# Finite mixtures of multivariate Wrapped Normal distributions for model based clustering of $p$-torus data

Luca Greco
University G. Fortunato, Benevento, Italy
and
Pier Luigi Novi Inverardi
Department of Economics, University of Trento, Italy
and
Claudio Agostinelli
Department of Mathematics, University of Trento, Italy

September 1, 2022

**Abstract**

We consider a finite mixture model of multivariate Wrapped Normal distributions to handle non homogeneous circular data on a $p-$dimensional torus ($p \geq 2$). The Wrapped Normal distribution is a valid alternative to model multivariate circular or directional data on a $p-$torus. Parameter estimation is carried out through a nested (classification) EM algorithm, by exploiting the ideas of unwrapping circular data. The source of incompleteness in the outer E-step is represented by unobserved group memberships, whereas the source of incompleteness in the inner E-step is given by the unobserved vectors of wrapping coefficients. The finite sample behavior of the proposed method has been investigated by Monte Carlo numerical studies and real data examples. Supplemental materials for the article, including data and `R` codes for implementing methods, running simulations and replicate data analyses, are available online.

*Keywords:* Circular, Classification, EM, Likelihood.

# 1 Introduction

Statistical analysis of circular or directional data have found applications in a variety of fields, including the analysis of wind directions (Lund, 1999; Agostinelli, 2007), animal movements (Ranalli and Maruotti, 2020), handwriting recognition (Bahlmann, 2006), people orientation (Baltieri et al, 2012), and protein bioinformatics (Mardia et al, 2007, 2012; Eltzner et al, 2018). Circular data can also arise with measurements on a periodic scale, such as arrival times measured on a clock scale, for instance. There exists a substantial literature on circular data. The reader is pointed to Mardia and Jupp (2000); Jammalamadaka and SenGupta (2001); Pewsey et al (2013) for a review on the topic and several examples in different contexts. In particular, the problem of modeling circular data has been tackled through suitable distributions, among which two of the the most popular are the von Mises and Wrapped Normal, with the latter being particularly advantageous, since it inherits many nice properties from the normal distribution and allows an effective generalization to circular multivariate variables and processes (Fisher and Lee, 1994; Coles, 1998; Jona Lasinio et al, 2012). It should be noted that the models considered in this paper are distinct from those defined on the surface of the unit sphere, such as the von Mises - Fisher distribution (Hornik and Grün, 2014).

Here, we are mainly interested in $p$-dimensional torus data and the multivariate Wrapped Normal (WN) distribution. The contribution of this paper is to introduce a likelihood based approach to handle the estimation of a finite mixture of multivariate Wrapped Normal distributions and the problem of (model based) clustering of circular data. Examples of finite mixture models for torus data can be found in Peel et al (2001); Banerjee et al (2005); Dortet-Bernadet and Wicker (2008); Mardia et al (2012); Mardia and Voss (2014). It is also worth to mention very recent papers dealing with clustering of time-dependent circular data (Mastrantonio et al, 2019; Ranalli and Maruotti, 2020). We stress that the authors of all these contributions did not use the Wrapped Normal distribution but they adopted a different choice for the components of the mixture. In particular, in Mardia et al (2012) the attention has been focused on the multivariate sine von Mises distribution. Although out of scope for this paper, we further point out a Bayesian approach to inference for (mix-

tures of) bivariate angular data, that has been developed in Chakraborty and Wong (2021), both for the Wrapped Normal and von Mises distributions. Furthermore, an agglomerative clustering technique for multivariate circular data has been proposed in Abraham et al (2013).

In this paper, model parameters are estimated by maximum likelihood, that is performed according to a nested (classification) expectation–maximization (EM) algorithm. Nesting comes from the implementation of a classical outer E-step, where the source of incompleteness is represented by unobserved group memberships, and an inner E-step characterized by the unobserved vectors of wrapping coefficients, appearing in the density function of the WN variate, as detailed in Section 2. The method relies on a data augmentation approach for circular data (Fisher and Lee, 1994; Coles, 1998; Jona Lasinio et al, 2012; Kurz and Hanebeck, 2015; Nodehi et al, 2020).

The remainder of the paper is organized as follows. We first introduce the multivariate Wrapped Normal distribution and related inferential issues in Section 2. Then, maximum likelihood estimation of the finite mixture of Wrapped Normal distributions is discussed in Section 3, providing some computational details and synthetic examples. We next explore the finite sample behavior of the proposed methodology through some numerical studies in Section 4 and real data analyses in Section 5. Concluding remarks end the paper in Section 6. Supplemental materials are available, that contain an Appendix with further results on all the examples, data and `R` code: a brief description is given in Section 7.

# 2 The multivariate Wrapped Normal distribution

The multivariate WN is obtained by component-wise wrapping of a $p$-variate Normal distribution on a $p$-dimensional torus (Johnson and Wehrly, 1978; Baba, 1981; Jammalamadaka and SenGupta, 2001). Each component is wrapped around the unit circle by the transformation $Y = X \bmod 2\pi$. Formally, let $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ be a random vector for which we assume a multivariate normal distribution $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)$ is the mean vector and $\Sigma$ is the $p \times p$ variance-covariance matrix. Then, the distribution of $\boldsymbol{Y} = \boldsymbol{X} \bmod 2\pi$ is a $p-$variate Wrapped Normal, in symbols $\boldsymbol{Y} \sim WN_p(\boldsymbol{\mu}, \Sigma)$, and

the modulus operator mod is performed component-wise. The density function $\phi_p^\circ(\boldsymbol{y})$ of $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_p)$ takes the form of an infinite sum over $\mathbb{Z}^p$, that is

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

with $\boldsymbol{y} = (y_1, y_2, \ldots, y_p) \in [0, 2\pi)^p$, $\boldsymbol{j} = (j_1, j_2, \ldots, j_p) \in \mathbb{Z}^p$, $\Omega = (\boldsymbol{\mu}, \Sigma)$ and $\phi_p(\cdot)$ denotes the density function of the normally distributed random vector $\boldsymbol{X}$. Without loss of generality, we let $\boldsymbol{\mu} \in [0, 2\pi)^p$ to ensure identifiability. The $p$-dimensional vector $\boldsymbol{j}$ is the vector of wrapping coefficients, that, if it were known, would describe how many times each component of the $p$-toroidal data point was wrapped. In other words, if we observed $\boldsymbol{j}$, we would obtain the unwrapped data $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ as $\boldsymbol{x} = \boldsymbol{y} + 2\pi\boldsymbol{j}$.

## 2.1 Maximum Likelihood Estimation

Evaluation of the multivariate WN density function can appear difficult (even in the univariate scenario) because it involves an infinite series. This aspect also makes inference more challenging. Nevertheless, it is well known that the density function (1) can be approximated with only a few terms (Mardia and Jupp, 2000), so that $\mathbb{Z}^p$ is replaced by the Cartesian product $\mathcal{C}_J = \otimes_{s=1}^p \boldsymbol{\mathcal{J}}$ where $\boldsymbol{\mathcal{J}} = (-J, -J+1, \ldots, 0, \ldots, J-1, J)$ for some $J$ providing a good approximation.

Here, maximum likelihood estimation is performed exploiting the concept of data augmentation and implementing an appropriate EM algorithm. The idea is to consider the wrapping coefficients $\boldsymbol{j}$ as latent variables and the observed circular data $\boldsymbol{y}$ as being incomplete. According to this approach, each circular data point $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ip})$ is assumed to be one component of the pair $(\boldsymbol{y}_i, \boldsymbol{v}_i)$, where $\boldsymbol{v}_i = (v_{i1}, v_{i2}, \ldots, v_{iq})$ is the associated wrapping coefficients label vector that is not available, with $q$ denoting the number of different wrapping coefficients vectors $\boldsymbol{j} \in \mathcal{C}_J$ and $\boldsymbol{v}_{iq} = 1$ or $\boldsymbol{v}_{iq} = 0$ according to whether $\boldsymbol{y}_i$ has the $q$-th wrapping coefficients vector.

Consider a sample of $p$-dimensional torus data of size $n$. Then, the EM algorithm to perform maximum likelihood estimation of the parameters of a $p-$variate WN distribution works with the complete (approximated, since $\mathbb{Z}^p$ is replaced by the Cartesian product $\mathcal{C}_J$)

log–likelihood function

$$\ell_c(\Omega) = \sum_{i=1}^{n} \sum_{\boldsymbol{j}_i \in \mathcal{C}_J} v_{i\boldsymbol{j}_i} \log \phi_p(\boldsymbol{y}_i + 2\pi \boldsymbol{j}_i; \Omega), \tag{2}$$

where $v_{i\boldsymbol{j}_i}$ is the index of the $i$-th observation having $\boldsymbol{j}_i$ as vector of wrapping coefficients. The wrapping coefficients label vector is assumed to be a realization of a multinomial distribution $\boldsymbol{V}_i$. Then, the conditional expected value of the complete log-likelihood in (2) given the observed data and the current parameters value in the E-step is obtained through the computation of the conditional probability that $\boldsymbol{y}_i$ has $\boldsymbol{j}_i$ as vector of wrapping coefficients, i.e.

$$v_{i\boldsymbol{j}_i} = \text{Prob}(V_{i\boldsymbol{j}_i} = 1 | \boldsymbol{Y} = \boldsymbol{y}_i, \Omega) = \frac{\phi_p(\boldsymbol{y}_i + 2\pi \boldsymbol{j}_i; \Omega)}{\sum_{\boldsymbol{j} \in \mathcal{C}_J} \phi_p(\boldsymbol{y}_i + 2\pi \boldsymbol{j}; \Omega)}. \tag{3}$$

In the M-step, parameters estimates are updated as weighted averages, that is

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \frac{1}{n} \sum_{i=1}^{n} \sum_{\boldsymbol{j}_i \in \mathcal{C}_J} v_{i\boldsymbol{j}_i} (\boldsymbol{y}_i + 2\pi \boldsymbol{j}_i) \\ \hat{\Sigma} &= \frac{1}{n} \sum_{i=1}^{n} \sum_{\boldsymbol{j}_i \in \mathcal{C}_J} v_{i\boldsymbol{j}_i} (\boldsymbol{y}_i + 2\pi \boldsymbol{j}_i)(\boldsymbol{y}_i + 2\pi \boldsymbol{j}_i)^\top . \end{aligned} \tag{4}$$

The algorithm alternates between the E- and the M-step until convergence of the sequence of log–likelihood values given by (2). Fisher and Lee (1994) suggested an EM algorithm to fit an autoregressive model to time series of circular data. In a similar fashion, Coles (1998) proposed a Metropolis–Hastings MCMC scheme to obtain samples for the wrapping coefficients in a Bayesian setting and Jona Lasinio et al (2012) extended further that technique to fit spatial processes. Recently, effective computational improvements have been discussed in Nodehi et al (2020) in a genuine likelihood framework. The authors also suggested an alternative strategy that allows a classification step after the evaluation of the conditional probabilities $v_{i\boldsymbol{j}_i}$, that is, we have

$$\tilde{\boldsymbol{j}}_i = \text{argmax}_{\boldsymbol{j}_i \in \mathcal{C}_J} v_{i\boldsymbol{j}_i}$$

and $\tilde{v}_{i\tilde{\boldsymbol{j}}_i} = 1$ and $\tilde{v}_{i\boldsymbol{j}_i} = 0$ otherwise. We call this step a CE-step, where C stands for classification. Then, $\tilde{v}_{i\boldsymbol{j}_i}$ is used in the M-step in place of $v_{i\boldsymbol{j}_i}$.

# 3 Maximum Likelihood estimation of a finite mixture of multivariate Wrapped Normal distributions

In order to handle the presence of possible clusters and multi-modality among the sample $p$-torus data, we assume a finite mixture model whose components are distributed according to a $p-$dimensional WN variate. Its density function is given by

$$f^{\circ}(\boldsymbol{y};\tau) = \sum_{g=1}^{G} \delta_g \phi_p^{\circ}(\boldsymbol{y};\Omega_g) \tag{5}$$

where we set $\tau = (\delta_1, \ldots, \delta_G, \Omega_1, \ldots, \Omega_G)$, $G$ denotes the number of groups, $\delta_g$ are membership probabilities and $\Omega_g$ are component specific parameters.

The operation of mixing and wrapping commute in general (Jammalamadaka and Kozubowski, 2017). It is straightforward to obtain

$$\begin{aligned} f^{\circ}(\boldsymbol{y};\tau) &= \sum_{g=1}^{G} \delta_g \left[ \sum_{\boldsymbol{j}\in\mathbb{Z}^p} \phi_p(\boldsymbol{y}+2\pi\boldsymbol{j};\Omega_g) \right] \\ &= \sum_{\boldsymbol{j}\in\mathbb{Z}^p} \left[ \sum_{g=1}^{G} \delta_g \phi_p(\boldsymbol{y}+2\pi\boldsymbol{j};\Omega_g) \right] \\ &= \sum_{\boldsymbol{j}\in\mathbb{Z}^p} f(\boldsymbol{y}+2\pi\boldsymbol{j};\tau) \ . \end{aligned}$$

where $f(\boldsymbol{y}+2\pi\boldsymbol{j};\tau) = \sum_{g=1}^{G} \delta_g \phi_p(\boldsymbol{y}+2\pi\boldsymbol{j};\Omega_g)$ is a usual mixture of multivariate normal distributions in $\mathbb{R}^p$.

Maximum likelihood estimation is based on maximizing the mixture log-likelihood function

$$\ell(\tau;\boldsymbol{y}) = \sum_{i=1}^{n} \log f^{\circ}(\boldsymbol{y};\tau) \ . \tag{6}$$

We do not pursue direct maximization but we implement an EM algorithm instead, stemming from two sources of incompleteness: both group memberships and wrapping coefficients are now treated as latent variables. In a standard fashion, let $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iG})$ denote the $i$-th component label binary vector and assume they are the realizations of multinomial random variables $\boldsymbol{Z}_i$ (McLachlan and Peel, 2004). Now, each vector $\boldsymbol{y}_i$ carries its own pair $(\boldsymbol{z}_i, \boldsymbol{v}_i)$ of missing components. The proposed algorithm works with the

complete (approximated) log–likelihood function

$$\ell_c(\tau) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left\{\log\delta_g + \sum_{\boldsymbol{j}_{ig}\in\mathcal{C}_J} v_{i\boldsymbol{j}_{ig}}\log\phi_p(\boldsymbol{y}_i + 2\pi\boldsymbol{j}_{ig};\Omega_g)\right\} \tag{7}$$

where, now, $z_{ig}$ is the indicator of the $i$-th observation belonging to the $g$-th cluster and $v_{i\boldsymbol{j}_{ig}}$ is the indicator of the $i$-th observation having $\boldsymbol{j}_i$ as wrapping coefficients vector in the $g$-th group. Let $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n)$ and $\boldsymbol{V} = (\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_n)$. The conditional expected value of the complete log–likelihood in (7) should be evaluated with respect to the joint distribution of $(\boldsymbol{Z}, \boldsymbol{V})$ given the observed data and the current parameter value $\tau'$. The problem can be reformulated as computing

$$Q(\tau|\tau') = E_{\boldsymbol{Z}}\left\{E_{\boldsymbol{V}|\boldsymbol{Z}}[\ell_c(\tau)|\boldsymbol{y},\tau']\right\}.$$

Then, based on the current parameter value, the E-step articulates in an outer and an inner step. The outer step requires the calculation of the conditional expectation of each $\boldsymbol{Z}_i$ given the data, whereas the inner step concerns the calculation of the conditional expectation of each $\boldsymbol{V}_i$ given the data and the group memberships, that is the conditional probabilities in (3) are evaluated for each group.

**Remark**. We notice that the number of unknowns $(\boldsymbol{z}_i, \boldsymbol{v}_i)$, $i = 1, \ldots, n$, rapidly becomes quite large and equal to $[1+(2J+1)^p]\times n\times G$. The computation of conditional probabilities about group membership and wrapping coefficients could have been interchanged but with a consequent even larger number of unknowns equal to $[1 + G] \times n \times (2J + 1)^p$, as long as $G < (2J + 1)^p$.

In the M-step, the parameter value is updated according to $\tau = \text{argmax}\, Q(\tau|\tau')$. As well as in (4), closed form expressions are available, that are detailed in Algorithm 1 (see page 31).

The classification EM (CEM, Celeux and Govaert (1992)) represents a valid alternative in mixture model fitting and model based clustering. According to this approach, the E-step of the classical EM algorithm can be enhanced by an outer classification step for which

we let

$$g_i = \text{argmax}_g z_{ig}$$

and $z_{ig} = 1$ for $g = g_i$ and zero otherwise. In Section 2, we also stressed the possibility to use a classification approach in the fitting process of a multivariate Wrapped Normal distribution (Nodehi et al, 2020). Therefore, one could also consider an inner classification step, that is setting

$$\tilde{\boldsymbol{j}}_{ig} \quad = \quad \text{argmax}_{\boldsymbol{j}_i \in \mathcal{C}_J} v_{i\boldsymbol{j}_i g},$$

and $v_{i\boldsymbol{j}_{ig}} = 1$ for $\boldsymbol{j}_{ig} = \tilde{\boldsymbol{j}}_{ig}$, and zero otherwise. As a result, feasible different estimation strategies to fit the WN mixture model (5) are

i. Expected EM (E-EM) with both outer and inner E step;

ii. Expected CEM (E-CEM) with outer E step and inner CE step;

iii. Classification EM (CE-EM) with outer CE step and inner E step;

iv. Classification CEM (CE-CEM) with both outer and inner CE step.

The generic iteration, allowing all the possible strategies listed above, is summarized in Algorithm 1, on page 31.

For what concerns the task of clustering, the use of an outer CE step automatically leads to classify all the sample units, whereas, when using the outer E step, cluster assignments follow from the application of a classification step after the last E-step, that is units are assigned to one group according to the largest conditional probability $z_{ig}$ at convergence.

## 3.1 Initialization issues

The algorithm may be initialized by partitioning the data according to an agglomerative hierarchical clustering technique. Since the classical statistics in use for linear variables are inappropriate for circular data, one should make use of circular distance measures between (vector of) angles $(\boldsymbol{y}', \boldsymbol{y})$ to define reliable dissimilarities (Lund, 1999). Suitable distance measures, that are available from the `R` package `circular` (Agostinelli and Lund, 2017) are:

a. angular separation: $d(\boldsymbol{y'}, \boldsymbol{y}) = \sum_{j=1}^{p} \left[1 - \cos(y'_j - y_j)\right]$;

b. chord: $d(\boldsymbol{y'}, \boldsymbol{y}) = \sum_{j=1}^{p} \sqrt{2\left[1 - \cos(y'_j - y_j)\right]}$;

c. geodesic: $d(\boldsymbol{y'}, \boldsymbol{y}) = \sum_{j=1}^{p} \left(\pi - |\pi - |y'_j - y_j||\right)$.

The initial clustering will also depend on the choice of the agglomerative technique, that could be one among Ward, complete link, single link, average link and even more techniques available from the rich literature about hierarchical clustering and available from the `R` packages `cluster` (Maechler et al, 2021) and `NbClust` (Charrad et al, 2014). As an alternative approach, $k-$means is also feasible. Once an initial partition has been made available, initial parameters values can obtained through moments based cluster-wise estimates (Mardia et al, 2012) or maximum likelihood estimation under the WN model. When the task of finding an initial partition becomes computationally demanding, a possible solution could be to resort to subsampling and obtain an initial partition from a randomly chosen subset of the available data.

In order to avoid the algorithm to be dependent on the selected starting parameters' values and the likelihood to be trapped in local maxima, it is advised to run Algorithm 1 from several initial solutions, especially when subsampling has been employed. In those circumstances in which initialization by hierarchical clustering is feasible, an appealing and effective strategy to obtain different starting values is to run the algorithm from some initial partition, as described above and, then, to perturb randomly the estimate at convergence to get new starting values (Farcomeni and Greco, 2015). As several initial values are used, the estimate leading to the largest mixture log-likelihood (6) will be selected, in a very common fashion.

## 3.2 Synthetic examples

Here, we consider some illustrative examples based on samples of synthetic data of size $n = 300$. We only present the results stemming from the E-EM algorithm. The findings from the other methods are illustrated in the Appendix included in the supplemental materials.

In the first pair of synthetic examples, data are drawn from a bivariate three components heteroskedastic mixture model with equal membership probabilities, vector means corresponding to the rows of

$$\boldsymbol{\mu} = \begin{pmatrix} m & 2\pi - m \\ 2m & 2m \\ 2\pi - m & m \end{pmatrix} \tag{8}$$

with $m = \pi/6$ and variance-covariance matrices $\Sigma_g = \sigma_g^2 R(\rho_g)$, where $R(\rho_g)$ is a bivariate correlation matrix with correlation $\rho_g$, $(\rho_1, \rho_2, \rho_3) = (0.3, -0.6, -0.3)$ and $(\sigma_1, \sigma_2, \sigma_3) = (\sigma, \sigma, \sigma/2)$. The parameter $\sigma$ allows to control the degree of overlapping over the three groups. In the first synthetic dataset $\sigma = \pi/8$, whereas $\sigma = \pi/6$ in the second.

The left panel of Figure 1 displays the Ramachandran plot of the angles $\boldsymbol{y}$ over $[-\pi, \pi)^2$, with the true component's densities overimposed (in the form of tolerance ellipses) and true assignments. We need to note that this scatterplot cannot show the wraparound nature of the angles. This representation is not unique and its appearance depends on how the angles are represented, e.g., in $[0, 2\pi)$ instead of $[-\pi, \pi)$. In the former situation, the scatterplot would look very different. One should be aware of the fact that the top and bottom boundaries in these plots join together, as do the left and right boundaries. In order to highlight these features, the fitted model has been represented on a flat torus in the right panel of Figure 1, after that the circular data have been unwrapped over $\mathcal{C}_J$. The results are very similar among all algorithms. The over-imposed component-specific tolerance ellipses are based on the 0.95-level quantile of the $\chi^2_2$ distribution. The dotted lines give the partition of both axes by multiples of $\pi$. The result is quite accurate, both in terms of fitting and classification. Figure 2 gives the results stemming from the second synthetic sample, that is characterized by a more challenging overlapping. The result is still quite satisfactory.

In the third illustrative example, as suggested by one anonymous referee, we considered the case in which one of the clusters has high variance, setting $(\sigma_1, \sigma_2, \sigma_3) = (\sigma, 8\sigma, \sigma/2)$ for $\sigma = \pi/8$. The location matrix is the same as before. In this situation, it is likely to suffer some lack of accuracy in the estimation of the variance-covariance structure of the

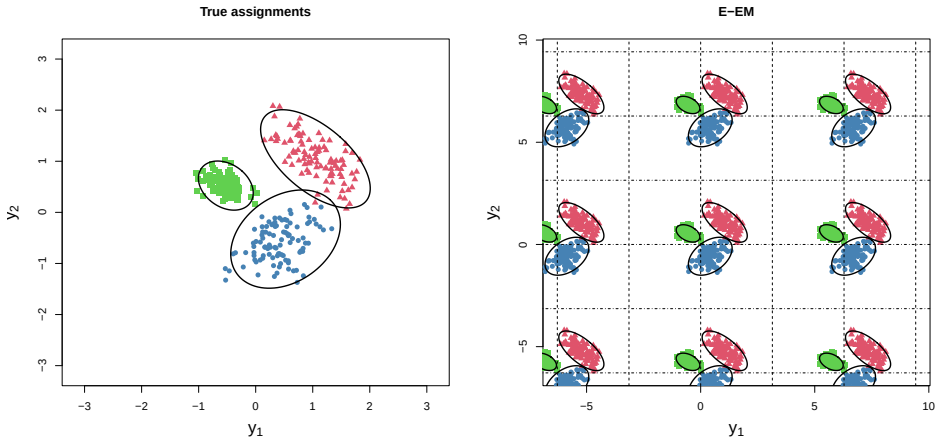


Figure 1: First synthetic example. Ramachandran plot over $[-\pi, \pi)^2$ with true model and cluster assignments (left). Fitted components and model based classification by E-EM on a flat torus (right). The dotted lines give the partition of both axes by multiples of $\pi$.

more dispersed cluster. However, the use of both the outer and inner E step in maximum likelihood estimation of the WN mixture model mitigates the problem, as it can be seen from the comparison between the alternative strategies given in the Appendix. It is worth noting that each estimation strategy still leads to a noticeable classification accuracy. The results are displayed in Figure 3. We also notice that data points belonging to the high variance cluster act as they were background noise in the Ramachandran plot.

We further explored the extreme case of two (well separated) clusters and a genuine uniform background noise. The clusters have mean vectors corresponding to the first two rows of (8) with $m = \pi/3$, respectively, and scatter matrices equal to those with non inflated variance in the previous example. We still fitted a three components mixture model. Also in this case, the three Wrapped Normal components mixture model successfully fits the two genuine clusters and accurately detect background noise: the results are displayed in Figure 4.

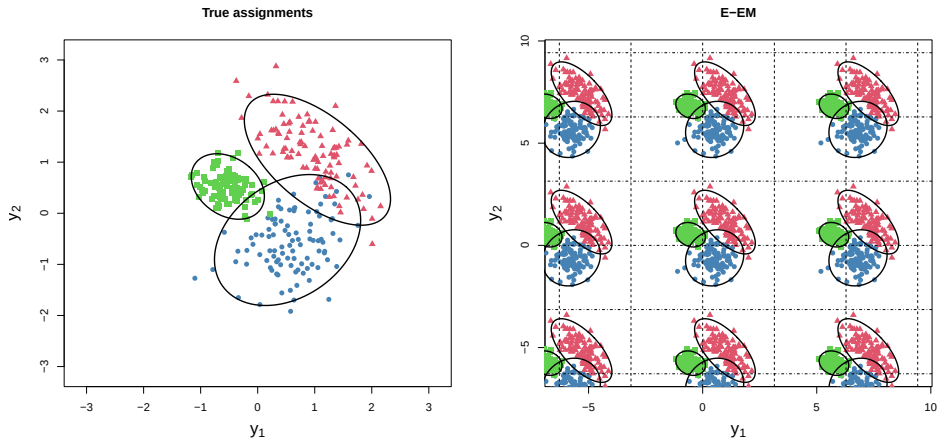


Figure 2: Second synthetic example. Ramachandran plot over $[-\pi, \pi)^2$ with true model and cluster assignments (left). Fitted components and model based classification by E-EM on a flat torus (right). The dotted lines give the partition of both axes by multiples of $\pi$.

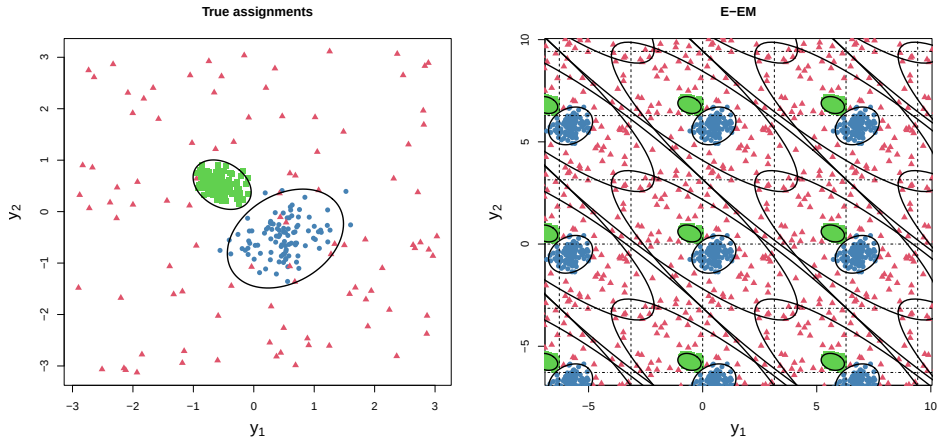


Figure 3: Third synthetic example. Ramachandran plot over $[-\pi, \pi)^2$ with true model and cluster assignments (left). Fitted components and model based classification by E-EM on a flat torus (right). The dotted lines give the partition of both axes by multiples of $\pi$.

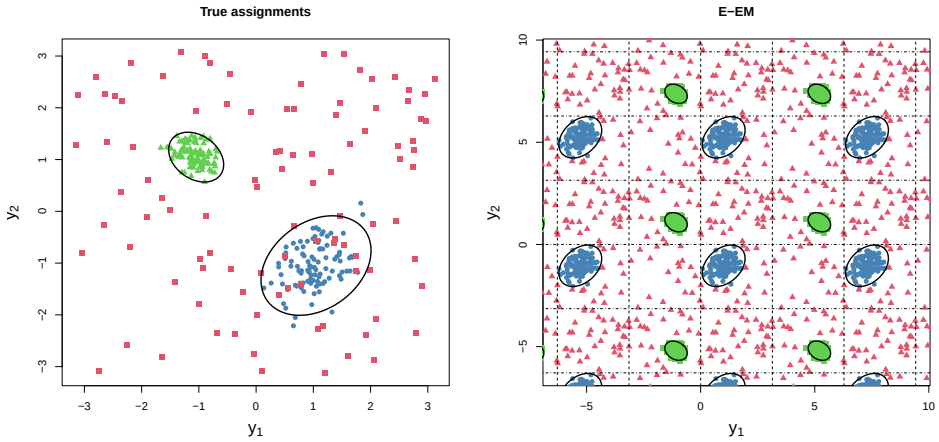


Figure 4: Fourth synthetic example. Ramachandran plot over $[-\pi, \pi)^2$ with true model and cluster assignments (left). Fitted components and model based classification by E-EM on a flat torus (right). The dotted lines give the partition of both axes by multiples of $\pi$.

# 4 Numerical studies

In this section, we investigate the finite sample behavior of the proposed techniques to fit a mixture of multivariate WN distributions. The fitting procedure is still based on non optimized `R` code and is available from the supplemental materials, according to the descriptions in Section 7. As suggested by one anonymous referee, we compared the proposed methods, that we developed under the WN mixture model with $J = 3$, to the existing alternatives based on the multivariate von Mises distribution (Mardia et al, 2012; Mardia and Voss, 2014): maximum likelihood estimation is carried out according to both a CEM (vM-CEM) and an EM (vM-EM) strategy (we remark that only the latter has been outlined in Mardia et al (2012)), whose M-step consists of moments based estimating equations. In order to provide a fair comparison of the different techniques, we selected a pair of data generation schemes. The first one has been described in Subsection 3.2 for the case $p = 2$ and what concerns the first couple of synthetic examples. Here, in addition, we also considered the situation with $p = 5$. In the latter case, data have been obtained by adding three uncorrelated components with null mean vector and variance equal to 0.5. Both overlapping scenarios have been considered, that is with $\sigma = \pi/8$ (scenario A) and $\sigma = \pi/6$ (scenario B).

The other data generation scheme allows to generate samples from a multivariate von Mises distribution according to an acceptance-rejection sampling algorithm, as implemented in the `R` package `BAMBI` (Chakraborty and Wong, 2021; Mardia et al, 2007; Mardia and Voss, 2014)). We set $\boldsymbol{\mu}$ as in (8) with $m = \pi/8$ (scenario C) and $m = \pi/6$ (scenario D), whereas $\Sigma_g^{-1} = \begin{pmatrix} k_{1g} & k_{3g} \\ k_{3g} & k_{2g} \end{pmatrix}$ with $\boldsymbol{k_1} = \boldsymbol{k_2} = (10, 10, 20)$ and $\boldsymbol{k_3} = (2, -2, 4)$. For this latter data configuration, the smaller $m$, the larger the overlapping among clusters. We stress that this approach allows to investigate properly the behavior of the proposed methodology under misspecification.

Numerical studies are based on 500 Monte Carlo trials. All the algorithms are assumed to reach convergence when the difference among consecutive mixture log-likelihood values is below the tolerance value $10^{-6}$. The initial partition has been obtained using the geodesic or angular separation distance and the Ward agglomerative method. Starting values are driven from cluster-wise moments based estimating equations. Fitting accuracy has been evaluated according to the following measures:

1. $\text{ave}_g \left\{\text{ave}_p[1 - \cos(\hat{\boldsymbol{\mu}}_g - \boldsymbol{\mu}_g)]\right\}$,
2. $\text{ave}_g \left[\text{tr}\left(\hat{\Sigma}_g \Sigma_g^{-1}\right) - \log\left(\det\left(\hat{\Sigma}_g \Sigma_g^{-1}\right)\right) - p\right]$,
3. $||\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}||$,

where $\text{ave}_g$ denotes averaging over the $G$ clusters and $\text{tr}(A)$ and $\det(A)$ are the trace and the determinant of the matrix $A$, respectively. On the other hand, classification accuracy has been measured by the Adjusted Rand Index (ARI, Hubert and Arabie (1985)) between the obtained partitions and the true component memberships of the observations. In order to avoid problems due to label switching issues, cluster labels have been sorted according to the first entry of the fitted location vectors. Figure 5 displays the distribution of the accuracy measures introduced above for both overlapping scenarios, A and B, when $p = 2$. The results are satisfactory and computational time always lies in a feasible range (see Table 1). We notice that there are minor differences between the four algorithms based on the WN mixture, albeit the use of the outer E step always leads to a better performance,

and those based on the von Mises mixture model. In scenario B, the larger overlap leads to a slightly worse accuracy, both in parameters estimation and classification. In general, the results from E-EM and E-CEM are always slightly better than those stemming from the use of the von Mises mixture model, both in terms of fitting and classification accuracy, especially under scenario B. Figure 6 gives the results for $p = 5$. Fitting and classification accuracies are still remarkable for all considered techniques. The findings are very similar to those obtained for $p = 2$, with the best performances from the estimation of a WN mixture model and the use of the outer E-step. However, computational time increased remarkably, as the number of parameters and unknowns grew. Table 1 gives the median times in seconds needed to reach convergence from one starting point under scenario B, with a 3.4 GHz Intel Core I5 processor, whereas Figure 9 displays the corresponding box-plots. The CE-CEM algorithm leads to a remarkable gain in computational time. Since we observed only very small differences in the value of the log-likelihood at convergence among the four methods, the reduction in computational burden driven by the CE-CEM makes the method particularly appealing, especially for moderate and large sizes and/or dimensions. It is worth noting that computational time can be reduced by setting $J = 1$.

Figures 7 and 8 display the distribution of the accuracy measures for both overlapping scenarios, C and D, when $p = 2, 5$, respectively. Despite the different data generation scheme, the results are in strong agreement with those illustrated above. In all cases, the proposed method based on the WN mixture model provides satisfactory fitting and classification accuracies, that are not different from those stemming from estimating the von Mises mixture model.

# 5 Real data examples

In the following, we consider two real data examples of increasing complexity. The first concerns bivariate circular data, whereas, the second involves data on a seven dimensional torus. In both cases the number of groups is selected according to a BIC criterion. Here, we only fit a multivariate WN mixture model to the data. The reader is pointed to the Appendix in the supplemental materials for a comparison with the results stemming from

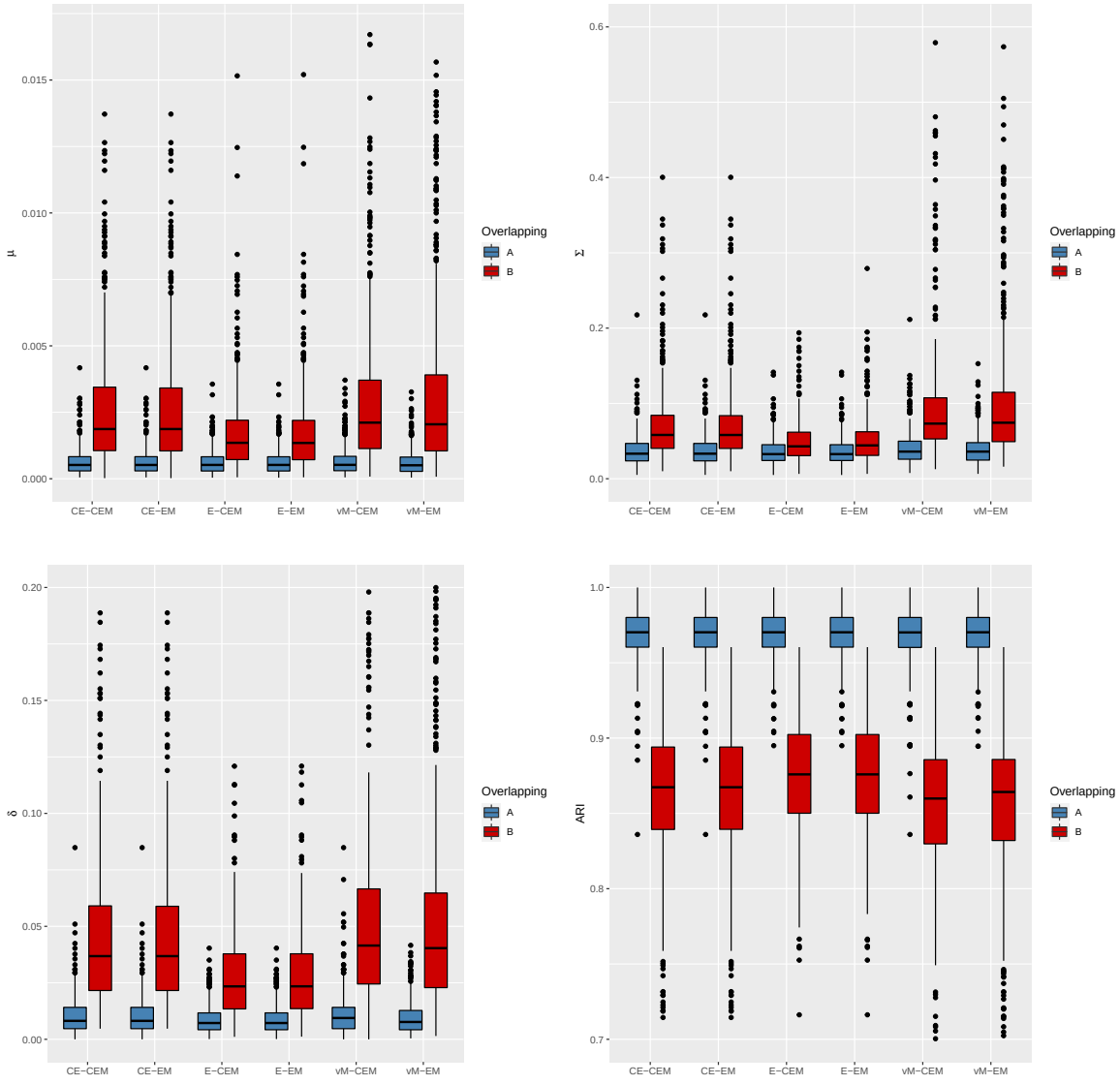


Figure 5: Numerical studies. Box-plots for accuracy measures with $n = 300$, $G = 3$, $p = 2$, $\boldsymbol{\delta} = (1/3, 1/3, 1/3)$, $\sigma = \pi/8$ (A) and $\sigma = \pi/6$ (B). Data are generated under the WN mixture model.

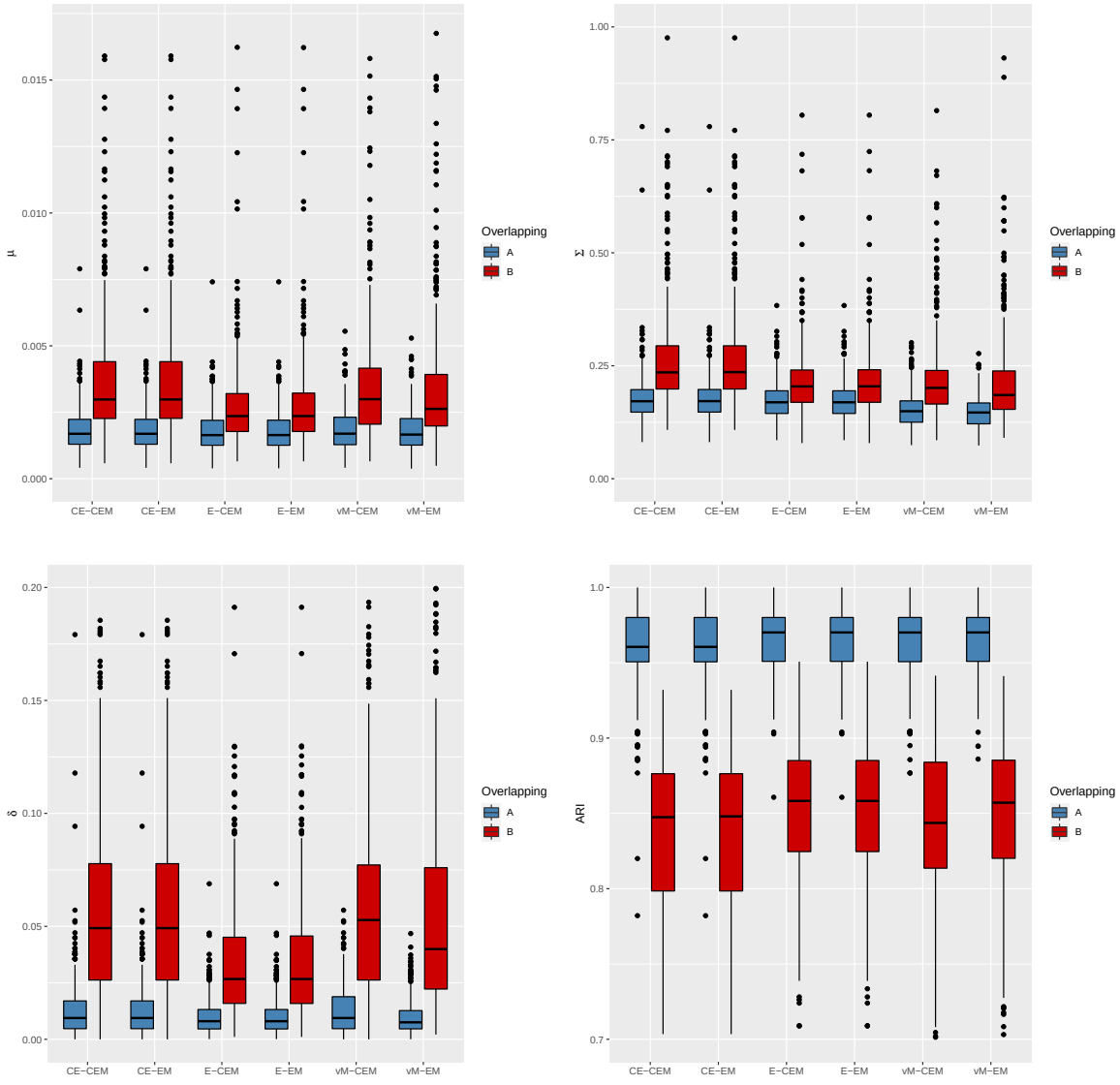


Figure 6: Numerical studies. Box-plots for accuracy measures with $n = 300$, $G = 3$, $p = 5$, $\boldsymbol{\delta} = (1/3, 1/3, 1/3)$, $\sigma = \pi/8$ (A) and $\sigma = \pi/6$ (B). Data are generated under the WN mixture model.

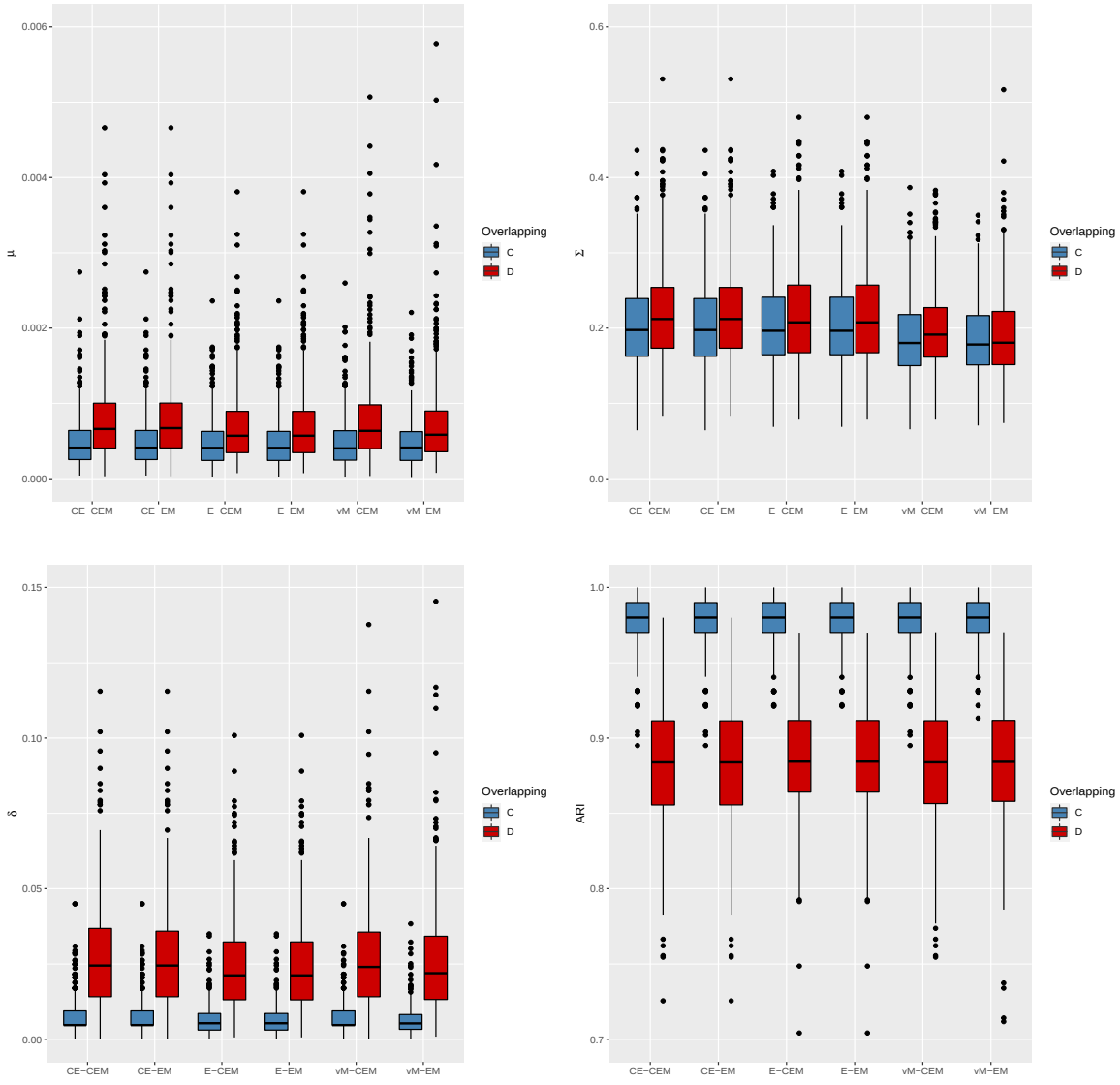


Figure 7: Numerical studies. Box-plots for accuracy measures with $n = 300$, $G = 3$, $p = 2$, $\boldsymbol{\delta} = (1/3, 1/3, 1/3)$, $m = \pi/6$ (C) and $m = \pi/b$ (D). Data are generated under the von Mises mixture model.

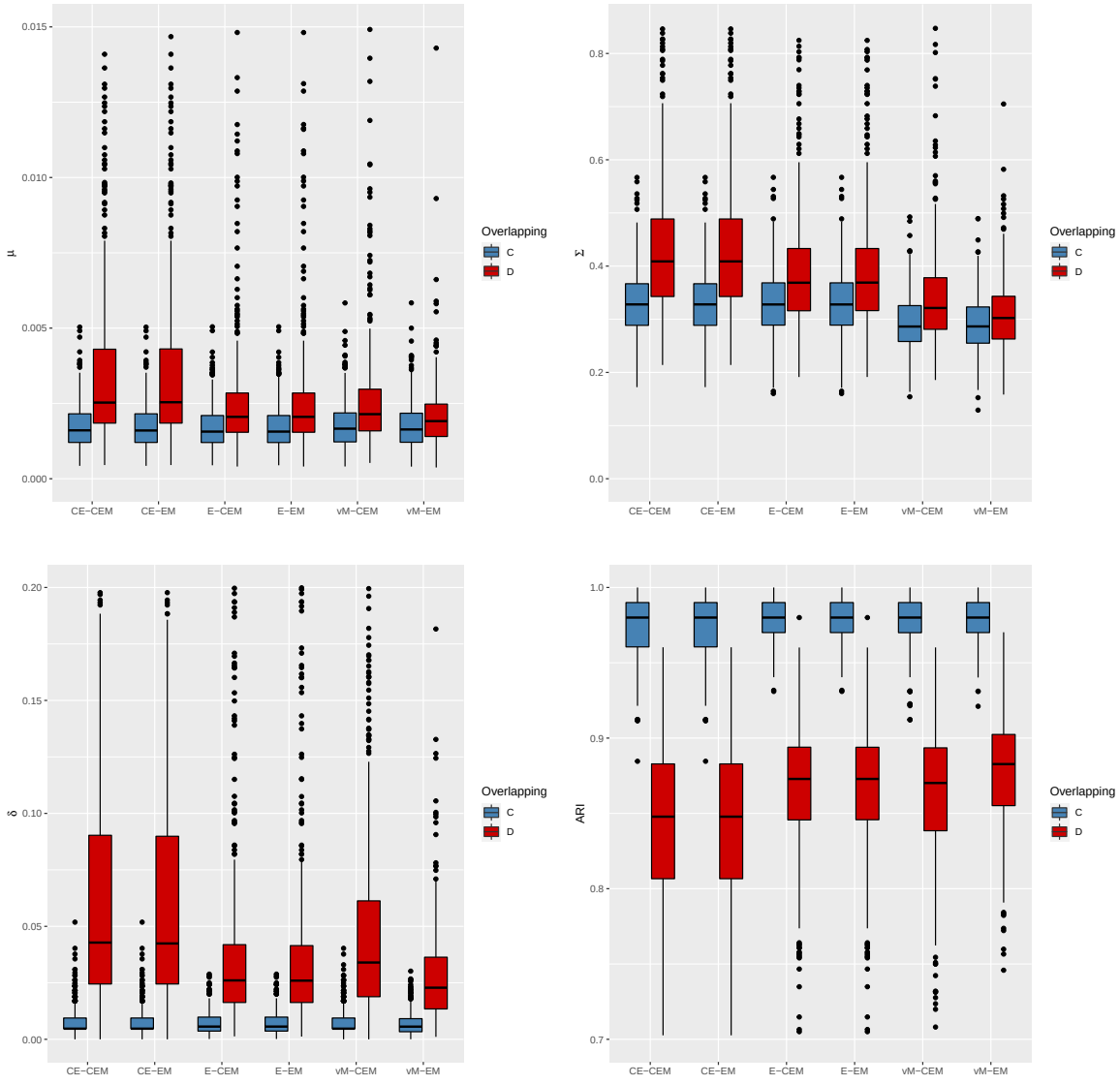


Figure 8: Numerical studies. Box-plots for accuracy measures with $n = 300$, $G = 3$, $p = 2$, $\boldsymbol{\delta} = (1/3, 1/3, 1/3)$, $m = \pi/6$ (C) and $m = \pi/b$ (D). Data are generated under the von Mises mixture model.

Table 1: Numerical studies. Median (with mad) computational time in seconds from one starting point under scenario B.

| | CE-CEM | CE-EM | E-CEM | E-EM |
|---|---|---|---|---|
| $p = 2$ | **1.507** | 1.627 | 2.340 | 2.886 |
| | (0.414) | (0.450) | (0.893) | (1.142) |
| $p = 5$ | **43.358** | 74.010 | 71.864 | 78.197 |
| | (6.313) | (37.424) | (32.378) | (23.148) |

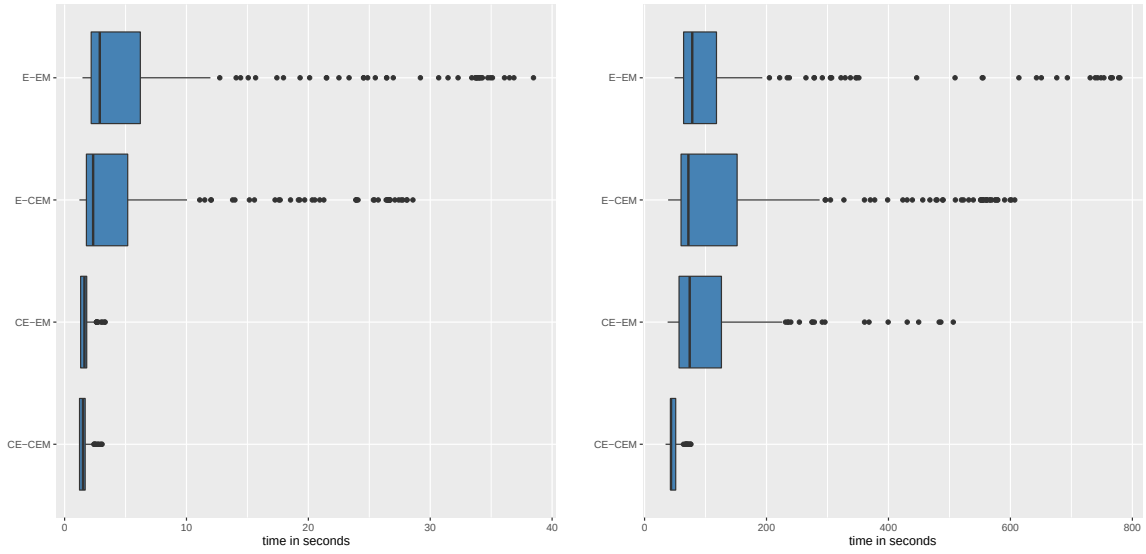


Figure 9: Numerical studies. Box-plots of computational time in seconds with $p = 2$ (left) and $p = 5$ (right) with $G = 3$, $J = 3$, under scenario B to fit the WN mixture model.

the use of a von Mises mixture model, instead. Data and `R` code to replicate them are also provided. We anticipate that results are always very close.

## 5.1 Protein data

We consider data about a protein domain extracted from the Structural Classification of Proteins (SCOP), that provides a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. The measurements give $n = 244$ pairs of critical backbone angles (Nodehi et al, 2020). We fit

Table 2: Protein data. BIC values for the selection of the number of clusters by different estimation methods. For $G = 1$, CE-CEM is CEM and E-EM is EM. All the methods suggest $G = 3$ (in bold), with the lowest BIC returned by E-CEM.

| | Number of clusters ($G$) | | | |
|---|---|---|---|---|
| Method | 1 | 2 | 3 | 4 |
| CE-CEM | 870.09 | 483.38 | **462.58** | 467.00 |
| CE-EM | – | 483.04 | **462.46** | 466.61 |
| E-CEM | – | 482.284 | **460.65** | 464.77 |
| E-EM | 869.947 | 482.549 | **461.73** | 472.77 |

a mixture of bivariate Wrapped Normal distributions: maximum likelihood estimation is pursued according to Algorithm 1 and the four strategies listed in Section 3. The methods developed in Nodehi et al (2020) have been used for $G = 1$. The number of clusters has been selected according to the BIC. The entries in Table 2 give the BIC values for each method. For all four considered estimation techniques, the BIC leads to select three clusters. This result is in close agreement with the robust analyses in Saraceno et al (2021); Greco et al (2021), where the non homogeneity of the data and the presence of sub-populations was claimed. The E-CEM always returned the largest likelihood. Cluster specific estimates are given in Table 3. The classification agrees with the visual inspection of the data. Cluster 1, whose size is 92, is composed by the subset of those bivariate points (steelblue in the color version) that highlight the smallest variability on both marginal distributions. The other two clusters are composed by 48 (red) and 104 (green) points, respectively. They are both characterized by a larger variability, especially with respect to the second critical angle. In particular, it is worth noting that the mean vectors on the second critical angle characterizing cluster 1 and cluster 3 have almost opposite directions. The fitted model and the resulting classification are also displayed in the Ramachandran plot in Figure 10: we remark on the periodic boundary conditions for a better comprehension of the plot.

Table 3: Protein data. Cluster specific mean direction, variance-covariance entries and size by E-CEM.

| Cluster | $\hat{\mu}_1$ | $\hat{\mu}_2$ | $\hat{\Sigma}_{11}$ | $\hat{\Sigma}_{22}$ | $\hat{\Sigma}_{12}$ | size |
|---|---|---|---|---|---|---|
| 1 | 1.593 | 0.874 | 0.005 | 0.011 | -0.003 | 92 |
| 2 | 1.656 | 1.186 | 0.016 | 0.405 | -0.017 | 48 |
| 3 | 2.099 | 3.894 | 0.061 | 0.770 | 0.038 | 104 |

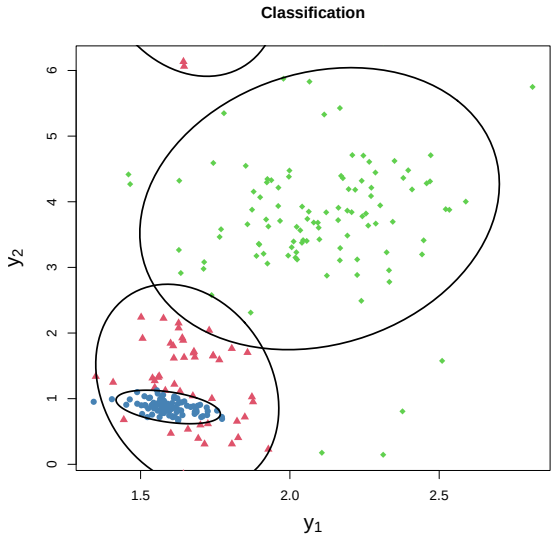


Figure 10: Protein data. Ramachandran plot: fitted components and model based classification by E-CEM.

## 5.2 RNA data

RNA molecules are made by stringing together individual ribonucleotides, always by adding the 5-phosphate carbon-sugar group of one nucleotide onto the 3-hydroxyl group of the previous nucleotide. Moreover, each RNA strand is also composed of nitrogenous bases covalently bound to the carbon-sugar phosphate backbone (Clancy, 2008; Eltzner et al, 2018). RNA has a *rotameric* nature governed largely by bond torsions: for each nucleotide, there are seven independent torsion angles, of which six backbone torsion angles and one angle that describes the rotation of the base relative to the sugar (Hershkovitz et al, 2006). Here, we consider all seven dimensions of RNA to investigate possible grouping elements

of interaction and conformation. In order to illustrate the potentialities of the proposed methodology, we focus on a (sub-)sample of size $n = 920$ taken from the large RNA dataset from Wadley et al (2007) and also used in Sargsyan et al (2012) and Nodehi et al (2020).

The data have been previously classified based on a couple of *pseudo*-torsion angles $(\eta, \theta)$ and the sugar pucker, instead of the seven conventional torsion angles, $\alpha$, $\beta$, $\gamma$, $\delta$, $\epsilon$, $\xi$, $\chi$. According to the classification stemming from pseudo-torsion angles, there are three clusters. We first performed a confirmatory analysis to study the agreement between the existing classification and that driven by the use of the proposed model based clustering technique. Maximum likelihood estimation of the WN mixture model has been run according to the E-EM algorithm. The three clusters that we obtained are similar to those stemming from the pseudo-torsion angles. The ARI is 0.946, whereas only 26 nucleotides are assigned to different clusters, as displayed in Table 4.

Then, we moved to an exploratory analysis, in order to investigate the presence of possible, otherwise hidden, substructures in the data at hand. Table 5 gives the BIC values to select the number of clusters stemming from the application of the E-EM algorithm over the 7-dimensional torus data: the lowest BIC is returned for $G = 5$. From the inspection of Table 6, we notice that the groups based on pseudo-torsion angles have been split, hence unveiling new elements of interaction and conformation among data.

The entries in Table 7 give the cluster specific mean vectors of torsion angles in degrees. Each group exhibits some peculiarity. We notice that the first group is mainly characterized by the largest centroid value for $\beta$ and $\gamma$ but the lowest for $\alpha$, whereas the second group differs from the first just in the values for $\alpha$ and $\beta$. The third cluster shows the lowest centroid for $\gamma$. The units in the fourth cluster exhibit a similar pattern to those in the third group but for the torsion angle $\alpha$. The fifth group is characterized by the largest values for $\alpha$ and $\chi$. In summary, we can observe that the five groups are well separated by the torsion angle $\alpha$. In a different fashion, the torsion angles $\beta$ and $\gamma$ are able to discriminate among at least three and two sub-groups, respectively, whereas the other centroids profiles allow only a poor separation among clusters.

Further differences can be unveiled by looking at the cluster specific correlation matrices

Table 4: RNA data. Confusion matrix comparing the classification from E-EM based on the seven torsion angles and that based on the pair of pseudo-torsion angles, for $G = 3$.

| | Pseudo-Torsion Angles | | |
|---|---|---|---|
| Cluster | 1 | 2 | 3 |
| 1 | 477 | 0 | 0 |
| 2 | 0 | 190 | 7 |
| 3 | 0 | 19 | 225 |

Table 5: RNA data. BIC values for the selection of the number of clusters by E-EM, for $G = 5$.

| G | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| BIC | 2222.42 | 1258.23 | 1126.13 | **738.56** | 845.92 |

displayed in Figure 11. In the first cluster, we highlight the large negative correlations between $(\chi, \epsilon)$, $(\xi, \epsilon)$, $(\gamma, \beta)$ and the positive correlation $(\delta, \gamma)$. The correlation structures in the second and third clusters are characterized by small correlations but for the negative value for the pair $(\alpha, \gamma)$. Correlations are generally low or mild in the remaining clusters.

# 6 Concluding remarks

We proposed an effective strategy to fit a mixture model to multivariate circular data that lies on a $p$-dimensional torus. The methodology relies on the assumption that the mixture components are Wrapped Normal and exploits a data augmentation estimation approach that treats the wrapping coefficients vectors, as well as cluster memberships, as latent variables. Then, maximum likelihood estimation can be performed through a suitable extension of the classical EM algorithm for finite mixture models, characterized by an outer and inner E-step for cluster membership and wrapping coefficients, respectively. A useful alternative is provided by the Classification EM algorithm: here, a classification step can enhance both the outer and inner E-step before the M-step. This choice ap-

Table 6: RNA data. Confusion matrix comparing the classification from E-EM based on the seven torsion angles and that based on the pair of pseudo-torsion angles, for $G = 5$.

| | Pseudo-Torsion Angles | | |
|---|---|---|---|
| Cluster | 1 | 2 | 3 |
| 1 | 66 | 0 | 0 |
| 2 | 410 | 0 | 0 |
| 3 | 0 | 116 | 0 |
| 4 | 0 | 80 | 71 |
| 5 | 1 | 15 | 161 |

pears very appealing, especially for large datasets, when the number of unknowns grows quickly. Actually, the use of a classification step always lead to mixture likelihood values at convergence that are close enough (or even identical) to those stemming from the EM algorithm, while being definitely less computationally demanding. In addition, outer-CEM based estimates represent desirable starting values to run the outer-EM type algorithms. The proposed method gives accurate results both for the task of parameters estimation and model based clustering of circular data. It is an effective strategy when dealing with small to moderate sizes and/or dimensions, but we also guess it may represent a valid starting point to develop further classification tools for multivariate circular data that lies on a high dimensional torus. Moreover, it also provides accurate fit and classification under misspecification, when data are sampled from a multivariate von Mises mixture model, according to a rejection sampling methodology, and represents more than a valid alternative. Actually, the results would even warrant to state that the E-CEM and E-EM variants generally appear to outperform or at least match the fitting techniques based on the von Mises mixture model. Finally, we point out that the same methodology can be easily extended to all wrapped distributions, e.g. Wrapped Cauchy, with location and scatter parametrization.

Table 7: RNA data. Cluster specific marginal mean vector of torsion angles (in degrees) and sizes by E-EM, for $G = 5$.

| | Cluster | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| $\alpha$ | 100.44 | 155.36 | 167.94 | 235.58 | 290.72 |
| $\beta$ | 232.31 | 196.75 | 161.24 | 185.18 | 166.63 |
| $\gamma$ | 189.77 | 174.60 | 52.78 | 53.28 | 60.48 |
| $\delta$ | 85.09 | 83.82 | 85.01 | 84.02 | 83.34 |
| $\epsilon$ | 226.58 | 225.94 | 220.65 | 214.26 | 216.66 |
| $\xi$ | 281.57 | 283.78 | 289.32 | 294.71 | 266.21 |
| $\chi$ | 182.77 | 184.30 | 195.66 | 189.48 | 212.82 |
| size | 66 | 410 | 116 | 151 | 177 |

# 7 Supplementary Materials

**Appendix:** An appendix containing further results from the synthetic and real data examples (Appendix.pdf)

**Data:** A folder containing all datasets used in the paper (Examples)

**Code:** A folder with different .R files containing R code to replicate the numerical studies and the examples described in the paper and the source archive file `torus.0.0-21.tar.gz` (Rcode). A `README` file has been also provided.

**Acknowledgments**. The authors wish to thank the Editor, one Associate Editor and an anonymous referee, whose comments and suggestions helped to improve the paper.

# References

Abraham C, Molinari N, Servien R (2013) Unsupervised clustering of multivariate circular data. Statistics in medicine 32(8):1376–1382

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). URL https://r-forge.r-project.org/projects/circular/

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

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

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

Banerjee A, Dhillon I, Ghosh J, Sra S (2005) Clustering on the unit hypersphere using von mises-fisher distributions. Journal of Machine Learning Research 6(Sep):1345–1382

Celeux G, Govaert G (1992) A classification EM algorithm for clustering and two stochastic versions. Computational statistics & Data analysis 14(3):315–332

Chakraborty S, Wong SWK (2021) BAMBI: An R package for fitting bivariate angular mixture models. Journal of Statistical Software 99(11):1–69

Charrad M, Ghazzali N, Boiteau V, Niknafs A (2014) NbClust: An R package for determining the relevant number of clusters in a data set. Journal of Statistical Software 61(6):1–36

Clancy S (2008) Chemical structure of RNA. Nature Education 1(1):223

Coles S (1998) Inference for circular distributions and processes. Statistics and Computing 8(2):105–113

Dortet-Bernadet J, Wicker N (2008) Model-based clustering on the unit sphere with an illustration using gene expression profiles. Biostatistics 9(1):66–80

Eltzner B, Huckermann S, Mardia K (2018) Torus principal component analysis with applications to RNA structure. Annals of Applied Statistics 12(2):1332–1359

Farcomeni A, Greco L (2015) S-estimation of hidden markov models. Computational Statistics 30(1):57–80

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

Greco L, Saraceno G, Agostinelli C (2021) Robust fitting of a wrapped normal model to multivariate circular data and outlier detection. Stats 4(2):454–471

Hershkovitz E, Sapiro G, Tannenbaum A, Williams L (2006) Statistical analysis of RNA backbone. IEEE/ACM transactions on computational biology and bioinformatics 3(1):33–46

Hornik K, Grün B (2014) movmf: an r package for fitting mixtures of von mises-fisher distributions. Journal of Statistical Software 58(10):1–31

Hubert L, Arabie P (1985) Comparing partitions. Journal of classification 2(1):193–218

Jammalamadaka S, Kozubowski T (2017) A general approach for obtaining wrapped circular distributions via mixtures. Sankhya A 79(1):133–157

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

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

Jona Lasinio G, Gelfand A, Jona Lasinio M (2012) Spatial analysis of wave direction data using wrapped gaussian processes. The Annals of Applied Statistics 6(4):1478–1498

Kurz G, Hanebeck U (2015) Parameter estimation for the bivariate wrapped normal distribution. In: 2015 54th IEEE Conference on Decision and Control (CDC), IEEE, pp 1192–1198

Lund U (1999) Cluster analysis for directional data. Communications in Statistics – Simulation and Computation 28(4):1001–1009

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2021) cluster: Cluster Analysis Basics and Extensions. URL https://CRAN.R-project.org/package=cluster

Mardia K, Jupp P (2000) Directional Statistics. Wiley, New York

Mardia K, Voss J (2014) Some fundamental properties of a multivariate von mises distribution. Communications in Statistics – Theory and Methods 43(6):1132–1144

Mardia K, Taylor C, Subramaniam G (2007) Protein bioinformatics and mixtures of bivariate von mises distributions for angular data. Biometrics 63(2):505–512

Mardia K, Kent J, Zhang Z, Taylor C, Hamelryck T (2012) Mixtures of concentrated multivariate sine distributions with applications to bioinformatics. Journal of Applied Statistics 39(11):2475–2492

Mastrantonio G, Jona Lasinio G, Maruotti A, Calise G (2019) Invariance properties and statistical inference for circular data. Statistica Sinica 29(1):67–80

McLachlan G, Peel D (2004) Finite mixture models. John Wiley & Sons

Nodehi A, Golalizadeh M, Maadooliat M, Agostinelli C (2020) Estimation of parameters in multivariate wrapped models for data on a p-torus. Computational Statistics DOI 10.1007/s00180-020-01006-x

Peel D, Whiten W, McLachlan G (2001) Fitting mixtures of kent distributions to aid in joint set identification. Journal of the American Statistical Association 96(453):56–63

Pewsey A, Neuhäuser M, Ruxton G (2013) Circular statistics in R. Oxford University Press

Ranalli M, Maruotti A (2020) Model-based clustering for noisy longitudinal circular data, with application to animal movement. Environmetrics 31(2):e2572

Saraceno G, Agostinelli C, Greco L (2021) Robust estimation for multivariate wrapped models. Metron 79(2):225–240

Sargsyan K, Wright J, Lim C (2012) GeoPCA: a new tool for multivariate analysis of dihedral angles based on principal component geodesics. Nucleic acids research 40(3):e25–e25

Wadley L, Keating K, Duarte C, Pyle A (2007) Evaluating and learning from rna pseudotorsional space: quantitative validation of a reduced representation for rna structure. Journal of molecular biology 372(4):942–957

**Algorithm 1** Maximum likelihood estimation of model (5), described in Section 3

**outer E step**

**for** $i = 1, 2, \ldots, n$ and $g = 1, \ldots, G$ **do**

$$z_{ig} = \frac{\delta_g \phi_p^{\circ}(\boldsymbol{y}_i + 2\pi \boldsymbol{j}_{ig}; \Omega_g)}{\sum_{g=1}^{G} \delta_g \phi_p^{\circ}(\boldsymbol{y}_i + 2\pi \boldsymbol{j}_{ig}; \Omega_g)}$$

**if** Outer CE = **TRUE**

$$\begin{aligned} g_i &= \text{argmax}_g z_{ig} \\ z_{ig} &= 1 \text{ for } g = g_i \\ z_{ig} &= 0 \text{ for } g \neq g_i \end{aligned}$$

**inner E step**

**for** $\boldsymbol{j}_{ig} \in \sum_{\mathcal{C}_J}$ **do**

$$v_{i\boldsymbol{j}_{ig}} = \frac{\phi_p(\boldsymbol{y}_i + 2\pi \boldsymbol{j}_{ig}; \Omega_g)}{\sum_{\boldsymbol{j}_{ig} \in \mathcal{C}_J} \phi_p(\boldsymbol{y}_i + 2\pi \boldsymbol{j}_{ig}; \Omega_g)}$$

**if** Inner CE = **TRUE**

$$\begin{aligned} \tilde{\boldsymbol{j}}_{ig} &= \text{argmax}_{\boldsymbol{j}_i \in \mathcal{C}_J} v_{i\boldsymbol{j}_i g} \\ v_{i\boldsymbol{j}_{ig}} &= 1 \text{ for } \boldsymbol{j}_{ig} = \tilde{\boldsymbol{j}}_{ig} \\ v_{i\boldsymbol{j}_{ig}} &= 0 \text{ for } \boldsymbol{j}_{ig} \neq \tilde{\boldsymbol{j}}_{ig} \end{aligned}$$

**end for**

**end for**

**M step**

**for** $g = 1, \ldots, G$ **do**

$$\begin{aligned} \delta_g &= \frac{\sum_{i=1}^{n} z_{ig}}{n} \\ \boldsymbol{\mu_g} &= \frac{\sum_{i=1}^{n} z_{ig} \sum_{\boldsymbol{j}_i \in \mathcal{C}_J} (\boldsymbol{y}_i + 2\pi \boldsymbol{j}_{ig}) v_{i\boldsymbol{j}_{ig}}}{\sum_{i=1}^{n} z_{ig}} \\ \Sigma_g &= \frac{\sum_{i=1}^{n} z_{ig} \sum_{\boldsymbol{j}_i \in \mathcal{C}_J} (\boldsymbol{y}_i + 2\pi \boldsymbol{j}_{ig})(\boldsymbol{y}_i + 2\pi \boldsymbol{j}_{ig})^{\top} v_{i\boldsymbol{j}_{ig}}}{\sum_{i=1}^{n} z_{ig}} \end{aligned}$$



**end for**

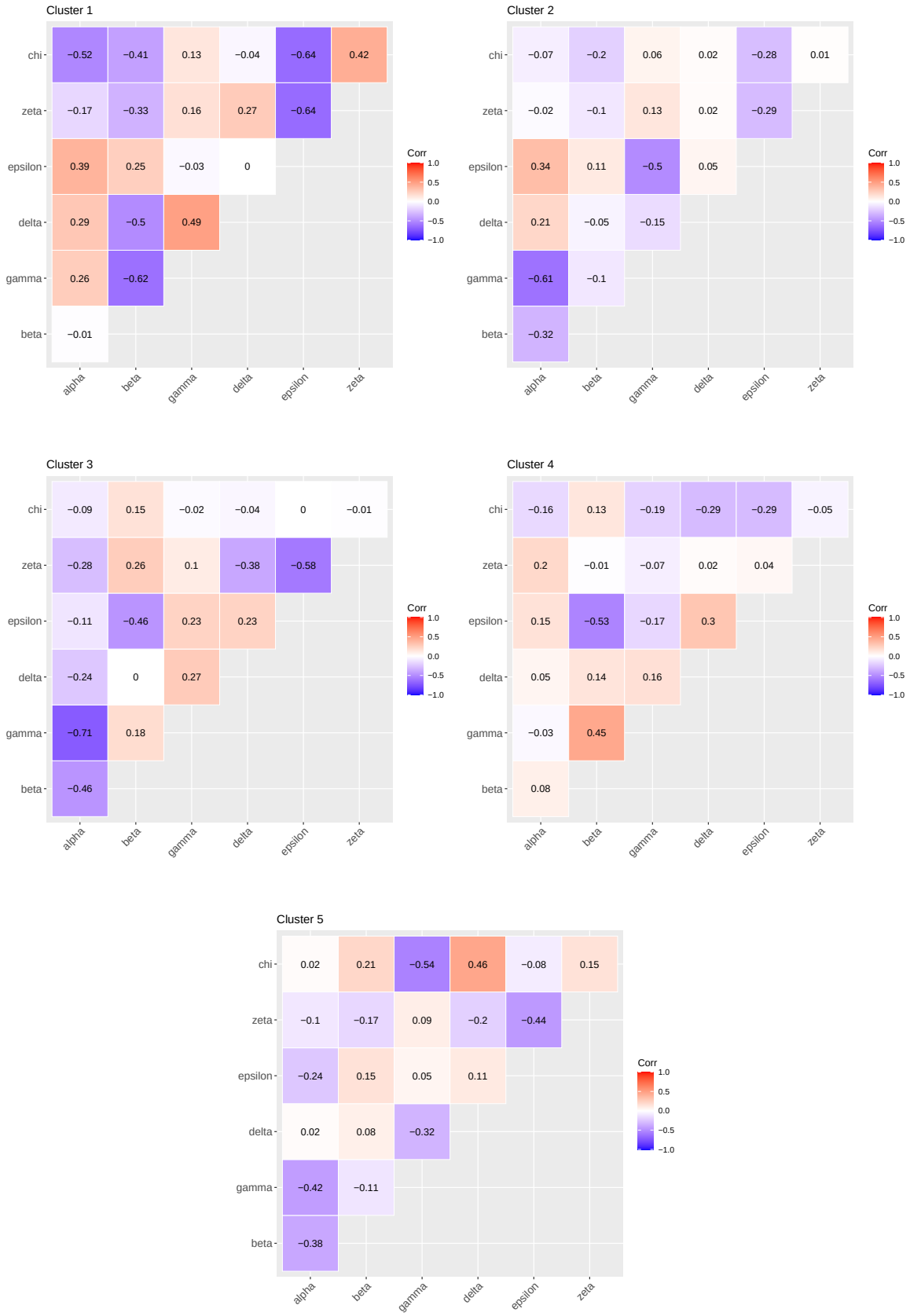




Figure 11: RNA data. Cluster specific correlation matrices by E-EM, for $G = 5$.