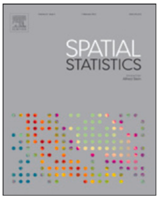
Spatial Statistics 41 (2021) 100497

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta



Inference in cylindrical models having latent Markovian classes—With an application to ocean current data

Henrik Syversveen Lie, Jo Eidsvik [*]

Department of Mathematical Sciences, NTNU, 7491 Trondheim, Norway

ARTICLE INFO

Article history:
Received 24 August 2020
Received in revised form 20 December 2020
Accepted 25 January 2021
Available online 6 February 2021

Keywords:
Composite likelihood
Cylindrical data
Markov random field
Ocean surface currents

ABSTRACT

Spatial direction vector data can be represented cylindrically by linear magnitudes and circular angles. We analyze such data by using a hierarchical Markov random field model with latent discrete classes and conditionally independent cylindrical data given the classes. The structure of a Potts model segments the spatial domain, and each class defines a cylindrical density that represents a specific structure. We consider two types of cylindrical distributions; the Weibull sine-skewed von Mises distribution, which is skewed in the circular part, and the generalized Pareto-type wrapped Cauchy distribution, which is heavy-tailed in the linear part. In this setting, we develop a statistically efficient block composite likelihood method for parameter estimation. The method is shown to provide much faster convergence than an expectation–maximization approach. However, the convergence is less stable for the block composite likelihood method, and we suggest a hybrid estimation approach for practical use. We apply the approach to study ocean surface currents in the Norwegian Sea. The models are able to describe the currents in terms of interpretable local regimes of two–three classes. Scoring rules are used to measure predictive performance of the two cylindrical densities. Results indicate that there is clearly skew angular components, and possibly also some heavy tails in magnitude.

[*] Corresponding author.
E-mail address: jo.eidsvik@ntnu.no (J. Eidsvik).

https://doi.org/10.1016/j.spasta.2021.100497

1. Introduction

Several applications of spatial statistics include the analysis of cylindrical data, e.g., wave direction and height (Wang et al., 2015), wind direction and speed (Modlin et al., 2012), animal migration direction and intensity (Hanks et al., 2015), or ocean currents (Lagona and Picone, 2016). Such data can be represented as a bivariate spatial series of angles and magnitudes at locations in a domain of interest. The term 'cylindrical data' refers to the combination of a circular angle and a linear magnitude, where the pair can be interpreted as a point on a cylinder. The special topology of the cylindrical support for the data is challenging for the statistical modeling, and cylindrical models may need to account for skewness, heavy tails, multimodality and asymmetry in the marginal distributions.

Circular and cylindrical data are two forms of directional data. A comprehensive treatment of all forms of directional data was given by Mardia and Jupp (1999), with Pewsey and García-Portugués (2020) providing a review of more recent advances. The literature for cylindrical data revolves mainly around conditional modeling, i.e., the linear variable depends on the circular variable (circular–linear regression) or the circular variable depends on the linear variable (linear–circular regression). Johnson and Wehrly (1978) developed distributions for cylindrical data based on a maximum entropy principle, whereas Mardia and Sutton (1978) created a distribution by conditioning in a trivariate Gaussian distribution. The former distribution was improved by Abe and Ley (2017), who invoked a power transformation to the linear part and a perturbation to the circular part to allow for asymmetric "sine-skewing" . Recently, Imoto et al. (2019) also proposed a way to model cylindrical data with heavy-tailed linear parts.

Approaches for spatial cylindrical data have been more limited because it is difficult to (i) phrase realistic spatial models for such data, and (ii) conduct reliable inference of model parameters. Wang and Gelfand (2014) utilized the projected Gaussian process to model circular variables on a spatial domain. Their model handles asymmetric data and accounts for spatio-temporal dependencies. This could be advantageous compared with the wrapped Gaussian process proposed by Jona-Lasinio et al. (2012), which only models symmetric spatially dependent circular data. See also Markov chain Monte Carlo implementations for inference of spatio-temporal circular data in such models (Jona Lasinio et al., 2020). The projected Gaussian process was further extended by Wang et al. (2015) to cylindrical data, resulting in a framework for joint modeling of the circular and linear components. Their model specifies a conditional Gaussian distribution for height, given the direction, and a marginal spatio-temporal projected Gaussian distribution for the direction. Another type of hierarchical model was developed by Modlin et al. (2012), based on a circular conditional auto-regressive model and a spatial auto-regressive model for the logarithm of the linear part. Other relevant works include that of Lagona et al. (2015b) who introduced a latent model giving the von Mises spatial field, and that proposed by Mastrantonio (2018) forming a distribution available for poly-cylindrical data.

In this paper, we follow an approach for segmenting spatial patterns into a small number of specific local regimes. This is done by employing a Markov random field (MRF), see e.g., Besag (1974), Guyon (1995). In particular, we use the one-parameter Potts model, see e.g., Wu (1982), for the latent classes. The cylindrical data are conditionally independent, given the classes. Thus, the data are represented by a hidden Markov random field (HMRF), which provides a way of classifying typical *states* represented by the cylindrical densities. Related hidden Markov models have been studied for cylindrical time-series by e.g., Holzmann et al. (2006), Bulla et al. (2012) and Lagona et al. (2015a). The HMRF model is however more complicated because of the intractable likelihood function of the model. Our contribution relates to that of Lagona and Picone (2016), who used a cylindrical HMRF. They developed a computationally intensive expectation–maximization (EM) algorithm by utilizing a mean-field approximation of the likelihood function. The method is numerically unstable and it was improved by Ranalli et al. (2018), who instead of considering the full likelihood, took a pairwise composite-likelihood approach, resulting in a more stable algorithm. We extend and improve their algorithm by calculating the exact likelihoods for blocks of observations rather than just pairs. This is combined with an EM algorithm to provide both computational and statistical efficiency. Our implementation using R code is available at https://github.com/henrisl/Cylindrical_HMM.

The paper is organized as follows: Section 2 presents theory behind MRFs and introduces the Potts model that is used. Two cylindrical probability distributions are also introduced and discussed, before the combined cylindrical HMRF model is defined. Section 3 presents new techniques for doing inference on this HMRF model, including parameter estimation and model selection. Section 4 presents results from a simulation study to compare inference methods. Section 5 presents results and discussion from fitting the HMRF model to a set of ocean current observations from the Norwegian Sea. Section 6 concludes and brings up topics for further work.

2. Models

The cylindrical HMRF used here is obtained by combining a parametric MRF (the Potts model) with a cylindrical density (two choices considered). In this section, we first present the Potts model and the two cylindrical densities, before describing the combined HMRF model.

2.1. The potts model

The Potts model is a special type of MRF. It is defined on a discrete lattice $\mathcal{L}$ consisting of $|\mathcal{L}| = n$ sites, which are indexed by $i = 1, \dots, n$. Here, we consider a rectangular $n_1 \times n_2$ grid and $n = n_1 n_2$. The random outcome (class) l_i at site i takes discrete values, i.e., $l_i \in \mathbb{L} = \{1, \dots, K\}$, where K is the number of possible classes. The probability mass function (pmf) gives the probability that the spatial variable $\mathbf{l} = (l_1, \dots, l_n)$ equals some value. We denote the pmf by $\Pr(\mathbf{l} = \mathbf{l}') = p(\mathbf{l})$. To define the Potts model, we follow the pioneering MRF work of Besag (1974), connecting a formulation of the distribution via a neighborhood system and that of a joint pmf formed by clique potentials. A set $\mathbf{N}_L : \{\mathbf{N}_1, \dots, \mathbf{N}_n\}$ is a neighborhood system for the lattice $\mathcal{L}$ if $\mathbf{N}_i \subseteq \mathcal{L} \setminus \{i\}$ for all $i \in \mathcal{L}$, and $i \in \mathbf{N}_j \iff j \in \mathbf{N}_i$ for all pairs $i, j \in \mathcal{L}$. A set $\mathbf{c} \subseteq \mathcal{L}$ is called a clique if $i \in \mathbf{N}_j$ for all pairs $i, j \in \mathbf{c}$. Then $\mathbf{c}_L : \{\mathbf{c}_1, \dots, \mathbf{c}_{n_c}\}$ is the set of all maximal cliques on the lattice, i.e., all cliques that cannot be extended by including another point, and n_c is the number of maximal cliques. With a first order neighborhood, the Potts model has clique systems defined by all two closest neighbors. In the special case of $K = 2$, i.e., only two classes, this Potts model is identical to the well-known Ising model.

The so-called Gibbs formulation of the Potts model defines the joint pmf by

$$\mathbf{l} \sim p(\mathbf{l}) = C(\rho)^{-1} \exp\Big(\rho \sum_{\mathbf{c} \in \mathbf{c}_L} I(l_i = l_j)\Big), \quad C(\rho) = \sum_{\mathbf{l}' \in \mathbb{L}^n} \exp\Big(\rho \sum_{\mathbf{c} \in \mathbf{c}_L} I(l_i' = l_j')\Big). \tag{1}$$

Here, $I(\cdot)$ is the indicator function that takes the value 1 if the argument inside is true, and 0 if it is false. The model has only one parameter ρ, hence it assumes symmetry in all directions and indifference to the numbering of classes. The normalizing constant $C(\rho)$ is a sum over all K outcomes for each entry in the vector $\mathbf{l}$. This sum involves K^n terms, and it is not feasible to compute for large spatial lattices.

The first order Markov formulation of the Potts model is

$$[l_i | l_j; j \in \mathbf{N}_i] \sim p(l_i | l_j; j \in \mathbf{N}_i) = \frac{\exp\big(\rho \sum_{j \in \mathbf{N}_i} I(l_i = l_j)\big)}{\sum_{l_i' \in \mathbb{L}} \exp\big(\rho \sum_{j \in \mathbf{N}_i} I(l_i' = l_j)\big)}, \tag{2}$$

for all $i \in \{1, \dots, n\}$. The neighborhood system includes only the four closest nodes, with less on the boundaries. The normalizing constant in the denominator is clearly tractable, as the sum is over only K components.

Note that the formulation has no prior preference for any colors, and in particular the marginal probability $p(l_i)$ is uniform for all colors. An extension of this model is a Potts model with (location-wise) external field, see e.g., Ameijeiras-Alonso et al. (2019). Even though most computations on the MRF still work similarly for this external field model, the interpretation is more difficult as the marginal probabilities are no longer uniform. In fact, they can only be extracted by summing over all other variables of the grid.

The Potts model shows very different behavior with changing values of the spatial interaction parameter ρ. For values of ρ above $\rho_{\text{crit}} \approx \log(1 + \sqrt{K})$, there is a phase transition and almost all values of the spatial variable $\mathbf{l}$ are equal (Barkema and de Boer, 1991). Hence, the pmf has K modes extremely located in the sample space. The spatial coupling decreases for smaller values of ρ. With $\rho = 0$, the model simplifies to one of spatial independence, and all lattice variables are independent multinomial distributed with $1/K$ probability for each class.

2.2. Cylindrical probability distributions

Following Abe and Ley (2017) and Ranalli et al. (2018) cylindrical data are represented on the form $\mathbf{z} = (x, \phi)$, where $x \in [0, \infty)$ is magnitude (linear part) and $\phi \in (-\pi, \pi]$ is angle (circular part). We present and discuss two cylindrical probability density functions (pdfs).

We first consider the Weibull sine-skewed von Mises (WSSVM) pdf proposed by Abe and Ley (2017). This equals

$$p(x, \phi) = \frac{\alpha\beta^{\alpha}}{2\pi \cosh \kappa}\big(1 + \lambda \sin(\phi - \mu)\big)x^{\alpha-1} \exp\Big(-(\beta x)^{\alpha}\big(1 - \tanh(\kappa)\cos(\phi - \mu)\big)\Big), \quad \text{WSSVM model.} \tag{3}$$

Here, $\alpha > 0$ and $\beta > 0$ are shape and scale parameters for the linear magnitude, $\mu \in (-\pi, \pi]$ is the circular location and $\lambda \in [-1, 1]$ is the circular skewness. Finally, the parameter $\kappa \geq 0$ represents the circular concentration and dependence between the circular and linear parts. For $\kappa = 0$, the circular and linear parts are independent.

A prominent property of this distribution is that the normalizing constant is numerically tractable, which is not always the case for cylindrical densities (Kato and Shimizu, 2008). Furthermore, the marginal and conditional distributions exist in closed forms, which for instance facilitates direct Monte Carlo sampling routines. To generate a random sample from the pdf, one first draws ϕ from its marginal distribution, which is a sine-skewed wrapped Cauchy distribution with location μ and concentration $\tanh(\kappa/2)$, and then draws from the conditional of x, given ϕ, which is a Weibull distribution with shape α and scale $\beta(1 - \tanh(\kappa)\cos(\phi - \mu))^{1/\alpha}$, see Abe and Ley (2017).

In Fig. 1, we display the WSSVM pdf for some values of (κ, λ) and with $(\alpha, \beta, \mu) = (2, 1, 0)$. Bear in mind that the variables are actually on a cylinder, so the angle ϕ on the second axis is wrapped around a circle, and x represents the cylinder height. Notice the high flexibility in both the circular and linear parts of the pdfs in the different displays. The skewness grows when increasing λ, and it is skewed towards negative values of ϕ when λ is negative. The circular part is uniform for $\kappa = \lambda = 0$.

We next consider another type of cylindrical distribution with a generalized Pareto-type model for the linear part and a wrapped Cauchy for the circular part (GPTWC), as proposed by Imoto et al. (2019). Its pdf is

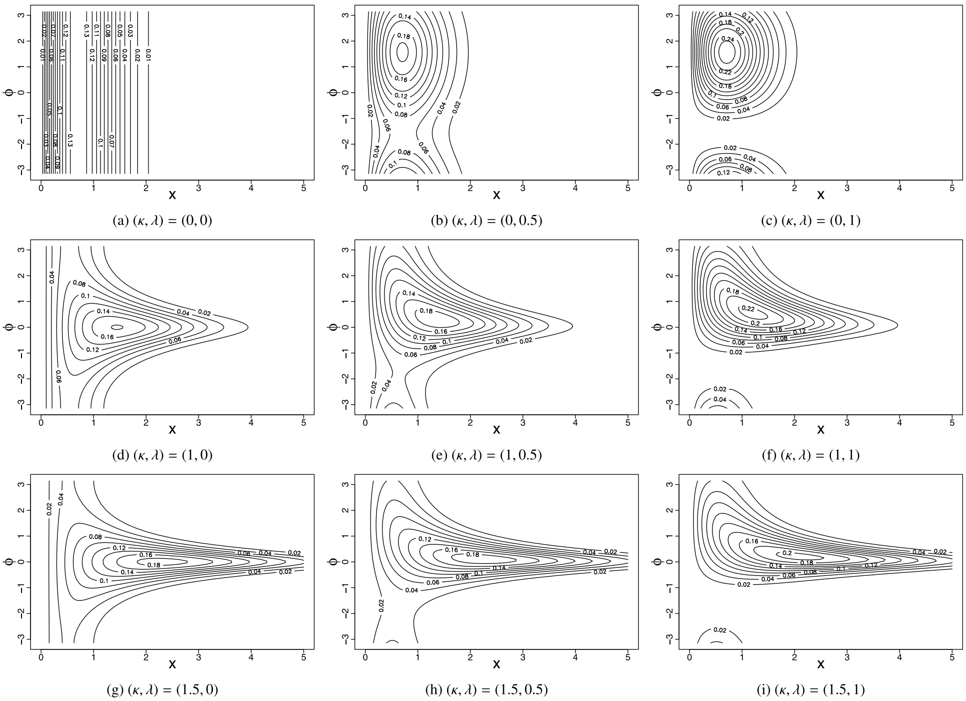
$$p(x, \phi) = \frac{\sqrt{1-\kappa^2}}{2\pi\beta\alpha}\left(\frac{x}{\beta}\right)^{1/\alpha-1}\left(1 + \frac{\tau}{\alpha}\left(\frac{x}{\beta}\right)^{1/\alpha}\big(1 - \kappa\cos(\phi - \mu)\big)\right)^{-(\alpha/\tau+1)}, \quad \text{GPTWC model.} \tag{4}$$

Again, $\mu \in (-\pi, \pi]$ is the circular location parameter, $\alpha > 0$ and $\beta > 0$ are linear shape and scale parameters and $\kappa \in [0, 1]$ acts as circular concentration and circular–linear dependence. Finally, $\tau > 0$ determines the tail behavior of the linear part. When $\tau \to 0$, the pdf in Eq. (4) reduces to

$$p(x, \phi) = \frac{\sqrt{1-\kappa^2}}{2\pi\beta\alpha}\left(\frac{x}{\beta}\right)^{1/\alpha-1} \exp\left(-\left(\frac{x}{\beta}\right)^{1/\alpha}\big(1 - \kappa\cos(\phi - \mu)\big)\right), \tag{5}$$

which is actually the same density as the WSSVM density in Eq. (3) with $\lambda = 0$ and a reparameterization of α, β and κ. Crucially, the normalizing constant of the GPTWC pdf is explicitly defined, as well as both marginal and conditional distributions. The conditional pdf of x, given ϕ, corresponds to a generalized Pareto-type, and it is straightforward to generate a sample from the GPTWC distribution, as suggested by Imoto et al. (2019).

In Fig. 2, we plot the GPTWC pdf for some values of (α, τ) and with $(\beta, \mu, \kappa) = (1, 0, 0.75)$. The tails become heavier as τ increases. By also varying β and κ, this pdf is flexible in both the



(a) $(\kappa, \lambda) = (0, 0)$ (b) $(\kappa, \lambda) = (0, 0.5)$ (c) $(\kappa, \lambda) = (0, 1)$

(d) $(\kappa, \lambda) = (1, 0)$ (e) $(\kappa, \lambda) = (1, 0.5)$ (f) $(\kappa, \lambda) = (1, 1)$

(g) $(\kappa, \lambda) = (1.5, 0)$ (h) $(\kappa, \lambda) = (1.5, 0.5)$ (i) $(\kappa, \lambda) = (1.5, 1)$

Fig. 1. Contour plots of the WSSVM density given in Eq. (3) for $(\alpha, \beta, \mu) = (2, 1, 0)$ and (κ, λ) as indicated.

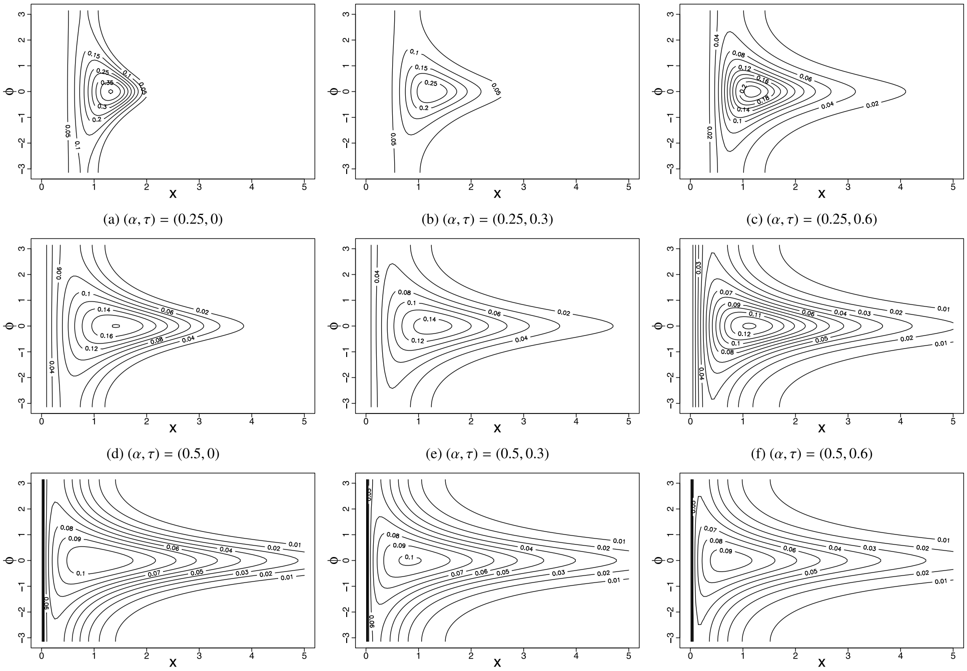
circular and linear part, and this enables one to fit both distributions that are concentrated in the circular part and distributions that are more evenly distributed across the circular variable. Unlike the WSSVM distribution, this pdf is not able to model data with skewness in the circular part. Rather, it is tailored to fit data with heavy tails in the linear part.

2.3. Cylindrical hidden Markov random field

We now combine the Potts MRF in Eq. (1) and cylindrical distributions of Eqs. (3) and (4) in a joint model. The classes of the MRF, l_i, $i \in \{1, \dots, n\}$, are assumed hidden, or latent, and a class determines the parameters for the cylindrical pdf. Hence, this can be seen as a HMRF with mixed cylindrical distributed observations. The Potts pmf is denoted $p_\rho(\mathbf{l})$, where the subscript highlights the influence of the interaction parameter ρ, which is assumed fixed but unknown. We further define $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$, $\mathbf{z}_i = (x_i, \phi_i)$ as the vector of bivariate observations. Each observation $\mathbf{z}_i$ is assumed to be conditionally independent, given the latent spatial class l_i. The cylindrical model parameters, denoted $\boldsymbol{\theta}_k$, for each class $k \in \{1, \dots, K\}$, are assumed fixed but unknown. Hence, if grid point i takes latent class $l_i = k$, the observation $\mathbf{z}_i$ will have a cylindrical pdf denoted by $p_{\boldsymbol{\theta}_k}(\mathbf{z}_i)$, which is either from Eq. (3) or (4). The subscript notation for the parameters $\boldsymbol{\theta}_k$ indicates their influence and that they are fixed but unknown. Because the observations are assumed conditionally independent, the joint pdf of all cylindrical data, given the lattice variables, is

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{l}) = \prod_{i=1}^{n} \prod_{k=1}^{K} p_{\boldsymbol{\theta}_k}(\mathbf{z}_i)^{I(l_i=k)}, \tag{6}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$ is the collection of all cylindrical model parameters.



(a) $(\alpha, \tau) = (0.25, 0)$

(b) $(\alpha, \tau) = (0.25, 0.3)$

(c) $(\alpha, \tau) = (0.25, 0.6)$

(d) $(\alpha, \tau) = (0.5, 0)$

(e) $(\alpha, \tau) = (0.5, 0.3)$

(f) $(\alpha, \tau) = (0.5, 0.6)$

(g) $(\alpha, \tau) = (0.75, 0)$

(h) $(\alpha, \tau) = (0.75, 0.3)$

(i) $(\alpha, \tau) = (0.75, 0.6)$

Fig. 2. Contour plots of the GPTWC density given in Eq. (4) for $(\beta, \mu, \kappa) = (1, 0, 0.75)$ and (α, τ) as indicated.

The cylindrical HMRF model has joint model for the observed data and latent classes given by

$$p_{\boldsymbol{\theta},\rho}(\mathbf{z}, \mathbf{l}) = p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{l})p_{\rho}(\mathbf{l}). \tag{7}$$

We get the marginal likelihood of the model parameters by summing over all the hidden spatial class variables:

$$L(\boldsymbol{\theta}, \rho|\mathbf{z}) = p_{\boldsymbol{\theta},\rho}(\mathbf{z}) = \sum_{\mathbf{l}' \in \mathbb{L}^n} p_{\boldsymbol{\theta},\rho}(\mathbf{z}, \mathbf{l}'). \tag{8}$$

In our presentation and examples, we use a regular grid for the latent field and available data. However, because the Markovian model is defined on a nearest-neighbor graph (Tjelmeland and Austad, 2012), the extension to non-regular structures works similarly, with some more bookkeeping in the implementation. One can also choose to keep a regular grid for the latent field, and simply have missing data at some of the nodes.

The model is visualized in Fig. 3. Here, the arrows indicate the conditional relations between variables. At one location, the model for cylindrical observations is defined via the class of the latent variable at that same location. For the MRF, the neighborhood relation does not have a natural direction, but in a sequential model formulation going from grid cell 1 to n, the dependence for l_i is in a length-m buffer as indicated by the gray cells in Fig. 3, where m is the smallest of the two grid dimensions; $\min(n_1, n_2)$.

3. Inference

We now consider ways of estimating unknown model parameters and predicting the latent variables in this HMRF. In Section 3.1, we outline an algorithm for exact likelihood computation for small grid sizes. Section 3.2 provides composite likelihood approaches for approximating the

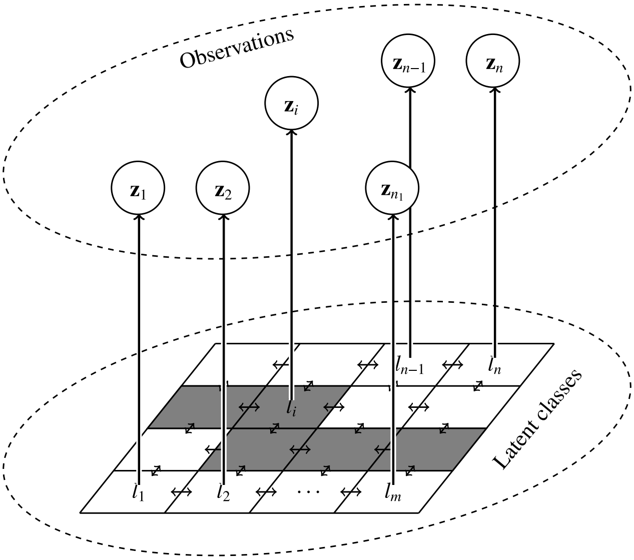


Fig. 3. Visualization of the HMRF model. Arrows indicate conditional dependence. Latent classes are spatially dependent, whereas observations are conditionally independent, given the latent classes. The gray cells indicate the buffer influencing cell i in a sequential 1 to n ordering.

likelihood for larger grids, which is coupled with an EM algorithm in Section 3.3. We discuss the asymptotic properties of these estimators and criteria for model selection in Section 3.4.

3.1. Exact likelihood

The likelihood can be defined sequentially by

$$L(\boldsymbol{\theta}, \rho|\mathbf{z}) = p_{\boldsymbol{\theta},\rho}(\mathbf{z}) = p_{\boldsymbol{\theta},\rho}(\mathbf{z}_1) \prod_{i=2}^{n} p_{\boldsymbol{\theta},\rho}(\mathbf{z}_i|\mathbf{z}_{1:(i-1)}), \tag{9}$$

where $\mathbf{z}_{1:(i-1)} = (\mathbf{z}_1, \dots, \mathbf{z}_{i-1})$. To evaluate this likelihood, we need to compute the sequential pdfs $p_{\boldsymbol{\theta},\rho}(\mathbf{z}_i|\mathbf{z}_{1:(i-1)})$, which involves sums over the latent variables l_i, for different i. Due to the computational complexity of the normalizing constant of the Potts model in Eq. (1), this is not feasible for large spatial grids. Yet, through a restriction on either the number of rows or columns, $m = \min(n_1, n_2)$, and a clever indexation on the grid cells, one can compute the normalizing constant exactly for relatively small m. This sequential calculation extends the classical forward–backward algorithm from hidden Markov chains (MacDonald and Zucchini, 1997) into a spatial setting (Bartolucci and Besag, 2002; Reeves and Pettitt, 2004). The forward–backward algorithm was first introduced by Baum et al. (1970) and it computes i) the likelihood of the data in a forward pass, ii) the posterior marginal probabilities of all latent variables, given the observed data, in an additional backward pass.

In the forward–backward algorithm, all ensuing probability distributions for the latent variables and the observations are conditioned on the fixed model parameters $(\boldsymbol{\theta}, \rho)$, but we leave this notation out to make the derivations easier to follow. We initialize the algorithm by finding the marginal probabilities of the first latent variable, given only the first data observation

$$p(l_1|\mathbf{z}_1) = \frac{p(\mathbf{z}_1|l_1)p(l_1)}{p(\mathbf{z}_1)} = C_1 p(\mathbf{z}_1|l_1)p(l_1), \quad 1/C_1 = p(\mathbf{z}_1) = \sum_{l_1'=1}^{K} p(\mathbf{z}_1|l_1')p(l_1'), \tag{10}$$

with $p(\mathbf{z}_1|l_1)$ being a cylindrical density from either Eq. (3) or (4), and $p(l_1)$ being the uniform pmf over the possible outcomes $l_1 \in \mathbb{L}$. The normalizing constant defines the first term in the sequential likelihood in Eq. (9).

The algorithm then moves forward through the spatial grid. Neighbors of grid point i are the points $\{i-m, i-1, i+1, i+m\}$. Hence, at each step i in the forward recursion, the algorithm needs to keep track of the probability of every outcome of the vector $\mathbf{l}_{(i-m):(i-1)} = (l_{i-m}, \dots, l_{i-1})$. By using the conditional independence assumption of the data, $p(l_i'|l_{1:i-1}', \mathbf{z}_{1:i-1}) = p(l_i'|l_{i-1}', l_{i-m}')$ and $p(\mathbf{z}_i|\boldsymbol{l}_{1:i}, \mathbf{z}_{1:i-1}) = p(\mathbf{z}_i|l_i)$, the forward recursion at step i involves

$$\begin{aligned}
1/C_i &= p(\mathbf{z}_i|\mathbf{z}_{1:(i-1)}) = \sum_{l_i'=1}^{K} \cdots \sum_{l_{i-m}'=1}^{K} p(\mathbf{z}_i|l_i') p(l_i'|l_{i-1}', l_{i-m}') p(\mathbf{l}_{(i-m):(i-1)}'|\mathbf{z}_{1:(i-1)}), \qquad (11)\\
p(\mathbf{l}_{(i-m):i}|\mathbf{z}_{1:i}) &= C_i p(\mathbf{z}_i|l_i) p(l_i|l_{i-1}, l_{i-m}) p(\mathbf{l}_{(i-m):(i-1)}|\mathbf{z}_{1:(i-1)}),\\
p(\mathbf{l}_{(i-m+1):i}|\mathbf{z}_{1:i}) &= \sum_{l_{i-m}'=1}^{K} p(\mathbf{l}_{(i-m+1):i}, l_{i-m}'|\mathbf{z}_{1:i}).
\end{aligned}$$

Here, we need to compute the forward probabilities $p(\mathbf{l}_{(i-m):i}|\mathbf{z}_{1:i})$ for all possible outcomes $\mathbf{l}_{(i-m):i} \in \mathbb{L}^{m+1}$, see gray cells in Fig. 3. This means that we need to compute and store K^{m+1} probabilities. Overall, this recursion for computing the exact likelihood has complexity $\mathcal{O}\big((n-m+1)K^{m+1}\big)$, as investigated by Reeves and Pettitt (2004). Clearly, the algorithm is unfeasible when the number of rows m is large. With only two latent classes, $K = 2$, Friel and Rue (2007) report that computation of the likelihood is feasible for grid sizes up to $m = 19$, while $m = 12$ for $K = 3$ and $m = 9$ for $K = 4$. The feasible grid sizes have increased slightly since then, but not substantially. In addition, we are not interested in merely computing the likelihood once, but seek to maximize it to estimate the model parameters. As a consequence, the feasible grid sizes for our application are small.

By stepping recursively backwards, one can compute the marginal probabilities $p(l_i|\mathbf{z})$ of each latent variable, given the complete set of observations. The recursion starts with $p(l_n|\mathbf{z}_{1:n}) = \sum_{l_{n-1}'=1}^{K} \cdots \sum_{l_{n-m}'=1}^{K} p(\mathbf{l}_{(n-m):n}'|\mathbf{z}_{1:n})$. The probabilities are used to make marginal predictions of the latent classes. For $i = n, \dots, 2$,

$$p(l_{i-1}, l_i|\mathbf{z}_{1:n}) = \frac{p(l_{i-1}, l_i|\mathbf{z}_{1:i})}{p(l_i|\mathbf{z}_{1:i})} p(l_i|\mathbf{z}_{1:n}), \quad p(l_{i-1}|\mathbf{z}_{1:n}) = \sum_{l_i'=1}^{K} p(l_{i-1}, l_i'|\mathbf{z}_{1:n}), \qquad (12)$$

where the forward calculations in Eq. (11) are re-used. Conveniently, we can also use these backwards probabilities to draw conditional samples of the latent classes, given $\mathbf{z} = \mathbf{z}_{1:n}$.

3.2. Composite likelihood

As a way to mitigate the unfeasible computational complexity involved with large spatial grids, we propose to substitute the likelihood in Eq. (8) with a composite likelihood. The composite likelihood is formed by multiplying component likelihoods, which correspond to marginal or conditional likelihoods of small subsets of data (Lindsay, 1988; Cox and Reid, 2004; Varin and Vidoni, 2005). The composite likelihood approach may be less statistically efficient than the full likelihood, but it is substantially cheaper to calculate, and hence poses a useful trade-off between efficiency and computational demand.

The simplest case of a composite likelihood is that of only considering pairwise interactions. For each clique $\mathbf{c} = (c_1, c_2) \in \mathbf{c}_L$ we define the clique likelihood

$$L_{\mathbf{c}}(\boldsymbol{\theta}, \rho|\mathbf{z}_{\mathbf{c}}) = \sum_{\mathbf{l}_{\mathbf{c}} \in \mathbb{L}^2} p_{\boldsymbol{\theta},\rho}(\mathbf{z}_{\mathbf{c}}, \mathbf{l}_{\mathbf{c}}), \qquad (13)$$

where $\mathbf{z_c}$ is the observed data on the grid points corresponding to clique $\mathbf{c}$, and the sum is over all K^2 possible outcomes for the clique latent classes $\mathbf{l_c}$. Their joint distribution is defined via

$$p_{\boldsymbol{\theta},\rho}(\mathbf{z_c}, \mathbf{l_c}) = p_\rho(\mathbf{l_c}) \prod_{i\in\mathbf{c}} \prod_{k=1}^{K} p_{\theta_k}(\mathbf{z}_i)^{I(l_i=k)}, \tag{14}$$

$$p_\rho(\mathbf{l_c}) = C(\rho)^{-1} \exp\big(\rho I(l_{c_1} = l_{c_2})\big), \quad C(\rho) = \sum_{\mathbf{l}'_\mathbf{c}\in\mathbb{L}^2} \exp\Big(\rho I(l'_{c_1} = l'_{c_2})\Big).$$

The composite log-likelihood function is given as the sum of the log-likelihood of each clique,

$$cl_p(\boldsymbol{\theta}, \rho|\mathbf{z}, \mathbf{c}_L) = \sum_{\mathbf{c}\in\mathbf{c}_L} \log\big(L_\mathbf{c}(\boldsymbol{\theta}, \rho|\mathbf{z_c})\big). \tag{15}$$

The pairwise likelihood is feasible to calculate as it involves only few terms in the required sums. It was used by Ranalli et al. (2018) for cylindrical data. However, as will be shown later, the direct optimization approach suffers from a small radius of convergence, i.e., the starting values for the optimization of the log-likelihood need to be close to the true solution in order to converge satisfactorily. To mitigate this problem, we connect this pairwise likelihood with the EM algorithm which is presented in Section 3.3.

Recall that exact likelihood is feasible when either the number of rows or columns is small. This method can hence be used to approximate the likelihood for subsets of larger grids. In this way, we can form a composite likelihood for entire rows or columns, or sets of a few rows or columns. Because the evaluation of such blocks goes beyond the pairwise calculations, it is expected to give a more statistically efficient approximation, closer to the exact likelihood. To avoid storage problems, the blocks must have either a small number of rows or a small number of columns, so the forward–backward algorithm runs effectively for each block. Similar ideas of splitting a spatial grid into smaller blocks for effective computations have been studied by e.g., Caragea and Smith (2007), Allard et al. (2011), Eidsvik et al. (2014) in various contexts.

In the same way as for the pairwise likelihood, the block log-likelihood is computed by summing the block components. Each block $\mathbf{b}$ has observations denoted by $\mathbf{z_b}$ and exact likelihood given by Eq. (9). The resulting block log-likelihood is on the form

$$cl_b(\boldsymbol{\theta}, \rho|\mathbf{z}, \mathbf{b}_L) = \sum_{\mathbf{b}\in\mathbf{b}_L} l(\boldsymbol{\theta}, \rho|\mathbf{z_b}), \tag{16}$$

where $\mathbf{b}_L$ represents the collection of all blocks. The maximum composite likelihood estimates optimize this expression for $\boldsymbol{\theta}$ and ρ.

3.3. EM algorithm

To define the iterations of the EM algorithm, we start by the complete-data composite likelihood, which is the likelihood of both the observations and the latent classes. This is defined by Eqs. (13) and (14). The distribution of the clique latent classes can be written $p_\rho(\mathbf{l_c}) = \prod_{\mathbf{l}'_\mathbf{c}\in\mathbb{L}^2} p_\rho(\mathbf{l_c})^{I(\mathbf{l}'_\mathbf{c}=\mathbf{l_c})}$, and the complete-data composite log-likelihood is then defined by

$$cl_{cd}(\boldsymbol{\theta}, \rho|\mathbf{z}, \mathbf{c}_L) = \sum_{\mathbf{c}\in\mathbf{c}_L} \big(cl^{\mathbf{c}}_{cd}(\boldsymbol{\theta}) + cl^{\mathbf{c}}_{cd}(\rho)\big), \tag{17}$$

$$cl^{\mathbf{c}}_{cd}(\boldsymbol{\theta}) = \sum_{i\in\mathbf{c}} \sum_{k=1}^{K} \log\big(p_{\theta_k}(\mathbf{z}_i)\big) I(l_i = k), \qquad cl^{\mathbf{c}}_{cd}(\rho) = \sum_{\mathbf{l}'_\mathbf{c}\in\mathbb{L}^2} \log\big(p_\rho(\mathbf{l_c})\big) I(\mathbf{l}'_\mathbf{c} = \mathbf{l_c}).$$

The E-step in the algorithm reduces to predicting the latent spatial classes, or rather the probability of each possible outcome for each clique. At iteration s of the EM algorithm, we have given parameter estimates $\hat{\boldsymbol{\theta}}_s$ and $\hat{\rho}_s$, and we compute the probability $\hat{\mathbf{l}}_\mathbf{c}$ for each latent configuration

$\mathbf{l}_{\mathbf{c}} \in \mathbb{L}^2$, for each clique $\mathbf{c} \in \mathbf{c}_L$ as

$$\hat{\mathbf{l}}_{\mathbf{c}} = p_{\hat{\boldsymbol{\theta}}_s,\hat{\rho}_s}(\mathbf{l}_{\mathbf{c}}|\mathbf{z}_{\mathbf{c}}) = \frac{p_{\hat{\rho}_s}(\mathbf{l}_{\mathbf{c}})p_{\hat{\boldsymbol{\theta}}_s}(\mathbf{z}_{\mathbf{c}})}{\sum_{\mathbf{l}'_{\mathbf{c}}\in\mathbb{L}^2} p_{\hat{\rho}_s}(\mathbf{l}'_{\mathbf{c}})p_{\hat{\boldsymbol{\theta}}_s}(\mathbf{z}_{\mathbf{c}})}, \quad p_{\hat{\boldsymbol{\theta}}_s}(\mathbf{z}_{\mathbf{c}}) = \prod_{i\in\mathbf{c}}\prod_{k=1}^{K} p_{\hat{\boldsymbol{\theta}}_{s,k}}(\mathbf{z}_i)^{I(l_i=k)}. \tag{18}$$

By marginalizing, we can obtain probabilities $\hat{l}_{ik} = p_{\hat{\boldsymbol{\theta}}_s,\hat{\rho}_s}(l_i = k|\mathbf{z}_{\mathbf{c}})$, $i \in \mathbf{c}$ for each grid point in the clique belonging to each latent class. The EM algorithm is in this way applied to the pairwise composite likelihood method, because the joint for blocks is typically too computer demanding.

The M-step consists of maximizing the expected value of the complete-data composite log-likelihood in Eq. (17). Conveniently, this is the sum of two components that are independent with respect to the model parameters. Hence, they can be maximized separately over their respective parameters, and we define two functions to be maximized

$$g(\boldsymbol{\theta}|\mathbf{z},\mathbf{c}_L) = E(cl_{cd}^{\mathbf{c}}(\boldsymbol{\theta})) = \sum_{\mathbf{c}\in\mathbf{c}_L}\sum_{i\in\mathbf{c}}\sum_{k=1}^{K}\hat{l}_{ik}\log\big(p_{\boldsymbol{\theta}_k}(\mathbf{z}_i)\big), \tag{19}$$

$$h(\rho|\mathbf{z},\mathbf{c}_L) = E(cl_{cd}^{\mathbf{c}}(\rho)) = \sum_{\mathbf{c}\in\mathbf{c}_L}\sum_{\mathbf{l}_{\mathbf{c}}\in\mathbb{L}^2}\hat{\mathbf{l}}_{\mathbf{c}}\log\big(p_{\rho}(\mathbf{l}_{\mathbf{c}})\big). \tag{20}$$

The parameter estimates are then obtained by $\hat{\boldsymbol{\theta}}_{s+1} = \arg\max_{\boldsymbol{\theta}}\{g(\boldsymbol{\theta}|\mathbf{z},\mathbf{c}_L)\}$ and $\hat{\rho}_{s+1} = \arg\max_{\rho}$ $\{h(\rho|\mathbf{z},\mathbf{c}_L)\}$.

To maximize Eq. (19) we use a handy reparameterization:

$$\boldsymbol{\theta}_k = \big(\theta_{1k},\ldots,\theta_{5k}\big) = \big(\log(\alpha_k), \log(\beta_k), \tan(\mu_k/2), \log(\kappa_k), \tanh^{-1}(\lambda_k)\big), \quad \text{WSSVM model},$$
$$\boldsymbol{\theta}_k = \big(\theta_{1k},\ldots,\theta_{5k}\big) = \big(\log(\alpha_k), \log(\beta_k), \tan(\mu_k/2), \log(\tau_k), \log(\kappa_k/(1-\kappa_k))\big), \quad \text{GPTWC model.}$$

This parameterization is identical to that used in Ranalli et al. (2018), and it enables optimization without any constraints because $\boldsymbol{\theta}_k \in \mathbb{R}^5$. The optimization can then be carried out by quasi-Newton methods, and we have used the `BFGS` method from the function `optim` in R. An alternative optimization approach with constraints is used by Abe and Ley (2017). Note that there could be potential challenges with the angle μ_k in either parameterization as a circular variable is transformed to a linear variable. It can hit $-\pi$ or π from one side in the optimization, but we did not experience problems with this here.

For the parameter ρ we need to impose a constraint based on the phase transition at the critical value ρ_{crit}, so that $\rho \in (0, \rho_{\text{crit}})$. We carry out the optimization through the method `L-BFGS-B` in `optim`, which allows for box-like constraints.

3.4. Asymptotic theory and derived quantities

For composite likelihoods, each component is multiplied, even though they are not independent. This means that composite likelihoods can be seen as likelihoods from a misspecified model that assumes independence between the components. As such, the assumptions of regular maximum likelihood estimators do not necessarily hold (Varin et al., 2011). Denoting the score of the composite likelihood by $\mathbf{s}_{cl}(\boldsymbol{\theta}|\mathbf{z}) = \nabla_{\boldsymbol{\theta}} cl(\boldsymbol{\theta}|\mathbf{z})$, the sensitivity and variability matrices

$$H(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\Big(-\nabla_{\boldsymbol{\theta}}\mathbf{s}_{cl}(\boldsymbol{\theta}|\mathbf{z})\Big), \quad J(\boldsymbol{\theta}) = \text{Var}_{\boldsymbol{\theta}}\Big(\mathbf{s}_{cl}(\boldsymbol{\theta}|\mathbf{z})\Big), \tag{21}$$

are not equal as in the full likelihood case. The Fisher information matrix is substituted by the Godambe (sandwich) information matrix (Godambe, 1960), given by

$$G(\boldsymbol{\theta}) = H(\boldsymbol{\theta})J(\boldsymbol{\theta})^{-1}H(\boldsymbol{\theta}). \tag{22}$$

If the composite likelihood was a true likelihood and not a misspecification, we would have $G(\boldsymbol{\theta}) = H(\boldsymbol{\theta}) = J(\boldsymbol{\theta})$. Under mild regularity conditions (Varin et al., 2011), the maximum composite likelihood estimator is asymptotically normally distributed,

$$(\hat{\boldsymbol{\theta}}_{cl} - \boldsymbol{\theta}) \xrightarrow{d} N(0, G(\boldsymbol{\theta})^{-1}). \tag{23}$$

Table 1
True model parameters for the two cases in consideration with the WSSVM density.

Case 1: Low separation			Case 2: High separation		
$\alpha_1 = 2$	$\alpha_2 = 2$	$\alpha_3 = 2$	$\alpha_1 = 3$	$\alpha_2 = 5$	$\alpha_3 = 1$
$\beta_1 = 1$	$\beta_2 = 1$	$\beta_3 = 0.6$	$\beta_1 = 1$	$\beta_2 = 5$	$\beta_3 = 0.8$
$\mu_1 = 0$	$\mu_2 = 0$	$\mu_3 = 0$	$\mu_1 = 0$	$\mu_2 = 0$	$\mu_3 = 0$
$\kappa_1 = 0$	$\kappa_2 = 0$	$\kappa_3 = 1.5$	$\kappa_1 = 0.21$	$\kappa_2 = 0.21$	$\kappa_3 = 1.7$
$\lambda_1 = 1$	$\lambda_2 = -1$	$\lambda_3 = 0$	$\lambda_1 = 0.8$	$\lambda_2 = 0$	$\lambda_3 = -0.8$

Unfortunately, the naive estimator of the variability matrix $J(\boldsymbol{\theta})$ tends to become singular when evaluated at the maximum composite likelihood estimate (Varin and Vidoni, 2005), and the Godambe matrix gets numerically unstable. Instead, a parametric bootstrap procedure is used for uncertainty quantification here. Random samples are then drawn from the HMRF model using the parameters specified by maximum composite likelihood. For each of the random samples we again use maximum composite likelihood to estimate the parameters for the bootstrap sampled data. These new estimates are used to assess the variability in the estimators. Note that such bootstrap sample approaches (Givens and Hoeting, 2013) could be affected by label switching issues, occurring because the likelihood is invariant to permutations of the classes. A distance-based re-permutation could potentially be used, but in our case the optimization for bootstrap estimates starts at the estimated parameters and the right mode of the likelihood surface was detected.

One key assumption in the model we have developed is that the number of latent classes K is known. In practice, we need model selection criteria to determine K. We use the composite BIC (Gao and Song, 2010) for model comparison:

$$\text{C-BIC} = -2cl(\boldsymbol{\theta}|\mathbf{z}) + \log(n)d_s^*. \tag{24}$$

Here, d_s^* is termed the effective degrees of freedom and given by $d_s^* = \text{tr}(\hat{J}(\boldsymbol{\theta})\hat{H}(\boldsymbol{\theta})^{-1})$. Note that $\hat{J}(\boldsymbol{\theta})$ and $\hat{H}(\boldsymbol{\theta})$ are estimators for $J(\boldsymbol{\theta})$ and $H(\boldsymbol{\theta})$, respectively, see e.g., Varin and Vidoni (2005). We use the observed score and the observed information matrix to calculate the effective degrees of freedom, i.e., we let $\hat{J}(\boldsymbol{\theta}) = \mathbf{s}_{cl}(\hat{\boldsymbol{\theta}}_{cl}|\mathbf{z})\mathbf{s}_{cl}(\hat{\boldsymbol{\theta}}_{cl}|\mathbf{z})^T$ and $\hat{H}(\boldsymbol{\theta}) = -\nabla^2_{\boldsymbol{\theta}} cl(\hat{\boldsymbol{\theta}}_{cl}|\mathbf{z})$. The score and Hessian of the composite log-likelihood can be computed by e.g., the R package `numDeriv`.

4. Simulation study

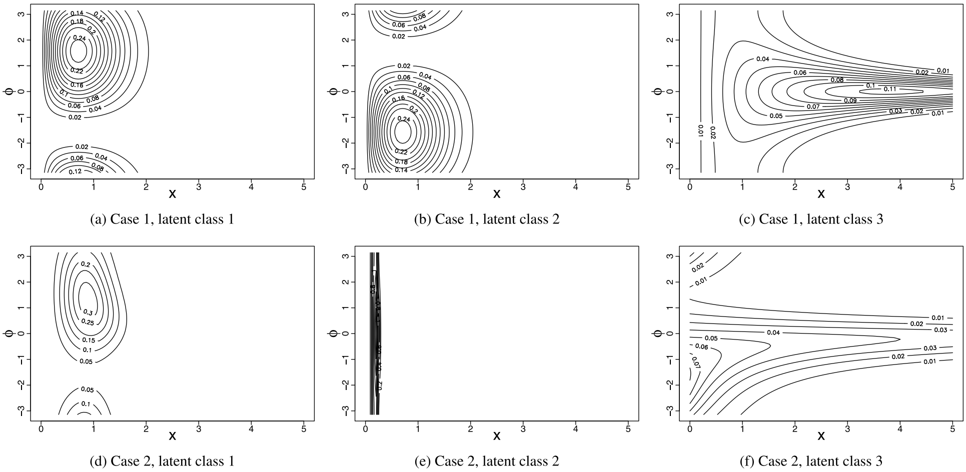
Simulation studies are carried out to compare the performance of various methods with regard to convergence radius of the optimization, parameter estimation accuracy and run time. The favored method is finally used to investigate bias and variance in the parameter estimation. This is only reported for the WSSVM distribution, but similar results also apply to the GPTWC distribution.

4.1. Experimental setup

We consider two different cases of parameter sets and two values for the spatial dependence parameter. In both cases, we simulate data from a 24 × 24 grid with $K = 3$ latent classes. The two cases have varying degree of separation between parameters in the latent classes. With low separation, the cylindrical densities corresponding to each latent class are rather similar. Hence, the method needs to be statistically robust in order to find the correct parameters. For the high separation case, the cylindrical densities are further apart, and the inference should be less demanding for this case.

Table 1 lists the true cylindrical model parameters for the two cases of the WSSVM density. Contour plots of the corresponding cylindrical densities are displayed in Fig. 4. Observe the difference in separation between the two cases. In Case 1 with low separation (top), the cylindrical densities overlap to a greater extent than for Case 2 with high separation (bottom).

To simulate data, we exploit the structure of the model. First, we generate a realization of the latent field by the Swendsen–Wang algorithm (Swendsen and Wang, 1987) with the true value



(a) Case 1, latent class 1

(b) Case 1, latent class 2

(c) Case 1, latent class 3

(d) Case 2, latent class 1

(e) Case 2, latent class 2

(f) Case 2, latent class 3

Fig. 4. Contour plots of the true sample distributions for each of the two sets of cylindrical parameters with the WSSVM density.

for the spatial dependence parameter ρ. Then, given this latent field, observations $\mathbf{z}_i = (x_i, \phi_i)$, $i = 1, \dots, n$, are generated by the WSSVM distribution according to the proposed algorithm, with model parameters determined by the latent class. The parameter inference is done from this simulated cylindrical data. We repeat this process for several independent replicate realizations of data.

In Sections 4.2 and 4.3, this setup is used to compare the composite likelihood methods regarding convergence radius and run time. The same setup is also applied to investigate properties of the parameters for the WSSVM density in Section 4.4. For the pairwise likelihood we consider both direct maximization of the composite likelihood function in Eq. (15) and the EM algorithm. Direct maximization can be carried out by for example the BFGS method from the `optim` function in R. For the block likelihood approach, we split the grid into all possible blocks of size $m \times 24$ horizontally and $24 \times m$ vertically. This means that e.g., for $m = 2$ we have in total 23 vertical and 23 horizontal overlapping blocks. For $m = 1$ we get 24 horizontal and 24 vertical blocks of size 1×24. The block log-likelihood is then computed based on this separation and maximized directly by the BFGS method from the `optim` function in R.

4.2. Convergence radius

To compare convergence radius of the optimization procedures, we consider three methods, the EM approach to the pairwise likelihood, the direct pairwise likelihood and the block composite likelihood with single rows or columns $m = 1$. For both cases and both values of the spatial dependence parameter, we draw 50 data samples. Thus, we get 200 data samples in total, 50 for each combination of cylindrical parameter set and spatial dependence parameter. For each data sample, we then use all three methods to estimate the parameters, using the same starting point for all three methods. The starting points of the optimization are chosen randomly some distance away from the true optimum, within a comparable domain for the different cases, so the method needs to be robust in order to find the true optimum. In this way, less robust methods will converge to local optimums that are not the true solution.

We count the number of times each method converges to the true parameter values. To decide if the optimization has converged to a local optimum or the true solution, we compute the RMSE versus the true parameters. If $\boldsymbol{\theta}$ are the true parameters, $\hat{\boldsymbol{\theta}}$ are the estimates and $\boldsymbol{\theta}^0$ are the starting

Table 2
Number of times each method converged to the true solution.

Method	Case 1		Case 2	
	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$
EM pairwise	45	38	40	35
Direct pairwise	20	17	16	15
Block, $m = 1$	17	20	17	17

values, the respective RMSE is given by

$$\mathrm{RMSE}^0 = \sqrt{\frac{1}{q}\sum_{i=1}^{q}(\boldsymbol{\theta}_i^0 - \boldsymbol{\theta}_i)^2}, \quad \widehat{\mathrm{RMSE}} = \sqrt{\frac{1}{q}\sum_{i=1}^{q}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^2}, \tag{25}$$

where q is the length of the parameter vector. Then, we say that the optimization converged to the true solution if the estimated parameters are closer to the true solution than the starting values, or $\widehat{\mathrm{RMSE}} < \mathrm{RMSE}^0$. On the other hand, if $\widehat{\mathrm{RMSE}} > \mathrm{RMSE}^0$ the algorithm has converged to a local optimum that is further away from the true solution than the starting values.

In Table 2 we display the number of times each method converged to the true solution out of the 50 data samples for each case. In total, the EM algorithm converged to the true solution for 158 of the 200 data samples, the direct pairwise converged 68 times and the block likelihood converged 71 out of a possible 200 times. This clearly shows that the EM algorithm has a larger convergence radius than the direct optimization strategies. Moreover, these results indicate that the convergence radius of the two direct methods is comparable. The block likelihood method comprises less approximations and should in theory be more statistically robust. However in practice the effect is limited.

We compare the run time in minutes for each method. As stopping criterion for the optimization, we use a relative increase tolerance of 10^{-5} in pairwise or block log-likelihood. The results show that the EM algorithm is computationally slow. In comparison, direct pairwise optimization reduces the run time of the EM algorithm by about 30%, which is also what Ranalli et al. (2018) reported. The block likelihood with $m = 1$ further reduces the computational time by 25% compared to the direct pairwise likelihood. Hence, considering these computation times as well as the convergence radius in Table 2, we claim that if the starting point is close to the true solution, the block likelihood is favored. For $m = 2$ the computation time is about 5 times higher than for $m = 1$ due to the exponential dependency on m. Further, tests of moderate sample sizes show no indication of significant increase in convergence radius for $m = 2$ compared to $m = 1$. Hence, the added computational complexity of increased m is not really worth the effort here, even though in theory the method is closer to the full likelihood.

As we have seen, both the direct pairwise likelihood and the block likelihood are susceptible to two possible issues, namely only converging to local maximums and sensitivity to initial parameters. Hence, when considering real data in the next section, we follow the short-run strategy of Ranalli et al. (2018) to prevent convergence to local maxima. The strategy revolves around running the EM-algorithm for 50 random starting values, stopping before full convergence, i.e., when the relative increase in composite log-likelihood is less than some moderate tolerance set to 10^{-2} in our case. Of all the 50 resulting parameter estimates, we choose the one maximizing the composite likelihood as starting point for the next part of the algorithm, which is to use a direct likelihood maximization to speed up convergence. Direct optimization is then run until full convergence, with convergence tolerance 10^{-5}. Hence, we get a hybrid algorithm that combines the large area of convergence for the EM algorithm with the computational efficiency of the direct likelihood methods.

4.3. Parameter estimation accuracy

We now assume that the starting point is relatively close to the true solution, so that direct optimization is guaranteed to converge properly. We then investigate different methods' accuracy

Table 3
Average computation time of each method for the optimization routine (in minutes).

Method	Case 1		Case 2	
	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$
Direct pairwise	1.79	1.89	2.32	2.00
Block, $m = 1$	1.22	1.34	1.44	1.42
Block, $m = 2$	6.50	8.70	7.83	6.78

Table 4
Average RMSE of parameter estimates.

Method	Case 1		Case 2	
	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$
Direct pairwise	0.200	0.203	0.194	0.199
Block, $m = 1$	0.182	0.165	0.190	0.194
Block, $m = 2$	0.184	0.163	0.188	0.188

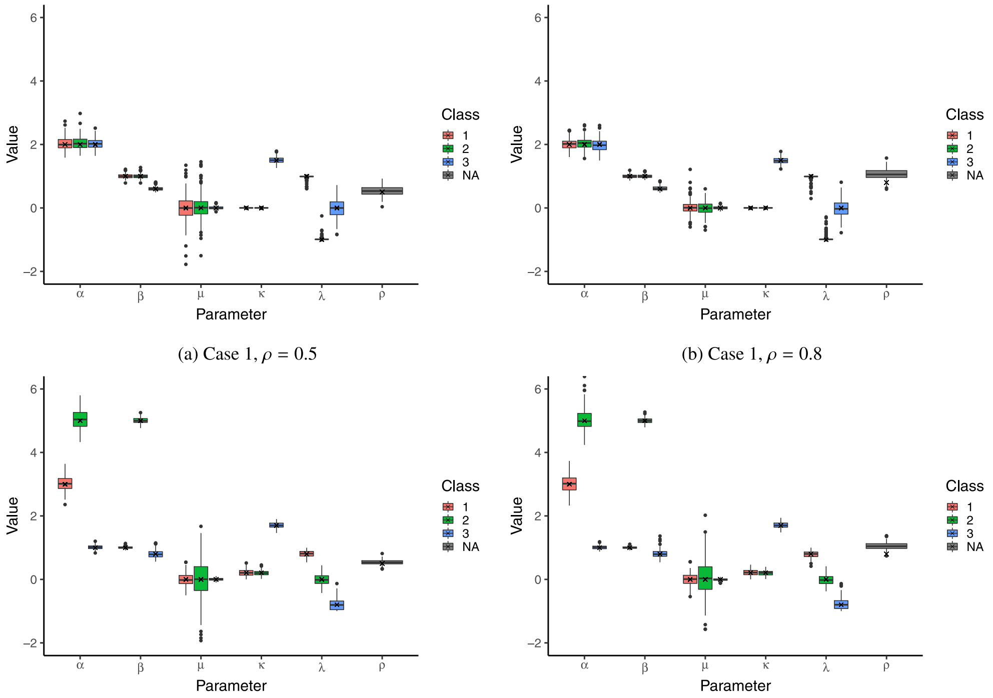
in parameter estimates and their computational efficiency. This is done to decide which of the two methods should be used for the second part of the hybrid algorithm. Since convergence is ensured for the second part of the hybrid algorithm, we prefer methods that estimate the parameters accurately while also maintaining a low run time. We consider the direct pairwise likelihood and block likelihood methods with $m = 1$ and $m = 2$. To compare the performance of the methods, we consider the RMSE of the parameter estimates as in Eq. (25) and average this over the random draws. We draw 50 replicates for both cases and both values of the spatial dependence parameter. As initial values for the optimizations we set the true parameters, in order to ensure and speed up the convergence. The average run times are shown in Table 3. On average the block composite likelihood method with $m = 1$ leads to roughly 30% reduction in computational time compared to the pairwise likelihood. Also, $m = 2$ is 5.5 times slower than $m = 1$, showcasing the exponential growth in computational complexity as m increases.

In Table 4 we list the average RMSE of the parameter estimates in all four cases for all three methods. The block composite likelihood approach with $m = 1$ outperforms the direct pairwise likelihood, although to a limited extent. Moreover, $m = 2$ performs slightly better than $m = 1$, as we would expect. Overall, the improvement is marginal and factoring in run time, we favor $m = 1$ for practical implementation. Tests of smaller sample size for $m = 3$ and $m = 4$ support the claim that increasing the number of rows only slightly improves performance, with a significant increase in run time.

4.4. Behavior of parameter estimates

We investigate the behavior of the estimates of each individual parameter by studying bias and variance. For each of the 4 cases, we draw 200 realizations of the latent field, each one followed by drawing observations from the WSSVM density. We use the block composite likelihood method with $m = 1$ to estimate parameters for each set of observations.

In Fig. 5 we display boxplots of the parameter estimates for all four cases in consideration. The density class for each parameter is indicated by the color. As there is only one true spatial interaction parameter ρ in each case, this has a gray color and class "NA". From the display, we see that overall all parameters look symmetric and exhibit little or no bias. Only the parameters κ and λ are asymmetric. These parameters can only take values greater than or equal to 0 or between -1 and 1, respectively, so when the true value is on or close to this limit, the estimates naturally get asymmetric. Further, there is a clear bias for $\rho = 0.8$, but not for $\rho = 0.5$. For $\rho = 0.8$, the average of all 200 estimates is roughly 1 in both cases, close to the critical value ρ_{crit}. There is a bias in ρ for larger values.



(a) Case 1, $\rho = 0.5$

(b) Case 1, $\rho = 0.8$

(c) Case 2, $\rho = 0.5$

(d) Case 2, $\rho = 0.8$

Fig. 5. Box plots of parameter estimates for Case 1 (top) and Case 2 (bottom) of the WSSVM density for $\rho = 0.5$ (left) and $\rho = 0.8$ (right). Parameter classes are indicated by the color. The crosses indicate the true parameter values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We compare these results to a similar experiment carried out by Ranalli et al. (2018). They used the direct pairwise likelihood method to find parameter estimates. Also, they only tested for low values of the spatial dependence parameter, $\rho = 0, 0.34, 0.5$. For these values, they found little bias. Similarly to our findings they report on asymmetry in λ and κ. Contrary to our results of little or no bias in all WSSVM parameters, they find that the bias of κ and λ is typically larger than the other estimates. Proposing the WSSVM density, Abe and Ley (2017) report no problems in finding maximum likelihood estimates with independently simulated data from a single WSSVM density. These results are more in line with our findings.

To highlight and ease the comparison of variability in parameter estimates, we plot the empirical variance of all 200 parameter estimates for all four cases in Fig. 6. The variability in parameter estimates is in this case a proxy on performance of the algorithm. Because all parameter estimates are unbiased (except for $\rho = 0.8$), small variability means that the algorithm finds the true value consistently. On the other hand, large variability implies that the parameter estimates deviate from the true value to a greater extent, although their average coincide with the true value. This means that the parameters with large variability are harder to estimate. Ranalli et al. (2018) claim that weakly separated classes lead to larger variability in parameter estimates. Judging from Fig. 6 we find no evidence for this claim, and we even observe much higher variance for some parameters in the case of large separation between classes. Our impression is that the variability is instead strongly linked to the properties of each class density: When densities are constructed such that change in a parameter value does not alter the density much, the variance of the estimate for that parameter becomes large. For instance, this is the case for μ_2 in Case 2 and to a lesser extent α_2 in Case 2 and λ_3 in Case 1. Density class 2 for Case 2 is approximately uniform in the circular component

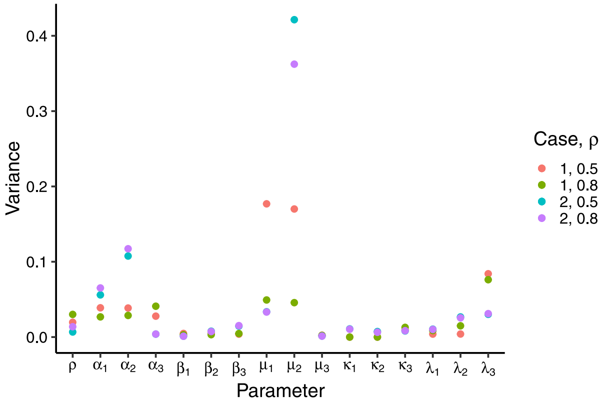


Fig. 6. Empirical variance in parameter estimates for the WSSVM density for the two cases and two values of the dependence parameter ρ.

due to $\lambda_2 = 0$ and κ_2 being low. Because μ_2 determines the circular location of this approximately circular uniform distribution, the density does not change much when μ_2 varies. Hence, μ_2 gets a large variability in Case 2. The same argument can be made for both Case 2 α_2 and Case 1 λ_3, small increments to these parameters have little effect on the corresponding densities.

A low value of ρ yields considerably larger variance in μ_1 and μ_2 for Case 1 and μ_2 for Case 2. If we consider the class 1 and 2 densities of Case 1 displayed in Figs. 4(a) and 4(b), we see that the combination of these two densities are almost uniform in the circular part. Hence, if we disregard the latent classes, or assume all data points to be independent, we are able to fit the data almost equally well by perturbing both densities in the circular part, i.e., vary both μ_1 and μ_2 simultaneously. Crucially, to keep with the uniform distribution μ_1 and μ_2 need to be about equal. This results in the optimization algorithm estimating values for μ_1 and μ_2 away from 0, and the effect is more prevalent for smaller values of ρ. For Case 1, the estimates of μ_1 and μ_2 tend to be heavily correlated, especially so for small spatial interaction parameter ρ. Since larger values of ρ imply a greater probability that neighboring points take the same latent class, we get better predictions of the latent classes. This leads to better estimation of the circular location parameters μ_1 and μ_2, and breaks up some of the correlation in these estimates.

A similar study was done for the GPTWC density, and we summarize some of the most important results. First, the same bias in the spatial dependence parameter ρ is observed for the GPTWC density. Also, there is positive bias in the heavy tail parameter τ when it is small, and a negative bias when it is large. The magnitudes of the biases are small, typically of size 0.02, but they are consistent and also in line with findings by Imoto et al. (2019), who report that the transition from overestimation to underestimation occurs for $\tau \geq 0.3$. All other parameters were unbiased.

5. Application to ocean surface currents data in the Norwegian Sea

We analyze a real data set of ocean surface currents (OSC). The data is presented in Section 5.1. The WSSVM pdf is considered in Section 5.2 and the GPTWC pdf in Section 5.3. Then we use model selection criteria in Section 5.4 to discuss the circular model results.

Modeling of OSC structures typically rely on computer intensive methods in the form of ocean models that solve large differential equation systems, see e.g., Slagstad and McClimans (2005). To capture the uncertainty involved with the complex ocean processes, statistical models have been shown to be very useful (Wikle et al., 2013). A key property of the models we develop is the need for compactness, i.e., sufficiently realistic, but relatively low-fidelity models that can be incorporated onboard for instance autonomous vehicles and drifters. These vessels are unable to run large-scale

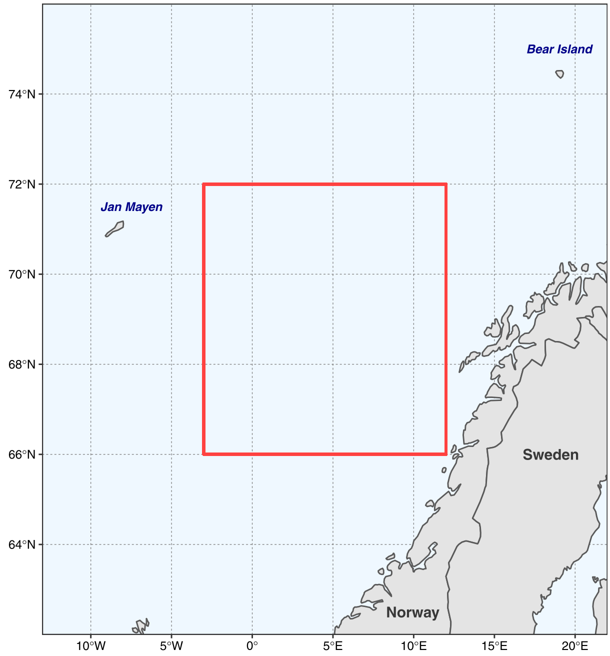


Fig. 7. Map of the Norwegian Sea. The area in consideration is indicated by the red rectangle.

numerical ocean models. Instead, one relies on models that can be computed on shore and then updated with information captured on the go (Fossum et al., 2019; Holm et al., 2020).

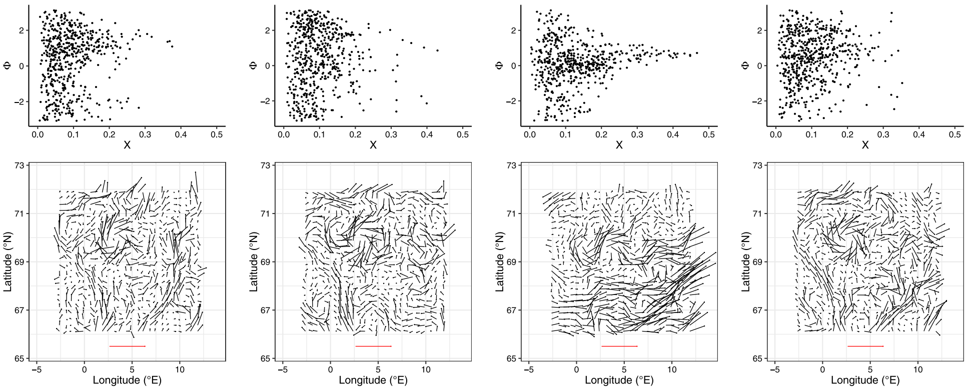
5.1. Data description

The OSC data we consider consist of measurements of the combined geostrophic and Ekman currents, and are provided by the GlobCurrent project, which is funded by the European Space Agency.[1] The area ranges from latitude 66°N to 72°N and longitude 3°W to 12°E (Fig. 7). The area is of interest because the inflow of Atlantic Water to the Arctic Oceans influences climate and has a great impact on the biological production in a vast region.

Ingvaldsen et al. (2002) describe temporal variability in the total flow of Atlantic Water from the Norwegian Sea into the Barents Sea, and conclude that there are large temporal and spatial variations in the current patterns. The mean state and variations in inflow of Atlantic Water into the Norwegian Sea was studied by Skagseth et al. (2008). Common for all, is that they used several time series of mooring observations to classify the states in the areas around the Norwegian Sea. By using spatial statistics on synoptic snapshot images, we can go further to classify typical current patterns in the Norwegian Sea.

The data are mapped to an equal angle grid of 0.25 degrees latitude by 0.25 degrees longitude. The distance between grid cells in the south–north direction is 28 km. However, due to the high latitude, distance between grid cells in the west–east direction is only approximately 11 km. To satisfy the Potts model assumption of equal spatial dependence in both directions, the distance between grid points needs to be the same in both directions. Hence, we thin the longitudinal locations. As a consequence, the grid points will not be evenly spaced, but it serves as a good

[1] The data are available to the public and downloadable through the GlobCurrent website: http://globcurrent.ifremer.fr/



(a) 2005 summer

(b) 2015 summer

(c) 2005 fall

(d) 2015 fall

Fig. 8. Display of the cylindrical observations (top) and the observations mapped to vectors at their grid points to create vector fields (bottom) for the four dates considered. The red arrow is plotted for reference and represents a current of angle 0 and speed 0.5 m/s. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

approximation. This gives a total size of the area of 667 km in the south–north direction and 676 km in the west–east direction. With a spatial resolution of 0.25 degrees, this means that the area is of size 24 × 24 and there are $n = 576$ observations. We consider OSC data from Summer (July 5) and Fall (October 3) in both 2005 and 2015.

Fig. 8 (top) shows scatter plots of the OSC speeds on the first axis, measured in m/s and angle of the OSC on the second axis. Here, an angle of 0 implies a current straight east and an angle of $\pi/2$ implies a current pointing straight north. Fig. 8 (bottom) shows the OSC as vectors on the grid. The vector displays give an overview of the spatial structure of the OSC data. For reference and to better relate to the length of the arrows, the plots also show a red arrow that represents a current pointing directly east, with speed 0.5 m/s beneath the spatial grid. In the displays of the vector fields, the Barents Sea entrance is close to the north-east corner. A major part of the flow in this corner points north, with less flow going east.

As noted by both Ingvaldsen et al. (2002) and Kwok et al. (2013), the current circulation patterns in the Norwegian Sea change with the seasons. The seasonal variability in the currents is in our model captured by the allocation of latent classes, with e.g., one season dominated by a circulation pattern corresponding to one latent class, and the other season displaying more of other classes. For one season, the data in 2005 and 2015 are treated as independent realizations from a model with the same class parameter settings. We model Summer and Fall data separately.

Because the number of ocean states K is unknown, it needs to be specified. For each season and for both WSSVM and GPTWC models, we estimate four different models by varying the number K of latent classes from 2 to 5. The hybrid algorithm starts by optimizing 50 random starting points with the EM algorithm, and then proceeds by block composite likelihood in the optimization.

5.2. WSSVM

For the WSSVM cylindrical density, the best model appears to be $K = 2$ (Summer) and $K = 3$ (Fall). In Summer, the C-BIC of $K = 2$ is 2026.4, compared to 2347.1 for $K = 3$ and 2412.8 for $K = 4$. For the Fall model, the C-BIC of $K = 2$ is 2515.1, compared to 2212.1 for $K = 3$ and 2511.3 for $K = 4$.

In Fig. 9, the latent class predictions are displayed for both seasons, and for year 2005 and 2015. The colors of the observations and current vectors represent the latent class predictions with a maximum marginal posterior prediction criterion. Notice that there are overall fairly large areas with equal latent class predictions. This is a result of large spatial dependency parameter ρ, and that

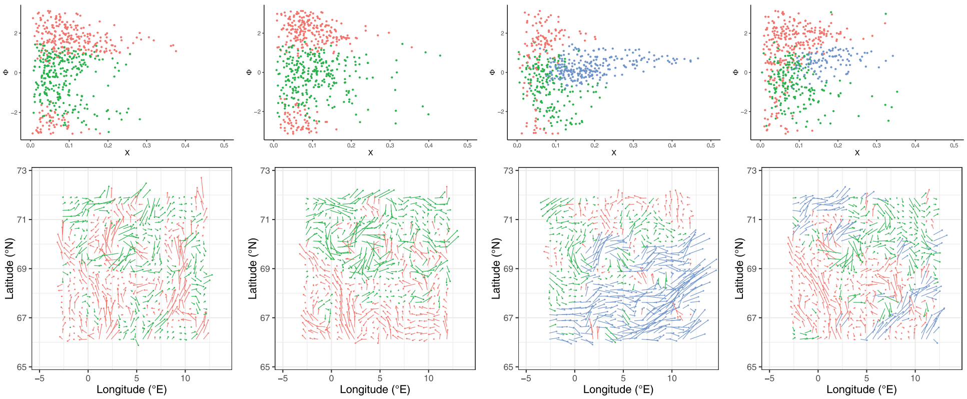


Fig. 9. Resulting prediction of the latent classes for the WSSVM density. From left to right: Summer 2005, Summer 2015, Fall 2005, Fall 2015. A maximum probability prediction criterion is used to predict the classes. The latent classes are indicated by the color of the points/arrows. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

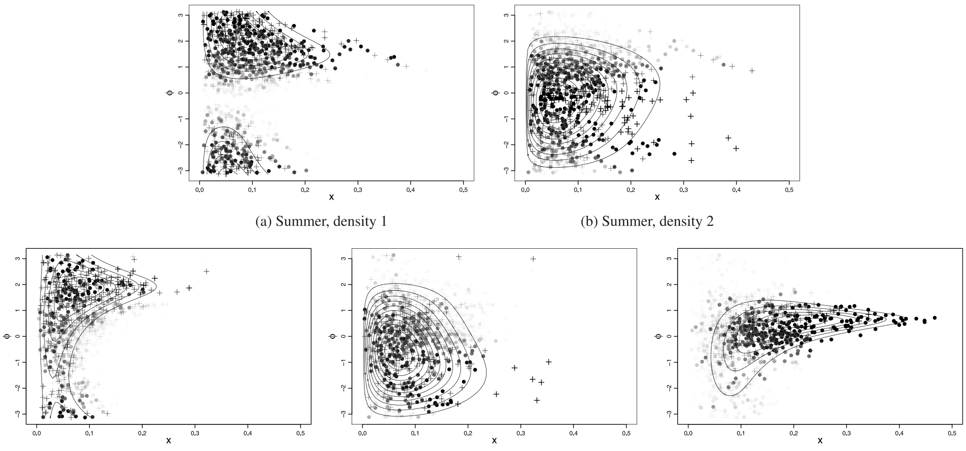
Table 5
Parameter estimates and bootstrap quantiles of the best Summer model with the WSSVM density.

Parameter	α_1	β_1	μ_1	κ_1	λ_1	α_2	β_2	μ_2	κ_2	λ_2	ρ
2.5% quantile	1.63	9.67	0.15	0.69	−0.26	1.48	7.94	0.02	0.01	−1.00	1.58
Estimate	1.94	12.17	1.32	0.80	1.00	1.58	9.12	1.21	0.13	−1.00	1.75
97.5% quantile	2.00	13.53	1.59	0.92	1.00	1.74	12.03	2.02	0.32	−1.00	1.90

the OSC is fairly homogeneous in large areas. By inspecting the displays of the observations in Fig. 8, we observe that the similarity between observations from 2005 and 2015 is higher for Summer than for Fall, implying that the Summer observations are more homogeneous in the cylindrical domain. The south-east corner of 2005 Fall contains currents of high speed that are highly concentrated eastwards. The same pattern is not observed for 2015. Hence, it makes sense that the Fall model includes more latent classes to account for increased variability in current patterns.

In Fig. 10, contour plots of the densities corresponding to the latent classes are displayed. We also show the OSC observations, using dots for 2005 and plus for 2015. The transparencies of the dots or plus represent how likely they are to take the latent class of the density displayed. These displays should be interpreted together with the maximum block composite likelihood estimates of the model parameters that are presented for the Summer model in Table 5 and the latent class predictions in Fig. 9. The Summer season observations are described in terms of two densities that represent unique current regimes. The first density is associated with large absolute angles, i.e., currents going west. The speeds are higher for positive angles than negative, implying larger speeds for currents pointing north-west, and lower speeds for currents pointing south-west. This density is skewed towards positive angles ($\lambda_1 = 1$) with moderate circular concentration ($\kappa_1 = 0.8$). This current pattern is most prominent in the south part of the 2015 data set, and the south-western corner of the 2005 data set. The second density is associated with small absolute angles, i.e., currents flowing east. This density is negatively skewed ($\lambda_2 = -1$) towards currents flowing south-east and the circular concentration is lower than for the first density ($\kappa_2 = 0.13$). A low circular concentration implies that the density is not as concentrated around a single directional mode, and the current directions are more spread out. Because of the low value of κ_2 the circular–linear dependence is lower for the second density, which is also observable from the displayed density contours. From the density displays we notice many outliers in the two distributions. This suggests that the compact model may be too simple to capture all realistic current patterns.

For the Fall model, density 1 and 2 appear fairly similar to those of Summer, and it is interesting to note that even though we did not enforce any physical structure in the modeling, the algorithm



(a) Summer, density 1

(b) Summer, density 2

(c) Fall, density 1

(d) Fall, density 2

(e) Fall, density 3

Fig. 10. Estimated densities of the best model for each season with the WSSVM density. The observations from 2005 and 2015 are plotted as dots and pluses respectively, with transparency of each dot/plus representing the probability of belonging to that class. Black dots/pluses represent a probability of 1 of belonging to that class.

recognizes this interpretable aspect. The most striking difference is that they are skewed in opposite ways, with $\lambda_1 = -0.26$ and $\lambda_2 = 1$ for the Fall model. The additional third density for Fall accounts for the large current speeds observed in the south-east corner of the 2005 observations, and also to some extent in the south-east and north-west corners of 2015. In these corners, the high-speed OSC are concentrated in the north-east direction. This makes the circular concentration very high ($\kappa_3 = 1.5$) around direction $\mu_3 = 0.69$, and it is also negatively skewed ($\lambda_3 = -1$). Even though the modal direction is north-east, the negative skewing means that the currents point more to the east than to the north.

We use parametric bootstrap to estimate empirical quantiles of the distribution of the maximum block composite likelihood estimates. The resulting empirical 2.5% and 97.5% quantiles, along with the estimates, are displayed in Table 5, for Summer only. All circular–linear dependence parameters κ are statistically significant. Further, the skewness parameter λ is significant in the second density, but not in the first. This indicates that the OSC data are skewed to some degree and justifies our choice of a skewed cylindrical distribution. The spatial dependence parameter ρ is clearly significant, substantiating the need for a hidden spatial process to explain the variability and dependence between OSC observations.

5.3. GPTWC

For the GPTWC model, the number of latent classes is similarly decided by varying K from 2 to 5. The lowest C-BIC of 2164.7 for the Summer model is achieved for $K = 2$ latent classes, and the C-BIC values are 2382.8 and 2406.9 for $K = 3$ and $K = 4$, respectively. For the Fall model, the lowest C-BIC of 2494.1 is attained for $K = 3$, with $K = 2$ having C-BIC of 2614.9 and $K = 4$ of 2630.9. This means that we get the same number of latent classes for both seasonal models as with the WSSVM density.

Similarly to what was done for the WSSVM model, Fig. 11 displays the results from latent class predictions with a maximum marginal posterior prediction criteria. As for the WSSVM model, the GPTWC model also has large areas with equal latent class predictions, suggesting a homogeneous current field and large spatial coupling.

Contour plots of the densities corresponding to the latent classes are displayed in Fig. 12. For the Summer model, the first density is associated with large absolute angles. This density resembles the first density of the WSSVM model, but also includes observations of lower speed with lower

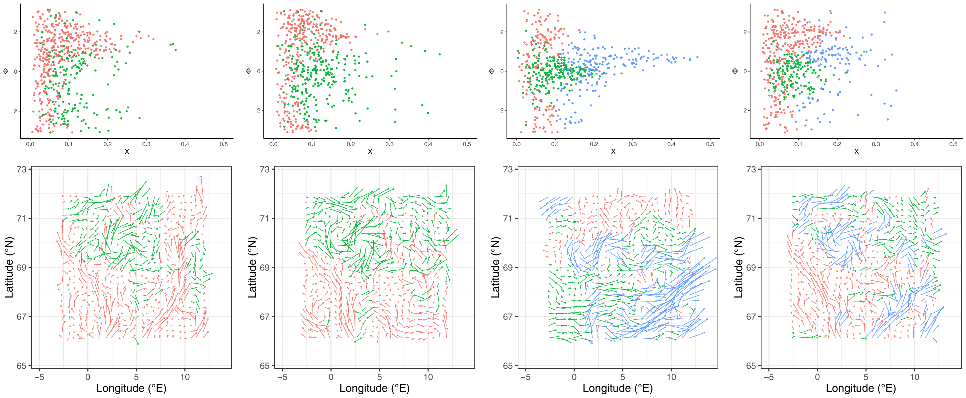
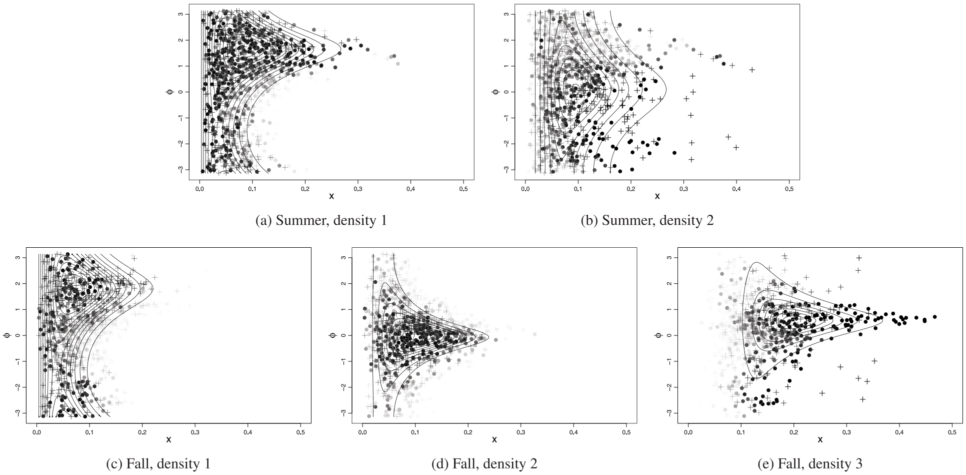


Fig. 11. Resulting prediction of the latent classes for the GPTWC density. From left to right: Summer 2005, Summer 2015, Fall 2005, Fall 2015. A maximum probability prediction criterion is used to predict the classes. The latent classes are indicated by the color of the points/arrows. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Summer, density 1

(b) Summer, density 2

(c) Fall, density 1

(d) Fall, density 2

(e) Fall, density 3

Fig. 12. Estimated densities of the best model for each season with the GPTWC density. The observations from 2005 and 2015 are plotted as dots and pluses respectively, with transparency of each dot/plus representing the probability of belonging to that class. Black dots/pluses represent a probability of 1 of belonging to that class.

absolute angles. The circular concentration is lower for this density than for the first density of the WSSVM model, which may explain the inclusion of observations with lower absolute angles. The circular part of the density is concentrated around $\mu_1 = 1.66$, which corresponds roughly to currents flowing straight north. Moreover, this density has virtually no heavy tailedness ($\tau_1 = 0.03$). This is induced by the fact that there are very few outliers in this density.

The second density is concentrated around $\mu_2 = 0.1$, i.e., currents flowing east, and has a lower circular concentration than the first density ($\kappa_2 = 0.45$). This means that the current directions are more spread out and not as concentrated around a single directional mode. This property was also observed for the WSSVM model, with the second density having lower circular concentration. Furthermore, the second density displays some degree of heavy tailedness ($\tau_2 = 0.27$) with more non-transparent outliers to the distribution. The GPTWC density is designed to deal with extreme linear observations, but the fitted model includes heavy tailedness purely in the second density. For

Table 6
Parameter estimates and bootstrap quantiles of the best Summer model with the GPTWC density.

Parameter	α_1	β_1	μ_1	τ_1	κ_1	α_2	β_2	μ_2	τ_2	κ_2	ρ
2.5% quantile	0.47	0.06	1.58	0.00	0.68	0.32	0.09	0.01	0.15	0.31	1.86
Estimate	0.52	0.07	1.66	0.03	0.72	0.36	0.10	0.10	0.27	0.45	2.47
97.5% quantile	0.56	0.07	1.78	0.11	0.78	0.46	0.11	0.54	0.33	0.52	2.97

the WSSVM Summer model, we also saw that only one density was significantly skewed, implying that the data to some extent displays both skewedness and heavy tailedness.

For the GPTWC model, as was also noted for the WSSVM, we see that density 1 and 2 for Summer resemble density 1 and 2 for Fall, and the parameter estimates are fairly similar. Again, the algorithm recognized this feature without any enforced constraints. The biggest differences are that the second density in the Fall model displays less heavy tailedness ($\tau_2 = 0.06$) and the circular concentration is larger ($\kappa_2 = 0.88$).

For the Fall model, only the third density is significantly heavy tailed ($\tau_3 = 0.21$). This third density is associated with current observations of high speed, centered around the modal direction $\mu_3 = 0.58$, i.e., currents pointing north-east. The circular concentration is fairly high ($\kappa_3 = 0.87$), and the many outliers are dealt with by the heavy tail property of the distribution. For the WSSVM model, these outliers were accounted for by the first two densities instead of allowing the third density to have heavy linear tails. This density clearly resembles the third density of the WSSVM model, with currents pointing north-east at high speed and high circular concentration.

We display quantiles achieved by parametric bootstrap (Table 6) for Summer only. Both circular–linear dependence parameters κ and the spatial coupling parameter ρ are significant.

5.4. Model comparison

We compare models using the continuous ranked probability score (CRPS), see e.g., Gneiting et al. (2008). We consider this scoring rule for univariate probabilistic forecasts. It is defined as,

$$\mathrm{CRPS}(P, x^o) = \int_{-\infty}^{\infty} \left(F(y) - I(y \geq x^o)\right)^2 dy, \tag{26}$$

where P denotes the predictive probability distribution, x^o is the observed value and F is the cumulative distribution function associated with P. The best model is the one that minimizes the CRPS.

For cylindrical data, the circular part has to be handled differently, and Grimit et al. (2006) proposed an analogue to the linear CRPS by considering the angular distance instead of absolute difference. The angular distance, wrapped around 2π, is then used in the CRPS calculation.

We randomly draw 50 hold-out grid points for each data set to conduct model comparison. For all the selected grid points, we calculate the CRPS for each model. To evaluate the scoring rules, we first need to compute the predictive distributions for each model at each selected grid point. Let $\mathbf{z}_{-i}$ denote all observations except $\mathbf{z}_i$, the hold-out probabilities $p(x_i|\mathbf{z}_{-i})$ are conveniently defined by a re-normalization,

$$p(l_i|\mathbf{z}_{-i}) = C\frac{p(l_i|\mathbf{z})}{p(\mathbf{z}_i|l_i)}, \tag{27}$$

where C is a normalizing constant dependent only on $\mathbf{z}$, $p(l_i|\mathbf{z})$ are the backward probabilities defined in Eq. (12) and $p(\mathbf{z}_i|l_i)$ is either from the WSSVM or GPTWC distribution. These probabilities are then used to define the predictive distribution of observing $\mathbf{z}_i^o$, given all the other observations $\mathbf{z}_{-i}$, as a mixture over the latent classes,

$$p(\mathbf{z}_i^o|\mathbf{z}_{-i}) = \sum_{l_i \in \mathbb{L}} p(\mathbf{z}_i^o|l_i)p(l_i|\mathbf{z}_{-i}). \tag{28}$$

Table 7
Average circular and linear CRPS for the estimated models. The model with the lowest CRPS is indicated with **bold** for each data set.

Model	Circular		Linear	
	WSSVM	GPTWC	WSSVM	GPTWC
Summer 2005	**0.1245**	0.1330	0.0074	**0.0073**
Summer 2015	**0.1298**	0.1461	0.0083	**0.0073**
Fall 2005	**0.1263**	0.1270	0.0114	**0.0104**
Fall 2015	**0.1211**	0.1289	0.0085	**0.0079**

In Table 7, the average linear and circular CRPS for the 50 grid points are displayed for each data set. To make an even comparison of the two models, we use $K = 2$ classes for the Summer data and $K = 3$ classes for the Fall data. Lower CRPS is a result of either less bias in the prediction or a sharper predictive distribution. From the listed values, we observe that the WSSVM models are better at predicting the circular part, while the GPTWC models are better at predicting the linear part. These are encouraging results, bearing in mind the design of the two cylindrical distributions; WSSVM is designed to handle skewness in the circular part, while GPTWC is designed to account for heavy tails in the linear part.

6. Conclusions

We extend a method for inferring model parameters in the setting of spatially dependent cylindrical data, where a Potts model is used for discrete latent classes. The method of inference is investigated for two cylindrical densities; one allowing skew circular data and another allowing a heavy-tailed linear part. The main statistical methodological contribution involves a hybrid algorithm bridging a fast and statistically efficient block composite likelihood method with a stable converging expectation–maximization algorithm. The block composite likelihood method conducts exact evaluations for spatial blocks of variables by leveraging a recursive calculation on these subsets of the data. Simulation studies show that the block composite likelihood method takes only about 50 % of the time to converge, compared with the expectation–maximization algorithm. However, simulation studies also show that the convergence radius of the complex optimization problem is clearly smaller for the block composite likelihood, so it is useful to start with expectation–maximization runs and then switch to block composite likelihood evaluations in the optimization procedure.

We apply the method to ocean surface current data from the Norwegian Sea. Here, the model serves as a parsimonious representation in that it breaks down current patterns into a discrete number of local regimes, which can be interpreted and could be useful in real-time operations onboard an autonomous vehicle. For the data that we studied, the composite information criterion indicated that 2 or 3 latent classes are appropriate in the Potts model. The skew cylindrical model provides best fit to the circular components (current direction), while there might be heavy-tails for the linear component (current strength).

An unexplored path for future work is to combine the skew and heavy tailed distributions to form a density that enables both skew circular part and heavy-tailed linear part. This could be done by sine-skewing the wrapped Cauchy distribution for the circular part in the heavy-tailed distribution, but the theoretical results related to this new density are not yet available and would have to be derived. Also, the issue concerning reliable parameter estimation would have to be tested.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We acknowledge support from Norwegian Research Council, Norway (RCN) project number 305445.

References

Abe, T., Ley, C., 2017. A tractable, parsimonious and flexible model for cylindrical data, with applications. Econom. Stat. 4, 91–104. http://dx.doi.org/10.1016/j.ecosta.2016.04.001.

Allard, D., d'Or, D., Froidevaux, R., 2011. An efficient maximum entropy approach for categorical variable prediction. Eur. J. Soil Sci. 62 (3), 381–393.

Ameijeiras-Alonso, J., Lagona, F., Ranalli, M., Crujeiras, R.M., 2019. A circular nonhomogeneous hidden Markov field for the spatial segmentation of wildfire occurrences. Environmetrics 30 (2), e2501.

Barkema, G., de Boer, J., 1991. Numerical study of phase transitions in Potts models. Phys. Rev. A 44, 8000–8005. http://dx.doi.org/10.1103/PhysRevA.44.8000.

Bartolucci, F., Besag, J., 2002. A recursive algorithm for Markov random fields. Biometrika 89 (3), 724–730.

Baum, L.E., Petrie, T., Soules, G., Weiss, N., 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Stat. 41 (1), 164–171. http://dx.doi.org/10.1214/aoms/1177697196.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B Stat. Methodol. 36 (2), 192–236, URL http://www.jstor.org/stable/2984812.

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. http://dx.doi.org/10.1007/s13253-012-0110-1.

Caragea, P.C., Smith, R.L., 2007. Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. J. Multivariate Anal. 98 (7), 1417–1440.

Cox, D.R., Reid, N., 2004. A note on pseudolikelihood constructed from marginal densities. Biometrika 91 (3), 729–737. http://dx.doi.org/10.1093/biomet/91.3.729.

Eidsvik, J., Shaby, B.A., Reich, B.J., Wheeler, M., Niemi, J., 2014. Estimation and prediction in spatial models with block composite likelihoods. J. Comput. Graph. Statist. 23 (2), 295–315. http://dx.doi.org/10.1080/10618600.2012.760460.

Fossum, T.O., Ryan, J., Mukerji, T., Eidsvik, J., Maughan, T., Ludvigsen, M., Rajan, K., 2019. Compact models for adaptive sampling in marine robotics. Int. J. Robot. Res. http://dx.doi.org/10.1177/0278364919884141.

Friel, N., Rue, H., 2007. Recursive computing and simulation-free inference for general factorizable models. Biometrika 94 (3), 661–672. http://dx.doi.org/10.1093/biomet/asm052.

Gao, X., Song, P.X.-K., 2010. Composite likelihood Bayesian information criteria for model selection in high-dimensional data. J. Amer. Statist. Assoc. 105 (492), 1531–1540. http://dx.doi.org/10.1198/jasa.2010.tm09414.

Givens, G.H., Hoeting, J.A., 2013. Computational Statistics, second ed. John Wiley & Sons, Ltd, ISBN: 9781118555552, pp. 287–321. http://dx.doi.org/10.1002/9781118555552.ch9, Chapter 9.

Gneiting, T., Stanberry, L., Grimit, E., Held, L., Johnson, N., 2008. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. TEST 17, 211–235. http://dx.doi.org/10.1007/s11749-008-0114-x.

Godambe, V.P., 1960. An optimum property of regular maximum likelihood estimation. Ann. Math. Stat. 31 (4), 1208–1211. http://dx.doi.org/10.1214/aoms/1177705693.

Grimit, E.P., Gneiting, T., Berrocal, V.J., Johnson, N.A., 2006. The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. Q. J. R. Meteorol. Soc. 132 (621C), 2925–2942. http://dx.doi.org/10.1256/qj.05.235.

Guyon, X., 1995. Random Fields on a Network. In: Probability Theory and Stochastic Modelling, Springer.

Hanks, E.M., Hooten, M.B., Alldredge, M.W., 2015. Continuous-time discrete-space models for animal movement. Ann. Appl. Stat. 9 (1), 145–165. http://dx.doi.org/10.1214/14-AOAS803.

Holm, H.H., Sætra, M.L., Van Leeuwen, P.J., 2020. Massively parallel implicit equal-weights particle filter for ocean drift trajectory forecasting. J. Comput. Phys. X 100053.

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. http://dx.doi.org/10.1007/s10651-006-0015-7.

Imoto, T., Shimizu, K., Abe, T., 2019. A cylindrical distribution with heavy-tailed linear part. Jpn. J. Stat. Data Sci. http://dx.doi.org/10.1007/s42081-019-00031-5.

Ingvaldsen, R., Loeng, H., Asplin, L., 2002. Variability in the Atlantic inflow to the Barents Sea based on a one-year time series from moored current meters. Cont. Shelf Res. 22 (3), 505–519. http://dx.doi.org/10.1016/S0278-4343(01)00070-X.

Johnson, R.A., Wehrly, T.E., 1978. Some angular-linear distributions and related regression models. J. Amer. Statist. Assoc. 73 (363), 602–606.

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. http://dx.doi.org/10.1214/12-aoas576.

Jona Lasinio, G., Santoro, M., Mastrantonio, G., 2020. CircSpaceTime: an R package for spatial and spatio-temporal modelling of circular data. J. Stat. Comput. Simul. 90 (7), 1315–1345.

Kato, S., Shimizu, K., 2008. Dependent models for observations which include angular ones. J. Statist. Plann. Inference 138 (11), 3538–3549. http://dx.doi.org/10.1016/j.jspi.2006.12.009.

Kwok, R., Spreen, G., Pang, S., 2013. Arctic sea ice circulation and drift speed: Decadal trends and ocean currents. J. Geophys. Res. Oceans 118 (5), 2408–2425. http://dx.doi.org/10.1002/jgrc.20191.

Lagona, F., Picone, M., 2016. Model-based segmentation of spatial cylindrical data. J. Stat. Comput. Simul. 86 (13), 2598–2610. http://dx.doi.org/10.1080/00949655.2015.1122791.

Lagona, F., Picone, M., Maruotti, A., 2015a. A hidden Markov model for the analysis of cylindrical time series. Environmetrics 26 (8), 534–544.

Lagona, F., Picone, M., Maruotti, A., Cosoli, S., 2015b. 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.

Lindsay, B., 1988. Composite likelihood methods. Contemp. Math. 80, 221–239. http://dx.doi.org/10.1090/conm/080/999014.

MacDonald, I.L., Zucchini, W., 1997. Hidden Markov And Other Models for Discrete-Valued Time Series, Vol. 110. CRC Press.

Mardia, K.V., Jupp, P.E., 1999. Directional Statistics. In: Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd., ISBN: 9780471953333, http://dx.doi.org/10.1002/9780470316979.

Mardia, K.V., Sutton, T.W., 1978. A model for cylindrical variables with applications. J. R. Stat. Soc. Ser. B Stat. Methodol. 40 (2), 229–233.

Mastrantonio, G., 2018. The joint projected normal and skew-normal: A distribution for poly-cylindrical data. J. Multivariate Anal. 165, 14–26.

Modlin, D., Fuentes, M., Reich, B., 2012. Circular conditional autoregressive modeling of vector fields. Environmetrics 23 (1), 46–53. http://dx.doi.org/10.1002/env.1133.

Pewsey, A., García-Portugués, E., 2020. Recent advances in directional statistics. arXiv preprint arXiv:2005.06889.

Ranalli, M., Lagona, F., Picone, M., Zambianchi, E., 2018. Segmentation of sea current fields by cylindrical hidden Markov models: a composite likelihood approach. J. R. Stat. Soc. Ser. C. Appl. Stat. 67 (3), 575–598. http://dx.doi.org/10.1111/rssc.12240.

Reeves, R., Pettitt, T., 2004. Efficient recursions for general factorisable models. Biometrika 91 (3), 751–757. http://dx.doi.org/10.1093/biomet/91.3.751.

Skagseth, Ø., Furevik, T., Ingvaldsen, R., Loeng, H., Mork, K.A., Orvik, K.A., Ozhigin, V., 2008. Volume and heat transports to the Arctic Ocean via the Norwegian and Barents Seas. In: Dickson, R.R., Meincke, J., Rhines, P. (Eds.), Arctic–Subarctic Ocean Fluxes. Springer, Dordrecht, pp. 45–64. http://dx.doi.org/10.1007/978-1-4020-6774-7_3, Chapter 3.

Slagstad, D., McClimans, T.A., 2005. Modeling the ecosystem dynamics of the Barents Sea including the marginal ice zone: I. Physical and chemical oceanography. J. Mar. Syst. 58 (1), 1–18. http://dx.doi.org/10.1016/j.jmarsys.2005.05.005.

Swendsen, R.H., Wang, J.-S., 1987. Nonuniversal critical dynamics in Monte Carlo simulations. Phys. Rev. Lett. 58, 86–88. http://dx.doi.org/10.1103/PhysRevLett.58.86.

Tjelmeland, H., Austad, H.M., 2012. Exact and approximate recursive calculations for binary Markov random fields defined on graphs. J. Comput. Graph. Statist. 21 (3), 758–780.

Varin, C., Reid, N., Firth, D., 2011. An overview of composite likelihood methods. Statist. Sinica 21 (1), 5–42.

Varin, C., Vidoni, P., 2005. A note on composite likelihood inference and model selection. Biometrika 92 (3), 519–528.

Wang, F., Gelfand, A.E., 2014. Modeling space and space-time directional data using projected Gaussian processes. J. Amer. Statist. Assoc. 109 (508), 1565–1580. http://dx.doi.org/10.1080/01621459.2014.934454.

Wang, F., Gelfand, A., 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. Statist. Sinica 25, 25–39. http://dx.doi.org/10.5705/ss.2013.204w.

Wikle, C.K., Milliff, R.F., Herbei, R., Leeds, W.B., 2013. Modern statistical methods in oceanography: A hierarchical perspective. Statist. Sci. 466–486.

Wu, F.Y., 1982. The Potts model. Rev. Modern Phys. 54, 235–268. http://dx.doi.org/10.1103/RevModPhys.54.235.