A hidden semi-Markov model for segmenting environmental toroidal data

Francesco Lagona [1,3] & Antonello Maruotti [2,3]

[1] *University of Roma Tre, via Chiabrera 199 00145 Rome (Italy) - francesco.lagona@uniroma3.it*
[2] *LUMSA University, Via della Traspontina, 21 - 00193 Rome (Italy) - a.maruotti@lumsa.it*
[3] *Dept. of Mathematics, University of Bergen, Allégaten 41, 5007 Bergen (Norway)*

Abstract. Toroidal time series are temporal sequences of bivariate angular observations that often arise in environmental and ecological studies. A hidden semi-Markov model is proposed for segmenting these data according to a finite number of latent classes, associated toroidal densities. The model conveniently integrates circular correlation, multimodality and temporal auto-correlation. A computationally efficient EM algorithm is proposed for parameter estimation. The proposal is illustrated on a time series of wind and sea wave directions.

Keywords. hidden semi-Markov model, EM algorithm, model-based clustering, toroidal data

1 Introduction

Bivariate sequences of angles are often referred to as toroidal time series, because the pair of two angles can be represented as a point on a torus. These data often arise in environmental and ecological studies. Examples include time series of wind and wave directions [8], time series of wind mean directions and directions of the maximum gust observed each day [2] and time series of turning angles in studies of animal movement [10].

The analysis of toroidal time series is complicated by the difficulties in modeling the dependence between angular measurements over time [7]. An additional complication is given by the multimodality of the marginal distribution of the data, because environmental toroidal data are observed under time-varying heterogeneous conditions.

This paper introduces a toroidal hidden semi-Markov model (HSMM) that simultaneously accounts for dependence across circular measurements, temporal auto-correlation, multimodality and latent time-varying heterogeneity. Under this model, the distribution of toroidal data is approximated by a mixture of toroidal densities, whose parameters depend on the evolution of a latent semi-Markov process. While the toroidal density

accommodates dependence between two circular variables, a mixture of toroidal densities allows for multimodality and, finally, a latent semi-Markov process accounts for temporal correlation and, simultaneously, for time-varying heterogeneity.

Our proposal extends previous approaches that are based on toroidal hidden Markov models [9, 1]. Under a toroidal hidden Markov model, the data are approximated by a mixture of toroidal densities, whose parameters depend on the evolution of a latent, first-order Markov chain with a finite number of states. The sojourn times of each state of a Markov chain are distributed according a geometric distribution. Hence the most likely dwell time for every state of a hidden Markov model with underlying first-order Markov chain is 1. Our proposal relaxes this restrictive assumption by replacing the latent Markov chain with a latent semi-Markov model, allowing for sojourn times that are not necessarily geometrically distributed.

2 A hidden semi-Markov model for toroidal data

Let $\mathbf{z} = (x, y)$ be a pair of angles, $x, y \in [0, 2\pi)$. Moreover, let $f(x; \alpha)$ and $f(y; \beta)$ be two circular densities, respectively known up to the parameters α and β. Further, let $F(x; \alpha)$ and $F(y; \beta)$ be the two cumulative distribution functions of x and y, defined with respect to a fixed, although arbitrary, origin. Finally, let $g(u; \gamma), u \in [0, 2\pi)$ be a parametric circular density, known up to a parameter γ. Then,

$$f_q(\mathbf{z}; \theta) = 2\pi g\left(2\pi\left(F(x; \alpha) - qF(y; \beta)\right)\right) f(x; \alpha) f(y; \beta)) \quad q = \pm 1 \tag{1}$$

is a parametric toroidal density with support $[0, 2\pi)^2$, known up to the parameter vector $\theta = (\alpha, \beta, \gamma)$, having the marginal densities $f(x; \alpha)$ and $f(y; \beta)$ [3]. Equation (1) is a typical example of a copula-based construction of a bivariate density, obtained by decoupling the margins from the joint distribution. When the binding density g is the uniform circular distribution, say $g(x) = (2\pi)^{-1}$, then equation (1) reduces to the product of the marginal densities. Otherwise, the dependence between x and y is captured by the concentration of g: when g is highly concentrated, the dependence is high; when g is more diffuse, dependence is low. Finally, the constant $q = \pm 1$ determines whether the dependence between x and y is positive ($q = 1$) or negative ($q = -1$).

The proposed hidden semi-Markov model can be described as a dynamic mixture of copula-based toroidal densities. To illustrate, let $\mathbf{z} = (\mathbf{z}_t, t = 1, \ldots, T)$, $\mathbf{z}_t = (x_t, y_t)$, $x_t, y_t \in [0, 2\pi)$, be a toroidal time series. We assume that the distribution of the data is driven by the evolution of an unobserved semi-Markov process with K states, which represents (time-varying) latent classes and can be specified as a sequence $\mathbf{u} = (\mathbf{u}_t, t = 1, \ldots, T)$ of multinomial variables $\mathbf{u}_t = (u_{t1} \ldots u_{tK})$ with one trial and K classes, whose binary components represent class membership at time t. The joint distribution $p(\mathbf{u}; \pi)$ of the chain is fully known up to a parameter π that includes K initial probabilities $\pi_k = P(u_{1k} = 1), k = 1, \ldots, K, \sum_k \pi_k = 1$, $K^2 - K$ transition probabilities $\pi_{hk} = P(u_{tk} =$

$1|u_{t-1,h} = 1), h, k = 1, \ldots, K, \sum_k \pi_{hk} = 1,\ h \neq k$ (whereas $\pi_{kk} = 0,\ k = 1 \ldots K$), and, finally, p parameters of the dwell time distributions of each state.

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

$$f(\mathbf{z}|\mathbf{u}; \theta_1, \ldots, \theta_K) = \prod_{t=1}^{T} \prod_{k=1}^{K} f(\mathbf{z}_t; \theta_k)^{u_{tk}}, \tag{2}$$

where $f(\mathbf{z}; \theta_k)$, $k = 1, \ldots, K$ are the K cylindrical densities defined by (1) and known up to a vector of parameters θ_k.

The likelihood function of the model is therefore obtained by integrating the joint density of the observed data and the unobserved class memberships with respect to the segmentation $\mathbf{u}$, namely

$$L(\pi, \theta; \mathbf{z}) = \sum_{\mathbf{u}} p(\mathbf{u}; \pi) f(\mathbf{z}|\mathbf{u}; \theta_1, \ldots, \theta_K). \tag{3}$$

By computing the maximum likelihood estimate $\hat{\theta}$ [11, chapter 12], the toroidal time series can be segmented according to the posterior probabilities of class membership

$$\hat{\pi}_{tk} = P(u_{tk} = 1 \mid \mathbf{z}; \hat{\theta}), \tag{4}$$

based on $\hat{\theta}$. More precisely, the observation at time t can be allocated to class $k^\star$ if $\hat{\pi}_{tk^\star} \geq \hat{\pi}_{th}$, for each $h = 1 \ldots K$ (maximum a posteriori, MAP, allocation).

When the dwell distribution of each latent state is geometric, the model reduces to a hidden Markov model that ignores alternative dwell time distribution. If, additionally, the transition probability matrix of the model has equal rows, the model reduces to a mixture model where observations are clustered by ignoring the information redundancy that is due to temporal correlation.

3 An application to marine data

The proposed methods have been implemented to segment a time series of $T = 1326$ semi-hourly wind and wave directions, taken in wintertime by the buoy of Ancona, which is located in the Adriatic Sea at about 30 km from the coast. 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 observations are plotted on the plane, although data points are actually on a torus. The interpretation of these data is not easy. While in the ocean wind and wave directions are strongly correlated, this is not necessarily true in the Adriatic Sea, due to the orography of the basin and the

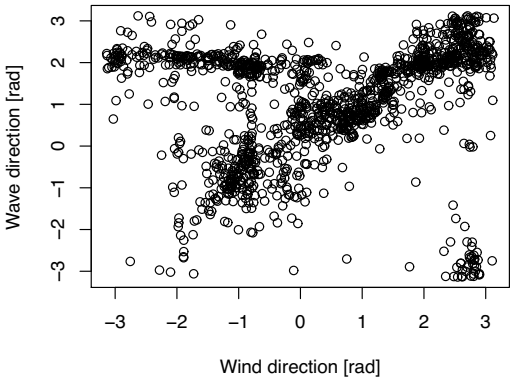

Figure 1: Wave directions and heights, as observed by the buoy of Ancona in wintertime ($-\pi$, $-pi/2$, 0, $\pi/2$ respectively indicate South, West, North, East). For simplicity, the data are plotted on the plane, although they are points on the torus $[-pi/2, \pi/2)^2$.

location of the buoy. Coastal winds generate synchronized waves only when the waves travel unobstructed, that is, either northwesterly or southeasterly, along the major axis of the basin. When western and south-western winds blow from the coast, waves are not synchronized with wind and travel along the major axis of the Adriatic basin from SE to NW. This explains the clusters shown in Figure 1 and suggests the occurrence of two latent wind–wave regimes. Accordingly, a HSMM with two states have been estimated from these data.

The proposed HMM requires a parametric specification of the toroidal density (1), which reduces to the choice of the binding density g and the choice of the marginal densities $f(x;\alpha)$ and $f(y;\beta)$ that respectively model the marginal distribution of the wind and wave direction. However, depending on the choice of the binding density, the density (1) can be multimodal [4]. Using multimodal densities in segmentation and classification problems, such as the one motivating this paper, may unnecessarily complicate the interpretation of the results. Unimodal densities can however be obtained by using the wrapped Cauchy as a binding density g [4].

Accordingly, for this study, the binding density has been specified as a centered wrapped Cauchy

$$g(u;\gamma) = \frac{1}{2\pi}\frac{1-\gamma^2}{1+\gamma^2-2\gamma\cos(u)} \quad u \in [0, 2\pi).$$

This circular density depends on a single concentration parameter $\gamma \in [0, 1)$ and reduces to the uniform circular density when $\gamma = 0$.

Wrapped Cauchy densities that include additional location parameters α_1 and β_1 have

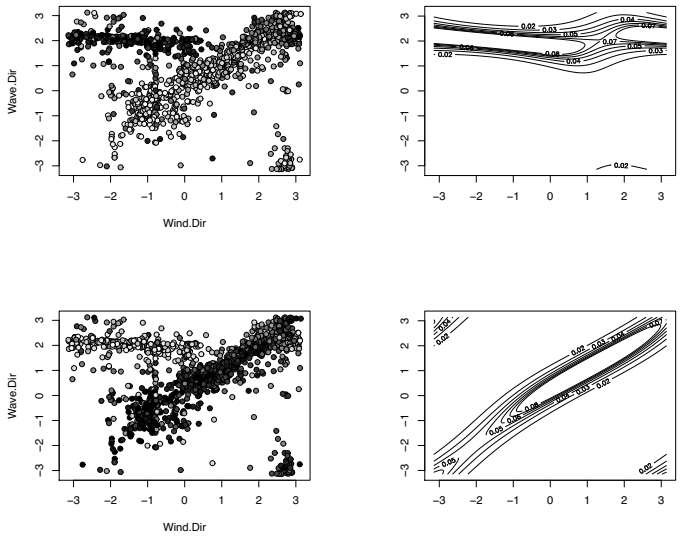

Figure 2: Segmentation of a time series of wind and wave directions. Left: observations colored with grey levels according to the estimated membership probabilities of each class (black indicates a probability equal to 1). Right: contour plot of state-specific toroidal densities.

been instead exploited to model the marginal distributions of wind and wave direction, say

$$f(x;\alpha) = \frac{1}{2\pi}\frac{1-\alpha_2^2}{1+\alpha_2^2-2\alpha_2\cos(y-\alpha_1)} \quad x \in [0,2\pi) \tag{5}$$

$$f(y;\beta) = \frac{1}{2\pi}\frac{1-\beta_2^2}{1+\beta_2^2-2\beta_2\cos(y-\beta_1)} \quad y \in [0,2\pi) \tag{6}$$

The proposed toroidal density is therefore obtained by taking a wrapped Cauchy density that binds wrapped Cauchy marginals, a model known as the bivariate wrapped Cauchy model [5].

Figure 2 shows the shapes of the two state-specific toroidal distributions and the segmented observations. The model successfully segment the observations according to to clusters, and offers a clear-cut indication of the distribution of the data under each regime. Under state 1, wind and wave directions are essentially independent, because coastal winds do not generate waves. Under state 2, winds blows along the major axis of the Adriatic basin and their directions are highly correlated with the directions of the wave that they generate.

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 con-

centrations 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 only under a specific regime (state 2).

References

[1] Bulla J, Lagona F, Maruotti A, Picone M (2012) A Multivariate Hidden Markov Model for the Identification of Sea Regimes from Incomplete Skewed and Circular Time Series, Journal of Agricultural, Biological, and Environmental Statistics, 17: 544-567

[2] Coles S (1998) Inference for circular distributions and processes, Statistics and Computing, 8: 105-113.

[3] Johnson RA, Wehrly TE (1978) Some angular-linear distributions and related regression models. Journal of the American Statistical Association 73: 602-606.

[4] Jones MC, Pewsey A, Kato S (2015). On a class of circulas: copulas for circular distributions. Annals of the Institute of Statistical Mathematics 67: 843-862.

[5] Kato S, Pewsey A (2015) A Möbius transformation-induced distribution on the torus, Biometrika, 102: 359-370

[6] Lagona F (2019) Copula-based segmentation of cylindrical time series, Statistical and Probability Letters, 144: 16-22.

[7] Lagona F (2018) Correlated cylindrical data. In: C. Ley and T. Verdebout (Eds) Applied Directional Statistics: Modern Methods and Case Studies, Chapman & Hall/CRC: New York, 45-59.

[8] Lagona F, Picone M, Maruotti A, Cosoli S (2014) A hidden Markov approach to the analysis of space-time environmental data with linear and circular components, Stochastic Environmental Research and Risk Assessment 29: 397-409.

[9] Lagona F, Picone M (2013) Maximum likelihood estimation of bivariate circular hidden Markov models from incomplete data, Journal of Statistical Computation and Simulation, 83: 1223-1237

[10] Mastrantonio G (2018) The joint projected normal and skew-normal: A distribution for poly-cylindrical data, Journal of Multivariate Analysis, 165: 14-26.

[11] Zucchini W, Macdonald IL and Langrock R (2016) Hidden Markov models for time series, Chapman and Hall, Boca Raton FL (US)