UNIVERSITA' DEGLI STUDI DI ROMA TRE

Facoltà di Economia

Scuola Dottorale in Economia e Metodi Quantitativi
Sezione Metodi Statistici per l'Economia e l'Impresa

Coordinatrice Prof.ssa Julia Mortera

# MODEL BASED CLUSTERING OF MIXED LINEAR AND CIRCULAR DATA

Tutor:
Prof. FRANCESCO LAGONA

Dottorando:
MARCO PICONE

Correlatore:
Prof. JAN BULLA

XXV Ciclo - Roma, Giugno 2013

# Contents

# Introduction

The interest in the construction of types is gradually increased over time. The possibility to define *types* or *classes* or *states* of elements starting from a plurality of measurements is one of the general purpose of the analysis of complex phenomena and data mining. The general purpose of clustering is the identification of groups of objects (or clusters) such that objects in a cluster are very similar and objects in different clusters are quite distinct [11].

Literature is scattered by many publications on a great numbers of clustering algorithms in diversificated areas such as pattern recognition [1], artificial intelligence [19], image processing [4], biology [6], medicine and psychology [18], marketing [15], climate [9], oceanography [2] and meteorology [16], etc.

Over the last 40 years, a wealth of algorithms and computer programs has been developed for cluster analysis. The reason for the variety of methods is twofold. On the one side, automatic classification is a very young scientific discipline in vigorous development, as it can be seen from the thousands of articles scattered over many periodicals. Nowadays, automatic classification is establishing itself as an independent scientific discipline. Even if a general theory on cluster analysis does not exist, several authors attempted to unify all methods under an unified approach. On the other side a general definition of a cluster does not exist, and in fact there are several kinds of them: spherical clusters, drawn-out clusters, linear clusters, circular clusters, and so on. Moreover, different applications make use of different data types, such as continuous variables, discrete variables, similarities and dissimilarities. Therefore, one needs different clustering methods in order to adapt to the kind of application and the type of clusters under study.

In order to organize a collection of data items into clusters, such that items within a cluster are more *similar* to each others, the notion of similarity can be expressed in very different ways, according to the purpose of the study, to domain-specific assumptions and to prior knowledge of the problem.

The supervised clustering approach makes use of a small amount of *guides* or *adjustment* or *a priori* informations to address the clustering process. As a result, the *similarity* is addressed by the initial selection of training data

used to infer the model. Unsupervised clustering is performed when no information is available concerning the membership of data items to predefined classes. Under this setting, clustering is traditionally seen as part of unsupervised learning. In this case, the most important methods suggested by a large number of papers, include [13]:

- Hierarchical models, that aim at obtaining a hierarchy of clusters, called dendrograms, which show how the clusters are related to each other. These methods proceed either by iteratively merging small clusters into larger ones (agglomerative algorithms, by far the most common) or by splitting large clusters (divisive algorithms). A partition of the data items can be obtained by cutting the dendrogram at a desired level.

- Distance-based methods, that identify each cluster by minimizing a cost function that is the sum over all the data items of the squared distance between the item and the prototype of the cluster assigned. In general, the prototypes are the cluster centroids, as in the popular k-means algorithm [17].

- Density-based methods view clusters as dense sets of data items, separated by less dense regions; clusters may have arbitrary shape and data items can be arbitrarily distributed. Many methods have been developed in this framework, such as DBSCAN [21], [8], that relies on the study of the density of items in the neighbourhood of each item, or grid-based methods that quantize the space of the data items into a finite number of cells.

- Mixture-based methods assume that the data items in a cluster are drawn from one of several distributions (usually Gaussian) and attempt to estimate the parameters of all these distributions. The introduction of the expectation maximization (EM) algorithm in [5] was an important step in solving the parameter estimation problem.

Traditionally, techniques of wind-wave data clustering in meteorology and oceanography are based on distance-based methods. Recent proposals require the use of a finite number of target distributions, defined as cluster centroids, and an optimization algorithm that associates the observed data to the closest centroid [3]. Hierarchical agglomerative clustering methods [12] have been also suggested to avoid the specification of a family of target distributions. The statistical properties of distance-based methods are generally unknown [10], precluding the possibility of formal inference on the clustering results. This is a critical issue in marine studies, because the identification of wave regimes

without a measure of the statistical uncertainty of regime specific parameters is of little practical use. In addition, there is little systematic guidance associated with distance-based methods for solving basic questions that arise in cluster analysis, such as the choice of an optimal number of clusters and the choice of an optimal clustering algorithm. A general framework to address these issues is provided by latent-class models, which cluster multivariate data according to a finite number of classes, by approximating the joint distribution of the data by a mixture of parametric densities, which represent the distributional shape of the data within each cluster. From a methodological viewpoint, a latent-class approach allows to solve the clustering problem as a missing value problem, by treating the unknown cluster membership of each observation as a missing value, to be estimated from the data. From a technical viewpoint, the clustering algorithm reduces to likelihood maximization and the choice of the optimal number of clusters reduces to a model selection problem in parametric inference.

# Overview

Cluster analysis is a useful data mining method to obtain detailed information on the physical state of the ocean. Clusters discovery from marine data is one of the very promising subfields of data mining because increasingly large volumes of marine data are collected from gauges (buoys, drifters, platforms, etc.) and satellites and need to be analysed. The focus of the marine data mining is to maximize the information that can be derived from data of a marine environment. In the literature, only few works ( [20], [7], [14]) present modern techniques for the analysis of datasets collected in oceanography.

The primary objective of this thesis is the development of an algorithm to classify physical oceanographic data and identify relevant sea regimes. Typical oceanographic data used to describe sea regimes are wave height, wave direction, wind speed and wind direction. These data are collected in Italy by the Italian Data Buoy Network in several points along the coasts and are characterised by a mixed linear-circular support (linear support for wave height and wind speed, circular support for wave and wind directions). Furthermore, time series are often incomplete, due to unmoorings, transmission errors or maintenance operations of the gauges. All these issues suggest the use of mixture-based algorithms, that appear more flexible than others in order to model data with different supports and missing values.

Chapter 1 includes relevant issues about mixed linear-circular data and latent class methods for classification. In particular, we focus on the features of circular data, especially in the case of incomplete datasets. We further

introduce the general theoretical framework of mixture models and hidden Markov models.

In Chapter 2 we display the paper *A latent-class model for clustering incomplete linear and circular data in marine studies*, which illustrates a latent-class model that allows to jointly model wave and wind data by a finite mixture of conditionally independent Gamma and von Mises distributions. In this case all the variables are assumed conditionally independent given the latent state, and each measurement profile is independent from other profiles in time. In a maximum-likelihood framework, a flexible approach to handle missing values and mixed type data is presented in order to identify relevant sea regimes.

*Model-based clustering of multivariate skew data with circular components and missing values* is displayed in Chapter 3. In this paper, the independence assumption between variables is relaxed. Linear and circular data are modelled as a finite mixture of bivariate circular densities and bivariate skew linear densities to capture the association between toroidal clusters of circular observations and planar clusters of linear observations. 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.

In Chapter 4, in the paper *Maximum likelihood estimation of bivariate circular hidden Markov models from incomplete data*, we propose a hidden Markov model for the analysis of the time series of bivariate circular observations (wind and wave directions), by assuming that the data are sampled from bivariate circular densities, whose parameters are driven by the evolution of a latent Markov chain. Therefore, data are considered not independent in time, but the temporal correlation is modelled by a latent variable that satisfy the Markov property. In this paper a computationally feasible expectation-maximization (EM) algorithm is provided for the maximum likelihood estimation of the model from incomplete data, by treating the missing values and the states of the latent chain as two different sources of incomplete information.

*A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series* is proposed in Chapter 5. In this paper, wind and wave data are clustered by pursuing a hidden Markov approach, where toroidal clusters are defined by a class of bivariate von Mises densities, while skew-elliptical clusters are defined by mixed linear models with positive random effects. Several computational aspects regarding the evaluation of standard errors and the estimation step in the presence of missing values are discussed.

Conclusions and further remarks are discussed in Chapter 6.

# Bibliography

[1] J.C. Bezdek and S.K. Pal. *Fuzzy Models for Pattern Recognition, Methods That Search for Structures in Data*. Ieee, 1992.

[2] D. Birant and A. Kut. Cluster analysis for physical oceanographic data and oceanographic surveys in turkish seas. *Journal of Marine Research*, 64(5):651–668, 2006.

[3] A.V. Boukhanovsky, L.J. Lopatouhkin, and C. Guedes Soares. Spectral wave climate of the north sea. *Applied Ocean Research*, 29:146–154, 2007.

[4] L. Chen, M.W. Berry, and W.W. Hargrove. Using dendronal signatures for feature extraction and retrieval. *International Journal of Imaging Systems and Technology*, 11 (4):243–253, 2000.

[5] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39 (1):1–38, 1977.

[6] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome wide expression patterns. *Proceedings of the National Academy of Sciences USA*, 95 (25):14863–14868, 1998.

[7] W. J. Emery and R.E. Thomson. *Data Analysis Methods in Physical Oceanography*. Elsevier Science, 2nd edition, 2001.

[8] M. Ester, H.P. Kriegel, J. Sander, and X. Xu. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, chapter A density-based algorithm for discovering clusters in large spatial databases with noise. AAAI Press, 1996.

[9] R.G. Fovell and M.C. Fovell. Climate zones of the conterminous united states defined using cluster analysis. *J. Climate*, 6(11):2103–2135, 1993.

[10] C. Fraley and A.E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of American Statistical Association*, 97:611–631, 2002.

[11] G. Goujon, C. Ma, and J. Wu. *Data Clustering: Theory, Algorithms, and Applications*. ASA SIAM Series on statistics and applied probability, n.20, 2007.

[12] L.J. Hamilton. Characterising spectral sea wave conditions with statistical clustering of actual spectra. *Applied Ocean Research*, 32:332–342, 2010.

[13] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31 (3):264–323, 1999.

[14] G. Jona-Lasinio, A. Gelfand, and M. Jona-Lasinio. Spatial analysis of wave direction data using wrapped gaussian processes. *The Annals of Applied Statstics*, 2012.

[15] D.J. Ketchen and C. L. Shook. The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal*, 17 (6):441–458, 1996.

[16] C. Marzban and S. Sandgathe. Cluster analysis for verification of precipitation fields. *Weather Forecasting*, 21(5):824–838, 2006.

[17] J. McQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[18] C. Quantin, E. Sauleau, P. Bolard, C. Mousson, M. Kerkri, P. Brunet Lecomte, T. Moreau, and L. Dusserre. Modeling of high-cost patient distribution within renal failure diagnosis related group. *Journal of Clinical Epidemiology*, 52 (3):251–258, 1999.

[19] M. Rezazadeh and S.G. Choobary. Synergetic use of cluster analysis and artificial intelligence techniques to improve reservoir upscaling. In *International Oil and Gas Conference and Exhibition in China*, 2010.

[20] A.R. Robinson and M. Golnaraghi. The physical and dynamical oceanography of the mediterranean sea. In Paola Malanotte-Rizzoli and AllanR. Robinson, editors, *Ocean Processes in Climate Dynamics: Global and Mediterranean Examples*, volume 419 of *NATO ASI Series*, pages 255–306. Springer Netherlands, 1994.

[21] J. Sander, M. Ester, H.P. Kriegel, and X. Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data Mining and Knowledge Discovery*, 2 (2):169–194, 1998.

# Mixture-based classification methods for linear-circular data

Most of the literature on mixture-based classification methods is associated with the analysis of data whose components share the same support. Continuous observations are typically clustered by mixtures of univariate normal distributions [12], [2]. More generally, multivariate continuous data are clustered by mixtures of multivariate normal distribution [1]. In the case of skew or non-negative data, mixtures of univariate gamma [31], t [20] or Weibull [11] distributions have been proposed for robust classification. These proposal have been extended in a multivariate context, exploiting mixtures of multivariate skew normal [19] and t distributions [21], or non-elliptically contoured distributions [16]. In directional statistics, while mixtures of Kent distributions [25] and Von Mises distributions [23] are popular in the analysis of spherical data, toroidal data have been recently modelled by mixtures of bivariate circular densities [24]. Unsupervised classification of mixed type multivariate data has been studied only in the case of mixed linear and categorical data [13] [17].

## 1.1 Finite mixture models for multivariate data

Finite mixtures of distributions represent a mathematical-based approach to the statistical modeling of a wide range of phenomena. Because of their flexibility, finite mixture models have continued to receive increasing attention over the past years, from both theoretical and applied viewpoints. Finite mixture models have often been proposed and studied in the context of clustering. More recently, it has been recognized that these models can provide a principled statistical approach to the practical questions that arise in applying clustering methods [1], [9] [10]. Under a finite mixture model, each probability distribution corresponds to a cluster. The problems of determining the number of clusters and of choosing an appropriate clustering method can be recast as model choice problems, and models that differ in number of components and/or in component distributions can be compared [10]. One of the first

major analysis involving the use of mixture models was undertaken by Karl Pearson in 1894 [26] who fitted a mixture of two normal probability density functions to biological data. In 1977 Dempster et al. formalized the iterative procedure of the Expectation-Maximization (EM) algorithm that converges to the Maximum-Likelihood solution of the mixture problem. The impact of the EM, and the advent of high-speed computers, gave a great impulse to latent-class classification methods.

Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be a sample of multivariate observations that can be clustered into $K$ groups (or classes). Let $\boldsymbol{Z} = (Z_1, ..., Z_K)$ be a latent (unobserved) multinomial random vector with one trial and cell probabilities $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$.

In a finite mixture context, for each observation $i$ an unobserved vector $\boldsymbol{z_i}$ indicates whether $\boldsymbol{X}_i$ belong or not belong to the $k$th class ($k = 1, ..., K$). We assume that each $\boldsymbol{X}_i$ is independently sampled from a mixture of $K$ distributions, say

$$f(\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i), \tag{1.1}$$

$$\sum_{k=1}^{K} \pi_k = 1 \tag{1.2}$$

where $f_k(\boldsymbol{x}_i)$ denotes multivariate densities and the $\pi_k$ are probabilities that sum to one.

Each probability $\pi_k$ represents the $k$th mixing proportion or the probability that the observation $\boldsymbol{X}_i$ belongs to the $k$th subpopulation with corresponding density $f_k(\boldsymbol{x})$ called the $k$th mixing or component density. $K$ is the total number of components.

This is the most general form of a mixture: usually the $f_k$'s are assumed to be of parametric form i.e. $f_k(x) = f_k(x; \theta_k)$, where the functional form of $f_k(\cdot;\cdot)$ is known the parameter vector $\boldsymbol{\theta}_k$. Thus, (1.1) can be written in the form

$$f(\boldsymbol{x}_i, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i, \boldsymbol{\theta}_k), \tag{1.3}$$

The response $\boldsymbol{x}_1, ...., \boldsymbol{x}_n$ are assumed to be conditionally independent given $\boldsymbol{z}_i$, that is

$$f(\boldsymbol{x}_1, ..., \boldsymbol{x}_n \mid \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K, \boldsymbol{z}_1, ..., \boldsymbol{z}_n) = \prod_{i=1}^{n} f(\boldsymbol{x}_i \mid \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K, \boldsymbol{z}_i) \tag{1.4}$$

where $f(\boldsymbol{x}_i \mid \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K, \boldsymbol{z}_i)$ is written according to equation (1.3). Under this setting, the likelihood function is given by:

$$\mathcal{L}(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K, \boldsymbol{z}_1, ..., \boldsymbol{z}_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k). \tag{1.5}$$

### 1.1.1 EM Algorithm

The EM algorithm is the most widely used method for estimating the parameters of a finite mixture probability density. The EM algorithm is an iterative method for performing maximum likelihood estimation when some of the data are missing, and exploits the fact that the complete-data log-likelihood may be straightforward to maximize even if the likelihood of the observed data is not. The "complete-data log-likelihood" (CDLL) is the log-likelihood of the parameters of interest, based on both the observed data and the missing data [32]. In the mixture framework the observed vector is viewed as being incomplete as the associated component-label vector $\boldsymbol{z} = \boldsymbol{z}_1, ..., \boldsymbol{z}_n$ are not available, so they are "missing". The complete-data vector is therefore declared to be

$$\boldsymbol{x}_c = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{z}_1, ..., \boldsymbol{z}_n) \tag{1.6}$$

The component-label vectors $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ are taken to be the realization of the random vector $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n$, where, for independent feature data, they are distributed unconditionally as a multinomial distribution. This assumption implies that the complete-data log likelihood is given by

$$\log L_c(\boldsymbol{\theta}_c) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \{\log \pi_k + \log f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)\}, \tag{1.7}$$

where

$$\boldsymbol{\theta}_c = (\pi_1, ..., \pi_{k-1}, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K) \tag{1.8}$$

is the vector containing all the unknown parameters to be estimated.

The EM algorithm proceeds iteratively in two steps, the *E step* (for Expectation) and the *M step* (for Maximization).

The E step evaluates the conditional expectation of the complete-data log likelihood $\log L_c(\boldsymbol{\theta}_c)$, given the observed data $\boldsymbol{x}$, using the current estimate of $\boldsymbol{\theta}_c^{(t)}$ available from the last M step.

$$Q(\boldsymbol{\theta}_c; \boldsymbol{\theta}_c^{(t)}) = E\{\log L_c(\boldsymbol{\theta}_c) | \boldsymbol{x}\}. \tag{1.9}$$

The M-step on the $(k+1)$th iteration requires the global maximization of $Q(\boldsymbol{\theta}_c; \boldsymbol{\theta}_c^{(t)})$ with respect to $\theta_c$ over the parameter space, and yields the updated estimate $\boldsymbol{\theta}_c^{(t+1)}$. The E and M steps are alternated repeatedly until the difference

$$L(\boldsymbol{\theta}_c^{(t+1)}) - L(\boldsymbol{\theta}_c^{(t)}) \tag{1.10}$$

changes by an arbitrarily small amount in the case of convergence of the sequence of likelihood values $\{L(\boldsymbol{\theta}_c^{(t)})\}$.

The EM algorithm needs the specification of the entire set of unknown parameters $\boldsymbol{\theta}^{(0)}$ at step 0. Different starting strategies and stopping rules can lead to quite different estimates in the contest of fitting mixtures via EM algorithm [30].

The slow convergence of the algorithms will be exacerbated by a poor choice of the unknown parameters. In some cases where the likelihood is unbounded on the edge of the parameter space, the sequence of estimates may diverge if the initial values is chosen too close to the boundary. Furthermore, the likelihood equation will usually have multiple roots corresponding to local maxima, especially in case of great complexity of the model (e.g. great number of components/states). So, the algorithms should be applied starting from a wide choice of initial values.

Usually, the random starts are setted as initial partitions $\boldsymbol{z}^{(0)}$. Data are random divided into $K$ groups corresponding to the $K$ components or states of the model. That is, for each observation $\boldsymbol{x}_i$, we randomly generate an integer between 1 and $K$, both inclusive. If this random integer is equal to $h$, the $k$th element of $\boldsymbol{z}_i^{(0)}$ is equal to 1 for $i = h$ and equal to zero for $i \neq h$ $(k = 1, ..., K)$.

In order to avoid local maxima short-runs strategies can be applied [4], by running the algorithms from a number of random initializations, and stopping the them without waiting for full convergence. The convergence to spurious maxima is fast (a phenomenon that is well known in the case of mixtures of multivariate normal densities; [14]) and can be detected within short runs by monitoring the class proportions.

In Chapter 2 and 3, EM algorithms are developed in case of mixture models with univariate and multivariate linear-circular densities.

### 1.1.2 Missing Values

Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ denote a set of incomplete multivariate observations. The $n \times J$ matrix measurement $X$ can be split in the observed part $X_O$ and the missing part $X_M$, so that $X = (X_O, X_M)$. It can be assume that the missing values are *missing at random* (MAR), so the probability that an observation is

missing may depends on $X_O$ but not on $X_M$. This assumption is less restrictive than MCAR (*missing completely at random*) in which the missing data are a simple random sample of all data values but it is more restrictive than NIM (*non-ignorable missing*) in which the value of the missing observations depends on the value of the missing data. Known the reasons of the missing mechanism, it is reasonable to assume the MAR hypothesis as missing mechanism.

For mixture-based data clustering, maximum-likelihood estimation could be carried out by discarding incomplete data profiles from the sample and using the complete cases to build up the likelihood function to be maximized (CC; complete case analysis).

If the joint distribution of the variables of interest is correctly specified and the data are missing at random, CC-based maximum-likelihood estimation is known to be (asymptotically) unbiased but inefficient. Loss of efficiency is due to the fact that incomplete profiles are informative of the parameters of the joint distribution of several variables, especially when these variables are strongly correlated. Efficient maximum-likelihood estimation can be carried out from MAR multivariate data often through data-augmentation or multiple-imputation methods [29].

Let us introduce a $n \times J$ matrix $R$ whose generic component $r_{ij} = 1$ if $x_{ij}$ is missing and 0 otherwise. Accordingly, the row-sums of $R$, say $r_{i\cdot} = \sum_{j=1}^{J} r_{ij}$, indicate the number of missing values within each $i$th profile. If the data are MAR, i.e. the probability of a missing value does not depend on the value that is missing, maximum likelihood estimates of model (1.1) can be found by maximizing the marginal log-likelihood function

$$\begin{aligned}
\log L(\boldsymbol{\theta}, \boldsymbol{\pi}) = & l(\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \log \int_{\boldsymbol{X}_{M,i}} \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i \mid \theta_k) d\boldsymbol{X}_{M,i} = \\
& = \sum_{i: r_{i\cdot}=0} \log \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i \mid \theta_k) + \\
& + \sum_{i > r_{i\cdot}=0} \log \int_{\boldsymbol{X}_{M,i}} \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i \mid \theta_k) d\boldsymbol{X}_{M,i} = \\
& = l_{CC}(\boldsymbol{\theta}, \boldsymbol{\pi}) + l_{IC}(\boldsymbol{\theta}, \boldsymbol{\pi});
\end{aligned} \tag{1.11}$$

which is the sum of the log-likelihood contributions of the complete (CC) and incomplete cases (IC).

### 1.1.3 Standard errors

Another critical issue is the provision of standard errors associated to the parameter estimates. One of the criticism of the EM algorithm is that it does not automatically provide an estimate of the covariance matrix of the estimates. A number of methods have been proposed in order to asses the covariance matrix of the parameter vector $\boldsymbol{\theta}_c$ obtained via the EM algorithm. One way is to approximate the covariance matrix of the parameters by the inverse of the observed information matrix $\boldsymbol{I}(\hat{\boldsymbol{\theta}}_c, \boldsymbol{x})$. The matrix $\boldsymbol{I}(\hat{\boldsymbol{\theta}}_c, \boldsymbol{x})$ can be directly evaluated after the computation of the MLE $\hat{\boldsymbol{\theta}}_c$. However, analytical evaluation of the second-order derivatives of the log likelihood may be difficult for most mixture models, in particular in multivariate cases. Louis [22] showed that $\boldsymbol{I}(\hat{\boldsymbol{\theta}}_c, \boldsymbol{x})$, the negative of the Hessian of the incomplete-data log likelihood can be computed in terms of the conditional moments of the gradient and curvature of the complete-data log likelihood.

$$\boldsymbol{I}(\hat{\boldsymbol{\theta}}_c, \boldsymbol{x}) = \boldsymbol{\mathcal{I}_c}(\hat{\boldsymbol{\theta}}_c, \boldsymbol{x}) \tag{1.12}$$

The calculation of $\boldsymbol{\mathcal{I}_c}(\hat{\boldsymbol{\theta}}_c, \boldsymbol{x})$ can be facilitated if the complete-data density belongs to the regular exponential family.

On the other hand a resampling approach may be considered. Standard error estimation of $\hat{\boldsymbol{\theta}}_c$ may be implemented according to the bootstrap [7] technique, that permits to evaluate the variability of a random quantity using just the data at hand. It follows three steps:

- a new set of data $\boldsymbol{x}^*$, called *bootstrap sample*, is generated according to $\hat{F}$, an estimate of the distribution function of $X$ formed from the original observed data $x$;
- the latent-class model is estimated from $\boldsymbol{x}^*$, to compute the maximum likelihood estimates for this dataset, say $\hat{\boldsymbol{\theta}}_c^*$;
- the bootstrap covariance matrix of $\hat{\boldsymbol{\theta}}_c^*$ can be approximated by the sample covariance matrix of these $B$ bootstrap replications:

$$cov(\hat{\boldsymbol{\theta}}_c^*) \approx \sum_{b=1}^{B} (\hat{\boldsymbol{\theta}}_{c,b}^* - \overline{\hat{\boldsymbol{\theta}}_c^*})(\hat{\boldsymbol{\theta}}_{c,b}^* - \overline{\hat{\boldsymbol{\theta}}_c^*})^T/(B-1) \tag{1.13}$$

where

$$\overline{\hat{\boldsymbol{\theta}}_c^*} = \sum_{b=1}^{B} \hat{\boldsymbol{\theta}}_{c,b}^*/B \tag{1.14}$$

The standard error of the $i$th element of $\hat{\theta}_c$ can be estimated by the positive square root of the $i$th diagonal element of (1.13).

It has been shown that 50 to 100 bootstrap replications are generally sufficient for standard error estimation [8]. Peel [27] compares the bootstrap and information-based approaches for some normal mixture models. He found that unless the sample size is very large, the standard errors found by an information-based approach were too unstable to be recommended.

### 1.1.4 Model Selection

Testing the number of components or states $K$ in a latent class model and the comparison between different models are important but very difficult problems which have not been completely resolved. In a clustering context, the choice of $K$ arises with the question of how many clusters or classes there are. There are two common ways to evaluate the number of clusters. One way is based on a penalized form of the likelihood. As the likelihood increases with the addition of a component, the criteria penalizes the model (usually, the log-likelihood) by a term that depends on the number of parameters. On the other hand it is possible to carry out an hypothesis test, using the likelihood ratio as test statistic. Penalized likelihood criteria are less demanding than the likelihood ratio test, which requires bootstrapping in order to obtain an assessment of the P-value [25]. The commonly used penalized likelihood criteria that would appear to be adequate are AIC and BIC, because they do not underestimate the true number of components, asymptotically [18]. The Akaike's Information Criterion (AIC) selects the model that minimizes

$$-2\log L(\boldsymbol{\theta}_c) + 2d \tag{1.15}$$

where $d$ is the total number of parameters $\boldsymbol{\theta}_c$ in the model. In the literature, many authors observe that AIC tends to overfit the data.

The Bayesian Information Criterion (BIC) penalizes the likelihood by the total number of parameters $d$ and the dimension of the sample size $n$:

$$-2\log L(\boldsymbol{\theta}_c) + d\log n. \tag{1.16}$$

In general BIC tends to penalize complex model more heavily than AIC, so it reduces the tendency of the AIC to fit too many components.

Additionally to AIC and BIC criterion, many others criteria have been developed in literature, but in a clustering context some classification-based information criteria are assuming an increasing relevance. The Integrated classification likelihood (ICL) is the easiest to apply and it selects the true model [25], whose components appears to be well separated. The ICL (in

its simplified ICL-BIC version) penalises the likelihood function by the total number of unknown parameters, the sample size and the entropy:

$$-2\log L(\boldsymbol{\theta}_c) + 2EN(\hat{\pi}) + d\log n. \tag{1.17}$$

where

$$EN(\hat{\pi}) = -\sum_{k=1}^{K}\sum_{i=1}^{n} \pi_{ki}\log \pi_{ki}, \tag{1.18}$$

is the entropy index, computed by summing up the posterior classification probabilities.

## 1.2 Hidden Markov Models

Although initially introduced and studied in the late 1960s and early 1970s, hidden Markov models have become increasingly popular in the last fifteen years, due to their rich mathematical structure that make them very flexible. Actually they can form the theoretical basis for use in a wide range of applications. HMM belongs to a class of Markovian models for which the dynamics of the stochastic process are completely or partially governed by a Markov chain or a Markov process. The model is *hidden* in the sense that the stochastic process is only partially observable. These models are used for investigating the properties of a given signal or time-series. As pointed out by Rabiner [28], real-world processes generally produce observable outputs which can be characterised as signals that can be discrete or continuous, stationary or nonstationary, pure or corrupted by transmission, distortions, reverberations, etc. There are two main approaches for characterizing these time-series. Deterministic models generally exploit some known properties of the signal, such as periodicity, decomposability, etc, that completely summarize the signal features. The second broad class is the set of statistical models (Gaussian models, Markov processes), under which it is possible to characterize only the statistical properties of the signal. The underlying assumption of the statistical model is that the signal can be well characterized as a parametric random process and that the parameters of the stochastic process can be determined (estimated) in a precise, well-defined manner.

For this reason, hidden Markov models have been used to handle a variety of real-world time dependent data in a large number of scientific fields, such as speech recognition, health science, environmental science, biology, etc.

A hidden Markov model can be considered a generalization of a mixture model where the hidden variables (or latent variables), which control the mix-

ture component to be selected for each observation, are related through a Markov process rather than independent of each other. Thus, the HMM provides a convenient way of formulating an extension of a mixture model to allow for dependent data.

### 1.2.1 Discrete-state Markov process

Consider a system which may be described at any time as being in one of a set of $K$ distinct states $S = \{1, ..., K\}$. According to a set of probabilistic rules, the system may, at certain discrete instants of time, undergo *changes of state* (or *state transitions*). Let $S_t(j)$ be the event that the system is in state $j$ at time $t$. The probability of this event can be written as $Pr(S_t(t))$. Each event may be described by transition probabilities as follow

$$Pr(S_t = j | S_{t-1} = a, S_{t-2} = b, ..., S_0 = c) \tag{1.19}$$

where $t = 1, 2, 3, ...$ and $1 \leq j, a, b, c, ..., \leq K$ are possible values from a countable set $\mathcal{S}$ called the *state space* that can be written as $s_t$, realization of $S_t$. These transition probabilities specify the probabilities associated with each instant, and they are conditional on the entire past history of the process. If the transition probabilities for a series of dependent trials satisfy the *Markov condition*, namely:

$$\begin{aligned} Pr(S_t = s_t | S_{t-1} = s_{t-1}, S_{t-2} = s_{t-2}, ..., S_0 = s_0) = \\ Pr(S_t = s_t | S_{t-1} = s_{t-1}) \quad \forall t, \end{aligned} \tag{1.20}$$

then the system is said to be a *discrete-state discrete-transition Markov process* [6]. If the state of the system at time $t - 1$ is known, the Markov condition requires that the conditional transition probabilities describing in time $t$ do not depend on any additional past history of the process [6].

The *state transition probabilities* for a discrete Markov process can be defined as follow:

$$\gamma_{ij} = Pr(S_t = j) | S_{t-1} = i) \quad 1 \leq i, j \leq K. \tag{1.21}$$

The two obvious properties of the transition probabilities are:

- $\gamma_{i,j} \geq 0$
- $\sum_{j=1}^{K} \gamma_{i,j} = 1$ for $i = 1, ..., K$.

It is convenient to display these transition probabilities as member of a $K \times K$ transition matrix $\Gamma$ for which $\gamma_{ij}$ is the entry in the $i$th row and $j$th column, namely:

$$\Gamma = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1K} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2K} \\ \cdots & \cdots & \cdots & \cdots \\ \gamma_{K1} & \gamma_{K2} & \cdots & \gamma_{KK} \end{bmatrix}. \tag{1.22}$$

To completely define the Markov chain the initial distribution of the states $\delta = (\delta_1, \delta_2, ..., \delta_K)$ must be defined as follow:

$$\delta_j = Pr(S_0 = j), \quad j = 1, ..., K. \tag{1.23}$$

### 1.2.2 Likelihood function and parameter estimation

Let $\{X_t\}_{t\geq 0}$ be a stochastic process, $y_t$ its realization, that corresponds to the observed response at time $t$, and a discrete, homogeneous, aperiodic, irreducible Markov chain $\{S_t\}_{t\geq 0}$ as described before with a transition matrix $\Gamma$. The model for the observed process $X_t$ can be defined as follow:

$$f_j(X_t|\theta) = Pr(X_t|S_t = j), \tag{1.24}$$

where $\theta$ denote the corresponding parameter set. The stochastic process is linked directly to the Markov chain $S_t$ that drives the distribution of the corresponding $X_t$. The observed process must satisfy two conditions:

- conditional independence condition: random variables $X_{0:T} = (X_0, ..., X_T)$ are conditionally independent given the latent states $S_{0:T} = (S_0, ..., S_T)$;
- contemporary dependence condition: the distribution of any $X_t$, given the state variables $(S_0, ..., S_T)$, depends only on the current state $S_t$.

Taking into account these assumptions, the likelihood function $L(\theta, \Gamma, \delta; x_{0:T})$ can be defined as a function of the model parameters $\theta_c = (\theta, \Gamma, \delta)$ when the observation sequence $x_{0:T}$ is given, as follow:

$$\begin{aligned} L(\theta_c; x_{0:T}) &= \sum_{s_{0:T}\in\mathcal{S}} Pr(X_{0:T} = x_{0:T}, S_{0:T} = s_{0:T}|\theta_c) = \\ &= \sum_{\mathcal{S}} \delta_{s_0} \prod_{t=1}^{T} \gamma_{s_{t-1},s_t} \prod_{t=0}^{T} f_{s_t}(x_t|\theta_{s_t}) \\ &= \delta\Gamma P(x_1)\Gamma P(x_2)\Gamma P(x_3)...\Gamma P(x_T)1'. \end{aligned} \tag{1.25}$$

In this form, the likelihood expression can not be evaluated because the number of operations involved is of order $TK^2$, making the evaluation of the likelihood quite feasible even for large $T$. In the case of discrete state-dependent distributions, the likelihood function, being made up of products of probabilities, become progressively smaller as $t$ increases, and are eventually rounded to zero.

The likelihood function given the observed data can be efficiently computed by a Baum-Welch (BW) procedure, by specifying the posterior probabilities in terms of suitably normalized functions, which can be computed recursively. In the literature, this approach is known as the Forward-Backward (FB) recursion and it can be implemented in a number of different ways (details can be found in Chapter 6, appendix B).

Another computational problem in HMM design concerns the estimation of the parameters $\theta_c$. In the literature, two common approaches may be used to maximize the likelihood function with respect to the parameters. Parameter estimation can be performed by direct numerical maximization of the likelihood with respect to the parameters. There are several problems that need to be addressed when the likelihood is maximized numerically. Relevant issues include numerical underflow, re-parametrization of the model in terms of unconstrained parameters, and multiple local maxima in the likelihood function [32]. Since the sequence of states visited by the Markov-chain component of an HMM is not observed, a very natural approach to parameter estimation in HMMs is to treat those states as missing data and to employ the EM algorithm in order to find maximum likelihood estimates of the parameters. As already described in Section 1.1.1, in the EM framework the observed vector $X_{0:t}$ is viewed as being incomplete, $S_{0:T}$ is called the "missing" data, while $(X_{0:t}, S_{0:t})$ is the complete data vector. The complete-data log-likelihood function can be easily computed as:

$$\begin{aligned} \log L_c(\theta_c) &= \log Pr(x_{0:T}, s_{0:T}|\theta_c) = \\ &= \sum_{j=1}^{K} \log \delta_j + \sum_{j=1}^{K}\sum_{k=1}^{K} \gamma_{jk} + \sum_{t=0}^{T}\sum_{j=1}^{K} f_j(x_t|\theta_{s_t}). \end{aligned} \tag{1.26}$$

As full convergence of the EM is slower then numerical maximization algorithms, it is convenient to combine the EM algorithm and numerical optimization. This approach is often called a hybrid algorithm [5] and provides a compromise between the large circle of convergence provided by the EM algorithm and the high speed of direct numerical maximization. The approach worked well in the applications considered in this thesis as we observed that

direct maximization of the log-likelihood is numerically stable and rapid when initial parameters are in the neighbourhood of a maximum (more details can be found in Chapter 5).

## 1.3 Linear-Circular data in marine datasets

Marine data are useful to describe the physical condition of a given basin. A set of atmospheric conditions, in particular wind conditions, generates different wave regimes depending on the physical, orographic, geographic features. Usually wave regimes are evaluated through numerical and semi-deterministic models starting from physical models of global atmospheric circulation. In particular complex situations such as coastal areas or semi-enclosed basins (e.g. basins in Mediterranean Sea) numerical wind-wave models, traditionally used for ocean waves, can give inaccurate results [3].

The data motivating this work are time series of wave and wind directions (*circular data*), as well as wind speeds and significant wave heights (*linear data*), recorded in wintertime by the Italian data buoy network, owned and managed by the Institute for Environmental Protection and Research. These time series are typically incomplete: the measurements taken by gauges can be missing because of unmoorings, devices maintenance or discontinuous functioning. This time series of environmental measurements were recorded at Ancona, where the buoy is located in the Adriatic Sea about 30 km from the coast in high water. The Adriatic Sea is a semi-enclosed basin. This area is subject to three relevant wind events: Bora, Maestral and Sirocco winds. Sirocco arises from a warm, dry, tropical air mass that is pulled northward by low-pressure cells moving eastward across the Mediterranean Sea. It typically blows from March to October and generates effects along the major axis of the Adriatic basin (along southeast-northwest direction). Bora episodes occur when a polar highpressure area sits over the snow-covered mountains of the interior plateau behind the coastal mountain range and a calm low-pressure area lies further south over the warmer Adriatic. It transfers a great amount of energy to the northern portion of the Adriatic basin essentially in wintertime. Finally, the Maestral is a sea-breeze wind blowing north-westerly when the east Adriatic coast gets warmer than the sea. While Bora and Sirocco episodes are usually associated with high-speed flows, Maestral is in general linked with good meteorological conditions. The orography of the Adriatic Sea plays a key role in this case study and most of the waves tend to travel from north-northwest and south-easterly along the major axis of the basin, where they can travel freely, without being obstructed by physical obstacles, such as coastlines. In wintertime (from November to March) the dominant

winds are Maestral and Bora. The first one generates low waves that travel along the Italian coast from northwest to southeast. The Bora wind generates high waves coming from North, reaching heights up to four to five meters. In summertime (April to October) the dominant winds are Maestral and Sirocco with waves never greater than three meters.

### 1.3.1 Circular Data

Directional measurements arise in many scientific fields. For examples, biologists measure the direction of flight of birds, while geologist may be interested in the direction of the earth's magnetic pole, geneticists evaluate the attack direction of proteins, and meteorologists predict wind and waves directions.

A directional data can be represented in a plane like a two-dimensional measure given by an angle measured with respect to some suitably chosen "zero direction", i.e., the starting point and a "sense of rotation", i.e., whether clockwise or anti-clockwise, is taken as positive direction [15]. The simple directional data (without any magnitude information) can be conveniently represented as points on a circumference of a unit circle centered at the origin. Because of this circular representation, directional data are also called *circular data*.

Directional data have many unique and novel features both in terms of modeling and in their statistical treatment [15]. For instance, since the angular value depends on the choice of the zero direction and the sense of rotation, the numerical representation of a direction is not unique. There exist $\infty^2$ different representations of the same values. Some relevant reference systems are adopted in literature. In mathematical analysis, the current reference system takes East as zero-direction and anti-clockwise as positive direction. In practical applications, the reference system use North as zero direction and clockwise as positive direction. It is possible to change the reference system through linear transformations of the circular observations (given $\alpha$ measured in the second reference system, it is possible to convert it into the first reference system by evaluating $\pi/2 - \alpha$).

Directional data are usually measured in degrees. However, it is sometimes useful to measure in radians. Angular measurements may be converted from degree to radians by multiplying by $\pi/180$.

Since the "beginning" coincides with the "end" i.e., $0 = 2\pi$ (in radians) or $0 = 360°$ (in degrees), directional data are assumed to have a bounded support, but the measurement is also periodic with $\alpha$ being the same as $\alpha + p2\pi$ for any integer $p$, with unbounded support.

It is therefore important to make sure that all statistical inferences and summaries are functions of the given observations and do not depend either

on the reference system or on the periodicity of the value.

These features make directional analysis substantially different from the standard "linear" statistical analysis of univariate and multivariate data. Commonly used summary measures on the real line turn out to be inappropriate. For instance, it is not possible to evaluate the euclidean distance between measurements. A reasonable measure of the *circular distance* between data is the smaller of the two arc lengths between the points along the circumference [15], i.e. for any two angles $\alpha$ and $\beta$

$$d_0(\alpha, \beta) = min(\alpha - \beta, 2\pi - (\alpha - \beta)) = \pi - |\pi - |\alpha - \beta||. \tag{1.27}$$

The circular mean is defined as the direction of the resultant vector of the sample

$$\bar{\alpha} = arctan \frac{\sum_{i=1}^{n} \sin \alpha_i}{\sum_{i=1}^{n} \cos \alpha_i}. \tag{1.28}$$

Even the classical representation of time series must be redefined. The classical histogram can give misleading informations because it does not represent the periodic feature of the data. Grouped circular data can be represented by circular histograms, which are analogous to histograms on the real line, but "wrapped" on a circle. A useful variant of the circular histogram is the rose diagram, in which the bars of the circular histogram are replaced by sectors. The area of each sector is proportional to the frequency in the corresponding groups.

As a result, also the probability distributions must be rewritten. A circular distribution is a probability distribution that has the following basic properties:

- $f(\alpha) \geq 0$;
- $\int_0^{2\pi} f(\alpha) d\alpha = 1$;
- $f(\alpha) = f(\alpha + 2p\pi)$ for any integer $p$.

In literature, many circular models may be generated from known probability distributions on the real line or on the plane by a variety of mechanism (wrapping methods, offset distributions, stereographic projection methods, etc.), and a number of circular distribution has been studied, such as the uniform circular distribution:

$$f(\alpha) = \frac{1}{2\pi}, \quad 0 \leq \alpha \leq 2\pi. \tag{1.29}$$

The most used circular distribution is the von Mises (or Circular Normal) distribution. A circular random variable $\alpha$ is said to have a von Mises distribution if it has the density function:

$$f(\alpha;\mu,\kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa\cos(\alpha-\mu)}, \quad 0 \leq \alpha \leq 2\pi \tag{1.30}$$

where $0 \leq \mu \leq 2\pi$ and $\kappa \geq 0$ are parameters and $I_0(\kappa)$ in the normalizing constant is the modified Bessel function of the first kind and order zero and is given by

$$I_0(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} exp(\kappa\cos(\alpha))d\alpha = \sum_{r=0}^{\infty}\left(\frac{\kappa}{2}\right)^{2r}\left(\frac{1}{r!}\right)^2. \tag{1.31}$$

The von Mises distribution verify the following properties:

- Symmetry: by the symmetry of the cosine function, the density is symmetric about the direction $\mu$;

- Mode at $\mu$: the cosine function has a maximum value at zero, so the density is maximum at $\alpha = \mu$, i.e., $\mu$ is the modal direction with the maximum value
$$f(\mu) = \frac{e^{\kappa}}{2\pi I_0(\kappa)}. \tag{1.32}$$

- Anti-mode at $(\mu \pm \pi)$: the density is minimum at $\alpha = \mu \pm \pi$,
$$f(\mu \pm \pi) = \frac{e^{-\kappa}}{2\pi I_0(\kappa)}. \tag{1.33}$$

- Role of $\kappa$: from the previous properties,
$$\frac{f(\mu)}{f(\mu \pm \pi)} = e^{2\kappa}. \tag{1.34}$$
Hence, the larger the value of $\kappa$, the larger will be the ratio of $f(\mu)$ to $f(\mu \pm \pi)$ indicating higher concentration towards the population mean direction $\mu$. Thus, $\kappa$ is a parameter which measure the concentration towards the mean direction $\mu$.

## 1.4 Mixture-based classification methods for marine datasets

The multivariate data described in Section 1.3 can be represented as $n$ vectors $\boldsymbol{d}_i = (d_{i1}, ..., d_{iJ})$, $i = 1, ..., n$, drawn from the multivariate distribution of $J$ variables $D_j$, $j = 1, ..., J$, measured on linear and circular supports. To account the different linear and circular support, we split the complete data vector $\boldsymbol{d}_i = (x_{i1}, ..., x_{iJ_1}; y_{i1}, ..., y_{iJ_2})$ into a vector $\boldsymbol{x}_i$ of $J_1$ circular data and a vector $\boldsymbol{y}_i$ of $J_2$ linear data, with $J = J_1 + J_2$.

In Chapters 2 and 3 we define two different mixture models starting from different conditional independence assumptions.

The Univariate Mixture Model (UMM, Chapter 2) simplifies the classification problem by assuming that the dependence structure among variables is well approximated by a conditional independence assumption. Within this conditional independence assumption, the equation (1.1) can be rewritten by specifying $K \times J_1$ distributions $f_k(x_j|\theta_{x,kj})$ and $K \times J_2$ distributions $f_k(y_j|\theta_{y,kj})$, each known up to the parameter vectors $\theta_{x,kj}, \theta_{y,kj}$, as a finite mixture of $J$-dimensional product densities, say

$$f(\boldsymbol{d}_i) = \sum_{k=1}^{K} \pi_k \Big( \prod_{j=1}^{J_1} f_k(x_{ij}|\theta_{x,kj}) \Big) \Big( \prod_{j=1}^{J_2} f_k(y_{ij}|\theta_{y,kj}) \Big), \tag{1.35}$$

where $f_k(x_{ij}|\theta_{x,kj})$ and $f_k(y_{ij}|\theta_{y,kj})$ denote the univariate conditional distributions of $X_j$ and $Y_j$ within the $k$th latent class. In the case of wind and wave data, circular univariate distributions must be used for the $J_1$ circular data (i.e. von Mises distribution) and skew univariate positive distributions for the $J_2$ linear data (i.e. Gamma or Weibull distributions).

In order to relax the conditional independence assumption, a Multivariate Mixture Model (MMM, Chapter 3) can be used to classify marine data. The joint distribution of multivariate linear-circular data would require the specification of densities that lie on a multi-dimensional hyper-cylinder. In order to avoid problems in the specification and identification of these type of densities, the joint distribution of hyper-cylindrical data can be approximated by the mixture of products of toroidal and planar densities, by specifying $K$ distributions $f_k(\boldsymbol{x}|\theta_{x,k})$ and $K$ distributions $f_k(\boldsymbol{y}|\theta_{y,k})$, each known up to the parameter vectors $\theta_{x,k}, \theta_{y,k}$. Equation 1.3 can be rewritten as follow:

$$f(\boldsymbol{d}_i) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i|\theta_{x,k}) f_k(\boldsymbol{y}_i|\theta_{y,k}), \tag{1.36}$$

where $f_k(\boldsymbol{x}_i|\theta_{x,k})$ and $f_k(\boldsymbol{y}_i|\theta_{y,k})$ denote the multivariate conditional distribu-

tion of $X$ and $Y$ within the $k$th latent class. A circular multivariate distribution must be used for circular data (i.e., a multivariate von Mises distribution) and a multivariate skew distribution for linear data (i.e., multivariate Gamma or multivariate skew normal distribution).

Both UMM and MMM start from the main hypothesis in mixture models that observed data are temporally independent. In Chapter 4 and 5 a multivariate hidden Markov model (MHMM) has been proposed for circular data (Chapter 4) and mixed linear-circular data (Chapter 5). Given a Markov chain, viewed as as multinomial process in discrete time $\boldsymbol{\xi}_{0:T} = (\boldsymbol{\xi}_t, t = 0, \ldots, T)$ with $\boldsymbol{\xi}_t = (\xi_{t1}, \ldots, \xi_{tK})$, and assuming that the observations are conditionally independent given a realization of the Markov chain, the conditional distribution of the observed process, given the latent process, is

$$f(\boldsymbol{d}_{0:T}|\boldsymbol{\xi}_{0:T}) = \prod_{t=0}^{T}\prod_{k=1}^{K}\left(f(\boldsymbol{x}_t|\theta_{x,k})f(\boldsymbol{y}_t|\theta_{y,k})\right)^{\xi_{tk}} \tag{1.37}$$

where $f_k(\boldsymbol{x}_t|\theta_{x,k})$ and $f_k(\boldsymbol{y}_t|\theta_{y,k})$ denote the multivariate conditional distribution of $X$ and $Y$ within the $k$th latent states as seen in MMM.

# Bibliography

[1] J.D. Banfield and A.E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49 (3):803–821, 1993.

[2] J.P. Baudry, A.E. Raftery, G. Celeux, K. Lo, and R. Gottardo. Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, 19 (2):332–352, 2010.

[3] L. Bertotti and L. Cavaleri. Wind and wave predictions in the adriatic sea. *Journal of Marine Systems*, 78:227–234, 2009.

[4] C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics and Data Analysis*, 41:561–575, 2003.

[5] J. Bulla and A. Berzel. Computational issues in parameter estimation for stationary hidden markov models. *Computational Statistics*, 23(1):1–18, January 2008.

[6] A.W. Drake. *Fundamentals of applied probability theory.* McGraw-Hill series in probability and statistics. McGraw-Hill, 1967.

[7] B. Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

[8] B. Efron and R. Tibshirani. *An introduction to the Bootstrap*. Chapman and Hall, 1993.

[9] C. Fraley and A.E. Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 1998.

[10] C. Fraley and A.E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of American Statistical Association*, 97:611–631, 2002.

[11] Y.K. Gu, D.L. Ge, and Y.G. Xiong. A reliability data analysis method using mixture weibull distribution model. *Applied Mechanics and Materials*, 2012.

[12] C. Hennig. Methods for merging gaussian mixture components. *Advances in Data Analysis and Classification*, 2010.

[13] L. Hunt and M. Jorgensen. Clustering mixed data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):352–361, 2011.

[14] S. Ingrassia and R. Rocci. Degeneracy of the em algorithm for the mle of multivariate gaussian mixtures and dynamic constraints. *Computational Statistics and Data Analysis*, 55:1715–1725, 2011.

[15] S.R. Jammalamadaka and A. SenGupta. *Topics in Circular Statistics*. Series on Multivariate Analysis Series. World Scientific Publishing Company, 2001.

[16] D. Karlis and A. Santourian. Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing*, 2009.

[17] C.J. Lawrence and W.J. Krzanowski. Mixture separation for mixed-mode data. *Statistics and Computing*, 1996.

[18] B.G. Leroux. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20 (3):1350–1360, 1992.

[19] T.I. Lin. Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis*, 2009.

[20] T.I. Lin and J.C. Lee. A robust approach to t linear mixed models applied to multiple sclerosis data. *Statistics in medicine*, 2006.

[21] T.I. Lin, J.C. Lee, and W.J. Hsieh. Robust mixture modeling using the skew t distribution. *Statistics and Computing*, 2007.

[22] T. A. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society, Series B*, 44:226–233, 1982.

[23] K.V. Mardia and P.E. Jupp. *Directional statistics*. John Wiley, 2010.

[24] K.V. Mardia, C.C. Taylor, and G.K. Subramaniam. Protein bioinformatics and mixtures of bivariate von mises distributions for angular data. *Biometrics*, 2007.

[25] G. McLachlan and D. Peel. *Finite mixture models*. John Wiley, 2000.

[26] K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, 1894.

[27] D. Peel. Mixture model clustering and related topics. Unplublished PhD thesis, University of Queensland, Brisbane.

[28] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 1989.

[29] J.L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapmanan and Hall London, 1997.

[30] W. Seidel, K. Mosler, and M. Alker. A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics*, 52 (3):481–487, 2000.

[31] M. Wiper, D.R. Insua, and F. Ruggeri. Mixtures of gamma distributions with applications. *Journal of Computational and Graphical Statistics*, 2001.

[32] W. Zucchini and I.L. MacDonald. *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor and Francis, 2009.

Journal of Data Science **9**(2011), 585-605

# A Latent-Class Model for Clustering Incomplete Linear and Circular Data in Marine Studies

Francesco Lagona* and Marco Picone
*Roma Tre University*

*Abstract*: Identification of representative regimes of wave height and direction under different wind conditions is complicated by issues that relate to the specification of the joint distribution of variables that are defined on linear and circular supports and the occurrence of missing values. We take a latent-class approach and jointly model wave and wind data by a finite mixture of conditionally independent Gamma and von Mises distributions. Maximum-likelihood estimates of parameters are obtained by exploiting a suitable EM algorithm that allows for missing data. The proposed model is validated on hourly marine data obtained from a buoy and two tide gauges in the Adriatic Sea.

*Key words*: Circular data, cross-validation, EM algorithm, Gamma distribution, latent classes, marine data, missing values, Von Mises distribution.

## 1. Introduction

Wave regimes are specific shapes that the distribution of wave attributes (such as wave height and direction) takes under latent environmental conditions. The identification of relevant regimes in a particular area is often necessary to estimate the drift of floating objects and oil spills (Huang *et al*, 2011), in the design of offshore structures (Faltinsen, 1990) and in studies of sediment transport (Jin and Ji, 2004) and coastal erosion (Pleskachevsky *et al.*, 2009). The description of wave data in terms of regimes is also useful in the analysis of coastal areas and enclosed seas, where numerical wind-wave models, traditionally used for ocean waves, can give inaccurate results (Bertotti and Cavalieri, 2004). For these reasons, the Assembly of the International Maritime Organization has repeatedly encouraged the publication of wave data atlas that include a description of representative wave regimes in specific areas, characterized by probability of occurrence, and corresponding to dominant environmental conditions (e.g., wind conditions) over

*Corresponding author.

the area of interest. This has motivated an increasing interest in methods for clustering wave data according to a finite number of regimes.

Traditionally, techniques of wave data clustering are based on distance-based methods. Recent proposals require the use of a finite number of target distributions, defined as cluster centroids, and an optimization algorithm that associates the observed data to the closest centroid (Boukhanovsky *et al.*, 2007). Hierarchical agglomerative clustering methods (Hamilton, 2010) have been also suggested to avoid the specification of a family of target distributions.

The limitations of distance-based based methods are well known (Fraley and Raftery, 2002). The statistical properties of these methods are generally unknown, precluding the possibility of formal inference on the clustering results. This is a critical issue in marine studies, because the identification of wave regimes without a measure of the statistical uncertainty of regime-specific parameters is of little practical use. In addition, there is little systematic guidance associated with distance-based methods for solving basic questions that arise in cluster analysis, such as the choice of an optimal number of clusters and the choice of an optimal clustering algorithm.

A general framework to address these issues is provided by latent-class models (Hagenaars and McCutcheon, 2002), which cluster multivariate data according to a finite number of classes, approximating the joint distribution of the data by a mixture of parametric densities, which represent the distributional shape of the data within each cluster. From a methodological viewpoint, a latent-class approach allows to solve the clustering problem as a missing value problem, by treating the unknown cluster membership of each observation as a missing value, to be estimated from the data. From a technical viewpoint, the clustering algorithm reduces to likelihood maximization and the choice of the optimal number of clusters reduces to a model selection problem in parametric inference.

In this paper we take a latent-class approach to describe sea conditions in terms of wave regimes, by clustering multivariate environmental profiles in a finite number of classes. Specifically, we model the data by a mixture of product densities, i.e. a particular latent-class model where the observed variables are assumed conditionally independent, given a latent multinomial variable. This model is tailored to identify wave regimes in practical settings that often arises in marine studies, where (1) environmental profiles include measurements taken on linear and circular supports and (2) some of these observations are missing, due to malfunctioning of the devices that provide the data.

While there is an extensive literature on modelling multivariate continuous, categorical and mixed continuous-categorical variables by multivariate normal models, log-linear models or a combination of both, the joint modelling of variables on linear and circular supports is still an open area of research. Recent

attempts include multivariate circular distributions defined on toroidal supports (Mardia *et al.*, 2008), distributions on cylinders that are based on nonnegative trigonometric sums (Fernández-Durán, 2007) and multivariate distributions with specified marginals on cylinders, discs and tori (Kato and Shimizu, 2008). When however the goal of an analysis is the identification of typical wave regimes, the specification of the joint distribution of marine variables should aim at clustering the data according to a finite number of classes in a way that the dependence structure between the data is well approximated by this partitioning of the sample. Mixtures of product densities provide such clustering of the data and flexibly accommodate for the mixed supports on which linear and circular data are taken. Moreover, the semi-parametric nature of the model allows for a parsimonious specification of the association structure between linear and circular measurements, which is of great help in marine studies, where too little is often known about the data generating process to assume a fully parametric specification.

Wave regimes identification is additionally complicated by the occurrence of missing values. Marine databases are often incomplete because of device malfunctioning or maintenance-related reasons. For mixture-based data clustering, maximum-likelihood estimation could be carried out by discarding incomplete data profiles from the sample and using the complete cases to build up the likelihood function to be maximized (CC; complete case 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; Rubin, 1987), CC-based maximum-likelihood estimation is known to be (asymptotically) unbiased but inefficient (Rotnitzky and Wypij, 1994). Loss of efficiency is due to the fact that incomplete profiles are informative of the parameters of the joint distribution of several variables, especially when these variables are strongly correlated. Efficient maximum-likelihood estimation from MAR multivariate data often requires data-augmentation or multiple-imputation methods (Shafer, 1997). Mixture of product densities, instead, can be efficiently estimated by including both complete and incomplete profiles into the likelihood, because likelihood contributions of incomplete profiles are available in closed form and data-augmentation/imputation methods are not necessary.

Mixtures of product densities have been already suggested in the statistical literature to cluster multivariate categorical data (Vermunt *et al.*, 2008) and mixed linear and categorical data (Hunt and Jorgensen, 2003) in the presence of missing values. From a technical viewpoint, therefore, our application extends this strand of literature to the case of linear and circular data with missing values. On the methodological side, our proposal is an alternative to the existing distance-based methods for wave regime identification, with three practical advantages. First,

it is based on an EM algorithm that is less computationally demanding than the algorithms currently in use for distance-based identification of wave regimes. Second, missing values are efficiently handled, while distance-based methods normally require complete data information. Third, while formal inference is not possible with a distance-based approach to clustering, mixture-based clustering is carried out within a parametric inferential framework and, as a result, it can be validated by using traditional methods of parametric inference.

Relevant details on the data that motivated this work are presented in Section 2, while Section 3 is devoted to maximum-likelihood estimation of mixture of product densities in the case of missing observations. In Section 4 we specify the Gamma-von Mises latent-class model that was exploited to examine the data presented in Section 2. Estimation and model validation results are summarized in Section 5. Relevant points of discussion are listed in Section 6.

## 2. Data

The Italian Institute for Environmental Research and Protection (ISPRA; www.isprambiente.it) maintains a network of buoys to monitor wave direction and height at various points of the Italian seas. A network of ISPRA tide gauges, located along the Italian coast, additionally provide data about wind direction and speed.

The data that we have exploited in this work include hourly measurements of wave height and direction, taken in the period 11/18/2002-01/17/2003 by the buoy of Ancona, which is located in the Adriatic sea at about 30 Km from the coast (Figure 1). During the same period, hourly data on wind speed and direction were obtained from the two nearest tide gauges, respectively located at Ancona (about 30 Km from the buoy) and at Ravenna (about 120 Km from the buoy). 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 eight hours.

Table 1 reports the percentages of missing data observed during the study period. Measurements taken by buoys and tide gauges can be missing because of devices maintenance or discontinuous functioning. Occurrence of missing values on wave measurements is more frequent than the occurrence of missing wind data because buoys are more exposed to transmission errors than tide gauges. We remark that our data are in the form of hourly profiles of six observations. As a result, different patterns of missing values occur: while about the 28% of the data profiles include at least one missing value, the modal missingness pattern (15.3%) includes a missing circular and a missing linear variable. During the study period, there is a very small portion (about 0.1%) of hourly profiles with no information.

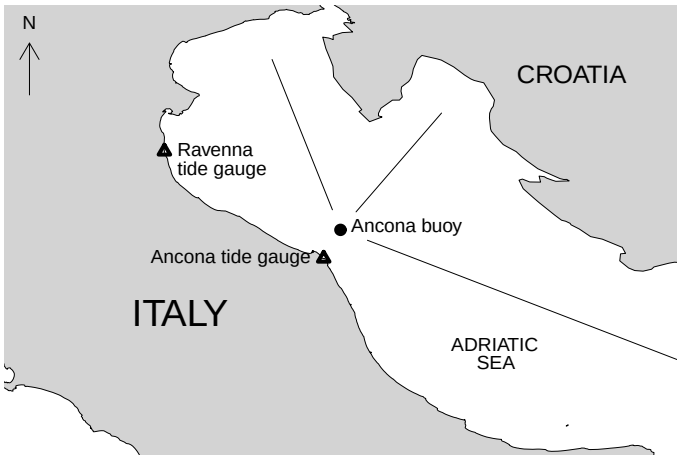


Figure 1: Locations of the buoy and the two tide gauges, from which the data displayed in Figure 2 were obtained; segments indicate the three directions of maximal fetch, i.e. the distance between the buoy and the closest coastline

Table 1: Percentages of missing values

| Site | Measurement (Unit) | Percentages |
|---|---|---|
| Ancona (buoy) | Wave Height (Meters) | 16.3% |
| | Wave Direction (Radians) | 16.3% |
| Ancona (tide gauge) | Wind Speed (Meter/Sec) | 1.1% |
| | Wind Direction (Radians) | 2.2% |
| Ravenna (tide gauge) | Wind Speed(Meter/Sec) | 10.4% |
| | Wind Direction (Radians) | 2.3% |

Univariate distributions of the available data are displayed in Figure 2. Rose diagrams indicate the distribution of directions from which the wind and the wave come from. As expected, waves mostly come from two modal directions (south-east and north-east), which relate to two of the three angles at which the distance between the buoy and the nearest coast (fetch) is maximum (Figure 1). Waves from North-West (along the third maximum-fetch direction) are rarely observed in wintertime, because winter winds do not typically blow from this direction. As displayed by the circular wind distributions at the two tide gauges of Ancona and Ravenna, two are the winds that dominate the Adriatic Sea in wintertime: bora, a typical cold wind, blowing from West/North-West, and Sirocco, blowing from South-East, and responsible for the storm surges in the northern part of the Adriatic sea, and hence for the famous floods of Venice.

The histograms on the right side of Figure 2 show the distributions of wave height, as observed at the buoy of Ancona, and wind speed, as measured at the

two tide gauges of Ancona and Ravenna. The multi-modal shape of these distributions is less apparent than that displayed by directional data. This is typical of wave and wind data that are observed in enclosed seas, such as the Adriatic, where the geometry of the coastline makes it difficult to separate components of dominant wind speeds and wave heights and is responsible for the inaccurate results provided by numerical wind-wave models that are normally used for modelling ocean waves.

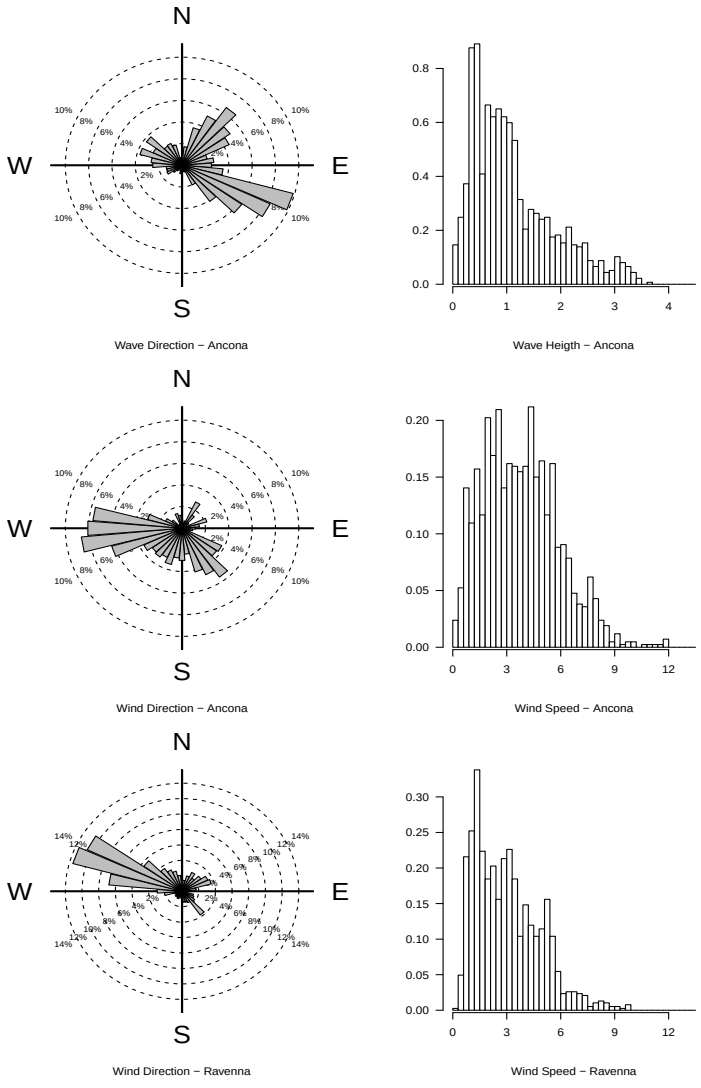


Figure 2: Distribution of the available wave metric data at the buoy of Ancona and wind data at the two nearest tide gauges (Ancona and Ravenna)

## 3. Estimation of Mixtures of Product Densities from Incomplete Mixed Data

The multivariate data described in Section 2 can be represented as $n$ vectors $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iJ})$, $i = 1, \cdots, n$, drawn from the multivariate distribution of $J$ variables $Y_j, j = 1, \cdots, J$, measured on different supports (e.g., linear or circular). We assume that these vectors can be clustered into $K$ groups (or classes) and that the association structure between the variables $Y_j$ is well approximated by this partitioning of the sample. Formally, we introduce a latent (unobserved) multinomial random vector $\boldsymbol{Z} = (Z_1, \cdots, Z_K)$ with one trial and cell probabilities $(\pi_1, \cdots, \pi_K)$, and assume that the $J$ variables $Y_j$ are conditionally independent given $Z$. Within this conditional independence assumption, we specify $K \times J$ distributions $f_k(y|\boldsymbol{\beta}_{kj})$, each known up to a parameter vector $\boldsymbol{\beta}_{kj}$, and model the multivariate distribution of vector $\boldsymbol{y}_i$ as a finite mixture of $J$-dimensional product densities, say

$$f(\boldsymbol{y}_i) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} f_k(y_{ij}|\boldsymbol{\beta}_{kj}), \tag{3.1}$$

where $f_k(y|\boldsymbol{\beta}_{kj})$ denotes the conditional distribution of $Y_j$ within the $k$th latent class. We observe that (3.1) specifies a multivariate distribution without imposing consistency constraints on the conditional densities $f_k(y|\boldsymbol{\beta}_{kj})$, which, hence, do not necessarily need to be member of the same parametric family. This flexibility is of great help in the modelling of mixed linear and circular data. Given the number $K$ of classes, mixtures of product densities are furthermore strictly identifiable, provided that the densities $f_k(\boldsymbol{y}) = \prod_{j=1}^{J} f_k(y_j|\boldsymbol{\beta}_{kj})$ are linearly independent (Teicher, 1967; Yakowitz and Spragins, 1968).

Mixture (3.1) is a particular latent-class model and is often presented in the literature as a model-based alternative to the traditional cluster-analysis methods that are based on distance-based procedures, such as hierarchical agglomerative clustering or iterative relocation procedures. Typically exploited in social science studies and marketing research, mixtures of product densities such as (3.1) have been successfully implemented in the classification of mixed profiles that include quantitative (continuous or discrete) and categorical (nominal or ordinal) observations.

Maximum-likelihood estimation of a mixture model is normally based on an EM algorithm. Hunt and Jorgensen (2003) developed an EM algorithm for estimating latent-class models from MAR data, in the case of mixed multi-normal and categorical data. In the case of mixtures of product densities, such as (3.1), their algorithm can be greatly simplified as follows.

We account for the occurrence of missing values by splitting the complete

data vector $\boldsymbol{y}_i = (\boldsymbol{y}_{O(i)}, \boldsymbol{y}_{M(i)})$ into a vector $\boldsymbol{y}_{O(i)}$ of observed data and a vector $\boldsymbol{y}_{M(i)}$ of missing values, $O(i) \cup M(i) = \{1, \cdots, J\}$. We furthermore introduce a $n \times J$ matrix $R$, whose generic component $r_{ij} = 1$ if $y_{ij}$ is missing and 0 otherwise. Accordingly, the row-sums of $R$, say $r_{i\cdot} = \sum_{j=1}^{J} r_{ij}$, indicate the number of missing values within each $i$th profile.

If the data are MAR, i.e. the probability of a missing value does not depend on the value that is missing, maximum likelihood estimates of model (3.1) can be found by maximizing the marginal log-likelihood function

$$\begin{aligned} l(\boldsymbol{\beta}, \pi) &= \sum_{i=1}^{n} \log \int_{\boldsymbol{y}_{M(i)}} \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} f_k(y_{ij}|\boldsymbol{\beta}_{kj}) d\boldsymbol{y}_{M(i)} \\ &= \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} (f_k(y_{ij}|\boldsymbol{\beta}_{kj}))^{1-r_{ij}} \\ &= \sum_{i:r_{i\cdot}=0} \log \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} f_k(y_{ij}|\boldsymbol{\beta}_{kj}) + \sum_{i:r_{i\cdot}>0} \log \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} (f_k(y_{ij}|\boldsymbol{\beta}_{kj}))^{1-r_{ij}} \\ &= l_{CC}(\boldsymbol{\beta}, \pi) + l_{IC}(\boldsymbol{\beta}, \pi), \end{aligned} \tag{3.2}$$

which is the sum of the log-likelihood contributions of the complete (CC) and incomplete cases (IC). We observe that the log-likelihood contribution of a completely missing profile, i.e. such that $r_{i\cdot} = J$, is given by $\log \sum_k \pi_k = 0$. Under a CC strategy, the log-likelihood contribution $l_{IC}$ is ignored, leading to inefficient estimates.

Local maximum points of the log-likelihood (3.2) can be found by an EM algorithm (Dempster, Laird and Rubin, 1977) that iteratively maximizes the expectation of the complete data log-likelihood function. In the case of MAR data drawn from a mixture of product densities, the complete log-likelihood can be written as follows

$$l_{\text{comp}}(\boldsymbol{\beta}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left( \log \pi_k + \sum_{j=1}^{J} (1 - r_{ij}) \log f_k(y_{ij}|\boldsymbol{\beta}_{jk}) \right), \tag{3.3}$$

where $(z_{i1}, \cdots, z_{ik})$ is the $i$th realization of the multinomial random variable $Z$. At the $h$th step of the algorithm, the expectation of $l_{\text{comp}}(\boldsymbol{\beta}, \boldsymbol{\pi})$ with respect to the conditional distribution $p(Z|\boldsymbol{y})$ is computed on the basis of the estimates

$\hat{\boldsymbol{\beta}}^{(h-1)}$ and $\hat{\boldsymbol{\pi}}^{(h-1)}$, obtained at the previous iteration, by evaluating (E-step)

$$
\begin{aligned}
&Q(\boldsymbol{\pi}, \boldsymbol{\beta}|\hat{\boldsymbol{\pi}}^{(h-1)}, \hat{\boldsymbol{\beta}}^{(h-1)}) = \mathbb{E}(l_{\text{comp}}(\boldsymbol{\beta}, \boldsymbol{\pi})) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\left(\log \pi_k + \sum_{j=1}^{J}(1-r_{ij})\log f_k(y_{ij}|\boldsymbol{\beta}_{jk})\right)\mathbb{E}(z_{ik}|\hat{\boldsymbol{\beta}}^{(h-1)}, \hat{\boldsymbol{\pi}}^{(h-1)}) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{\pi}_{ik}^{(h-1)}\log \pi_k + \sum_{j=1}^{J}\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}^{(h-1)}(1-r_{ij})\log f_k\left(y_{ij}|\boldsymbol{\beta}_{jk}\right) \\
&= Q\big(\boldsymbol{\pi}|\hat{\beta}^{(h-1)}, \hat{\pi}^{(h-1)}\big) + \sum_{j=1}^{J} Q_j\big(\boldsymbol{\beta}_{kj}|\hat{\beta}^{(h-1)}, \hat{\pi}^{(h-1)}\big),
\end{aligned}
\tag{3.4}
$$

where

$$
\begin{aligned}
\hat{\pi}_{ik}^{(h-1)} &= \mathbb{E}(z_{ik}|\hat{\boldsymbol{\beta}}^{(h-1)}, \hat{\boldsymbol{\pi}}^{(h-1)}) \\
&= \frac{\int_{\boldsymbol{y}_{M(i)}} \hat{\pi}_k^{(h-1)} \prod_{j=1}^{J} f_k(y_{ij}|\hat{\boldsymbol{\beta}}_{kj}^{(h-1)}) d\boldsymbol{y}_{M(i)}}{\int_{\boldsymbol{y}_{M(i)}} \sum_{k=1}^{K} \hat{\pi}_k^{(h-1)} \prod_{j=1}^{J} f_k(y_{ij}|\hat{\boldsymbol{\beta}}_{kj}^{(h-1)}) d\boldsymbol{y}_{M(i)}} \\
&= \frac{\hat{\pi}_k^{(h-1)} \prod_{j=1}^{J} \left(f_k(y_{ij}|\hat{\boldsymbol{\beta}}_{kj}^{(h-1)})\right)^{1-r_{ij}}}{\sum_{k=1}^{K} \hat{\pi}_k^{(h-1)} \prod_{j=1}^{J} \left(f_k(y_{ij}|\hat{\boldsymbol{\beta}}_{kj}^{(h-1)})\right)^{1-r_{ij}}}
\end{aligned}
\tag{3.5}
$$

indicates the conditional probability of vector $\boldsymbol{y}_{O(i)}$ to belong to the $k$th latent class. The previous E-step is followed by an M-step where vector $(\hat{\boldsymbol{\beta}}^{(h-1)}, \hat{\boldsymbol{\pi}}^{(h-1)})$ is updated by a new vector $(\hat{\boldsymbol{\beta}}^{(h)}, \hat{\boldsymbol{\pi}}^{(h)})$ that maximizes the expected log-likelihood (3.4). We observe that (3.4) is the sum of $J+1$ functions, which depend on independent sets of parameters, and, as a result, the M-step can be carried out by separately solving $J+1$ maximization problems. In particular, the maximum point of $Q(\boldsymbol{\pi}|\hat{\beta}^{(h-1)}, \hat{\pi}^{(h-1)})$ is available in closed form and it is equal to

$$
\hat{\pi}_k^{(h)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_{ik}^{(h-1)}.
$$

The form of the updating equations for parameters $\boldsymbol{\beta}$ that maximize the remaining $J$ functions $Q_j(\boldsymbol{\beta}_{kj}|\hat{\beta}^{(h-1)}, \hat{\pi}^{(h-1)})$ depend on the form of the densities $f_k(y_j|\boldsymbol{\beta}_{kj})$. In Section 4 we derive these updates under Gamma and von Mises densities.

The algorithm alternates the E-step and the M-step up to convergence of the estimates, whose limit (Wu, 1983) is a local maximum point of the likelihood function (3.2).

## 4. A Gamma-Von Mises Latent-Class Model

The $J = 6$ variables of our case study can be clustered in two groups according to the scale on which they are measured. A first group includes three circular variables, say $Y_1$ (wave direction at the Ancona buoy), $Y_2$ (wind direction at the Ancona tide gauge) and $Y_3$ (wind direction at the Ravenna tide gauge). A second group includes three variables on a linear support, say $Y_4$ (wave height at the Ancona buoy), $Y_5$ (wind speed at the Ancona tide gauge) and $Y_6$ (wind speed at the Ravenna tide gauge).

The mixture model presented in Section 3 allows for a flexible choice of the univariate distributions that can be placed within each latent class.

We have decided to model wave and wind directions by exploiting three von Mises distributions, i.e.

$$f_k(y|\boldsymbol{\beta}_{kj}) = \mathrm{VM}(\beta_{kj0}, \beta_{kj1}) = \frac{\exp(\beta_{kj1}\cos(y - \beta_{kj0}))}{2\pi I_0(\beta_{kj1})}, \quad j = 1, 2, 3, \tag{4.1}$$

where the parameters $\beta_{kj0}$ and $\beta_{kj1}$, $j = 1, 2, 3$, respectively indicate the mean (or modal) direction and the concentration of each conditional circular distribution, given the $k$th latent class, and $I_0$ is the modified Bessel function of order 0.

Wave height at the buoy and wind speeds at the two tide gauges have been instead modeled by three Gamma distributions, i.e.

$$f_k(y|\boldsymbol{\beta}_{kj}) = \mathrm{Gam}(\beta_{kj0}, \beta_{kj1}) = \frac{\beta_{kj0}^{\beta_{kj1}} y^{\beta_{kj1}-1} \exp -(y/\beta_{kj0})}{\Gamma(\beta_{kj1})}, \quad j = 4, 5, 6, \tag{4.2}$$

where parameters $\beta_{kj0}$ and $\beta_{kj1}$, $j = 4, 5, 6$, respectively indicate the scale and shape of the conditional distributions, given the latent class.

Under the above distributional assumptions, the mixture of product densities

$$f(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} f_k(y_j|\boldsymbol{\beta}_{kj})$$

is a multivariate distribution on a six-dimensional hyper-cylinder. According to the sufficient conditions stated by Teicher (1967) and Yakowitz and Spragins (1968), identifiability of this mixture follows by the linear independence of the families of the Gamma and the von Mises densities. Moreover, the marginal distribution of each variable on a linear support is approximated by a mixture of $K$ Gamma densities and the marginal distribution of each circular variable is approximated by a mixture of $K$ von Mises densities. As a result, the $J$-dimensional profiles of wave and wind data ($J = 6$ in our application) are clustered according to $K$ wind-wave regimes. Because von Mises and Gamma densities are known

up to 2 parameters, each regime is defined on the basis of $2J$ parameters, which indicate not only class-specific modal directions of waves and winds and their average heights and speeds, but also the amount of variation of the circular and linear measurements around these means. In particular, the association between each variable and the remaining variables is semi-parametrically described by conditional densities that take the following mixture form

$$f(y_j|y_l, l \neq j; \boldsymbol{\beta}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \frac{\pi_k \prod_{h \neq j} f_k(y_h|\boldsymbol{\beta}_{hk})}{\sum_{k=1}^{K} \pi_k \prod_{h \neq j} f_k(y_h|\boldsymbol{\beta}_{hk})} f_k(y_j|\boldsymbol{\beta}_k j). \tag{4.3}$$

Class-specific parameters of the above Gamma-von Mises mixture model can be separately updated by the EM algorithm within the M-step. In particular, standard derivative computations show that contributions to the expected log-likelihood function given by the circular data, $Q_j(\boldsymbol{\beta}_j|\hat{\boldsymbol{\pi}}^{(h-1)}, \hat{\boldsymbol{\beta}}^{(h-1)}), j = 1, 2, 3$ are separately maximized by

- an update of the modal directions, given by

$$\hat{\beta}_{kj0}^{(h)} = \operatorname{arctg} \frac{\sum_{i=1}^{n} (1 - r_{ij}) \hat{\pi}_{ik}^{(h-1)} \sin y_{ij}}{\sum_{i=1}^{n} (1 - r_{ij}) \hat{\pi}_{ik}^{(h-1)} \cos y_{ij}}, \tag{4.4}$$

- and by the roots $\hat{\beta}_{kj1}^{(h)}$ of the three equations

$$\frac{I_0(\beta_{kj1})}{I_0'(\beta_{kj1})} = \frac{\sum_{i=1}^{n} (1 - r_{ij}) \hat{\pi}_{ik}^{(h-1)} \cos(y_{ij} - \hat{\beta}_{kj0}^{(h)})}{\sum_{i=1}^{n} (1 - r_{ij}) \hat{\pi}_{ik}^{(h-1)}}, \tag{4.5}$$

  which are the updated concentrations of wave and wind directions on the circle.

Analogous derivative computations show that the remaining three functions $Q_j\big(\boldsymbol{\beta}_j|\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\beta}}\big), j = 4, 5, 6$, i.e. the contributions of the linear data to the expected log-likelihood function, are separately maximized by

- an update of the shape parameters, given by

$$\hat{\beta}_{kj0}^{(h)} = \frac{\sum_{i=1}^{n} (1 - r_{ij}) \hat{\pi}_{ik}^{(h-1)} y_{ij}}{\sum_{i=1}^{n} (1 - r_{ij}) \hat{\pi}_{ik}^{(h-1)}},$$

- and by the roots $\hat{\beta}_{kj1}^{(h)}$ of the three equations

$$\begin{aligned} \log(\beta_{kj1}) - \psi(\beta_{kj1}) = & \log \left( \frac{\sum_{i=1}^{n} (1 - r_{ij}) \hat{\pi}_{ik}^{(h-1)} y_{ij}}{\sum_{i=1}^{n} \hat{\pi}_{ik}^{(h-1)}} \right) \\ & - \left( \frac{\sum_{i=1}^{n} (1 - r_{ij}) \hat{\pi}_{ik}^{(h-1)} \log y_{ij}}{\sum_{i=1}^{n} \hat{\pi}_{ik}^{(h-1)}} \right), \end{aligned}$$

where $\psi(\beta_{kj1})$ is the Digamma function.

## 5. Results and Model Validation

The proposed model was estimated from the data illustrated in Section 2, by considering $K = 4, \cdots, 10$ classes. According to the BIC criterion, a model with $K = 7$ classes is needed to adequately describe the data (results not reported here and available upon request to the corresponding author).

Table 2 displays the maximum likelihood estimates and the standard errors of the $7 \times 12 + 7 = 91$ parameters of the model with minimum BIC. The last row of the table indicate the estimated class probabilities $\hat{\boldsymbol{\pi}}$. While point estimates were computed by exploiting the EM algorithm of Section 3, standard errors were computed by taking the square root of the diagonal elements of the inverse observed information matrix, obtained by extracting the observed information from the complete log-likelihood (Louis, 1982). These estimates (all significant at a 95% significance level) can be directly exploited for a variety of applications that include for example the computation of the expected wave load to ships and off-shore structures. In addition, these estimates have an immediate physical interpretation, which can be summarized with the help of Figure 3, which displays the $6 \times 7$ densities that have been estimated under model (3.1). To draw this picture, we have used seven different colors (listed in Table 2) to show the grouping of the conditional densities according to the seven latent classes. Latent classes can be interpreted with the help of the map in Figure 1. Components 1, 2 and 7 cluster S-E waves of high (comp. 1), medium (comp. 2) and low (comp. 7) average heights, respectively. As expected, components 1, 2 and 7 are respectively associated with Sirocco winds of high, medium and low speed at both the tide gauges considered for analysis. Components 3 and 5 cluster N-W waves of medium (comp. 3) and low (comp. 5) height, associated with Bora winds of medium (comp. 3) and low (comp. 5) speed, blowing from west and north-west at the two tide gauges. Components 4 and 6 cluster waves with a direction that is perpendicular to the coast (coastal waves) and, as expected, are of moderate/medium heights. However, while waves within latent class 4 are associated with winds blowing along the same direction as waves, waves within latent class 6 are associated with winds coming from north. We note that the occurrence of coastal waves of moderate heights, regardless of wind and speed direction, is responsible for numerical wind-wave models giving inaccurate results in coastal areas. Our mixture model correctly separates coastal waves and wind-generated waves moving along maximal fetch directions. The results additionally suggest that regimes that generate coastal waves cannot be ignored in the analysis of sea conditions, because the probability of occurrence of classes 4 and 6 is about 0.22. We also remark that the model seems able to separate regimes that drive

severe and moderate conditions of the sea. Component 1 detects the distributional shape of wave height and direction of sea storms and identifies the wind conditions under which this event occur.

Table 2: Parameter estimates and standard errors (within brackets)

| | Parameters | Component 1 (red) | 2 (blue) | 3 (green) | 4 (yellow) | 5 (cyan) | 6 (magenta) | 7 (orange) |
|---|---|---|---|---|---|---|---|---|
| Wave Dir [a] (radians) | mean | 1.935 | 2.102 | 5.927 | 0.820 | 5.097 | 0.821 | 2.409 |
| | | (0.013) | (0.020) | (0.076) | (0.035) | (0.051) | (0.036) | (0.100) |
| | concentration | 99.278 | 16.778 | 1.215 | 9.377 | 6.951 | 6.269 | 0.898 |
| | | (17.476) | (1.862) | (0.108) | (1.343) | (1.341) | (0.784) | (0.116) |
| Wind Dir [b] (radians) | mean | 2.339 | 2.864 | 4.632 | 1.195 | 4.534 | 5.701 | 3.411 |
| | | (0.022) | (0.049) | (0.018) | (0.058) | (0.052) | (0.103) | (0.063) |
| | concentration | 30.995 | 3.122 | 11.111 | 3.181 | 7.057 | 1.240 | 1.656 |
| | | (5.223) | (0.335) | (0.964) | (0.447) | (1.626) | (0.139) | (0.122) |
| Wind Dir [c] (radians) | mean | 2.305 | 2.697 | 5.103 | 1.065 | 5.322 | 5.942 | 5.319 |
| | | (0.022) | (0.169) | (0.012) | (0.020) | (0.034) | (0.064) | (0.082) |
| | concentration | 33.105 | 0.632 | 24.259 | 24.716 | 13.778 | 2.467 | 1.064 |
| | | (6.116) | (0.130) | (2.357) | (3.908) | (2.543) | (0.241) | (0.110) |
| Wave Height [a] (meters) | shape | 99.226 | 5.782 | 12.556 | 35.778 | 26.081 | 10.147 | 3.031 |
| | | (20.640) | (0.701) | (1.148) | (5.125) | (5.116) | (1.360) | (0.247) |
| | scale | 0.029 | 0.174 | 0.078 | 0.055 | 0.014 | 0.179 | 0.146 |
| | | (0.006) | (0.021) | (0.007) | (0.008) | (0.003) | (0.024) | (0.014) |
| Wind Speed [b] (meters/sec) | shape | 10.372 | 5.140 | 11.779 | 9.358 | 9.768 | 7.685 | 2.276 |
| | | (1.756) | (0.592) | (0.915) | (1.289) | (1.630) | (1.013) | (0.177) |
| | scale | 0.517 | 0.628 | 0.420 | 0.386 | 0.192 | 0.771 | 0.896 |
| | | (0.089) | (0.070) | (0.033) | (0.056) | (0.033) | (0.099) | (0.084) |
| Wind Speed [c] (meters/sec) | shape | 22.656 | 4.497 | 10.517 | 12.732 | 13.004 | 4.159 | 6.232 |
| | | (4.712) | (0.586) | (0.817) | (1.683) | (2.354) | (0.452) | (0.577) |
| | scale | 0.223 | 0.626 | 0.296 | 0.448 | 0.083 | 0.969 | 0.210 |
| | | (0.045) | (0.077) | (0.023) | (0.060) | (0.015) | (0.110) | (0.021) |
| | probability | 0.053 | 0.166 | 0.250 | 0.087 | 0.060 | 0.135 | 0.249 |
| | | (0.006) | (0.013) | (0.012) | (0.008) | (0.007) | (0.010) | (0.015) |

[a] Ancona buoy - [b] Ancona tide gauge - [c] Ravenna tide gauge

Figures 4 and 5 display the classification of the multivariate profiles, as obtained by modal allocation, i.e. assigning each profile $i$ to the latent class $k$ with the highest probability $\hat{\pi}_{ik}$. Fiducial intervals for each single observation $y_{ij}$ were obtained on the basis of the estimated conditional distribution (5.1) whose expectation

$$\mathbb{E}(y_{ij}|y_{il}, l \neq j; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}) = \sum_{k=1}^{K} \frac{\hat{\pi}_k \prod_{h \neq j} f_k(y_{ih}|\hat{\boldsymbol{\beta}}_{hk})}{\sum_{k=1}^{K} \hat{\pi}_k \prod_{h \neq j} f_k(y_{ih}|\hat{\boldsymbol{\beta}}_{hk})} \mathbb{E}_k(y_{ij}|\hat{\boldsymbol{\beta}}_k j) \tag{5.1}$$

was exploited to impute missing values (the black dots in Figures 4 and 5). The model gives an adequate fit of the observed data (right-hand histograms in Figures

4 and 5 and, simultaneously, operates an intuitively appealing classification of complete and incomplete profiles of wind and wave measurements.

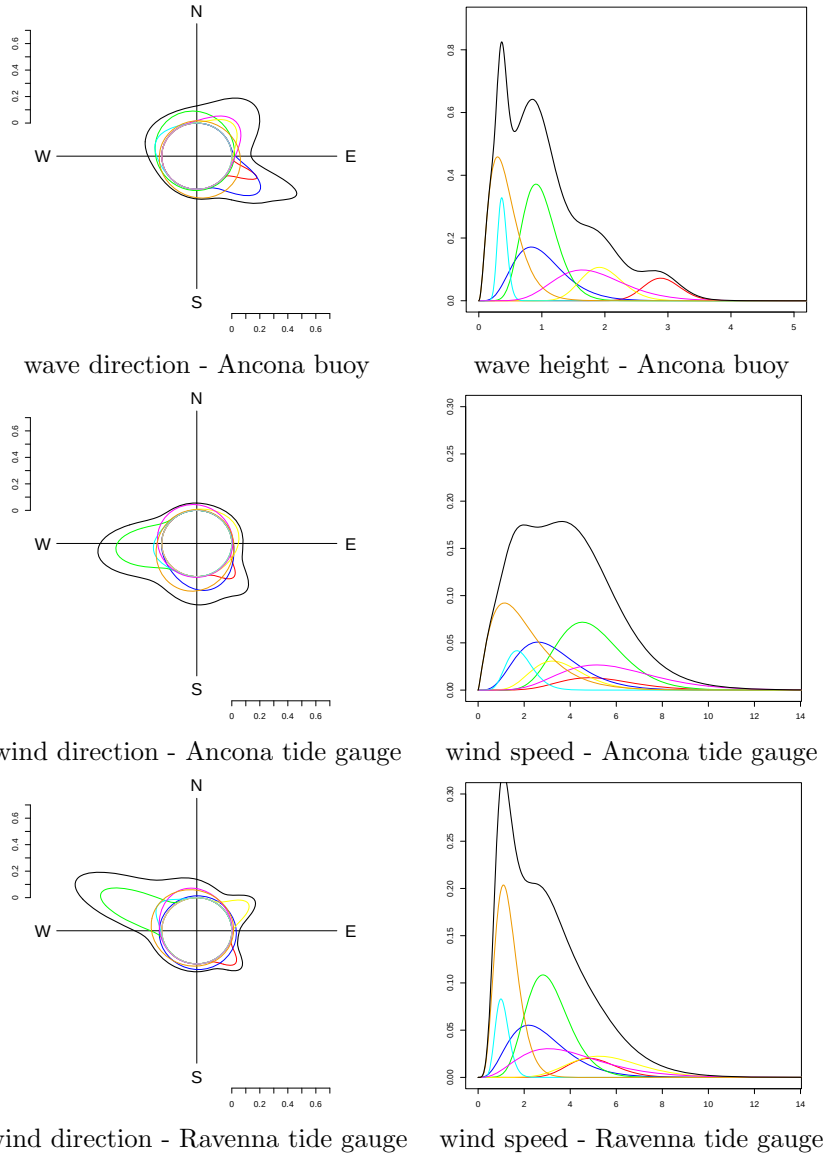


Figure 3: Densities of wave direction and height at the Ancona buoy (top) and wind direction and speed (middle: Ancona tide gauge; bottom: Ravenna tide gauge), as estimated by a 7-components LC cluster model; coloured lines indicate conditional densities and black lines indicate mixture densities

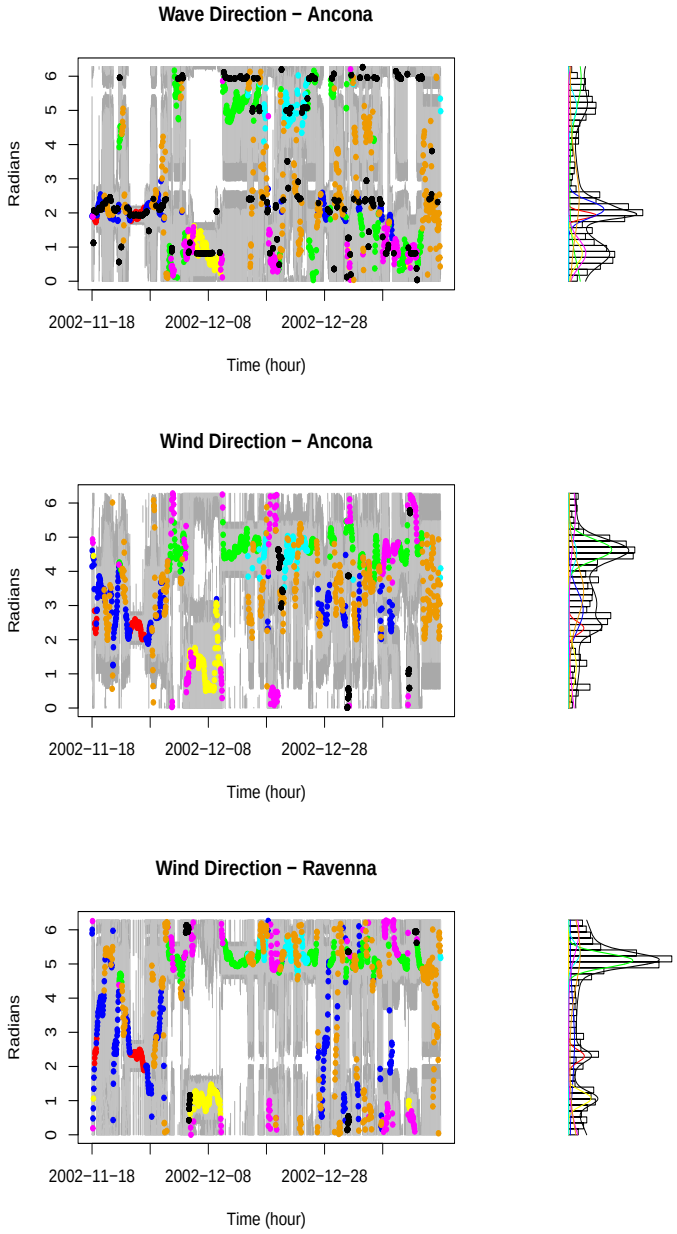


Figure 4: Left: directional data, clustered into seven latent classes and 95% (grey) and 99% (dark grey) fiducial intervals, as estimated by a 7-components mixture model. Black dots indicate missing values, imputed by the expectation of the conditional distribution of the missing values given the observed data, as estimated by the model. Right: histograms of complete data fitted by the model

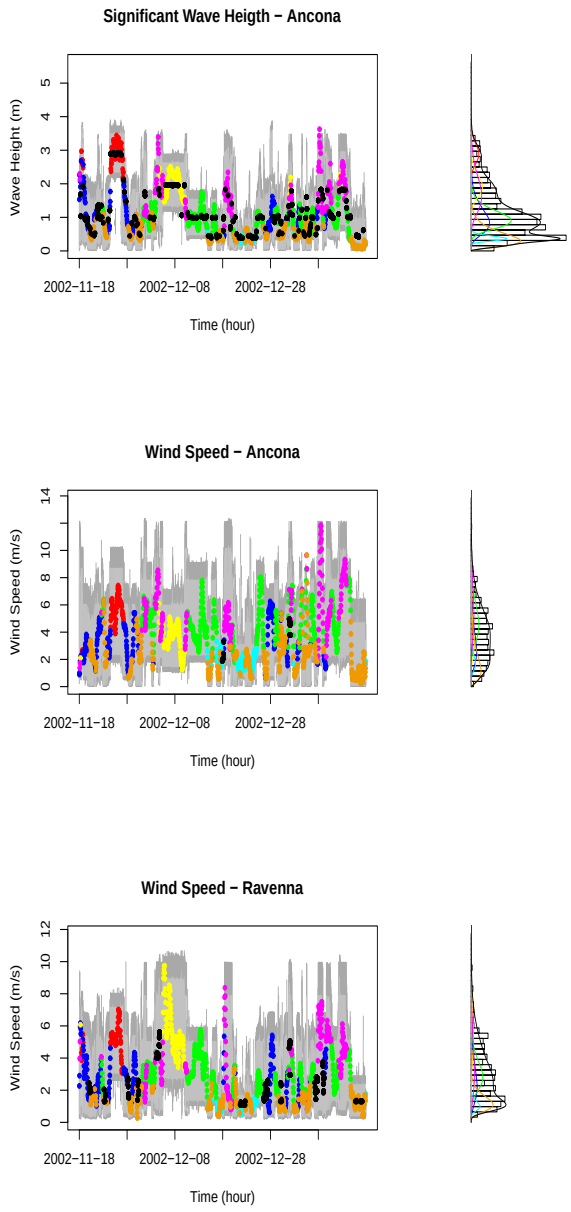


Figure 5: Left: linear data, clustered into seven latent classes and 95% (grey) and 99% (dark grey) fiducial intervals, as estimated by a 7-components mixture model. Black dots indicate missing values, imputed by the expectation of the conditional distribution of the missing values given the observed data, as estimated by the model. Right: histograms of complete data fitted by the model

Goodness of fit was also evaluated by comparing the squared cross-correlations between the observed data and those expected by the mixture model. To compute the empirical correlation between intensity observations (wind speed and wave height), we have used the standard Pearson correlation. The empirical correlation between circular data (wind and wave direction) was computed by exploiting the Fisher-Lee correlation index (Fisher and Lee, 1983). Finally, we computed the cross-correlation between linear and circular data (e.g., between wind direction and wave height) by exploiting the Mardia's linear-circular correlation index (Mardia, 1976). Table 3 displays a reasonable matching between the empirical correlations against their expected counterparts, under the estimated mixture model, showing that the conditional independence assumption of model (3.1) (coupled with the choice of 7 latent classes) explains a significant part of data variability.

Table 3: Observed and expected squared correlations

| | Wave H. | Wind S. [a] | Wind S. [b] | Wave D. | Wind D. [a] | Wind D. [b] |
|---|---|---|---|---|---|---|
| Wave Heigth | 1 | | | | | |
| (expected) | (1) | | | | | |
| Wind Speed [a] | 0.191 | 1 | | | | |
| (expected) | (0.320) | (1) | | | | |
| Wind Speed [b] | 0.385 | 0.142 | 1 | | | |
| (expected) | (0.517) | (0.199) | (1) | | | |
| Wave Direction | 0.199 | 0.111 | 0.168 | 1 | | |
| (expected) | (0.193) | (0.117) | (0.165) | (1) | | |
| Wind Direction [a] | 0.233 | 0.108 | 0.193 | 0.002 | 1 | |
| (expected) | (0.222) | (0.165) | (0.111) | (0.005) | (1) | |
| Wind Direction [b] | 0.184 | 0.007 | 0.119 | 0.008 | 0.017 | 1 |
| (expected) | (0.194) | (0.003) | (0.156) | (0.011) | (0.027) | (1) |

[a] Tide gauge: Ancona - [b] Tide gauge: Ravenna

We also evaluated the predictive accuracy of the model by non-parametric cross-validation (Gelman *et al.*, 1998). More precisely, we randomly split the sample in 10 subsamples. From each subsample, we discarded the 10% of the observations and (1) use the remaining portion of the subsample to fit a new model and (2) draw 5 imputations for each discarded vector of data, from the estimated conditional distribution of the discarded values given the observed data. If multiple imputations were of good quality, then we would expect than the actual outcome and the multiple imputations to have the same distributions, so that if one ranked the actual response along to the 5 imputations, then all 6 possible

orderings (actual outcome lowest, second lowest,$\cdots$, highest) would be equally likely. Figure 6 displays the cumulative distribution functions of the 6 ranks of circular and linear outcomes (overlapped to that of the uniform distribution), showing the good predictive accuracy of the model.

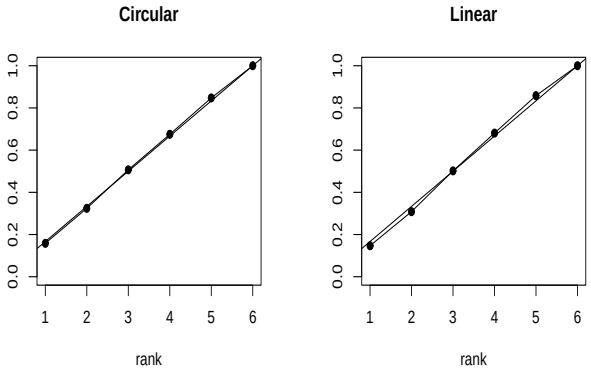


Figure 6: Rank cumulative distribution of the actual outcome with respect to 5 multiple imputations in a cross-validation experiment and cumulative distribution function of a uniform distribution

## 6. Discussion

We propose a latent-class approach to identify wave regimes under various wind conditions and estimate regime-specific wave parameters, such as modal wave directions and average wave heights, in the case of incomplete data, observed at different locations.

Using mixtures of product densities to model multivariate data allows for a simple specification of the dependence structure between variables that are measured on different supports (e.g. linear and circular) and, simultaneously, provides a flexible framework within which a variety of different parametric families can be exploited to model the univariate distribution of each single variable, given the latent class. We exploited von Mises and Gamma distributions, but the estimation procedure of Section 3 can be implemented by choosing different parametric families that can be more suitable in different case studies. By assuming a mixture densities, moreover, missing values are efficiently handled in a maximum-likelihood framework.

Modelling flexibility and computational efficiency in the case of incomplete data information come at the price of a simplifying constraint on the dependence structure among variables, given by the conditional independence assumption. In marine studies, this assumption can be often motivated by empirical evidence of a number of latent sea regimes and by the need of clustering the data in a

way that the association structure between the observed variables is well approximated by this partitioning of the sample. Nevertheless, issues of goodness of fit should be carefully addressed. Rigorous goodness-of-fit methods are however problematic with missing values. We have obtained reassuring results by computing case-wise fiducial intervals, overlaying the estimated marginal densities of the variables on the observed histograms (Figures 4 and 5) and comparing expected and empirical squared correlations between the variables (Table 3). These results should be interpreted with care, because empirical histograms and correlations are computed after discarding the missing values and because having most of the observed values within fiducial intervals says little about their ability to include missing values. These issues motivated our cross-validation experiment, whose results indicate that the proposed model was capable to explain most of the data variability and to re-impute artificially-removed values with a reasonable accuracy.

## References

Bertotti, L. and Cavalieri, L. (2009). Wind and wave predictions in the Adriatic Sea. *Journal of Marine Systems* **78**, S227-S234.

Boukhanovsky, A. V., Lopatouhkin, L. J. and Guedes Soares, C. (2007). Spectral wave climate of the North Sea. *Applied Ocean Research* **29**, 146-154.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1-38.

Faltinsen, O. M. (1990). *Sea Loads on Ships and Offshore Structures*. Cambridge University Press, Cambridge.

Fernández-Durán, J. J. (2007). Models for circular-linear and circular-circular data constructed from circular distributions based on nonnegative trigonometric sums. *Biometrics* **63**, 579-585.

Fisher, N. I. and Lee, A. J. (1983). A correlation coefficient for circular data. *Biometrika* **70**, 327-332.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of American Statistical Association* **97**, 611-631.

Gelman, A., King, G. and Liu, C. (1998). Not asked or not answered: multiple imputation for multiple surveys. *Journal of the American Statistical Association* **93**, 846-874.

Hagenaars, J. A. and McCutcheon, A. L. (2002). *Applied Latent Class Analysis*. Cambridge University Press, Cambridge.

Hamilton, L. J. (2010). Characterising spectral sea wave conditions with statistical clustering of actual spectra. *Applied Ocean Research* **32**, 332-342.

Huang, G., Wing-Keung Law, A. and Huang, Z. (2011). Wave-induced drift of small floating objects in regular waves. *Ocean Engineering* **38**, 712-718.

Hunt, L. and Jorgensen, M. (2003). Mixture model clustering for mixed data with missing information. *Computational Statistics and Data Analysis* **41**, 429-440.

Jin, K. R. and Ji, Z. G. (2004). Case study: modeling of sediment transport and wind-wave impact in Lake Okeechobee. *Journal of Hydraulic Engineering* **130**, 1055-1067.

Kato, S. and Shimizu, K. (2008). Dependent models for observations which include angular ones. *Journal of Statistical Planning and Inference* **138**, 3538-3549.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44**, 226-233.

Mardia, K. V. (1976). Linear-circular correlation coefficients and rhythmometry. *Biometrika* **63**, 403-405.

Mardia, K. V., Hughes, G., Taylor, C. C. and Singh, H. (2008). A multivariate von Mises distribution with applications to bioinformatics. *Canadian Journal of Statistics* **36**, 99-109.

Pleskachevsky, A., Eppel, D. P. and Kapitza, H. (2009). Interaction of waves, currents and tides, and wave-energy impact on the beach area of Sylt Island. *Ocean Dynamics* **59**, 451-461.

Rotnitzky, A. and Wypij, D. (1994). A note on the bias of estimators with missing data. *Biometrics* **50**, 1163-1170.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapmanan and Hall, London.

Teicher, H. (1967). Identifiability of mixtures of product measures. *Annals of Mathematical Statistics* **38**, 1300-1302.

Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A. and Sijtsma, K. (2008). Multiple imputation of categorical data using latent class analysis. *Sociological Methodology* **33**, 369-297.

Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *Annals of Mathematical Statistics* **39**, 209-214.

Wu, C. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* **11**, 95-103.

Received November 18, 2010; accepted April 21, 2011.

Francesco Lagona
DIPES and GRASPA Unit of Rome
Roma Tre University
Chiabrera 199, 00145 Rome, Italy
lagona@uniroma3.it

Marco Picone
Department of Economics and GRASPA Unit of Rome
Roma Tre University
Chiabrera 199, 00145 Rome, Italy
marco.picone@uniroma3.it

*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

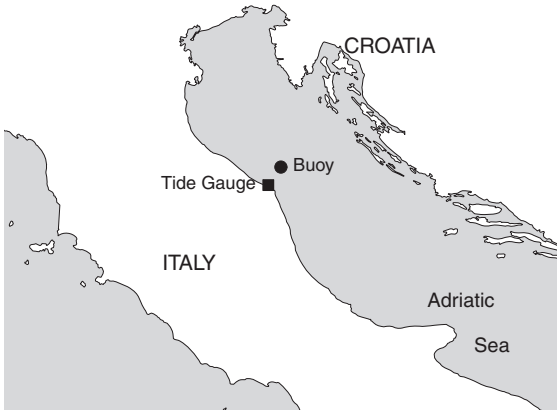


Figure 1. Locations of the buoy (circle) and tide gauge (square) at Ancona.

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 |

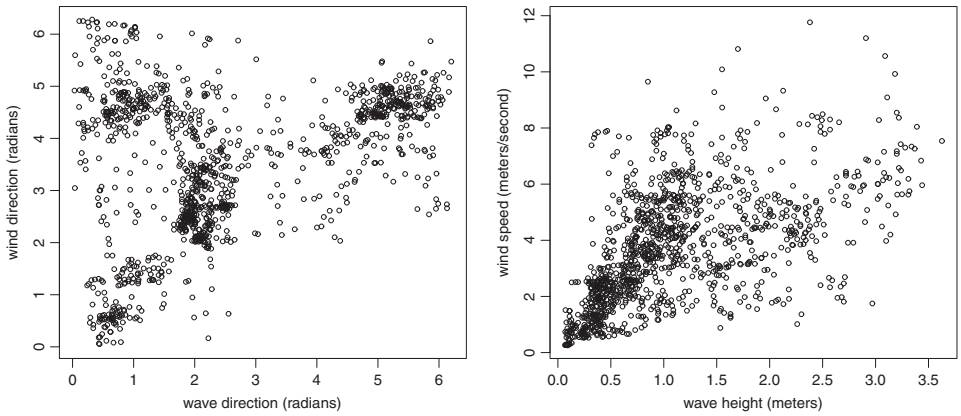


Figure 2. Complete profiles of wind and wave direction (left) and complete profiles of wind speed and wave height (right).

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

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

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

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 $k$th 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 $\beta_{12}$ 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 $\beta_i$, $i = 1, 2$, 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 $\beta_i$ and marginal concentrations

$\beta_{ii}$. 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 $(\beta_{1k}, \beta_{2k})$ 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, $\Gamma$ 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

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

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

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

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 $i$th 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

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

where $L_i(\boldsymbol{\theta})$ is the likelihood contribution of the $i$th 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 $i$th 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:

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

where $\boldsymbol{s}_i(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_t)$ is the $i$th score vector, obtained by deriving the $i$th 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 |
| $\beta_{1k}$ | 2.06 | 1.08 | 5.60 |
| (Wave mean direction) | (0.07) | (0.03) | (0.08) |
| $\beta_{2k}$ | 3.13 | 1.33 | 4.61 |
| (Wind mean direction) | (0.05) | (0.06) | (0.02) |
| $\beta_{11k}$ | 1.61 | 4.57 | 1.15 |
| (Wave directional concentration) | (0.11) | (0.51) | (0.13) |
| $\beta_{22k}$ | 2.14 | 0.76 | 8.57 |
| (Wind directional concentration) | (0.16) | (0.09) | (0.72) |
| $\beta_{12k}$ | −0.19 | 3.09 | 1.23 |
| (Wind/wave directional inverse correlation) | (0.23) | (0.28) | (0.28) |
| $\gamma'_{1k}$ | 0.38 | 1.85 | 0.70 |
| (Wave mean height) | (0.05) | (1.69) | (0.10) |
| $\gamma'_{2k}$ | 1.51 | 3.35 | 3.12 |
| (Wind mean speed) | (0.22) | (0.38) | (0.22) |
| $\gamma_{11k}$ | 0.06 | 0.41 | 0.12 |
| (Wave height variance) | (0.01) | (0.29) | (0.03) |
| $\gamma_{22k}$ | 1.29 | 2.65 | 2.85 |
| (Wind speed variance) | (0.29) | (0.64) | (0.55) |
| $\gamma_{12k}$ | 0.22 | 0.58 | 0.54 |
| (Wind/wave covariance) | (0.03) | (0.10) | (0.06) |
| $\gamma''_{1k}$ | 0.21 | 0.18 | 0.20 |
| (Wave skewness) | (0.07) | (2.15) | (0.13) |
| $\gamma''_{2k}$ | 0.93 | 1.54 | 1.68 |
| (Wind skewness) | (0.26) | (0.44) | (0.25) |
| $\pi$ | 0.32 | 0.32 | 0.36 |
| (Component weight) | (0.02) | (0.02) | (0.01) |

estimated. In our application, the BIC takes the form

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

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

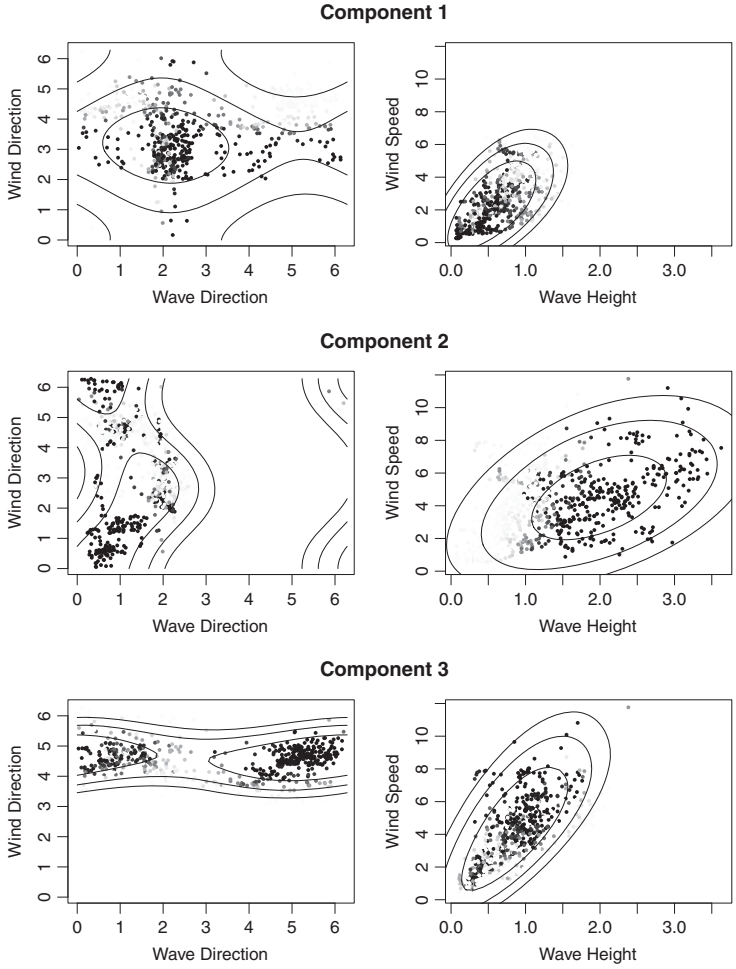


Figure 3. Contour plots of the conditional circular and linear bivariate log-densities, as estimated by fitting a three-component mixture model, at levels 4, 6 and 8; for each component, points are filled with a gray color that is proportional to the estimated probability that each observation belongs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

the estimated mean entropy

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

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 ($\beta_{121}$ 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.

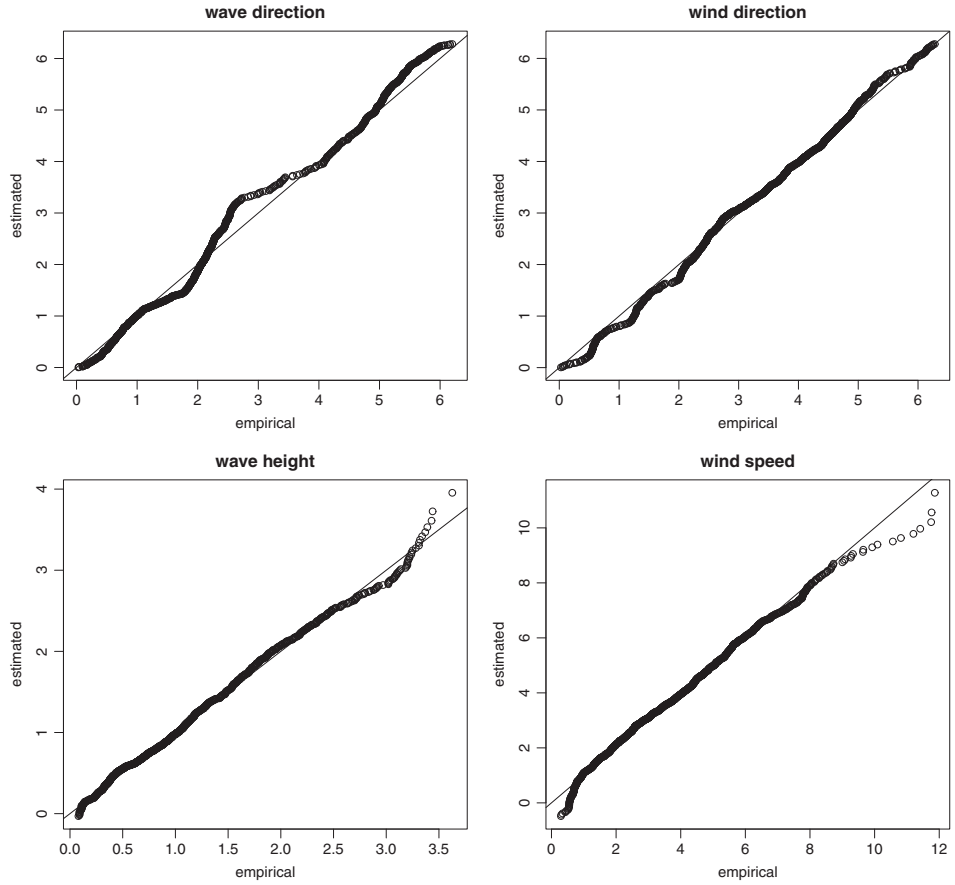


Figure 4. Quantile–quantile plots of the marginal distribution of wave direction and height (left) and wind direction and speed (right), as estimated by a model with three components.

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

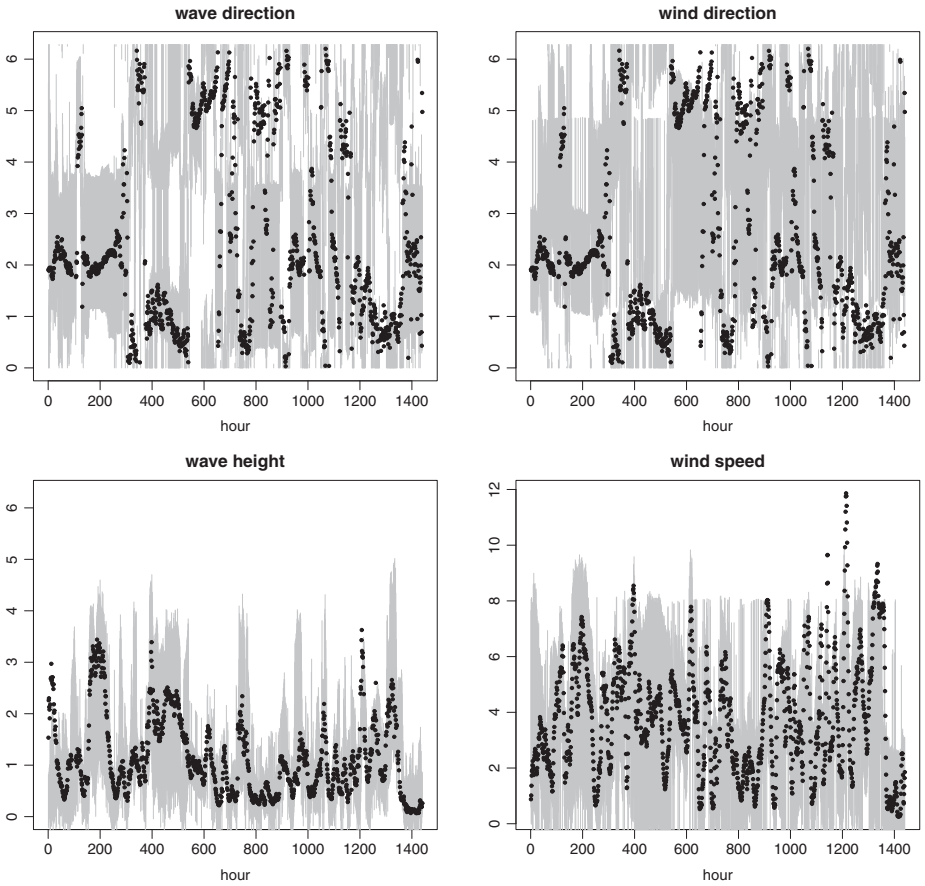


Figure 5. The 99% predictive intervals of wave direction and height (left) and wind direction and speed (right), as estimated by a model with three components.

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 $x_2$.

*Journal of Statistical Computation and Simulation*
iFirst, 2012, 1–15

Taylor & Francis
Taylor & Francis Group

# Maximum likelihood estimation of bivariate circular hidden Markov models from incomplete data

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

[a]*DIPES, University of Roma Tre, Via G. Chiabrera 199, 00164 Rome, Italy;* [b]*Department of Economics, University of Roma Tre and National Institute for Environmental Protection – Marine Service, Italy*

(*Received 23 July 2011; final version received 9 January 2012*)

In this paper, we propose a hidden Markov model for the analysis of the time series of bivariate circular observations, by assuming that the data are sampled from bivariate circular densities, whose parameters are driven by the evolution of a latent Markov chain. The model segments the data by accounting for redundancies due to correlations along time and across variables. A computationally feasible expectation maximization (EM) algorithm is provided for the maximum likelihood estimation of the model from incomplete data, by treating the missing values and the states of the latent chain as two different sources of incomplete information. Importance-sampling methods facilitate the computation of bootstrap standard errors of the estimates. The methodology is illustrated on a bivariate time series of wind and wave directions and compared with popular segmentation models for bivariate circular data, which ignore correlations across variables and/or along time.

**Keywords:** bivariate circular data; EM algorithm; hidden Markov model; importance sampling; missing values; mixtures; sine model; bivariate von Mises; wind; wave

## 1. Introduction

Circular hidden Markov models (HMMs) have been recently introduced as flexible frameworks for the statistical analysis of univariate circular time series, which arise in environmental, biological and ecological studies [1]. Univariate circular time series are temporal sequences of angles and can be represented as trajectories of points on the circle. The marginal distribution of these data is often multimodal, suggesting that the data are drawn from different distributions, associated with different latent regimes. This often motivates the use of mixture models in circular data analysis. Mixture models are helpful for recovering latent regimes from the data and for estimating the parameters of the data distribution under each regime. HMMs are particular mixture models that account for temporal autocorrelation, by assuming that the temporal transitions between latent regimes occur according to the transition probability matrix of an unobserved (i.e. hidden) Markov chain. Circular HMMs are particular HMMs that accommodate for the circular structure of the data, by modelling observations as

*Corresponding author. Email: lagona@uniroma3.it

ISSN 0094-9655 print/ISSN 1563-5163 online

http://dx.doi.org/10.1080/00949655.2012.656642
http://www.tandfonline.com

samples drawn from a mixture of circular densities, that is, probability densities with circular support. The von Mises HMM is obtained by considering mixtures of von Mises densities and it provides an intuitively appealing tool in the analysis of univariate time series of circular observations.

We generalize the von Mises HMM to handle bivariate circular time series. Bivariate circular times series are temporal sequences of bivariate angular observations and can be represented as trajectories on a torus, that is, a surface generated by revolving a circle in the three-dimensional space. Examples include time series of hourly wind and wave directions [2] in environmetrics and sequences of dihedral angles in bioinformatics [3]. An HMM approach has been already proposed in the literature to model bivariate circular data, through a conditional independence (CI) assumption [1,3]: at each time, the two observed angles are assumed to be conditionally independent given the latent state of the Markov chain. As a result, a bivariate circular HMM is obtained by modelling bivariate observations with the product of two univariate densities. CI is often exploited in the specification of multivariate HMMs for non-normal observations [4] and can be motivated by borrowing arguments from the latent-class literature, where multivariate densities are approximated by mixtures of product densities and latent states capture the association structure of multivariate observations. In this paper, we show that CI is not necessary in the modelling of bivariate circular time series, by introducing the bivariate sine HMM (BSHMM). The BSHMM is specified by modelling pairs of angles by a mixture of sine bivariate densities [5], whose parameters depend on the states of a latent Markov chain. The sine bivariate density can be viewed as the circular counterpart of a bivariate Gaussian density and depends on five parameters: two mean directions, two concentrations that indicate the spread of the angles around their means and, finally, a parameter that indicates the dependence between the two angular observations. As a result, the BSHMM introduced in this paper can be viewed as the circular counterpart of a bivariate normal HMM, well known in the literature as a powerful tool to study bivariate time series. The BHMM allows us to segment the data by accounting for possible redundancies due to the dependence across variables and the autocorrelation along time. In addition, a number of models, often employed in the analysis of bivariate circular data, can be obtained as particular cases of a BSHMM. For example, if the dependence parameter is equal to zero, a sine bivariate density is equal to the product of two univariate von Mises densities and the BSHMM reduces to the aforementioned CI-based circular HMM. Furthermore, if all the rows of the transition probability matrix in a BSHMM are equal to the initial probability distribution of the chain, then the Markov chain reduces to a Bernoulli scheme and, as a result, the BSHMM reduces to a mixture of sine bivariate densities, recently considered in the literature to cluster independent toroidal data [6,7]. Finally, if both the dependence parameter is equal to zero and all the rows of the transition probability matrix are equal, a BSHMM reduces to a mixture whose components are products of von Mises univariate densities, recently exploited in environmental studies [2].

In this paper, we focus on the maximum likelihood estimation of the BSHMM from incompletely observed, bivariate circular time series, because missing values often occur in environmental and biological studies. Specifically, we present a computationally feasible expectation maximization (EM) algorithm that treats the unobserved states of the Markov chain and the missing values as two different sources of incomplete information. We also show that the BSHMM can be simulated by exploiting standard importance-sampling methods and, as a result, estimate uncertainty can be efficiently assessed by computing bootstrap standard errors.

The rest of the paper is organized as follows. Section 2 summarizes relevant details of the BSHMM, while Section 3 presents the technical details of the EM algorithm. Section 4 illustrates a feasible simulation procedure to compute the standard errors of the estimates. Section 5 is devoted to an application on marine data and, finally, Section 6 summarizes relevant points of discussion.

## 2. The bivariate sine hidden Markov model

A bivariate circular time series takes the form of a vector $\boldsymbol{x}_T = (\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$, with coordinates $\boldsymbol{x}_t = (x_{1t}, x_{2t}) \in (0, 2\pi)^2$. To specify the distribution of $\boldsymbol{x}_T$, we introduce a homogeneous Markov chain with $K$ states, defined by a vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ of initial probabilities and a transition probability matrix $\boldsymbol{P}$, whose $(h, k)$th element $\pi_{h,k}$ indicates the conditional probability of visiting state $k$ at time $t$, given that the chain is in state $h$ at time $t - 1$.

We also introduce a parametric family of toroidal densities $f(\boldsymbol{x}; \boldsymbol{\beta}), \boldsymbol{x} \in (0, 2\pi)^2$, indexed by a multivariate vector $\boldsymbol{\beta}$. Under this setting, the time series $\boldsymbol{x}_T$ is distributed according to a bivariate circular HMM if its joint density is given by

$$f(\boldsymbol{x}_T) = \sum_{k_0, k_1, \ldots, k_T}^{1,\ldots,K} p(k_0, k_1, \ldots, k_T; \boldsymbol{\pi}, \boldsymbol{P}) \prod_{t=0}^{T} f(\boldsymbol{x}_t; \boldsymbol{\beta}_{k_t}), \tag{1}$$

where the $(T + 1)$-fold summation runs over all the possible sequences $(k_0, k_1, \ldots, k_T)$, $k_t \in \{1, \ldots, K\}$, $t \in \{0, 1, \ldots, T\}$, that can be visited by a $K$-state Markov chain, with probability

$$p(k_0, k_1, \ldots, k_T; \boldsymbol{\pi}, \boldsymbol{P}) = \pi_{k_0} \prod_{t=1}^{T} \pi_{k_{t-1}, k_t}. \tag{2}$$

Note that if $\boldsymbol{x}$ was a point in the plane $\mathbb{R}^2$ and $f(\boldsymbol{x}; \boldsymbol{\beta})$ was a bivariate normal density $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then Equation (1) would be the distribution of a bivariate normal HMM.

Parametric families of toroidal densities $f(\boldsymbol{x}; \boldsymbol{\beta})$ can be specified in a number of different ways [6,8] and plugged into Equation (1) to obtain several bivariate circular HMMs. To our knowledge, however, none of the existing toroidal densities has been ever considered under an HMM setting. The BSHMM, proposed here, exploits a bivariate sine model to specify a toroidal density. The sine density [5] is a parametric distribution on the torus $(0, 2\pi)^2$ which naturally imbeds the bivariate normal distribution when the range of observations is small. Given a point $\boldsymbol{x} \in (0, 2\pi)^2$, the sine density is given by

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

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

$$I_m(\beta) = \frac{1}{\pi} \int_0^{\pi} \mathrm{e}^{\beta \cos u} \cos(mu)\, \mathrm{d}u$$

is the modified Bessel function of order $m$. The sine model can be viewed as a bivariate generalization of the von Mises density, where $\beta_{12}$ accounts for the statistical dependence between $x_1$ and $x_2$, $\beta_1$ and $\beta_2$ are the modal directions of $x_1$ and $x_2$ and, finally, $\beta_{11}$ and $\beta_{22}$ are the concentration parameters of $x_1$ and $x_2$, indicating the spread of the data around the mode.

The two univariate marginal densities

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

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

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

We observe that, in general, these marginal densities are not von Mises and can have a bimodal shape. As has been proved in [5], unimodality holds under the sufficient condition

$$\beta_{12}^2 < 2 \min\{\beta_{11}, \beta_{22}\},$$

and in this case, the marginal densities are well approximated by von Mises densities. 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, that is,

$$f(x_i; \boldsymbol{\beta}) = \frac{\exp(\beta_{ii} \cos(x_i - \beta_i))}{2\pi I_0(\beta_{ii})} = f_{\text{vm}}(x_i; \boldsymbol{\beta}_i),$$

where $\boldsymbol{\beta}_i = (\beta_i, \beta_{ii})$ includes the marginal mode and the marginal concentration parameter of $x_i$. Regardless of the value taken by $\beta_{12}$, the conditional distribution of one component given the other one follows a von Mises law, namely

$$f(x_i|x_j; \boldsymbol{\beta}) = \frac{\exp(a(x_i) \cos(x_i - \beta_i - b(x_j)))}{2\pi I_0(a(x_i))} = f_{\text{vm}}(x_i; \beta_i + b(x_j), a(x_i)), \tag{6}$$

where

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

A BSHMM parsimoniously allows for possible correlations across variables (through the parameter $\beta_{12}$) and along time (through the transition probability matrix $\boldsymbol{P}$). Popular models for bivariate circular data ignore some of these correlations and can be obtained as particular cases of the BSHMM, by appropriately restricting these two parameters. For example, if $\beta_{12} = 0$, then Equation (1) reduces to a CI-based HMM:

$$f(\boldsymbol{x}_T) = \sum_{k_0, k_1, \ldots, k_T}^{1, \ldots, K} p(k_0, k_1, \ldots, k_T; \boldsymbol{\pi}, \boldsymbol{P}) \prod_{t=0}^{T} f_{\text{vm}}(x_{1t}; \boldsymbol{\beta}_{1,k_t}) f_{\text{vm}}(x_{2t}; \boldsymbol{\beta}_{2,k_t}). \tag{8}$$

If, instead, $\boldsymbol{P} = \boldsymbol{1}\boldsymbol{\pi}^{\mathsf{T}}$, that is, all the rows of $\boldsymbol{P}$ are equal to the initial distribution of the chain, then the data are assumed as independent samples, drawn from a mixture of bivariate sine densities. As a result, the joint distribution of the time series reduces to the product

$$\begin{aligned} f(\boldsymbol{x}_T) &= \sum_{k_0, k_1, \ldots, k_T}^{1, \ldots, K} \pi_{k_0}, \ldots, \pi_{k_T} \prod_{t=0}^{T} f(\boldsymbol{x}_t; \boldsymbol{\beta}_{k_t}) \\ &= \prod_{t=0}^{T} \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}_t; \boldsymbol{\beta}_k). \end{aligned} \tag{9}$$

If, finally, $\beta_{12} = 0$ and $\boldsymbol{P} = \boldsymbol{1}\boldsymbol{\pi}^{\mathsf{T}}$, then the BSHMM reduces to a CI-based mixture model, whose components are specified as products of von Mises densities, as follows:

$$f(\boldsymbol{x}_T) = \prod_{t=0}^{T} \sum_{k=1}^{K} \pi_k f_{\text{vm}}(x_{1t}; \boldsymbol{\beta}_{1k}) f_{\text{vm}}(x_{2t}; \boldsymbol{\beta}_{2k}). \tag{10}$$

## 3. Maximum likelihood estimation from incomplete data

We allow for missing values and, accordingly, refer to $\boldsymbol{x}_{t,\text{mis}}$ and $\boldsymbol{x}_{t,\text{obs}}$ as the missing and observed circular components at time $t$, respectively. We further introduce pairs $\boldsymbol{r}_t = (r_{t1}, r_{t2})$ of binary missing indicators, where $r_{tj} = 1$, if $x_{tj}$ is missing and zero otherwise, $j = 1, 2$. If the data are missing at random (i.e. the conditional probability of not observing a value, given the observed data, does not depend on the unobserved value [9]), the missing data mechanism can be ignored and the maximum likelihood estimate of parameter $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{P}, \boldsymbol{\beta})$ is the maximum point of the marginal log-likelihood function:

$$\begin{aligned}\log L(\boldsymbol{\theta}) &= \log \sum_{k_0,k_1,\ldots,k_T}^{1,\ldots,K} p(k_0, k_1, \ldots, k_T; \boldsymbol{\pi}, \boldsymbol{P}) \prod_{t=0}^{T} \int f(\boldsymbol{x}_t; \boldsymbol{\beta}_{k_t})\, \mathrm{d}\boldsymbol{x}_{\text{mis}} \\ &= \log \sum_{k_0,k_1,\ldots,k_T}^{1,\ldots,K} p(k_0, k_1, \ldots, k_T; \boldsymbol{\pi}, \boldsymbol{P}) \prod_{t=0}^{T} L_t(\boldsymbol{\beta}_{k_t}),\end{aligned} \tag{11}$$

where

$$L_t(\boldsymbol{\beta}_{k_t}) = f(\boldsymbol{x}_t; \boldsymbol{\beta}_{k_t})^{(1-r_{t1})(1-r_{t2})} f(x_{t1}; \boldsymbol{\beta}_{k_t})^{(1-r_{t1})r_{t2}} f(x_{t2}; \boldsymbol{\beta}_{k_t})^{r_{t1}(1-r_{t2})}$$

is the conditional likelihood contribution of the $t$th incomplete observation, given the state $k_t$, whereas $f(x_{tj}; \boldsymbol{\beta}_{k_t})$, $j = 1, 2$, are the marginal densities, defined in Equation (4), and evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_{k_t}$. As a result, the contribution of a fully missing profile is identically 1.

Because direct maximization of Equation (11) can be computationally problematic, we describe an EM algorithm that generates a sequence $(\hat{\boldsymbol{\theta}}_p, p = 1, 2, \ldots)$ of estimates such that $L(\hat{\boldsymbol{\theta}}_p) \geq L(\hat{\boldsymbol{\theta}}_{p-1})$. 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 (e.g. the latent states of the Markov chain and the unobserved directions), given the observed data.

To specify the complete-data log-likelihood function, the states of the hidden Markov chain can be conveniently defined as samples drawn from a multinomial process $(\boldsymbol{\xi}_t, t \geq 0)$ in discrete time, where $\boldsymbol{\xi}_t = (\xi_{t1}, \ldots, \xi_{tK})$ is a multinomial random variable with one trial and $K$ classes. Under Equation (2), the distribution of $\boldsymbol{\xi}_0$ is given by

$$p(\boldsymbol{\xi}_0) = \prod_{k=1}^{K} \pi_k^{\xi_{0k}},$$

while the conditional distribution of $\boldsymbol{\xi}_t$ given $\boldsymbol{\xi}_{t-1}$ is given by

$$p(\boldsymbol{\xi}_t | \xi_{t-1,h} = 1) = \prod_{k=1}^{K} \pi_{h,k}^{\xi_{t-1,h}\xi_{tk}}.$$

Treating both the latent states and the unobserved directions as missing values, we define the complete-data log-likelihood function as

$$\log L_{\text{comp}}(\boldsymbol{\theta}) = \sum_{k=1}^{K} \xi_{0k} \log \pi(k) + \sum_{t=1}^{T} \sum_{h=1}^{K} \sum_{k=1}^{K} \xi_{t-1,h}\xi_{t,k} \log \pi_{h,k} + \sum_{t=0}^{T} \sum_{k=1}^{K} \xi_{tk} \log f(\boldsymbol{x}_t; \boldsymbol{\beta}_k). \tag{12}$$

Given the estimate $\hat{\boldsymbol{\theta}}_p$, provided by the algorithm at step $p$, the expected value of $\log L_{\text{comp}}(\boldsymbol{\theta})$ with respect to the conditional distribution of the missing values $(\boldsymbol{\xi}, \boldsymbol{x}_{\text{mis}}) = (\boldsymbol{\xi}_t, \boldsymbol{x}_{t,\text{mis}}, t =$

$0, \ldots, T$) given the observed data $\boldsymbol{x}_{\text{obs}} = (\boldsymbol{x}_{t,\text{obs}}, t = 0, \ldots, T)$, evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_p$, say

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_p) = \mathbb{E}(\log L_{\text{comp}}(\boldsymbol{\theta})|\boldsymbol{x}_{\text{obs}}), \tag{13}$$

can be evaluated by observing that the distribution of the missing values given the observed data can be factorized into two components, as follows:

$$\begin{aligned} f(\boldsymbol{\xi}, \boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}}; \hat{\boldsymbol{\theta}}_p) &= p(\boldsymbol{\xi}|\boldsymbol{x}_{\text{obs}}; \hat{\boldsymbol{\theta}}_p) f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{\xi}, \boldsymbol{x}_{\text{obs}}; \hat{\boldsymbol{\beta}}_p) \\ &= p(\boldsymbol{\xi}|\boldsymbol{x}_{\text{obs}}; \hat{\boldsymbol{\theta}}_p) \prod_{t=0}^{T} \prod_{k=1}^{K} (f(\boldsymbol{x}_{t,\text{mis}}|\xi_{tk} = 1, \boldsymbol{x}_{t,\text{obs}}; \hat{\boldsymbol{\beta}}_k))^{\xi_{tk}}, \end{aligned} \tag{14}$$

where the conditional density

$$f(\boldsymbol{x}_{t,\text{mis}}|\xi_{tk} = 1, \boldsymbol{x}_{t,\text{obs}}; \hat{\boldsymbol{\beta}}_k) = \frac{f(\boldsymbol{x}_t; \hat{\boldsymbol{\beta}}_k)}{L_t(\hat{\boldsymbol{\beta}}_k)} \tag{15}$$

is identically 1, if $r_{i1} = r_{i2} = 0$, it reduces to the conditional univariate von Mises densities (6) with $\boldsymbol{\beta} = \boldsymbol{\beta}_k$, if either $(r_{i1}, r_{i2}) = (0, 1)$ or $(1, 0)$, and it is finally equal to the bivariate circular density (3) with $\boldsymbol{\beta} = \boldsymbol{\beta}_k$, if $(r_{i1}, r_{i2}) = (1, 1)$. As a result, the expected value of the complete log-likelihood function with respect to the conditional distribution of the missing values given the observed data (at the $(p + 1)$th step of the algorithm) is given by

$$\begin{aligned} Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_t) = &\sum_{k=1}^{K} \hat{\pi}_{0k(p+1)} \log \pi_k + \sum_{t=1}^{T} \sum_{h=1}^{K} \sum_{k=1}^{K} \hat{\pi}_{t-1,t,hk(p+1)} \log \pi_{h,k} \\ &+ \sum_{t=0}^{T} \sum_{k=1}^{K} \hat{\pi}_{tk(p+1)} \mathbb{E}_{(p+1)}(\log f(\boldsymbol{x}_t; \boldsymbol{\beta}_{k_t})|\boldsymbol{x}_{t,\text{obs}}, \xi_{tk} = 1), \end{aligned} \tag{16}$$

where the univariate and bivariate posterior probabilities of the latent states given the observed data and the output of the $p$th step of the EM algorithm, namely

$$\begin{aligned} \hat{\pi}_{tk(p+1)} &= P(\xi_{tk} = 1|\boldsymbol{x}_{\text{obs}}, \hat{\boldsymbol{\theta}}_p), \\ \hat{\pi}_{t-1,t,hk(p+1)} &= P(\xi_{th} = 1, \xi_{tk} = 1|\boldsymbol{x}_{\text{obs}}, \hat{\boldsymbol{\theta}}_p), \end{aligned}$$

can be computed by standard backward–forward iterations that account for incomplete profiles and prevent from overflows. To illustrate the recursion that we exploited, let

$$L_{0:t}(\boldsymbol{\theta}) = \sum_{k_0, k_1, \ldots, k_t}^{1,\ldots,K} p(k_0, k_1, \ldots, k_t; \boldsymbol{\pi}, \boldsymbol{P}) \prod_{t=0}^{t} L_t(\boldsymbol{\beta}_{k_t})$$

be the contribution of the first $t$ incomplete profiles to the marginal likelihood and let $\boldsymbol{x}_{0:t} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t)$ be the time series, observed up to time $t$. During the forward iteration, we exploit the output of the $p$th step of the EM algorithm to compute the probabilities $\psi^{(t)}(k) = P(\xi_{tk} = 1|\boldsymbol{x}_{0:t-1})$, the likelihood ratios $c_t = L_{0:t}(\hat{\boldsymbol{\theta}}_p)/L_{0:t-1}(\hat{\boldsymbol{\theta}}_p)$ and the forward probabilities $\bar{\alpha}_t(k) = P(\xi_{tk} = 1|\boldsymbol{x}_{0:t})$, as follows:

*Forward recursion*

- Initialization:

$$\psi^{(0)}(k) = \hat{\pi}_{kp},$$

$$c_0 = \sum_{k=1}^{K} \psi^{(0)}(k) L_0(\hat{\boldsymbol{\beta}}_{kp}),$$

$$\bar{\alpha}_0(k) = \frac{\psi^{(0)}(k) L_0(\hat{\boldsymbol{\beta}}_{kp})}{c_0}.$$

- For $t = 1, \ldots, T$,

$$\psi^{(t)}(k) = \sum_{h=1}^{K} \bar{\alpha}_{t-1}(h) \hat{\pi}_{hkp},$$

$$c_t = \sum_{k=1}^{K} \psi^{(t)}(k) L_t(\hat{\boldsymbol{\beta}}_{kp}),$$

$$\bar{\alpha}_t(k) = \frac{\psi^{(t)}(k) L_t(\hat{\boldsymbol{\beta}}_{kp})}{c_t}.$$

At the end of the forward recursion, we store the values $c_0, \ldots, c_T$ and $\bar{\alpha}_0(k), \ldots, \bar{\alpha}_T(k)$ and run a backward recursion, by computing the ratios $\bar{\gamma}_t(k) = f(\boldsymbol{x}_{t+1:T} | \xi_{tk} = 1) / \prod_{l=t}^{T} c_l$, as follows.

*Backward recursion*

- Initialization: $\bar{\gamma}_T(k) = 1/c_T$.
- For $t = T-1, T-2, \ldots, 0$,

$$\bar{\gamma}_t(k) = \frac{\sum_{h=1}^{K} \hat{\pi}_{khp} L_{t+1}(\hat{\boldsymbol{\beta}}_{kp}) \bar{\gamma}_{t+1}(h)}{c_t}.$$

At the end of the backward recursion, we store the values of $\bar{\gamma}_0(k), \ldots, \bar{\gamma}_T(k)$ and compute the posterior univariate state probabilities as

$$\hat{\pi}_{tk(p+1)} = \frac{\bar{\alpha}_t(k) \bar{\gamma}_t(k)}{\sum_{k=1}^{K} \bar{\alpha}_t(k) \bar{\gamma}_t(k)}.$$

The bivariate posterior probabilities can be instead computed as

$$\hat{\pi}_{t-1,t,hk(p+1)} = \bar{\alpha}_t(k) \hat{\pi}_{hkp} L_{t+1}(\hat{\boldsymbol{\beta}}_{kp}) \bar{\gamma}_{t+1}(k).$$

The E step of the algorithm is completed by evaluating the expected value of $\log f(\boldsymbol{x}_t; \boldsymbol{\beta}_k)$ with respect to Equation (15). This expectation takes the simple form

$$\mathbb{E}(\log f(\boldsymbol{x}_t; \boldsymbol{\beta}_k) | \boldsymbol{x}_{t,\text{obs}}, \xi_{tk} = 1) = -\log C(\hat{\boldsymbol{\beta}}_{kp}) + \hat{\beta}_{11kp} E_{tk1p} + \hat{\beta}_{22kp} E_{tk2p} + \hat{\beta}_{12kp} E_{tk3p},$$

where

$$E_{tk1p} = \mathbb{E}(\cos(x_{t1} - \hat{\beta}_{1kp}) | \boldsymbol{x}_{t,\text{obs}}, \xi_{tk} = 1) = \begin{cases} \cos(x_{t1} - \hat{\beta}_{1kp}), & r_{t1} = 0, \\ \cos \hat{b}_{kp}(x_{t2}) \dfrac{I_1(\hat{a}_{kp}(x_{t2}))}{I_0(\hat{a}_{kp}(x_{t2}))}, & r_{t1} = 1,\ r_{t2} = 0, \end{cases}$$

$$E_{tk2p} = \mathbb{E}(\cos(x_{t2} - \hat{\beta}_{2kp}) | \boldsymbol{x}_{t,\text{obs}}, \xi_{tk} = 1) = \begin{cases} \cos(x_{t2} - \hat{\beta}_{2kp}), & r_{t2} = 0, \\ \cos \hat{b}_{kp}(x_{t1}) \dfrac{I_1(\hat{a}_{kp}(x_{t1}))}{I_0(\hat{a}_{kp}(x_{t1}))}, & r_{t1} = 0,\ r_{t2} = 1, \end{cases}$$

$$E_{tk3p} = \mathbb{E}(\sin(x_{t1} - \hat{\beta}_{1kp}) \sin(x_{t2} - \hat{\beta}_{2kp}) | \boldsymbol{x}_{t,\text{obs}}, \xi_{tk} = 1)$$
$$= \begin{cases} \sin(x_{t1} - \hat{\beta}_{1kp}) \sin(x_{t2} - \hat{\beta}_{2kp}), & r_{t1} = r_{t2} = 0, \\ \sin \hat{b}_{kp}(x_{t2}) \dfrac{I_1(\hat{a}_{kp}(x_{t2}))}{I_0(\hat{a}_{kp}(x_{t2}))} \sin(x_{t2} - \hat{\beta}_{2kp}), & r_{t1} = 1,\ r_{t2} = 0, \\ \sin(x_{t1} - \hat{\beta}_{1kp}) \sin \hat{b}_{kp}(x_{t1}) \dfrac{I_1(\hat{a}_{kp}(x_{t1}))}{I_0(\hat{a}_{kp}(x_{t1}))}, & r_{t1} = 0,\ r_{t2} = 1. \end{cases}$$

We then carry out the M step of the algorithm by maximizing $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_p)$, which is the sum of two functions of independent parameters and can be then maximized separately. Maximization with respect to the transition probabilities $\pi_{hk}$ provides the well-known updated values

$$\hat{\pi}_{hk(p+1)} = \frac{\sum_{t=1}^{T} \hat{\pi}_{t-1,t,hk(p+1)}}{\sum_{t=1}^{T} \hat{\pi}_{t-1,hk(p+1)}}, \quad h, k = 1, \ldots, K.$$

Maximization with respect to the parameters $\boldsymbol{\beta}$ is instead obtained by separately solving the following $K$ systems of score equations:

$$\begin{aligned} \frac{\sum_{i=1}^{n} \hat{\pi}_{tk(p+1)} E_{tk1p}}{\sum_{i=1}^{n} \hat{\pi}_{tk(p+1)}} &= \frac{1}{C(\boldsymbol{\beta}_k)} \frac{\partial C(\boldsymbol{\beta}_k)}{\partial \beta_{11k}}, \\ \frac{\sum_{i=1}^{n} \hat{\pi}_{tk(p+1)} E_{tk2p}}{\sum_{i=1}^{n} \hat{\pi}_{tk(p+1)}} &= \frac{1}{C(\boldsymbol{\beta}_k)} \frac{\partial C(\boldsymbol{\beta}_k)}{\partial \beta_{22k}}, \\ \frac{\sum_{i=1}^{n} \hat{\pi}_{tk(p+1)} E_{tk3p}}{\sum_{i=1}^{n} \hat{\pi}_{tk(p+1)}} &= \frac{1}{C(\boldsymbol{\beta}_k)} \frac{\partial C(\boldsymbol{\beta}_k)}{\partial \beta_{12k}}, \\ \frac{\beta_{11k} \sum_{i=1}^{n} \hat{\pi}_{tk(p+1)} A_{tk1(p)} - \beta_{12k} \sum_{i=1}^{n} \hat{\pi}_{tk(p+1)} C_{tk1p}}{\beta_{11k} \sum_{i=1}^{n} \hat{\pi}_{tk(p+1)} B_{tk1p} + \beta_{12k} \sum_{i=1}^{n} \hat{\pi}_{tk(p+1)} D_{tk1p}} &= \tan \beta_{1k}, \\ \frac{\beta_{22k} \sum_{i=1}^{n} \hat{\pi}_{tk(p+1)} A_{tk2p} - \beta_{12k} \sum_{i=1}^{n} \hat{\pi}_{tk(p+1)} C_{tk2p}}{\beta_{22k} \sum_{i=1}^{n} \hat{\pi}_{tk(p+1)} B_{tk2p} + \beta_{12k} \sum_{i=1}^{n} \hat{\pi}_{tk(p+1)} D_{tk2p}} &= \tan \beta_{2k}, \end{aligned} \tag{17}$$

where the expected sufficient statistics in the last two score equations are given by

$$A_{tk1p} = \mathbb{E}(\sin x_{i1} | \boldsymbol{x}_{i,\text{obs}}, \xi_{ik} = 1) = \begin{cases} \sin x_{i1}, & r_{i1} = 0, \\ \sin(\hat{\beta}_{1kp} + \hat{b}_{kp}(x_{i2})) \dfrac{I_1(\hat{a}_{kp}(x_{i2}))}{I_0(\hat{a}_{kp}(x_{i2}))}, & r_{i1} = 0,\ r_{i2} = 1, \end{cases}$$

$$B_{tk1p} = \mathbb{E}(\cos x_{i1} | \boldsymbol{x}_{i,\text{obs}}, \xi_{ik} = 1) = \begin{cases} \cos x_{i1}, & r_{i1} = 0, \\ \cos(\hat{\beta}_{1kp} + \hat{b}_{kp}(x_{i2})) \dfrac{I_1(\hat{a}_{kp}(x_{i2}))}{I_0(\hat{a}_{kp}(x_{i2}))}, & r_{i1} = 1,\ r_{i2} = 0, \end{cases}$$

$$C_{tk1p} = \mathbb{E}(\sin(x_{i2} - \hat{\beta}_{2kp}) \cos x_{i1} | \boldsymbol{x}_{i,\text{obs}}, \xi_{ik} = 1)$$
$$= \begin{cases} \sin(x_{i2} - \hat{\beta}_{2kp}) \cos x_{i1}, & r_{i1} = r_{i2} = 0, \\ \cos(\hat{\beta}_{1kp} + \hat{b}_{kp}(x_{i2})) \dfrac{I_1(\hat{a}_{kp}(x_{i2}))}{I_0(\hat{a}_{kp}(x_{i2}))} \sin(x_{i2} - \hat{\beta}_{2kp}), & r_{i1} = 1,\ r_{i2} = 0, \\ \cos x_{i1} \sin \hat{\beta}_{2kp} \dfrac{I_1(\hat{a}_{kp}(x_{i1}))}{I_0(\hat{a}_{kp}(x_{i1}))}, & r_{i1} = 0,\ r_{i2} = 1, \end{cases}$$

$$D_{tk1p} = \mathbb{E}(\sin(x_{i2} - \hat{\beta}_{2kp}) \sin x_{i1} | \boldsymbol{x}_{i,\text{obs}}, \xi_{ik} = 1)$$
$$= \begin{cases} \sin(x_{i2} - \hat{\beta}_{2kp}) \sin x_{i1}, & r_{i1} = r_{i2} = 0, \\ \sin(\hat{\beta}_{1kp} + \hat{b}_{kp}(x_{i2})) \dfrac{I_1(\hat{a}_{kp}(x_{i2}))}{I_0(\hat{a}_{kp}(x_{i2}))} \sin(x_{i2} - \hat{\beta}_{2kp}), & r_{i1} = 1,\ r_{i2} = 0, \\ \sin x_{i1} \sin \hat{\beta}_{2kp} \dfrac{I_1(\hat{a}_{kp}(x_{i1}))}{I_0(\hat{a}_{kp}(x_{i1}))}, & r_{i1} = 0,\ r_{i2} = 1, \end{cases}$$

and where $A_{tk2p}, B_{tk2p}, C_{tk2p}$ and $D_{tk2p}$ can be derived in a similar way, by exchanging $x_1$ with $x_2$. Because the sine model belongs to the exponential family, the derivatives on the right-hand side of the equations are given by

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

and, respectively, indicate the conditional expectations of $\cos(x_1 - \beta_{1k})$, $\cos(x_2 - \beta_{2k})$ and $\sin(x_1 - \beta_{1k}) \sin(x_2 - \beta_{2k})$ under the hidden state $k$.

In the case of fully observed, independent profiles, these equations reduce to those used in [10] for estimating mixtures of toroidal densities. We solve system (17) iteratively, by solving first the third equation and plugging the estimate of $\beta_{12k}$ in the first and second equations to obtain the estimates of the concentration parameters. We finally solve the last two equations to find the updated estimates of the directional means. Remarkably, this part of the EM algorithm is computationally fast and numerically stable, if an efficient method to compute Bessel functions (such as the `BesselI` function of R) is available.

It is well known that the EM algorithm can get stuck in 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. In the classification literature, 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 [11], by running the EM algorithm from a number of random initializations, stopping the iterations without waiting for full convergence. Because convergence to spurious maxima is fast (a phenomenon that is well documented in the case of mixtures of multivariate normal densities), it can be detected within those short EM runs, by monitoring both the class proportions $\hat{\pi}_{kp}$ and the eigenvalues of the covariance matrices:

$$\begin{pmatrix} \hat{\beta}_{11kp} & \hat{\beta}_{12kp} \\ \hat{\beta}_{12kp} & \hat{\beta}_{22kp} \end{pmatrix}^{-1}.$$

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. This strategy does not avoid convergence to local maxima and must be repeated several times.

## 4. Variance estimation

We computed the standard errors of the parameter estimates using parametric bootstrap, as standard errors based on the observed information matrix are often unstable (see, e.g. [12]). Specifically, we

re-fitted the model to the bootstrap data that were simulated from the estimated model. Simulation of a BSHMM is straightforward. We first simulate a sequence of states from the Markov chain. Given a sequence of states, a bivariate circular observation is at each time $t$ drawn according to the appropriate sine density, evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_{k_t}$, where $k_t$ is the state that has been drawn at time $t$. To obtain a bivariate circular sample, we first draw a sample $x_1$ from marginal density (4) with $\boldsymbol{\beta} = \boldsymbol{\beta}_{k_t}$ and then use $x_1$ to draw a value from the conditional distribution of $x_2$ given $x_1$, according to the von Mises density (6). There are several well-known routines for sampling from a von Mises distribution. Samples from the marginal circular density can be instead obtained by exploiting an acceptance–rejection algorithm, with a von Mises density as a proposal. Specifically, we first compute

$$\hat{\beta}_{11} = \max_{x\in(0,2\pi)} \left\{ \frac{f(x;\boldsymbol{\beta}_1)}{f_{\mathrm{vm}}(x;\beta_1,\beta_{11})} \right\},$$

where $f(x;\boldsymbol{\beta}_1)$ is the marginal distribution of $x_1$. We then evaluate

$$\hat{p} = \frac{1}{\max_{x\in(0,2\pi)}\{f(x;\boldsymbol{\beta}_1)/f_{\mathrm{vm}}(x;\beta_1,\hat{\beta}_{11})\}}.$$

Finally, we draw a sample $u$ from the uniform distribution on the unit interval and a sample $y$ from the von Mises $f_{\mathrm{vm}}(x;\beta_1,\hat{\beta}_{11})$. If

$$u \le \left\{ \frac{f(y)}{(1/\hat{p})f_{\mathrm{vm}}(y;\beta_1,\beta_{11})} \right\},$$

then we accept $y$ as $x_1$, otherwise, we reject the value and repeat the procedure.

Model re-fitting was repeated $R$ times, and the approximate standard error of each model parameter $\theta$ was computed by

$$\mathrm{se}_R(\hat{\theta}) = \left( \frac{1}{R-1} \sum_{r=1}^{R} (\hat{\theta}_r - \bar{\theta}_R)^2 \right)^{1/2},$$

where $\hat{\theta}_r$ is the estimate from the $r$th bootstrap sample and $\bar{\theta}_R$ is the mean of the bootstrap estimates.

In a general framework, there are at least three different methods for computing the standard errors of HMM parameters, namely likelihood profiling, bootstrapping and a method based on a finite difference approximation to the Hessian [13]. In this paper, we adopt the parametric bootstrap approach generating bootstrap samples according to the parametric model using the maximum likelihood estimates of the parameters. Our choice is due to both the simplicity of implementing the parametric bootstrap and the results produced by this procedure. As shown in [13], in the context of long time series computing, the exact Hessian is not feasible and, via a simulation study, it can be proved that likelihood profiling and bootstrapping produce similar results, whereas the standard errors from the finite-difference approximation of the Hessian are mostly too small.

## 5. Application

We illustrate the maximum likelihood estimation of a BSHMM using a time series of semi-hourly wind and wave directions, taken during the period 1 January 2010–21 February 2010 by the buoy of Ancona, which is located in the Adriatic Sea at about 30 km from the coast. Because of transmission errors, about 20% of the wind directions are missing and about 12% of the wave directions are missing. Finally, about 4% of the observed profiles are fully missing.

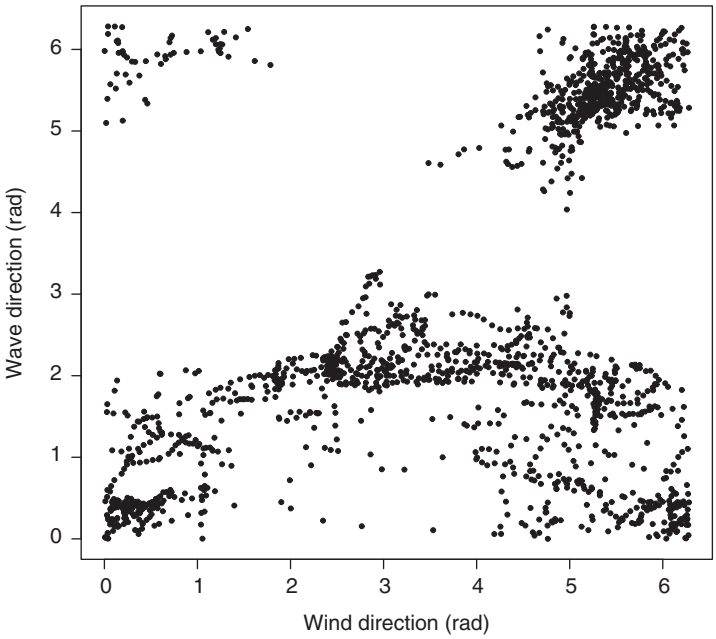


Figure 1. Adriatic Sea: wind and wave directions, taken by a buoy in wintertime.

Figure 1 displays the scatter plot of the data. Point coordinates indicate the direction (in radians) *from* which winds blow and waves travel. For simplicity, these bivariate circular data are plotted on the plane, although data points are actually on a torus.

Although in the ocean, wind and wave directions are strongly correlated, this is not necessarily true in the Adriatic Sea, due to the complex orography of the basin. The Adriatic Sea 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. In winter, relevant wind events in the Adriatic Sea are typically generated by the Bora wind, which in the Ancona area blows north–northwesterly along the major axis of the basin, and by the Sirocco wind, which blows southeasterly. Waves generated by these winds travel in the same direction of the winds or slightly rotate along the major axis of the basin. In addition, there are winds that blow northwesterly, westerly and southwesterly from the Italian coast, along the minor axis of the basin. Coastal winds generate synchronized waves only when the waves travel unobstructed, that is, either northwesterly or southeasterly, along the major axis of the basin. In the case of western winds, waves travel southwesterly. When, however, coastal winds rotate clockwise, waves tend to travel from north. This explains the clusters shown in Figure 1 and suggests the occurrence of a number of latent wind–wave regimes. Estimation of an HMM from these data can be helpful in clustering the data into a number of toroidal clusters, each associated with a specific wind–wave distribution, and assessing the temporal persistence of these regimes, accounting for the temporal autocorrelation in the classification process.

We estimated a number of BSHMMs and selected a model with $K = 3$ components by the Bayesian information criterion (BIC). Maximum likelihood estimates were computed by exploiting the short-run strategy, illustrated in Section 3. EM short runs were started from random initialization points and stopped as soon as the relative difference of the likelihood

$$\frac{L(\hat{\boldsymbol{\theta}}_p) - L(\hat{\boldsymbol{\theta}}_{p-1})}{L(\hat{\boldsymbol{\theta}}_p) - L(\hat{\boldsymbol{\theta}}_0)}$$

Table 1. Estimates and standard errors of the toroidal distribution parameters.

| Parameter | Component | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| $\beta_{1k}$ (wave mean direction) | 1.476(0.028) | 5.660 (0.015) | 1.203 (0.015) |
| $\beta_{2k}$ (wind mean direction) | 4.993(0.030) | 5.499 (0.014) | 1.470 (0.017) |
| $\beta_{11k}$ (wave directional concentration) | 2.373(0.102) | 8.128 (0.392) | 3.910 (0.154) |
| $\beta_{22k}$ (wind directional concentration) | 2.261(0.108) | 8.367 (0.416) | 1.980 (0.099) |
| $\beta_{12k}$ (wind/wave directional dependence) | −0.823(0.129) | 5.100 (0.373) | 6.002 (0.182) |

was less than $10^{-3}$. Divergence towards spurious solutions was normally detected within the first 10 iterations of the algorithm. Otherwise, EM short runs typically required between 40 and 60 iterations, depending on the dimension $K$ of the model and the starting point of the algorithm. The subsequent long-run EM run typically required between 130 and 200 iterations to reach convergence (we stopped the algorithm when the log-likelihood difference between successive iterations was less than $10^{-6}$). Standard errors were computed on the basis of 250 bootstrap samples.

Table 1 displays the estimates of the parameters of the three toroidal densities and Figure 2 shows the shapes of the related distributions. We can observe that the dependence parameters of the three densities are all significant at a 95% level, confirming that in this data set, the CI assumption is difficult to motivate. In addition, the estimated transition probability matrix (Table 2) is essentially diagonal, suggesting that the assumption of independent samples (i.e. a transition probability matrix with equal rows) is, in this example, unrealistic.

The model clusters the data into well-separated groups, which can be interpreted as latent wind–wave regimes. Components 2 and 3 are, respectively, associated with Bora and Sirocco events and cluster highly correlated data. In particular, component 3 (Figure 2, bottom) is bimodal and the two modes are, respectively, located at NW and SE, which are the two opposite directions of the major axis of the basin. On the other side, component 2 (Figure 2, middle) captures the northern wind and wave directions that are highly synchronized and concentrated around their modes. Component 1 is instead associated with coastal winds, which generate waves that tend to travel along the major axis of the basin. As a result, waves travel in a direction that is weakly correlated with the wind direction. Overall, the model describes the plasticity of the wind–wave interaction in the Adriatic Sea, indicating that the joint distribution of wind and wave data changes under different environmental regimes. Regime switching changes not only the modal directions and concentrations around these modes but also, and more interestingly, the correlation structure of the data. As a result, on the one side, the (marginal) weak correlation between wind and wave directions is explained by the presence of coastal winds (component 1). On the other side, the model indicates that the wind direction is an accurate predictor of the wave direction during Bora and Sirocco episodes, but that the level of accuracy decreases in the presence of coastal winds. In summary, wind directions should not be used to predict wave directions, without accounting for the latent, environmental heterogeneity of the data under study.

Table 3 compares the overall fit of the BSHMM with that obtained when estimating more parsimonious models, such as a mixture of conditionally independent von Mises densities, a mixture of bivariate sine densities and a CI-based bivariate HMM, respectively, described by Equations (10), (9) and (8) given in Section 2. As expected, the log-likelihood dramatically increases as correlations across variables and along time are progressively introduced. However, the monotonic increase of the log-likelihood is only partially explained by the increasing number of parameters, as shown by the BIC statistic, which suggests the least parsimonious BSHMM as

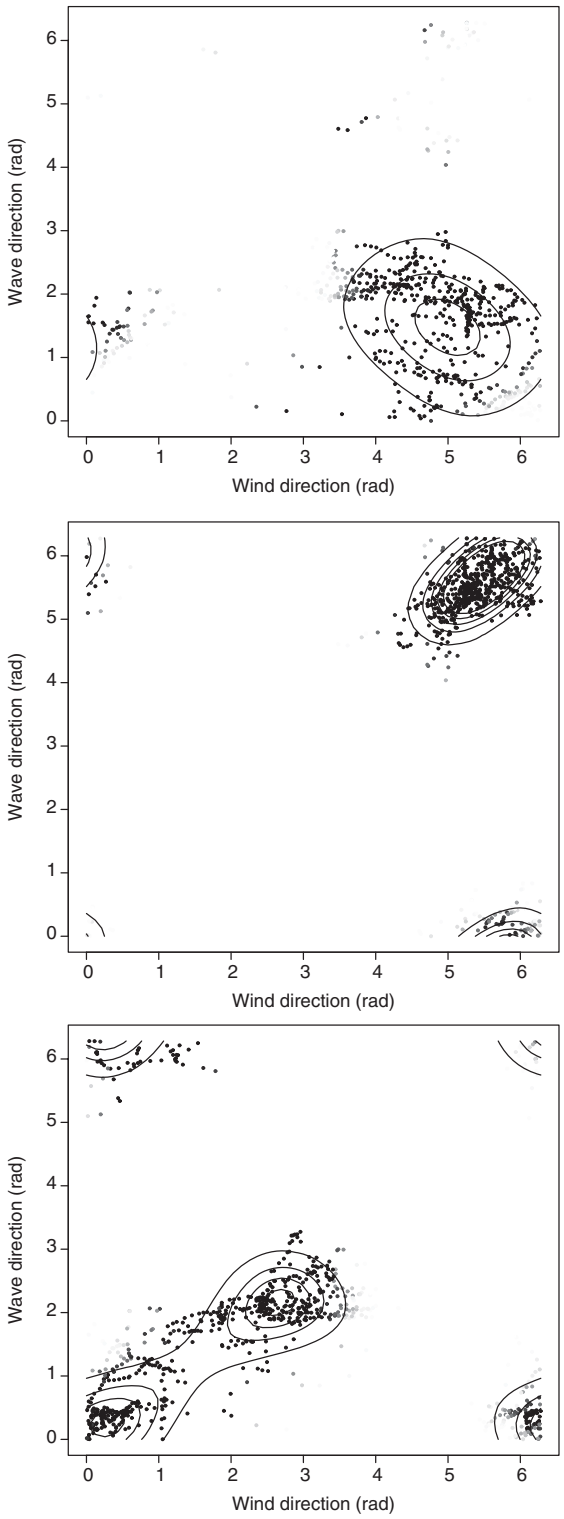


Figure 2. Contour plots of the three toroidal densities, as estimated by a three-component bivariate circular HMM; points are filled on a grey level scale according to their posterior probability of class membership.

Table 2. Estimates and standard errors of the transition probability matrix.

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 0.946 (0.009) | 0.021 (0.006) | 0.033 (0.006) |
| 2 | 0.014 (0.004) | 0.976 (0.005) | 0.010 (0.004) |
| 3 | 0.026 (0.005) | 0.008 (0.003) | 0.966 (0.006) |

Table 3. Overall fit and classification output.

| Model | Log-likelihood | BIC | Classification entropy |
|---|---|---|---|
| Mixture of conditionally independent von Mises densities | −6403.01 | 12,915.56 | 1256.28 |
| Mixture of bivariate sine densities | −5792.71 | 11,718.44 | 1148.90 |
| CI-based HMM | −4465.84 | 9088.17 | 228.35 |
| Bivariate sine HMM | −3947.80 | 8052.09 | 140.10 |

the model of choice. Table 3 furthermore includes the values obtained by the entropy index:

$$-\sum_{t=0}^{T}\sum_{k=1}^{K}\hat{\pi}_{tk}\log\hat{\pi}_{tk},$$

computed by summing up the posterior classification probabilities, obtained by the four models. This index increases with classification uncertainty, reaching a maximum when $\hat{\pi}_{tk} = K^{-1}$ for each time $t$. Table 3 shows that ignoring correlations typically leads to an unnecessary large uncertainty in the final classification. On the contrary, the progressive introduction of correlations across variables and along time strongly influences the final classification, reducing classification uncertainty.

## 6. Discussion

In this paper, we have introduced a new HMM for the analysis of bivariate time series of circular data. The model is useful for clustering biological and environmental data in a number of latent classes or regimes, associated with different toroidal distributions. Clustering is carried out by accounting for both the temporal autocorrelation of the data and the special structure of bivariate circular data, which are wrapped around a torus. The model is based on the sine model, introduced in [5]. The advantages of this distributional choice include a simple specification of the dependence structure between variables and the computational feasibility of a mixture-based classification strategy, where missing values can be efficiently handled within a likelihood framework. In addition, temporal transitions between regimes are addressed by means of simple Markov transition probabilities. An application to marine data motivates the model as a general tool for studying bivariate circular time series and demonstrates that the introduction of relevant correlations in a segmentation model may significantly reduce the uncertainty of the final classification.

We have focused on the maximum likelihood estimation from incomplete time series and proposed an EM algorithm where missing values and hidden states are treated as different sources of incomplete information. The expected values of complete-data sufficient statistics and hidden states are evaluated simultaneously through closed-form expressions. The M step of the algorithm

hence reduces to a battery of score equations that are available in a closed form. As a result, the EM algorithm is computationally feasible and numerically stable. Standard importance-sampling strategies allow the straightforward simulation of the model and, as a result, estimate uncertainty can be easily assessed by bootstrap standard errors.

## References


[1] H. Holzmann, A. Munk, M. Suster, and W. Zucchini, *Hidden Markov models for circular and linear-circular time series*, Environ. Ecol. Stat. 13 (2006), pp. 325–347. doi:10.1007/s10651-006-0015-7.

[2] 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.

[3] T. Edgoose and L. Allison, *MML Markov classification of sequential data*, Stat. Comput. 9 (1999), pp. 269–278. doi:10.1023/A:1008907921792.

[4] F. Lagona, A. Maruotti, and M. Picone, *A non-homogeneous hidden Markov model for the analysis of multi-pollutant exceedances data*, in *Hidden Markov Models, Theory and Applications*, P. Dymarsky, ed., InTech, Rijeka, Croatia, 2011, Chap. 10, pp. 207–222.

[5] H. Singh, V. Hnizdo, and E. Demchuk, *Probabilistic model for two dependent circular variables*, Biometrika 89 (3) (2002), pp. 719–723.

[6] 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.

[7] F. Lagona and M. Picone, *Model-based clustering of multivariate skew data with circular components and missing values*, J. Appl. Stat. (to appear), pp. 1–19. doi:10.1080/02664763.2011.626850. Available at http://www.tandfonline.com/doi/abs/10.1080/02664763.2011.626850.

[8] S. Kato and K. Shimizu, *Dependent models for observations which include angular ones*, J. Statist. Plann. Inference 138 (2008), pp. 3538–3549. Special Issue in Honor of Junjiro Ogawa (1915–2000): Design of Experiments, Multivariate Analysis and Statistical Inference.

[9] D. Rubin, *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York, 1987.

[10] 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.

[11] 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.

[12] G. McLachlan and D. Peel, *Finite Mixture Models*, Wiley, New York, 2000.

[13] I. Visser, M. Raijmakers, and P. Molenaar, *Confidence intervals for hidden Markov model parameters*, British J. Math. Statist. Psych. 53 (2000), pp. 317–327.

# A Multivariate Hidden Markov Model for the Identification of Sea Regimes from Incomplete Skewed and Circular Time Series

J. BULLA, F. LAGONA, A. MARUOTTI, and M. PICONE

The identification of sea regimes from environmental multivariate times series is complicated by the mixed linear–circular support of the data, by the occurrence of missing values, by the skewness of some variables, and by the temporal autocorrelation of the measurements. We address these issues simultaneously by a hidden Markov approach, and segment the data into pairs of toroidal and skew-elliptical clusters by means of the inferred sequence of latent states. Toroidal clusters are defined by a class of bivariate von Mises densities, while skew-elliptical clusters are defined by mixed linear models with positive random effects. The core of the classification procedure is an EM algorithm accounting for missing measurements, unknown cluster membership, and random effects as different sources of incomplete information. Moreover, standard simulation routines allow for the efficient computation of bootstrap standard errors. The proposed procedure is illustrated for a multivariate marine time series, and identifies a number of wintertime regimes in the Adriatic Sea.

**Key Words:** Circular data; EM algorithm; Hidden Markov model; Model-based clustering; Skewness; Unsupervised classification; Wave; Wind.

## 1. INTRODUCTION

A major goal in marine research is the development of models that help scientists to understand how air–sea interactions influence the sea surface. These models are useful in a variety of application areas, including studies of the drift of floating objects and

J. Bulla is Maitre de Conferénce at LMNO, CNRS UMR 6139, Université de Caen, Caen, France. F. Lagona (✉) is Associate Professor of Statistics at DIPES (E-mail: *lagona@uniroma3.it*), Università Roma Tre, via G. Chiabrera 199, 00145 Rome, Italy. A. Maruotti is Assistant Professor at DIPES, Università Roma Tre, Rome, Italy and Lecturer in Medical Statistics at the Southampton Statistical Sciences Research Institute & School of Mathematics, University of Southampton, Building 39, Southampton SO17 1BJ, UK. M. Picone is a Marine Engineer at the Marine Service, Institute for Environmental Protection and Research, Via Curtatone 3, 00185 Rome, Italy.


*Journal of Agricultural, Biological, and Environmental Statistics, Volume 17, Number 4, Pages 544–567*
*DOI: 10.1007/s13253-012-0110-1*

oil spills (Huang, Wing-Keung Law, and Huang 2011), the design of off-shore structures (Faltinsen 1990), and studies of sediment transport (Jin and Ji 2004) and coastal erosion (Pleskachevsky, Eppel, and Kapitza 2009). These applications are especially important in coastal areas and semi-enclosed basins, where wind–wave interactions are influenced by the orography of the site.

Studies of air–sea interactions involve the analysis of multivariate, often incomplete time series of marine data, that include hourly or semi-hourly measurements of mixed type variables, like wind and wave direction (i.e., circular variables) and wind speed and wave height (i.e., linear variables). These data are traditionally examined through numerical wind–wave models. Although well suited for oceans, numerical wind–wave models may provide inaccurate results under complex orography conditions (Bertotti and Cavalieri 2009). This has motivated the use of statistical models for the analysis of time series of wind and wave data (Monbet, Ailliot, and Prevosto 2007). Widely exploited are autoregressive and Markov-switching autoregressive models for wind speed (Ailliot and Monbet 2012), spectral models for the analysis of wave height time series in the frequency domain (Hamilton 2010; Reikard and Rogers 2011), and autoregressive and hidden Markov chain models for wind speed and direction (Holzmann et al. 2006). Most of these statistical models have been however specified in a univariate setting, where wind and wave time series are modeled separately. Multivariate extensions are challenging because they should account for a number of nonstandard features of the data, including correlation in time and across variables, mixed supports (circular and linear) of the data, the special nature of circular measurements, typically skewness of wind speed and wave height, and the occurrence of missing values.

We specify a multivariate hidden Markov model (HMM) by describing wind–wave data in terms of latent environmental regimes, i.e., specific distributions that the data take under latent environmental conditions. This approach is particularly convenient under complex orography conditions, such as closed basins or coastal areas, where the correlation structure of the data can be decomposed according to a finite number of easily interpretable distributions.

More precisely, we approximate the joint distribution of the data by a mixture of multivariate densities, each specified as the product of a bivariate von Mises and a bivariate skew-normal density, the parameters of which depend on the states of a latent Markov chain. In this setting, wind and wave directions are segmented by toroidal clusters, while bivariate observations of wind speed and wave height are clustered within skewed ellipses. This allows clustering mixed linear and circular data separately, avoiding the definition of possibly hardly interpretable hyper-cylindrical clusters. In addition, toroidal and skew-elliptical clusters are paired according to the states of the Markov chain, which can be hence interpreted as latent environmental regimes. Finally, the transition probabilities matrix of the Markov chain captures regime-switching in time, accounting for temporal autocorrelation. Taking a multivariate HMM approach to classification, observations are clustered according to the latent state that is conditionally expected each time, given the observed data. As a result, classification is not only based on similarities in the variables space, but also on similarities that occur in a temporal neighborhood. This can be particularly useful when clustering incomplete time-series data, a specific issue in our case

study, because the missing information in the variables space can be partially recovered by similarities observed in time.

Parametric families of bivariate von Mises (Singh, Hnizdo, and Demchuk 2002) and skew-normal distributions (Sahu, Dey, and Branco 2003) have been recently exploited as mixture components in classification studies (Lin 2009; Mardia, Taylor, and Subramaniam 2007; Cabral, Lachos, and Prates 2012; Lagona and Picone 2012b). Within this strand of literature, however, model-based clustering is usually developed by assuming that the data are in the form of independent multivariate samples. The independence assumption is a shortcoming when classification is based on data collected in the form of multivariate time series, such as in marine classification studies, because a clustering procedure should account for the potential redundancy of the data resulting from temporal autocorrelation. An HMM-based approach provides a natural extension of mixture models to allow for temporal dependence (Cappé, Moulines, and Rydén 2005). However, the literature on HMM-based classification studies is dominated by Gaussian HMMs for multivariate continuous data. Multivariate time series with non-normal components of mixed type are instead traditionally modeled by approximating the joint distribution of the data with a mixture having products of univariate probability distributions as components, therefore assuming that measurements in a multivariate profile are conditionally independent, given the states of the hidden Markov chain. In the context of linear–circular data, this approach to the specification of an HMM has already been proposed for the analysis of bivariate time series with one circular and one linear component (Holzmann et al. 2006). Conditional independence facilitates both the specification and the estimation of a bivariate non-normal HMM, and can be motivated by borrowing arguments from the latent-class literature. However, the number of the observed variables is often larger than two in marine classification studies, and conditional independence should be used with parsimony: products of univariate distributions may be too restrictive to accommodate for the complex shape of multivariate clusters and therefore an unnecessary large number of latent classes (states) may be required to obtain a reasonable goodness of fit (Lagona and Picone 2011). A large number of states is acceptable to some extent if the purpose of an analysis is density estimation. It nevertheless complicates the interpretation of the results, especially when the purpose is to cluster the data into meaningful groups. This has motivated a number of efforts in order to relax the conditional independence assumption in non-normal HMMs at least partially. Noticeable examples in the analysis of categorical data are provided by Zucchini and Guttorp (1991) and Zhang et al. (2010) as well as Lagona and Picone (2012a) in the case of circular data. We extend this strand of the literature in the context of multivariate, incomplete, mixed linear–circular data by allowing for correlation within latent classes of circular and linear observations.

The rest of the paper is organized as follows. Section 2 summarizes some relevant details on the data that motivated this work. In Section 3 we illustrate the multivariate HMM, exploited for the unsupervised classification of the data, while Section 4 is devoted to the computational aspects arising in the estimation of model parameters and standard errors (technical details are postponed to the Appendix at the end of the paper). Section 5 summarizes the results obtained when applying the model to marine data recorded in the Adriatic Sea. A list of relevant discussion points is finally included in Section 6.

Table 1. Missing values patterns.

| Count | Wave direction | Wave height | Wind speed | Wind direction |
|---|---|---|---|---|
| 1043 | obs | obs | obs | obs |
| 75 | mis | obs | obs | obs |
| 77 | obs | obs | obs | mis |
| 75 | obs | mis | obs | obs |
| 80 | obs | obs | mis | obs |
| 7 | mis | obs | obs | mis |
| 10 | mis | mis | obs | obs |
| 3 | obs | mis | obs | mis |
| 2 | mis | obs | mis | obs |
| 73 | obs | obs | mis | mis |
| 3 | obs | mis | mis | obs |
| 1 | mis | mis | obs | mis |
| 5 | mis | obs | mis | mis |
| 8 | obs | mis | mis | mis |
| 39 | mis | mis | mis | mis |

## 2. WINTERTIME REGIMES IN THE ADRIATIC SEA

The data that motivated this work are time series of semi-hourly wave and wind directions, as well as wind speeds and wave heights, recorded in the period 12/12/2009–12/1/2010 by the buoy of Ancona, located in the Adriatic Sea at about 30 km from the coast. Compared to outputs of meteorological models, known to be smoother than in situ observations and to underestimate extreme events (Caires and Sterl 2005; Izquierdo and Guedes Soares 2005), buoy data are of better quality but often include missing values.

Table 1 displays the distribution of the missing values patterns, observed during the study period. About 30 % of the profiles includes at least one missing value, while a small portion of the sample (about 2 %) includes fully unobserved profiles. 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 often occur because of transmission errors or malfunctioning of the device. As buoys are normally constructed to transmit data even in the case of severe environmental conditions, missing values in marine studies are often missing completely at random (MCAR), i.e., the missingness probability depends neither on observed nor on unobserved data. Assuming that the data are missing at random (MAR), we relax the MCAR assumption and allow for the missingness probability to depend on the observed data, which seems reasonable for marine data that are collected in semi-enclosed seas such as the Adriatic Sea, where severe environmental conditions seldom occur.

Figure 1 displays the univariate distributions of the data. The salient features of these histograms can be interpreted by recalling that the Adriatic Sea is a semi-enclosed, long narrow basin, bordered by mountains on three sides, and extending for about 800 km along the major axis from SE to NW, with a width of about 200 km. In wintertime, relevant wind events in the Adriatic Sea are typically generated by the southeastern Sirocco, the northern Bora and the northwestern Maestral. These conditions can be associated with the three

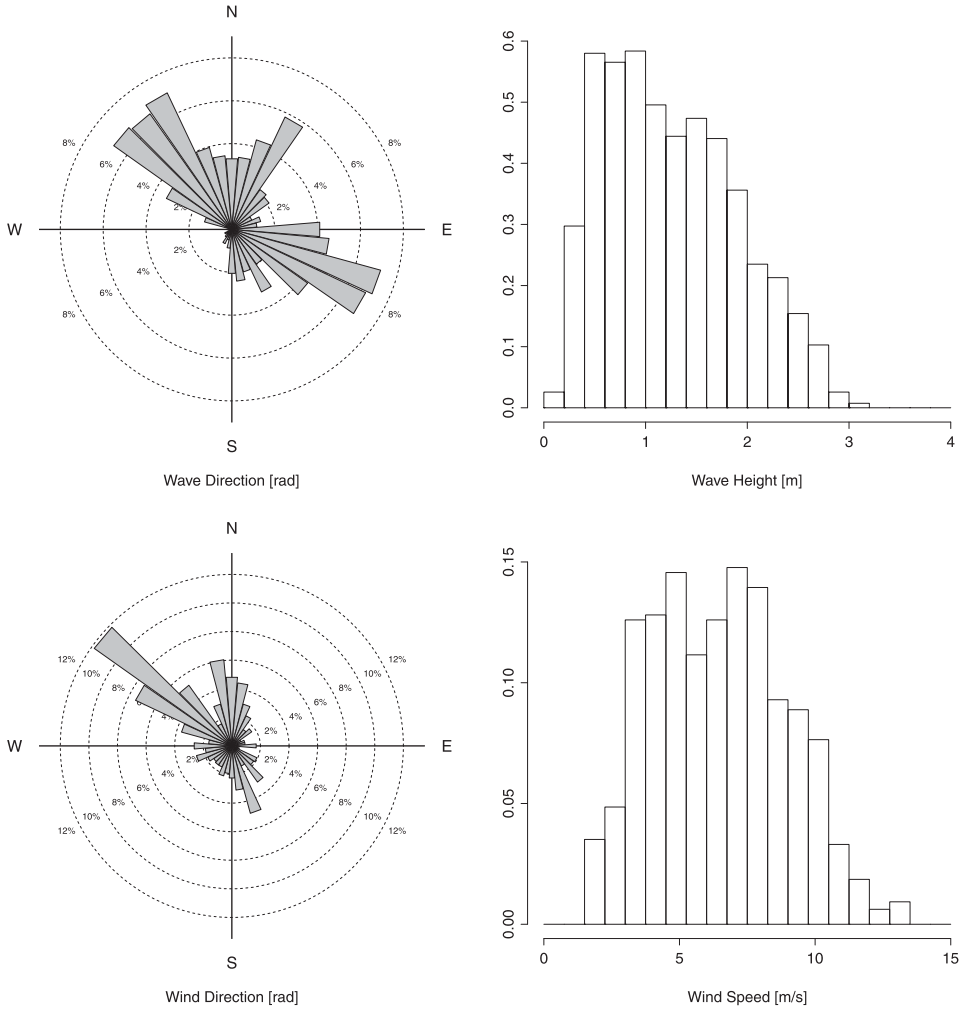


Figure 1. Univariate distributions of the circular (left) and linear (right) variables, observed during the study period. Circular histograms display the frequency of directions from which the winds blow and the waves travel, respectively (the area of each sector is proportional to the group frequency).

modes of the bottom-left circular histogram of wind directions in Figure 1. Sirocco arises from a warm, dry, tropical air mass that is pulled northwards by low-pressure cells moving eastwards across the Mediterranean Sea. By contrast, Bora episodes occur when a polar high-pressure area sits over the snow-covered mountains of the interior plateau behind the coastal mountain range and a calm low-pressure area lies further south over the warmer Adriatic. Finally, the Maestral is a sea-breeze wind blowing northwesterly when the east Adriatic coast gets warmer than the sea. While Bora and Sirocco episodes are usually associated with high-speed flows, Maestral is in general linked with good meteorological conditions. Hence, the marginal distribution of wind speed (Figure 1, bottom-right) may be interpreted as the result of mixing different wind-speed regimes.

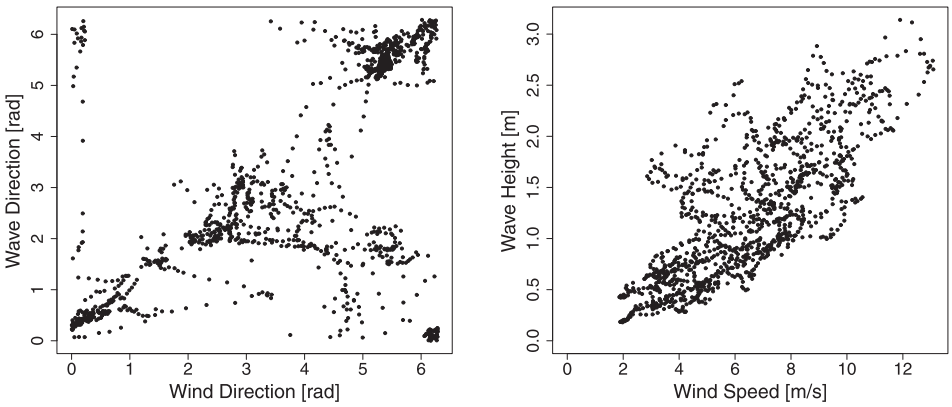


Figure 2. Wind and wave direction (left) as well as wind speed and wave height (right), observed at a buoy of the Adriatic Sea in wintertime. Points on the left-hand side picture indicate the directions from which wind blows and wave travels (North = 0 rad).

Height and direction of waves are only partially influenced by wind conditions in semi-enclosed basins. The orography of the Adriatic Sea plays a key role in this case and most of the waves tend to travel from north-northwest and southeasterly along the major axis of the basin, where they can travel freely, without being obstructed by physical obstacles, such as coastlines. As a result, the marginal distributions of wind and wave directions (Figure 1, left-column pictures) are poorly synchronized. This complicates the interpretation of the marginal distribution of wave heights (Figure 1, top-right). While this distribution is likely the outcome of mixing different sea conditions, the association structure between wind and wave regimes is difficult to visualize.

In traditional wave atlases, marine data are typically depicted in terms of univariate distributions, such as in Figure 1. The complex wind–wave interaction structure in the Adriatic Sea is, however, better shown by Figure 2, which displays the scatter plots of the circular and the linear available observations. For simplicity, bivariate circular data are plotted on the plane, although data points are actually on a torus. Point coordinates are measured in radians; 0 and $2\pi$ indicate North.

Although a number of patterns appear in these scatter plots, their interpretation is difficult due to the weak correlation of the circular measurements and the skewness of the linear observations. Weak correlation and skewness are traditionally explained as the result of the complex orography of the Adriatic Sea and often held responsible for the inaccuracy of numerical wind–wave models (Bertotti and Cavalieri 2009). Nevertheless, the observations might result from mixing of a number of latent environmental regimes, conditionally on which the distribution of the data takes a shape that is easier to interpret than the shape taken by the marginal distributions. From a technical viewpoint, the identification of relevant regimes is complicated by the toroidal nature of the data in the left-hand plot, and by the skewness of the data in the right-hand plot. By taking an HMM approach, we cluster directional and planar data separately to account for the different nature of the data, and simultaneously pair these clusters into a number of latent classes evolving in time according to a Markov chain, being interpretable as time-varying regimes of air–sea interactions.

## 3. A MULTIVARIATE HIDDEN MARKOV MODEL

The data considered in this paper are in the form of a time series $z_{0:T} = (z_t, t = 0, \ldots, T)$, with bivariate circular and linear components, say $z_t = (\boldsymbol{x}_t, \boldsymbol{y}_t)$, $\boldsymbol{x}_t = (x_{1t}, x_{2t}) \in (0, 2\pi]^2$ and $\boldsymbol{y}_t = (y_{1t}, y_{2t}) \in \mathbb{R}^2$. In mixture-based classification studies, class membership is conveniently treated as the value taken by a latent multinomial variable with one trial and $K$ classes. In HMM-based classification studies, the temporal evolution of class membership is driven by a latent Markov chain, which can be conveniently described as a multinomial process in discrete time. Accordingly, we introduce a sequence $\boldsymbol{\xi}_{0:T} = (\boldsymbol{\xi}_t, t = 0, \ldots, T)$ of multinomial variables $\boldsymbol{\xi}_t = (\xi_{t1} \ldots \xi_{tK})$ with one trial and $K$ classes, whose binary components represent class membership at time $t$. The joint distribution $p(\boldsymbol{\xi}_{0:T}; \boldsymbol{\pi})$ of the chain is fully known up to a parameter $\boldsymbol{\pi}$ that includes $K$ initial probabilities $\pi_k = P(\xi_{0k} = 1),\ k = 1, \ldots, K,\ \sum_k \pi_k = 1$, and $K^2$ transition probabilities $\pi_{hk} = P(\xi_{tk} = 1 | \xi_{t-1,h} = 1),\ h, k = 1, \ldots, K,\ \sum_k \pi_{hk} = 1$. Formally, we assume that

$$p(\boldsymbol{\xi}_{0:T}; \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{\xi_{0k}} \prod_{t=1}^{T} \prod_{h=1}^{K} \prod_{k=1}^{K} \pi_{hk}^{\xi_{t-1,h}\xi_{tk}}. \tag{3.1}$$

The specification of a multivariate HMM is completed by assuming that the observations are conditionally independent, given a realization of the Markov chain. As a result, the conditional distribution of the observed process, given the latent process, takes the form of a product density, say

$$f(z_{0:T} | \boldsymbol{\xi}_{0:T}) = \prod_{t=0}^{T} \prod_{k=1}^{K} \big(f_k(z_t)\big)^{\xi_{tk}},$$

where $f_k(z), k = 1, \ldots, K$ are $K$ multivariate densities. For classification purposes, these densities are usually assumed to be known up to a number of parameters that indicate the locations and the shapes of $K$ clusters. In the context of multivariate continuous data, multivariate normal distributions have been widely adopted. In the case of mixed linear and circular data, observations can be conveniently clustered by grouping linear and circular components into a number of toroidal and planar clusters, respectively, and then associating these clusters to $K$ latent classes. Accordingly, we assume that circular and linear observations are conditionally independent given a realization of the Markov chain, and introduce (1) a family of bivariate densities $f(\boldsymbol{x}; \boldsymbol{\beta})$ on the torus, indexed by a parameter $\boldsymbol{\beta}$ that indicates the location and the shape of a toroidal cluster, and (2) a family of bivariate densities on the plane, $f(\boldsymbol{y}; \boldsymbol{\gamma})$, indexed by a parameter $\boldsymbol{\gamma}$, which indicates the location and the shape of a planar cluster. Formally, we assume that

$$f(z_{0:T} | \boldsymbol{\xi}_{0:T}) = \prod_{t=0}^{T} \prod_{k=1}^{K} \big(f(\boldsymbol{x}_t | \boldsymbol{\beta}_k) f(\boldsymbol{y}_t | \boldsymbol{\gamma}_k)\big)^{\xi_{tk}}. \tag{3.2}$$

Integrating $f(z_{0:T} | \boldsymbol{\xi}_{0:T}) p(\boldsymbol{\xi}_{0:T})$ with respect to $\boldsymbol{\xi}_{0:T}$, we obtain the marginal distribution of the observed data, known up to a parameter $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, on which our classification

procedure is based. In particular, we first maximize the likelihood function

$$L(\boldsymbol{\theta}; z_{0:T}) = \sum_{\boldsymbol{\xi}_{0:T}} p(\boldsymbol{\xi}_{0:T}; \boldsymbol{\pi}) f(z_{0:T}|\boldsymbol{\xi}_{0:T}), \tag{3.3}$$

and find the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$. Secondly, we cluster the data according to the posterior probabilities of class membership

$$\hat{p}_{tk} = P(\xi_{tk} = 1|z_{0:T}; \hat{\boldsymbol{\theta}}) = \mathbb{E}(\xi_{tk}|z_{0:T}; \hat{\boldsymbol{\theta}}), \tag{3.4}$$

based on $\hat{\boldsymbol{\theta}}$. The computational complexity of the estimation of both $\hat{\boldsymbol{\theta}}$ and $\hat{p}_{tk}$ depends on the choice of the bivariate densities which are used to model the circular and the linear components of the observations. We exploit bivariate von Mises and skew-normal densities, described below, as a compromise between numerical complexity and modeling flexibility.

The bivariate von Mises density in the form introduced by Singh, Hnizdo, and Demchuk (2002) is a parametric distribution on the torus, which naturally embeds the bivariate normal distribution when the range of observations is small. Its density is given by

$$f(\boldsymbol{x}; \boldsymbol{\beta}) = \frac{\exp(\beta_{11}\cos(x_1 - \beta_1) + \beta_{22}\cos(x_2 - \beta_2) + \beta_{12}\sin(x_1 - \beta_1)\sin(x_2 - \beta_2))}{C(\boldsymbol{\beta})}, \tag{3.5}$$

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

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

is the modified Bessel function of order $m$.

This density can be viewed as a bivariate generalization of the von Mises distribution, where $\beta_{12}$ accounts for the statistical dependence between $x_1$ and $x_2$. The two univariate marginal densities

$$f(x_i; \boldsymbol{\beta}) = \int_0^{2\pi} f(\boldsymbol{x}; \boldsymbol{\beta})\, dx_j = \frac{2\pi}{C(\boldsymbol{\beta})} I_0\big(a(x_i)\big) \exp\big(\beta_{ii}\cos(x_i - \beta_i)\big), \quad i = 1, 2, \tag{3.6}$$

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

$$a(x_i) = \big(\beta_{jj}^2 + \beta_{12}^2 \sin^2(x_i - \beta_i)\big)^{1/2}, \quad i, j = 1, 2,\ i \neq j. \tag{3.7}$$

For $\beta_{12} = 0$, $a(x_i) = \beta_{jj}$, $i = 1, 2$, and, as a result, $x_1$ and $x_2$ are independent and each of them follows a von Mises distribution with marginal mean angles $\beta_i$ and marginal concentrations $\beta_{ii}$. The conditional distributions

$$f(x_i|x_j; \boldsymbol{\beta}) = \frac{f(\boldsymbol{x}; \boldsymbol{\beta})}{f(x_j; \boldsymbol{\beta})} = \frac{\exp(a(x_i)\cos(x_i - \beta_i - b(x_j)))}{2\pi I_0(a(x_i))}, \quad i, j = 1, 2,\ i \neq j, \tag{3.8}$$

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

$$b(x_j) = \arctan\left(\frac{\beta_{12}}{\beta_{jj}} \sin(x_j - \beta_j)\right). \tag{3.9}$$

A bivariate skew-normal distribution is employed to define skew-elliptical clusters of wind speeds and wave heights. Following Lin (2009), we specify a bivariate skew-normal density as a linear mixed model with positive random effects. More precisely, let $\varphi(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the bivariate normal density function with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We first introduce a bivariate random effect $\boldsymbol{v} = (v_1, v_2)$ with independent components, distributed according to two standard normal distributions truncated at 0, say

$$\boldsymbol{v} \sim f(\boldsymbol{v}) = \frac{\varphi(\boldsymbol{v}; \boldsymbol{0}, \boldsymbol{I})}{\int_{(0,+\infty)^2} \varphi(\boldsymbol{u}; \boldsymbol{0}, \boldsymbol{I})\, d\boldsymbol{u}} = \frac{2}{\pi} \exp\left(-\frac{1}{2}\boldsymbol{v}^\top \boldsymbol{v}\right), \quad \boldsymbol{v} \in [0, +\infty)^2. \tag{3.10}$$

Second, we assume that $\boldsymbol{y}$ follows a bivariate normal distribution conditionally on $\boldsymbol{v}$,

$$f(\boldsymbol{y}|\boldsymbol{v}; \boldsymbol{\gamma}) = \varphi\big(\boldsymbol{y}; \boldsymbol{\mu}(\boldsymbol{v}; \boldsymbol{\gamma}), \boldsymbol{\Sigma}(\boldsymbol{\gamma})\big), \tag{3.11}$$

with mean

$$\mu(\boldsymbol{v}; \boldsymbol{\gamma}) = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} + \begin{pmatrix} \gamma_1' & 0 \\ 0 & \gamma_2' \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

and covariance matrix

$$\boldsymbol{\Sigma}(\boldsymbol{\gamma}) = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{12} & \gamma_{22} \end{pmatrix}.$$

In this setting, a bivariate skew-normal distribution is obtained as

$$f(\boldsymbol{y}; \boldsymbol{\gamma}) = \int_{(0,+\infty)^2} \varphi(\boldsymbol{y}|\boldsymbol{v}; \boldsymbol{\gamma}) f(\boldsymbol{v})\, d\boldsymbol{v} \tag{3.12}$$

and reduces to a bivariate normal distribution when the skewness parameters $\gamma_1' = \gamma_2' = 0$. Otherwise, the skewness parameters perturb both the mean and the covariance matrix of $\boldsymbol{y}$ as follows:

$$\mathbb{E}\boldsymbol{y} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} + \sqrt{\frac{2}{\pi}} \begin{pmatrix} \gamma_1' \\ \gamma_2' \end{pmatrix},$$

$$\mathbb{E}(\boldsymbol{y} - \mathbb{E}\boldsymbol{y})(\boldsymbol{y} - \mathbb{E}\boldsymbol{y})^\top = \boldsymbol{\Sigma}(\boldsymbol{\gamma}) + \left(1 - \frac{2}{\pi}\right) \begin{pmatrix} \gamma_1' & 0 \\ 0 & \gamma_2' \end{pmatrix}^2.$$

In our HMM, we use a family of $K$ bivariate von Mises densities $f(\boldsymbol{x}|\boldsymbol{\beta}_k)$, indexed by the five parameters $\boldsymbol{\beta}_k = (\beta_{1k}, \beta_{2k}, \beta_{11k}, \beta_{22k}, \beta_{12k})$, to define $K$ toroidal clusters of wind and wave directions, centered at $(\beta_{1k}, \beta_{2k})$ and shaped by the parameters $(\beta_{11k}, \beta_{22k}, \beta_{12k})$. $K$ skew-normal densities, indexed by $K$ vectors $\boldsymbol{\gamma}_k$, serve to define $K$ skew-elliptical clusters of wind speed and wave height.

## 4. LIKELIHOOD INFERENCE

As our data are in the form of incomplete profiles, we refer to $\boldsymbol{x}_{t,\text{mis}}$ and $\boldsymbol{x}_{t,\text{obs}}$, respectively, as the missing and observed circular components at time $t$. Analogously, $\boldsymbol{y}_{t,\text{mis}}$ and $\boldsymbol{y}_{t,\text{obs}}$, respectively, represent the missing and observed linear components at time $t$. Accordingly, $z_{t,\text{mis}} = (\boldsymbol{x}_{t,\text{mis}}, \boldsymbol{y}_{t,\text{mis}})$ and $z_{t,\text{obs}} = (\boldsymbol{x}_{t,\text{obs}}, \boldsymbol{y}_{t,\text{obs}})$, respectively, indicate the missing and observed parts of an observation at time $t$. In this setting, we define $z_{0:T,\text{obs}} = (z_{t,\text{obs}}, t = 0, \ldots, T)$ and $z_{0:T,\text{mis}} = (z_{t,\text{mis}}, t = 0, \ldots, T)$. If the data are missing at random (MAR), the missing data mechanism can be ignored and the maximum likelihood estimate of parameter $\boldsymbol{\theta}$ is the maximum point of the marginal likelihood function

$$L(\boldsymbol{\theta}|z_{0:T,\text{obs}}) = \sum_{\boldsymbol{\xi}_{0:T}} p(\boldsymbol{\xi}_{0:T}; \boldsymbol{\pi}) \prod_{t=0}^{T} \int f(z_t|\boldsymbol{\xi}_t; \boldsymbol{\beta}, \boldsymbol{\gamma})\, dz_{t,\text{mis}},$$

$$f(z_t|\boldsymbol{\xi}_t; \boldsymbol{\beta}, \boldsymbol{\gamma}) = f(\boldsymbol{x}_t|\boldsymbol{\xi}_t; \boldsymbol{\beta}) \int_{\boldsymbol{v}} f(\boldsymbol{y}_t|\boldsymbol{\xi}_t, \boldsymbol{v}; \boldsymbol{\gamma}) f(\boldsymbol{v})\, d\boldsymbol{v}, \tag{4.1}$$

which reduces to (3.3) in the case of complete data.

### 4.1. ESTIMATION

In order to estimate $\boldsymbol{\theta}$, we maximize $L(\boldsymbol{\theta})$ by using a version of the EM algorithm. EM algorithms are based on the definition of a complete-data log-likelihood function, obtained by considering the sampling distribution of both the observed and the unobserved quantities. As our HMM is a mixture which integrates circular densities and mixed-effects normal models, the unobserved quantities are not only the missing measurements, but also the unknown class memberships and the values taken by the skewness random effects. Treating all these unobserved quantities as missing values reflecting different sources of incomplete information, we define the complete-data log-likelihood function as follows:

$$\begin{aligned}\log L_{\text{comp}}(\boldsymbol{\theta}, \boldsymbol{\xi}_{0:T}, z_{0:T}, \boldsymbol{v}_{0:T}) = {} & \sum_{k=1}^{K} \xi_{0k} \log \pi_k + \sum_{t=1}^{T} \sum_{h=1}^{K} \sum_{k=1}^{K} \xi_{t-1,h} \xi_{t,k} \log \pi_{hk} \\ & + \sum_{t=0}^{T} \sum_{k=1}^{K} \xi_{tk} \log f(\boldsymbol{x}_t; \boldsymbol{\beta}_k) \\ & + \sum_{t=0}^{T} \sum_{k=1}^{K} \xi_{tk} \big(\log f(\boldsymbol{y}_t|\boldsymbol{v}_t; \boldsymbol{\gamma}_k) + \log f(\boldsymbol{v}_t)\big). \end{aligned} \tag{4.2}$$

The algorithm is iterated by alternating the expectation (E) and maximization (M) steps. Given the estimate $\hat{\boldsymbol{\theta}}_s$, obtained at the end of the $s$th iteration, the $(s+1)$th iteration is initialized by an E-step, which evaluates the expected value of (4.2) with respect to the conditional distribution of the missing values given the observed data. For the HMM, this distribution takes a complex, but tractable, form, because it factorizes as follows:

$$f(\boldsymbol{\xi}_{0:T}, z_{0:T,\text{mis}}, \boldsymbol{v}_{0:T}|z_{\text{obs}}; \hat{\boldsymbol{\theta}}_s) = p(\boldsymbol{\xi}_{0:T}|z_{0:T,\text{obs}}; \hat{\boldsymbol{\theta}}_s) f(z_{0:T,\text{mis}}, \boldsymbol{v}_{0:T}|\boldsymbol{\xi}_{0:T}, z_{0:T,\text{obs}}; \hat{\boldsymbol{\theta}}_s), \tag{4.3}$$

where

$$f(\boldsymbol{z}_{0:T,\text{mis}}, \boldsymbol{v}_{0:T}|\boldsymbol{\xi}_{0:T}, \boldsymbol{z}_{0:T,\text{obs}}; \hat{\boldsymbol{\theta}}_s) = f(\boldsymbol{x}_{0:T,\text{mis}}|\boldsymbol{\xi}_{0:T}, \boldsymbol{x}_{0:T,\text{obs}}; \hat{\boldsymbol{\beta}}_s) \times f(\boldsymbol{y}_{0:T,\text{mis}}|\boldsymbol{\xi}_{0:T}, \boldsymbol{y}_{0:T,\text{obs}}, \boldsymbol{v}_{0:T}; \hat{\boldsymbol{\gamma}}_s) \times f(\boldsymbol{v}_{0:T}|\boldsymbol{\xi}_{0:T}, \boldsymbol{y}_{0:T,\text{obs}}; \hat{\boldsymbol{\gamma}}_s)$$

and

$$f(\boldsymbol{x}_{0:T,\text{mis}}|\boldsymbol{\xi}_{0:T}, \boldsymbol{x}_{0:T,\text{obs}}; \hat{\boldsymbol{\beta}}_s) = \prod_{t=0}^{T}\prod_{k=1}^{K}\big(f(\boldsymbol{x}_{t,\text{mis}}|\xi_{tk}=1, \boldsymbol{x}_{t,\text{obs}}; \hat{\boldsymbol{\beta}}_{ks})\big)^{\xi_{tk}}, \quad (4.4)$$

$$f(\boldsymbol{y}_{0:T,\text{mis}}|\boldsymbol{\xi}_{0:T}, \boldsymbol{y}_{0:T,\text{obs}}, \boldsymbol{v}_{0:T}; \hat{\boldsymbol{\gamma}}_s) = \prod_{t=0}^{T}\prod_{k=1}^{K}\big(f(\boldsymbol{y}_{t,\text{mis}}|\xi_{tk}=1, \boldsymbol{y}_{t,\text{obs}}; \hat{\boldsymbol{\gamma}}_{ks})\big)^{\xi_{tk}}, \quad (4.5)$$

$$f(\boldsymbol{v}_{0:T}|\boldsymbol{\xi}_{0:T}, \boldsymbol{y}_{0:T,\text{obs}}; \hat{\boldsymbol{\gamma}}_s) = \prod_{t=0}^{T}\prod_{k=1}^{K}\big(f(\boldsymbol{v}_{t,\text{mis}}|\xi_{tk}=1, \boldsymbol{y}_{t,\text{obs}}; \hat{\boldsymbol{\gamma}}_{ks})\big)^{\xi_{tk}}. \quad (4.6)$$

Each distribution $f(\boldsymbol{x}_{t,\text{mis}}|\xi_{tk}=1, \boldsymbol{x}_{t,\text{obs}}; \hat{\boldsymbol{\beta}}_{ks})$ in (4.4) is equal to 1 if the observed profile at time $t$ is complete; it is otherwise equal to the bivariate von Mises (3.5), evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{ks}$, if both measurements at time $t$ are missing; it finally reduces to the univariate conditional von Mises (3.8), evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{ks}$, if only one observation is missing. Analogously, each conditional distribution $f(\boldsymbol{y}_{t,\text{mis}}|\xi_{tk}=1, \boldsymbol{y}_{t,\text{obs}}; \hat{\boldsymbol{\gamma}}_{ks})$ in (4.5) is identically equal to 1 in the case of a fully observed profile; it reduces to the bivariate normal distribution (3.11), at $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}_{ks}$, if both the observations are missing; it reduces to a univariate, conditional normal density, if only one observation is missing. Finally, the conditional densities in (4.6) take the form of truncated distributions, with parameters that depend on the missing pattern in the linear profile. More precisely, we introduce a $o_t \times 2$ binary indicator matrix $\boldsymbol{O}_t$ for each profile $\boldsymbol{y}_t$ with $o_t$ observed values. This indicator matrix is obtained by extracting the rows of a $2 \times 2$ identity matrix, associated with the row positions of the observed values. In the case of a fully observed profile, $\boldsymbol{O}_t$ is equal to the $2 \times 2$ identity matrix; it otherwise reduces to the row vectors $(1, 0)$ and $(0, 1)$, respectively, if $y_{1t}$ is observed and $y_{2t}$ is missing; and vice versa. Furthermore, we define the matrix

$$\boldsymbol{C}_t(\hat{\boldsymbol{\gamma}}_{ks}) = \boldsymbol{O}_t^{\mathsf{T}}\big(\boldsymbol{O}_t\boldsymbol{\Omega}(\hat{\boldsymbol{\gamma}}_{ks})\boldsymbol{O}_t^{\mathsf{T}}\big)^{-1}\boldsymbol{O}_t,$$

$$\boldsymbol{\Omega}(\hat{\boldsymbol{\gamma}}_{ks}) = \begin{pmatrix} \hat{\gamma}_{11ks} & \hat{\gamma}_{12ks} \\ \hat{\gamma}_{12ks} & \hat{\gamma}_{22ks} \end{pmatrix} + \begin{pmatrix} \hat{\gamma}'_{1ks} & 0 \\ 0 & \hat{\gamma}'_{2ks} \end{pmatrix}^2,$$

for each latent class $k$. By Bayes' theorem,

$$f(\boldsymbol{v}_t|\xi_{tk}=1, \boldsymbol{y}_{t,\text{obs}}; \hat{\boldsymbol{\gamma}}_{ks}) = \frac{\phi_2(\boldsymbol{v}; \boldsymbol{a}(\hat{\boldsymbol{\gamma}}_{ks}), \boldsymbol{B}(\hat{\boldsymbol{\gamma}}_{ks}))}{\int_{(0,+\infty)^2}\phi_2(\boldsymbol{v}; \boldsymbol{a}(\hat{\boldsymbol{\gamma}}_{ks}), \boldsymbol{B}(\hat{\boldsymbol{\gamma}}_{ks}))\,d\boldsymbol{v}},$$

where

$$\boldsymbol{a}(\hat{\boldsymbol{\gamma}}_{ks}) = \begin{pmatrix} \hat{\gamma}'_{1ks} & 0 \\ 0 & \hat{\gamma}'_{2ks} \end{pmatrix}\boldsymbol{C}_t(\hat{\boldsymbol{\gamma}}_{ks})\begin{pmatrix} y_{1t} - \hat{\gamma}_{1ks} \\ y_{2t} - \hat{\gamma}_{2ks} \end{pmatrix}$$

and

$$\boldsymbol{B}(\hat{\boldsymbol{\gamma}}_{ks}) = \boldsymbol{I} - \begin{pmatrix} \hat{\gamma}'_{1ks} & 0 \\ 0 & \hat{\gamma}'_{2ks} \end{pmatrix} \boldsymbol{C}_t(\hat{\boldsymbol{\gamma}}_{ks}) \begin{pmatrix} \hat{\gamma}'_{1ks} & 0 \\ 0 & \hat{\gamma}'_{2ks} \end{pmatrix}.$$

The factorization (4.3) facilitates the evaluation of the expected complete-data log-likelihood, which can be computed in terms of iterated expectations as follows:

$$\begin{aligned} Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_s) &= \mathbb{E}\big(\log L_{\text{comp}}(\boldsymbol{\theta}, \boldsymbol{\xi}_{0:T}, \boldsymbol{z}_{0:T}, \boldsymbol{v}_{0:T}|\boldsymbol{z}_{0:T,\text{obs}}; \hat{\boldsymbol{\theta}}_s)\big) \\ &= \sum_{k=1}^{K} \mathbb{E}(\xi_{0k}|\boldsymbol{z}_{0:T,\text{obs}}, \hat{\boldsymbol{\theta}}_s) \log \pi_k \qquad (4.7) \\ &\quad + \sum_{t=1}^{T}\sum_{h=1}^{K}\sum_{k=1}^{K} \mathbb{E}(\xi_{t-1,h}\xi_{tk}|\boldsymbol{z}_{0:T,\text{obs}}, \hat{\boldsymbol{\theta}}_s) \log \pi_{h,k} \qquad (4.8) \\ &\quad + \sum_{t=0}^{T}\sum_{k=1}^{K} \mathbb{E}(\xi_{tk}|\boldsymbol{z}_{0:T,\text{obs}}, \hat{\boldsymbol{\theta}}_s)\mathbb{E}\big(\log f(\boldsymbol{x}_t; \boldsymbol{\beta}_k)|\boldsymbol{x}_{t,\text{obs}}, \hat{\boldsymbol{\beta}}_{ks}\big) \qquad (4.9) \\ &\quad + \sum_{t=0}^{T}\sum_{k=1}^{K} \mathbb{E}(\xi_{tk}|\boldsymbol{z}_{0:T,\text{obs}}, \hat{\boldsymbol{\theta}}_s)\mathbb{E}\big\{\mathbb{E}\big(\log f(\boldsymbol{y}_t; \boldsymbol{\gamma}_k) \\ &\quad + \log f(\boldsymbol{v}_t)|\boldsymbol{y}_{t,\text{obs}}, \hat{\boldsymbol{\gamma}}_{ks}, \boldsymbol{v}_t\big)|\boldsymbol{y}_{t,\text{obs}}\big\}. \qquad (4.10) \end{aligned}$$

The function $Q$ generalizes the familiar form of the expected complete-data log-likelihood $Q$, traditionally used in EM algorithms for estimating HMMs (Cappé, Moulines, and Rydén 2005), by allowing for incomplete observations of mixed type and random effects. Given the distributional assumptions of the model, all the expectations that appear in the above function $Q$ can be computed exactly (see Appendices A and B for details).

The M-step of the algorithm updates the estimate $\hat{\boldsymbol{\theta}}_s$ with a new estimate $\hat{\boldsymbol{\theta}}_{s+1}$, which maximizes the above function $Q$. This function is the sum of three functions that depend on independent sets of parameters and can thus be then maximized separately. Maximization of (4.8) with respect to the transition probabilities $\pi_{hk}$ provides the closed-form updating formula

$$\hat{\pi}_{hk(s+1)} = \frac{\sum_{t=1}^{T} \hat{p}_{t-1,t,hk}(\hat{\boldsymbol{\theta}}_s)}{\sum_{t=1}^{T} \hat{p}_{t-1,h}(\hat{\boldsymbol{\theta}}_s)}, \quad h, k = 1, \ldots, K.$$

Maximization of (4.9) reduces to $K$ separate nonlinear systems of five equations, which may be solved following, e.g., the iterative procedure suggested by Mardia et al. (2008). This requires a numerically efficient routine for the computation of Bessel functions, such as the `BesselI` function of R. Maximization of (4.10) reduces to a system of seven equations, for which Lin (2009) suggests an iterative procedure.

### 4.2. COMPUTATIONAL ASPECTS

The EM algorithm may converge to local maxima of the log-likelihood function or singularities at the edge of the parameter space, where the log-likelihood is unbounded (Wu

1983). The presence of multiple local and spurious maxima is well documented in the case of mixtures of heteroscedastic normal distributions (McLachlan and Peel 2000) and less widely known in the case of bivariate circular distributions (Mardia, Taylor, and Subramaniam 2007). 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-runs strategy (known as the emEM algorithm; Biernacki, Celeux, and Govaert 2003), by running the EM algorithm from a number of random initializations, and stopping the algorithm without waiting for full convergence. 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; Ingrassia and Rocci 2011) and can be detected within short EM runs by monitoring the class proportions.

We selected the ten outputs of the EM short run maximizing the log-likelihood and checked for spurious solutions, where this effect did not occur. Then, these ten parameter sets were used to initialize longer runs of the EM algorithm. As full convergence of the EM algorithm typically requires an inconveniently large number of iterations, we replaced the final steps of the algorithm by a faster direct numerical maximization of the log-likelihood function (Zucchini and MacDonald 2009, p. 72). We stopped the optimization when the increase of two successive log-likelihoods fell below $10^{-4}$ %, as this stopping criterion produced stable parameter estimates in preliminary experiments. This combination of the EM algorithm and numerical optimization is often called a hybrid algorithm (Lange and Weeks 1989; Redner and Walker 1984; Bulla and Berzel 2008) and provides a compromise between the large circle of convergence provided by the EM algorithm and the high speed of direct numerical maximization. The approach worked well as we observed that direct maximization of the log-likelihood is numerically stable and rapid when initial parameters are in the neighborhood of a maximum. In order to numerically maximize the likelihood with respect to the parameters, one needs to take care of some technical problems, such as avoiding numerical underflow and re-parameterizing the model in terms of unconstrained parameters (if an unconstrained maximization algorithm, e.g., $nlm()$ in R, is used). For details on how to deal with these problems, see, e.g., Chapter 3 of Zucchini and MacDonald (2009).

The procedure outlined above does not produce standard errors of the estimates, because approximations based on the observed information matrix often require a very large sample size (McLachlan and Peel 2000, p. 68). Visser, Raijmakers, and Molenaar (2000, 2002) investigate the reliable estimation of confidence bands in the context of HMMs and recommend bootstrap-based techniques. We followed their proposal and implemented a parametric bootstrap approach: we re-fitted the model to $R = 200$ bootstrap samples, which were simulated from the estimated model parameters. The approximate standard error of each model parameter $\theta$ was computed by

$$\mathrm{se}_R(\hat{\theta}) = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\theta}_r - \bar{\theta}_R)^2},$$

where $\hat{\theta}_r$ is the estimate from the $r$th bootstrap sample and $\bar{\theta}_R$ is the mean of all bootstrap estimates. For performance reasons, the model estimation for the bootstrap samples was

carried out by the same quasi-Newton optimization of the log-likelihood function $L(\boldsymbol{\theta})$ that served for the hybrid algorithm, taking the maximum likelihood estimates as initial values.

Simulation of the HMM proposed in this paper is straightforward. We first simulate a sequence of states from the Markov chain. Given a sequence of states, a bivariate circular observation is at each time $t$ drawn according to the appropriate bivariate von Mises density, evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_{k_t}$, where $k_t$ is the state that has been drawn at time $t$. Simultaneously, a bivariate linear observation is drawn from a bivariate skew-normal density, evaluated at $\boldsymbol{\gamma} = \boldsymbol{\gamma}_{k_t}$. To obtain a bivariate circular sample, we first draw a sample $x_1$ from the marginal density (3.6) with $\boldsymbol{\beta} = \boldsymbol{\beta}_{k_t}$ and then use $x_1$ to draw a value from the conditional distribution of $x_2$ given $x_1$, according to the von Mises density (3.8). There are several well-known routines for sampling from a von Mises distribution. Samples from the marginal circular density can be instead obtained by exploiting an acceptance–rejection algorithm, with a von Mises density as a proposal. Finally, we sample from a bivariate skew distribution by first drawing a bivariate random effect from the truncated distribution (3.10), and subsequently drawing a sample from the conditional normal density (3.11) evaluated at $\boldsymbol{\gamma} = \boldsymbol{\gamma}_{k_t}$.

## 5. RESULTS

We have estimated a number of HMMs from the data illustrated in Section 2, by varying the number of components from 2 to 4. To select the number of components, we computed both the Bayesian Information Criterion (BIC) and the Integrated Complete Likelihood (ICL) statistics. We intend to select the appropriate number of clusters, and therefore use the ICL as selection criterion. Alternatively, the commonly used BIC might have been an option. However, this criterion rather selects the correct number of components, which are in our case overlapping, thus not adding value for the physical interpretation, but just improving the fitted density in different clusters. The BIC statistic suggests a model with $K = 4$ components, while the minimum ICL is attained by a model with three components, which is the model we consider to analyze the data (Table 2). Figures 3 and 4 show the obtained solutions according to the BIC and the ICL, respectively. Circular and linear components are displayed as log-densities through contour lines. Each scatter plot includes the data points, filled with gray levels according to the posterior membership probabilities $\hat{p}_{tk}$ (black indicates $\hat{p}_{tk} = 1$). A model with 4 components distinguishes the same three clusters provided by a model with 3 components, using overlapping components to approximate the distribution of the data. This behavior of BIC has been extensively discussed in Baudry et al. (2010) in the context of mixture models. In our case study, however, overlapping components lack of physical interpretation, and cluster separation is more important than goodness of fit. We therefore use the ICL criterion, which includes cluster separation as an additional criterion for model choice (Biernacki, Celeux, and Govaert 2000).

Table 3 displays the estimates and the standard errors of the von Mises and the skew-normal densities for each state in the final three-state HMM. Regardless of the state, the

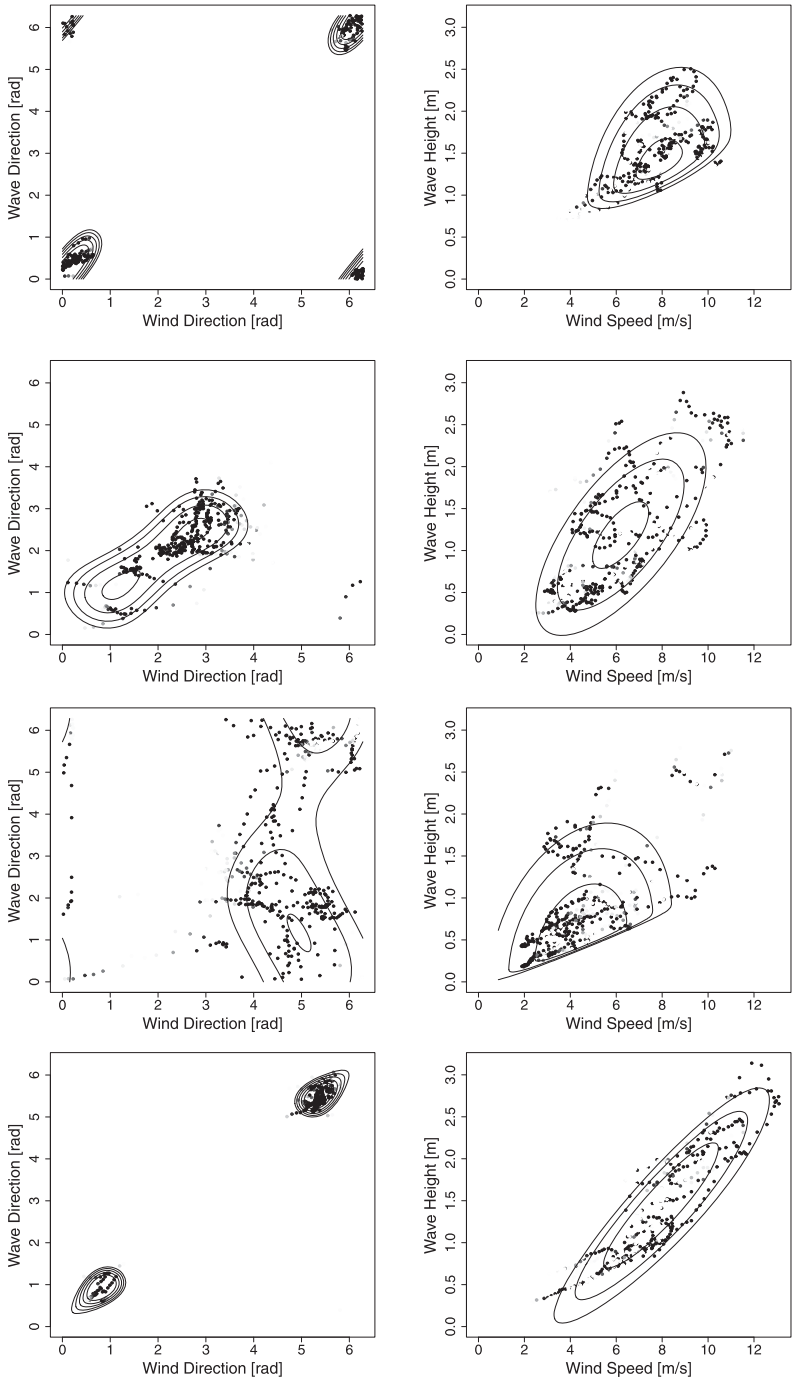


Figure 3. Log-densities of the circular (left) and linear (right) components of a four-state hidden Markov model. Contour lines are computed at the levels –0.5, –1.25, –2, –2.75, –3.5, and points are filled on a gray level scale according to their posterior probability of class membership, where black is associated with probability 1.

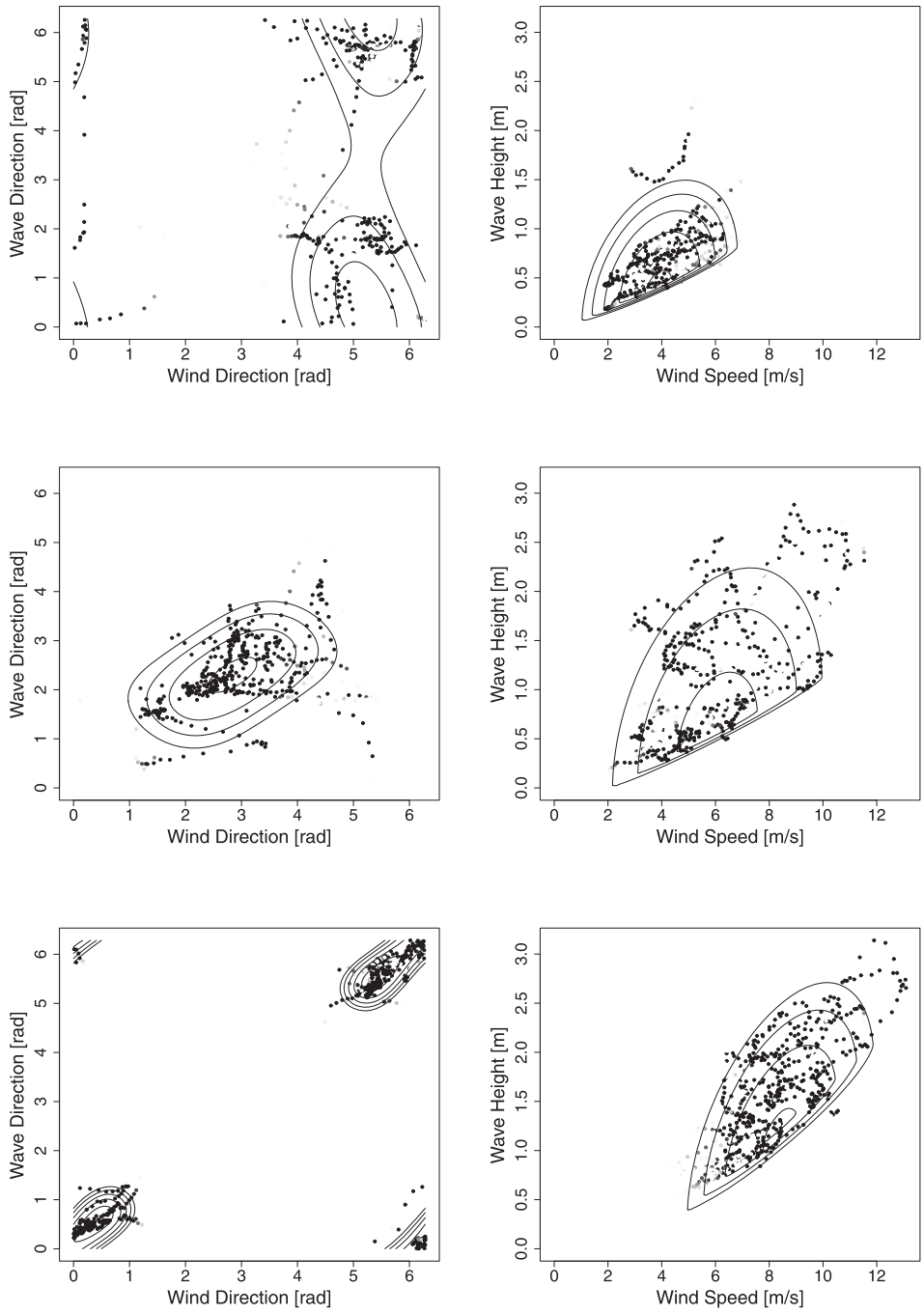


Figure 4. Log-densities of the circular (left) and linear (right) components of a three-state hidden Markov model. Contour lines are computed at the levels –0.5, –1.25, –2, –2.75, –3.5, and points are filled on a gray level scale according to their posterior probability of class membership, where black is associated with probability 1.

Table 2. Model selection results.

| Number of components | Number of parameters | Log-likelihood | BIC | ICL |
|---|---|---|---|---|
| 2 | 27 | –6812.9 | 13828.4 | 13861.5 |
| 3 | 44 | –5493.8 | 11309.3 | 11414.9 |
| 4 | 63 | –5373.7 | 11208.2 | 11419.7 |

Table 3. Estimated parameters.

| | | | State 1 | State 2 | State 3 |
|---|---|---|---|---|---|
| Circular parameters | Wave mean direction | $\beta_1$ | 0.341 | 2.305 | 6.199 |
| | | (se) | (0.062) | (0.024) | (0.012) |
| | Wind mean direction | $\beta_2$ | 5.227 | 2.840 | 6.070 |
| | | (se) | (0.028) | (0.028) | (0.010) |
| | Wave directional concentration | $\beta_{11}$ | 0.860 | 3.235 | 14.483 |
| | | (se) | (0.074) | (0.187) | (0.888) |
| | Wind directional concentration | $\beta_{22}$ | 2.700 | 2.154 | 15.627 |
| | | (se) | (0.185) | (0.129) | (0.913) |
| | Wave/wind directional dependence | $\beta_{12}$ | –0.669 | 1.758 | 18.840 |
| | | (se) | (0.152) | (0.155) | (1.084) |
| Linear parameters | Wave average height | $\gamma_1$ | 0.400 | 0.514 | 1.119 |
| | | (se) | (0.014) | (0.020) | (0.024) |
| | Wind average speed | $\gamma_2$ | 3.665 | 5.740 | 7.979 |
| | | (se) | (0.096) | (0.183) | (0.095) |
| | Wave height variance | $\gamma_{11}$ | 0.022 | 0.085 | 0.144 |
| | | (se) | (0.003) | (0.008) | (0.012) |
| | Wind speed variance | $\gamma_{22}$ | 1.319 | 4.257 | 2.462 |
| | | (se) | (0.098) | (0.264) | (0.158) |
| | Wind/wave covariance | $\gamma_{12}$ | 0.171 | 0.602 | 0.596 |
| | | (se) | (0.014) | (0.043) | (0.041) |
| | Wave height skewness | $\gamma_1'$ | 0.415 | 0.872 | 0.629 |
| | | (se) | (0.017) | (0.036) | (0.018) |
| | Wind speed skewness | $\gamma_2'$ | 0.320 | 0.409 | 0.553 |
| | | (se) | (0.094) | (0.179) | (0.077) |
| | Destination state | | 1 | 2 | 3 |
| | Origin state | 1 | 0.961 | 0.025 | 0.014 |
| | | | (0.012) | (0.009) | (0.007) |
| | | 2 | 0.011 | 0.981 | 0.007 |
| | | | (0.005) | (0.007) | (0.004) |
| | | 3 | 0.013 | 0.002 | 0.986 |
| | | | (0.006) | (0.002) | (0.006) |
| | Initial state Distribution | | 0.240 | 0.355 | 0.405 |
| | | | (0.074) | (0.102) | (0.118) |

dependence between circular variables and the covariance between linear variables is significant, indicating that in this case a conditional independence assumption between univariate distributions is unrealistic.

The first component of the model is associated with periods of calm sea: weak winds ($\gamma_2 = 3.665$) generate small waves ($\gamma_1 = 0.400$). In this regime, the shape of the joint distribution of wave and wind directions is essentially spherical ($\beta_{12}$ is barely significant) and centered at the average wind direction $\hat{\beta}_2 = 5.227$, corresponding to northwesterly Maestral episodes. As expected, wind and wave directions are poorly synchronized under this regime, because wave direction is more influenced by marine currents than by wind direction during weak wind episodes.

The second component is associated with Sirocco episodes ($\beta_2 = 2.840$). Compared to the first regime, wind and wave directions appear more synchronized ($\beta_{12} = 1.758$) and characterized by winds of higher speed ($\gamma_2 = 5.740$) and higher waves ($\gamma_1 = 0.514$). In this second regime, waves travel southeasterly along the major axis of the basin ($\beta_1 = 2.305$), driven by winds that blow from a similar directional angle ($\beta_2 = 2.840$). As there are neither coastlines nor mountains, there is little dispersion of energy in the interaction between wind and wave and, as a result, waves can reach significant heights. In studies of the Adriatic Sea, detection of Sirocco regimes is very important because it exposes Venice to the famous flooding tides when occurring in combination with lunisolar astronomical forces.

A similar phenomenon, although in the opposite direction, is captured by the third component of the model. In this regime, northern Bora jets ($\beta_2 = 6.070$) generate high waves ($\gamma_2 = 1.119$) that travel along the major axis of the basin ($\beta_1 = 6.199$). Compared to the other two regimes, waves and winds are much more synchronized ($\beta_{12} = 18.840$) and highly concentrated around one modal direction. Most of the wind energy is transferred to the sea surface and, as a result, the correlation between wind speed and wave height is larger than that observed under Sirocco or Maestral episodes. As expected, most of the profiles with the highest waves in the sample are clustered in this regime.

A somewhat unexpected pattern of point is observed in Figure 4, second plot in the first row. It is associated with an atypical condition of good sea conditions with waves of significant height, occurring right after a Sirocco storm.

The rows at the bottom of Table 3 include the estimated transition probabilities and initial probabilities of the latent Markov chain. As expected, the transition probability matrix is essentially diagonal, reflecting the temporal persistence of the states. Furthermore, the small off-diagonal transition probabilities between states 2 and 3 ($\hat{\pi}_{23} = 0.007$ and $\hat{\pi}_{32} = 0.002$) indicate that direct transitions between Sirocco and Bora episodes are very unlikely. The model hence confirms that the Adriatic Sea typically alternates relevant marine events with periods of good sea conditions.

The model describes the plasticity of the wind–wave interaction in the Adriatic Sea, indicating that the joint distribution of wind and wave data changes under different environmental regimes. Regime-switching does not only change directional and linear averages but also, and more interestingly, the correlation structure of the data. As a result, on the one side the weak (marginal) correlation between wind and wave observations is explained by the presence of a Maestral-specific regime of good weather conditions. On the other side, the model indicates that wind is an accurate predictor of wave-metric processes during a Bora episode, but that the level of accuracy decreases under Sirocco and almost vanishes

under Maestral episodes. In summary, weather conditions should not be used to predict wave direction and height, without accounting for the latent, environmental heterogeneity of the data under study.

## 6. DISCUSSION

We have illustrated an HMM-based classification method for multivariate mixed-type time series, focusing on skewed and circular variables. The data are clustered according to bivariate skew-normal and von Mises distributions, which are associated with the states of a latent Markov chain. The resulting model captures several sources of heterogeneity: time-dependence, through the hidden Markov chain; unobserved (or spurious) association, by defining latent regimes; observed correlations, estimated along with all other model parameters; skewness, by exploiting positive random effects. Our classification procedure is motivated by issues that arise in marine studies, but can be easily adapted to a wide range of real-world cases, including for example ecological studies of animal behavior, where direction and speed of movements are recorded (Holzmann et al. 2006), and bioinformatics applications, where sequences of protein dihedral angles (Mardia, Taylor, and Subramaniam 2007) are recorded with a number of continuous variables.

The model relies on a latent-class approach to the analysis of multivariate mixed-type data by assuming that circular and linear variables are conditionally independent given the states visited by a latent Markov chain. Part of the dependence structure between variables is therefore non-parametrically captured by the association between planar and toroidal clusters. This seems a convenient strategy in marine studies, where the dependence between circular and linear measurements is the result of complex environmental conditions. This conditional independence assumption between pairs of linear and circular variables could in principle be avoided by taking a fully parametric approach and replacing the product of von Mises and skew-normal densities by quadrivariate densities with a hyper-cylindrical support such as those proposed by Kato and Shimizu (2008). Their use in an HMM setting is however problematic, because little is known about efficient estimation procedures and identifiability issues under hyper-cylindrical parametric models. In addition, mixtures of hyper-toroidal densities would group our data according to quadrivariate clusters of difficult visualization and interpretation, without necessarily improving the fit of the model. On the contrary, a conditional independence assumption between pairs of variables offers a number of advantages: clusters interpretation is intuitively appealing; parameter estimation can be carried out by combining EM algorithms that have been developed for data with homogeneous supports. Finally, identifiability issues do not arise because a sufficient condition for the identifiability of mixtures of product densities is the linear independence of the mixture components (Teicher 1967; Yakowitz and Spragins 1968) and, as a result, the identifiability of the model we propose follows from the linear independence of the bivariate circular densities (Mardia et al. 2008) and the linear independence of the bivariate skew-normal densities (Sahu, Dey, and Branco 2003).

We focused on the analysis of a single multivariate time series. Further developments may include the analysis of multiple time-series, i.e. observations collected at different sites (for a review on longitudinal data in a HMM framework see, e.g., Maruotti 2011), and the specification of spatial effects aiming at identifying geographically homogeneous groups likely to share the same environmental regimes.

A limit of the proposed approach is represented by the assumptions on the hidden chain. We assume that the dwell time in each state is geometrically distributed as a consequence of the Markov property of the hidden Markov chain. However, the probability of a state change may exhibit different patterns or even depend on the time spent in the current state. Thus, more flexible assumptions could be analyzed, such as considering a hidden semi-Markov chain, which follows more general dwell time distributions (see, e.g., Barbu and Limnios 2005; Bulla, Bulla, and Nenadić 2010; Langrock and Zucchini 2011).

The use of the skew-normal distribution for wind speed and wave height theoretically allows for negative values. Under the estimated model, however, the probability of negative observations is smaller than 0.0001, a value that we believe as acceptable to justify the model. In other studies where observations of wind speed and wave height are very close to zero, this probability can however be (though not necessarily) inadequately large. In these cases the model can be fit by first taking a logarithm transformation of wind speed and wave height (see, e.g., Marchenko and Genton 2010 for an analysis of precipitation log-skew-normal data) and then using the EM algorithm on the transformed data. This is a straightforward strategy to recast the theoretical validity of the model, at a price of difficulties in the presentation and the interpretation of the results, which have to be back-transformed to the original scale of the data.

## APPENDIX A

All the expectations in the above $Q$ function can be computed exactly. In particular,

$$\begin{aligned} \hat{p}_{t,k}(\hat{\boldsymbol{\theta}}_s) &= \mathbb{E}(\xi_{tk}|\boldsymbol{z}_{\text{obs}}, \hat{\boldsymbol{\theta}}_s), \\ \hat{p}_{t-1,t,hk}(\hat{\boldsymbol{\theta}}_s) &= \mathbb{E}(\xi_{t-1,h}\xi_{tk}|\boldsymbol{z}_{\text{obs}}, \hat{\boldsymbol{\theta}}_s) \end{aligned} \tag{A.1}$$

are the conditional first- and second-order expectations of the latent Markov chain, given the observed data, and can be efficiently computed by a Baum–Welch (BW) procedure. We exploited a BW procedure that allows for incomplete time series, illustrated in Appendix B. The expectations

$$\mathbb{E}\big(\log f(\boldsymbol{x}_t; \boldsymbol{\beta}_k)|\boldsymbol{x}_{t,\text{obs}}, \hat{\boldsymbol{\beta}}_{ks}\big)$$

can be computed by replacing the statistics $\cos(x_{t1} - \beta_{1k})$, $\cos(x_{t2} - \beta_{2k})$ and $\sin(x_{t1} - \beta_{1k})\sin(x_{t2} - \beta_{2k})$ by their expected values, with respect to the conditional distributions (4.4) evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{ks}$. The expectations

$$\mathbb{E}\big(\log f(\boldsymbol{y}_t; \boldsymbol{\gamma}_k) + \log f(\boldsymbol{v}_t)|\boldsymbol{y}_{t,\text{obs}}, \hat{\boldsymbol{\gamma}}_{ks}, \boldsymbol{v}_t\big)$$

can be evaluated by computing the expected values of $\boldsymbol{y}_t$ and $\boldsymbol{y}_t\boldsymbol{y}_t^\top$ with respect to the normal distributions (4.5). Finally, the expectations

$$\mathbb{E}\big\{\mathbb{E}\big(\log f(\boldsymbol{y}_t;\boldsymbol{\gamma}_k)+\log f(\boldsymbol{v}_t)|\boldsymbol{y}_{t,\text{obs}},\hat{\boldsymbol{\gamma}}_k,\boldsymbol{v}_t\big)|\boldsymbol{y}_{t,\text{obs}}\big\}$$

can be obtained by computing the expected values of $\boldsymbol{v}_t$ and $\boldsymbol{v}_t\boldsymbol{v}_t^\top$, with respect to the truncated normal distributions (4.6), using the efficient computation of the moments of the truncated normal proposed by Lin (2009).

## APPENDIX B

The task of computing the posterior probabilities from an estimate $\hat{\boldsymbol{\theta}}_s$ is generally referred to as the HMM-smoothing numerical issue and it is typically solved by specifying the posterior probabilities in terms of suitably normalized functions, which can be computed recursively, avoiding unpractical summations over the state space of latent Markov chain and numerical under- and over-flows. In the literature, this approach is known as the Forward-Backward (FB) recursion and it can be implemented in a number of different ways (Cappé, Moulines, and Rydén 2005; Chapter 3). We describe below an FB recursion that allows for incomplete observations.

Let

$$L_t(\hat{\boldsymbol{\beta}}_s,\hat{\boldsymbol{\gamma}}_s)=\int_{\boldsymbol{z}_{t,\text{mis}}}\prod_{k=1}^{K}\big(f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_{ks})f(\boldsymbol{y}_t;\hat{\boldsymbol{\gamma}}_{ks})\big)^{\xi_{tk}}d\boldsymbol{z}_{t,\text{mis}}$$

be the conditional contribution of $\boldsymbol{z}_{t,\text{obs}}$ to the likelihood function. In addition, let

$$L_{0:t}(\hat{\boldsymbol{\theta}}_s)=\sum_{\boldsymbol{\xi}_0}\cdots\sum_{\boldsymbol{\xi}_t}p(\boldsymbol{\xi}_{0:t};\hat{\boldsymbol{\pi}}_s)\prod_{\tau=0}^{t}L_\tau(\hat{\boldsymbol{\beta}}_s,\hat{\boldsymbol{\gamma}}_s)$$

be the contribution of the first $t$ profiles to the likelihood and let $\boldsymbol{z}_{0:t,\text{obs}}$ be the time series, observed up to time $t$. We run a forward and a backward iteration.

During the forward iteration, we exploit the output of the $s$th step of the EM algorithm to compute the probabilities $\psi^{(t)}(k)=P(\xi_{tk}=1|\boldsymbol{z}_{0:t-1,\text{obs}})$, the likelihood ratios $c_t=\frac{L_{0:t}(\hat{\boldsymbol{\theta}}_s)}{L_{0:t-1}(\hat{\boldsymbol{\theta}}_s)}$ and the forward probabilities $\bar{\alpha}_t(k)=P(\xi_{tk}=1|\boldsymbol{z}_{0:t,\text{obs}})$, as follows.

Forward recursion:

- initialization:

$$\psi^{(0)}(k)=\hat{\pi}_{ks}$$

$$c_0=\sum_{k=1}^{K}\psi^{(0)}(k)L_0(\hat{\boldsymbol{\beta}}_s,\hat{\boldsymbol{\gamma}}_s)$$

$$\bar{\alpha}_0(k)=\frac{\psi^{(0)}(k)L_0(\hat{\boldsymbol{\beta}}_s,\hat{\boldsymbol{\gamma}}_s)}{c_0}$$

- for $t = 1, \ldots, T$:

$$\psi^{(t)}(k) = \sum_{h=1}^{K} \bar{\alpha}_{t-1}(h)\hat{\pi}_{hks}$$

$$c_t = \sum_{k=1}^{K} \psi^{(t)}(k) L_t(\hat{\boldsymbol{\beta}}_{ks}, \hat{\boldsymbol{\gamma}}_s)$$

$$\bar{\alpha}_t(k) = \frac{\psi^{(t)}(k) L_t(\hat{\boldsymbol{\beta}}_{ks}, \hat{\boldsymbol{\gamma}}_s)}{c_t}$$

At the end the forward recursion, we store the values $c_0 \ldots c_T$ and $\bar{\alpha}_0(k) \ldots \bar{\alpha}_T(k)$. The sequence $c_0 \ldots c_T$ can be exploited to compute the value taken by the log-likelihood at the $s$th step of the EM algorithm, as follows:

$$\log L(\hat{\boldsymbol{\theta}}_s) = \sum_{t=0}^{t} \log c_t.$$

We then run a backward recursion, by computing the ratios $\bar{\varphi}_t(k) = \frac{f(z_{t+1:T,\text{obs}} | \xi_{tk}=1)}{\prod_{l=t}^{T} c_l}$, as follows.

Backward recursion:

- initialization: $\bar{\varphi}_T(k) = \frac{1}{c_T}$
- for $t = T-1, T-2, \ldots, 0$:

$$\bar{\varphi}_t(k) = \frac{\sum_{h=1}^{K} \hat{\pi}_{khs} L_{t+1}(\hat{\boldsymbol{\beta}}_{ks}, \hat{\boldsymbol{\gamma}}_s) \bar{\varphi}_{t+1}(h)}{c_t}.$$

At the end of the backward recursion, we store the values of $\bar{\varphi}_0(k) \ldots \bar{\varphi}_T(k)$ and compute the posterior univariate state probabilities as

$$\hat{p}_{tk}(\hat{\boldsymbol{\theta}}_s) = \frac{\bar{\alpha}_t(k)\bar{\varphi}_t(k)}{\sum_{k=1}^{K} \bar{\alpha}_t(k)\bar{\varphi}_t(k)}.$$

The bivariate posterior probabilities can be instead computed as

$$\hat{p}_{t-1,t,hk}(\hat{\boldsymbol{\theta}}_s) = \bar{\alpha}_t(k)\hat{\pi}_{hks} L_{t+1}(\hat{\boldsymbol{\beta}}_{ks}, \hat{\boldsymbol{\gamma}}_{ks})\bar{\varphi}_{t+1}(k).$$

## ACKNOWLEDGEMENTS

Thanks are expressed to the Fédération Normandie-Mathématiques FR CNRS 3335 for the continuous support. Also, to the Italian Institute for Environmental Protection and Research for providing the data and the financial support.

*[Received February 2012. Accepted August 2012. Published Online September 2012.]*

# REFERENCES


Ailliot, P., and Monbet, V. (2012), "Markov-Switching Autoregressive Models for Wind Time Series," *Environmental Modelling & Software*, 30, 92–101.

Barbu, V., and Limnios, N. (2005), "Maximum Likelihood Estimation for Hidden Semi-Markov Models," *Comptes Rendus Mathematique*, 342, 201–205.

Baudry, J.-P., Raftery, A. E., Celeux, G., Lo, K., and Gottardo, R. (2010), "Combining Mixture Components for Clustering," *Journal of Computational and Graphical Statistics*, 19, 332–353.

Bertotti, L., and Cavalieri, L. (2009), "Wind and Wave Predictions in the Adriatic Sea," *Journal of Marine Systems*, 78, S227–S234.

Biernacki, C., Celeux, G., and Govaert, G. (2000), "Assessing a Mixture Model for Clustering With the Integrated Completed Likelihood," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 719–725.

——— (2003), "Choosing Starting Values for the EM Algorithm for Getting the Highest Likelihood in Multivariate Gaussian Mixture Models," *Computational Statistics & Data Analysis*, 41, 561–575.

Bulla, I., Bulla, J., and Nenadić, O. (2010), "`hsmm`—An **R** Package for Analyzing Hidden Semi-Markov Models," *Computational Statistics & Data Analysis*, 54, 611–619.

Bulla, J., and Berzel, A. (2008), "Computational Issues in Parameter Estimation for Stationary Hidden Markov Models," *Computational Statistics*, 23, 1–18.

Cabral, C. R. B., Lachos, V. H., and Prates, M. O. (2012), "Multivariate Mixture Modeling Using Skew-Normal Independent Distributions," *Computational Statistics & Data Analysis*, 56, 126–142.

Caires, S., and Sterl, A. (2005), "A New Non-Parametric Method to Correct Model Data: Application to Significant Wave Height From the ERA-40 Reanalysis," *Journal of Atmospheric and Oceanic Technology*, 22, 443–459.

Cappé, O., Moulines, E., and Rydén, T. (2005), *Inference in Hidden Markov Models*, Berlin: Springer.

Faltinsen, O. (1990), *Sea Loads on Ships and Offshore Structures*, Cambridge: Cambridge University Press.

Hamilton, L. (2010), "Characterising Spectral Sea Wave Conditions With Statistical Clustering of Actual Spectra," *Applied Ocean Research*, 32, 332–342.

Holzmann, H., Munk, A., Suster, M., and Zucchini, W. (2006), "Hidden Markov Models for Circular and Linear–Circular Time Series," *Environmental and Ecological Statistics*, 13, 325–347. doi:*10.1007/s10651-006-0015-7*.

Huang, G., Wing-Keung Law, A., and Huang, Z. (2011), "Wave-Induced Drift of Small Floating Objects in Regular Waves," *Ocean Engineering*, 38, 712–718.

Ingrassia, S., and Rocci, R. (2011), "Degeneracy of the EM Algorithm for the MLE of Multivariate Gaussian Mixtures and Dynamic Constraints," *Computational Statistics & Data Analysis*, 55, 1715–1725.

Izquierdo, P., and Guedes Soares, C. (2005), "Analysis of Sea Waves and Wind From X-Band Radar," *Ocean Engineering*, 32, 1404–1419.

Jin, K.-R., and Ji, Z.-G. (2004), "Case Study: Modeling of Sediment Transport and Wind–Wave Impact in Lake Okeechobee," *Journal of Hydraulic Engineering*, 130, 1055–1067.

Kato, S., and Shimizu, K. (2008), "Dependent Models for Observations Which Include Angular Ones," *Journal of Statistical Planning and Inference*, 138, 3538–3549. Special Issue in Honor of Junjiro Ogawa (1915–2000): Design of Experiments, Multivariate Analysis and Statistical Inference.

Lagona, F., and Picone, M. (2011), "A Latent-Class Model for Clustering Incomplete Linear and Circular Data in Marine Studies," *Journal of Data Science*, 9, 585–605

——— (2012a), "Maximum Likelihood Estimation of Bivariate Circular Hidden Markov Models From Incomplete Data," *Journal of Statistical Computation and Simulation*. Available online at *http://www.tandfonline.com/doi/pdf/10.1080/00949655.2012.656642*.

——— (2012b), "Model-Based Clustering of Multivariate Skew Data With Circular Components and Missing Values," *Journal of Applied Statistics*, 39, 927–945.

Lange, K., and Weeks, D. E. (1989), "Efficient Computation of LOD Scores: Genotype Elimination, Genotype Redefinition, and Hybrid Maximum Likelihood Algorithms," *Annals of Human Genetics*, 53, 67–83.

Langrock, R., and Zucchini, W. (2011), "Hidden Markov Models With Arbitrary State Dwell-Time Distributions," *Computational Statistics & Data Analysis*, 55, 715–724.

Lin, T. I. (2009), "Maximum Likelihood Estimation for Multivariate Skew Normal Mixture Models," *Journal of Multivariate Analysis*, 100, 257–265.

Marchenko, Y. V., and Genton, M. G. (2010), "Multivariate Log-Skew-Elliptical Distributions With Applications to Precipitation Data," *Environmetrics*, 21, 318–340.

Mardia, K., Taylor, C., and Subramaniam, G. (2007), "Protein Bioinformatics and Mixtures of Bivariate von Mises Distributions for Angular Data," *Biometrics*, 63, 505–512.

Mardia, K. V., Hughes, G., Taylor, C. C., and Singh, H. (2008), "A Multivariate von Mises Distribution With Applications to Bioinformatics," *Canadian Journal of Statistics*, 36, 99–109.

Maruotti, A. (2011), "Mixed Hidden Markov Models for Longitudinal Data: An Overview," *International Statistical Review*, 79, 427–454.

McLachlan, G., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

Monbet, V., Ailliot, P., and Prevosto, M. (2007), "Survey of Stochastic Models for Wind and Sea-State Time Series," *Probabilistic Engineering Mechanics*, 22, 113–126.

Pleskachevsky, A., Eppel, D., and Kapitza, H. (2009), "Interaction of Waves, Currents and Tides, and Wave-Energy Impact on the Beach Area of Sylt Island," *Ocean Dynamics*, 59, 451–461.

Redner, R. A., and Walker, H. F. (1984), "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, 26, 195–239.

Reikard, G., and Rogers, W. E. (2011), "Forecasting Ocean Waves: Comparing a Physics-Based Model With Statistical Models," *Coastal Engineering*, 58, 409–416.

Sahu, S., Dey, D., and Branco, M. (2003), "A New Class of Multivariate Skew Distributions With Applications to Bayesian Regression Models," *Canadian Journal of Statistics*, 31, 129–150.

Singh, H., Hnizdo, V., and Demchuk, E. (2002), "Probabilistic Model for Two Dependent Circular Variables," *Biometrika*, 89 (3), 719–723.

Teicher, H. (1967), "Identifiability of Mixtures of Product Measures," *Annals of Mathematical Statistics*, 38, 1300–1302.

Visser, I., Raijmakers, M., and Molenaar, P. (2000), "Confidence Intervals for Hidden Markov Model Parameters," *British Journal of Mathematical & Statistical Psychology*, 53, 317–327.

Visser, I., Raijmakers, M. E. J., and Molenaar, P. C. M. (2002), "Fitting Hidden Markov Models to Psychological Data," *Scientific Programming*, 10, 185–199.

Wu, C. (1983), "On the Convergence Properties of the EM Algorithm," *Annals of Statistics*, 11, 95–103.

Yakowitz, S., and Spragins, J. (1968), "On the Identifiability of Finite Mixtures," *Annals of Mathematical Statistics*, 39, 209–214.

Zhang, Q., Snow Jones, A., Rijmen, F., and Ip, E. (2010), "Multivariate Discrete Hidden Markov Models for Domain-Based Measurements and Assessment of Risk Factors in Child Development," *Journal of Computational and Graphical Statistics*, 19, 746–765.

Zucchini, W., and Guttorp, P. (1991), "A Hidden Markov Model for Space-Time Precipitation," *Water Resources Research*, 27, 1917–1923.

Zucchini, W., and MacDonald, I. (2009), *Hiddden Markov Models for Time Series: An Introduction Using R*, London: Chapman & Hall.

# Conclusions

The contribution of this work is to provide a rigorous and flexible approach to the classification of mixed linear and circular incomplete data, based on mixture models. The advantages of these methods include a simple specification of the dependence structure between variables and a good computation feasibility that allows to efficiently handle missing values within a likelihood framework.

Three mixture-based models have been proposed, an univariate mixture model, a multivariate mixture model and a multivariate hidden Markov model. All these models allow to explain the correlation structure of wind and wave measurements in the Adriatic Sea, in terms of different latent regimes that reflect the heterogeneity of marine events.

In particular, in semi-enclosed basins, where the observed phenomenon is driven by a number of variables, such as the orography of the coasts, the bathymetry, the internal circulation and currents, etc., mixture models can better explain the relation between wind and wave data than semi-deterministic and physical-based model, that are very common in the marine literature.

The thesis proceeds along successive extensions, by extending the flexibility of the data dependence structures.

The simplest model is the univariate mixture model (UMM). It assumes that multivariate data profiles are temporally independent and that contemporary measurements are conditionally independent given a number of latent classes. As a results, we obtain a large number of clusters, which are partially overlapping and of difficult physical interpretation.

By relaxing the conditional independence hypothesis in the multivariate mixture model (MMM), the cluster interpretation becomes intuitively and physically appealing even with a small number of latent classes.

By further removing the temporal independence hypothesis in the multivariate hidden Markov model (MHMM), clusters are better separated than the previous models and the uncertainty of the final classification significantly reduces.

The conditional independence assumption should be used with parsimony. Under a UMM, products of univariate distributions may be too restrictive

to accommodate for the complex shape clusters and therefore an unnecessary large number of latent classes may be required to obtain a reasonable goodness of fit. It nevertheless complicates the interpretation of the results, especially when the purpose is to cluster the data into meaningful groups.

Complex models, such as MHMM, are more flexible and account for the potential redundancy of the data resulting from temporal autocorrelation. The price for a good classification is however the computational complexity of the estimation step.