Journal of Applied Statistics
Vol. 39, No. 5, May 2012, 927–945

Taylor & Francis
Taylor & Francis Group

Model-based clustering of multivariate skew data with circular components and missing values

Francesco Lagona[a]* and Marco Picone[b]

[a]*DIPES, University Roma Tre, Via Gabriello Chiabrera 199, 00145 Rome, Italy;* [b]*Department of Economics, University Roma Tre, Italy*

(*Received 5 February 2010; final version received 21 September 2011*)

Motivated by classification issues that arise in marine studies, we propose a latent-class mixture model for the unsupervised classification of incomplete quadrivariate data with two linear and two circular components. The model integrates bivariate circular densities and bivariate skew normal densities to capture the association between toroidal clusters of bivariate circular observations and planar clusters of bivariate linear observations. Maximum-likelihood estimation of the model is facilitated by an expectation maximization (EM) algorithm that treats unknown class membership and missing values as different sources of incomplete information. The model is exploited on hourly observations of wind speed and direction and wave height and direction to identify a number of sea regimes, which represent specific distributional shapes that the data take under environmental latent conditions.

Keywords: circular data; EM algorithm; latent classes; missing values; skew normal; unsupervised classification; von Mises; wave; wind

1. Introduction

Sea conditions are often monitored by taking circular and linear measurements such as wave and wind direction, wind speed and wave height. Model-based clustering of these data is helpful in identifying relevant sea regimes, that is, specific shapes that the distribution of wind and wave data takes under latent environmental conditions. In a multivariate analysis, mixture models [19] provide a general approach to classification: the joint distribution of the data is approximated by a mixture of tractable multivariate distributions, which represent cluster locations and shapes, and the clustering problem is solved as a missing value problem, by treating the unknown cluster membership of each observation as a missing value, to be estimated from the data.

Mixture-based clustering of marine data is, however, complicated by the concurrence of different supports on which the data are observed. While a pair of wind speed and wave height is a

*Corresponding author. Email: lagona@uniroma3.it

ISSN 0266-4763 print/ISSN 1360-0532 online

http://dx.doi.org/10.1080/02664763.2011.626850
http://www.tandfonline.com

point in the plane, the profiles of wind and wave directions are points in a torus, that is, a surface generated by revolving a circle in a three-dimensional space.

Most of the literature on mixture-based classification methods is associated with the analysis of multivariate data whose components share the same support. Linear observations are typically clustered by mixtures of multivariate normal distributions [2], although mixtures of multivariate skew normal [13] and t distributions [15], or, more generally, non-elliptically contoured distributions [9], have been recently proposed for robust classification. Multivariate categorical observations are instead typically clustered by using latent-class models that involve mixtures of multinomial distributions [6]. In directional statistics, while mixtures of Kent distributions are popular in the analysis of spherical data [19], toroidal data that arise in bioinformatics have been recently modeled by mixtures of bivariate circular densities [18].

Unsupervised classification of multivariate data of mixed type has been studied only in the case of mixed linear and categorical data [7,12]. We extend this strand of literature by taking a latent-class approach to cluster mixed linear and circular data. Latent-class models approximate the joint distribution of the data by a mixture of products of low-dimensional densities, by assuming that the groups of observed variables are conditionally independent given a latent class, drawn from an unobserved multinomial random variable (conditional independence assumption).

In the modeling of mixed-type multivariate data, a latent approach has a number of advantages. First, latent classes non-parametrically capture part of the data dependence structure, which is difficult to describe with a fully parametric specification. In marine studies, the dependence between circular and linear measurements is the result of complex environmental conditions. On the one side, latent classes can be then used to capture the association between toroidal clusters of wave and wind directions and planar clusters of wind speed and wave height. On the other side, we take a fully parametric approach to detect locations and shapes of both toroidal clusters, which are modeled directly by bivariate circular densities, and planar clusters, which are modeled by bivariate skew normal densities.

Secondly, the conditional independence assumption facilitates maximum-likelihood estimation from mixed multivariate data. Maximum-likelihood estimation of mixture models is often based on the expectation maximization (EM) algorithms, which iteratively estimate the expected class membership and simultaneously update the parameters of the mixture components. Under a conditional independence assumption, an EM procedure for classifying mixed-type data can be easily obtained by combining EM algorithms that have been developed for data with homogeneous supports.

Thirdly, the identifiability of mixtures of product densities can be easily addressed. In general, identifiability issues may arise when variables on different supports are mixed together. Direct modeling of the joint distribution of multivariate linear–circular data would require the specification of densities that lie on a multi-dimensional hyper-cylinder [10], and identifiability conditions for mixtures of densities of this type have recently appeared in the literature and have not been studied, yet. By taking a latent-class approach, on the contrary, the joint distribution of hyper-cylindrical data is approximated by the mixture of products of toroidal and planar densities, and a sufficient condition for the identifiability of mixtures of product densities is the linear independence of the mixture components [25,26]. The identifiability of the model that we propose then follows from the linear independence of the bivariate circular densities [17] and the linear independence of the bivariate skew normal densities [22].

An additional complication in marine classification studies is the presence of missing values. Marine databases are often incomplete because of device malfunctioning or maintenance-related reasons. In the case of incomplete data, maximum-likelihood estimation of a mixture model could be carried out by discarding the incomplete profiles from the sample and using the complete cases (CCs) to build up the likelihood function to be maximized (CC analysis). If the joint distribution of the variables of interest is correctly specified and the data are missing at random (MAR; i.e.

the conditional probability of not observing a value, given the observed data, does not depend on the unobserved value [21]), the CC-based maximum-likelihood estimation is known to be (asymptotically) unbiased but inefficient [20]. Loss of efficiency is due to the fact that incomplete data profiles are informative of the parameters of the joint distribution of several variables. When data are MAR, mixture models can be estimated by EM algorithms that account for missing class membership and missing measurements as different sources of incomplete information. Efficient algorithms of this type are well known for the unsupervised classification of incomplete normal and incomplete categorical data [23] and have been extended to mixture-based classification studies for clustering incomplete skew normal or *t* distributed continuous data [14,16] and mixed continuous and categorical data [7]. The EM algorithm that we propose in this paper is based on an extension of the EM algorithms for mixtures of bivariate circular densities [18] to the case of incomplete data, which is then combined with EM iterations that have been developed for the estimation of mixtures of multivariate skew distributions from incomplete data [16].

After summarizing relevant details on the data that motivated this study (Section 2), the latent-class model that we propose for clustering mixed linear and circular data is illustrated in Section 3. Likelihood-based inference from incomplete data is presented in Section 4, while Section 5 illustrates an application to marine data. Relevant points of discussion are finally summarized in Section 6.

2. Data

The Adriatic Sea (Figure 1) is a semi-enclosed, long narrow basin, extending for about 800 km along the major axis from SE to NW, with a width of about 200 km. The basin is also bordered by mountains on three sides. Relevant wind events in the Adriatic Sea are typically generated by the sirocco wind, which blows from SE along the major basin axis, and by the bora flow, which creates fine-structured jets within the Dinaric Alps on the eastern Adriatic coast. These jets typically cross the Adriatic Sea along the NE–SW minor axis of the basin, but sometimes they rotate anticlockwise toward SE as soon as they approach the topographic barrier of the Apennines. High-speed winds generate high waves only when they persistently blow from directions that are highly concentrated around one modal angle. As a result, when the above rotation episodes occur, offshore winds blow from multi-modal directions and generate waves of modest size. Wind–wave data are traditionally examined by exploiting numerical wind–wave models. These models, well suited for the analysis of ocean waves, are not flexible enough to account for the complex orography of semi-enclosed basins and, as a result, give biased results in Adriatic studies [3]. When numerical wind–wave

Figure 1. Locations of the buoy (circle) and tide gauge (square) at Ancona.

models are problematic, sea conditions can be alternatively described in terms of representative wave regimes in specific areas, characterized by the probability of occurrence and corresponding to dominant environmental conditions (e.g. wind conditions), acting in the area and during a period of interest [11]. The data normally exploited for this purpose are environmental observations taken by buoys or tide gauges, located within the study area.

The data that motivated this paper are hourly, quadrivariate profiles with two linear and two circular components: wind speed and wave height, wind direction and wave direction. Hourly wave height and direction were taken in the period 18 November 2002–17 January 2003 by the buoy of Ancona, which is located in the Adriatic Sea at about 30 km from the coast (Figure 1). Hourly wind speed and direction were obtained from the nearest tide gauge, located at Ancona. To account for the cumulative effect that wind has on waves, wind data were smoothed by taking, for each hour, the average of wind speeds and the circular average of wind directions, observed during the last 8 h.

Of the resulting 1440 hourly profiles of wind and wave observations, about 20% include at least a missing value (Table 1). As expected, missing values on wave measurements are more frequent than missing wind data because buoys are more exposed to transmission errors than tide gauges. During the study period, only two are the profiles with no information.

In this paper, we assume that missing values occur at random. Under this hypothesis, the contribution of missing patterns to the likelihood can be ignored, facilitating model-based clustering of the data. In marine studies, missing values occur because of device transmission errors or malfunctioning. Because buoys and tide gauges are normally equipped in a way that they are able to transmit data even in the case of severe environmental conditions, missing values in marine studies are often missing completely at random (MCAR), that is, the missingness probability does not depend on observed and unobserved data. The MCAR assumption is a particular case of the MAR hypothesis and is often likely for marine data that are obtained in semi-enclosed seas, such as the Adriatic Sea, where severe environmental conditions seldom occur. The MAR assumption is violated when the conditional probability of device malfunctioning, given the observed data, depends on the value that the device has not transmitted, and in this case, the missing mechanism may not be ignored. For example, high-speed wind and high waves might increase the probability of a buoy transmission error, leading to a non-ignorable missing value when both are missing. We, however, have only six cases (Table 1) where both wind speed and wave height are missing. In all the other cases, when either wind speed or wave height is observed, the MAR assumption seems to be reasonable.

Figure 2 displays the scatter plots of the circular and the linear observations, after discarding the incomplete profiles. For simplicity, bivariate circular data are plotted on the plane, although data points are actually in a torus. In particular, point coordinates on the left-hand-side plot of the figure indicate hourly directions from which the wind blows and the wave travels, respectively.

Table 1. Missing value distribution.

Wind speed	Wind direction	Wave direction	Wave height	Count
obs	obs	obs	obs	1173
obs	mis	obs	obs	21
mis	obs	obs	obs	6
obs	obs	mis	mis	227
mis	mis	obs	obs	6
obs	mis	mis	mis	3
mis	obs	mis	mis	2
mis	mis	mis	mis	2

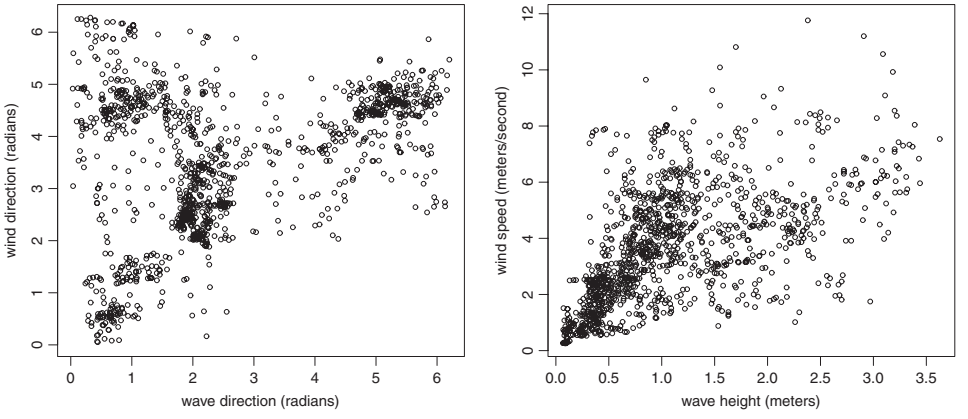

Figure 2. Complete profiles of wind and wave direction (left) and complete profiles of wind speed and wave height (right).

The interpretation of these data is complicated by the complex orography of the Adriatic Sea and by the different locations (tide gauge and buoy) where wind and wave data are observed.

A number of clusters appear in the directional scatter plot on the left-hand side of Figure 2. Points around the $3/4(\pi, \pi)$ centroid indicate sirocco events (waves travel along the major axis of the basin, driven by a southeasterly wind), whereas points centered around the centroid $(\pi/4, \pi/4)$ can be interpreted as bora episodes (waves travel along the minor axis of the basin, driven by a northwesterly wind). The remaining two clusters at the top of the scatter plot can be interpreted by recalling that the buoy and the tide gauge are located about 30 km apart. These points are bora episodes where some NE jets rotate anticlockwise and blow from NW. As a result, on the one side, the buoy detects waves that travel northeasterly and northwesterly, either driven by the offshore bora winds that blow from the east side of the basin or driven by bora winds that rotate along the major axis of the basin. On the other side, offshore northeasterly winds are not observed at the coast, where the tide gauge is located.

The right-hand-side plot shown in Figure 2 shows that wind speed and wave height are (marginally) skewed and weakly correlated. Both skewness and weak correlation are traditionally explained as the result of the orography of the Adriatic Sea and they are often held responsible for the inaccuracy of numerical wind–wave models. It is, however, possible that the marginal skewness and weak correlation can be explained, at least in part, as a result of latent data heterogeneity. What we observe, in other words, could be the result of the mixing of a number of latent regimes of the sea, conditionally to which the distribution of the data takes a shape that is easier to interpret than the shape taken by the marginal distribution. By taking a latent-class approach, we try to identify these latent regimes by associating toroidal and planar clusters that provide an intuitively appealing partitioning of the two scatter plots shown in Figure 2 and, when mixed together, adequately approximate the marginal distribution of the data.

3. A latent-class model for linear and circular data

The data described in Section 2 are gathered in the form of n profiles $\boldsymbol{z}_i = (\boldsymbol{x}_i, \boldsymbol{y}_i), i = 1 \ldots n$, which include two circular components, say $\boldsymbol{x}_i = (x_{i1}, x_{i2})$, and two linear components, say $\boldsymbol{y}_i = (y_{i1}, y_{i2})$. We model these data by exploiting the mixture

$$f(\boldsymbol{z}|\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{k=1}^{K} \pi_k f_c(\boldsymbol{x}|\boldsymbol{\beta}_k) f_l(\boldsymbol{y}|\boldsymbol{\gamma}_k), \tag{1}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ are the unknown mixing weights, $\pi_1 + \cdots + \pi_K = 1$, while $f_c(\boldsymbol{x}|\boldsymbol{\beta}_k)$ and $f_l(\boldsymbol{y}|\boldsymbol{\gamma}_k)$ are the bivariate densities, respectively, defined on the torus and on the plane, and known up to two independent vectors of parameters, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K)$. In mixture-based classification studies, mixing weights can be conveniently interpreted as the cell probabilities of a latent multinomial vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)$. As a result, the above mixture can be described as a two-level hierarchical model

$$\boldsymbol{\xi} \sim \prod_{k=1}^{K} \pi_k^{\xi_k}$$

$$\boldsymbol{z}|\boldsymbol{\xi} \sim \prod_{k=1}^{K} (f_c(\boldsymbol{x}|\boldsymbol{\beta}_k) f_l(\boldsymbol{y}|\boldsymbol{\gamma}_k))^{\xi_k}.$$

At the upper level of the hierarchy, directions (e.g. wind and wave directions) and intensities (e.g. wind speed and wave height) are modeled separately by parametric distributions. These distributions are then non-parametrically associated to K latent classes at the lower level of the hierarchy. This hierarchy allows to transform the data clustering problem into a missing value problem, where missing class membership $\boldsymbol{\xi}_i$ of each profile can be predicted by its expectation $\mathbb{E}(\boldsymbol{\xi}_i|\boldsymbol{z}_i)$, whose kth component is given by

$$\pi_{ik} = \mathbb{E}(\xi_{ik}|\boldsymbol{z}_i) = \frac{\pi_k f_c(\boldsymbol{x}_i|\boldsymbol{\beta}_k) f_l(\boldsymbol{y}_i|\boldsymbol{\gamma}_k)}{\sum_{k=1}^{K} \pi_k f_c(\boldsymbol{x}_i|\boldsymbol{\beta}_k) f_l(\boldsymbol{y}_i|\boldsymbol{\gamma}_k)}. \tag{2}$$

The distribution $f_c(\boldsymbol{x}|\boldsymbol{\beta})$ of the bivariate circular data can be specified in a number of different ways [18]. The sine model [24] is a parametric distribution on the torus which imbeds naturally the bivariate normal distribution when the range of observations is small. Its density is given by

$$f_c(\boldsymbol{x}; \boldsymbol{\beta}) = \frac{\exp(\beta_{11}\cos(x_1 - \beta_1) + \beta_{22}\cos(x_2 - \beta_2) + \beta_{12}\sin(x_1 - \beta_1)\sin(x_2 - \beta_2))}{C(\boldsymbol{\beta})}, \tag{3}$$

with normalizing constant

$$C(\boldsymbol{\beta}) = 4\pi^2 \sum_{m=0}^{\infty} \binom{2m}{m} \left(\frac{\beta_{12}^2}{4\beta_{11}\beta_{22}}\right)^m I_m(\beta_{11}) I_m(\beta_{22}),$$

where

$$I_m(x) = \frac{1}{\pi} \int_0^{\pi} \mathrm{e}^{x\cos t} \cos(mt)\, \mathrm{d}t$$

is the modified Bessel function of order m.

The sine model can be viewed as a bivariate generalization of the von Mises distribution, where β_{12} accounts for the statistical dependence between x_1 and x_2. The two univariate marginal densities

$$f_c(x_i; \boldsymbol{\beta}) = \int_{-\pi}^{\pi} f_c(\boldsymbol{x}; \boldsymbol{\beta})\, \mathrm{d}x_j = \frac{2\pi}{C(\boldsymbol{\beta})} I_0(a(x_i)) \exp(\beta_{ii}\cos(x_i - \beta_i)), \quad i = 1, 2, \tag{4}$$

depend on the marginal mean angles β_i, $i = 1, 2$, and on the shape parameters

$$a(x_i) = (\beta_{jj}^2 + \beta_{12}^2 \sin^2(x_i - \beta_i))^{1/2}, \quad i = 1, 2. \tag{5}$$

If $\beta_{12} = 0$, then $a(x_i) = \beta_{jj}$, $i = 1, 2$, and, as a result, x_1 and x_2 are independent and each of them assumes the von Mises distribution with marginal mean angles β_i and marginal concentrations

β_{ii}. The conditional distributions

$$f_c(x_i|x_j;\boldsymbol{\beta}) = \frac{f_c(\boldsymbol{x};\boldsymbol{\beta})}{f_c(x_j;\boldsymbol{\beta})} = \frac{\exp(a(x_i)\cos(x_i-\beta_i-b(x_j)))}{2\pi I_0(a(x_i))} \tag{6}$$

are von Mises with conditional mean angles $\beta_i + b(x_j)$ and conditional concentrations $a(x_i)$, where

$$b(x_j) = \arctan\left(\frac{\beta_{12}}{\beta_{jj}}\sin(x_j-\beta_j)\right). \tag{7}$$

In model (1), we use a family of K sine models $f_c(\boldsymbol{x}|\boldsymbol{\beta}_k)$, indexed by the five parameters $\boldsymbol{\beta}_k = (\beta_{1k}, \beta_{2k}, \beta_{11k}, \beta_{22k}, \beta_{12k})$, to define K toroidal clusters centered at (β_{1k}, β_{2k}) and shaped by the parameters $(\beta_{11k}, \beta_{22k}, \beta_{12k})$.

To model the joint distribution of wind speed and wave height, we use seven parameters, arranged in a triplet:

$$\boldsymbol{\gamma} = (\boldsymbol{\gamma}', \Gamma, \mathbf{D}(\boldsymbol{\gamma}'')) = \left(\begin{pmatrix}\gamma_1'\\ \gamma_2'\end{pmatrix}, \begin{pmatrix}\gamma_{11} & \gamma_{12}\\ \gamma_{12} & \gamma_{22}\end{pmatrix}, \begin{pmatrix}\gamma_1'' & 0\\ 0 & \gamma_2''\end{pmatrix}\right),$$

where $\boldsymbol{\gamma}'$ is a location vector, Γ is a positive definite, scale covariance matrix and, finally, $\mathbf{D}(\boldsymbol{\gamma}'')$ is a diagonal matrix that includes two skewness parameters. These parameters are exploited to specify a bivariate skew normal density [22], namely

$$f_l(\boldsymbol{y};\boldsymbol{\gamma}) = 2^2\phi_2(\boldsymbol{y};\boldsymbol{\gamma}', \Gamma+\mathbf{D}^2(\boldsymbol{\gamma}''))\Phi_2(\mathbf{D}(\boldsymbol{\gamma}'')(\Gamma+\mathbf{D}^2(\boldsymbol{\gamma}''))^{-1}(\boldsymbol{y}-\boldsymbol{\gamma}'); (\mathbf{I}+\mathbf{D}(\boldsymbol{\gamma}'')\Gamma^{-1}\mathbf{D}(\boldsymbol{\gamma}''))^{-1}), \tag{8}$$

where $\phi_p(\cdot;\boldsymbol{\mu},\Sigma)$ indicates the density of a p-variate normal distribution $N_p(\mu,\Sigma)$ and $\Phi_p(\cdot;\Sigma)$ indicates the cdf of a centered, p-variate normal distribution $N_p(\mathbf{0},\Sigma)$. Under Equation (8), the mean vector and the covariance matrix of $\boldsymbol{y}$ are, respectively, given by

$$\mathbb{E}\boldsymbol{y} = \boldsymbol{\gamma}' + \sqrt{\frac{2}{\pi}}\mathbf{D}(\boldsymbol{\gamma}'')\mathbf{1}, \quad \mathbb{E}(\boldsymbol{y}-\mathbb{E}\boldsymbol{y})(\boldsymbol{y}-\mathbb{E}\boldsymbol{y})^{\mathsf{T}} = \Gamma + \left(1-\frac{2}{\pi}\right)\mathbf{D}^2(\boldsymbol{\gamma}''),$$

where $\mathbf{1}$ is a vector of ones. When the skewness parameters $\gamma_1'' = \gamma_2'' = 0$, Equation (8) reduces to a bivariate normal distribution $N_2(\boldsymbol{\gamma}', \Gamma)$. Moreover, the marginal distribution of y_i is a univariate skew normal distribution with parameters $(\gamma_i', \gamma_{ii}, \gamma_i'')$, say

$$f_l(y_i;\boldsymbol{\gamma}) = 2\phi_1(y_i;\gamma_i', \gamma_{ii}+(\gamma_i'')^2)\Phi_1\left(\frac{\gamma_i''}{\gamma_{ii}+(\gamma_i'')^2}(y_i-\gamma_i'); \frac{\gamma_{ii}}{\gamma_{ii}+(\gamma_i'')^2}\right), \tag{9}$$

while the conditional distribution of y_j given y_i is given by

$$\begin{aligned} f_l(y_j|y_i;\boldsymbol{\gamma}) &= \frac{f_l(\boldsymbol{y};\boldsymbol{\gamma})}{f_l(y_i;\boldsymbol{\gamma})} \\ &= 4\phi_1\left(y_j;\gamma_j' + \frac{\gamma_{ij}}{\gamma_{ii}+(\gamma_i'')^2}(y_i-\gamma_i'), \frac{\gamma_{ij}^2}{\gamma_{ii}+(\gamma_i'')^2}\right) \\ &\quad \times \frac{\Phi_2(\mathbf{D}(\boldsymbol{\gamma}'')(\Gamma+\mathbf{D}(\boldsymbol{\gamma}'')^2)^{-1}(\boldsymbol{y}-\boldsymbol{\gamma}'); (\mathbf{I}+\mathbf{D}(\boldsymbol{\gamma}'')\Gamma^{-1}\mathbf{D}(\boldsymbol{\gamma}''))^{-1})}{\Phi_1(\gamma_i''/(\gamma_{ii}+(\gamma_i'')^2)(y_i-\gamma_i'); \gamma_{ii}/(\gamma_{ii}+(\gamma_i'')^2))}. \end{aligned} \tag{10}$$

A bivariate skew normal density can be conveniently represented [1] as the convolution

$$f_l(\boldsymbol{y};\boldsymbol{\gamma}) = \int_0^{+\infty}\int_0^{+\infty} f_l(\boldsymbol{y}|\boldsymbol{v};\boldsymbol{\gamma}) f_{\mathrm{HN}}(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v},$$

where $f_{\mathrm{HN}}(\boldsymbol{v})$ is a standard half-normal distribution:

$$f_{\mathrm{HN}}(\boldsymbol{v}) = \frac{2}{\pi}\exp\left(-\frac{1}{2}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{v}\right) \quad \boldsymbol{v}\in[0,+\infty)^2,$$

while $f_l(\boldsymbol{y}|\boldsymbol{v}) = \phi_2(\boldsymbol{y};\boldsymbol{\gamma}' + \mathbf{D}(\boldsymbol{\gamma}'')\boldsymbol{v},\Gamma)$. This random-effect specification of the multivariate skew distribution facilitates the implementation of EM algorithms for the maximum-likelihood estimation in the mixtures of multivariate skew normal distributions [13].

In model (1), we use a family of K skew normal densities $f_l(\boldsymbol{y}|\boldsymbol{\gamma}_k)$, indexed by the seven parameters included in the vector $\boldsymbol{\gamma}_k = (\boldsymbol{\gamma}'_k, \Gamma_k, \mathbf{D}(\boldsymbol{\gamma}''_k))$, to define K skew clusters centered at $\boldsymbol{\gamma}'_k + \sqrt{2/\pi}\mathbf{D}(\boldsymbol{\gamma}''_k)\mathbf{1}$ and shaped by the covariance matrices $\Gamma_k + (1-2/\pi)\mathbf{D}(\boldsymbol{\gamma}''_k)^2$.

4. Maximum-likelihood estimation from incomplete data

Because our data are in the form of incomplete profiles, we, respectively, refer to $\boldsymbol{x}_{i,\mathrm{mis}}$ and $\boldsymbol{x}_{i,\mathrm{obs}}$ as the missing and observed circular components of profile i and, analogously, to $\boldsymbol{y}_{i,\mathrm{mis}}$ and $\boldsymbol{y}_{i,\mathrm{obs}}$ as the missing and observed linear components. Accordingly, $\boldsymbol{z}_{i,\mathrm{mis}} = (\boldsymbol{x}_{i,\mathrm{mis}}, \boldsymbol{y}_{i,\mathrm{mis}})$ and $\boldsymbol{z}_{i,\mathrm{obs}} = (\boldsymbol{x}_{i,\mathrm{obs}}, \boldsymbol{y}_{i,\mathrm{obs}})$ indicate the missing and observed parts of the ith profile. We further introduce a vector $\boldsymbol{r}_i = (r_{i1}, r_{i2}, r_{i3}, r_{i4})$ of binary missing indicators, where $r_{ij} = 1$ if z_{ij} is missing and 0 otherwise. If the data are MAR, the missing data mechanism can be ignored and the maximum-likelihood estimate of parameter $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is the maximum point of the marginal log-likelihood function

$$\begin{aligned}
\log L(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log\left(\int \sum_{k=1}^{K} \pi_k f_c(\boldsymbol{x}_i|\boldsymbol{\beta}_k) f_l(\boldsymbol{y}_i|\boldsymbol{\gamma}_k)\,\mathrm{d}\boldsymbol{z}_{i,\mathrm{mis}}\right)\\
&= \sum_{i=1}^{n} \log L_i(\boldsymbol{\theta})\\
&= \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k L_{ic}(\boldsymbol{\beta}_k) L_{il}(\boldsymbol{\gamma}_k),
\end{aligned} \tag{11}$$

where $L_i(\boldsymbol{\theta})$ is the likelihood contribution of the ith profile and

$$\begin{aligned}
L_{ic}(\boldsymbol{\beta}_k) &= f_c(\boldsymbol{x}_i;\boldsymbol{\beta}_k)^{(1-r_{i1})(1-r_{i2})} f_c(x_{i1};\boldsymbol{\beta}_k)^{(1-r_{i1})r_{i2}} f_c(x_{i2};\boldsymbol{\beta}_k)^{r_{i1}(1-r_{i2})},\\
L_{il}(\boldsymbol{\gamma}_k) &= f_l(\boldsymbol{y}_i;\boldsymbol{\gamma}_k)^{(1-r_{i3})(1-r_{i4})} f_l(y_{i1};\boldsymbol{\gamma}_k)^{(1-r_{i3})r_{i4}} f_l(y_{i2};\boldsymbol{\gamma}_k)^{r_{i3}(1-r_{i4})}
\end{aligned}$$

are the conditional likelihood contributions of the circular and linear components of the ith profile, given the latent class k.

Because direct maximization of (11) can be computationally problematic, we describe an EM algorithm that generates a sequence $(\hat{\boldsymbol{\theta}}_t, t = 1, 2, \ldots)$ of estimates such that $L(\hat{\boldsymbol{\theta}}_t) \geq L(\hat{\boldsymbol{\theta}}_{t-1})$. The algorithm is based on the iterative maximization of the expected value of a complete-data log-likelihood function, computed with respect to the conditional distribution of the unobserved quantities given the observed data. More precisely, we treat the unknown class membership $\boldsymbol{\xi}_i$,

the unobserved data $(\boldsymbol{x}_{i,\text{mis}}, \boldsymbol{y}_{i,\text{mis}})$ and the skewness random effects $\boldsymbol{v}_i$ as missing values and define the complete log-likelihood function as

$$\log L_{\text{comp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log L_{i,\text{comp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \xi_{ik} \begin{pmatrix} \log \pi_k \\ \log f_c(\boldsymbol{x}_i; \boldsymbol{\beta}_k) \\ \log f_l(\boldsymbol{y}_i | \boldsymbol{v}_i; \boldsymbol{\gamma}_k) + \log f_{\text{HN}}(\boldsymbol{v}_i) \end{pmatrix}.$$

Given the estimate $\hat{\boldsymbol{\theta}}_t$, provided by the algorithm at step t, a new point $\hat{\boldsymbol{\theta}}_{t+1}$ is computed within step $t+1$, as follows. We first compute (E step) the expected value of $\log L_{i,\text{comp}}(\boldsymbol{\theta})$ with respect to the conditional distribution of the missing values $(\boldsymbol{\xi}_i, \boldsymbol{x}_{i,\text{mis}}, \boldsymbol{y}_{i,\text{mis}}, \boldsymbol{v}_i)$ given the observed data $\boldsymbol{y}_{i,\text{obs}}$, evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_t$, say

$$\text{(Estep)} \quad Q_i(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_t) = \mathbb{E}_t(\log L_{i,\text{comp}}(\boldsymbol{\theta}) | \boldsymbol{y}_{i,\text{obs}}), \quad i = 1, \ldots, n. \tag{12}$$

We then (M step) maximize $Q(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_t) = \sum_{i=1}^{n} Q_i(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_t)$ by finding the roots $\hat{\boldsymbol{\theta}}_{t+1}$ of the expected complete data score equations:

$$\text{(Mstep)} \quad \frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_t) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} Q_i(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_t) = \sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_t) = \boldsymbol{0}, \tag{13}$$

where $\boldsymbol{s}_i(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_t)$ is the ith score vector, obtained by deriving the ith contribution to the expected complete log-likelihood with respect to the parameters.

Variances of the estimates can be found on the diagonal of the inverse of the information matrix $\mathbf{I}(\boldsymbol{\theta})$, which can be consistently estimated by the empirical information matrix:

$$\hat{\mathbf{I}} = \sum_{i=1}^{n} \boldsymbol{s}_i(\hat{\boldsymbol{\theta}}_T) \boldsymbol{s}_i^{\mathsf{T}}(\hat{\boldsymbol{\theta}}_T),$$

where $\hat{\boldsymbol{\theta}}_T$ is the last parameter update, as provided by the algorithm upon convergence.

The practical implementation of both the E step and the M step of the algorithm is facilitated by the conditional independence assumption between circular and linear data, which holds under (1). For the purpose of illustration, we observe that the distribution of the missing values given the observed data can be factorized into three components, as follows:

$$\begin{aligned} f(\boldsymbol{v}_i, \boldsymbol{z}_{i,\text{mis}}, \boldsymbol{\xi}_i | \boldsymbol{z}_{i,\text{obs}}; \hat{\boldsymbol{\theta}}_t) &= \prod_{k=1}^{K} \left(\frac{\hat{\pi}_{tk} f_c(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}_{tk}) f_l(\boldsymbol{y}_i | \boldsymbol{v}_i; \hat{\boldsymbol{\gamma}}_{tk}) f_l(\boldsymbol{v}_i)}{L_i(\hat{\boldsymbol{\theta}}_t)} \right)^{\xi_{ik}} \\ &= \prod_{k=1}^{K} \left(\frac{\hat{\pi}_{tk} f_c(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}_{tk}) f_l(\boldsymbol{y}_i | \boldsymbol{v}_i, \hat{\boldsymbol{\gamma}}_{tk}) f_l(\boldsymbol{v}_i)}{\hat{\pi}_{tk} L_{ic}(\hat{\boldsymbol{\beta}}_{tk}) L_{il}(\hat{\boldsymbol{\gamma}}_{tk})} \frac{\hat{\pi}_{tk} L_{ic}(\hat{\boldsymbol{\beta}}_{tk}) L_{il}(\hat{\boldsymbol{\gamma}}_{tk})}{L_i(\hat{\boldsymbol{\theta}}_t)} \right)^{\xi_{ik}} \\ &= f(\boldsymbol{x}_{i,\text{mis}} | \boldsymbol{\xi}_i, \boldsymbol{x}_{i,\text{obs}}; \hat{\boldsymbol{\beta}}_t) f(\boldsymbol{y}_{i,\text{mis}} | \boldsymbol{\xi}_i, \boldsymbol{y}_{i,\text{obs}}; \hat{\boldsymbol{\gamma}}_t) p(\boldsymbol{\xi}_i | \boldsymbol{z}_{i,\text{obs}}; \hat{\boldsymbol{\theta}}_t), \end{aligned} \tag{14}$$

where

- the conditional density

$$f(\boldsymbol{x}_{i,\text{mis}} | \xi_{ik} = 1, \boldsymbol{x}_{i,\text{obs}}; \hat{\boldsymbol{\beta}}_t) = \frac{f_c(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}_{tk})}{L_{ic}(\hat{\boldsymbol{\beta}}_{tk})} \tag{15}$$

  is identically 1 if $r_{i1} = r_{i2} = 0$, it reduces to the conditional univariate von Mises densities (6) with $\boldsymbol{\beta} = \boldsymbol{\beta}_{tk}$ if either $(r_{i1}, r_{i2}) = (0, 1)$ or $(1, 0)$ and it is finally equal to the bivariate circular density (3) with $\boldsymbol{\beta} = \boldsymbol{\beta}_{tk}$, if $(r_{i1}, r_{i2}) = (1, 1)$;

- the conditional density

$$f(\boldsymbol{y}_{i,\text{mis}}|\xi_{ik}=1,\boldsymbol{y}_{i,\text{obs}};\hat{\boldsymbol{\gamma}}_t)=\frac{f_l(\boldsymbol{y}_i;\hat{\boldsymbol{\gamma}}_{tk})}{L_{il}(\hat{\boldsymbol{\gamma}}_{tk})} \tag{16}$$

  is identically 1 if $r_{i3}=r_{i4}=0$, it reduces to the conditional univariate skew normal densities (10) with $\boldsymbol{\gamma}=\boldsymbol{\gamma}_{tk}$ if either $(r_{i3},r_{i4})=(0,1)$ or $(1,0)$ and it is equal to the bivariate skew normal density (8) with $\boldsymbol{\gamma}=\boldsymbol{\gamma}_{tk}$ if $(r_{i3},r_{i4})=(1,1)$
- and, finally,

$$\hat{\pi}_{tik}=P(\xi_{ik=1}|\boldsymbol{z}_{i,\text{obs}};\hat{\boldsymbol{\theta}}_t)=\frac{\hat{\pi}_{tk}L_{ic}(\hat{\boldsymbol{\beta}}_{tk})L_{il}(\hat{\boldsymbol{\gamma}}_{tk})}{L_i(\hat{\boldsymbol{\theta}}_t)} \tag{17}$$

  are the conditional cell probabilities of the multinomial class membership vector, given the observed data; when profile $\boldsymbol{z}_i$ is fully observed, these probabilities reduce to (2), evaluated at $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_t$.

Upon convergence of the algorithm ($t=T$), if desired, distributions (15)–(16) can be exploited to impute the missing circular and linear observations, respectively. Probabilities (17) can be instead exploited to cluster incomplete profiles into K groups by modal allocation, that is, assigning each profile i to the latent class with the highest probability $\hat{\pi}_{Tik}$.

Given the factorization (14), the expected value of the complete log-likelihood function with respect to the conditional distribution of the missing values given the observed data is (at the $(t+1)$th step of the algorithm) given by

$$\begin{aligned}Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_t)&=\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{\pi}_{tik}\begin{pmatrix}\log\pi_k\\ \mathbb{E}_t(\log f_c(\boldsymbol{x}_i;\boldsymbol{\beta}_k)|\boldsymbol{x}_{i,\text{obs}},\xi_{ik}=1)\\ \mathbb{E}_t(\log f_l(\boldsymbol{y}_i;\boldsymbol{\gamma}_k)|\boldsymbol{y}_{i,\text{obs}},\xi_{ik}=1)\end{pmatrix}\\ &=\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{\pi}_{tik}\begin{pmatrix}\log\pi_k\\ Q_{ic}(\boldsymbol{\beta}_k|\hat{\boldsymbol{\beta}}_{tk})\\ Q_{il}(\boldsymbol{\gamma}_k|\hat{\boldsymbol{\gamma}}_{tk})\end{pmatrix},\end{aligned}$$

where $Q_{ic}(\boldsymbol{\beta}_k|\boldsymbol{\beta}_{tk})=\mathbb{E}_t(\log f_c(\boldsymbol{x}_i;\boldsymbol{\beta}_k)|\boldsymbol{x}_{i,\text{obs}},\xi_{ik}=1)$ indicates the expected value of $\log f_c(\boldsymbol{x}_i;\boldsymbol{\beta}_k)$ with respect to (15) and $Q_{il}(\boldsymbol{\gamma}_k|\hat{\boldsymbol{\gamma}}_{tk})=\mathbb{E}_t(\log f_l(\boldsymbol{y}_i;\boldsymbol{\gamma}_k)|\boldsymbol{y}_{i,\text{obs}},\xi_{ik}=1)$ indicates the expected value of $\log f_l(\boldsymbol{y}_i;\boldsymbol{\gamma}_k)$ with respect to (16). Therefore, both the E step and the M step of the algorithm essentially reduce to the evaluation of three updating functions, namely

$$\begin{aligned}Q_1(\boldsymbol{\pi})&=\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{\pi}_{tik}\log\pi_k,\\ Q_2(\boldsymbol{\beta})&=\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{\pi}_{tik}Q_{ic}(\boldsymbol{\beta}_k|\hat{\boldsymbol{\beta}}_{tk}),\\ Q_3(\boldsymbol{\gamma})&=\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{\pi}_{tik}Q_{il}(\boldsymbol{\gamma}_k|\hat{\boldsymbol{\gamma}}_{tk}),\end{aligned}$$

which can then be maximized separately within the M step. Function Q_1 is maximized by solving the $K-1$ score equations

$$\frac{\partial}{\partial\pi_k}Q_1(\boldsymbol{\pi})=\sum_{i=1}^{n}\frac{\pi_k}{\hat{\pi}_{tik}}-\frac{\pi_K}{\hat{\pi}_{tiK}}=0,\quad k=1,\ldots,K-1,$$

which have the following closed-form roots:

$$\hat{\pi}_{t+1,k} = \frac{\sum_{i=1}^{n} \hat{\pi}_{tik}}{n}, \quad k = 1, \ldots, K.$$

Function Q_2 is maximized by separately solving K systems of the score equations

$$\sum_{i=1}^{n} \hat{\pi}_{tik} \frac{\partial}{\partial \boldsymbol{\beta}_k} Q_{ic}(\boldsymbol{\beta}_k | \hat{\boldsymbol{\beta}}_{tk}) = \mathbf{0}, \quad k = 1, \ldots, K.$$

In the appendix, we derive the analytical form taken by the expectations $Q_{ic}(\boldsymbol{\beta}_k | \hat{\boldsymbol{\beta}}_{tk})$ and display a computationally tractable form of the score equations to update the circular parameters $\boldsymbol{\beta}$. Finally, function Q_3 can be maximized by separately solving K systems of the score equations

$$\sum_{i=1}^{n} \hat{\pi}_{tik} \frac{\partial}{\partial \boldsymbol{\gamma}_k} Q_{il}(\boldsymbol{\gamma}_k | \hat{\boldsymbol{\gamma}}_{tk}) = \mathbf{0}, \quad k = 1, \ldots, K,$$

according to the expressions derived in [16] for the unsupervised classification of incomplete, multivariate skew normal data.

The EM algorithm can get stuck in the local maxima of the log-likelihood function or can be attracted by singularities at the edge of the parameter space, where the log-likelihood is unbounded [27]. The presence of multiple local and spurious maxima is well documented in the case of mixtures of heteroscedastic normal distributions [19] and less widely known in the case of bivariate circular distributions [18]. A number of strategies have been proposed to select a local maximizer and detect a spurious maximizer. To avoid local maxima, we follow a short-run strategy (known as the emEM algorithm [5]), by running the EM algorithm from a number of random initializations, stopping at iteration t as soon as

$$\frac{\log L(\hat{\boldsymbol{\theta}}_t) - \log L(\hat{\boldsymbol{\theta}}_{t-1})}{\log L(\hat{\boldsymbol{\theta}}_t) - \log L(\hat{\boldsymbol{\theta}}_0)} \leq \eta.$$

We have observed that convergence to spurious maxima is fast (a phenomenon that is well known in the case of mixtures of multivariate normal densities [8]) and can be detected within short EM runs, by monitoring both the class proportions $\hat{\pi}_{tk}$ and the eigenvalues of the covariance matrices

$$\begin{pmatrix} \hat{\beta}_{t11k} & \hat{\beta}_{t12k} \\ \hat{\beta}_{t12k} & \hat{\beta}_{t22k} \end{pmatrix}^{-1} \quad \left(\hat{\Gamma}_{tk} + \left(1 - \frac{2}{\pi} \right) \mathbf{D}(\hat{\boldsymbol{\gamma}}''_{tk}) \right).$$

After excluding spurious solutions, we select the output of the EM short run that maximizes the log-likelihood, which is then used to initialize a long run of the EM algorithm.

5. Results

We have estimated a number of mixture models from the data given in Section 2, by varying the number of components from two to five. The computer code is available from the corresponding author upon request. EM short runs were stopped by using a threshold $\eta = 10^{-3}$, typically reached between 50 and 100 iterations, depending on the dimension K of the model. The subsequent long EM run typically required between 1000 and 2000 iterations to reach convergence (we stopped the algorithm when the log-likelihood difference between the successive iterations was less than 10^{-6}).

EM short runs were initialized as in [7]. We randomly split the observations into K groups. The first M step was then performed on the basis of these initial groupings. Circular parameters were estimated from the available data by the method of moments, as suggested in [17]. Means, covariance matrices and skewness parameters of the skew normal components were estimated by their empirical counterparts, using the available data, by following Lin [13].

To select the number of components, we computed both the Bayesian information criterion (BIC) and the integrated complete likelihood (ICL) statistics (Table 2). The BIC statistic is a traditional approximation of the log-likelihood function, integrated with respect to a non-informative prior distribution of the unknown parameters, and reduces to the maximum value attained by the log-likelihood function, penalized by a function of the number of unknown parameters $\boldsymbol{\theta}$ to be

Table 2. Model selection results.

Number of components	Number of parameters	BIC	ICL
2	25	15557.2	15952.3
3	38	14945.4	15550.8
4	51	14865.1	15750.2
5	64	15040.1	16240.0

Table 3. Estimates and standard errors (within brackets).

	Component		
Parameter	1	2	3
β_{1k}	2.06	1.08	5.60
(Wave mean direction)	(0.07)	(0.03)	(0.08)
β_{2k}	3.13	1.33	4.61
(Wind mean direction)	(0.05)	(0.06)	(0.02)
β_{11k}	1.61	4.57	1.15
(Wave directional concentration)	(0.11)	(0.51)	(0.13)
β_{22k}	2.14	0.76	8.57
(Wind directional concentration)	(0.16)	(0.09)	(0.72)
β_{12k}	−0.19	3.09	1.23
(Wind/wave directional inverse correlation)	(0.23)	(0.28)	(0.28)
γ'_{1k}	0.38	1.85	0.70
(Wave mean height)	(0.05)	(1.69)	(0.10)
γ'_{2k}	1.51	3.35	3.12
(Wind mean speed)	(0.22)	(0.38)	(0.22)
γ_{11k}	0.06	0.41	0.12
(Wave height variance)	(0.01)	(0.29)	(0.03)
γ_{22k}	1.29	2.65	2.85
(Wind speed variance)	(0.29)	(0.64)	(0.55)
γ_{12k}	0.22	0.58	0.54
(Wind/wave covariance)	(0.03)	(0.10)	(0.06)
γ''_{1k}	0.21	0.18	0.20
(Wave skewness)	(0.07)	(2.15)	(0.13)
γ''_{2k}	0.93	1.54	1.68
(Wind skewness)	(0.26)	(0.44)	(0.25)
π	0.32	0.32	0.36
(Component weight)	(0.02)	(0.02)	(0.01)

estimated. In our application, the BIC takes the form

$$\text{BIC}\left(\hat{\boldsymbol{\theta}}, K\right) = -\log L_K(\hat{\boldsymbol{\theta}}) + \frac{K(5+7+1)}{2}\log n$$

and suggests a model with $K = 4$ components. However, this model distinguishes the same three clusters provided by a model with three components, using two overlapping components to approximate the distribution of the data under a single latent regime. This behavior of the BIC has been extensively discussed in [4], and in our application, it arises because the distribution of the data under one latent class is not very well approximated by the model. In our case study, however, overlapping components' lack of physical interpretation and cluster separation are more important than goodness of fit. We, therefore, used the ICL criterion, which approximates the integrated complete log-likelihood [4] and reduces to a BIC statistic, penalized by substracting

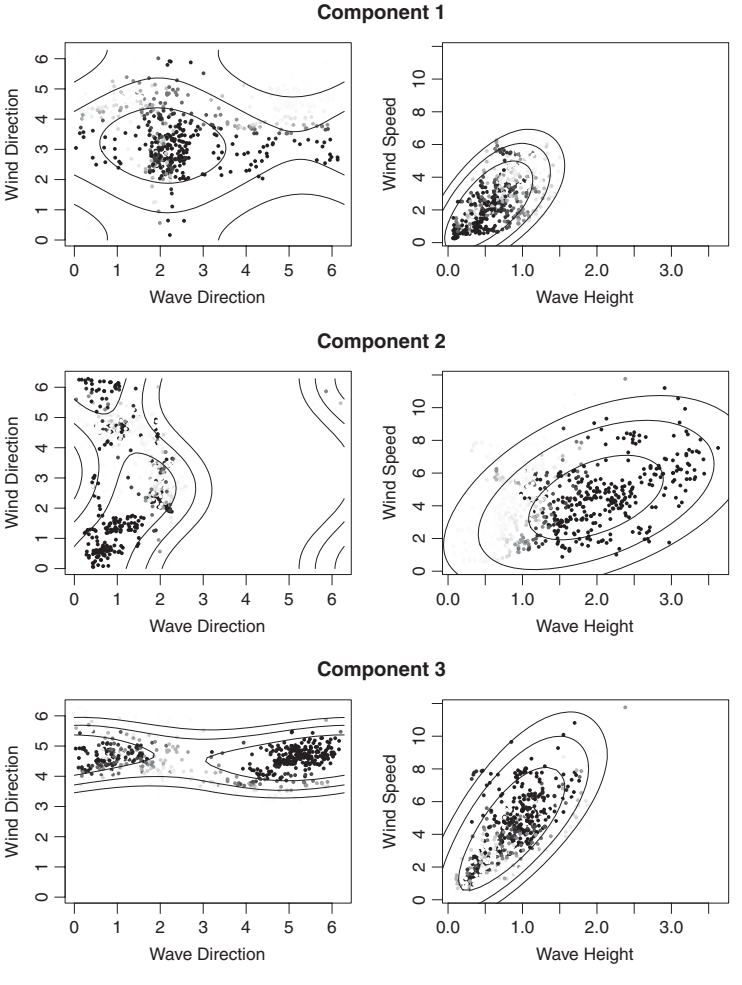

Figure 3. Contour plots of the conditional circular and linear bivariate log-densities, as estimated by fitting a three-component mixture model, at levels 4, 6 and 8; for each component, points are filled with a gray color that is proportional to the estimated probability that each observation belongs to that component.

the estimated mean entropy

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{\pi}_{\bar{T}ik}\log\hat{\pi}_{\bar{T}ik}.$$

Because the ICL includes cluster separation as an additional criterion for model choice, the minimum ICL is attained by a model with three components, which is the model that we considered to analyze the data.

The model estimates, displayed in Table 3, indicate the locations and shapes of three pairs of toroidal and planar clusters, depicted in Figure 3 through contour lines of bivariate log-densities.

The first component of the model includes about one-third of the sample ($\hat{\pi}_1 = 0.32$) and is associated with periods of calm sea: weak winds generate small waves. Under this regime, the shape of the joint distribution of wave and wind directions is essentially spherical (β_{121} is not significant at a 95% confidence level) and centered at the directional mean vector $(\hat{\beta}_{11}, \hat{\beta}_{21}) = (2.06, 3.13)$ that summarizes sirocco episodes (southeasterly winds and waves traveling along the major axis of the basin). As expected, wind and wave directions are poorly synchronized under good sea conditions, because if wind episodes are weak, then wave direction is more influenced by marine currents than by wind direction.

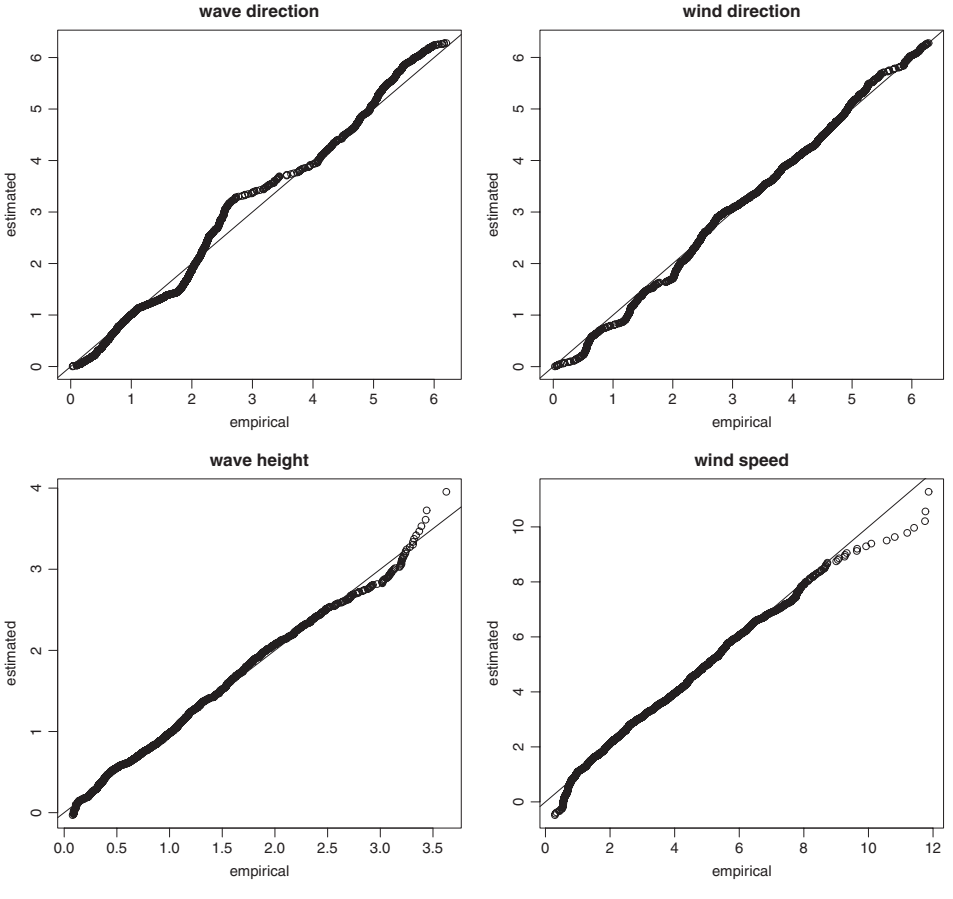

Figure 4. Quantile–quantile plots of the marginal distribution of wave direction and height (left) and wind direction and speed (right), as estimated by a model with three components.

The second and the third components are instead associated with bora episodes. Under the second component, bora jets blowing from a modal direction, $\hat{\beta}_{22} = 1.33$, drive high waves that travel along the major axis of the basin. Compared with episodes of calm sea, wind and wave directions appear to be strongly synchronized now. The third component is instead associated with episodes of bora jets that rotate at the Apennines barrier. Under this regime, the wind direction at the tide gauge is poorly synchronized with directions taken by waves, which travel according to offshore winds that are only partially captured by the tide gauge. These winds can blow at a considerable speed but with multi-modal directions and, as a result, generate waves of modest height.

Overall, the model indicates that the influence of coastal wind on offshore waves changes under different environmental regimes. The (marginal) weak correlation between wind speed and wave height can be then explained by the presence of a regime under which coastal winds do not generate waves of significant height. Under all the three latent regimes, moreover, wind skewness is modest, but significant at a 95% confidence level. On the contrary, and interestingly, wave skewness is either barely or not significant, indicating that the marginal skewness of wave height is essentially due to latent heterogeneity.

To identify latent sea regimes, the model tries to cluster the data, by providing an adequate fit of the univariate marginal and conditional distributions of the data. To check the marginal features

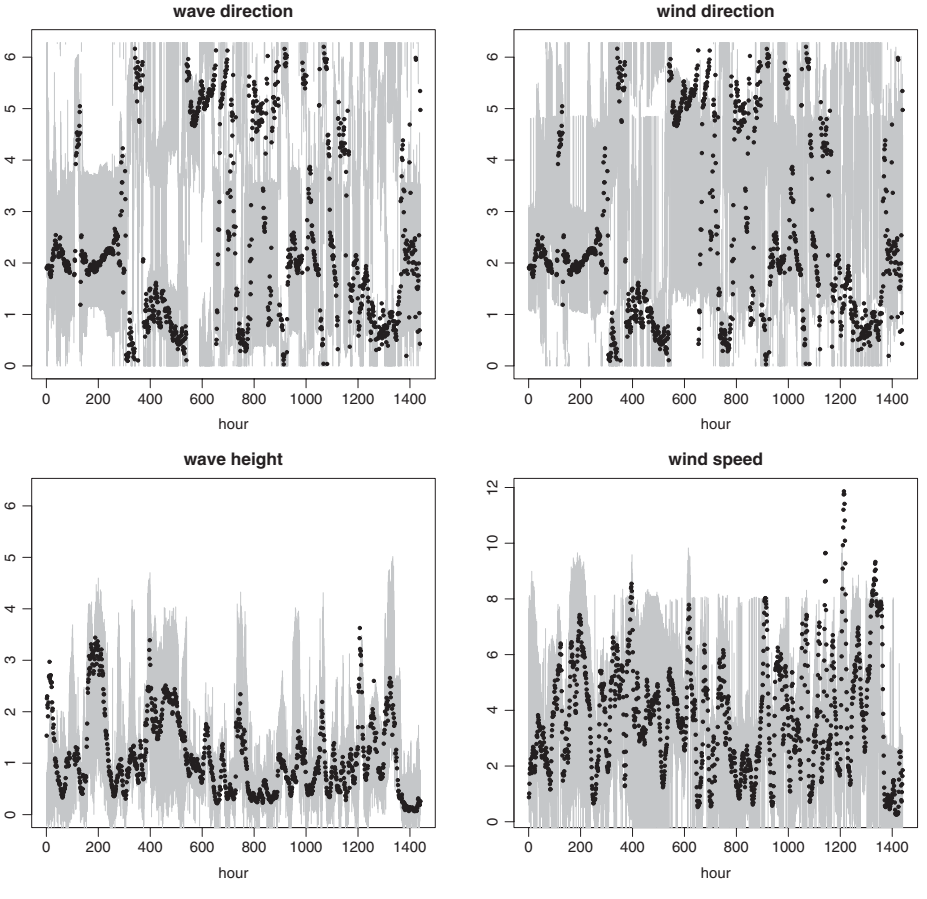

Figure 5. The 99% predictive intervals of wave direction and height (left) and wind direction and speed (right), as estimated by a model with three components.

of the model, we have computed the four quantile–quantile plots of the empirical quantiles of each variable versus the theoretical quantiles, as estimated by the model (Figure 4). These plots indicate a reasonable goodness of fit of the marginal distributions of the data. Departures from the $y = x$ line are due to the oversmoothing of the data, carried out by a mixture model that uses only there components. Oversmoothing can be alleviated by a larger number of components, at a price of overlapping components that are difficult to interpret, as discussed previously.

Conditional features of the model have been examined by computing predictive intervals (Figure 5) from the conditional distributions of each variable given the values of the remaining variables, as estimated by the three-component model. For simplicity, the two pictures at the top of the figure are drawn as rectangles, although they are cylinders. The model shows a good accuracy in predicting wave height and direction and wind speed. It seems to be less accurate in the prediction of wind direction, due to the extreme variability of this variable, only partially captured by three latent classes.

6. Discussion

We clustered multivariate data with circular components by associating toroidal and planar clusters into a finite number of latent classes. This classification strategy relies on a conditional independence assumption between the linear and the circular variables, given a latent multinomial variable. The advantages of this approach include a simple specification of the dependence structures between variables that are observed on different supports and the computational feasibility of a mixture-based classification strategy where missing values can be efficiently handled within a likelihood framework. These advantages have been illustrated with respect to the wind–wave data that are difficult to examine by means of traditional ocean numerical models [3].

To identify sea regimes, we exploited bivariate sine and skew normal distributions. While the EM algorithm given in Section 4 can be easily generalized to allow for multivariate skew normal distribution of any dimension, circular densities of a dimension larger than two are difficult to handle, because the normalizing constant is not known in a closed form. A first option could be to rely on specific M steps, based on the maximization of a complete pseudo-likelihood function. The pseudo-likelihood function provides good results in the maximum-likelihood estimation of trivariate circular densities [17], but its performance in a mixture context is at present not known. A second option could be to use a stochastic M step, based on the Markov Chain Monte Carlo methods that avoid the direct computation of the normalizing constant.

References

[1] R.B. Arellano-Valle, H. Bolfarine, and V.H. Lachos, *Bayesian inference for skew-normal linear mixed models*, J. Appl. Stat. 34 (2007), pp. 663–682.

[2] J.D. Banfield and A.E. Raftery, *Model-based Gaussian and non-Gaussian clustering*, Biometrics 49 (1993), pp. 803–821.

[3] L. Bertotti and L. Cavalieri, *Wind and wave predictions in the Adriatic Sea*, J. Mar. Syst. 78 (2009), pp. S227–S234.

[4] C. Biernacki, G. Celeux, and G. Govaert, *Assessing a mixture model for clustering with the integrated completed likelihood*, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000), pp. 719–725.

[5] C. Biernacki, G. Celeux, and G. Govaert, *Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models*, Comput. Statist. Data Anal. 41 (2003), pp. 561–575.

[6] J. Hagenaars and A. McCutcheon (eds.), *Applied Latent Class Analysis*, Cambridge University Press, Cambridge, 2002.

[7] L. Hunt and M. Jorgensen, *Mixture model clustering for mixed data with missing information*, Comput. Statist. Data Anal. 41 (2003), pp. 429–440.

[8] S. Ingrassia and R. Rocci, *Degeneracy of the EM algorithm for the MLE of multivariate Gaussian mixtures and dynamic constraints*, Comput. Statist. Data Anal. 55 (2011), pp. 1715–1725.

[9] D. Karlis and A. Santourian, *Model-based clustering with non-elliptically contoured distributions*, Stat. Comput. 19 (2009), pp. 73–83.
[10] S. Kato and K. Shimizu, *Dependent models for observations which include angular ones*, J. Statist. Plann. Inference 138 (2008), pp. 3538–3549.
[11] F. Lagona and M. Picone, *A latent-class model for clustering incomplete linear and circular data in marine studies*, J. Data Sci. 9 (2011), pp. 585–605.
[12] C.J. Lawrence and W.J. Krzanowski, *Mixture separation for mixed-mode data*, Stat. Comput. 6 (1996), pp. 85–92.
[13] T.I. Lin, *Maximum likelihood estimation for multivariate skew normal mixture models*, J. Multivariate Anal. 100 (2009), pp. 257–265.
[14] T.I. Lin, H. Ho, and P. Shen, *Computationally efficient learning of multivariate t mixture models with missing information*, Comput. Statist. 24 (2009), pp. 375–392.
[15] T. Lin, J. Lee, and W. Hsieh, *Robust mixture modeling using the skew t distribution*, Stat. Comput. 17 (2007), pp. 81–92.
[16] T.C. Lin and T.I. Lin, *Supervised learning of multivariate skew normal mixture models with missing information*, Comput. Statist. 25 (2010), pp. 183–201.
[17] K.V. Mardia, G. Hughes, C.C. Taylor, and H. Singh, *A multivariate von Mises distribution with applications to bioinformatics*, Canad. J. Statist. 36 (2008), pp. 99–109.
[18] K. Mardia, C. Taylor, and G. Subramaniam, *Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data*, Biometrics 63 (2007), pp. 505–512.
[19] G. McLachlan and D. Peel, *Finite mixture models*, Wiley, New York, 2000.
[20] A. Rotnitzky and D. Wypij, *A note on the bias of estimators with missing data*, Biometrics 50 (1994), pp. 1163–1170.
[21] D. Rubin, *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York, 1987.
[22] S.K. Sahu, D.K. Dey, and M.D. Branco, *A new class of multivariate skew distributions with applications to Bayesian regression models*, Canad. J. Statist. 31 (2003), pp. 129–150.
[23] J.L. Shafer, *Analysis of Incomplete Multivariate Data*, Chapman and Hall, London, 1997.
[24] H. Singh, V. Hnizdo, and E. Demchuk, *Probabilistic model for two dependent circular variables*, Biometrika 89(3) (2002), pp. 719–723.
[25] H. Teicher, *Identifiability of mixtures of product measures*, Ann. Math. Statist. 38 (1967), pp. 1300–1302.
[26] S. Yakowitz and J. Spragins, *On the identifiability of finite mixtures*, Ann. Math. Statist. 39 (1968), pp. 209–214.
[27] C.F.J. Wu, *On the convergence properties of the EM algorithm*, Ann. Statist. 11 (1983), pp. 95–103.

Appendix

To derive the analytical form taken by expectations $Q_{ic}(\boldsymbol{\beta}_k|\hat{\boldsymbol{\beta}}_{tk})$, we first observe that

$$\frac{\partial \log C(\boldsymbol{\beta}_k)}{\partial \beta_{11k}} = \frac{1}{C(\boldsymbol{\beta}_k)}\frac{\partial C(\boldsymbol{\beta}_k)}{\partial \beta_{11k}} = \frac{4\pi^2 \sum_{m=0}^{\infty}\binom{2m}{m}(\beta_{12}^2/(4\beta_{11}\beta_{22}))^m I_{m+1}(\beta_{11})I_m(\beta_{22})}{C(\boldsymbol{\beta}_k)}$$

$$\frac{\partial \log C(\boldsymbol{\beta}_k)}{\partial \beta_{22k}} = \frac{1}{C(\boldsymbol{\beta}_k)}\frac{\partial C(\boldsymbol{\beta}_k)}{\partial \beta_{22k}} = \frac{4\pi^2 \sum_{m=0}^{\infty}\binom{2m}{m}(\beta_{12}^2/(4\beta_{11}\beta_{22}))^m I_m(\beta_{11})I_{m+1}(\beta_{22})}{C(\boldsymbol{\beta}_k)}$$

$$\frac{\partial \log C(\boldsymbol{\beta}_k)}{\partial \beta_{12k}} = \frac{1}{C(\boldsymbol{\beta}_k)}\frac{\partial C(\boldsymbol{\beta}_k)}{\partial \beta_{12k}} = \frac{4\pi^2 \beta_{12}^{-1} \sum_{m=1}^{\infty}\binom{2m}{m}2m(\beta_{12}^2/(4\beta_{11}\beta_{22}))^m I_m(\beta_{11})I_m(\beta_{22})}{C(\boldsymbol{\beta}_k)},$$

respectively, indicate the marginal expectations of $\cos(x_1-\beta_{1k})$, $\cos(x_2-\beta_{2k})$ and $\sin(x_1-\beta_{1k})\sin(x_2-\beta_{2k})$ with respect to $f_c(\boldsymbol{x};\boldsymbol{\beta}_k)$. Furthermore, let a_k and b_k, respectively, be the functions (5) and (7), with $\boldsymbol{\beta}=\boldsymbol{\beta}_k$. We observe that

$$\frac{\partial \log a_k(x_1)}{\partial a_k(x_1)} = \frac{I_1(a_k(x_1))}{I_0(a_k(x_1))},$$

$$\frac{\partial \log a_k(x_2)}{\partial a_k(x_2)} = \frac{I_1(a_k(x_2))}{I_0(a_k(x_2))},$$

respectively, indicate the conditional expectation of $\cos(x_1-\beta_{1k}-b_k(x_2))$ with respect to $f_c(x_1|x_2;\boldsymbol{\beta}_k)$ and the conditional expectation of $\cos(x_2-\beta_{2k}-b_k(x_1))$ with respect to $f_c(x_2|x_1;\boldsymbol{\beta}_k)$.

Standard integration procedures and trigonometric identities allow to write $Q_{ik}(\boldsymbol{\beta}_k|\boldsymbol{\beta}_{kt})$ as a linear combination of expected sufficient statistics, whose value depends on the pattern $\boldsymbol{r}_i$ of the missing values within each profile. Precisely,

$$Q_{ik}(\boldsymbol{\beta}_k|\boldsymbol{\beta}_{tk}) = -\log C(\boldsymbol{\beta}_k) + \beta_{11k}E_{tik1} + \beta_{22k}E_{tik2} + \beta_{12k}E_{tik3},$$

where

$$E_{tik1} = \mathbb{E}(\cos(x_{i1} - \beta_{1k})|\boldsymbol{x}_{i,\text{obs}}, \xi_{ik} = 1)) = \begin{cases} \cos(x_{i1} - \beta_{1k}) & r_{i1} = 0, \\ \dfrac{1}{C(\boldsymbol{\beta}_k)}\dfrac{\partial C(\boldsymbol{\beta}_k)}{\partial \beta_{11k}} & r_{i1} = r_{i2} = 1, \\ \cos b_k(x_{i2})\dfrac{I_1(a_k(x_{i2}))}{I_0(a_k(x_{i2}))} & r_{i1} = 1, r_{i2} = 0, \end{cases}$$

$$E_{tik2} = \mathbb{E}(\cos(x_{i2} - \beta_{2k})|\boldsymbol{x}_{i,\text{obs}}, \xi_{ik} = 1)) = \begin{cases} \cos(x_{i2} - \beta_{2k}) & r_{i2} = 0, \\ \dfrac{1}{C(\boldsymbol{\beta}_k)}\dfrac{\partial C(\boldsymbol{\beta}_k)}{\partial \beta_{22k}} & r_{i1} = r_{i2} = 1, \\ \cos b_k(x_{i1})\dfrac{I_1(a_k(x_{i1}))}{I_0(a_k(x_{i1}))} & r_{i1} = 0, r_{i2} = 1, \end{cases}$$

$$E_{tik3} = \mathbb{E}\left(\sin(x_{i1} - \beta_{1k})\sin(x_{i2} - \beta_{2k})|\boldsymbol{x}_{i,\text{obs}}, \xi_{ik} = 1\right)$$

$$= \begin{cases} \sin(x_{i1} - \beta_{1k})\sin(x_{i2} - \beta_{2k}) & r_{i1} = r_{i2} = 0, \\ \dfrac{1}{C(\boldsymbol{\beta}_k)}\dfrac{\partial C(\boldsymbol{\beta}_k)}{\partial \beta_{12k}} & r_{i1} = r_{i2} = 1, \\ \sin b_k(x_{i2})\dfrac{I_1(a_k(x_{i2}))}{I_0(a_k(x_{i2}))}\sin(x_{i2} - \beta_{2k}) & r_{i1} = 1, r_{i2} = 0, \\ \sin(x_{i1} - \beta_{1k})\sin b_k(x_{i1})\dfrac{I_1(a_k(x_{i1}))}{I_0(a_k(x_{i1}))} & r_{i1} = 0, r_{i2} = 1. \end{cases}$$

Function $Q_2(\boldsymbol{\beta})$ is, therefore, maximized by separately solving the following system of score equations, for each k:

$$\begin{aligned} \frac{\sum_{i=1}^n \hat{\pi}_{tik}E_{tik1}}{\sum_{i=1}^n \hat{\pi}_{tik}} &= \frac{1}{C(\boldsymbol{\beta}_k)}\frac{\partial C(\boldsymbol{\beta}_k)}{\partial \beta_{11k}}, \\ \frac{\sum_{i=1}^n \hat{\pi}_{tik}E_{tik2}}{\sum_{i=1}^n \hat{\pi}_{tik}} &= \frac{1}{C(\boldsymbol{\beta}_k)}\frac{\partial C(\boldsymbol{\beta}_k)}{\partial \beta_{22k}}, \\ \frac{\sum_{i=1}^n \hat{\pi}_{tik}E_{tik3}}{\sum_{i=1}^n \hat{\pi}_{tik}} &= \frac{1}{C(\boldsymbol{\beta}_k)}\frac{\partial C(\boldsymbol{\beta}_k)}{\partial \beta_{12k}}, \\ \frac{\beta_{11k}\sum_{i=1}^n \hat{\pi}_{tik}A_{tik1} - \beta_{12k}\sum_{i=1}^n \hat{\pi}_{tik}C_{tik1}}{\beta_{11k}\sum_{i=1}^n \hat{\pi}_{tik}B_{tik1} + \beta_{12k}\sum_{i=1}^n \hat{\pi}_{tik}D_{tik1}} &= \tan\beta_{1k}, \\ \frac{\beta_{22k}\sum_{i=1}^n \hat{\pi}_{tik}A_{tik2} - \beta_{12k}\sum_{i=1}^n \hat{\pi}_{tik}C_{tik2}}{\beta_{22k}\sum_{i=1}^n \hat{\pi}_{tik}B_{tik2} + \beta_{12k}\sum_{i=1}^n \hat{\pi}_{tik}D_{tik2}} &= \tan\beta_{2k}, \end{aligned}$$

where

$$A_{tik1} = \mathbb{E}\left(\sin x_{i1} | \boldsymbol{x}_{i,\text{obs}}, \xi_{ik} = 1\right)) = \begin{cases} \sin x_{i1} & r_{i1} = 0, \\ \sin(\beta_{1k} + b_k(x_{i2})) \dfrac{I_1(a_k(x_{i2}))}{I_0(a_k(x_{i2}))} & r_{i1} = 0, r_{i2} = 1, \\ \sin \beta_{1k} \dfrac{1}{C(\boldsymbol{\beta}_k)} \dfrac{\partial C(\boldsymbol{\beta}_k)}{\partial \beta_{1kk}} & r_{i1} = r_{i2} = 1, \end{cases}$$

$$B_{tik1} = \mathbb{E}\left(\cos x_{i1} | \boldsymbol{x}_{i,\text{obs}}, \xi_{ik} = 1\right)) = \begin{cases} \cos x_{i1} & r_{i1} = 0, \\ \cos(\beta_{1k} + b_k(x_{i2})) \dfrac{I_1(a_k(x_{i2}))}{I_0(a_k(x_{i2}))} & r_{i1} = 1, r_{i2} = 0, \\ \cos \beta_{1k} \dfrac{1}{C(\boldsymbol{\beta}_k)} \dfrac{\partial C(\boldsymbol{\beta}_k)}{\partial \beta_{1kk}} & r_{i1} = r_{i2} = 1, \end{cases}$$

$$C_{tik1} = \mathbb{E}\left(\sin(x_{i2} - \beta_{2k}) \cos x_{i1} | \boldsymbol{x}_{i,\text{obs}}, \xi_{ik} = 1\right))$$

$$= \begin{cases} \sin(x_{i2} - \beta_{2k}) \cos x_{i1} & r_{i1} = r_{i2} = 0, \\ \cos(\beta_{1k} + b_k(x_{i2})) \dfrac{I_1(a_k(x_{i2}))}{I_0(a_k(x_{i2}))} \sin(x_{i2} - \beta_{2k}) & r_{i1} = 1, r_{i2} = 0, \\ \cos x_{i1} \sin \beta_{2k} \dfrac{I_1(a(x_{i1}))}{I_0(a(x_{i1}))} & r_{i1} = 0, r_{i2} = 1, \\ \cos \beta_{1k} \dfrac{1}{C(\boldsymbol{\beta}_k)} \dfrac{\partial C(\boldsymbol{\beta}_k)}{\partial \beta_{1kk}} & r_{i1} = r_{i2} = 1, \end{cases}$$

$$D_{tik1} = \mathbb{E}\left(\sin(x_{i2} - \beta_{2k}) \sin x_{i1} | \boldsymbol{x}_{i,\text{obs}}, \xi_{ik} = 1\right))$$

$$= \begin{cases} \sin(x_{i2} - \beta_{2k}) \sin x_{i1} & r_{i1} = r_{i2} = 0, \\ \sin(\beta_{1k} + b_k(x_{i2})) \dfrac{I_1(a_k(x_{i2}))}{I_0(a_k(x_{i2}))} \sin(x_{i2} - \beta_{2k}) & r_{i1} = 1, r_{i2} = 0, \\ \sin x_{i1} \sin \beta_{2k} \dfrac{I_1(a(x_{i1}))}{I_0(a(x_{i1}))} & r_{i1} = 0, r_{i2} = 1, \\ -\sin \beta_{1k} \dfrac{1}{C(\boldsymbol{\beta}_k)} \dfrac{\partial C(\boldsymbol{\beta}_k)}{\partial \beta_{1kk}} & r_{i1} = r_{i2} = 1, \end{cases}$$

and where A_{tik2}, B_{tik2}, C_{tik2}, D_{tik2} can be derived in a similar way, by exchanging x_1 with x_2.