ROMA TRE
UNIVERSITÀ DEGLI STUDI

Scuola dottorale in "Economia e Metodi Quantitativi"
Sezione "Metodi Statistici per l'Economia e l'Impresa"
XVIII ciclo

TEMPORAL AND SPATIO-TEMPORAL MODELS FOR CIRCULAR AND CIRCULAR-LINEAR DATA

Tutor:
Prof. Antonello Maruotti

Dottorando:
Gianluca Mastrantonio

Correlatore:
Prof.sa Giovanna Jona Lasinio

ROMA TRE
UNIVERSITÀ DEGLI STUDI

Scuola dottorale in "Economia e Metodi Quantitativi"
Sezione "Metodi Statistici per l'Economia e l'Impresa"
XVIII ciclo

TEMPORAL AND SPATIO-TEMPORAL MODELS FOR CIRCULAR AND CIRCULAR-LINEAR DATA

Tutor:
Prof. Antonello Maruotti

Dottorando:
Gianluca Mastrantonio

Correlatore:
Prof.sa Giovanna Jona Lasinio

Contents

Chapter 1

Introduction

Circular (or directional) data, i.e. observations with support the unit circle, arise naturally in many scientific fields, for example meteorology (Bulla *et al.*, 2012), social sciences (Gill and Hangartner, 2010), biology (Morales *et al.*, 2004; Patterson *et al.*, 2008) and musicology (Beran, 2004; Resa, 2010).

Due to the circular domain, to the sensitivity of descriptive and inferential results to the starting point and orientation on the circle (Mastrantonio *et al.*, 2015), analysis of circular data is more challenging than for linear (or inline) ones. Standard statistical methods, as the sample mean and variance, loose their meaning if used on circular data and must be replaced by their circular counterparts; for a general discussion see the books of Fisher (1996), Mardia and Jupp (1999) or Jammalamadaka and SenGupta (2001).

Dating back to Von Mises (1918), the attention in circular data has increased over time (Mardia, 1972; Fisher, 1996; Mardia and Jupp, 1999; Jammalamadaka and SenGupta, 2001; Pewsey *et al.*, 2013), leading to important probability distributions theory and inferential results. Many univariate distributions have been proposed to model circular data (Mardia and Jupp, 1999), and there are few multivariate extensions, generally bivariate: the bivariate circular distributions proposed by Mardia (1975a), Mardia (1975b) Singh *et al.* (2002), the bivariate wrapped normal of Johnson and Wehrly (1977), the multivariate von Mises of Mardia *et al.* (2008) and the multivariate wrapped normal of Baba (1981), while Ong and SenGupta (2012) propose a mixture approach to built multivariate distributions. In the late years the analysis became less descriptive and more inferential: for example have been proposed linear models (Harrison and Kanji, 1988; Kato *et al.*, 2008), spatial models (Modlin *et al.*, 2012; Jona Lasinio *et al.*, 2012; Wang and Gelfand, 2014), temporal models (Holzmann *et al.*, 2006; Bulla *et al.*, 2012; Maruotti *et al.*, 2015) and spatio-temporal models (Lagona *et al.*, 2014; Mastrantonio *et al.*, 2015b,a).

Often circular variables are recorded along with linear ones, e.g. wind direction and velocity or wave height and direction, and hence a joint modeling of mixed type variables is often needed. The circular and linear variables live on different spaces and there is not an obvious way to define multivariate circular-linear distributions. Up to now there are few proposals and among all of them the most interesting are capable to model cylindrical

data (one linear variable and one circular), for example SenGupta (2004) and Sengupta and Ong (2014) offer various methods to construct circular-linear distributions and Abe and Ley (2015) proposed a distribution based on the Weibull and the wrapped Cauchy.

Although the Bayesian framework has proved to be well suited to model complex data, there is little in the way of formal theory for circular variables in the fully Bayesian setting, some examples are the spatial models of Jona Lasinio *et al.* (2012) and Mastrantonio *et al.* (2015a), the spatio-temporal models of Wang *et al.* (2015) and Mastrantonio *et al.* (2015b) or the model for longitudinal data of Nuñez-Antonio and Gutiérrez-Peña (2014).

1.1 Overview

The primary objective of this Thesis is to propose models for temporal and spatio-temporal circular and circular-linear data. We show that, under a Bayesian framework, the complex nature of circular data and the difficulties in a joint modelling of circular-linear variables, can be easily overcome. In the Thesis two main research frameworks are touched. The first deals with the build of spatio-temporal models for circular variables, while the seconds addresses topics in the joint temporal classification of circular and linear variables. In all the models proposed, exploiting data augmentation techniques, we are able to define efficient, and easy to implement, Markov chain Monte Carlo (MCMC) algorithms.

This Thesis is based on a selection of 4 articles produced during the three years Ph.D. In these papers, among other things, we introduce processes with non-separable spatio-temporal correlation function and nugget effect (Chapter 3.1), a new circular processes (Chapter 3.2), a cylindrical hidden Markov model (HMM) (Chapter 4.1), a new multivariate circular-linear distribution, a non-parametric HMM based on the Dirichlet process, and a new algorithm to estimate the projected normal parameters (Chapter 4.2).

The Thesis is organized as follows. In Chapter 2 we formalize the circular random variables and we show how to obtain a circular distribution starting from a linear one, namely the wrapping (Chapter 2.1) ad the projection (Chapter 2.2). In Chapter 3 we first introduce the basic ideas in spatio-temporal modeling of circular data necessary to understand the new developments obtained in Chapters 3.1 and 3.2.

- Chapter 3.1 - *"Spatio-temporal circular models with non-separable covariance structure"*.
  We extend the wrapped (Jona Lasinio *et al.*, 2012) and projected Gaussian (Wang and Gelfand, 2014) processes by i) introducing a flexible correlation structure, i.e. the Gneiting non-separable function (Gneiting, 2002), ii) a nugget effect for circular variables and iii) modelling the circular mean and variance with linear covariates. The predictive performances of the models proposed, are evaluated and compared on a real data example.

- Chapter 3.2 - *"The wrapped skew Gaussian process for analyzing spatio-temporal data"*.

In Chapter 3.1 we show that the projected Gaussian process is more flexible than the wrapped Gaussian, but it is more difficult to implement and the associated parameters have not a straightforward interpretation. In this Chapter, we introduce a new spatio-temporal process based on the wrapped skew normal (Pewsey, 2000), i.e. the wrapped skew Gaussian process. The new process retains straightforward parametric interpretation and it is more flexible than the wrapped Gaussian one. We show, with simulated and real data examples, that, in terms of predictive ability, the wrapped skew Gaussian process outperforms the wrapped Gaussian, even if the data are simulated from the latter.

In Chapter 3 we formalize the HMM for multivariate linear variables and then we show how to obtain the circular-linear extensions of Chapters 4.1 and 4.2.

- Chapter 4.1 - *"Bayesian hidden Markov modelling using circular-linear general projected normal distribution"*.
  In this work we introduce an HMM suitable to model cylindrical data. The emission distribution of the HMM is based on the normal and the general projected normal (Wang and Gelfand, 2013). The distribution allows to have dependence among the circular and linear variables and a bimodal marginal circular distribution. If the circular-linear dependence is ignored, we empirically demonstrate that the number of latent states are generally overestimated

- Chapter 4.2 - *"A Bayesian hidden Markov model for telemetry data"*.
  The work is motivated by a real data example of six free-ranging Maremma sheepdogs. The data are a time series of six linear variables (step-lengths) and six circular variables (turning-angles). The multivariate nature of the data requires the definition of a multivariate circular-linear distribution with multivariate interaction. The time series is modelled using a non-parametric HMM based on the hierarchical Dirichlet process. We show that our proposed emission distribution outperforms the most used in the literature.

The Thesis ends with a discussion (Chapter 5).

Bibliography

Abe, T. and Ley, C. (2015). A tractable, parsimonious and highly flexible model for cylindrical data, with applications. *ArXiv e-prints*.

Baba, Y. (1981). Statistics of angular data : wrapped normal distribution model (in japanese). In *Proceedings of the Institute of Statistical mathematics*, volume 28, pages 179–195.

Beran, J. (2004). *Statistics in Musicology*. Interdisciplinary statistics. Chapman and Hall/CRC, Boca Raton.

Bulla, J., Lagona, F., Maruotti, A., and Picone, M. (2012). A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series. *Journal of Agricultural, Biological, and Environmental Statistics*, **17**(4), 544–567.

Fisher, N. I. (1996). *Statistical Analysis of Circular Data*. Cambridge University Press, Cambridge.

Gill, J. and Hangartner, D. (2010). Circular data in political science and how to handle it. *Political Analysis*, **18**(3), 316–336.

Gneiting, T. (2002). Nonseparable, Stationary Covariance Functions for Space-Time Data. *Journal of the American Statistical Association*, **97**(458), 590–600.

Harrison, D. and Kanji, G. K. (1988). The Development of Analysis of Variance for Circular Data. *Journal of Applied Statistics*, **15**, 197–224.

Holzmann, H., Munk, A., Suster, M., and Zucchini, W. (2006). Hidden Markov models for circular and linear-circular time series. *Environmental and Ecological Statistics*, **13**(3), 325–347.

Jammalamadaka, S. R. and SenGupta, A. (2001). *Topics in Circular Statistics*. World Scientific, Singapore.

Johnson, R. and Wehrly, T. (1977). Measures and models for angular correlation and angular-linear correlation. *Journal of the Royal Statistical Society Series B*, **39**(222-229).

Jona Lasinio, G., Gelfand, A., and Jona Lasinio, M. (2012). Spatial analysis of wave direction data using wrapped Gaussian processes. *Annals of Applied Statistics*, **6**(4), 1478–1498.

Kato, S., Shimizu, K., and Shieh, G. S. (2008). A circular-circular regression model. *Statistica Sinica*, **18**, 633–645.

Lagona, F., Picone, M., Maruotti, A., and Cosoli, S. (2014). A hidden Markov approach to the analysis of space-time environmental data with linear and circular components. *Stochastic Environmental Research and Risk Assessment*, **to appear**.

Mardia, K. (1975a). Characterizations of directional distributions. In G. Patil, S. Kotz, and J. Ord, editors, *A Modern Course on Statistical Distributions in Scientific Work*, volume 17 of *NATO Advanced Study Institutes Series*, pages 365–385. Springer Netherlands.

Mardia, K. V. (1972). *Statistics of Directional Data*. Academic Press, London.

Mardia, K. V. (1975b). Statistics of directional data. *Journal of the Royal Statistical Society Series B*, **37**, 349–393.

Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics*. John Wiley and Sons, Chichester.

Mardia, K. V., Hughes, G., Taylor, C. C., and Singh, H. (2008). A multivariate von Mises distribution with applications to bioinformatics. *Canadian Journal of Statistics*, **36**(1), 99–109.

Maruotti, A., Punzo, A., Mastrantonio, G., and Lagona., F. (2015). A time-dependent extension of the projected normal regression model for longitudinal circular data based on a hidden Markov heterogeneity structure. *Stochastic Environmental Research and Risk Assessment*, **To appear**.

Mastrantonio, G., Jona Lasinio, G., Maruotti, A., and Calise, G. (2015). On initial direction, orientation and discreteness in the analysis of circular variables. *ArXiv e-prints*.

Mastrantonio, G., Jona Lasinio, G., and Gelfand, A. E. (2015a). Spatio-temporal circular models with non-separable covariance structure. *TEST*, **To appear**.

Mastrantonio, G., Gelfand, A. E., and Jona Lasinio, G. (2015b). The wrapped skew Gaussian process for analyzing spatio-temporal data. *Stochastic Environmental Research and Risk Assessment*, **To appear**.

Modlin, D., Fuentes, M., and Reich, B. (2012). Circular conditional autoregressive modeling of vector fields. *Environmetrics*, **23**(1), 46–53.

Morales, J. M., Haydon, D. T., Frair, J., Holsinger, K. E., and Fryxell, J. M. (2004). Extracting more out of relocation data: building movement models as mixtures of random walks. *Ecology*, **85**(9), 2436–2445.

Nuñez-Antonio, G. and Gutiérrez-Peña, E. (2014). A Bayesian model for longitudinal circular data based on the projected normal distribution. *Computational Statistics and Data Analysis*, **71**(C), 506–519.

Ong, S. H. and SenGupta, A. (2012). Bivariate and multivariate circular distributions by mixtures. *Journal of Indian Statistical association*, **50**, 193–204.

Patterson, T., Thomas, L., Wilcox, C., Ovaskainen, O., and Matthiopoulos, J. (2008). State-space models of individual animal movement. *Trends in Ecology & Evolution*, **23**(2), 87–94.

Pewsey, A. (2000). The wrapped skew-normal distribution on the circle. *Communications in Statistics - Theory and Methods*, **29**(11), 2459–2472.

Pewsey, A., Neuhäuser, M., and Ruxton, G. D. (2013). *Circular Statistics in R*. Oxford University Press, Croydon.

Resa, Z. (2010). *Towards Time-aware Contextual Music Recommendation: An Exploration of Temporal Patterns of Music Listening Using Circular Statistics*. Master's thesis.

SenGupta, A. (2004). On the construction of probability distributions for directional data. *Bulletin of Calcutta Mathematical Society*, **96**, 139–154.

Sengupta, A. and Ong, S. H. (2014). A unified approach for construction of probability models for bivariate linear and directional data. *Communications in Statistics - Theory and Methods*, **43**(10-12), 2563–2569.

Singh, H., Hnizdo, V., and Demchuk, E. (2002). Probabilistic model for two dependent circular variables. *Biometrika*, **89**(3), 719–723.

Von Mises, R. (1918). Über die ganzzahligkeit der atomgewicht und verwandte fragen. *Phys. Z*, **19**, 490–500.

Wang, F. and Gelfand, A. E. (2013). Directional data analysis under the general projected normal distribution. *Statistical Methodology*, **10**(1), 113–127.

Wang, F. and Gelfand, A. E. (2014). Modeling space and space-time directional data using projected Gaussian processes. *Journal of the American Statistical Association*, **109**(508), 1565–1580.

Wang, F., Gelfand, A., and Jona Lasinio, G. (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**(1), 25–39.

Chapter 2

Circular variables and distributions

Let $\{\mathbb{S}, \mathcal{A}, P\}$ be a probability space, where the sample space $\mathbb{S} = \{(u_1, u_2) : u_1^2 + u_2^2 = 1\}$ is the unit circle, $\mathcal{A}$ is the $\sigma-$algebra on $\mathbb{S}$ and $P : \mathbb{S} \to [0, 1]$ is the normalized Lebesgue measure on the measurable space $\{\mathbb{S}, \mathcal{A}\}$. Let $\mathbb{D}$ be a subset of $\mathbb{R}$, such that its length is 2π, and consider the measurable function $\Theta : \mathbb{S} \to \mathbb{D}$, with $\Theta^{-1}(d) = (u_1, u_2) = (\cos d, \sin d)$, $d \in \mathbb{D}$. Let $\mathcal{D} = \sigma(\mathbb{D}))$ be the $\sigma-$algebra of $\mathbb{D}$ induced by Θ and $A_{\Theta,D} \equiv \{(x, y) : \Theta(x, y) \in D\}$ and $\mathbb{P}_\Theta(D) = P(\Theta^{-1}(D)) = P\left(A_{\Theta,D}\right), \forall D \in \mathcal{D}$. The measurable space induced by Θ is $(\mathbb{D}, \mathcal{D}, \mathbb{P}_\Theta)$ with

1. $\mathbb{P}_\Theta(D) = P\left(A_{\Theta,D}\right) \geq 0, \forall D \in \mathcal{D}$;
2. $\mathbb{P}_\Theta(\mathbb{D}) = P\left(A_{\Theta,\mathbb{D}}\right) = 1$;
3. for any countable sequence of disjoint sets $\{D_j\}_{j=1}^{\infty}$ of $\mathcal{D}$,

$$\mathbb{P}_\Theta\left(\cup_{j=1}^{\infty} D_j\right) = P\left(A_{\Theta,\cup_{j=1}^{\infty} D_j}\right) = P\left(\cup_{j=1}^{\infty} A_{\Theta,D_j}\right) = \sum_{j=1}^{\infty} P\left(A_{\Theta,D_j}\right) = \sum_{j=1}^{\infty} \mathbb{P}_\Theta(D_j),$$

i.e. $(\mathbb{D}, \mathcal{D}, \mathbb{P}_\Theta)$ is a probability space. It follows that Θ is a random variable and $\mathbb{P}_\Theta$ is its *probability distribution*. Θ represents an angle over the unit circle and it is called a *circular random variable*. Accordingly, for all $d \in \mathbb{D}$, $\Theta^{-1}(d) = \Theta^{-1}(d \mod 2\pi)$ and, without loss of generality, we can represent any circular variable in $[0, 2\pi)$. $\mathbb{D}$ can be either continuous or discrete. In the latter case, it is generally assumed that it is composed of l distinct points equally spaced, e.g. $\mathbb{D} \equiv \{2\pi j/l\}_{j=0}^{l-1}$. If $\mathbb{D}$ is a continuous domain, Θ is a continuous circular variable and $\mathbb{P}_\Theta$ is the Lebesgue measure. On the other hand, if $\mathbb{D}$ is discrete, Θ is a discrete circular variable or a *lattice* circular variable (see Mardia and Jupp, 1999), and $\mathbb{P}_\Theta$ is the counting measure. In both cases $f_\Theta = d\mathbb{P}_\Theta/dP_\Theta : \mathbb{D} \to \mathbb{R}^+$ is the Radon-Nicodym derivative of $\mathbb{P}_\Theta$, i.e. $\mathbb{P}_\Theta(D) = \int_D f_\Theta dP_\Theta$.

There are different approaches to specify valid circular distributions, see for example Jammalamadaka and SenGupta (2001). In this Thesis we will focus on two methods that allow to built a circular distribution starting from a linear one, namely the wrapping (Chapter 2.1) and the projection (Chapter 2.2). Under both methods, the resulting distribution has a complex functional forms but introducing a suitable latent variable, the joint distribution of observed and latent variables are easy to handle in a fully Bayesian framework.

2.1 The wrapping approach

Let $Y \in \mathbb{R}$ be a linear random variable with probability density function (pdf) $f_Y(\cdot|\boldsymbol{\psi})$, where $\boldsymbol{\psi}$ is a generic vector of parameters. We can obtain a circular random variable using the following transformation:

$$\Theta = Y \bmod 2\pi \in [0, 2\pi). \tag{2.1}$$

The pdf of Θ is

$$f_\Theta(\theta|\boldsymbol{\psi}) = \sum_{k=-\infty}^{\infty} f_Y(\theta + 2\pi k|\boldsymbol{\psi}). \tag{2.2}$$

Between Y and Θ there is the following relation: $Y = \Theta + 2\pi K$, where K is called the *winding number*. Equation (2.2) wraps $f_Y(\cdot|\boldsymbol{\psi})$ around the unit circle and Θ is called the *wrapped* version of $\mathbf{Y}$ with period 2π, e.g. if Y is normally distributed, then Θ follows a *wrapped normal* (WN) distribution.

It is not easy to work directly with equation (2.2), since it requires the evaluation of an infinite sum. Following Coles and Casson (1998), if we consider K as (latent) random variable we can see that $f_{\Theta,K}(\theta, k|\boldsymbol{\psi}) = f_Y(\theta + 2\pi k|\boldsymbol{\psi})$, i.e. $f_Y(\theta + 2\pi k|\boldsymbol{\psi})$ is the joint density of (Θ, K), and a marginalization over K gives equation (2.2). The marginal distribution of K is

$$f_K(k|\boldsymbol{\psi}) = \int_0^{2\pi} f_Y(\theta + 2\pi k|\boldsymbol{\psi}) d\theta.$$

The conditional distributions of $K|\Theta, \boldsymbol{\psi}$ and $\Theta|K, \boldsymbol{\psi}$ are respectively

$$\frac{f_Y(\theta + 2\pi k|\boldsymbol{\psi})}{\sum_{k=-\infty}^{\infty} f_Y(\theta + 2\pi k|\boldsymbol{\psi})},$$

and

$$\frac{f_Y(\theta + 2\pi k|\boldsymbol{\psi})}{\int_0^{2\pi} f_Y(\theta + 2\pi k|\boldsymbol{\psi}) d\theta}.$$

It is generally easier to work with the joint density of $\Theta, K|\boldsymbol{\psi}$, with respect to the one of $\Theta|\boldsymbol{\psi}$, since the former does not require the evaluation of the infinite sum. For example if Y is Gaussian, then the joint density of $(\Theta, K|\boldsymbol{\psi})$ is the normal pdf evaluated at $\theta + 2\pi k$.

The wrapping approach can be easily extended to a multivariate setting (Jona Lasinio

et al., 2012). Let $\mathbf{Y} = (Y_1, \ldots, Y_p)'$ be a p-variate vector with pdf $f_{\mathbf{Y}}(\cdot|\boldsymbol{\psi})$, then $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_p)'$, with

$$\Theta_i = Y_i \bmod 2\pi \in [0, 2\pi)$$

is a vector of circular variables with pdf

$$f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\boldsymbol{\psi}) = \sum_{k_1=-\infty}^{\infty} \cdots \sum_{k_p=-\infty}^{\infty} f_{\mathbf{Y}}(\boldsymbol{\theta} + 2\pi\mathbf{k}|\boldsymbol{\psi}). \tag{2.3}$$

Here again, it is easier to work with the joint density of $\boldsymbol{\Theta}, \mathbf{K}|\boldsymbol{\psi}$ with respect to the one of $\boldsymbol{\Theta}|\boldsymbol{\psi}$, treating $\mathbf{K}$ as a vector of latent random variables.

2.2 The projection approach

Let $\mathbf{Y} = (Y_1, Y_2)$ be a bivariate vector of linear variables with pdf $f_{\mathbf{Y}}(\cdot|\boldsymbol{\psi})$. The unit vector

$$\mathbf{U} = \frac{\mathbf{Y}}{||\mathbf{Y}||}$$

represents a point over the unit circle and the associated angle Θ, where $U_1 = \cos(\Theta)$ and $U_2 = \sin(\Theta)$, is a circular random variable. We have that

$$\tan(\Theta) = \frac{Y_2}{Y_1} = \frac{U_2}{U_1}. \tag{2.4}$$

Since the period of the tangent is π, inversion of this function, to obtain Θ in equation (2.4), requires some care. A common choice is the atan*, formally defined in Jammalamadaka and SenGupta (2001), pag. 13, that takes into account the signs of Y_1 and Y_2 to determine the right portion of the unit circle where Θ is located. Between Θ and $\mathbf{Y}$ the following relation exists

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = R \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} = R\mathbf{U},$$

with $R = ||\mathbf{Y}||$.

The pdf of $\Theta|\boldsymbol{\psi}$ is

$$f_{\Theta}(\theta|\boldsymbol{\psi}) = \int_{\mathbb{R}^+} r f_{\mathbf{Y}}((r\cos(\theta), r\sin(\theta))'|\boldsymbol{\psi})dr. \tag{2.5}$$

The integral in equation (2.5) is not easy to solve and, even when a closed form exists, the resulting pdf has a complicated functional structure. For example if $\mathbf{Y} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix},$$

then (Wang and Gelfand, 2013)

$$f_\Theta(\theta|\boldsymbol{\psi}) = \frac{\phi_2(\boldsymbol{\mu}|\mathbf{0}_2, \boldsymbol{\Sigma}) + aD(\theta)\Phi(D(\theta)|\mathbf{0}, \mathbf{I}_2)\phi(aC(\theta)^{-\frac{1}{2}}(\mu_1 \sin(\theta)) - \mu_2 \cos(\theta))}{C(\theta)},$$

where $\phi_n(\cdot|\cdot,\cdot)$ and $\Phi_n(\cdot|\cdot,\cdot)$ are, respectively, the $n-$variate normal pdf and cumulative density function, with

$$\begin{aligned}
a &= \left(\sigma_1\sigma_2\sqrt{1-\rho^2}\right)^{-1}, \\
C(\theta) &= a^2\left(\sigma_2^2\cos^2(\theta) + \sigma_1^2\sin^2(\theta) - \rho\sigma_1\sigma_2\sin(2\theta)\right), \\
D(\theta) &= a^2 C(\theta)^{-\frac{1}{2}}\left(\mu_1\sigma_2(\sigma_2\cos(\theta) - \rho\sigma_1\sin(\theta)) + \mu_2\sigma_1(\sigma_1\sin(\theta) - \rho\sigma_2\cos(\theta))\right).
\end{aligned}$$

Note that the joint density of $\Theta, R|\boldsymbol{\psi}$ is

$$f_{\mathbf{Y}}((r\cos(\theta), r\sin(\theta))|\boldsymbol{\psi}),$$

see equation (2.5) and, for example, if $\mathbf{Y} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then

$$f_{\Theta,R}(\theta, r|\boldsymbol{\psi}) = r\phi_2\left((r\cos(\theta), r\sin(\theta))'|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right).$$

Since the distribution of Θ does not change if $\mathbf{Y}$ is scaled by $c > 0$, i.e.

$$\Theta = \text{atan}^*\left(\frac{Y_2}{Y_1}\right) = \text{atan}^*\left(\frac{cY_2}{cY_1}\right), \tag{2.6}$$

a constraint on the density parameters is needed for identifiability purpose, e.g. if $\mathbf{Y}$ is bivariate normal, it is sufficient to set $\sigma_2^2 = 1$.

The projection approach can be easily adapted to obtain a distribution for multivariate circular variable (Wang and Gelfand, 2014). If $\mathbf{Y}$ is a $2p-$variate linear variable, a $p-$variate vector of (projected) circular variables is obtained with the following transformation:

$$\Theta_i = \text{atan}^*\left(\frac{Y_{2i}}{Y_{2i-1}}\right), i = 1, \ldots, p. \tag{2.7}$$

The pdf of $\boldsymbol{\Theta}|\boldsymbol{\psi}$, where $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_p)'$, is

$$f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\boldsymbol{\psi}) = \int_{\mathbb{R}^+} \ldots \int_{\mathbb{R}^+} \prod_{i=1}^{p} r_i f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\psi}) dr_1 \ldots dr_p, \tag{2.8}$$

where $r_i = ||(y_{2i-1}, y_{2i})'||$ and, in equation (2.8), $\mathbf{y}$ is a function of $\boldsymbol{\theta}$ and $\mathbf{r} = (r_1, \ldots, r_p)$.

Although this procedure is straightforward, the p constraints needed to identify the density, one for each transformation (2.7), poses difficulties in the parameter estimations. Those difficulties can be overcome if the multivariate dependence is structured, e.g. spatially or temporally, but until our work, see Mastrantonio (2015) or Section 4.32, in the

general case no feasible proposals for the estimation were available.

In the multivariate case, as in the univariate one, it is generally easier to work with the joint density of $\boldsymbol{\Theta}, \mathbf{R}|\boldsymbol{\psi}$ (the integrand in (2.8)) than the one of $\boldsymbol{\Theta}|\boldsymbol{\psi}$.

Bibliography

Coles, S. and Casson, E. (1998). Extreme value modelling of hurricane wind speeds. *Structural Safety*, **20**(3), 283 – 296.

Jammalamadaka, S. R. and SenGupta, A. (2001). *Topics in Circular Statistics*. World Scientific, Singapore.

Jona Lasinio, G., Gelfand, A., and Jona Lasinio, M. (2012). Spatial analysis of wave direction data using wrapped Gaussian processes. *Annals of Applied Statistics*, **6**(4), 1478–1498.

Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics*. John Wiley and Sons, Chichester.

Mastrantonio, G. (2015). A Bayesian hidden Markov model for telemetry data. *ArXiv e-prints - Submitted to "Annals of Applied Statistics"*.

Wang, F. and Gelfand, A. E. (2013). Directional data analysis under the general projected normal distribution. *Statistical Methodology*, **10**(1), 113–127.

Wang, F. and Gelfand, A. E. (2014). Modeling space and space-time directional data using projected Gaussian processes. *Journal of the American Statistical Association*, **109**(508), 1565–1580.

Chapter 3

Spatio-temporal processes for circular variables

A stochastic process can be defined through its finite dimensional distribution, i.e. the distribution of an $n-$dimensional realization, that is a multivariate pdf (Gelfand *et al.*, 2010). As we showed in the previous Chapter, starting from a distribution for linear variables, we can use the wrapping or the projection approach to obtain circular distributions. Then from an $n-$dimensional realization of a linear process, we can obtain an $n-$dimensional realization of a circular one.

More precisely, let $\mathbf{Y}(\mathbf{s}) \in \mathbb{R}^p$, with $\mathbf{s} \in \mathcal{S} \subset \mathbb{R}^d$, be a $p-$variate stochastic process, defined over a $d-$dimensional domain, and suppose that an $n-$dimensional realization of the process $\mathbf{Y}(\mathbf{s})$, $\mathbf{y}$, has pdf $f_{\mathbf{Y}}(\cdot|\boldsymbol{\psi})$.

Wrapped circular process Let $p = 1$ and let $\mathbf{y} = (y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n))$ be the $n-$dimensional realization of $\mathbf{Y}(\mathbf{s})$. If we apply the transformation (2.1) to each component of $\mathbf{y}$, we obtain a vector of dimension n of wrapped circular variables: $\boldsymbol{\theta} = (\theta(\mathbf{s}_1), \ldots, \theta(\mathbf{s}_n))'$, with pdf (2.3). Jona Lasinio *et al.* (2012) show that the vector $\boldsymbol{\theta}$ is the $n-$dimensional realization of the circular process

$$\Theta(\mathbf{s}) = \mathbf{Y}(\mathbf{s}) \bmod 2\pi,$$

with $\mathbf{Y}(\mathbf{s}) = \Theta(\mathbf{s}) + 2\pi\mathbf{K}(\mathbf{s})$.

The evaluation of the $n-$dimensional circular distribution requires the evaluation of the infinite sum, equation (2.3). Here again, we can bypass the problem by introducing the realization of the (latent) discrete process $K(\mathbf{s})$, i.e. $\mathbf{k} = (k(\mathbf{s}_1), \ldots, k(\mathbf{s}_n))'$, and working with the joint density of $\boldsymbol{\theta}, \mathbf{k}|\boldsymbol{\psi}$.

Projected circular process Let $p = 2$, then $\mathbf{Y}(\mathbf{s}) = (Y_1(\mathbf{s}), Y_2(\mathbf{s}))'$ is a bivariate process and $\mathbf{y} = (\mathbf{y}(\mathbf{s}_1), \ldots, \mathbf{y}(\mathbf{s}_n))'$, its finite-dimensional realization, is a vector of bivariate

variables, i.e. $\mathbf{y}(\mathbf{s}_i) = (y_1(\mathbf{s}_i), y_2(\mathbf{s}_i))' \in \mathbb{R}^2$. The projected circular process is obtained as

$$\Theta(\mathbf{s}) = \text{atan}^* \left(\frac{Y_2(\mathbf{s})}{Y_1(\mathbf{s})} \right), \tag{3.1}$$

i.e. we apply the transformation (3.1) to the process $\mathbf{Y}(\mathbf{s})$ (Wang and Gelfand, 2014). The finite dimensional realization of the circular process is $\boldsymbol{\theta} = (\theta(\mathbf{s}_1), \ldots, \theta(\mathbf{s}_n))'$, where

$$\theta(\mathbf{s}_i) = \text{atan}^* \left(\frac{y_2(\mathbf{s}_i)}{y_1(\mathbf{s}_i)} \right), i = 1, \ldots, n.$$

The pdf of $\boldsymbol{\theta}|\boldsymbol{\psi}$ is given by (2.5).

Let $R(\mathbf{s}) = ||\mathbf{Y}(\mathbf{s})|| \in \mathbb{R}^+$ be a (latent) process and $\mathbf{r} = (r(\mathbf{s}_1), \ldots, r(\mathbf{s}_n))'$ be its $n-$dimensional realization. Then, instead of $\boldsymbol{\theta}$, we can work with the joint density of $\boldsymbol{\theta}, \mathbf{r}|\boldsymbol{\psi}$, that is the integrand in (2.8).

Bibliography

Gelfand, A., Diggle, P., Fuentes, M., and Guttorp, P. (2010). *Handbook of Spatial Statistics*. Chapman and Hall.

Jona Lasinio, G., Gelfand, A., and Jona Lasinio, M. (2012). Spatial analysis of wave direction data using wrapped Gaussian processes. *Annals of Applied Statistics*, **6**(4), 1478–1498.

Wang, F. and Gelfand, A. E. (2014). Modeling space and space-time directional data using projected Gaussian processes. *Journal of the American Statistical Association*, **109**(508), 1565–1580.

TEST
DOI 10.1007/s11749-015-0458-y

ORIGINAL PAPER

Spatio-temporal circular models with non-separable covariance structure

Gianluca Mastrantonio[1] · **Giovanna Jona Lasinio**[2] · **Alan E. Gelfand**[3]

Received: 7 September 2014 / Accepted: 11 June 2015

Abstract Circular data arise in many areas of application. Recently, there has been interest in looking at circular data collected separately over time and over space. Here, we extend some of this work to the spatio-temporal setting, introducing space–time dependence. We accommodate covariates, implement full kriging and forecasting, and also allow for a nugget which can be time dependent. We work within a Bayesian framework, introducing suitable latent variables to facilitate Markov chain Monte Carlo model fitting. The Bayesian framework enables us to implement full inference, obtaining predictive distributions for kriging and forecasting. We offer comparison between the less flexible but more interpretable wrapped Gaussian process and the more flexible but less interpretable projected Gaussian process. We do this illustratively using both simulated data and data from computer model output for wave directions in the Adriatic Sea off the coast of Italy.

Keywords Average prediction error · Continuous ranked probability score · Kriging · Markov chain Monte Carlo · Projected distribution · Wrapped distribution

Electronic supplementary material The online version of this article (doi:10.1007/s11749-015-0458-y) contains supplementary material, which is available to authorized users.

✉ Gianluca Mastrantonio
gianluca.mastrantonio@yahoo.it; gianluca.mastrantonio@uniroma3.it

[1] Roma Tre University, Via Silvio D'Amico 77, 00145 Rome, Italy

[2] Sapienza University of Rome, P.le Aldo Moro 5, 00185 Rome, Italy

[3] Duke University, 223-A Old Chemistry Building, Box 90251, Durham, NC 27708-0251, USA

Published online: 30 June 2015

1 Introduction

Circular data, i.e., observations with support on the unit circle, arise in many contexts. Examples include natural directions, such as wind directions (meteorology), animal movement directions (biology) and rock fracture orientations (geology). Another type of circular data arises by wrapping periodic time data with period L (say, day or week) onto a circle with circumference L and then rescaling the circumference to 2π, that of the unit circle. Two-dimensional directional data may be observed in space and time, along with linear variables, as in marine studies where, for example wave heights and directions are jointly observed, or in atmospheric modeling where wind fields are represented by wind intensity as well as direction. Due to the restriction of the domain to the circle, analysis of circular data must be treated differently from linear data. Customary statistical summaries are replaced with their circular counterparts. For a discussion of inference with circular variables see, e.g., Fisher (1996), Mardia and Jupp (1999), Jammalamadaka and SenGupta (2001) or the recent paper by Lee (2010).

The contribution of this paper is to extend recent spatial and spatio-temporal circular data models. In particular, Jona Lasinio et al. (2012) consider the use of the wrapped normal approach by developing the wrapped Gaussian process while Wang and Gelfand (2014) consider the use of the projected normal approach by developing the projected Gaussian process. Here we: (i) extend both processes to the spatio-temporal setting, introducing space–time dependence; (ii) introduce space and time varying covariate information; (iii) show how to implement fully model-based kriging and forecasting; (iv) allow a nugget which can be time dependent, and (v) provide an extensive comparison between the more sparsely parametrized wrapped Gaussian process with the more flexible projected Gaussian process. We do this illustratively using simulation, as a proof of concept, as well as with data in the form of computer model output for wave directions in the Adriatic Sea off the coast of Italy. The models are fitted under a Bayesian framework, introducing suitable latent variables, enabling full inference.

Modeling of circular data has progressed well beyond the i.i.d. case. Examples include linear models (Harrison and Kanji 1988; Fisher 1996; Kato and Shimizu 2008), linear models in a Bayesian context (Guttorp and Lockhart 1988; Damien and Walker 1999), models for circular time series (Breckling 1989; Fisher and Lee 1992; Coles 1998; Holtzman et al. 2006; Ravindran and Ghosh 2011), and hidden Markov models to address classification issues (Lagona and Picone 2011; Bulla et al. 2012; Mastrantonio et al. 2015). In Kato (2010) a Markov process for circular variables is presented. Jona Lasinio et al. (2012) consider a spatial wrapped Gaussian process. Wang and Gelfand (2013) explore the general projected normal model while in Wang and Gelfand (2014) Bayesian analysis of space–time circular data is developed using projected Gaussian processes. In Wang et al. (2015), directional wave data is modeled jointly with linear wave height data.

The format of the remainder of the paper is as follows. In Sect. 2, we review the wrapping approach and offer a non-separable space–time model for circular data. In Sect. 3, an analogous model is presented using the projected normal process. Section 4 presents several simulation examples giving insight into the inferential performance

of the models, while Sect. 5 analyzes the behavior of the models for wave directions. Section 6 extends the modeling approach to enable space–time varying covariates reflecting sea state at a location and time. Some concluding remarks are provided in Sect. 7. Implementation details, further simulated examples and more details on the real data application are available in the Supplementary Online Material, Sections S1, S2 and S3.

2 A brief review of the wrapped modeling approach

Let $Y \in \mathbb{R}$ be a random variable on the real line and let $g(y)$ and $G(y)$ be, respectively, its probability density function and cumulative distribution function. The random variable

$$X = Y \bmod 2\pi, \quad 0 \leq X < 2\pi$$

is the wrapped version of Y having period 2π. The probability density function of X, $f(x)$, is obtained by wrapping the probability density function of Y, $g(y)$, around a circle of unit radius via the transformation $Y = X + 2\pi K$, with $K \in \mathbb{Z} \equiv \{0, \pm 1, \pm 2, \ldots\}$, and takes the form

$$f(x) = \sum_{k=-\infty}^{\infty} g(x + 2\pi k), \tag{1}$$

that is, a doubly infinite sum.

Equation (1) shows that $g(x + 2\pi k)$ is the joint distribution of (X, K). Hence, the marginal distribution of K is $P(K = k) = \int_0^{2\pi} g(x + 2\pi k)dx$, the conditional distributions $P(K = k|X = x) = g(x+2\pi k)/\sum_{j=-\infty}^{\infty} g(x+2\pi j)$ and the distribution of $X|K = k$ is $g(x+2\pi k)/\int_0^{2\pi} g(x+2\pi k)dx$. The introduction of K as latent variable facilitates model fitting (Jona Lasinio et al. 2012).

Following Coles (1998), we can extend the wrapping approach to multivariate distributions. Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p) \sim g(\cdot)$, with $g(\cdot)$ a p-variate distribution on $\mathbb{R}^p$ indexed by say $\boldsymbol{\theta}$ and let $\mathbf{K} = (K_1, K_2, \ldots, K_p)$ be such that $\mathbf{Y} = \mathbf{X} + 2\pi\mathbf{K}$. Then the distribution of $\mathbf{X}$ is

$$f(\mathbf{X}) = \sum_{k_1=-\infty}^{+\infty} \sum_{k_2=-\infty}^{+\infty} \ldots \sum_{k_p=-\infty}^{+\infty} g(\mathbf{X} + 2\pi\mathbf{K}). \tag{2}$$

From (2) we see, as in the univariate case, that the joint density of $(\mathbf{X}, \mathbf{K})$ is $g(\mathbf{X} + 2\pi\mathbf{K})$. If $g(\cdot; \boldsymbol{\theta})$ is a p-variate normal density, with $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{X}$ has a p-variate wrapped normal distribution with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Here, we introduce the latent random vector of winding numbers $\mathbf{K}$ to facilitate model fitting. Mardia and Jupp (1999) point out that only a few values of K are needed to obtain a reasonable approximation of the wrapped distribution and Jona Lasinio et al. (2012) show, when $g(\cdot; \boldsymbol{\theta})$ is Gaussian, how to choose the set of values of K based on the variance of the associated conditional distribution.

Let $Y(\mathbf{s})$ be a Gaussian process (GP) with $\mathbf{s} \in \mathbb{R}^2$, mean function $\mu(\mathbf{s})$ and covariance function say $\sigma^2\rho(||\mathbf{s}_i - \mathbf{s}_j||; \boldsymbol{\psi})$, where $\boldsymbol{\psi}$ is a set of parameters. For a set of locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$, $\mathbf{Y} = (Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots, Y(\mathbf{s}_n)) \sim N(\boldsymbol{\mu}, \sigma^2 C(\boldsymbol{\psi}))$, where $\boldsymbol{\mu} = (\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_n))$ and $C(\boldsymbol{\psi})_{ij} = \rho(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\psi})$. As a consequence $\mathbf{X} = (X(\mathbf{s}_1), X(\mathbf{s}_2), \ldots, X(\mathbf{s}_n)) \sim WrapN(\boldsymbol{\mu}, \sigma^2\mathbf{C}(\boldsymbol{\psi}))$ (Jona Lasinio et al. 2012), where $WrapN(\cdot, \cdot)$ indicates the wrapped normal distribution.

2.1 Space–time model specification and model fitting

Turning to space and time, suppose we seek $\{X(\mathbf{s}, t) \in [0, 2\pi), \mathbf{s} \in \mathcal{S} \subseteq \mathbb{R}^2, t \in \mathcal{T} \subseteq \mathbb{Z}^+\}$, a spatio-temporal process of angular variables. We can model $X(\mathbf{s}, t)$ as a spatio-temporal wrapped Gaussian process through its linear counterpart $Y(\mathbf{s}, t)$, extending the above approach. We assume that the linear process is a spatio-temporal Gaussian process having non-separable covariance structure with variance σ^2 and the stationary correlation function due to Gneiting [see equation (14) in Gneiting 2002]:

$$\text{Cor}(Y(\mathbf{s}, t), Y(\mathbf{s}', t')) \equiv \rho(\mathbf{h}, u) = \frac{1}{(a|u|^{2\alpha} + 1)^{\tau}} \exp\left(-\frac{c\|\mathbf{h}\|^{2\gamma}}{(a|u|^{2\alpha} + 1)^{\beta\gamma}}\right), \quad (3)$$

where $(\mathbf{h}, u) \in \mathbb{R}^d \times \mathbb{R}$, $\mathbf{h} = \mathbf{s} - \mathbf{s}'$ and $u = t - t'$. Here $d = 2$, a and c are non-negative scaling parameters for time and space, respectively. The smoothness parameters α and γ take values in (0, 1], the space–time interaction parameter β is in [0, 1], and $\tau \geq d/2 = 1$ is, in fact, fixed at 1 following Gneiting (2002). Attractively, as β decreases toward zero, we tend to separability in space and time.

We write the linear GP $Y(\mathbf{s}, t)$ as $Y(\mathbf{s}, t) = \mu_Y + \omega_Y(\mathbf{s}, t) + \tilde{\varepsilon}_Y(\mathbf{s}, t)$ where μ_Y is a constant mean function, $\omega_Y(\mathbf{s}, t)$ is a zero mean space–time GP with covariance function $\sigma^2\rho(\mathbf{h}, u)$, and $\tilde{\varepsilon}(\mathbf{s}, t) \overset{iid}{\sim} N(0, \phi_Y^2)$, i.e., is pure error. It is convenient to work with the marginalized model where we integrate over all of the $\omega_Y(\mathbf{s}, t)$, see Banerjee et al. (2014). That is,

$$Y(\mathbf{s}, t) = \mu_Y + \varepsilon_Y(\mathbf{s}, t). \quad (4)$$

Then, $\varepsilon(\mathbf{s}, t)$ is a zero mean Gaussian process with covariance function

$$\text{Cov}(\varepsilon_Y(\mathbf{s}_i, t_j), \varepsilon_Y(\mathbf{s}_{i'}, t_{j'})) = \sigma_Y^2 \text{Cor}(\mathbf{h}_{i,i'}, u_{j,j'}) + \phi_Y^2 1_{(i=i')} 1_{(j=j')}.$$

To complete the model specification, we need to specify prior distributions. We suggest the following choices. Since a and c are positive, a and $c \sim G(\cdot, \cdot)$ where $G(\cdot, \cdot)$ denotes a gamma distribution. Since α, β, and γ are bounded between 0 and 1, we adopt a beta distribution ($B(\cdot, \cdot)$). Priors for the variances and the mean direction are given the usual normal-inverse gamma form, i.e., $\sigma_Y^2, \phi_Y^2 \sim IG(\cdot, \cdot)$, where $IG(\cdot, \cdot)$ denotes the inverse gamma, and $\mu_y \sim WrapN(\cdot, \cdot)$. In the sequel, this model will be denoted by *WN*.

2.2 Kriging and forecasting

We clarify prediction of the process at a new location and time, say $(\mathbf{s}_0, t_0)$, given what we have observed. We provide a full predictive distribution, extending Jona Lasinio et al. (2012) who only provide a posterior mean. Let $\mathcal{D} \subset \mathbb{R}^2 \times \mathbb{Z}^+$ be the set of n observed points. Let $\mathbf{X} = \{X(\mathbf{s}, t), (\mathbf{s}, t) \in \mathcal{D}\}$ be the vector of observed circular variables. Let $\mathbf{Y} = \{Y(\mathbf{s}, t), (\mathbf{s}, t) \in \mathcal{D}\}$ be the associated linear ones and let $\mathbf{K} = \{K(\mathbf{s}, t), (\mathbf{s}, t) \in \mathcal{D}\}$ be the associated vector of winding numbers. The predictive distribution we seek is $g(X(\mathbf{s}_0, t_0)|\mathbf{X})$. We use usual composition sampling within Markov chain Monte Carlo (MCMC) to obtain samples from it. Here, again we move from the circular process to the linear one, i.e., a sample from the distribution of $Y(\mathbf{s}_0, t_0)|\mathbf{X}$ can be considered as a sample from $X(\mathbf{s}_0, t_0)$, $K(\mathbf{s}_0, t_0)|\mathbf{X}$. If we let $\boldsymbol{\Psi}_Y$ be the vector of all parameters, we can write

$$g(X(\mathbf{s}_0, t_0), K(\mathbf{s}_0, t_0)|\mathbf{X}) = \sum_{\mathbf{K}\in\mathbb{Z}^n} \int_{\boldsymbol{\Psi}_Y} g(X(\mathbf{s}_0, t_0), K(\mathbf{s}_0, t_0)|\boldsymbol{\Psi}_Y, \boldsymbol{K}, \mathbf{X}) g(\boldsymbol{\Psi}_Y, \boldsymbol{K}|\mathbf{X}) d\boldsymbol{\Psi}_Y.$$

So, suppose, for each posterior sample of $\mathbf{K}$ and $\boldsymbol{\Psi}_Y$ in $\{\mathbf{K}_l^*, \boldsymbol{\Psi}_{Y,l}^*, l = 1, 2, \ldots, L\}$ we generate a value from the distribution of $X(\mathbf{s}_0, t_0)$, $K(\mathbf{s}_0, t_0)|\, \boldsymbol{\Psi}_Y, \boldsymbol{K}, \mathbf{X}$. Then, we will obtain the set of posterior samples $\{X_l^*(\mathbf{s}_0, t_0), K_l^*(\mathbf{s}_0, t_0), l = 1, 2, \ldots, L\}$ from $X(\mathbf{s}_0, t_0)$, $K(\mathbf{s}_0, t_0)|\mathbf{X}$. If, we retain the set $\{X_l^*(\mathbf{s}_0, t_0), l = 1, 2, \ldots, L\}$, we will have samples from the desired predictive distribution.

Therefore, we need to sample from the distribution of $X(\mathbf{s}_0, t_0)$, $K(\mathbf{s}_0, t_0)|$ $\boldsymbol{\Psi}_Y, \boldsymbol{K}, \mathbf{X}$ or equivalently $Y(\mathbf{s}_0, t_0)|\mathbf{Y}, \boldsymbol{\Psi}_Y$. Let $\mathbf{1}_m$ be the $m \times 1$ vector of 1s, let $\boldsymbol{C}_\mathbf{Y}$ be the correlation matrix of $\boldsymbol{Y}$, and let $\boldsymbol{C}_{\mathbf{Y},Y(\mathbf{s}_0,t_0)}$ be the correlation vector between $\mathbf{Y}$ and $Y(\mathbf{s}_0, t_0)$. Then, the joint distribution of $Y(\mathbf{s}_0, t_0)$, $\mathbf{Y}|\boldsymbol{\Psi}_Y$ is

$$\begin{pmatrix} Y(\mathbf{s}_0, t_0) \\ \mathbf{Y} \end{pmatrix} |\boldsymbol{\Psi}_Y \sim N\left(\begin{pmatrix} \mu_Y \\ \mu_Y \mathbf{1}_n \end{pmatrix}, \sigma_Y^2 \begin{pmatrix} 1 & \boldsymbol{C}'_{\mathbf{Y},Y(\mathbf{s}_0,t_0)} \\ \boldsymbol{C}_{\mathbf{Y},Y(\mathbf{s}_0,t_0)} & \mathbf{C}_\mathbf{Y} \end{pmatrix} + \phi_Y^2 \mathbf{I}_{n+1}\right).$$

As a result, the conditional distribution of $Y(\mathbf{s}_0, t_0)|\mathbf{Y}, \boldsymbol{\Psi}_Y$ is Gaussian with mean

$$\mathrm{M}_{Y(\mathbf{s}_0,t_0)} = \mu_Y + \sigma_Y^2 \boldsymbol{C}'_{\mathbf{Y},Y(\mathbf{s}_0,t_0)} \left(\sigma_Y^2 \mathbf{C}_\mathbf{Y} + \phi_Y^2 \mathbf{I}_n\right)^{-1} (\mathbf{Y} - \mu_Y \mathbf{1}_n)$$

and variance

$$\mathrm{V}_{Y(\mathbf{s}_0,t_0)} = \sigma_Y^2 + \phi_Y^2 - \sigma_Y^2 \boldsymbol{C}'_{\mathbf{Y},Y(\mathbf{s}_0,t_0)} \left(\sigma_Y^2 \mathbf{C}_\mathbf{Y} + \phi_Y^2 \mathbf{I}_n\right)^{-1} \sigma_Y^2 \boldsymbol{C}_{\mathbf{Y},Y(\mathbf{s}_0,t_0)}.$$

Finally, suppose, for each posterior sample, we simulate $Y_l^*(\mathbf{s}_0, t_0)$ from $N(\mathrm{M}^*_{Y(\mathbf{s}_0,t_0),l}, \mathrm{V}^*_{Y(\mathbf{s}_0,t_0),l})$, where $\mathrm{M}^*_{Y(\mathbf{s}_0,t_0),l}$ and $\mathrm{V}^*_{Y(\mathbf{s}_0,t_0),l}$ are $\mathrm{M}_{Y(\mathbf{s}_0,t_0)}$ and $\mathrm{V}_{Y(\mathbf{s}_0,t_0)}$ computed with the lth sample. then, $X_l^*(\mathbf{s}_0, t_0) = Y_l^*(\mathbf{s}_0, t_0) \bmod 2\pi$ is a posterior sample from the predictive distribution.

3 The spatio-temporal projected normal process

Let (Z_1, Z_2) be a bivariate vector normally distributed with mean $\boldsymbol{\mu}_Z = (\mu_{Z_1}, \mu_{Z_2})$ and covariance matrix

$$\tilde{\mathbf{V}} = \begin{pmatrix} \sigma^2_{Z_1} & \sigma_{Z_1}\sigma_{Z_2}\rho_z \\ \sigma_{Z_1}\sigma_{Z_2}\rho_z & \sigma^2_{Z_2} \end{pmatrix}.$$

The vector $\mathbf{Z}$ is mapped into an angular variable Θ by the transformation $\Theta = \text{atan}^*(Z_2/Z_1)$, where the function $\text{atan}^*(S/C)$ is defined as $\text{atan}(S/C)$ if $C > 0$ and $S \geq 0$, $\pi/2$ if $C = 0$ and $S > 0$, $\text{atan}(S/C) + \pi$ if $C < 0$, $\text{atan}(S/C) + 2\pi$ if $C \geq 0$ and $S < 0$, undefined if $C = S = 0$. Θ is referred to as a projected normal random variable (Mardia 1972, p. 52) with parameters $\boldsymbol{\mu}_Z$ and $\tilde{\mathbf{V}}$. Wang and Gelfand (2013) note that the distribution of Θ does not change if we multiply (Z_1, Z_2) by a positive constant, so, following their lead, to identify the distribution we set $\sigma^2_{Z_2} = 1$ and the covariance matrix becomes

$$\mathbf{V} = \begin{pmatrix} \sigma^2_{Z_1} & \sigma_{Z_1}\rho_z \\ \sigma_{Z_1}\rho_z & 1 \end{pmatrix}.$$

Again, it is convenient to introduce a latent variable. Here, it is $R = ||\mathbf{Z}||$, obtaining the joint density of (Θ, R):

$$(2\pi)^{-1}|\mathbf{V}|^{1/2} \exp\left(-\frac{(r(\cos\theta, \sin\theta)' - \boldsymbol{\mu}_Z)'\mathbf{V}^{-1}(r(\cos\theta, \sin\theta)' - \boldsymbol{\mu}_Z)}{2}\right) r.$$

We can move back and forth between the linear variables and the pair (Θ, R) using the transformation $Z_1 = R\cos\Theta$, $Z_2 = R\sin\Theta$ and the equation $\Theta = \text{atan}^*(Z_2/Z_1)$.

Consider a bivariate spatio-temporal process $\mathbf{Z}(\mathbf{s}, t) = (Z_1(\mathbf{s}, t), Z_2(\mathbf{s}, t))$ with constant mean $\boldsymbol{\mu}_Z$ and cross covariance function $\text{C}\left(\mathbf{Z}(\mathbf{s}_i, t_j), \mathbf{Z}(\mathbf{s}_{i'}, t_{j'})\right) = \text{Cor}(\mathbf{s}_i - \mathbf{s}_{i'}, t_j - t_{j'})\mathbf{V}$ where $\text{Cor}(\cdot, \cdot)$ is a given space–time correlation function and $\mathbf{V}$ is as above. Then the circular process $\Theta(\mathbf{s}, t)$ induced by $\mathbf{Z}(\mathbf{s}, t)$ with the atan* transformation is a projected Gaussian process with mean $\boldsymbol{\mu}_Z$ and covariance function induced by $\text{C}\left(\mathbf{Z}(\mathbf{s}_i, t_j), \mathbf{Z}(\mathbf{s}_{i'}, t_{j'})\right)$. More details on the properties of the process can be found in Wang and Gelfand (2014). Now, latent $R(\mathbf{s}, t)$s are introduced to facilitate model fitting.

3.1 Model specification and model fitting

We define the bivariate linear process as

$$Z_\ell(\mathbf{s}, t) = \mu_{Z_\ell} + \omega_{Z_\ell}(\mathbf{s}, t) + \tilde{\varepsilon}_{Z_\ell}(\mathbf{s}, t), \quad \ell = 1, 2, \tag{5}$$

where $\boldsymbol{\mu}_Z = (\mu_{Z_1}, \mu_{Z_2})'$ is the mean level, $\boldsymbol{\omega}_Z(\mathbf{s}, t) = (\omega_{Z_1}(\mathbf{s}, t), \omega_{Z_2}(\mathbf{s}, t))'$ is a bivariate Gaussian process with zero mean and covariance $\text{Cov}(\boldsymbol{\omega}_Z(\mathbf{s}_i, t_j), \boldsymbol{\omega}_Z(\mathbf{s}_{i'}, t_{j'})) = \text{Cor}(\mathbf{h}_{i,i'}, u_{j,j'})\mathbf{V}$ where $\text{Cor}(\mathbf{h}_{i,i'}, u_{j,j'})$ is defined in (3). Finally,

$\tilde{\boldsymbol{\varepsilon}}_Z(\mathbf{s},t) = (\tilde{\varepsilon}_{Z_1}(\mathbf{s},t), \tilde{\varepsilon}_{Z_2}(\mathbf{s},t))$ is bivariate pure error with zero mean, independent components, and variance ϕ_Z^2. Marginalizing over the ω process in (5) yields

$$Z_\ell(\mathbf{s},t) = \mu_{Z_\ell} + \varepsilon_{Z_\ell}(\mathbf{s},t), \quad \ell = 1,2,$$

where $\boldsymbol{\varepsilon}_Z(\mathbf{s},t)$ is a mean zero bivariate Gaussian process with covariance function $\text{Cov}(\boldsymbol{\varepsilon}_Z(\mathbf{s}_i,t_j), \boldsymbol{\varepsilon}_Z(\mathbf{s}_{i'},t_{j'}) = \text{Cor}(\mathbf{h}_{i,i'}, u_{j,j'})\mathbf{V} + \phi_Z^2\mathbf{I}_2 1_{(i=i')}1_{(j=j')}$.

$\Theta(\mathbf{s},t) = \text{atan}^*(Z_2(\mathbf{s},t)/Z_1(\mathbf{s},t))$ is a circular process and, as in the WN setting, correlation between the circular variables is induced by the Gneiting spatio-temporal correlation function. To specify the prior distributions for μ_{Z_1}, μ_{Z_2}, $\sigma_{Z_1}^2$ and ϕ_Z^2, we adopt the customary normal-inverse gamma specification. That is, $\mu_{Z_1}, \mu_{Z_2} \sim N(\cdot,\cdot)$, $\sigma_{Z_1}^2, \phi_Z^2 \sim IG(\cdot,\cdot)$ while, since ρ_Z is a correlation parameter, we adopt a truncated normal: $\rho_Z \sim N(\cdot,\cdot)I(-1,1)$. In the sequel, this model will be denoted by *PN*.

We seek the predictive distribution at an unobserved location and time, $(\mathbf{s}_0,t_0)$. Let $\boldsymbol{\Theta}$ be the vector of observed circular values and $\mathbf{Z} = \{\mathbf{Z}(\mathbf{s},t), (\mathbf{s},t) \in \mathcal{D}\}$ be the associated linear ones. Let $\mathbf{Z}(\mathbf{s}_0,t_0) = (Z_1(\mathbf{s}_0,t_0), Z_2(\mathbf{s}_0,t_0))'$, $\mathbf{R} = \{R(\mathbf{s},t), (\mathbf{s},t) \in \mathcal{D}\}$ and let $\boldsymbol{\Psi}_Z$ be all the parameters of the projected model.

Specifically, the predictive distribution we seek is $\Theta(\mathbf{s}_0,t_0)|\boldsymbol{\Theta}$. If we sample from the distribution of $\mathbf{Z}(\mathbf{s}_0,t_0)|\boldsymbol{\Theta}$ then $\Theta(\mathbf{s}_0,t_0) = \text{atan}^*(Z_2(\mathbf{s}_0,t_0)/Z_1(\mathbf{s}_0,t_0))$ is a sample from the desired predictive distribution. We have that

$$g(\mathbf{Z}(\mathbf{s}_0,t_0)|\boldsymbol{\Theta}) = \int_{\mathbf{R}}\int_{\boldsymbol{\Psi}_Z} g(\mathbf{Z}(\mathbf{s}_0,t_0)|\boldsymbol{\Psi}_Z,\mathbf{R},\boldsymbol{\Theta})g(\boldsymbol{\Psi}_Z,\mathbf{R}|\boldsymbol{\Theta})\text{d}\boldsymbol{\Psi}_Z\text{d}\mathbf{R}.$$

So, we need to obtain $g(\mathbf{Z}(\mathbf{s}_0,t_0)|\boldsymbol{\Psi}_Z,\mathbf{R},\boldsymbol{\Theta})$ and be able to sample from it. We start from the joint distribution of $\mathbf{Z}(\mathbf{s}_0,t_0), \mathbf{Z}|\boldsymbol{\Psi}_Z$:

$$\begin{pmatrix}\mathbf{Z}(\mathbf{s}_0,t_0)\\ \mathbf{Z}\end{pmatrix}|\boldsymbol{\Psi}_Z$$
$$\sim N\left(\begin{pmatrix}\boldsymbol{\mu}_Z\\ \mathbf{1}_n\otimes\boldsymbol{\mu}_Z\end{pmatrix}, \begin{pmatrix}1 & \boldsymbol{C}'_{\mathbf{Z},\mathbf{Z}(\mathbf{s}_0,t_0)}\\ \boldsymbol{C}_{\mathbf{Z},\mathbf{Z}(\mathbf{s}_0,t_0)} & \mathbf{C}_\mathbf{Z}\end{pmatrix}\otimes\mathbf{V} + \phi_Z^2\mathbf{I}_{2n+2}\right),$$

where $\mathbf{C}_\mathbf{Z}$ and $\boldsymbol{C}_{\mathbf{Z},\mathbf{Z}(\mathbf{s}_0,t_0)}$ are the analogous of $\mathbf{C}_\mathbf{Y}$ and $\boldsymbol{C}_{\mathbf{Y},Y(\mathbf{s}_0,t_0)}$ for the process $\mathbf{Z}(\mathbf{s},t)$. The conditional distribution of $\mathbf{Z}(\mathbf{s}_0,t_0)|\mathbf{Z},\boldsymbol{\Psi}_Z$ (equivalently $\mathbf{Z}(\mathbf{s}_0,t_0)|\boldsymbol{\Theta},\mathbf{R},\boldsymbol{\Psi}_Z$) is bivariate normal with mean

$$\text{M}_{\mathbf{Z}(\mathbf{s}_0,t_0)} = \boldsymbol{\mu}_Z + \boldsymbol{C}'_{\mathbf{Z},\mathbf{Z}(\mathbf{s}_0,t_0)}\otimes\mathbf{V}\left(\mathbf{C}_\mathbf{Z}\otimes\mathbf{V} + \phi_Z^2\mathbf{I}_{2n}\right)^{-1}(\mathbf{Z} - \mathbf{1}_n\otimes\boldsymbol{\mu}_Z)$$

and variance

$$\text{V}_{\mathbf{Z}(\mathbf{s}_0,t_0)} = \mathbf{V} - \boldsymbol{C}'_{\mathbf{Z},\mathbf{Z}(\mathbf{s}_0,t_0)}\otimes\mathbf{V}\left(\mathbf{C}_\mathbf{Z}\otimes\mathbf{V} + \phi_Z^2\mathbf{I}_{2n}\right)^{-1}\boldsymbol{C}_{\mathbf{Z},\mathbf{Z}(\mathbf{s}_0,t_0)}\otimes\mathbf{V}.$$

Using the posterior samples $\{\mathbf{R}_l^*, \boldsymbol{\Psi}_{Z,l}^*, l = 1,2,\ldots,L\}$ we can collect samples of $\Theta_l^*(\mathbf{s}_0,t_0)$ from its posterior predictive distribution.

4 Simulated examples

The Gneiting correlation function (3) has not been widely investigated within a Bayesian framework. The aim of this simulation study is essentially to provide a proof of concept. If space–time dependence, captured through the Gneiting correlation function, is driving an observed spatio-temporal circular dataset, can we learn about this dependence and can we demonstrate improved predictive performance by incorporating it in our modeling? We explore several different choices of parameters in (3).

For each proposed model, we simulated 48 datasets with $n = 240$ (20 locations and 12 time points) with spatial coordinates uniformly generated in $[0, 10] \times [0, 10]$. 24 datasets for the WN model were simulated from all possible combinations of $(a, c) = \{(1, 0.2), (0.2, 1)\}$, $\beta = \{0, 0.5, 1\}$, $\alpha = \{0.5, 0.8\}$, $\gamma = \{0.5, 0.8\}$ and $(\mu_Y, \sigma^2_Y, \phi^2_y) = (\pi, 0.1, 0.01)$. In the other 24 datasets we used the same combinations of correlation parameters but with $(\mu_Y, \sigma^2_Y, \phi^2_y) = (\pi, 1, 0.1)$. The datasets cover a wide range of situations in terms of spatio-temporal correlation: strong spatial correlation with weak temporal correlation ($(a, c) = (1, 0.2)$), weak spatial correlation with strong temporal correlation ($(a, c) = (0.2, 1)$), fully separable spatio-temporal correlation ($\beta = 0$), non-separable ($\beta = \{0.5, 0.9\}$) and two levels for the smoothing parameters. The difference between the two collections of 24 datasets is that the first 24 have smaller circular variance than the remaining ones, where the circular variance was computed as one minus the mean resultant length divided by the sample size (Jammalamadaka and SenGupta 2001, p. 15).

The projected normal datasets were built according to the same rationale adopted for the wrapped normal, i.e., we built 24 datasets with small circular variance and 24 datasets with large circular variance. We simulated from unimodal projected distributions adopting the following sets of parameters:

- all possible combinations of $(a, c) = \{(1, 0.2), (0.2, 1)\}$, $\beta = \{0, 0.5, 1\}$, $\alpha = \{0.5, 0.8\}$, $\gamma = \{0.5, 0.8\}$ with $(\mu_{Z_1}, \mu_{Z_2}, \sigma^2_{Z_1}, \rho_Z, \phi^2_Z) = (2.5, 2.5, 1, 0, 0.01)$ which yields a circular variance close to the WN examples with $\sigma^2_Y = 0.1$.
- all possible combinations of $(a, c) = \{(1, 0.2), (0.2, 1)\}$, $\beta = \{0, 0.5, 1\}$, $\alpha = \{0.5, 0.8\}$, $\gamma = \{0.5, 0.8\}$ with $(\mu_{Z_1}, \mu_{Z_2}, \sigma^2_{Z_1}, \rho_Z, \phi^2_Z) = (0.85, 0.85, 1, 0, 0.1)$ which, again, yields a circular variance close to the WN examples with $\sigma^2_Y = 1$.

The parameters for the prior distributions were chosen so that the priors were centered on the "true" values used to simulate each dataset:

- correlation parameters: $a = 0.2 \Rightarrow a \sim G(2, 5)$, $a = 1 \Rightarrow a \sim G(5, 4)$, $c = 0.2 \Rightarrow c \sim G(2, 5)$, $c = 1 \Rightarrow c \sim G(5, 4)$, $\alpha = 0.5 \Rightarrow \alpha \sim B(5, 5)$, $\alpha = 0.8 \Rightarrow \alpha \sim B(6, 1.5)$, $\beta = 0 \Rightarrow \beta \sim B(1, 4)$, $\beta = 0.5 \Rightarrow \beta \sim B(5, 5)$, $\beta = 0.9 \Rightarrow \beta \sim B(6, 1.5)$, $\gamma = 0.5 \Rightarrow \gamma \sim B(5, 5)$, $\gamma = 0.8 \Rightarrow \gamma \sim B(6, 1.5)$;
- parameters of the WN model: $\mu_Y = 5 \Rightarrow \mu_Y \sim WN(\pi, 5)$, $\sigma^2_Y = 0.1 \Rightarrow \sigma^2_Y \sim IG(4.5, 0.55)$, $\sigma^2_Y = 1 \Rightarrow \sigma^2_Y \sim IG(2.01, 4.01)$, $\phi^2_Y = 0.01 \Rightarrow \phi^2_Y \sim IG(2.001, 0.03)$, $\phi^2_Y = 0.1 \Rightarrow \phi^2_Y \sim IG(4.5, 0.55)$;
- parameters of the PN model: $\mu_{Z_1} = 2.5 \Rightarrow \mu_{Z_1} \sim N(2.5, 5)$, $\mu_{Z_1} = 0.85 \Rightarrow \mu_{Z_1} \sim N(0.85, 5)$ $\mu_{Z_2} = 2.5 \Rightarrow \mu_{Z_2} \sim N(2.5, 5)$, $\mu_{Z_2} = 0.85 \Rightarrow$

$\mu_{Z_2} \sim N(0.85, 5)$, $\sigma^2_{Z_1} = 1 \Rightarrow \sigma^2_{Z_1} \sim IG(2.01, 4.01)$, $\rho_Z = 0 \Rightarrow \rho_Z \sim N(0,1)I(-1,1)$, $\phi^2_Z = 0.01 \Rightarrow \phi^2_Z \sim IG(2.001, 0.03)$, $\phi^2_Z = 0.1 \Rightarrow \phi^2_Z \sim IG(4.5, 0.55)$.

Among the 240 simulated observations in each dataset, 170 points, chosen between the first and tenth time points, were used for estimation and the remaining 70 points were set aside for validation purposes. The predictive performance was evaluated using two criteria. We computed an *average prediction error* (APE), defined as the average circular distance between a validation dataset and model predicted values, where we adopted as circular distance $d(\alpha, \beta) = 1 - \cos(\alpha - \beta)$ (Jammalamadaka and SenGupta 2001, p. 15). In particular, suppose the validation set has n^* observations, the APE for the models based on the wrapped normal is $\frac{1}{n^*}\sum_{(\mathbf{s}_0,t_0)} d(\mu(\mathbf{s}_0, t_0|\mathbf{X}), x(\mathbf{s}_0, t_0))$ and $\frac{1}{n^*}\sum_{(\mathbf{s}_0,t_0)} d(\mu(\mathbf{s}_0, t_0|\boldsymbol{\Theta}), \theta(\mathbf{s}_0, t_0))$ for the projected normal ones. Here, $x(\mathbf{s}_0, t_0)$ and $\theta(\mathbf{s}_0, t_0)$ are the realizations of the processes at $(\mathbf{s}_0, t_0)$ and $\mu(\mathbf{s}_0, t_0|\mathbf{X})$ and $\mu(\mathbf{s}_0, t_0|\boldsymbol{\Theta})$ are the posterior mean directions.

We also computed the *continuous ranked probability score* (CRPS) for circular variables as defined in Grimit et al. (2006):

$$\text{CRPS}(F, \delta) = E(d(\Delta, \delta)) - \frac{1}{2}E(d(\Delta, \Delta^*)), \tag{6}$$

where F is a predictive distribution, δ is a holdout value, and Δ and Δ^* are independent copies of a circular variable with distribution F. In this form, small values of CRPS are preferred.

For both models we do not know F in closed form, but we can compute a Monte Carlo approximation of (6). For the wrapped model, the CRPS for a held-out space–time point $(\mathbf{s}_0, t_0)$ is

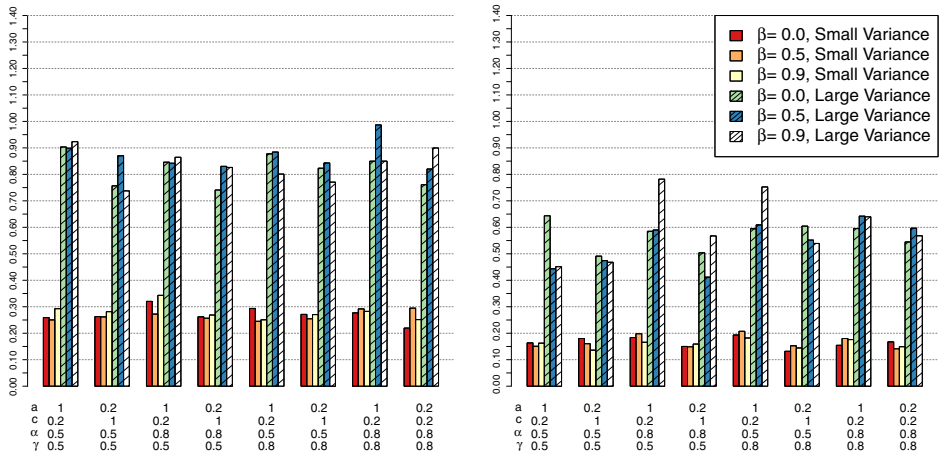
$$\frac{1}{L}\sum_{l=1}^{L} d(x_l^*(\mathbf{s}_0, t_0), x(\mathbf{s}_0, t_0)) - \frac{1}{2L^2}\sum_{l=1}^{L}\sum_{j=1}^{L} d(x_l^*(\mathbf{s}_0, t_0), x_j^*(\mathbf{s}_0, t_0))$$

and for the projected model it is

$$\frac{1}{L}\sum_{l=1}^{L} d(\theta_l^*(\mathbf{s}_0, t_0), \theta(\mathbf{s}_0, t_0)) - \frac{1}{2L^2}\sum_{l=1}^{L}\sum_{j=1}^{L} d(\theta_l^*(\mathbf{s}_0, t_0), \theta_j^*(\mathbf{s}_0, t_0)).$$

For each of the 48 simulated datasets, the values of the mean CRPS under the two models, computed over the set of points used for model validation, are shown in Fig. 1. For both models we see that the CRPS depends heavily on the variance of the process, but seems unaffected by changes in the other parameters.

A potentially important difference between the two models is the computational time required to fit them. The WN model is computationally more efficient than the PN model; the main issue is computational complexity (see Supplementary Online Material, Section S1). The PN requires, at each MCMC iteration, roughly 8 times as



(a) WN **(b)** PN

Fig. 1 Simulation study: CRPS comparing performances of the two proposed models. **a** WN. **b** PN

many operations as the WN to be fitted. If computational time is a relevant issue, then the WN may be more attractive.

5 Real data

We model wave directions obtained as outputs from a deterministic computer model implemented by Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA). The computer model starts from a wind forecast model predicting the surface wind over the entire Mediterranean. The hourly evolution of sea wave spectra is obtained by solving energy transport equations using the wind forecast as input. Wave spectra are locally modified using a source function describing the wind energy, the energy redistribution due to nonlinear wave interactions, and energy dissipation due to wave fracture. The model produces estimates every hour on a grid with 10×10 km cells (Speranza et al. 2004, 2007). The ISPRA dataset has forecasts for a total of 4941 grid points over the Italian Mediterranean. Over the Adriatic Sea area, there are 1494 points.

Our aim is to compare the performance of the WN and PN models. From a phenomenological perspective, the PN model is arguably the more natural choice since we are not wrapping a linear scale to obtain the directions. However, the WN model does provide a suitable model and, as suggested above, it may be attractive in terms of computational efficiency and interpretability of parameters. In the selected dataset, the three sea states, *calm, transition* and *storm* are present. The sea state is defined through the wave height (which is also supplied by the computer model output): when this height is below 1 m, we have *calm*, when it is between 1 and 2 m we have *transition* (between calm and storm) and when it is greater than 2 m we have a *storm*. Wave directions vary more in calm than in storm. Here, we seek to learn about the spatio-temporal structure of the data relying only on the specification of the correla-

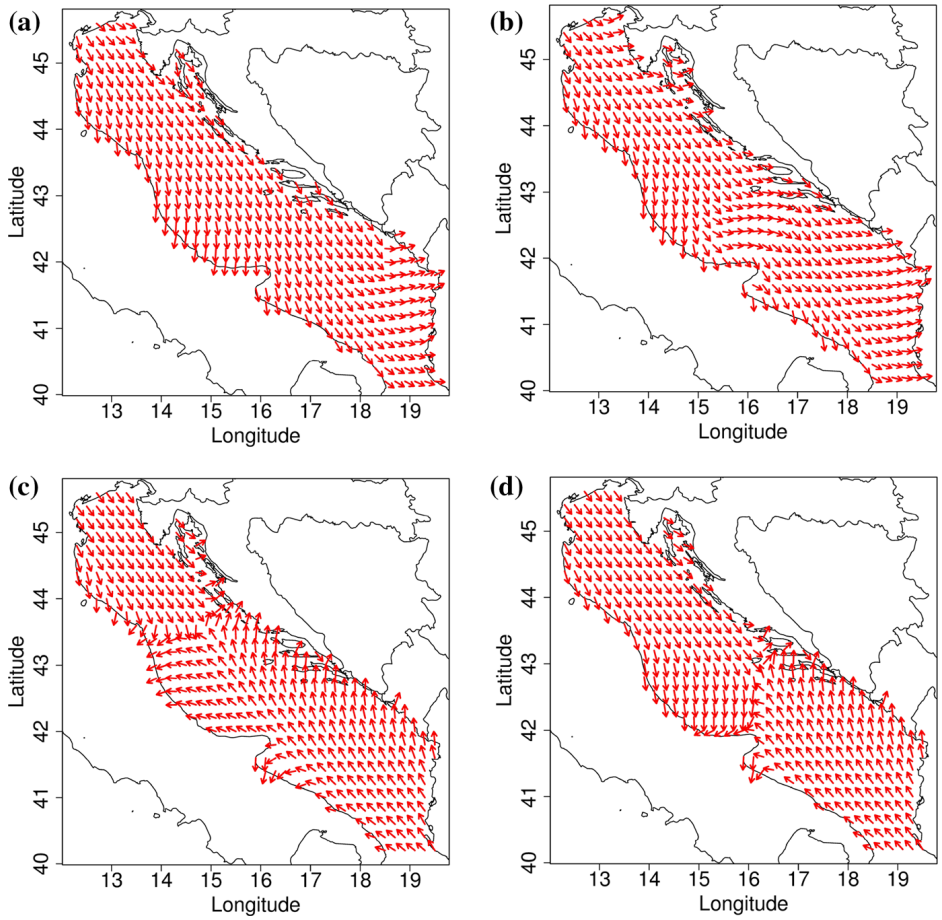


Fig. 2 Time windows for different sea states used for validation. The *four panels* represent the observed wave direction over the entire area at: **a** 12:00 on 5/5/2010 (storm); **b** 00:00 on 6/5/2010 (transition between storm and calm); **c** 00:00 on 7/5/2010 (calm); **d** 12:00 on 7/5/2010 (one-step prediction, calm)

tion function. We will use the information given by the wave heights in the models proposed in Sect. 6.1.

We fitted the model using 100 spatial points $\times$ 10 time points 6 h apart (1000 observations in total) in order to have a dataset including all sea states. Notice that spatial distances are evaluated in kilometers. Then, we developed four validation datasets, each with 350 spatial points and 1 time point. Specifically, we have one dataset for each sea state plus one for a one-step forward prediction. Finally, we used the model fitted over the 1000 points to predict each validation dataset. Three of the datasets are inside the time window used for model estimation, one in calm sea, one in transition and one during a storm. The fourth validation set is at 12:00 on May 7, 2010, 6 h after the last time used for model fitting. The observed circular process in each of these four time windows can be seen in Fig. 2. For each time window and model we computed the mean CRPS and APE, see Table 1. Furthermore, we computed the mean CRPS and APE over the four time windows.

Table 1 Real data example: CRPS and APE for the WN and PN models computed on each validation dataset

	WN	PN
Average		
CRPS	0.655	0.629
APE	0.437	0.421
Calm		
CRPS	1.450	1.398
APE	0.995	0.973
Transition		
CRPS	0.082	0.074
APE	0.033	0.028
Storm		
CRPS	0.063	0.042
APE	0.026	0.009
One-step prediction		
CRPS	1.024	1.001
APE	0.693	0.674

Following our discussion in Sects. 2.1 and 3, we used the following priors: $a \sim G(1.5, 1)$, $c \sim G(1.5, 1)$, $\alpha \sim B(2, 2.5)$, $\beta \sim B(1.1, 2)$, $\gamma \sim B(2, 2.5)$, $\sigma_Y^2 \sim IG(2, 2)$, $\phi_Y^2 \sim IG(1, 0.25)$, $\mu_Y \sim WrapN(\pi, 10)$, $\mu_{Z_1} \sim N(0, 10)$, $\mu_{Z_2} \sim N(0, 10)$, $\rho_Z \sim N(0, 5)I(-1, 1)$, $\sigma_Z^2 \sim IG(2, 2)$ and $\phi_Z^2 \sim IG(1, 0.25)$. Notice that all distributions are weakly informative. Also, the prior for β is centered near 0.1, i.e., close to the separable model. Decay parameters in space and time are related to the minimum and maximum distances in space and time, chosen to ensure that they concentrate the probability mass over such intervals.

As we expected, the predictive capability of the two models, in terms of both CRPS and APE, is poorest in a calm state, the variance being larger than in other states. On the other hand, it is very accurate during a storm or a transition for both models as we can see in Table 1. The PN always performs better that the WN. The largest difference between the APE values of the two models (0.022) is observed during the calm sea time window.

In Table 2 we give credible intervals and posterior mean estimates for the value of the parameters of the correlation function. For both models non-separable correlation structure is strongly supported. The point estimates of the spatial (c) and temporal (a) decay are smaller in the PN model. Notice that data are bimodal whenever the wave directions look like those in Fig. 2c, d, i.e., when over a large region at a given time a storm is rotating or two different weather systems are meeting. Then, scalar statistics, such as the overall mean direction or the overall concentration, may not be informative regarding this behaviour.

In the Supplementary Online Material, we provide the parameter estimates for the wrapped and projected distributions with associated 95 % credible intervals (Table S1). Since μ_Y is defined on a circular domain (recall that the prior on μ_Y is $WrapN(\cdot, \cdot)$), following Jona Lasinio et al. (2012), we can compute a 95 % credible interval as the arc that contains the central 95 % of the posterior samples.

Table 2 Real data example: mean point estimate (PE) and 95 % credible interval (CI) for the correlation parameters for the WN and PN models

	WN	PN
a		
PE	0.076	0.009
(CI)	(0.019, 0.200)	(0.005, 0.019)
c		
PE	3.2×10^{-4}	1.4×10^{-4}
(CI)	$(1.3 \times 10^{-4}, 7.1 \times 10^{-4})$	$(7.0 \times 10^{-4}, 2.9 \times 10^{-4})$
α		
PE	0.495	0.693
(CI)	(0.288, 0.744)	(0.562, 0.819)
β		
PE	0.592	0.430
(CI)	(0.158, 0.915)	(0.101, 0.774)
γ		
(PE)	0.797	0.872
(CI)	(0.697, 0.897)	(0.779, 0.939)

6 Extending the models

In the framework of the wrapped and projected normal models, introducing covariate information to explain the angular response is straightforward. For the wrapped approach we revise the linear version (4) to $Y(\mathbf{s}, t) = \mu_Y(\mathbf{s}, t) + \varepsilon_Y(\mathbf{s}, t)$.

The external variables can be introduced by modeling the mean of the circular process. Linear specification induces a circular likelihood for the regression coefficients that has infinitely many maxima of comparable size since this model wraps the line infinitely many times around the circle, (see for example Johnson and Wehrly 1978; Fisher and Lee 1992). To address this problem it is customary to limit the domain of $\mu_Y(\mathbf{s}, t)$ using a link function, i.e., $\mu_Y(\mathbf{s}, t) = \mathcal{L}(\mathbf{H}(\mathbf{s}, t)\,\boldsymbol{\eta})$, where $\mathcal{L}(\cdot) : \mathbb{R} \to I$ is the *link function* and I is some interval of $\mathbb{R}$ of length equal to the circular variable period, in our case 2π. We employ the *inverse tan link* (Fisher and Lee, 1992).

If only categorical covariates are available we do not need a link function; we can adopt an ANOVA representation for the relation between circular response and discrete covariates. This is computationally more efficient (see Supplementary Online Material, Section S1). Illustratively, suppose we have two predictors, with m_1 and m_2 levels, respectively, say $\mathbf{H}_1 = (H_{1,1}, \ldots, H_{1,m_1})$ and $\mathbf{H}_2 = (H_{2,1}, \ldots, H_{2,m_2})$. Then, to simplify the condition ensuring $\mu_Y(\mathbf{s}, t) \in I$, we use the following parametrization:

$$\mu_Y(\mathbf{s}, t) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mu_{Y,im_2+j} 1_{(H_1(\mathbf{s},t)=H_{1,i})} 1_{(H_2(\mathbf{s},t)=H_{2,j})}.$$

Table 3 Real data example: CRPS and APE for WNR, WNA, PNR and PNA models computed on each validation dataset

	WNR	WNA	PNR	PNA
Average				
CRPS	0.668	0.644	0.507	0.588
APE	0.502	0.431	0.496	0.450
Calm				
CRPS	1.548	1.409	1.129	1.342
APE	1.158	0.997	0.985	0.984
Transition				
CRPS	0.095	0.094	0.092	0.093
APE	0.033	0.030	0.046	0.038
Storm				
CRPS	0.057	0.054	0.118	0.053
APE	0.016	0.013	0.110	0.012
One-step prediction				
CRPS	0.971	1.018	0.689	0.866
APE	0.802	0.685	0.841	0.765

We can also introduce the covariates into the specifications for the variances, creating $\sigma_Y^2(\mathbf{s},t)$ and $\phi_Y^2(\mathbf{s},t)$. Again, we consider ANOVA-type models, e.g., $\sigma_Y^2(\mathbf{s},t) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\sigma_{Y,im_2+j}^2 1_{(H_1(\mathbf{s},t)=H_{1,i})}\ 1_{(H_2(\mathbf{s},t)=H_{2,j})}$ and $\phi_Y^2(\mathbf{s},t) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\phi_{Y,im_2+j}^2 1_{(H_1(\mathbf{s},t)=H_{1,i})} 1_{(H_2(\mathbf{s},t)=H_{2,j})}$.

We investigate two models, both with an ANOVA parametrization for $\sigma_Y^2(\mathbf{s},t)$ and $\phi_Y^2(\mathbf{s},t)$ while for the mean, one has an ANOVA parametrization (*WNA*) and the other has a regression form (*WNR*). Below, we obtain an ANOVA form if we work with sea state and a regression form if we work with wave height. As prior distributions we propose: $N(\cdot,\cdot)$ for $\eta_{Y,i}, i = 1,2,\ldots$, that is, a customary prior for a regression coefficient; $WrapN(\cdot,\cdot)$ for $\mu_{Y,i}, i = 1,2,\ldots$, the circular equivalent of a normal prior over mean level; and $IG(\cdot,\cdot)$ for $\sigma_{Y,i}^2$ and $\phi_{Y,i}^2$, $i = 1,2,\ldots$, that is, the customary prior for a variance. To sample from the predictive distribution, we adopt the same procedure used above for the WN model.

To introduce dependence on covariates in the projected normal model, following Wang and Gelfand (2013), we revise Eq. (5) to $Z_\ell(\mathbf{s},t) = \mu_{Z_\ell}(\mathbf{s},t) + \omega_{Z_\ell}(\mathbf{s},t) + \tilde{\varepsilon}_{Z_\ell}(\mathbf{s},t)$, $\ell = 1,2$ where the mean of the linear bivariate process is a function of space and/or time and $\tilde{\varepsilon}_{Z_\ell}(\mathbf{s},t) \overset{iid}{\sim} N(0,\phi_Z^2(\mathbf{s},t))$. Then we marginalize over $\boldsymbol{\omega}_Z(\mathbf{s},t)$ to obtain $Z_\ell(\mathbf{s},t) = \mu_{Z_\ell}(\mathbf{s},t) + \varepsilon_{Z_\ell}(\mathbf{s},t)$, $\ell = 1,2$. We write $\mu_{Z_\ell}(\mathbf{s},t) = \mathbf{H}(\mathbf{s},t)\boldsymbol{\eta}_{Z_\ell}$, $\ell = 1,2$ and $\phi_Z^2(\mathbf{s},t) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\phi_{Z,im_2+j}^2 1_{(H_1(\mathbf{s},t)=H_{1,i})} 1_{(H_2(\mathbf{s},t)=H_{2,j})}$, where $\boldsymbol{\eta}_{Z_\ell} = (\eta_{Z_\ell,1}, \eta_{Z_\ell,2},\ldots)'$. Note that, depending on the types of variables in $\mathbf{H}(\mathbf{s},t)$, continuous or categorical, we can specify a (projected normal) regression (*PNR*) or (projected normal) ANOVA (*PNA*). As noted in Wang and Gelfand (2014), there is complex interaction among the parameters in the general projected normal, complicating interpretation of the behavior of the resulting projected normal distributions as we vary

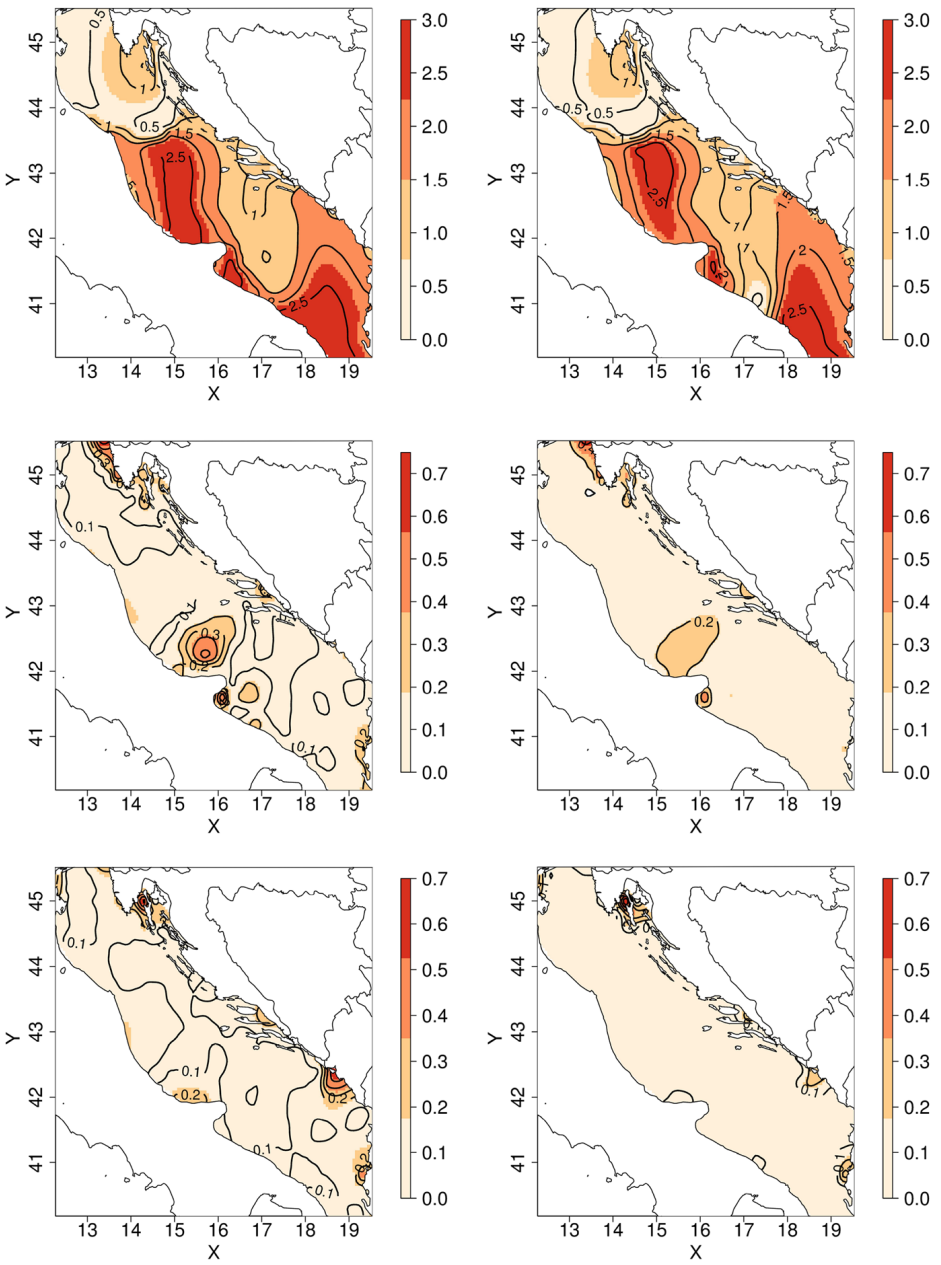


Fig. 3 Real data example: CRPS surfaces for the WN (*first column*) and PN (*second column*) models, under calm (*first row*), transition (*second row*) and storm (*third row*) states. Scales differ across states

them. With the same rationale used for the priors of the WNA and WNR models, we propose $\eta_{Z_\ell,i} \sim N(\cdot,\cdot)$, $l = 1, 2$, $i = 1, 2, \ldots$ and $\phi_{Z,i} \sim IG(\cdot,\cdot)$, $i = 1, 2, \ldots$. Here, again, we can sample from the predictive distribution adopting the same procedure as illustrated in Sect. 3.1.

Table 4 Real data example: mean point estimate (PE) and 95 % credible interval (CI) for the correlation parameters of the WNA, WNR, PNA and PNR models

	WNR	WNA
a		
PE	0.015	0.008
(CI)	(0.005, 0.035)	(0.003, 0.020)
c		
PE	6.1×10^{-5}	4.0×10^{-5}
(CI)	$(2.0 \times 10^{-5}, 1.4 \times 10^{-4})$	$(2.0 \times 10^{-5}, 7.0 \times 10^{-5})$
α		
PE	0.620	0.611
(CI)	(0.445, 0.786)	(0.434, 0.765)
β		
PE	0.396	0.539
(CI)	(0.070, 0.830)	(0.181, 0.868)
γ		
(PE)	0.705	0.936
(CI)	(0.620, 0.794)	(0.880, 0.976)
	PNR	PNA
a		
PE	0.119	0.108
(CI)	(0.042, 0.267)	(0.042, 0.225)
c		
PE	3.0×10^{-3}	1.0×10^{-3}
(CI)	$(1.01 \times 10^{-3}, 1.35 \times 10^{-3})$	$(4.60 \times 10^{-4}, 3.46 \times 10^{-3})$
α		
PE	0.575	0.506
(CI)	(0.345, 0.763	(0.340, 0.706)
β		
PE	0.082	0.063
(CI)	(0.000, 0.402)	(0.000, 0.300)
γ		
(PE)	0.561	0.541
(CI)	(0.435, 0.677)	(0.441, 0.645)

6.1 Application to the wave data

We fitted the new models using the same dataset as in Sect. 5. For the ANOVA representation we used, as a categorical variable, the state of the sea while for the regression setting we used the significant wave height. Adopting the same rationale as in Sect. 5, the prior distributions for the regression coefficients ($\eta_{Y,j,i}$ and

Table 5 Real data example: mean point estimate (PE) and 95 % credible interval (CI) of the parameters of the WNA and WNR models

	WNA			
	$\mu_{Y,\text{calm}}$	$\sigma^2_{Y,\text{calm}}$	$\phi^2_{Y,\text{calm}}$	
PE	0.095	1.524	0.051	
(CI)	(5.232, 1.328)	(0.959, 2.387)	(0.039, 0.068)	
	$\mu_{Y,\text{tran}}$	$\sigma^2_{Y,\text{tran}}$	$\phi^2_{Y,\text{tran}}$	
PE	5.998	0.541	0.018	
(CI)	(5.278, 0.490)	(0.332, 0.876)	(0.013, 0.026)	
	$\mu_{Y,\text{storm}}$	$\sigma^2_{Y,\text{storm}}$	$\phi^2_{Y,\text{storm}}$	
PE	5.860	0.385	0.009	
(CI)	(5.254, 0.281)	(0.246, 0.582)	(0.007, 0.012)	
	WNR			
	$\eta_{Y,0,\text{calm}}$	$\eta_{Y,1,\text{calm}}$	$\sigma^2_{Y,\text{calm}}$	$\phi^2_{Y,\text{calm}}$
PE	0.997	4.918	5.000	0.041
(CI)	(0.360, 1.901)	(2.433, 7.619)	(2.313, 9.494)	(0.027, 0.058)
	$\eta_{Y,0,\text{tran}}$	$\eta_{Y,1,\text{tran}}$	$\sigma^2_{Y,\text{tran}}$	$\phi^2_{Y,\text{tran}}$
PE	3.166	2.526	1.825	0.018
(CI)	(0.763, 5.894)	(0.174, 6.844)	(1.013, 3.046)	(0.012, 0.025)
	$\eta_{Y,0,\text{storm}}$	$\eta_{Y,1,\text{storm}}$	$\sigma^2_{Y,\text{storm}}$	$\phi^2_{Y,\text{storm}}$
PE	3.470	1.933	1.322	0.010
(CI)	(0.666, 6.445)	(0.064, 5.870)	(0.660, 2.167)	(0.007, 0.013)

$\eta_{Z_\ell,j,i},\ j=0,1,\ i=calm,trans,storm)$ were all $N(0,10)$. For the ANOVA coefficients, $\mu_{Y,i}$ and $\mu_{Z_\ell,i}$, they were all $WrapN(\pi,10)$. For the $\sigma^2_{Y,i}$, they were all $IG(2,2)$ and for the $\phi_{Y,i}$ and $\phi_{Z,i}$ they were all $IG(1,0.25)$. The prior distributions for the other parameters were the same as those used in Sect. 5.

From Table 3 we see that the WNA model is generally preferred to the WNR. For the projected models, APE and CRPS are almost indistinguishable between PNA and PNR during transition. With one-step ahead predictions, the two criteria return contradicting choices; PNR is preferred with CRPS, PNA with the APE. With the calm sea state, the CRPS chooses PNR while APE does not yield a clear decision. With the storm state, both criteria are lower with the PNA model. Overall, our selection would be the PNA model but, more importantly, we value the informative comparison our approach enables. In fact, the remarkable improvement of PNA over PNR in storm is likely due to the very high predictability of direction during a storm period. In this regard, the PN models are generally preferred to the WN models except in storm where WNR, WNA, and PNA are essentially equivalent.

To analyze the local behavior of model fitting, in Fig. 3 we report CRPS surfaces, evaluated in calm, transition and storm for the two "best average APE" models, the WNA (see Table 3) and PN (see Table 1). We see that the local behavior of the

Table 6 Real data example: mean point estimate (PE) and 95 % credible interval (CI) for the parameters of the PNA and PNR models

	PNA				
	$\mu_{Z_1,\text{calm}}$	$\mu_{Z_2,\text{calm}}$	$\phi^2_{Z,\text{calm}}$		
PE	0.841	−0.404	0.027		
(CI)	(−1.112, 2.706)	(−2.408, 1.427)	(0.014, 0.051)		
	$\mu_{Z_1,\text{tran}}$	$\mu_{Z_2,\text{tran}}$	$\phi^2_{Z,\text{tran}}$		
PE	0.697	−0.724	0.047		
(CI)	(−1.281, 2.640)	(−2.600, 1.173)	(0.018, 0.099)		
	$\mu_{Z_1,\text{storm}}$	$\mu_{Z_2,\text{storm}}$	$\phi^2_{Z,\text{storm}}$		
PE	0.615	−0.615	0.037		
(CI)	(−1.376, 2.615)	(−2.543, 1.289)	(0.016, 0.076)		
	$\sigma^2_{Z,1}$	ρ_Z			
(PE)	2.072	−0.161			
(CI)	(1.425, 2.938)	(−0.320, 0.003)			
	PNR				
	$\eta_{Z_1,0,\text{calm}}$	$\eta_{Z_1,1,\text{calm}}$	$\eta_{Z_2,0,\text{calm}}$	$\eta_{Z_2,1,\text{calm}}$	$\phi^2_{Z,\text{calm}}$
PE	0.997	0.875	−0.925	0.840	0.110
(CI)	(−0.989, 3.026)	(−1.160, 2.927)	(−2.878, 1.091)	(−1.162, 2.798)	(0.033, 0.250)
	$\eta_{Z_1,0,\text{tran}}$	$\eta_{Z_1,1,\text{tran}}$	$\eta_{Z_2,0,\text{tran}}$	$\eta_{Z_2,1,\text{tran}}$	$\phi^2_{Z,\text{tran}}$
PE	0.916	0.976	−1.117	−0.554	0.127
(CI)	(−1.195, 3.015)	(−1.258, 3.117)	(−3.322, 0.893)	(−2.601, 1.649)	(0.037, 0.322)
	$\eta_{Z_1,0,\text{storm}}$	$\eta_{Z_1,1,\text{storm}}$	$\eta_{Z_2,0,\text{storm}}$	$\eta_{Z_2,1,\text{storm}}$	$\phi^2_{Z,\text{storm}}$
PE	0.768	1.088	−0.974	−1.190	0.091
(CI)	(−1.424, 2.899)	(−1.083, 3.235)	(−3.146, 1.177)	(−3.281, 0.955)	(0.031, 0.201)
	$\sigma^2_{Z,1}$	ρ_Z			
PE	2.293	−0.191			
(CI)	(1.602, 3.212)	(−0.358, −0.013)			

models is very similar. The worst predictions are found around the Gargano peninsula during calm. This is consistent with the physics of wave movement since, around the peninsula, local winds play a more relevant role, inducing very high variability in wave directions. The same behavior is shown with the other models. In terms of parameter estimation the WNA and PN models suggest a non-separable model (Tables 2, 4) with very strong spatial (c) and temporal (a) dependence. WNA suggests that a different nugget is necessary for each sea state. In fact analyzing the credible intervals of these parameters we observe that, for each sea state, nuggets are significantly different among them as their credible intervals do not overlap (Table 5). For the projected normal models (Table 6), all nugget credible intervals are substantially overlapping, suggesting that one nugget should be enough to model all sea states.

7 Conclusions

We have presented a range of models for spatio-temporal circular data based on the wrapped and projected normal distributions, incorporating space–time dependence, allowing explanatory variables, introducing a nugget, implementing kriging and forecasting. The models based on the projected normal are more flexible since they allow bimodal and asymmetric distributions while the wrapped normal is unimodal and symmetric. On the other hand, the wrapped normal models are easy to interpret and are computationally better behaved and more efficient. Predictions obtained under the two models are very close and almost indistinguishable when data are roughly unimodal and symmetric (see Supplementary Online Material, Section S2). Then, if fast computation is sought, WN models become attractive.

The projected normal process can be straightforwardly extended to general directional fields on the sphere since the projected normal distribution is well defined in this case, see Mardia and Jupp (1999). The wrapped Gaussian process is not easily extended to a sphere. In fact, we are unaware of any approach to wrap multivariate linear data onto spheres. Conceptually, such wrapping would *not* appear to be well defined.

Future work will find us enriching wrapped modeling to allow asymmetry through the use of skewed distributions. Skewness is easy to introduce by wrapping skew normal distributions. In a completely different direction, we are also extending the modeling to explore spatio-temporal data consisting of geo-coded locations with periodic (in time) behaviour that can be represented as a circular variable. There, we work with trivariate GPs in space and time, incorporating temporal projection.

Acknowledgments The authors thank INFN Bari CED for allowing the use of their high-performance grid computing infrastructure Bc2S. The authors thank ISPRA for the use of data output from the wave model of its SIMM hydro-meteo-marine forecasting system.

References

Banerjee S, Gelfand AE, Carlin BP (2014) Hierarchical modeling and analysis for spatial data, 2nd edn. Chapman and Hall/CRC, New York

Breckling J (1989) The analysis of directional time series: applications to wind speed and directions. Lecture notes in statistics. Springer-Verlag, Berlin

Bulla J, Lagona F, Maruotti A, Picone M (2012) A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series. J Agric Biol Environ Stat 17:544–567

Coles S (1998) Inference for circular distributions and processes. Stat Comput 8:105–113

Damien P, Walker S (1999) A full Bayesian analysis of circular data using the von Mises distribution. Can J Stat 27:291–298

Fisher NI (1996) Statistical analysis of circular data. Cambridge University Press, Cambridge

Fisher NI, Lee AJ (1992) Regression models for an angular response. Biometrics 48:665–677

Gneiting T (2002) Nonseparable, stationary covariance functions for space-time data. J Am Stat Assoc 97:590–600

Grimit EP, Gneiting T, Berrocal VJ, Johnson NA (2006) The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. Q J R Meteorol Soc 132:2925–2942

Guttorp P, Lockhart RA (1988) Finding the location of a signal: a Bayesian analysis. J Am Stat Assoc 83:322–330

Harrison D, Kanji GK (1988) The development of analysis of variance for circular data. J Appl Stat 15:197–224
Holtzman H, Munk A, Suster M, Zucchini W (2006) Hidden Markov models for circular and linear-circular time series. Environ Ecol Stat 13:325–347
Jammalamadaka SR, SenGupta A (2001) Topics in circular statistics. World Scientific, Singapore
Johnson RA, Wehrly TE (1978) Some angular-linear distributions and related regression models. J Am Stat Assoc 73:602–606
Jona Lasinio G, Gelfand AE, Jona Lasinio M (2012) Spatial analysis of wave direction data using wrapped Gaussian processes. Ann Appl Stat 6:1478–1498
Kato S (2010) A Markov process for circular data. J R Stat Soc Ser B 72:655–672
Kato S, Shimizu K (2008) Dependent models for observations which include angular ones. J Stat Plan Infer 138:3538–3549
Lagona F, Picone M (2011) A latent-class model for clustering incomplete linear and circular data in marine studies. J Data Sci 9:585–605
Lee A (2010) Circular data. Wiley Interdiscip Rev Comput Stat 2:477–486
Mardia KV (1972) Statistics of directional data. Academic Press, London
Mardia KV, Jupp PE (1999) Directional statistics. Wiley, Chichester
Mastrantonio G, Maruotti A, Jona Lasinio G (2015) Bayesian hidden Markov modelling using circular-linear general projected normal distribution. Environmetrics 26:145–158
Ravindran P, Ghosh SK (2011) Bayesian analysis of circular data using wrapped distributions. Stat Sin 5:547–561
Speranza A, Accadia C, Casaioli M, Mariani S, Monacelli G, Inghilesi R, Tartaglione N, Ruti PM, Carillo A, Bargagli A, Pisacane G, Valentinotti F, Lavagnini A (2004) Poseidon: an integrated system for analysis and forecast of hydrological, meteorological and surface marine fields in the Mediterranean area. Nuovo Cimento 27:329–345
Speranza A, Accadia C, Mariani S, Casaioli M, Tartaglione N, Monacelli G, Ruti PM, Lavagnini A (2007) Simm: an integrated forecasting system for the Mediterranean area. Meteorol Appl 14:337–350
Wang F, Gelfand AE (2013) Directional data analysis under the general projected normal distribution. Stat Methodol 10:113–127
Wang F, Gelfand AE (2014) Modeling space and space–time directional data using projected Gaussian processes. J Am Stat Assoc 109:1565–1580
Wang F, Gelfand AE, Jona Lasinio G (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. Stat Sin 25:25–39

Stoch Environ Res Risk Assess
DOI 10.1007/s00477-015-1163-9

ORIGINAL PAPER

The wrapped skew Gaussian process for analyzing spatio-temporal data

Gianluca Mastrantonio[1] · **Alan E. Gelfand**[2] · **Giovanna Jona Lasinio**[3]

Abstract We consider modeling of angular or directional data viewed as a linear variable wrapped onto a unit circle. In particular, we focus on the spatio-temporal context, motivated by a collection of wave directions obtained as computer model output developed dynamically over a collection of spatial locations. We propose a novel wrapped skew Gaussian process which enriches the class of wrapped Gaussian process. The wrapped skew Gaussian process enables more flexible marginal distributions than the symmetric ones arising under the wrapped Gaussian process and it allows straightforward interpretation of parameters. We clarify that replication through time enables criticism of the wrapped process in favor of the wrapped skew process. We formulate a hierarchical model incorporating this process and show how to introduce appropriate latent variables in order to enable efficient fitting to dynamic spatial directional data. We also show how to implement kriging and forecasting under this model. We provide a simulation example as a proof of concept as well as a real data example. Both examples reveal consequential improvement in predictive performance for the wrapped skew Gaussian specification compared with the earlier wrapped Gaussian version.

✉ Gianluca Mastrantonio
gianluca.mastrantonio@yahoo.it;
gianluca.mastrantonio@uniroma3.it

[1] Roma Tre University, Via Silvio D'Amico 77, 00145 Rome, Italy

[2] Duke University, 223-A Old Chemistry Building, Box 90251, Durham, NC 27708-0251, USA

[3] Sapienza University of Rome, P.le Aldo Moro 5, 00185 Rome, Italy

Published online: 28 September 2015

Keywords Directional data · Hierarchical model · Kriging · Markov chain Monte Carlo · Space–time data · Wave directions

1 Introduction

There is increasing interest in analyzing directional data which are collected over space and time. Examples arise, for instance, in oceanography (wave directions), meteorology (wind directions), biology (study of animal movement). They also arise from periodic data, e.g., event times might be wrapped according to a daily period to give a circular view (eliminating *end* effects). We wrap time around a circle by a modulus transformation and, without loss of generality, can rescale to degrees or angles on a unit circle. Time wrapping with spatial data occurs naturally in applications such as locations and times of crime events, locations and times of automobile accidents, and residence address with time of admission for hospitalizations.

Jona Lasinio et al. (2012) introduced a Bayesian hierarchical model to handle angular data, enabling full inference regarding all model parameters and prediction under the model. Their context was multivariate directional observations arising as angular data measurements taken at spatial locations, anticipating structured dependence between these measurements. They proposed the wrapped spatial Gaussian process, induced from a linear spatial Gaussian process. They explored dependence structure and showed how to implement kriging of mean directions and concentrations in this setting.

The current state of the art for modeling circular space–time data includes the wrapped Gaussian process and the projected Gaussian process. The second, although more flexible, is based upon a four parameter model such that

complex interactions among the parameters make interpretation difficult. In this paper our contribution is to overcome a key limitation of the wrapped Gaussian process, that the marginal distributions at all locations are symmetric. Here we introduce the wrapped skew Gaussian process. This new circular process allows for asymmetric marginal distributions while retaining straightforward parametric interpretation. Our wrapping approach is developed from the skew normal distribution proposed by Azzalini (1985) and the process extension constructed by Zhang et al. (2010).

By now, there is a fairly rich literature on skew multivariate normal models (Azzalini 2005; Sahu et al. 2003; Ma and Genton 2004; Wang et al. 2004) but all are *inline*, i.e., on a linear scale.

The first attempt to wrap the skew normal distribution for circular data can be found in Pewsey (2000) where its basic properties are derived. Follow-on work appears in Pewsey (2006), Hernández-Sánchez and Scarpa (2012).

To our knowledge, we propose the first extension to multivariate wrapped skew distributions, in particular, to a spatial and spatio-temporal setting. In what follows we review the univariate wrapped skew normal distribution, showing the flexibility of shapes and do the same for bivariate wrapped skew normal distributions. Then, we turn to a hierarchical model for dynamic spatial data and show how, using suitable latent variables, to fit it efficiently. We also show how to implement kriging under this model.

A critical point emerges: though we can fit both models with a single sample of spatially referenced directions, in terms of kriging performance, we can not criticize the wrapped spatial Gaussian process in favor of the wrapped skew spatial Gaussian process. This is not surprising. Consider the linear situation. With a single sample of data from a set of locations, it is difficult to criticize the Gaussian process in favor of a more complex stochastic process specification, i.e., it is difficult to criticize a multivariate normal model with a single sample of multivariate data. However, with replicates, we are able to demonstrate substantially improved predictive performance for the wrapped skew Gaussian process. We do this both with simulated data, as a proof of concept, and with real data, making direct comparison. In our setting replicates arise through a dynamic spatial data where we envision i.i.d. spatial increment processes.

Inference for spatial data is challenging due to the restriction of support to the unit circle, $[0, 2\pi)$, and to the sensitivity of descriptive and inferential results to the starting point on the circle. There exists a substantial early literature on circular data [see e.g. Mardia (1972) and Mardia and Jupp (1999), Jammalamadaka and SenGupta (2001) or Fisher (1996)] primarily confined to descriptive statistics and limited inference for simple univariate models.

Computational procedures such as MCMC methods and the EM algorithm, have substantially advanced inference opportunities for directional data. Some examples include linear models (Harrison and Kanji 1988; Fisher 1996; Kato and Shimizu 2008), linear models in a Bayesian context (Guttorp and Lockhart 1988; Damien and Walker 1999), models for circular time series (Breckling 1989; Coles 1998; Mardia and Jupp 1999; Ravindran and Ghosh 2011; Hughes 2007; Fisher and Lee 1992; Holzmann et al. 2006) or model for space–time circular-linear data (Lagona et al. 2015). Recently, Kato (2010), building upon earlier work (Kato et al. 2008), proposed a discrete time Markov process for circular data. He uses the Möbius circle transformation, connecting it with an early Markov process model of Fisher and Lee (1994).

With regard to multivariate theory for circular data, particularly in the fully Bayesian setting, the work of Coles (1998) is foundational for ours. He also employs wrapped distributions, noting that, in the Gaussian case, they can be readily given a multivariate extension. Coles mostly works with independent replicates of multivariate circular data in low dimension with an unknown covariance matrix and develops some theory and examples for the time series setting. He mentions possible extensions to the spatial setting but offers no development, in particular, no thoughts on regression or kriging (Sects. 3.5 and 3.6 below). Coles and Casson (1998) include spatial dependence in looking at the direction of maximum wind speed. With little detail, they propose conditionally independent directions modeled with a von Mises distribution, introducing spatial structure in the modal direction and concentration parameters, a second stage specification. Our view, again following Jona Lasinio et al. (2012), is to introduce spatial structure at the first stage of the modeling, directly on the angular variables, resulting in a spatial process model with smooth process realizations.

Following a different strand, the projected normal and the associated projected Gaussian process (Wang and Gelfand 2013, 2014) have generated recent interest. In particular, a general bivariate normal distribution is projected to an angle, extending work of Presnell et al. (1998) and Nuñez-Antonio and Gutiérrez-Peña (2005). The extension to a stochastic process for variables on the circle over a continuous spatial domain, the *projected Gaussian process*, is induced from a linear bivariate spatial Gaussian process. The projected Gaussian process has marginal distributions that can be asymmetric, possibly bimodal, an advantage over the wrapped Gaussian process. Wang and Gelfand (2014) also investigate properties of this process, including the nature of joint distributions for pairs of directions at different locations. Working within a hierarchical Bayesian framework, they show that model fitting is straightforward using suitable latent variable augmentation

in the context of Markov chain Monte Carlo (MCMC). In very recent work, Mastrantonio et al. (2015) offer comparison between the wrapping and the projection modeling approaches.

We remark that we have explored the possibility of introducing skewness into the projected Gaussian process. The overall process model is induced by a bivariate skewed Gaussian process. This is a more challenging process to work with; the resulting directional process model is extremely messy and has proved very difficult to fit. It likely exceeds what the data is capable of supporting. We do not discuss it further.

Our motivating example is drawn from marine data. Wave heights and outgoing wave directions, the latter being measured in degrees relative to a fixed orientation, are the main outputs of marine forecasts. Numerical models for weather and marine forecasts need statistical post-processing. Wave directions, being angular variables, cannot be treated through standard post-processing techniques [see Engel and Ebert (2007); Bao et al. 2009), and references therein]. In Bao et al. (2009) bias correction and ensemble calibration forecasts of surface wind direction are proposed. The authors use circular–circular regression as in Kato et al. (2008) for bias correction and Bayesian model averaging with the von Mises distribution for ensemble calibration. However, their approach does not explicitly account for spatial structure.

Lastly, it is worth commenting that, in our setting, wave direction data is viewed differently from wind direction data. The former is only available as an angle while the latter is customarily associated with wind speed, emerging as the resultant of North–South and East–West wind speed components.

The format of the paper is as follows. In Sect. 2 we review, develop and illustrate the univariate wrapped skew normal distribution. Section 3 extends to the wrapped skew Gaussian process, including distribution theory, model fitting, and kriging. Section 4 provides the dynamic version which we then pursue through simulation in Sect. 5 and a wave direction data analysis in Sect. 6. Section 7 offers a brief summary and some future research possibilities.

2 The wrapped skew normal

2.1 The univariate case

We begin with the univariate wrapped skew normal distribution. Let X and W be two independent standard normal variables, let $\sigma^2 \in \mathbb{R}^+$ and $\lambda \in \mathbb{R}$. Then, the random variable

$$Z = \mu + \frac{\sigma\lambda}{\sqrt{1+\lambda^2}}|X| + \frac{\sigma}{\sqrt{1+\lambda^2}}W - \frac{\sigma\lambda\sqrt{2}}{\sqrt{\pi(1+\lambda^2)}} \quad (1)$$

is said to be distributed as a skew normal variable (Azzalini 1985) with parameters μ, σ^2 and λ; i.e., $Z|\boldsymbol{\Psi} \sim SN(\mu, \sigma^2, \lambda)$, where $\boldsymbol{\Psi}$ denotes the vector of parameters. Let $\phi(\cdot)$ and $\Phi(\cdot)$ be the probability density function (pdf) and the cumulative density function (cdf), respectively, of a standard normal. Then, the pdf of $Z|\boldsymbol{\Psi}$ is

$$\frac{2}{\sigma}\phi\left(\frac{z-\mu+\frac{\sigma\lambda\sqrt{2}}{\sqrt{\pi(1+\lambda^2)}}}{\sigma}\right)\Phi\left(\lambda\left(\frac{z-\mu+\frac{\sigma\lambda\sqrt{2}}{\sqrt{\pi(1+\lambda^2)}}}{\sigma}\right)\right)$$

and from (1) we can easily derive the mean and the variance of Z, respectively. They are μ (the definition in (1) was made in order to center Z at μ) and

$$\sigma^2\lambda^2/(1+\lambda^2)(1-2/\pi) + \sigma^2/(1+\lambda^2).$$

With the transformation

$$\Theta = Z \bmod 2\pi, \text{ implying } \Theta \in [0, 2\pi), \quad (2)$$

we obtain a random variable with support on the unit circle. We can express the inline variable as $Z = \Theta + 2\pi K$, where K, the *winding number*, assumes values in $\mathbb{Z} = \{0, \pm1, \pm2, \ldots\}$. The transformation (2) defines what is called a *wrapped skew normal* (WSN) distribution, as introduced in Pewsey (2000). It wraps the skew normal distribution, defined on the real line, onto the unit circle. Details on the wrapping approach can be found in Jammalamadaka and SenGupta (2001) or Mardia and Jupp (1999).

The pdf of $\Theta|\boldsymbol{\Psi}$ is

$$\sum_{k\in\mathbb{Z}}\frac{2}{\sigma}\phi\left(\frac{\theta+2\pi k-\mu+\frac{\sigma\lambda\sqrt{2}}{\sqrt{\pi(1+\lambda^2)}}}{\sigma}\right) \times \Phi\left(\lambda\left(\frac{\theta+2\pi k-\mu+\frac{\sigma\lambda\sqrt{2}}{\sqrt{\pi(1+\lambda^2)}}}{\sigma}\right)\right). \quad (3)$$

The infinite sum in (3) is impossible to evaluate but, to display the density, as with the wrapped normal case, we can obtain an accurate approximation by appropriately truncating the sum. Figure 1 illustrates the effect of introduction of skewness into the wrapped normal density. To obtain a sample from a wrapped skew normal we first obtain a sample from the skew normal and then transform it to a circular variable via (2). Also, note that, if we let K be a random variable, the density inside the sum in (3) is the joint density of $(\Theta, K|\boldsymbol{\Psi})$ whence, we marginalize over K to obtain the density of the circular variable.

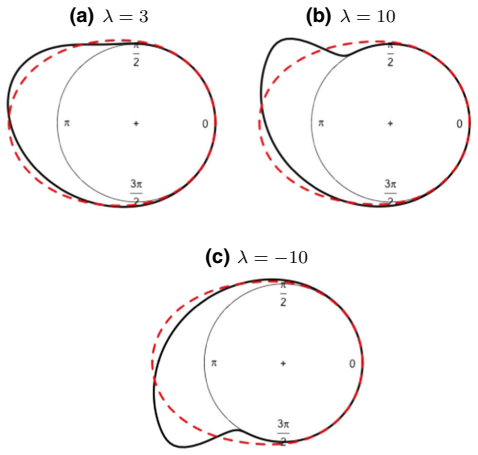


Fig. 1 Densities of the wrapped skew normal (*solid line*) with $\mu = \pi$, $\sigma^2 = 1$ and different values of λ along with the associated densities of the wrapped normal (*dashed line*) having the same circular mean and variance

Pewsey (2000) gives the fundamental properties of the WSN along with closed forms for the cosine and sine moments. Let $\mu^* = \mu - \frac{\sigma\lambda\sqrt{2}}{\sqrt{\pi(1+\lambda^2)}}$ and $\mathcal{J}(a) = \int_0^a \sqrt{\frac{2}{\pi}} \exp\left(\frac{u^2}{2}\right) du$ the cosine and sine moments become

$$\alpha_p = E(\cos p\Theta|\boldsymbol{\Psi}) = \exp\left(-\frac{p^2\sigma^2}{2}\right) \times \left(\cos(p\mu^*) - \mathcal{J}\left(\frac{\lambda\sigma p}{\sqrt{1+\lambda^2}}\right)\sin(p\mu^*)\right) \quad (4)$$

and

$$\beta_p = E(\sin p\Theta|\boldsymbol{\Psi}) = \exp\left(-\frac{p^2\sigma^2}{2}\right) \times \left(\sin(p\mu^*) + \mathcal{J}\left(\frac{\lambda\sigma p}{\sqrt{1+\lambda^2}}\right)\cos(p\mu^*)\right). \quad (5)$$

The trigonometric moments (4) and (5) are useful to compute the circular mean of Θ, $\bar{\mu} = \text{atan}^* \frac{\alpha_1}{\beta_1}$[1], and the circular concentration, $\bar{c} = \sqrt{\alpha_1^2 + \beta_1^2}$. However, unfortunately we need to compute $\mathcal{J}(\cdot)$, which is not available in closed form. Pewsey (2000) suggests to use deterministic numerical integration methods but we note that α_p and β_p can be computed using Monte Carlo approximation.

[1] For the definition of atan* see Jammalamadaka and SenGupta (2001), p. 13

Indeed, from (1) we can see that

$$Z|X, \boldsymbol{\Psi} \sim N\left(\mu + \frac{\sigma\lambda}{\sqrt{1+\lambda^2}}|X| - \frac{\sigma\lambda\sqrt{2}}{\sqrt{\pi(1+\lambda^2)}}, \frac{\sigma^2}{1+\lambda^2}\right) \quad (6)$$

and as a consequence

$$\Theta|X, \boldsymbol{\Psi} \sim WN\left(\mu + \frac{\sigma\lambda}{\sqrt{1+\lambda^2}}|X| - \frac{\sigma\lambda\sqrt{2}}{\sqrt{\pi(1+\lambda^2)}}, \frac{\sigma^2}{1+\lambda^2}\right), \quad (7)$$

where $WN(\cdot)$ indicates the wrapped normal distribution. Let $\{X^b\}_{b=1}^B$ be a set of B samples from the distribution of X. Then, we can write the cosine moments as $\alpha_p = E(\cos p\Theta|\boldsymbol{\Psi}) = E_{X|\boldsymbol{\Psi}} E_{\Theta|X,\boldsymbol{\Psi}}(\cos p\Theta|X, \boldsymbol{\Psi})$, since $E_{\Theta|X,\boldsymbol{\Psi}}(\cos p\Theta|X, \boldsymbol{\Psi})$ is the cosine moment of $\Theta|X, \boldsymbol{\Psi}$.

Following Jona Lasinio et al. (2012), a Monte Carlo approximation of α_p is

$$\hat{\alpha}_p \approx \frac{\exp\left(-p^2\frac{\sigma^2}{2(1+\lambda^2)}\right)}{B} \times \sum_{b=1}^{B} \cos\left(p\left(\mu + \frac{\sigma\lambda}{\sqrt{1+\lambda^2}}|X| - \frac{\sigma\lambda\sqrt{2}}{\sqrt{\pi(1+\lambda^2)}}\right)\right).$$

Similarly, we find

$$\hat{\beta}_p \approx \frac{\exp\left(-p^2\frac{\sigma^2}{2(1+\lambda^2)}\right)}{B} \times \sum_{b=1}^{B} \sin\left(p\left(\mu + \frac{\sigma\lambda}{\sqrt{1+\lambda^2}}|X| - \frac{\sigma\lambda\sqrt{2}}{\sqrt{\pi(1+\lambda^2)}}\right)\right)$$

and then $\hat{\bar{\mu}} = \text{atan}^* \frac{\hat{\alpha}_1}{\hat{\beta}_1}$ and $\hat{\bar{c}} = \sqrt{\hat{\alpha}_1^2 + \hat{\beta}_1^2}$.

2.2 The bivariate case

Let Z_1 and Z_2 be two random variables skew normal distributed with, respectively, parameters $(\mu_1, \sigma_1^2, \lambda_1)$ and $(\mu_2, \sigma_2^2, \lambda_2)$:

$$Z_1 = \mu_1 + \frac{\sigma_1\lambda_1}{\sqrt{1+\lambda_1^2}}|X_1| + \frac{\sigma_1}{\sqrt{1+\lambda_1^2}}W_1 - \frac{\sigma_1\lambda_1\sqrt{2}}{\sqrt{\pi(1+\lambda_1^2)}},$$

$$Z_2 = \mu_2 + \frac{\sigma_2\lambda_2}{\sqrt{1+\lambda_2^2}}|X_2| + \frac{\sigma_2}{\sqrt{1+\lambda_2^2}}W_2 - \frac{\sigma_2\lambda_2\sqrt{2}}{\sqrt{\pi(1+\lambda_2^2)}}.$$

We introduce dependence between Z_1 and Z_2 by letting $\text{Cor}(X_1, X_2|\boldsymbol{\Psi}) = \rho_x$ and $\text{Cor}(W_1, W_2|\boldsymbol{\Psi}) = \rho_w$. Then, we

say that $(Z_1, Z_2|\boldsymbol{\Psi})$ is distributed as a bivariate skew normal with the additional parameters, ρ_x and ρ_w. This specification of the bivariate skew normal, due to Zhang and El-Shaarawi (2010), differs from the one that can be derived using the multivariate normal of Azzalini and Valle (1996) and it is more suitable to built a stationary process, see Sect. 3.

Using the transformation (2) we can obtain the circular variables $\Theta_1 = Z_1 \bmod 2\pi$ and $\Theta_2 = Z_2 \bmod 2\pi$ associated with (Z_1, Z_2). The parameters ρ_x and ρ_w govern the dependence between Θ_1 and Θ_2 and if both are 0, Θ_1 and Θ_2 are independent as with the associated linear variables.

Let $g(\cdot|\boldsymbol{\Psi})$ be the density of $(Z_1, Z_2|\boldsymbol{\Psi})'$, let $\mathbf{K} = (K_1, K_2)'$ be the vector of winding numbers and $\boldsymbol{\Theta} = (\Theta_1, \Theta_2)'$, with $\mathbf{Z} = \boldsymbol{\Theta} + 2\pi\mathbf{K}$. As in the univariate case, we obtain the density of $\boldsymbol{\Theta}$, a bivariate wrapped skew normal, through marginalization over $\mathbf{K}$ of the joint density of $(\boldsymbol{\Theta}, \mathbf{K}|\boldsymbol{\Psi})$:

$$f(\boldsymbol{\theta}|\boldsymbol{\Psi}) = \sum_{k_1 \in \mathbb{Z}} \sum_{k_2 \in \mathbb{Z}} g(\boldsymbol{\theta} + 2\pi\mathbf{k}|\boldsymbol{\Psi}).$$

In Fig. 2 we show plots of the bivariate wrapped skew normal distributions.

3 The wrapped skew Gaussian process

A natural way to construct a wrapped skew Gaussian process $\Theta(\mathbf{s}), \mathbf{s} \in \mathbb{R}^d$ is to start from a skew Gaussian process $Z(\mathbf{s})$ on the line and define, for each $\mathbf{s}$, $\Theta(\mathbf{s}) = Z(\mathbf{s}) \bmod 2\pi$, following the approach of Jona Lasinio et al. (2012). To capture stationarity we use the following stationary skew Gaussian process, proposed by Zhang and El-Shaarawi (2010):

$$Z(\mathbf{s}) = \mu + \frac{\sigma\lambda}{\sqrt{1+\lambda^2}}|X(\mathbf{s})| + \frac{\sigma}{\sqrt{1+\lambda^2}}W(\mathbf{s}) - \frac{\sigma\lambda\sqrt{2}}{\sqrt{\pi(1+\lambda^2)}}. \tag{8}$$

Here, $X(\mathbf{s})$ and $W(\mathbf{s})$ are independent zero mean Gaussian process with isotropic parametric correlation functions, $\boldsymbol{\rho}_x(h;\boldsymbol{\psi}_x)$ and $\boldsymbol{\rho}_w(h;\boldsymbol{\psi}_w)$, respectively.

The process in (8) is not the only stationary skew Gaussian process proposed in the literature. However, Minozzo and Ferracuti (2012) point out that most of them are in fact not stationary. For example Kim and Mallick (2004) or Allard and Naveau (2007) built stochastic skew normal processes where the n-finite dimensional distributions have, as special case, the multivariate skew normal of Azzalini and Capitanio (1999). But, the class of multivariate skew normal of Azzalini and Capitanio (1999) is not closed under marginalization. Each marginal is still a skew normal but not of the same form, and Minozzo and Ferracuti (2012) demonstrate that the stationarity property of an n-dimensional finite distribution in this case is not passed onto the marginals. Note that if in (8) we let the process $X(\mathbf{s})$ to be spatially constant, i.e. $X(\mathbf{s}) \equiv X$, the associated n-finite dimensional distributions are the Azzalini and Capitanio (1999)'s multivariate skew normal and then, from above, the process is not stationary. On the other hand, if the process $W(\mathbf{s})$ is spatially constant, it is easy to demonstrate that (8) can be written as

$$Z(\mathbf{s}) = \mu + \frac{\sigma\lambda}{\sqrt{1+\lambda^2}}|X^*(\mathbf{s})| - \frac{\sigma\lambda\sqrt{2}}{\sqrt{\pi(1+\lambda^2)}},$$

where $X^*(\mathbf{s})$ is a process with finite dimensional distributions that are a mixture of folded normal with mode at $\mathbf{0}$ and covariance matrix that depends on the covariance matrix of $X(\mathbf{s})$ and on the parameters σ^2 and λ. As a consequence the resulting process is not a skew Gaussian process.

The correlation in each of the $X(\mathbf{s})$ and $W(\mathbf{s})$ processes induces association for the $\Theta(\mathbf{s})$ process. However, because circular variables have no *magnitude* (they only acquire a numerical value given an orientation), there is no unique way to define the correlation between two circular variables $\Theta(\mathbf{s})$ and $\Theta(\mathbf{s}')$. A common choice, which exhibits most of the desirable properties of a correlation, is the one proposed by Jammalamadaka and Sarma (1988), that is,

$$\frac{E[\cos(\Theta(\mathbf{s}) - \Theta(\mathbf{s}')|\boldsymbol{\Psi}) - \cos(\Theta(\mathbf{s}) + \Theta(\mathbf{s}') + 2\tilde{\mu}|\boldsymbol{\Psi})]}{2\sqrt{E(\sin^2(\Theta(\mathbf{s}) - \tilde{\mu})|\boldsymbol{\Psi})E(\sin^2(\Theta(\mathbf{s}') - \tilde{\mu})|\boldsymbol{\Psi})}}. \tag{9}$$

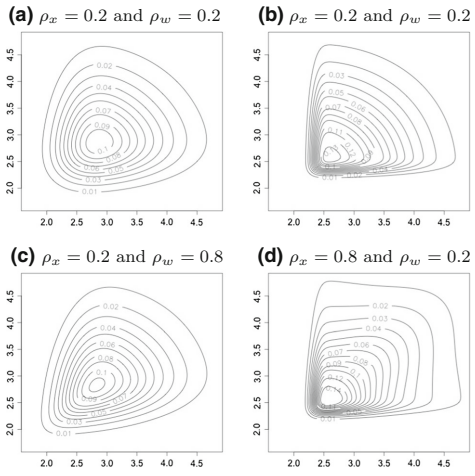


Fig. 2 Bivariate densities of the wrapped skew normal with $\mu = \pi$, $\sigma^2 = 1$, $\lambda = 3$ in the first column and $\lambda = 10$ in the second column and several values of ρ_x and ρ_w

In our setting (9) is not a valid correlation function; it is not a positive definite function. Moreover, we cannot compute (9) in closed form but, again, we can resort to Monte Carlo approximation. Figure 3 provides an illustrative display of the inline and corresponding circular correlations arising from the exponential correlation functions $\boldsymbol{\rho}_x(h;\boldsymbol{\psi}_x) = \exp(-h\psi_x)$ and $\boldsymbol{\rho}_w(h;\boldsymbol{\psi}_w) = \exp(-h\psi_w)$.

3.1 Implementation and kriging

Working directly with the wrapped skew Gaussian process is not feasible since the likelihood for a n-dimensional realization of the circular process involves n doubly infinite sums, i.e. let $\boldsymbol{\Theta} = (\Theta(\mathbf{s}_1),\ \Theta(\mathbf{s}_2), \ldots, \Theta(\mathbf{s}_n))'$ and $\mathbf{K} = (K(\mathbf{s}_1),\ K(\mathbf{s}_2), \ldots, K(\mathbf{s}_n))'$, the density of $\boldsymbol{\Theta}|\boldsymbol{\psi}$ is

$$f(\boldsymbol{\theta}|\boldsymbol{\Psi}) = \sum_{k_1\in\mathbb{Z}}\sum_{k_2\in\mathbb{Z}}\cdots\sum_{k_n\in\mathbb{Z}} g(\boldsymbol{\theta} + 2\pi\mathbf{k}|\boldsymbol{\Psi}),$$

where $g(\cdot|\boldsymbol{\Psi})$ is the density of $\mathbf{Z} = \boldsymbol{\Theta} + 2\pi\mathbf{K}$, the realization of the skew Gaussian process. When dealing with wrapped distributions the winding numbers are treated as latent random variables (see Jona Lasinio et al. 2012; Coles 1998, for details and ideas). Hence, the joint distribution of the circular variables and the winding numbers coincides with the joint distribution of the associated linear variables, i.e., $g(\cdot|\boldsymbol{\Psi})$, and we can work directly with the process $Z(\mathbf{s})$.

A critical point is the following. To simplify the model fitting, recalling (6) and (7) and extending them to n-variate random variables, $\mathbf{Z}|\mathbf{X}, \boldsymbol{\Psi}$ is normal, hence the process $Z(\mathbf{s})|X(\mathbf{s}), \boldsymbol{\Psi}$ is Gaussian and $\Theta(\mathbf{s})|X(\mathbf{s}), \boldsymbol{\Psi}$ is wrapped Gaussian. This implies that, in the model fitting, if we further introduce the realization of the latent Gaussian process, $X(\mathbf{s})$, along with the set of winding numbers, the $K(\mathbf{s}_i)$s, then the MCMC implementation follows directly from the work of Jona Lasinio et al. (2012) on the wrapped Gaussian process. In this setting, kriging is straightforward. More precisely, let $\mathbf{s}_0$ be the spatial location where we want to predict the circular process and let $\mathbf{X} = (X(\mathbf{s}_1), X(\mathbf{s}_2), \ldots, X(\mathbf{s}_n))'$. As is customary in the Bayesian framework, to perform kriging we draw samples from the predictive distribution of $\Theta(\mathbf{s}_0)|\boldsymbol{\Theta}$:

$$f(\Theta(\mathbf{s}_0)|\boldsymbol{\Theta}) = \sum_{\mathbf{K}\in\mathbb{Z}^n}\int_{\boldsymbol{\Psi}} f(\Theta(\mathbf{s}_0)|X(\mathbf{s}_0), \mathbf{X}, \mathbf{K}, \boldsymbol{\Psi}, \boldsymbol{\Theta}) \times f(X(\mathbf{s}_0)|\mathbf{X}, \boldsymbol{\Psi}) f(\mathbf{X}, \mathbf{K}, \boldsymbol{\Psi}|\boldsymbol{\Theta}) d\boldsymbol{\Psi}. \quad (10)$$

Let $\boldsymbol{\Psi}^b$, $\mathbf{X}^b$ and $\mathbf{K}^b$ be the b^{th} sample from the posterior distribution $f(\mathbf{X}, \mathbf{K}, \boldsymbol{\Psi}|\boldsymbol{\Theta})$. We can sample from (10) with composition sampling. That is, if for each posterior sample we simulate $X^b(\mathbf{s}_0)$ from the distribution $X(\mathbf{s}_0)|\mathbf{X}^b, \boldsymbol{\Psi}^b$ and $\Theta^b(\mathbf{s}_0)$ from the distribution $\Theta(\mathbf{s}_0)|X^b(\mathbf{s}_0),\ \mathbf{X}^b, \mathbf{K}^b, \boldsymbol{\Psi}^b, \boldsymbol{\Theta}$, then each $\Theta^b(\mathbf{s}_0)$ can be considered as a sample from (10).

We can easily simulate $X^b(\mathbf{s}_0)$ since $X(\mathbf{s}_0), \mathbf{X}^b|\boldsymbol{\Psi}^b$ is Gaussian and then $X(\mathbf{s}_0)|\mathbf{X}^b, \boldsymbol{\Psi}^b$ is univariate normal with mean and covariance that can be derived using standard results. If we simulate $Z^b(\mathbf{s}_0)$ from $Z(\mathbf{s}_0)|\mathbf{Z}^b$, $X^b(\mathbf{s}_0), \mathbf{X}^b, \boldsymbol{\Psi}^b$, where $\mathbf{Z}^b = \boldsymbol{\Theta} + 2\pi\mathbf{K}^b$, we can immediately obtain $\Theta^b(\mathbf{s}_0)$ as $\Theta^b(\mathbf{s}_0) = Z^b(\mathbf{s}_0) \bmod 2\pi$, that is a sample from $\Theta(\mathbf{s}_0)|X^b(\mathbf{s}_0),\ \mathbf{X}^b, \mathbf{K}^b, \boldsymbol{\Psi}^b, \boldsymbol{\Theta}$. Remark that to obtain a sample of $Z^b(\mathbf{s}_0)$ is really easy since

$$\begin{pmatrix} Z(\mathbf{s}_0) \\ \mathbf{Z} \end{pmatrix} |\mathbf{X}, X(\mathbf{s}_0), \boldsymbol{\Psi} \sim N\left(\begin{pmatrix} \mu^* + \dfrac{\sigma\lambda}{\sqrt{1+\lambda^2}}|X(\mathbf{s}_0)| \\ \mu^*\mathbf{1}_n + \dfrac{\sigma\lambda}{\sqrt{1+\lambda^2}}|\mathbf{X}| \end{pmatrix}, \ \frac{\sigma^2}{1+\lambda^2}\begin{pmatrix} 1 & \boldsymbol{\rho}'_{0,w} \\ \boldsymbol{\rho}_{0,w} & \Upsilon \end{pmatrix} \right)$$

where $\mathbf{1}_n$ is a vector of 1s of dimension n, $(\Upsilon)_{ij} = \rho_w(||\mathbf{s}_i - \mathbf{s}_j||;\boldsymbol{\psi}_w)$ and $(\boldsymbol{\rho}_{0,w})_i = \rho_w(||\mathbf{s}_i - \mathbf{s}_0||;\boldsymbol{\psi}_w)$. Then the distribution of $Z(\mathbf{s}_0)|\mathbf{Z}^b, X^b(\mathbf{s}_0), \mathbf{X}^b, \boldsymbol{\Psi}^b$ is normal.

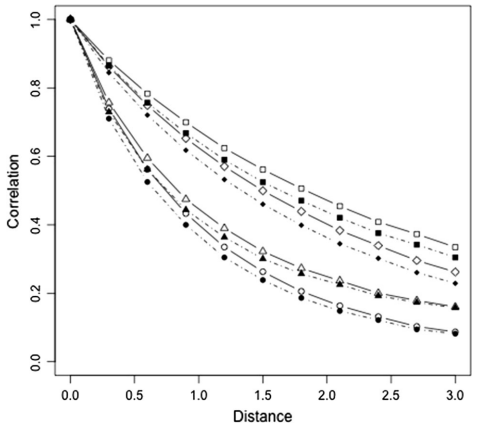


Fig. 3 Correlation functions for the inline (*empty symbols*) and circular (*solid symbols*) process with $\sigma^2 = 1$, $\delta = 0.95$ and exponential correlation function for the processes $X(\mathbf{s})$ and $W(\mathbf{s})$ with respectively decays parameters 0.5 and 0.5 (*circle*), 0.5 and 0.2 (*triangle*), 0.2 and 0.5 (*rhombus*), 0.2 and 0.2 (*square*)

4 A dynamic extension of the wrapped skew Gaussian process

We extend our model to the dynamic setting following ideas in Banerjee et al. (2014). We start by specifying an inline process $Z_t(\mathbf{s})$, $t \in [1, \ldots, T]$, as

$$Z_1(\mathbf{s}) = \mu + \frac{\sigma\lambda}{\sqrt{1+\lambda^2}}|X_1(\mathbf{s})| + \frac{\sigma}{\sqrt{1+\lambda^2}}W_1(\mathbf{s}) - \frac{\sigma\lambda\sqrt{2}}{\sqrt{\pi(1+\lambda^2)}}, \quad (11)$$

$$Z_t(\mathbf{s}) = \mu + \gamma(Z_{t-1}(\mathbf{s}) - \mu) + \frac{\sigma\lambda}{\sqrt{1+\lambda^2}}|X_t(\mathbf{s})| + \frac{\sigma}{\sqrt{1+\lambda^2}}W_t(\mathbf{s}) - \frac{\sigma\lambda\sqrt{2}}{\sqrt{\pi(1+\lambda^2)}}, \; t \neq 1, \quad (12)$$

where $\gamma \in [-1, 1]$, $\forall t$ we have $X_t(\mathbf{s})|\boldsymbol{\Psi} \sim GP(0, \boldsymbol{\rho}_x(h; \boldsymbol{\psi}_x))$, $W_t(\mathbf{s})|\boldsymbol{\Psi} \sim GP(0, \boldsymbol{\rho}_w(h; \boldsymbol{\psi}_w))$ and $\text{Cov}\,(X_t(\mathbf{s}), X_{t'}(\mathbf{s}')|\boldsymbol{\Psi}) = \text{Cov}\,(W_t(\mathbf{s}), W_{t'}(\mathbf{s}')|\boldsymbol{\Psi}) = 0$ if $t \neq t'$. Expressions (11) and (12) provide a mean-centered, first order auto-regressive model with i.i.d. process increments. Moreover, the process increments are skew GP's with parameters $\sigma, \lambda, \rho_x, \rho_w$. Equivalently, we see that $Z_1(\mathbf{s})|\boldsymbol{\Psi} \sim SGP(\mu, \sigma^2, \lambda)$ and $Z_t(\mathbf{s})|Z_{t-1}(\mathbf{s}), \boldsymbol{\Psi} \sim SGP(\mu + \gamma(Z_{t-1}(\mathbf{s}) - \mu), \sigma^2, \lambda)$.

Under the dynamic spatial setting, we are generally interested in predicting the process (i) at an observed spatial location at time $T + h$, $h \in \mathbb{Z}^+$ (usually $h = 1$) or (ii) at an unobserved spatial location $\mathbf{s}_0$ inside the observed time window. Suppose we let μ^b, $(\sigma^2)^b$, λ^b and γ^b be the samples of the parameters of the b^{th} iteration of the MCMC algorithm, $(\mu^*)^b = \mu^b - \sigma^b\lambda^b\sqrt{2}/\sqrt{\pi(1+(\lambda^b)^2)}$, $X_t^b(\mathbf{s})$ and $K_t^b(\mathbf{s})$ the b^{th} realization of the processes $X_t(\mathbf{s})$ and $K_t(\mathbf{s})$ at site $\mathbf{s}$ and time t and $Z_t^b(\mathbf{s}) = X_t(\mathbf{s}) + 2\pi K_t^b(\mathbf{s})$. B samples from the predictive distribution $\Theta_{T+h}(\mathbf{s})|\boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is the observed circular data, can be obtained if, for each MCMC sample, we draw a value $Z_{T+h}^b(\mathbf{s})$ from a normal distribution with mean

$$(\mu^*)^b + (\gamma^b)^h(Z_T^b(\mathbf{s}) - \mu^b) + \frac{\sigma^b\lambda^b}{\sqrt{1-(\lambda^b)^2}}|X_{T+h}^b(\mathbf{s})|$$

and variance

$$\frac{(\sigma^2)^b}{1-(\lambda^b)^2}.$$

The set $\{\Theta_{T+1}^b(\mathbf{s})\}_{b=1}^B$ is from the desired predictive distribution.

To obtain the b^{th} posterior sample of the predictive distribution of $\Theta_t(\mathbf{s}_0)|\boldsymbol{\Theta}$ we adopt the usual composition sampling by first sampling $X_t^b(\mathbf{s}_0)$ from the distribution of $X_t(\mathbf{s}_0)|\mathbf{X}, \boldsymbol{\Psi}^b$ and then sampling $Z_t^b(\mathbf{s}_0)$ from $Z_t(\mathbf{s}_0)|\mathbf{Z}, \mathbf{X}, X_t^b(\mathbf{s}_0), \boldsymbol{\Psi}^b$. Finally, $\Theta_t^b(\mathbf{s}_0) = Z_t^b(\mathbf{s}_0) \bmod 2\pi$ is a draw from the predictive distribution $\Theta_t(\mathbf{s}_0)|\boldsymbol{\Theta}$.

The distribution of $Z_t(\mathbf{s}_0), \mathbf{Z}|\mathbf{X}, X_t^b(\mathbf{s}_0), \boldsymbol{\Psi}^b$ is again multivariate normal and for spatial locations $\mathbf{s}_i, i = 1, 2, \ldots, n$, let $\mathbf{Z}_t = (Z_t(\mathbf{s}_1), Z_t(\mathbf{s}_2), \ldots, Z_t(\mathbf{s}_n))'$, $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_T)'$ and $\mathbf{X}$ be the associated realization of the process $X(\mathbf{s})$. Let $\boldsymbol{\Gamma}$ be a $T \times T$ correlation matrix with i, jth element equal to $\gamma^{|i-j|}$, $\boldsymbol{\Gamma}_l$ be the lower triangular part of $\boldsymbol{\Gamma}$ and $\mathbf{C}$ be the correlation matrix of $\mathbf{W}_t = (W_t(\mathbf{s}_1), W_t(\mathbf{s}_2), \ldots, W_t(\mathbf{s}_n))'$. Let $\mathbf{D}$ be a vector of length n with i^{th} element equal to $\text{Cor}\,(W_t(\mathbf{s}_0), W(\mathbf{s}_i))$, $\mathbf{F}_t$ be a vector of length T with ith element equal to $\gamma^{|t-i|}$, $\mathbf{I}_n$ be the identity matrix of dimension n and let $\otimes$ indicates the Kronecker product. Altogether, we have that

$$\begin{pmatrix} Z_t(\mathbf{s}_0) \\ \mathbf{Z} \end{pmatrix} | \mathbf{X}, X_t(\mathbf{s}_0), \boldsymbol{\Psi} \sim N\left(\begin{pmatrix} \mu^* + \frac{\sigma\lambda}{\sqrt{1-\lambda^2}}|X_t(\mathbf{s}_0)| \\ \boldsymbol{\delta} \end{pmatrix}, \frac{\sigma^2}{1-\lambda^2}\begin{pmatrix} 1 & (\mathbf{F}_t \otimes \mathbf{D})' \\ \mathbf{F}_t \otimes \mathbf{D} & \boldsymbol{\Gamma} \otimes \mathbf{C} \end{pmatrix} \right)$$

where

$$\boldsymbol{\delta} = \mu\mathbf{1}_{nT} + \sigma\lambda/\sqrt{1-\lambda^2}(\boldsymbol{\Gamma}_l \otimes \mathbf{I}_n)|\mathbf{X}| - \sigma\lambda\sqrt{2}/\sqrt{\pi(1+\lambda^2)}(\boldsymbol{\Gamma}_l \otimes \mathbf{I}_n)\mathbf{1}_{nT}.$$

Here, again $Z_t(\mathbf{s}_0)|\mathbf{Z}, \mathbf{X}, X_t^b(\mathbf{s}_0), \boldsymbol{\Psi}^b$ is univariate normal and sampling from it is easy.

5 A brief simulation study

We fit and estimate the model proposed in Sect. 4 to 8 datasets simulated with $\mu = \pi$, $\sigma^2 = 1$ and 4 levels of the skew parameter $\lambda = \{0.0, 1.5, 3, 10\}$. For the AR(1) parameter we chose $\gamma = 0.5$; we experimented with several values of $\gamma \in (0, 1)$ obtaining similar results, so we report estimates using the central value of the interval. We work with 2 sample sizes, 110 spatial locations and 60 time points, ($N = n \times T = 110 \times 60$), 220 spatial locations and 60 time points, $N = 220 \times 60$), to assess whether there are differences in the parameter estimates when the sample size increases. The coordinates are uniformly generated over $[0, 10]^2$ and for both processes, $X_t(\mathbf{s})$ and $W_t(\mathbf{s})$, we adopt exponential correlation functions. We choose $\psi_x = 0.5$ and $\psi_w = 0.2$ and notice that, as λ varies, we obtain different spatial correlations as shown in Fig. 4.

The model is estimated with 90 % of the spatial locations, i.e. 100 for the first sample size and 200 for the second, using the first 50 time points. Therefore, the training set is made of 100 × 50 and 200 × 50 points. We select observations using simple random sampling on the spatial locations (probability of inclusion in the training set 1 / n). The remaining spatial locations and 10 final time points are used to define two types of validation sets: (i) prediction at observed times, i.e. we use observations

between time 1 and time 50 not used to estimate the models. To simplify we call this set the *spatial validation set*; (ii) prediction at unobserved times, i.e. we use observations from time 51 to time 60 at spatial locations used to estimate the models. We call this set the *temporal validation set.* We repeat the sampling procedure 40 times.

As prior distributions we use $\mu \sim U(0, 2\pi)$, $\gamma \sim U(-1, 1)$, $\psi_x \sim U(0.1, 1)$ and $\psi_w \sim U(0.1, 1)$. To choose the prior on σ^2 and λ we note that, as for the wrapped Normal case (Jona Lasinio et al. 2012), if the variance of the associated inline distribution increases we are unable to tell the difference between the wrapped skew normal and a circular uniform. As we noted in Sect. 2.1, the variance of the skew normal is

$$\sigma^2\lambda^2/(1+\lambda^2)(1-2/\pi) + \sigma^2/(1+\lambda^2),$$

i.e., it is a function of both σ^2 and λ. In this regard, when $\sigma^2 = 10$, with sample size of 200, independently of λ, the Rayleigh test of (circular) uniformity fails to discriminate between the wrapped skew normal and the circular uniform. So, we chose $\sigma^2 \sim U(0, 10)$ and a weak informative prior for λ, $\lambda \sim N(0, 100)$.

For each dataset we also fit a wrapped normal model (setting $\lambda = 0$) and we compare the models with regard to posterior point estimates and predictive ability. The predictive ability of the models is evaluated by computing the continuous rank probability score (CRPS) for circular variables (Grimit et al. 2006). The CRPS is a proper scoring rules defined, for circular variables, as

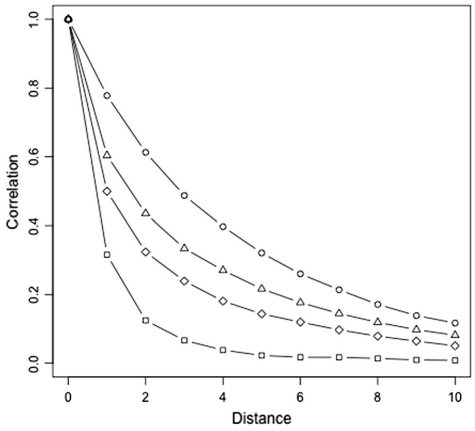


Fig. 4 Spatial correlation functions for the simulated datasets: circles are associated to Data1 ($\lambda = 0$), triangles to Data2 ($\lambda = 1.5$), diamonds to Data3 ($\lambda = 3$) and squares to Data4 ($\lambda = 10$)

$$CRPS(F, \xi) = E(d(\Xi, \xi)) - \frac{1}{2}E(d(\Xi, \Xi^*)), \tag{13}$$

where F is the predictive distribution, ξ is the holdout value, Ξ and Ξ^* are independent copies of a circular variable with distribution F, and $d(\Xi, \Xi^*) = 1 - \cos(\Xi - \Xi^*)$, the circular distance (Jammalamadaka and SenGupta 2001, p.15). Exact calculation of (13) is not possible since we can not obtain the predictive distribution under the skew or the non skew Gaussian process in closed form. However, for the validation point $\theta_t(\mathbf{s}_0)$ we can compute a Monte Carlo approximation as

$$\frac{1}{B}\sum_{b=1}^{B} d(\theta_t^b(\mathbf{s}_0), \theta_t(\mathbf{s}_0)) - \frac{1}{2B^2}\sum_{l=1}^{B}\sum_{b=1}^{B} d(\theta_t^l(\mathbf{s}_0), \theta_t^b(\mathbf{s}_0))$$

where $\theta_t^b(\mathbf{s}_0)$ denotes the simulated value of $\theta_t(\mathbf{s}_0)$ using the bth posterior parameters and B is the total number of posterior samples.

As an example, in Tables 1 and 2 we present the posterior mean estimates and credible intervals for all the parameters in all simulated datasets using one training set, i.e. the same locations and times for each dataset. For the fourth dataset and for both sample sizes, the skew model well estimates the parameters [the true value is inside the credible interval (C.I.)]. In the first dataset λ is far from 0. The wrapped skew normal process shows a substantial gain relative to the wrapped Gaussian process in terms of predictive ability for locations inside the observed time windows, even if the true model used to simulate the data is the wrapped Gaussian (Data1), see Table 3. As for forecasting (temporal validation set), we see that there is no difference between the models in terms of CPRS. Illustrative comparison of the predictive distributions under the two models can be seen in Fig. 5. As we expect,in the fourth dataset the predictive distribution is highly skewed while, in the first, it is essentially symmetric.

6 The wave direction data example

The real data we use come from a deterministic wave model implemented by Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA) that gives hourly prediction over a grid of about 12.5 × 12.5 Km on the Adriatic sea (Speranza et al. 2004). Over the Adriatic Sea area, there are 1494 points, with minimum and maximum distance of about 7 and 852 km respectively. The computer model starts from a wind forecast model predicting the surface wind over the entire Mediterranean and then the prediction of the wave direction is obtained solving energy transport equations using the wind forecast as input.

We developed two datasets. The first spans the period April 2010 between the 2nd at 00:00 and the 4th at 22:00, a

Table 1 Parameter estimates (mean) and credible intervals (C.I.) for the wrapped skew Gaussian model in the 4 simulated datasets

	Data1 ($\lambda = 0$)	Data2 ($\lambda = 1.5$)	Data3 ($\lambda = 3$)	Data4 ($\lambda = 10$)
$n = 110$				
$\hat{\mu}$	3.03	3.365	3.217	3.109
C.I.	(2.762 3.321)	(3.205 3.533)	(3.106 3.334)	(3.044 3.177)
$\hat{\sigma}^2$	1.715	1.213	1.061	0.962
C.I.	(1.390 2.186)	(1.080 1.388)	(0.976 1.177)	(0.888 1.046)
$\hat{\lambda}$	0.931	1.690	3.278	9.864
C.I.	(0.689 1.275)	(1.498 1.924)	(2.881 3.716)	(8.572 11.282)
$\hat{\gamma}$	0.388	0.446	0.499	0.488
C.I.	(0.35 0.42)	(0.421 0.470)	(0.479 0.518)	(0.475 0.502)
$\hat{\psi}_x$	0.234	0.399	0.472	0.528
C.I.	(0.139 0.483)	(0.332 0.473)	(0.413 0.528)	(0.475 0.589)
$\hat{\psi}_w$	0.144	0.254	0.191	0.210
C.I.	(0.109 0.186)	(0.195 0.318)	(0.141 0.251)	(0.137 0.307)
$n = 220$				
$\hat{\mu}$	2.981	3.353	3.209	3.094
C.I.	(2.713 3.261)	(3.209 3.504)	(3.067 3.346)	(3.031 3.161)
$\hat{\sigma}^2$	1.448	1.087	1.097	0.956
C.I.	(1.266 1.701)	(0.994 1.196)	(1.005 1.211)	(0.887 1.034)
$\hat{\lambda}$	-0.716	1.383	2.501	9.619
C.I.	(-0.869 -0.589)	(1.242 1.532)	(2.227 2.777)	(8.449 10.771)
$\hat{\gamma}$	0.370	0.436	0.488	0.499
C.I.	(0.349 0.390)	(0.418 0.452)	(0.474 0.503)	(0.490 0.507)
$\hat{\psi}_x$	0.430	0.558	0.500	0.511
C.I.	(0.323 0.625)	(0.485 0.639)	(0.444 0.558)	(0.467 0.555)
$\hat{\psi}_w$	0.152	0.286	0.192	0.152
C.I.	(0.119 0.186)	(0.235 0.340)	(0.143 0.245)	(0.112 0.212)

Table 2 Parameter estimates (mean) and credible intervals (C.I.) for the wrapped Gaussian model in the 4 simulated datasets

	Data1 ($\lambda = 0$)	Data2 ($\lambda = 1.5$)	Data3 ($\lambda = 3$)	Data4 ($\lambda = 10$)
$n = 110$				
$\hat{\mu}$	2.986	3.313	3.208	3.138
C.I.	(2.752 3.222)	(3.211 3.409)	(3.123 3.290)	(3.082 3.199)
$\hat{\sigma}^2$	1.141	0.596	0.465	0.369
C.I.	(0.993 1.340)	(0.556 0.645)	(0.438 0.497)	(0.35 0.39)
$\hat{\gamma}$	0.415	0.417	0.489	0.488
C.I.	(0.388 0.441)	(0.392 0.441)	(0.465 0.514)	(0.463 0.514)
$\hat{\psi}_w$	0.225	0.67	0.796	1.182
C.I.	(0.189 0.261)	(0.611 0.726)	(0.731 0.862)	(1.099 1.265)
$n = 220$				
$\hat{\mu}$	3.023	3.308	3.181	3.143
C.I.	(2.833 3.210)	(3.216 3.403)	(3.103 3.254)	(3.090 3.205)
$\hat{\sigma}^2$	1.061	0.602	0.473	0.370
C.I.	(0.937 1.209)	(0.564 0.647)	(0.449 0.501)	(0.354 0.390)
$\hat{\gamma}$	0.365	0.426	0.468	0.503
C.I.	(0.346 0.384)	(0.407 0.444)	(0.452 0.487)	(0.486 0.519)
$\hat{\psi}_w$	0.273	0.678	0.867	1.152
C.I.	(0.237 0.309)	(0.626 0.730)	(0.809 0.923)	(1.081 1.218)

Table 3 Simulated datasets: mean CRPSs over 40 validation sets. Models based on the wrapped skew normal (WS) and the wrapped normal (W)

		Data1	Data2	Data3	Data4
Spatial					
n = 110	WS	0.173	0.146	0.118	0.116
	W	0.221	0.179	0.176	0.160
n = 220	WS	0.170	0.149	0.116	0.091
	W	0.205	0.179	0.168	0.148
Temporal					
n = 110	WS	0.348	0.266	0.188	0.181
	W	0.349	0.265	0.191	0.184
n = 220	WS	0.350	0.275	0.193	0.181
	W	0.349	0.272	0.194	0.183

calm period. The second spans the period April 2010 between the 5th at 00:00 and the 7th at 22:00, a storm period.

We randomly select 220 spatial locations; the same spatial locations are used for the calm and storm period dataset.

Similarly to what we did in the simulated examples, we use 90 % of the spatial locations, taking the first 48 time points to estimate models while the remaining locations and times are included in the building of the two types of validation sets. Again, for each training set, we fitted a skew Gaussian model and a wrapped Gaussian model. We repeat the splitting procedure into training and validation sets 40 times and each time we compute the CRPS to compare the performance of the models.

As prior distributions we used the same choices as in Sect. 5 with the exception of the spatial decays; for ψ_w we adopt a $U(10^{-3}, 10^{-1})$ which corresponds to a maximum and minimum practical range of 3000 and 30km while for ψ_x we adopt a $U(5^{-4}, 5^{-2})$ which roughly corresponds to the same practical spatial range for the process $|X(\mathbf{s})|$.

In Table 4 we provide the parameters estimates for the first selected training sets. The estimated spatial dependence (ψ_w) of the $W(\mathbf{s})$ process is stronger during the storm for both models while (ψ_x) seems to remain the same in both sea states for $X(\mathbf{s})$. Again, employing the CRPS, for both validation sets under both sea states, the wrapped skew Gaussian process shows a consequential gain in predictive ability compared with the standard wrapped Gaussian, see Table 5.

Finally, Fig. 6 shows examples of predictive distributions for a holdout sample during a calm and a storm state. We showed in Fig. 1 that with $|\lambda| < 3$ there is little difference between the (symmetric) wrapped normal and the (asymmetric) wrapped skew normal. Since, in these two examples $|\hat{\lambda}| < 1.5$, the predictive distributions under the skew normal models are roughly symmetric.

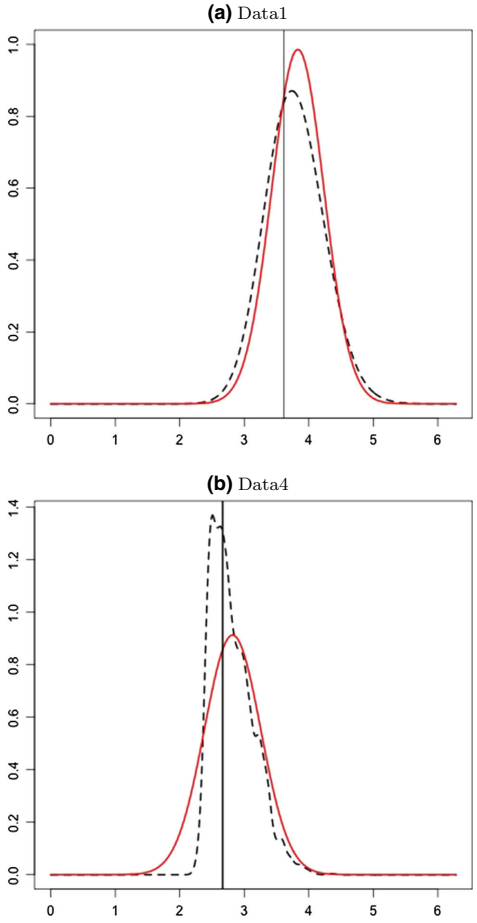


Fig. 5 Illustrative predictive distributions for a holdout site in the first (**a**) and in the fourth simulated dataset (**b**). The *solid line* is the predictive distribution under the wrapped Gaussian model while the *dashed one* is under the wrapped skew Gaussian model. The *vertical line* represents the true holdout simulated value

7 Summary and future work

We have presented a novel process model for dynamic spatial directional data. That is, we have a conceptual time series of directions at each spatial location in the region

Table 4 Parameter estimates (mean) and credible intervals (C.I.) for the wave direction data

	Calm WS	Calm W	Storm WS	Storm W
$\hat{\mu}$	3.372	3.19	3.398	3.39
C.I.	(2.610 4.150)	(2.905 3.500)	(2.498 4.274)	(2.939 3.851)
$\hat{\sigma}^2$	5.246	1.827	5.015	1.283
C.I.	(4.214 6.883)	(1.526 2.276)	(4.029 6.581)	(1.130 1.477)
$\hat{\lambda}$	1.432	·	1.159	·
C.I.	(1.068 1.762)	(· ·)	(0.868 1.496)	(· ·)
$\hat{\gamma}$	0.438	0.567	0.377	0.479
C.I.	(0.406 0.471)	(0.540 0.594)	(0.350 0.406)	(0.453 0.504)
$\hat{\psi}_x$	0.006	·	0.006	·
C.I.	(0.005 0.008)	(· ·)	(0.005 0.008)	(· ·)
$\hat{\psi}_w$	0.002	0.013	0.001	0.007
C.I.	(0.001 0.003)	(0.011 0.015)	(0.001 0.001)	(0.005 0.008)

Table 5 Wave data: mean CRPSs over 40 validation sets. Models based on the wrapped skew normal (WS) and the wrapped normal (W)

	Calm WS	Calm W	Storm WS	Storm W
Spatial	0.426	0.494	0.528	0.567
Temporal	0.520	0.628	0.446	0.476

and we observe these series for a finite collection of locations. The model, referred to as a wrapped skew Gaussian process, enables more flexible marginal distributions for the locations than the symmetric ones that are available under the previously published wrapped Gaussian process. Using both simulation and a wave direction dataset, we are able to show improved out-of-sample prediction with the former.

Future work offers several opportunities. One is to note that wave heights are available in addition to wave directions. Wave heights inform about the sea state and therefore whether we are in a calm, storm, or transition state. In particular, predictive uncertainty varies with wave height and/or sea state, e.g., prediction is more precise during storm. So, we can attempt to extend the proposed model to introduce covariates into the mean model and also into the variance model for the wrapped skew Gaussian process. Another possibility is to model temporal data, where the time of the observed event is treated as random. Then, upon wrapping, we would have circular times. In addition, the locations of the events are random. The data would be treated as a point pattern over space and (circular) time.

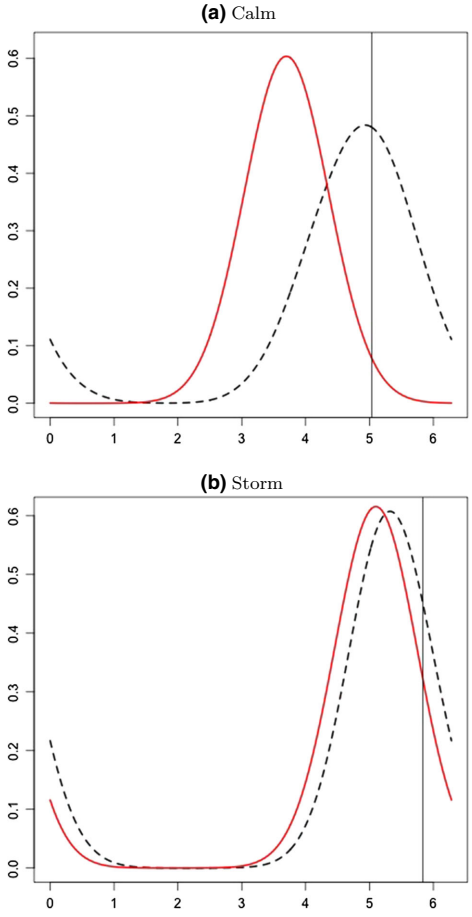


Fig. 6 Examples of predictive distributions for one of the holdout site in calm (**a**) and storm (**b**) sea state. The *solid line* is the predictive distribution under the wrapped Gaussian model while the *dashed one* is under the skew Gaussian model. The *vertical line* represents the true holdout observed value

References

Allard D, Naveau P (2007) A new spatial skew-normal random field model. Commun Stat 36(9):1821–1834

Azzalini A (1985) A class of distributions which includes the normal ones. Scand J Stat 12:171–178
Azzalini A (2005) The skew-normal distribution and related multivariate families. Scand J Stat 32(2):159–188
Azzalini A, Capitanio A (1999) Statistical applications of the multivariate skew normal distribution. J R Stat Soc Ser B 61(3):579–602
Azzalini A, Valle AD (1996) The multivariate skew-normal distribution. Biometrika 83(4):715–726
Banerjee S, Gelfand AE, Carlin BP (2014) Hierarchical modeling and analysis for spatial data, 2nd edn. Chapman and Hall/CRC, New York
Bao L, Gneiting T, Grimit EP, Guttorp P, Raftery AE (2009) Bias correction and bayesian model averaging for ensemble forecasts of surface wind direction. Mon Weather Rev 138(5):1811–1821
Breckling J (1989) The analysis of directional time series: applications to wind speed and direction. Lecture notes in statistics. Springer, Berlin Heidelberg
Coles S (1998) Inference for circular distributions and processes. Stat Comput 8(2):105–113
Coles S, Casson E (1998) Extreme value modelling of hurricane wind speeds. Struct Saf 20(3):283–296
Damien P, Walker S (1999) A full Bayesian analysis of circular data using the von Mises distribution. Can J Stat 27:291–298
Engel C, Ebert E (2007) Performance of hourly operational consensus forecasts (OCFs) in the australian region. Weather Forecast 22(6):1345–1359
Fisher N, Lee A (1994) Time series analysis of circular data. J R Stat Soc Ser B 56(327):332
Fisher NI (1996) Statistical analysis of circular data. Cambridge University Press, Cambridge
Fisher NI, Lee AJ (1992) Regression models for an angular response. Biometrics 48(3):665–677
Grimit EP, Gneiting T, Berrocal VJ, Johnson NA (2006) The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. Quart J R Meteorol Soc 132(621C):2925–2942
Guttorp P, Lockhart RA (1988) Finding the location of a signal: a Bayesian analysis. J Am Stat Assoc 83(402):322–330
Harrison D, Kanji GK (1988) The development of analysis of variance for circular data. J Appl Stat 15:197–224
Hernández-Sánchez E, Scarpa B (2012) A wrapped flexible generalized skew-normal model for a bimodal circular distribution of wind direction. Chil J Stat 3(2):129–141
Holzmann H, Munk A, Suster M, Zucchini W (2006) Hidden Markov models for circular and linear-circular time series. Environ Ecol Stat 13(3):325–347
Hughes G (2007) Multivariate and time series models for circular data with applications to protein conformational angles. PhD thesis, University of Leeds, Leeds
Jammalamadaka S, Sarma Y (1988) A correlation coefficient for angular variables. Stat Theory Data Anal II:349–364
Jammalamadaka SR, SenGupta A (2001) Topics in circular statistics. World Scientific, Singapore
Jona Lasinio G, Gelfand A, Jona Lasinio M (2012) Spatial analysis of wave direction data using wrapped Gaussian processes. Ann Appl Stat 6(4):1478–1498
Kato S (2010) A Markov process for circular data. J R Stat Soc Ser B (Stat Methodol) 72(5):655–672
Kato S, Shimizu K (2008) Dependent models for observations which include angular ones. J Stat Plan Inference 138(11):3538–3549
Kato S, Shimizu K, Shieh GS (2008) A circular-circular regression model. Stat Sin 18:633–645
Kim HM, Mallick BK (2004) A Bayesian prediction using the skew gaussian distribution. J Stat Plan Inference 120(1–2):85–101
Lagona F, Picone M, Maruotti A, Cosoli S (2015) A hidden Markov approach to the analysis of space–time environmental data with linear and circular components. Stoch Environ Res Risk Assess 29(2):397–409
Ma Y, Genton MG (2004) A flexible class of skew-symmetric distributions. Scand J Stat 31:459–468
Mardia KV (1972) Statistics of directional data. Academic Press, London, New York
Mardia KV, Jupp PE (1999) Directional statistics. Wiley, Chichcster
Mastrantonio G, Jona Lasinio G, Gelfand AE (2015) Spatio-temporal circular models with non-separable covariance structure. TEST To appear
Minozzo M, Ferracuti L (2012) On the existence of some skew-normal stationary process. Chil J Stat 3(2):157–170
Nuñez-Antonio G, Gutiérrez-Peña E (2005) A Bayesian analysis of directional data using the projected normal distribution. J Appl Stat 32(10):995–1001
Pewsey A (2000) The wrapped skew-normal distribution on the circle. Commun Stat 29(11):2459–2472
Pewsey A (2006) Modelling asymmetrically distributed circular data using the wrapped skew-normal distribution. Environ Ecol Stat 13(3):257–269
Presnell B, Morrison SP, Littell RC (1998) Projected multivariate linear models for directional data. J Am Stat Assoc 93(443):1068–1077
Ravindran P, Ghosh SK (2011) Bayesian analysis of circular data using wrapped distributions. J Stat Theory Pract 5(4):547–561
Sahu SK, Dey DK, Branco MD (2003) A new class of multivariate skew distributions with applications to bayesian regression models. Can J Stat 31(2):129–150
Speranza A, Accadia C, Casaioli M, Mariani S, Monacelli G, Inghilesi R, Tartaglione N, Ruti PM, Carillo A, Bargagli A, Pisacane G, Valentinotti F, Lavagnini A (2004) Poseidon: an integrated system for analysis and forecast of hydrological, meteorological and surface marine fields in the Mediterranean area. Nuovo Cimento 27(C):329–345
Wang F, Gelfand AE (2013) Directional data analysis under the general projected normal distribution. Stat Methodol 10(1):113–127
Wang F, Gelfand AE (2014) Modeling space and space–time directional data using projected Gaussian processes. J Am Stat Assoc 109(508):1565–1580
Wang J, Boyer J, Genton MG (2004) A skew-symmetric representation of multivariate distributions. Stat Sin 14:1259–1270
Zhang H, El-Shaarawi A (2010) On spatial skew-Gaussian processes and applications. Environmetrics 21(1):33–47
Zhang Q, Snow Jones A, Rijmen F, Ip EH (2010) Multivariate discrete hidden Markov models for domain-based measurements and assessment of risk factors in child development. J Comput Gr Stat 19(3):746–765

Chapter 4

Hidden Markov models for circular-linear data

Often circular data take the form of a time series and they are observed along with linear variables (see for example Holzmann *et al.* (2006) or Bulla *et al.* (2012)). In this Chapter we propose HMMs for circular-linear time series. Similarly to what we did for the stochastic processes (Chapter 3), we specify the models in term of multivariate linear variables and then, using the transformation (2.6) to a subset of them, we obtain a model for circular-linear variables.

Let $\boldsymbol{\pi}_k = (\pi_{k1}, \ldots, \pi_{kK})$ be a vector of probabilities, i.e. $\sum_{j=1}^{K} \pi_{kj} = 1$, let $z_t \in [1, \ldots, K]$ be a discrete variable and assume that $\mathbf{z} = (z_1, \ldots, z_n)'$ follows a Markov process of the first order, i.e.

$$P(z_t|z_{t-1}, \ldots, z_1) = P(z_t|z_{t-1}) = \pi_{z_{t-1}z_t},$$

Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_T)'$, where $\mathbf{y}_t = (y_{t1}, \ldots, y_{t2p})' \in \mathbb{R}^{2p}$, and $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)'$, where $\mathbf{x}_t = (x_{t1}, \ldots, x_{tq})' \in \mathbb{R}^q$, be time series of multivariate linear variables, and suppose that the pdf of $\mathbf{y}_t, \mathbf{x}_t|\boldsymbol{\psi}_{z_t}$ is $f_{\mathbf{Y}_t}(\cdot|\mathbf{X}_t, \boldsymbol{\psi}_{z_t}) f_{\mathbf{X}_t}(\cdot|\boldsymbol{\psi}_{z_t})$.

The HMM (Zucchini and MacDonald, 2009) is specified as

$$f_{\mathbf{Y},\mathbf{X}}(\mathbf{y}, \mathbf{x}|\mathbf{z}, \{\boldsymbol{\psi}_k\}_{k=1}^K) = \prod_{t=1}^{T} \prod_{k=1}^{K} [f_{\mathbf{Y}_t}(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\psi}_k) f_{\mathbf{X}_t}(\mathbf{x}_t|\boldsymbol{\psi}_k)]^{I_{(z_t=k)}}$$
$$z_t|z_{t-1} \sim \boldsymbol{\pi}_{z_t-1},$$

where $I_{(z_t=k)}$ is a function that assumes value 1 if $z_t = k$, and 0 otherwise.

Using (2.6), i.e. the projection, we can obtain an HMM for $p-$variate circular and $q-$variate linear variables from the HMM for $(2p + q)$-variate linear variables specified

above. More precisely let

$$\theta_{ti} = \text{atan}^* \left(\frac{y_{2i}}{y_{2i-1}} \right), i = 1, \ldots, p,$$

$\boldsymbol{\theta}_t = (\theta_{t1}, \ldots \theta_{tp})' \in [0, 2\pi)^p$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T)'$. The HMM that arises by applying the projection is

$$\begin{aligned} f_{\boldsymbol{\theta},\mathbf{x}}(\boldsymbol{\theta}, \mathbf{x}|\mathbf{z}, \{\boldsymbol{\psi}_k\}_{k=1}^K) &= \prod_{t=1}^T \prod_{k=1}^K [f_{\boldsymbol{\Theta}_t}(\boldsymbol{\theta}_t|\mathbf{x}_t, \boldsymbol{\psi}_k) f_{\mathbf{X}_t}(\mathbf{x}_t|\boldsymbol{\psi}_k)]^{I_{(z_t=k)}} \\ z_t|z_{t-1} &\sim \boldsymbol{\pi}_{z_t-1}, \end{aligned} \tag{4.1}$$

The circular-linear density that arises by projecting is not easy to work with, see Chapter 2. Let $r_{ti} = ||(y_{2i}, y_{2i-1})'||$, the inference can be simplified if we introduce the time series of p-variate (latent) variables $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_n)'$, where $\mathbf{r}_i = (r_{i1}, \ldots, r_{ip})'$. The augmented HMM is

$$\begin{aligned} f_{\boldsymbol{\theta},\mathbf{r},\mathbf{x}}(\boldsymbol{\theta}, \mathbf{r}, \mathbf{x}|\mathbf{z}, \{\boldsymbol{\psi}_k\}_{k=1}^K) &= \prod_{t=1}^T \prod_{k=1}^K \left[f_{\mathbf{Y}_t}(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\psi}_k) \prod_{j=1}^p r_{tp} f_{\mathbf{X}_t}(\mathbf{x}_t|\boldsymbol{\psi}_k) \right]^{I_{(z_t=k)}} \\ z_t|z_{t-1} &\sim \boldsymbol{\pi}_{z_t-1}, \end{aligned} \tag{4.2}$$

where $\mathbf{y}_t$ must be seen as a function of $\boldsymbol{\theta}_t$ and $\mathbf{r}_t$. A marginalization over $\mathbf{r}$ in (4.2) gives (4.1).

Bibliography

Bulla, J., Lagona, F., Maruotti, A., and Picone, M. (2012). A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series. *Journal of Agricultural, Biological, and Environmental Statistics*, **17**(4), 544–567.

Holzmann, H., Munk, A., Suster, M., and Zucchini, W. (2006). Hidden Markov models for circular and linear-circular time series. *Environmental and Ecological Statistics*, **13**(3), 325–347.

Zucchini, W. and MacDonald, I. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Research Article **Environmetrics**

Received: 1 September 2014, Revised: 18 November 2014, Accepted: 18 November 2014, Published online in Wiley Online Library: 9 January 2015

(wileyonlinelibrary.com) DOI: 10.1002/env.2326

Bayesian hidden Markov modelling using circular-linear general projected normal distribution

Gianluca Mastrantonio[a]*, Antonello Maruotti[b,c] and Giovanna Jona-Lasinio[d]

We introduce a multivariate hidden Markov model to jointly cluster time-series observations with different support, that is, circular and linear. Relying on the general projected normal distribution, our approach allows for bimodal and/or skewed cluster-specific distributions for the circular variable. Furthermore, we relax the independence assumption between the circular and linear components observed at the same time. Such an assumption is generally used to alleviate the computational burden involved in the parameter estimation step, but it is hard to justify in empirical applications. We carry out a simulation study using different data-generation schemes to investigate model behavior, focusing on well recovering the hidden structure. Finally, the model is used to fit a real data example on a bivariate time series of wind speed and direction.

Keywords: directional data; hidden Markov models; Markov chain Monte Carlo; wind data; multivariate time series; projected normal distribution

1. INTRODUCTION

Hidden Markov models (HMMs) have become more frequently used to provide a natural and flexible framework for univariate and multivariate time-dependent data (e.g. time-series and longitudinal data). They are a class of mixture models in which the data-generation distribution depends on the state of an underlying and unobserved Markov process. Hidden Markov modelling has been used as a statistical tool for density estimation (Langrock *et al.*; Dannemann, 2012), supervised and unsupervised classification (Lagona and Picone, 2012; Alfò and Maruotti, 2010; Frühwirth Schnatter, 2006), and a wide range of empirical problems in environmetrics (Martinez-Zarzoso and Maruotti, 2013; Langrock *et al.*, 2012), medicine (Langrock *et al.*, 2013; Lagona *et al.*, 2014), and education Bartolucci *et al.* (2011). For a comprehensive introduction to fundamental theory of HMMs encountered in practice, see the review papers of Bartolucci *et al.* (2014), Maruotti (2011), and monographs by Bartolucci *et al.* (2012), Zucchini and MacDonald (2009) and Cappé *et al.* (2005).

The literature on multivariate hidden Markov modelling is dominated by Gaussian HMMs (Spezia, 2010; Bartolucci and Farcomeni, 2010; Geweke and Amisano, 2011). Modelling multivariate time series with non-normal components of mixed-type is challenging. The joint distribution of multivariate (mixed-type) data is usually specified as a mixture having products of univariate distributions as components (see e.g. Lagona *et al.*, 2011; Lagona and Picone, 2011; Zhang *et al.*, 2010). Bartolucci and Farcomeni (2009) is a notable exception. In other words, random variables are assumed conditionally independent given the latent structure. Although conditional independence facilitates parameters estimation, it is a too restrictive assumption in many empirical applications and may not properly accommodate for the complex shape of multivariate distributions (Baudry *et al.*, 2010). Moreover, an unnecessary number of latent states is often needed to obtain reasonable fit, at the price of an increased computational burden and difficulties in interpreting results, as shown in the simulation study section.

In this paper, we propose a bivariate distribution for circular-linear time-series in a HMM framework. We accommodate for nonstandard features of data including correlation in time and across variables, mixed supports (circular and linear) of the data, the special nature of circular measurements, and the occurrence of missing values. We relax the conditional independence assumption between circular and linear variables by taking a fully parametric approach.

* *Correspondence to: G. Mastrantonio, Department of Economics, University of Roma Tre, Rome, Italy. E-mail: gianluca.mastrantonio@uniroma3.it*

a *Department of Economics, University of Roma Tre, Rome, Italy*

b *Southampton Statistical Science Research Institute, University of Southampton, Southampton, U.K.*

c *Department of Political Science, University of Roma Tre, Rome, Italy*

d *Department of Statistical Science, Sapienza University of Rome, Rome, Italy*

This is not the first attempt to jointly modelling circular and linear variables in an HMM framework. Bulla *et al.* (2012) introduced a latent-class approach to the analysis of multivariate mixed-type data by assuming that circular and linear variables are conditionally independent given the states visited by a latent Markov chain while Kato *et al.* (2008) propose a hyper-cylindrical distribution. The latter is problematic (in the HMM setting in particular), because little is known about efficient estimation procedures and identifiability issues under hyper-cylindrical parametric models. In addition, mixtures of hyper-toroidal densities would group data according to clusters of difficult interpretation, without necessarily improving the fit of the model.

We introduce a flexible structure, relying on the general projected normal (PN) distribution Wang and Gelfand (2014), to model circular measurements, and extending Bulla *et al.* (2012) to a more general setting, allowing for (conditional) correlation between circular and linear variables. We treat the circular response as projection onto the unit circle of a bivariate variable and define the joint circular-linear distribution through the specification of a multivariate model in a multivariate linear setting, extending Wang *et al.* (2014) to a clustering framework.

The resulting hidden Markov model parameters are estimated in a Bayesian framework. We provide details on how to fit the model by using Markov chain Monte Carlo (MCMC) methods, and we point out possible drawbacks in the implementation of the algorithm. Advantages of the Bayesian approach, with respect to the expectation maximization (EM) algorithm, include a convenient framework to simultaneously account for several data features, adjust for identifiability issues, and produce natural measures of uncertainty for model parameters. For a general discussion see, for example, (Rydén and Titterington, 1998; Rydén, 2008; Yildirim *et al.*, 2014).

We illustrate the proposal by a large-scale simulation study in order to investigate the empirical behavior of the proposed approach with respect to several factors, such as the number of observed times, the association structure between the circular and linear variables, and the fuzziness of the classification. We evaluate model performance in recovering the true model structure, we compare several models on the basis of their ability to accurately estimate the vectors of state-dependent parameters and hidden parameters. Finally, we test the proposal by analysing time series of semi-hourly wind directions and speeds, recorded in the period 12/12/2009, 12/1/2010 by the buoy of Ancona, located in the Adriatic Sea at about 30 km from the coast.

The rest of the paper is organized as follows. In Section 2, we briefly review relevant aspects necessary for the introduction of our approach and outline some results about the PN distribution. Section 3 discusses the specification of the circular-linear general PN hidden Markov model and provides Bayesian inference. Computational details and parameters estimation are discussed as well. Section 4 presents a large-scale simulation study. In Section 5, the application of the proposed methodology is illustrated through a real-world data set. Some concluding remarks are given in Section 6.

2. PRELIMINARIES

Circular data are a particular class of directional data, specifically, they are directions in two dimensions. To analyze circular data is challenging because usual statistics, which have been developed for linear data (for example, the mean and variance), will not be meaningful and will be misleading when applied to directional data without taking into account the particular definition of the domain. There are many ways to define distributions in a circular domain; see the book of Mardia and Jupp (1999) for a comprehensive overview. The one we used in this paper is to radially project onto the circle a probability distribution originally defined on the plane. Let $\mathbf{Z} = [Z_1, Z_2]'$ be a 2D random vector such that $\Pr(\mathbf{Z} = \mathbf{0}) = 0$. Then, its radial projection $\mathbf{W} = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} = \frac{\mathbf{Z}}{||\mathbf{Z}||}$ is a random vector on the unit circle, which can be converted to a random angle X relative to some direction treated as 0 via the transformation $X = \arctan^* \frac{W_2}{W_1} = \arctan^* \frac{Z_2}{Z_1} \in [0, 2\pi)$, where the function $\arctan^*$ is a quadrant specific inverse of the tangent function, sometimes called *atan2*, that takes into account the signs of W_1 and W_2 to identify the right quadrant of X; for a formal definition, see Jammalamadaka and SenGupta (2001), pag. 13. Note that $\mathbf{W} = \begin{bmatrix} \cos X \\ \sin X \end{bmatrix}$ and let $R = ||Z||$ the following relation holds: $\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} R\cos X \\ R\sin X \end{bmatrix} = R\mathbf{W}$.

By assuming $\mathbf{Z} \sim N_2(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix}$ and $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, X is said to have a 2D PN distribution, denoted by $PN_2(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Because the distribution of X does not change if we multiply $\mathbf{Z}$ for a positive constant $c > 0$, for identifiability purposes, following Wang and Gelfand (2012), σ_2^2 is set to be 1. The PN distribution is specified as a four-parameter distribution: $PN_2(\cdot|\mu_1, \mu_2, \sigma_1^2, \rho)$.

We provide some examples to illustrate the flexibility of the PN distribution. The PN density can be symmetric, asymmetric, or possibly bimodal, and apart from some special case, the interpretation of the parameters can be difficult. The number of modes and the shape depend on the value of the all parameters and different sets of parameters can give really similar shapes. As a general comment, we highlight that μ_1 and μ_2 are the means of the two Cartesian coordinates z_1 and z_2 and are, respectively, connected to the cosine and sine of the circular variable. By fixing $\sigma_1^2 = 1$ and $\rho = 0$, the resulting distribution is unimodal, and symmetric and if $\mu_1 = \mu_2 = 0$, the distribution becomes a circular uniform; see Figure 1 in the online supplement material. Departure from zero for the two means, in the case of identity covariance matrix, creates one mode in the trigonometric quadrant with the same sign of the means, for example, if $\mu_1 > 0$ and $\mu_2 < 0$, then the mode is in the quadrant with positive cosine and negative sine; higher values of a mean attract the mode to its correspondent axis; see Figure 1 in the online supplement material. By allowing the ρ parameter to vary, we obtain very flexible shapes. The resulting distribution shows asymmetry with more mass of probability near the axis with the highest μ. By increasing $|\rho|$, bimodality is detected; Figure 1. Moreover, for $\sigma_1^2 < 1$, the modes are closer to the sine axis; whereas for $\sigma_1^2 > 1$, the modes are closer to the cosine axis; see Figures in the online supplement material.

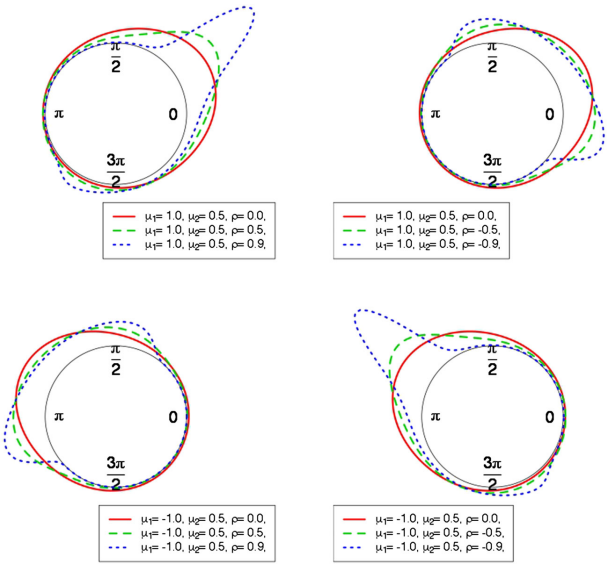


Figure 1. Shape of the projected normal distribution for $\sigma^2 = 1$ and different values of ρ, μ_1 and μ_2

3. THE CIRCULAR-LINEAR GENERAL PROJECTED NORMAL HIDDEN MARKOV MODEL

3.1. The model

Let $\mathcal{T} = \{0, 1, \ldots, T-1, T\}$, in this paper, we consider a bivariate time series $[\mathbf{x}, \mathbf{y}] = \{[x_t, y_t]; t \in \mathcal{T}\backslash\{0\}\}$ with circular, x_t, and linear, y_t, components. Our aim is to jointly classify $[x_t, y_t]$ in K classes, generally called regimes or states, with a HMM-based classifier.
Let $\pi_{k,h}$ indicate the probability to move from state k to state h, and let ξ_{tk} be an indicator variable such that if we are in state k on time t it is 1, otherwise is 0. Then, $f(\xi_{th} = 1|\xi_{t-1k} = 1) = \pi_{k,h}$ and we set $f(\xi_{0k}) = \pi_k$. We indicate with

$$\boldsymbol{\pi} = \begin{bmatrix} \pi_{1,1} & \pi_{1,2} & \cdots & \pi_{1,K} \\ \pi_{2,1} & \pi_{2,2} & \cdots & \pi_{2,K} \\ \cdots & \cdots & \cdots & \cdots \\ \pi_{K,1} & \pi_{K,2} & \cdots & \pi_{K,K} \end{bmatrix}, \sum_{h=1}^{K} \pi_{k,h} = 1,\ k = 1, 2, \ldots, K,$$

the transition matrix that governs the evolution of the Markov chain, $\boldsymbol{\pi_0} = [\pi_1, \pi_2, \ldots, \pi_K]'$ and $\boldsymbol{\xi} = [\boldsymbol{\xi}_0, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_T]'$ where $\boldsymbol{\xi}_t = [\xi_{t1}, \xi_{t2}, \ldots, \xi_{tK}]'$.

Let $n_{k,h} = \sum_{t=1}^{T} \xi_{t-1k}\xi_{th}$ be the number of times we move from state k to state h, the joint density of the vector of states is $f(\boldsymbol{\xi}|\boldsymbol{\pi}, \boldsymbol{\pi_0}) = \prod_{k=1}^{K} \pi_k^{\xi_{0k}} \prod_{k=1}^{K} \prod_{h=1}^{K} \pi_{k,h}^{n_{k,h}}$. In the literature on HMM for circular-linear variables, see for example Bulla *et al.* (2012) and Holzmann *et al.* (2006), it is generally assumed that conditioning to the latent vector $\boldsymbol{\xi}$, the pairs $[x_t, y_t]$ and $[x_g, y_g]$ are independent if $g \neq t$ and at the same time $x_t \perp y_t$. As a result, the conditional distribution of the observed process, given the latent process, takes the form of a product density, say $f(\mathbf{x}, \mathbf{y}|\boldsymbol{\xi}) = \prod_{k=1}^{K} \prod_{t=1}^{T} [f(x_t|\xi_{tk} = 1) f(y_t|\xi_{tk} = 1)]^{\xi_{tk}}$. We maintain the so-called conditional independence

property: given the hidden state at time t, the distribution of the observation at this time is fully determined, but we relax the assumption on independence between the circular and linear variables observed at the same time. Thus, we get a multivariate conditional distribution $f(\mathbf{x},\mathbf{y}|\boldsymbol{\xi}) = \prod_{k=1}^{K}\prod_{t=1}^{T} f(x_t, y_t|\xi_{tk}=1)^{\xi_{tk}}$.

Let $\mathbf{Z}_t|\xi_{tk} = 1 \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, with $\mathbf{Z}_t = \begin{bmatrix} Z_{t1} \\ Z_{t2} \end{bmatrix}$, $\boldsymbol{\mu}_k = \begin{bmatrix} \mu_{k1} \\ \mu_{k2} \end{bmatrix}$ $\boldsymbol{\Sigma}_k = \begin{bmatrix} \sigma_{k1}^2 & \sigma_{k1}\rho_k \\ \sigma_{k1}\rho_k & 1 \end{bmatrix}$ and let $R_t = ||\mathbf{Z}_t||$. We define X_t as the radial projection of $\mathbf{Z}_t$: $X_t = \arctan^* \frac{Z_1}{Z_2}$ and then $X_t|\xi_{tk} = 1 \sim PN_2(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. We can write easily the joint density of $[X_t, R_t]$ that is the density that arises by a variable transformation from the bivariate normal Z_t to its polar system representation. Let $\phi_h(\zeta|\mathbf{M},\mathbf{V})$ be the probability density function of an h-variate normal distribution with mean $\mathbf{M}$ and covariance matrix $\mathbf{V}$ evaluated in ζ, then

$$f(x_t, r_t|\xi_{tk} = 1) = \phi_2(r_t\mathbf{w}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)r_t \tag{1}$$

We built the (conditional) joint density $f(x_t, y_t|\xi_{tk} = 1) = f(y_t|x_t, \xi_{tk} = 1)f(x_t|\xi_{tk} = 1)$ as a marginalization over the latent variable R_t: $f(x_t, y_t|\xi_{tk} = 1) = \int_{r_t} f(y_t|x_t, r_t, \xi_{tk} = 1)f(x_t, r_t|\xi_{tk} = 1)dr_t$, where $f(x_t, r_t|\xi_{tk} = 1)$ is specified in Eqn. 1, and $y_t|x_t, r_t, \xi_{tk} = 1$ is defined through a circular-linear regression. However, there is not an obvious and standard way to formalize the relations between circular and linear variables. Jammalamadaka and SenGupta (2001) propose a flexible approach using trigonometric polynomials while Mardia (1976) and Johnson and Wehrly (1978) proposed a regression where the covariates are the sine and cosine components of the circular variable. Here, following Wang *et al.* (2014), we specify the relation as $y_t = \gamma_{k0} + \gamma_{k1}r_t\cos x_t + \gamma_{k2}r_t\sin x_t + \epsilon_{tk}$, with $\epsilon_{tk} \sim N(0, \sigma_{ky})$. Thus, $y_t|x_t, r_t, \xi_{tk} = 1$ is distributed as a normal variable with mean $\gamma_{k0} + \gamma_{k1}r_t\cos x_t + \gamma_{k2}r_t\sin x_t$ and variance σ_{ky}^2. Note that the regression can be seen as a linear regression between y_t and the inline variables $r_t\cos x_t$ and $r_t\sin x_t$. This type of representation gives more flexibility to the circular linear regression than the ones proposed by Mardia (1976) and Johnson and Wehrly (1978). Notice that with the r_t variable, for a given value of the circular variable at different time point, say $x_t = x_{t'}, t \neq t'$, the relation between x_t and y_t and $x_{t'}$ and $y_{t'}$ can be different as it depends on the realization of the non observed variable r_t.

Then, we have that

$$f(x_t, y_t|\xi_{tk} = 1) = \frac{\phi_1(y_t|\gamma_{k0}, \sigma_{ky}^2)\phi_2(\boldsymbol{\mu}_k|\mathbf{0}_2, \boldsymbol{\Sigma}_k)\left[m_{tk}\Phi\left(\frac{m_{tk}}{\sqrt{v_{tk}}}\right) + \phi_1(m_{tk}|0, v_{tk})\right]}{\phi_1(m_{tk}|0, v_{tk})}, \tag{2}$$

where Φ is the cumulative density function of a standard normal distribution, $\mathbf{w} = \begin{bmatrix} \cos x_t \\ \sin x_t \end{bmatrix}$, $v_{tk} = \left[\frac{c_{tk}^2}{\sigma_{ky}^2} + \mathbf{w}_t'\Sigma_k^{-1}\mathbf{w}_t\right]^{-1}$, $m_{tk} = v_{tk}\left[\frac{c_{tk}(y_t-\gamma_{k0})}{\sigma_{ky}^2} + \mathbf{w}_t'\Sigma_k^{-1}\boldsymbol{\mu}_k\right]$ and $c_{tk} = \mathbf{w}_t'\begin{bmatrix} \gamma_{k1} \\ \gamma_{k2} \end{bmatrix}$. For the computation of (2), see Section 1 in the online supplement material. The circular linear general PN (CL-GPN) distribution with parameters $\left[\mu_{k1}, \mu_{k2}, \sigma_{k1}^2, \rho_k, \gamma_{k0}, \gamma_{k1}, \gamma_{k2}, \sigma_{ky}^2\right]'$ is thus defined in (2). In this setting, the parameter γ_{k1} and γ_{k2} govern the dependency between the two variables (linear and circular), γ_{k1} is connected to the correlation between the linear variable and the cosine of the circular, γ_{k1} is connected to the correlation between the linear variable and the sine of the circular.

Wang *et al.* (2014) and Wang and Gelfand (2012) argue that working with the PN density is not easy, and its form is practically intractable (to see the closed form of the PN density, see Wang and Gelfand (2012)). because the CL-GPN is based on the PN is itself an intractable distribution and the implementation of the MCMC algorithm can be difficult. However, the introduction of r_t is of practical use as it simplifies the implementation of the MCMC algorithm; see Section 3.3.

3.2. Posterior inference

Let $\boldsymbol{\Psi}$ be the vector of all the parameters of the CL-GPN in all the K regimes, we have the following posterior distribution

$$f(\boldsymbol{\pi}, \boldsymbol{\xi}, \boldsymbol{\pi}_0, \boldsymbol{\Psi}, \mathbf{r}|\mathbf{x}, \mathbf{y}) = \frac{f(\mathbf{r}, \mathbf{x}, \mathbf{y}|\boldsymbol{\Psi}, \boldsymbol{\xi})\, f(\boldsymbol{\xi}_{-0}|\boldsymbol{\xi}_0, \boldsymbol{\pi})\, f(\boldsymbol{\pi})\, f(\boldsymbol{\xi}_0|\boldsymbol{\pi}_0)\, f(\boldsymbol{\pi}_0)\, f(\boldsymbol{\Psi})}{f(\mathbf{x}, \mathbf{y})}$$

where $\mathbf{r} = [r_1, \ldots, r_t]'$ and $f(\mathbf{x}, \mathbf{y}, \mathbf{r}|\boldsymbol{\xi}) = \prod_{k=1}^{K}\prod_{t=1}^{T} f(x_t, r_t, y_t|\xi_{tk} = 1)^{\xi_{tk}}$. As prior distribution, we assume $\mu_{ki} \sim N(\cdot,\cdot)$, $\sigma_{k1}^2 \sim IG(\cdot,\cdot)$, $\rho_k \sim N(\cdot,\cdot)I(-1,1)$, $\sigma_{ky}^2 \sim IG(\cdot,\cdot)$, $\gamma_{kj} \sim N(\cdot,\cdot)$ for $k = 1, \ldots, K$, $i = 1, 2$, $j = 1, 2, 3$, where $IG(\cdot,\cdot)$ indicates the inverse gamma distribution, $\boldsymbol{\pi}_0 \sim Dir(\cdot)$ and $\boldsymbol{\pi}_{k,\cdot} \sim Dir(\cdot)$ where $Dir(\cdot)$ indicates the Dirichlet distribution and $\boldsymbol{\pi}_{k,\cdot}$ is the k^{th} row of $\boldsymbol{\pi}$: we assume $\boldsymbol{\pi}_k \perp \boldsymbol{\pi}_{k'}$ if $k \neq k'$. The prior specification allows us to marginalize over $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_0$ reducing of K^2 the number of parameters to simulate and leads to a more efficient and stable algorithm (see Banerjee *et al.* (2004) and Section 3.3). Note that we can always sample from $f(\boldsymbol{\pi}_{k,\cdot}|\mathbf{x}, \mathbf{y}) = \sum_{\boldsymbol{\xi}} f(\boldsymbol{\pi}_{k,\cdot}|\boldsymbol{\xi})f(\boldsymbol{\xi}|\mathbf{x}, \mathbf{y})$ and $f(\boldsymbol{\pi}_0|\mathbf{x}, \mathbf{y}) = \sum_{\boldsymbol{\xi}} f(\boldsymbol{\pi}_0|\boldsymbol{\xi})f(\boldsymbol{\xi}|\mathbf{x}, \mathbf{y})$ given the set of B posterior samples $\left\{\boldsymbol{\xi}^b\right\}_{b=1}^{B}$ of $\boldsymbol{\xi}$, with an MCMC integration. For each sample $\boldsymbol{\xi}^b$, we draw a sample from $\boldsymbol{\pi}_{k,\cdot}^b|\boldsymbol{\xi}^b \sim Dir\left(\beta + \sum_{t=2}^{T}\xi_{t-1k}^b\xi_{t1}^b, \ldots, \beta + \sum_{t=2}^{T}\xi_{t-1k}^b\xi_{tK}^b\right)$ and one from $\boldsymbol{\pi}_0^b|\boldsymbol{\xi}^b \sim Dir\left(\beta + \xi_{01}^b, \ldots, \beta + \xi_{0K}^b\right)$. The sets $\left\{\boldsymbol{\pi}_{k,\cdot}^b\right\}_{b=1}^{B}$ and $\left\{\boldsymbol{\pi}_0^b\right\}_{b=1}^{B}$ are draw from their, respectively, marginal posterior distributions. The posterior distribution we will work with is then

$$f(\boldsymbol{\xi}, \boldsymbol{\Psi}, \mathbf{r}|\mathbf{x}, \mathbf{y}) = \frac{f(\mathbf{r}, \mathbf{x}, \mathbf{y}|\boldsymbol{\Psi}, \boldsymbol{\xi})\, f(\boldsymbol{\Psi})\, f(\boldsymbol{\xi}_{-0}|\boldsymbol{\xi}_0)f(\boldsymbol{\xi}_0)}{f(\mathbf{x}, \mathbf{y})}.$$

where $f(\boldsymbol{\xi}_{-0}|\boldsymbol{\xi}_0) = \int_{\boldsymbol{\pi}} f(\boldsymbol{\xi}_{-0}|\boldsymbol{\xi}_0, \boldsymbol{\pi})\, f(\boldsymbol{\pi})\, d\boldsymbol{\pi}$ and $f(\boldsymbol{\xi}_0) = \int_{\boldsymbol{\pi}_0} f(\boldsymbol{\xi}_0|\boldsymbol{\pi}_0)\, f(\boldsymbol{\pi}_0)\, d\boldsymbol{\pi}_0$ can be computed in closed form: $f(\boldsymbol{\xi}_{-0}|\boldsymbol{\xi}_0) = \frac{\Gamma(K\beta)^K}{\Gamma(\beta)^{K^2}} \frac{\prod_{k=1}^K \prod_{h=1}^K \Gamma(n_{k,h}+\beta)}{\prod_{k=1}^K \Gamma(n_k - \xi_{T,k} + K\beta)}$ and $f(\boldsymbol{\xi}_0) = \frac{1}{K}$ (for computational details, we refer the reader to the online supplement material, Section 2).

3.3. Computational details

Model parameters are estimated with a MCMC algorithm. More precisely, the μs, γs, σ_y^2s and ξ are simulated with a Gibbs sampler while the remaining parameters require the introduction of a Metropolis step. The full conditionals of μs and γs are normal distributions, those of σ_y^2s are inverse gamma. The full conditionals for the latent variables $\boldsymbol{\xi}_t$, $t \in \mathcal{T}$ are multinomial, and the vector of probabilities depends on the entire vector of $\boldsymbol{\xi}_{-t} = \boldsymbol{\xi} \backslash \{\boldsymbol{\xi}_t\}$. More precisely, let s^- and s^+ be the regimes on time $t-1$ and $t+1$, that is, $\xi_{t-1s^-} = 1$ and $\xi_{t+1s^+} = 1$, respectively, let $n_k^{-t'} = \sum_{\substack{t=0 \\ t \neq t'}}^{T} \xi_{tk}$ and $n_{k,h}^{-t'} = \sum_{\substack{t=1 \\ t \neq t', t \neq t'+1}}^{T} \xi_{t-1k}\xi_{t,h}$. On the online supplement material Section 3, we show that if $t \in \mathcal{T} \backslash \{0, T\}$.

$$f(\boldsymbol{\xi}_t|\mathbf{r}, \mathbf{x}, \mathbf{y}, \boldsymbol{\xi}_{-t}, \boldsymbol{\Psi}) \propto \prod_{k=1}^{K} \frac{\left(n_{s^-,k}^{-t} + \beta + a_{s^-,k,s^+}\right)\left(n_{k,s^+}^{-t} + \beta\right)}{\left(n_k^{-t} - \xi_{T,k} + K\beta\right)} f(r_t, x_t, y_t|\boldsymbol{\xi}_t, \boldsymbol{\Psi})$$

where a_{s^-,k,s^+} assumes value 1 if $s^- = k = s^+$, 0 otherwise, whereas

$$f(\boldsymbol{\xi}_T|\mathbf{r}, \mathbf{x}, \mathbf{y}, \boldsymbol{\xi}_{-T}, \boldsymbol{\Psi}) \propto \prod_{k=1}^{K} \left(n_{s^-,k}^{-T} + \beta\right) f(r_T, x_T, y_T|\boldsymbol{\xi}_T, \boldsymbol{\Psi}).$$

and

$$f(\boldsymbol{\xi}_0|\mathbf{r}, \mathbf{x}, \mathbf{y}, \boldsymbol{\xi}_{-0}, \boldsymbol{\Psi}) \propto \prod_{k=1}^{K} \frac{\left(n_{k,s^+}^{-0} + \beta\right)}{\left(n_k^{-0} - \xi_{T,k} + K\beta\right)}.$$

It is well known that the MCMC sampler for HMM tends to mix really slow (Andrieu *et al.*, 2010). To speed up the convergence, we try to find an optimal proposal distribution for the Metropolis step, which samples K σ_1^2 variables, K ρ variables, and T r variables, using the algorithm described in Robert and Casella (2009), page 258. With the goal to speed up the MCMC convergence, as a general advice, is suggested to decrease the dimension of the parameters space, that is, do as much marginalization as possible (Banerjee *et al.*, 2004). In our model, we found it convenient to marginalize over the vectors $\boldsymbol{\pi}_{k.}, k = 1, \ldots, K$ and $\boldsymbol{\pi}_0$ but not over $\mathbf{r}$. Marginalization over $\mathbf{r}$ decreases significantly the number of random variables to simulate but does not allow to have closed form for full conditional distributions of γ_{k1}, γ_{k2} and $\boldsymbol{\mu}_k$. Without employing the Gibbs step, the MCMC algorithm becomes considerably slower in moving toward its stationary distribution, and then the computational burden increases as a larger number of iterations is required. On the other hand, marginalization over $\boldsymbol{\pi}_{k.}, k = 1, \ldots, K$ and $\boldsymbol{\pi}_0$ has impact only on the way we simulate $\xi_t, t = 0, 1, \ldots, T$, but their simulation is simple in both cases, with or without $\boldsymbol{\pi}_{k.}, k = 1, \ldots, K$ and $\boldsymbol{\pi}_0$, and can be carried out in a Gibbs step.

In the estimation step, we take into account the label-switching issue, common to all latent-class-based models. This problem occurs when exchangeable priors are used for the state specific parameters, which is common practice if there are not prior informations about the hidden states. In these cases, the posterior distribution is invariant to permutations of the state labels and, hence, the marginal posterior distributions of the state specific parameters are identical for all states. Therefore, direct inferences about the state specific parameters are not available from the MCMC output. Various approaches to deal with the label switching problem in finite mixture models have been proposed in the literature; see Jasra *et al.* (2005) for a recent review. To tackle the label switching, we decide to use the post processing technique called *pivotal reordering*, proposed in Spezia (2009) or in Marin and Robert (2013), Chapter 6.5.

3.4. Model selection

To decide the number of regimes, we considered the idea of use the reversible jump (Green, 1995) or a non-parametric approach, as the one proposed by Teh *et al.* (2004). However, our main goal is to demonstrate that the CL-GPN is suitable in an HMM Bayesian framework to model circular-linear variables. Thus, we do not want to further increase the complexity of an already highly complex model by introducing K as random variable.

Common model choice criteria are Akaike information criterion (AIC), Bayesian information criterion (BIC), integrated classification likelihood (ICL), and different classification-based information criteria, which are minimized among a set of potential models. We evaluate these criteria using the set of parameters, among the MCMC draws, that maximize the posterior distribution (called maximum at posterior, MAP or MAP estimator) (Frühwirth Schnatter, 2006, Section 4.4.2, 7.1.4) Let $\tilde{\boldsymbol{\Psi}}$ be the MAP estimator, we compute the BIC and AIC as $BIC = -2\log\left(f\left(\mathbf{x}, \mathbf{y}|\tilde{\boldsymbol{\Psi}}\right)\right) + \#parameters \times \log(T)$ and $AIC = -2\log\left(f\left(\mathbf{x}, \mathbf{y}|\tilde{\boldsymbol{\Psi}}\right)\right) + 2 \times \#parameters$.

The BIC and AIC are generally criticized because they do not take into account the quality of classification of the variables in the K regimes. For classification purpose, Biernacki *et al.* (2000) propose to use the ICL; an index based on the likelihood of observed variables and the vector of regimes indicator that is used by Celeux and Durand (2008) in a HMM context. We compute a BIC approximation of the $ICL = f(\mathbf{x}, \mathbf{y}|\tilde{\boldsymbol{\xi}}, \tilde{\boldsymbol{\Psi}}) - 2\log f(\tilde{\boldsymbol{\xi}}) + \#parameters \times \log(T)$, (see for example Frühwirth Schnatter, 2006, p. 214) in the latter case, as

suggested by McLachlan and Peel (2000), p. 216, we first obtain an estimator of $\boldsymbol{\xi}$, that is, the MAP $\tilde{\boldsymbol{\xi}}$, and then, as for the BIC and AIC, we compute the ICL using the MAP estimator of $\boldsymbol{\Psi}$ conditioning on the value $\tilde{\boldsymbol{\xi}}$.

4. SIMULATION STUDY

In this section, we carried out a simulation study to investigate the performance of the proposed approach in recovering model parameters and the hidden structure of the data. We empirically demonstrate that the CL-GPN can be used in presence of both unimodal or bimodal state-dependent circular distributions and that ignoring the dependencies between the circular and linear variable at a given time leads to a higher number of states.

We plan the simulation study to cover schemes with different underlying *null* models assuming bimodal or almost uniform shapes for the circular variable, and overlapping or well-separated state-dependent distributions for the linear variable. On each simulated datasets, we estimate three models: (1) the CL-GPN model; (2) a constrained model, defined as diagonal CL-GPN (CL-DPN), with $\boldsymbol{\Sigma}_k = \mathbf{I}_2$, so that the state-dependent circular distribution is symmetric and unimodal; (3) a CL-GPN model with all the γ_{k1} and γ_{k2} equal to zero, that is, assuming independence between circular and linear variable given the latent state (indicated as Ind-CL-GPN).

4.1. Designing the simulation study

For each null model, we simulated 200 datasets considering two time-series lengths, $T = 500$ and $T = 2000$, with $K = 3$, $\boldsymbol{\xi}_0 = 1$ and transition matrix $\boldsymbol{\pi}$ with diagonal elements equal to 0.8 and extradiagonal elements equal to 0.1. The considered schemes are summarized in Figure 2 and are characterized by the following settings:

(a). distributions C1 and L1, C2 and L2, and C3 and L3 are considered as state-dependent distributions for the first, the second, and the third regime, respectively. The joint representation through scatters is displayed in Figure 3. This scheme has bimodal state-dependent circular distributions and well as separated linear ones. The following parameters are used to generate data:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{21} & \mu_{22} & \mu_{23} \end{bmatrix} = \begin{bmatrix} 0.1 & 0.1 & 0.0 \\ 0.1 & -1.0 & -0.1 \end{bmatrix}$$

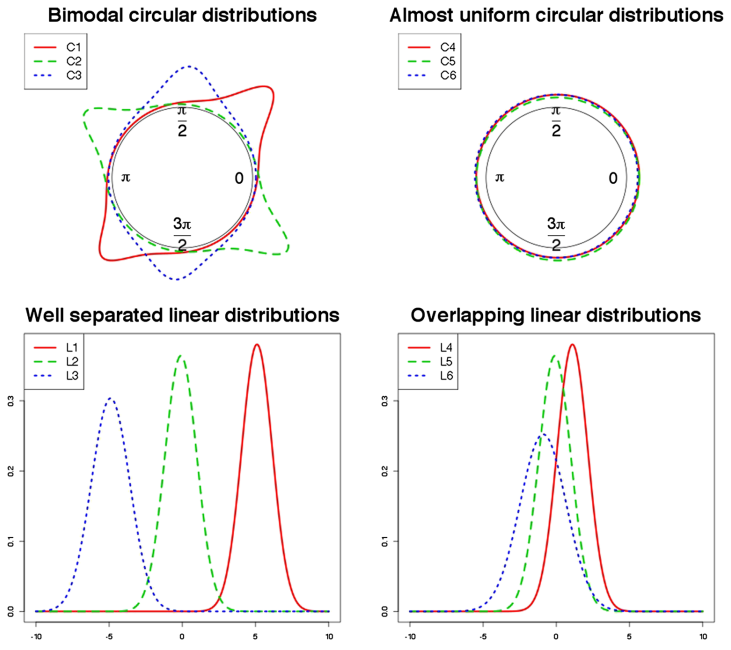


Figure 2. Marginal distributions used in the simulation examples

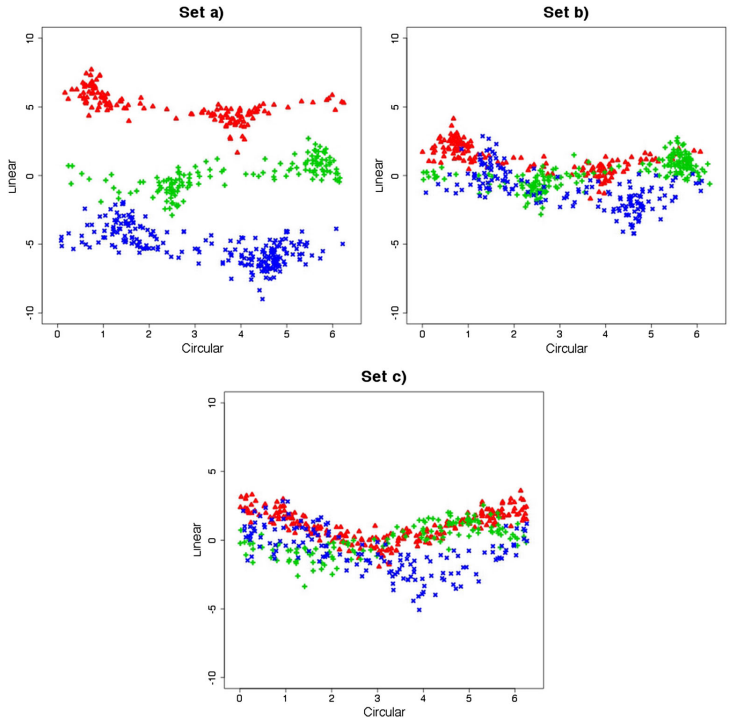


Figure 3. Scatter plot of one simulated dataset for each set of parameters $(T = 500)$

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_{01} & \gamma_{02} & \gamma_{03} \\ \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \end{bmatrix} = \begin{bmatrix} 5.0 & 0.0 & -5.0 \\ 1.0 & 0.0 & 1.0 \\ 0.0 & -1.0 & 1.0 \end{bmatrix}$$

$$\sigma^2_{k1} = \begin{cases} 1 & k = 1 \\ 2 & k = 2 \\ 0.1 & k = 3 \end{cases}; \quad \sigma^2_{ky} = \begin{cases} 0.1 & k = 1 \\ 0.2 & k = 2 \\ 0.5 & k = 3 \end{cases}; \quad \rho_k = \begin{cases} 0.9 & k = 1 \\ -0.9 & k = 2 \\ 0.2 & k = 3 \end{cases}$$

(b). this setting shares circular distributions with scheme (a) while the state-dependent linear distributions are, respectively, the density L4, L5, and L6 of Figure 2. The joint representation through scatters is displayed in Figure 3. With respect to scheme (a), we change the values of $\boldsymbol{\gamma}$:

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_{01} & \gamma_{02} & \gamma_{03} \\ \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \end{bmatrix} = \begin{bmatrix} 1.0 & 0.0 & -1.0 \\ 1.0 & 0.0 & 1.0 \\ 0.0 & -1.0 & 1.0 \end{bmatrix}$$

to have more overlapping state-dependent distributions for the linear variable.

(c). the state-dependent distributions for the linear variable are the same as in scheme (b), whereas the circular ones are, respectively, the density C4, C5, and C6 of Figure 2. The joint representation through scatters is displayed in Figure 3. In this case, we simulate from a CL-DPN because we use $\sigma^2_{11} = \sigma^2_{12} = \sigma^2_{13} = 1$ and $\rho_1 = \rho_2 = \rho_3 = 0$, that is, the circular variable has state-dependent unimodal (almost uniforms) distributions.

Table 1. Frequency distribution of predicted number of regimes

			Predicted K (AIC)					Predicted K (BIC)					Predicted K (ICL)				
T	Model	Scheme	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
500	CL-GPN	(a)	0.00	0.62	0.35	0.03	0.00	0.00	0.82	0.18	0.00	0.00	0.00	0.91	0.09	0.01	0.00
500	CL-GPN	(b)	0.00	0.68	0.25	0.07	0.00	0.00	0.67	0.30	0.03	0.00	0.00	0.90	0.09	0.01	0.00
500	CL-GPN	(c)	0.00	0.98	0.02	0.00	0.00	0.00	0.98	0.02	0.00	0.00	0.00	0.99	0.01	0.00	0.00
500	CL-DPN	(a)	0.00	0.59	0.37	0.04	0.00	0.00	0.80	0.18	0.02	0.00	0.00	0.87	0.10	0.03	0.00
500	CL-DPN	(b)	0.00	0.61	0.31	0.08	0.00	0.00	0.63	0.33	0.04	0.00	0.00	0.86	0.12	0.02	0.00
500	CL-DPN	(c)	0.00	0.98	0.01	0.01	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
500	Ind-CL-GPN	(a)	0.00	0.00	0.01	0.08	0.91	0.00	0.00	0.00	0.09	0.91	0.00	0.00	0.08	0.26	0.66
500	Ind-CL-GPN	(b)	0.00	0.00	0.02	0.08	0.90	0.00	0.00	0.05	0.03	0.92	0.03	0.00	0.00	0.07	0.90
500	Ind-CL-GPN	(c)	0.00	0.00	0.00	0.09	0.91	0.00	0.00	0.00	0.09	0.91	0.48	0.41	0.06	0.03	0.02
2000	CL-GPN	(a)	0.00	0.97	0.03	0.00	0.00	0.00	0.98	0.01	0.01	0.00	0.00	0.98	0.00	0.00	0.00
2000	CL-GPN	(b)	0.00	0.97	0.01	0.02	0.00	0.00	0.98	0.02	0.00	0.00	0.00	0.98	0.02	0.00	0.00
2000	CL-GPN	(c)	0.00	0.96	0.03	0.01	0.00	0.00	0.99	0.01	0.00	0.00	0.00	1.00	0.00	0.00	0.00
2000	CL-DPN	(a)	0.00	0.90	0.08	0.02	0.00	0.00	0.91	0.05	0.04	0.00	0.00	0.93	0.07	0.00	0.00
2000	CL-DPN	(b)	0.00	0.91	0.07	0.02	0.00	0.00	0.93	0.03	0.04	0.00	0.00	0.91	0.09	0.00	0.00
2000	CL-DPN	(c)	0.00	0.98	0.02	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
2000	ind-CL-GPN	(a)	0.00	0.00	0.02	0.31	0.67	0.00	0.00	0.02	0.27	0.71	0.00	0.04	0.21	0.31	0.44
2000	ind-CL-GPN	(b)	0.00	0.00	0.03	0.07	0.90	0.00	0.00	0.02	0.08	0.90	0.03	0.06	0.22	0.41	0.28
2000	ind-CL-GPN	(c)	0.00	0.00	0.00	0.29	0.71	0.00	0.00	0.00	0.25	0.75	0.06	0.27	0.39	0.24	0.04

CL-GPN, circular linear general projected normal.

Table 2. posterior median estimates of the parameter ($\hat{}$) and credibility intervals for scheme (c) and $T = 500$

	CL-GPN			CL-DPN		
	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
$\hat{\mu}_{k1}$	0.12	0.12	−0.05	0.13	0.11	−0.06
CI	(−0.04 0.28)	(−0.09 0.33)	(−0.26 0.15)	(−0.03 0.31)	(−0.11 0.34)	(−0.27 0.16)
$\hat{\mu}_{k2}$	0.04	−0.08	0.07	0.03	−0.09	0.09
CI	(−0.12 0.19)	(−0.30 0.12)	(−0.14 0.27)	(−0.13 0.18)	(−0.32 0.14)	(−0.12 0.29)
$\hat{\rho}_k$	0.06	−0.09	0.14	·	·	·
CI	(−0.13 0.22)	(−0.32 0.16)	(−0.11 0.37)	(· ·)	(· ·)	(· ·)
$\hat{\sigma}^2_{k1}$	0.94	0.76	1.01	·	·	·
CI	(0.65 1.35)	(0.47 1.22)	(0.63 1.57)	(· ·)	(· ·)	(· ·)
$\hat{\gamma}_{k0}$	0.98	−0.04	−1.04	0.98	−0.04	−1.05
CI	(0.89 1.06)	(−0.19 0.13)	(−1.22 −0.86)	(0.89 1.06)	(−0.20 0.12)	(−1.23 −0.88)
$\hat{\gamma}_{k1}$	1.04	0.14	0.82	1.01	0.15	0.86
CI	(0.88 1.22)	(−0.04 0.35)	(0.60 1.07)	(0.91 1.13)	(−0.01 0.33)	(0.68 1.06)
$\hat{\gamma}_{k2}$	−0.04	−1	1.04	−0.02	−0.96	1.05
CI	(−0.11 0.05)	(−1.18 −0.83)	(0.82 1.28)	(−0.09 0.05)	(−1.12 −0.81)	(0.87 1.26)
$\hat{\sigma}^2_{ky}$	0.14	0.32	0.42	0.14	0.3	0.41
CI	(0.09 0.19)	(0.18 0.52)	(0.26 0.67)	(0.10 0.19)	(0.18 0.49)	(0.25 0.63)

CI, credibility intervals; CL-GPN, circular linear general projected normal.

On each dataset, we estimate models with K from 2 to 6 and assuming the following prior distributions: $\mu_{ki} \sim N(0, 5)$, $\gamma_{kj} \sim N(0, 5)$, $\rho_k \sim N(0, 5)I(-1, 1)$, $\sigma^2_{k1} \sim IG(2, 1)^{\dagger}$, $\sigma^2_{ky} \sim IG(2, 1)$, $\beta = 1$ with $i = 1, 2$ and $j = 1, 2, 3$; that is, they do not depend on the regime.

4.2. Simulation study results

To evaluate the performance of AIC, BIC, and ICL as selection criteria for the number of regimes, in Table 1, we report the frequency distribution of the predicted K under each simulation setting considered for the CL-GPN, CL-DPN, and Ind-CL-GPN models. With respect

[†]The two parameters are the shape and rate, respectively

to the CL-GPN model, we can observe that ICL performs considerably well in all cases. In fact, the predicted K is only occasionally different from the true one, and, when this happens, the former is always larger than the latter. On the other hand, AIC and BIC have an excellent behavior with the exception of the cases $T = 500$, schemes (a) and (b). As may be expected, these criteria perform better as the amount of information in the data increases.

Ignoring the (state-dependent) correlation between circular and linear measurements may strongly affect the hidden structure. Indeed, by looking at information criteria for the Ind-CL-GPN model, we have that the latent structure is not well recovered and a higher number of regimes than expected is estimated. Of course, this affects parameter estimates and results interpretation, as a not needed number of (latent) regimes is identified in the data.

Here, we briefly summarize the results of the simulation study for scheme c). By looking at parameters estimates (see Table 2), we have that the CL-GPN and the CL-DPN models lead essentially to the same results. Point estimates and credibility intervals are very close, suggesting that the CL-GPN distribution can be used whenever we cast doubts on the unimodality of circular distributions. Indeed, the CL-DPN distribution is a specific case of the CL-GPN one, in which conditional circular distribution are constrained to be unimodal.

To further resemble empirical situations, we randomly drop 10% observation of a randomly selected dataset simulated accordingly to the scheme (c) with $T = 500$ and estimate the CL-GPN and a CL-DPN models. Along with model parameters, we also simulate the missing observations. We compute the average continuous ranked probability score (CRPS) for both the circular (Grimit *et al.*, 2006) and linear variable (Gneiting and Raftery, 2007) from the posterior samples of the missing observations, as well as the average prediction error (APE) for the circular variable (Jona Lasinio *et al.*, 2012) and the mean squared error for the linear ones (MSE). With the CRPS, we evaluate the model performance regarding the entire predictive distribution. APE and MSE allow us to measure the distance between the true values and the simulated ones. The CPRPs for the circular variable and the MSE are identical under the two models, whereas the CRPS for the linear one is 0.66 under the CL-GPN and 0.67 under the CL-DPN, and the APE is, respectively, 0.76 and 0.75 for the CL-GPN and the CL-DPN. Then, the two models have the same performances in dealing with the missing values as well.

From the computational point of view, in the datasets with $T = 2000$ our C++ implementation of the model needs 1 000 000 iterations with a burnin of 700 000 and a thin of 100, whereas with $T = 500$, the iterations needed are 800 000, with a burnin of 400 000 and again a thin of 100. The computational work for the simulation study and the real data application of Section 5 has been executed on the IT resources made available by ReCaS, a project financed by the MIUR (Italian Ministry for Education, University and Research) in the 'PON Ricerca e Competività 2007–2013 - Azione I - Interventi di rafforzamento strutturale PONa3_00052, Avviso 254/Ric. The computational time are of the order of 1 h for $T = 500$ and 5 h for $T = 2000$. All the results shown are from MCMC chains that reach the convergence, checked using the standard tool on the R package coda.

5. REAL DATA EXAMPLE

Finally, we apply the CL-GPN hidden Markov model to a bivariate time series of wind directions and (log-transformed) speeds. Data are recorded on a semi-hourly basis from 12/12/2009 to 12/1/2010 in Ancona (Italy) at a bouy located in the Adriatic Sea 30 km from the coast (Figure 4). Data are recorded on $T = 1500$ times and have been previously analysed by Bulla *et al.* (2012).

As often arise in environmental studies, data are not complete. Recorded for directions and speeds, respectively, are 213 and 210 missing values; 125 profiles are completely missing.

During wintertime, relevant wind events in the Adriatic Sea are typically generated by the south-eastern Sirocco, the north-eastern Bora and the north-western Maestral. Sirocco arises from a warm, dry, tropical air mass that is pulled northwards by low-pressure cells moving eastwards 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. Finally, the Maestral is a sea breeze wind blowing northwesterly when the east Adriatic coast gets warmer than the sea. While Bora and Sirocco episodes are usually associated with high-speed flows, Maestral is in general linked with good meteorological conditions. Hence, the marginal distribution of (log-transformed) wind speed may be interpreted as the result of mixing different wind-speed regimes.

As for the simulation examples, we look at the AIC, BIC, and ICL to select the appropriated number of components. The ICL suggest to use $K = 3$, whereas the AIC and BIC $K = 4$. To help decide between the two number of regimes, we look ar their predictive ability, the CRPS_c, and APE highlight loss of predictive ability on the circular variable if we choose $K = 4$ (CRPS_c=0.59 and APE=0.75 with $K = 4$ while $\text{CRPS}_c = 0.34$ and $\text{APE} = 0.35$ with $K = 3$). For the linear variable, looking at the values of CRPS_l and MSE, there is a small difference between $K = 3$ and $K = 4$; however, both CRPS_l and MSE favor $K = 3$ ($\text{CRPS}_l = 0.17$ and $\text{APE} = 0.39$ with $K = 4$, whereas CRPS_l=0.16 and APE = 0.34 with $K = 3$). We decide to adopt $K = 3$, that is also the choice of Bulla *et al.* (2012) following their suggestion that three regimes provide well-separated and more interpretable states. The resulting classification is displayed in Figure 5 and all the credibility intervals and point estimates of the parameters are in Table 3. The estimated transition probabilities are displayed in Table 4. As expected, the transition probability matrix is essentially diagonal, reflecting the temporal persistence of the regimes, that is, of wind conditions. Furthermore, the small off-diagonal transition probabilities between states indicate that direct transitions between Sirocco and Bora episodes are very unlikely. The model hence confirms that the Adriatic Sea typically alternates relevant wind events with periods of good conditions.

For a more clear interpretation of the state dependent distributions, we compute some feature of the CL-GPN distribution. In detail, we look at the posterior marginal mean and variance of the linear distribution for each regime ($\tilde{\mu}_{ky}$ and $\tilde{\sigma}^2_{ky}$), the circular mean ($\tilde{\mu}_{kx}$) and

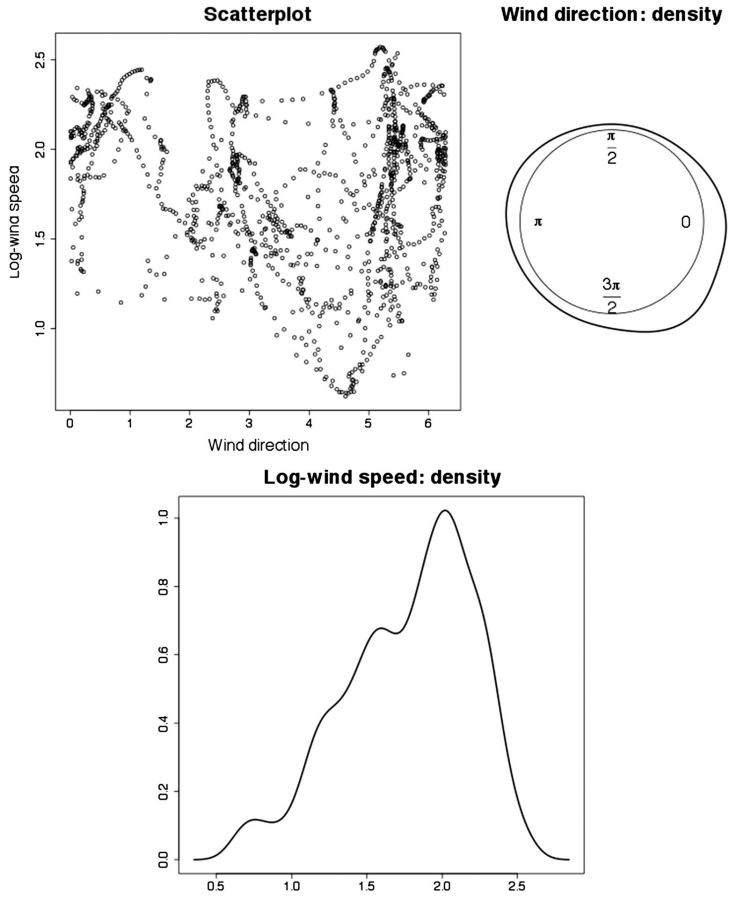


Figure 4. Real data

concentration ($\bar{g}_{kx}$) of the circular variable and a measure of correlation between the circular and linear variables ($\bar{\rho}^2_{kxy}$) as in Mardia and Jupp (1999), p. 245. Point estimates and credibility intervals are provided in Table 5). Further details can be found in the online material.

The regimes are ordered according to the marginal log-wind speed. In the three regimes, the point estimates are, respectively, $\hat{\mu}_{ky} = 1.42, 1.70, 2.11$ (which correspond to 4.14, 5.47, and 8.25m/s in the natural scale). With the increases of the velocity, the distribution becomes more concentrated: the marginal linear variance, $\hat{\bar{\sigma}}^2_{ky}$, is, respectively, 0.17, 0.11, 0.04, for a plot of the distributions, see Figure 5. The circular mean is 4.98 in the first regime, north-westerly Maestral episodes, 2.70 in the second, south-eastern Sirocco, and 6.04 in the third, northern Bora jets, on the first regime, the circular marginal distribution is less concentrated than in the others (0.78 for $k = 1$, 0.83 for $k = 2$, and 0.82 for $k = 3$).

The correlations between the circular and linear variables are weak in all the regimes: $\hat{\bar{\rho}}^2_{kxy}$ is 0.02 in the first and 0.05 in the others. Under the hypothesis of no correlation, that is, $\bar{\rho}^2_{kxy} = 0$, the statistic $\tilde{F} = \frac{\bar{\rho}^2_{kxy}(n-1)}{1-\bar{\rho}^2_{kxy}}$ is distributed as a $F_{2,T-3}$ where in our case $T = 1500$. The lower limits of the 95% credibility intervals of the posterior distributions of $\tilde{F}$ are 1.24, 4.80, and 4.95 in the calm, transition, and storm conditions, respectively, and the 95% percentile of $F_{2,T-3}$ is 3.00. Accordingly, circular-linear correlations are significant in the transition and storm conditions only. This result is not present at all in previous analyses. This can be seen also with the value of γ_{k1} and γ_{k2} in Table 3. In the first regime $\hat{\gamma}_{11} = 0.01$ and $\hat{\gamma}_{12} = -0.04$, both credibility intervals contain the 0. In the second, there is a negative relation

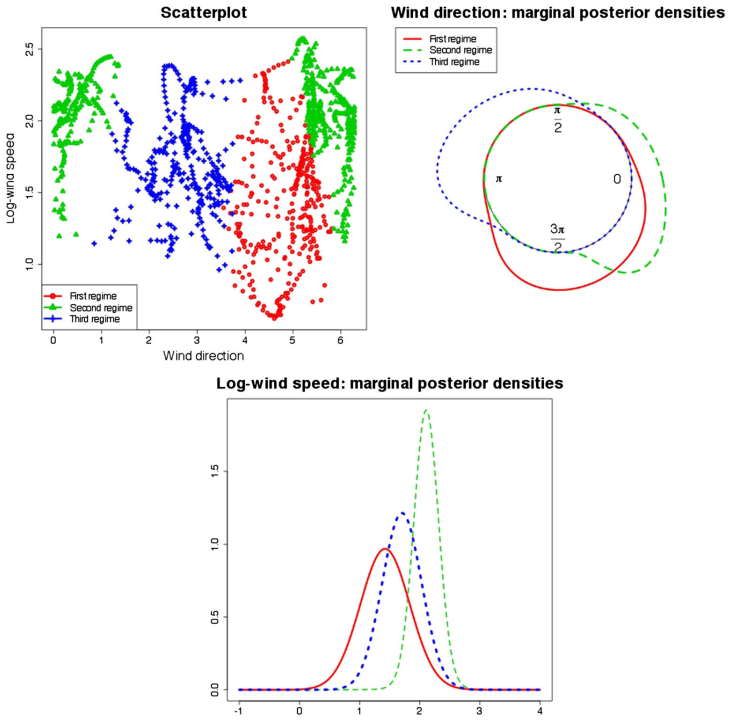


Figure 5. Real data classification

Table 3. Real data: posterior median estimates of the parameter ($\hat{}$) and credibility intervals

	$k = 1$	$k = 2$	$k = 3$
$\hat{\mu}_{k1}$	0.45	−1.62	1.19
CI	(0.28 0.63)	(−1.83 −1.41)	(1.02 1.32)
$\hat{\mu}_{k2}$	−1.80	0.67	−0.41
CI	(−2.07 −1.57)	(0.52 0.81)	(−0.51 −0.30)
$\hat{\rho}_k^2$	0.36	0.56	−0.27
CI	(0.14 0.54)	(0.38 0.71)	(−0.46 −0.06)
$\hat{\sigma}_{k1}^2$	2.03	0.94	0.15
CI	(1.46 2.85)	(0.71 1.28)	(0.10 0.23)
$\hat{\gamma}_{k0}$	1.34	1.47	2.36
CI	(1.21 1.49)	(1.33 1.62)	(2.25 2.45)
$\hat{\gamma}_{k1}$	0.01	−0.09	−0.22
CI	(−0.03 0.05)	(−0.16 −0.03)	(−0.30 −0.13)
$\hat{\gamma}_{k2}$	−0.04	0.12	−0.02
CI	(−0.11 0.03)	(0.06 0.18)	(−0.05 0.01)
$\hat{\sigma}_{ky}^2$	0.17	0.09	0.03
CI	(0.14 0.19)	(0.08 0.11)	(0.03 0.06)

CI, credibility interval.

Table 4. Real data: transition matrix

Destination state		1	2	3
Origin state	1	0.96 (0.94, 0.97)	0.02 (0.01, 0.04)	0.02 (0.01, 0.04)
	2	0.02 (0.01 0.04)	0.97 (0.95, 0.98)	0.00 (0.00, 0.01)
	3	0.02 (0.01, 0.03)	0.00 (0.00, 0.01)	0.98 (0.97, 0.99)

Table 5. Real data: posterior median estimates ($\hat{}$) and credibility intervals of the features of the distribution circular linear general projected normal

	$k = 1$	$k = 2$	$k = 3$
$\hat{\mu}_{kx}$	4.98	2.7	6.04
CI	(4.81 5.16)	(2.55 2.83)	(5.90 6.19)
$\hat{g}_{kx}$	0.78	0.83	0.82
CI	(0.71 0.84)	(0.77 0.87)	(0.76 0.86)
$\hat{\mu}_{ky}$	1.42	1.70	2.11
CI	(1.34 1.50)	(1.63 1.78)	(2.02 2.20)
$\hat{\sigma}^2_{ky}$	0.17	0.11	0.04
CI	(0.15 0.20)	(0.09 0.12)	(0.03 0.07)
$\hat{\rho}^2_{kxy}$	0.02	0.05	0.05
CI	(0.00 0.10)	(0.00 0.18)	(0.00 0.17)

CI, credibility interval.

between the linear variable and the cosine of the circular one ($\gamma_1 = -0.09$) and a positive relation with the sine ($\gamma_2 = 0.12$). In the third regime, the dependence between the linear and circular variable is on the cosine direction ($\gamma_1 = -0.22$).

We estimate the model using 1 000 000 iterations, a burnin of 700 000 and a thin of 100. Here, again, we checked the convergence of the MCMC chain using the standard tool on the R package coda.

6. DISCUSSION

In this work, we introduce, for the first time, the CL-GPN distribution in a Bayesian HMM framework, and we present the explicit expression of the CL-GPN likelihood (Section 2 and online materials Section 1 for the details). This approach allows to easily model multivariate processes with mixed support (circular-linear), by combining the bivariate representation of the circular component (i.e. the PN distribution) and a Gaussian distribution for the linear part. Here, we considered one circular and one linear variable, although it is fairly easy to extend the proposed model to more than one linear component.

The Bayesian framework allows us to overcome identifiability issues and computational problems that may arise in the classical setting. Several implementation novelties are introduced to speed up algorithms convergence. We use an adaptive Metropolis whenever a Gibbs sampler is not implementable (Section 3.3). Furthermore, we marginalize the transition matrix so to avoid its estimation to reduce the problem size obtaining it as an a posteriori byproduct (Section 3.2) and we provide evidence that the marginalization does not affect parameters estimation. We also demonstrate that assuming conditional independence between the circular and linear variable can make difficult to correctly estimate the number of regimes.

We applied this methodology to wind data confirming previously obtained results and highlighting new data features. Circular parameters interpretation is not straightforward; however, this does not limit the inferential richness of the model. Using MCMC simulations posterior circular mean and concentration can be derived, as well as the circular-linear correlation. Of course, different areas of application can be considered for the proposed approach, for example, animal movement modelling (Langrock *et al.*) and driving behavior (Jackson *et al.*, 2014).

Further developments will include the extension to more than one circular variable. This extension requires a careful definition of correlation between circular variables that is not straightforward under the PN distribution. Another interesting extension of the proposed approach is to allow the estimation of the number of states along with the model parameters. The latter can be obtained using a hierarchical Dirichlet process on the states or a reversible jump.

A crucial assumption of our model is that the temporal dependence is well described by a first-order Markov chain, that is, the sojourn time is geometrical. If we want to allow for different sojourn time distributions with finite support, the HMM formulation is exact. Similarly,

by allowing the number of hidden states to grow with the sample size, we can allow for continuous time, that is, the hidden distribution can be approximated with arbitrary accuracy using the proposed model. This can be seen as a possible solution to computational issues arising with continuous-valued latent models (Langrock *et al.*, 2012).

REFERENCES

Alfò M, Maruotti A. 2010. A hierarchical model for time dependent multivariate longitudinal data, *Data analysis and classification*, Palumbo F, Lauro CN, Greenacre MJ (eds), Studies in Classification, Data Analysis, and Knowledge Organization. Springer: Berlin, Heidelberg,271–279.

Andrieu C, Doucet A, Holenstein R. 2010. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(3):269–342.

Banerjee S, Gelfand AE, Carlin BP. 2004. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC: Boca Raton, Florida.

Bartolucci F, Farcomeni A. 2009. A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *Journal of the American Statistical Association* **104**(486):816–831.

Bartolucci F, Farcomeni A. 2010. A note on the mixture transition distribution and hidden Markov models. *Journal of Time Series Analysis* **31**(2):132–138.

Bartolucci F, Farcomeni A, Pennoni F. 2012. *Latent Markov Models for Longitudinal Data*. Chapman and Hall: Boca Raton, Florida.

Bartolucci F, Farcomeni A, Pennoni F. 2014. Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates. *TEST*. to appear.

Bartolucci F, Pennoni F, Vittadini G. 2011. Assessment of school performance through a multilevel latent Markov Rasch model. *Journal of educational and behaviour statistics* **36**(4):491–522.

Baudry JP, Raftery AE, Celeux G, Lo K, Gottardo R. 2010. Combining mixture components for clustering. *Journal of Computational and Graphical Statistics* **19**(2):332–353.

Biernacki C, Celeux G, Govaert G. 2000-07. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(7):719–725.

Bulla J, Lagona F, Maruotti A, Picone M. 2012. A multivariate hidden Markov model for the identification of sea regimes from Incomplete skewed and circular time series. *Journal of Agricultural, Biological, and Environmental Statistics* **17**(4):544–567.

Cappé O, Moulines E, Rydén T. 2005. *Inferencein Hidden Markov Models*, Springer Series in Statistics. Springer: New York.

Celeux G, Durand J-B. 2008-10. Selecting hidden Markov model state number with cross-validated likelihood. *Computational Statistics* **23**(4):541–564.

Dannemann J. 2012. Semiparametric hidden Markov models. *Journal of Computational and Graphical Statistics* **21**(3):677–692.

Frühwirth Schnatter S. 2006. *Finite Mixture and Markov Switching Models*. Springer, Verlag: New York.

Geweke J, Amisano G. 2011. Hierarchical Markov normal mixture models with applications to financial asset returns. *Journal of Applied Econometrics* **26**(1):1–29.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**(477):359–378.

Green PJ. 1995. Reversible jump Markov chain monte carlo computation and Bayesian model determination. *Biometrika* **82**(4):711–732.

Grimit EP, Gneiting T, Berrocal VJ, Johnson NA. 2006. The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society* **132**:2925–2942.

Holzmann H, Munk A, Suster M, Zucchini W. 2006. Hidden Markov models for circular and linear-circular time series. *Environmental and Ecological Statistics* **13**(3):325–347.

Jackson J, Albert P, Zhiwei Z. 2014. A two-state mixed hidden Markov model for risky teenage driving behavior. *The Annals of Applied Statistics*. To appear.

Jammalamadaka SR, SenGupta A. 2001. *Topics in Circular Statistics*. World Scientific: New Jersey.

Jasra A, Holmes CC, Stephens DA. 2005. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science* **20**(1):50–67.

Johnson RA, Wehrly TE. 1978. Some Angular-Linear Distributions and Related Regression Models. *Journal of the American Statistical Association* **73**(363):602–606.

Jona Lasinio G, Gelfand A, Jona Lasinio M. 2012. Spatial analysis of wave direction data using wrapped Gaussian processes. *Annals of Applied Statistics* **6**(4):1478–1498.

Kato S, Shimizu K, Shieh GS. 2008. A circular-circular regression model. *Statistica Sinica* **18**:633–645.

Lagona F, Maruotti A, Picone M. 2011-04. *A Non-homogeneous Hidden Markov Model for the Analysis of Multi-pollutant Exceedances Data – InTechOpen,* Hidden Markov models: Theory and application, Dymarski P (ed.). INTECH: University Campus STeP Ri Slavka Krautzeka 83/A 51000 Rijeka Croatia,207–222. chapter 10.

Lagona F, Picone M. 2011. A latent-class model for clustering incomplete linear and circular data in marine studies. *Journal of Data Science* **9**(4):585–605.

Lagona F, Picone M. 2012. Model-based clustering of multivariate skew data with circular components and missing values. *Journal of Applied Statistics* **39**(5):927–945.

Lagona F, Picone M, Maruotti A, Cosoli S. 2014. A hidden Markov approach to the analysis of space-time environmental data with linear and circular components. *Stochastic Environmental Research and Risk Assessment*. DOI: 10.1007/s00477-014-0919-y.

Langrock R, King R, Matthiopoulos J, Thomas L, Fortin D, Morales JM. 2012. Flexible and practical modeling of animal telemetry data: hidden Markov models and extensions. *Ecology* **93**(11):2336–2342.

Langrock R, Kneib T, Michelot T: Markov-switching generalized additive models. 2014, *ArXiv e-prints*, http://arxiv.org/abs/1406.3774

Langrock R, MacDonald IL, Zucchini W. 2012. Some nonstandard stochastic volatility models and their estimation using structured hidden Markov models. *Journal of Empirical Finance* **19**(1):147 –161.

Langrock R, Swihart BJ, Caffo BS, Punjabi NM, Crainiceanu CM. 2013. Combining hidden Markov models for comparing the dynamics of multiple sleep electroencephalograms. *Statistics in Medicine* **32**(19):3342–3356.

Mardia KV. 1976. Linear-circular correlation coefficients and rhythmometry. *Biometrika* **63**(2):403–405.

Mardia KV, Jupp PE. 1999. *Directional Statistics*. John Wiley and Sons.

Marin JM, Robert C. 2013. *Bayesian Essentials with R*, Springer Texts in Statistics. Springer: New York.

Martinez-Zarzoso I, Maruotti A. 2013. The environmental kuznets curve: functional form, time-varying heterogeneity and outliers in a panel setting. *Environmetrics* **24**(7):461–475.

Maruotti A. 2011. Mixed hidden Markov models for longitudinal data: An overview. *International Statistical Review* **79**(3):427–454.

McLachlan C, Peel D. 2000. *Finite Mixture Models*. John Wiley and Sons: New York.

Robert CP, Casella G. 2009. *Introducing Monte Carlo Methods with R*. Springer: Berlin, Heidelberg.

Rydén T. 2008. EM versus Markov chain Monte Carlo for estimation of hidden Markov models: a computational perspective. *Bayesian Analysis* **3**(4): 659–688.

Rydén T, Titterington DM. 1998. Computational Bayesian analysis of hidden Markov models. *Journal of Computational and Graphical Statistics* **7**(2): 194–211.
Spezia L. 2009. Reversible jump and the label switching problem in hidden Markov models. *Journal of Statistical Planning and Inference* **139**(7): 2305–2315.
Spezia L. 2010. Bayesian analysis of multivariate Gaussian hidden Markov models with an unknown number of regimes. *Journal of Time Series Analysis* **31**(1):1–11.
Teh YW, Jordan MI, Beal MJ, Blei DM. 2004. Hierarchical dirichlet processes. *Journal of the American Statistical Association* **101**:1566–1581.
Wang F, Gelfand A. 2012. Directional data analysis under the general projected normal distribution. *Statistical Methodology* **10**(1):113 –127.
Wang F, Gelfand A. 2014. Modeling space and space-time directional data using projected Gaussian processes. *Journal of the American Statistical Association*. DOI: 10.1080/01621459.2014.934454.
Wang F, Gelfand A, Jona Lasinio G. 2014. 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*. DOI: 10.5705/ss.2013.204w.
Yildirim S, Singh SS, Dean T, Jasra A. 2014. Parameter estimation in hidden Markov models with intractable likelihoods using sequential monte carlo. *Journal of Computational and Graphical Statistics* **41**(4):970–987.
Zhang Q, Snow Jones A, Rijmen F, Ip EH. 2010. Multivariate discrete hidden Markov models for domain-based measurements and assessment of risk factors in child development. *Journal of Computational and Graphical Statistics* **19**(3):746–765.
Zucchini W, MacDonald IL. 2009. *Hidden Markov Models For Time Series: An Introduction Using R*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis: Boca Raton.

SUPPORTING INFORMATION

Supporting information may be found in the online version of this article.

A Bayesian hidden Markov model for telemetry data

Gianluca Mastrantonio
Department of Economics, University of Roma Tre

Abstract

We introduce a new multivariate circular linear distribution suitable for modeling direction and speed in (multiple) animal movement data. To properly account for specific data features, such as heterogeneity and time dependence, a hidden Markov model is used. Parameters are estimated under a Bayesian framework and we provide computational details to implement the Markov chain Monte Carlo algorithm.

The proposed model is applied to a dataset of six free-ranging Maremma Sheepdogs. Its predictive performance, as well as the interpretability of the results, are compared to those given by hidden Markov models built on all the combinations of von Mises (circular), wrapped Cauchy (circular), gamma (linear) and Weibull (linear) distributions

Keywords: Animal movement, Circular-linear distribution, Multivariate projected normal, Multivariate skew normal, Dirichlet Process, Hidden Markov model

1 Introduction

In the recent literature, the interest in modelling animal movement data, with the goal to understand the animals behaviour, is increasing. Animal movement modeling has a long history, dating back to the diffusion model of Brownlee (1912). A wide range of different models have been proposed, such as stochastic differential equations (Blackwell, 2003), mixture of random walks (Morales *et al.*, 2004), Brownian bridge (Horne *et al.*, 2007), agent-based model (Hooten *et al.*, 2010), mechanistic approach (McClintock *et al.*, 2012) or the continuous-time discrete-space model (Hanks *et al.*, 2015).

Animal movement data often take the form of a bivariate time series of spatial coordinates obtained by equipping an animal with a tracking device, e.g. a GPS collar, that records locations at different times. These type of data are called *telemetry data*. From telemetry data, movements are measured by computing the so-called *movement metrics* (Patterson *et al.*, 2008), such as *step-length* and *turning-angle*, see for example D'Elia (2001), Jonsen *et al.* (2005) or Ciucci *et al.* (2009). Observed metrics are random variables and, accordingly, a parametric distribution is often needed to model these data.

Improved communication systems, shrinking battery sizes and the prices drop of GPS devices, have led to an increasing availability of datasets (Cagnacci *et al.*, 2010). The data, often freely available (see for example the *movebank data repository* at `www.movebank.org`), have a complex structure because the animal behaviour changes over time and the occurrences of behavioural modes are not time-independent (Houston and Mcnamara, 1999) (temporal dependence) (Morales *et al.*, 2004), each behaviour is characterized by different distributions of the associated movement metrics (heterogeneity) and there is dependence in the movement metrics between and within animals (multivariate associations) (Langrock *et al.*, 2014).

Time dependence and heterogeneity have been addressed, in the literature, by using hidden Markov models (HMMs), see for example Franke *et al.* (2004), Holzmann *et al.* (2006), Jonsen *et al.* (2007), Eckert *et al.* (2008), Patterson *et al.* (2009), Schliehe-Diecks *et al.* (2012) or Langrock *et al.* (2014). In most of the animal movement applications, the number of behavioural modes is fixed at priori using external knowledge.

The multivariate interactions between animals have been modeled in several ways. Jonsen *et al.* (2006) assume a common distribution for some individual-level parameters that allows inference about population-level parameters. Langrock *et al.* (2014) propose a parent-child structure, where the animals (the children) are all attracted to an abstract point (the parent) while in Morales *et al.* (2010) the animals are treated as independent assuming that the movement of one animal is representative of the group's overall movement.

A natural way to model the multivariate interactions is to define a joint distribution

for the movement metrics that, generally, are composed by measures of speed (e.g. the step-length) and direction (e.g the turning-angle). The direction is a *circular variable*, it represents an angle or a point over a circumference and due to the particular topology of the circle must be treated differently from the linear (or inline) ones, e.g. variables defined over $\mathbb{R}$ or $\mathbb{R}^+$; for an introduction on circular variables see the book of Mardia and Jupp (1999) or Jammalamadaka and SenGupta (2001). A joint modelling of step-lengths and turning-angles requires a multivariate distribution for circular-linear variables but in the literature have been proposed only distribution for cylindrical data (SenGupta, 2004; Sengupta and Ong, 2014; Mastrantonio *et al.*, 2015a; Abe and Ley, 2015), i.e. one circular and one linear variable.

In this work we are interested in finding the behavioural modes of six free-ranging sheepdogs attending livestock (van Bommel and Johnson, 2014b,a) and understand how they interact. Motivated by the data at hand we introduce a new flexible multivariate circular-linear distribution with dependent components, called the *projected-skew normal*, based on the skew normal of Sahu *et al.* (2003) and on a multivariate extension of the projected normal (Wang and Gelfand, 2013). This distribution allows us to model jointly the movement metrics, introducing dependence among animals. The proposal is used as emission distribution in an HMM. We propose to estimate the parameters in a non-parametric Bayesian framework, relying on the *sticky hierarchical Dirichlet process-hidden Markov model* (sHDP-HMM) of Fox *et al.* (2011). This allows to jointly estimate model parameters and the number of behavioural modes without fixing it a priori. We show how to estimate the parameters using a Markov chain Monte Carlo (MCMC) algorithm. As a by-product, our MCMC implementation solves the well-known identification problem of the univariate projected normal distribution (Wang and Gelfand, 2013).

The paper is organized as follows. In Section 2 we introduce the proposed distribution and in Section 3 we show how to estimate its parameters in a Bayesian framework. In Section 4 we introduce the HMM (Section 4.1) and the non-parametric extension (Section 4.2). In Section 5 we apply the model to the real data example and in Section 5.3 we compare the proposed emission distribution with the most used in the literature. The paper ends with some conclusion remarks in Section 6.

2 The multivariate circular-linear distribution

In this Section we introduce the projected normal, its multivariate extension and the skew normal of Sahu *et al.* (2003) used to built the new multivariate circular-linear distribution.

2.1 The multivariate projected normal distribution

Let $\mathbf{W}_i = (W_{i1}, W_{i2})'$ be a 2-dimensional random variable, normally distributed with mean vector $\boldsymbol{\mu}_{wi}$ and covariance matrix $\boldsymbol{\Sigma}_{wi}$. The random variable

$$\Theta_i = \text{ atan}^* \frac{W_{i2}}{W_{i1}} \in [0, 2\pi), \text{[1]} \tag{1}$$

is a circular variable, i.e. a variable that represents an angle over the unit circle, distributed as a *projected normal* (PN): $\Theta \sim PN(\boldsymbol{\mu}_{wi}, \boldsymbol{\Sigma}_{wi})$. Let $\mathbf{U}_i = (U_{i1}, U_{i2})'$, where $U_{i1} = \cos\Theta_i$ and $U_{i2} = \sin\Theta_i$, the following explicit relation exists between $\mathbf{W}_i$ and Θ_i:

$$\mathbf{W}_i = R_i \begin{pmatrix} \cos\Theta_i \\ \sin\Theta_i \end{pmatrix} = R_i \mathbf{U}_i,\ R_i = ||\mathbf{W}_i||. \tag{2}$$

The couple (Θ_i, R_i) is the representation in polar coordinates of $\mathbf{W}_i$.

A natural way to define an $n-$variate projected normal is to consider a $2n$-dimensional vector $\mathbf{W} = \{\mathbf{W}_i\}_{i=1}^n$ distributed as a $2n$-variate normal with mean vector $\boldsymbol{\mu}_w$ and covariance matrix $\boldsymbol{\Sigma}_w$. The random vector $\boldsymbol{\Theta} = \{\Theta_i\}_{i=1}^n$, of associated circular variables, is said to be distributed as an $n-$variate projected normal (PN_n): $\boldsymbol{\Theta} \sim PN_n(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$.

The projected normal distribution is often considered in a univariate setting. Multivariate extensions have been developed in a spatial or spatio-temporal framework only, see for example Wang and Gelfand (2014) or Mastrantonio *et al.* (2015b).

2.2 The skew normal

To model the linear part of telemetry data, we consider a skew normal distribution. Let $\boldsymbol{Y} = \{Y_j\}_{j=1}^q$ be a $q-$variate random variable, let $\boldsymbol{\mu}_y$ be a vector of length q, $\boldsymbol{\Sigma}_y$ be a $q \times q$ covariance matrix and $\boldsymbol{\Lambda}$ be a $q \times q$ matrix with elements belonging to $\mathbb{R}$. $\boldsymbol{Y}$ is distributed as a $q-$variate skew normal (Sahu *et al.*, 2003) with parameters $\boldsymbol{\mu}_y$, $\boldsymbol{\Sigma}_y$ and $\boldsymbol{\Lambda}$ ($\boldsymbol{Y} \sim SN_q(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y, \boldsymbol{\Lambda})$) and it has probability density function (pdf)

$$2^q \phi_q\left(\mathbf{y}|\boldsymbol{\mu}_y, \boldsymbol{\Upsilon}\right) \Phi_q\left(\boldsymbol{\Lambda}'\boldsymbol{\Upsilon}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y)|\mathbf{0}_q, \boldsymbol{\Gamma}\right),$$

where $\phi_q(\cdot|\cdot,\cdot)$ and $\Phi_q(\cdot|\cdot,\cdot)$ indicate respectively the $q-$variate normal pdf and cumulative distribution function, $\mathbf{0}_q$ is a vector of 0s of dimension q, $\boldsymbol{\Upsilon} = \boldsymbol{\Sigma}_y + \boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ and $\boldsymbol{\Gamma} = \mathbf{I}_q - \boldsymbol{\Lambda}'\boldsymbol{\Upsilon}^{-1}\boldsymbol{\Lambda}$. The parameter $\boldsymbol{\Lambda}$ is generally called the *skew parameter* and if all its elements are 0, then $\mathbf{Y} \sim N_q\left(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y\right)$.

[1] atan* is a modified arctangent function defined in Jammalamadaka and SenGupta (2001) pag. 13.

The skew normal distribution has a nice stochastic representation, that follows from Proposition 1 of Arellano-Valle *et al.* (2007). Let $\mathbf{D} \sim HN_q(\mathbf{0}, \mathbf{I}_q)$, where $HN_q(\cdot, \cdot)$ indicates the $q-$dimensional half normal distribution, and $\mathbf{H} \sim N_q(\mathbf{0}, \boldsymbol{\Sigma}_y)$, then

$$\mathbf{Y} = \boldsymbol{\mu}_y + \boldsymbol{\Lambda}\mathbf{D} + \mathbf{H}, \tag{3}$$

and $Y \sim SN_q(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y, \boldsymbol{\Lambda})$. The mean vector and covariance matrix of $\mathbf{Y}$ are given by:

$$E(\mathbf{Y}) = \boldsymbol{\mu}_y + \boldsymbol{\Lambda}\sqrt{\frac{2}{\pi}},$$
$$\text{Var}(\mathbf{Y}) = \boldsymbol{\Sigma}_y + \left(1 + \frac{2}{\pi}\right)\boldsymbol{\Lambda}\boldsymbol{\Lambda}'.$$

In the general case, the (multivariate or univariate) marginal distributions of $\mathbf{Y}$ are not skew normal (Sahu *et al.*, 2003) but if $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = \{\lambda_i\}_{i=1}^q$, then all the marginal distributions are skew normal and λ_i affects only the mean and variance of Y_i.

2.3 The joint linear-circular distribution

The new multivariate circular-linear distribution proposed is obtained as follows. Let

$$(\mathbf{W}, \mathbf{Y})' \sim SN_{2n+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \text{diag}((\mathbf{0}_{2n}, \boldsymbol{\lambda}))),$$

with $\boldsymbol{\mu} = (\boldsymbol{\mu}_w, \boldsymbol{\mu}_y)$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_w & \boldsymbol{\Sigma}_{wy} \\ \boldsymbol{\Sigma}'_{wy} & \boldsymbol{\Sigma}_y \end{pmatrix}$, where $\boldsymbol{\Sigma}$ is a $(2n+q) \times (2n+q)$ covariance matrix. The marginal distribution of $\mathbf{W}$ is a $2n$-variate normal with mean $\boldsymbol{\mu}_w$ and covariance matrix $\boldsymbol{\Sigma}_w$, since the associate skew parameters are all zeros, while $\mathbf{Y} \sim SN_q(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y, \text{diag}(\boldsymbol{\lambda}))$.

If we apply the transformation (1) to the components $\mathbf{W}_i$ of $(\mathbf{W}, \mathbf{Y})$, then $(\boldsymbol{\Theta}, \mathbf{Y})'$ is a multivariate vector of n circular and q linear variables and we say that is distributed as an (n, q)-variate projected-skew normal ($PSN_{n,q}$) with parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\lambda}$: $(\boldsymbol{\Theta}, \mathbf{Y})' \sim PSN_{n,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$. A closed form for the joint distribution is available by introducing suitable latent variables (see Section 3)

As the skew matrix is diagonal, each marginal distribution of $(\mathbf{W}, \mathbf{Y})'$ is still a skew normal with parameters given by the appropriate subset of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\lambda}$. Accordingly all the marginal distributions of $(\boldsymbol{\Theta}, \mathbf{Y})'$ are projected-skew normals.

The interpretation of the parameters $\boldsymbol{\mu}_y$, $\boldsymbol{\Sigma}_y$ and $\boldsymbol{\lambda}$ is straightforward. The interpretation of $(\boldsymbol{\mu}_{wi}, \boldsymbol{\Sigma}_{wi})'$, i.e. the parameters of the univariate marginal projected distribution, is not easy because there is a complex interaction between them and it is not clear how a single component of $\boldsymbol{\mu}_w$ or $\boldsymbol{\Sigma}_w$ affects the shape of the univariate density, that can be symmetric, asymmetric, unimodal or bimodal (for a discussion see Wang and Gelfand (2013)).

However, in a Bayesian framework we can compute Monte Carlo approximations of all the features of the marginal univariate circular distribution (Mastrantonio *et al.*, 2015a), such as the directional mean, the circular concentration and the posterior predictive density, bypassing the difficult in the interpretation of the parameters $(\boldsymbol{\mu}_{wi}, \boldsymbol{\Sigma}_{wi})'$.

The two components of $\mathbf{U}_i$ are respectively the cosine and sine of the circular variable Θ_i, see (2), and the correlation matrix of $(\mathbf{W}, \mathbf{Y})'$, $\boldsymbol{\Omega}$, is the same of $(\mathbf{U}, \mathbf{Y})'$, where $\mathbf{U} = \{\mathbf{U}_i\}_{i=1}^n$. We can easily interpret the circular-circular and circular-linear dependence in terms of the correlation between the linear variables and the sine and cosine of the circular ones.

The parameters of the projected-skew normal are not identifiable, since $\mathbf{W}_i$ and $c_i\mathbf{W}_i$, with $c_i > 0$, produce the same Θ_i, and hence the same $\boldsymbol{\Theta}$. As consequence the distribution of $(\boldsymbol{\Theta}, \mathbf{Y})'$ is unchanged if the parameters $(\boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ are replaced by $(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}, \boldsymbol{\lambda})$, where $\mathbf{C} = \text{diag}(\mathbf{c}, \mathbf{1}_q)$ with $\mathbf{c} = \{(c_i, c_i)'\}_{i=1}^n$; to identify the parameters a constraint is needed. Without loss of generality, following and extending Wang and Gelfand (2013), we can fix the scale of each $\mathbf{W}_i$ by setting to a constant, say 1, each second element of the diagonals of the $\boldsymbol{\Sigma}_{wi}$s. The constrains create some difficult in the estimation of $\boldsymbol{\Sigma}$ since we have to ensure that it is a positive definite (PD) matrix. To avoid confusion, from now to go on we indicate with $\tilde{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\mu}}$ the identifiable version of $\boldsymbol{\Sigma} = \mathbf{C}\tilde{\boldsymbol{\Sigma}}\mathbf{C}$ and $\boldsymbol{\mu} = \mathbf{C}\tilde{\boldsymbol{\mu}}$.

3 The Bayesian inference

Suppose to have T observations drawn from an (n, q)-variate projected-skew normal, $(\boldsymbol{\Theta}_t, \mathbf{Y}_t)' \sim PSN_{n,q}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\lambda})$ with $t = 1, \ldots, T$, where $\boldsymbol{\Theta}_t = \{\Theta_{ti}\}_{i=1}^n$ and $\mathbf{Y}_t = \{Y_{tj}\}_{j=1}^q$. With a slight abuse of notation, let $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_t\}_{t=1}^T$ and $\mathbf{Y} = \{\mathbf{Y}_t\}_{t=1}^T$ and suppose that given $(\boldsymbol{\Theta}, \mathbf{Y})$ we want to learn about $\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\Sigma}}$ and $\boldsymbol{\lambda}$ in a Bayesian perspective using an MCMC algorithm, i.e. obtaining samples from the posterior distribution

$$f(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\lambda}|\boldsymbol{\theta}, \mathbf{y}) \propto \prod_{t=1}^{T} f(\boldsymbol{\theta}_t, \mathbf{y}_t|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\lambda}) f(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\lambda}). \tag{4}$$

We cannot work directly with the posterior (4) since $f(\boldsymbol{\theta}_t, \mathbf{y}_t|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\lambda})$ is not known in closed form and it is not easy to find an appropriate prior distribution $f(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\lambda})$, even if we assume independence between the parameters, i.e. $f(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\lambda}) = f(\tilde{\boldsymbol{\mu}}) f(\tilde{\boldsymbol{\Sigma}}) f(\boldsymbol{\lambda})$, because $f(\tilde{\boldsymbol{\Sigma}})$ must be a valid distribution for a PD matrix with some of its diagonal elements constrained.

To solve both problems let $\mathbf{W}_{ti}$ be the bivariate linear variable associated with Θ_{ti}, let $R_{ti} = ||\mathbf{W}_{ti}||$ and let $\mathbf{D}_t$ be the $q-$variate half normal random variable associated with $\mathbf{Y}_t$ in the stochastic representation given in (3). Let $\mathbf{R}_t = \{R_{ti}\}_{i=1}^n$, $\mathbf{R} = \{\mathbf{R}_t\}_{t=1}^T$ and, again

with a slight abuse of notation, $\mathbf{D} = \{\mathbf{D}_t\}_{t=1}^T$. Instead of the posterior (4) we evaluate, i.e. we obtain posterior samples, from the posterior

$$f(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \mathbf{r}, \mathbf{d}|\boldsymbol{\theta}, \mathbf{y}) \propto \prod_{t=1}^{T} f(\boldsymbol{\theta}_t, \mathbf{r}_t, \mathbf{y}_t, \mathbf{d}_t|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) f(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}), \tag{5}$$

where the joint density of $(\boldsymbol{\Theta}_t, \mathbf{R}_t, \mathbf{Y}_t, \mathbf{D}_t)$ is

$$2^q \phi_{2n+q}((\mathbf{w}_t, \mathbf{y}_t - \text{diag}(\boldsymbol{\lambda})\mathbf{d}_t)'|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \phi_q(\mathbf{d}_t|\mathbf{0}, \mathbf{I}_q) \prod_{i=1}^{n} r_{ti}. \tag{6}$$

Equation (6) is the density that arises by transforming each $\mathbf{W}_{ti}$, in the joint density of $(\mathbf{W}_t, \mathbf{Y}_t, \mathbf{D}_t)$, to its representation in polar coordinate (equation (2)). Note that in (5) we work with the unconstrained PD matrix $\boldsymbol{\Sigma}$ and then the definition of the prior distribution is easier with respect to (4). The posterior distribution (5) is not identifiable, but nevertheless we can obtain samples from it. Suppose to have B samples from (5), i.e. $\{\boldsymbol{\mu}^b, \boldsymbol{\Sigma}^b, \boldsymbol{\lambda}^b, \mathbf{r}^b, \mathbf{d}^b\}_{b=1}^B$. The subset $\{\boldsymbol{\mu}^b, \boldsymbol{\Sigma}^b, \boldsymbol{\lambda}^b\}_{b=1}^B$ are samples from the posterior distribution $f(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}|\boldsymbol{\theta}, \mathbf{y})$ and if we transform the set $\{\boldsymbol{\mu}^b, \boldsymbol{\Sigma}^b, \boldsymbol{\lambda}^b\}_{b=1}^B$ to the set $\{\tilde{\boldsymbol{\mu}}^b, \tilde{\boldsymbol{\Sigma}}^b, \mathbf{C}^b, \boldsymbol{\lambda}^b\}_{b=1}^B$, the latter is a set of samples from $f(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}, \mathbf{C}, \boldsymbol{\lambda}|\mathbf{u}, \mathbf{y})$. As consequence the subset $\{\tilde{\boldsymbol{\mu}}^b, \tilde{\boldsymbol{\Sigma}}^b, \boldsymbol{\lambda}^b\}_{b=1}^B$ are samples from (4), the posterior distribution of interest. From a practical point of view, we can work with (5) and put a prior distribution over $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$. The posterior samples $\{\boldsymbol{\mu}^b, \boldsymbol{\Sigma}^b, \boldsymbol{\lambda}^b\}_{b=1}^B$ are transformed to the set $\{\tilde{\boldsymbol{\mu}}^b, \tilde{\boldsymbol{\Sigma}}^b, \boldsymbol{\lambda}^b\}_{b=1}^B$ that are posterior samples from (4). The prior distribution $f(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\lambda})$ in (4) is induced by $f(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ in (5). To verify what is the real advantage of this approach, let assume $f(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = f(\boldsymbol{\mu}, \boldsymbol{\Sigma}) f(\boldsymbol{\lambda})$. The full conditional of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is proportional to $\prod_{t=1}^T \phi_{2n+q}((\mathbf{w}_t, \mathbf{y}_t - \text{diag}(\boldsymbol{\lambda})\mathbf{d}_t)'|\boldsymbol{\mu}, \boldsymbol{\Sigma}) f(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, i.e. the product of a $(2n+q)$-variate normal density and a prior distribution over its mean and covariance matrix. We can then use the standard prior for the normal likelihood that gives the possibility to find in closed form the full conditional of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

We suggest $\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim NIW(\cdot, \cdot, \cdot, \cdot)$, where $NIW(\cdot, \cdot, \cdot, \cdot)$ indicates the normal inverse Wishart (NIW) distribution. This induces a full conditional for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ that is NIW and then it is easy to simulate with a Gibbs step. We can apply our approach to obtain posterior samples with a Gibbs step even when we have only circular variables, i.e. we are working with the multivariate projected normal, and also in the univariate case where, till now, the components of $\boldsymbol{\Sigma}_{wi}$ were sampled using Metropolis steps, see for example Wang and Gelfand (2013), Wang and Gelfand (2014), Mastrantonio *et al.* (2015a) or Mastrantonio *et al.* (2015b). Under the NIW we are not able to compute, in closed form, the induced prior on $(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ but, if needed, it can always be evaluated through simulation. Of course the NIW it is not the only possible choice, for example can be used the prior proposed by

Huang and Wand (2013) or the one of O'Malley and Zaslavsky (2008), but we think that the NIW is easiest to implement.

To conclude the MCMC specification, we have to show how to sample the remaining parameters and latent variables. Let $\boldsymbol{\mu}_{y_t|w_t} = \boldsymbol{\mu}_y + \boldsymbol{\Sigma}'_{wy}\boldsymbol{\Sigma}_w^{-1}(\mathbf{w}_t - \boldsymbol{\mu}_w)$ and $\boldsymbol{\Sigma}_{y|w} = \boldsymbol{\Sigma}_y - \boldsymbol{\Sigma}'_{wy}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{wy}$. The full conditional of $\boldsymbol{\lambda}$ is proportional to $\prod_{t=1}^{T}\phi_q(\mathbf{y}_t|\boldsymbol{\mu}_{y_t|w_t}+\text{diag}(\mathbf{d}_t)\boldsymbol{\lambda}, \boldsymbol{\Sigma}_{y|w})f(\boldsymbol{\lambda})$ and if we use a multivariate normal prior over $\boldsymbol{\lambda}$, we obtain a multivariate normal full conditional. Let $\mathbf{V}_d = \left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}_{y|w}^{-1}\boldsymbol{\Lambda} + \mathbf{I}_q\right)^{-1}$ and $\mathbf{M}_{d_t} = \mathbf{V}_d\boldsymbol{\Lambda}'\boldsymbol{\Sigma}_{y|w}^{-1}\left(\mathbf{y}_t - \boldsymbol{\mu}_{y|w}\right)$ then the full condition of $\mathbf{d}_t$ is $N_q\left(\mathbf{M}_{d_t}, \mathbf{V}_q\right) I_{\mathbf{0_q},\boldsymbol{\infty}}$, where $N_q\left(\cdot,\cdot\right) I_{\mathbf{0_q},\boldsymbol{\infty}}$ is a $q-$dimensional truncated normal distribution with components having support $\mathbb{R}^+$. We are not able to find in closed form the full conditionals of the r_{ti}s and then we sample them with Metropolis steps.

4 The hidden Markov model

In this Section, we introduce the HMM and its Bayesian non-parametric version, the sHDP-HMM.

4.1 The model

Let $z_t \in \mathcal{K} \subseteq \mathbb{Z}^+\backslash\{0\}$ be a discrete variable which indicates the latent behaviour at time t and let $\boldsymbol{\psi}_k$ be the vector of parameters of the PSN in the behaviour k, i.e. $\boldsymbol{\psi}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\lambda}_k)$.

In the HMM the observations are time independent given $\{z_t\}_{t=1}^T$ and $\{\boldsymbol{\psi}_k\}_{k\in\mathcal{K}}$, i.e.:

$$f(\boldsymbol{\theta}, \mathbf{y}|\{z_t\}_{t\in\mathcal{T}}, \{\boldsymbol{\psi}_k\}_{k\in\mathcal{K}}) = \prod_{t\in\mathcal{T}}\prod_{k\in\mathcal{K}}\left[f(\boldsymbol{\theta}_t, \mathbf{y}_t|\boldsymbol{\psi}_{z_t})\right]^{I(z_t,k)},$$

where $I(z_t, k)$, the indicator function, is equal to 1 if $z_t = k$, 0 otherwise. The hidden variables $\{z_t\}_{t=1}^T$ follow a first-order Markov chain with $P(z_t = k|z_{t-1} = j) = \pi_{jk}$ and

$$z_t|z_{t-1}, \boldsymbol{\pi}_{z_{t-1}} \sim \boldsymbol{\pi}_{z_{t-1}},$$

where $\boldsymbol{\pi}_j = \{\pi_{jk}\}_{k\in\mathcal{K}}$. As pointed out by Cappé *et al.* (2005), the initial state z_0 cannot be estimated consistently since we have no observation at time 0 and then we set $z_0 = 1$. The distribution of $\boldsymbol{\Theta}_t, \mathbf{Y}_t|\boldsymbol{\psi}_{z_t}$, that in the HMM literature it is called the *emission distribution*, is the projected-skew normal, i.e. $\boldsymbol{\Theta}_t, \mathbf{Y}_t|z_t, \boldsymbol{\psi}_{z_t} \sim PSN_{n,q}(\boldsymbol{\mu}_{z_t}, \boldsymbol{\Sigma}_{z_t}, \boldsymbol{\lambda}_{z_t})$. We remark that, although the model is specified with respect to $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, we can only estimate $(\tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k)$

We can equivalently express the hidden process in a more suitable way for the specification of the sHDP-HMM. Let $\boldsymbol{\eta}_t = \boldsymbol{\psi}_{z_t}$, with $\boldsymbol{\psi}_k \in \boldsymbol{\Psi}$, $k \in \mathcal{K}$, and suppose that each element of the sequence $\{\boldsymbol{\eta}_t\}_{t\in\mathcal{T}}$, is drawn from a discrete space $\boldsymbol{\Xi} = \{\boldsymbol{\psi}_k\}_{k\in\mathcal{K}}$ and the probability

of drawing $\boldsymbol{\eta}_t$ depends only on the value $\boldsymbol{\eta}_{t-1}$. We let $P(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1} = \boldsymbol{\psi}_l) \sim G_{\boldsymbol{\eta}_{t-1}} \equiv G_{\boldsymbol{\psi}_l}$, with $G_{\boldsymbol{\psi}_l} = \sum_{k\in\mathcal{K}} \pi_{lk}\delta_{\boldsymbol{\psi}_k}$ where $\delta_{\boldsymbol{\psi}_k}$ is the *Dirac delta function* placed on $\boldsymbol{\psi}_k$. The above HMM can be expressed as

$$\begin{aligned} f(\boldsymbol{\theta}, \mathbf{y}|\{\boldsymbol{\eta}_t\}_{t\in\mathcal{T}}) &= \prod_{t\in\mathcal{T}} f(\boldsymbol{\theta}_t, \mathbf{y}_t|\boldsymbol{\eta}_t), \\ \boldsymbol{\Theta}_t, \mathbf{Y}_t|\boldsymbol{\eta}_t &\sim PSN_{n,q}(\boldsymbol{\mu}_{z_t}, \boldsymbol{\Sigma}_{z_t}, \boldsymbol{\lambda}_{z_t}), \\ \boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1}, G_{\eta_{t-1}} &\sim G_{\eta_{t-1}}. \end{aligned}$$

The standard way to estimate the cardinality of $\mathcal{K}$ (K^*) is to set it a priori and then run models with different values of K^*. The models are compared using informational criteria such as the AIC, BIC or ICL and the model that has the better value of the selected criterion is chosen. They can suggest different values of K^* and moreover they are used to obtain the optimal K^* but without any measurement of uncertainty.

In a Bayesian framework there are several ways to deal with an unknown number of behaviours. We can use the Reversible Jump proposed by Green (1995) or the HDP by Teh *et al.* (2006) and its modification, the sticky hierarchical Dirichlet process (sHDP), proposed by Fox *et al.* (2011). Here we use the sHDP because it is easier to implement. This method let $K^* \to \infty$ and estimates from the data the number of non-empty behaviours, K. K is a random variable and we can have a measurement of uncertainty on its estimate.

4.2 The sHDP-HMM

In the sHDP-HMM is assumed the following:

$$G_{\boldsymbol{\eta}_t}|\tau, \rho, \gamma, H \sim sHDP(\tau, \gamma, \rho, H), \quad \tau > 0, \gamma > 0, \rho \in [0, 1], \tag{7}$$

where with $sHDP(\cdot)$ we indicate the sticky hierarchical Dirichlet process (Fox *et al.*, 2011) with first level concentration parameter τ, second level concentration parameter γ, self-transition parameter ρ and base measure H, where the base measure is a distribution over the space $\boldsymbol{\Psi}$. Fox *et al.* (2011) show that equation (7) can be written equivalently as

$$\begin{aligned} G_{\boldsymbol{\eta}_t}|\rho, \gamma &\sim DP(\gamma, (1-\rho)G_0 + \rho\delta_{\boldsymbol{\eta}_t}), \\ G_0|\tau, H &\sim DP(\tau, H), \end{aligned}$$

where $DP(v, L)$ indicates the Dirichlet process with base measure L and concentration parameter a.

We can write the sHDP-HMM as

$$
\begin{aligned}
f(\boldsymbol{\theta}, \mathbf{y} | \{\boldsymbol{\eta}_t\}_{t \in \mathcal{T}}) &= \prod_{t \in \mathcal{T}} f(\boldsymbol{\theta}_t, \mathbf{y}_t | \boldsymbol{\eta}_t), \\
\boldsymbol{\Theta}_t, \mathbf{Y}_t | \boldsymbol{\eta}_t &\sim PSN_{n,q}(\boldsymbol{\mu}_{z_t}, \boldsymbol{\Sigma}_{z_t}, \boldsymbol{\lambda}_{z_t}), \\
\boldsymbol{\eta}_t | \boldsymbol{\eta}_{t-1}, G_{\eta_{t-1}} &\sim G_{\eta_{t-1}}, \\
G_{\boldsymbol{\eta}_t} | \rho, \gamma &\sim DP(\gamma, (1-\rho) G_0 + \rho \delta_{\boldsymbol{\eta}_t}), \\
G_0 | \tau, H &\sim DP(\tau, H).
\end{aligned}
$$

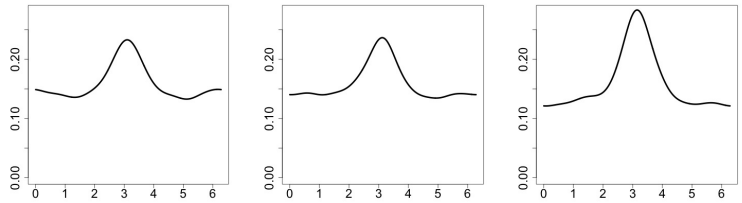
To simplify the implementation we write the model using the stick-breaking representation of the Dirichlet process (Sethuraman, 1994) and we introduce the latent variables $\mathbf{r} = \{\mathbf{r}_t\}_{t=1}^T$ and $\mathbf{d} = \{\mathbf{d}_t\}_{t=1}^T$. Then, let $\boldsymbol{\beta} = \{\beta_k\}_{k=1}^\infty$, the model we estimate with the MCMC algorithm is

$$
\begin{aligned}
f(\boldsymbol{\theta}, \mathbf{y}, \mathbf{r}, \mathbf{d} | \{z_t\}_{t \in \mathcal{T}}, \{\boldsymbol{\psi}_k\}_{k \in \mathcal{K}}) &= \prod_{t \in \mathcal{T}} \prod_{k \in \mathcal{K}} \left[f(\boldsymbol{\theta}_t, \mathbf{r}_t, \mathbf{y}_t, \mathbf{d}_t | \boldsymbol{\psi}_k) \right]^{I(k = z_t)}, \\
z_t | z_{t-1}, \boldsymbol{\pi}_{z_{t-1}} &\sim \boldsymbol{\pi}_{z_{t-1}}, \\
\boldsymbol{\pi}_k | \rho, \gamma, \boldsymbol{\beta}, \boldsymbol{\psi}_k &\sim DP\left(\gamma, (1-\rho)\boldsymbol{\beta} + \rho \delta_{\boldsymbol{\psi}_k}\right), \\
\boldsymbol{\beta} | \tau &\sim GEM(\tau), \\
\boldsymbol{\psi}_k | H &\sim H,
\end{aligned}
$$

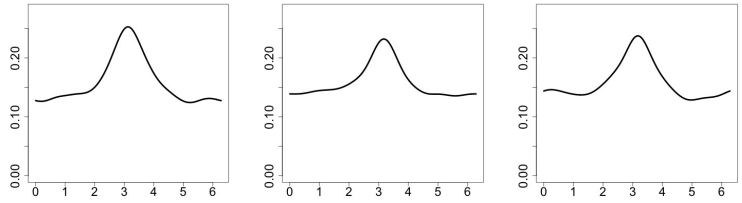
where $GEM(\cdot)$ indicates the stick-breaking process (Sethuraman, 1994).

To complete the model we have to specify the base measure H, that in the stick-breaking representation acts as a prior distribution over the parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\lambda}_k)$. We use $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \sim NIW(\boldsymbol{\mu}_0, \eta, \varsigma, \boldsymbol{\Psi})$ and $\boldsymbol{\lambda} \sim N_q(\mathbf{M}, \mathbf{V})$ because, as noted in Section 2.3, these choices lead to full conditionals easy to simulate.

We assume that the parameters of the sHDP, τ, γ and ρ, are random quantities and following Fox *et al.* (2011) we choose as priors: $\tau \sim G(a_\tau, b_\tau)$, $\gamma \sim G(a_\gamma, b_\gamma)$ and $\rho \sim B(a_\rho, b_\rho)$, where $G(\cdot, \cdot)$ indicates the gamma distribution, expressed in terms of shape and scale, and $B(\cdot, \cdot)$ is the beta. Since we treat ρ as a random variable we can estimate through the data the strength of the self-transition. For the MCMC sampling of the behaviour indicator variables we use the *beam sampler* (Van Gael *et al.*, 2008). Despite the complexity of the model, in terms of emission distribution and the underline Markov structure, with the exception of the r_{it}s, all the other unknown quantities can be updated in the MCMC with Gibbs steps.



(a) First dog (b) Second dog (c) Third dog



(d) Fourth dog (e) Fifth dog (f) Sixth dog

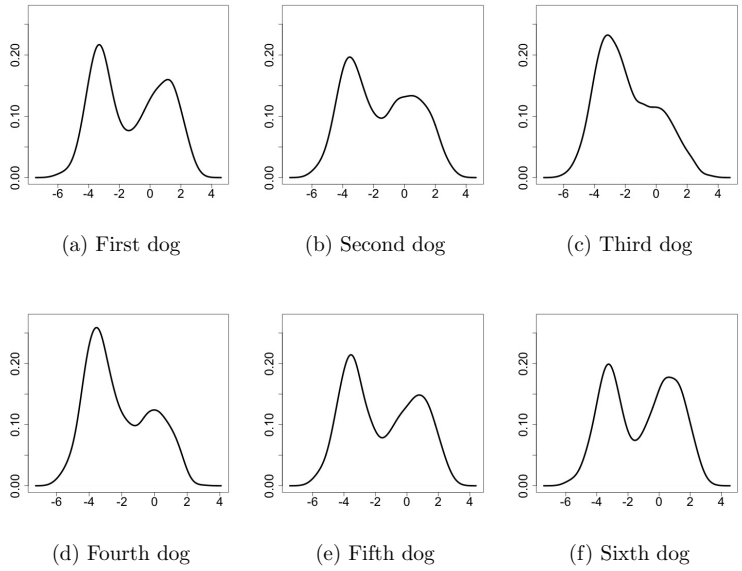
Figure 1: Marginal densities of the turning-angles.

5 Real data example

In this Section we apply our proposal on a real dataset, taken from the movebank website (van Bommel and Johnson, 2014a).

5.1 Data description

Data on free-ranging Maremma sheepdogs positions are recorded by tracking collars every 30 minutes. The behaviour of the dogs is unknown because there is minimal supervision by their owners and the animals are allowed to range freely. The dataset was first analyzed in van Bommel and Johnson (2014b) with the aim to understand how much space the dogs utilize and the portion of time that the dogs spent with livestock. Even if the primary purpose was not to identify behavioural modes, van Bommel and Johnson (2014b) results show that the dogs can be clustered in two states, one characterized by low speeds and tortuous path at the core of their home ranges (we call it state VB1), when they are resting or attending livestock, and large step-lengths (i.e. high movement speeds) in relatively straight lines, related to boundary patrolling or seeing off predators (we call it state VB2), at the edge of their home ranges.

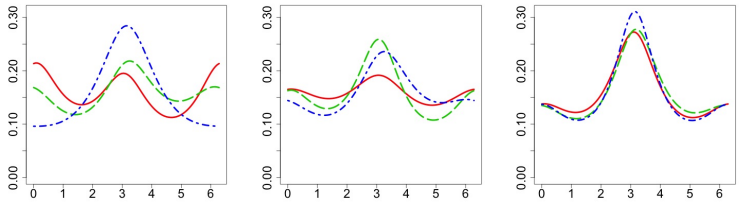


(a) First dog (b) Second dog (c) Third dog

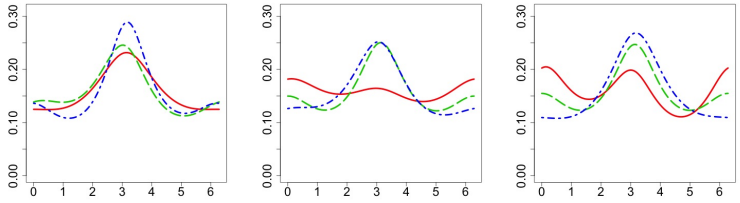
(d) Fourth dog (e) Fifth dog (f) Sixth dog

Figure 2: Marginal densities of the log-step-lengths.

We characterize the hidden behaviours by analyzing the turning-angles and the logarithm of the step-lengths[2] (log-step-lengths) for each dog taken into consideration. We model movement metrics belonging to dogs sharing the same property and observed on a common time period. We select the data from the "Heatherlie" property where between the 08/02/2012 5:30 and 10/03/2012 17:00, six dogs are observed, having then a time series of 3000 points with 6 circular and 6 linear variables. The six dogs have respectively 107, 63, 231, 43, 31 and 63 missing circular observations and 95, 53, 117, 32, 22 and 45 linear ones. The density estimates of the turning-angles and log-step-lengths can be seen in Figures 1 and 2. In the selected property and observational interval, van Bommel and Johnson (2014b) note that four of the six dogs form one social group responsible for protecting all livestock, one dog is old and mostly solitary and the last one suffers of an extreme social exclusion which severely restrict its movements. The animals that are part of the social group are often found together, but regularly they split into sub-groups.



(a) First dog (b) Second dog (c) Third dog

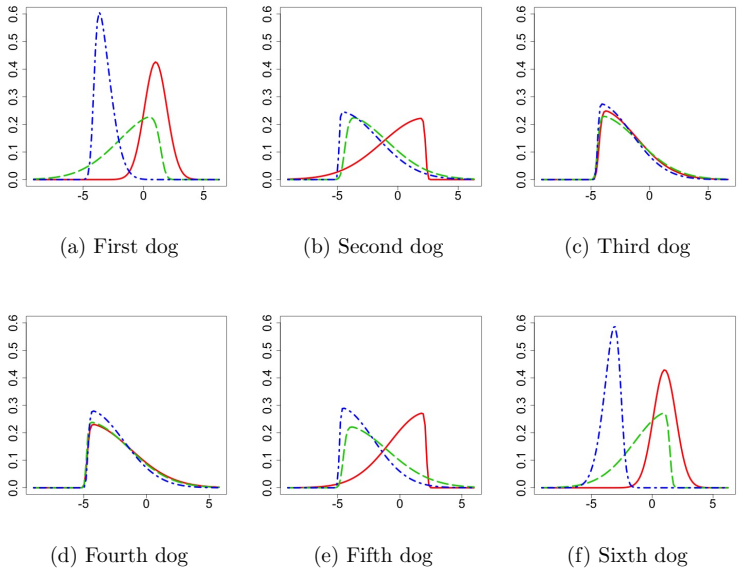


(d) Fourth dog (e) Fifth dog (f) Sixth dog

Figure 3: Posterior marginal densities of the turning-angles. The dashed-dotted line is the marginal density of the first behaviour, the dashed one is the second and the full the third. $\hat{K} = 3$.

	1	2	3
1	0.711	0.181	0.108
	(0.679 0.741)	(0.154 0.209)	(0.085 0.132)
2	0.141	0.672	0.187
	(0.121 0.164)	(0.640 0.704)	(0.161 0.215)
3	0.083	0.209	0.708
	(0.065 0.103)	(0.180 0.239)	(0.676 0.739)

Table 1: Posterior mean estimates and 95 % credible intervals for the transition probability matrix: $\hat{K} = 3$.



(a) First dog (b) Second dog (c) Third dog

(d) Fourth dog (e) Fifth dog (f) Sixth dog

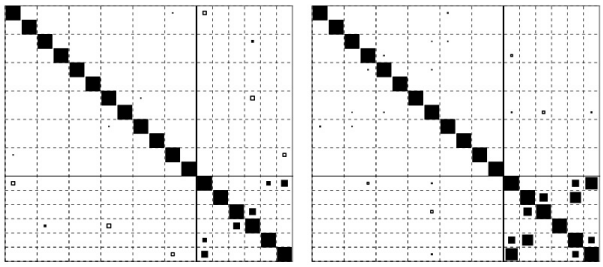
Figure 4: Posterior marginal densities of the log-step-lengths. The dashed-dotted line is the marginal density of the first behaviour, the dashed one is the second and the full the third. $\hat{K} = 3$.

5.2 Results

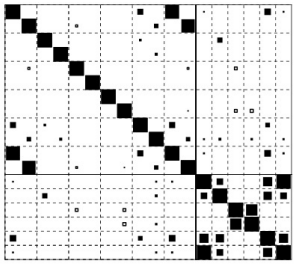
The model is estimated considering 400000 iterations, burnin 300000, thin 20 and by taking 5000 samples for inferential purposes. As prior distributions we choose $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \sim NIW(\mathbf{0}_6, 0.001, 25, \mathbf{I}_6)$ and $\boldsymbol{\lambda}_k \sim N_2(\mathbf{0}_2, 50\mathbf{I}_2)$ that are standard weak informative distributions. For the parameter ρ, that governs the self-transition probabilities, we decide to use $\rho \sim B(1,1)$, that is equivalent to a uniform distribution over $[0,1]$, while $\tau \sim G(1, 0.01)$ and $\gamma \sim G(1, 0.01)$. The priors of ρ, τ and γ induce a prior over K (Fox *et al.*, 2011) that we evaluated through simulation, by using the *degree ℓ weak limit approximation* (Ishwaran and Zarepour, 2002) with $\ell = 1000$; we found that $K \in [4, 465]$ with a coverage of 95%.

The model estimates 3 behavioural modes with $P(K = 3|\boldsymbol{\theta}, \mathbf{y}) = 1$. From the analysis of the posterior marginal distributions, Figures 3 and 4, and the correlation matrices $\boldsymbol{\Omega}_k$, Figure 5, we can easily interpret the behaviour modes, we can confirm what van Bommel and Johnson (2014b) found, we find connections between our estimated behavioural modes

[2]The logarithm is needed since the linear components of the projected-skew normal must be defined over $\mathbb{R}$.



(a) First behaviour (b) Second behaviour



(c) Third behaviour

Figure 5: Graphical representation of the posterior means of the correlation matrices $\mathbf{\Omega}_k$s. A filled square indicates a positive value while an empty one a negative; a square is depicted only if the associated correlation is significantly different from 0. The dimension of the square is proportional to the absolute value of the associated mean correlation coefficient. The full lines separate the correlation matrix of the cosine and sine of the circular variables (top-left), the correlation matrix of the linear ones (bottom-right) and the correlation matrix between the linear variables and the sine and cosine of the circular ones (top-right). $\hat{K} = 3$.

and the states VB1 and VB2 hypothesized by van Bommel and Johnson (2014b), and we add some new results. There are three groups of dogs that share similar marginal distributions: the dogs group one (DG1) composed by the dogs three and four, Figure 3 (c) and (d) and Figure 4 (c) and (d), the dog group two (DG2) is composed by the dogs two and five, Figure 3 (b) and (e) and Figure 4 (d) and (e), and the dogs one and six form the third group (DG3), Figure 3 (a) and (f) and Figure 4 (a) and (f).

In the first behaviour, all the dogs are in state VB1. They have small log-step-lengths and there are few movements in a straight line, i.e. the circular densities have low values

at 0. There are few correlations that differ significantly from 0[3], see Figure 5 (a), and they have all mean posterior values (PEs) below 0.5. The stronger correlation (PE 0.474) is between the linear variables of the first and sixth dogs, i.e. the one in the DG3.

In the second behaviour, the linear distributions of the dogs in the DG3 have more mass of probability at higher values, but still having mass at low ones, and the correlation between the log-step-lengths increases (PE 0.840). The dogs in the DG3 have more movements in a straight line. The linear and circular distributions of the dogs in the DG1 and DG2 are similar to the ones of the first behaviour. The log-step-length of the dog two is correlated with the one of the dog three (PE 0.568) and five (PE 0.757) and the linear variables of the dogs five and six are correlated with PE 0.450. In this behaviour the dogs in the DG3 move to the state VB2 while the ones of the DG1 and DG2 remain in the state VB1.

In the third behaviour the linear distributions of the dogs in the DG3 put all the mass of probability on high values and the circular distributions have two modes, with more or less the same heights, at about 0 and 3.141. The correlation between the linear variables of the dogs in the DG3 increases, with respect to the second behaviour, and it is almost one (PE 0.982). The two dogs change direction one accordingly to the other, since the cosine and sine of the circular variables are highly correlated (PEs 0.958 and 0.916). The linear distributions of the dogs in the DG2 move their mass of probability to higher values and the distributions resemble the ones of the dogs in the DG3, second behaviour. The circular distributions of the dogs in the DG2 are close to the circular uniform. The circular and linear distributions of the dogs in the DG1 are similar to the one of the first and second behaviours but their linear variables are now correlated (PE 0.781). In this behaviour the dogs in the DG3 remain in the state VB2 but, with respect to the second behaviour, they increase the amount of movement (in terms of higher step-length and more tortuous path). The dogs in the DG2 are in the state VB2 while the one in the DG1 remain in the state VB1. The dogs in the DG2 and DG3 have all the linear variables correlated.

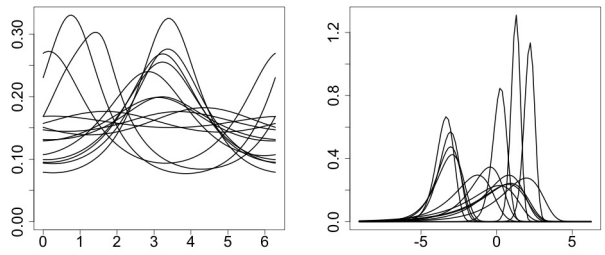
The fourth dog is the one that suffers of an extreme social exclusion since its variables (circular and linear) are never correlated with the ones of the other dogs, with the exception of the third. The third dog is, probably, the old one since it does not move a lot, see Figure 4 (c), it is solitary, i.e. it bonds (in terms of correlation) only with the socially excluded one and, occasionally, with the dog two (second behaviour). The dogs in the DG2 and DG3 form one social group in the third behaviour, i.e. all the linear variables are correlated, and they split into subgroup in the behaviours one and two.

From Table 1 we see that there is a strong self-transition in all three behaviours (re-

[3]A correlation differs significantly from 0 if its 95% credible interval (CI) does not contain the 0.

	PSN	VMLG	VMLW	WCLG	WCLW
$\hat{K}$	3	14	14	14	13
$CRPSc$	0.489	0.491	0.499	0.492	0.501
$CRPSl$	2.342	2.782	3.225	2.466	2.965

Table 2: Estimated number of non-empty behaviours ($\hat{K}$), mean CRPS for circular ($CRPSc$) and linear ($CRPSl$) variables, for the five models based on the PSN, VMLG, VMLW, WCLG and WCLW.



(a) (b)

Figure 6: Posterior marginal densities of the turning-angle (a) and the log-step-length (b) of the first dog estimated with the emission distribution WCLG.

spectively PE 0.711, 0.672 and 0.708). The CIs of the probabilities to move to a new empty behaviour ($\sum_{k=4}^{\infty} \pi_{jk}$, $j = 1, 2, 3$), not shown in Table 1, have always right side limit below 0.00001.

5.3 Comparisons with other emission distributions

In this Section we show that our proposed emission distribution performs better, on the data we used in Section 5.2, than the standard distributions used in the literature. We are going to estimate sHDP-HMMs with different emission distributions and we compare the results. Unfortunately there are not measures of goodness of fit, such as the AIC or the BIC, when the model is based on the sHDP. Then we decide to base our comparison between models in terms of missing observations estimate, i.e. predictive ability, and behavioural modes interpretability.

To have a measure of how the model estimates the missing, we randomly select, for each circular and linear variables, 10% of the observations. We treat them as missing and,

using the *continuous ranked probability score* (CRPS) (Matheson and Winkler, 1976), we compare the holdout values with the associated posterior distributions. The CRPS is a proper scoring rule that can be easily computed for both circular (Grimit *et al.*, 2006) and linear (Gneiting and Raftery, 2007) variables using the MCMC output. Let $\mathcal{C}_i$ be the set of time points where the i^{th} value of the circular variable is setted as missing, $\mathcal{L}_j$ be the ones of the j^{th} linear variable and let θ_{ti}^b, $t \in \mathcal{C}_i$ and y_{tj}^b, $t \in \mathcal{L}_j$ be respectively the b^{th} posterior sample of θ_{ti} and y_{tj}. A Monte Carlo approximation of the CRPS for a circular variable is computed as

$$CRPSc_i \approx \frac{1}{B}\sum_{b=1}^{B} d(\theta_{ti}, \theta_{ti}^b) - \frac{1}{2B^2}\sum_{b=1}^{B}\sum_{b'=1}^{B} d(\theta_{ti}^b, \theta_{ti}^{b'}),\ t \in \mathcal{C}_i,$$

where $d(\cdot,\cdot)$ is the angular distance. The CRPS for a linear variable is approximated with

$$CRPSl_j \approx \frac{1}{B}\sum_{b=1}^{B} |y_{tj} - y_{tj}^b| - \frac{1}{2B^2}\sum_{b=1}^{B}\sum_{b'=1}^{B} |y_{tj}^b - y_{tj}^{b'}|,\ t \in \mathcal{L}_j.$$

We then compute the overall mean CRPS for the circular variables, $CRPSc = \frac{1}{n}\sum_{i=1}^{n} CRPSc_i$, and the linear ones, $CRPSl = \frac{1}{q}\sum_{j=1}^{q} CRPSl_j$, and we use these two indeces to measure the ability of the model in estimating the missing observations.

It is generally supposed, in the literature, that the turning-angle is distributed as a von Mises (Langrock *et al.*, 2012; Holzmann *et al.*, 2006; Eckert *et al.*, 2008) or a wrapped Cauchy (Langrock *et al.*, 2012; Eckert *et al.*, 2008; Morales *et al.*, 2004; Holzmann *et al.*, 2006) while the gamma (Langrock *et al.*, 2012; Holzmann *et al.*, 2006) or the Weibull (Langrock *et al.*, 2012; Morales *et al.*, 2004) are used for the step-length; these distributions are compared with our proposal. In the model specification, Section 4.1, we assume that each linear variable belongs to $\mathbb{R}$ and then, instead of the gamma and Weibull, we use the log-gamma and log-Weibull, i.e. the distributions that arise by taking the log of, respectively, a random variable gamma or Weibull distributed. The model in Section 4 is compared with the ones based on the von Mises and the log-gamma (VMLG), the von Mises and the log-Weibull (VMLW), the wrapped Cauchy and the log-gamma (WCLG), the wrapped Cauchy and the log-Weibull (WCLW). There is not an obvious way to introduce dependence between the movement metrics on the model VMLG, VMLW, WCLG and WCLW, and, in the literature, they are generally supposed to be independent (see for example Morales *et al.* (2004) or Langrock *et al.* (2012)). Then, in these models, we assume the following:

$$f(\boldsymbol{\theta}, \mathbf{y} | \{z_t\}_{t\in\mathcal{T}} \{\boldsymbol{\psi}_k\}_{k\in\mathcal{K}}) = \prod_{t\in\mathcal{T}}\prod_{k\in\mathcal{K}} \left[\prod_{i=1}^{n} f(\theta_{ti}|\boldsymbol{\psi}_{z_t}) \prod_{j=1}^{q} f(y_{tj}|\boldsymbol{\psi}_{z_t})\right]^{I(z_t,k)}.$$

We use a $G(1, 0.5)$ as prior for the *shape* and *rate* parameters of the log-gamma and the log-Weibull, that is a standard weak-informative prior. For the two parameters of the wrapped Cauchy, one defined over $[0, 2\pi)$ and one over $[0, 1]$, we use uniform distributions in the respective domains while on the two parameters of the von Mises, one defined over $[0, 2\pi)$ and one over $\mathbb{R}^+$, we use respectively the non-informative $U(0, 2\pi)$ and the weak informative $G(1, 0.5)$. As prior distributions for the sHDP parameters, we use the same used in Section 5.

In Table 2 we can see the estimated number of non-empty behaviours ($\hat{K}$), $CRPSc$s and $CRPSl$s. The predictive ability of our model outperforms all the others in both CRPS for circular and linear variables. The models based on the VMLG, VMLW, WCLG and WCLW estimate a larger number of behaviours, with respect to our proposal, i.e. in three of them $\hat{K}$ =14 and in one $\hat{K}$ =13; we can see an example of the estimated behaviours in Figure 6. It is challenging to give an interpretation to these behaviours and moreover such a large number of behaviours does not increase the predictive ability of the models, see Table 2.

A possible explanation of why the models based on the VMLG, VMLW, WCLG and WCLW estimate 13 or 14 behaviours can be found in Mastrantonio *et al.* (2015a). They simulate datasets using bivariate emission distributions with dependent components, bimodal marginals for the circular variable, with the aim to understand what happens if, on the simulated datasets, are estimated HMMs with emission distributions that do not allow for dependent components and bimodality in the circular marginal. They found that the number of behaviours is generally overestimated, since each behaviour is separated into two, one for each mode of the circular distribution.
In our real data application, we have six circular and six linear variables, some of them are correlated and often the marginal circular distributions have two modes, see Figure 3. If we assume independence between the circular and linear variables, as we did in the model base on the VMLG, VMLW, WCLG and WCLW, and the marginal circular distributions are unimodal, as the von Mises and the wrapped Cauchy, then we can expect a large number of behaviours.

6 Conclusions

The primary objective of this work, motivated by our real data, was to introduce an HMM capable of modelling a group of animals, taking into account possible correlations between the associated movement metrics. For this reason we introduced a new multivariate circular-linear distribution, namely the projected-skew normal. The new distribution has dependent components and it is used as emission distribution in the HMM. The HMM was estimated

under a non-parametric Bayesian framework and we showed how to implement the MCMC algorithm.

The model, applied to the real data example, confirmed known results and added new ones. We showed that our emission distribution outperforms the most used in the literature in terms of predictive ability and the estimated behaviours are more easily interpretable.

Future work will lead us to incorporate covariates to model the circular-linear mean and variance of the projected-skew normal and we will explore different temporal dependence structures, such as the semi-HMM or the autoregressive-HMM.

References

Abe, T. and Ley, C. (2015). A tractable, parsimonious and highly flexible model for cylindrical data, with applications. *ArXiv e-prints*.

Arellano-Valle, R., Bolfarine, H., and Lachos, V. (2007). Bayesian inference for skew-normal linear mixed models. *Journal of Applied Statistics*, **34**(6), 663–682.

Blackwell, P. G. (2003). Bayesian inference for markov processes with diffusion and discrete components. *Biometrika*, **90**(3), 613–627.

Brownlee, J. (1912). The mathematical theory of random migration and epidemic distribution. *Proceedings of the Royal Society of Edinburgh*, **31**, 262–289.

Cagnacci, F., Boitani, L., Powell, R. A., and Boyce, M. S. (2010). Animal ecology meets gps-based radiotelemetry: a perfect storm of opportunities and challenges. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **365**(1550), 2157–2162.

Cappé, O., Moulines, E., and Ryden, T. (2005). *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer.

Ciucci, P., Reggioni, W., Maiorano, L., and Boitani, L. (2009). Long-distance dispersal of a rescued wolf from the northern apennines to the western alps. *The Journal of Wildlife Management*, **73**(8), 1300–1306.

D'Elia, A. (2001). A statistical model for orientation mechanism. *Statistical Methods and Applications*, **10**(1-3), 157–174.

Eckert, S. A., Moore, J. E., Dunn, D. C., van Buiten, R. S., Eckert, K. L., and Halpin, P. N. (2008). Modeling loggerhead turtle movement in the mediterranean: importance of body size and oceanography. *Ecological Applications*, **18**(2), 290–308.

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011). A sticky hdp-hmm with application to speaker diarization. *The Annals of Applied Statistics*, **5**(2A), 1020–1056.

Franke, A., Caelli, T., and Hudson, R. J. (2004). Analysis of movements and behavior of caribou (rangifer tarandus) using hidden markov models. *Ecological Modelling*, **173**(2â3), 259 – 270.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**(477), 359–378.

Green, P. J. (1995). Reversible jump Markov chain monte carlo computation and Bayesian model determination. *Biometrika*, **82**(4), 711–732.

Grimit, E. P., Gneiting, T., Berrocal, V. J., and Johnson, N. A. (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, **132**(621C), 2925–2942.

Hanks, E. M., Hooten, M. B., and Alldredge, M. W. (2015). Continuous-time discrete-space models for animal movement. *Ann. Appl. Stat.*, **9**(1), 145–165.

Holzmann, H., Munk, A., Suster, M., and Zucchini, W. (2006). Hidden Markov models for circular and linear-circular time series. *Environmental and Ecological Statistics*, **13**(3), 325–347.

Hooten, M., Johnson, D., Hanks, E., and Lowry, J. (2010). Agent-based inference for animal movement and selection. *Journal of Agricultural, Biological, and Environmental Statistics*, **15**(4), 523–538.

Horne, J. S., Garton, E. O., Krone, S. M., and Lewis, J. S. (2007). Analyzing animal movements using brownian brigdes. *Ecology*, **88**(9), 2354–2363.

Houston, A. I. and Mcnamara, J. M. (1999). *Models of Adaptive Behaviour: An approach based on state*. Cambridge University Press, Cambridge, United Kingdom.

Huang, A. and Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, **1**, 1–14.

Ishwaran, H. and Zarepour, M. (2002). Exact and approximate sum representations for the dirichlet process. *Canadian Journal of Statistics*, **30**(2), 269–283.

Jammalamadaka, S. R. and SenGupta, A. (2001). *Topics in Circular Statistics*. World Scientific, Singapore.

Jonsen, I. D., Flemming, J. M., and Myers, R. A. (2005). Robust state-space modeling of animal movement data. *Ecology*, **86**(11), 2874–2880.

Jonsen, I. D., Myers, R. A., and James, M. C. (2006). Robust hierarchical state-space models reveal diel variation in travel rates of migrating leatherback turtles. *Journal of Animal Ecology*, **75**(5), 1046–1057.

Jonsen, I. D., Myers, R. A., and James, M. C. (2007). Identifying leatherback turtle foraging behaviour from satellite telemetry using a switching state-space model. *Marine Ecology Progress Series*, **337**, 255–264.

Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D., and Morales, J. M. (2012). Flexible and practical modeling of animal telemetry data: hidden Markov models and extensions. *Ecology*, **93**(11), 2336–2342.

Langrock, R., Hopcraft, G., Blackwell, P., Goodall, V., King, R., Niu, M., Patterson, T., Pedersen, M., Skarin, A., and Schick, R. (2014). Modelling group dynamic animal movement. *Methods in Ecology and Evolution*, **5**(2), 190–199.

Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics*. John Wiley and Sons, Chichester.

Mastrantonio, G., Maruotti, A., and Jona Lasinio, G. (2015a). Bayesian hidden markov modelling using circular-linear general projected normal distribution. *Environmetrics*, **26**, 145–158.

Mastrantonio, G., Jona Lasinio, G., and Gelfand, A. E. (2015b). Spatio-temporal circular models with non-separable covariance structure. *TEST*, **To appear**.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, **22**(10), 1087–1096.

McClintock, B. T., King, R., Thomas, L., Matthiopoulos, J., McConnell, B. J., and Morales, J. M. (2012). A general discrete-time modeling framework for animal movement using multistate random walks. *Ecological Monographs*, **82**(3), 335–349.

Morales, J. M., Haydon, D. T., Frair, J., Holsinger, K. E., and Fryxell, J. M. (2004). Extracting more out of relocation data: building movement models as mixtures of random walks. *Ecology*, **85**(9), 2436–2445.

Morales, J. M., Moorcroft, P. R., Matthiopoulos, J., Frair, J. L., Kie, J. G., Powell, R. A., Merrill, E. H., and Haydon, D. T. (2010). Building the bridge between animal movement and population dynamics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**(1550), 2289–2301.

O'Malley, A. J. and Zaslavsky, A. M. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association*, **103**(484), 1405–1418.

Patterson, T., Thomas, L., Wilcox, C., Ovaskainen, O., and Matthiopoulos, J. (2008). State-space models of individual animal movement. *Trends in Ecology & Evolution*, **23**(2), 87–94.

Patterson, T. A., Basson, M., Bravington, M. V., and Gunn, J. S. (2009). Classifying movement behaviour in relation to environmental conditions using hidden markov models. *Journal of Animal Ecology*, **78**(6), 1113–1123.

Sahu, S. K., Dey, D. K., and Branco, M. D. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *Canadian Journal of Statistics*, **31**(2), 129–150.

Schliehe-Diecks, S., Kappeler, P. M., and Langrock, R. (2012). On the application of mixed hidden markov models to multiple behavioural time series. *Interface Focus*, **2**(2), 180–189.

SenGupta, A. (2004). On the construction of probability distributions for directional data. *Bulletin of Calcutta Mathematical Society*, **96**, 139–154.

Sengupta, A. and Ong, S. H. (2014). A unified approach for construction of probability models for bivariate linear and directional data. *Communications in Statistics - Theory and Methods*, **43**(10-12), 2563–2569.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, **101**(476), 1566–1581.

van Bommel, L. and Johnson, C. (2014a). Data from: Where do livestock guardian dogs go? movement patterns of free-ranging maremma sheepdogs. movebank data repository.

van Bommel, L. and Johnson, C. (2014b). Where do livestock guardian dogs go? movement patterns of free-ranging maremma sheepdogs. *PLoS ONE*, **9**(10).

Van Gael, J., Saatci, Y., Teh, Y. W., and Ghahramani, Z. (2008). Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1088–1095, New York, NY, USA. ACM.

Wang, F. and Gelfand, A. E. (2013). Directional data analysis under the general projected normal distribution. *Statistical Methodology*, **10**(1), 113–127.

Wang, F. and Gelfand, A. E. (2014). Modeling space and space-time directional data using projected Gaussian processes. *Journal of the American Statistical Association*, **109**(508), 1565–1580.

Chapter 5

Discussion

The main contribution of this work is to provide new models for temporal and spatio-temporal circular and circular-linear data, showing that complex models can be fitted in a fully Bayesian framework, with efficient and easy to implement algorithms and introducing joint and conditional modelling of circular and linear variables.

In Chapter 3.1, we extend the two state of the art models for spatio-temporal data, namely the wrapped Gaussian and the projected Gaussian processes, introducing a flexible correlation function and adding a nugget effect. We also use linear (discrete and continuous) covariates to model the circular mean and variance of the circular processes. The predictive performances of the models proposed are compared on a real data example.

The projected Gaussian process is highly flexible but parameters interpretation is challenging. On the other hand, the wrapped Gaussian has a straightforward parameters interpretation but, since its univariate distribution is unimodal and symmetric, it is not flexible as the projected one. In Chapter 3.2, we propose a new spatio-temporal process, the wrapped skew Gaussian, that is a generalization of the wrapped Gaussian. With a simulation study we show that the new process is more flexible than the wrapped Gaussian and, differently from the projected Gaussian process, retains and easy and straightforward parameters interpretation.

In Chapter 4.1 we propose and HMM for cylindrical data. The marginal circular distribution is based on the projected normal. With a simulation study we show the important improvement obtained by adopting our proposal versus the common practice of ignoring linear-circular dependence and/or circular bimodality. We are consistently able to avoid the over estimation in the number of regimes that otherwise can easily occur.

The model in Chapter 4.2 extends the one of Chapter 4.1, allowing to model a multivariate time series of circular and linear data. We propose a new circular-linear distribution with multivariate dependence, based on the projected normal (circular variables) and the skew normal (linear variables) and we show how to estimate its parameters in a Bayesian framework. The proposed HMM is based on the hierarchical Dirichlet process that allows us to not fix at prior the number of latent regimes. The proposed model is estimated on a real dataset and its predictive performance, as well as the posterior interpretation of the

obtained regimes, are compare with the emission distributions most used in the literature.

Future work will be devoted to the development and improvement of models for discrete circular variables as at present, a lack of efficient and general models for these type of problems is in the literature. We will also focus in the use of circular variables to describe periodic phenomena.