Henrik Syversveen Lie

Cylindrical hidden Markov random field models with applications to ocean surface currents

Master's thesis in Applied Physics and Mathematics

Supervisor: Jo Eidsvik

June 2020

Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Henrik Syversveen Lie

Cylindrical hidden Markov random field models with applications to ocean surface currents

Master's thesis in Applied Physics and Mathematics
Supervisor: Jo Eidsvik
June 2020

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Abstract

Observations of ocean surface currents are represented by direction vectors and give rise to spatial cylindrical data, which are bivariate representations of a linear magnitude and a circular angle. To analyse such data we develop a hidden Markov random field model. The model segments the spatial domain into latent classes, with the structure of a Potts model. Each class defines a cylindrical density that represents a specific circulation pattern, or state of the ocean. The classes decide the patterns by supplying a set of fixed parameters to the cylindrical densities. 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 way, the model can parsimoniously account for various commonly observed features in cylindrical data, such as circular-linear dependence, multimodality, skewness, and heavy-tailedness.

Because the likelihood function of the model is computationally unfeasible, approximations are needed to estimate the model parameters. Hence, we consider two approaches towards forming a composite-likelihood. First, we regard pairs of observations as components of the likelihood. This method reduces to an expectation-maximization algorithm that is simple to implement and works iteratively by first predicting the probability of the latent classes and then maximizing the composite-likelihood based on these probabilities. The second method regards larger blocks of observations and computes the exact likelihood of each block by a spatial extension of the forward–backward algorithm for hidden Markov models. The properties of the two methods are investigated in a simulation study, indicating that the former has a larger area of convergence, whereas the latter approximately halves the computation time. Based on the results, we develop a hybrid algorithm that combines the large area of convergence of the expectation-maximization algorithm with the computational efficiency of the block-likelihood.

We employ the hybrid algorithm to study ocean surface currents at two locations in the Norwegian Sea. In both cases, the model is able to describe the currents in terms of interpretable local regimes. We apply scoring rules to measure how well the two cylindrical densities match the two data sets. Results indicate that both densities have merit for the first data set, whereas the second does not display heavy tails.

Sammendrag

Observasjoner av overflatestrømmer i havet gir opphav til romlige sylindriske data, som er bivariate representasjoner av en lineær styrke og en sirkulær vinkel. For å kunne analysere disse dataene utvikler vi en skjult Markov tilfeldig feltmodell. Modellen segmenterer det romlige domenet i latente klasser, med strukturen til en Potts modell. Hver klasse definerer en sylindrisk tetthet som representerer et spesifikt sirkulasjonsmønster. Klassene bestemmer mønstrene ved å tilføre et sett fikserte parametere til de sylindriske tetthetene. Vi vurderer to ulike sylindriske fordelinger, Weibull sine-skewed von Mises fordelingen med mulighet for skjevhet i den sirkulære delen, og generalized Pareto-type wrapped Cauchy fordelingen med mulighet for tunge haler i den lineære delen. På denne måten er modellen i stand til å redegjøre for forskjellige typiske trekk i sylindriske data, slik som sirkulær-lineær avhengighet, multimodalitet, skjevhet og tunge haler.

Fordi modellens likelihood funksjon er for krevende å beregne, er det behov for tilnærminger for å estimere modellparametrene. Vi vurderer to tilnærminger til en sammensatt likelihood funksjon. Først ser vi på par av observasjoner som komponenter i likelihood funksjonen. Denne metoden reduseres til en forventnings-maksimeringsalgoritme som er enkel å implementere og fungerer iterativt ved først å predikere sannsynligheten for de latente klassene og deretter maksimere den sammensatte likelihood funksjonen basert på disse sannsynlighetene. Den andre metoden betrakter større blokker av observasjoner og beregner den eksakte likelihood funksjonen for hver blokk ved hjelp av en romlig utvidelse av forward–backward algoritmen for skjulte Markov-modeller. En simuleringsstudie indikerer at førstnevnte metode har et større konvergensområde, mens sistnevnte halverer beregningstiden. Basert på resultatene utvikler vi en hybrid algoritme som kombinerer det store konvergensområdet til forventnings-maksimaliseringsalgoritmen med beregningseffektiviteten til blokk likelihood algoritmen.

Vi benytter den hybride algoritmen for å studere overflatestrømmer to steder i Norskehavet. I begge tilfeller er modellen i stand til å bryte strømningene ned i tolkbare lokale regimer. For å kunne sammenligne egnetheten til de to sylindriske tetthetene for hvert datasett, benytter vi såkalte ”scoring rules”. Resultatene indikerer at begge tetthetene kan være nyttige for det første datasettet, mens det andre datasettet ikke viser tegn til tunge haler.

Preface

This thesis concludes my M.Sc. degree in Physics and Mathematics at NTNU. Over the course of the last five years I have chosen to specialize in industrial mathematics, focusing on statistics. However, my interest in spatial statistics started merely last year, thanks to the always entertaining lectures with Henning Omre. This interest is what eventually lead me to the topic of this thesis.

First and foremost, I would like to thank my supervisor, Jo Eidsvik, for providing advice, engaging in discussion over the thesis and taking interest in my success. It has been a pleasure working with you over the last year. I would also like to thank Laurent Bertino at the Nansen Environmental and Remote Sensing Center for providing me with ocean current data from the GlobCurrent project, Annette Stahl from the Department of Engineering Cybernetics at NTNU for discussing the topics of my thesis, and Morten Omholt Alver from the Department of Engineering Cybernetics at NTNU for preparing data from the SINMOD model. Thanks also to all my friends who have supported me and kept me in good spirits over these past five years. A special mention goes to Erik, Martin, Lars, Mikal & Bjørnulf. Finally, I thank my partner and family for their support and keeping me company during these last few months of social distancing.

Henrik Syversveen Lie,
Bærum, June 2020

Table of Contents

Chapter 1

Introduction

In this chapter, we motivate the thesis and provide a general introduction to ocean surface currents (OSC) and how they are monitored. OSC data are represented by direction vectors or equivalently by their components as angles and magnitudes, and we also provide an introduction to such kind of cylindrical data in a statistical setting. Then we present a problem description and an outline for the chapters in the thesis.

1.1 Ocean Surface Currents

The ocean is ever-changing, chaotic in nature, and always in motion. The motion is visually manifested in breaking waves at the ocean surface, but these are merely one of many forms of motion. Other kinds of motion include eddies, narrow currents, filaments, and turbulence. These span a vast diversity of spatial scales, from over $100\,\mathrm{km}$ to less than $1\,\mathrm{m}$. Temporal scales also range from decadal climate-related changes to hourly tidal currents. Global surface motions are induced by the earth's rotation, global winds, and the continent borders. Also, interplay between global and local conditions makes the motions even more complex.

OSC can be described as *a coherent horizontal and vertical movement of water – in contact with the surface and over a specific depth regime – with a given velocity that persists over a region and time period* [Chapron et al., 2015]. Going forward, we consider OSC as only water moving horizontally close to the surface. OSC are the results of local and external characteristics such as winds, waves, tides, mixed layer depth, and turbulence. When these random characteristics interact with flows of larger scale they create complex current structures. Hence, stochastic models are needed to estimate OSC drift.

There are several influential forces affecting the water masses that make up the oceans. Some of them include forces relating to variations in surface elevation, wind friction, Coriolis force connected to Earth's rotation, and forces due to density disparity in the water. These forces generate several flows and currents, among others tidal motion, internal waves, Kelvin waves, and Rossby waves, but our focus is on the geostrophic and Ekman currents.

For large scales, the ocean is approximately hydrostatic, meaning that the height and density of the water column determine pressure. Then, the balance of force is between horizontal pressure differences and the Coriolis effect [Dohan and Maximenko, 2010], and currents are parallel to lines of constant pressure, so-called isobars. These currents are known as geostrophic currents and can be calculated from gradients of the sea surface height, because the pressure depends on the surface height. In other words, geostrophic currents come as a result of small differences in the surface height. In the Northern (Southern) hemisphere, high pressure is to the right (left) of the flow. Large scales dominate the open ocean, thus the mean ocean motion is treated as approximately geostrophic. Major currents such as the Gulf Stream and the Agulhas Current are examples of geostrophic currents.

As already stated, geostrophic currents dominate large-scale regions, but this does not mean that all large-scale motions are geostrophic. In the Ekman balance, introduced by Ekman [1905], the Coriolis effect is balanced by the vertically decaying stress induced by wind friction. Assuming steady winds, infinitely deep ocean, and constant vertical eddy viscosity, the theoretical Ekman current flows at 45° to the right (left) of the wind direction in the Northern (Southern) hemisphere at the surface [Dohan and Maximenko, 2010]. The angle increases and the speed decreases with depth. Ekman currents can be computed by estimates of wind stress from measurements of wind speed and direction.

In this thesis, we consider OSC in the Norwegian Sea, which is part of the Arctic Ocean. The Arctic plays a crucial role in the climate system and has been substantially affected by the ongoing climate change. Changes to the Arctic are easily observable, with increased ocean and air temperatures, stark shrinking of sea ice areas, and decreased sea ice thickness. Because of this, the OSC in the Arctic Ocean is expected to endure changes, but quantification is difficult [Johannessen et al., 2014]. Undoubtedly, changes to the Arctic environment will pose severe consequences for its surroundings, and both local and global ecosystems will get affected. Also, flow of warm and saline Atlantic water to the Arctic Ocean has great impact on the Arctic climate. This is why broad knowledge and quantification of ocean circulation and its variability is needed for the Arctic Ocean.

The ocean circulation in the Arctic region has been extensively studied ever since the first approaches by Helland-Hansen and Nansen [1909]. The dominating current in the Norwegian Sea is the warm Norwegian Atlantic Current, extending the North Atlantic Current, which again is a continuation of the Gulf Stream. The Norwegian Atlantic Current comprises three currents: one at the coast, one at the edge of the continental shelf, and one further offshore [Mork and Blindheim, 2000]. These three currents merge west of Lofoten and Vesterålen. The Norwegian Sea also gets inflow from the East Icelandic Current, which is a mix between Atlantic water (warm and salty) and Arctic water (cold and fresh). At the Barents Sea entrance, the Norwegian Atlantic Current splits into two branches, one going eastwards into the Barents Sea, and the other continuing north towards Spitsbergen. The inflow from the Norwegian Atlantic Current to the Barents Sea has been studied to examine the current velocity field and describe its seasonal variability [Ingvaldsen et al., 2002, 2004]. The Norwegian Coastal Current flows along the Norwegian coast and transports fresh water from the Baltic Sea towards the Barents Sea. The interaction between the coastal and the Atlantic currents is also of high importance for the circulation in the Norwegian Sea.

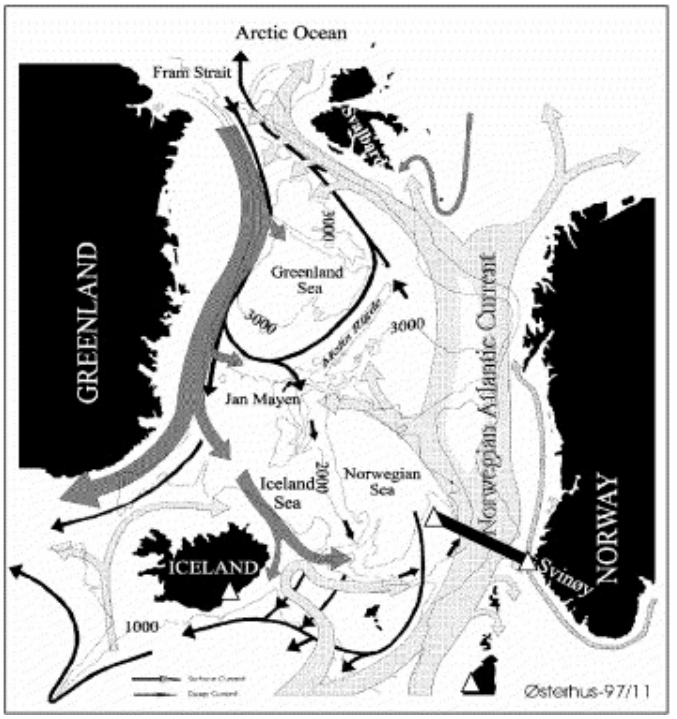


Figure 1.1: Schematic presentation of the circulation pattern in the Nordic Seas. Deep currents are represented by black arrows. Cold and fresh surface currents are represented by dark grey arrows, whereas the light gray arrows represent warm and saline currents. Taken from Østerhus et al. [1996].

Figure 1.1 is taken from Østerhus et al. [1996] and presents the large scale currents in the Nordic Seas in a schematic way. The northward flowing Norwegian Atlantic Current is depicted in light grey to indicate that it contains warm and saline Atlantic water. The Norwegian Coastal Current flows closer along the Norwegian coast. It is important to note that this image merely displays the general large scale flow of the currents. In practice and at smaller scales, the currents behave in a chaotic manner and even interact with each other. This calls for stochastic models to accurately represent the current patterns.

1.2 Monitoring

Initially, sailors and oceanologists were dependent on their personal observations to map and study OSC. Over the past few decades there has been an extensive deployment of various types of satellites, buoys, gliders, and drifters. These provide synoptic images as well as *in situ* measurements, and major advances have been made towards combining these two forms of observations [Chapron et al., 2015]. As we discussed in the previous section, OSC have a complex behaviour. They span a broad range of both spatial and temporal scales and depend on both global and local factors. Thus, remote sensing

through satellites is an ideal tool to study the dynamics of ocean surfaces. Satellites offer the possibility to consistently and systematically do repeated measurements of specific areas, which is unfeasible with *in situ* measurements. However, the repeated synoptic view comes at the cost of lower spatial and temporal resolution. Consequently, remote sensing observations should ideally be accompanied by observations with higher resolution, and capturing synergies from such a diverse range of observations is an important challenge. Because accuracy varies in all types of observations, it is of high importance to quantify and model the uncertainty in data involving ocean processes. In light of these demands, Oceanographic research formed in a statistical framework has been very productive [Wikle et al., 2013].

Satellite altimetry is a way of measuring the sea surface height, and one of the most mature techniques for mapping ocean currents [Chapron et al., 2015]. Numerous measurements and calculations are needed to calculate the surface height. In addition to altimetry data, this includes accurate information about the satellite orbit and tide water. Crucially, the height is measured relative to the so-called geoid, the sea surface at resting state, and precise knowledge about the geoid is also necessary [Dohan and Maximenko, 2010].

However, altimeters are limited to relatively coarse spatial grids or low temporal resolution. The satellite can pass over an area frequently, but this comes at the cost of large distances between each tracking line. Hence, a compromise needs to be made between the temporal and spatial resolution. Deploying multiple altimeters solves some of the problems, but resolution is still restricted to approximately $100\,\mathrm{km}$ and 10 days. In specific areas, attempts have been made towards combining the altimeter data with other sources of higher-resolution observations. These include *in situ* observations and microwave images of sea surface temperature.

Several types of instruments are able to monitor and measure winds to calculate Ekman currents. These include scatterometers, passive polarimetric sensors and synthetic aperture radar (SAR). For a more detailed overview of these types of sensors we refer to Bourassa et al. [2010]. The Ekman current data we use in this thesis have been calculated by using trajectories from SVP-type drifting buoys and Argo floats [Chapron et al., 2015]. From the drifter velocities, the Ekman currents are extracted by subtracting the geostrophic current component measured by altimetry.

The first set of data we consider in this thesis are collected by the GlobCurrent project. The data, as well as extensive documentation on collection procedures and details on calculations, are publicly available on the GlobCurrent web page.[1] OSC are not merely a simple addition of current components, but the combined current data we consider are the sum of the geostrophic and the Ekman current, given by

$$\mathbf{u} = \mathbf{u}_{\text{geost}} + \mathbf{u}_{\text{Ek}}. \tag{1.1}$$

Current data are provided both at the significant wave height and at $15\,\mathrm{m}$ depth, but we will only use the former. The data span all the world's major oceans and seas with a spatial resolution of $0.25°$ latitude and longitude. Additionally, data are provided on a 3 hourly time scale from 1993 to 2017.

[1] `http://globcurrent.ifremer.fr/`

The second data set originates from the SINMOD[2] ocean model, which is a numerical ocean model that has been developed by SINTEF since the 1980s. The model is built on a 3D hydrodynamic model that is based on the primitive Navier-Stokes equations. It incorporates atmospheric forcings, such as wind, air temperature and air pressure, and also freshwater runoff and tidal components. Currents are used as boundary conditions, and the model is nested, enabling lower resolution setups to produce boundary values for higher resolution implementations. A detailed description of the model is given by Slagstad and McClimans [2005]. The model originally covered most of the Barents Sea, but has later been extended to other areas. The horizontal resolution is equal in both direction and varies from $12\,\mathrm{km}$ to $32\,\mathrm{m}$. The vertical layers have varying thickness, with higher resolution closer to the ocean surface. In this thesis, we use data with $800\,\mathrm{m}$ horizontal resolution. We use the upmost vertical layer, which due to the vertical layer setup is $3\,\mathrm{m}$ below the ocean surface. Critically, this data set differs from the first one in that the data are simulated from an ocean model, rather than calculated directly from satellite observations and drifter trajectories.

1.3 Cylindrical data

Circular data arise in applications involving directions, and are usually expressed in terms of angles relative to a fixed reference point. When the data also contain a linear magnitude, they are referred to as cylindrical, because the pair of an angle and a magnitude can be interpreted as a point on a cylinder, with the magnitude representing the height of the cylinder. As such, the angle is called the circular part and resides on the unit circle, where it is measured for the most part in radians, but also sometimes in degrees. The magnitude is called the linear part and takes values on the non-negative real line. Examples of cylindrical data include ocean wave direction and height [Wang et al., 2015], wind direction and speed [Modlin et al., 2012], X-ray imaging [Abraham et al., 2013], and animal migration direction and intensity [Hanks et al., 2015].

The analysis of circular data has seen major development since the 1970s. Early summarizing work include the books by Mardia [1972], Batschelet [1981], and Fisher [1993]. A more recent review was provided by Lee [2010], in addition to a comprehensive bibliography on the subject. Ravindran and Ghosh [2011] proposed to model circular data taking a Bayesian approach, wrapping probability distributions defined on the real line to obtain circular distributions. For cylindrical data, the literature revolves mainly around conditional modelling, 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 from a trivariate Gaussian distribution. The former distribution was improved by Abe and Ley [2017] to increase its flexibility. They invoked a power transformation to the linear part and a perturbation to the circular part to allow for asymmetric "sine-skewing". Recently, Tomoaki et al. [2019] also proposed a way to model cylindrical data with heavy-tailed linear parts. Circular and cylindrical data are two forms of

[2]More information about SINMOD is available on the web page: `https://www.sintef.no/en/ocean/initiatives/sinmod/`

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 in this field.

Data from the GlobCurrent project and the SINMOD model are measured across space and decomposed into a Cartesian representation of u- and v-components of the currents at each observation point of the spatial lattice. In this representation, u corresponds to the current speed in west–east direction, whereas v corresponds to the current speed in south–north direction. Both components are measured in metres per second, and can be transformed to a spatial cylindrical data series by the usual polar transform,

$$x = \sqrt{u^2 + v^2}, \tag{1.2}$$

$$\phi = \arctan2(v, u), \tag{1.3}$$

where $\arctan2(\cdot, \cdot)$ is the usual 2-argument inverse tangent function for determining angles in the Euclidean plane. In this way, we get cylindrical data, with the current speed $x \in [0, \infty)$ measured in metres per second and the direction of the current $\phi \in (-\pi, \pi]$ measured in radians. The direction is anti-clockwise from east ($\phi = 0$) to north ($\phi = \pi/2$).

The reason why we use a cylindrical representation for the OSC data instead of a Cartesian representation is because of the complications that appear with the latter. The correlation between the u and v components varies over the spatial domain [Reich and Fuentes, 2007]. It is therefore beneficial to consider the cylindrical representation, as we are able to model circular-linear dependence. However, challenges may also arise with this representation because of the special topology of the cylindrical support for the data. Consequently, cylindrical models need to be able to account for skewness, multimodality, and asymmetry in the marginal distributions.

Approaches towards modelling spatial cylindrical data have been limited. Even the field of spatial circular data is relatively unexplored, with Morphet [2009] developing methods for Kriging as well as simulation of circular spatial data. Wang and Gelfand [2014] proposed a Bayesian setting to model spatial dependence of circular data in the form of ocean wave directions. They utilized the projected Gaussian process to model stochastic circular variables on a continuous spatial domain. This enables them to model asymmetric data and account for spatio-temporal dependencies. This is advantageous compared to the wrapped Gaussian process proposed by Jona-Lasinio et al. [2012], which only models symmetric spatially dependent circular data.

The projected Gaussian process was further extended by Wang et al. [2015] to cylindrical data by also including wave height, resulting in a framework for joint modelling of the circular and linear component. 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 Bayesian 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. Here, the spatial dependency is defined on a spatial lattice, contrary to a continuous spatial domain in Wang et al. [2015]. The spatial lattice is suitable for applications with limited spatial resolution, as in our case.

1.4 Problem description

The incentives for studying and understanding ocean currents obviously has its roots in long-established enterprises such as shipping and fishing. However, the importance of ocean currents is far-reaching and does not only concern maritime industries. Surface currents impact marine life and define migration of e.g., larvae and fish eggs, sediment transport, and dispersal of pollution. Hence, knowledge of ocean circulation and drift is essential for maintaining the marine environment and operating marine businesses. Also, search and rescue missions and recreational activities depend on surface current patterns. Classifying and understanding these patterns is therefore of utmost importance.

In this thesis, we intend to provide a way to model the OSC in the Norwegian Sea to increase the understanding of current circulation patterns in the area. We propose a way to segment the global pattern into a small number of specific local regimes, which are easier to interpret than the global circulation. This is done by employing a Markov random field (MRF) to segment the spatial domain into latent, or hidden classes. This means that the conditional distribution of the latent classes only depend on their neighbours and they possess the Markov property. Each latent class supplies a corresponding set of fixed parameters to the cylindrical densities for the observations. In this way, we get a cylindrical hidden Markov random field (HMRF). The cylindrical HMRF provides a way of classifying typical *states of the ocean*, represented by the cylindrical densities, based on observations from different dates.

The observation sites are assigned a probability of attaining each latent class. Essentially, this means that the distribution of the OSC observations is a mixture of the cylindrical densities corresponding to the latent classes. This provides flexible mixture distributions that allow for multimodality. The mixture weights vary according to the evolution of a latent process across space, which captures spatial heterogeneity and correlation.

A key property of the models we develop is the need for compactness, i.e., compact models that are sufficiently simple to be incorporated on-board autonomous marine agents such as autonomous underwater vehicles (AUVs) and drifters. These vessels are unable to run, for instance, full-scale differential equation systems in the form of a numerical ocean model. Instead, we want to develop models that can be largely computed on shore and then be updated with information captured on the go. This means that the compact model is static, and the fixed parameters are not modified during autonomous missions. Methodologically this drives requirement for using a frequentist mind-set, with fixed parameters, as opposed to the more computationally intensive Bayesian approach that assigns probability distributions to the parameters.

The research that is closest to this thesis includes work by Holzmann et al. [2006] and Bulla et al. [2012] on methods to model cylindrical time-series data with a hidden Markov approach. Sea current circulation, which are spatial series of cylindrical data similarly to our OSC data, have been modelled by Lagona and Picone [2016] using a cylindrical HMRF. They developed a computationally intensive expectation-maximization algorithm by utilizing a mean-field approximation of the likelihood function. However, the method is numerically unstable and was improved by Ranalli et al. [2018], who instead of considering the full likelihood, took a composite-likelihood approach, resulting in a computationally efficient and more stable algorithm. We intend to further improve their algorithm in this thesis. Segmentation of OSC data based on sea surface temperature images was also

studied by Tandeo et al. [2013], who proposed a model to discriminate between hidden spatio-temporal classes of currents at the ocean surface. The latent classes were predicted and tracked by using local satellite observations as well as patches of surface temperature measurements.

1.5 Outline

The thesis is organized as follows:

- Chapter 2 presents theory behind MRFs and the Potts model. Two cylindrical probability distributions are also introduced and discussed before the combined cylindrical HMRF model is defined.
- Chapter 3 discusses techniques for doing inference on the HMRF model, including parameter estimation and model selection.
- Chapter 4 presents a simulation study to compare inference models and investigate accuracy in parameter estimation.
- Chapter 5 provides results and discussion from fitting the HMRF model to real data sets of OSC observations from the GlobCurrent project.
- Chapter 6 provides results and discussion from fitting the HMRF model to simulated OSC data from the SINMOD model.
- Chapter 7 gives concluding remarks and discusses possible extensions and topics for further work.

Chapter 2

Models

In this chapter we present a cylindrical HMRF, where the classes determining the parameters of the cylindrical distribution are considered unknown or hidden. In Section 2.1, we first give a general introduction to MRFs and specify the Potts model that determines the underlying spatial classes. Then, Section 2.2 introduces two cylindrical distributions and presents the probability density functions (pdf) in contention. Finally, in Section 2.3 we describe how the models can be combined to form a cylindrical HMRF.

2.1 Markov Random Fields

We denote by $\mathbf{s}$ a reference variable in continuous two dimensional space, $\mathbf{s} \in \mathcal{D} \subset \mathbb{R}^2$. A spatial variable, $l(\mathbf{s})$, with reference $\mathbf{s}$ can then be interpreted probabilistically by defining a random field (RF). First, we define the pdf of the spatial variable to be

$$(l(\mathbf{s}_1), l(\mathbf{s}_2), \ldots, l(\mathbf{s}_n)) \sim p(l(\mathbf{s}_1), l(\mathbf{s}_2), \ldots, l(\mathbf{s}_n)), \tag{2.1}$$

for all possible $(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n) \in \mathcal{D}$ and all $n \in \mathbb{N}_+$. Then $l(\mathbf{s})$ is an RF if its values are random for every $\mathbf{s} \in \mathcal{D} \subset \mathbb{R}^2$ [Abrahamsen, 1997]. The RF is called stationary if the pdf $p(l(\mathbf{s}_1), l(\mathbf{s}_2), \ldots, l(\mathbf{s}_n))$ is invariant to shift in the reference $\mathbf{s}$.

The reference variable can also be discretized into $\mathbf{s} \in \mathcal{L}$, where $\mathcal{L}$ is a lattice of grid points, or cells, in two-dimensional space, consisting of $|\mathcal{L}| = n$ grid points. Going forward, we simplify notation by denoting the vector $\mathbf{l} = (l(\mathbf{s}_1), l(\mathbf{s}_2), \ldots, l(\mathbf{s}_n))$. Also, we denote the grid point $\mathbf{s}_i \in \mathcal{L}$ by i, and the spatial variable at grid point i is denoted l_i.

The spatial variable l_i is called a mosaic spatial variable if it takes discrete values, i.e., $l_i \in \mathbb{L} = \{1, \ldots, K\}$. Here, K is the number of possible discrete values, or classes that the variable can take. When the spatial variable takes discrete values, its distribution is a probability mass function (pmf) instead of a pdf, which gives the probability that the spatial variable is exactly equal to some value. We use the following notation for the pmf of the spatial variable $\mathbf{l}$,

$$\mathbf{l} \sim \Pr(\mathbf{l} = \mathbf{l}') = p(\mathbf{l}), \tag{2.2}$$

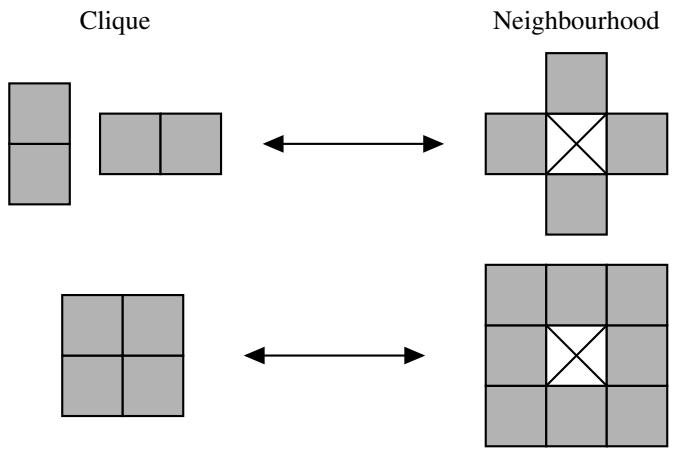


Figure 2.1: Two examples of neighbourhood systems to the node marked with a cross (right) with the corresponding clique system (left).

where $\Pr(\mathbf{l} = \mathbf{l}')$ is the probability that the spatial variable $\mathbf{l}$ is exactly equal to $\mathbf{l}'$, i.e., the pmf of $\mathbf{l}$.

MRFs have been extensively studied since the seminal papers by Besag [1974, 1975]. General introductions to MRFs are provided by, e.g., Cressie [1993] and Guyon [1995]. To define the MRF, we follow the lines of Besag [1974] by first defining a neighbourhood system. A set $\mathbf{N}_{\mathcal{L}} : \{\mathbf{N}_1, \ldots, \mathbf{N}_n\}$ is a neighbourhood 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 $i, j \in \mathcal{L}$. Such pairs are called neighbours. Having defined a neighbourhood system, we can define the corresponding cliques. A set $\mathbf{c} \subseteq \mathcal{L}$ is called a clique if $i \in \mathbf{N}_j$ for all pairs $i, j \in \mathbf{c}$. We then denote by $\mathbf{c}_{\mathcal{L}} : \{\mathbf{c}_1, \ldots, \mathbf{c}_{n_c}\}$ the set of all maximal cliques on the lattice, i.e., all cliques that cannot be extended by including another point. Two examples of neighbourhood systems and the corresponding clique system is displayed in Figure 2.1.

Having defined cliques and neighbourhoods, we can now define the MRF. If

$$\mathbf{l} \sim p(\mathbf{l}) = C \times \prod_{i=1}^{n} v_{0i}(l_i) \times \prod_{\mathbf{c} \in \mathbf{c}_{\mathcal{L}}} v_{1\mathbf{c}}(l_j; j \in \mathbf{c}), \tag{2.3}$$

with $v_{0i}(\cdot), v_{1\mathbf{c}}(\cdot) \in \mathbb{R}_{\oplus}$, and C being a normalizing constant, then $\{l(\mathbf{s}); \mathbf{s} \in \mathcal{L}\}$ is the Gibbs formulation of a MRF on the grid $\mathcal{L}$ with clique system $\mathbf{c}_{\mathcal{L}}$. The function $v_{0i}(\cdot)$ is a single-cell potential function, and $v_{1\mathbf{c}}(\cdot)$ represents the clique potential function, or spatial coupling between between cells in a clique. By the Hammersley-Clifford theorem [Hammersley and Clifford, 1971], there exists a corresponding neighbourhood system $\mathbf{N}_{\mathcal{L}}$, such that for all $\mathbf{s}_i \in \mathcal{L}$ we have

$$[l_i | \mathbf{l}_{-i}] \sim p(l_i | \mathbf{l}_{-i}) = p(l_i | l_j; j \in \mathbf{N}_i). \tag{2.4}$$

This is called the Markov formulation of the MRF on the grid $\mathcal{L}$ with neighbourhood system $\mathbf{N}_{\mathcal{L}}$. Note here that the Gibbs formulation in Equation (2.3) consists of one n-

dimensional pmf, whereas the Markov formulation in Equation (2.4) consists of n univariate conditional pmfs. The normalizing constant C is cumbersome to calculate for the Gibbs formulation, as one needs to sum over K^n terms. This can only be done exactly for spatial grids of limited size, and special algorithms are necessary [Friel and Rue, 2007]. However, it is tractable for the Markov formulation, for which the pmf is given by

$$\begin{aligned}[l_i|\mathbf{l}_{-i}] \sim p(l_i|\mathbf{l}_{-i}) &= \left[\sum_{l_i' \in \mathbb{L}} p(l_i', \mathbf{l}_{-i})\right]^{-1} p(\mathbf{l}) = p(l_i|l_j; j \in \mathbf{N}_i) \\ &= \left[\sum_{l_i' \in \mathbb{L}} v_{0i}(l_i') w_i(l_i'|l_j; j \in \mathbf{N}_i)\right]^{-1} \times v_{0i}(l_i) w_i(l_i|l_j; j \in \mathbf{N}_i),\end{aligned} \tag{2.5}$$

where $v_{0i}(\cdot)$ is the same single-cell potential function as for the Gibbs formulation in Equation (2.3). The interaction functions $w_i(\cdot|\cdot) \in \mathbb{R}_\oplus$ are related to the clique potential functions by the formula

$$w_i(l_i|l_j; j \in \mathbf{N}_i) \propto \prod_{\mathbf{c} \in \mathbf{c}_{L|i}} v_{1\mathbf{c}}(l_j; j \in \mathbf{c}), \tag{2.6}$$

where $\mathbf{c}_{L|i}$ denotes all cliques containing grid point i.

2.1.1 The Potts model

We will next present a special form of a MRF with its particular clique potential functions, known as the Potts model. The Potts model was introduced in a Ph.D. thesis by Potts [1951], but for a tidy tutorial review we refer to Wu [1982]. The clique system for the Potts model is defined as all two closest pairs, and is denoted by $\mathbf{c}_\mathcal{L}$. This means that the cliques and neighbourhoods correspond to the structure at the top of Figure 2.1. Moreover, the clique coupling functions are

$$v_{1\mathbf{c}}(l_i, l_j; \mathbf{c}) = \exp\big(\rho I(l_i = l_j)\big), \tag{2.7}$$

where $\mathbf{c} \in \mathbf{c}_\mathcal{L}$ represents any clique from the clique system, $\rho \in \mathbb{R}_\oplus$ is the coupling parameter, and $I(\cdot)$ 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, hence it assumes symmetry in all classes and the model is indifferent to the numbering of classes. Because there is only one coupling, or dependence parameter ρ, the model also assumes equal dependence in both spatial directions. In addition, all the single-site potential functions $v_{0i}(\cdot)$ are assumed constant and equal. Consequently, the Potts model is a stationary model, except for boundary effects on the coupling. In the special case of $K = 2$, i.e., only two spatial classes, the Potts model is identical to the Ising model, which is prominent in several applications, including neuroscience [Schneidman et al., 2006] and biophysics [Liu and Dilger, 1993]. It is also one of the most studied models in modern physics [Niss, 2005].

The Gibbs formulation of the Potts model is given by

$$\mathbf{l} \sim p(\mathbf{l}) = C(\rho)^{-1} \exp\Big(\rho \sum_{\mathbf{c} \in \mathbf{c}_\mathcal{L}} I(l_i = l_j)\Big), \tag{2.8}$$

where $C(\rho)$ is a normalizing constant, depending on ρ, given by

$$C(\rho) = \sum_{\mathbf{l}' \in \mathbb{L}^n} \exp\Big(\rho \sum_{\mathbf{c} \in \mathbf{c}_{\mathcal{L}}} I(l'_i = l'_j)\Big). \tag{2.9}$$

This implies that the normalizing constant is a sum over all K possible outcomes for each entry in the n-vector $\mathbf{l}$. This is obviously not feasible to compute for large spatial lattices and a large number of classes. Hence, exact computation of the normalizing constant is only possible for small grids and requires special algorithms, see e.g., Bartolucci and Besag [2002], Reeves and Pettitt [2004], or Friel and Rue [2007]. For large grids, the normalizing constant needs to be approximated.

The Markov formulation of the Potts model is

$$[l_i | l_j; j \in \mathbf{N}_i] \sim p(l_i | l_j; j \in \mathbf{N}_i) = \frac{\exp\left(\rho \sum_{j \in \mathbf{N}_i} I(l_i = l_j)\right)}{\sum_{l'_i \in \mathbb{L}} \exp\left(\rho \sum_{j \in \mathbf{N}_i} I(l'_i = l_j)\right)}, \tag{2.10}$$

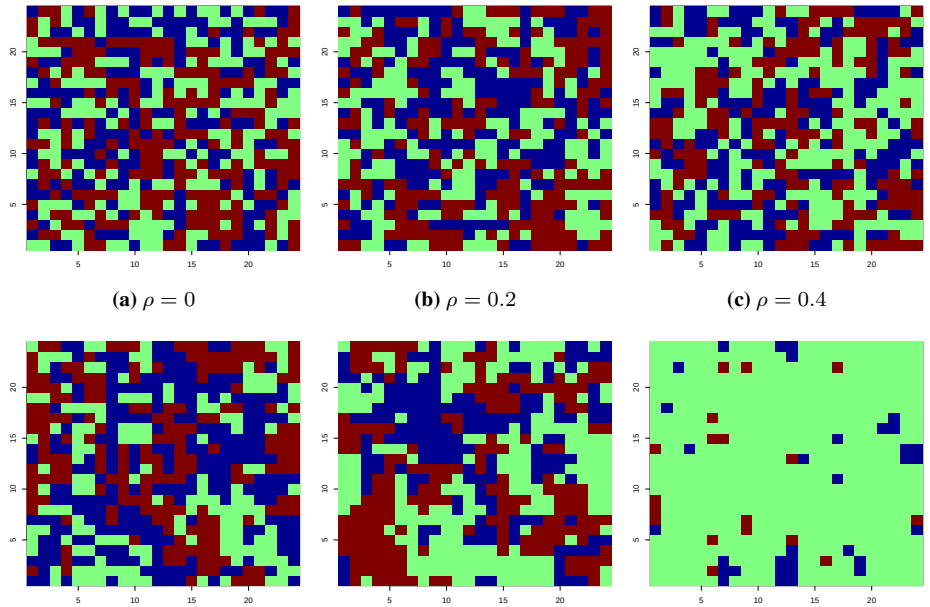
for all $i \in \{1, \ldots, n\}$. The Potts model has a neighbourhood system that includes the four closest nodes. For this formulation the normalizing constant is tractable, as the sum is only over K components.

One disadvantage of the Potts model is that it behaves very differently for varying values of ρ. For values of ρ above $\rho_{\text{crit}} \approx \ln(1 + \sqrt{K})$, there is a phase transition and almost all the values of the spatial variable $\mathbf{l}$ become 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. For values below the critical value, the field then appears with a mixing of the classes, gradually with less spatial coupling as ρ goes to 0. However, the most characteristic feature of the Potts model is the transition between the two regimes at the critical value ρ_{crit}. For our application, this implies that we need to impose the restriction $\rho \in [0, \rho_{\text{crit}})$, because we do not want all the grid cells to take the same class. As a final note, with $\rho = 0$ the Potts model simplifies to a spatially independent model. Then, all grid points are independently multinomial with equal probability for each class.

Figure 2.2 shows random samples from the Potts model with different values for the interaction parameter ρ on a 24×24 lattice with $K = 3$ classes. The samples are generated using the Swendsen–Wang algorithm [Swendsen and Wang, 1987], which samples from the Potts model using McMC. The algorithm is available through the `potts` package in R. From the image with $\rho = 0$, we observe that all points are essentially independent. Notice also that the grid points display increasing spatial dependence and larger clusters of equal classes form for larger values of the interaction parameter ρ. For values of ρ close to and above the critical value, almost all grid points take the same value.

2.2 Cylindrical probability distributions

Cylindrical data are on the form $\mathbf{z} = (x, \phi)$, where $x \in [0, \infty)$ represents magnitude and $\phi \in (-\pi, \pi]$ represents angle. The magnitude is also called the linear part of the distribution, whereas the angle is called the circular part. Hence, the cylinder is a combination of a linear and a circular part. Examples of applications for cylindrical data include ocean



(a) $\rho = 0$ **(b)** $\rho = 0.2$ **(c)** $\rho = 0.4$

(d) $\rho = 0.6$ **(e)** $\rho = 0.8$ **(f)** $\rho = \rho_{\text{crit}}$

Figure 2.2: Samples from the Potts model over a 24×24 grid with $K = 3$ classes for different values of the interaction parameter ρ.

wave height and direction, wind speed and direction, or in our case, OSC speed and direction. In this section, we present and discuss two cylindrical densities we will use to model cylindrical data.

2.2.1 Weibull sine-skewed von Mises

We first consider the Weibull sine-skewed von Mises density (WSSVM) proposed by Abe and Ley [2017]. The distribution combines a Weibull density for the magnitude and a circular sine-skewed von Mises distribution for the angle. The pdf is given by

$$p(x, \phi) = \frac{\alpha\beta^{\alpha}}{2\pi \cosh \kappa} \big(1 + \lambda \sin(\phi - \mu)\big) x^{\alpha-1} \exp\Big(- (\beta x)^{\alpha} \big(1 - \tanh(\kappa) \cos(\phi - \mu)\big)\Big). \tag{2.11}$$

The interpretation of the parameters is as follows: $\alpha > 0$ 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.

A prominent property of the 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 makes for simple sampling routines. The marginal distribution of the angle ϕ is a sine-skewed

wrapped Cauchy distribution with location μ and concentration $\tanh(\kappa/2)$, with pdf,

$$p(\phi) = \frac{1-\tanh^2(\kappa/2)}{2\pi} \frac{1+\lambda\sin(\phi-\mu)}{1+\tanh^2(\kappa/2)-2\tanh(\kappa/2)\cos(\phi-\mu)}. \tag{2.12}$$

In turn, the linear component x has a marginal distribution with pdf,

$$p(x) = \frac{I_0\big((\beta x)^\alpha \tanh(\kappa)\big)}{\cosh(\kappa)} \alpha\beta^\alpha x^{\alpha-1} \exp\big(-(\beta x)^\alpha\big), \tag{2.13}$$

where $I_0(\cdot)$ is the modified Bessel function of the first kind and order zero.

For $\kappa = 0$, there is independence between the circular and linear part. Then, the joint distribution simplifies to the product between a linear Weibull distribution with shape α and scale β, given by

$$p(x) = \alpha\beta^\alpha x^{\alpha-1} \exp\big(-(\beta x)^\alpha\big), \tag{2.14}$$

and a circular cardioid distribution with location $\mu + \pi/2$ and concentration λ, given by

$$p(\phi) = \frac{1}{2\pi}\big(1+\lambda\sin(\phi-\mu)\big). \tag{2.15}$$

Moreover, the conditional distribution of the magnitude x given the angle ϕ is a Weibull distribution as in Equation (2.14) with shape α and scale $\beta(1-\tanh(\kappa)\cos(\phi-\mu))^{1/\alpha}$, or,

$$p(x|\phi) = \alpha\beta^\alpha(1-\tanh(\kappa)\cos(\phi-\mu))x^{\alpha-1} \exp\big(-(1-\tanh(\kappa)\cos(\phi-\mu))(\beta x)^\alpha\big). \tag{2.16}$$

The conditional distribution of the circular part ϕ given the linear part x has a sine-skewed von Mises distribution with concentration $(\beta x)^\alpha \tanh(\kappa)$ and pdf,

$$p(\phi|x) = \frac{1+\lambda\sin(\phi-\mu)}{2\pi I_0\big((\beta x)^\alpha\tanh(\kappa)\big)} \exp\big((\beta x)^\alpha \tanh(\kappa)\cos(\phi-\mu)\big). \tag{2.17}$$

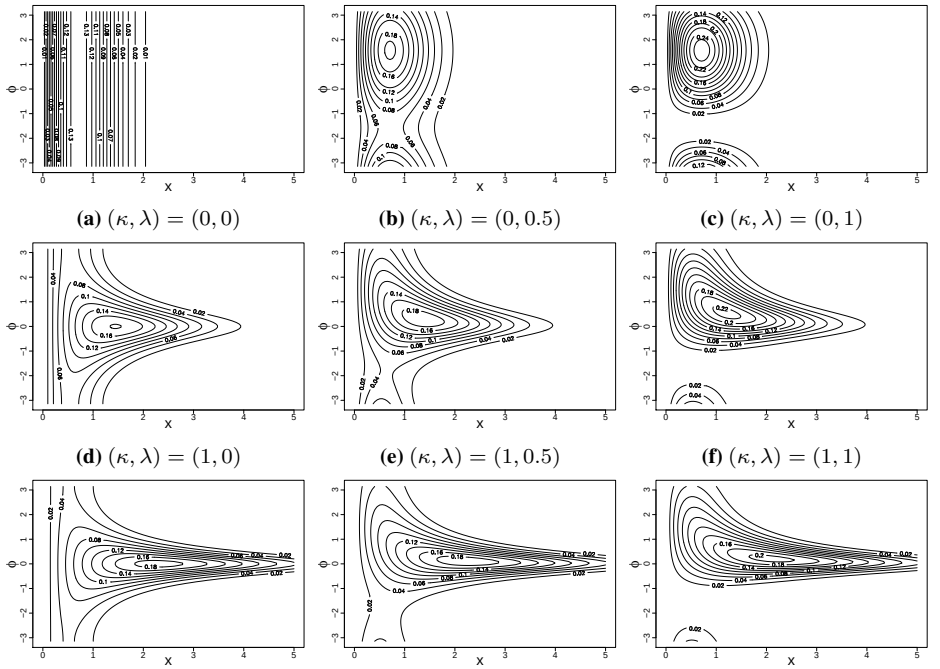
These distributions can be utilized to define a simple sampling routine. Sampling is conducted by first sampling from the marginal distribution of ϕ given by Equation (2.12) and then the conditional of x given ϕ as in Equation (2.16). Abe and Ley [2017] proposed the following algorithm to generate a sample (X, Φ) from the WSSVM distribution:

1. Generate a random variable Φ_1 from a wrapped Cauchy distribution with location μ and concentration $\tanh(\kappa/2)$.

2. Draw $U \sim \text{Uniform}[0,1]$, and let

$$\Phi = \begin{cases} \Phi_1, & \text{if } U < \big(1+\lambda\sin(\Phi_1-\mu)\big)/2, \\ -\Phi_1, & \text{if } U \geq \big(1+\lambda\sin(\Phi_1-\mu)\big)/2. \end{cases}$$

   Then, Φ has a sine-skewed wrapped Cauchy distribution with location μ and concentration $\tanh(\kappa/2)$.



(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)** $(\kappa, \lambda) = (1, 1)$

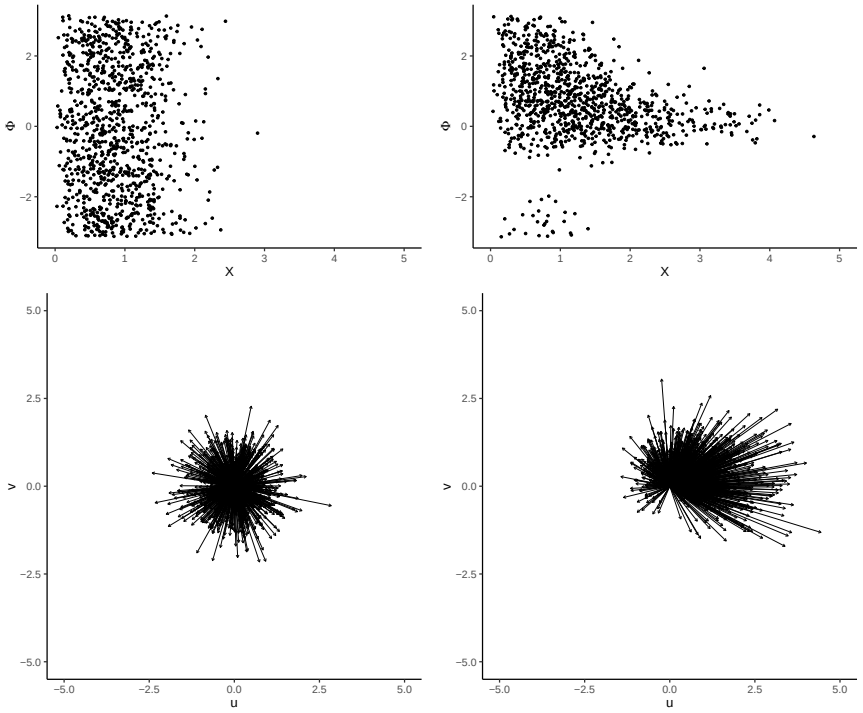
(g) $(\kappa, \lambda) = (1.5, 0)$ **(h)** $(\kappa, \lambda) = (1.5, 0.5)$ **(i)** $(\kappa, \lambda) = (1.5, 1)$

Figure 2.3: Contour plots of the WSSVM density given in Equation (2.11) for $(\alpha, \beta, \mu) = (2, 1, 0)$ and (κ, λ) as indicated.

3. Generate X from from a Weibull distribution with shape α and scale $\beta\big(1 - \tanh(\kappa)\cos(\Phi - \mu)\big)^{1/\alpha}$.

In Figure 2.3 we display the WSSVM density for some values of (κ, λ) and with $(\alpha, \beta, \mu) = (2, 1, 0)$. The densities, and all future plots in this thesis of cylindrical densities or data, should be interpreted bearing in mind that they are actually on a cylinder, so the variable 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 part of the distribution, and the skewness of the distribution achieved by increasing λ. The distribution is of course skewed towards negative values of ϕ when λ is negative. Figure 2.3a shows that the circular part is uniform for $\kappa = \lambda = 0$. This is also observed from the marginal distribution of the angle in Equation (2.12), the distribution becomes uniform when $\kappa = \lambda = 0$. As either κ or λ increases, the circular part becomes less uniform.

To make future discussions easier to follow and to familiarize the structure of the WSSVM model and cylindrical data, we present some additional visualizations of data from this model and discuss some critical properties of the density. First, in Figure 2.4 we have drawn 1000 samples from the WSSVM distribution with $(\alpha, \beta, \mu) = (2, 1, 0)$ for two different values of (κ, λ). At the top of the figure, these samples are plotted with the magnitude X on the first axis and the angle Φ on the second axis, as was done for the



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

Figure 2.4: Samples from the WSSVM density for $(\alpha, \beta, \mu) = (2, 1, 0)$ and (κ, λ) as indicated.

densities in Figure 2.3. At the bottom of the display, the same samples are transformed to Cartesian coordinates and plotted as vectors originating from the origin. In the left display, all angles are equally likely, because the circular part is uniform. Smaller angles, $\Phi \in (-2, -1.5)$, are much less likely in the right display. Displays like these may help familiarize the cylindrical data structure.

All moments of the linear part are explicitly defined and the qth moment is available through the expression

$$E(X^q) = \left(\frac{\cosh(\kappa)}{\beta^\alpha}\right)^{q/\alpha} \Gamma\Big(\frac{q+\alpha}{\alpha}\Big) P^0_{q/\alpha}\big(\cosh(\kappa)\big), \tag{2.18}$$

where $\Gamma(\cdot)$ is the Gamma function and $P^0_{q/\alpha}(\cdot)$ is the associated Legendre function of the first kind of order 0 and degree q/α [Abe and Ley, 2017]. Note that all linear moments are independent of the circular location μ and the skewness λ. Both the expectation and variance of the magnitude are easily computable from this expression. For example, the analytical expression for the expectation of the magnitude in the left display of Figure 2.4 is $\sqrt{\pi}/2 \approx 0.89$, and for the right display it is approximately 1.31. Also the theoretical variance of the linear part is $1 - \pi/4 \approx 0.21$ for the sample to the left, and 0.67 for the sample to the right, verifying that the right sample has a higher observed variance in the

linear part.

The circular part of the distribution is obtained by wrapping a sine-skewed Cauchy distribution onto the unit circle. Because the distribution is wrapped, it is not appropriate to define linear moments, as this ignores the similarity of angles near $-\pi$ and near π. Hence, we consider trigonometric moments, and these are available and can be derived from the original distribution defined on the line [Abe and Pewsey, 2011]. The trigonometric expectations and variances are given by

$$E(\cos(\Phi)) = \tanh\left(\frac{\kappa}{2}\right), \quad E(\sin(\Phi)) = \frac{\lambda}{2\cosh^2(\frac{\kappa}{2})}, \tag{2.19}$$

$$\text{Var}(\cos(\Phi)) = \frac{1}{2\cosh^2(\frac{\kappa}{2})}, \quad \text{Var}(\sin(\Phi)) = \frac{1}{2\cosh^2(\frac{\kappa}{2})}\left(1 - \frac{\lambda^2}{2\cosh^2(\frac{\kappa}{2})}\right). \tag{2.20}$$

Expressions for the covariance between the linear part and trigonometric circular part are also explicitly given by,

$$\text{Cov}(X, \cos(\Phi)) = \frac{\Gamma(\frac{1}{\alpha})P^1_{1/\alpha}(\cosh(\kappa)) - \Gamma(\frac{1+\alpha}{\alpha})P^0_{1/\alpha}(\cosh(\kappa))\tanh(\frac{\kappa}{2})}{(\cosh(\kappa))^{-1/\alpha}\beta}, \tag{2.21}$$

$$\text{Cov}(X, \sin(\Phi)) = \frac{\Gamma(\frac{1+\alpha}{\alpha})P^0_{1/\alpha}(\cosh(\kappa))\tanh^2(\frac{\kappa}{2}) - \Gamma(\frac{1-\alpha}{\alpha})P^2_{1/\alpha}(\cosh(\kappa))}{2\lambda^{-1}(\cosh(\kappa))^{-1/\alpha}\beta}. \tag{2.22}$$

Notably, for $\kappa = 0$ both these covariance expressions equal zero. This is as expected because $\kappa = 0$ yields independence between the circular and linear parts. The correlation between the two parts also has an explicit expression, for which we refer to Abe and Ley [2017] for a detailed derivation. Essentially, they state that the correlation does not depend on β, μ, or the sign of λ. Moreover, the correlation increases as α or $|\lambda|$ increases. For small values of κ, the correlation increases with κ until it reaches its maximum, and then decreases with κ.

2.2.2 Generalized Pareto-type wrapped Cauchy

We also consider a distribution that consists of a generalized Pareto-type distribution for the linear part and a wrapped Cauchy distribution for the circular part (GPTWC). As opposed to the previous one, this distribution is not skewed, but it has heavy tails in the linear part, and was proposed by Tomoaki et al. [2019]. The joint pdf of the magnitude and angle is given by

$$p(x, \phi) = \frac{\sqrt{1-\kappa^2}}{2\pi\beta\alpha}\left(\frac{x}{\beta}\right)^{1/\alpha-1}\left(1 + \frac{\tau}{\alpha}\left(\frac{x}{\beta}\right)^{1/\alpha}\big(1 - \kappa\cos(\phi - \mu)\big)\right)^{-(\alpha/\tau+1)}. \tag{2.23}$$

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 behaviour of the linear part. As a limiting case of this distribution, when $\tau \to 0$, the distribution in Equation (2.23) reduces to the

distribution

$$p(x,\phi) = \frac{\sqrt{1-\kappa^2}}{2\pi\beta\alpha}\left(\frac{x}{\beta}\right)^{1/\alpha-1} \exp\left(-\left(\frac{x}{\beta}\right)^{1/\alpha}\big(1-\kappa\cos(\phi-\mu)\big)\right), \tag{2.24}$$

which is actually the same density as the WSSVM density in Equation (2.11) with $\lambda = 0$ and a reparametrization of α, β, and κ.

Crucially, the normalizing constant of the GPTWC distribution is explicitly defined, as well as all marginal and conditional distributions. The marginal distribution of the circular part ϕ is a standard wrapped Cauchy distribution with location μ and concentration $\kappa/(1+\sqrt{1-\kappa^2})$, with pdf,

$$p(\phi) = \frac{\sqrt{1-\kappa^2}}{2\pi\big(1-\kappa\cos(\phi-\mu)\big)}. \tag{2.25}$$

This is the same distribution as the marginal distribution for the angle in the WSSVM distribution, except that it is not skewed. By letting $\lambda = 0$ and reparametrizing κ in the expression for the marginal distribution of the angle in the WSSVM distribution in Equation (2.12), we arrive at the expression for the marginal of the GPTWC distribution in Equation (2.25). The marginal pdf of the linear component x is expressed by,

$$\begin{aligned} p(x) =& \frac{\sqrt{1-\kappa^2}}{\alpha\beta}\left(\frac{x}{\beta}\right)^{1/\alpha-1}\left(1+\frac{\tau}{\alpha}\left(\frac{x}{\beta}\right)^{1/\alpha}\right)^{-(\alpha/\tau+1)} \\ & \times {}_2F_1\left(\frac{\alpha/\tau+1}{2}, \frac{\alpha/\tau+2}{2}; 1; \left(\frac{\kappa(x/\beta)^{1/\alpha}}{\alpha/\tau+(x/\beta)^{1/\alpha}}\right)^2\right), \end{aligned} \tag{2.26}$$

where ${}_2F_1(\cdot,\cdot;\cdot;\cdot)$ is the Gauss hypergeometric function [Tomoaki et al., 2019].

For $\kappa = 0$, the circular and linear parts are independent, as was the case for the WSSVM distribution. Then, the joint distribution simplifies to the product between a circular uniform distribution for the angle ϕ and a generalized Pareto-type distribution for the linear variable x with pdf,

$$p(x) = \frac{1}{\beta\alpha}\left(\frac{x}{\beta}\right)^{1/\alpha-1}\left(1+\frac{\tau}{\alpha}\left(\frac{x}{\beta}\right)^{1/\alpha}\right)^{-(\alpha/\tau+1)}. \tag{2.27}$$

In the limiting case of $\tau \to 0$, the marginal distribution of the linear part x in Equation (2.26) simplifies to,

$$p(x) \to \frac{\sqrt{1-\kappa^2}}{\alpha\beta}\left(\frac{x}{\beta}\right)^{1/\alpha-1} \exp\big(-(x/\beta)^{1/\alpha}\big) I_0\big(\kappa(x/\beta)^{1/\alpha}\big), \tag{2.28}$$

which is the same distribution as the marginal of x for the WSSVM distribution in Equation (2.13), with a reparametrization of α, β, and κ.

The conditional distribution of ϕ given x is listed by Tomoaki et al. [2019], and we only write the conditional distribution of the magnitude x given the angle ϕ. This distribution

corresponds to a generalized Pareto-type distribution with density,

$$p(x|\phi) = \left(\frac{x}{\beta}\right)^{1/\alpha-1} \frac{1-\kappa\cos(\phi-\mu)}{\alpha\beta} \left(1+\frac{\tau}{\alpha}\left(\frac{x}{\beta}\right)^{1/\alpha}\big(1-\kappa\cos(\phi-\mu)\big)\right)^{-(\alpha/\tau+1)}. \tag{2.29}$$

Conveniently, its cumulative distribution function,

$$F_{x|\phi}(x|\phi) = 1 - \left(1+\frac{\tau}{\alpha}\left(\frac{x}{\beta}\right)^{1/\alpha}\big(1-\kappa\cos(\phi-\mu)\big)\right)^{-\alpha/\tau} \tag{2.30}$$

has an inverse on closed form. This gives a simple procedure to generate a sample (X, Φ) from the GPTWC distribution, as suggested by Tomoaki et al. [2019]:

1. Generate Φ from a wrapped Cauchy distribution with location μ and concentration $\kappa/(1+\sqrt{1-\kappa^2})$.
2. Draw $U \sim \text{Uniform}[0,1]$, and let

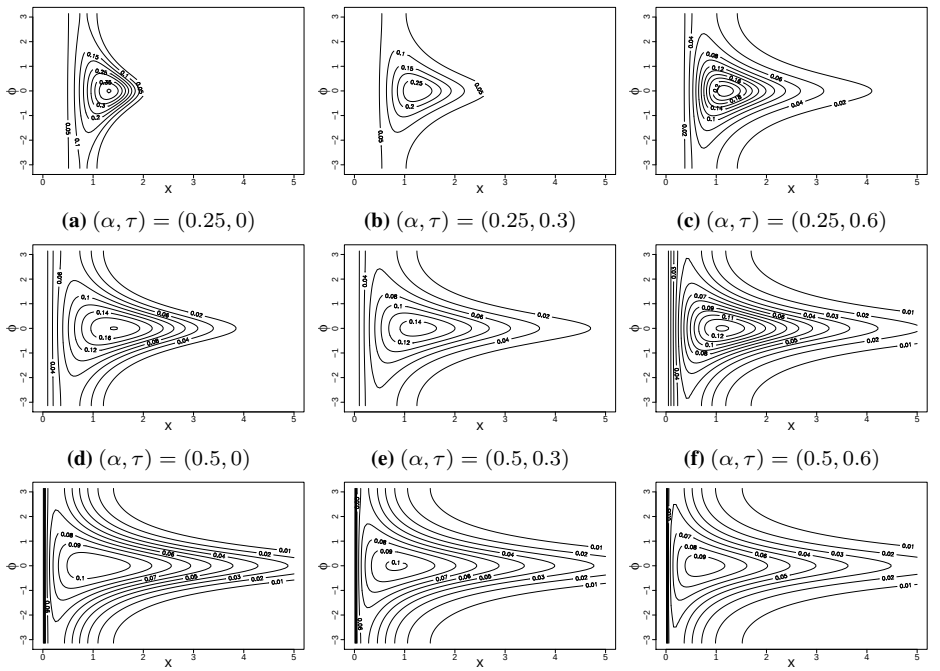
$$X = \begin{cases} \beta\Big(\frac{(1-U)^{-\tau/\alpha}-1}{\tau(1-\kappa\cos(\Phi-\mu))/\alpha}\Big)^{\alpha}, & \text{if } \tau > 0, \\ \beta\Big(-\frac{\log(1-U)}{1-\kappa\cos(\Phi-\mu)}\Big)^{\alpha}, & \text{if } \tau = 0. \end{cases}$$

Figure 2.5 shows the GPTWC density for some values of (α, τ) and with $(\beta, \mu, \kappa) = (1, 0, 0.75)$. From the displays it is clear that the tails become heavier as τ increases. By also varying β and κ, this distribution is flexible in both the circular and linear part, and able to fit both distributions that are concentrated in the circular part and distributions that are more evenly distributed across the circular variable. The parameter μ decides the circular location, allowing also for distributions that are not symmetric about $\phi = 0$. However, unlike the WSSVM distribution, this distribution is not able to model data with skewness in the circular part. Rather, this distribution is tailored to fit data with heavy tails in the linear part.

In Figure 2.6 we display samples similarly to what was done for the WSSVM density with parameters $(\beta, \mu, \kappa) = (1, 0, 0.75)$ and two different sets of values for (α, τ). In the left display there is no heavy tailedness, and there are no clear outliers. The right display, on the other hand, displays several outliers that come as a result of the heavy tails, even with a moderate value $\tau = 0.3$. The heavy tails cause some of the magnitudes to be very large, and to include all samples, the linear scale in this figure is twice as large as it was for Figure 2.4.

We summarize some of the most important properties for this distribution. For a more detailed discussion of its characteristics, the reader should consult Tomoaki et al. [2019]. Crucially, as the linear part of the distribution is heavy-tailed, not all linear moments are defined. The marginal linear moments $E(X^q)$ are only defined for $0 \le q < 1/\tau$. This implies that the expectation of the linear part only exists for $0 \le \tau < 1$, for which it is given by,

$$E(X) = \sqrt{1-\kappa^2}\beta\Big(\frac{\alpha}{\tau}\Big)^{\alpha} \frac{\Gamma(\alpha/\tau-\alpha)\Gamma(\alpha+1)}{\Gamma(\alpha/\tau)} \times {}_2F_1\left(\frac{\alpha+1}{2}, \frac{\alpha+2}{2}; 1; \kappa^2\right). \tag{2.31}$$

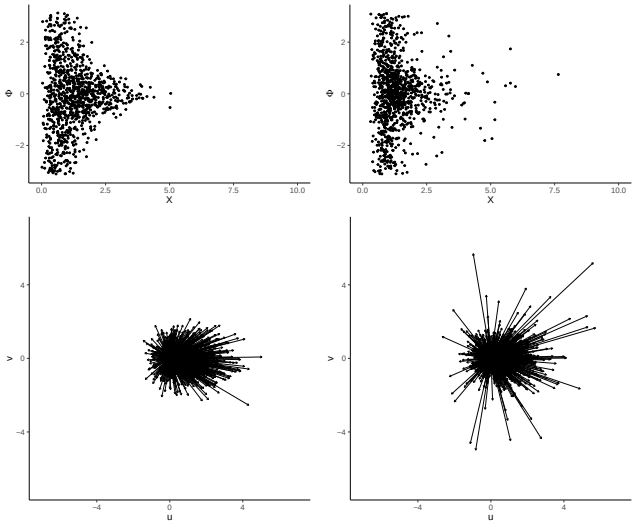


(a) $(\alpha, \tau) = (0.25, 0)$ **(b)** $(\alpha, \tau) = (0.25, 0.3)$ **(c)** $(\alpha, \tau) = (0.25, 0.6)$

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

Figure 2.5: Contour plots of the GPTWC density given in Equation (2.23) for $(\beta, \mu, \kappa) = (1, 0, 0.75)$ and (α, τ) as indicated.



(a) $(\alpha, \tau) = (0.5, 0)$ **(b)** $(\alpha, \tau) = (0.25, 0.3)$

Figure 2.6: Samples from the GPTWC density for $(\beta, \mu, \kappa) = (1, 0, 0.75)$ and (α, τ) as indicated.

Observe that this expression is independent of the circular location μ. Further, the variance only exists for $0 \leq \tau < 1/2$. This means that the correlation between angle and magnitude is only defined for $0 \leq \tau < 1/2$. Then, correlation decreases as α or τ increases and increases as κ increases. When the correlation does not exist, it is beneficial to consider the conditional mode and median of the linear part, given the circular part to represent the relation between the two. The conditional mode is given by,

$$\text{Mode}(X|\Phi = \phi) = \begin{cases} \beta\left(\frac{1-\alpha}{(1+\tau)(1-\kappa\cos(\phi-\mu))}\right)^{\alpha}, & \alpha < 1, \\ 0, & \alpha \geq 1, \end{cases} \tag{2.32}$$

and the conditional median by,

$$\text{Median}(X|\Phi = \phi) = \beta\left(\frac{\alpha}{\tau}\frac{2^{\tau/\alpha}-1}{1-\kappa\cos(\phi-\mu)}\right)^{\alpha}. \tag{2.33}$$

2.3 Cylindrical Hidden Markov Random Field

We now combine the MRF and cylindrical distributions into a joint model. The Potts model for the MRF has an interaction parameter ρ, which is assumed fixed but unknown. The classes of the MRF, l_i, $i \in \{1, \ldots, n\}$, are assumed hidden, or unknown, and they determine the parameters for the cylindrical distribution. Hence, this can be seen as a HMRF with cylindrical distributed observations. As such, we define $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, $\mathbf{z}_i = (x_i, \phi_i)$ 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 $\boldsymbol{\theta}_k$, $k \in \{1, \ldots, K\}$, are assumed fixed but unknown. If grid point i takes latent class k, or $l_i = k$, the observation $\mathbf{z}_i$ will have a cylindrical distribution with parameters $\boldsymbol{\theta}_k$. In this way, we write the conditional distribution of the observation $\mathbf{z}_i$ given the latent class l_i as,

$$p_{\boldsymbol{\theta}}(\mathbf{z}_i|l_i) = \prod_{k=1}^{K} p_{\boldsymbol{\theta}_k}(\mathbf{z}_i)^{I(l_i=k)}, \tag{2.34}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ is the vector of K different sets of cylindrical model parameters, and $p_{\boldsymbol{\theta}_k}(\mathbf{z}_i)$ is a cylindrical density from either Equation (2.11) or (2.23). The subscript notation for the parameters $\boldsymbol{\theta}_k$ indicates that they are fixed. Because the observations are assumed conditionally independent, the joint conditional distribution of all observations is,

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{l}) = \prod_{i=1}^{n}\prod_{k=1}^{K} p_{\boldsymbol{\theta}_k}(\mathbf{z}_i)^{I(l_i=k)}. \tag{2.35}$$

This gives a joint distribution for the observed data points and latent classes of

$$p_{\boldsymbol{\theta},\rho}(\mathbf{z}, \mathbf{l}) = p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{l})p_{\rho}(\mathbf{l}), \tag{2.36}$$

where the distribution of the latent classes $p_\rho(\mathbf{l})$ is according to the Potts model with fixed ρ, defined in Equation (2.8). We can then marginalize by summing out the hidden spatial

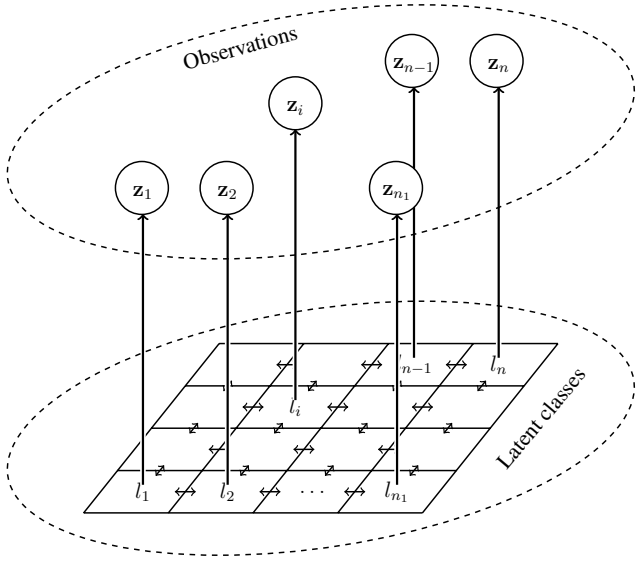


Figure 2.7: Visualization of the HMRF model. Arrows indicate conditional dependence. Latent classes are spatially dependent, whereas observations are conditionally independent, given the latent classes.

classes and achieve the likelihood of the model parameters, given the observed data,

$$L(\boldsymbol{\theta}, \rho|\mathbf{z}) = p_{\boldsymbol{\theta},\rho}(\mathbf{z}) = \sum_{\mathbf{l}' \in \mathbb{L}^n} p_{\boldsymbol{\theta},\rho}(\mathbf{z}, \mathbf{l}'). \tag{2.37}$$

The combined cylindrical HMRF model is conceptualized in Figure 2.7. Cylindrical observations are achieved based on a latent spatial model. Arrows from one variable to another indicate that the latter is conditionally dependent on the former, and there are no edges for variables that are conditionally independent. The latent classes in each cell in turn depend on all their neighbours, due to the Potts model neighbourhood structure. Here, n_1 represents the first of the grid dimensions.

Chapter 3

Inference

Having presented the models we intend to employ, we now consider ways of doing inference on the model parameters. First, in Section 3.1 we demonstrate an algorithm to exactly compute the likelihood of the HMRF model for small grid sizes. Section 3.2 provides two ways of approximating the likelihood of the HMRF for large grid sizes, which are used for parameter estimation. We continue by discussing the asymptotic properties of these estimators, before posing criteria that can be used to do model selection in Section 3.3. Finally, we present metrics for evaluating probabilistic predictions made by the models in Section 3.4.

3.1 Exact likelihood

In this section, we provide an algorithm for computing the exact likelihood of the HMRF model defined in Equation (2.37). First, note that the likelihood can be defined sequentially by,

$$p_{\boldsymbol{\theta},\rho}(\mathbf{z}) = p_{\boldsymbol{\theta},\rho}(\mathbf{z}_1) \prod_{t=2}^{n} p_{\boldsymbol{\theta},\rho}(\mathbf{z}_t | \mathbf{z}_{1:(t-1)}), \tag{3.1}$$

where the notation $\mathbf{z}_{1:(t-1)}$ represents the vector $(\mathbf{z}_1, \ldots, \mathbf{z}_{t-1})$. To find the likelihood, we therefore need to compute the sequential probabilities $p_{\boldsymbol{\theta},\rho}(\mathbf{z}_t | \mathbf{z}_{1:(t-1)})$. However, due to the computational complexity of the normalizing constant of the Potts model defined in Equation (2.9), this is not feasible for large spatial grids. Yet, through a restriction on either the number of rows or columns and a clever indexation on the grid cells, we are able to compute the normalizing constant exactly for relatively small grid sizes. The prerequisite for this exact algorithm to be computationally tractable is that one of the grid dimensions is small, either rows or columns, and the other can be large. The smallest dimension of the grid is denoted m. Further, the computation time is directly proportional to the largest spatial dimension. When the number of rows is small, the ordering of the spatial grid is as indicated in Figure 3.1 to create the vector representing the latent classes $\mathbf{l} = (l_1, \ldots, l_n)$ and the observations $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$.

1	5	9	13	17	21
2	6	10	14	18	22
3	7	11	15	19	23
4	8	12	16	20	24

Figure 3.1: Illustration of the indexation of cells on a 4×6 grid.

By utilizing this indexation, we are able to extend the classical forward–backward algorithm from hidden Markov chains [MacDonald and Zucchini, 1997] into a spatial setting [Reeves and Pettitt, 2004]. The forward–backward algorithm was first introduced by Baum et al. [1970] and computes the posterior marginal probabilities of all latent variables given the observed data. The algorithm also conveniently computes the full likelihood of the data after the forward pass. Following the lines of Friel and Rue [2007] we now describe the extension of the forward–backward algorithm to a spatial grid indexed as in Figure 3.1. Note that all ensuing probabilities of the latent variables and the observations are defined with fixed model parameters $(\boldsymbol{\theta}, \rho)$ denoted by subscripts as in Equation (3.1). In this section we drop writing the subscripts to make the derivations easier to follow.

The forward–backward algorithm moves recursively through the observations, first by a forward pass and then by a backward pass through the observation vector. We initialize the algorithm by finding the marginal distribution of the first latent variable, given only the first data observation

$$p(l_1|\mathbf{z}_1) = \frac{p(\mathbf{z}_1|l_1) \times p(l_1)}{p(\mathbf{z}_1)} = C_1 \times p(\mathbf{z}_1|l_1) \times p(l_1), \tag{3.2}$$

with $p(\mathbf{z}_1|l_1)$ being a cylindrical density from either Equation (2.11) or (2.23). As no information about the neighbours is available, $p(l_1)$ takes a discrete uniform distribution over the possible outcomes $l_1 \in \mathbb{L}$. The normalizing constant for the first latent variable is calculated as,

$$C_1 = \left(\sum_{l_1'=1}^{K} p(\mathbf{z}_1|l_1') \times p(l_1') \right)^{-1}, \tag{3.3}$$

and is actually the inverse of the first term in the sequential likelihood defined in Equation (3.1), i.e., $C_1^{-1} = p_{\boldsymbol{\theta},\rho}(\mathbf{z}_1)$.

The algorithm then moves forward through the spatial grid. In the Potts model, the neighbours of grid point t are the points $\{t-m, t-1, t+1, t+m\}$. This means that the latent variable at point t depends on all latent variables from $t-m$ to $t-1$. Hence, at each step t in the forward recursion, the algorithm needs to keep track of the probability of every outcome of the vector $\mathbf{l}_{(t-m):(t-1)} = (\mathbf{l}_{t-m}, \ldots, \mathbf{l}_{t-1})$, explaining the need for one of the spatial dimensions to be small. The following forward recursion for $t = 2, \ldots, n$ is

carried out,

$$C_t = \Bigg(\sum_{l'_t=1}^{K} \cdots \sum_{l'_{t-m}=1}^{K} p(\mathbf{z}_t|l'_t) p(l'_t|l'_{t-1}, l'_{t-m}) p(\mathbf{l}'_{(t-m):(t-1)}|\mathbf{z}_{1:(t-1)}) \Bigg)^{-1}, \tag{3.4}$$

$$p(\mathbf{l}_{(t-m):t}|\mathbf{z}_{1:t}) = C_t p(\mathbf{z}_t|l_t) p(l_t|l_{t-1}, l_{t-m}) p(\mathbf{l}_{(t-m):(t-1)}|\mathbf{z}_{1:(t-1)}), \tag{3.5}$$

$$p(\mathbf{l}_{(t-m+1):t}|\mathbf{z}_{1:t}) = \sum_{l'_{t-m}=1}^{K} p(\mathbf{l}'_{(t-m):t}|\mathbf{z}_{1:t}), \tag{3.6}$$

$$p(l_t|\mathbf{z}_{1:t}) = \sum_{l'_{t-1}=1}^{K} \cdots \sum_{l'_{t-m}=1}^{K} p(\mathbf{l}'_{(t-m):t}|\mathbf{z}_{1:t}). \tag{3.7}$$

Again, $p(\mathbf{z}_t|l_t)$ is a cylindrical density of either of the two forms (2.11) or (2.23). The distribution of l_t conditioned on l_{t-1} and l_{t-m} is given by the Potts model,

$$p(l_t|l_{t-1}, l_{t-m}) = \frac{\exp\Big[\rho\big(I(l_t = l_{t-1}) + I(l_t = l_{t-m})\big)\Big]}{\sum_{l'_t \in \mathbb{L}} \exp\Big[\rho\big(I(l'_t = l_{t-1}) + I(l'_t = l_{t-m})\big)\Big]}. \tag{3.8}$$

Observe now that the normalizing constants are actually the sequential conditional distributions of the observations,

$$C_t^{-1} = p(\mathbf{z}_t|\mathbf{z}_{1:(t-1)}). \tag{3.9}$$

This implies that the full likelihood of the observations, as given by Equation (3.1), is handily computed after the forward pass as,

$$L(\boldsymbol{\theta}, \rho|\mathbf{z}) = p_{\boldsymbol{\theta},\rho}(\mathbf{z}) = \prod_{t=1}^{n} C_t^{-1}. \tag{3.10}$$

To find estimates for the model parameters, we optimize the log-likelihood, given by

$$l(\boldsymbol{\theta}, \rho|\mathbf{z}) = -\sum_{t=1}^{n} \log(C_t). \tag{3.11}$$

This can be carried out by, e.g., the `optim` function in R [R Core Team, 2020].

At each step t in the forward recursion, we need to compute the forward probabilities $p(\mathbf{l}_{(t-m):t}|\mathbf{z}_{1:t})$ for all possible outcomes $\mathbf{l}_{(t-m):t} \in \mathbb{L}^{m+1}$, i.e., we need to compute K^{m+1} probabilities at each iteration. Also, to compute the normalizing constant in Equation (3.4) we sum $m+1$ variables over K outcomes, totalling K^{m+1} summation terms. Finally, the sum in Equation (3.6) involves K terms and the sum in Equation (3.7) involves K^m terms. Hence, we get an algorithm for computing the exact likelihood with complexity $O\big((n-m+1)K^{m+1}\big)$, as was reported by Reeves and Pettitt [2004]. Clearly, the algorithm is unfeasible when either the number of latent classes K or the number of rows m is large.

By iterating backwards, we are able to compute the marginal probabilities $p(l_t|\mathbf{z})$ of each latent variable, given the complete set of observations. These are important for making predictions of the true latent classes. The backward pass of the algorithm then consists of the following operations, iterated backwards for $t = n, \ldots, 2$,

$$p(l_{t-1}, l_t|\mathbf{z}) = \frac{p(l_{t-1}, l_t|\mathbf{z}_{1:t})}{p(l_t|\mathbf{z}_{1:t})} \times p(l_t|\mathbf{z}), \tag{3.12}$$

$$p(l_{t-1}|\mathbf{z}) = \sum_{l'_t=1}^{K} p(l_{t-1}, l'_t|\mathbf{z}), \tag{3.13}$$

where $p(l_{t-1}, l_t|\mathbf{z}_{1:t})$ can be found by summing Equation (3.5) over all outcomes of the latent spatial vector $\mathbf{l}_{(t-m):(t-2)} \in \mathbb{L}^{m-2}$, i.e., marginalize out all variables except (l_{t-1}, l_t). Finally, the forward–backward algorithm is very useful in practice. Essentially, the algorithm reuses the same design twice, by first iterating in increasing index order to calculate the forward probabilities in Equation (3.7), and then iterating by decreasing the index to calculate the backward probabilities in Equation (3.13).

Conveniently, we can also use the backwards probabilities to draw samples of the latent classes. First we define the sequential conditional probabilities,

$$p(l_t|l_{t-1}, \mathbf{z}) = \frac{p(l_{t-1}, l_t|\mathbf{z})}{p(l_{t-1}|\mathbf{z})}. \tag{3.14}$$

Using these probabilities, the full posterior latent model can be written sequentially as

$$p(\mathbf{l}|\mathbf{z}) = p(l_1|\mathbf{z}) \prod_{t=2}^{n} p(l_t|l_{t-1}, \mathbf{z}). \tag{3.15}$$

The sequential structure is utilized to develop an algorithm for drawing a sample of the latent segmentation $\mathbf{l}^s$ from the HMRF:

1. Simulate $l_1^s \sim p(l_1|\mathbf{z})$
2. For $t = 2, \ldots, n$, iteratively simulate $l_t^s \sim p(l_t|l_{t-1}^s, \mathbf{z})$

The resulting sample $\mathbf{l}^s$ is then distributed according to the full posterior model in Equation (3.15).

Another useful property of the forward–backward algorithm is that the hold-out probabilities are easily defined. These are the probabilities of the latent classes for each grid point, conditioned on all observations except the observation in that grid point. For grid point i, we denote by $\mathbf{z}_{-i}$ all observations except $\mathbf{z}_i$, or $\mathbf{z}_{-i} = (\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}, \mathbf{z}_{i+1}, \ldots, \mathbf{z}_n)$. The hold-out probabilities $p(l_i|\mathbf{z}_{-i})$ are then defined by,

$$p(l_i|\mathbf{z}_{-i}) = C_i \frac{p(l_i|\mathbf{z})}{p(\mathbf{z}_i|l_i)}, \tag{3.16}$$

where C_i is a normalizing constant dependent only on the observations $\mathbf{z}$, $p(l_i|\mathbf{z})$ are the backward probabilities defined in Equation (3.13), and $p(\mathbf{z}_i|l_i)$ are the cylindrical distributions defined in Equation (2.34). The hold-out probabilities are used to define the

predictive distribution of observing $\mathbf{z}_i^o$, given all the other observations $\mathbf{z}_{-i}$, as

$$p(\mathbf{z}_i^o|\mathbf{z}_{-i}) = \sum_{l_i' \in \mathbb{L}} p(\mathbf{z}_i^o|l_i')p(l_i'|\mathbf{z}_{-i}), \tag{3.17}$$

where again $p(\mathbf{z}_i^o|l_i')$ is a cylindrical distribution defined in Equation (2.34), and $p(l_i'|\mathbf{z}_{-i})$ are the hold-out probabilities from Equation (3.16).

3.2 Composite-likelihood

In the previous section, we developed an algorithm for computing the full likelihood of the HMRF model defined in Equation (2.37), with complexity $O\big((n-m+1)K^{m+1}\big)$. With only two latent classes, $K = 2$, Friel and Rue [2007] report that computation of the likelihood is feasible for grid sizes up to $m = 19$. Increasing the number of latent classes reduces the feasible grid size. With $K = 3$ and $K = 4$ latent classes, they report that the feasible grid size decreases to $m = 12$ and $m = 9$, respectively. What is considered tractable grid sizes have increased slightly since then, but not substantially. In addition, we are not interested in merely computing the likelihood, we seek to maximize it to estimate the model parameters. This reduces the feasible grid size for our application significantly.

There exists several approaches towards mitigating the unfeasible computational complexity involved with large spatial grids. Possible solutions include estimations by MCMC algorithms [Geyer and Thompson, 1995, Gelman and Meng, 1998] or approximation [Rydén and Titterington, 1998, Austad and Tjelmeland, 2016]. Another possibility is to substitute the full likelihood with a composite-likelihood. The composite-likelihood, also called pseudo-likelihood, is formed by multiplying component likelihoods, that correspond to marginal or conditional likelihoods of small subsets of data. Composite-likelihoods are used to reduce computational complexity to deal with large datasets or complex models [Varin and Vidoni, 2005] and have been thoroughly discussed; see, e.g., Besag [1974], Lindsay [1988], or Cox and Reid [2004]. The composite-likelihood approach may be less statistically efficient than the full likelihood, but it is much less expensive to calculate and may even be more robust to model misspecification [Lindsay et al., 2011]. Hence, the composite-likelihood can pose a valuable trade-off between statistical efficiency and computational tractability.

3.2.1 Pairwise-likelihood

Initially, we consider the components to be the set of all cliques defined in the Potts model in Section 2.1, i.e., the set of all pairs of neighbours. Hence, the likelihood reduces to a pairwise-likelihood [Varin et al., 2011], and we denote this method the pairwise-likelihood. Recall that this set of cliques is denoted $\mathbf{c}_{\mathcal{L}}$. Similarly to the full likelihood in Equation (2.37), we define the likelihood of each clique $\mathbf{c} = (c_1, c_2) \in \mathbf{c}_{\mathcal{L}}$ as,

$$L_{\mathbf{c}}(\boldsymbol{\theta}, \rho|\mathbf{z}_{\mathbf{c}}) = p_{\boldsymbol{\theta},\rho}(\mathbf{z}_{\mathbf{c}}) = \sum_{\mathbf{l}_{\mathbf{c}}' \in \mathbb{L}^2} p_{\boldsymbol{\theta},\rho}(\mathbf{z}_{\mathbf{c}}, \mathbf{l}_{\mathbf{c}}'), \tag{3.18}$$

where $\mathbf{z}_\mathbf{c} = (\mathbf{z}_{c_1}, \mathbf{z}_{c_2})$ 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 $\mathbf{l}_\mathbf{c} = (l_{c_1}, l_{c_2})$. Furthermore, the joint distribution of the clique observations $\mathbf{z}_\mathbf{c}$ and the clique latent classes $\mathbf{l}_\mathbf{c}$ is given as in Equation (2.36) by

$$p_{\boldsymbol{\theta},\rho}(\mathbf{z}_\mathbf{c}, \mathbf{l}_\mathbf{c}) = p_\rho(\mathbf{l}_\mathbf{c}) \prod_{i \in \mathbf{c}} \prod_{k=1}^{K} p_{\boldsymbol{\theta}_k}(\mathbf{z}_i)^{I(l_i = k)}. \tag{3.19}$$

Because the cliques are neighbouring pairs and necessarily neighbours in the Potts model, the distribution of the clique latent classes is,

$$p_\rho(\mathbf{l}_\mathbf{c}) = C(\rho)^{-1} \exp\big(\rho I(l_{c_1} = l_{c_2})\big), \tag{3.20}$$

with normalizing constant

$$C(\rho) = \sum_{\mathbf{l}'_\mathbf{c} \in \mathbb{L}^2} \exp\big(\rho I(l'_{c_1} = l'_{c_2})\big). \tag{3.21}$$

This normalizing constant is fast to evaluate because the sum is only over the outcome of two points in a pair, as each pair is assumed independent. The composite log-likelihood function is given as the sum of the log-likelihood of each clique,

$$cl(\boldsymbol{\theta}, \rho | \mathbf{z}, \mathbf{c}_\mathcal{L}) = \sum_{\mathbf{c} \in \mathbf{c}_\mathcal{L}} \log\big(L_\mathbf{c}(\boldsymbol{\theta}, \rho | \mathbf{z}_\mathbf{c})\big). \tag{3.22}$$

Because of the fast computation of each normalizing constant, the composite-likelihood is also cheap to calculate. This means that the function can be maximized directly by e.g., the `optim` function in R. However, as will be investigated and discussed later, the direct approach suffers from a small radius of convergence. The starting values for the optimization of the log-likelihood need to be close to the true solution in order for the optimization procedure to converge to the correct solution. To mitigate this problem we use the expectation-maximization (EM) algorithm, as was done in [Ranalli et al., 2018].

The EM algorithm presents an alternative way of finding maximum-likelihood estimators of model parameters with incomplete or missing data [Bilmes, 2000]. The algorithm appears naturally in problems where the data contains missing values. Also, it is useful in applications where the likelihood function is difficult to maximize analytically, but can be simplified by introducing additional hidden parameters. Our problem is an example of the latter, and we therefore exploit the EM algorithm to find the model parameters.

To define the iterations of the EM algorithm, we start by defining the complete-data composite-likelihood, which is the joint likelihood of the observations and the latent classes,

$$p_{\boldsymbol{\theta},\rho}(\mathbf{z}, \mathbf{l}) = \prod_{\mathbf{c} \in \mathbf{c}_\mathcal{L}} p_{\boldsymbol{\theta},\rho}(\mathbf{z}_\mathbf{c}, \mathbf{l}_\mathbf{c}), \tag{3.23}$$

with $p_{\boldsymbol{\theta},\rho}(\mathbf{z}_\mathbf{c}, \mathbf{l}_\mathbf{c})$ given by Equation (3.19). Then, note that the distribution of the clique latent classes can be written as,

$$p_\rho(\mathbf{l}_\mathbf{c}) = \prod_{\mathbf{l}'_\mathbf{c} \in \mathbb{L}^2} p_\rho(\mathbf{l}_\mathbf{c})^{I(\mathbf{l}'_\mathbf{c} = \mathbf{l}_\mathbf{c})}. \tag{3.24}$$

The complete-data composite log-likelihood is then defined by,

$$cl_{cd}(\boldsymbol{\theta}, \rho|\mathbf{z}, \mathbf{c}_{\mathcal{L}}) = \sum_{\mathbf{c}\in\mathbf{c}_{\mathcal{L}}} \big(cl^{\mathbf{c}}_{cd}(\boldsymbol{\theta}) + cl^{\mathbf{c}}_{cd}(\rho)\big), \tag{3.25}$$

with

$$cl^{\mathbf{c}}_{cd}(\boldsymbol{\theta}) = \sum_{i\in\mathbf{c}} \sum_{k=1}^{K} \log\big(p_{\boldsymbol{\theta}_k}(\mathbf{z}_i)\big) I(l_i = k), \tag{3.26}$$

$$cl^{\mathbf{c}}_{cd}(\rho) = \sum_{\mathbf{l}'_{\mathbf{c}}\in\mathbb{L}^2} \log\big(p_{\rho}(\mathbf{l}_{\mathbf{c}})\big) I(\mathbf{l}'_{\mathbf{c}} = \mathbf{l}_{\mathbf{c}}). \tag{3.27}$$

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 $t+1$, we have given parameter estimates $\hat{\boldsymbol{\theta}}_t$ and $\hat{\rho}_t$, 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 $\mathbf{c} \in \mathbf{c}_{\mathcal{L}}$ as

$$\hat{\mathbf{l}}_{\mathbf{c}} = p_{\hat{\boldsymbol{\theta}}_t, \hat{\rho}_t}(\mathbf{l}_{\mathbf{c}}|\mathbf{z}_{\mathbf{c}}) = \frac{p_{\hat{\rho}_t}(\mathbf{l}_{\mathbf{c}}) p_{\hat{\boldsymbol{\theta}}_{tk}}(\mathbf{z}_{\mathbf{c}})}{\sum_{\mathbf{l}'_{\mathbf{c}}\in\mathbb{L}^2} p_{\hat{\rho}_t}(\mathbf{l}'_{\mathbf{c}}) p_{\hat{\boldsymbol{\theta}}_{tk}}(\mathbf{z}_{\mathbf{c}})}, \tag{3.28}$$

with

$$p_{\hat{\boldsymbol{\theta}}_{tk}}(\mathbf{z}_{\mathbf{c}}) = \prod_{i\in\mathbf{c}} \prod_{k=1}^{K} p_{\hat{\boldsymbol{\theta}}_{tk}}(\mathbf{z}_i)^{I(l_i=k)}. \tag{3.29}$$

By marginalizing, we can obtain probabilities $\hat{l}_{ik} = p_{\hat{\boldsymbol{\theta}}_t, \hat{\rho}_t}(l_i = k|\mathbf{z}_{\mathbf{c}})$, $i \in \mathbf{c}$ of each grid point in the clique belonging to each latent class. The M-step consists of maximizing the expected value of the complete-data composite log-likelihood in Equation (3.25). 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

$$g(\boldsymbol{\theta}|\mathbf{z}, \mathbf{c}_{\mathcal{L}}) = E(cl^{\mathbf{c}}_{cd}(\boldsymbol{\theta})) = \sum_{\mathbf{c}\in\mathbf{c}_{\mathcal{L}}} \sum_{i\in\mathbf{c}} \sum_{k=1}^{K} \hat{l}_{ik} \log\big(p_{\boldsymbol{\theta}_k}(\mathbf{z}_i)\big), \tag{3.30}$$

$$h(\rho|\mathbf{z}, \mathbf{c}_{\mathcal{L}}) = E(cl^{\mathbf{c}}_{cd}(\rho)) = \sum_{\mathbf{c}\in\mathbf{c}_{\mathcal{L}}} \sum_{\mathbf{l}_{\mathbf{c}}\in\mathbb{L}^2} \hat{\mathbf{l}}_{\mathbf{c}} \log\big(p_{\rho}(\mathbf{l}_{\mathbf{c}})\big). \tag{3.31}$$

The parameter estimates are then obtained by $\hat{\boldsymbol{\theta}}_{t+1} = \arg\max_{\boldsymbol{\theta}}\{g(\boldsymbol{\theta}|\mathbf{z}, \mathbf{c}_{\mathcal{L}})\}$ and $\hat{\rho}_{t+1} = \arg\max_{\rho}\{h(\rho|\mathbf{z}, \mathbf{c}_{\mathcal{L}})\}$. We summarize the steps in the following simple algorithm:

1. Initialize values $(\hat{\boldsymbol{\theta}}_1, \hat{\rho}_1)$ for the model parameters and set iteration counter $t = 1$.
2. E-step: Given parameter estimates $(\hat{\boldsymbol{\theta}}_t, \hat{\rho}_t)$, compute the probabilities $\hat{\mathbf{l}}_{\mathbf{c}}$ for each latent configuration $\mathbf{l}_{\mathbf{c}} \in \mathbb{L}^2$ and for each clique $\mathbf{c} \in \mathbf{c}_{\mathcal{L}}$ by Equation (3.28). Also, marginalize over each grid point to obtain probabilities $\hat{l}_{ik} = p_{\hat{\boldsymbol{\theta}}_t, \hat{\rho}_t}(l_i = k|\mathbf{z}_{\mathbf{c}})$.
3. M-step: Update parameter estimates based on new latent probabilities by maximizing Equations (3.30) and (3.31).

4. Increase iteration counter, $t = t+1$ and repeat Steps 2 and 3 until some convergence criterion is satisfied.

To maximize Equation (3.30), we make use of a handy reparametrization. By defining the parameter vectors

$$\boldsymbol{\theta}_k = \big(\theta_{1k}, \ldots, \theta_{5k}\big) = \big(\log(\alpha_k), \log(\beta_k), \tan(\mu_k/2), \log(\kappa_k), \tanh^{-1}(\lambda_k)\big), \quad (3.32)$$

$$\boldsymbol{\theta}_k = \big(\theta_{1k}, \ldots, \theta_{5k}\big) = \big(\log(\alpha_k), \log(\beta_k), \tan(\mu_k/2), \log(\tau_k), \log(\frac{\kappa_k}{1-\kappa_k})\big), \quad (3.33)$$

in the case of WSSVM and GPTWC, respectively, we can optimize without any constraints because then $\boldsymbol{\theta}_k \in \mathbb{R}^5$. The optimization can then be carried out by quasi-Newton methods, and we have used the `BFGS` method from the function `optim` in R.

For the parameter ρ, we need to impose a constraint. As stated in Section 2.1, there is a phase transition above some critical value ρ_{crit}, and because we want mixing of the latent classes rather than all equal, we impose the constraint $\rho \in (0, \rho_{\text{crit}})$. We carry out the optimization through the method `L-BFGS-B` in `optim`, which allows for box-like constraints.

3.2.2 Block-likelihood

As we have discussed, calculating the exact likelihood is only feasible when either the number of rows or columns is small. Still, this method can be used to approximate the likelihood for grids where both spatial dimensions are large. This is done by separating the spatial grid into blocks. Crucially, these blocks need to have either a small number of rows or a small number of columns. Figure 3.2 represents a toy example of this separation. A 4×4 grid is separated into four blocks of size 1×4. Of course, the separation can also be done vertically to create blocks of size 4×1.

Because these blocks have either few rows or columns, the forward–backward algorithm can be used to calculate the exact likelihood of each block. The blocks can then be used as components to form a composite-likelihood as an approximation to the full likelihood, which we call the block-likelihood. The idea of splitting the spatial grid into smaller blocks that can be evaluated separately was also studied by Eidsvik et al. [2014]. In the same way as the pairwise-likelihood, the block log-likelihood is computed by summing the exact log-likelihoods of all blocks. Each block $\mathbf{b}$ has corresponding observations $\mathbf{z}_\mathbf{b}$ and exact log-likelihood given by Equation (3.11). The resulting block log-likelihood is on the form

$$bl(\boldsymbol{\theta}, \rho | \mathbf{z}, \mathbf{b}_\mathcal{L}) = \sum_{\mathbf{b} \in \mathbf{b}_\mathcal{L}} l(\boldsymbol{\theta}, \rho | \mathbf{z}_\mathbf{b}), \quad (3.34)$$

where $\mathbf{b}_\mathcal{L}$ represents the collection of all blocks. This block log-likelihood can be used as a substitute for the full log-likelihood and optimized by, e.g., the `optim` function in R. In the implementation, we use the same reparametrization of the parameters for the cylindrical distribution given in Equations (3.32) and (3.33), but we do not put a constraint on the spatial dependence parameter ρ. In this way, we get an unconstrained optimization problem and we use the `BFGS` from the `optim` function in R. In theory, the block-likelihood is more statistically efficient than the pairwise-likelihood, because it accounts for dependencies between entire blocks rather than dependence only between pairs.

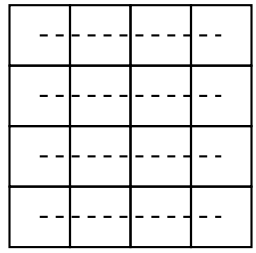


Figure 3.2: Toy example of the separation into blocks on a 4×4 grid. Dashed lines represents grid cells in the same block.

The pairwise-likelihood in Section 3.2.1 conveniently reduces to an EM algorithm that is simple to implement and relatively computationally efficient. This reduction is not as simple for the block-likelihood. The latent configuration probabilities can be computed in the E-step by the spatial extension of the forward-backward algorithm from Section 3.1, but the M-step is problematic. The inner sum in Equation (3.31) is over all clique configurations, $\mathbf{l_c} \in \mathbb{L}^2$, i.e., a sum over K^2 components. For an EM approach to the block-likelihood, this sum is instead over all possible configurations of the block, $\mathbf{l_b} \in \mathbb{L}^n$, i.e., complexity K^n, where n is the number of grid cells in the block. Clearly, this is intractable and an approximation would be needed. A natural solution is, for example, to consider only pairs in the block. This, however, leads us back to the original pairwise EM algorithm in the M-step, with simply a new way of computing the latent configuration probabilities in the E-step. Hence, not much is gained in comparison with the pairwise EM algorithm and we only consider direct maximization of the block-likelihood.

3.3 Asymptotic theory and derived quantities

For standard maximum-likelihood problems, the maximum-likelihood estimator $\hat{\boldsymbol{\theta}}$ can be found by solving the score function, which is the gradient of the log-likelihood, $l(\boldsymbol{\theta}|\mathbf{z})$,

$$\mathbf{s}(\boldsymbol{\theta}|\mathbf{z}) = \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{z}). \tag{3.35}$$

Under some regularity conditions [Gurland, 1954], the maximum-likelihood estimator is asymptotically normally distributed, or

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_q(0, I(\boldsymbol{\theta})^{-1}), \tag{3.36}$$

where q is the number of model parameters, or the length of the vector $\boldsymbol{\theta}$, and $I(\boldsymbol{\theta})$ is the expected Fisher information matrix, given by

$$I(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\Big(-\nabla^2_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{z})\Big). \tag{3.37}$$

Here, $\nabla^2_{\boldsymbol{\theta}}$ is the Hessian with respect to the model parameters $\boldsymbol{\theta}$. For more on the asymptotic properties of standard maximum-likelihood estimators, see, e.g., Millar [2011].

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 no necessarily hold. Consequently, we want to describe the asymptotic properties for these likelihoods and derive quantities that can be used for model selection. As a consequence of the assumed independence between each component, the second Bartlett identity [Bartlett, 1953],

$$I(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\Big(-\nabla^2_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{z})\Big) = \text{Var}_{\boldsymbol{\theta}}\Big(\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{z})\Big), \tag{3.38}$$

does not hold for composite-likelihoods [Varin et al., 2011]. Hence, 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 matrix,

$$H(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\Big(-\nabla_{\boldsymbol{\theta}} \mathbf{s}_{cl}(\boldsymbol{\theta}|\mathbf{z})\Big), \tag{3.39}$$

and the variability matrix,

$$J(\boldsymbol{\theta}) = \text{Var}_{\boldsymbol{\theta}}\Big(\mathbf{s}_{cl}(\boldsymbol{\theta}|\mathbf{z})\Big), \tag{3.40}$$

are not equal. Then, the Fisher information matrix is substituted by the Godambe information matrix [Godambe, 1960], given by

$$G(\boldsymbol{\theta}) = H(\boldsymbol{\theta}) J(\boldsymbol{\theta})^{-1} H(\boldsymbol{\theta}). \tag{3.41}$$

This matrix is commonly named the sandwich matrix due to its sandwich-like form, where $H(\boldsymbol{\theta})$ is multiplied on both sides of $J(\boldsymbol{\theta})^{-1}$. It is now evident that if the composite-likelihood was a true likelihood and not a misspecification, we would have $G(\boldsymbol{\theta}) = H(\boldsymbol{\theta}) = I(\boldsymbol{\theta})$. Finally, we arrive at the result that, under regularity conditions, the maximum-composite-likelihood estimator is asympotically normally distributed,

$$(\hat{\boldsymbol{\theta}}_{cl} - \boldsymbol{\theta}) \xrightarrow{d} N_q(0, G(\boldsymbol{\theta})^{-1}). \tag{3.42}$$

One key assumption in the model we have developed is that the number of latent classes K is known. Hence, we need a way to determine K, or in other words a way to do model selection. For regular maximum-likelihood a common way is to consider the Akaike information Criterion (AIC), introduced by Akaike [1973], which is an estimator for the quality of a model, relative to each of the other models. The AIC is given by,

$$\text{AIC} = -2l(\boldsymbol{\theta}|\mathbf{z}) + 2d_s, \tag{3.43}$$

where d_s is the number of estimated parameters. We consider the counterpart to AIC for composite-likelihoods, introduced by Varin and Vidoni [2005]. They developed a selection criteria that maximizes prediction power by minimizing the Kullback–Leibler distance to a future random variable. AIC and other information criteria combine goodness of fit (the log-likelihood) and a penalty term for the model complexity. For composite-likelihoods, the AIC is given by,

$$\text{C-AIC} = -2cl(\boldsymbol{\theta}|\mathbf{z}) + 2d^*_s, \tag{3.44}$$

where d_s^* is termed the effective degrees of freedom and given by $d_s^* = \text{tr}(\hat{J}(\boldsymbol{\theta})\hat{H}(\boldsymbol{\theta})^{-1})$. When the composite-likelihood is an ordinary likelihood, the effective degrees of freedom are equal to the number of parameters in the model, and the C-AIC coincides with the regular AIC.

Notice that $\hat{J}(\boldsymbol{\theta})$ and $\hat{H}(\boldsymbol{\theta})$ are estimators for $J(\boldsymbol{\theta})$ and $H(\boldsymbol{\theta})$, respectively, and these estimators need to be efficient. We refer to Varin and Vidoni [2005] for more elaborate details on how to define efficient estimators, but in this thesis we use the naive observed score and 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`.

The AIC is designed to select models with the best prediction power. Hence, it penalizes complex models mildly and as a result models that over-fit the data may be preferred. This thesis revolves around finding states of the ocean that are able to compactly explain ocean dynamics. Predictive power is therefore not critical and we rather want a model that parsimoniously can explain the complex surface currents. Thus, we also consider the composite-likelihood equivalent of the Bayesian information criterion (BIC) [Schwarz, 1978], which was proposed by Gao and Song [2010]. The composite-likelihood BIC is given by

$$\text{C-BIC} = -2cl(\boldsymbol{\theta}|\mathbf{z}) + \log(n)d_s^*, \tag{3.45}$$

where $\log(\cdot)$ is the natural logarithm. Compared to AIC, BIC penalizes complex models more heavily and generally leads to selection of less complex models. In contrast to AIC, BIC has been shown to be a consistent model selection criterion [Gao and Song, 2010].

Unfortunately, our naive estimator of the variability matrix $J(\boldsymbol{\theta})$ vanishes, i.e., it becomes singular, and the Godambe matrix becomes numerically unstable. Hence, we need a different procedure to estimate uncertainty in the maximum-composite-likelihood parameter estimates. A parametric bootstrap procedure is used in this thesis, with again Varin and Vidoni [2005] listing other possible approaches. The parametric bootstrap is based on the assumption that the original data set is a realization of a random sample from a parametric model [Givens and Hoeting, 2013], in our case the cylindrical HMRF. Random samples are then drawn from this model using the parameters estimated by maximum-composite-likelihood. Then, for each of the random samples we again use maximum-composite-likelihood to estimate the model parameters. These new maximum-composite-likelihood estimates are then used to assess the variability in the original estimate. The parametric bootstrap is a powerful tool when the parametric model is a good representation of reality, and allows for inference in intractable situations like this.

3.4 Performance measures

A key purpose of our models is to make predictions or forecasts for future observations and to assess the uncertainty in these forecasts. Due to the uncertainty element, the predictions take the form of probability distributions over possible outcomes. The goal is to maximize the sharpness of this predictive distribution, while maintaining unbiased predictions. Sharpness is a property of probabilistic forecasts and represents the concentration of the predictive distribution. Sharper predictive distributions are less uncertain, giving

better predictions. In this section, we develop scoring rules for predictions of cylindrical data. These rules are used to assess the quality of probabilistic forecasts by assigning a numerical score based on the forecast distribution and the observed value. The scoring rules are negatively oriented penalties, implying that the best forecast is the one minimizing the scoring rule. Scoring rules are discussed in more detail in, e.g., Gneiting et al. [2008].

In this thesis, we consider the continuous ranked probability score (CRPS), which is a scoring rule for univariate probabilistic forecasts. It is defined as,

$$\text{CRPS}(P, x^o) = \int_{-\infty}^{\infty} \big(F(y) - I(y \geq x^o)\big)^2 dy, \tag{3.46}$$

where P is the predictive probability distribution, x^o is the observed value, and F is the cumulative distribution function associated with P. Gneiting and Raftery [2007] showed that the CRPS can be equivalently defined by

$$\text{CRPS}(P, x^o) = E(|X - x^o|) - \frac{1}{2}E(|X - X'|), \tag{3.47}$$

where X and X' are independent random variables with distribution P.

The CRPS can be extended to multivariate linear variables by the energy score (ES), which is defined by,

$$\text{ES}(P, \mathbf{x}^o) = E(\|\mathbf{X} - \mathbf{x}^o\|) - \frac{1}{2}E(\|\mathbf{X} - \mathbf{X}'\|), \tag{3.48}$$

where $\|\cdot\|$ denotes the Euclidean norm. However, the circular part in the cylindrical data has to be handled differently. Grimit et al. [2006] proposed a circular analogue to the linear CRPS by considering the angular distance instead of absolute difference. The angular distance between the two angles ϕ and ϕ' is defined by

$$a(\phi, \phi') = \begin{cases} |\phi - \phi'|, & \text{if } |\phi - \phi'| \leq \pi, \\ 2\pi - |\phi - \phi'|, & \text{if } \pi < |\phi - \phi'| \leq 2\pi. \end{cases} \tag{3.49}$$

This is used to define the circular CRPS,

$$\text{CRPS}_{\text{circ}}(P, \phi^o) = E\big(a(\Phi, \phi^o)\big) - \frac{1}{2}E\big(a(\Phi, \Phi')\big), \tag{3.50}$$

where again Φ and Φ' are independent random variables with distribution P and ϕ^o is the observed angle.

As scoring rule for the cylindrical data, we consider the linear and circular part separately and deploy the appropriate version of CRPS separately. Another possible approach is to transform the cylindrical data to Cartesian coordinates. Then, the observations would be bivariate linear and the energy score would be an appropriate scoring rule.

Chapter 4

Simulated data

In this chapter, we demonstrate the proposed methods on simulated data sets. Section 4.1 begins by describing how the experiments are set up. In Section 4.2, we compare performance of the methods with regard to convergence radius of the optimization, parameter estimation accuracy, and run time. In Section 4.3, we use the favoured method to investigate bias and variance in the parameter estimation with both cylindrical densities.

4.1 Experimental setup

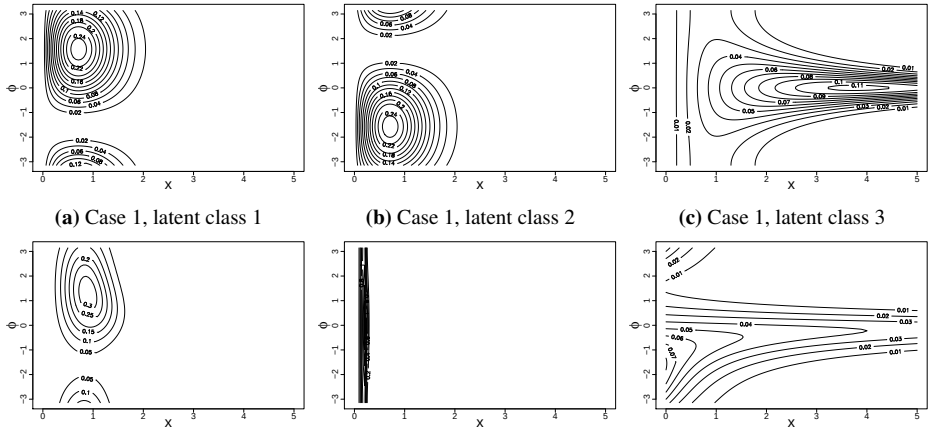
In this section we describe how the succeeding simulation experiments are set up. We first present the parameter sets that are used to simulate data from the WSSVM density and describe how data are simulated. We then display one simulated sample, before doing the same for the GPTWC density.

WSSVM: To study the methods for parameter estimation with the WSSVM density, 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 the true model parameters corresponding to each latent class. In the case of low separation, finding the true parameters and latent classes is difficult because the cylindrical densities corresponding to each latent class are close to each other. Hence, the method needs to be statistically robust to find the correct parameters. For larger separation, it is easier to determine the latent class because the cylindrical densities are further apart. As a result, the demand for robustness decreases.

To simulate data, we exploit the conditional structure of the model. First, we generate a random realization of the latent field by the Swendsen–Wang algorithm [Swendsen and Wang, 1987] with the true value for the spatial dependence parameter ρ. Then, given this latent field, observations $\mathbf{z}_i = (x_i, \phi_i)$ are generated by the WSSVM distribution from Section 2.2.1 according to the proposed algorithm, with model parameters determined by the latent class. Table 4.1 lists the true cylindrical model parameters for the two cases of the WSSVM density. Contour plots of the corresponding cylindrical densities are displayed

Table 4.1: True model parameters for the two cases in consideration with the WSSVM density.

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



(a) Case 1, latent class 1 **(b)** Case 1, latent class 2 **(c)** Case 1, latent class 3

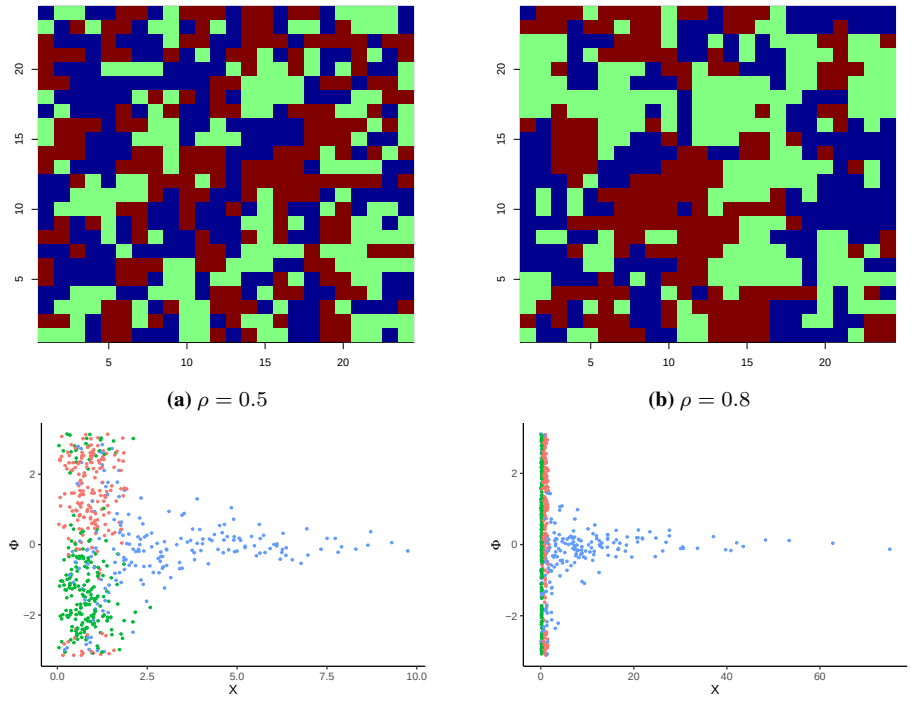
(d) Case 2, latent class 1 **(e)** Case 2, latent class 2 **(f)** Case 2, latent class 3

Figure 4.1: Contour plots of the true sample distributions for each of the two sets of cylindrical parameters with the WSSVM density.

in Figure 4.1. 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).

One random sample of the data is displayed in Figure 4.2. Here, we display both the true latent classes for both values of the spatial dependence parameter and the observed cylindrical data for both sets of parameters. In the display of the cylindrical data, the latent class is indicated by the colour of the points. From the displays of the true latent fields at the top, we observe that the right display, with $\rho = 0.8$, has larger patches of equal classes than the left display, with $\rho = 0.5$. Greater spatial dependence implies that the neighbours are more likely to take the same latent class. For the cylindrical observations at the bottom we see that the observations from each class are less separated in the display to the left. Consequently, the three cylindrical densities are harder to distinguish, and the parameters more demanding to estimate.

In Section 4.2, this setup is used to compare the pairwise-likelihood method from Section 3.2.1 and the block-likelihood method from Section 3.2.2. We also apply the same setup in Section 4.3.1 to investigate how the parameters of the WSSVM density behave. For the pairwise-likelihood, we consider both direct maximization of the composite-likelihood function in Equation (3.22) and the EM algorithm. Direct maximization can



(a) $\rho = 0.5$ **(b)** $\rho = 0.8$

(c) Case 1 **(d)** Case 2

Figure 4.2: One random realization of the latent classes (top) and observations (bottom) for the cases in consideration with the WSSVM density.

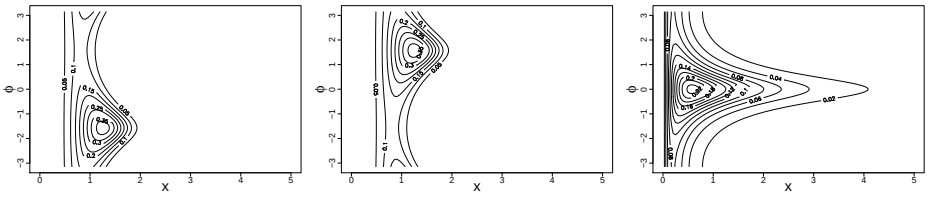
be carried out by, for example, the `BFGS` method from the `optim` function in R. For the block-likelihood approach, we split the grid into all possible blocks of size $m \times 24$ horizontally and $24 \times m$ vertically. This means that, e.g., for $m = 2$ we have in total 23 vertical and 23 horizontal 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. The two methods are compared regarding three critical issues: convergence radius, parameter estimation accuracy, and run time.

GPTWC: For the GPTWC density, we again construct two cases for the cylindrical parameter sets and consider the same two values for the spatial dependence parameter. Data are simulated on a 24×24 grid with $K = 3$ latent classes. The two cases are constructed with greater and lesser extent of separation between the densities linked to the latent classes. Again, Case 1 has low separation, implying that the true parameters and also the true latent classes are more difficult to find. Table 4.2 lists the true model parameters for the two cases. Contour plots of the resulting cylindrical densities are displayed in Figure 4.3. Again, we clearly observe that there is more overlap between the densities with low separation in Case 1.

To generate data, we apply the same algorithm as for the data generation with the

Table 4.2: True model parameters for the two cases in consideration with the GPTWC density.

Case 1: Low separation			Case 2: High separation		
$\alpha_1 = 0.25$	$\alpha_2 = 0.25$	$\alpha_3 = 0.5$	$\alpha_1 = 0.25$	$\alpha_2 = 0.25$	$\alpha_3 = 0.25$
$\beta_1 = 1$	$\beta_2 = 1$	$\beta_3 = 0.5$	$\beta_1 = 1$	$\beta_2 = 3$	$\beta_3 = 0.5$
$\mu_1 = -\frac{\pi}{2}$	$\mu_2 = \frac{\pi}{2}$	$\mu_3 = 0$	$\mu_1 = 0$	$\mu_2 = 0$	$\mu_3 = 0$
$\tau_1 = 0$	$\tau_2 = 0$	$\tau_3 = 0.8$	$\tau_1 = 0$	$\tau_2 = 0$	$\tau_3 = 0$
$\kappa_1 = 0.7$	$\kappa_2 = 0.7$	$\kappa_3 = 0.8$	$\kappa_1 = 0.6$	$\kappa_2 = 0.6$	$\kappa_3 = 0.2$



(a) Case 1, latent class 1 **(b)** Case 1, latent class 2 **(c)** Case 1, latent class 3

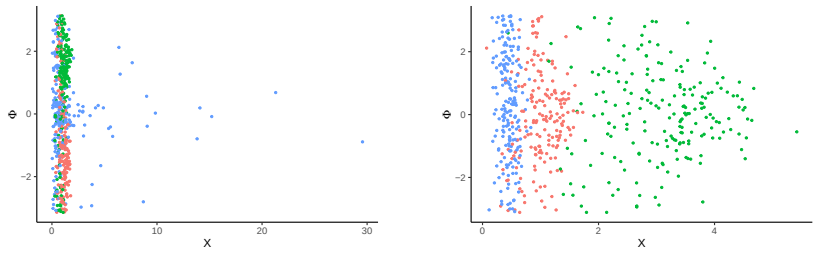
(d) Case 2, latent class 1 **(e)** Case 2, latent class 2 **(f)** Case 2, latent class 3

Figure 4.3: Contours of the true sample distributions for each of the two cases with GPTWC density.

WSSVM density, but the cylindrical observations are instead simulated from the procedure for the GPTWC density presented in Section 2.2.2. One random realization of the observations is displayed for both cases in Figure 4.4. The colour of the dots indicate the latent class. The latent fields are similar to the ones presented in Figure 4.2 and are hence not displayed. Again, notice that the observations from each latent class are more separated for Case 2 in the right display than for Case 1 in the display to the left.

4.2 Model comparison

Here, we compare the methods presented in the previous chapter by applying them to simulated data. We investigate convergence radius, parameter estimation accuracy and run time. This is only reported for the WSSVM density, but similar results also apply to the GPTWC density. When comparing models, it is advantageous to use simulated instead of real data. This is because the models can be judged on how well they estimate the known truth. Also, simulated data are easily controllable. This allows for the design of experiments to test specific cases or highlight particular properties of the models.



(a) Case 1 **(b)** Case 2

Figure 4.4: One random sample of the observations for the two cases in consideration with the GPTWC density.

4.2.1 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-likelihood with $m = 1$. We draw 50 data samples for both cases and both values of the spatial dependence parameter. 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, so the method needs to be robust to find the correct solution. In this way, less robust methods will converge to local optimums that are not the true solution. Finally, 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 implement the following strategy. For each starting point we compute the RMSE versus the true parameters. If $\boldsymbol{\theta}$ are the true parameters and $\boldsymbol{\theta}^0$ are the starting values, the starting point RMSE is given by

$$\text{RMSE}^0 = \sqrt{\frac{1}{q}\sum_{i=1}^{q}(\boldsymbol{\theta}_i^0 - \boldsymbol{\theta}_i)^2}, \tag{4.1}$$

where q is the length of the parameter vector, i.e., the number of parameters. Similarly, we compute the RMSE of the parameter estimates, $\hat{\boldsymbol{\theta}}$, versus the true parameters as,

$$\widehat{\text{RMSE}} = \sqrt{\frac{1}{q}\sum_{i=1}^{q}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^2}. \tag{4.2}$$

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 $\widehat{\text{RMSE}} < \text{RMSE}^0$. On the other hand, if $\widehat{\text{RMSE}} > \text{RMSE}^0$, the algorithm has converged to a local optimum that is further away from the true solution than the starting values.

Table 4.3: Number of times each method converged to the true solution.

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

Table 4.4: Average computation time of each method in minutes.

	Case 1		Case 2	
Method	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$
EM pairwise	5.47	5.22	4.31	5.25
Direct pairwise	4.08	3.92	3.30	3.10
Block, $m = 1$	3.12	2.89	2.37	2.34

Convergence results: In Table 4.3, we display the number of times each method converged to the true solution from 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 parwise converged 69 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 two direct optimization strategies. Moreover, these results indicate that the convergence radius of the two direct methods are comparable. The block-likelihood method comprises less approximations and should in theory be more statistically robust. However, the effect is limited in practice.

Run time: Table 4.4 displays the average run time in minutes for each method in all four cases using an Intel Core m3 processor ($0.9\,\mathrm{GHz}$). As stopping criterion for the optimization, we use a relative tolerance of $T = 10^{-5}$. This means that when the relative increase in pairwise or block log-likelihood is less than 10^{-5}, the optimization procedure is stopped. The EM algorithm has a large area of convergence, but it is computationally slow. Consequently, the direct pairwise approach was designed to accelerate and reduce run time of the EM algorithm. The results show that direct pairwise optimization reduces the run time of the EM algorithm by approximately 30%, which is also what Ranalli et al. [2018] reported. However, this comes at the cost of a smaller convergence radius. The block-likelihood further reduces the computational time by 25% compared to the direct pairwise-likelihood. Each component of the block-likelihood takes longer to compute than the components of the pairwise-likelihood. However, the pairwise-likelihood is a summation over all pairs, and because there are so many pairs, it is slower than the block-likelihood. Hence, if the starting point is close to the true solution so that convergence is guaranteed, the block-likelihood is favoured.

By increasing the number of rows in each block, we expect to get a method that is more

exact and hence more robust. Although we did not consider m values larger than 1 in this test, tests of smaller sample size indicate that the convergence radius does not increase significantly by increasing m. Hence, the added computational complexity of increased m does not pay off in a larger convergence radius.

Hybrid algorithm: 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 chapter, we develop a hybrid algorithm that follows the short-run strategy of Ranalli et al. [2018] to prevent converging to local maximums. 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 T_{short}. In our case we use $T_{\text{short}} = 10^{-2}$. 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. As we have seen, direct likelihood maximization requires that the starting point is sufficiently close to the maximum, which is ensured by the 50 random starting points scheme. Direct optimization is then run until full convergence, with convergence tolerance $T_{\text{long}} = 10^{-5}$. 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.

4.2.2 Parameter estimation accuracy

In this section, we assume that the starting point is close to the true solution and that direct optimization is guaranteed to converge to the true solution. We then investigate if the pairwise or the block-likelihood yield the most accurate parameter estimates and again compare 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.

As direct likelihood methods, we consider again the direct pairwise-likelihood and block-likelihood methods, but this time we also include $m = 2$. To compare the performance of the methods, we consider the RMSE of the parameter estimates as in Equation (4.2) and average this over the random draws. We draw 50 replicates for both cases and both values of the spatial dependence parameter. We set the true parameters as initial values for the optimization procedure to ensure and speed up the convergence.

Run time: The average run times are shown in Table 4.5. Observe that these run times are much shorter than for the convergence radius test in Table 4.4. This is because the starting point is much closer to the optimum in this test, leading to faster convergence. On average the block-likelihood method with $m = 1$ leads to roughly 30% reduction in computational time compared to the pairwise-likelihood, which is about the same as in the test for convergence radius. Also, $m = 2$ is 5.5 times slower than $m = 1$, showcasing the exponential growth in computational complexity as m increases.

Table 4.5: Average computation time of each method in minutes.

	Case 1		Case 2	
Method	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$
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 4.6: Average RMSE of parameter estimates.

	Case 1		Case 2	
Method	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$
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

Accuracy: In Table 4.6, we list the average RMSE of the parameter estimates in all four cases for all three methods. Table 4.7 lists the average proportion of correctly predicted latent classes for each method and each case with a maximum probability prediction criterion. This means that for each grid point we predict the latent class that maximizes the marginal probability, given all data, defined in Equation (3.13).

From the results, we see that the block-likelihood approach outperforms the direct pairwise-likelihood, although to a limited extent, in both estimating the parameters and predicting the latent classes. Moreover, $m = 1$ has a lower average run time than the direct pairwise-likelihood. As we would expect, $m = 2$ performs better than $m = 1$. However, the improvement is marginal and factoring in run time, we favour $m = 1$. Again, 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. As a result, we use the block-likelihood with $m = 1$ when estimating parameters going forward.

As expected, greater separation between the densities results in a larger proportion of correctly predicted latent classes. That is, strong separation makes for easier prediction of the latent class. Likewise, prediction of the latent classes improves for larger values of the spatial dependence parameter ρ. For larger values of ρ, the dependence between grid

Table 4.7: Average proportion of correctly predicted latent classes.

	Case 1		Case 2	
Method	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$
Direct pairwise	73.9%	76.0%	90.6%	91.7%
Block, $m = 1$	75.4%	78.8%	91.5%	93.2%
Block, $m = 2$	76.8%	78.1%	92.3%	93.2%

points is higher, implying that the prediction of the latent class of each grid point is more dependent on the observations in neighbouring grid points. As a result, each latent class prediction gets more support from neighbouring observations and accuracy improves.

Interestingly, the prediction of the latent classes improves with stronger separation, but the same can not be said for the RMSE of parameter estimates. The pairwise-likelihood exhibits lower RMSE for Case 2, but for the block-likelihood, the RMSE is lower for Case 1. Hence, larger separation does not necessarily imply more accurate estimates of each individual parameter. The accuracy of parameter estimates depends on both the bias and the variability in these estimates. This is investigated in the next section.

4.3 Behaviour of parameter estimates

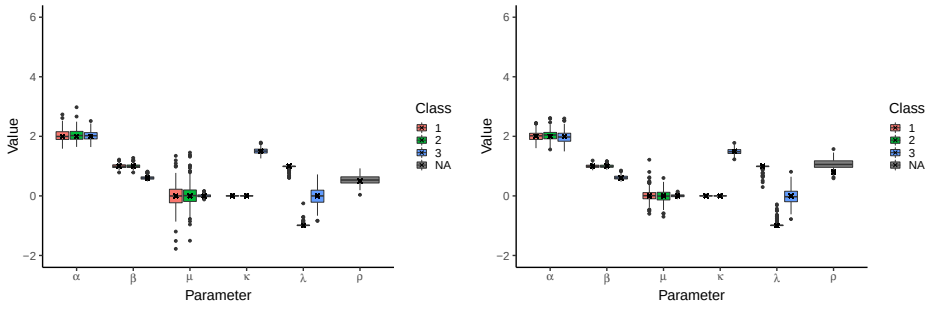
In this section, we investigate the behaviour of the estimates of each individual parameter. This is done for both the WSSVM density and the GPTWC density. First, we study bias and variance of all parameters, before checking if the normality assumption of maximum-likelihood estimators holds.

4.3.1 WSSVM

To examine the properties of the parameter estimates with the WSSVM density, we consider the two cases of parameter sets from Table 4.1 and the two values for the spatial dependency parameter ρ presented in Section 4.1. For each of the 4 cases, we draw 200 realizations of the latent field. For each latent field, we draw observations from the WSSVM density, resulting in 200 sets of observations for each case. As discussed in the previous sections, the block-likelihood method with $m = 1$ is favoured, and we use this method to estimate parameters for each set of observations. The true parameters are used as initial values to speed up convergence and prevent convergence to local maximums.

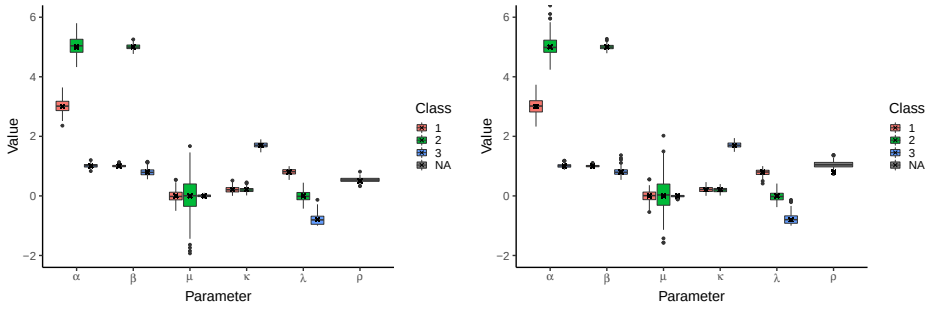
Bias: In Figure 4.5, we display box plots of the parameter estimates for all four cases in consideration. The density class for each parameter is indicated by the colour. As there is only one spatial interaction parameter ρ, this has a grey colour and class "NA". The true parameter values are indicated by crosses. 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. Recall from Section 2.2.1 that the boundaries for these parameters are $\kappa \geq 0$ and $\lambda \in [-1, 1]$. When the true value is at or close to these limits, 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 $\rho_{\text{crit}} \approx 1.005$. With $\rho = 0.5$, the averages are both close to 0.5. This clearly shows that there is a bias in the parameter ρ for larger values.

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, $\rho \in \{0, 0.34, 0.5\}$. 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. In the paper where



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

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



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

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

Figure 4.5: 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 colour. The crosses indicate the true parameter values.

they proposed 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.

Variance: To better 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 Figure 4.6. The variability in parameter estimates are 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 values to a greater extent, although their averages coincide with the truth. 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 Figure 4.6, 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 theory is that the variability is instead strongly linked to the properties of each class density. When densities are constructed in a way such that changing a parameter value does not alter the density much, the variance of the estimate for that

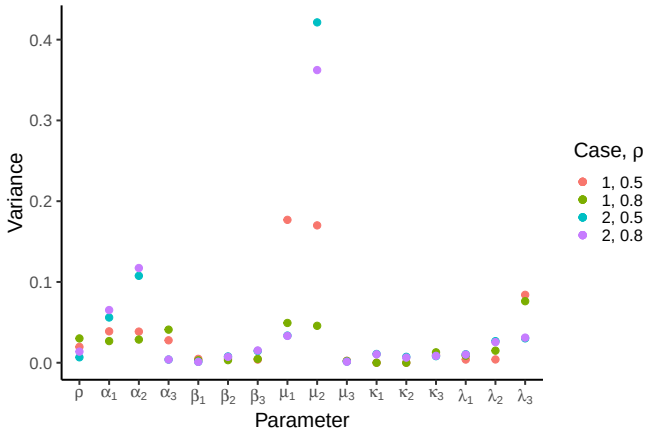


Figure 4.6: Empirical variance in parameter estimates for the WSSVM density for the two cases and two values of the spatial dependence parameter ρ.

parameter becomes large. As an example, 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 because of $\lambda_2 = 0$ 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 α_2 in Case 2 and λ_3 in Case 1, small increments to these parameters have little effect on the corresponding densities.

Another interesting feature observed from the variance display is that a low value of ρ yields considerably larger variance in μ_1 and μ_2 for Case 1 and μ_2 for Case 2. The rationale is not obvious and calls for a thorough explanation. If we consider the class 1 and 2 densities of Case 1 displayed in Figures 4.1a and 4.1b, and also the corresponding data sample in Figure 4.2c, 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 approximately 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.

In Figure 4.7, we display all estimates of μ_1 versus μ_2 for Case 1 for three different values of ρ. We observe that the estimates are heavily correlated. Moreover, there is a clear trend that lower values of ρ lead to larger variation in both parameter estimates and also a larger correlation between the two parameters. This supports the claim that μ_1 and μ_2 need to be approximately equal to fit the data correctly. Computing the empirical correlation between estimates of μ_1 and μ_2 gives a correlation of 0.95 for $\rho = 0.2$, 0.86 for $\rho = 0.5$ and 0.58 for $\rho = 0.8$. The correlation is clearly larger for smaller values of ρ.

Now, to explain why this is the case, we restate the fact that larger values of ρ imply larger spatial dependency and a greater probability that neighbouring points take the same

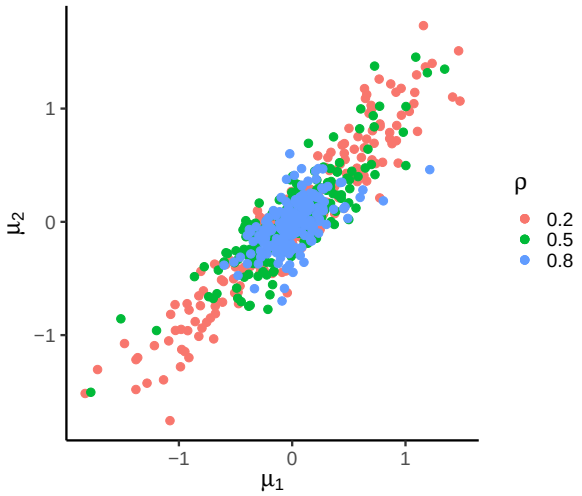


Figure 4.7: Parameter estimates for μ_1 (x-axis) and μ_2 (y-axis) for Case 1 for different values of the spatial dependence parameter ρ.

latent class. Hence, we get better predictions of the latent classes, leading to more information about which observations belong to which class density. Then, better segmentation of the approximately uniformly circular distributed observations into their correct latent classes makes for better estimations of the circular location parameters μ_1 and μ_2.

The same line of reasoning also applies to μ_2 in Case 2. Better predictions of the latent classes makes for better estimates of the parameters. In this case, the class 2 density in Case 2 is almost uniform in the circular part, making μ_2 very sensitive to the segmentation of the observations. This is why the variability increases when ρ decreases. For all other parameters, the variance is approximately equal for the two values of ρ. As a result, these parameters are not as sensitive to changes in the observations or the predicted segmentation and ρ does not affect the estimates to the same degree.

As previously stated, there was no evidence of bias for $\rho = 0.5$, but $\rho = 0.8$ gave a bias towards larger values. Therefore, we also check for bias with smaller values for ρ. In Figure 4.8, we display a box plot of the parameter estimates for Case 1 with $\rho = 0.2$. These estimates were also used to support the claim for increasing variation in μ_1 and μ_2 for lower values of ρ. Apart from the increased variability in μ_1 and μ_2, there is not much difference between the parameter estimates for different values of ρ, implying that ρ does not affect the other parameters to the same extent.

For $\rho = 0.2$, we observe no bias in the spatial dependence parameter. We have also run the same experiment with $\rho = 0.95$ and observed a mean of 2.14 and an empirical variance of 1.58. The critical parameter value for $K = 3$ is $\rho_{\text{crit}} \approx 1.005$, meaning that the model severly overestimates ρ, even beyond the critical parameter value. Clearly, the bias in ρ becomes larger as it approaches the critical value ρ_{crit} and the variance also increases. It is hard to tell why this bias occurs and we see no obvious explanation. Hence, care needs to be taken when we move to real data.

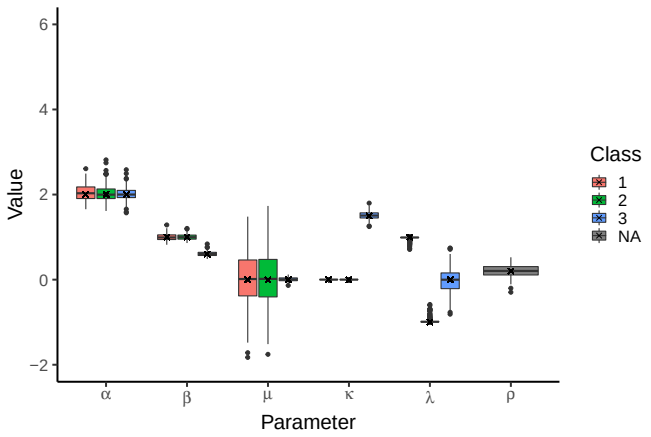


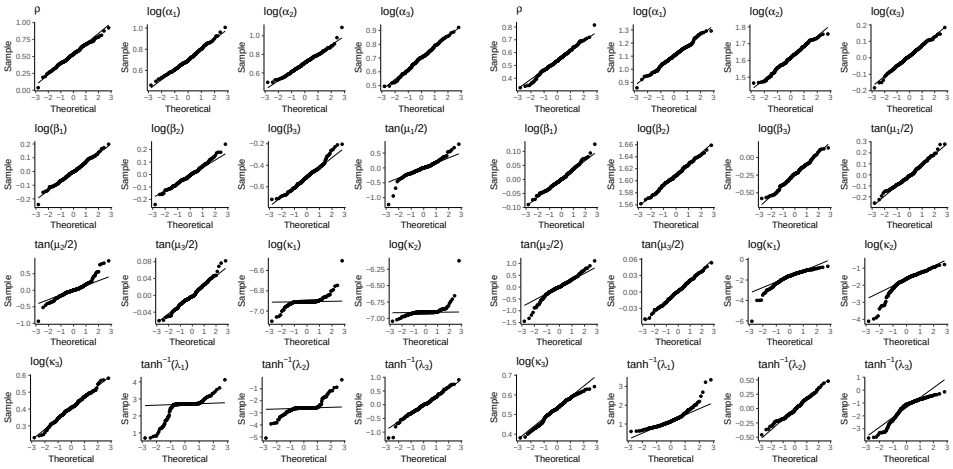
Figure 4.8: Box plots of parameter estimates for Case 1 of the WSSVM density for $\rho = 0.2$. Parameter classes are indicated by the colour. The crosses indicate the true parameter values.

Asymptotic normality: In Section 3.3, we stated that standard maximum-likelihood estimators are asymptotically normally distributed with variance as the inverse of the expected Fisher information matrix. We also stated that maximum-composite-likelihood estimators are asymptotically normal with variance as the inverse of the Godambe sandwich matrix. Thus, we check if this statement is true for our problem, i.e., check the computed parameter estimates for normality.

To check the normality assumption, we make Q-Q plots of the parameter estimates. Figure 4.9 displays Q-Q plots of all parameter estimates for Case 1 and Case 2 with $\rho = 0.5$. The results are similar for $\rho = 0.2$ and $\rho = 0.8$, so these are not included. Note that the Q-Q plots display the transformed parameters $\boldsymbol{\theta}$ that we optimized for, and not the original parameters. Recalling Equation (3.32), this means that, e.g., the estimates for $\log(\alpha)$ are displayed instead of the estimates for α. The display shows that, with some exceptions, most parameters are approximately normally distributed. The major outliers are the circular skewness parameter λ and the circular-linear dependence parameter κ. Critically, when these parameters are at or close to their boundaries, $|\lambda| = 1$ and $\kappa = 0$, the normality assumption no longer holds. For Case 1, we see that at the boundary, these parameter estimates become extremely heavy tailed. For Case 2, the parameters are close to the boundary and the estimates become skewed.

4.3.2 GPTWC

We also study properties of the parameter estimates for the GPTWC density. With the setup described in Section 4.1, we draw 200 sets of observations for each case and each value of the spatial dependence parameter. As for the WSSVM density, the block-likelihood method with $m = 1$ is used to estimate parameters for each set of observations, with the true parameters used as initial values for the optimization.



(a) Case 1, $\rho = 0.5$ **(b)** Case 2, $\rho = 0.5$

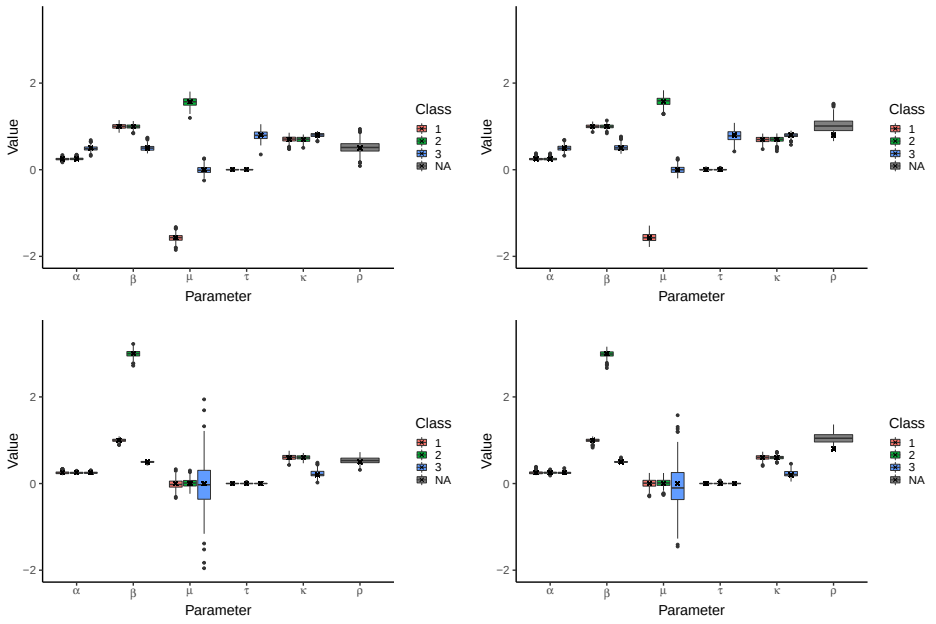
Figure 4.9: Q-Q plots of parameter estimates for Case 1 (left) and Case 2 (right) of the WSSVM density for $\rho = 0.5$. The parameter is indicated at the top left corner of each plot.

Bias: Figure 4.10 shows box plots of parameter estimates for all four cases considered. Again, the density class for each parameter is indicated by the colour, with a grey "NA" for ρ. Comparing the display to the true parameters, there is little evidence of bias in most parameters. However, we again observe the same effect that there is a bias for $\rho = 0.8$, with the mean of all estimates being roughly 1 in both cases. Noticeably, for this density, all parameter estimates seem to be symmetric, even those at the limit $\tau = 0$.

Although it is not apparent from the box plots, we can observe a small bias in the parameter τ by computing the means for the estimates of this parameter. For small values, $\tau = 0$, there is a positive bias, meaning that estimates of the parameter τ are too large. For large values of τ, here for Case 1, where $\tau_3 = 0.8$, there is a negative bias, i.e., the estimates are too small. The biases are very limited, typically of magnitude 0.02, but importantly they are consistent across all cases and classes. These results support the findings by Tomoaki et al. [2019], who computed maximum-likelihood estimates of parameters with data generated from a single GPTWC density. When τ is low, they found that τ was overestimated, whereas larger values of τ lead to underestimation of the parameter. For $\tau = 0.1$, they observed overestimation and for $\tau \geq 0.3$ they observed underestimation.

Variance: As we did for the WSSVM distribution, we display empirical variances of the parameter estimates in Figure 4.11. For both densities, but even more so for the GPTWC density, we observe that the variance in ρ is greater for weakly separated classes than it is for strongly separated classes. The intuition is that for weakly separated classes, the latent classes are harder to predict accurately. As a consequence, the spatial dependence parameter ρ becomes harder to estimate and variability increases, although the bias remains the same for both cases.

From the display of the empirical variances, we also observe that μ_3 in Case 2 has



(a) $\rho = 0.5$ **(b)** $\rho = 0.8$

Figure 4.10: Box plots of parameter estimates for Case 1 (top) and Case 2 (bottom) of the GPTWC density for $\rho = 0.5$ (left) and $\rho = 0.8$ (right). Parameter classes are indicated by the colour. The crosses indicate the true parameter values.

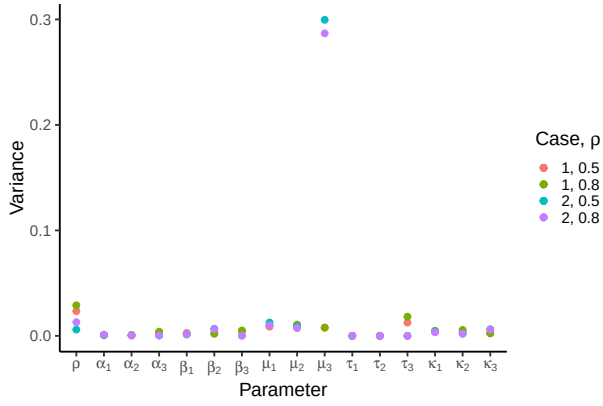


Figure 4.11: Empirical variance in parameter estimates for the GPTWC density for the two cases and two values of the spatial dependence parameter ρ.

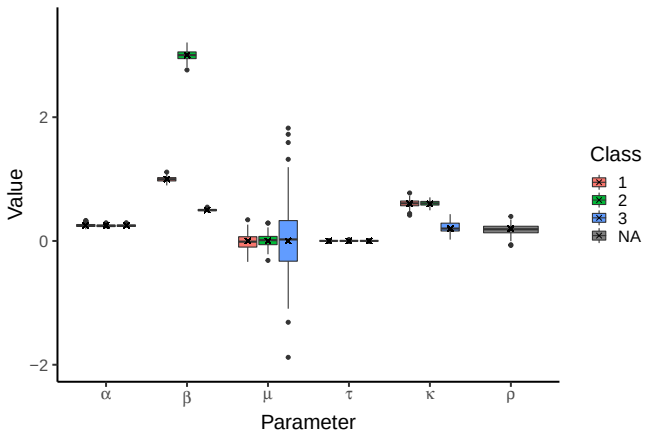


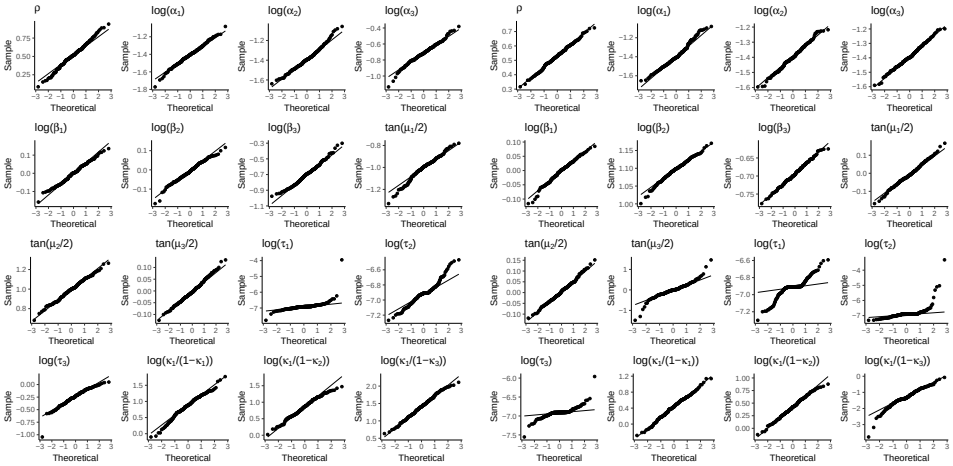
Figure 4.12: Box plots of parameter estimates for Case 2 of the GPTWC density for $\rho = 0.2$. Parameter classes are indicated by the colour. The crosses indicate the true parameter values.

substantially larger variance than the other parameters. The interpretation is the same as for μ_2 in Case 2 of the WSSVM density. Because of κ_3 being low and $\tau_3 = 0$ in Case 2, the class 3 density is almost uniform in the circular part. Hence, the parameter μ_3 is harder to estimate correctly because perturbations in the parameter have limited impact on the corresponding density. This leads to a large variance in the parameter estimates. For μ_3 in Case 2, we also observe increasing variance as ρ decreases. Again, the reason behind this is that this parameter is highly sensitive to the latent segmentation and smaller values of ρ makes the prediction of latent classes harder.

Interestingly, apart from μ_3 in Case 2 and ρ, there seems to be less variance for the GPTWC densities than there was for the WSSVM densities. This may be because the GPTWC densities are more sensitive than the WSSVM densities to changes in the parameters. This would explain why the variance is lower for the GPTWC density.

To again check for bias for lower values of the spatial dependence parameter, we run the Case 2 experiment with $\rho = 0.2$. The box plots of these parameter estimates are displayed in Figure 4.12. We can then verify that there is no bias for $\rho = 0.2$ for the GPTWC density. From the display we also observe that the variance in μ_3 is larger than it was for $\rho = 0.5$.

Asymptotic normality: Similarly to what we did for the WSSVM density, we also display Q-Q plots of the transformed parameter estimates for the GPTWC density in Figure 4.13. We observe that most parameters fulfil the asymptotic normality assumption, except for the heavy-tailedness parameter τ. Crucially, when the true value of this parameter is at its limit, $\tau = 0$, the normality assumption no longer holds. We also observe some heavy tails for μ_3 for Case 2. The estimates for κ are closer to the normal distribution than the estimates for κ with the WSSVM distribution in Figure 4.9. This is because the κ values for the WSSVM distribution were closer to the lower boundary $\kappa = 0$.



(a) Case 1, $\rho = 0.5$ **(b)** Case 2, $\rho = 0.5$

Figure 4.13: Q-Q plots of parameter estimates for Case 1 (left) and Case 2 (right) of the GPTWC density for $\rho = 0.5$. The parameter is indicated at the top left corner of each plot.

Chapter 5

Real data

This chapter is dedicated to the analysis of a real set of ocean surface current (OSC) observations in the Norwegian Sea. We first present and motivate the data set in Section 5.1, before estimating parameters for both densities for different number of latent classes. Then we use model selection criteria to decide the number of latent classes and discuss the results. We estimate two separate models for summer and fall in Section 5.2 and a single model for both seasons in Section 5.3. Then, in Section 5.4 the performance of all estimated models is compared.

5.1 Data description

The OSC data we consider in this chapter consist of measurements of the combined geostrophic and Ekman currents, as described in Chapter 1, and are provided by the GlobCurrent project, which is funded by the European Space Agency.[1] The area is part of the Norwegian Sea and ranges from 66°N to 72°N latitude and 3°W to 12°E longitude. Figure 5.1 displays the area. It is of particular interest to study the OSC in this area because the inflow of Atlantic Water has a great impact on climate and biological production in the Nordic Seas. From here, the Atlantic Water flows further into the Arctic Oceans, hence impacting the climatic state of a vast region.

Motivation: 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 at the opening of the Barents Sea. These fluctuations were analyzed and decomposed using orthogonal functions by Ingvaldsen et al. [2004]. The mean state and variations in inflow of Atlantic Water into the Norwegian Sea was studied by Skagseth et al. [2008], with a focus on describing the flow of heat and volume. Common for all, is that they used several time series of mooring observations to classify the general total flow in the areas around the Norwegian Sea.

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

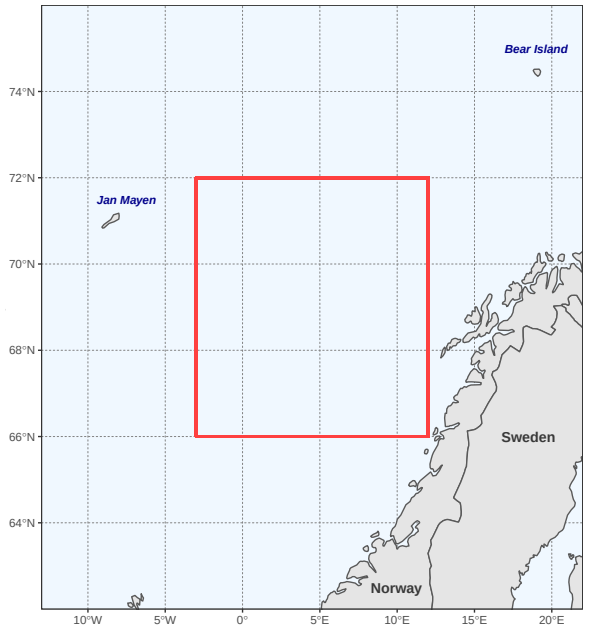
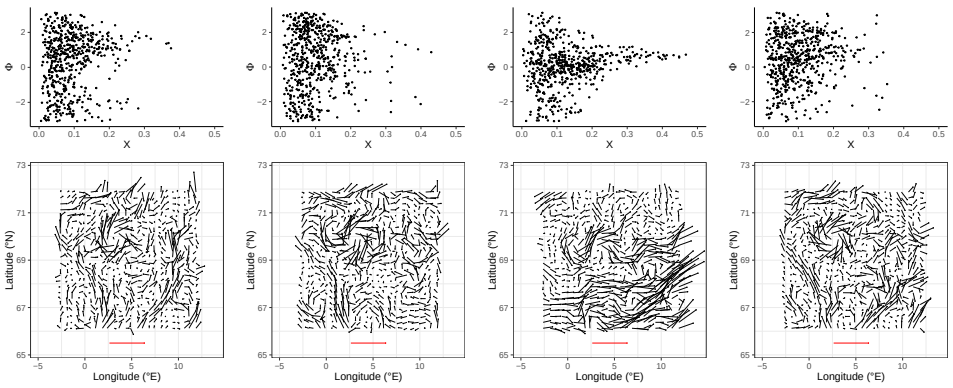


Figure 5.1: Map of the Norwegian Sea. The area in consideration is indicated by the red rectangle.

We want to go further by building a model that can classify typical current patterns in the Norwegian Sea. This is done by using synoptic snapshot images of larger areas, instead of time series from multiple single-site moorings. The aim is to build a compact model that can be run on-board AUVs or simple autonomous buoys in real time, similarly to the works of, e.g., Fossum et al. [2019], or Holm [2020]. This puts constraints on the complexity of the model. Hence, instead of looking at the entire water column, we only consider the currents at the surface layer, and use the spatial model to borrow information from several points in space to form classes with similar currents. These classes, or distinct current regimes, are then allocated to positions in space for snapshots of the constantly changing current circulation. Then, the OSC can at all times be described by the same current patterns, with the spatial configuration of the classes changing with time.

Data set description: 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 $28\,\mathrm{km}$. However, due to the high latitude, distance between grid cells in the west–east is only approximately $11\,\mathrm{km}$. 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 and use only 2 out of every 5 observation sites, i.e., we use only longitudinal grid point 1, 3, 6, 8, and so forth. As a consequence, the grid points will not be evenly spaced, but it serves as a good approximation. This gives a total size of the area of $667\,\mathrm{km}$ in the south–north direction and $676\,\mathrm{km}$ in the west–east direction.



(a) 2005 summer **(b)** 2015 summer **(c)** 2005 fall **(d)** 2015 fall

Figure 5.2: 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 $0.5\,\mathrm{m/s}$.

With a spatial resolution of 0.25 degrees, this means that the area is of size 24×24, and for each data set there are $n = 576$ observations. We consider data from four dates taken from two different decades and two seasonal periods. For 2005 and 2015 we extract data on July 5 at 12.00 and denote this by summer. We also have data from October 3 at 12.00, and these are denoted by fall. In conclusion, we get data from four dates, and these are displayed in Figure 5.2. The scatter plots on the top display the surface current speeds on the first axis, measured in $\mathrm{m/s}$. The second axis contains the angle of the surface current direction, with an angle of $\Phi = 0$ implying a current straight east and an angle of $\Phi = \pi/2$ implying a current pointing straight north. Observe that the speeds vary from around $0\,\mathrm{m/s}$ to $0.5\,\mathrm{m/s}$. These measurements comply with typical current speeds in the area, with, e.g., Johannessen et al. [2014] reporting mean surface current speeds of $0.2\,\mathrm{m/s}$ for the Norwegian Sea. Ingvaldsen et al. [2002] report that the maximal current velocities in the Norwegian Atlantic Current are higher during fall than during summer. While this is true for the data from 2005, the same can not be said for the 2015 data.

In the bottom part of the display, the surface currents are plotted as vectors on a grid. The arrows point in the direction of the current at the corresponding grid point, and the length of the arrows represent the current speeds. The vector displays give an overview of the spatial structure of the surface current 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, i.e., angle $\Phi = 0$, with speed $0.5\,\mathrm{m/s}$ beneath the spatial grid.

From the displays of the vector fields, we clearly observe the formation of eddies. Most prominently, there is an eddy roughly between $0°$E and $5°$E longitude and $69°$N and $71°$N latitude for all dates. Eddies are frequently occurring in this area, and similar patterns have been observed in earlier studies, using both drifters [Poulain et al., 1996] and mooring buoys [Ingvaldsen et al., 2002]. As described in Section 1.1, the Norwegian Atlantic Current splits into two branches at the opening to the Barents Sea, with one going east and the other north. Investigations into this bifurcation with numerical mod-

els suggests that 65–70% of the flow goes north toward Spitsbergen, and 30–35% of the flow goes eastwards into the Barents Sea [Ingvaldsen, 2003]. These proportions changed during winter to approximately 30% northwards and 70% eastwards, and clearly there are seasonal variations. In the displays of the vector fields, the Barents Sea entrance is close to the north–east corner. Observe that a major part of the flow in this corner points north, with less flow going east.

Two strategies: The problem of this thesis concerns classifying states of the ocean for different seasonal periods or specific dates. As noted by both Ingvaldsen [2003] and Kwok et al. [2013], the current circulation patterns in the Norwegian Sea change with the seasons. We implement and study two different strategies for capturing the seasonal variations in the OSC. First, we estimate two separate models to characterize the surface currents during summer and fall. Hence, we consider data from the same season in 2005 and 2015 to be two independently observed data sets that have been generated from the same seasonal model. This means that we estimate one model for both data sets from the same season. Observations from summer 2005 are denoted by $\mathbf{z}^{\mathrm{S}}_{2005}$ and fall 2005 by $\mathbf{z}^{\mathrm{F}}_{2005}$. The same is done for the observations from 2015. Because the observations from different years are assumed independently drawn from the same model, the total log-likelihood for each seasonal model is a sum of the log-likelihood for the two individual years, or

$$\begin{aligned} l(\boldsymbol{\theta}, \rho | \mathbf{z}^{\mathrm{S}}_{2005}, \mathbf{z}^{\mathrm{S}}_{2015}) &= l(\boldsymbol{\theta}, \rho | \mathbf{z}^{\mathrm{S}}_{2005}) + l(\boldsymbol{\theta}, \rho | \mathbf{z}^{\mathrm{S}}_{2015}), \\ l(\boldsymbol{\theta}, \rho | \mathbf{z}^{\mathrm{F}}_{2005}, \mathbf{z}^{\mathrm{F}}_{2015}) &= l(\boldsymbol{\theta}, \rho | \mathbf{z}^{\mathrm{F}}_{2005}) + l(\boldsymbol{\theta}, \rho | \mathbf{z}^{\mathrm{F}}_{2015}). \end{aligned} \tag{5.1}$$

To estimate the parameters of each seasonal model, this log-likelihood is optimized by the hybrid algorithm described in Chapter 4.

The second strategy implements a single model based on all the observations. The seasonal variability in the currents is then 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. The four dates are assumed independent, implying that the total log-likelihood of the single model is a sum of the four log-likelihoods,

$$l(\boldsymbol{\theta}, \rho | \mathbf{z}^{\mathrm{S}}_{2005}, \mathbf{z}^{\mathrm{S}}_{2015}, \mathbf{z}^{\mathrm{F}}_{2005}, \mathbf{z}^{\mathrm{F}}_{2015}) = \sum_{\mathbf{z}' \in \{\mathbf{z}^{\mathrm{S}}_{2005}, \mathbf{z}^{\mathrm{S}}_{2015}, \mathbf{z}^{\mathrm{F}}_{2005}, \mathbf{z}^{\mathrm{F}}_{2015}\}} l(\boldsymbol{\theta}, \rho | \mathbf{z}'). \tag{5.2}$$

The total log-likelihood is again optimized by the hybrid algorithm to estimate the model parameters.

5.2 Seasonal model

Oceans are always in motion, and as we previously noted, the current patterns change with the season. Hence, the first methodology we consider is to develop separate models for the two seasons in question. The variability in currents from season to season is then accounted for by each seasonal model having different current classes.

5.2.1 WSSVM

Model estimation: The cylindrical HMRF model is first estimated with a WSSVM density for the OSC observations. Recall from Section 2.2.1 that the WSSVM is a flexible distribution, designed to incorporate skewness in the circular part. It is not evident from the OSC observations in Figure 5.2 that they possess skewness. Nevertheless, skewness is commonly observed in cylindrical data and the WSSVM distribution is a natural choice.

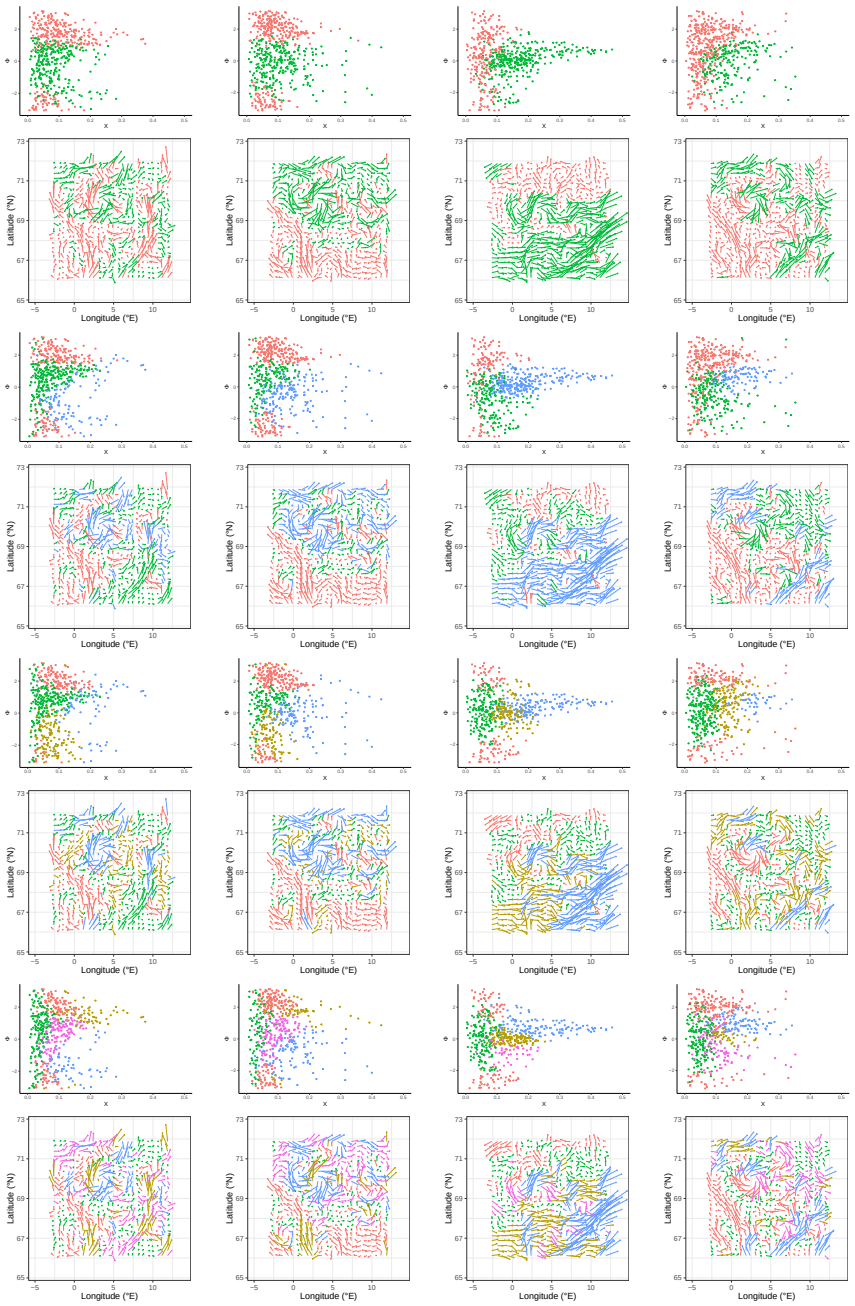
Because the number of ocean states, i.e., the number of latent classes K, is unknown, it needs to be decided. We estimate four different models by varying the number K of latent classes from 2 to 5. The hybrid algorithm starts by optimizing 50 random starting points with the EM algorithm. The initial parameters are drawn uniformly with the following boundaries,

$$\begin{aligned} 0 \leq &\rho^0 \leq \rho_{\text{crit}}, \\ 0.9 \leq &\alpha_k^0 \leq 3, \\ 5 \leq &\beta_k^0 \leq 25, \\ -\pi \leq &\mu_k^0 \leq \pi, \\ 0 \leq &\kappa_k^0 \leq 3, \\ -0.9 \leq &\lambda_k^0 \leq 0.9, \end{aligned} \tag{5.3}$$

for each $k \in \{1, \ldots, K\}$. The large number of short runs makes the algorithm start from a broad range of initial parameters, and the boundaries are selected empirically to ensure that the algorithm does not converge to a local optimum. Yet, there is no guarantee for reaching the global maximum after the second part of the algorithm. Unfortunately, we have observed in some cases that the parameters maximizing the pairwise-likelihood after the EM algorithm do not always result in the highest block-likelihood. In some cases, starting from a point with lower pairwise-likelihood actually results in a higher block-likelihood. As such, care needs to be taken when finding the parameter estimates, especially as K increases.

Figure 5.3 displays results from the model fit with K ranging from 2 to 5. This figure is intended to give an overview over how varying K alters the latent class predictions. The colour of the observations and current vectors represents the latent class prediction with a maximum posterior prediction criterion. Again, this means that the class with the largest marginal probability, given all the observations, is predicted. These probabilities are given by Equation (3.13). The figure shows how the segmentation develops as K goes from 2 to 5. Notice that there are overall fairly large areas with equal latent class predictions. This means that the spatial dependency parameter ρ is large, and that the OSC is fairly homogeneous in large areas.

Model selection: We utilize the information criteria presented in Section 3.3 to set K. Table 5.1 presents the block log-likelihood, C-AIC and C-BIC values for all K and both seasons. The best model (lowest information criterion) is indicated by bold text. Observe that $K = 5$ yields the lowest C-AIC for both seasons, whereas $K = 2$ gives lowest C-BIC for summer and $K = 3$ for fall. As already discussed, the C-AIC is designed for prediction power and penalizes model complexity mildly. Hence, it tends to favour complex models.



(a) 2005 summer **(b)** 2015 summer **(c)** 2005 fall **(d)** 2015 fall

Figure 5.3: Resulting prediction of the latent classes with K ranging from 2 (top) to 5 (bottom) with the WSSVM density. A maximum probability prediction criterion is used to predict the classes. The latent classes are indicated by the colour of the points/arrows.

Table 5.1: Computed values for the block log-likelihood (bl) defined in Equation (3.34), C-BIC from Equation (3.45) and C-AIC from Equation (3.44) for the summer and fall model with the WSSVM density. The best model is indicated by **bold**.

	Summer			Fall		
K	bl	C-BIC	C-AIC	bl	C-BIC	C-AIC
2	-762.3	**2026.4**	1675.5	-976.1	2515.1	2112.0
3	-559.0	2347.1	1466.7	-737.7	**2212.1**	1684.4
4	-445.4	2412.8	1322.6	-666.2	2511.3	1666.8
5	-331.7	2182.4	**1094.3**	-499.7	2595.3	**1452.2**

Our goal is not to develop a model with good prediction power for individual surface current observations, but rather a model that parsimoniously represents states of the ocean. Consequently, we prefer C-BIC and decide $K = 2$ for summer and $K = 3$ for fall.

By inspecting the displays of the observations in Figure 5.2 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 in one direction. The same pattern is not observed for 2015. Hence, it makes sense that the model includes more latent classes for fall to account for more current patterns.

Model properties: Now that we have made a decision for the number of latent classes, we can move on to visualizing the models and discussing their properties. In Figure 5.4, contour plots of the densities corresponding to the latent classes are displayed. On top of the densities we also show the OSC observations, with observations from 2005 plotted as dots and 2015 plotted as pluses. The transparency of the observations represents how likely they are to take the latent class. With parameter estimates $\hat{\rho}$ and $\hat{\boldsymbol{\theta}}$, the probability of belonging to class k for grid point t is $p_{\hat{\boldsymbol{\theta}},\hat{\rho}}(l_t = k|\mathbf{z})$. Black points indicate $p_{\hat{\boldsymbol{\theta}},\hat{\rho}}(l_t = k|\mathbf{z}) = 1$, and for $p_{\hat{\boldsymbol{\theta}},\hat{\rho}}(l_t = k|\mathbf{z}) = 0$ the point is completely transparent.

This figure should be interpreted together with the maximum-block-likelihood estimates of the model parameters that are presented in Table 5.2. The summer 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. Some predictions of this class are also scattered across the rest of the area, mostly for 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 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

this means that the circular-linear dependence is lower for the second density, which is also observable from the displayed densities. For the 2015 data set there is a large patch displaying this current pattern in the northern part of the spatial domain. The 2005 data set contains several smaller areas with this latent class across the spatial grid. From the density display we further observe that there are many outliers to the two distributions. This suggests that this compact model may be under-fitting, and that it is too simple to capture all realistic current patterns.

The fall observations, on the other hand, are described by three distributions. Notably, density 1 and 2 for summer and fall appear fairly similar, and the parameter estimates are much alike. Even though we did not enforce any physical structure in the modelling, the algorithm 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. Density 1 is most prominent in the north–east corner for 2005 and south–east corner of the 2015 data set, whereas density 2 dominates the north–east corner for 2015.

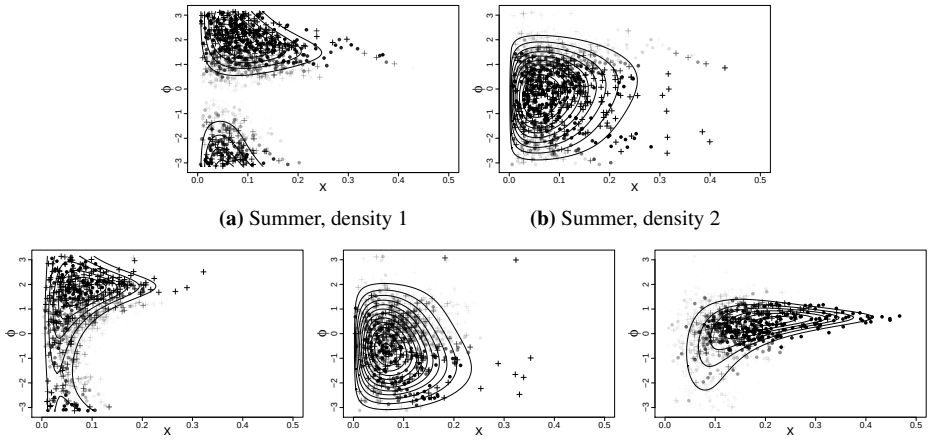
The fall model also includes an additional third density to account for the large current speeds observed in the south–east corner of the 2005 fall observations, and also to some extent in the south–east and north–west corners of 2015. In these corners, the OSC are highly concentrated in the north–east direction with high speeds. This makes the circular concentration of the third density very high ($\kappa_3 = 1.5$), around a north–east modal 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.

The current pattern described by this density, i.e., currents flowing north–east, represents an inflow to the Barents Sea. Ingvaldsen [2003] showed that the inflow of Atlantic Water is stronger during winter and fall than during summer. Our results agree with this statement. Because the inflow is stronger during fall than during summer, the fall model includes an additional density that represents high-speed currents flowing north–east.

Figures 5.3 and 5.4 show that the summer model largely splits the observations based on direction. Observations with low absolute angles are likely to take latent class 2, and observations with high absolute angle are likely to take latent class 1. The same can not be said for the fall model. Here, observations with roughly the same angle are predicted to take all three different classes based on their speed and the spatial dependence structure. Still, although to a lesser extent than for the summer model, we observe that the first two densities in the fall model split the data set into small and large absolute angles. The third density corresponds to high speeds concentrated in the north–east direction.

From the displays in Figure 5.4 we also see the spatial coupling come into play. Clearly, the observations in the centre of a cylindrical density have high probability of belonging to that class, and are consequently not transparent. However, we also observe that some observations with low cylindrical density are less transparent than observations with higher cylindrical density. This means that even though the cylindrical density is low, the probability of belonging to that class is high, due to the spatial dependency.

Model uncertainty: We seek to investigate the uncertainty in the parameter estimates. Recall from Section 3.3 that the maximum-block-likelihood estimates are asymptotically normally distributed with variance as the inverse of the Godambe matrix. This means that



(a) Summer, density 1 **(b)** Summer, density 2

(c) Fall, density 1 **(d)** Fall, density 2 **(e)** Fall, density 3

Figure 5.4: 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.

theoretically, standard errors of the parameter estimates can be obtained by approximating the observed Godambe matrix numerically . Unfortunately, this matrix suffers from numerical instability, specifically the variability matrix is difficult to estimate because it vanishes at the maximum-block-likelihood estimate. A workable alternative is to use parametric bootstrap to estimate empirical quantiles for the distribution of parameter estimates. These quantiles are used to assess uncertainty in the maximum-block-likelihood estimates.

To perform parametric bootstrap, we simulate observations from the estimated models following a similar procedure as in the previous chapter: First, latent classes are simulated and then observations are drawn from the cylindrical density corresponding to the latent class. For both models we have $\rho > \rho_{\text{crit}}$. As a result, it is not realistic to sample a true latent field directly using the Swendsen–Wang algorithm. Because of the inherent phase transition of the Potts model, sampling with spatial dependence parameter larger than the critical value produces fields that have almost all grid points taking the same latent class. This does not accurately represent bootstrap samples of the true observations. Instead, we use the forward-backward algorithm to sample latent classes, as presented in Section 3.1. This algorithm allows for sampling of latent classes for each block. However, each block is assumed independent, so the dependence between blocks is lost. Hence, to increase the dependence without making a complicated algorithm we do the following. First, one horizontal block is simulated at the bottom of the lattice. Then, all vertical blocks are simulated, starting from the already simulated bottom grid point. In this way, the only grid point that is completely independent is the bottom left corner, compared to the even simpler procedure of sampling 24 horizontal or 24 vertical independent blocks. We believe this approximate sampling procedure is sufficient for our purposes.

For each bootstrap sample, two latent fields are sampled, one based on the 2005 ob-

Table 5.2: Parameter estimates and bootstrap quantiles of the best model for each seasonal period with the WSSVM density.

	Summer			Fall		
Parameter	2.5% quantile	Estimate	97.5% quantile	2.5% quantile	Estimate	97.5% quantile
α_1	1.63	1.94	2.00	1.66	1.95	1.96
β_1	9.67	12.17	13.53	14.01	17.02	17.12
μ_1	0.15	1.32	1.59	1.85	1.95	2.08
κ_1	0.69	0.80	0.92	0.51	1.02	1.37
λ_1	0.26	1.00	1.00	-0.57	-0.26	-0.18
α_2	1.48	1.58	1.74	1.57	1.78	1.88
β_2	7.94	9.12	12.03	9.40	9.69	11.70
μ_2	0.02	1.21	2.02	-3.13	-2.08	-1.57
κ_2	0.01	0.13	0.32	0.00	0.18	0.30
λ_2	-1.00	-1.00	-1.00	0.59	1.00	1.00
α_3				2.28	2.85	3.03
β_3				7.53	8.61	8.78
μ_3				0.58	0.69	0.72
κ_3				0.83	1.50	1.73
λ_3				-1.00	-1.00	-0.99
ρ	1.58	1.75	1.90	1.90	2.39	3.01

servations and one based on the 2015 observations. Having sampled the latent classes, observations are then drawn from the corresponding WSSVM density with the maximum-block-likelihood parameters. These parameters are used as initial values and new parameter estimates are found for the simulated sample. In this thesis we use 200 bootstrap samples. The empirical 2.5% and 97.5% quantiles, along with the maximum-block-likelihood estimates, are displayed in Figure 5.2.

Observe from the quantiles that all circular-linear dependence parameters κ are statistically significant, except for κ_2 in the fall model. This supports the decision to model the data by a cylindrical distribution. Hence, assuming independence between the current angle and speed is not appropriate, and these variables should be modelled jointly. Further, for both models and each density the skewness parameter λ is significant. This indicates that the current data are skewed and justifies our choice of a skewed cylindrical distribution. Also worth noting is that the spatial dependence parameter ρ is significant in both models, substantiating the need for a hidden spatial process to explain the variability and dependence between current measurements. Clearly, the current observations are tightly connected in the spatial domain. Finally, the spatial dependence parameter is larger in fall than in summer. This implies that the currents in summer are more heterogeneous than in fall in the spatial domain, but more homogeneous in the cylindrical domain as the summer model includes only two separate densities.

We also observe that some of the uncertainty intervals for the parameter λ are very narrow. Recall that the optimization procedure estimates a transformed version of the

parameters, given in Equation (3.32). The narrow intervals come as a result of the transformation back to the original parameters and the behaviour of the hyperbolic tangent function when the estimate for λ is close to the limits $\lambda \in [-1, 1]$.

In Section 2.1.1, we discussed the phase transition in the Potts model for values of ρ above $\rho_{\text{crit}} \approx \log(1 + \sqrt{K})$. For values above this threshold, almost all grid cells take the same latent class. Clearly, the estimated spatial dependence parameters are above this threshold for both seasons. Still, we observe from the latent class predictions in Figure 5.3 that the spatial domain appears with mixing of the latent classes, rather than all equal. This is because of the bias in the spatial interaction parameter discussed in Chapter 4. Here we observed that both the bias and the variance in the parameter estimation increase dramatically when the true value is close to the critical value. With a true value of $\rho = 0.95$ we observed a mean of 2.14 in the predictions. Hence, we conclude that for both models the true value of the spatial dependence parameter must be close to the critical value, but not above the threshold, as the latent classes appear with some mixing.

5.2.2 GPTWC

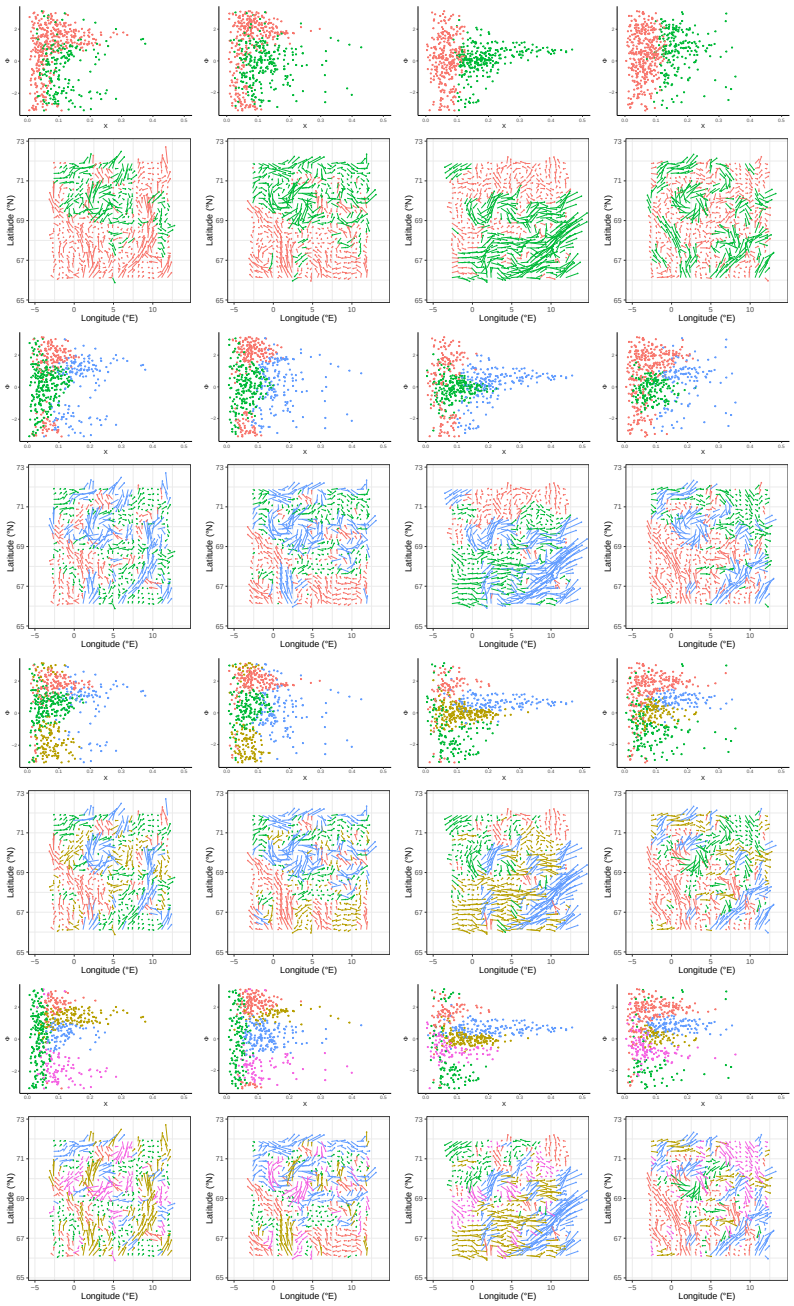
Model estimation: Figure 5.4 showed that there were several outliers in the linear parts of the estimated skewed densities. To deal with these outliers, we estimate a cylindrical HMRF model with the GPTWC density. This density can account for heavy tails in the linear part, but not skewness. We explore whether this model is more suitable for the data or not.

Again, the number of latent classes needs to be decided, so the model is estimated with K varying from 2 to 5. The initial parameters for the hybrid algorithm are drawn uniformly with boundaries

$$\begin{aligned} 0 \leq &\rho^0 \leq \rho_{\text{crit}}, \\ 0.3 \leq &\alpha_k^0 \leq 1.1, \\ 0.04 \leq &\beta_k^0 \leq 0.2, \\ -\pi \leq &\mu_k^0 \leq \pi, \\ 0 \leq &\tau_k^0 \leq 3, \\ 0.5 \leq &\kappa_k^0 \leq 0.99, \end{aligned} \tag{5.4}$$

for each $k \in \{1, \ldots, K\}$. As for the WSSVM distribution, we use 50 random starting points.

Figure 5.5 displays results from the parameter estimation in a similar fashion as was done for the WSSVM density. Latent classes are predicted with a maximum posterior prediction criteria for the observations and vectors mapped to the spatial grid for both seasons. We observe that the predicted latent classes are remarkably similar to those achieved with the WSSVM density in Figure 5.3. The biggest differences are for $K = 5$ in both seasons and $K = 4$ for the fall model. Also for the GPTWC model, large areas with equal latent class predictions are formed, suggesting a homogeneous current field and large spatial coupling.



(a) 2005 summer **(b)** 2015 summer **(c)** 2005 fall **(d)** 2015 fall

Figure 5.5: Resulting prediction of the latent classes with K ranging from 2 (top) to 5 (bottom) with the GPTWC density. A maximum probability prediction criterion is used to predict the classes. The latent classes are indicated by the colour of the points/arrows.

Table 5.3: Computed values for the block log-likelihood (bl) defined in Equation (3.34), C-BIC from Equation (3.45) and C-AIC from Equation (3.44) for the summer and fall model with the GPTWC density. The best model is indicated by **bold**.

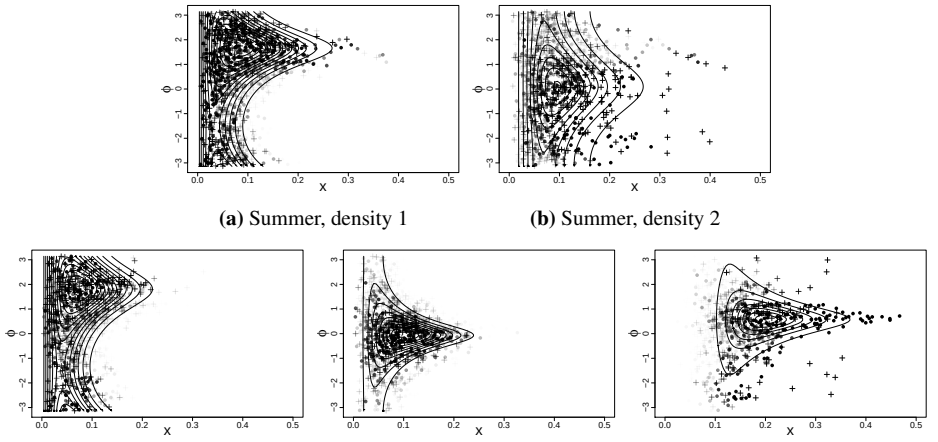
	Summer			Fall		
K	bl	C-BIC	C-AIC	bl	C-BIC	C-AIC
2	-878.1	**2164.7**	1872.1	-1040.4	2614.9	2232.3
3	-727.8	2382.8	1718.6	-821.2	**2494.1**	1884.1
4	-543.5	2406.9	1461.4	-589.7	2630.9	1591.2
5	-440.2	2739.8	**1407.9**	-494.8	2626.3	**1454.0**

Model selection: The block log-likehood, C-BIC, and C-AIC values are collected in Table 5.3 for all values of K and both seasons. The results are identical to the WSSVM density. The lowest C-AIC is achieved with $K = 5$ for both seasons, and $K = 2$ and $K = 3$ give the lowest C-BIC for summer and fall, respectively. This again verifies that the fall observations are more heterogeneous in the cylindrical domain because an additional density is required to model the data accurately. We select the number of latent classes by minimizing the C-BIC similarly to the WSSVM density.

Model properties: Figure 5.6 displays contour plots of the densities corresponding to the latent classes of the best model for summer and fall. Again, observations from 2005 are plotted as dots and observations from 2015 are plotted as pluses. Transparency of the observations also represent the probability of belonging to the given latent class. Further, the maximum-block-likelihood estimates of the model parameters are shown in Table 5.4 and the latent class predictions in Figure 5.5. The densities should also be compared to the densities achieved by the WSSVM density in Figure 5.4. Overall, it is observed that the densities produced by the GPTWC density resemble the densities produced by the WSSVM density to a great extent.

For the summer data, the best model includes $K = 2$ latent classes. The first density is associated with large absolute angles, similarly to the first density of the WSSVM model, but also includes observations of lower speed with lower absolute angles. Noticeably, 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 roughly corresponds 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 to this density with a non-negligible probability of belonging to this density. The first density is mostly observed in the southern parts of both the 2005 and 2015 data set, with some smaller patches further north in 2005.

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 de-



(a) Summer, density 1 **(b)** Summer, density 2

(c) Fall, density 1 **(d)** Fall, density 2 **(e)** Fall, density 3

Figure 5.6: 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.

gree of heavy tailedness ($\tau_2 = 0.27$) with more non-transparent outliers to the distribution. This density dominates the northern parts for both years, with 2015 displaying larger patterns of this density than 2005. The GPTWC density is designed to deal with outliers in the linear part, but the fitted model includes heavy tailedness purely in the second density. For the WSSVM summer model, both densities were significantly skewed, implying that it may be more beneficial to include skewness than linear heavy tails to model the data.

Also for the GPTWC model, we observe 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 feat without any prior knowledge. The biggest differences between the first two densities for the summer and fall model is that the second density in the fall model displays little heavy tailedness ($\tau_2 = 0.06$) and the circular concentration of the second density is larger ($\kappa_2 = 0.88$). Apart from these two parameters, the two pairs of densities are very similar. As was the case with the WSSVM density, the first density dominates the north–east corner of the 2005 observations and the south–west corner of the 2015 observations. The allocation of the second density is also fairly similar to the WSSVM density, with the north–eastern corner of the 2015 data set taking this class.

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, centred 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. This density is prevalent in the

Table 5.4: Parameter estimates and bootstrap quantiles of the best model for each seasonal period with the GPTWC density.

	Summer			Fall		
Parameter	2.5% quantile	Estimate	97.5% quantile	2.5% quantile	Estimate	97.5% quantile
α_1	0.47	0.52	0.56	0.43	0.48	0.53
β_1	0.06	0.07	0.07	0.06	0.07	0.07
μ_1	1.58	1.66	1.78	1.64	1.81	1.98
τ_1	0.00	0.03	0.11	0.00	0.04	0.13
κ_1	0.68	0.72	0.78	0.55	0.62	0.70
α_2	0.32	0.36	0.46	0.32	0.37	0.42
β_2	0.09	0.10	0.11	0.05	0.06	0.06
μ_2	0.01	0.10	0.54	-0.14	-0.07	0.01
τ_2	0.15	0.27	0.33	0.00	0.06	0.16
κ_2	0.31	0.45	0.52	0.84	0.88	0.91
α_3				0.16	0.18	0.24
β_3				0.14	0.15	0.16
μ_3				0.48	0.58	0.66
τ_3				0.15	0.21	0.25
κ_3				0.82	0.87	0.89
ρ	1.86	2.47	2.97	2.06	2.27	2.46

south–east corner of the 2005 observations, and also for smaller patches in the south–east and north–west corners of 2015, much like the WSSVM third density.

Model uncertainty: To investigate uncertainty in the parameter estimates, we apply a similar parametric bootstrap procedure as was done for the WSSVM density. The resulting parameter quantiles are displayed in Figure 5.4. Note first that also for this model, all circular-linear dependence parameters κ are statistically significant. This supports the use of a cylindrical density to model the data. Interestingly, for both seasons only one density is significantly heavy tailed, and it is the distribution associated with the largest speeds that is heavy tailed. This suggests that it may not be necessary to use a heavy tailed distribution to model the data. For the WSSVM density, all skewness parameters were significant, indicating that it makes more sense to use a skewed distribution rather than a heavy tailed distribution. Furthermore, the spatial dependence parameter ρ is significant in both models, supporting our decision to include a hidden spatial process in the model. Contrary to the WSSVM model, the summer model has higher spatial dependence than the fall model with the GPTWC density. Hence, the GPTWC model indicates that the fall observations are more heterogeneous in the spatial domain than the summer observations, the opposite of what was indicated by the WSSVM model, but the results are not very significant.

5.3 Single model

In the previous section, we saw that the first two densities of the summer and fall model were very similar for both cylindrical densities. The model identified this trait, even though we did not supply any prior information about the current regimes and the data sets were collected independently. This motivates us to estimate a single model that accounts for the observable circulation patterns in both seasons simultaneously. Then, the seasonal fluctuations can be captured by the allocation of classes, rather than the classes themselves varying. This gives a less complex model.

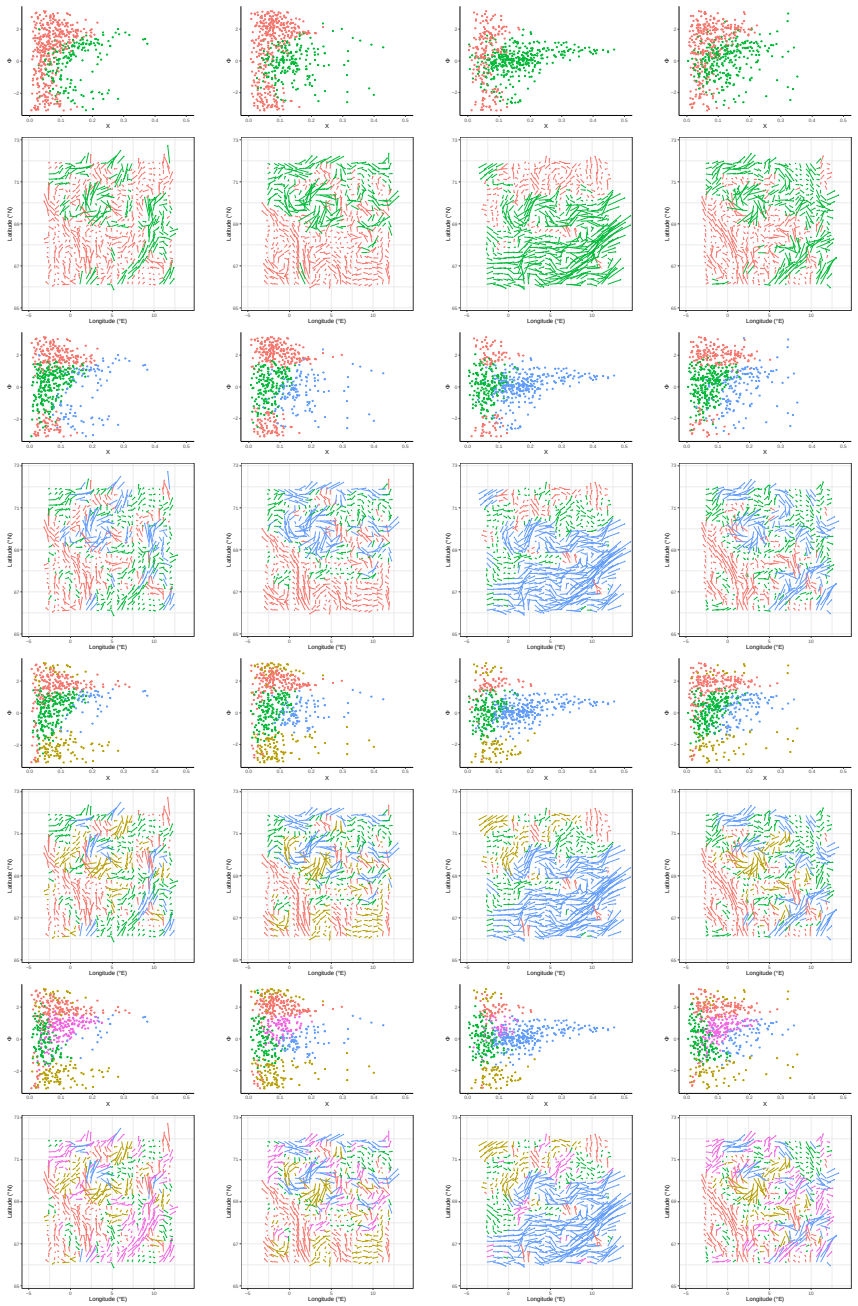
5.3.1 WSSVM

Model estimation: For the seasonal models, we saw that all skewness parameters were significant. On the other hand, only one density in each of the GPTWC models were significantly heavy tailed. Thus, we start by implementing a model with the skewed WSSVM density. Parameters are estimated using the hybrid algorithm, and we draw 50 sets of initial parameters using the same boundaries as in Equation (5.3). The model is estimated with the number of latent classes K varying from 2 to 5.

Figure 5.7 displays the latent class predictions with K ranging from 2 to 5. Notice the striking similarity between this display and the corresponding display with the seasonal model in Figure 5.3, especially for small values of K. Both the cylindrical observations and the spatial locations are classified similarly for the two models. This further suggests that we do not need to specify separate models for the two seasons.

Model selection: Table 5.5 displays the total block log-likelihood, C-BIC and C-AIC values for all values of K. Observe that for all models, the total block log-likelihood is strictly smaller than the sum of the block log-likelihoods for the summer and fall model with the same number of latent classes. The seasonal models were fitted to smaller data sets than the single model, thus providing better fits and larger block log-likelihood values than the single model. However, the larger block log-likelihood comes at the cost of two models instead of one. The C-BIC values suggest that $K = 2$ classes are sufficient to parsimoniously represent the various circulation patterns. The seasonal model included two classes for the summer observations and 3 classes for the fall observations. For the single model, though, adding a third density is penalized more than what is achieved in terms of additional explanatory power.

Model properties: Figure 5.8 displays contour plots of the two cylindrical densities corresponding to the latent classes. In this display, observations from 2005 are plotted as dots, observations from 2015 are plotted as pluses, observations from summer are black and observations from fall are red. Transparency of the observations represents the probability of belonging to that class, as before. The densities should be interpreted together with the maximum-block-likelihood estimates of the model parameters in Table 5.6 and the latent class predictions in Figure 5.7, and compared to the densities from the seasonal model in Figure 5.4. Certainly, these two densities compare very well to the first two densities of the summer and fall model.

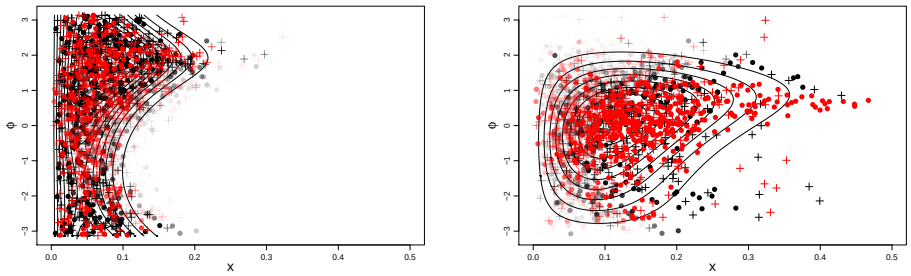


(a) 2005 summer **(b)** 2015 summer **(c)** 2005 fall **(d)** 2015 fall

Figure 5.7: Resulting prediction of the latent classes with K ranging from 2 (top) to 5 (bottom) with the WSSVM density. A maximum probability prediction criterion is used to predict the classes. The latent classes are indicated by the colour of the points/arrows.

Table 5.5: Computed values for the block log-likelihood (bl) defined in Equation (3.34), C-BIC from Equation (3.45) and C-AIC from Equation (3.44) for the single model with the WSSVM density. The best model is indicated by **bold**.

K	bl	C-BIC	C-AIC
2	-2000.8	**5228.2**	4318.4
3	-1597.5	5312.1	3741.9
4	-1266.9	5388.1	3271.1
5	-1063.5	5566.4	**3015.5**



(a) Density 1 **(b)** Density 2

Figure 5.8: Estimated densities of the best single model with the WSSVM density. The observations from 2005 and 2015 are plotted as dots and pluses respectively and the observations from summer and fall as black and red, respectively. Transparency of each dot/plus represents the probability of belonging to that class. No transparency represent a probability of 1 of belonging to that class.

The first density represents currents with low speed, and also currents pointing north–west with higher speeds. The density has modal direction to the north–west ($\mu_1 = 1.93$) and is slightly negatively skewed ($\lambda_1 = -0.11$), with modest circular concentration ($\kappa_1 = 0.69$). The density is naturally very similar to the first density of both the GPTWC and WSSVM model for both seasons. Compared to the first density of the seasonal WSSVM models, this density has a lower circular concentration. This enables it to capture low speed currents in all directions, and not only the currents close to the modal direction. From the latent class predictions, we see that this density dominates the southern part of the summer observations, the northern part of the 2005 fall observations, and the south–west corner of the 2015 fall observations. These are areas where the speed is low, with a few high-speed currents flowing north.

The second density is associated with currents of higher speeds. The circular concentration is low ($\kappa_2 = 0.47$), meaning that the directions are fairly spread out, with a north–east modal direction ($\mu_2 = 1.17$). However, the density is negatively skewed ($\lambda_2 = -1.00$), implying that the density also includes currents flowing east and south–east. Consequently, this density contains currents on the eastern semicircle, with higher speeds than the first density. Furthermore, to include the high-speed currents, the linear scale parameter is low ($\beta_2 = 6.78$). For both the seasonal models, the high-speed observations from 2005 fall were captured by including a third density in the fall model. In

Table 5.6: Parameter estimates and bootstrap quantiles of the best model with the WSSVM density.

Parameter	2.5% quantile	Estimate	97.5% quantile
α_1	1.89	2.02	2.14
β_1	10.03	14.40	15.84
μ_1	0.22	1.93	2.44
κ_1	0.37	0.69	0.87
λ_1	-0.41	-0.11	0.29
α_2	1.83	1.99	2.17
β_2	5.68	6.78	10.51
μ_2	0.03	1.17	2.08
κ_2	0.28	0.47	0.57
λ_2	-1.00	-1.00	-1.00
ρ	1.51	2.18	2.95

this case, we only have two densities, and the high-speed observations are accounted for by lowering the linear scale parameter β_2, making the density include currents of higher speed. Hence, this density can be seen as a compromise between the second density of the summer and fall models and the third density in the fall model.

Critically, the second density contains several outliers with significant probability. For the first density, most visible observations (those with significant probability) are within the main density area displayed by the contours. This, however, is not the case for the second density, and several visible observations are outside the displayed contours. The large amount of outliers suggests that the model may be too simple to represent all possible circulation patterns. Even though the C-BIC advocated only two latent classes, better results would perhaps be achieved by adding a third class to include the high-speed currents.

Model uncertainty: Parameter quantiles achieved by a parametric bootstrap procedure, along with the maximum-block-likelihood estimates are displayed in Figure 5.6. Again, both circular-linear dependence parameters κ are statistically significant. However, we observe that the first skewness parameter λ_1 is not statistically significant. For the seasonal models all skewness parameters were statistically significant. Consequently, the data do not display the same degree of skewness when considering the seasons collectively, rather than separately. Finally, the spatial interaction parameter is significant, and its value is between the parameter estimate for the summer and fall WSSVM models.

5.3.2 GPTWC

Model estimation: Figure 5.8 showed a large number of outliers in density 2 of the single WSSVM model. Similarly to the seasonal models, we consider the single model with the GPTWC density to handle the outliers. Again, the hybrid algorithm is used with 50 sets of initial parameters drawn from the boundaries in Equation (5.4). Models are estimated with the number of latent classes K varying from 2 to 5. Latent class predictions for all

Table 5.7: Computed values for the block log-likelihood (bl) defined in Equation (3.34), C-BIC from Equation (3.45) and C-AIC from Equation (3.44) for the single model with the GPTWC density. The best model is indicated by **bold**.

K	bl	C-BIC	C-AIC
2	-2202.2	5841.4	4775.6
3	-1803.0	**5707.8**	4149.0
4	-1365.8	5781.7	3519.5
5	-1102.7	6008.0	**3187.7**

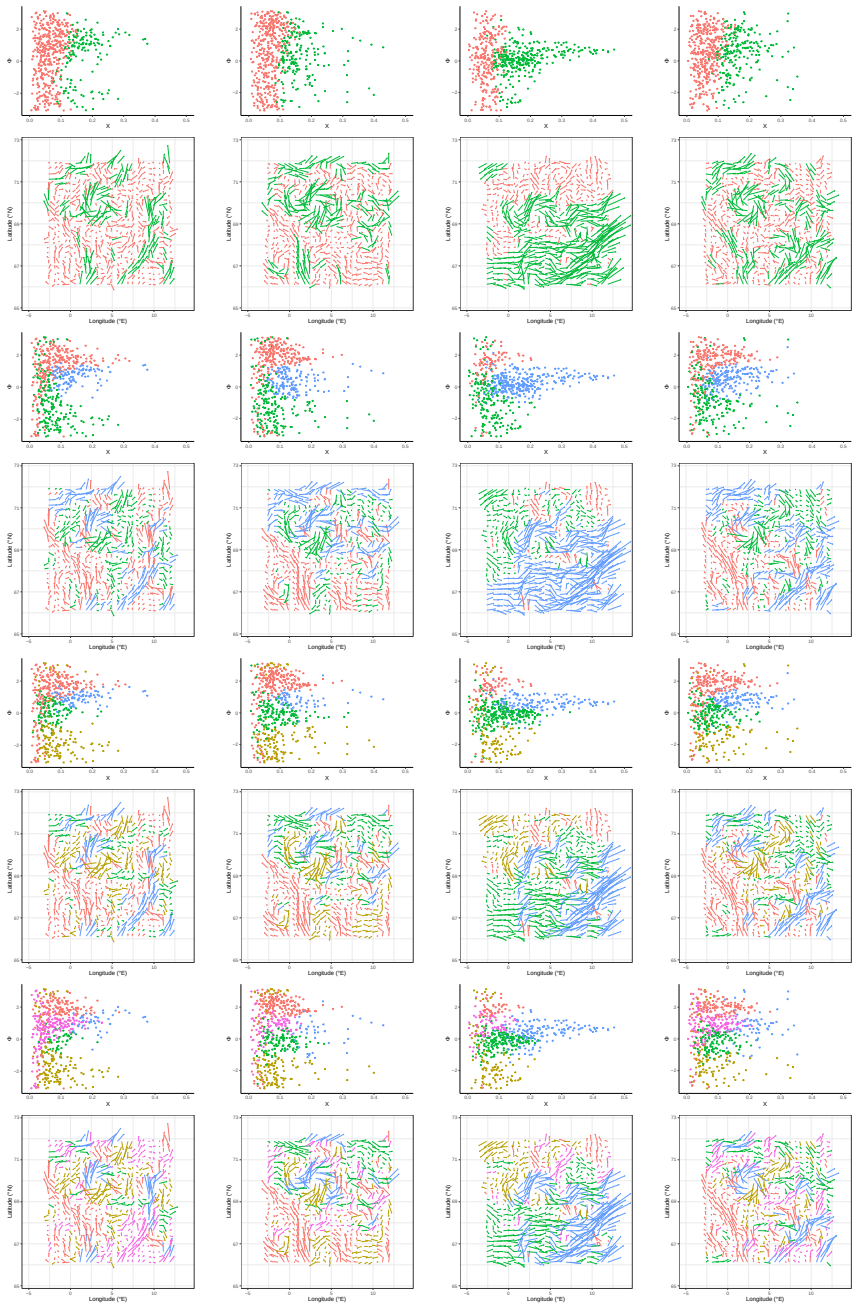
fitted models are displayed in Figure 5.9. Overall, the class predictions are similar to the ones obtained by the single WSSVM model. However, the clustering of the observations does not display the same skewed structure.

Model selection: Values for the block log-likelihood, C-BIC and C-AIC are displayed for all the models in Table 5.7. Again observe that these block log-likelihood values are smaller than the sums of the summer and fall models with the same number of latent classes. The lowest C-BIC is achieved by including $K = 3$ latent classes, which is different from the single WSSVM model. For the seasonal models, we saw that the WSSVM and GPTWC densities agreed on the number of latent classes, but this is not the case for the single model. Instead, the C-BIC suggests to add an additional density for the GPTWC model. This makes sense, considering the large number of outliers observed for the WSSVM model with $K = 2$.

Model properties: Contour plots of the three densities corresponding to the latent classes with $K = 3$ are displayed in Figure 5.10. Similarly to what was done earlier, the observations from 2005 are plotted as dots and observations from 2015 as pluses. The observations from summer are black, and observations from fall are red. The transparency of each observations represents the probability of belonging to the given class. The maximum-block-likelihood parameter estimates corresponding to the densities are displayed in Table 5.8, with latent class predictions in Figure 5.9. The first two densities are similar to the first two densities of both seasonal models with the GPTWC density. In addition, the first density resembles the first density of the single WSSVM model.

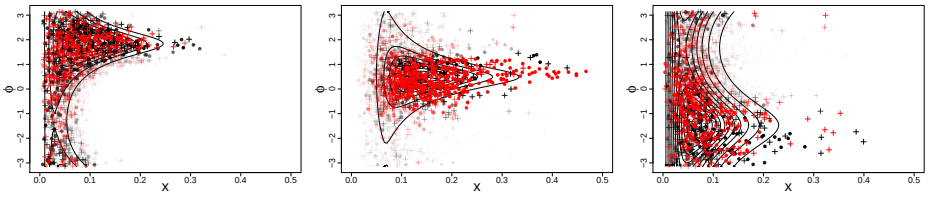
First and foremost, the first density accounts for currents flowing north–west with high speed, but it also includes currents with low speed, regardless of direction. Clearly, the modal direction is to the north–west ($\mu_1 = 1.84$), but the circular concentration is modest ($\kappa_1 = 0.85$). Comparing this density to the first density of the two GPTWC seasonal models, we see that they are remarkably similar, but the single model has a slightly higher circular concentration. By inspecting the latent class prediction, we observe that this density is prevalent in the southern part of the summer observations and the south–west corner of the 2015 fall observations. This density is hardly represented in the 2005 fall observations, but some small areas exist in the north–east corner.

The second density constitutes currents pointing in direction north–east and east, with higher speeds. The density serves as a middle ground between densities 2 and 3 from the



(a) 2005 summer **(b)** 2015 summer **(c)** 2005 fall **(d)** 2015 fall

Figure 5.9: Resulting prediction of the latent classes with K ranging from 2 (top) to 5 (bottom) with the GPTWC density. A maximum probability prediction criterion is used to predict the classes. The latent classes are indicated by the colour of the points/arrows.



(a) Density 1 **(b)** Density 2 **(c)** Density 3

Figure 5.10: Estimated densities of the best model with the GPTWC density. The observations from 2005 and 2015 are plotted as dots and pluses respectively and the observations from summer and fall as black and red, respectively. Transparency of each dot/plus represents the probability of belonging to that class. No transparency represent a probability of 1 of belonging to that class.

GPTWC fall model, and to a lesser extent density 2 from the GPTWC summer model. The circular concentration is approximately equal to the two fall model densities ($\kappa_2 = 0.89$). Also, the linear shape and scale and the circular location are all between the values for density 2 and 3 from the GPTWC fall model. The density is heavy tailed in the linear part ($\tau_2 = 0.27$), which enables it to take care of the many visible outliers. The density dominates the southern part of the 2005 fall observations, and smaller clusters of the density are scattered across spatial domain for the other data sets. However, the density is more present in the fall than during summer, which is due to the increased inflow to the Barents Sea that was discussed previously.

The third density comprises currents with direction on the southern semi-circle, i.e., clockwise from west to east. The speeds are higher than the first density, but generally lower than the second. The circular location is southward ($\mu_3 = -1.47$) with a low circular concentration ($\kappa_3 = 0.55$), making the density include a wide range of directions. This density is different from all the other estimated densities in that the modal direction is southward with relatively high speeds. Moreover, the density is heavy tailed in the linear part ($\tau_3 = 0.23$) to also capture the observed high-speed currents flowing south. The circulation pattern described by this density is present in small areas across the grid for the summer observations, and larger areas in the northern part of the domain for the fall observations.

Model uncertainty: To investigate the uncertainty in the parameter estimates, we display quantiles achieved by parametric bootstrap, as well as the maximum-block-likelihood estimates in Table 5.8. As before, we observe that all circular-linear dependence parameters κ and the spatial coupling parameter ρ are significant. Also, the heavy tail parameter τ is significant for both the second and third density. For the single WSSVM model, we saw that only one of the densities were significantly skewed, which may lead to the conclusion that the data display more heavy tailedness than skewness, and that the GPTWC distribution is better suited to model the data. This conclusion is the opposite of what was observed for the seasonal model. With the seasonal model, only one density of each model was heavy tailed, whereas all estimated densities were skewed.

Table 5.8: Parameter estimates and bootstrap quantiles of the best model with the GPTWC density.

Parameter	2.5% quantile	Estimate	97.5% quantile
α_1	0.44	0.48	0.51
β_1	0.05	0.05	0.06
μ_1	1.78	1.84	1.90
τ_1	0.00	0.03	0.08
κ_1	0.82	0.85	0.88
α_2	0.23	0.26	0.30
β_2	0.09	0.09	0.10
μ_2	0.45	0.49	0.55
τ_2	0.22	0.27	0.32
κ_2	0.87	0.89	0.92
α_3	0.39	0.44	0.49
β_3	0.06	0.07	0.07
μ_3	-1.64	-1.47	-1.33
τ_3	0.12	0.23	0.30
κ_3	0.44	0.55	0.62
ρ	2.04	2.22	2.40

5.4 Model comparison

In this section, we compare the performance of the estimated models. The models represent a probabilistic forecast of OSC observations in terms of cylindrical densities corresponding to latent spatial classes. Hence, to evaluate the performance of the probabilistic forecasts, we deploy the scoring rules presented in Section 3.4. Now, the models are designed to parsimoniously represent the typical current patterns and to segment the global circulation into less complex local regimes. Hence, they are not developed to make predictions of the OSC. Yet, there is some merit to comparing the predictive performance of the models as more accurate representations of the typical patterns lead to better predictions.

To carry out the model comparison, we randomly draw 50 grid points. The same grid points are used for all data sets. For all the selected grid points we compare the prediction of each model by evaluating a scoring rule. To evaluate the scoring rules, we first need to compute the predictive distributions for each model at the selected observation sites, and these are computed as in Equation (3.17). Once we have computed the predictive distributions, we use these in combination with the actual observations to compute CRPS for the linear and circular part of the observations separately. Table 5.9 displays the average linear CRPS for the 50 grid points for each model and each data set. Table 5.10 lists the average circular CRPS. Similar computations for other sets of 50 random grid points gave similar results. Hence, instead of computing the CRPS for all 576 grid points, it was deemed sufficient for our purpose to draw only 50 sites.

CRPS is designed such that lower values indicate better predictive distributions, i.e., better models. Lower CRPS is a result of either less bias in the prediction or a sharper predictive distribution. From the listed values, we first note that overall the WSSVM

Table 5.9: Average linear CRPS for each model and each data set. The model with the lowest CRPS is indicated with **bold** for each data set.

	Summer		Fall	
Model	2005	2015	2005	2015
Seasonal WSSVM	0.0074	0.0083	0.0114	0.0085
Seasonal GPTWC	**0.0073**	**0.0073**	**0.0104**	**0.0079**
Single WSSVM	0.0082	0.0089	0.0114	0.0093
Single GPTWC	**0.0073**	0.0077	0.0108	0.0082

Table 5.10: Average circular CRPS for each model and each data set. The model with the lowest CRPS is indicated with **bold** for each data set.

	Summer		Fall	
Model	2005	2015	2005	2015
Seasonal WSSVM	**0.1245**	**0.1298**	**0.1263**	**0.1211**
Seasonal GPTWC	0.1330	0.1461	0.1270	0.1289
Single WSSVM	0.1302	0.1373	0.1394	0.1356
Single GPTWC	0.1258	0.1305	0.1294	0.1290

models are better at predicting the circular part, whereas 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, whereas GPTWC is designed to account for heavy tails in the linear part.

Model complexity plays a crucial role in prediction performance. Both seasonal models include two latent classes in the summer model and three latent classes in the fall model, whereas the single WSSVM model includes two latent classes and the single GPTWC model includes three latent classes. Including more latent classes makes the model more complex, and hence better prediction performance is expected. For the single models, we observe that the GPTWC model has better prediction performance than the WSSVM model also for the circular part, and this is because the model is more complex. This clearly illustrates the trade-off between parsimonious models and accurate predictions. We also observe that the single GPTWC model has good performance on the summer data, especially for the circular part. This is because the three other models include only two latent classes for the summer data, and the performance of the single GPTWC model is thus enhanced by the inclusion of a third latent class, making it more complex.

Finally, we argue that estimating a single model for both seasons is favoured over estimating separate models. To make a fair comparison between the single and seasonal models, we need to compare the WSSVM models for the summer data and the GPTWC models for the fall data, as these have the same number of latent classes. We observe that the single GPTWC model performs slightly worse than the seasonal GPTWC model on the fall data, whereas the discrepancy between seasonal and single is somewhat larger for the WSSVM models. Still, we consider the added complexity of separate models for the two seasons as too large compared to the small gain in predictive power.

Chapter 6

SINMOD data

This chapter analyses an OSC data set on the border between the North Sea and the Norwegian Sea, closer to the Norwegian coast. The data are presented in Section 6.1, before the models are fitted. We apply the same methodology as in Chapter 5, by estimating a range of models and then using model selection criteria to decide the number of latent classes. The WSSVM density is considered in Section 6.2 and the GPTWC density in Section 6.3. Finally, Section 6.4 compares the predictive performance of the two models.

6.1 Data description

Measurements of OSC rely heavily on satellite images. The data acquisition process requires sufficient satellite coverage, but is still prone to cloud cover over the region of interest. These factors limit the ability to systematically do repeated measurements. To mitigate these difficulties, we demonstrate in this chapter that the models can also be built on data acquired from numerical ocean models. These models can offer repeated estimates of surface currents at different scales and resolutions, without bearing the risk of cloud cover hampering the data collection.

The OSC data considered originate from SINMOD[1], a numerical ocean model system, which has been under development at SINTEF for over 25 years. The model can deliver a broad range of applications and services, but our data set consists of estimates of surface currents. SINMOD models the entire water column, and due to the vertical layer setup, estimates of the currents are only available $3\,\mathrm{m}$ below the ocean surface. These are mapped to an equal-distance grid, with a spatial horizontal resolution of $800\,\mathrm{m}$. At these small scales, the variability in currents is low. Hence, to cover a large enough area to capture any meaningful variance and still maintain a small grid size, we thin the data. We use only every third data point, so that the horizontal resolution decreases to $2.4\,\mathrm{km}$. We again use a grid size of 24×24, implying that the area covers $57.6\,\mathrm{km}$ in both spatial dimensions.

[1] More information about SINMOD is available on the web page: `https://www.sintef.no/en/ocean/initiatives/sinmod/`

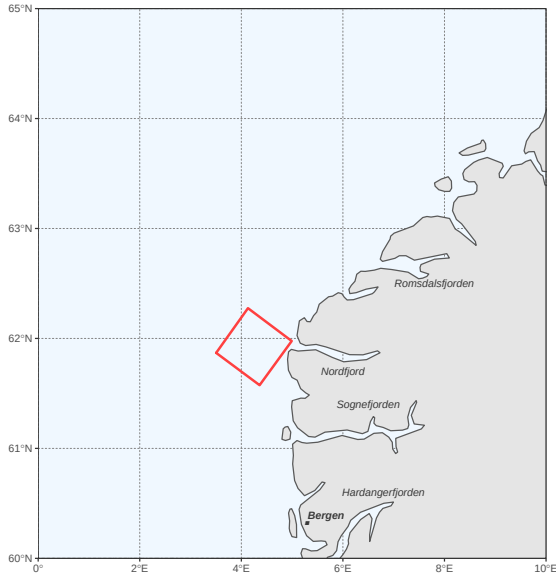
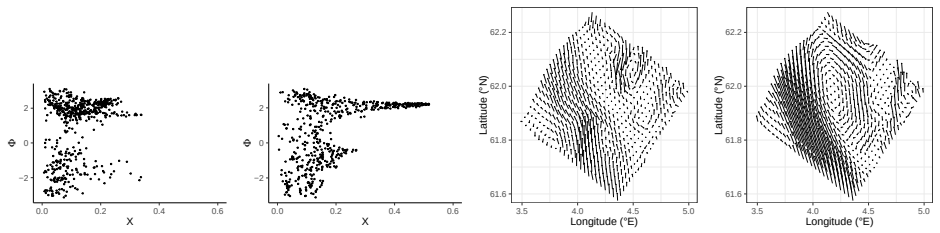


Figure 6.1: Map of the west coast of Norway and the Norwegian Sea, with the area in consideration indicated by the red rectangle.

The area lies just outside the mouth of Nordfjord and is depicted in Figure 6.1. We extract data from two dates in 2017, one in spring and one in fall. The data are displayed in Figure 6.2. The scatter plots on the left display the surface current speeds on the first axis, measured in $\mathrm{m/s}$. The second axis represents the angle of the surface current direction, with $\Phi = 0$ for currents flowing east. We observe that the maximal current speed is approximately $0.5\,\mathrm{m/s}$, which is the same as for the data set in Chapter 5. The surface currents are plotted as vectors at their respective spatial location on the right. From this display, it is evident that the currents are spatially more homogeneous than the previous data set. This comes as a result of the increased spatial resolution. When grid points are closer, the observations naturally become more dependent, leading to increased spatial homogeneity. We also observe that the maximum velocities are larger in spring than during fall. Moreover, the spring measurements seem to be more spatially homogeneous. Similar findings have been proposed by Mork and Skagseth [2010], who studied the Norwegian Atlantic Current at the Svinøy section, just north of our area.

The research related to surface currents in the vicinity of fjords is limited. Surface currents in the Porsanger fjord was studied by Stramska et al. [2016], while Fraser et al. [2018] studied the circulation in Isfjorden at the west coast of Spitsbergen. A more general overview of externally forced processes and current circulation modes in fjords was provided by Stigebrandt [2012]. Yet, circulation patterns at and just outside of the fjord mouth is relatively unexplored. Our study intends to segment the estimated currents into a small number of local regimes, or states of the ocean, which are easy to interpret, and in this way increase the understanding of the current patterns in this area. To estimate the local regimes, we use the same methodology as in Chapter 5, where we saw that not



(a) Fall **(b)** Spring **(c)** Fall **(d)** Spring

Figure 6.2: Display of the cylindrical observations (left) and the observations mapped to vectors at their grid points to create vector fields (bottom) for the fall and spring data.

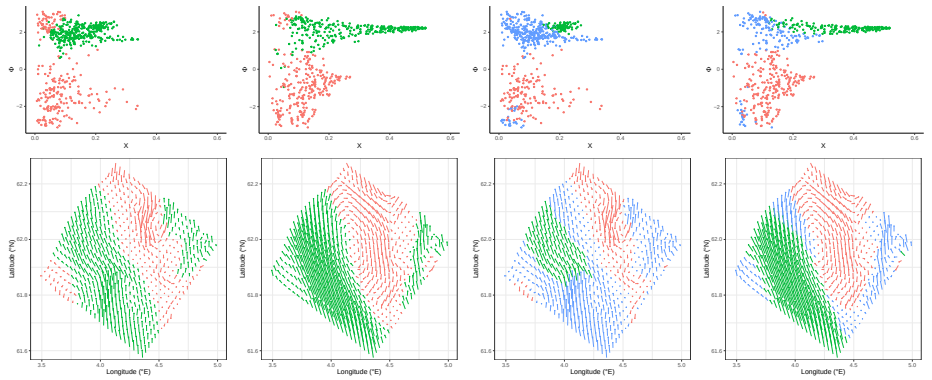
much is gained in terms of explanatory power by estimating separate models for different seasons. Hence, we estimate one model for both the spring and fall data. The variability in the currents is then captured by the allocation of latent classes.

6.2 WSSVM

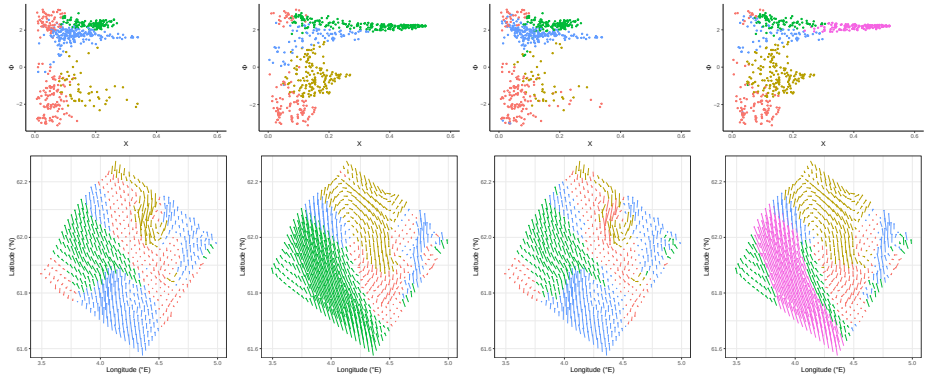
Model estimation: We observe from Figure 6.2a that for large angles, the data seem to be positively skewed. This observable skewed pattern motivates the implementation of a model with the skewed WSSVM density. Parameters are again estimated using the hybrid algorithm, with 50 sets of initial values drawn uniformly within the usual boundaries listed in Equation (5.3). As always we estimate models with the number of latent classes ranging from 2 to 5, and the resulting latent class predictions are displayed in Figure 6.3. We observe that large areas are predicted to take the same latent class. This is because the currents are spatially homogeneous, and the spatial dependence parameter ρ is therefore large. For $K = 5$, none of the fall observations are predicted to take the pink latent class. This means that the pink latent class represents a current pattern that is present only during spring, indicating that the model is over-fitted.

Model selection: To make an informed decision about the number of latent classes, we display the block log-likelihood, C-BIC and C-AIC values for all models in Table 6.1. Like with the previous data set, the most complex model yields the lowest C-AIC. However, as our focus is on constructing parsimonious models, rather than maximizing predictive power, we choose to include $K = 3$ latent classes, as this minimizes the C-BIC.

Model properties: Contour plots of the three densities representing the chosen model are displayed in Figure 6.4. The observations from the fall data set are plotted as dots and the observations from spring as pluses. The probability of each observation belonging to a given latent class is represented by the transparency of the observation. Further, we have the maximum-block-likelihood parameter estimates in Table 6.2 and the latent class predictions in Figure 6.3. Observe firstly that there are very few outliers in all the three densities. This suggests that the model fits the data well, and that the three densities together are able to represent all plausible current patterns. Further, we observe that



(a) $K = 2$ **(b)** $K = 3$

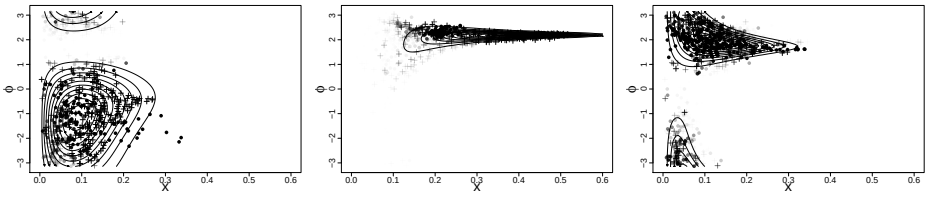


(c) $K = 4$ **(d)** $K = 5$

Figure 6.3: Resulting prediction of the latent classes with K ranging from 2 (top left) to 5 (bottom right) with the WSSVM density. A maximum probability prediction criterion is used to predict the classes. The latent classes are indicated by the colour of the points/arrows.

Table 6.1: Computed values for the block log-likelihood (bl) defined in Equation (3.34), C-BIC from Equation (3.45) and C-AIC from Equation (3.44) for the SINMOD data with the WSSVM density. The best model is indicated by **bold**.

K	bl	C-BIC	C-AIC
2	-620.0	2011.2	1458.7
3	154.8	**833.5**	14.8
4	201.8	873.4	-41.3
5	604.5	885.9	**-667.8**



(a) Density 1 **(b)** Density 2 **(c)** Density 3

Figure 6.4: Estimated densities of the best model with the WSSVM density. The observations from fall and spring are plotted as dots and pluses, respectively. The transparency of each dot/plus represents the probability of belonging to that class. No transparency represent a probability of 1 of belonging to that class.

the densities generally have higher circular concentration than the previous data set. The circular concentration is particularly large for the second and third density.

The density corresponding to the first latent class mostly comprises currents on the southern semi-circle. Compared to the other two densities, the circular concentration is very low ($\kappa_1 = 0.28$), making the density spread out in the circular domain. The modal direction is east ($\mu_1 = 0.13$), but the low circular concentration combined with maximal negative skewness ($\lambda_1 = -1.00$) makes it so that the major part of the density represents southward flowing currents. This density is predicted in a large patch, stretching from the north corner to the middle of the area, for the spring measurements. The fall measurements contain a similar, but thinner patch of this density.

The second density is highly concentrated in the circular part ($\kappa_2 = 3.03$) and represents currents flowing north-northwest with high speeds. The circular location is to the north-west ($\mu_2 = 2.20$), but the negative skewness ($\lambda_2 = -1.00$) makes the density comprise currents pointing north-northwest. This density dominates the western part of the spring measurements, and is less prevalent in the fall. As previously mentioned, Mork and Skagseth [2010] found that the maximal current speeds are larger in spring than fall. Hence, it is logical for this density to be dominating the spring measurements, and occur less frequently in fall.

The third density constitutes currents flowing north and west. The speeds are lower than for the second density, and the circular concentration is also lower ($\kappa_3 = 1.61$), making the density include a greater spread of directions. With a modal direction to the north ($\mu_3 = 1.71$) and a strong positive skewness ($\lambda_3 = 1.00$), the density is shaped so that it includes currents flowing north with higher speeds, and currents pointing west with lower speeds. The density dominates large parts of the fall measurements, and is present in smaller areas in the spring.

Model uncertainty: As we did for the previous data set, we also conduct parametric bootstrap to estimate the uncertainty in the parameter estimates. The resulting parameter quantiles are displayed in Table 6.2. Firstly, we observe that the spatial interaction parameter ρ is larger than for the previous data set. As we previously discussed, this data set has a higher spatial resolution, and the measurements are consequently more spatially homogeneous, leading to a larger value for the dependence parameter ρ. Moreover, we

Table 6.2: Parameter estimates and bootstrap quantiles of the best model with the WSSVM density.

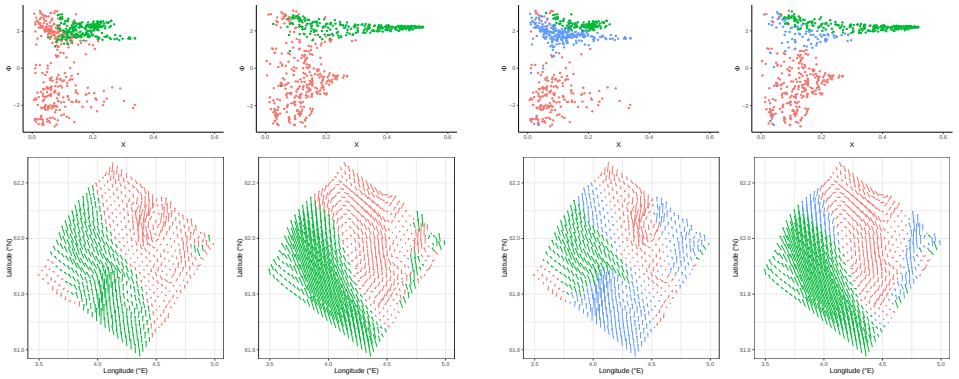
Parameter	2.5% quantile	Estimate	97.5% quantile
α_1	1.76	1.98	2.51
β_1	6.20	8.05	8.96
μ_1	-0.07	0.13	0.31
κ_1	0.12	0.28	0.47
λ_1	-1.00	-1.00	-1.00
α_2	2.60	3.84	4.65
β_2	7.32	9.31	12.94
μ_2	2.16	2.20	2.22
κ_2	2.21	3.03	3.47
λ_2	-1.00	-1.00	-1.00
α_3	2.09	2.27	2.47
β_3	14.27	17.11	20.72
μ_3	1.36	1.71	2.80
κ_3	0.84	1.61	2.11
λ_3	1.00	1.00	1.00
ρ	2.19	3.04	3.40

observe that all circular-linear dependence parameters κ are statistically significant. This supports the decision to employ a cylindrical model. Also, all skewness parameters λ are significant, justifying the use of a skewed cylindrical distribution. Interestingly, all three distributions attain maximal skewness, with all three skewness parameters, and their respective quantiles, being at the parameter boundary with two decimals precision.

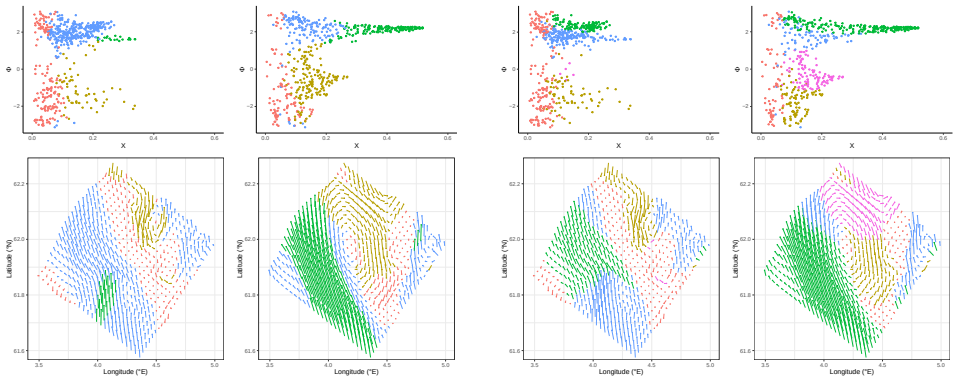
6.3 GPTWC

Model estimation: Although the WSSVM model displayed few outliers, we also fit a model with the GPTWC density to check if any of the heavy tail parameters become significant. We employ the hybrid algorithm and draw 50 sets of initial parameters with the usual boundaries from Equation (5.4). Models are estimated with the number of latent classes K ranging from 2 to 5. The latent class predictions for each model are displayed in Figure 6.5. Observe that these are very similar to the latent classes predicted with the WSSVM density, especially for small values of K, but they do not display a skewed structure. The largest discrepancies are found in the fifth (pink) class with $K = 5$.

Model selection: Values for the block log-likelihood, C-BIC and C-AIC for each model are displayed in Table 6.3. As always, the C-AIC is lowest for the model with the most latent classes, $K = 5$. We also observe that the C-BIC suggests the same number of latent classes, $K = 3$, as with the WSSVM density. This confirms that using $K = 3$ latent classes provides the best fit to the SINMOD data, while simultaneously keeping the model relatively simple.



(a) $K = 2$ **(b)** $K = 3$

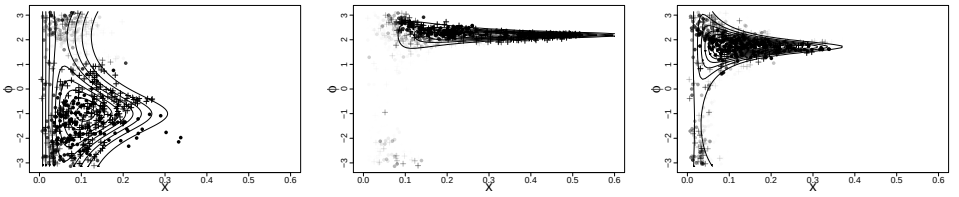


(c) $K = 4$ **(d)** $K = 5$

Figure 6.5: Resulting prediction of the latent classes with K ranging from 2 (top left) to 5 (bottom right) with the GPTWC density. A maximum probability prediction criterion is used to predict the classes. The latent classes are indicated by the colour of the points/arrows.

Table 6.3: Computed values for the block log-likelihood (bl) defined in Equation (3.34), C-BIC from Equation (3.45) and C-AIC from Equation (3.44) for the SINMOD data with the GPTWC density. The best model is indicated by **bold**.

K	bl	C-BIC	C-AIC
2	-937.8	2441.8	2036.3
3	7.3	**1058.6**	289.9
4	271.8	1195.4	-50.3
5	351.9	1199.6	**-163.7**



(a) Density 1 **(b)** Density 2 **(c)** Density 3

Figure 6.6: Estimated densities of the best model with the GPTWC density. The observations from fall and spring are plotted as dots and pluses, respectively. The transparency of each dot/plus represents the probability of belonging to that class. No transparency represent a probability of 1 of belonging to that class.

Model properties: Figure 6.6 displays contour plots of the three densities from the chosen model to study the current patterns that the model represents. Measurements from the fall data are displayed as dots and measurements from spring as pluses. The transparency of the measurements indicates the probability of belonging to that class. These contour plots will be interpreted together with the maximum-block-likelihood parameter estimates and their quantiles from Table 6.4 and the latent class predictions in Figure 6.5. Take note that these densities are very similar to the WSSVM model, and the predicted latent classes are overall alike. Also, similarly to the WSSVM model, there are very few outliers to the main distributional areas. Even though this density is designed to handle outliers in the linear part by imposing heavy tails, none of the distributions display heavy tailedness.

The first density has a circular location to the south–southeast ($\mu_1 = -1.00$), but the circular concentration is low ($\kappa_1 = 0.64$), making the density account for a wide range of directions on the southern semi-circle. The density display virtually no heavy tailedness ($\tau_1 = 0.02$). However, judging from the plot, the measurements seem to display skewness in that a larger proportion of the measurements are located at lower angles than the circular location compared to higher angles. That is, the measurements are not symmetric about the modal direction, but rather skewed towards negative angles. Predictions of this class are located similarly to the first density of the WSSVM model. There is a large patch stretching from the north corner to the middle of the area for the spring measurements, and a similar, but thinner patch for the fall measurements.

The second and third densities are very similar to the second and third densities of the WSSVM model. They have the exact same modal directions ($\mu_2 = 2.20$, $\mu_3 = 1.71$), display high circular concentration ($\kappa_2 = 0.99$, $\kappa_3 = 0.97$), and the latent class predictions are fairly equal. However, they are not able to handle the skewness observed in the data, similarly to the first density. This, combined with the fact that neither of the densities are significantly heavy tailed ($\tau_2 = 0.00$, $\tau_3 = 0.02$), leads us to the conclusion that the WSSVM density is better suited for this data set.

Table 6.4: Parameter estimates and bootstrap quantiles of the best model with the GPTWC density.

Parameter	2.5% quantile	Estimate	97.5% quantile
α_1	0.44	0.49	0.54
β_1	0.08	0.09	0.10
μ_1	-1.10	-1.00	-0.90
τ_1	0.00	0.02	0.09
κ_1	0.56	0.64	0.70
α_2	0.33	0.35	0.37
β_2	0.06	0.07	0.08
μ_2	2.18	2.20	2.21
τ_2	0.00	0.00	0.00
κ_2	0.99	0.99	0.99
α_3	0.39	0.44	0.47
β_3	0.04	0.04	0.05
μ_3	1.68	1.71	1.74
τ_3	0.00	0.02	0.09
κ_3	0.96	0.97	0.98
ρ	3.02	3.14	3.22

6.4 Model comparison

We now compare the predictive performance of the WSSVM and GPTWC models. We already concluded that WSSVM is better suited to model this data set, because the skewness parameters are significant, whereas the heavy tail parameters of the GPTWC model are not significant. From Section 2.2 it is apparent that the WSSVM distribution with no skewness and the GPTWC distribution without heavy tails are the same distribution. Hence, the WSSVM distribution with significant skewness can be seen as a generalization of the GPTWC distribution without heavy tails. Consequently, we want to verify that the WSSVM model has better predictive performance than the GPTWC model.

To do the model comparison, we carry out the same procedure as in Section 5.4. First, we draw 50 grid points for the spring and fall data sets. For all selected grid points we compare prediction of both models by evaluating the linear and circular CRPS. These are evaluated by first computing the predictive distributions as in Equation (3.17). The predictive distributions are then used together with the actual observations to compute linear and circular CRPS separately. Table 6.5 displays the average linear and circular CRPS scores for both data sets and both models.

From the table it is evident that the WSSVM model outperforms the GPTWC model both on the circular and the linear CRPS. The skewness in the model only affects the circular part, but apparently the prediction of the linear part is also better with the skewed model. These results verify that it is appropriate to employ the WSSVM distribution to model this data set.

By comparing these CRPS values with the CRPS values from Section 5.4, we ob-

Table 6.5: Average linear and circular CRPS for the SINMOD data. The model with the lowest CRPS is indicated with **bold** for each data set.

	Linear		Circular	
Model	Fall	Spring	Fall	Spring
WSSVM	**0.0132**	**0.0214**	**0.1045**	**0.0970**
GPTWC	0.0149	0.0221	0.1085	0.1034

serve that the linear CRPS are higher for this data set, and the circular CRPS are lower. This means that the model is better at predicting the angle for the SINMOD data set, but worse at predicting the speed. One possible explanation is the large values for the circular concentration parameter κ. Increased concentration in the circular part also increases the circular sharpness, which lowers the CRPS values. From the expression of the linear moments of the WSSVM distribution in Equation (2.18) it can be confirmed that the linear variance increases as κ increases. When the linear variance is increased, the sharpness is reduced and this leads to higher values for the linear CRPS.

Chapter 7

Conclusion

This chapter summarizes by giving a short review of the work. We also discuss and make concluding remarks on the main contributions of the thesis. Finally, we propose possible extensions and future research.

In this thesis, we have studied a cylindrical HMRF model for spatial cylindrical data. The model hierarchically combines a cylindrical density with a latent Potts model. It serves as a parsimonious representation in that it breaks down a global pattern into a discrete number of local regimes. These regimes, represented by a set of parameters supplied to a cylindrical density, are easy to interpret with traditional properties such as location, concentration, scale, shape, skewness, and heavy tails. The regimes are assigned to all observation sites, and in doing so, the model accounts for both spatial dependence and the cylindrical observation at each site.

The structure of the model is motivated by the call for compact models during autonomous missions at sea. For these data collection procedures, the main interest is typically *in-situ* observation of a range of ocean parameters, such as sea surface temperature, salinity and ocean color, as opposed to remote sensing which covers larger areas. If the vessel also observes the surface currents, the model can make probabilistic forecasts of surrounding circulation patterns on-board based on model parameters estimated on shore. To be computationally feasible in real time, the model requires relatively small grid sizes. Another application for the model is to make predictions of surface currents based on sparse *in-situ* observations. This is crucial to fill in gaps of missing data, caused by, e.g., cloud cover, in remote sensing observations.

Assuming a Potts model for the latent process imposes two main restrictions on the model. First, the Potts model and subsequently the composite and block-likelihood approaches assume a nearest neighbourhood structure on the spatial lattice. Disregarding more complex dependencies may be a too strict assumption in some cases. The second limitation is the assumption of a stationary latent field that depends on a single spatial coupling parameter, which implies that the spatial dependence is constant across space and equal for both spatial directions. In Chapter 5, we had to thin the data in order to accommodate the assumption of equal dependence in both directions. In our case, we were able to achieve reasonable results with these two restrictions. However, other studies might

require the implementation of a non-stationary Potts model, in which the spatial interaction parameters vary by direction or across space according to some external factors.

We have shown that the cylindrical HMRF model is capable of offering an interpretable description of surface current patterns by decomposing them into local regimes, both for direct measurements of OSC, but also for estimations provided by an ocean model. For the direct OSC measurements, we estimated separate models for observations from different seasons and a joint model for both seasons. Comparisons by scoring rules suggest that the gain in estimating separate models is limited, and to reduce model complexity, a single model for both seasons should be used. Further, this data set displayed both skewness in the circular part and heavy-tails in the linear part. The ocean model data set gave significant skewness parameters, but the heavy-tail parameters were non-significant. By comparing the two densities through scoring rules, we found that the skewed density provided more accurate predictions both in the circular and linear part.

Although the model is motivated by marine applications, specifically to study speed and direction of OSC, it is readily applicable to various forms of spatial cylindrical data. An obvious example is to study wave height and direction [Wang et al., 2015]. Other possibilities include environmental studies of wind fields [Modlin et al., 2012] or ecological studies of animal movement [Hanks et al., 2015]. These are both examples in which direction and speed are measured across space.

An important contribution of this thesis is the development of the hybrid algorithm for estimating model parameters. Through a simulation study we showed that the EM algorithm as an optimization procedure for a composite-likelihood approach has a large area of convergence. However, this comes at the cost of a slow convergence rate. By deploying a block-likelihood method, with the likelihood in each block computed exactly through a spatial extension of the forward–backward algorithm for hidden Markov models, we were able to reduce the run time of the EM algorithm by 50%. The block-likelihood method also reduced the run time of a direct optimization of the composite-likelihood by 25%. Additionally, the block-likelihood is more statistically robust, as it accounts for more dependencies in the latent structure. The hybrid algorithm exploits the large area of convergence of the EM algorithm and the fast convergence of the block-likelihood.

One obvious unexplored path for future work is to combine the WSSVM and GPTWC distribution to form a density that is both skewed in the circular part and heavy-tailed in the linear part. This is simply done by sine-skewing the wrapped Cauchy distribution for the circular part in the GPTWC distribution and in the process adding a skewness parameter λ. This creates an even more flexible distribution that allows for skewness and heavy-tails simultaneously. However, theoretical results about this new density are not yet available and would have to be derived. Also, the issue concerning parameter estimation would have to be tested. A simpler approach towards combining the skewness and heavy-tails is to estimate a model where some latent classes correspond to the WSSVM distribution and some latent classes correspond to the GPTWC distribution.

A more ambitious extension of the model is to add a temporal element in the latent process. One possibility in this regard is to allow the latent process to evolve in time according to a Markov model, dependent on a temporal interaction parameter. What makes this interesting is that it allows for probabilistic forecasts in time and enables studies of surface currents as a space-time evolving process. However, adding a temporal dimension to the latent process severely increases the computational demand for parameter estimation.

Bibliography

T. Abe and C. Ley. A tractable, parsimonious and flexible model for cylindrical data, with applications. *Econometrics and Statistics*, 4:91–104, 2017. doi:10.1016/j.ecosta.2016.04.001.

T. Abe and A. Pewsey. Sine-skewed circular distributions. *Statistical Papers*, 52:683–707, 08 2011. doi:10.1007/s00362-009-0277-x.

C. Abraham, N. Molinari, and R. Servien. Unsupervised clustering of multivariate circular data. *Statistics in Medicine*, 32(8):1376–1382, 2013. doi:10.1002/sim.5589.

P. Abrahamsen. A review of Gaussian random fields and correlation functions, 04 1997. Unpublished.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Proc. Second International Symposium on Information Theory*, pages 267–281. Akademiai Kiado, Budapest, 1973.

H. Austad and H. Tjelmeland. Approximate computations for binary Markov random fields and their use in Bayesian models. *Statistics and Computing*, 27, 08 2016. doi:10.1007/s11222-016-9685-7.

G. Barkema and J. de Boer. Numerical study of phase transitions in Potts models. *Physical Review A*, 44:8000–8005, Dec 1991. doi:10.1103/PhysRevA.44.8000.

M. S. Bartlett. Approximate confidence intervals. *Biometrika*, 40(1/2):12–19, 1953.

F. Bartolucci and J. Besag. A recursive algorithm for Markov random fields. *Biometrika*, 89(3):724–730, August 2002. URL `http://www.jstor.org/stable/4140615`.

E. Batschelet. *Circular Statistics in Biology*. Mathematics in biology. Academic Press, 1981. URL `https://books.google.no/books?id=UxNXvQEACAAJ`.

L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 02 1970. doi:10.1214/aoms/1177697196.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974. URL http://www.jstor.org/stable/2984812.

J. Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 24(3):179–195, 1975. URL http://www.jstor.org/stable/2987782.

J. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *Technical Report ICSI-TR-97-021, University of Berkeley*, 4, 06 2000.

M. A. Bourassa, S. T. Gille, D. L. Jackson, J. B. Roberts, and G. A. Wick. Ocean winds and turbulent air-sea fluxes inferred from remote sensing. *Oceanography*, 23, December 2010. doi:10.5670/oceanog.2010.04.

J. Bulla, F. Lagona, A. Maruotti, and M. Picone. 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*, 17:544–567, 12 2012. doi:10.1007/s13253-012-0110-1.

B. Chapron, J. A. Johannessen, and C. Donlon. Technical note (TN-1). Deliverable D-140 of the GlobCurrent project. https://globcurrent.nersc.no/system/files/pubdeliver/GlobCurrent_D-140_TN-1-Year-1_v0.5-signed.pdf, 2015. Online; accessed 24 March 2020.

D. R. Cox and N. Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737, 09 2004. doi:10.1093/biomet/91.3.729.

N. A. C. Cressie. *Statistics for Spatial Data*. Wiley Series in Probability and Statistics. John Wiley and Sons, 1993.

K. Dohan and N. Maximenko. Monitoring ocean currents with satellite sensors. *Oceanography*, 23, 12 2010. doi:10.5670/oceanog.2010.08.

J. Eidsvik, B. A. Shaby, B. J. Reich, M. Wheeler, and J. Niemi. Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics*, 23(2):295–315, 2014. doi:10.1080/10618600.2012.760460.

V. Ekman. On the influence of the earth's rotation on ocean-currents. *Archives of Mathematics, Astronomy, and Physics*, 2(11), 1905.

N. I. Fisher. *Statistical Analysis of Circular Data*. Cambridge University Press, 1993. doi:10.1017/CBO9780511564345.

T. O. Fossum, J. Ryan, T. Mukerji, J. Eidsvik, T. Maughan, M. Ludvigsen, and K. Rajan. Compact models for adaptive sampling in marine robotics. *The International Journal of Robotics Research*, 2019. doi:10.1177/0278364919884141.

N. J. Fraser, R. Skogseth, F. Nilsen, and M. E. Inall. Circulation and exchange in a broad arctic fjord using glider-based observations. *Polar Research*, 37(1), 2018. doi:10.1080/17518369.2018.1485417.

N. Friel and H. Rue. Recursive computing and simulation-free inference for general factorizable models. *Biometrika*, 94(3):661–672, 08 2007. doi:10.1093/biomet/asm052.

X. Gao and P. X.-K. Song. Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105 (492):1531–1540, 2010. doi:10.1198/jasa.2010.tm09414.

A. Gelman and X.-L. Meng. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 05 1998. doi:10.1214/ss/1028905934.

C. Geyer and E. Thompson. Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of The American Statistical Association*, 90:909–920, 09 1995. doi:10.1080/01621459.1995.10476590.

G. H. Givens and J. A. Hoeting. *Computational Statistics*, chapter 9, pages 287–321. John Wiley & Sons, Ltd, 2 edition, 2013. doi:10.1002/9781118555552.ch9.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi:10.1198/016214506000001437.

T. Gneiting, L. Stanberry, E. Grimit, L. Held, and N. Johnson. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, 17:211–235, 02 2008. doi:10.1007/s11749-008-0114-x.

V. P. Godambe. An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 31(4):1208–1211, 12 1960. doi:10.1214/aoms/1177705693.

E. P. Grimit, T. Gneiting, V. J. Berrocal, and N. A. Johnson. 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, 2006. doi:10.1256/qj.05.235.

J. Gurland. On regularity conditions for maximum likelihood estimators. *Scandinavian Actuarial Journal*, 1954(1):71–76, 1954. doi:10.1080/03461238.1954.10414197.

X. Guyon. *Random fields on a network*. Probability Theory and Stochastic Modelling. Springer, 1995.

J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices, 1971. Unpublished.

E. M. Hanks, M. B. Hooten, and M. W. Alldredge. Continuous-time discrete-space models for animal movement. *The Annals of Applied Statistics*, 9(1):145–165, 03 2015. doi:10.1214/14-AOAS803.

B. Helland-Hansen and F. Nansen. The Norwegian Sea - its physical oceanography based upon the Norwegian researches 1900–1904. *Report on Norwegian Fishery and Marine Investigations*, 2(2), 1909.

H. H. Holm. *Efficient Forecasting of Drift Trajectories using Simplified Ocean Models and Nonlinear Data Assimilation on GPUs*. PhD thesis, NTNU, 4 2020. URL `https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2652592`.

H. Holzmann, A. Munk, M. Suster, and W. Zucchini. Hidden Markov models for circular and linear-circular time series. *Environmental and Ecological Statistics*, 13(3):325–347, Sep 2006. doi:10.1007/s10651-006-0015-7.

R. Ingvaldsen. *The Atlantic inflow to the Barents Sea*. PhD thesis, University of Bergen, 2 2003. URL `http://web.gfi.uib.no/publikasjoner/rmo/RMO-2003-1.pdf`.

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

R. B. Ingvaldsen, L. Asplin, and H. Loeng. Velocity field of the western entrance to the Barents Sea. *Journal of Geophysical Research: Oceans*, 109(C3), 2004. doi:10.1029/2003JC001811.

J. A. Johannessen, R. P. Raj, J. E. Ø. Nilsen, T. Pripp, P. Knudsen, F. Counillon, D. Stammer, L. Bertino, O. B. Andersen, N. Serra, and N. Koldunov. Toward improved estimation of the dynamic topography and ocean circulation in the high latitude and arctic ocean: The importance of GOCE. *Surveys in Geophysics*, 35:661–679, 2014. doi:10.1007/978-94-017-8789-5_9.

R. A. Johnson and T. E. Wehrly. Some angular-linear distributions and related regression models. *Journal of the American Statistical Association*, 73(363):602–606, 1978.

G. Jona-Lasinio, A. Gelfand, and M. Jona-Lasinio. Spatial analysis of wave direction data using wrapped Gaussian processes. *The Annals of Applied Statistics*, 6(4):1478–1498, Dec 2012. doi:10.1214/12-aoas576.

S. Kato and K. Shimizu. Dependent models for observations which include angular ones. *Journal of Statistical Planning and Inference*, 138(11):3538–3549, 2008. doi:10.1016/j.jspi.2006.12.009.

R. Kwok, G. Spreen, and S. Pang. Arctic sea ice circulation and drift speed: Decadal trends and ocean currents. *Journal of Geophysical Research: Oceans*, 118(5):2408–2425, 2013. doi:10.1002/jgrc.20191.

F. Lagona and M. Picone. Model-based segmentation of spatial cylindrical data. *Journal of Statistical Computation and Simulation*, 86(13):2598–2610, 2016. doi:10.1080/00949655.2015.1122791.

A. Lee. Circular data. *WIREs Computational Statistics*, 2(4):477–486, 2010. doi:10.1002/wics.98.

B. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80:221–239, 01 1988. doi:10.1090/conm/080/999014.

B. Lindsay, G. Yi, and J. Sun. Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21:71–105, 01 2011.

Y. Liu and J. Dilger. Application of the one- and two-dimensional ising models to studies of cooperativity between ion channels. *Biophysical journal*, 64:26–35, 02 1993. doi:10.1016/S0006-3495(93)81337-7.

I. L. MacDonald and W. Zucchini. *Hidden Markov and other models for discrete-valued time series*, volume 110. CRC Press, 1997.

K. Mardia. *Statistics of Directional Data*. Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Academic Press, 1972. doi:10.1016/B978-0-12-471150-1.50018-2.

K. V. Mardia and P. E. Jupp. *Directional Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, 1999. doi:10.1002/9780470316979.

K. V. Mardia and T. W. Sutton. A model for cylindrical variables with applications. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(2):229–233, 1978.

R. B. Millar. *Maximum Likelihood Estimation and Inference*. John Wiley & Sons, Ltd, 2011. doi:10.1002/9780470094846.scard.

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

K. A. Mork and J. Blindheim. Variations in the Atlantic inflow to the Nordic Seas, 1955–1996. *Deep Sea Research Part I: Oceanographic Research Papers*, 47(6):1035–1057, 2000. doi:10.1016/S0967-0637(99)00091-6.

K. A. Mork and Ø. Skagseth. A quantitative description of the Norwegian Atlantic Current by combining altimetry and hydrography. *Ocean Science*, 6(4):901–911, 2010. doi:10.5194/os-6-901-2010.

W. J. Morphet. *Simulation, kriging, and visualization of circular-spatial data*. PhD thesis, Utah State University, 5 2009.

M. Niss. History of the Lenz-Ising model 1920–1950: From ferromagnetic to cooperative phenomena. *Archive for History of Exact Sciences*, 59:267–318, 2005. doi:10.1007/s00407-004-0088-3.

A. Pewsey and E. García-Portugués. Recent advances in directional statistics, 2020.

R. B. Potts. *The mathematical investigation of some cooperative phenomena*. PhD thesis, University of Oxford, 1951.

P.-M. Poulain, A. Warn-Varnas, and P. P. Niiler. Near-surface circulation of the Nordic seas as measured by Lagrangian drifters. *Journal of Geophysical Research: Oceans*, 101(C8):18237–18258, 1996. doi:10.1029/96JC00506.

R Core Team. R: A language and environment for statistical computing. `https://www.r-project.org/`, 2020. Online; accessed 29 May 2020.

M. Ranalli, F. Lagona, M. Picone, and E. Zambianchi. Segmentation of sea current fields by cylindrical hidden Markov models: a composite likelihood approach. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 67(3):575–598, 2018. doi:10.1111/rssc.12240.

P. Ravindran and S. Ghosh. Bayesian analysis of circular data using wrapped distributions. *Journal of Statistical Theory and Practice*, 5, 12 2011. doi:10.1080/15598608.2011.10483731.

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

B. J. Reich and M. Fuentes. A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *The Annals of Applied Statistics*, 1(1): 249–264, 06 2007. doi:10.1214/07-AOAS108.

T. Rydén and D. M. Titterington. Computational Bayesian analysis of hidden Markov models. *Journal of Computational and Graphical Statistics*, 7(2):194–211, 1998.

E. Schneidman, M. J. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087): 1007–1012, Apr 2006. doi:10.1038/nature04701.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 03 1978. doi:10.1214/aos/1176344136.

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

D. Slagstad and T. A. McClimans. Modeling the ecosystem dynamics of the Barents Sea including the marginal ice zone: I. physical and chemical oceanography. *Journal of Marine Systems*, 58(1):1–18, 2005. doi:10.1016/j.jmarsys.2005.05.005.

A. Stigebrandt. Hydrodynamics and circulation of fjords. In R. L. Bengtsson, Herschy and R. Fairbridge, editors, *Encyclopedia of Lakes and Reservoirs*, pages 327–344. Springer, 01 2012. doi:10.1007/978-1-4020-4410-6_247.

M. Stramska, A. Jankowski, and A. Cieszyńska. Surface currents in the Porsanger fjord in northern Norway. *Polish Polar Research*, 37, 04 2016. doi:10.1515/popore-2016-0018.

R. H. Swendsen and J.-S. Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58:86–88, Jan 1987. doi:10.1103/PhysRevLett.58.86.

P. Tandeo, S. Ba, R. Fablet, B. Chapron, and E. Autret. Spatio-temporal segmentation and estimation of ocean surface currents from satellite sea surface temperature fields. In *2013 IEEE International Conference on Image Processing*, pages 2344–2348, Sep. 2013. doi:10.1109/ICIP.2013.6738483.

I. Tomoaki, K. Shimizu, and T. Abe. A cylindrical distribution with heavy-tailed linear part. *Japanese Journal of Statistics and Data Science*, 02 2019. doi:10.1007/s42081-019-00031-5.

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

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

F. Wang and A. E. Gelfand. Modeling space and space-time directional data using projected Gaussian processes. *Journal of the American Statistical Association*, 109(508): 1565–1580, 2014. doi:10.1080/01621459.2014.934454.

F. Wang, A. Gelfand, and G. Jona Lasinio. Joint spatio-temporal analysis of a linear and a directional variable: space-time modeling of wave heights and wave directions in the adriatic sea. *Statistica Sinica*, 25:25–39, 01 2015. doi:10.5705/ss.2013.204w.

C. K. Wikle, R. F. Milliff, R. Herbei, and W. B. Leeds. Modern statistical methods in oceanography: A hierarchical perspective. *Statistical Science*, pages 466–486, 2013.

F. Y. Wu. The Potts model. *Reviews of Modern Physics*, 54:235–268, Jan 1982. doi:10.1103/RevModPhys.54.235.

S. Østerhus, T. Gammelsrød, and R. Hogstad. Ocean weather ship station M (66°N, 2°E): The longest homogeneous time series from the deep ocean. *WOCE Newsletter*, 96(24): 31, 1996.

Henrik Syversveen Lie

Cylindrical hidden Markov random field models with applications to ocean surface currents

NTNU
Norwegian University of
Science and Technology