Sigma))}
\end{aligned}
$$

and

$$
\begin{aligned}
\nabla_{\Sigma^{-1}} \ell^\circ(\boldsymbol{\mu}, \Sigma) &= \frac{n}{2}\Sigma^\top + \sum_{i=1}^n \nabla_{\Sigma^{-1}} \log \sum_{\boldsymbol{j}\in\mathbb{Z}^p} h(d_{ij}(\boldsymbol{\mu}, \Sigma)) \\
&= \frac{n}{2}\Sigma + \sum_{i=1}^n \frac{\sum_{\boldsymbol{j}\in\mathbb{Z}^p} \nabla_{\Sigma^{-1}} h(d_{ij}(\boldsymbol{\mu}, \Sigma))}{\sum_{\boldsymbol{k}\in\mathbb{Z}^p} h(d_{ik}(\boldsymbol{\mu}, \Sigma))} \\
&= \frac{n}{2}\Sigma + \sum_{i=1}^n \frac{\sum_{\boldsymbol{j}\in\mathbb{Z}^p} h'(d_{ij}(\boldsymbol{\mu}, \Sigma)) (\boldsymbol{y}_i + 2\pi\boldsymbol{j} - \boldsymbol{\mu})(\boldsymbol{y}_i + 2\pi\boldsymbol{j} - \boldsymbol{\mu})^\top}{\sum_{\boldsymbol{k}\in\mathbb{Z}^p} h(d_{ik}(\boldsymbol{\mu}, \Sigma))},
\end{aligned}
$$

where $h'(d) = \partial h(d)/\partial d$. Let

$$v_{ij} = \frac{h'(d_{ij}(\boldsymbol{\mu}, \Sigma))}{\sum_{\boldsymbol{k}\in\mathbb{Z}^p} h(d_{ik}(\boldsymbol{\mu}, \Sigma))}.$$

then, the MLE $(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ is the solution to the (set of) fixed point equations

$$\boldsymbol{\mu} = \frac{\sum_{i=1}^n \sum_{\boldsymbol{j}\in\mathbb{Z}^p} v_{ij}(\boldsymbol{y}_i + 2\pi\boldsymbol{j})}{\sum_{i=1}^n \sum_{\boldsymbol{k}\in\mathbb{Z}^p} v_{ik}}$$

$$\Sigma = -\frac{2}{n}\sum_{i=1}^n \sum_{\boldsymbol{j}\in\mathbb{Z}^p} v_{ij}(\boldsymbol{y}_i + 2\pi\boldsymbol{j} - \boldsymbol{\mu})(\boldsymbol{y}_i + 2\pi\boldsymbol{j} - \boldsymbol{\mu})^\top.$$

The WN distribution corresponds to $h(t) = \exp\left(-\frac{t}{2}\right)$. Since, $h'(t) = -\frac{1}{2}h(d)$ then

$$v_{ij} = -\frac{1}{2}\frac{h(d_{ij})}{\sum_{\boldsymbol{k}\in\mathbb{Z}^p} h(d_{ik})} = -\frac{1}{2}\frac{m(\boldsymbol{y}_i + 2\pi\boldsymbol{j}; \boldsymbol{\mu}, \Sigma)}{\sum_{\boldsymbol{k}\in\mathbb{Z}^p} m(\boldsymbol{y}_i + 2\pi\boldsymbol{k}; \boldsymbol{\mu}, \Sigma)}.$$

and the estimating equations simplify to

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^n \sum_{\boldsymbol{j}\in\mathbb{Z}^p} \omega_{ij}(\boldsymbol{y}_i + 2\pi\boldsymbol{j})$$

$$\Sigma = \frac{1}{n}\sum_{i=1}^n \sum_{\boldsymbol{j}\in\mathbb{Z}^p} \omega_{ij}(\boldsymbol{y}_i + 2\pi\boldsymbol{j} - \boldsymbol{\mu})(\boldsymbol{y}_i + 2\pi\boldsymbol{j} - \boldsymbol{\mu})^\top .$$

with

$$\omega_{ij} = \frac{m(\boldsymbol{y}_i + 2\pi\boldsymbol{j}; \boldsymbol{\mu}, \Sigma)}{\sum_{\boldsymbol{k}\in\mathbb{Z}^p} m(\boldsymbol{y}_i + 2\pi\boldsymbol{k}; \boldsymbol{\mu}, \Sigma)}.$$

Appendix B: EM algorithm for WN estimation

Given, an i.i.d. sample $(\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)$ from a WN distribution, in the EM algorithm the wrapping coefficients $\boldsymbol{j}$ are considered as latent variables and the observed torus data $\boldsymbol{y}_i$s as being incomplete, that is $\boldsymbol{y}_i$ assumed to be one component of the pair $(\boldsymbol{y}_i, \boldsymbol{\omega}_i)$, where $\boldsymbol{\omega}_i = (\omega_{ij} : \boldsymbol{j} \in \mathbb{Z}^p)$ is the associated latent wrapping coefficients label vector. Then, the MLE for $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma)$ is the result of the EM algorithm based on the complete log-likelihood function

$$\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{\boldsymbol{j}\in\mathbb{Z}^p} \omega_{ij} \log m(\boldsymbol{y}_i + 2\pi\boldsymbol{j}; \boldsymbol{\theta}). \tag{25}$$

In the expectation step (E-step), we evaluate the conditional expectation of (25) given the observed data and the current parameters value $\boldsymbol{\theta}$ by computing the conditional probability that $\boldsymbol{y}_i$ has $\boldsymbol{j}$ as wrapping coefficients vector, that is

$$\omega_{ij} = \frac{m(\boldsymbol{y}_i + 2\pi\boldsymbol{j};\boldsymbol{\theta})}{\sum_{\boldsymbol{k}\in\mathbb{Z}^p} m(\boldsymbol{y}_i + 2\pi\boldsymbol{k};\boldsymbol{\theta})}, \forall\boldsymbol{j}\in\mathbb{Z}^p.$$

parameters estimation is carried out in the maximization step (M-step) solving the set of (complete) likelihood equations

$$\sum_{i=1}^{n}\sum_{\boldsymbol{j}\in\mathbb{Z}^p}\omega_{ij}u(\boldsymbol{y}_i + 2\pi\boldsymbol{j};\boldsymbol{\theta}) = \boldsymbol{0}$$

with $u(\boldsymbol{y}_i + 2\pi\boldsymbol{j};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log m(\boldsymbol{y} + 2\pi\boldsymbol{j};\boldsymbol{\theta})$. An alternative estimation strategy can be based on a CEM algorithm leading to an approximated solution. At each iteration, a Classification step (C-step) is performed after the E-step, that provides crispy assignments. Let

$$\hat{\boldsymbol{j}}_i = \text{argmax}_{\boldsymbol{j}\in\mathbb{Z}^p}\omega_{ij},$$

then, set $\omega_{ij} = 1$ when $\boldsymbol{j} = \hat{\boldsymbol{j}}_i$, $\omega_{ij} = 0$ otherwise. As a result, the torus data $\boldsymbol{y}_i$ are *unwrapped* to (fitted) linear data $\hat{\boldsymbol{x}}_i = \boldsymbol{y}_i + 2\pi\hat{\boldsymbol{j}}_i$. It is easy to see that the M-step simplifies to

$$\sum_{i=1}^{n} u(\hat{\boldsymbol{x}}_i;\boldsymbol{\theta}) = \boldsymbol{0}.$$

both the procedures are iterated until some convergence criterion is fulfilled, that could be based on the changes in the likelihood or in fitted parameter values (Nodehi et al. 2021).

Acknowledgements The authors wish to thank the Associate Editor who supported and encouraged the reviewing process and two anonymous referees whose comments helped improving the quality of the paper.

References

Agostinelli, C.: Robust estimation for circular data. Comput. Stat. Data Anal. **51**(12), 5867–5875 (2007)

Agostinelli, C., Greco, L.: 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)

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

Azzalini, A., Menardi, G.: Clustering via nonparametric density estimation: the R package pdf Cluster. J. Stat. Softw. **57**(11), 1–26 (2014)

Bahlmann, C.: Directional features in online handwriting recognition. Pattern Recognit. **39**(1), 115–125 (2006)

Baltieri, D., Vezzani, R., Cucchiara, R.: People orientation recognition by mixtures of wrapped distributions on random trees. In: European Conference on Computer Vision, Springer, pp. 270–283 (2012)

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

Beran, R.: Minimum hellinger distance estimates for parametric models. Ann. Stat., pp. 445–463 (1977)

Bourne, P.E.: The protein data bank. Nucleic Acids Res. **28**, 235–242 (2000)

Chakraborty, S., Wong, S.W.K.: BAMBI: an R package for fitting bivariate angular mixture models. J. Stat. Softw. **99**(11), 1–69 (2021)
Chang, M., Artymiuk, P., Wu, X., et al.: Human triosephosphate isomerase deficiency resulting from mutation of phe-240. Am J Hum Genet **52**, 1260 (1993)
Coles, S.: Inference for circular distributions and processes. Stat. Comput. **8**(2), 105–113 (1998)
Cremers, J., Klugkist, I.: One direction? A tutorial for circular data analysis using r with examples in cognitive psychology. Front. Psychol., p. 2040 (2018)
Davies, P.L., Gather, U.: Breakdown and groups. Ann. Stat. **33**(3), 977–1035 (2005)
Davies, P.L., Gather, U.: Addendum to the discussion of "breakdown and groups". Ann. Stat., pp. 1577–1579 (2006)
Eltzner, B., Huckermann, S., Mardia, K.: Torus principal component analysis with applications to RNA structure. Ann. Appl. Stat. **12**(2), 1332–1359 (2018)
Farcomeni, A., Greco, L.: Robust Methods for Data Reduction. CRC Press (2016)
Fisher, N., Lee, A.: Time series analysis of circular data. J. R. Stat. Soc. B **56**, 327–339 (1994)
Greco, L., Agostinelli, C.: Weighted likelihood mixture modeling and model-based clustering. Stat. Comput. **30**(2), 255–277 (2020)
Greco, L., Lucadamo, A., Agostinelli, C.: Weighted likelihood latent class linear regression. Stat. Methods Appl., pp. 1–36 (2020)
Greco, L., Saraceno, G., Agostinelli, C.: Robust fitting of a wrapped normal model to multivariate circular data and outlier detection. Stats **4**(2), 454–471 (2021)
Greco, L., Novi Inverardi, P., Agostinelli, C.: Finite mixtures of multivariate wrapped normal distributions for model based clustering of p-torus data. J. Comput. Graph. Stat. **32**(3), 1215–1228 (2022)
He, X., Simpson, D.G.: Robust direction estimation. Ann. Stat. **20**(1), 351–369 (1992)
Huber, P., Ronchetti, E.: Robust Statistics. Wiley, London (2009)
Jammalamadaka, S., SenGupta, A.: Topics in Circular Statistics, Multivariate Analysis, vol. 5. World Scientific, Singapore (2001)
Jona Lasinio, G., Gelfand, A., Jona Lasinio, M.: Spatial analysis of wave direction data using wrapped Gaussian processes. Ann. Appl. Stat. **6**(4), 1478–1498 (2012)
Ko, D., Guttorp, P.: Robustness of estimators for directional data. Ann. Stat., pp. 609–618 (1988)
Kurz, G., Gilitschenski, I., Hanebeck, U.D.: Efficient evaluation of the probability density function of a wrapped normal distribution. In: 2014 Sensor Data Fusion: Trends, pp. 1–5. Solutions, Applications (SDF), IEEE (2014)
Lenth, R.V.: Robust measures of location for directional data. Technometrics **23**(1), 77–81 (1981)
Lindsay, B.: Efficiency versus robustness: the case for minimum hellinger distance and related methods. Ann. Stat. **22**, 1018–1114 (1994)
Lund, U.: Cluster analysis for directional data. Commun. Stat. Simul. Comput. **28**(4), 1001–1009 (1999)
Mardia, K.: Statistics of Directional Data. Academic Press (1972)
Mardia, K., Jupp, P.: Directional Statistics. Wiley, New York (2000)
Mardia, K., Taylor, C., Subramaniam, G.: Protein bioinformatics and mixtures of bivariate von mises distributions for angular data. Biometrics **63**(2), 505–512 (2007)
Mardia, K., Kent, J., Zhang, Z., et al.: Mixtures of concentrated multivariate sine distributions with applications to bioinformatics. J. Appl. Stat. **39**(11), 2475–2492 (2012)
Mardia, K.V., Frellsen, J.: Statistics of bivariate von mises distributions. In: Bayesian Methods in Structural Bioinformatics. Springer, p. 159–178 (2012)
Mardia, K.V., Jupp, P.E.: Directional Statistics. Wiley Online Library (2000b)
Markatou, M., Basu, A., Lindsay, B.G.: Weighted likelihood equations with bootstrap root search. J. Am. Stat. Assoc. **93**(442), 740–750 (1998)
Maronna, R.A., Martin, R.D., Yohai, V.J., et al.: Robust Statistics: Theory and Methods (with R). Wiley, London (2019)
Munkres, J.R.: Elements of Algebraic Topology. CRC Press (2018)
Nodehi, A., Golalizadeh, M., Maadooliat, M., et al.: Estimation of parameters in multivariate wrapped models for data on ap-torus. Comput. Stat. **36**, 193–215 (2021)
Park, C., Basu, A.: The generalized Kullback–Leibler divergence and robust inference. J. Stat. Comput. Simul. **73**(5), 311–332 (2003)
Park, C., Basu, A., Lindsay, B.: The residual adjustment function and weighted likelihood: a graphical interpretation of robustness of minimum disparity estimators. Comput. Stat. Data Anal. **39**(1), 21–33 (2002)
Pewsey, A., Neuhäuser, M., Ruxton, G.: Circular Statistics in R. Oxford University Press, Oxford (2013)

Prestele, C.: Credit portfolio modelling with elliptically contoured distributions. Ph.D. thesis, Institute for Finance Mathematics, University of Ulm (2007)
R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2021), https://www.R-project.org/
Ranalli, M., Maruotti, A.: Model-based clustering for noisy longitudinal circular data, with application to animal movement. Environmetrics **31**(2), e2572 (2020)
Rao, B.: Nonparametric Functional Estimation. Academic Press (2014)
Rivest, L.P., Duchesne, T., Nicosia, A., et al.: A general angular regression model for the analysis of data on animal movement in ecology. J. R. Stat. Soc.: Ser. C (Appl. Stat.) **65**(3), 445–463 (2016)
Rousseeuw, P.J., Hampel, F.R., Ronchetti, E.M., et al.: Robust Statistics: The Approach Based on Influence Functions. Wiley, London (2011)
Rutishauser, U., Ross, I.B., Mamelak, A.N., et al.: Human memory strength is predicted by theta-frequency phase-locking of single neurons. Nature **464**(7290), 903–907 (2010)
Saraceno, G., Agostinelli, C., Greco, L.: Robust estimation for multivariate wrapped models. Metron **79**(2), 225–240 (2021)
Serfling, R.J.: Approximation Theorems of Mathematical Statistics. Wiley, London (2009)
Wadley, L., Keating, K., Duarte, C., et al.: Evaluating and learning from rna pseudotorsional space: quantitative validation of a reduced representation for rna structure. J. Mol. Biol. **372**(4), 942–957 (2007)
Warren, W.H., Rothman, D.B., Schnapp, B.H., et al.: Wormholes in virtual space: from cognitive maps to cognitive graphs. Cognition **166**, 152–163 (2017)