A multilevel hidden Markov model for space-time cylindrical data

Un modello multilivello a classi markoviane latenti per dati cilindrici spazio-temporali

Francesco Lagona and Monia Ranalli

Abstract Motivated by segmentation issues in marine studies, a novel hidden Markov model is proposed for the analysis of cylindrical space-time series, that is, bivariate space-time series of intensities and angles. The model is a multilevel mixture of cylindrical densities, where the parameters of the mixture vary at the spatial level according to a latent Markov random field, while the parameters of the hidden Markov random field evolve at the temporal level according to the states of a hidden Markov chain. Due to the numerical intractability of the likelihood function, parameters are estimated by a computationally efficient EM algorithm based on the specification of a weighted composite likelihood. The proposal is tested in a case study that involves speeds and directions of marine currents in the Gulf of Naples.

Riassunto *Motivati da problemi di classificazione in studi marini, proponiamo un nuovo modello a classi markoviane latenti per l'analisi di serie cilindriche spazio-temporali, ovvero serie spazio-temporali bivariate di intensità ed angoli. Il modello è una mistura multi-livello di densità cilindriche, con parametri che variano al livello spaziale secondo un campo markoviano latente, i cui parametri variano nel tempo secondo una catena markoviana latente. A causa dell'intrattabilità numerica della funzione di verosimiglianza, i parametri sono stimati da un algoritmo EM basato sulla definizione di una funzione di pseudo-verosimiglianza composita. Il modello è applicato all'analisi di una serie spazio-temporale contenente le velocità e le direzioni delle correnti marine del golfo di Napoli.*

Key words: Cylindrical data, hidden Markov model, EM algorithm, Composite likelihood

Francesco Lagona
Department of Political Sciences, Roma Tre University, e-mail: francesco.lagona@uniroma3.it

Monia Ranalli
Department of Political Sciences, Roma Tre University e-mail: monia.ranalli@uniroma3.it

1 Introduction

A detailed knowledge of coastal currents is crucial for a valid integrated coastal zone management. Among the different available ocean observing technologies, high-frequency radars (HFRs) have unique characteristics, that make them play a key role in coastal observatories. HFR data can be conveniently described as space-time bivariate arrays of angles and intensities that respectively indicate the directions and the speeds of sea currents across space and over time. Data with a mixed circular-linear support are often referred to as *cylindrical* data [1], because the pair of an angle and an intensity can be represented as a point on a cylinder.

The statistical analysis of cylindrical space-time series is complicated by the unconventional topology of the cylinder and by the difficulties in modeling the cross-correlations between angular and linear measurements across space and over time. Additional complications arise from the skewness and the multimodality of the marginal distributions of the data. As a result, specific methods for the analysis of space-time cylindrical data have been relatively unexplored. Proposals in this context are limited to geostastical models, where cylindrical data are assumed conditionally independent given a latent process that varies continuously across space and time [15]. Geostatistical models give good results in open sea areas, where waves and currents can move freely without obstacles. Sea motion in coastal areas provides, however, a different setting. Coastal currents are shaped and constrained by the orography of the site. As a result, coastal circulation is much more irregular than ocean-type patterns and it is inaccurately represented by traditional geostatistical models, which do not incorporate orographic information. The development of a physical model that well represents sea motion in coastal areas can be a formidable task if the orography of the site is irregular. A more practical approach relies on decomposing an observed circulation pattern into a small number of local regimes whose interpretation is easier than the global pattern.

To accomplish this goal, we propose a model that segments coastal data according to finitely many latent classes that vary across space and time and are associated with the distribution of the data under specific, space-time varying, environmental conditions. Specifically, we assume that the joint distribution of the data is well approximated by a multi-level mixture of cylindrical densities. At each time, the parameters of the mixture vary according to a latent Markov field, whose parameters evolve over time according to a latent Markov chain. The idea of using hidden Markov models to segment cylindrical data is not completely novel. Hidden Markov models have been proposed for segmenting cylindrical time series [6] and hidden Markov fields have been proposed to segment spatial cylindrical data [11]. Our proposal integrates these specifications in a space-time setting.

A potential disadvantage of the model is the intractability of the likelihood function. We address estimation issues by relying on composite likelihood (CL) methods [14, 7]. This estimation strategy, on one hand, provides feasible and fast estimation methods. On the other hand, some dependence among observations is lost, resulting in a loss of statistical efficiency. However, consistency of the CL estimators still holds under regularity conditions [9]. Under these conditions, furthermore, CLEs are asymptotically normal with covariance matrix given by the inverse of a sandwich matrix, known as Godambe information [4] rather than the usual Fisher information matrix for maximum likelihood estimators (MLEs). CL methods have been successfully applied in spatial and space-time settings [11, 3].

2 Marine currents in the Gulf of Naples

The Gulf of Naples is a semienclosed marginal basin of the central Tyrrhenian Sea. It is a coastal area characterized by striking environmental contrasts: one of the most intensely urbanized coastlines in the whole Mediterranean, with massive industrial settlements, the very polluted Sarno river mouth, a number of distributed sewage outlets, coexisting with the extremely scenic coastal landscapes of the Sorrento Peninsula, of the Islands of Capri, Procida and Ischia and with unique underwater archaeological treasures (e.g. Baiae and Gaiola). For this reason, the Gulf of Naples has been subject to intense monitoring of its meteorological and oceanographic conditions. In particular, starting in 2004 an HFR system has been installed along its coastline, consisting first of two, and from 2008 of three, transceiving antennas operating at 25 MHz, providing hourly data of the surface current field at 1-km2 horizontal resolution. Such a system has shed light on very rich, multiple-scale surface dynamics and on the mechanisms driving water renewal of individual subbasins of the gulf [8, 13, 2]. Moreover, these data have been exploited in numerical models to enhance their predictive skills through state of the art assimilation schemes [5]. The functioning principle of HFRs is based on resonant backscatter, resulting from coherent reflection of a transmitted electromagnetic wave by ocean surface waves whose wavelength is half of the transmitted electromagnetic wavelength. As a result, every station can provide only the radial component of the surface currents with respect to the antenna location. Two, at least, or even better more stations (to ensure better statistics, to minimize gaps due to physical obstacles or to electromagnetic disturbances, to lower geometric dilution of precision) are needed to combine the radial information to obtain a current vector field. A vector map (or field) decomposes the current's field into the u- and v-components (Cartesian representation) of the sea surface at each observation point of a spatial lattice, where u corresponds to the west–east and v to the south–north current component. Joint modelling of u and v is, however, typically complicated by cross-correlations that vary dramatically in different parts of the spatial domain [12]. We therefore model sea current fields by using polar coordinates. Specifically, the observed current field is represented as a cylindrical spatial series, obtained by computing for each observation site the speed $y = \sqrt{u^2 + v^2} \in [0, +\infty)$ of the current (meters per second) and the direction $x = \tan2^{-1}(u, v) \in [0, 2\pi)$ of the current (radians), where $\tan2^{-1}$ is the inverse tangent function with two arguments and x follows the geographical convention, clockwise from North (0) to East ($\pi/2$). The data that motivated this paper include current speed and direction across a grid of 300 points, observed every hour during March 2009 in the Gulf of Naples.

3 A cylindrical space-time hidden Markov model

The data that motivated this work are in the form of an $n \times T$ array of cylindrical data, say $(\mathbf{z}_{it}, i = 1 \dots n, t = 1 \dots T)$, where $\mathbf{z}_{it} = (x_{it}, y_{it})$ is a pair of an angle $x_{it} \in [0, 2\pi)$ and an intensity $y_{it} \in [0, +\infty)$, observed at time t and in the spatial site i. We assume that the temporal evolution of these data is driven by a multinomial process in discrete time $\boldsymbol{\xi} = (\boldsymbol{\xi}_t, t = 1 \dots T)$, where $\boldsymbol{\xi}_t = (\xi_{t1}, \dots, \xi_{tK})$ is a multinomial random variable with K classes. We specifically assume that such process is distributed as a Markov chain, whose distribution, say $p(\boldsymbol{\xi}; \boldsymbol{\pi})$, is known up to a vector of parameters $\boldsymbol{\pi}$ that includes the initial probabilities and the transition probabilities of the chain. Conditionally on the value assumed each time by the Markov chain, the spatial distribution of the data at time t depends on a multinomial process in discrete space $\mathbf{u}_t = (\mathbf{u}_{it}, i = 1 \dots n)$, where $\mathbf{u}_{it} = (u_{it1}, \dots, u_{itG})$ is a multinomial variable with G classes. We assume that such spatial process is distributed as a G-parameter Potts model, whose parameters depend on the value taken at time t by the latent Markov chain $p(\boldsymbol{\xi}; \boldsymbol{\pi})$. This model depends on $G-1$ sufficient statistics

$$n_g(\boldsymbol{u}_t) = \sum_{i=1}^{n} u_{itg}, \quad g = 1 \dots G-1,$$

that indicate the frequencies of each latent class across the study area, and one sufficient statistic

$$n(\boldsymbol{u}_t) = \sum_{i=1}^{n} \sum_{j>i: j \in N(i)} \sum_{g=1}^{G_{t-1}} u_{itg} u_{jtg},$$

which indicates the frequency of neighboring sites which share the same class (for each i, $N(i)$ indicates the sets of neighboring sites of i). Precisely, we assume that the joint distribution of a sample $\mathbf{u}_t$, conditionally on $\boldsymbol{\xi}_t$, is known up to an array of class-specific parameters $\boldsymbol{\alpha} = (\alpha_{gk}, g = 1 \dots G-1, k = 1 \dots K)$ and a vector of auto-correlation parameters $\boldsymbol{\rho} = (\rho_1, \dots, \rho_k)$, and given by

$$p(\boldsymbol{u}_t \mid \boldsymbol{\xi}_t; \boldsymbol{\alpha}, \boldsymbol{\rho}) = \frac{\exp\left(\sum_{g=1}^{G_{t-1}} n_g(\boldsymbol{u}_t)\alpha_{gt} + n(\boldsymbol{u}_t)\rho_t\right)}{W(\boldsymbol{\alpha}, \boldsymbol{\rho})}, \tag{1}$$

where

$$\alpha_{gt} = \sum_{k=1}^{K} \xi_{tk} \alpha_{gk}$$

and

$$\rho_t = \sum_{k=1}^{K} \xi_{tk} \rho_k.$$

Our proposal is completed by assuming that, conditionally on the values taken by the Markov chain and the Potts model, the observed cylindrical data are independently distributed according to cylindrical densities, known up to a vector of parameters that depends on the latent class taken by the latent Markov random field at time t in site i. Precisely, we assume that

$$f(\mathbf{z} \mid \boldsymbol{\xi}, \mathbf{u}) = \prod_{i=1}^{n} \prod_{t=1}^{T} f(\mathbf{z}_{it}; \boldsymbol{\theta}_{itg}),$$

where

$$\boldsymbol{\theta}_{itg} = \sum_{g=1}^{G} u_{itg} \boldsymbol{\theta}_g,$$

and $\boldsymbol{\theta}_g$ is the gth entry of a vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_G)$. Under this setting, we follow [1] and exploit the following parametric cylindrical distribution, namely

$$f(\mathbf{z}; \boldsymbol{\theta}) = \frac{\alpha \beta^{\alpha}}{2\pi \cosh(\kappa)} (1 + \lambda \sin(x - \mu)) y^{\alpha - 1} \exp(-(\beta y)^{\alpha} (1 - \tanh(\kappa) \cos(x - \mu))), \tag{2}$$

known up to five parameters $\boldsymbol{\theta} = (\alpha, \beta, \kappa, \lambda, \mu)$, where $\alpha > 0$ is a shape parameter, $\beta > 0$ is a scale parameter, $\mu \in [0, 2\pi)$ is a circular location parameter, $\kappa > 0$ is a circular concentration parameter, while $\lambda \in [-1, 1]$ is a circular skewness parameter.

The joint distribution of the observed and the latent variables is therefore given by

$$f(\mathbf{z}, \mathbf{u}, \boldsymbol{\xi}; \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\rho}) = f(\mathbf{z} \mid \boldsymbol{u}; \boldsymbol{\theta}) p(\mathbf{u}; \boldsymbol{\rho}, \boldsymbol{\alpha}) p(\boldsymbol{\xi}; \boldsymbol{\pi}). \tag{3}$$

By integrating this distribution with respect to the unobserved variables, we obtain the likelihood function of the unknown parameters

$$L(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\rho}; \mathbf{z}) = \sum_{\boldsymbol{\xi}} \sum_{\boldsymbol{u}} f(\mathbf{z}, \boldsymbol{u}, \boldsymbol{\xi}; \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\rho}). \tag{4}$$

The maximization of the corresponding complete log-likelihood through an EM algorithm is unfeasible. As a result, we propose to estimate the parameters by maximizing a surrogate function, namely a composite log-likelihood function. Our proposal is based on the specification of a cover $\mathbb{A}$ of the set $S = \{1 \dots n\}$ of the observation sites, i.e. a family of (not necessarily disjoint) subsets $A \subseteq S$ such that $\cup_{A \in \mathbb{A}} = S$. For each subset A, we respectively define $\mathbf{z}_A = (\mathbf{z}_{it}, i \in A, t = 1 \dots T)$, $\mathbf{u}_A = (\mathbf{u}_{it}, i \in A, t = 1 \dots T)$, and

$$L^A(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\rho}; \mathbf{z}_A) = \sum_{\boldsymbol{\xi}} \sum_{\boldsymbol{u}_A} f(\mathbf{z}_A, \boldsymbol{u}_A, \boldsymbol{\xi}; \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\rho}) \tag{5}$$

as the contribution of the data in A to the composite likelihood (CL), where CL=$\prod_{A \in \mathbb{A}} L^A$. This composite likelihood function involves summations over all the possible values that $\boldsymbol{u}_A$ can take. As a result, the numerical tractability of these steps dramatically decreases with the cardinality of the largest subset of the cover $\mathbb{A}$. On the one side, this would suggest to choose a cover with many small subsets. On the other side, a cover that includes a few large subsets is expected to provide a CL function that is a better approximation of the likelihood function. Because summations over $\boldsymbol{u}_A$ become cumbersome for $\mid A \mid \geq 3$, a natural strategy is a cover that includes subsets with 2 elements. When $\mathbb{A}$ include all the subsets of two elements, then composite likelihood reduces to the pairwise likelihood function [14]. In a spatial setting, a pairwise likelihood can be further simplified by discarding all the pairs (i, j) that are not in the neighborhood structure $N(i), i = 1 \ldots n$. This choice provides a computationally efficient EM algorithm, without sacrificing the good distributional properties that are expected by a CL estimator.

Acknowledgements The authors were supported by the 2015 PRIN project 'Environmental processes and human activities: capturing their interactions via statistical methods', funded by the Italian Ministry of Education, University and Research.

References

1. Abe, T. and C. Ley (2017). A tractable, parsimonious and flexible model for cylindrical data, with applications. Econometrics and Statistics 4, 91-104.
2. Cianelli, D., Falco, P., Iermano, I., Mozzillo, P., Uttieri, M., Buonocore, B., Zambardino, G. and Zambianchi, E. (2015) Inshore/offshore water exchange in the Gulf of Naples. Journal of Marine Systems, 145, 37-52.
3. Eidsvik, J., B. A. Shaby, B. J. Reich, M. Wheeler, and J. Niemi (2014). Estimation and prediction in spatial models with block composite likelihoods. Journal of Computational and Graphical Statistics 23(2), 295-315.
4. Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. The Annals of Mathematical Statistics 31(4), 1208-1211.
5. Iermano, I., Moore, A. and Zambianchi, E. (2016) Impacts of a 4-dimensional variational data assimilation in a coastal ocean model of southern Tyrrhenian Sea. Journal of Marine Systems, 154, 157-171.
6. Lagona, F., M. Picone, and A. Maruotti (2015). A hidden markov model for the analysis of cylindrical time series. Environmetrics 26, 534–544.
7. Lindsay, B. (1988) Composite likelihood methods. Contemporary Mathematics, 80, 221-239.
8. Menna, M., Mercatini, A., Uttieri, M., Buonocore, B. and Zambianchi, E. (2007). Wintertime transport processes in the Gulf of Naples investigated by HF radar measurements of surface currents. Nuovo Cimento C, 30, 605-622.
9. Molenberghs, G. and G. Verbeke (2005). Models for discrete longitudinal data. Springer Series in Statistics. Springer Science+Business Media, Incorporated New York.
10. Okabayashi, S., L. Johnson, and C. Geyer (2011). Extending pseudo-likelihood for Potts models. Statistica Sinica 21(1), 331-347.
11. Ranalli, M., F. Lagona, M. Picone, and E. Zambianchi (2018). Segmentation of sea current fields by cylindrical hidden Markov models: a composite likelihood approach. Journal of the Royal Statistical Society C 67(3), 575-598.
12. Reich, B. and Fuentes, M. (2007) A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. Annals of Applied Statistics, 1, 249-264.
13. Uttieri, M., Cianelli, D., Nardelli, B. B., Buonocore, B., Falco, P., Colella, S. and Zambianchi, E. (2011) Multiplatform observation of the surface circulation in the Gulf of Naples (southern Tyrrhenian sea). Ocean Dynamics, 61, 779-796.
14. Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. Statistica Sinica 21(1), 1–41.
15. Wang, F., A. Gelfand, and G. Jona Lasinio (2015). Joint spatio-temporal analysis of a linear and a directional variable: space-time modeling of wave heights and wave directions in the Adriatic sea. Statistica Sinica 25, 25–39.