Web Working Papers
by
The Italian Group of Environmental Statistics

Gruppo di Ricerca per le Applicazioni della Statistica
ai Problemi Ambientali

www.graspa.org

A mixture-based approach to multiple imputation in incomplete linear-circular datasets

Francesco Lagona and Marco Picone

GRASPA Working paper n.35, December 2009

A mixture-based approach to multiple imputation in incomplete linear-circular datasets

Francesco Lagona * and Marco Picone

DIPES and GRASPA Unit of Rome - University Roma Tre
Via G. Chiabrera 199, 00145 Rome, Italy

Model-based methods of missing value imputation in incomplete multivariate datasets require the definition of an imputation model that specifies the predictive distribution of the missing values given the observed data. Mixture models for multivariate data provide a flexible approach to specify imputation models when variables are measured on different scales. We propose to specify the joint distribution of linear and circular data through a finite mixture of conditionally independent Gamma and von Mises distributions. The procedure is illustrated on an incomplete dataset that includes measurements of wind speed and direction and significant wave height and direction, taken by a buoy and two tide gauges of the Italian wave-metric network.

Keywords: E-M algorithm, Gamma distribution, marine data, mixture, model-based clustering, Von Mises distribution

1. Introduction

Environmental multivariate data are disseminated by environmental protection agencies for a variety of purposes, ranging from the computation of simple descriptive summaries that communicate environment conditions to the public, to the estimation of sophisticated statistical models that detect significant relationships between different environmental measurements. Incomplete datasets, where some of the measurements are missing, pose a serious obstacle in the fulfillment of these purposes.

There is an extensive literature about the estimation of statistical models in the presence of missing values. However, these methods often require the expertise of a trained statistician, as they involve both computational and methodological issues that can be challenging, depending on the nature of the mechanism that generate the missing values and the complexity of the model that is exploited for analysis. To reduce the workload of the data analyst, incomplete data should be provided in a way that they can be analyzed by "standard" methods, i.e. methods that require the availability of complete data information.

Environmental data could be completed by imputing missing values according to an imputation model. This approach is referred to as single imputation. It is however well known that if the data analyst uses complete-data methods for analyzing the completed dataset by treating imputed values as if they were real data, this generally leads to variance estimates that are too low, confidence intervals which are too narrow, and wrong tests (real significance level above nominal level).

* Corresponding author. Email: lagona@uniroma3.it

Intuitively, this is due to the fact that imputation does not generate information which is not already present in the data; hence the sample size is the same for the incomplete data and for the imputed (pseudo-complete) data.

Multiple imputation (MI; Rubin, 1987) has been suggested as a way of overcoming the variance estimation problem that arises under a single-imputation strategy. Under a MI protocol, the data-base constructor (or imputer) and the end user (or data analyst) are though as distinct entities (Rubin, 1996). The data-base constructor draws a number of imputed values from the predictive distribution of the missing values, given the observed data, computed on the basis of an imputation model. The resulting completed datasets are appended together to provide an augmented dataset to the data analyst, who can exploit standard methods to simultaneously examine these datasets and, appropriately pooling the results, use them to correct for the variability in the imputations, which differs from the variability in the observed data. Directions about the pooling procedure are provided by the imputer and involve simple calculations, which can be carried out by a data-analyst who is not necessarily a trained statistician.

Under a MI strategy, imputation is typically carried out by estimating a parametric model from the complete cases and using the predictive distribution of the missing values given the observed data to draw a number of imputations for each missing value in the incomplete dataset. Under a Bayesian approach, the posterior distribution of the parameters is computed from the observed data and the imputations are drawn from the posterior predictive distribution of the missing values, via Markov chain Monte Carlo methods (Schafer 1996; Tanner and Wong 1987). Under a frequentist approach (Wei and Tanner, 1990), a model is fitted to the observed data by maximum likelihood and imputations are drawn from the conditional distribution of the missing values, given the observed data, evaluated at the maximum likelihood estimate of the parameters that have been obtained from the observed data.

The key to the success of multiple imputation is the concept of proper multiple imputation. In general, multiple imputations are proper if the pooling procedure (to be carried out by the data analyst) yields a consistent, asymptotically normal estimator of the unknown parameters and a consistent estimator of its asymptotic variance. Although it is generally believed that Bayesian MIs are proper if the imputer's and the analyst's models are compatible, while frequentist MIs are usually improper, both approaches give proper imputations if the imputation model is not misspecified (Robins and Wang, 1998; Wang and Robins, 2000).

The specification of an imputation model typically depends on the scale types of the variables in the incomplete data set. For continuous variables that take values on the real line, the most widely used imputation model is the multivariate normal model, while log-linear models are popular for imputing categorical variables (Shafer, 1997). For data sets containing both categorical and continuous variables, Schafer (1997) proposed imputation using the general location model, a combination of a log-linear and a multivariate normal model.

More recently, finite-mixture models appeared in the literature as a flexible approach to specify imputation models in multivariate datasets. Di Zio *et al.* (2007), for example, use mixtures of multivariate normal distributions for continuous variables, while Vermunt *et al.* (2008) suggest a latent-class model for the imputation of categorical variables. Mixture models for multivariate data that include both categorical and continuous variables have been suggested by Hunt and Jorgensen (2003).

We present an application where the mixture-based approach to multiple imputation is extended to the case of incomplete datasets that include both linear and circular variables, which typically arise in environmental research. We take a frequentist approach, and specify a multivariate mixture of Gamma and von Mises distributions. Parameters of the mixture model are estimated by maximization of the likelihood from the observed data, by a suitable E-M algorithm. Imputations are drawn from the multivariate conditional distribution of the missing values given the observed data, evaluated at the maximum likelihood estimate. We discuss the performance of this imputation strategy on marine incomplete data that include wave direction and hight and wind speed and direction.

The rest of the paper is organized as follows. In Section 2 we introduce a mixture model that can be exploited (1) to specify the joint distribution of a multivariate random variable, whose components are measured on different scales (e.g., linear and circular), and (2) to draw multiple imputations to complete multivariate mixed data that are partially observed. We then illustrate the data that motivated this work in Section 3 and apply the proposed procedure in Section 4. Relevant points of discussion are briefly mentioned in Section 5.

2. Mixture models for incomplete mixed data

We assume that the data are gathered in the form of n independent vectors $\boldsymbol{y}_i = (y_{i1} \dots y_{ij} \dots y_{iJ})$, $i = 1 \dots n$, drawn from the multivariate distribution of J variables $Y_j, j = 1 \dots J$, that are measured on different scales (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) variable Z that takes K discrete values, say $z_1 \dots z_k \dots z_K$, with distribution $P(Z = z_k) = \pi_k$, 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 K multivariate components, say

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

where $f_k(y|\boldsymbol{\beta}_{kj})$ denotes the conditional distribution of Y_j within the kth latent class. Mixture-based specifications of multivariate distributions (McLachlan and Peel, 2000) such as (1), aim at determining the inner structure of clustered data when no information other than the observed values is available. Mixtures of the type (1) provide a model-based clustering of multivariate observations and are an alternative to classical cluster analysis that is based on distance-based procedures, such as hierarchical agglomerative clustering or iterative relocation procedures.

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, $O(i) \cup M(i) = \{1 \dots J\}$. We further let $m(j) \subseteq \{1, \dots i \dots n\}$ indicate the subset of the observations where the jth variable have been observed. If the data are missing at random (MAR; Rubin 1987), i.e. the probability of a missing value does not depend on the value that is missing, maximum likelihood estimates of model (1) can be found by maximizing the marginal likelihood function

$$\begin{aligned} L(\boldsymbol{\beta}, \pi) =& \prod_{i=1}^{n} \int_{\boldsymbol{y}_{M(i)}} \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} f_k(y_{ij}|\boldsymbol{\beta}_{kj}) d\boldsymbol{y}_{M(i)} \\ =& \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \prod_{j \in O(i)} f_k(y_{ij}|\boldsymbol{\beta}_{kj}). \end{aligned} \tag{2}$$

Likelihood (2) can be maximized by an E-M algorithm (McLachlan and Krishnan, 2008) that iteratively maximizes an updating function $Q(\boldsymbol{\pi}, \boldsymbol{\beta}|\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\beta}})$ of the parameters $\boldsymbol{\pi}, \boldsymbol{\beta}$, evaluated on the basis of a previous estimate $\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\beta}}$. Hunt and Jorgensen (2003) developed an E-M algorithm for model-based clustering of incomplete mixed categorical and continuous data. Their method can be easily adapted for estimating the parameters of model (1). To illustrate, we first introduce a binary random variable δ_{ik} with expected value

$$\mathbb{E}\delta_{ik} = \pi_{ik} = \frac{\pi_k \prod_{j \in O(i)} f_k(y_{ij}|\boldsymbol{\beta}_{kj})}{\sum_{k=1}^{K} \pi_k \prod_{j \in O(i)} f_k(y_{ij}|\boldsymbol{\beta}_{kj})}. \tag{3}$$

Expectation $\pi_{ik} = P(Z = k|\boldsymbol{y}_{O(i)}; \boldsymbol{\pi}, \boldsymbol{\beta})$ is known up the parameters $(\boldsymbol{\pi}, \boldsymbol{\beta})$ of model (1) and indicates the conditional probability of vector $\boldsymbol{y}_{O(i)}$ to belong to the kth latent class. If the value δ_{ik} was observed, the likelihood function would be given by

$$L_c(\beta, \pi) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\pi_k \prod_{j \in O(i)} f_k(y_{ij}|\boldsymbol{\beta}_{kj}) \right)^{\delta_{ik}}.$$

Because the δs variables are not observed, L_c is a random variable, known as the complete data likelihood function in the E-M terminology. Given an estimate $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\beta}})$, we can therefore evaluate the expected value of $\log L_c$, which is given by

$$\begin{aligned} Q(\beta, \pi|\hat{\beta}, \hat{\pi}) =& \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{\pi}_{ik} \left(\log \pi_k + \sum_{j \in O(i)} \log f_k(y_{ij}|\boldsymbol{\beta}_{jk}) \right) \\ =& Q(\boldsymbol{\pi}|\hat{\beta}, \hat{\pi}) + \sum_{j=1}^{J} Q_j(\boldsymbol{\beta}_{kj}|\hat{\beta}, \hat{\pi}) \end{aligned}$$

where

$$\begin{aligned} Q(\boldsymbol{\pi}|\hat{\beta}, \hat{\pi}) =& \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{\pi}_{ik} \log \pi_k \\ Q_j(\boldsymbol{\beta}_{kj}|\hat{\beta}, \hat{\pi}) =& \sum_{i \in m(j)} \sum_{k=1}^{K} \hat{\pi}_{ik} \log f_k(y_{ij}|\boldsymbol{\beta}_{kj}), \end{aligned}$$

and

$$\hat{\pi}_{ik} = \frac{\hat{\pi}_k \prod_{j \in O(i)} f_k(y_{ij}|\hat{\boldsymbol{\beta}}_{kj})}{\sum_{k=1}^{K} \hat{\pi}_k \prod_{j \in O(i)} f_k(y_{ij}|\hat{\boldsymbol{\beta}}_{kj})}. \tag{4}$$

At any step, the E-M algorithm exploits the previous estimate $\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\beta}}$ to compute the probabilities $\hat{\pi}_{ik}$ (E-step) and (M-step) maximizes $Q(\boldsymbol{\pi}, \boldsymbol{\beta}|\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\beta}})$ to obtain a new estimate. The algorithm is iterated up to convergence of the estimates, whose limit (Wu, 1983) is the maximum likelihood estimate of model (1). We remark that, due to the conditional independence assumption, the M-step can be carried out by maximizing separately $Q(\boldsymbol{\pi}|\hat{\beta}, \hat{\pi})$ and $Q_j(\boldsymbol{\beta}_{kj}|\hat{\beta}, \hat{\pi})$. In particular, the maximum point of $Q(\boldsymbol{\pi}|\hat{\beta}, \hat{\pi})$ is available in closed form and it is equal to

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

The updating equations for parameters $\boldsymbol{\beta}$ depend on the form of the densities $f_k(y_j|\boldsymbol{\beta}_{kj})$.

On the basis of the maximum likelihood estimates obtained after the last iteration of the algorithm, the multivariate conditional distribution of the missing values given the observed data can be evaluated in closed form as follows:

$$f(\boldsymbol{y}_{M(i)}|\boldsymbol{y}_O(i)) = \sum_{k=1}^{K} \hat{\pi}_{ik} \prod_{j \in M(i)} f_k(y_{ij}|\hat{\beta}_{kj}). \tag{5}$$

We propose (5) as a flexible imputation model in incomplete datasets where relevant covariates are not available and variables are measured on different scales, such as in the case study that is described in Section 3.

Because these conditional distributions take the form of a mixture, we can easily draw the desired number H of imputations for each $\boldsymbol{y}_{M(i)}$, by first sampling H binary vectors $\boldsymbol{d}_1(i) \dots \boldsymbol{d}_h(i) \dots \boldsymbol{d}_H(i)$ from a multinomial distribution with probability masses $\hat{\pi}_{i1} \dots \hat{\pi}_{iK}$, where $\boldsymbol{d}_h(i) = (d_{h1}(i) \dots d_{hk}(i) \dots d_{hK}(i))$. We secondly draw H independent vectors $\boldsymbol{y}^h_{M(i)}, h = 1 \dots H$, from

$$\prod_{k=1}^{K} \left(\prod_{j \in M(i)} f_k(y_{ij}|\hat{\beta}_{kj}) \right)^{d_{hk}(i)},$$

in order to obtain H complete versions of the vectors $(\boldsymbol{y}_{M(i)}, \boldsymbol{y}_{O(i)})$ that include missing values. The resulting H complete datasets can be then used for subsequent analysis, by following the theory of MI-based inference, developed by Robins and Wang (1998, 2000).

3. Data

The Italian Institute for Environmental Research and Protection (ISPRA) 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 coast, additionally provide data about wind direction and speed. Wave-metric data are often

Table 1. Number of missing values

Site	Measurement (Unit)	Percentages of missing values
Mazara del Vallo (buoy)	Wave Height (Meters)	13%
	Wave Direction (Radians)	13%
Porto Empedocle (tide gauge)	Wind Speed (Meter/Sec)	3%
	Wind Direction (Radians)	8%
Lampedusa (tide gauge)	Wind Speed(Meter/Sec)	1%
	Wind Direction (Radians)	1%

incomplete for a variety of reasons, which include buoy maintenance, discontinuous devices functioning and transmission errors. On the contrary, tide gauges are generally capable to provide data that are essentially complete. Imputation of missing values is complicated by the different locations at which wind and wave data are measured. The relationship between data observed at a buoy and those observed at the nearest tide gauges is highly non-linear, depending not only on the distance between the tide gauge and the buoy, but also on the oleography of the site where the tide gauge is located. As a result, linear correlations between wind and wave can be significantly small, in contrast to the high correlation level between wind and wave data that is normally observed when these variables are observed at the same location.

The data that we have exploited in this work include hourly measurements of wave height and direction, taken in the period 10/13/2003-11/11/2003 by the buoy of Mazara del Vallo, which is located at about 10 Km from the southern coast of Sicily. Eight-hours moving averages of wind speed and direction were obtained from the two nearest tide gauges, respectively located at Porto Empedocle (Sicilian coast, about 100 Km from the buoy) and at Lampedusa Island (about 250 Km from the buoy).

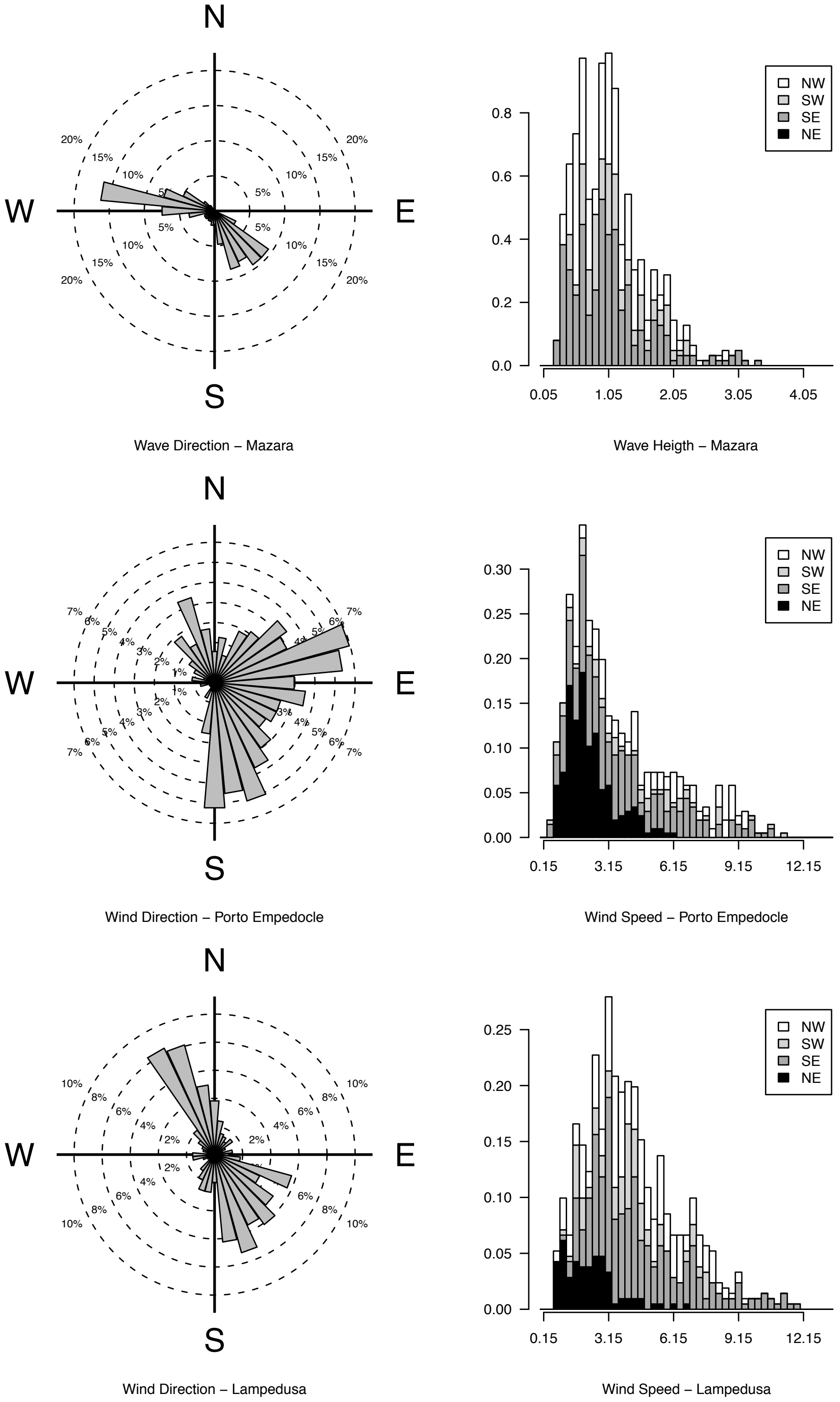
Table 1 reports the percentages of missing data in the study period. Missing values typically occur in wave-metric data, while observations on wind that are taken at the tide gauges are essentially complete.

Univariate distributions of the available data are displayed in Figure 1. Rose diagrams indicate the distribution of directions from where the wind and the wave come from. Waves are concentrated on two modal directions (west and south-east). North-west and south-east are the modal directions of wind at Lampedusa, while a three-mode distribution is observed at Porto Empedocle. The stacked histograms show the conditional distribution of wave height and wind speed, given wave direction, clustered into four quadrants (north-east (NE), north-west (NW), south-west (SW) and south-east (SE).). Visual comparison between the histograms and the rose diagrams is a first-glance evidence of the existence of a number of latent clusters. A mixture model of the type considered in Section 2 takes advantage of this clustering to impute missing values from the observed data.

4. Application

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 buoy), Y_2 (wind direction at Porto Empedocle) and Y_3 (wind direction at Lampedusa). A second group includes three linear variables, say Y_4 (wave height at the buoy), Y_5 (wind speed at Porto Empedocle) and Y_6 (wind speed at Lampedusa).

The procedure of Section 2 is flexible enough to allow each random variable to be modeled in a different way. The choice of specific parametric distributions is therefore left to the imputer and depends on the application.



Wave Direction – Mazara

Wave Heigth – Mazara

Wind Direction – Porto Empedocle

Wind Speed – Porto Empedocle

Wind Direction – Lampedusa

Wind Speed – Lampedusa

Figure 1. Distribution of the available wave metric data at the buoy of Mazara del Vallo and wind data at the two nearest tide gauges (Porto Empedocle and Lampedusa).

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

$$f_k(y|\boldsymbol{\beta}_{kj}) = \mathrm{VM}(\beta_{kj0}, \beta_{kj1}) = \frac{\exp(\beta_{kj1}\cos(y - \beta_{kj0}))}{2\pi I_0(\beta_{kj1})}, \quad j = 1, 2, 3 \tag{6}$$

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

		Component				
	Parameters	1	2	3	4	5
Wave Direction	mean	2.4839	2.6117	4.8271	4.1006	5.1694
(radians)		(0.0281)	(0.0291)	(0.0124)	(0.0809)	(0.0214)
	concentration	12.9903	9.9137	60.9952	2.1765	31.7242
		(2.0065)	(1.4160)	(8.3103)	(0.2390)	(5.1457)
Wind Direction [a]	mean	2.8588	1.8103	6.0255	2.0685	0.7845
(radians)		(0.0266)	(0.0683)	(0.0680)	(0.1045)	(0.0497)
	concentration	13.0170	2.0389	2.7433	1.2663	5.0242
		(1.8801)	(0.2183)	(0.3705)	(0.1675)	(0.7518)
Wind Direction [b]	mean	2.7962	2.1582	5.6115	4.5549	0.0111
(radians)		(0.0330)	(0.0422)	(0.0866)	(0.3186)	(0.0484)
	concentration	8.1217	4.5904	1.7237	0.4065	5.2683
		(1.0559)	(0.5828)	(0.2258)	(0.1202)	(0.8137)
Wave Height	shape	8.1192	5.0405	11.3646	14.4668	14.2928
(meters)		(1.1076)	(0.5673)	(1.5026)	(2.0493)	(2.9057)
	scale	0.1958	0.1398	0.1366	0.0686	0.0428
		(0.0272)	(0.0170)	(0.0180)	(0.0097)	(0.0094)
Wind Speed [a]	shape	10.3998	7.1708	3.7892	2.7154	7.2170
(meters/sec)		(1.3219)	(0.8604)	(0.4708)	(0.3058)	(1.0448)
	scale	0.5949	0.2821	1.2745	0.8886	0.3840
		(0.0757)	(0.0360)	(0.1608)	(0.1138)	(0.0573)
Wind Speed [b]	shape	7.1961	6.2779	3.9368	5.4507	2.8335
(meters/sec)		(0.9824)	(0.7086)	(0.4967)	(0.6527)	(0.3888)
	scale	0.9000	0.5529	1.0773	0.5496	1.4104
		(0.1233)	(0.0645)	(0.1407)	(0.0690)	(0.2089)
	component prob.	0.1884	0.2272	0.1988	0.2357	0.1500
		(0.0153)	(0.0169)	(0.0173)	(0.0184)	(0.0147)

[a] Tide gauge: Porto Empedocle

[a] Tide gauge: Lampedusa

where the parameters β_{kj0} and β_{kj1}, $j = 1, 2, 3$, respectively indicate the mean direction and concentration of each conditional circular distribution, given the 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.

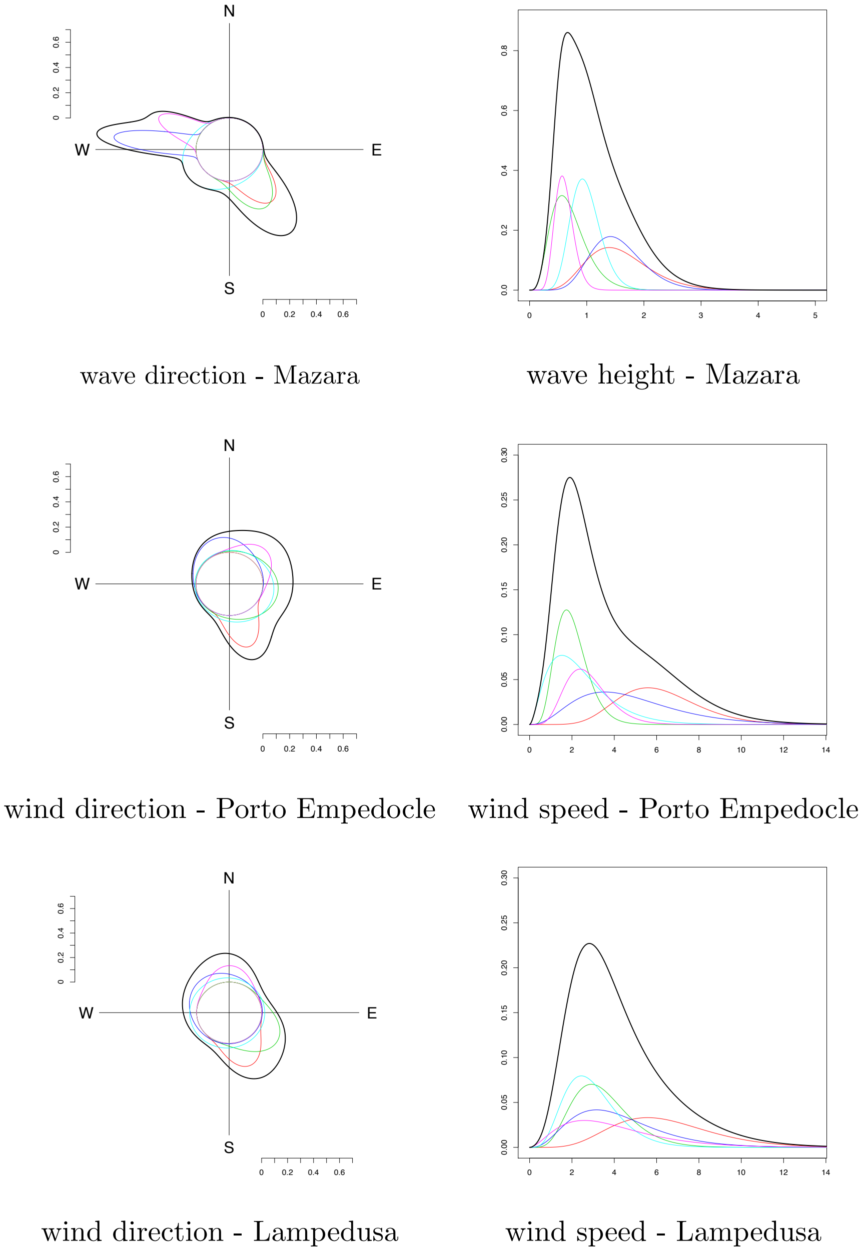
$$f_k(y|\boldsymbol{\beta}_{kj}) = \mathrm{Gam}(\beta_{kj0}, \beta_{kj1}) = \frac{\beta_{kj0}^{\beta_{kj1}} y^{\beta_{kj1}-1} \exp -(y/\beta_{kj0})}{\Gamma(\beta_{kj1})}, \quad j = 4, 5, 6 \qquad (7)$$

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

According to the BIC criterion, we selected $K = 5$ latent classes. Table 2 displays the maximum likelihood estimates and the standard errors of the $K \times 12 + K = 60+5$ parameters of the model. Point estimates were computed by exploiting the E-M algorithm of Section 2 (details about the M-step are reported in the Appendix). Standard errors were computed by taking the square root of the diagonal elements of the inverse observed Fisher information matrix, by numerically approximating the second derivatives of the marginal likelihood function (2).

The similar sizes of the five latent classes (bottom row of Table 2) and the strong significance of most of the parameters reflect the high level of latent heterogeneity of the data, partly explained by the absence of covariates in our model.

Figure 2 displays the 12×5 densities that have been estimated under the mixture model and helps to interpret the estimates of Table 2. To draw this picture, we have used five different colors to show the grouping of the conditional densities according to the five latent class of the multivariate mixture models. Components 1 and 2 cluster S-E wave and wind directions and distinguish between lower levels of wind speed and wave height (component 1) and higher levels of wind speed and wave



wave direction - Mazara

wave height - Mazara

wind direction - Porto Empedocle

wind speed - Porto Empedocle

wind direction - Lampedusa

wind speed - Lampedusa

Figure 2. wave and wind direction and intensity distributions, as estimated by a 5-component mixture model; red, green, bleu, azure, magenta indicate components 1-5, respectively.

height (component 2). N-W wave directions are associated to components 3 and 5, and distinguish between lower levels of wind speed and wave height (component 5) and higher levels of wind speed and wave height (component 3). Component 4 cluster S-W waves of medium height that are associated with moderate wind speed.

This model-based clustering of the data is a typical outcome of mixture models such as that proposed in this paper. Specifically, under the mixture model (1), either complete or incomplete observations can be clustered by assigning each observation i to the latent class k with the highest probability $\hat{\pi}_{ik}$ (modal allocation). In our case, modal-allocation criterion of the data yielded the classification displayed in Figures 3 and 4, which additionally show 95% fiducial intervals and the overall fit of the model on each of the six variables of interest. Colors are consistent with

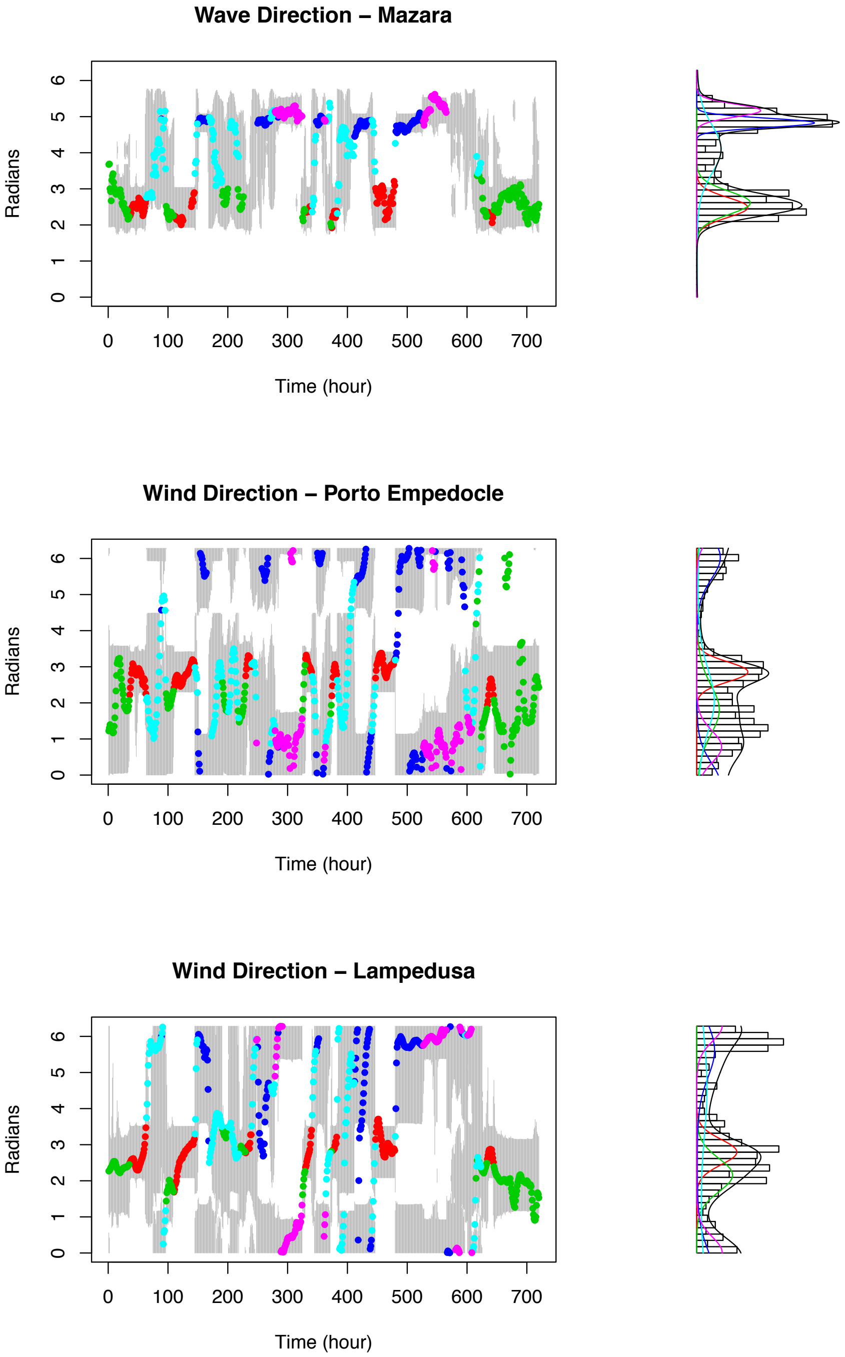


Figure 3. Left: directional data, clustered into five latent classes (red, green, bleu, azure, magenta indicate components 1-5, respectively) and 95% fiducial intervals (grey), as estimated by a 5-components mixture model: wave direction (top) and wind direction at Porto Empedocle (middle) and Lampedusa (bottom). Right: histograms of complete data fitted by the model.

those used for the figures above. The 95% fiducial intervals cover most of the observations, indicating a good fit of the model, although some extreme events are not well predicted.

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

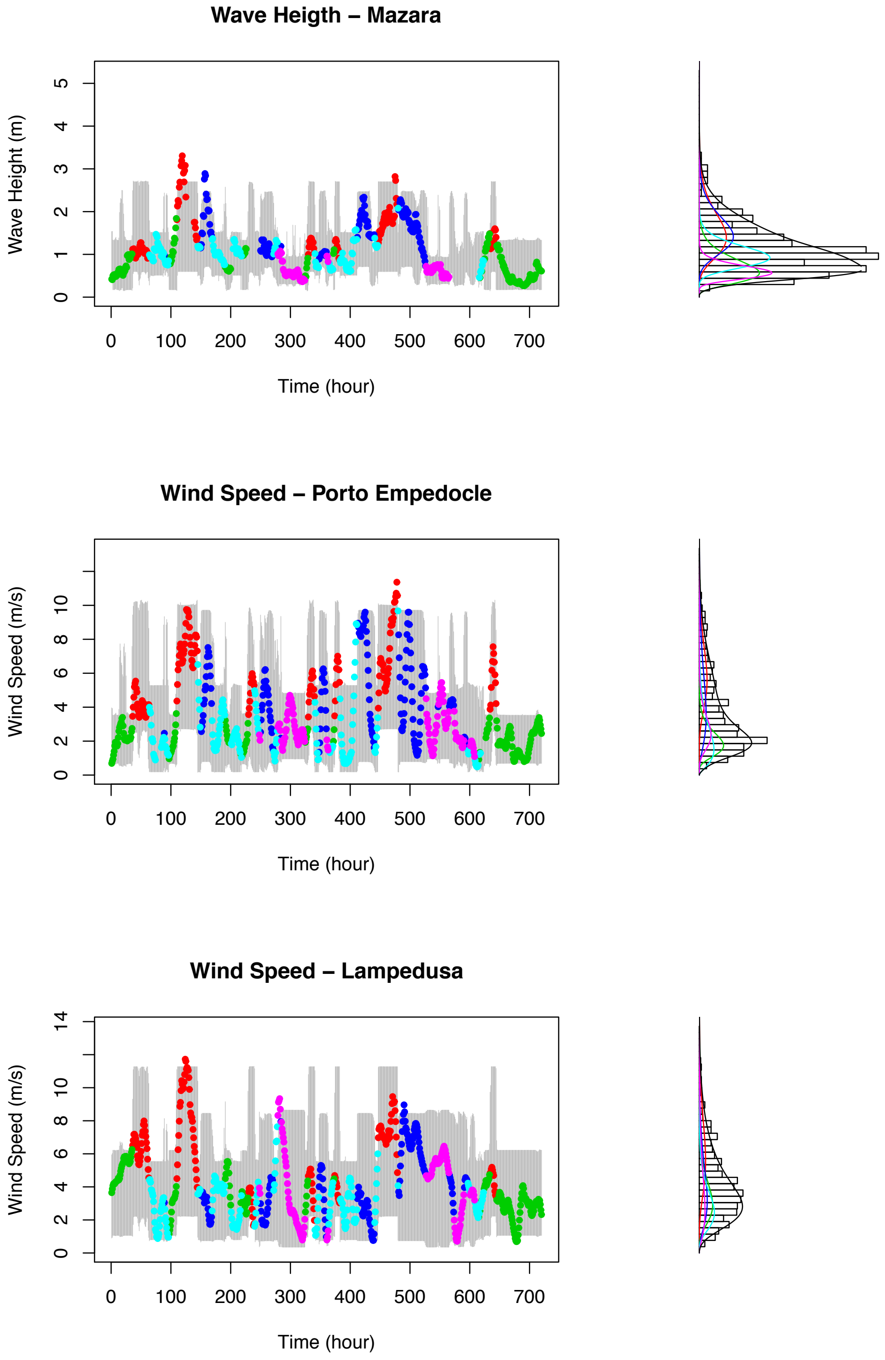


Figure 4. Left: intensity data, clustered into five latent classes (red, green, bleu, azure, magenta indicate components 1-5, respectively) and 95% fiducial intervals (grey), as estimated by a 5-components mixture model: wave height (top) and wind speed at Porto Empedocle (middle) and Lampedusa (bottom). Right: histograms of complete data fitted by the model.

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). The expected counterparts of these empirical correlations, under model (1),

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]	0.212	1				
(expected)	(0.344)	(1)				
Wind Speed [b]	0.068	0.112	1			
(expected)	(0.149)	(0.100)	(1)			
Wave Direction	0.002	0.015	0.056	1		
(expected)	(0.002)	(0.022)	(0.045)	(1)		
Wind Direction [a]	0.153	0.106	0.050	0.065	1	
(expected)	(0.087)	(0.223)	(0.046)	(0.037)	(1)	
Wind Direction [b]	0.075	0.038	0.033	0.000	0.022	1
(expected)	(0.111)	(0.092)	(0.057)	(0.000)	(0.024)	(1)

[a] Tide gauge: Porto Empedocle

[b] Tide gauge: Lampedusa

are given in the Appendix. Table 4 displays a reasonable matching between the empirical correlations against their expected counterparts, showing that the conditional independence assumption of model (1) explains a significant part of data variability.

Because we propose model (1) as an imputation model, we also evaluated the predictive accuracy of multiple imputations by cross-validation. More precisely, we randomly split the sample in 10 subsamples. From each subsample, we discarded the 10% of the observations and use the remaining portion of the subsample to draw 5 imputations for each discarded vector of 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, ... highest) would be equally likely. Figure 5 displays the cumulative distribution functions of the 6 ranks of circular and linear outcomes, showing the good accuracy of the multiple imputation procedure suggested in this paper.

5. Discussion

We propose a mixture model to draw multiple imputations in multivariate datasets with circular and linear variables that are partially observed. A conditional independence assumption between variables allows for a simple specification of the dependence structure between variables that are measured on different scales 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. Although this assumption can be unrealistic in certain settings, in our case study it was capable to explain most of the data variability and to multiply impute missing value with a reasonable accuracy. A serious shortcoming of the mixture model (1) is the assumption of temporal independence between observations. On one hand, the comparison between expected and empirical cross-correlations suggests that a reasonably large number of latent classes is capable to remove most of the temporal dependence between observations. On the other hand, accounting for temporal dependence should lead to a more parsimonious model, which yields multiple imputations of good quality, with a reduced number of latent classes. Intuitively, the inclusion of temporal dependence in the model should in particular enhance the performance of the mixture approach to imputation when long sequences of consecutive missing data occur. Hidden Markov models (Holtzmann *et*

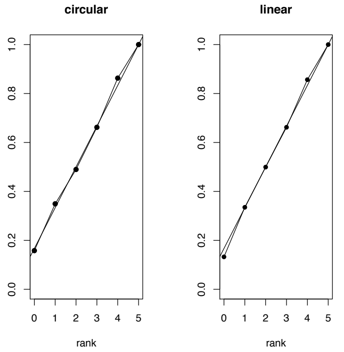


Figure 5. Rank cumulative distribution of the actual outcome with respect to 5 multiple imputations in a cross-validation experiment and cumulative distribution function of a uniform distribution.

al., 2006) for linear-circular data are a natural extension of the mixture considered in this work and could be exploited to multiply impute incomplete, multivariate time series of mixed environmental data. In our case study, we did not have long series of consecutive missing data and the simpler mixture model for independent observations was considered.

Appendix

M-step

The $K \times 12 + K$ parameters of the model were obtained by the E-M algorithm of Section 2. The 6 functions $Q_j(\boldsymbol{\beta}_j|\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\beta}})$ can be maximized separately as follows. For $j = 1, 2, 3$ (von Mises components),

- $\hat{\beta}_{kj0} = \text{arctg} \frac{\sum_{i \in m(j)} \hat{\pi}_{ik} \sin y_{ij}}{\sum_{i \in m(j)} \hat{\pi}_{ik} \cos y_{ij}}$
- while $\hat{\beta}_{kj1}$ is the root of

$$\frac{I_0(\beta_{kj1})}{I_0'(\beta_{kj1})} = \frac{\sum_{i \in m(j)} \hat{\pi}_{ik} \cos(y_{ij} - \hat{\beta}_{kj0})}{\sum_{i \in m(j)} \hat{\pi}_{ik}}$$

For $j = 4, 5, 6$ (Gamma components), instead:

- $\hat{\beta}_{kj0} = \frac{\sum_{i \in m(j)} \hat{\pi}_{ik} y_{ij}}{\hat{\beta}_{kj1} \sum_{i \in m(j)} \hat{\pi}_{ik}}$
- while $\hat{\beta}_{kj1} = A - \frac{\log A - \psi(A) - T}{\frac{1}{A} - \psi'(A)}$

where $A = \frac{3 - T + \sqrt{(T-3)^2 + 24T}}{12T}$ and $T = \log \frac{\sum_{i \in m(j)} \hat{\pi}_{ik} y_{ij}}{\sum_{i \in m(j)} \hat{\pi}_{ik}} - \frac{\sum_{i \in m(j)} \hat{\pi}_{ik} \log y_{ij}}{\sum_{i \in m(j)} \hat{\pi}_{ik}}$

Linear-circular cross-correlations

Under model (1), the marginal linear correlation between two Gamma components X and Y is given by

$$r(X,Y)=\frac{\sum_{k=1}^{K}\pi_k\mu_{k,X}\mu_{k,Y}-(\sum_{k=1}^{K}\pi_k\mu_{k,X})(\sum_{k=1}^{K}\pi_k\mu_{k,Y})}{\sqrt{(\sum_{k=1}^{K}\pi_k(\sigma^2_{k,X}+\mu^2_{k,X})-(\sum_{k=1}^{K}\pi_k\mu_{k,X})^2)(\sum_{k=1}^{K}\pi_k(\sigma^2_{k,Y}+\mu^2_{k,Y})-(\sum_{k=1}^{K}\pi_k\mu_{k,Y})^2)}}, \tag{8}$$

where $\mu_{k,X}=\mathbb{E}X$ and $\mu_{k,Y}=\mathbb{E}Y$, while $\sigma^2_{k,X}$ and $\sigma^2_{k,Y}$ are the variances of X and Y.

Generalizing the circular-coefficient by Fisher and Lee (1983) to the case of multivariate mixtures, we obtain that, under model (1), the marginal circular correlation between two von Mises components X and Y, is given by

$$r(X,Y)=\frac{[R^2(X-Y)-R^2(X+Y)]}{\sqrt{(1-R^2(2X))(1-R^2(2Y))}}$$

where

$$\begin{aligned}
R^2(X-Y)=&\Bigg[\sum_{k=1}^{K}\pi_k\frac{I_1(\nu_{Xk})I_1(\nu_{Yk})}{I_0(\nu_{Xk})I_0(\nu_{Yk})}\sin(\mu_{Xk}-\mu_{Yk})\Bigg]^2+\\
&+\Bigg[\sum_{k=1}^{K}\pi_k\frac{I_1(\nu_{Xk})I_1(\nu_{Yk})}{I_0(\nu_{Xk})I_0(\nu_{Yk})}\cos(\mu_{Xk}-\mu_{Yk})\Bigg]^2\\
R^2(X+Y)=&\Bigg[\sum_{k=1}^{K}\pi_k\frac{I_1(\nu_{Xk})I_1(\nu_{Yk})}{I_0(\nu_{Xk})I_0(\nu_{Yk})}\sin(\mu_{Xk}+\mu_{Yk})\Bigg]^2+\\
&+\Bigg[\sum_{k=1}^{K}\pi_k\frac{I_1(\nu_{Xk})I_1(\nu_{Yk})}{I_0(\nu_{Xk})I_0(\nu_{Yk})}cos(\mu_{Xk}+\mu_{Yk})\Bigg]^2\\
R^2(2X)=&\Bigg[\sum_{k=1}^{K}\pi_k\frac{I_2(\nu_{Xk})}{I_0(\nu_{Xk})}\sin(2\mu_{Xk})\Bigg]^2+\Bigg[\sum_{k=1}^{K}\pi_k\frac{I_2(\nu_{Xk})}{I_0(\nu_{Xk})}\cos(2\mu_{Xk})\Bigg]^2\\
R^2(2Y)=&\Bigg[\sum_{k=1}^{K}\pi_k\frac{I_2(\nu_{Yk})}{I_0(\nu_{Yk})}\sin(2\mu_{Yk})\Bigg]^2+\Bigg[\sum_{k=1}^{K}\pi_k\frac{I_2(\nu_{Yk})}{I_0(\nu_{Yk})}cos(2\mu_{Yk})\Bigg]^2,
\end{aligned}$$

where μ_{Xk} and μ_{Yk} are the directional means of the two conditional von Mises densities, given latent class k and ν_{Xk} and ν_{Yk} are the conditional concentration parameters.

Finally, following the Mardia's (1983) proposal for a linear- circular correlation index, the squared marginal linear-circular correlation between a Gamma component X and a von Mises component Y is given by

$$r^2(X,Y)=\frac{[r^2(X,\cos Y)+r^2(X,\sin Y)-2r(X,\cos Y)r(X,\sin Y)r(\cos Y,\sin Y)]}{1+r^2(\cos Y,\sin Y)}.$$

Exploiting (8) to compute r in the formula above, we obtain the marginal correlation between a von Mises and a Gamma component, under model (1).

References

[1] M. Di Zio, U. Guarnera, O. Luzi (2007) Imputation through finite Gaussian mixture models *Computational Statistics & Data Analysis*, 51, pp. 5305-5316.

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

[3] H. Holzmann, A. Munk, M. Suster and W. Zucchini (2006) Hidden Markov models for circular and linear-circular time series, *Environmental and Ecological Statistics*, 13, pp. 325347.

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

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

[7] G.J. McLachlan and T. Krishnan (2008) *The EM Algorithm and Extensions*, Wiley: New York

[7] G.J. McLachlan and D. Peel (2000). *Finite Mixture Models.* New York: John Wiley.

[8] J.M. Robins and N Wang (2000) Inference for imputation estimators, *Biometrika*, 87, pp. 113-124.

[9] D.B. Rubin (1987) *Multiple Imputation for Nonresponse in Surveys* New York: John Wiley.

[10] D.B. Rubin (1996) *Multiple Imputation After 18+ Years* Journal of the American Statistical Association, 91, pp. 473-489.

[11] J.L. Schafer (1997) *Analysis of Incomplete Multivariate Data*, London: Chapmanan d Hall.

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

[13] N. Wang and J.M. Robins (1998) Large-sample theory for parametric multiple imputation procedures, *Biometrika*, 85, pp. 935-948.

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