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

Estimation of Multivariate Wrapped Models for Data on a p-Torus

Anahita Nodehi · Mousa Golalizadeh* · Mehdi Maadooliat · Claudio Agostinelli

Received: date / Accepted: date

Abstract Multivariate circular observations, i.e. points on a torus arise frequently in fields where instruments such as compass, protractor, weather vane, sextant or theodolite are used. Multivariate wrapped models are often appropriate to describe data points scattered on $p-$dimensional torus. However, the statistical inference based on such models is quite complicated since each contribution in the log-likelihood function involves an infinite sum of indices in $\mathbb{Z}^p$, where p is the dimension of the data. To overcome this problem, for moderate dimension p, we propose two estimation procedures based on Expectation-Maximisation and Classification Expectation-Maximisation algorithms. We study the performance of the proposed techniques on a Monte Carlo simulation and further illustrate the advantages of the new procedures on three real-world data sets.

Keywords CEM Algorithm · EM Algorithm · Estimation Procedures · Multivariate Wrapped Distributions · Torus.

1 Introduction

There are many problems in applied sciences where a quantity of interest is measured as a direction. Mardia (1972) is one of the first references in this field describing how to deal with this kind of data in many subjects, e.g. physics, psychology,

Anahita Nodehi
Department of Statistics, Tarbiat Modares University, Tehran, Iran

* Corresponding author: Mousa Golalizadeh
Department of Statistics, Tarbiat Modares University, Tehran, Iran.
Tel. +98-21-82884705
golalizadeh@modares.ac.ir,

Mehdi Maadooliat
Department of Mathematical and Statistical Sciences, Marquette University, Milwaukee, USA

Claudio Agostinelli
Department of Mathematics, University of Trento, Trento, Italy

image analysis, medicine and astronomy. As a simple example of directional data, one may consider a unit vector of length n. Clearly, such a vector can represent an angle on a unit n-sphere, provided we choose an initial direction and orientation for the n-sphere. This type of data are often referred to as circular (directional) data. One important aspect of the directional data is that they cannot be analysed using standard methods/models developed in the Euclidean space. Mardia and Jupp (2000) and Jammalamadaka and SenGupta (2001) are two commonly used resources that provide a comprehensive review of tools and techniques used in the directional (circular) statistics. Another important reference is Batschelet (1981).

The Wrapped Normal and the von Mises distributions are two important probability distributions defined on the unit circle. They often play a similar role as the Normal model on the Euclidean space. For instance, the von Mises distribution belongs to the Exponential family and it is a natural circular analogue of the univariate Normal distribution, when the variability in the circular domain is small. As for the multivariate von Mises distribution, its conditional distributions are also von Mises while its marginal distributions are not (see, e.g. Mardia and Jupp, 2000, p. 55). The Wrapped Normal is another circular distribution similar to the univariate Normal. It is symmetric and obtained by wrapping a Normal distribution around the unit circle. Although it does not belong to the Exponential family, the convolution of two Wrapped Normal variables is also Wrapped Normal (Jammalamadaka and SenGupta, 2001). The associated conditional and marginal distributions are Wrapped Normal, too. The Wrapped Normal distribution also appears in the central limit theorem on the unit circle and in connection with Brownian Motion on the unit circle see, (see, Stephens, 1963, for more details).

The von Mises distribution is perhaps more famous, and its fame is due to the analytical tractability of the maximum likelihood estimators (MLEs) in univariate framework. Despite this, deriving MLEs for the multivariate case is still an open problem. In some recent works, Mardia et al (2007, 2008) and Mardia (2010) introduced applications of bivariate and trivariate von Mises distributions. Furthermore, Mardia and Voss (2014) studied some properties of the multivariate von Mises distribution but statistical inference requires evaluation of a quite complex estimation scheme (Mardia et al, 2008). These are some of the reasons that one may investigate utilising the Wrapped Normal distribution as an alternative to the von Mises in the multivariate setting. Moreover, for most of the distributions versus their wrapped versions, there is a correspondence between the non-wrapped (line) parameters and the associated wrapped ones. This allows for direct interpretation of the circular parameters as well as the respective inference results.

The bivariate Wrapped Normal distribution is proposed by Johnson and Wehrly (1978) while multivariate Wrapped Normal distribution is presented in Baba (1981). Estimation of the Wrapped Normal parameters even in univariate case leads to a complex numerical optimisation, since evaluation of the associated likelihood function requires dealing with an infinite series. That is why some authors, e.g. Fisher (1987) and Breckling (1989) proposed approximating this distribution by the von Mises distribution. Kent (1978) showed that, the von Mises and Wrapped Normal distributions can be well approximated by one another. Agostinelli (2007) proposed an iterative reweighted maximum likelihood (IRML) estimating equations algorithm for univariate Wrapped Normal which is also available in the R

package `circular` (Agostinelli and Lund, 2017). Fisher and Lee (1994) used the Expectation-Maximisation (EM) algorithm to obtain parameter estimates from the Wrapped Normal distribution for an autoregressive model with low order. The E-step involves ratios of large infinite sums, which need to be approximated at each step, making the algorithm computationally inefficient. Moreover, Coles (1998), Ravindran and Ghosh (2011) and Ferrari (2009) adopted a data augmentation approach to estimate the missing unobserved wrapping coefficients and the other parameters in a Bayesian framework.

As mentioned before, the main difficulty of working with the Wrapped Normal distribution is that the density function consists of an infinite sum, which does not allow the exact evaluation. The likelihood-based inference for such distribution can be very complicated and computationally intensive. That leads to our contribution in this work on the estimation problem for both univariate and multivariate Wrapped Normal distributions using two innovative algorithms.

The remainder of this paper is organised as follows. Section 2 describes the multivariate wrapped model in a general framework. Section 3 introduces two new algorithms based on Expectation-Maximisation and Classification Expectation-Maximisation methods for the estimation of the parameters when dealing with the wrapped multivariate normal model. These approaches can easily be extended to other multivariate wrapped models. This section also includes a description of a method to obtain initial values and a discussion on how to extend the proposed techniques to the case we are simultaneously dealing with circular and non-circular observations. Section 4 reports the results of an extensive Monte Carlo experiment, while Section 5 provides two illustrative examples based on real-world datasets; Section 6 gives final comments and remarks. A third example, further illustrative results on the two real-world datasets and a complete summary of the Monte Carlo experiment can be found in the Supplementary Material.

2 Multivariate Wrapped Normal distribution

Wrapping consists of the geometric translation of a standard distribution to a space defined on a *circular* domain, e.g., a unit circle. In other words, a distribution with support on the entire real-line is translated to one with support on a circle of finite circumference. A rich class of distributions on the unit circle can be obtained using the wrapping technique. The procedure is as follows: given a random variable (r.v.) X defined on $\mathbb{R}$, $Y = X \mod 2\pi$ is a r.v. on the unit circle, by accumulating the probability densities over all points $X = (Y + 2\pi j)$ where $j \in \mathbb{Z}$. If G represents the cumulative distribution function (CDF) on $\mathbb{R}$, the resulting wrapped distribution F on the unit circle is given by

$$F(y) = \sum_{j=-\infty}^{+\infty} \left[G(y + 2\pi j) - G(2\pi j)\right] , \qquad y \in (0, \pi] .$$

In particular, for any r.v. X with the density function g, and support on $\mathbb{R}$, the r.v. Y with a circular density function f and support on $(0, 2\pi]$ can be defined as

$$f(y) = \sum_{j=-\infty}^{+\infty} g(y + 2\pi j) , \qquad y \in (0, \pi] .$$

By this translation, both discrete and continuous wrapped distributions can be constructed (Mardia and Jupp, 2000). Among the continuous wrapped distributions, the Wrapped Normal and the Wrapped Cauchy play an important role in data analysis. A Wrapped Normal distribution, denoted $WN(\mu, \sigma^2)$, is obtained by wrapping a $N(\mu, \sigma^2)$ distribution around the unit circle. This distribution is unimodal and symmetric about the mean μ. The mean resultant length ρ (see, e.g. Mardia and Jupp, 2000, pp. 28-29 and 50), which is the length of the first trigonometric moment, is given by $\rho = \exp[-\sigma^2/2]$ for the Wrapped Normal distribution. As $\rho \to 0$ the distribution converges to the uniform distribution on the unit circle while as $\rho \to 1$, it tends to a point mass distribution at μ.

For the present paper we concentrate on the multivariate Wrapped Normal distribution which is obtained by component-wise wrapping of a p-variate Normal distribution on a p-dimensional torus (often called the p-torus or hypertorus for short). In general a multivariate wrapped distribution can be obtained as follows. Let $\boldsymbol{X} \sim G$ represents the r.v. on $\mathbb{R}^p$ space, then the resulting wrapped r.v. $\boldsymbol{Y} \sim F$ on the p-torus has a CDF, F, given by

$$F(\boldsymbol{y}) = \sum_{\boldsymbol{j} \in \mathbb{Z}^p} [G(\boldsymbol{y} + 2\pi \boldsymbol{j}) - G(2\pi \boldsymbol{j})] \ , \qquad \boldsymbol{y} \in (0, 2\pi]^p \ ,$$

where the sum is extended to all vectors $\boldsymbol{j} \in \mathbb{Z}^p$. If $\boldsymbol{X}$ has the density function g defined on $\mathbb{R}^p$, then $\boldsymbol{Y}$ has a density function f on the p-torus given by

$$f(\boldsymbol{y}) = \sum_{\boldsymbol{j} \in \mathbb{Z}^p} g(\boldsymbol{y} + 2\pi \boldsymbol{j}) \ , \qquad \boldsymbol{y} \in (0, 2\pi]^p.$$

For the multivariate Normal distribution, let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the random vector $\boldsymbol{Y} = \boldsymbol{X} \mod 2\pi$ is called a multivariate Wrapped Normal distribution $WN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where the modulus operator is applied component-wise.

The bivariate Wrapped Normal distribution is proposed by Johnson and Wehrly (1978) and multivariate Wrapped Normal distribution is introduced by Baba (1981). Evaluation of the Wrapped Normal density function can be difficult, even in univariate case, because it involves an infinite series.

2.1 Equivariance in Wrapped Normal models

Let $\boldsymbol{X}$ be a p-variate random vector. For a given $\boldsymbol{b}$, a vector of length p, and a full rank $p \times p$ matrix $\boldsymbol{A}$ consider the affine transformation $\boldsymbol{W} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$. A location estimate T is affine equivariant if $T(\boldsymbol{W}) = \boldsymbol{A}T(\boldsymbol{X}) + \boldsymbol{b}$, while a scatter estimate S is affine equivariant if $S(\boldsymbol{W}) = \boldsymbol{A}S(\boldsymbol{X})\boldsymbol{A}^\top$. Now let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ so that $\boldsymbol{W} \sim N_p(\boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W)$, where $\boldsymbol{\mu}_W = \boldsymbol{A}\boldsymbol{\mu}_X + \boldsymbol{b}$ and $\boldsymbol{\Sigma}_W = \boldsymbol{A}\boldsymbol{\Sigma}_X\boldsymbol{A}^\top$. Define $\boldsymbol{U} = \boldsymbol{X} \mod 2\pi$ and $\boldsymbol{V} = \boldsymbol{W} \mod 2\pi$ as two multivariate Wrapped Normal models on the p-torus. In Section SM–1 of the Supplementary Material we show that the likelihood $L(\boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W | \boldsymbol{v}_1, \cdots, \boldsymbol{v}_n)$ of the parameters $\boldsymbol{\mu}_W$, $\boldsymbol{\Sigma}_W$ based on the samples $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n$ is proportional to $L(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X | \boldsymbol{u}_1^*, \cdots, \boldsymbol{u}_n^*)$, where $\boldsymbol{u}_i^*$ is a sample from $\boldsymbol{U}^* = \boldsymbol{X} \mod (2\pi \boldsymbol{A}^{-1}\boldsymbol{j})$ and $\boldsymbol{j}$ is a length p vector of ones. An immediate result is that $L(\boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W | \boldsymbol{v}_1, \cdots, \boldsymbol{v}_n)$ is not proportional to $L(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X | \boldsymbol{u}_1, \cdots, \boldsymbol{u}_n)$. This fact shows that MLEs are not affine equivariant for the multivariate Wrapped Normal model.

3 Parameters estimation

Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be an independent and identically distributed (i.i.d.) random sample from a multivariate Wrapped Normal model on the p-torus with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{Y} \sim WN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. As discussed in previous section, we can consider $\boldsymbol{y}_i$, equivalently, as $\boldsymbol{y}_i = \boldsymbol{x}_i \mod 2\pi$ where $\boldsymbol{x}_i$ is a sample from, i.e. $\boldsymbol{X}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The log-likelihood of the unknown parameters $\boldsymbol{\Omega} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ of a multivariate Wrapped Normal model is represented by

$$\ell(\boldsymbol{\Omega}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \sum_{i=1}^{n} \log \left[\sum_{\boldsymbol{j} \in \mathbb{Z}^p} \phi_p(\boldsymbol{y}_i + 2\pi \boldsymbol{j}; \boldsymbol{\Omega}) \right] , \tag{1}$$

where ϕ_p is the multivariate Normal density in $\mathbb{R}^p$ and $\boldsymbol{j}$ is a vector of indices in $\mathbb{Z}^p$. For the univariate case Agostinelli (2007) proposed an iteratively reweighted least squares (IRLS) algorithm to maximise the log-likelihood function. The details of this method are provided in Agostinelli (2007) and the implementation is available in the function `mle.wrappednormal` in the R (R Core Team, 2019) package `circular` (Agostinelli and Lund, 2017). For the multivariate case ($p > 1$), a similar approach seems infeasible and alternative techniques are required. A direct maximisation of the log-likelihood function for small to moderate dimensional problems, says $p \leq 5$, is possible provided an appropriate parametrisation of the model is considered. For instance, the variance-covariance matrix $\boldsymbol{\Sigma}$ can be reparametrised as described in Pinheiro and Bates (1996). In this work we use the log-Cholesky parameterisation, which allows for unconstrained optimisation while ensuring the positive definiteness of the estimated $\boldsymbol{\Sigma}$ is achieved. Let $\boldsymbol{\sigma}$ be the set of $p(p+1)/2$ parameters that we introduce to represent $\boldsymbol{\Sigma}$ uniquely, and let $\boldsymbol{\Sigma}(\boldsymbol{\sigma}) = \boldsymbol{R}(\boldsymbol{\sigma})^\top \boldsymbol{R}(\boldsymbol{\sigma})$ be the associated Cholesky decomposition. The $p \times p$ upper triangular matrix $\boldsymbol{R}(\boldsymbol{\sigma})$ is full rank and parameterised via $\boldsymbol{\sigma}$. To ensure that the diagonal elements of $\boldsymbol{\Sigma}$ are positive, we incorporate the logarithms of the diagonal elements of $\boldsymbol{R}(\boldsymbol{\sigma})$ in our parameterisation. Besides direct maximisation, we propose two algorithms, based on Expectation-Maximisation (EM) method (Dempster et al, 1977), to maximise (1) in subsections 3.1, and 3.2.

Fisher and Lee (1994) used the EM algorithm to obtain parameter estimates of an autoregressive model for circular data with Wrapped Normal distribution. Their procedure suffers from high computational complexity, as it involves ratios of large multivariate infinite sums in each iteration of the E-step, and that makes the algorithm computationally inefficient. Alternatively, Coles (1998), Ravindran and Ghosh (2011) and Ferrari (2009) adopted a data augmentation approach to estimate the missing unobserved wrapping coefficients in a Bayesian framework which is not our main objective in this paper.

3.1 EM algorithm

Instead of the log-likelihood in Equation (1), the EM algorithm maximises the complete log-likelihood function given by

$$\ell_C(\boldsymbol{\Omega}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \sum_{i=1}^{n} \log \left[\sum_{\boldsymbol{j} \in \mathbb{Z}^p} v_{i\boldsymbol{j}} \phi_p(\boldsymbol{y}_i + 2\pi \boldsymbol{j}; \boldsymbol{\Omega}) \right] , \tag{2}$$

where $v_{i\boldsymbol{j}}$ is an indicator of the ith unit having the $\boldsymbol{j}$ vector as wrapping coefficients. The algorithm alternates between two steps: Expectation (E) and Maximisation (M) as follows.

- **E** step: In this step we compute the conditional expectation of the complete log-likelihood function by setting $v_{i\boldsymbol{j}}$ equal to the probability that $\boldsymbol{y}_i$ has $\boldsymbol{j}$ as wrapping coefficients, i.e.

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

- **M** step: In this step we update the estimates of $\boldsymbol{\Omega}$ by maximising the complete log-likelihood conditional on $v_{i\boldsymbol{j}}$'s, $i = 1, \ldots, n$ given in **E** step.

In practical implementation, $\mathbb{Z}^p$ 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 large enough J (see, e.g. Mardia and Jupp, 2000, pp. 50).

Note that the complete log-likelihood is non-decreasing at each iteration of the EM algorithm, and hence the algorithm converges to a local maximum. See Section SM–2 of the Supplemental Material for details. In practice we declare convergence of the algorithm if both $\max_{r=1,\ldots,p}(2(1-\cos(\mu_r^{(k+1)} - \mu_r^{(k)})))^{1/2}$ and $\max_{r,s=1,\ldots,p}(|\sigma_{rs}^{(k+1)} - \sigma_{rs}^{(k+1)}|)$ are smaller than a fixed threshold (10^{-6} in our case), where $\mu_r^{(k)}$ are the elements of the vector $\boldsymbol{\mu}$ at the kth step of the algorithm and $\sigma_{rs}^{(k)}$'s are defined similarly for the elements of the matrix $\boldsymbol{\Sigma}$. An alternative criterion can be obtained based on the increments of the log-likelihood functions between two subsequent iterations.

Since the M step involves a complicated maximisation problem, we introduce a modification based on the law of total variance. We call the new algorithm the Total Variance EM algorithm. For a fixed J let $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\Sigma}^{(k)}$ be the estimates at the kth step of the algorithm; for the ith observation we recompute $\boldsymbol{y}_i$ ($i = 1, \cdots, n$) such that each components of $\boldsymbol{y}_i - \boldsymbol{\mu}^{(k)}$ are expressed in the interval $(-\pi, \pi]$. This translation prevents the use of large values of J in order to have a good approximation. We build a data matrix $\tilde{\boldsymbol{Y}}_i$ of dimension $(2 \times J + 1)^p \times p$ with the row entries of the form

$$\tilde{\boldsymbol{y}}_r = \boldsymbol{y}_i + 2\pi \boldsymbol{J}_r = (y_{i1} + 2\pi j_{r1},\ y_{i2} + 2\pi j_{r2},\ \ldots,\ y_{ip} + 2\pi j_{rp}), \qquad r = 1, \ldots, (2 \times J + 1)^p,$$

where the vector $\boldsymbol{J}_r = (j_{r1}, \cdots, j_{rp})$ is one of the $(2 \times J + 1)^p$ rows of the matrix obtained by the Cartesian product $\times_{s=1}^{p} \mathcal{J}$. Let $\tilde{\boldsymbol{w}}_i$ be a weight vector with entries $\tilde{w}_r = \phi_p(\tilde{\boldsymbol{y}}_r; \boldsymbol{\Omega}^{(k)})$ where $\boldsymbol{\Omega}^{(k)} = (\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})$.

We define $\tilde{\boldsymbol{\mu}}_i$ and $\tilde{\boldsymbol{\Sigma}}_i$ be the weighted sample mean and sample covariance based on the $\tilde{\boldsymbol{Y}}_i$ data matrices and the weight vectors $\tilde{\boldsymbol{w}}_i$. Now let $\boldsymbol{M}$ to be the

matrix with the row entries $\tilde{\boldsymbol{\mu}}_i$, and define $\boldsymbol{C}$ to be the sample covariance of the data matrix $\boldsymbol{M}$. Then, we can update the parameters by

$$\boldsymbol{\mu}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\boldsymbol{\mu}}_i \ ,$$
$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\boldsymbol{\Sigma}}_i + \boldsymbol{C} \ ,$$

where $\tilde{\boldsymbol{\mu}}_i$'s and $\tilde{\boldsymbol{\Sigma}}_i$'s are the conditional means and conditional (within) variance-covariance matrices respectively, while $\boldsymbol{C}$ is the between variance matrix. This algorithm can be easily implemented in a parallel fashion, since updating the parameters associated with the ith observation can be performed independently at each step. Furthermore, updating the elements of the weight vector $\tilde{\boldsymbol{w}}_i$ can also be performed in parallel.

3.2 Classification EM algorithm

An alternative algorithm for the aforementioned estimation problem is the Classification EM (CEM) algorithm (Celeux and Govaert, 1992) where the E step is followed by a C step (classification step) in which $v_{i\boldsymbol{j}}$ is estimated as either 0 or 1, so that the ith observation is associated to the most likely $\boldsymbol{j} \in \mathbb{Z}^p$ vector. In this context the CEM algorithm summarises the complete log-likelihood, given in (2), to the following "classification" log-likelihood function:

$$\ell^C(\boldsymbol{\Omega}, \boldsymbol{j}_1, \ldots, \boldsymbol{j}_n; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \sum_{i=1}^{n} \log \phi_p(\boldsymbol{y}_i + 2\pi\boldsymbol{j}_i; \boldsymbol{\Omega}), \tag{3}$$

in which the $\boldsymbol{j}_i$'s $\in \mathbb{Z}^p$ $(i = 1, \ldots, n)$ are treated as unknown parameters. The procedure at the kth step is then performed as follows.

- **E** step: Same as before we compute $v_{i\boldsymbol{j}}$ by

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

- **C** step: Let $\hat{\boldsymbol{j}}_i = \arg\max_{\boldsymbol{h}\in\mathbb{Z}^p} v_{i\boldsymbol{h}}$;
- **M** step: In this step we update the estimates of $\boldsymbol{\Omega}$ by maximising the classification log-likelihood, given in (3), conditional on $\hat{\boldsymbol{j}}_i$ $(i = 1, \ldots, n)$ given in C step.

In Section SM–2 of the Supplemental Material, we show that the classification log-likelihood function is non-decreasing at each iteration of the CEM algorithm, and hence the algorithm converges to a local optimum.

Similar to the the EM algorithm, in the practical implementation $\mathbb{Z}^p$ is replaced by the Cartesian product $\times_{s=1}^{p} \mathcal{J}$, for some large enough J. The $\hat{\boldsymbol{j}}_i$ plays the role of an offset in the classification log-likelihood and hence the M step is straightforward. Note that, at each iteration, the classification step can provide an estimate of the original unobserved samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, obtained as $\hat{\boldsymbol{x}}_i = \boldsymbol{y}_i + 2\pi\hat{\boldsymbol{j}}_i$, $(i = 1, \ldots, n)$.

3.3 Extending to joint circular and linear variables

An extension to the joint estimation of circular and linear (non-circular) observations can be obtained for both the EM and CEM algorithms. Suppose that

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \sim N_{p_1+p_2}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right).$$

Here our samples are from the joint vector $(\boldsymbol{Y}_1, \boldsymbol{X}_2)$ where $\boldsymbol{Y}_1 = \boldsymbol{X}_1 \mod 2\pi$, that is $\boldsymbol{Y}_1 \sim WN_{p_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$. For the CEM algorithm, first we suggest performing the algorithm as described in the subsection 3.2 to estimate the parameters associated to the first component, $\boldsymbol{Y}_1$. As a by-product we obtain $\hat{\boldsymbol{j}}_i$ and $\hat{\boldsymbol{x}}_{1i}$, $(i = 1, \ldots, n)$, in the C step of the algorithm. The $\hat{\boldsymbol{x}}_{1i}$'s can be used to achieve n pseudo-observed vectors of the length p_1+p_2 in the non-wrapped Gaussian setting, $(\hat{\boldsymbol{x}}_{1i}, \boldsymbol{x}_{2i})$'s. Now we perform the MLE for the multivariate Normal distribution on these pseudo-observed vectors, in order to obtain estimates of the remaining components $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_{12}$ and $\boldsymbol{\Sigma}_{22}$. One may note that, only the estimate of $\boldsymbol{\Sigma}_{12}$ depends on the joint vectors. As for the EM algorithm, similarly we suggest performing the algorithm as illustrated in the subsection 3.1 and obtain the final $\tilde{\boldsymbol{\mu}}_{1i}$, $(i = 1, \ldots, n)$. These $\tilde{\boldsymbol{\mu}}_{1i}$'s can then be used as estimates for the unknown observations, $\boldsymbol{x}_{1i}$'s. Likewise by considering the whole n pseudo-observed vectors of $(\tilde{\boldsymbol{\mu}}_{1i}, \boldsymbol{x}_{2i})$'s, an estimate of $\boldsymbol{\Sigma}_{12}$ can be obtained. The rest of the parameters can be estimated using $\boldsymbol{x}_{1i}$'s and $\boldsymbol{x}_{2i}$'s, separately.

3.4 Initial values

In order to intialize the algorithms introduced in this section, we require appropriate starting values. For these purposes, we suggest using the circular means and $-2\log(\hat{\rho}_r)$, where $\hat{\rho}_r$ is the sample mean resultant length for mean vector $\boldsymbol{\mu}^{(0)}$, and the variances $\sigma_{rr}^{(0)}$ $(r = 1, \ldots, p)$. Following Jammalamadaka and SenGupta (2001) the circular correlation coefficient between two circular samples $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined by

$$\rho_c(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{n} \sin(x_i - \bar{x}) \sin(y_i - \bar{y})}{\left(\sum_{i=1}^{n} \sin(x_i - \bar{x})^2 \sum_{i=1}^{n} \sin(y_i - \bar{y})^2\right)^{1/2}},$$

where $\bar{x}$ and $\bar{y}$ are the circular means. We let $\sigma_{rs}^{(0)} = \rho_c(\boldsymbol{y}_r, \boldsymbol{y}_s)\sigma_{rr}^{(0)}\sigma_{ss}^{(0)}$, for $r \neq s$, as the initial values for the covariance matrix at the start of the algorithms. This ensures that the initial matrix $\boldsymbol{\Sigma}^{(0)}$ is of full rank.

4 Monte Carlo experiments

To compare the performance of the proposed methods, we consider two Monte Carlo experiments, one for the univariate case; and the second one for the multivariate case. For the univariate case, the experiments have the following configurations: sample size $n = 10, 50, 100, 500$; $\mu_0 = 0$; $\sigma_0 = \pi/8, \pi/4, \pi/2, \pi, 3/2\pi, 2\pi$; and we set the number of Monte Carlo replications to 500. We compare the following methods: (a) the direct maximisation of the log-likelihood function performed via

the R function `optim` using the default settings; (b) the EM algorithm; (c) the CEM algorithm; and (d) we also consider the algorithm implemented in function `mle.wrappednormal`, available in the R package `circular` (Agostinelli and Lund, 2017), which is based on an IRML procedure to compute the maximum likelihood estimates. As for the initial values of these methods, we use the circular mean and $-2\log(\hat{\rho})$ respectively for μ and σ, where $\hat{\rho}$ is the sample mean resultant length. In this section, without loss of generality, we set J equal to 4, for all of the procedures in the univariate case, where $p = 1$.

For the multivariate case, the experiments has the following configurations: number of variables $p = 2, 5, 10$; sample sizes are selected from the range of $n = 10, 50, 100, 500$, depending on the value of p; $\boldsymbol{\mu}_0 = \mathbf{0}$; and we set the number of Monte Carlo replications to 500.

To account for the lack of affine equivariance of the Wrapped Normal model, we consider different covariance structures, $\boldsymbol{\Sigma}_0$, as in Agostinelli et al (2015). For each sample in our simulation setup, we create a different random correlation matrix with a fixed condition number (CN). We use the following procedure to obtain random correlations with condition numbers CN fixed at 20:

1. For a fixed condition number CN given, we first simulate a diagonal matrix $\boldsymbol{\Upsilon}^{(t)} = \text{diag}(\lambda_1, \ldots, \lambda_p)$, (where $t = 0$ and $\lambda_1 > \lambda_2 > \ldots > \lambda_p$) with the largest eigenvalue $\lambda_1 = \text{CN}$ and the smallest eigenvalue $\lambda_p = 1$. The remaining $(p-2)$ eigenvalues, $\lambda_2, \ldots, \lambda_{p-1}$, are sorted random samples from a uniform distribution in the interval $(1, \text{CN})$.
2. Next, we generate a $p \times p$ random matrix, $\boldsymbol{Y}$, in which its elements are chosen independently from standard normal distribution. Then, we use the eigendecomposition of the symmetric matrix $\boldsymbol{Y}^\top \boldsymbol{Y}$ to obtain a random orthogonal matrix $\boldsymbol{U}$, whose columns are the orthonormal eigenvectors of $\boldsymbol{Y}^\top \boldsymbol{Y}$.
3. Using the results of 1 and 2 above, we construct a random covariance matrix by the eigendecomposition, $\boldsymbol{\Sigma}^{(t)} = \boldsymbol{U}\boldsymbol{\Upsilon}^{(t)}\boldsymbol{U}^\top$. Note that the condition number of $\boldsymbol{\Sigma}^{(t)}$ is equal to the desired value, CN.
4. In this step, we convert the covariance matrix $\boldsymbol{\Sigma}^{(t)}$ into the correlation matrix $\boldsymbol{R}^{(t)}$, i.e.,
$$\boldsymbol{R}^{(t)} = \boldsymbol{D}^{-1/2}\boldsymbol{\Sigma}^{(t)}\boldsymbol{D}^{-1/2},$$
where $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_p)$, and d_i's are the main diagonal elements of $\boldsymbol{\Sigma}^{(t)}$.
5. We may note that the condition number of $\boldsymbol{R}^{(t)}$ is not necessarily equal to CN. To remedy this problem, we use the following spectral decomposition to obtain the updated covariance matrix, $\boldsymbol{\Sigma}^{(t+1)}$:
$$\boldsymbol{\Sigma}^{(t+1)} = \boldsymbol{U}\boldsymbol{\Upsilon}^{(t+1)}\boldsymbol{U}^\top \ , \tag{4}$$
where
$$\boldsymbol{\Upsilon}^{(t+1)} = \text{diag}\left(\tilde{\lambda}_1^{\boldsymbol{R}^{(t)}}, \lambda_2^{\boldsymbol{R}^{(t)}}, \ldots, \lambda_p^{\boldsymbol{R}^{(t)}}\right).$$
Here $\boldsymbol{\Upsilon}^{(t+1)}$ is a diagonal matrix formed by the eigenvalues of $\boldsymbol{R}^{(t)}$ ($\lambda_i^{\boldsymbol{R}^{(t)}}$s), and $\tilde{\lambda}_1^{\boldsymbol{R}^{(t)}} = \text{CN} \times \lambda_p^{\boldsymbol{R}^{(t)}}$. This adjustment of the first eigenvalue is to make sure the condition number of $\boldsymbol{\Sigma}^{(t+1)}$, given in (4), is equal to desired CN.
6. Now we set $t \leftarrow t + 1$ and repeat steps 4 and 5 until the condition number of $\boldsymbol{R}^{(t)}$ is within a tolerance level (or some iteration limit is reached). In our Monte Carlo study, in all of the cases, convergence was reached after a few iterations.

Once a desired correlation matrix is obtained, covariance matrices are constructed in such a way that variances in the main diagonal elements are the square of σ_0, chosen among the values $(\pi/8, \pi/4, \pi/2, \pi, 3/2\pi, 2\pi)$, as in the univariate case. Here we compare the following four methods:

(a) optim: where direct maximisation of the log-likelihood (1) is performed using the R function `optim` with the default setting;
(b) EM: where the EM algorithm is used;
(c) CEM: where the CEM algorithm is employed; and
(d) circular: where the IRML procedure implemented in function `mle.wrappednormal`, package `circular` is utilised.

As initial values for all these methods, we use the approach recommended in Section 3.4, which aims to be fast and effective. Furthermore, we also run the first three algorithms, providing the true values $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ as the initial values, in order to better understand the effect of different initial values in the performance of each method. The character "T" (as for the *true* parameters) is added at the end of the labels to single out these scenarios (i.e., optimT, EMT, CEMT) in the associated Figures. As of the remaining of this section, we set J equal to 3, for all the methods in the multivariate case, where $p > 1$. For evaluation of the log-likelihood function, the covariance matrix $\boldsymbol{\Sigma}$ is parametrised using the log-Cholesky parameterisation (Pinheiro and Bates, 1996) as described in Section 3, which allows for unconstrained optimisation while ensuring positive definiteness of $\boldsymbol{\Sigma}$.

In all cases the performance is evaluated using the following three measures:

(i) Likelihood-ratio test statistic (Λ): is expressed as difference between the log-likelihoods, i.e.
$$\Lambda(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = -2(\ell(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) - \ell(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})) \ ,$$
where $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are the true parameters.
(ii) Angular separation (AS):
$$\mathsf{AS}(\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{p} (1 - \cos(\hat{\mu}_i - \mu_{i0})) \ ,$$
is an angular measure to obtain the dissimilarity between the estimated and true mean vectors, with a range from 0 to $2p$.
(iii) Entropy loss (Δ): is a divergence measure, also appears in the likelihood ratio test statistics, for testing the null hypothesis that a multivariate Normal distribution has covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$.
$$\Delta(\hat{\boldsymbol{\Sigma}}) = \text{trace}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}_0^{-1}) - \log(|\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}_0^{-1}|) - p \ .$$
Entropy loss is essentially the Kullback-Leibler divergence between two Gaussian distributions with equal mean vectors and covariances $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}_0$, respectively.

In Figures 1 – 4, we report the results of the likelihood-ratio test statistic (Λ), the angular separation (AS), and the entropy loss (Δ) for all values of dimensions: $p = (1, 2, 5, 10)$, sample sizes: $n = (10, 50, 100, 500)$, standard deviations: $\sigma_0 = (\pi/8, \pi/4, \pi/2, \pi, 3/2\pi, 2\pi)$, and different methods (optim, optimT, EM, EMT, CEM,

CEMT, circular), as long as the results converge to a reasonable solution in a feasible time frame. Figure 5 provides information on the execution time for $n = 100, 500$ and $\sigma_0 = \pi/8, 3\pi/2$. Comprehensive results of the outputs are available in Section SM–6 of the Supplementary Material.

For dimension $p = 1$, and small values of σ_0 (e.g., smaller than $\pi/2$), all of the methods perform equally well, regardless of sample size. As σ_0 increases further, we notice that all algorithms with the initial values set to be the true ones (optimT, EMT, CEMT) have better performances, i.e., lower AS, and Δ in compare to the corresponding methods (optim, EM, CEM), as expected. For $\sigma_0 \geq \pi$, as sample size increases, optim and EM obtain smaller Δ values, whereas CEM performs better based on the Λ criterion.

For $p \geq 2$, the IRML algorithm in Agostinelli (2007) is not available anymore (as it is developed only for $p = 1$). Similar to the univariate case ($p = 1$), for all sample sizes and $\sigma_0 < \pi/2$, the remaining six methods perform equally well for $p = 2$. For $\sigma_0 = \pi$, EM and optim obtain smaller Δ and Λ. Like $p = 1$, the methods starting with the initial values set to be the true ones (optimT, EMT, CEMT), perform better in $p = 2$ as well. For larger σ_0, EM and optim show somewhat similar behaviour based on the Λ measure, while they do not have a huge difference with CEM using the AS performance.

As the dimension is increased to $p = 5$, for $\sigma_0 \leq \pi/2$, all of the methods have similar results based on the AS measure. In this case, EM, EMT, CEM, and CEMT perform better with respect to the measurements Δ and Λ. For $\sigma_0 > \pi/2$ and larger sample size ($n = 500$), these methods (EM, EMT, CEM, and CEMT) show similar performance as well. For $p = 10$, EM and CEM approaches are the only ones that do not fail, and their performances are similar to the case $p = 5$. Furthermore, as it is shown in Figure 5, on average optim is much slower than the EM and CEM approaches. In fact, in all four cases: $(n, \sigma_0) = \{(100, \pi/8), (500, \pi/8), (100, 3\pi/2), (500, 3\pi/2)\}$, optim fails to converge to a local maximum for $p > 5$. Also Figure 5 confirms that CEM is slightly faster than EM.

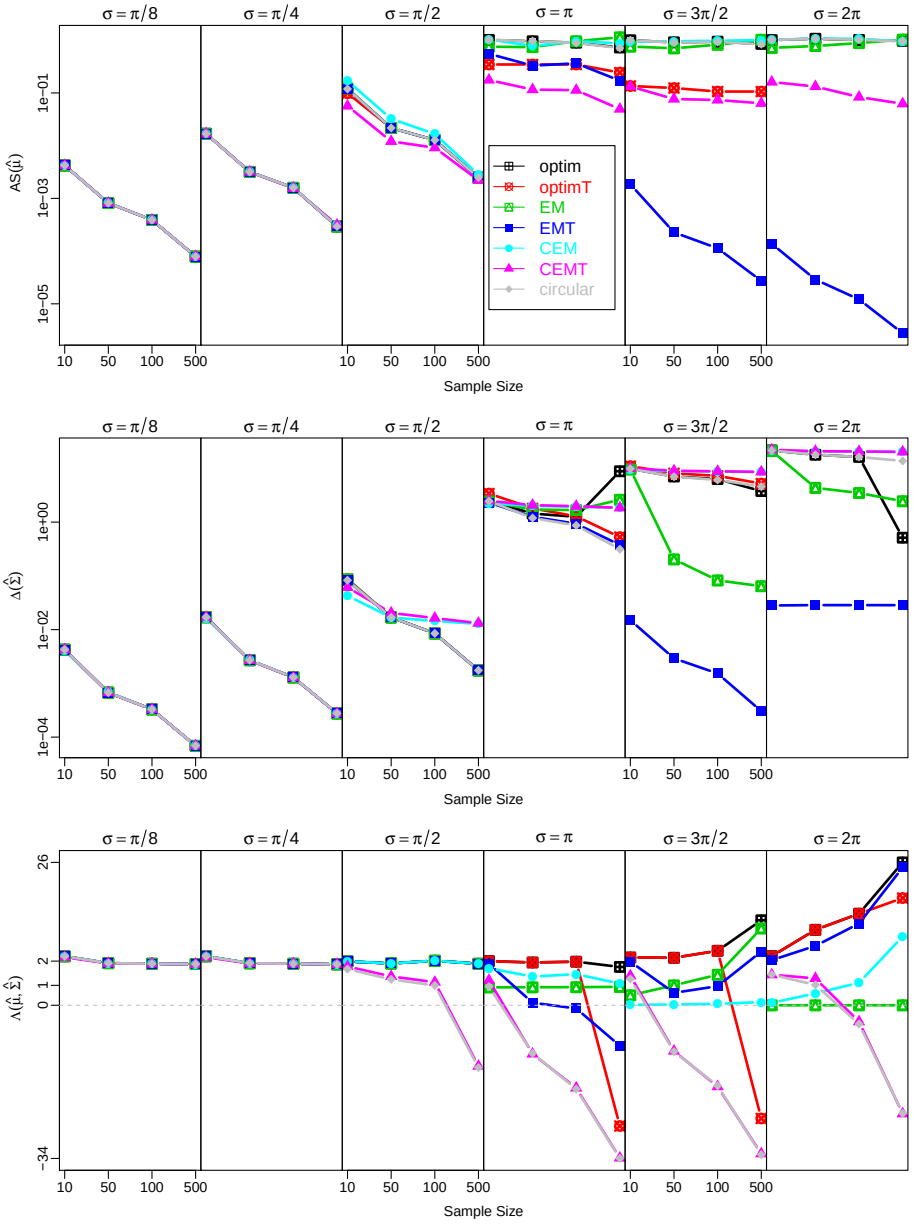

Fig. 1: Performance of the estimators in the univariate case $p = 1$. First row is corresponding to AS, second and third rows are based on Δ and Λ, respectively.

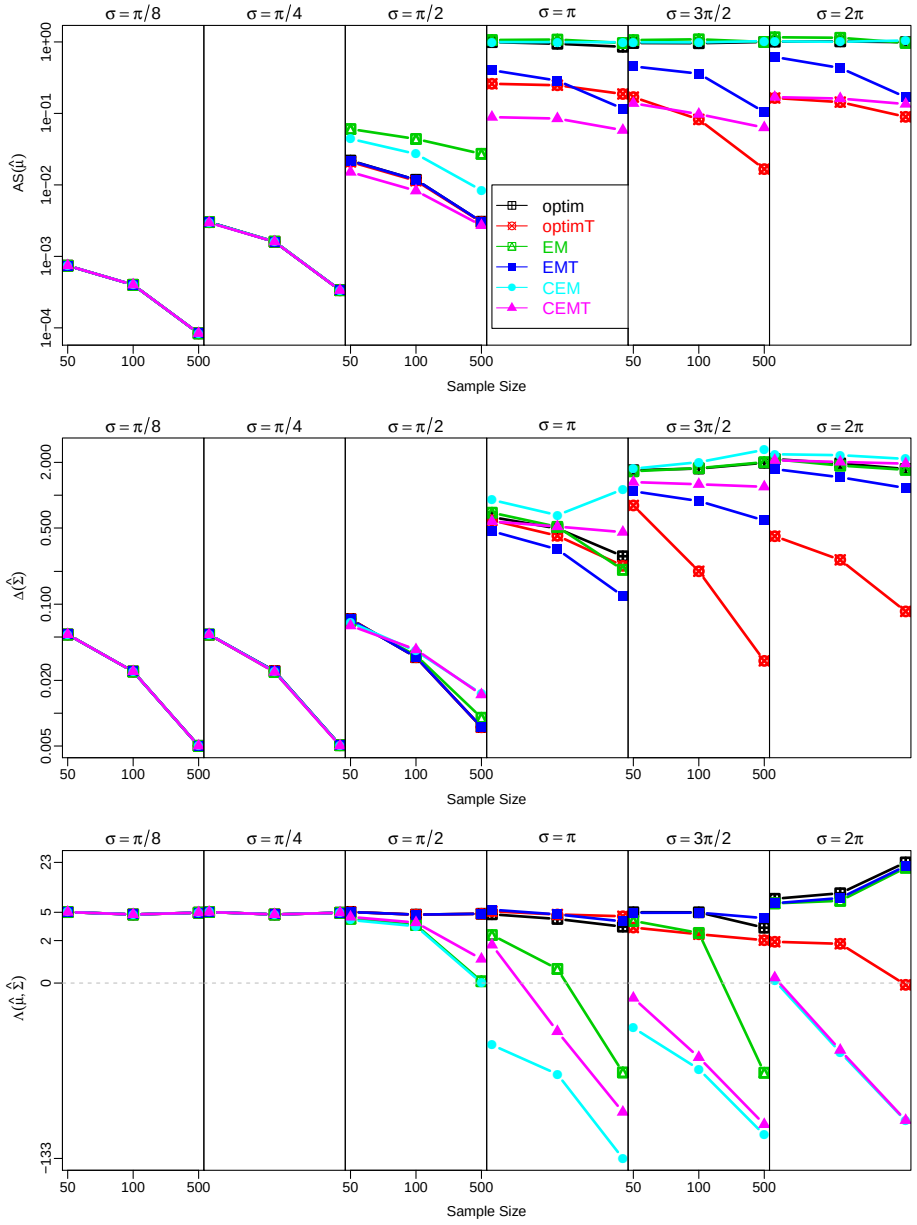

Fig. 2: Performance of the estimators in the bivariate case $p = 2$. First row is corresponding to AS, second and third rows are based on Δ and Λ, respectively.

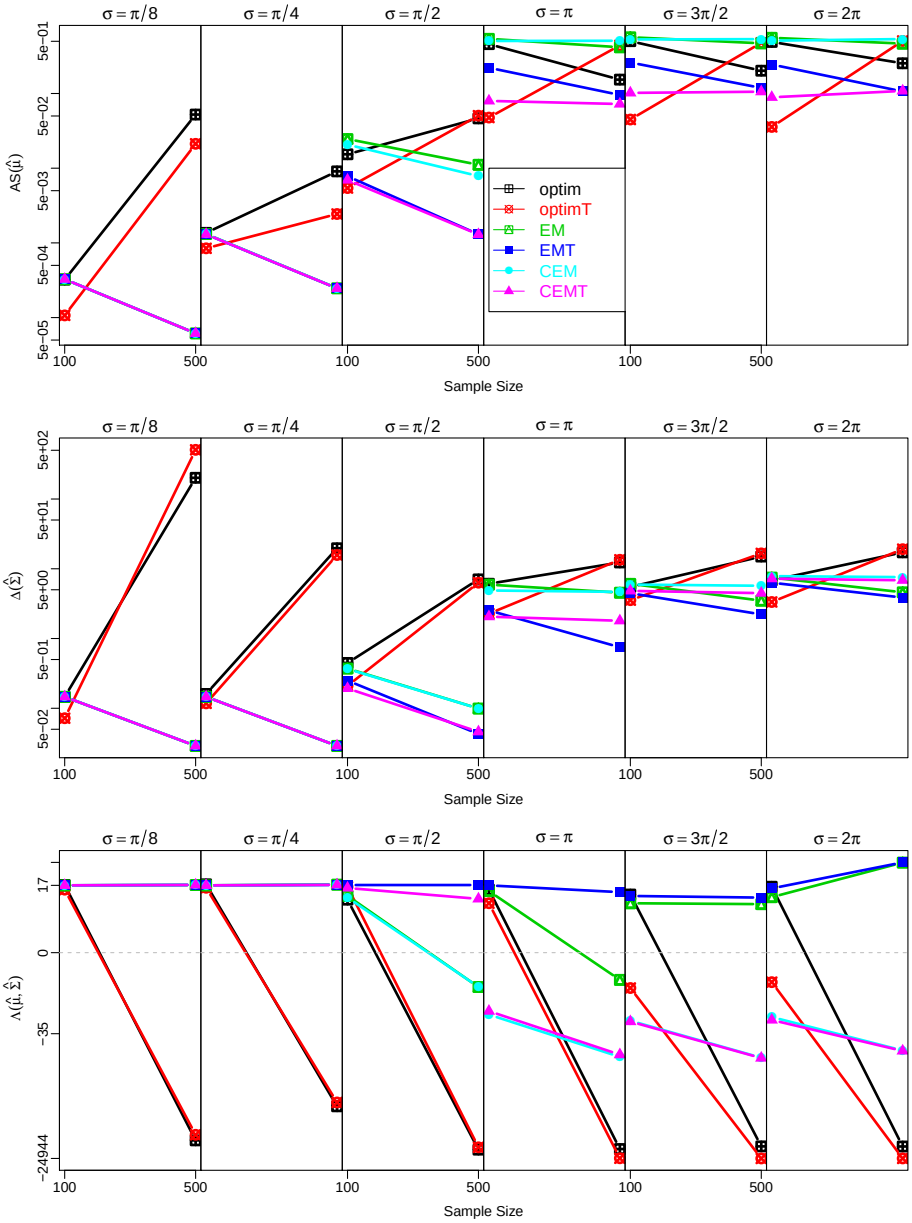

Fig. 3: Performance of the estimators in the case $p = 5$. First row is corresponding to AS, second and third rows are based on Δ and Λ, respectively.

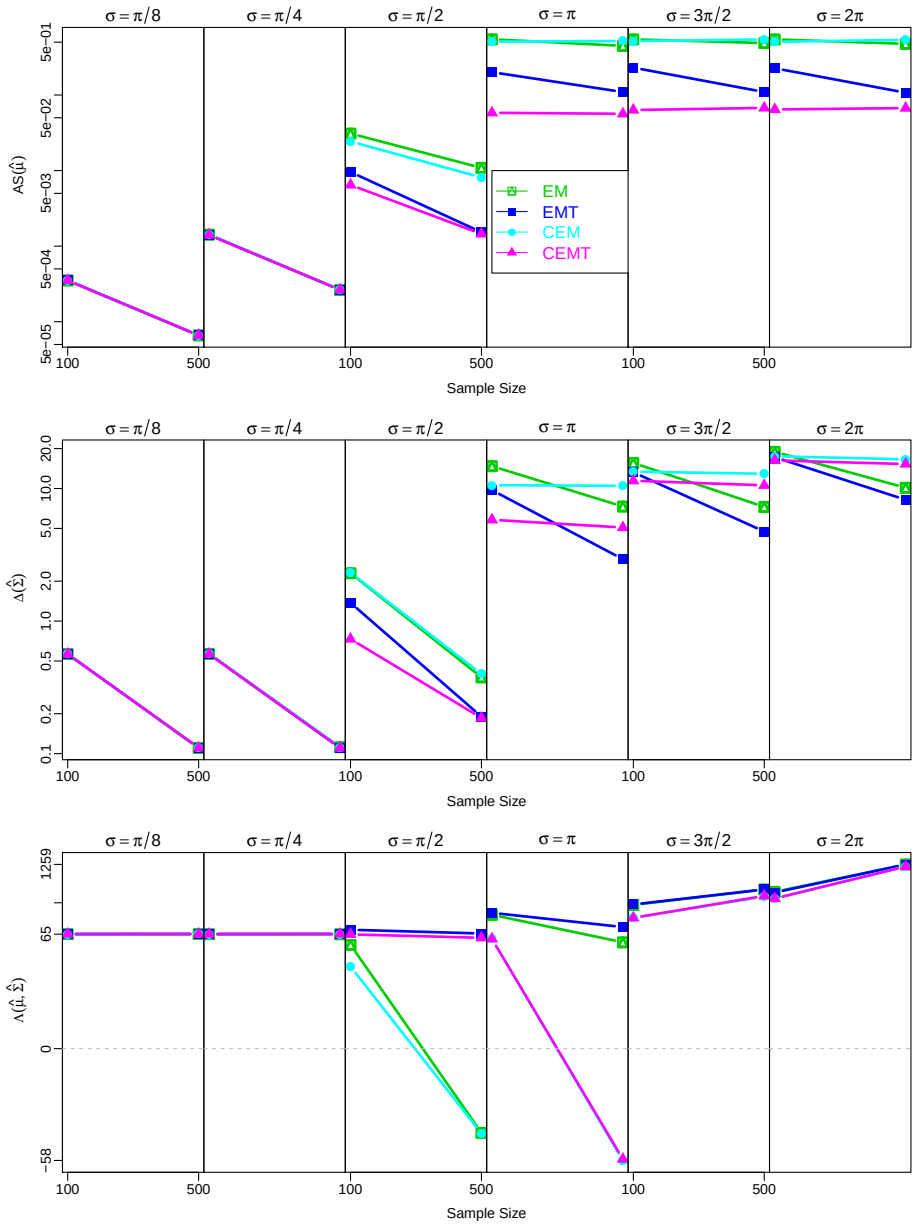

Fig. 4: Performance of the estimators in the case $p = 10$. First row is corresponding to AS, second and third rows are based on Δ and Λ, respectively.

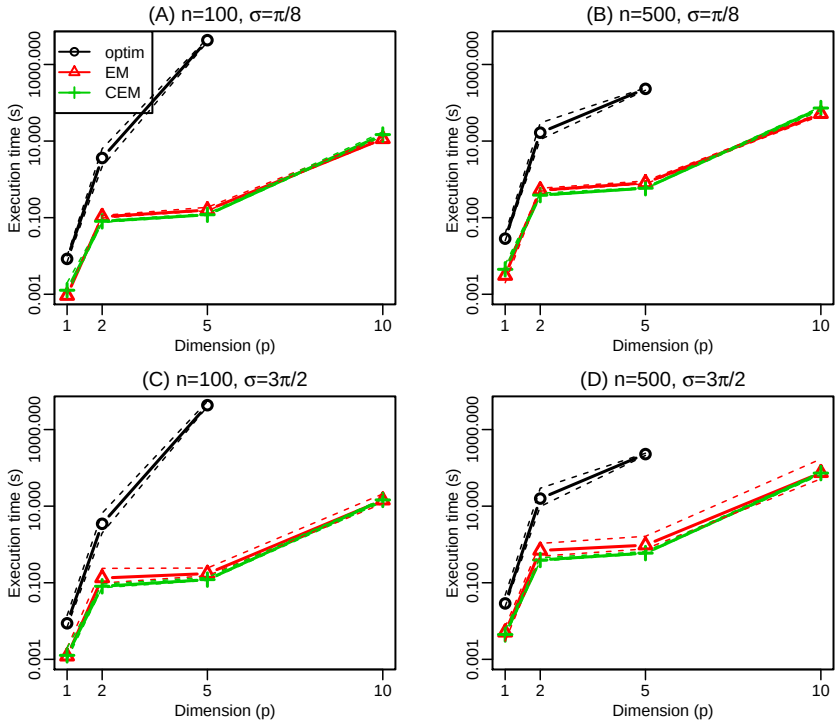

Fig. 5: Mean execution times for $n = 100, 500$ (first and second rows respectively), and $\sigma_0 = \pi/8, 3\pi/2$ (first and second columns respectively). Black line: optim, red line: EM, and green line: CEM. The point-wise confidence intervals of the execution times are shown by dash line.

5 Application to real-world data

In this section we consider two real data examples in bioinformatics. The first one deals with describing the protein structure using bivariate angles. The results of this example are reported in the subsection 5.1. The second example is on RNA data and it is analysed in subsection 5.2; in this case observations are on 7-torus, i.e. the variables are lying on a 7-dimensional torus. A third example is related to wind direction, a univariate example, where its analysis is reported in Section SM–3 of the Supplementary Material.

5.1 Protein structure: A bivariate case

One of the important topic in the field of structural biology is the determination of the three-dimensional (3D) structure of a protein. Protein backbone is what holds a protein together with a relatively simple chemical structure: a nitrogen atom, two carbon atoms, one or two oxygen atoms, and a few hydrogens. Amino acid is

an organic side chain (or residue), unique to 20 well-known amino acids, attached to the central carbon, C_α.

The backbone conformation of proteins can be represented equivalently by the Cartesian coordinates of C_α traces or the 2 pseudo-angles (θ, τ) between the two consecutive planes formed by 4 successive C_α (see Figure 6(A) for a clarification on (θ, τ) representation). The Ramachandran plot, a scatter plot of θ vs. τ, can reflect the allowed regions of conformational space available to protein chains, provide a path for distinctive classification of protein structures, and largely contribute to different applications in this area of research (Oldfield and Hubbard, 1994). A variety of techniques is used in the literature to estimate the bivariate density functions associated to the Ramachandran plots. Here we use the bivariate Wrapped Normal distribution, and three procedures to obtain the associated MLEs.

A collection of data sets called SCOP.1 about protein structure analysed in Najibi et al (2017) and described in their Supplementary Material are used as an illustrative example here. The collection contains bivariate information about 63 protein domains that were randomly selected from three remote Protein classes in the Structural Classification of Proteins (SCOP). Hence, each of these 63 collection is a bivariate data set. The class labels are available and we can check the homogeneity in each of the three clusters by applying the techniques developed in this paper.

In this example, three clusters are identified by the locations of the proteins in the SCOP tree. See Figure 6(B-D) for an illustration on three randomly selected bivariate datasets (three randomly selected protein structures are: 1BWW, 1SW7, and 1BRK, publicly available in the Protein Data Bank archive, known as PDB) associated to each cluster. The constituents of the collection of protein domains in SCOP.1 are as follows.

- Cluster 1: 19 domains from "All beta proteins/ Immunoglobulin-like beta-sandwich /Immunoglobulin/ V set domains (antibody variable domain-like)/ Immunoglobulin light chain kappa variable domain, VL-kappa/ Human (Homo sapiens)". Sample sizes are in the range between 104 and 106.
- Cluster 2: 26 domains from "Alpha and beta proteins (a/b)/TIM beta, alpha-barrel/Triosephosphate isomerase (TIM) / Triosephosphate isomerase (TIM) / Triosephosphate isomerase/ Chicken (Gallus gallus)". Sample sizes are in the range between 233 and 244.
- Cluster 3: 18 domains from "Alpha and beta proteins (a+b)/ Microbial ribonucleases/ Microbial ribonucleases/ Bacterial ribonucleases/ Barnase/ Bacillus amyloliquefaciens". Sample sizes are in the range between 104 and 107.

Three procedures are considered: (a) optim: direct maximisation of the log-likelihood using the the function `optim` in R with the default settings, (b) EM, and (c) CEM algorithms, as introduced in subsections 3.1 and 3.2 respectively. All the algorithms are initialized using the same values as described in subsection 3.4. For all of the methods we investigate the results for $J = 3$ and $J = 6$.

Figure 7 reports the results in each of the three clusters. In Figure 7(A), one can see the estimated means based on all methods in 3 different clusters. It is obvious that EM, and CEM have better performance to distinguish all clusters and by increasing J, from 3 to 6, optim improves as well. In Figures 7(B)– 7(D), all estimated means are marked by different signs (black circles for cluster 1, red

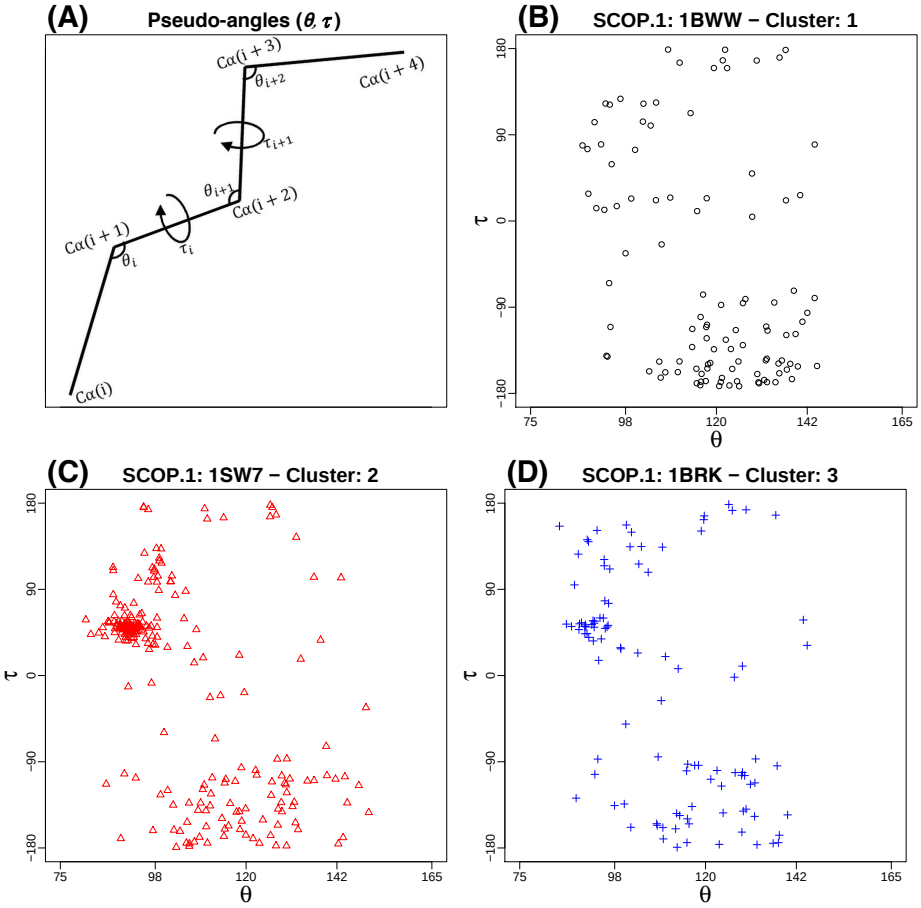

Fig. 6: Schematic representation of the protein backbone angles. (A) Angles along the C_α trace is denoted by (θ_i, τ_i), where θ_i is the pseudo-bond angle of three consecutive C_α atoms $\big(C_\alpha(i), C_\alpha(i+1), C_\alpha(i+2)\big)$, and τ_i is the pseudo-torsion angle of four consecutive C_α atoms $\big(C_\alpha(i), \cdots, C_\alpha(i+3)\big)$. The term pseudo is used for (θ, τ) here because the consecutive C_α atoms are not actually connected by a single chemical bond. (B-D) Ramachandran plots associated to three randomly selected bivariate samples from clusters $1-3$ in SCOP.1.

triangle and blue crosses for clusters 2 and 3, respectively) for different clusters. To further illustrate the behaviour of the second moment, for each density, ellipsoid confidence regions with nominal coverage probability 0.95 are also provided.

In Figure 8, we evaluate the performance of each method by comparing the log-likelihood functions at the estimated values in cluster 1 and cluster 3. Since the use of $J = 3$ and $J = 6$ in EM and CEM lead to similar results, we report only the $J = 3$ case. As for optim, the performance is very different for $J = 3$ and $J = 6$, and that is why we report both cases. While for $J = 6$ the direct maximisation (optim) provides homogeneous solutions within each clusters, for

1WTL.A

	Method		
	optim	EM	CEM
init	−2463.97	−306.97	−310.98
optim	−4918.87	−2464.61	−310.98
EM	−310.49	−306.97	−310.98
CEM	−2280.10	−306.97	−310.98

1BRE.B

	Method		
	optim	EM	CEM
init	−403.39	−314.50	−301.31
optim	−327.74	−314.72	−370.96
EM	−324.06	−314.50	−309.54
CEM	−576.15	−314.50	−301.31

1BRE.C

	Method		
	optim	EM	CEM
init	−309.27	−312.80	−301.10
optim	−375.34	−312.80	−299.58
EM	−312.80	−312.80	−306.51
CEM	−305.31	−312.80	−301.10

Table 1: Protein data set. Log-likelihood values of optim, EM, and CEM (columns) obtained after convergence using different starting values (rows).

$J = 3$ the algorithm seems very unstable. However as it is shown in Figure 8, the log-likelihood values provided by optimusing $J = 3$ is always larger than that of using $J = 6$. Consistent with this fact, for almost all situations the EM, and CEM report a larger value of the log-likelihood compare with the direct optimisation (optim); this is particularly the case for the third cluster. It is worth to note that, the whole results for the log-likelihood functions in each cluster have been further illustrated in Section SM–4 in Supplementary Material.

To study the stability of the procedures we investigate the performance of each technique under different starting values. Table 1 reports the log-likelihood of the methods (in columns: EM, CEM, and optim) using different starting values (in rows: initials values described in Section 3.4, init; and the estimates obtained from optim, EM, CEM) for three particular data sets namely **1WTL.A**, **1BRE.B**, **1BRE.C**(for more information, see Najibi et al (2017)). The solutions provided by EM, and CEM seem very stable, apart from the **1WTL.A**, where the EM starting value is chosen to be the optim solution. The optim shows high sensitivity on the choice of the initial values. Finally, we would like to remark that EM, and CEM are maximising different flavours of the log-likelihood and hence it can happen that CEM, after convergence, provides a larger value of the genuine log-likelihood.

5.2 RNA data set: An example of data on a 7-torus

We consider a data set on Ribonucleic acid (RNA). RNA is a polymeric molecule essential in various biological roles in coding, decoding, regulation, and expression

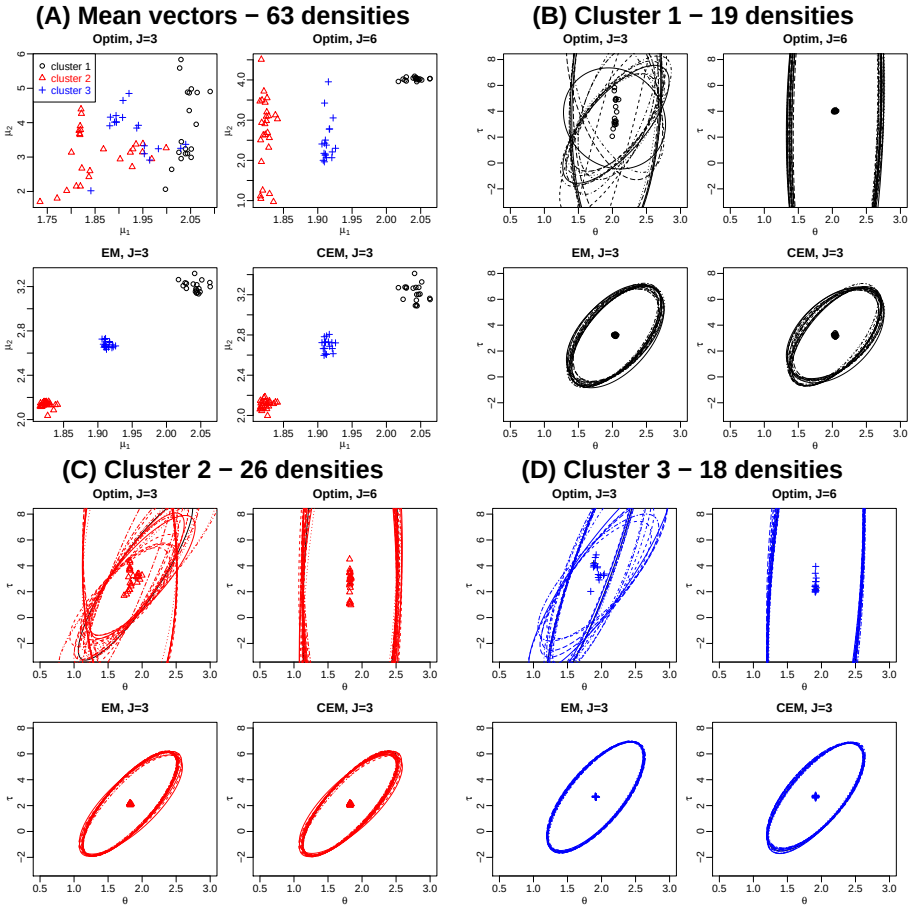

Fig. 7: Estimated densities based on EM, CEM, and optim. (A) Estimated means for three clusters (cluster 1: black circles, cluster 2: red triangle, and cluster 3: blue crosses), (B)–(D) Estimated means for different clusters are marked by different signs (cluster 1: black circles, cluster 2: red triangle, and cluster 3: blue crosses), plus the ellipsoid confidence regions with nominal coverage probability 0.95 for each density.

of genes. RNA and DNA are nucleic acids, and, along with lipids, proteins and carbohydrates, constitute the four major macromolecules essential for all known forms of life. In RNA, each nucleic base corresponds to a backbone segment described by 6 dihedral angles and one angle for the base, giving a total of 7 angles. The distribution of these 7 angles over large samples of RNA strands have been studied, among others, by Eltzner et al (2018) using Torus Principal Component Analysis. The original data set contains 8301 observations, but based on a clustering procedure the data set was split into 23 clusters and all of the observations with more than 50° in angular distance from their nearest neighbour were removed. So, the final data set contains 7390 observations grouped in 23 clusters. We apply the

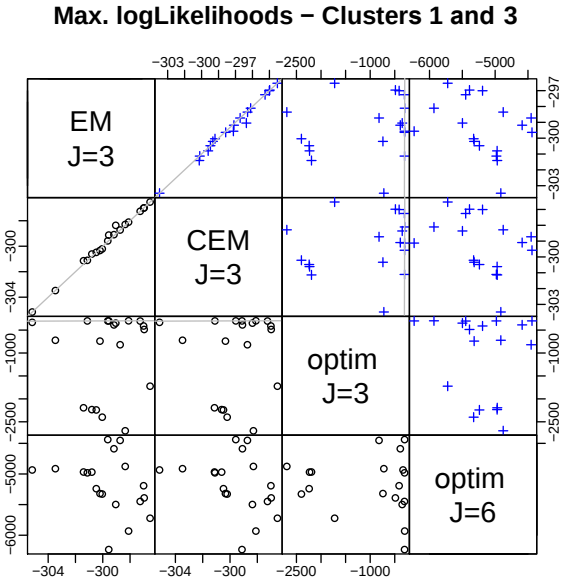

Fig. 8: Protein data set. Comparison of the log-likelihood functions at the MLEs for cluster 1 (lower triangle: black circles) and cluster 3 (upper triangle: blue crosses).

optim, EM, and CEM algorithms, with $J = 3$, in each of these clusters, to estimate the parameters of 23 multivariate Wrapped Normal models. The estimated parameters can be used to provide a qualitative measure to evaluate the homogeneity of the groups, i.e., a qualitative comparison of the estimated means and the shape of estimated correlation matrices using the associated graphical representations. Due to moderate dimension on both sample size and number of variables, direct optimisation of the log-likelihood, optim needs about 100 times the execution time of EM and between 150–400 times the execution time of CEM (see Figure SM–11). In Tables SM–2, SM–3 and SM–4 we report the estimated mean angles for each of the 23 clusters based on EM, CEM and optim algorithm respectively. Figures SM–2 – SM–10 represents the estimated correlation structures for all the 23 clusters.

As seen in both Tables SM–2 and SM–3, the means estimated within each cluster are closer to each other. Also, Figures SM–2 – SM–10 depict each correlation by an ellipse whose shape tends towards a line with slope 1 for positive linear correlations, to a circle for correlations near zero, and to a line with negative slope, −1, for negative linear correlations. In addition, a red colour indicates strong negative, and a blue colour implies strong positive correlations. A through inspection of these plots confirms the promising agreement between the EM and CEM algorithms.

To summarize our finding from Monte Carlo simulation and the analysis of real daata sets it is clear that the proposed methodology is computationally more efficient than the direct likelihood optimization. Furthermore,

the comparison between direct likelihood optimization and the proposed method for different values of J is interesting. The analysis suggests that $J = 3$ is not sufficient for purposes of evaluating the likelihood directly (Figure 7), while $J = 6$ provides a better fit but still not optimal. On contrary the introduced methodology show a good stability with respect to the choice of J. Finally, the discrepancy in likelihood values (Figure 8) suggests that the methodology is more appropriate from an optimisation perspective for handling the bahaviour of the present objective functions.

6 Discussion

We introduced two new algorithms based on Expectation-Maximisation and Classification Expectation-Maximisation methods for the estimation of the parameters in multivariate Wrapped Normal model to deal with circular data on torus. The proposed algorithms perform well in comparison with the direct maximisation of the log-likelihood function. The new EM, and CEM procedures converge to an acceptable solution in moderate to high dimensional setting, while this is not the case for the direct maximisation, (optim). Also real examples indicate that, for large dimensions, the new algorithms (EM, CEM) outperform the direct maximisation of the log-likelihood (optim), in finding the global maximum. The proposed methods can be easily extended to most wrapped multivariate elliptical symmetric distributions indexed by multivariate location and scatter matrix.

Acknowledgments

The authors thank Stephan Huckemann and Benjamin Eltzner for providing the RNA data set. We would also like to thank the editor, and two referees for their constructive and thoughtful comments which helped us tremendously in improving the manuscript.

References

Agostinelli C (2007) Robust estimation for circular data. Computational Statistics and Data Analysis 51(12):5867–5875

Agostinelli C, Lund U (2017) R package `circular`: Circular Statistics (version 0.4-93). CA: Department of Mathematics, University of Trento, Italy. UL: Department of Statistics, California Polytechnic State University, San Luis Obispo, California, USA, URL `https://r-forge.r-project.org/projects/circular/`

Agostinelli C, Leung A, Yohai VJ, Zamar RH (2015) Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. TEST 24(3):441–461

Baba Y (1981) Statistics of angular data: wrapped normal distribution model. In Proceedings of the Institute of Statistical Mathematics 28:41–54, (in Japanese)

Batschelet E (1981) Circular Statistics in Biology. Academic Press, NewYork

Breckling J (1989) The Analysis of Directional Time Series: Applications to Wind Speed and Direction. Springer-Verlag, Berlin, lecture Notes in Statistics 61

Celeux G, Govaert G (1992) A classification EM algorithm for clustering and two stochastic versions. Computational Statistics and Data Analysis 14:315–332

Coles S (1998) Inference for circular distributions and processes. Statistics and Computing 8:105–113

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B 39(1):1–38

Eltzner B, Huckermann S, Mardia KV (2018) Torus principal component analysis with applications to rna structure. Annals of Applied Statistics In press

Ferrari C (2009) The wrapping approach for circular data bayesian modeling. Phd thesis, Alma Mater Studiorum Universit di Bologna. Dottorato di ricerca in Metodologia statistica per la ricerca scientifica, 21 Ciclo.

Fisher NI (1987) Problem with the current definition of the standard deviation of wind direction. Journal of Climate and Applied Meteorology 26:1522–1529

Fisher NI, Lee AJ (1994) Time series analysis of circular data. Journal of the Royal Statistical Society Series B 56:327–339

Jammalamadaka SR, SenGupta A (2001) Topics in Circular Statistics, Multivariate Analysis, vol 5. World Scientific, Singapore

Johnson RA, Wehrly T (1978) Some angular-linear distributions and related regression models. Journal of the American Statistical Association 73:602–606

Kent JT (1978) Limiting behaviour of the von Mises-Fisher distribution. Mathematical Proc of the Cambridge Philosophical Society 84:531–536

Mardia KV (1972) Statistics of Directional Data. Academic Press, London

Mardia KV (2010) Bayesian analysis for bivariate von Mises distributions. Journal of Applied Statistics 37:515–528

Mardia KV, Jupp PE (2000) Directional Statistics. Wiley, New York

Mardia KV, Voss J (2014) Some fundamental properties of a multivariate von Mises distribution. Communications in Statistics-Theory and Methods 43:1132–1144

Mardia KV, Taylor CC, Subramaniam GK (2007) Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. Biometrics 63:505–512

Mardia KV, Hughes G, Taylor CC, Singh H (2008) A multivariate von Mises distribution with applications to bioinformatics. The Canadian Journal of Statistics 1:99–109

Najibi SM, Maadooliat M, Zhou L, Huang JZ, Gao X (2017) Protein structure classication and loop modeling using multiple Ramachandran distributions. Computational and Structutal Biotechnology Journal 15:243–254

Oldfield TJ, Hubbard RE (1994) Analysis of C_α geometry in protein structures. Proteins 18:324–337

Pinheiro JC, Bates DM (1996) Unconstrained parameterizations for variance-covariance matrices. Statistics and Computing 6(3):289–296

R Core Team (2019) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL `https://www.R-project.org/`

Ravindran P, Ghosh S (2011) Bayesian analysis of circular data using wrapped distributions. Journal of Statistical Theory and Practice 5:547–561

Stephens MA (1963) Random walk on a circle. Biometrika 50:385–390