Journal of the Royal Statistical Society Series C: Applied Statistics, 2024, **00**, 1–25
https://doi.org/10.1093/jrsssc/qlae049

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society Series C Applied Statistics

Original Article

Nonhomogeneous hidden semi-Markov models for toroidal data

Francesco Lagona and Marco Mingione

Department of Political Sciences, University of Roma Tre, Via G. Chiabrera, Rome 00145, Italy

Address for correspondence: Francesco Lagona, Department of Political Sciences, University of Roma Tre, Via G. Chiabrera, Rome 00145, Italy. Email: francesco.lagona@uniroma3.it

Abstract

A nonhomogeneous hidden semi-Markov model is proposed to segment bivariate time series of wind and wave directions according to a finite number of latent regimes and, simultaneously, estimate the influence of time-varying covariates on the process' survival under each regime. The model is a mixture of toroidal densities, whose parameters depend on the evolution of a semi-Markov chain, which is in turn modulated by time-varying covariates. It includes nonhomogeneous hidden Markov models and hidden semi-Markov models as special cases. Parameter estimates are obtained using an Expectation-Maximization algorithm that relies on an efficient augmentation of the latent process. Fitted on a time series of wind and wave directions recorded in the Adriatic Sea, the model offers a clear-cut description of sea state dynamics in terms of latent regimes and captures the influence of time-varying weather conditions on the duration of such regimes.

Keywords: circular data, dwell times, hidden semi-Markov model, model-based segmentation, wave, wind

1 Introduction

Pairs of circular observations are often referred to as toroidal data because they can be represented as points on a torus, the cartesian product of two circles. Toroidal data arise in numerous contexts. Examples include earthquake data consisting of the pre-earthquake direction of steepest descent and the direction of lateral ground movement (Rivest, 1997), peak systolic blood pressure times, converted to angles, during two separate time periods (Fisher & Lee, 1983), protein backbone conformational angles (Lennox et al., 2009), phase angles of circadian-related genes in two tissues (Liu et al., 2006), and orthologous genes shared by circular prokaryotic genomes (Shieh et al., 2011).

The statistical analysis of toroidal data is inherently different from the traditional analysis of bivariate continuous data, due to the wraparound nature of their domain and the difficulties in modelling the dependence between two angular measurements (circular correlation). Luckily, the spread of toroidal data across multiple disciplines has pushed the literature towards the definition of several distributions on the torus (see the reviews by Ley & Verdebout, 2017—Sections 2.4–2.5—and by Pewsey & García-Portugués, 2021—Section 3.2) and a whole library of toroidal distributions is nowadays available. As a result, when the data are in the form of a sequence of independent and identically distributed observations, they can be efficiently analyzed by fitting one of these distributions.

In most case studies, however, toroidal data are heterogeneous and dependent across space and/or time, motivating models where toroidal densities are just a building block among other model components. An important example arises in marine studies, where buoys routinely record time series of wind and wave directions to gain general knowledge about the evolution of sea conditions

Received: September 29, 2023. **Revised:** June 30, 2024. **Accepted:** September 14, 2024

and, more specifically, to address issues of coastal erosion, maritime transport and the drift of floating objects or oil spills (Monbet et al., 2007). In this context, the specification of a statistical model is quite intricate, because these data are not only heterogeneous with a distribution that dynamically varies across different latent sea regimes, but the duration of these regimes (dwell times) may vary according to time-varying weather conditions (e.g. wind speed), introducing complex auto-correlation structures.

In this paper, we model a time series of wind and wave directions by a novel nonhomogeneous hidden semi-Markov model (HSMM) that accounts for circular correlation and heterogeneity, simultaneously allowing for flexible dwell times. Under this model, the distribution of the data is approximated by a mixture of toroidal densities, whose parameters depend on the evolution of a latent, nonhomogeneous semi-Markov chain. While the toroidal density accommodates circular correlation, the mixture controls for heterogeneity and, finally, the semi-Markov chain allows flexible dwell times whose distribution is modulated by time-varying covariates and it is therefore nonhomogeneous.

The general idea of modelling environmental time series by finite mixtures that are driven by a nonhomogeneous latent chain is not new and dates back at least to the seminal work by Hughes et al. (1999), who modelled precipitation occurrences by a nonhomogeneous hidden Markov model (HMM). In that work, occurrences are modelled by a mixture of multivariate distributions: the parameters of these distributions depend on the evolution of a latent Markov chain and the transition probabilities of the chain depend on the evolution of time-varying weather conditions, thereby making the model nonhomogeneous. Under this setting, the states of the chain can be interpreted as weather states, and the model is capable of capturing the influence of time-varying weather conditions on state transitions. After two decades, nonhomogeneous HMMs are still routinely employed in environmental studies that require segmenting time series according to latent regimes and time-varying covariates are available (Ailliot et al., 2015; Jiang et al., 2023; Maruotti et al., 2017).

The dynamics of mixture models that are driven by nonhomogeneous latent chains are conveniently described by the state-specific dwell time hazard. This hazard is the discrete counterpart of the hazard function, traditionally exploited to characterize the distribution of absolutely continuous, positive random variables. Given the current state of the chain, the hazard indicates the (time-varying) conditional probability of switching state given the current dwell time, where the current dwell time is defined as the current amount of time since the last transition. The dynamics of a nonhomogeneous HMM, for example, is described by a hazard that is exogenously updated at each time by the current covariate information, regardless of the dwell time. This is a consequence of the memoryless property of the latent Markov chain in an HMM, and it is both a strength and a weakness of the model. On one side, it is conceptually appealing to assume that state switches should depend only on current conditions. On the other side, when these conditions do not change, the hazard is time-constant and dwell times are constrained to be geometrically distributed. This potential limitation of HMMs has opened the way to alternative approaches where the latent Markov chain is replaced by a semi-Markov chain, leading to the class of HSMMs. Under a semi-Markov chain, the hazard is endogenously updated by the current dwell time. This dissolves the geometric constraint on the dwell time distribution, by relaxing the Markov property to a setting where the evolution of the chain depends on its history since the last transition. Due to their flexibility, HSMMs have gained popularity in environmental research and they have been successfully implemented in studies of regime shifts in ocean density variability (Economou & Menary, 2019), GPS tracking data of animal movements (Pohle et al., 2022), multivariate pollutant concentrations (Merlo et al., 2022), and migratory bird count data (Nicol et al., 2023).

Dwell time hazards highlight the key difference between nonhomogeneous HMMs and HSMMs. In the first case, hazards vary exogenously with time according to time-varying covariate information. In the second case, hazards vary endogenously with time according to dwell times that are, at each time, updated by the chain itself. Our proposal relies on dwell time hazards that depend on both the dwell time and exogenous covariate information, allowing time-varying covariates to modulate the dwell time distribution in a novel way.

From one side, our proposal can be therefore viewed as an extension of current HSMMs to a setting that includes time-varying covariates. The literature on HSMMs typically relies on dwell

time mass probability functions that are known up to unknown parameters to be estimated, which are time-constant. As a result, these models are typically restricted to having time-constant covariates, which influence the dwell-time distribution at baseline. By working with hazards, we overcome this limitation.

On the other side, our model includes nonhomogeneous HMMs and HSMMs as special cases, which can then be viewed as complementary and reconciled under the same umbrella. Besides being of theoretical interest, integrating different models within a unified framework facilitates the development of a single Expectation-Maximization (EM) algorithm for estimating both nonhomogeneous HMMs and HSMMs.

The rest of the paper is organized as follows. The marine data that motivated this work are described in Section 2, while the model and the estimation procedures are respectively described in Sections 3, 4, and 5. Theg proposal is first tested on simulated data (Section 6) and then exploited to segment the marine data that motivated this study (Section 7). Section 8 finally summarizes relevant discussion points. The whole code to reproduce both the simulation experiments and the results on the real data is publicly available in a GitHub repository accessible at https://github.com/minmar94/HSMM_covariates.

2 Wind and wave direction

Time series of wind and wave directions are routinely collected by environmental agencies to identify sea regimes, that is, typical distributions that these data take under specific environmental conditions. Detecting environmental regimes and estimating their duration is crucial across several problems of marine research. Recent applications include studies of the drift of floating objects and oil spills (Gu et al., 2024), the design of marine structures (Fang et al., 2022), and the analysis of coastal erosion (Flor-Blanco et al., 2021), and extreme wave events at the coast (Kalisch et al., 2024).

The data that motivated this work are in the form of a time series of $T = 1{,}326$ semi-hourly wave and wind directions, recorded between 15th February and 16th March 2010 by the Ancona buoy, located in the Adriatic Sea at about 30 km from the coast (Figure 1a). This buoy is a floating device anchored to the seabed and it collects weather and oceanographic information. Precisely, the data are gathered in the form of angles that indicate the average direction *from* which the wave travels and the wind blows, during a period of 30 min. Along with wind and wave directions, the buoy also records wind speed (m/s) at the same time resolution. Figure 1b and 1c shows the two marginal distributions of wind and wave directions, in the form of rose diagrams. The data joint distribution is instead shown by Figure 2 in the form of two scatterplots, one with dots displayed on the original toroidal manifold (Figure 2b), and the other one with dots projected on the plane that is obtained by unwrapping the torus (Figure 2a). Dots are filled with colours according to the classes of wind speed.

Some features of the data can be straightforwardly interpreted by recalling the meteorology of the sea that surrounds the Ancona buoy. In wintertime, waves in the Adriatic Sea are typically generated by the southeastern Sirocco wind and the northern Bora wind. These conditions can be associated with the two modes of the rose diagram in Figure 1b. Sirocco arises from a warm, dry, tropical air mass that is pulled northward by low-pressure cells moving eastward 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. Remarkably, wind and wave directions are not always synchronized, as shown by the data scatterplot (Figure 2a). In the open sea, waves can travel freely, without being obstructed by physical obstacles, such as coastlines. As a result, the wind energy is fully transferred to the sea surface and wind and wave directions are highly correlated. In a semi-enclosed basin like the Adriatic Sea, instead, wave direction is modulated by the orography of the area and, as a result, wind and wave directions are not necessarily highly correlated. Orographic effects are often held responsible for the inaccuracy of numerical wave models in the Adriatic Sea (Bertotti & Cavaleri, 2009), motivating statistical methods that segment these data into a small number of latent classes, conditionally on which the distribution of the data takes a shape that is easier to interpret than the shape taken by the marginal distribution. Figure 2a, where observations are coloured according to

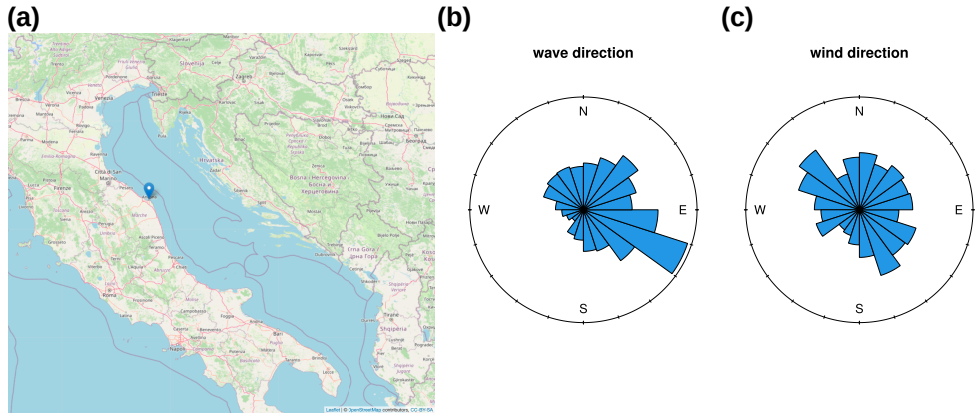

Figure 1. (a) The Adriatic sea and the location of the Ancona buoy (Latitude 43°37′29.16″ N; Longitude 13°30′23.46″ E); (b) observed wave directions and (c) wind directions.

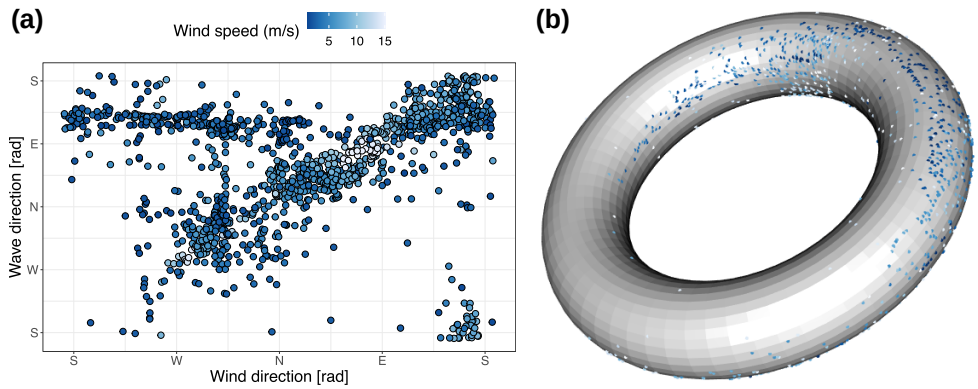

Figure 2. (a) Joint distribution of wave and wind directions, projected to a plane obtained by unwrapping a torus; (b) the same data displayed in their original manifold. Dots are coloured according to contemporaneous wind speed (m/s).

contemporaneous wind speeds, not only offers first-hand evidence of the presence of at least two clusters. It also shows that wind speed is not uniformly distributed across the plot, first-hand evidence of the influence of wind speed on the observed data.

By modelling these data by a mixture of toroidal densities with parameters driven by a nonhomogeneous semi-Markov chain, we aim to identify meaningful environmental regimes, simultaneously accomplishing two further goals. First, we estimate the distribution of the dwell time spent by the observed process in each regime. Regimes are therefore not only classified according to different shapes of the data distribution, as it is typically done in model-based segmentation, but also according to the distribution of the associated dwell times. Second, we estimate the influence of environmental time-varying covariates (e.g. wind speed) on the dynamics of the process.

3 A toroidal hidden semi-Markov model

Let $\boldsymbol{y} = (\boldsymbol{y}_t, t = 1, \ldots T)$ be a bivariate time series, where $\boldsymbol{y}_t = (y_{t1}, y_{t2})$ is a vector of two circular observations $-\pi < y_{t1}, y_{t2} \leq \pi$. Further, let $\boldsymbol{u} = (\boldsymbol{u}_t, t = 1, \ldots T)$ be a sequence of latent multinomial random variables $\boldsymbol{u}_t = (u_{t1} \ldots u_{tK})$ with one trial and K classes (or regimes), whose binary components represent class membership at time t. Multinomial processes in discrete time are often simply referred to as chains, and we follow this terminology. Our proposal is a hierarchical model where the joint distribution of the time series is obtained by integrating a parametric conditional

distribution $f(\boldsymbol{y} \mid \boldsymbol{u}; \boldsymbol{\theta})$ of the observed data given the latent classes (the observation process) with respect to a parametric distribution $p(\boldsymbol{u}; \boldsymbol{\eta})$ of the latent chain, namely

$$f(\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\eta}) = \sum_{\boldsymbol{u}} f(\boldsymbol{y} \mid \boldsymbol{u}; \boldsymbol{\theta}) p(\boldsymbol{u}; \boldsymbol{\eta}). \tag{1}$$

3.1 The observation process

We assume that the toroidal observations are conditionally independent given the latent classes. Under this setting, the observation process is driven by a family of K toroidal densities $f_k(\boldsymbol{y}) = f(\boldsymbol{y}; \boldsymbol{\theta}_k)$, which represent the conditional distribution of the data under each regime and are known up to an array $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ of parameters, namely

$$f(\boldsymbol{y} \mid \boldsymbol{u}; \boldsymbol{\theta}) = \prod_{t=1}^{T} \prod_{k=1}^{K} f(\boldsymbol{y}_t; \boldsymbol{\theta}_k)^{u_{tk}}. \tag{2}$$

As mentioned in Section 1, the literature offers a variety of toroidal densities that can be integrated in our model. The choice of a suitable parametric family should be a compromise between flexibility, ease of interpretation and numerical tractability. We rely on the particularly attractive bivariate wrapped Cauchy density (Kato & Pewsey, 2015) that optimally satisfies these three requirements. This density depends on a vector of five parameters, $\boldsymbol{\theta} = (\mu_1, \mu_2, \kappa_1, \kappa_2, \rho)$, namely two means, $-\pi \leq \mu_1, \mu_2 \leq \pi$, two concentrations $0 \leq \kappa_1, \kappa_2 < 1$ and a correlation parameter $-1 < \rho < 1$ and it can be written in terms of six coefficients $c, c_0 \ldots c_4$ that only depend on the three parameters κ_1, κ_2, ρ (Kato & Pewsey, 2015), namely

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = \frac{c}{c_0 - c_1 \cos(y_1 - \mu_1) - c_2 \cos(y_2 - \mu_2) - c_3 \cos(y_1 - \mu_1)\cos(y_2 - \mu_2) - c_4 \sin(y_1 - \mu_1)\sin(y_2 - \mu_2)}. \tag{3}$$

The density shares several properties with the bivariate normal density: it is unimodal and pointwise symmetric with respect to (μ_1, μ_2); κ_1 (κ_2) controls the dispersion of the marginal distribution of y_1 (y_2), which tends to a point mass at μ_1 (μ_2) when κ_1 (κ_2) approaches 1 and to a uniform distribution on the circle when κ_1 (κ_2) approaches 0; positive (negative) values of ρ correspond to positive (negative) correlation between the two angles, with the density for $-\rho$ being the reflection of the density associated with ρ and with $\rho = 0$ corresponding to independence.

It furthermore depends on a normalizing constant that is available in closed form (a property not necessarily shared by other toroidal densities) and it is also closed under marginalization and conditioning (the two marginal and conditional densities are univariate wrapped Cauchy). This property not only facilitates the evaluation of regression lines of one circular variable over the other one, but it also makes simulation straightforward. A random toroidal observation can be obtained by drawing the first coordinate from the marginal distribution and then using this value for sampling the second coordinate from the univariate conditional distribution given the first coordinate. Since both the marginal and the conditional distributions are univariate wrapped Cauchy, the whole task can be undertaken by calls to a single simulation wrapper of this distribution such as the function `rwrappedcauchy` in the R package `circular` (Lund et al., 2017). Figure 3 shows the shape taken by this density under specific parameter values and the related regression functions.

3.2 Latent chains

By the chain rule, the joint distribution of the latent chain $\boldsymbol{u} = (\boldsymbol{u}_t, t = 1, \ldots T)$ can be written as

$$p(\boldsymbol{u}) = p(\boldsymbol{u}_1) \prod_{t=2}^{T} p(\boldsymbol{u}_t \mid \boldsymbol{u}_{t-1}, \ldots, \boldsymbol{u}_1),$$

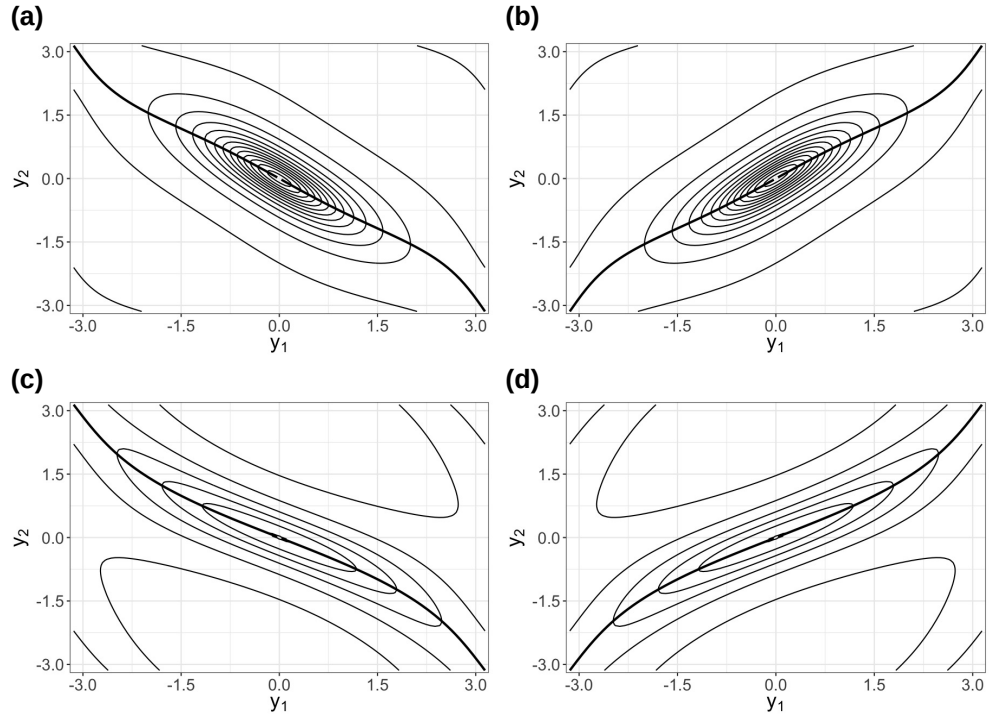

Figure 3. Contour plots of bivariate Wrapped Cauchy densities, centred at $\mu_1 = \mu_2 = 0$ and obtained by varying the concentrations κ_1, κ_2 and the correlation parameter ρ. Contours are projected on the plane obtained by unwrapping a torus. Curves (in bold) indicate the circular regression functions $\mathbb{E}(Y_2 \mid y_1)$. (a) $\kappa_1 = \kappa_2 = 0.3$, $\rho = -0.4$, (b) $\kappa_1 = \kappa_2 = 0.3$, $\rho = 0.4$, (c) $\kappa_1 = 0.01$, $\kappa_2 = 0.2$, $\rho = -0.4$, and (d) $\kappa_1 = 0.01$, $\kappa_2 = 0.2$, $\rho = 0.4$.

where $p(\boldsymbol{u}_1)$ is the initial distribution of the chain and $p(\boldsymbol{u}_t \mid \boldsymbol{u}_{t-1}, \ldots, \boldsymbol{u}_1)$ is the univariate conditional distribution of the chain at time t, given the past. The chain can be modelled in several ways, depending on specific assumptions on these conditional distributions. We focus below on two models, namely a nonhomogeneous Markov chain and a homogeneous semi-Markov chain because our proposal can be viewed as a combination of these two specifications.

Under a Markov hypothesis, the state of the process at time t is assumed conditionally independent of the past given the state at time $t-1$, namely

$$p(\boldsymbol{u}_t \mid \boldsymbol{u}_{t-1}, \ldots, \boldsymbol{u}_1) = p(\boldsymbol{u}_t \mid \boldsymbol{u}_{t-1}), \quad t = 2, \ldots, T \tag{4}$$

and the process is referred to as a Markov chain. If the conditional distribution $p(\boldsymbol{u}_t \mid \boldsymbol{u}_{t-1})$ is time-constant, then the Markov chain is referred to as homogeneous. A homogeneous Markov chain moves from state $\boldsymbol{u}_{t-1}$ to state $\boldsymbol{u}_t$ according to $K \times K$ time-constant transition probabilities $\gamma_{kh} = p(u_{th} = 1 \mid u_{t-1,k} = 1)$. Such probabilities can be conveniently reparametrized in terms of the (discrete) hazard of a state switch for a chain that is in state k at time $t-1$, say $q_k = \sum_{h \neq k} \gamma_{kh}$, and the conditional probability $\omega_{kh} = \gamma_{kh}/q_k$ of switching to a specific state h, given a switch, $\sum_{h \neq k} \omega_{kh} = 1$, namely

$$\gamma_{kh} = \begin{cases} 1 - q_k & h = k \\ q_k \omega_{kh} & h \neq k. \end{cases}$$

Under this reparametrization, the event 'leaving a state' is kept separated by the conditional event of reaching a specific state given the occurrence of a transition, similar to what is often done in

survival analysis when 'all-cause mortality' is separated by the cause of death. Under this setting, each conditional distribution is a multinomial distribution

$$p(\boldsymbol{u}_t \mid \boldsymbol{u}_{t-1}) = \prod_{k=1}^{K}\prod_{h=1}^{K} \gamma_{kh}^{u_{tkh}}, \tag{5}$$

where $u_{tkh} = u_{t-1,k}u_{th}$ indicates the transition event ($u_{tkh} = 1$ if the chain switches from state k to state h, and 0 otherwise). The joint distribution of the chain takes therefore the form

$$p(\boldsymbol{u}) = p(\boldsymbol{u}_1)\prod_{t=2}^{T}\prod_{k=1}^{K}\prod_{h=1}^{K} \gamma_{kh}^{u_{tkh}}. \tag{6}$$

The homogeneity assumption can be relaxed by introducing a $(T-1)\times Q$ design matrix $\boldsymbol{X}$, where each row $\boldsymbol{x}_t^T$, $t = 2 \dots T$, is associated to the transition from $t-1$ to t. Specifically, a nonhomogeneous Markov chain is obtained by allowing transition probabilities to depend on the row profiles $\boldsymbol{x}_t^T$, say

$$\text{(Markov chain)} \quad \gamma_{tkh} = \gamma_{kh}(\boldsymbol{x}_t^T) = \begin{cases} 1 - q_k(\boldsymbol{x}_t^T) & h = k \\ q_k(\boldsymbol{x}_t^T)\omega_{kh} & h \neq k. \end{cases} \tag{7}$$

Here, the relationship between the hazards and the covariates can be specified in terms of Q regression coefficients by a generalized linear model with link function g, say

$$g(q_k(\boldsymbol{x}_t^T)) = \boldsymbol{x}_t^T\boldsymbol{\beta}_k.$$

When the design matrix reduces to a column of 1's, then $g(q_{tk}) = \beta_k$ and a homogeneous Markov chain is obtained as a particular case. Otherwise, the chain is driven by hazards that are exogenously updated at each time t by the current profile $\boldsymbol{x}_t^T$.

Under a semi-Markov hypothesis, instead, the state of the chain at time t depends on the history of the process since the last transition, summarized by the dwell time d_t, that is the amount of time the chain has been dwelling in the same state by time t. Precisely, a semi-Markov chain moves from state $\boldsymbol{u}_{t-1}$ to state $\boldsymbol{u}_t$ according to time-varying transition probabilities

$$\gamma_{tkh} = \gamma_{kh}(d_t) = \begin{cases} 1 - q_k(d_t) & h = k \\ q_k(d_t)\omega_{kh} & h \neq k. \end{cases} \tag{8}$$

Under this setting, the chain is driven by hazards that at every time t are endogenously updated by the current dwell time d_t. As a result, the Markov property does not hold and the chain's univariate conditional distributions depend on the process's history since the last transition. Precisely, let t_1 and t_2 be the times of two consecutive transitions and let k and h be the states of the chain at times t_1 and t_2: for each time $t_1 < t \leq t_2$,

$$p(\boldsymbol{u}_t \mid \boldsymbol{u}_{t-1}, \dots, \boldsymbol{u}_1) = p(\boldsymbol{u}_t \mid \boldsymbol{u}_{t-1}, \dots, \boldsymbol{u}_{t_1}) = \begin{cases} 1 - q_k(t - t_1) & t \neq t_2 \\ q_k(t_2)\omega_{kh} & t = t_2, \end{cases} \tag{9}$$

where $t - t_1 = d_t$ is the current dwell time at time t. As above, a generalized linear model can be exploited to specify the relationship between the hazard and the dwell time, say

$$g(q_k(d_t)) = \beta_{0k} + \beta_{1k}d_t \quad \beta_{1k} \geq 0, \quad k = 1, \dots, K.$$

When $\beta_{1k} = 0$, then $g(q_{tk}) = \beta_{0k}$, and the semi-Markov chain reduces to a homogeneous Markov chain since in this case knowledge of the current state k suffices to predict the next move of the

chain. Otherwise, if $\beta_{1k} > 0$, such knowledge must be augmented with the information on the current dwell time.

Since $q_k(d)$ is a probability, any link function that maps the unit interval into the real line is admissible. Typical choices for g are the logit, the probit and the complementary log–log transformation. The complementary log–log is a natural choice for modelling discrete dwell times because it can be viewed as the discrete counterpart of an exponential hazard in continuous time. Specifically, if $\lambda_k(t) = \exp(\beta_{0k} + \beta_{1k}t)$ is the hazard function of a positive, continuous random variable T_k that indicates the time up to a transition from state k, then the survival function of T_k is given by

$$S_k(t) = P(T_k > t) = \exp\left(-\int_0^t \lambda_k(\tau)d\tau\right).$$

As a result, the discrete hazard is given by

$$\begin{aligned} q_k(d) &= 1 - \frac{S(d+1)}{S(d)} = 1 - \exp\left(-\int_d^{d+1} \lambda_k(t)dt\right) \\ &\approx 1 - \exp(-\exp(\beta_{0k} + \beta_{1k}(d + 0.5))), \end{aligned}$$

and therefore

$$\log(-\log(1 - (q_k(d_t)))) = \beta_{0k} + \beta_{1k}(d_t + 0.5) \quad \beta_{1k} \geq 0, \quad k = 1, \ldots, K.$$

Figure 4 displays the role played by β_{1k} on the shape of the dwell time hazard and the associated dwell time distribution

$$p_k(d) = \begin{cases} q_k(1) & d = 1 \\ q_k(d)\prod_{\delta=1}^{d-1}(1 - q_k(\delta)) & d > 1. \end{cases}$$

When $\beta_{1k} = 0$, the hazard is constant and the dwell time distribution is geometric. As a result, the most probable dwell time is $d = 1$. Increasing values of β_{1k} are associated with hazards that increase faster and more compressed dwell time distributions with lower right tails.

Two remarks are in order here. First, similarly to a nonhomogeneous Markov chain, a semi-Markov chain is defined by time-varying transition probabilities and it is hence intrinsically nonhomogeneous. However, while in a Markov chain these probabilities are exogenously updated by current covariate information, in a semi-Markov chain they are endogenously updated by the current dwell time. Second, similarly to the case of a homogeneous Markov chain, the joint

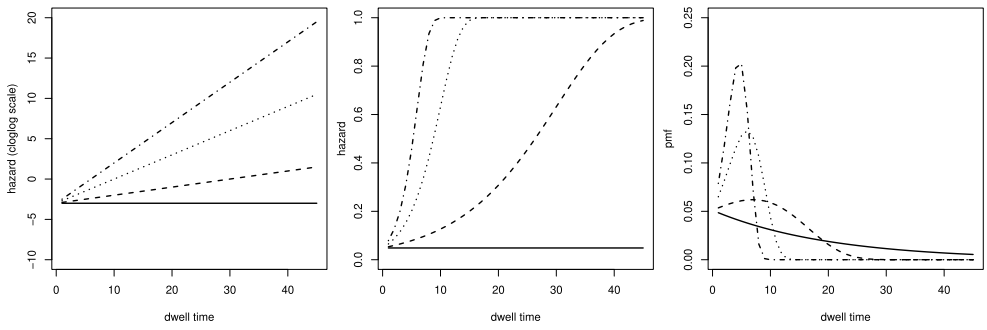

Figure 4. Baseline dwell time hazard curves on the complementary log–log scale (left) for $\beta_0 = -3$ and four values of β_1 (0=continuous, 0.1= dashed, 0.3=dotted, 0.5=dot-dashed) and the corresponding hazard curves (middle) and probability mass functions (right). Although dwell times are discrete, both hazards and distributions are displayed as continuous, for better visualization.

distributions of both a nonhomogeneous Markov chain and a semi-Markov chain take the product form

$$p(\boldsymbol{u}) = p(\boldsymbol{u}_1) \prod_{t=2}^{T} \prod_{k=1}^{K} \prod_{h=1}^{K} \gamma_{tkh}^{u_{tkh}}, \tag{10}$$

where the transition probabilities γ_{tkh} are given by (7) and (8), respectively. However, in the case of a Markov chain, (10) is the result of the Markov hypothesis and represents the product $p(\boldsymbol{u}_1) \prod_{t=2}^{T} p(\boldsymbol{u}_t \mid \boldsymbol{u}_{t-1})$. Differently, in the case of a semi-Markov chain, (10) results from the reparametrization of the transition probabilities in terms of hazards, and it represents the product $p(\boldsymbol{u}_1) \prod_{t=2}^{T} p(\boldsymbol{u}_t \mid \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{t-1})$. This representation of the joint distribution of a semi-Markov chain by means of transition probabilities that are augmented with dwell time information has long been known in probability (Anselone, 1960), but it has only recently been brought to the attention of the statistical community by Langrock and Zucchini (2011).

Our proposal combines the two specifications above in a single model. Specifically, we assume that the chain moves from state $\boldsymbol{u}_{t-1}$ to state $\boldsymbol{u}_t$ according to transition probabilities

$$\text{(semi} - \text{Markov chain)} \quad \gamma_{tkh} = \gamma_{kh}(d_t, \boldsymbol{x}_t^T) = \begin{cases} 1 - q_k(d_t, \boldsymbol{x}_t^T) & h = k \\ q_k(d_t, \boldsymbol{x}_t^T)\omega_{kh} & h \neq k, \end{cases} \tag{11}$$

where the hazard is specified in terms of the generalized linear model

$$g(q_k(d_t, \boldsymbol{x}_t^T)) = \beta_{0k} + \beta_{1k}(d_t + 0.5) + \boldsymbol{x}_t^T \boldsymbol{\beta}_k \quad \beta_{1k} \geq 0, \quad k = 1, \ldots, K. \tag{12}$$

These state-specific regression coefficients regulate the dynamics of the chain conditionally on the current state. If $\beta_{1k} = 0$, the state k is Markovian, that is, the next move of the chain can be predicted based on the current covariate information and the model reduces to a memory-less process (Ren & Barnett, 2023). Otherwise, the state is semi-Markovian, and predictions depend also on the current dwell time. By choosing a complementary log–log link function g, model (12) can be seen as the discrete-time counterpart of a continuous-time proportional hazards model $\lambda_k(t; \boldsymbol{x}_t) = \exp(\beta_{0k} + \beta_{1k} t + \boldsymbol{x}_t^T \boldsymbol{\beta}_k))$ (see, for instance, Kalbfleisch & Prentice, 1980, page 37; and Tutz & Schmid, 2016, Section 3.3).

All the above specifications restrict the conditional transition probabilities ω_{kh} to be time-constant. When time-varying covariates are available, however, this limitation can be overcome by an additional generalized linear model where time-varying conditional transition probabilities are linked to the covariates by a link function g' and an array of regression coefficients $\boldsymbol{\alpha}_{kh}$, say

$$g'(\omega_{tkh}) = \boldsymbol{z}_t^T \boldsymbol{\alpha}_{kh},$$

where $\boldsymbol{z}_t^T$ is the tth row of a $(T-1) \times Q'$ design matrix $\boldsymbol{Z}$, not necessarily equal to the design matrix $\boldsymbol{X}$ that modulates the hazards. The particular case of time-constant conditional probabilities is obtained when $\boldsymbol{z}_t^T = 1$, and, as a result, $g'(\omega_{tkh}) = \alpha_{kh}$. A natural choice of the link function g' is the multinomial logit, say

$$\log\left(\frac{\omega_{kh}}{\omega_{kh^\star}}\right) = \boldsymbol{z}_t^T \boldsymbol{\alpha}_{kh}$$

where $h^\star \neq k$ is a reference state. Under this setting, the Markov chain model (7) can be extended to the specification

$$\text{(extended Markov chain)} \quad \gamma_{tkh} = \gamma_{kh}(\boldsymbol{x}_t^T, \boldsymbol{z}_t^T) = \begin{cases} 1 - q_k(\boldsymbol{x}_t^T) & h = k \\ q_k(\boldsymbol{x}_t^T)\omega_{kh}(\boldsymbol{z}_t^T) & h \neq k, \end{cases} \tag{13}$$

while the semi-Markov chain model (11) can be extended to the specification

$$\text{(extended semi – Markov chain)} \quad \gamma_{tkh} = \gamma_{kh}(d_t, \boldsymbol{x}_t^T, \boldsymbol{z}_t^T) = \begin{cases} 1 - q_k(d_t, \boldsymbol{x}_t^T) & h = k \\ q_k(d_t, \boldsymbol{x}_t^T)\omega_{kh}(\boldsymbol{z}_t^T) & h \neq k. \end{cases} \tag{14}$$

In summary, models (7), (13), (11) and (14) represent four different ways to integrate time-varying covariates in the latent process distribution. Specifically, when combined with the observation process (2), the Markov chain (7) leads to an HMM with covariate-specific hazards and homogeneous transition probabilities, while the extended Markov chain (13) relaxes this homogeneity constrain, leading to an extended HMM where both the hazards and the conditional transitional probabilities depend on covariates. When instead the observation process (11) is combined with the semi-Markov chain model (2), we obtain an HSMM with covariate-specific hazards and homogeneous transition probabilities, while the extended semi-Markov (14) extends this specification to an extended HSMM where both the hazards and the conditional transitional probabilities depend on covariates.

4 Data augmentation and maximum likelihood estimation

All the models described above are particular cases of model (14), which depends on the vector $\boldsymbol{\theta}$ of the parameters of the state-specific toroidal distributions and on a parameter vector $\boldsymbol{\eta} = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\alpha})$, where $\boldsymbol{\pi} = (\pi_1, \ldots \pi_K)$ is the vector of the initial probabilities of the chain, $\boldsymbol{\beta}$ is the vector of the state-specific regression coefficients of the hazards and, finally, $\boldsymbol{\alpha}$ is the vector of the state-specific regression coefficients of the conditional transition probabilities. The likelihood function is obtained by integrating (2) with (10), namely

$$L(\boldsymbol{\theta}, \boldsymbol{\eta}) = \sum_{\boldsymbol{u}} \prod_{t=1}^{T} \prod_{k=1}^{K} f(\boldsymbol{y}_t; \boldsymbol{\theta}_k)^{u_{tk}} p(\boldsymbol{u}_1; \boldsymbol{\pi}) \prod_{t=2}^{T} \prod_{k=1}^{K} \prod_{h=1}^{K} \gamma_{tkh}(\boldsymbol{\alpha}, \boldsymbol{\beta})^{u_{tkh}}. \tag{15}$$

The maximization of (15) is complicated by the summation over all the possible paths of the latent semi-Markov chain. This is a typical difficulty in the literature of hierarchical models and it is often overcome by an EM algorithm that alternates the maximization of a weighted complete-data log-likelihood function (M step) with weights updates (E step). The complete-data log-likelihood function is defined by appropriately augmenting the missing information by pseudo-observations. The efficiency of the EM algorithm depends on the augmentation design because only an appropriate definition of the pseudo-observations facilitates the execution of both the E step and the M step of the algorithm.

Under (14), the missing information at each time t is not only the state of the chain, but also the current dwell time, and it is conveniently described by an array of multinomial pseudo-observations $\boldsymbol{u}_t = (u_{tkd}, k = 1, \ldots, K, d = 1, \ldots T - 1)$, where $u_{tkd} = 1$ if at time t the chain is in state k and the current dwell time is d, and 0 otherwise. Here, the dwell time runs up to $T - 1$, which is the maximum possible dwell time in a time series of length T, by assuming a state switch at time $t = 1$. Under this setting, the complete-data log-likelihood function is given by

$$\begin{aligned} \log L_c(\boldsymbol{\theta}, \boldsymbol{\eta}) = & \sum_{k=1}^{K} u_{1k1} \log \pi_k + \sum_{t=2}^{T} \sum_{k=1}^{K} \sum_{h=1}^{K} \sum_{d,d'=1}^{1,\ldots,T-1} u_{t-1,kd} u_{thd'} \log \gamma_{kh}(d, \boldsymbol{x}_t^T, \boldsymbol{z}_t^T; \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ & + \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{d=1}^{T-1} u_{tkd} \log f(\boldsymbol{y}_t; \boldsymbol{\theta}_k). \end{aligned} \tag{16}$$

It is worth remarking that most of the products $u_{t-1,kd} u_{thd'}$, indicating the transition event from $t - 1$ to t, are equal to zero. As an example, if the chain is in state k at time $t - 1$ and the current dwell time is d, then only K events are allowed: one of the possible $K - 1$ transitions to state $h \neq k$, which

reset the current dwell time at $d' = 1$; or no transition, thus updating the current dwell time to $d' = d + 1$. As a result, if $u_{t-1,kd} = 1$, then $u_{tkd'} = 1$ for $h = k$ and $d' = 1$ or for $h \neq k$ and $d' = d + 1$, and 0 otherwise. Though alleviated by such sparsity, the dimension of the pseudo-observations arrays depends on the size of the time series, and potential issues of storage memory may arise. A possible solution relies on truncating these arrays at a value $m < T - 1$ (Langrock et al., 2015), leading to the approximated complete-data log-likelihood function

$$\begin{aligned}\log L_c^{(m)}(\boldsymbol{\theta}, \boldsymbol{\eta}) = &\sum_{k=1}^{K} u_{1k1} \log \pi_k + \sum_{t=2}^{T} \sum_{k=1}^{K} \sum_{h=1}^{K} \sum_{d,d'=1}^{1,\ldots,m} u_{t-1,kd} u_{thd'} \log \gamma_{kh}(d, \boldsymbol{x}_t^T, z_t^T; \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &+ \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{d=1}^{m} u_{tkd} \log f(\boldsymbol{y}_t; \boldsymbol{\theta}_k).\end{aligned} \tag{17}$$

This method does not constrain the maximum dwell time to be m but, for $d > m$, the transition probabilities of the semi-Markov chain are approximated by those of a nonhomogeneous Markov chain, namely

$$\gamma_{tkh} = \gamma_{kh}(d, \boldsymbol{x}_t^T, z_t^T) = \begin{cases} \gamma_{kh}(d, \boldsymbol{x}_t^T, z_t^T) & d \leq m \\ \gamma_{kh}(m, \boldsymbol{x}_t^T, z_t^T) & d > m. \end{cases} \tag{18}$$

Note that, setting $m = 1$ reduces equation (17) to the complete-data log-likelihood of a hidden Markov model. Suitable tuning of m therefore allows to use (17) to estimate all the models in Section 3. Specifically, the proposed EM algorithm iteratively relies on the approximated complete-data log-likelihood (17). Given the parameter estimates, say $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\eta}}$, obtained at the end of the sth EM iteration, the $(s + 1)$th iteration is initialized by the E-step, which evaluates the expected value of the complete-data log-likelihood (17) with respect to the conditional distribution of the pseudo-observations u_{tkd} and their pairwise products $u_{t-1,kd} u_{thd'}$ given the observed data. The E step boils down to the computation of the univariate posterior probabilities $\hat{\pi}_{tkd} = P(u_{tkd} = 1 \mid \boldsymbol{y}, \boldsymbol{x}, z, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}})$, and the bivariate posterior probabilities $\hat{\pi}_{tkhdd'} = P(u_{t-1,kd} = 1, u_{thd'} = 1 \mid \boldsymbol{y}, \boldsymbol{x}, z, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}})$. The task of computing these posterior probabilities is generally referred to as the 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 semi-Markov chain and numerical under- and over-flows. Due to the product form (10) of the joint distribution of a semi-Markov chain, the smoothing problem can be efficiently solved by recycling the popular forward–backward algorithms that are exploited in the HMMs context (see Cappé et al., 2005 for an excellent review).

The M-step of the algorithm updates the parameter estimates, by maximizing the expected value of the complete-data log-likelihood (17), obtained from the previous E step. This expected value is the sum of functions that depend on independent sets of parameters and can therefore be maximized separately, namely

$$\begin{aligned}Q(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = &\underbrace{\sum_{k=1}^{K} \hat{\pi}_{1k} \log \pi_k + \sum_{t=2}^{T} \sum_{k=1}^{K} \sum_{h=1}^{K} \sum_{d,d'=1}^{m} \hat{\pi}_{tkhdd'} \log \gamma_{kh}(d, \boldsymbol{x}_t^T, z_t^T; \boldsymbol{\alpha}, \boldsymbol{\beta})}_{Q(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta})} \\ &+ \underbrace{\sum_{t=1}^{T} \sum_{k=1}^{K} \hat{\pi}_{tk} \log f(\boldsymbol{y}_t; \boldsymbol{\theta}_k)}_{Q(\boldsymbol{\theta})},\end{aligned} \tag{19}$$

where $Q(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = Q(\boldsymbol{\pi}) + Q(\boldsymbol{\alpha}) + Q(\boldsymbol{\beta})$, and

$$Q(\boldsymbol{\pi}) = \sum_{k=1}^{K} \hat{\pi}_{1k} \log \pi_k,$$

$$Q(\boldsymbol{\alpha}) = \sum_{t=2}^{T} \sum_{k=1}^{K} \sum_{h \neq k} \sum_{d,d'=1}^{m} \hat{\pi}_{tkhdd'} \log \omega_{kh}(z_t^T; \boldsymbol{\alpha}),$$

$$Q(\boldsymbol{\beta}) = \sum_{t=2}^{T} \sum_{k=1}^{K} \sum_{h \neq k} \sum_{d,d'=1}^{m} \hat{\pi}_{tkhdd'} \log q_k(d, \boldsymbol{x}_t^T; \boldsymbol{\beta}) + \sum_{t=2}^{T} \sum_{k=1}^{K} \sum_{d,d'=1}^{m} \hat{\pi}_{tkkdd'} \log (1 - q_k(d, \boldsymbol{x}_t^T; \boldsymbol{\beta})).$$

Function $Q(\boldsymbol{\pi})$ is a weighted multinomial log-likelihood and it is maximized by $\hat{\pi}_k = \hat{\pi}_{1k}$. Functions $Q(\boldsymbol{\alpha})$ and $Q(\boldsymbol{\beta})$ are instead a weighted multinomial and a weighted binomial likelihood, respectively. Upon choosing an appropriate link function, as suggested in Section 3, they can be therefore maximized by fitting a weighted multinomial regression and a weighted binomial regression model, respectively. Finally, maximization of the function $Q(\boldsymbol{\theta})$ can be undertaken by an unconstrained maximization algorithm, after a suitable re-parametrization of the involved parameters. In the case of the bivariate-wrapped Cauchy, we maximize $Q(\boldsymbol{\theta})$ over the parameter vector $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$ where $\theta_1 = \tan(\mu_1)$, $\theta_2 = \tan(\mu_2)$, $\theta_3 = (\tanh(\kappa_1) + 1)/2$, $\theta_4 = (\tanh(\kappa_2) + 1)/2$ and, finally, $\theta_5 = \tanh(\rho)$, exploiting a quasi-Newton procedure as provided by the function `optim` in R.

5 Computational details, uncertainty quantification, and model selection

In this section, we describe the computational challenges posed by the implementation of the EM algorithm (see Section 4) and discuss the adopted inferential strategies for uncertainty quantification and model selection.

Likelihood maximization of the proposed models is not trivial as it is well known that the EM algorithm may converge to local maxima or singularities at the edge of the parameter space, where the log-likelihood is unbounded. As a result, several strategies have been proposed to select a local maximizer and detect a spurious maximizer (Maruotti & Punzo, 2021). We pursue a short-run strategy, by running the EM algorithm from multiple random initializations and stopping it without waiting for full convergence. Then, we select the best output across these short runs, i.e. the one maximizing the log-likelihood, and use this solution to initialize longer runs to reach convergence. Eventually, we stop the optimization when the increase of two successive log-likelihoods falls below 10^{-4}, as this stopping criterion produced stable parameter estimates in preliminary experiments.

While $Q(\boldsymbol{\alpha})$ can be maximized by running any generalized linear model estimation routine that allows for case weights, such as the glm function in `R`, the same approach is not fully desirable for maximizing $Q(\boldsymbol{\beta})$, because of the constrain $\beta_{1k} \geq 0$. The constrained optimization of $Q(\boldsymbol{\beta})$ can be however transformed in an unconstrained optimization problem, by maximizing with respect to the unconstrained working parameters η_{1k}, where $\beta_{1k} = \exp(\eta_{1k})$, $k = 1, \ldots, K$.

The EM algorithm has been implemented by using a single value of m for geometrically approximating the tail of the dwell time distribution of every state after $d > m$. In principle, the algorithm could be extended to allow state-specific thresholds, say m_k, that could shorten execution times. However, this extension is practical only when an 'optimal' list of thresholds is known *a priori*. When this is not the case, searching for the optimal threshold set, whose size increases with the number of latent states, requires running the whole algorithm several times, therefore dissipating the computational advantage of working with a single threshold.

The dimension of the array $\gamma_{tkh}(d, \boldsymbol{x}_t^T, z_t^T)$ increases with m and K and may lead to potential storage issues. However, such issues were not encountered in our scenarios as $m \leq 75$ and $K \leq 5$. Nevertheless, other applications may require larger values for both, which can be handled by exploiting the sparsity of the array and using packages like `Matrix`, that support working with sparse objects.

Uncertainty quantification for the model parameters is based on a parametric bootstrap approach. Precisely, we re-fitted the model to $B = 1{,}000$ bootstrap samples, which were simulated from the MLEs, hence obtaining $\{\boldsymbol{\eta}\}_{b=1}^{B}$ and $\{\boldsymbol{\theta}\}_{b=1}^{B}$ sets of estimates, exploited to compute equal-tail confidence intervals. Within the proposed framework, this inferential approach proves convenient for two main reasons: first, simulation from the described HSMM (and its particular cases) is straightforward; second, it overcomes the computation of the observed information matrix, a nontrivial task. Methods to obtain the observed information matrix have only been proposed in the literature for the HMMs (Bartolucci & Farcomeni, 2015). They rely on the computation of the derivatives of each element of the forward recursion algorithm—used to obtain univariate and bivariate posterior probabilities—but suffer from numerical instability and require suitable normalizations, especially with large sample sizes (as it is our case). Eventually, to the best of the authors' knowledge, these methods have not yet been extended and tested to the HSMM framework. As an aside, we note that symmetric confidence intervals are not guaranteed to be within the admissible parameter space when an estimate is close to the space boundary. In an HSMM setting, this issue typically arises for the (conditional) transition probabilities, which may often be estimated to be close to either 0 or 1, making symmetric confidence intervals unreliable (Section 3.6.1 in Zucchini et al., 2016). Equal-tail, bootstrap-based confidence intervals do not suffer from this drawback.

Finally, model selection reduces to find the optimal number K of components, according to a suitable criterion. Like any other mixture model, our proposal can be either exploited for density estimation to get a good approximation of the data distribution or for clustering, where the goal is to find the *best* data segmentation (Fruhwirth-Schnatter et al., 2018). In our case study, segmentation is as important as density estimation and, therefore, we rely on the minimization of the integrated complete likelihood (ICL; Biernacki et al., 2000), where the goodness of fit is penalized by classification entropy.

6 Simulation study

We run a simulation experiment to assess the ability of an HSMM with covariate-specific hazards and homogeneous transition probabilities to recover the population parameter values under different scenarios, as well as to assign each observation a probability of coming from one of the latent regimes, and eventually classify it. The simulation was performed under $K \in \{2, 3, 4\}$ regimes and included a time-varying covariate, drawn from $N(0, 9)$. Table 1 and Figure 5 respectively display the population values of the parameters chosen for simulating three scenarios, and the contour plots of the resulting component densities. These scenarios—though not exhaustive—aim to mimic possible real data situations where mixture components may (or may not) overlap and show positive (or negative) correlations, while also accounting for different effects of a time-varying covariate on the dwell-time distribution and not necessarily uniform transition probabilities.

For each K, we simulated $N = 250$ samples with increasing size $T \in \{1{,}000, 2{,}000, 3{,}000\}$ to evaluate the consistency of the estimates. For each simulated sample, we know the maximum dwell time m_{obs} and such knowledge provides us with the opportunity to perform a sensitivity analysis of the geometric approximation of the dwell time distribution. Accordingly, parameter estimation was repeated by maximizing the log-likelihood (17) with $m = \lceil m_{\text{obs}} \times \delta \rceil$, with $\delta \in \{0.5, 1, 1.5\}$. The performance of the proposed estimation algorithm was finally assessed in terms of (i) Adjusted Rand Index (ARI) for the classification of the latent regimes and (ii) Root Mean Squared Error (RMSE) for the estimation of the parameters of both the latent part of the model, $\boldsymbol{\eta}$, and the observed part of the model, $\boldsymbol{\theta}$. Specifically, the RMSE associated with the circular means was computed by using the angular deviation (Mardia & Jupp, 2000) of the estimates. The results are summarized by Figure 6. The ability to recover the true value of the parameters for the density and the latent process is instead shown in Tables A2, A1, and A3 of the Appendix.

In this study, the median Adjusted Rand Index (ARI) was always above 0.8 (Figure 6), reassuring about the capability of the algorithm to perform a satisfactory data segmentation, even when the sample size is small and the maximum dwell time m is misspecified. In particular, the rightmost panel of the figure shows that most ARI values are above 0.8 also when the model dimension is large and the algorithm is challenged by a huge number of parameters. Under this setting, however, we obtained a few cases of incorrect data segmentation, associated with a sub-optimal local

Table 1. Simulation: parameter values chosen for each scenario

		observation process					latent process						
							dwell time			transition probabilities			
no. states	state k	μ_{1k}	μ_{2k}	κ_{1k}	κ_{2k}	ρ_k	β_{0k}	β_{1k}	β_{2k}	ω_{k1}	ω_{k2}	ω_{k3}	ω_{k4}
2	1	0.5	0.5	0.2	0.3	0.6	−8	0.35	−0.5	0	1	–	–
	2	2	2	0.2	0.8	0.1	−3	0.075	0.5	1	0	–	–
3	1	0.5	0.5	0.2	0.3	0.6	−8	0.4	−0.5	0	0.50	0.50	–
	2	2	2	0.2	0.8	0.1	−5	0.15	0.2	0.90	0	0.10	–
	3	2	−2	0.5	0.5	−0.6	−3	0.05	0.7	0.45	0.55	0	–
4	1	0.5	0.5	0.2	0.3	0.6	−8	0.4	−0.5	0	0.25	0.25	0.50
	2	2	2	0.2	0.8	0.1	−6	0.3	0.2	0.70	0	0.20	0.10
	3	−2	−2	0.7	0.9	−0.3	−4	0.05	0.7	0.15	0.25	0	0.60
	4	2	−2	0.5	0.5	−0.6	−2	0.15	−0.1	0.30	0.20	0.50	0

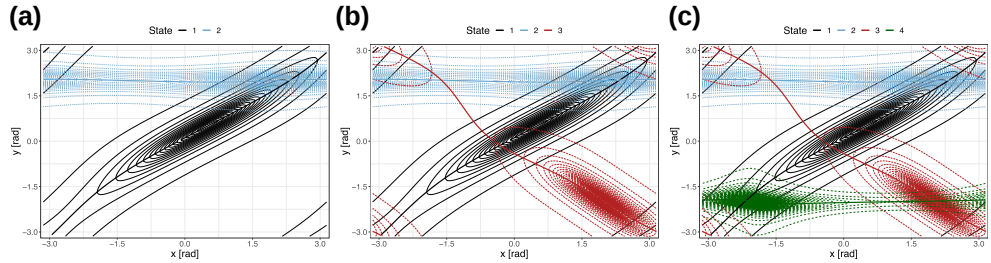

Figure 5. The three scenarios considered for the simulation study. (a) $K = 2$, (b) $K = 3$, and (c) $K = 4$.

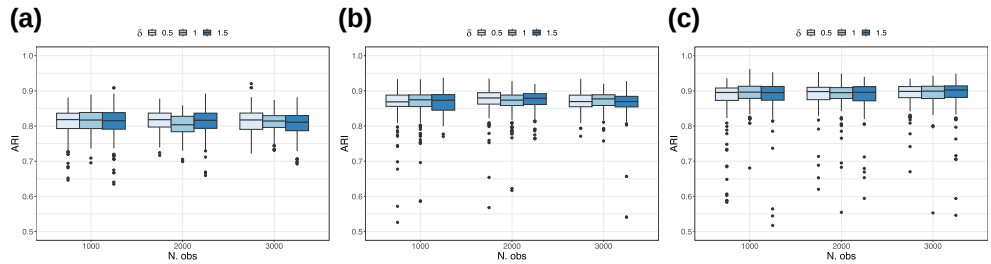

Figure 6. Simulation study: classification performances in terms of Adjusted Rand Index in each considered scenario. (a) $K = 2$, (b) $K = 3$, and (c) $K = 4$.

maximum of the log-likelihood. The chances of suboptimal solutions may increase with model complexity.

The RMSE values shown by Tables A2, A1, and A3 always decrease with sample size, indicating the consistency of the estimates obtained by the proposed algorithm. When the log-likelihood is misspecified by a value of m that is smaller than the true maximum dwell time, the RMSE is always larger, as expected. This difference is attenuated when instead m is overestimated, suggesting the somehow obvious strategy of choosing the largest m allowed by the available storage memory.

7 Segmenting marine conditions

It is intuitive that wind speed plays a role in the evolution of sea states across several regimes. However, a model that formalizes this role is far from being obvious and Section 3 describes four alternative ways of including wind speed as a driver of the process dynamics: an HMM and an HSMM with covariate-specific hazards and homogeneous transition probabilities and, respectively, an extended HMM and an extended HSMM with covariate-specific transition probabilities. Table 2 summarizes the results obtained by estimating these models across two to five states, reporting the BIC, the (negative) classification entropy and the ICL, obtained by summing BIC and negative entropy. Results are obtained by truncating the complete-data log-likelihood (17) at $m = 75$. In our application, the time resolution is 30 min and, therefore, setting $m = 75$ means geometrically approximating the dwell time distribution tail after 1.5 days. After several attempts, we noticed that in this application larger values of m yielded practically indistinguishable results.

As anticipated, we use ICL for model selection. Alternatively, the commonly used BIC might have been an option. In this case study, BIC tends to favour an extended HMM as the model that most parsimoniously offers the best fit. However, a good fit is often not necessarily associated with a good segmentation. As the number of the states increases, under an extended HMM, monotonically decreasing BIC values are counterbalanced by monotonically increasing entropy values, indicating that improvements in density estimation are obtained at the price of less separated latent classes. In this application, the accuracy of data segmentation, measured by the negative classification entropy, is as important as density estimation. As a result, we rely on ICL values, where BIC is further penalized by entropy. Under this criterion, the best candidate is a hidden semi-Markov model with four states. In terms of BIC, this model is better than an extended HMM with four states, but it is worse than an extended HMM with five states and an HSMM with five states. However, these two competitors perform worse in terms of entropy values and, as a result, the minimum ICL value is reached by an HSMM with four states, which seems to indicate a good compromise between goodness of fit, parsimony and latent class separation.

Therefore, Table 3 displays the estimates under an HSMM model with four states, along with bootstrap ET 95% confidence intervals, computed by simulating $B = 1{,}000$ samples, and here used to test whether estimates are significantly different from zero. The table displays three results that support the proposed model. First, the dependence parameter ρ is always significantly different from zero, supporting the choice of a toroidal distribution that accounts for circular correlation within latent regimes. The choice of using two univariate, conditionally independent circular densities, as it is often done, would have been a dispensable model restriction in this case study. Second, the regression coefficient β_1 of the dwell time is significant under three regimes (states 1, 2, and 4). This motivates the choice of a latent process where states are nonnecessarily Markovian. Third, the regression coefficient β_2 of wind speed is always significant, motivating wind speed as a relevant covariate in this study and indicating that a homogeneous version of our model would have been an unnecessary shortcoming.

The rest of the table can be interpreted with the help of Figures 7 and 8, which summarize the inferential output. Specifically, Figure 7a and c displays the data segmentation obtained in the variables space and, respectively, in the temporal domain, by associating each observation with the latent class k with the highest estimated posterior probability $\hat{\pi}_{tk}$. Figure 7b shows instead the four estimated toroidal densities in the unwrapped, Cartesian-like, toroidal space. Figure 7d shows the (conveniently smoothed) conditional distributions of the time-varying covariate (wind speed) within each regime, obtained by associating each observation with the most probable latent class. The regime-specific dwell-time hazards and distributions displayed by Figure 8 were obtained by setting wind speed at the three quartiles of these conditional distributions. The vertical line at $m = 75$ recalls that dwell time distributions are approximated by a geometric tail for times larger than 75.

Overall, the four estimated components appear well-separated and identify four distinct sea regimes or states. Three regimes (states 2, 3, and 4) are associated with positively correlated directions, clustered around three modal directions in the northwest, the northeast and the southeast quadrant. Specifically, while states 2 and 3 occur during episodes of northern Bora winds, state 4 is the result of Sirocco episodes. Both Bora and Sirocco are usually strong winds that drive

Table 2. Model dimension and goodness of fit for a varying number of latent states

Model	No. states	No. params.	BIC	Entropy	ICL
HMM	2	15	6,809.24	5,804.37	12,613.60
	3	26	6,056.05	3,874.66	9,930.71
	4	39	5,211.12	2,911.08	8,122.20
	5	54	4,934.61	2,334.55	7,269.16
Extended HMM	2	15	6,809.24	5,804.37	12,613.60
	3	29	6,019.02	3,869.57	9,888.59
	4	47	5,239.01	2,909.94	8,148.95
	5	69	4,872.46	2,335.00	7,207.47
HSMM	2	17	6,815.76	1,827.39	8,643.15
	3	29	6,056.23	1,716.90	7,773.13
	4	43	5,217.06	1,594.15	**6,811.21**
	5	59	5,052.75	1,870.05	6,922.79
Extended HSMM	2	17	6,815.76	1,827.39	8,643.15
	3	32	6,714.48	854.77	7,569.25
	4	51	6,060.42	900.76	6,961.18
	5	74	6,214.51	1,214.95	7,429.46

Note. HMM: a hidden Markov model with covariate-specific hazards and homogeneous transition probabilities. Extended HMM: an HMM with covariate-specific hazards and covariate-specific transition probabilities. HSMM: a hidden semi-Markov model with covariate-specific hazards and homogeneous transition probabilities. Extended HSMM: a hidden semi-Markov model with covariate-specific hazards and covariate-specific transition probabilities. The number in bold highlights the best model overall. ICL: integrated complete likelihood.

the direction of waves. Other winds in the Adriatic Sea blow at typically moderate speed and they are not able to change the direction of the southeasterly waves that travel towards the northwest along the major axis of the Adriatic basin. This regime is captured by the first component of the model, which clusters westerly winds blowing from the Italian coast that encounter waves that travel from southeast to northwest.

The temporal segmentation obtained in Figure 7c reflects the transition graph of the latent semi-Markov chain, as estimated by the conditional transition probabilities of Table 3. These estimates indicate that when the process is in state 2, it tends to switch to state 1. Indeed, under state 2, southeastern waves are generated by Sirocco episodes and keep such direction along the major basin of the Adriatic Sea as wind speed decreases (state 1). State 3 instead behaves as an intermediate state that is predicted to switch to states 2 and 4 (but rarely to state 1). State 3 features northeastern waves generated by a Bora episode that can change direction according to a rotation of wind toward the northwestern sector (state 4) or the southeastern sector (state 2). In general, by looking at the maximum value of the estimated conditional transition probabilities, the model seems to suggest $1 \to 4 \to 3 \to 2 \to 1$ as the most likely trajectory, associated with an anticyclonic atmospheric pattern.

The distinct advantage of an HSMM, compared to a more restrictive HMM, is that the temporal segmentation of the observed process (Figure 7c) can be interpreted not only in terms of transitions between latent regimes (as it is normally done with an HMM), but also in terms of duration of each regime. Figure 8a–d displays the four estimated, state-specific hazard functions, computed at three reference values of wind speed, namely the first quartile, the median and third quartile of the conditional distribution of the covariate given the most probable state, as estimated by the model. Such conditional computation avoids evaluations at covariate values that rarely occur under the state of interest. Figures 8e–8h display the resulting dwell time distributions, approximated by a geometric tail at dwell points greater than $m = 75$. The shape of the dwell time distributions under each state is driven by the coefficient β_{1k}, whose state-specific estimate is displayed by Table 3. As described in Section 3.2, large values of this parameter indicate departures from a geometric shape

Table 3. Parameters' estimates (bootstrap 95% confidence intervals) of the proposed 4-state HSMM

		State			
	Param.	**1**	**2**	**3**	**4**
Wind circular mean	μ_1	2.013 (1.921, 2.110)	2.414 (2.367, 2.464)	0.819 (0.752, 0.892)	2.196 (2.163, 2.230)
Wave circular mean	μ_2	1.985 (1.962, 2.011)	2.129 (2.093, 2.175)	0.804 (0.745, 0.854)	2.554 (2.491, 2.612)
Wind circular concentration	κ_1	0.532 (0.471, 0.593)	0.699 (0.665, 0.741)	0.688 (0.617, 0.747)	0.763 (0.734, 0.792)
Wave circular concentration	κ_2	0.861 (0.833, 0.882)	0.754 (0.723, 0.785)	0.769 (0.726, 0.809)	0.643 (0.609, 0.68)
Wind-wave circular correlation	ρ	−0.378 (−0.468, −0.287)	0.221 (0.148, 0.297)	0.287 (0.175, 0.389)	0.225 (0.155, 0.292)
Intercept	β_0	−4.577 (−6.826, −2.594)	−2.909 (−4.898, −0.77)	0.768 (−1.612, 4.5)	−0.385 (−2.783, 2.395)
Time	β_1	0.025 (0.001, 0.108)	0.102 (0.047, 0.191)	0.012 (−0.028, 0.118)	0.043 (0.006, 0.095)
Wind speed	β_2	0.313 (0.047, 0.749)	−0.66 (−1.435, −0.23)	−1.026 (−2.66, −0.4)	−1.214 (−2.28, −0.606)
Destination state					
Origin state	**Param**	**1**	**2**	**3**	**4**
1	ω_{1k}	0	0.205 (0.000, 0.450)	0.212 (0.000, 0.456)	0.583 (0.344, 0.810)
2	ω_{2k}	0.877 (0.642, 1.000)	0	0.108 (0.000, 0.355)	0.015 (0.000, 0.116)
3	ω_{3k}	0.009 (0.000, 0.131)	0.759 (0.408, 1.000)	0	0.232 (0.000, 0.550)
4	ω_{4k}	0.287 (0.000, 0.603)	0.270 (0.000, 0.569)	0.443 (0.122, 0.793)	0

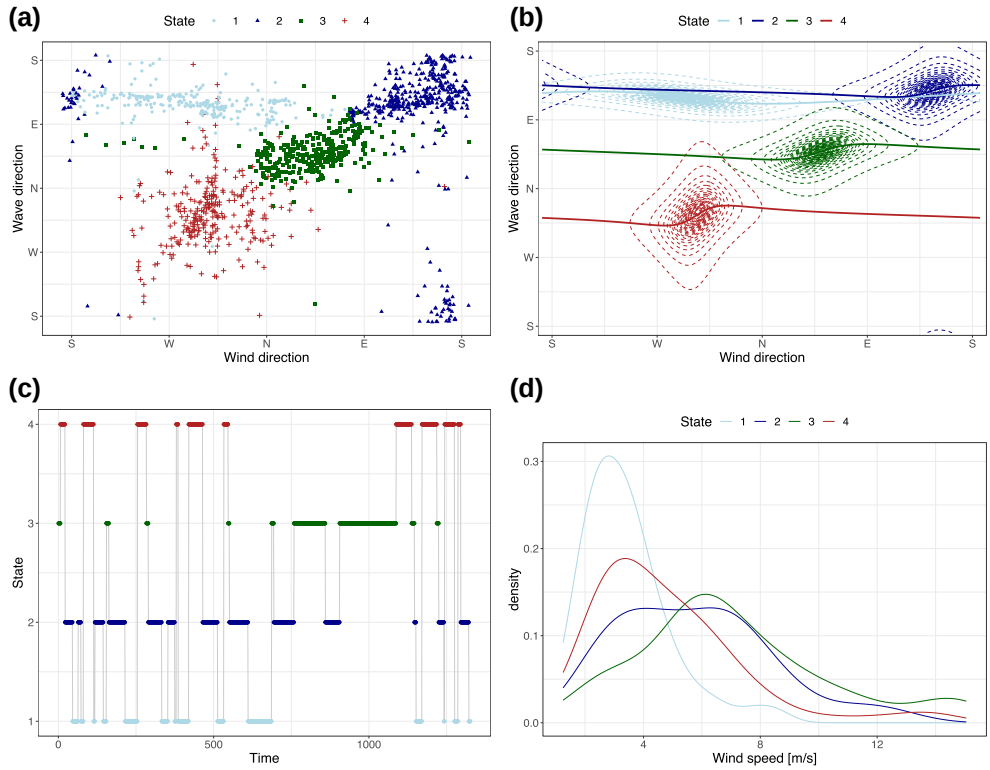

Figure 7. Top left: classification of the observed data (a); top right: estimated densities of the four mixture components with related circular regression lines (b); bottom left: estimated path of the hidden semi-Markov chain (c); the conditional distribution of wind speed within each latent regime (d).

in the dwell time distribution. Accordingly, the estimates under the intermediate regimes 1 and 3 (respectively $\hat{\beta}_1 = 0.025, 0.012$) are associated with dwell time distributions that are essentially geometric. Conversely, the estimates under states 2 and 4 (respectively $\hat{\beta}_1 = 0.102, 0.043$) reflect a significant departure from the geometric shape. According to these findings, the states associated with Bora and Sirocco episodes (2 and 4) tend to have longer durations than the states associated with intermediate regimes (1 and 3). More interestingly, wind speed is not always positively associated with regime duration. Under state 1, for example, $\hat{\beta}_2 = 0.313$, indicating that the hazard of a transition to another regime increases with wind speed. The conditional distribution of wind speed under this regime (Figure 7d) shows that this state is associated with moderate winds. An increase in wind speed expectedly indicates that a transition to another regime is approaching. All the other states (2, 3 and 4) are instead associated with stronger winds (Figure 7d) and an increase in wind speed tends to postpone the transition to another regime. Accordingly, the regression coefficients of wind speed under these states are negative. That association between wind speed and the duration of a specific sea regime may take different signs is well known in marine research (Holthuijsen, 2007), but difficult to measure without estimating a nonhomogeneous HSMM.

8 Discussion

Hierarchical models offer a useful strategy to interpret complex environmental phenomena, by allocating different data features to separate levels of a hierarchy, where several models can be efficiently integrated. Our proposal is a hierarchical model for toroidal time series that parsimoniously combines directional statistics and survival analysis tools. While directional statistics is exploited at the observation level of the hierarchy using toroidal densities, survival analysis is employed at the latent level of the hierarchy using hazard regression. The framework that makes such integration identifiable (and estimable) is provided by a class of hidden semi-Markov models

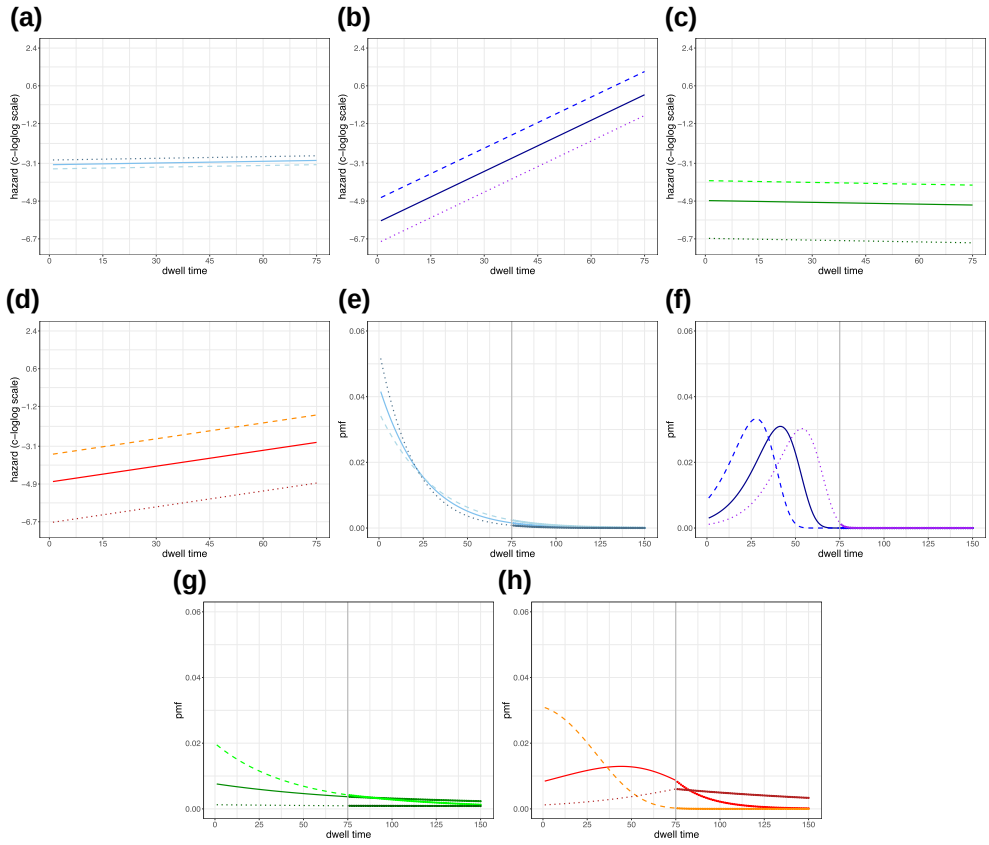

Figure 8. (a)–(d) State-specific hazards (in the complementary log-log scale) under three different scenarios obtained by fixing the covariate to its conditional first quartile (dashed), median (solid) and third quartile (dotted). (e)–(h) Corresponding dwell-time distribution in the three considered scenarios, where the points for $d > 75$ show the Geometric approximation. Although dwell times are discrete, both the hazards and the distributions are displayed as continuous for better visualization.

where the hazard of a transition is modelled separately from the conditional probability of reaching a specific state, given a transition. By working with the hazard function, we directly test whether a time-varying covariate is responsible for a regime shift, *given* the time spent in that regime. In other words, we are capable of identifying the conditions under which the chances of a regime shift are time-constant or time-varying.

In the considered case study, the model is capable of offering a clear-cut description of wind-wave interactions in terms of intuitively appealing environmental regimes. It provides a classification that reflects the orography of the study area, a feature often neglected by numerical models. It offers valid data segmentation by relaxing the restrictions that affect other models. It finally captures the influence of environmental conditions (wind speed) on the duration of specific regimes.

Tough tailored to issues that arise in marine studies, it can be easily extended to a wide range of real-world cases, where the interest is not only on the segmentation of the data according to time-varying latent classes, but also on the influence of covariates on sojourn times within each latent class.

The model is fully parametric and therefore exposed to misspecification issues. Non-parametric extensions can in principle be developed either at the observation level of the model hierarchy or at the latent level.

At the observation level, an option could be replacing the proposed mixture of toroidal densities with a nonparametric density. Methods of nonparametric estimation of toroidal densities are available (Di Marzio et al., 2011) and can in principle be integrated into an HSMM framework.

This approach could be pursued when the interest is on the accurate estimation of the marginal data density and appropriate resources for the additional computational burden are available. However, when the interest is on segmenting the data by means of physically meaningful parameters, such as in our case study, a fully parametric strategy seems preferable and easier to perform than a nonparametric approach.

At the latent level, a nonparametric approach can be pursued by assuming a nonparametric dwell time distribution and separately estimating the whole set of probabilities masses under each regime (Sansom & Thomson, 2001). This strategy is computationally tractable only when the study involves a long time series with short dwell times and it does not guarantee against wiggly dwell time distributions with implausible gaps and spikes (Bulla et al., 2010). Recently, Pohle et al. (2022) suggest penalized likelihood methods to estimate the dwell time distribution in a (homogeneous) HSMM. This idea could in principle be integrated into our proposal to estimate hazard functions nonparametrically, at the price of additional computational burden.

Acknowledgments

The authors thank Dr. Pierfrancesco Alaimo Di Loro for his insightful comments, which greatly enhanced the quality and rigour of this work.

Conflicts of interest: None declared.

Funding

This work has been supported by MIUR, grant number 2022XRHT8R—The SMILE project: Statistical Modelling and Inference for Living the Environment.

Data availability

The data underlying this article are available at https://github.com/minmar94/HSMM_covariates.

Appendix. Further results of the simulation study

Table A1. Results of the simulation study: RMSE for $\boldsymbol{\beta}$ in each considered scenario

	K	2			3			4		
	T	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$
β_{01}	1,000	2.031	1.336	1.332	2.353	1.682	1.776	4.262	1.952	3.847
	2,000	1.441	0.966	0.965	1.546	1.154	1.069	3.022	1.319	2.759
	3,000	1.182	0.820	0.819	1.049	0.812	0.806	3.107	1.040	2.219
β_{02}	1,000	0.626	0.543	0.539	2.144	0.853	0.926	3.010	1.738	2.174
	2,000	0.389	0.333	0.331	1.253	0.531	0.496	1.674	0.905	1.068
	3,000	0.312	0.273	0.275	0.984	0.399	0.389	1.511	0.687	0.814
β_{03}	1,000	–	–	–	1.718	1.627	1.682	1.294	0.925	1.288
	2,000	–	–	–	0.736	0.885	0.775	0.924	0.559	0.760
	3,000	–	–	–	0.552	0.499	0.543	0.758	0.418	0.627
β_{04}	1,000	–	–	–	–	–	–	1.204	0.577	0.925
	2,000	–	–	–	–	–	–	0.597	0.343	0.539
	3,000	–	–	–	–	–	–	0.549	0.288	0.466
β_{11}	1,000	0.235	0.082	0.082	0.235	0.127	0.146	0.556	0.157	0.355
	2,000	0.191	0.063	0.063	0.152	0.090	0.082	0.261	0.096	0.166
	3,000	0.163	0.048	0.048	0.110	0.062	0.059	0.208	0.084	0.128

(*continued*)

Table A1. Continued

	K	2			3			4		
	T	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$
β_{12}	1,000	0.058	0.056	0.055	0.215	0.051	0.051	0.441	0.145	0.302
	2,000	0.038	0.031	0.031	0.141	0.033	0.028	0.166	0.071	0.100
	3,000	0.031	0.026	0.026	0.104	0.022	0.023	0.134	0.053	0.134
β_{13}	1,000	–	–	–	0.199	0.299	0.183	0.137	0.077	0.111
	2,000	–	–	–	0.073	0.092	0.069	0.076	0.030	0.041
	3,000	–	–	–	0.056	0.054	0.053	0.045	0.025	0.036
β_{14}	1,000	–	–	–	–	–	–	0.179	0.150	0.181
	2,000	–	–	–	–	–	–	0.092	0.077	0.090
	3,000	–	–	–	–	–	–	0.085	0.063	0.089
β_{21}	1,000	0.133	0.117	0.118	0.161	0.135	0.175	0.330	0.219	0.431
	2,000	0.097	0.081	0.081	0.094	0.108	0.100	0.326	0.131	0.207
	3,000	0.099	0.066	0.066	0.079	0.080	0.079	0.153	0.091	0.115
β_{22}	1,000	0.133	0.122	0.122	0.105	0.103	0.117	0.179	0.146	0.161
	2,000	0.074	0.072	0.072	0.064	0.066	0.066	0.097	0.091	0.093
	3,000	0.061	0.059	0.059	0.056	0.058	0.057	0.075	0.068	0.073
β_{23}	1,000	–	–	–	0.647	0.643	0.521	0.230	0.182	0.232
	2,000	–	–	–	0.194	0.209	0.203	0.156	0.110	0.142
	3,000	–	–	–	0.140	0.135	0.140	0.139	0.082	0.117
β_{24}	1,000	–	–	–	–	–	–	0.235	0.106	0.186
	2,000	–	–	–	–	–	–	0.141	0.065	0.124
	3,000	–	–	–	–	–	–	0.178	0.049	0.105

Note. RMSE=root mean squared error.

Table A2. Results of the simulation study: RMSE for $\boldsymbol{\theta}$ in each considered scenario

	K	2			3			4		
	T	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$
μ_{11}	1,000	0.087	0.085	0.083	0.110	0.121	0.107	0.149	0.118	0.140
	2,000	0.062	0.062	0.063	0.070	0.076	0.074	0.089	0.087	0.086
	3,000	0.054	0.059	0.057	0.076	0.070	0.063	0.081	0.070	0.078
μ_{12}	1,000	0.146	0.162	0.148	0.166	0.189	0.189	0.311	0.205	0.253
	2,000	0.135	0.133	0.133	0.104	0.110	0.105	0.203	0.146	0.145
	3,000	0.098	0.098	0.099	0.089	0.083	0.088	0.159	0.116	0.161
μ_{13}	1,000	–	–	–	0.097	0.097	0.107	0.397	0.028	0.383
	2,000	–	–	–	0.068	0.063	0.068	0.314	0.022	0.247
	3,000	–	–	–	0.055	0.055	0.056	0.309	0.018	0.246
μ_{14}	1,000	–	–	–	–	–	–	0.442	0.087	0.398
	2,000	–	–	–	–	–	–	0.289	0.057	0.239
	3,000	–	–	–	–	–	–	0.308	0.046	0.235

(*continued*)

Table A2. Continued

	K	2			3			4		
	T	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$
μ_{21}	1,000	0.074	0.073	0.071	0.094	0.101	0.093	0.149	0.103	0.124
	2,000	0.053	0.054	0.055	0.011	0.011	0.010	0.076	0.074	0.073
	3,000	0.047	0.052	0.049	0.009	0.009	0.009	0.070	0.059	0.068
μ_{22}	1,000	0.019	0.019	0.019	0.176	0.013	0.171	0.223	0.019	0.130
	2,000	0.015	0.014	0.014	0.011	0.011	0.011	0.143	0.014	0.014
	3,000	0.011	0.011	0.011	0.010	0.011	0.010	0.116	0.012	0.116
μ_{23}	1,000	–	–	–	0.200	0.097	0.209	0.366	0.008	0.270
	2,000	–	–	–	0.067	0.063	0.067	0.173	0.006	0.054
	3,000	–	–	–	0.058	0.055	0.058	0.208	0.005	0.130
μ_{24}	1,000	–	–	–	–	–	–	0.097	0.091	0.094
	2,000	–	–	–	–	–	–	0.066	0.056	0.060
	3,000	–	–	–	–	–	–	0.063	0.046	0.052
κ_{11}	1,000	0.023	0.023	0.022	0.031	0.026	0.030	0.046	0.033	0.044
	2,000	0.018	0.018	0.018	0.021	0.020	0.021	0.028	0.024	0.029
	3,000	0.015	0.014	0.014	0.017	0.016	0.016	0.025	0.020	0.023
κ_{12}	1,000	0.045	0.045	0.044	0.051	0.033	0.049	0.063	0.048	0.050
	2,000	0.029	0.028	0.028	0.024	0.023	0.024	0.045	0.030	0.031
	3,000	0.025	0.024	0.024	0.019	0.019	0.019	0.030	0.022	0.029
κ_{13}	1,000	–	–	–	0.065	0.056	0.066	0.124	0.025	0.099
	2,000	–	–	–	0.040	0.036	0.039	0.084	0.017	0.066
	3,000	–	–	–	0.031	0.030	0.030	0.089	0.014	0.065
κ_{14}	1,000	–	–	–	–	–	–	0.090	0.046	0.078
	2,000	–	–	–	–	–	–	0.053	0.031	0.046
	3,000	–	–	–	–	–	–	0.056	0.028	0.048
κ_{21}	1,000	0.023	0.022	0.022	0.028	0.024	0.029	0.055	0.030	0.039
	2,000	0.018	0.017	0.017	0.020	0.019	0.021	0.025	0.021	0.026
	3,000	0.014	0.014	0.014	0.016	0.015	0.017	0.024	0.018	0.023
κ_{22}	1,000	0.020	0.018	0.018	0.032	0.014	0.035	0.095	0.018	0.074
	2,000	0.016	0.015	0.015	0.009	0.009	0.009	0.041	0.011	0.014
	3,000	0.012	0.010	0.010	0.008	0.007	0.008	0.014	0.010	0.012
κ_{23}	1,000	–	–	–	0.069	0.054	0.076	0.098	0.009	0.117
	2,000	–	–	–	0.038	0.038	0.039	0.117	0.006	0.104
	3,000	–	–	–	0.031	0.030	0.029	0.098	0.005	0.083
κ_{24}	1,000	–	–	–	–	–	–	0.128	0.050	0.114
	2,000	–	–	–	–	–	–	0.083	0.034	0.068
	3,000	–	–	–	–	–	–	0.088	0.027	0.070
ρ_1	1,000	0.018	0.018	0.018	0.022	0.024	0.022	0.044	0.027	0.044
	2,000	0.013	0.013	0.013	0.016	0.017	0.016	0.032	0.019	0.030
	3,000	0.011	0.010	0.010	0.014	0.013	0.015	0.033	0.015	0.026
ρ_2	1,000	0.054	0.055	0.057	0.044	0.047	0.045	0.083	0.055	0.080
	2,000	0.041	0.041	0.042	0.032	0.033	0.032	0.049	0.043	0.045
	3,000	0.036	0.032	0.036	0.029	0.027	0.028	0.042	0.040	0.037

(*continued*)

Table A2. Continued

	K	2			3			4		
	T	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$
ρ_3	1,000	–	–	–	0.088	0.055	0.087	0.080	0.050	0.074
	2,000	–	–	–	0.037	0.037	0.038	0.042	0.029	0.030
	3,000	–	–	–	0.030	0.029	0.030	0.049	0.023	0.034
ρ_4	1,000	–	–	–	–	–	–	0.094	0.045	0.087
	2,000	–	–	–	–	–	–	0.062	0.032	0.054
	3,000	–	–	–	–	–	–	0.063	0.026	0.050

Note. RMSE=root mean squared error.

Table A3. Results of the simulation study: RMSE for $\boldsymbol{\omega}$ in each considered scenario

	K	3			4		
	T	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$
ω_{12}	1,000	0.091	0.072	0.086	0.181	0.127	0.166
	2,000	0.053	0.054	0.052	0.112	0.078	0.092
	3,000	0.044	0.045	0.041	0.114	0.071	0.091
ω_{13}	1,000	0.148	0.137	0.148	0.157	0.110	0.161
	2,000	0.093	0.088	0.091	0.120	0.060	0.109
	3,000	0.072	0.072	0.072	0.113	0.051	0.103
ω_{14}	1,000	–	–	–	0.159	0.114	0.150
	2,000	–	–	–	0.105	0.066	0.097
	3,000	–	–	–	0.110	0.053	0.091
ω_{21}	1,000	0.109	0.119	0.104	0.169	0.122	0.159
	2,000	0.067	0.080	0.066	0.100	0.062	0.096
	3,000	0.063	0.059	0.061	0.095	0.056	0.080
ω_{23}	1,000	0.148	0.137	0.148	0.122	0.098	0.114
	2,000	0.093	0.088	0.091	0.072	0.068	0.070
	3,000	0.072	0.072	0.072	0.075	0.051	0.057
ω_{24}	1,000	–	–	–	0.128	0.093	0.119
	2,000	–	–	–	0.077	0.053	0.060
	3,000	–	–	–	0.082	0.051	0.076
ω_{31}	1,000	0.109	0.119	0.104	0.162	0.123	0.162
	2,000	0.067	0.080	0.066	0.133	0.072	0.123
	3,000	0.063	0.059	0.061	0.120	0.063	0.106
ω_{32}	1,000	0.091	0.072	0.086	0.148	0.100	0.130
	2,000	0.053	0.054	0.052	0.083	0.067	0.076
	3,000	0.044	0.045	0.041	0.093	0.057	0.063
ω_{34}	1,000	–	–	–	0.162	0.118	0.146
	2,000	–	–	–	0.099	0.073	0.091
	3,000	–	–	–	0.102	0.059	0.084

(*continued*)

Table A3. Continued

	K	3			4		
	T	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$
ω_{41}	1,000	–	–	–	0.211	0.166	0.204
	2,000	–	–	–	0.138	0.089	0.131
	3,000	–	–	–	0.132	0.075	0.116
ω_{42}	1,000	–	–	–	0.155	0.103	0.138
	2,000	–	–	–	0.100	0.066	0.075
	3,000	–	–	–	0.094	0.061	0.087
ω_{43}	1,000	–	–	–	0.172	0.133	0.164
	2,000	–	–	–	0.119	0.085	0.106
	3,000	–	–	–	0.113	0.070	0.096

Note. RMSE=root mean squared error.

References

Ailliot P., Bessac J., Monbet V., & Pène F. (2015). Non-homogeneous hidden Markov-switching models for wind time series. *Journal of Statistical Planning and Inference*, *160*, 75–88. https://doi.org/10.1016/j.jspi.2014.12.005

Anselone P. M. (1960). Ergodic theory for discrete semi-Markov chains. *Duke Mathematical Journal*, *27*(1), 33–40. https://doi.org/10.1215/S0012-7094-60-02703-4

Bartolucci F., & Farcomeni A. (2015). Information matrix for hidden Markov models with covariates. *Statistics and Computing*, *25*(3),515–526. https://doi.org/10.1007/s11222-014-9450-8

Bertotti L., & Cavaleri L. (2009). Wind and wave predictions in the adriatic sea. *Journal of Marine Systems*, *78*, S227–S234. https://doi.org/10.1016/j.jmarsys.2009.01.018

Biernacki C., Celeux G., & Govaert G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719–725. https://doi.org/10.1109/34.865189

Bulla J., Bulla I., & Nenadić O. (2010). hsmm—an R package for analyzing hidden semi-Markov models. *Computational Statistics & Data Analysis*, *54*(3), 611–619. https://doi.org/10.1016/j.csda.2008.08.025

Cappé O., Moulines E., & Rydén T. (2005). *Inference in hidden Markov models*. Springer. https://doi.org/10.1007/0-387-28982-8

Di Marzio M., Panzera A., & Taylor C. C. (2011). Kernel density estimation on the torus. *Journal of Statistical Planning and Inference*, *141*(6), 2156–2173. https://doi.org/10.1016/j.jspi.2011.01.002

Economou T., & Menary M. B. (2019). A hidden semi-Markov model for characterizing regime shifts in ocean density variability. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *68*(5), 1529–1553. https://doi.org/10.1111/rssc.12373

Fang C., Xu Y.-L., & Li Y. (2022). Optimized C-vine copula and environmental contour of joint wind-wave environment for sea-crossing bridges. *Journal of Wind Engineering and Industrial Aerodynamics*, *225*, 104989. https://doi.org/10.1016/j.jweia.2022.104989

Fisher N. I., & Lee A. J. (1983). A correlation coefficient for circular data. *Biometrika*, *70*(2), 327–332. https://doi.org/10.1093/biomet/70.2.327

Flor-Blanco G., Alcántara-Carrió J., Jackson D. W. T., Flor G., & Flores-Soriano C. (2021). Coastal erosion in NW Spain: Recent patterns under extreme storm wave events. *Geomorphology*, *387*, 107767. https://doi.org/10.1016/j.geomorph.2021.107767

Fruhwirth-Schnatter S., Celeux G., & Robert C. (2018). Model selection for mixture models—perspectives and strategies. In S. Fruhwirth-Schnatter, G. Celeux, & C.P. Robert (Eds.), *Handbook of mixture analysis*. Chapman and Hall/CRC.

Gu J., Zhang Y., Tuo P., Hu Z., Chen S., & Hu J. (2024). Surface floating objects moving from the Pearl River Estuary to Hainan Island: An observational and model study. *Journal of Marine Systems*, *241*, 103917. https://doi.org/10.1016/j.jmarsys.2023.103917

Holthuijsen L. H. (2007). *Waves in oceanic and coastal waters*. Cambridge University Press.

Hughes J. P., Guttorp P., & Charles S. P. (1999). A non-homogeneous hidden Markov model for precipitation occurrence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *48*(1), 15–30. https://doi.org/10.1111/1467-9876.00136

Jiang Q., Cioffi F., Conticello F. R., Giannini M., Telesca V., & Wang J. (2023). A stacked ensemble learning and non-homogeneous hidden Markov model for daily precipitation downscaling and projection. *Hydrological Processes*, *37*(9), e14992. https://doi.org/10.1002/hyp.v37.9

Kalbfleisch J. D., & Prentice R. L. (1980). *The statistical analysis of failure time data*. Wiley.

Kalisch H., Lagona F., & Roeber V. (2024). Sudden wave flooding on steep rock shores: A clear but hidden danger. *Natural Hazards*, *120*(3), 3105–3125. https://doi.org/10.1007/s11069-023-06319-w

Kato S., & Pewsey A. (2015). A Möbius transformation-induced distribution on the torus. *Biometrika*, *102*(2), 359–370. https://doi.org/10.1093/biomet/asv003

Langrock R., Kneib T., Sohn A., & DeRuiter S. L. (2015). Nonparametric inference in hidden Markov models using p-splines. *Biometrics*, *71*(2), 520–528. https://doi.org/10.1111/biom.12282

Langrock R., & Zucchini W. (2011). Hidden Markov models with arbitrary state dwell-time distributions. *Computational Statistics & Data Analysis*, *55*(1), 715–724. https://doi.org/10.1016/j.csda.2010.06.015

Lennox K. P., Dahl D. B., Vannucci M., & Tsai J. W. (2009). Density estimation for protein conformation angles using a bivariate von Mises distribution and Bayesian nonparametrics. *Journal of the American Statistical Association*, *104*(486), 586–596. https://doi.org/10.1198/jasa.2009.0024

Ley C., & Verdebout T. (2017). *Modern directional statistics*. Wiley. https://doi.org/10.1201/9781315119472

Liu D., Peddada S. D., Li L., & Weinberg C. R. (2006). Phase analysis of circadian-related genes in two tissues. *BMC Bioinformatics*, *7*, 1–10. https://doi.org/10.1186/1471-2105-7-1

Lund U., Agostinelli C., & Agostinelli M. C. (2017). Package 'circular'. *Repository CRAN*, *775*, 142. https://doi.org/10.32614/CRAN.package.circular

Mardia K. V., & Jupp P. E. (2000). *Directional statistics*. Wiley Online Library.

Maruotti A., Bulla J., Lagona F., Picone M., & Martella F. (2017). Dynamic mixtures of factor analyzers to characterize multivariate air pollutant exposures. *The Annals of Applied Statistics*, *11*(3), 1617–1648. https://doi.org/10.1214/17-AOAS1049

Maruotti A., & Punzo A. (2021). Initialization of hidden Markov and semi-Markov models: A critical evaluation of several strategies. *International Statistical Review*, *89*(3), 447–480. https://doi.org/10.1111/insr.v89.3

Merlo L., Maruotti A., Petrella L., & Punzo A. (2022). Quantile hidden semi-Markov models for multivariate time series. *Statistics and Computing*, *32*(4), 61. https://doi.org/10.1007/s11222-022-10130-1

Monbet V., Ailliot P., & Prevosto M. (2007). Survey of stochastic models for wind and sea state time series. *Probabilistic Engineering Mechanics*, *22*(2), 113–126. https://doi.org/10.1016/j.probengmech.2006.08.003

Nicol S., Cros M.-J., Peyrard N., Sabbadin R., Trépos R., Fuller R. A., & Woodworth B. K. (2023). Flywaynet: A hidden semi-Markov model for inferring the structure of migratory bird networks from count data. *Methods in Ecology and Evolution*, *14*(1), 265–279. https://doi.org/10.1111/mee3.v14.1

Pewsey A., & García-Portugués E. (2021). Recent advances in directional statistics. *TEST*, *30*(1), 1–58. https://doi.org/10.1007/s11749-021-00759-x

Pohle J., Adam T., & Beumer L. T. (2022). Flexible estimation of the state dwell-time distribution in hidden semi-Markov models. *Computational Statistics & Data Analysis*, *172*, 107479. https://doi.org/10.1016/j.csda.2022.107479

Ren B., & Barnett I. (2023). Combining mixed effects hidden Markov models with latent alternating recurrent event processes to model diurnal active–rest cycles. *Biometrics*, *79*(4), 3402–3417. https://doi.org/10.1111/biom.13865

Rivest L.-P. (1997). A decentred predictor for circular-circular regression. *Biometrika*, *84*(3), 717–726. https://doi.org/10.1093/biomet/84.3.717

Sansom J., & Thomson P. (2001). Fitting hidden semi-Markov models to breakpoint rainfall data. *Journal of Applied Probability*, *38*(A), 142–157. https://doi.org/10.1239/jap/1085496598

Shieh G. S., Zheng S., Johnson R. A., Chang Y.-F., Shimizu K., Wang C.-C., & Tang S.-L. (2011). Modeling and comparing the organization of circular genomes. *Bioinformatics*, *27*(7), 912–918. https://doi.org/10.1093/bioinformatics/btr049

Tutz G., & Schmid M. (2016). Basic regression models. In *Modeling discrete time-to-event data* (pp. 35–72). Springer.

Zucchini W., MacDonald I. L., & Langrock R. (2016). *Hidden Markov models for time series: An introduction using R*. Chapman and Hall/CRC.