Robust Inference for Mixtures of Multivariate Wrapped Normal Distributions: The Weighted Likelihood Approach

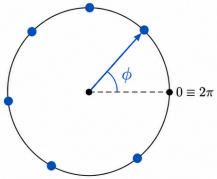
ICORS 2026, Istanbul, Türkiye

Giulia Bertagnolli*, L. Greco and C. Agostinelli

*Free University of Bozen-Bolzano

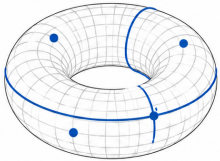
Directional data

1. Univariate directional data



$\phi \in [0, 2\pi)$

2. Bivariate directional data

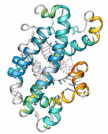


$(\phi, \psi)^\top \in [0, 2\pi) \times [0, 2\pi)$

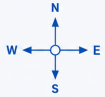
3. Multivariate directional data (p-dimensional torus)

$$\mathbb{T}^p = \underbrace{S^1 \times \cdots \times S^1}_{p}$$

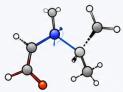
$$\phi = (\phi_1, \ldots, \phi_p)^\top \in [0, 2\pi)^p$$



Example:
Wind direction



Example:
Protein dihedral angles



Examples:
Wind directions at different stations

RNA (rotameric nature)

- Key feature of directional data: **periodicity**

Multivariate Wrapped Distributions I

Definition (Wrapped Distributions)

Consider $\boldsymbol{X} \in \mathbb{R}^p$ with density function $m(\boldsymbol{x}, \theta)$ and distribution function $M(\boldsymbol{x}, \theta)$ with $\theta \in \Theta$. The wrapped random variable $\boldsymbol{Y} = \boldsymbol{X} \mod 2\pi$ has density function

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

and distribution function

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

where $\boldsymbol{y} \in \mathbb{T}^p \equiv (0, 2\pi]^p$.

unibz

Multivariate Wrapped Distributions II

Wrapped Unimodal Elliptically Symmetric distributions:

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

where here we assume that h is a non-increasing function with continuous first derivative, and $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma)$.

The Wrapped Normal (WN) distribution corresponds to $h(t) = \exp\left(-\frac{t}{2}\right)$.

unibz

Maximum Likelihood I

Given a sample $(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)$ from $\boldsymbol{Y} \sim m^\circ(\cdot; \boldsymbol{\theta})$

maximise:
$$\ell^\circ(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log m^\circ(\boldsymbol{y}_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \sum_{\boldsymbol{j} \in \mathbb{Z}^p} m(\boldsymbol{y}_i + 2\pi \boldsymbol{j}; \boldsymbol{\theta}) = \ldots$$

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

solve:
$$\sum_{i=1}^{n} u^\circ(\boldsymbol{y}; \boldsymbol{\theta}) = 0$$

unibz

Maximum Likelihood II

- For the Wrapped Unimodal Elliptically Symmetric model: MLEs $\hat{\boldsymbol{\mu}}$, $\hat{\Sigma}$ are the solution of a pair of fixed point equations (through an IRW procedure) (Agostinelli, Greco, and Saraceno 2024)
- For the WN model ($h(t) = \exp\left(-\frac{t}{2}\right)$): same MLEs (approx. MLEs) from **EM** and (resp. **Classification EM**) algorithms as in (Nodehi et al. 2021), where wrapping coefficients $\boldsymbol{j} \in \mathbb{Z}^p$ are treated as **latent variables**

$$\ell_c(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{j}) = \sum_{i=1}^{n} \sum_{\boldsymbol{j}_i \in \mathbb{Z}^p} v_{i\boldsymbol{j}_i} \log m\left(\boldsymbol{y}_i + 2\pi \boldsymbol{j}_i; \boldsymbol{\theta}\right),$$

$$\text{with} \quad v_{i\boldsymbol{j}} = \frac{m(\boldsymbol{y}_i + 2\pi \boldsymbol{j}; \boldsymbol{\theta})}{\sum\limits_{\boldsymbol{k} \in \mathbb{Z}^p} m(\boldsymbol{y}_i + 2\pi \boldsymbol{k}; \boldsymbol{\theta})}$$

Mixtures of Wrapped Normal Distributions I

Consider a finite mixture model whose K components are distributed according to a $p-$dimensional WN variate (Greco, Inverardi, and Agostinelli 2023):

$$f^{\circ}(\boldsymbol{y};\boldsymbol{\theta}) = \sum_{k=1}^{K} \omega_k m_k^{\circ}(\boldsymbol{y};\boldsymbol{\theta}_k)$$

with $\boldsymbol{\theta} = (\omega_1, \ldots, \omega_G, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$:

- ω_k: membership probabilities
- $\boldsymbol{\theta}_k$: component specific parameters.

The operation of mixing and wrapping commute $f^{\circ}(\boldsymbol{y};\boldsymbol{\theta}) = \sum_{\boldsymbol{j}\in\mathbb{Z}^p} f(\boldsymbol{y} + 2\pi\boldsymbol{j};\boldsymbol{\theta})$.

unibz

Estimation of Mixtures of WNs I

Efficient EM algorithm based on the complete log-likelihood

$$\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\left\{\log\omega_k + \sum_{\boldsymbol{j}\in\mathbb{Z}^p} v_{ik}(\boldsymbol{j})\log m(\boldsymbol{y}_i + 2\pi\boldsymbol{j};\boldsymbol{\theta}_k)\right\} \tag{1}$$

- $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iK})$ denote the i-th observation component label binary vector (Assumpt.: realizations of multinomial random variables $\boldsymbol{Z}_i$)
- $\boldsymbol{v}_i = (v_{i1}, v_{i2}, \ldots, v_{iK})$ denotes the i-th component wrapping index vector, i.e. $v_{ik}(\boldsymbol{j}) = 1$ if obs. i has wrapping coefficients $\boldsymbol{j}$ in group k ($\boldsymbol{V}_i$)
- for each $\boldsymbol{y}_i$ we have $(\boldsymbol{z}_i, \boldsymbol{v}_i)$ missing components.

unibz

Estimation of Mixtures of WNs II

Evaluate the conditional expected value of the complete log-likelihood in (Equation 1) w.r.t. the joint distribution of $(\boldsymbol{Z}, \boldsymbol{V})$ given the data and current parameter values:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = \mathbb{E}_{\boldsymbol{Z}}\left[\mathbb{E}_{\boldsymbol{V}|\boldsymbol{Z}}\left[\ell_c(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta}'\right]\right]$$

- **E-step**: *outer* for group memberships $\boldsymbol{Z}_i$ and *inner* for $\boldsymbol{V}_i$ given the data, current parameter values, and $\boldsymbol{Z}_i$.
- **M-step**: parameter update according to $\arg\max_{\theta} Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$
  - For the (unimodal) elliptically symmetric model: in the M-step we solve a pair of fixed point equations for $\boldsymbol{\mu}_k$ and Σ_k for each $k = 1, \ldots, K$.

unibz

Outliers

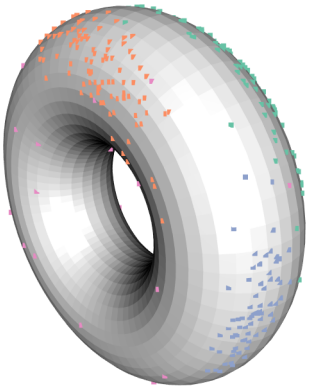


Figure 1: A three-components WN mixture and scattered outliers on a 2-torus.

The probabilistic approach:

- Outliers are *"observations that are highly unlikely to occur under the assumed model"* (*Markatou, Basu, and Lindsay 1998*; *Agostinelli 2007*)

Background on WLE I

Agreement with the model is quantified by Pearson's residuals (Lindsay 1994):

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

Solve weighted likelihood estimating equations (WLEEs):

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

where weights are obtained by *adjusting*[1] large residuals, so that outliers are down-weighted in the WLEE (Markatou, Basu, and Lindsay 1998).

[1] through a **R**esidual **A**djustment **F**unction (RAF) (Lindsay 1994)

A note on RAFs and weights

A Residual Adjustment Function (RAF) (Lindsay 1994) is a function $A(\delta)$, which is increasing and twice differentiable in $[-1, +\infty)$ and satisfies $A(0) = 0$ and $A'(0) = 1$.

- A RAF bounds the effect of large Pearson's residuals.
- RAFs connect WLE to *minimum divergence estimation*, see (Kuchibhotla and Basu 2018).

Weights are then defined as:

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

- Weights decline smoothly to zero as $\delta \to \infty$ (outliers) and, depending on the RAF, also for $\delta \to -1$ (inliers).

Robust estimation of mixtures of WNs I

Weights are specific for each observations $\boldsymbol{y}_i$ and label k of the group

$$\delta_k^\circ(\boldsymbol{y};\boldsymbol{\theta}) = \frac{\hat{f}_k{}^\circ(\boldsymbol{y})}{\hat{m}^\circ(\boldsymbol{y};\boldsymbol{\theta})} - 1$$

where

$$\hat{f}_k^\circ(\boldsymbol{y}) = \frac{1}{\sum_{i=1}^{n} z_{ik}} \sum_{i=1}^{n} z_{ik} k^\circ(\boldsymbol{y};\boldsymbol{y}_i, H)$$

with (kernel) density k° on the torus and bandwidth H, so that

$$w_{ik} = w(\delta_k^\circ(\boldsymbol{y}_i;\boldsymbol{\theta}_k))$$

unibz

Robust estimation of mixtures of WNs II

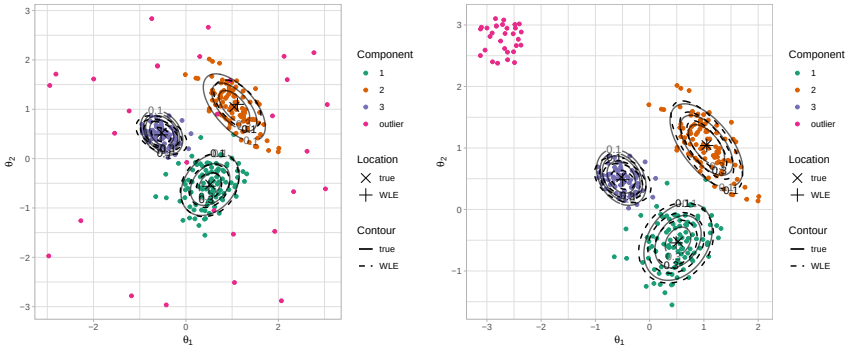
In the M-step we solve **weighted** fixed point equations:

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

$$\Sigma_k = -\frac{2}{\sum_{i=1}^{n} z_{ik} w_{ik}} \sum_{i=1}^{n} z_{ik} w_{ik} \sum_{\boldsymbol{j} \in \mathbb{Z}^p} v_{ik}(\boldsymbol{j})(\boldsymbol{y}_i + 2\pi \boldsymbol{j} - \boldsymbol{\mu}_k)(\boldsymbol{y}_i + 2\pi \boldsymbol{j} - \boldsymbol{\mu}_k)^\top$$

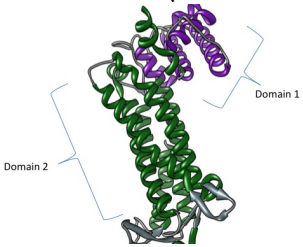
unibz

Synthetic Examples



Empirical Example I

Protein Data (Nodehi et al. 2021)



- A protein domain from the Structural Classification of Proteins (SCOP) database

- Dihedral angles, $(\phi, \psi)^\top$, determine how the protein chain bends locally

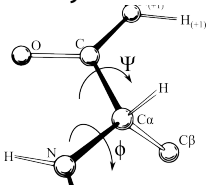


Figure 2: Adapted from: Dcrjsr, Protein backbone dihedral angles, Wikimedia Commons, CC BY 3.0.

Empirical Example II

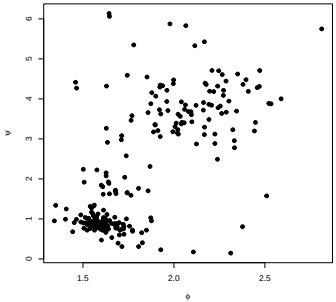


Figure 3: Ramachandran plot of (ϕ, ψ). $n = 244$ observations: residues (selected positions along the protein chain)

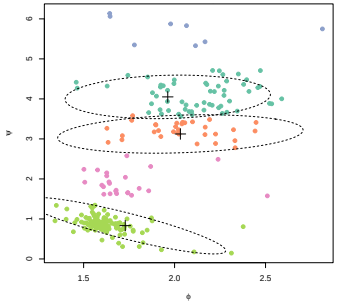


Figure 4: Hierachical clustering (angular sep. distance)

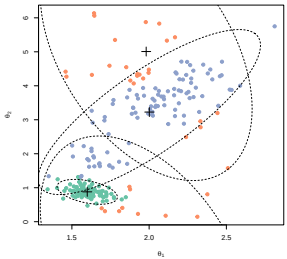


Figure 5: Robust estimates of $\boldsymbol{\mu}_k, \Sigma_k$ for $k = 1, 2, 3$ through an E-CEM.

Thanks for your attention!

Giulia Bertagnolli

giulia.bertagnolli@unibz.it

gbertagnolli.github.io

unibz

References I

Agostinelli, C. 2007. “Robust Estimation for Circular Data.” *Computational Statistics and Data Analysis* 51 (12): 5867–75.

Agostinelli, C., L. Greco, and G. Saraceno. 2024. “Weighted Likelihood Methods for Robust Fitting of Wrapped Models for p-Torus Data.” *AStA Advances in Statistical Analysis* 108: 853–88. https://doi.org/10.1007/s10182-024-00494-2.

Cressie, N., and T. R. C. Read. 1984. “Multinomial Goodness-of-Fit Tests.” *Journal of the Royal Statistical Society Series B: Statistical Methodology* 46 (3): 440–64. https://doi.org/10.1111/j.2517-6161.1984.tb01318.x.

Greco, L., P. L. Novi Inverardi, and C. Agostinelli. 2023. “Finite Mixtures of Multivariate Wrapped Normal Distributions for Model Based Clustering of p-Torus Data.” *Journal of Computational and Graphical Statistics* 32 (3): 1215–28. https://doi.org/10.1080/10618600.2022.2128808.

unibz

References II

Kuchibhotla, A. K., and A. Basu. 2018. “A Minimum Distance Weighted Likelihood Method of Estimation.” Interdisciplinary Statistical Research Unit (ISRU), Indian Statistical Institute, Kolkata, India. https://faculty.wharton.upenn.edu/wp-content/uploads/2018/02/attemptv4p1.pdf.

Lindsay, B. G. 1994. “Efficiency Versus Robustness: The Case for Minimum Hellinger Distance and Related Methods.” *The Annals of Statistics* 22: 1018–114.

Markatou, M., A. Basu, and B. G. Lindsay. 1998. “Weighted Likelihood Equations with Bootstrap Root Search.” *Journal of the American Statistical Association* 93 (442): 740–50.

Nodehi, A., M. Golalizadeh, M. Maadooliat, and C. Agostinelli. 2021. “Estimation of Parameters in Multivariate Wrapped Models for Data on Ap-Torus.” *Computational Statistics* 36: 193–215.

unibz

References III

Park, C., A. Basu, and B. G. Lindsay. 2002. “The Residual Adjustment Function and Weighted Likelihood: A Graphical Interpretation of Robustness of Minimum Disparity Estimators.” *Computational Statistics & Data Analysis* 39 (1): 21–33. https://doi.org/10.1016/s0167-9473(01)00047-0.

unibz

Appendix

Appendix I

RAFs

We choose a RAF stemming from the Generalised Kullback–Leibler (GKL) divergence

$$A(\delta, \tau) = \frac{\log(\tau\delta + 1)}{\tau}, \quad 0 \leq \tau \leq 1$$

But one can choose a RAF corresponding to Symmetric Chi-squared divergence (SCHI) (Markatou, Basu, and Lindsay 1998) or the class of Power Divergences (PWD) (Cressie and Read 1984; Park, Basu, and G. Lindsay 2002).

Appendix II

Approximated MLEs and Classification EM algorithm

An approximate MLE can be obtained using crispy assignments after the computation of $v_i(\boldsymbol{j})$, that is we let

$$\widehat{\boldsymbol{j}}_i = \arg\max_{\boldsymbol{j}\in\mathbb{Z}^p} v_i(\boldsymbol{j})$$

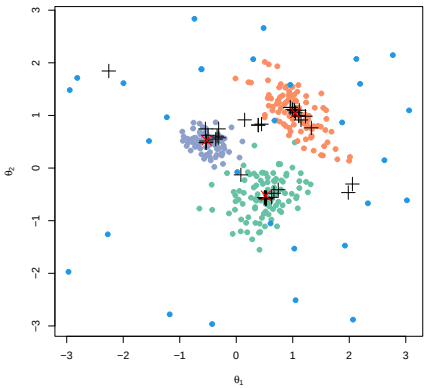
and solve the estimating equation

$$\sum_{i=1}^{n} u(\hat{\boldsymbol{x}}_i;\boldsymbol{\theta}) = \mathbf{0} \tag{2}$$

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

Appendix III

Inizialisation



1. Hierarchical clustering on circular distances (angular separation) allowing for “more” groups
2. Subsmpling inside of each *initial component*
3. For each subsample of cluster k we obtain MLE of $\boldsymbol{\mu}_k$ and Σ_k
4. For each component k we select the root $(\hat{\boldsymbol{\mu}}_k, \widehat{\Sigma}_k)$ minimising the GLK divergence between the data distribution and the WN distribution with parameters $(\hat{\boldsymbol{\mu}}_k, \widehat{\Sigma}_k)$.