SEGMENTING TOROIDAL TIME SERIES BY NONHOMOGENEOUS HIDDEN SEMI-MARKOV MODELS

Francesco Lagona [1], Marco Mingione[1]

[1] Department of Political Sciences, University of Roma Tre, (e-mail: francesco lagona@uniroma3.it, marco.mingione@uniroma3.it)

ABSTRACT: Motivated by classification issues in marine studies, we propose a hidden semi-Markov model to segment toroidal time series according to a finite number of latent regimes. The time spent in a given regime and the chances of a regime-switching event are separately modeled by a battery of regression models that depend on time-varying covariates.

KEYWORDS: hidden semi-Markov model, toroidal data, model-based classification, wave, wind.

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. Examples include time series of wind and wave directions and time series of turning angles in studies of animal movement.

We introduce a nonhomogeneous, toroidal hidden semi-Markov model (HSMM) that segments toroidal time series. Precisely, the distribution of toroidal data is approximated by a mixture of toroidal densities, whose parameters evolve according to a latent semi-Markov process with covariate-specific dwell times.

Our proposal extends previous approaches that are based on toroidal hidden Markov models (Lagona & Picone, 2013). Under a toroidal hidden Markov model, the sojourn times of the states of the latent process are distributed according to a geometric distribution. Our proposal relaxes this restrictive assumption by replacing the latent Markov chain with a latent, nonhomogeneous semi-Markov model, where the (non necessarily geometric) time spent in a given regime and the chances of a regime-switching event are separately modeled by a battery of regression models that allow the introduction of covariates.

2 A toroidal hidden semi-Markov model

Let $\boldsymbol{y} = (\boldsymbol{y}_t, t = 1, \dots T)$ be a bivariate time series, where $\boldsymbol{y}_t = (y_{t1}, y_{t2})$ is a vector of two circular observations. Further, let $\boldsymbol{u} = (\boldsymbol{u}_t, t = 1, \dots T)$ be a sequence of latent multinomial random variables $\boldsymbol{u}_t = (u_{t1} \dots u_{tK})$ with one trial and K classes (or states), whose binary components represent class membership at time t. Our proposal is a hierarchical model where the joint distribution of the time series is obtained by

$$f(\boldsymbol{y}) = \sum_{\boldsymbol{u}} f(\boldsymbol{y} \mid \boldsymbol{u}) p(\boldsymbol{u}).$$

The joint distribution $p(\boldsymbol{u})$ of the latent process is described by extending the notion of a Markov chain. If $\mathbf{u}$ is a Markov chain, then $p(\boldsymbol{u})$ is fully known up to a vector of K initial probabilities $\pi_k = P(u_{1k} = 1), k = 1, \dots, K, \sum_k \pi_k = 1$, and a $K \times K$ matrix of transition probabilities

$$\begin{pmatrix} \pi_{11} & \pi_{12} & \dots & \pi_{1K} \\ \pi_{21} & \pi_{22} & \dots & \pi_{2K} \\ \dots & \dots & \dots & \dots \\ \pi_{K1} & \pi_{K2} & \dots & \pi_{KK} \end{pmatrix} = \begin{pmatrix} 1-p_1 & p_1\omega_{12} & \dots & p_1\omega_{1K} \\ p_2\omega_{21} & 1-p_2 & \dots & p_2\omega_{2K} \\ \dots & \dots & \dots & \dots \\ p_K\omega_{K1} & p_K\omega_{K2} & \dots & 1-p_K \end{pmatrix}$$

where $p_k = \sum_{k' \neq k} \pi_{kk'}$ is the probability of a transition from k to a different state and $\omega_{kk'}$ is the conditional probability of a transition to state $k' \neq k$, given a transition from state k. Under this setting, if the process is in state k, the time τ_k up to a transition to a different state is geometric

$$P(\tau_k = \tau) = p_k(1-p_k)^{\tau-1}. \tag{1}$$

More generally, let $S_k(\tau) = P(\tau_k > \tau) = \exp\left(-\int_0^\tau h_k(v)dv\right)$ be the survival function of τ_k, where $h_k(\tau)$ is the associated hazard function. Then

$$p_k(\tau) = P(\tau_k \leq \tau + 1 \mid \tau_k > \tau) = 1 - \exp\left(-\int_\tau^{\tau+1} h_k(v)dv\right),$$

is the conditional probability of a transition at time $t+1$, given that the process has been in state k during a period of length t. Then

$$P(\tau_k = \tau) = p_k(\tau) \prod_{i=1}^{\tau-1} (1 - p_k(\tau)). \tag{2}$$

When the hazard h_k is time-constant, then (2) reduces to (1). Alternatively, (2) can be approximated with the desired accuracy by

$$P(\tau_k = \tau) = p_k(m)(1 - p_k(m))^{\tau - m} \prod_{i=1}^{m-1} (1 - p_k(i)). \tag{3}$$

Parametric hazard functions can be borrowed from the survival analysis literature and some of them are conveniently associated to a link function g that trasforms $p_k(\tau)$ to a linear function of time, say $g(p_k(\tau)) = \beta_{0k} + \beta_{1k}\tau$. Such a specification can be further extended by introducing a vector of q (possibly time-varying) covariates, say $\boldsymbol{x}_t$, which influence the dwell time distribution

$$g(p_k(\tau; \boldsymbol{x}_t)) = \beta_{0k} + \beta_{1k}\tau + \boldsymbol{x}_t^{\mathsf{T}} \boldsymbol{\beta}. \tag{4}$$

Similarly, covariates may be introduced to shape the conditional transition probabilities, say $\omega_{kk'} = \omega_{kk'}(\boldsymbol{x}_t)$, through a multinomial regression equation. The introduction of time-varying covariates makes the latent process nonhomogeneous, extending recent literature proposals.

Our proposal is completed by a conditional independence assumption on the observation process. Precisely,

$$f(\boldsymbol{y} \mid \boldsymbol{u}) = \prod_{t=1}^{T} \prod_{k=1}^{K} \prod_{i=1}^{m} f(\mathbf{y}_t; \boldsymbol{\theta}_k)^{u_{tki}}, \tag{5}$$

where $\boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_K$ is a sequence of unknown parameters. Parametric toroidal densities can be borrowed by the proposals available in the directional statistics literature. A convenient specification is for example the bivariate wrapped Cauchy distribution (Kato & Pewsey, 2015). It is unimodal, pointwise symmetric and has a closed-form expression for the conditional distribution. A single dependence parameter controls the relationship between the two component circular variables, ranging from independence to perfect correlation. The remaining four parameters respectively indicate the two marginal means and concentrations.

3 Results

Figure 1 shows the results obtained on 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. A 2-state hidden semi-Markov model has been used to segment the data. The model integrates

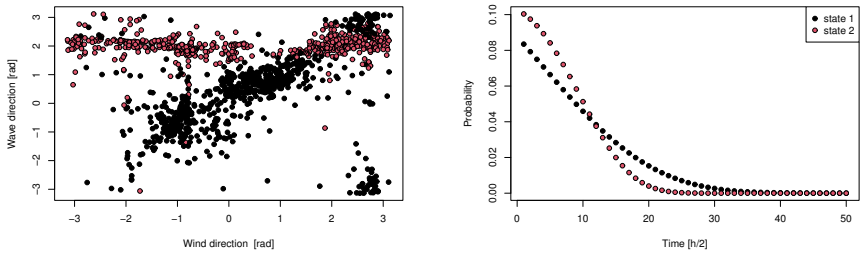

Figure 1. *Left: toroidal data clustered within state 1 (black) and state 2 (red). Right: state-specific dwell time distribution at baseline.*

bivariate wrapped Chaucy densities with dwell time regressions that depend on a baseline Gompertz hazard rate and a time-varying covariate, the *fetch*. The fetch is the closest coastal point following the direction from which the wave comes from and it is computed here by cyclical cubic smoothing splines (Wood, 2017) that appropriately smooth distances across the Adriatic basin.

The model successfully segments the observations according to two clusters, and offers a clear-cut indication of the distribution of the data under each regime. Under state 1, winds appear well syncronized with waves. Under state 2, wind and wave directions are essentially independent. Under state 1, the tail of the baseline dwell time distribution is larger than that one estimated under state 2, indicating that state 1 is more persistent than state 2. The regression coefficient of the fetch is equal to -1.38, indicating that the longer is the distance from the coast, the smaller is the probability of a state transition.

References

KATO, S., & PEWSEY, A. 2015. A Mobius transformation-induced distribution on the torus. *Biometrika*, **102**, 359–370.

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.

WOOD, SIMON N. 2017. *Generalized additive models: an introduction with R*. CRC Press.